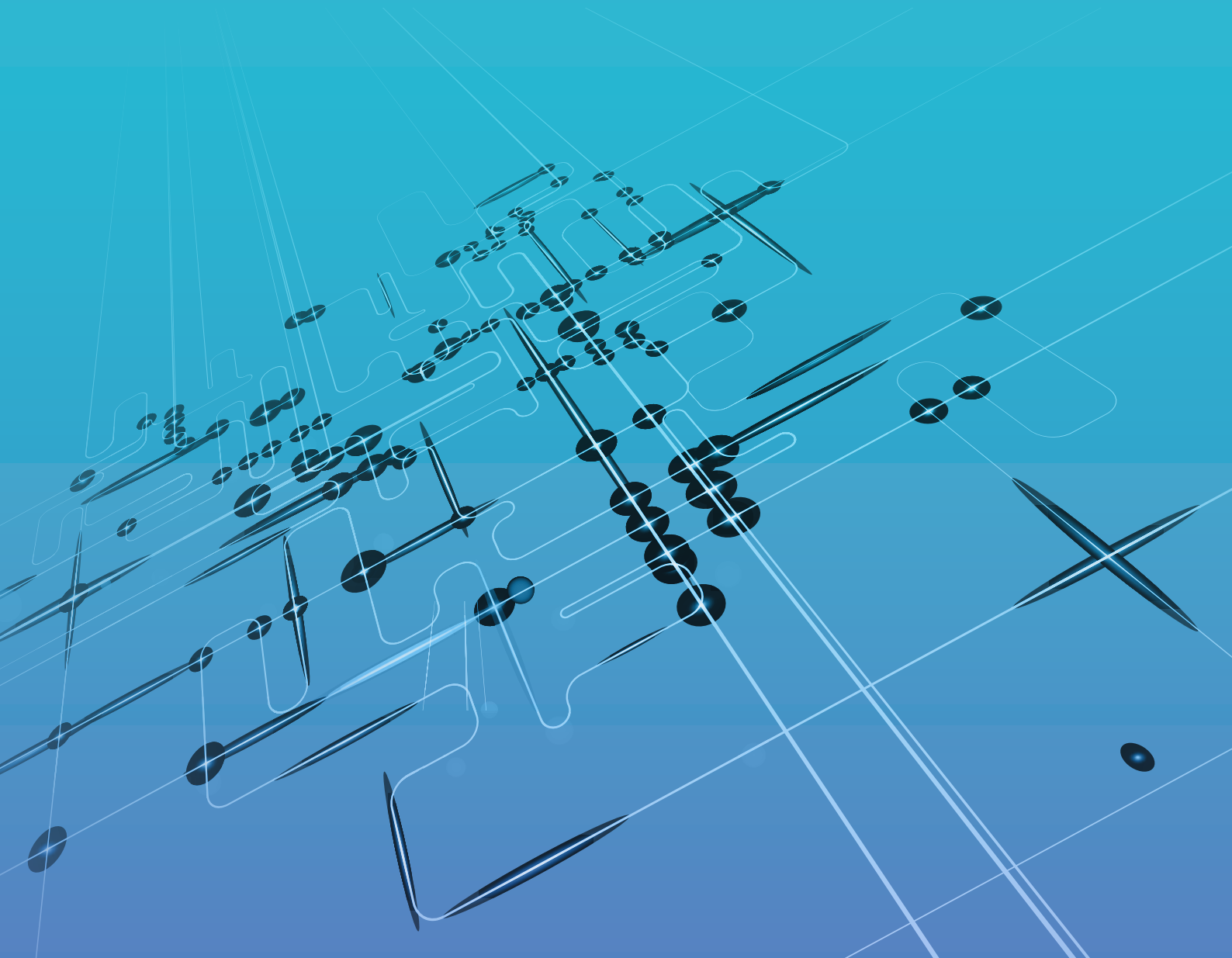
Kenneth E. Train

MÉTODOS DE ELECCIÓN DISCRETA CON SIMULACIÓN

SEGUNDA EDICIÓN (ESPAÑOL)



MÉTODOS DE ELECCIÓN DISCRETA CON SIMULACIÓN

Kenneth E. Train

Este libro describe la nueva generación de métodos de elección discreta, centrándose en los numerosos avances que han sido posibles gracias a la simulación. Investigadores de todo el mundo están usando estos métodos estadísticos para estudiar las elecciones que consumidores, hogares, empresas y otros agentes realizan. En este texto se tratan cada uno de los principales modelos existentes: logit, distribución generalizada del valor extremo (incluyendo logit jerárquico y logit jerárquico cruzado), probit y logit mixto, además de otras especificaciones desarrolladas a partir de estos modelos básicos. Se investigan y comparan los procedimientos de estimación basados en simulación, incluyendo el estimador de máxima verosimilitud simulada, el método de momentos simulados y el método de puntuaciones simuladas. También se describen procedimientos para extraer valores al azar de densidades de probabilidad, incluyendo técnicas de reducción de la varianza como el método de los opuestos y las extracciones de Halton. Del mismo modo, se exploran avances recientes en el terreno de los procedimientos Bayesianos, incluyendo el uso del algoritmo Metropolis-Hastings y su variante, el muestreo de Gibbs. En la segunda edición del presente libro se han añadido dos capítulos sobre endogeneidad y sobre algoritmos de maximización del valor esperado. Ningún otro libro incluye todos estos temas, que han ido surgiendo durante los últimos 25 años. Los procedimientos son aplicables en numerosos campos, incluyendo la energía, el transporte, estudios ambientales, salud, ocupación y marketing.

El profesor Kenneth E. Train imparte cursos sobre econometría, regulación y organización industrial en la Universidad de California, Berkeley. Asimismo ocupa la plaza de vicepresidente de la *National Economic Research Associates* (NERA), Inc., en San Francisco, California. Autor de *"Optimal Regulation: The Economic Theory of Natural Monopoly" (1991)* y *"Qualitative Choice Analysis" (1986)*, el Dr. Train ha escrito más de 60 artículos sobre teoría económica y regulación. Train presidió el *Center for Regulatory Policy* en la Universidad de California, Berkeley, desde 1993 hasta el 2000 y ha testificado como experto en procedimientos reguladores y casos judiciales. Ha recibido numerosos galardones por su actividad como docente e investigador.

Comentarios adicionales recibidos tras la publicación de la primera edición de *"Métodos de elección discreta con simulación":*

"El libro de Ken Train ofrece una cobertura excepcional a los elementos más avanzados de la estimación y el uso de modelos de elección discreta que requieren de simulación para tener en cuenta la aleatoriedad de la población objeto de estudio. Su escritura es clara y comprensible, proporcionando a los lectores, tanto noveles como experimentados, conocimientos y comprensión de todos los aspectos relativos a estos nuevos métodos, cada vez más importantes".

Frank S. Koppelman, Universidad Northwestern

"Se trata de un libro magistral, cuyo autor es uno de los principales contribuyentes al campo de los métodos y análisis de elección discreta. Ningún otro libro cubre este terreno con tal detalle hasta la fecha, tanto en el ámbito de la teoría como de la aplicación. Los capítulos sobre simulación y los recientes desarrollos como el método logit mixto son especialmente lúcidos. Como texto de referencia este trabajo debería tener vigencia durante mucho tiempo. Será de interés tanto para el profesional como para el investigador especializado que haya ejercido en este campo durante muchos años".

David Hensher, Universidad de Sidney

"La estimación basada en la simulación es un avance crucial en el campo de la econometría y de los modelos de elección discreta. Esta técnica ha revolucionado tanto el análisis clásico como el Bayesiano. Muchos de los trabajos de Ken Train han supuesto una gran contribución a la literatura en este ámbito. *"Métodos de elección discreta con simulación"* recopila los resultados obtenidos hasta la fecha de forma integral, dedicando capítulos a los fundamentos teóricos del comportamiento, a aspectos prácticos y teóricos de la estimación, así como a una gran variedad de aplicaciones. Este libro es, de principio a fin, una mezcla agradable de teoría, análisis y estudio de casos, así como una referencia completa para desarrolladores y profesionales".

William Greene, de la Universidad de Nueva York

El porqué de esta edición en castellano

Carlos Ochoa

Debo confesar que soy un intruso. He desarrollado mi carrera profesional en el sector de la investigación de mercado siendo ingeniero de telecomunicaciones. Y ahora, he dedicado los últimos meses de mi vida a traducir el libro el lector tiene frente a sí, sin ser traductor. De alguna manera, ambos hechos están relacionados.

Todo empezó en 2004. Aquel año me incorporé al proyecto de Netquest, con el objetivo de crear una empresa dedicada a la recolección de opiniones a través de Internet en los mercados de habla hispana y portuguesa. La idea era simple: trasladar a internet las encuestas que se llevaban a cabo de forma presencial o telefónica, con la ayuda de paneles online de personas dispuestas a compartir su opinión. La mayor parte de estas encuestas tenían un diseño clásico: un conjunto de preguntas acerca de temas variados acompañadas por una escala de respuesta para indicar preferencias. Pero de vez en cuando algún cliente se interesaba por un tipo de cuestionario diferente, los cuestionarios tipo *conjoint*. En este tipo de cuestionarios, en lugar de preguntar en qué medida el respondiente valora unos atributos descontextualizados, estos se agrupan formando productos sobre los que realmente se pide la opinión. En su modalidad más avanzada, los estudios *conjoint* enfrentan productos entre sí, haciendo que el respondiente escoja cuál de ellos prefiere. Son los *conjoint* basados en la elección (*choice based conjoint, CBC*).

La idea detrás de este tipo de estudios me sedujo de inmediato. Las cosas valiosas de la vida son difíciles de lograr y pocas cosas son más valiosas que una opinión sincera. Los cuestionarios clásicos en cierto modo son una simplificación ingenua del proceso mental que lleva a una persona a tomar una decisión. Los cuestionarios tipo CBC afrontan la complejidad que subyace en cada toma de decisión, permitiendo al investigador llegar tan lejos en su comprensión como esté dispuesto a llegar.

Dos son los principales hechos diferenciales de este tipo de estudios frente a los cuestionarios clásicos. En primer lugar, las personas no valoramos los atributos de los productos o servicios de forma independiente, las valoramos formando un todo. ¿Cómo de importante es la seguridad en un vehículo? ¿Y el precio? ¿Y el confort? Si preguntamos las cosas así, sólo podemos obtener una respuesta: todo es importante, todos queremos cualquier atributo deseable en un producto. La importancia de un atributo sólo tiene sentido en relación al resto de atributos. Los atributos deseables suelen ir acompañados de otros menos deseables, habitualmente un incremento de precio.

En segundo lugar, la mayor parte de las decisiones que toma el ser humano no son valoraciones, son elecciones. Nos pasamos el día eligiendo: comprar el producto A o B, ir al trabajo en transporte público o en automóvil… Los mecanismos que nos llevan a decidir una opción son procesos sofisticados, una parte importante de los cuales operan fuera del nivel consciente del individuo. Los cuestionarios tradicionales tratan de comprender estos mecanismos preguntando directamente por ellos. Es inútil en muchos casos: las respuestas que obtendremos son reconstrucciones racionales que el decisor hace sobre cómo cree que debería decidir, no sobre cómo decide realmente. Los ejemplos de esta divergencia entre lo que decimos y lo que hacemos son numerosos: la seguridad de un vehículo debería ser su atributo más importante, pero no parece ser el elemento más valorado en el momento de elegir un nuevo automóvil. Poca gente admite comprar un producto lujoso por el impacto que produce en su entorno social.

Los experimentos *conjoint* CBC tratan de comprender los procesos que operan en la toma de decisiones, a través de la observación de las elecciones de los individuos. La forma en que elegimos habla de la importancia relativa que otorgamos a cada atributo presente en las opciones que se nos ofrecen. Una

sucesión de elecciones puede ser suficientemente informativa como para asignar un peso – o utilidad – a cada uno de esos atributos. Dicho en otras palabras: no preguntemos, observemos.

Sin embargo, si este tipo de metodologías ofrecen mejor información que el cuestionario clásico, ¿por qué no se utilizan con más frecuencia? La respuesta debemos buscarla en la falta de difusión y conocimiento de los modelos estadísticos detrás de estas técnicas, los conocidos como métodos de elección discreta, que nos permiten acceder al peso de los atributos a partir de las elecciones observadas. Existe muy poca literatura accesible para personas fuera del ámbito académico, que ofrezca una visión clara y comprensible de estas metodologías.

El hallazgo del libro que tiene en sus manos fue una revelación para mí. El profesor Kenneth E. Train es una de aquellas personas que tiene el don hacer fáciles las cosas difíciles. Su obra es una revisión de las diferentes técnicas existentes para el análisis de decisiones discretas, desde lo más simple a lo más complejo, redactado de una manera comprensible para aquellos lectores menos avezados en la materia, sin renunciar al rigor y a la exhaustividad que un investigador experimentado espera encontrar en la obra de una persona del prestigio del profesor Train.

Pude acceder a este libro gracias a que el profesor Train decidió, de forma totalmente altruista, difundir una edición digital desde su página web personal. Un gesto que le honra, y que contribuye a la difusión de este conjunto de valiosas técnicas. Por mi parte, he querido contribuir a esta difusión traduciendo este libro y poniéndolo al alcance de investigadores y profesionales de habla hispana. Propuse la idea al profesor Train y encontré por su parte todas las facilidades para llevarla a cabo.

Espero que el lector disfrute de su lectura tanto como yo lo he hecho, y que pueda dar utilidad a los contenidos que aquí se explican. Tan sólo puedo añadir que desde que descubrí estas técnicas, además de dar soporte a nuestros clientes en la programación online de estudios tipo conjoint, en Netquest hemos podido emplearlas en relación a nuestra principal área de actividad: la creación y gestión de paneles de personas. Cuestiones como qué variables determinan la participación de una persona en una encuesta, qué método de incentivación logra mejor participación o qué factores determinan el canje de puntos por regalos en un sistema de incentivos, son algunas de las preguntas para las que hemos hallado respuesta con la ayuda de las técnicas expuestas aquí.

Tan sólo me queda desearle una agradable lectura y agradecerle nuevamente al profesor Train su generosidad y su colaboración para hacer posible esta edición de su obra en castellano.

Carlos Ochoa

Dedicado a Daniel McFadden
y en memoria de Kenneth Train , Sr.

MÉTODOS DE ELECCIÓN DISCRETA CON SIMULACIÓN

Segunda Edición

Autor: Kenneth E. Train 2009.

Traducción all castellano: Carlos Ochoa, 2014.

Primera publicación en inglés 2009.

Primera publicación en español 2014.

Contenidos

KENNETH E. TRAIN

El profesor Kenneth E. Train imparte cursos sobre econometría, regulación y organización industrial en la Universidad de California, Berkeley. Asimismo ocupa la plaza de vicepresidente de la National Economic Research Associates (NERA), Inc., en San Francisco, California. Autor de "Optimal Regulation: The Economic Theory of Natural Monopoly" (1991) y "Qualitative Choice Analysis" (1986), el Dr. Train ha escrito más de 60 artículos sobre teoría económica y regulación. Train presidió el Center for Regulatory Policy en la Universidad de California, Berkeley, desde 1993 hasta el 2000 y ha testificado como experto en procedimientos reguladores y casos judiciales. Ha recibido numerosos galardones por su actividad como docente e investigador.

TRADUCCIÓN AL CASTELLANO:

Carlos Ochoa, 2014

CON LA COLABORACIÓN DE

netquest
Campo Online Avanzado

1

Introducción

1.1 Motivación

Cuando escribí mi primer libro, *"Qualitative Choice Analysis"*, a mediados de los años 80, este campo del conocimiento había alcanzado un momento crítico. Los conceptos innovadores que lo definían habían sido descubiertos. Los modelos básicos – principalmente logit y logit jerárquico – habían sido introducidos, y las propiedades estadísticas y económicas de estos modelos se habían inferido. Estos conceptos habían sido aplicados con éxito en diferentes áreas, incluyendo transporte, energía, vivienda y marketing, por nombrar sólo unas cuantas.

Este ámbito está hoy en día en un momento similar en relación a una nueva generación de procedimientos. Los modelos de primera generación tenían limitaciones importantes que reducían su utilidad práctica y su realismo. Esas limitaciones fueron claramente identificadas en su momento, pero la forma de superarlas no había sido descubierta. A lo largo de los últimos veinte años se han realizado enormes progresos, lo que nos ha llevado a un cambio radical en los métodos de análisis de la elección. Los primeros modelos han sido complementados por nuevos métodos, más potentes y flexibles. Los nuevos conceptos han surgido gradualmente, gracias a investigadores edificando sobre el trabajo de otros investigadores. Sin embargo, en cierto modo, el cambio ha sido más parecido a un salto brusco que a una progresión gradual. La forma en que los investigadores piensan, especifican y estiman sus modelos, ha cambiado. Y lo que es más importante, un alto grado de consenso, o de comprensión, parece haber emergido en relación a la nueva metodología. Entre los investigadores que trabajan en este campo, un evidente sentido del propósito y del progreso prevalece.

Mi propósito al escribir este nuevo libro es reunir todas estas ideas, en una forma que ejemplifique la unificación de criterios que a mi parecer se ha logrado, y de una manera que haga estos métodos accesibles para una amplia audiencia. Los avances se han centrado principalmente en la simulación. En esencia, la simulación es la respuesta del investigador a la incapacidad de los ordenadores de realizar la operación de integración. O dicho de forma más precisa, la simulación proporciona una aproximación numérica a las integrales, existiendo diferentes métodos que ofrecen diferentes propiedades, siendo aplicable cada uno de ellos a diferentes tipos de integrandos.

La simulación permite la estimación de modelos intratables por otras vías. Prácticamente cualquier modelo puede ser estimado mediante alguna forma de simulación. El investigador se ve liberado de esta forma de antiguas restricciones sobre la especificación del modelo, restricciones que reflejaban más la conveniencia matemática que la realidad económica de la situación estudiada. Esta nueva flexibilidad es un tremendo impulso para la investigación. Hace posible una representación más realista de la enorme

variedad de situaciones relativas a la elección que aparecen en el mundo. Permite al investigador obtener más información a partir de un conjunto de datos y, en muchos casos, permite afrontar problemas hasta ahora inabordables. Esta flexibilidad supone, sin embargo, una nueva carga para el investigador. En primer lugar, los nuevos métodos son en sí mismos más complicados que los anteriores, y utilizan numerosos conceptos y procedimientos que no se estudian en cursos de econometría típicos. Entender las diferentes técnicas – sus ventajas y limitaciones, y las relaciones entre ellas – es importante para escoger el método apropiado para un caso práctico específico y para desarrollar nuevos métodos cuando ninguno de los existentes parece apropiado. El propósito de este libro es ayudar al lector a lo largo de este camino.

En segundo lugar, para implementar un nuevo método o una variante de un método existente, el investigador necesita ser capaz de programar el procedimiento mediante software. Esto significa que el investigador a menudo necesitará conocer cómo funciona desde un punto de vista computacional la estimación mediante máxima verosimilitud (*maximum likelihood*) y otros métodos de estimación, cómo programar modelos específicos y cómo modificar programas existentes para representar variaciones en el comportamiento. Algunos modelos, como por ejemplo el logit mixto o el probit puro (adicionalmente al logit estándar), están implementados en paquetes de software estadístico disponibles comercialmente. De hecho, el código de estos y otros modelos, así como manuales y datos de ejemplo, están disponibles (de forma gratuita) en mi página web http://elsa.berkeley.edu/~train. Cuando sea apropiado, los investigadores deberían usar código ya disponible en lugar de escribir su propio código. Sin embargo, el valor real del nuevo enfoque dado a los modelos de elección es la capacidad de crear modelos a medida. Las tareas de cálculo y programación que se necesitan para implementar un nuevo modelo no son difíciles por norma general. Un objetivo importante del libro es enseñar estas capacidades como parte integral de la explicación de los propios modelos. Personalmente, considero que programar es extremadamente valioso a nivel pedagógico. El proceso de programación de un modelo me ayuda a comprender cómo funciona exactamente, las motivaciones e implicaciones de su estructura, qué características constituyen los elementos esenciales que no pueden ser cambiados para preservar el enfoque básico, y qué características son arbitrarias y pueden ser fácilmente modificadas. Imagino que otras personas también aprenden de esta misma manera.

1.2 Probabilidades de elección e integración

Para centrar ideas, voy a establecer la base conceptual de los modelos de elección discreta y a mostrar cómo la integración entra en juego. Un agente (por ejemplo, una persona, una empresa, un decisor) afronta la necesidad de realizar una elección, o una serie de elecciones a lo largo del tiempo, entre varias opciones disponibles. Por ejemplo, un consumidor elige qué producto comprar entre varios disponibles; una empresa decide qué tecnología usar en su producción; un estudiante elige qué respuesta dar a un test de respuesta múltiple; un participante en una encuesta elige un número entero entre 1 y 5 en una pregunta con una escala tipo *likert*; un trabajador elige si debe continuar trabajando cada año o retirarse. Nos referiremos al resultado de la decisión o decisiones tomadas en cualquier situación de elección como y, indicando la opción elegida o la secuencia de opciones. Asumimos para los propósitos de este libro que la variable resultado es discreta en el sentido de que puede tomar un conjunto numerable de valores. Muchos de los conceptos que describimos son fácilmente generalizables a situaciones en las que la variable resultado es continua. Sin embargo, la notación y la terminología son diferentes cuando tratamos con variables continuas en lugar de discretas. Asimismo, las elecciones discretas generalmente revelan menos información sobre el proceso de elección que las elecciones con resultado continuo, por lo que habitualmente la econometría de la elección discreta es más compleja.

Nuestro objetivo es entender el proceso de comportamiento que conduce a la elección realizada por el agente. Tomamos para ello una perspectiva causal. Hay factores que colectivamente determinan, o causan, la elección del agente. Algunos de estos factores son observados por el investigador y otros no.

A los factores observados los llamaremos x, y a los factores no observados ε. Los factores se relacionen con la elección del agente a través de una función $y = h(x, \varepsilon)$. Esta función la denominaremos proceso de comportamiento (*behavioral process*). Es determinista en el sentido de que dado x y ε, la elección del agente está totalmente determinada.

Pero dado que ε no ha sido observado, la elección del agente no es determinista y no puede ser predicha exactamente. En su lugar, calculamos la *probabilidad* de cualquier posible resultado. Los términos no observados son considerados aleatorios con una densidad de probabilidad $f(\varepsilon)$. La probabilidad de que el agente elija un resultado particular entre el conjunto de todos los posibles resultados es simplemente la probabilidad de que los factores no observados sean tales que hagan que el proceso de comportamiento arroje un resultado concreto: $P(y|x) = Prob(\varepsilon\ s.t.\ h(x, \varepsilon) = y)$.

Podemos expresar esta probabilidad de una forma más práctica. Definamos una función indicadora $I\,[h(x, \varepsilon) = y]$ que toma el valor 1 cuando la expresión entre corchetes es verdadera y 0 cuando es falsa. Es decir, $I\,[\cdot] = 1$ si el valor de ε, combinado con x, induce al agente a elegir un resultado y, y $I\,[\cdot] = 0$ si el valor de ε, combinado con x, induce al agente a elegir otro resultado. De esta forma, la probabilidad de que el agente escoja el resultado y es simplemente el valor esperado de esta función indicadora, donde la esperanza se calcula respecto a todos los posibles valores de los factores no observados:

$$P(y \mid x) = Prob(I\,[h(x, \varepsilon) = y] = 1)$$

$$= \int I\,[h(x, \varepsilon) = y] f(\varepsilon) d\varepsilon \qquad (1.1)$$

Expresada de esta forma, la probabilidad es una integral, concretamente una integral de un indicador del resultado del proceso de comportamiento sobre todos los posibles valores de los factores no observados.

Para calcular esta probabilidad, debemos evaluar esta integral. Existen tres posibilidades para hacerlo.

1.2.1 Cálculo basado completamente en una expresión cerrada

Para ciertas especificaciones de h y f, la integral puede expresarse de forma cerrada. En esos casos, la probabilidad de elección puede calcularse de forma exacta a partir de dicha fórmula. Por ejemplo, consideremos un modelo logit binario relativo a si una persona realiza una acción o no, por ejemplo comprar un nuevo producto. El modelo de comportamiento se especifica de la siguiente manera. La persona obtendría cierto beneficio neto, o utilidad, en caso de realizar la acción. Esta utilidad, que puede ser positiva o negativa, está constituida por una parte que es observada por el investigador, $\beta' x$, donde x es un vector de variables y β es un vector de parámetros, y una parte que no es observada, ε: $U = \beta' x + \varepsilon$. La persona realiza la acción sólo si la utilidad es positiva, es decir, sólo si emprender la acción le proporciona un beneficio neto. La probabilidad de que la persona realice la acción, dada la información que el investigador puede observar, es por lo tanto $P = \int I\,[\beta' x + \varepsilon > 0]\, f\,(\varepsilon)\, d\varepsilon$, donde f es la densidad de probabilidad de ε. Asumamos que ε se distribuye logísticamente, de manera que su densidad es $f(\varepsilon) = \mathrm{e}^{-\epsilon}/(1 + \mathrm{e}^{-\epsilon})^2$ con una distribución de probabilidad acumulada $F(\varepsilon) = 1/(1 + \mathrm{e}^{\epsilon})$. En este caso, la probabilidad de que la persona realice la acción será:

$$P = \int I\,[\beta' x + \varepsilon > 0]\, f\,(\varepsilon)\, d\varepsilon$$

$$= \int I\,[\varepsilon > -\beta' x]\, f\,(\varepsilon)\, d\varepsilon$$

$$= \int_{\varepsilon = -\beta' x}^{\infty} f\ (\varepsilon) d\varepsilon$$

$$= 1 - F(-\beta' x) = 1 - \frac{1}{1 + e^{\beta' x}}$$

$$= \frac{e^{\beta' x}}{1 + e^{\beta' x}}$$

Para cualquier x, la probabilidad puede calcularse de forma exacta como $P = exp(\beta' x)/(1 + exp(\beta' x))$.

Otros modelos también tienen una expresión cerrada para las probabilidades. Los modelos logit multinomial (capítulo 3), logit jerárquico (capítulo 4) y logit ordenado (capítulo 7) son ejemplos destacados. Los métodos que describí en mi primer libro y que fueron la base del interés inicial que despertó el análisis de la elección discreta, se apoyaban casi exclusivamente en modelos con expresión cerrada para las probabilidades de elección. En general, sin embargo, la integral necesaria para el cálculo de probabilidades no puede ser expresada de forma cerrada. O siendo más precisos, debemos aplicar restricciones sobre el modelo de comportamiento h y la distribución de probabilidad de los términos aleatorios f para lograr que la integral tenga una expresión cerrada. Estas restricciones pueden hacer los modelos poco realistas en muchas situaciones.

1.2.2 Cálculo basado completamente en la simulación

En lugar de resolver la integral de forma analítica, es posible aproximar su resultado mediante simulación. La simulación es aplicable de una manera u otra a prácticamente cualquier especificación de h y f. La simulación se fundamenta en el hecho de que integrar sobre una densidad de probabilidad es una forma de promediar. Consideremos la integral $\bar{t} = \int t(\varepsilon) f(\varepsilon) d\varepsilon$, donde $t(\varepsilon)$ es un estadístico basado en ε con densidad de probabilidad $f(\varepsilon)$. Esta integral corresponde al valor esperado de t sobre todos los posibles valores de ε. Este promedio puede aproximarse de una forma intuitivamente directa. Tomemos múltiples realizaciones (valores al azar) de la variable aleatoria ε a partir de su distribución de probabilidad f, calculemos $t(\varepsilon)$ para cada valor, y promediemos los resultados. Este promedio simulado es un estimador no sesgado del promedio real. Este procedimiento aproxima el valor del promedio real a medida que se utilizan más y más valores en la simulación.

Este concepto de simulación de un promedio es la base de todos los métodos de simulación, por lo menos de todos los que consideramos en este libro. Tal y como se indica en la ecuación (1.1), la probabilidad de que se produzca un resultado concreto es un promedio del indicador $I\ [\cdot]$ sobre todos los posibles valores de ε. La probabilidad, cuando se expresa de esta forma, puede ser simulada directamente como sigue:

1. Extraemos un valor al azar de ε a partir de $f(\varepsilon)$. Etiquetamos este valor como ε^1, donde el superíndice indica que es la primera realización.
2. Determinamos si $h(x, \varepsilon^1) = y$ usando este valor de ε. Si es así, creamos $I^1 = 1$; en caso contrario fijamos $I^1 = 0$.
3. Repetimos los pasos 1 y 2 muchas veces, hasta un total de R valores. El indicador obtenido para cada realización se etiqueta como I^r donde $r = 1, \ldots, R$.

4. Calculamos el promedio de los I^r . Este promedio es la probabilidad simulada: $\check{P}(y|x) = \frac{1}{R}\sum_{r=1}^{R} I^r$. Es la proporción de veces que los valore extraidos al azar de los factores no observados, en combinación con las variables observadas x, han producido un resultado y.

Como veremos en los siguientes capítulos, este simulador, aunque es fácil de comprender, tiene algunas propiedades desafortunadas. Las probabilidades de elección a menudo pueden expresarse como promedios de otros estadísticos, en lugar de promedios de una función indicadora. Los simuladores basados en estos otros estadísticos se calculan de forma análoga, mediante la extracción de valores al azar de la densidad de probabilidad, calculando el estadístico, y promediando los resultados. El modelo probit (capítulo 5) es el ejemplo más representativo de un modelo estimado completamente por simulación. Varios métodos para simular las probabilidades del modelo probit han sido desarrollados basándose en promedios de varios estadísticos sobre varias densidades (relacionadas).

1.2.3 Cálculo basado parcialmente en la simulación, parcialmente en una expresión cerrada

Hasta ahora hemos presentado los dos polos opuestos: o resolvemos la integral analíticamente o mediante simulación. En muchas ocasiones, es posible hacer un poco de ambas cosas.

Supongamos que los términos aleatorios pueden descomponerse de dos partes, que llamaremos ε_1 y ε_2. La densidad de probabilidad conjunta de estos dos términos sería $f(\varepsilon) = f(\varepsilon_1, \varepsilon_2)$. La densidad conjunta puede expresarse como el producto de una densidad marginal y una densidad condicionada: $f(\varepsilon_1, \varepsilon_2) = f(\varepsilon_1|\varepsilon_2) \cdot f(\varepsilon_1)$. Usando esta descomposición, la probabilidad de la ecuación (1.1) puede expresarse como

$$P(y|x) = \int I\,[h(x,\varepsilon) = y] f(\varepsilon) d\varepsilon$$

$$\int_{\varepsilon_1} \left[\int_{\varepsilon_2} I\,[h(x,\varepsilon_1,\varepsilon_2) = y] f(\varepsilon_2|\varepsilon_1) d\varepsilon_2 \right] f(\varepsilon_1) d\varepsilon_1$$

Ahora supongamos que existe una expresión cerrada para la integral que se encuentra dentro de los corchetes grandes. Denominemos esta fórmula como $g(\varepsilon_1) \equiv \int_{\varepsilon_2} I\,[h(\mathrm{x},\varepsilon_1,\varepsilon_2) = \mathrm{y}] f(\varepsilon_2|\varepsilon_1)\mathrm{d}\varepsilon_2$, fórmula que está condicionada respecto al valor de ε_1. La probabilidad se puede simular extrayendo valores al azar de $f(\varepsilon_1)$, calculando $g(\varepsilon_1)$ para cada realización, y promediando posteriormente los resultados.

Este procedimiento se denomina *partición conveniente del error* (*convenient error partitioning*, Train, 1995). La integral respecto a ε_2 dado ε_1 se calcula exactamente, mientras que la integral respecto a ε_1 se calcula mediante simulación. Esta aproximación al problema presenta ventajas claras respecto a la simulación completa. Las integrales analíticas son más precisas y más fáciles de calcular que las integrales simuladas. Es útil por lo tanto, cuando es posible, descomponer los términos aleatorios de manera que una parte de ellos pueda ser integrada analíticamente, aun cuando el resto de términos deban ser simulados. Logit mixto (capítulo 6) es un ejemplo representativo de modelo que usa esta descomposición de forma efectiva. Otros ejemplos son el probit binario sobre datos de un panel, a cargo de Gourieroux and Monfort (1993), y el análisis de respuestas ordenadas de Bhat (1999).

1.3 Esquema del libro

El análisis de elecciones discretas consta de dos tareas interrelacionadas: la especificación del modelo de comportamiento y la estimación de los parámetros del modelo. La simulación juega un papel en ambas tareas. Por una parte, la simulación permite al investigador aproximar las probabilidades de

elección que surgen del modelo de comportamiento. Tal y como hemos mostrado, la capacidad de usar simulación da libertad al investigador para especificar modelos sin la restricción de tener que trabajar con probabilidades que tengan necesariamente una expresión cerrada. Por otra parte, la simulación también entra en juego en la tarea de estimación. Las propiedades de un estimador, como por ejemplo el estimador de máxima verosimilitud, pueden cambiar cuando se utilizan probabilidades simuladas en lugar de las probabilidades reales. Comprender estos cambios y mitigar los efectos negativos, es importante para el investigador. En algunos casos, como en los procedimientos Bayesianos, el estimador mismo es una integral sobre una densidad (en contraposición a los casos en los que la probabilidad de elección es una integral). La simulación permite implementar estos estimadores incluso cuando la integral que define el estimador no tiene una expresión cerrada.

Este libro se organiza en torno a estas dos tareas. La Parte I describe modelos de comportamiento que han sido propuestos para describir el proceso de elección. Los capítulos en esta sección van desde el modelo más simple, logit, hasta modelos progresivamente más generales y consecuentemente más complejos. Dedicamos un capítulo a cada uno de los siguientes modelos: logit, la familia de modelos generalizados de valor extremo (cuyo miembro más destacado es el logit jerárquico), probit y logit mixto. Esta parte del libro finaliza con un capítulo titulado “Variaciones sobre el tema”, que cubre una variedad de modelos que se construyen sobre los conceptos explicados en los capítulos precedentes. El objetivo de este capítulo va más allá de simplemente introducir varios modelos nuevos. El capítulo ilustra el concepto subyacente en todo el libro, a saber, que los investigadores necesitan no confiar únicamente en las pocas especificaciones comúnmente disponibles en software comercial, sino que pueden diseñar modelos que reflejen la singularidad de la configuración, los datos y los objetivos de su proyecto, escribiendo su propio código y usando simulación cuando se requiera.

La Parte II describe la estimación de los modelos de comportamiento. En primer lugar se aborda la maximización numérica, dado que la mayor parte de procedimientos de estimación implican la maximización de alguna función, como por ejemplo la función logaritmo de la verosimilitud (*log-likelihood*). A continuación describimos procedimientos para extraer valores al azar de diferentes tipos de densidades de probabilidad, lo cual es la base de la simulación. Este capítulo también describe diferentes tipos de extracciones de valores al azar, incluyendo variantes del método de antitéticos y las secuencias cuasi-aleatorias, que nos proporcionan mayor precisión en la simulación que el uso de valores aleatorios independientes. A continuación abordamos la estimación asistida por simulación, estudiando en primer lugar los procedimientos clásicos, incluyendo la máxima verosimilitud simulada, el método de momentos simulados y el método de puntuaciones simuladas, y posteriormente los procedimientos Bayesianos, incluyendo los métodos de Monte Carlo – Cadena de Markov. Hasta este punto del libro, asumimos que las variables explicativas son exógenas, es decir, independientes de factores no observados. El capítulo 13, que es nuevo en esta segunda edición, examina la endogeneidad, identificando situaciones en las que los factores no observados están correlacionados con las variables explicativas y describiendo métodos de estimación apropiados para estas situaciones, incluyendo el enfoque BLP, las funciones de control y la máxima verosimilitud con información completa. El capítulo final, que también es nuevo, muestra cómo los algoritmos EM, usados extensamente en otras áreas de la estadística, pueden ser de ayuda para modelos de elección complejos, incluyendo la estimación no paramétrica de la distribución de preferencias entre agentes. La simplicidad y la potencia de los algoritmos EM al ser aplicados a modelos de elección hacen de este capítulo un final apropiados para el libro.

1.4 Un par de notas

A lo largo de todo el libro, me refiero al investigador como “ella” y al decisor como “él”. Este uso, además de ser comparativamente neutral en relación al género (o al menos simétricamente no inclusivo), permite referirnos a ambos sujetos en la misma frase sin confusión.

Muchos colegas han proporcionado comentarios y sugerencias valiosas para este libro. Estoy muy agradecido por su ayuda. Gracias a Greg Allenby, Moshe Ben-Akiva, Chandra Bhat, Denis Bolduc, David Brownstone, Siddhartha Chib, Jon Eisen-Hecht, Florian Heiss, Stephane Hess, David Hensher, Joe Herriges, Rich Johnson, Frank Koppelman, Jordan Louviere, Aviv Nevo, Juan de Dios Ortúzar, John Rose, Ric Scarpa, Ken Small, Joan Walker, Cliff Winston, Joachim Winter y a los estudiantes de mi curso de econometría.

PARTE I

Modelos de comportamiento

2

Propiedades de los modelos de elección discreta

2.1 Resumen

El presente capítulo describe las características comunes de todo modelo de elección discreta. Empezamos con una exposición sobre el concepto de "conjunto de elección", es decir, el conjunto de las diferentes opciones disponibles para el decisor. A continuación obtenemos las probabilidades de elección y las especificamos a partir de un comportamiento encaminado a la maximización de la utilidad. En el contexto de esta especificación general introducimos y comparamos los modelos de elección discreta más representativos, concretamente los modelos logit, valor extremo generalizado (GEV), probit y logit mixto. La utilidad, como una medida construida del bienestar, no tiene una escala natural. Este hecho tiene importantes implicaciones en la especificación y normalización de los modelos de elección discreta, las cuales exploramos. A continuación mostramos cómo modelos a nivel individual pueden agregarse para obtener predicciones a nivel de mercado, y cómo los modelos pueden ser usados para hacer predicciones en el tiempo.

2.2 El conjunto de elección

Los modelos de elección discreta describen las elecciones que los decisores hacen entre diferentes alternativas. Los decisores pueden ser personas, hogares, empresas o cualquier otra unidad con capacidad de escoger, y las alternativas pueden representar productos que compiten entre ellos, acciones a emprender o cualesquiera otras opciones o ítems sobre los cuales las elecciones deben hacerse. Para encajar en un marco de elección discreta, el conjunto de alternativas, llamado *conjunto de elección*, tiene que presentar tres características. En primer lugar, las alternativas deben ser *mutuamente excluyentes* desde el punto de vista del decisor. Escoger una alternativa necesariamente implica no escoger ninguna de las alternativas restantes. El decisor elige sólo una alternativa del conjunto de elección. En segundo lugar, el conjunto de elección debe ser *exhaustivo*, en el sentido de que todas las posibles alternativas deben estar contempladas. El decisor necesariamente elige una de las alternativas. En tercer lugar, el número de alternativas debe ser *finito*. El investigador puede contar las alternativas y finalizar en algún momento el recuento.

El primer y segundo criterios no son restrictivos. Una definición apropiada de las alternativas puede asegurar, prácticamente en todos los casos, que las alternativas sean mutuamente excluyentes y que el conjunto de elección sea exhaustivo. Por ejemplo, supongamos que dos alternativas A y B no son

mutuamente excluyentes porque el decisor puede elegir las dos alternativas. Podemos redefinir las alternativas para que sean “sólo A”, “sólo B” y “tanto A como B”, la cuales son necesariamente mutuamente excluyentes. De forma similar, un conjunto de alternativas podría no ser exhaustivo porque el decisor tiene la opción de no escoger ninguna de ellas. En este caso, podemos definir una alternativa adicional “ninguna de las otras alternativas”. El conjunto de elección extendido, formado por las alternativas originales más esta nueva opción, es claramente exhaustivo.

A menudo el investigador puede satisfacer estas dos condiciones de varias maneras. La especificación apropiada del conjunto de elección en estas situaciones se rige principalmente por los objetivos del investigador y por los datos que están a su alcance. Consideremos la elección que realizan los hogares entre combustibles para calefacción, un tema que ha sido estudiado ampliamente en un esfuerzo para pronosticar el uso de energía y desarrollar programas efectivos para el cambio de combustibles y el ahorro energético. Los combustibles disponibles son generalmente el gas natural, la electricidad, el petróleo y la madera. Estas cuatro alternativas, tal y como se han listado, violan tanto el principio de exclusión mutua como el de exhaustividad. Las alternativas no son mutuamente exclusivas porque un hogar puede (y muchos lo hacen) disponer de dos tipos de calefacción, como por ejemplo una calefacción central de gas natural y calentadores eléctricos en las habitaciones, o una estufa de leña junto a una calefacción eléctrica de placas. Y el conjunto no es exhaustivo porque el hogar puede no tener calefacción (algo que desafortunadamente no es tan extraño como podría pensarse). El investigador puede manejar cada uno de estos problemas de diferentes maneras. Para lograr la exclusividad mutua, una solución pasa por definir como alternativas cada una de las posibles combinaciones de combustibles para calefacción. Las alternativas quedan definidas de esta forma como: “sólo electricidad”, “electricidad y gas natural, pero no otros combustibles”, etc. Otra aproximación es definir la elección como la elección entre combustibles para el sistema de calefacción “principal”. Mediante este procedimiento, el investigador define una regla para determinar qué combustible es el principal cuando un hogar usa varios combustibles para la calefacción. Por definición, sólo un combustible (electricidad, gas natural, petróleo o madera) es el principal. La ventaja de listar todas las posibles combinaciones de combustibles es que evita tener que definir el concepto de combustible “principal”, algo difícil y que representa una distinción un tanto arbitraria. Asimismo, usando todas las combinaciones posibles, el investigador tiene la posibilidad de examinar los factores que determinan que un hogar use múltiples combustibles. Sin embargo, para usar esta solución, el investigador necesita datos que distingan las alternativas, por ejemplo, el costo de calentar un hogar con gas natural y electricidad respecto el costo con gas natural únicamente. Si el investigador restringe el análisis a la elección del combustible principal, los requisitos de los datos son menos severos. Sólo son necesarios los costos asociados a cada combustible. A su vez, un modelo con cuatro alternativas es inherentemente más simple de estimar y de usar en predicciones, que uno con el gran número de alternativas que resulta de todas las posibles combinaciones de los combustibles considerados. El investigador necesitará valorar estos pros y contras para especificar el conjunto de elección.

El mismo tipo de problema surge en relación a la exhaustividad. En nuestro caso de la elección del combustible para calefacción, el investigador puede incluir la opción “sin calefacción” como una alternativa o puede redefinir el problema de elección para que sea la elección de combustible para calefacción condicionada a tener calefacción. La primera aproximación permite al investigador examinar los factores relacionados con el hecho de que un hogar tenga o no calefacción. Sin embargo, esta capacidad sólo es efectiva si el investigador dispone de datos que se relacionen de manera significativa con el hecho de que un hogar tenga o no tenga calefacción. Usando la segunda aproximación, el investigador excluye del análisis los hogares sin calefacción y, haciendo esto, se libera de la necesidad de datos que se refieran a estos hogares.

Como acabamos de describir, las condiciones de exclusividad mutua y exhaustividad generalmente pueden satisfacerse, y el investigador a menudo tiene varias posibilidades para hacerlo. En contraste, la tercera condición, es decir, que el número de alternativas sea finito, es realmente más restrictiva. Esta condición es la característica que define a los modelos de elección discreta y distingue su ámbito de aplicación de la de los modelos de regresión. En los modelos de regresión, la variable dependiente es continua, lo que significa que hay un número infinito de posibles resultados. El resultado podría ser elegido por un decisor, como la decisión de cuánto dinero mantener en cuentas de ahorro. Sin embargo, las alternativas disponibles para el decisor, que son cada posible valor monetario por encima de cero, no son finitas (al menos no lo son si se consideran todas las partes, un tema que al que volveremos luego). Cuando hay un número infinito de alternativas, los modelos de elección discreta no son aplicables.

A menudo, los modelos de regresión y los modelos de elección discreta se distinguen diciendo que las regresiones examinan elecciones de "cuánto" y los modelos de elección discreta elecciones de "cuál". Esta distinción, aunque quizá es ilustrativa, no es del todo precisa. Los modelos de elección discreta pueden y han sido utilizados para examinar las elecciones de "cuánto". Un ejemplo representativo es la elección que realizan los hogares sobre cuántos automóviles poseen. Las alternativas son 0, 1, 2 y así sucesivamente, hasta el número más grande que el investigador considere posible (u observa). Este conjunto de elección contiene un número finito de alternativas exhaustivas y mutuamente exclusivas, apropiadas para ser analizadas mediante un modelo de elección discreta. El investigador también puede definir el conjunto de elección de forma más sucinta como 0, 1 y 2 o más vehículos, si los objetivos de la investigación pueden lograrse con esta especificación.

Cuando se consideran de esta forma, muchas elecciones que implican “cuántos” pueden representarse mediante un modelo de elección discreta. En el ejemplo de las cuentas de ahorro, cada incremento de un dólar (o incluso cada incremento de un centavo) puede considerarse una alternativa y, siempre y cuando exista algún máximo finito, el conjunto de elección se ajustará a los criterios impuestos por un modelo de elección discreta. La conveniencia de usar en estas situaciones un modelo de regresión o un modelo de elección discreta es una cuestión de especificación que el investigador debe tener en consideración. Por lo general, un modelo de regresión es más natural y simple. Un modelo de elección discreta debería usarse en estas situaciones sólo si existen razones de peso para hacerlo. Como ejemplo, Train et al. (1987) analizaron el número y la duración de las llamadas telefónicas que los hogares hacen, usando un modelo de elección discreta en lugar de un modelo de regresión debido a que el modelo de elección discreta permitía una mayor flexibilidad en el manejo de las tarifas no lineales de precios que los hogares manejan. En general, el investigador debe tener en cuenta los objetivos de la investigación y las capacidades de los diferentes métodos en el momento de decidir si se debe aplicar un modelo de elección discreta.

2.3 Obtención de las probabilidades de elección

Los modelos de elección discreta se obtienen habitualmente bajo el supuesto de que el decisor se comporta de forma que maximiza la utilidad que percibe. Thurstone (1927) desarrolló en primer lugar estos conceptos en términos de estímulos psicológicos, dando lugar a un modelo probit binario relativo a si los encuestados pueden diferenciar el nivel de estímulo recibido. Marschak (1960) interpretó los estímulos como una utilidad y proporcionó una formulación a partir de la maximización de la utilidad. Siguiendo los pasos de Marschak, los modelos que pueden obtenerse de esta manera reciben el nombre de modelos de utilidad aleatoria (*random utility models,* RUMs). Sin embargo, es importante señalar que los modelos obtenidos a partir de la maximización de la utilidad también pueden ser usados para representar tomas de decisiones que no implican maximización de utilidad. La forma en que se obtiene el modelo garantiza su consistencia con la maximización de la utilidad, pero no se opone a que el modelo pueda ser coherente con otras formas de comportamiento. Los modelos también pueden ser

vistos como simples descripciones de la relación existente entre variables explicativas y el resultado de una elección, sin referencia exacta a cómo se realiza la elección.

Los modelos de utilidad aleatoria (RUMs) se obtienen de la siguiente manera. Un decisor, llamémosle n, se enfrenta a una elección entre J alternativas. El decisor obtendría un cierto nivel de utilidad (o ganancia) en caso de escoger cada alternativa. La utilidad que el decisor n obtiene de la alternativa j es U_{nj}, j = 1 … J. Esta utilidad es conocida para el decisor, pero no lo es, como veremos a continuación, para el investigador. El decisor elige la alternativa que le proporciona la mayor utilidad. Por lo tanto, el modelo de comportamiento es: elige la alternativa i si y sólo si $U_{ni} > U_{nj}$ ∀j ≠ i.

Consideremos ahora el rol del investigador. El investigador no observa la utilidad del decisor. El investigador observa algunos atributos de las alternativas que afronta el decisor, etiquetados como x_{nj}∀j, y algunos atributos del decisor, etiquetados como s_{nj}, y puede especificar una función que relaciona estos factores observados con la utilidad que percibe el decisor. Esta función se denota como $V_{nj} = \mathrm{V}(x_{nj}, s_n)$ ∀j y suele llamarse a menudo *utilidad representativa*. Por lo general, V depende de parámetros desconocidos para el investigador y que por lo tanto son estimados estadísticamente; sin embargo, suprimiremos esta dependencia por el momento.

Puesto que hay aspectos de la utilidad que el investigador no observa o no puede observar, $V_{nj} \neq U_{nj}$. Podemos descomponer la utilidad como $U_{nj} = V_{nj} + \varepsilon_{\mathrm{nj}}$, donde $\varepsilon_{\mathrm{nj}}$ captura factores que afectan a la utilidad, pero que no están incluidos en V_{nj}. Esta descomposición es totalmente general, ya que $\varepsilon_{\mathrm{nj}}$ se define simplemente como la diferencia entre la verdadera utilidad U_{nj} y la parte de la utilidad que el investigador captura en V_{nj}. Teniendo en cuenta su definición, las características de $\varepsilon_{\mathrm{nj}}$ tales como su distribución de probabilidad, dependen de forma crítica de la especificación que el investigador haga de V_{nj}. En particular, $\varepsilon_{\mathrm{nj}}$ no está definido para una situación de elección *per se*. Más bien, se define en relación a la representación que un investigador hace de esa situación de elección. Esta distinción se hace relevante al evaluar la idoneidad de diversos modelos específicos de elección discreta.

El investigador no conoce $\varepsilon_{\mathrm{nj}}$ $\forall j$, por lo que trata estos términos como variables aleatorias. La densidad de probabilidad conjunta del vector aleatorio $\varepsilon_n' = \langle \varepsilon_{n1}, \ldots, \varepsilon_{nJ} \rangle$ se denota como $f(\varepsilon_{\mathrm{n}})$. Con esta densidad, el investigador puede hacer afirmaciones probabilísticas acerca de la elección del decisor. La probabilidad de que el decisor n elija la alternativa i es

$$P_{ni} = Prob(U_{ni} > U_{nj}\ \forall j \neq i)$$

$$= Prob(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj}\ \forall j \neq i)$$

$$= Prob\big(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}\ \forall j \neq i\big). \qquad (2.1)$$

Esta probabilidad es una distribución acumulativa, es decir, es la probabilidad de que cada término aleatorio $\varepsilon_{\mathrm{nj}} - \varepsilon_{\mathrm{ni}}$ esté por debajo de la cantidad observada $V_{ni} - V_{nj}$. Usando la densidad $f(\varepsilon_{\mathrm{n}})$, esta probabilidad acumulativa puede reescribirse como

$$P_{ni} = Prob(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}\ \forall j \neq i)$$

$$= \int_{\varepsilon} I\big(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}\ \forall j \neq i\big) f(\varepsilon_n) d\varepsilon_n, \qquad (2.2)$$

donde $I(\cdot)$ es la función indicadora, igual a 1 cuando la expresión entre paréntesis es verdadera y 0 en caso contrario. Esta expresión es una integral multidimensional sobre la densidad de probabilidad de la

parte no observada de la utilidad, $f(\varepsilon_n)$. Diferentes modelos de elección discreta se obtienen mediante especificaciones diferentes de esta densidad, es decir, a partir de diferentes supuestos acerca de cómo se distribuye la densidad de probabilidad de la parte no observada de la utilidad. La integral tiene una forma cerrada sólo para ciertas especificaciones de $f(\cdot)$. Logit y logit jerárquico tienen una expresión cerrada para esta integral. Se obtienen bajo la suposición de que la parte no observada de la utilidad se distribuye de acuerdo a una distribución de tipo valor extremo independiente e idénticamente distribuida (iid en adelante) y una distribución de tipo valor extremo generalizada, respectivamente. Probit se obtiene bajo la suposición de que $f(\cdot)$ es una normal multivariada y logit mixto se basa en la asunción de que la parte no observada de la utilidad consiste en una parte que sigue cualquier distribución especificada por el investigador más una parte que es de tipo valor extremo iid. Tanto en el caso de probit como de logit mixto, la integral resultante no tiene una forma cerrada y debe evaluarse numéricamente mediante simulación. Cada uno de estos modelos se trata en detalle en los siguientes capítulos.

El significado de las probabilidades de elección es más sutil, y más revelador, de lo que podría parecer a primera vista. Un ejemplo nos sirve de ilustración. Consideremos una persona que puede ir al trabajo en automóvil o en autobús. El investigador observa el tiempo y el costo en que la persona incurre usando cada medio de transporte. Sin embargo, el investigador se da cuenta de que hay otros factores, además del tiempo y del costo, que afectan a la utilidad de la persona y por lo tanto a su elección. El investigador especifica

$$V_c = \alpha T_c + \beta M_c,$$

$$V_b = \alpha T_b + \beta M_b,$$

donde T_c y M_c son el tiempo y el costo (M en relación a *money*, dinero) en que la persona incurre al viajar al trabajo en automóvil, T_b y M_b se definen de forma análoga para el autobús, y el subíndice n que denota al individuo se omite por conveniencia. Los coeficientes α y β, o bien son conocidos, o bien son estimados por el investigador.

Supongamos que, dados α y β, y las medidas realizadas por los investigadores sobre el tiempo y el costo asociados a viajar en automóvil y autobús, resulta que $V_c = 4$ y $V_b = 3$. Esto significa que, basándonos en los factores observados, el automóvil es mejor para esta persona que el autobús por 1 unidad de diferencia (trataremos más adelante la normalización de la utilidad que define la dimensión de estas unidades). Este resultado no significa, sin embargo, que la persona necesariamente escoja el automóvil, ya que hay otros factores no observados por el investigador que afectan a la persona. La probabilidad de que la persona elija el autobús en lugar del automóvil es la probabilidad de que los factores no observados para el autobús sean suficientemente mejores que los del automóvil como para superar la ventaja que el automóvil tiene en los factores observados. En concreto, la persona va a elegir el autobús si la parte no observada de la utilidad es mayor que la del automóvil por lo menos en 1 unidad, superando así la ventaja de 1 unidad que el automóvil tiene sobre los factores observados. Por tanto, la probabilidad de que esta persona escoja el autobús es la probabilidad de que $\varepsilon_b - \varepsilon_c > 1$. Del mismo modo, la persona va a elegir el automóvil si la parte no observada de la utilidad del autobús no es mejor que la del automóvil por lo menos en 1 unidad , es decir, si $\varepsilon_b - \varepsilon_c < 1$. Dado que 1 es la diferencia entre V_c y V_b en nuestro ejemplo, las probabilidades se pueden expresar más explícitamente como

$$P_c = Prob(\varepsilon_b - \varepsilon_c < V_c - V_b)$$

y

$$P_b = Prob(\varepsilon_b - \varepsilon_c > V_c - V_b)$$

$$= Prob(\varepsilon_c - \varepsilon_b < V_b - V_c).$$

Estas ecuaciones son iguales a la ecuación (2.1), reformuladas para nuestro ejemplo automóvil-autobús.

La cuestión que aparece en la formulación de las probabilidades de elección es: ¿qué se entiende por la distribución de ε_n? La interpretación que el investigador hace sobre esta densidad afecta a la interpretación que hace de las probabilidades de elección. La manera más habitual de interpretar esta distribución es la siguiente. Consideremos una población de personas que perciben la misma utilidad observada $V_{nj}\ \forall j$ para cada persona n. Entre estas personas, los valores de los factores no observados difieren. La densidad $f(\varepsilon_n)$ es la distribución de la parte no observada de la utilidad dentro de la población de personas que perciben la misma porción observada de utilidad. Según esta interpretación, la probabilidad P_{ni} es la proporción o cuota (*share*) de personas que optan por la alternativa i respecto al total de la población de personas que perciben la misma utilidad observada para cada alternativa que la persona n. La distribución también puede considerarse en términos subjetivos, como la representación que el investigador hace de la probabilidad subjetiva de que la parte no observada de la utilidad de la persona tome ciertos valores. En este caso, P_{ni} es la probabilidad de que el investigador atribuya a la persona la elección de la alternativa i, dadas las ideas que el investigador tenga sobre la porción no observada de la utilidad de la persona. Como tercera posibilidad, la distribución puede representar el efecto de factores que son incomprensibles para el propio decisor (representando, por ejemplo, aspectos como una racionalidad limitada), de modo que P_{ni} sería la probabilidad de que estos factores incomprensibles induzcan a la persona a elegir la alternativa i dados los factores comprensibles/racionales observados.

2.4 Modelos específicos

Logit , GEV , probit y logit mixto se analizan en profundidad en los siguientes capítulos. Sin embargo, llegados a este punto, es útil dar un rápido vistazo a estos modelos con el fin de mostrar cómo se relacionan con la formulación general de todo modelo de elección y cómo difieren entre ellos dentro de esta formulación. Como se afirmó con anterioridad, diferentes modelos de elección se obtienen bajo diferentes especificaciones de la densidad de probabilidad de los factores no observados $f(\varepsilon_n)$. Por tanto, la cuestión es qué distribución se asume para cada modelo y cuál es la motivación para estas diferentes asunciones.

Logit (tratado en el capítulo 3) es de lejos el modelo de elección discreta más utilizado. Se obtiene bajo el supuesto de que $\varepsilon_{\mathrm{ni}}$ se distribuye con una densidad de probabilidad de tipo valor extremo iid para todo i . La parte más crítica del supuesto es asumir que los factores no observados no están correlacionados entre alternativas, así como aceptar que tienen la misma varianza para todas las alternativas. Esta hipótesis, aunque restrictiva, proporciona una forma muy conveniente para la probabilidad de elección. La popularidad del modelo logit se debe a esta conveniencia. Sin embargo, la hipótesis de independencia puede ser inadecuada en algunas situaciones. Los factores no observados relacionados con una alternativa concreta podrían ser similares a los relacionados con otra alternativa. Por ejemplo, una persona a la que no le gusta viajar en autobús a causa de la presencia de otros viajeros podría tener una reacción similar frente a los viajes en tren; si esto sucediese, los factores no observados que afectan a autobús y tren estarían correlacionados en lugar de ser independientes. El supuesto de independencia también entra en juego cuando se aplica un modelo logit a secuencias de elecciones en el tiempo. El modelo logit asume que cada elección es independiente de las demás. En muchos casos, es de esperar que factores no observados que afectan a la elección en un período

persistirán, al menos en parte, en el siguiente período, induciendo dependencia entre las elecciones a lo largo del tiempo.

El desarrollo de otros modelos ha venido motivado en gran medida con el fin de evitar el supuesto de independencia dentro de un logit. Los modelos generalizados de valor extremo (GEV, analizados en el capítulo 4) se basan, como su nombre indica, en una generalización de la distribución valor extremo. La generalización puede tomar muchas formas, pero el elemento común es que permite la correlación de factores no observados entre alternativas, de forma que colapsa en un modelo logit cuando esta correlación es cero. Dependiendo del tipo de modelo GEV, las correlaciones pueden ser más o menos flexibles. Por ejemplo, un modelo GEV comparativamente simple coloca las alternativas en varios grupos llamados nidos o jerarquías, con factores no observados que tienen la misma correlación para todas las alternativas dentro de un mismo nido y correlación nula para alternativas en diferentes nidos. Otras formas más complejas de estos modelos permiten esencialmente cualquier patrón de correlación. Los modelos GEV generalmente tienen expresiones cerradas para las probabilidades de elección, por lo que no se necesita simulación para su estimación.

Los modelos probit (capítulo 5) se basan en la suposición de que los factores no observados se distribuyen conjuntamente con una densidad de probabilidad normal: $\varepsilon_n{}' = \langle \varepsilon_{n1}, \dots, \varepsilon_{nJ} \rangle \sim N(0, \Omega)$. Con una matriz de covarianza completa Ω, se puede acomodar cualquier patrón de correlación y heterocedasticidad. Cuando se aplica a secuencias de elecciones a lo largo del tiempo, se asume que los factores no observados son conjuntamente normales entre periodos temporales, así como entre alternativas, con cualquier patrón de correlación temporal. La flexibilidad del modelo probit en el manejo de las correlaciones respecto a las alternativas y el tiempo es su principal ventaja. Su única limitación funcional proviene de su dependencia de la distribución normal. En algunas situaciones, los factores no observados pueden no distribuirse normalmente. Por ejemplo, la disposición de un cliente a pagar por un atributo deseable de un producto es necesariamente positiva. Asumir que este factor no observado se distribuye normalmente entra en contradicción con el hecho de que sea positivo, dado que la distribución normal tiene densidad en ambos lados del cero.

El modelo logit mixto (capítulo 6) permite que los factores no observados sigan cualquier distribución. La característica definitoria de un logit mixto es que los factores no observados pueden descomponerse en una parte que contiene toda la correlación y heterocedasticidad, y otra parte que se distribuye iid valor extremo. La primera parte puede seguir cualquier distribución, incluyendo distribuciones no normales. Demostraremos que el modelo logit mixto puede aproximar cualquier modelo de elección discreta posible y por lo tanto es completamente general.

Otros modelos de elección discreta (capítulo 7) han sido especificados por investigadores con propósitos específicos. A menudo, estos modelos se obtienen mediante la combinación de conceptos de otros modelos existentes. Por ejemplo, un probit mixto se obtiene mediante la descomposición de los factores no observados en dos partes, como en el logit mixto, pero dando a la segunda parte una distribución normal en lugar de valor extremo. Este modelo tiene la generalidad del logit mixto y sin embargo en algunas situaciones puede ser más fácil de estimar. Comprendiendo la formulación y la motivación de todos los modelos, cada investigador puede especificar un modelo a medida de la situación y los objetivos de su investigación.

2.5 Identificación de modelos de elección

Varios aspectos del proceso de comportamiento de la decisión afectan a la especificación y a la estimación de cualquier modelo de elección discreta. Los problemas se pueden resumir fácilmente en dos afirmaciones: “Sólo las diferencias de utilidad importan” y “la escala de la utilidad es arbitraria”. Las

implicaciones de estas afirmaciones son de largo alcance, sutiles y, en muchos casos, muy complejas. Las trataremos a continuación.

2.5.1 Sólo las diferencias de utilidad importan

El nivel absoluto de utilidad es irrelevante tanto para el comportamiento del decisor como para el modelo especificado por el investigador. Si añadimos una constante a la utilidad de todas las alternativas, la alternativa con la utilidad más alta no cambia. El decisor escoge la misma alternativa tanto con $U_{nj}\ \forall j$ como con $U_{nj}+k\ \forall j$ para cualquier constante k. Una forma coloquial de expresar este hecho es "cuando la marea sube, levanta todos los barcos".

El nivel de utilidad tampoco importa desde la perspectiva del investigador. La probabilidad de elección es $P_{ni}=Prob\left(U_{ni}>U_{nj}\forall j\neq i\right)=Prob\left(U_{ni}-U_{nj}>0\ \forall j\neq i\right)$, que sólo depende de la diferencia en la utilidad, no de su nivel absoluto. Cuando la utilidad se descompone en las partes observadas y no observadas, la ecuación (2.1) expresa la probabilidad de elección como $P_{ni}=Prob\left(\varepsilon_{nj}-\varepsilon_{ni}<V_{ni}-V_{nj}\ \forall j\neq i\right)$, que también depende sólo de las diferencias.

El hecho de que sólo las diferencias en utilidad importen tiene varias implicaciones para la identificación y especificación de modelos de elección discreta. En general, significa que los únicos parámetros que pueden estimarse (es decir, están identificados) son aquellos que capturan diferencias entre alternativas. Esta declaración general toma varias formas.

Constantes específicas de alternativa

A menudo es razonable especificar la parte observada de la utilidad como una función lineal respecto a los parámetros más una constante: $V_{nj}=x_{nj}'\beta+k_j$, donde x_{nj} es un vector de variables que describen la alternativa j tal y como es vista por el decisor n, β son los coeficientes de estas variables y k_j es una constante que es específica para la alternativa j. La constante específica de alternativa para una alternativa concreta captura el efecto promedio en la utilidad de todos los factores que no están incluidos en el modelo. Por lo tanto, esta constante realiza una función similar a la constante en un modelo de regresión, que también captura el efecto promedio de todos los factores no incluidos.

Si se incluyen constantes específicas de alternativas, la parte no observada de la utilidad ε_{nj} tiene media cero por la forma en que se construye. Si ε_{nj} tiene una media distinta de cero cuando no se han incluido constantes, añadir las constantes hace que el error remanente tenga media cero: es decir, si $U_{nj}=x_{nj}'\beta+\varepsilon_{nj}^{*}$ con $E(\varepsilon_{nj})^{*}=k_j\neq 0$, entonces $U_{nj}=x_{nj}'\beta+\mathrm{k}_j+\varepsilon_{nj}$ con $E(\varepsilon_{nj})=0$. Es razonable, por lo tanto, incluir una constante en V_{nj} para cada alternativa. Sin embargo, dado que sólo las diferencias en utilidad importan, sólo las diferencias entre las constantes específicas de alternativa son relevantes, no sus niveles absolutos. Para reflejar este hecho, el investigador debe establecer el nivel global de estas constantes.

El concepto se hace evidente en el ejemplo del automóvil y el autobús. Una especificación de la utilidad que tenga la forma

$$U_c=\alpha T_c+\beta M_c+k_c^0+\varepsilon_c,$$

$$U_b=\alpha T_b+\beta M_b+k_b^0+\varepsilon_b,$$

con $k_b^0-k_c^0=\mathrm{d}$, es equivalente a un modelo con

$$U_c=\alpha T_c+\beta M_c+k_c^1+\varepsilon_c,$$

$$U_b = \alpha T_b + \beta M_b + k_b^1 + \varepsilon_b,$$

donde la diferencia entre las nuevas constantes es la misma de las constantes iniciales, es decir, $k_b^1 - k_c^1 = \mathrm{d} = k_b^0 - k_c^0$. Cualquier modelo con la misma diferencia entre constantes será equivalente. En cuanto a la estimación, es imposible estimar las dos constantes simultáneamente dado que hay infinitas parejas de constantes (cualquier pareja de valores que tenga la misma diferencia) que dan lugar a las mismas probabilidades de elección.

Para tener en cuenta este hecho, el investigador debe normalizar los niveles absolutos de las constantes. El procedimiento habitual es normalizar una de las constantes a cero. Por ejemplo, el investigador podría normalizar la constante para la alternativa automóvil a cero:

$$U_c = \alpha T_c + \beta M_c + \varepsilon_c,$$

$$U_b = \alpha T_b + \beta M_b + k_b + \varepsilon_b,$$

En virtud de esta normalización, el valor de k_b es d, que es la diferencia entre las constantes originales (sin normalizar) . De esta forma, la constante de la alternativa autobús se interpreta como el efecto medio de los factores no incluidos en la utilidad de la alternativa autobús en relación a la alternativa automóvil.

Con J alternativas, como máximo podemos incluir $\mathrm{J} - 1$ constantes específicas de alternativa en el modelo, con una de las constantes normalizada a cero. Es irrelevante qué constante se normaliza: las otras constantes se interpretan como relativas a la que se ha fijado a cero. El investigador podría normalizar a un valor distinto de cero, por supuesto, sin embargo no existe ninguna razón para hacerlo ya que la normalización a cero es más sencilla (la constante simplemente se deja fuera del modelo) y tiene el mismo efecto.

Variables sociodemográficas

El mismo problema afecta a la forma en que las variables sociodemográficas entran en un modelo. Los atributos de las alternativas, como el tiempo y el costo de los viajes en los diferentes medios de transporte, por lo general varían entre alternativas. Sin embargo, los atributos del decisor no varían entre alternativas. Sólo pueden entrar en el modelo si se especifican de manera que produzcan diferencias entre la utilidad de las alternativas.

Consideremos por ejemplo el efecto del ingreso de una persona en la decisión de tomar el autobús o el automóvil para ir a trabajar. Es razonable suponer que la utilidad de una persona es mayor si tiene mayores ingresos, tanto si la persona toma el autobús como el automóvil. La utilidad se especifica como

$$U_c = \alpha T_c + \beta M_c + \theta_c^0 Y + \varepsilon_c,$$

$$U_b = \alpha T_b + \beta M_b + \theta_b^0 Y + k_b + \varepsilon_b,$$

donde Y es el ingreso y θ_c^0 y θ_b^0 capturan los efectos que tienen los cambios en ingresos en la utilidad de viajar en automóvil y en autobús, respectivamente. Esperamos que $\theta_c^0 > 0$ y $\theta_b^0 > 0$, dado que tener mayores ingresos hace a la gente más feliz sin importar qué medio de transporte usan. Sin embargo $\theta_c^0 \neq \theta_b^0$, ya que los ingresos probablemente tienen un efecto diferente sobre la persona en función del medio de transporte que elijan para viajar. Dado que sólo las diferencias en utilidad importan, los niveles absolutos de θ_c^0 y θ_b^0 no pueden ser estimados, sólo su diferencia. Para establecer el nivel, uno de estos parámetros se normaliza a cero. El modelo se convierte de esta forma en

$$U_c = \alpha T_c + \beta M_c + \varepsilon_c,$$

$$U_b = \alpha T_b + \beta M_b + \theta_b Y + k_b + \varepsilon_b,$$

donde $\theta_b = \theta_b^0 - \theta_c^0$ se interpreta como el efecto diferencial de los ingresos sobre la utilidad del autobús en comparación con el automóvil. El valor de θ_b puede ser positivo o negativo.

Las variables sociodemográficas pueden entrar en la utilidad de otras maneras. Por ejemplo, el costo a menudo se divide por los ingresos:

$$U_c = \alpha T_c + \beta M_c / Y + \varepsilon_c,$$

$$U_b = \alpha T_b + \beta M_b / Y + \theta_b Y + k_b + \varepsilon_b.$$

El coeficiente del costo en esta especificación es β / Y. Puesto que este coeficiente disminuye con Y, el modelo refleja el concepto de que el costo se vuelve menos importante en la toma de decisiones de una persona en comparación con otros factores, cuando aumentan los ingresos que percibe.

Cuando las variables sociodemográficas aparecen interactuando con los atributos de las alternativas, no hay necesidad de normalizar los coeficientes. Las variables sociodemográficas afectan a las diferencias en utilidad a través de su interacción con los atributos de las alternativas. La diferencia $U_c - U_b = \cdots \beta(M_c - M_b)/\mathrm{Y} \ldots \ldots$ varía con los ingresos, ya que los costos difieren entre alternativas.

Número de términos de error independientes

Tal y como establece la ecuación (2.2), las probabilidades de elección toman la forma

$$P_{ni} = \int_{\varepsilon} I\big(\varepsilon_{nj} - \varepsilon_{ni} < V_{ni} - V_{nj}\ \forall j \neq i\big) f(\varepsilon_n) d\varepsilon_n.$$

Esta probabilidad es una integral J-dimensional sobre la densidad de los J términos de error $\varepsilon_n{}' = \langle \varepsilon_{n1}, \ldots, \varepsilon_{nJ} \rangle$. No obstante, la dimensión puede reducirse reconociendo que sólo las diferencias de utilidad importan. Con J errores (uno para cada alternativa) hay $\mathrm{J} - 1$ diferencias de error. La probabilidad de elección puede ser expresada como una integral $(\mathrm{J} - 1)$-dimensional sobre la densidad de estas diferencias de error:

$$P_{ni} = Prob\big(U_{ni} > U_{nj}\ \forall j \neq i\big)$$

$$= Prob\big(\varepsilon_{nj} - \varepsilon_{ni} > V_{ni} - V_{nj}\ \forall j \neq i\big)$$

$$= Prob\big(\tilde{\varepsilon}_{nji} < V_{ni} - V_{nj}\ \forall j \neq i\big)$$

$$\int_{\varepsilon} I\big(\tilde{\varepsilon}_{nji} < V_{ni} - V_{nj}\ \forall j \neq i\big) g(\tilde{\varepsilon}_{ni}) d\tilde{\varepsilon}_{ni}.$$

donde $\tilde{\varepsilon}_{nji} = \varepsilon_{nj} - \varepsilon_{ni}$ es la diferencia entre errores de las alternativas i y j; $\tilde{\varepsilon}_{ni} = \langle \tilde{\varepsilon}_{n1i}, \ldots, \tilde{\varepsilon}_{nJi} \rangle$ es el vector $(\mathrm{J} - 1)$-dimensional de las diferencias de error, con el símbolo "..." refiriéndose a todas las alternativas excepto la i; y $g(\cdot)$ es la densidad de estas diferencias de error. Expresada de esta manera, la probabilidad de elección es una integral $(\mathrm{J} - 1)$-dimensional.

La densidad de las diferencias entre errores $g(\cdot)$ y la densidad de los errores originales $f(\cdot)$ se relacionan de una manera particular. Supongamos que un modelo se especifica con un error para cada alternativa: $\varepsilon_n = \langle \varepsilon_{n1}, \dots, \varepsilon_{nJ} \rangle$ con densidad $f(\varepsilon_n)$. Este modelo es equivalente a un modelo con $J-1$ errores definidos como $\tilde{\varepsilon}_{njk} = \varepsilon_{nj} - \varepsilon_{nk}$ para cualquier k y densidad $g(\tilde{\varepsilon}_{nk})$ obtenida a partir de $f(\varepsilon_n)$. Para cualquier $f(\varepsilon_n)$ es posible obtener la correspondiente $g(\tilde{\varepsilon}_{nk})$. Sin embargo, dado que ε_n tiene más elementos que $\tilde{\varepsilon}_{nk}$, hay un número infinito de densidades de los J términos de error que generan la misma densidad para las $J-1$ diferencias de error. Dicho de forma equivalente, cualquier $g(\tilde{\varepsilon}_{nk})$ es consistente con un número infinito de diferentes$f(\varepsilon_n)$s. Dado que las probabilidades de elección siempre se pueden expresar dependiendo sólo de $g(\tilde{\varepsilon}_{nk})$, una dimensión de la densidad de $f(\varepsilon_n)$ no puede identificarse y debe ser normalizada por el investigador.

La normalización de $f(\varepsilon_n)$ puede ser manejada de varias maneras. Para algunos modelos, como logit, la distribución de los términos de error es suficientemente restrictiva como para que la normalización se produzca de forma automática al aplicar los supuestos sobre la distribución. Para otros modelos, como probit, la normalización a menudo se obtiene especificando el modelo sólo en términos de diferencias de error, es decir, parametrizando $g(\cdot)$ sin referencia a $f(\cdot)$. En todos los modelos exceptuando los más simples, el investigador debe tener en cuenta el hecho de que sólo la densidad de las diferencias de error afecta a las probabilidades y por lo tanto puede identificarse. Al analizar los distintos modelos en los capítulos siguientes vamos a volver a hablar sobre este problema y cómo manejarlo.

2.5.2 La escala general de la utilidad es irrelevante

Así como la adición de una constante a la utilidad de todas las alternativas no cambia la elección del decisor, tampoco lo hace multiplicar la utilidad de cada alternativa por una constante. La alternativa de mayor utilidad sigue siendo la misma sin importar cómo se haya escalado la utilidad. El modelo $U_{nj}^0 = V_{nj} + \varepsilon_{nj}\ \forall j$ es equivalente a $U_{nj}^1 = \lambda V_{nj} + \lambda\varepsilon_{nj}\ \forall j$ para cualquier $\lambda > 0$. Para tener en cuenta este hecho, el investigador debe normalizar la escala de utilidad.

La forma estándar de normalizar la escala de utilidad es normalizar la varianza de los términos de error. Las escalas de la utilidad y de la varianza de los términos de error están vinculadas por definición. Cuando la utilidad se multiplica por λ, la varianza de cada uno de los ε_{nj} cambia por un factor λ^2: $Var(\lambda\varepsilon_{nj}) = \lambda^2 Var(\varepsilon_{nj})$. Por lo tanto, la normalización de la varianza de los términos de error es equivalente a la normalización de la escala de la utilidad.

Normalización con errores iid

Si se supone que los términos de error están distribuidos independientemente e idénticamente (iid), la normalización de la escala es directa. El investigador normaliza la varianza del error a cierta cantidad, que por lo general se elige por conveniencia. Dado que todos los errores tienen la misma varianza (debido al supuesto de partida) la normalización de la varianza de cualquiera de ellos establece la varianza para todos ellos.

Cuando la parte observada de la utilidad es lineal en relación a los parámetros, la normalización proporciona una manera de interpretar los coeficientes. Considere el modelo $U_{nj}^0 = x_{nj}'\beta + \varepsilon_{nj}^0$ donde la varianza de los términos de error es $Var(\varepsilon_{nj}^0) = \sigma^2$. Supongamos que el investigador normaliza la escala estableciendo la varianza del error a 1. El modelo original se convierte en la siguiente especificación equivalente: $U_{nj}^1 = x_{nj}'(\beta/\sigma) + \varepsilon_{nj}^1$ con $Var(\varepsilon_{nj}^1) = 1$. Los coeficientes β originales aparecen ahora divididos por la desviación estándar de la parte no observada de la utilidad. Los nuevos coeficientes (β/σ) reflejan, por lo tanto, el efecto de las variables observadas en relación con la desviación estándar de los factores no observados.

Los mismos conceptos aplican a cualquier cantidad que el investigador elija para la normalización. Como veremos en el próximo capítulo, las varianzas de error en un modelo logit estándar tradicionalmente se normalizan a $\pi^2/6$, que es aproximadamente 1.6. En este caso, el modelo anterior se convierte en $U_{nj} = x_{nj}'(\beta/\sigma)\sqrt{1.6} + \varepsilon_{nj}$ con $Var(\varepsilon_{nj}) = 1.6$. Los coeficientes todavía reflejan la varianza de la porción no observada de la utilidad. La única diferencia es que los coeficientes son mayores en un factor de $\sqrt{1.6}$.

Si bien es irrelevante qué cantidad es utilizada por el investigador para la normalización, la interpretación de los resultados del modelo debe tener en consideración la normalización. Supongamos, por ejemplo, que un modelo logit y un modelo probit independiente han sido estimados con los mismos datos. Como se ha mencionado recientemente, la varianza del error en un modelo logit está normalizada a 1.6. Supongamos que el investigador ha normalizado el modelo probit para tener varianzas de error de 1, algo que también es tradicional en modelos probit independientes. Es necesario tener en cuenta esta diferencia en la normalización a la hora de comparar las estimaciones de los dos modelos. En particular, los coeficientes en el modelo logit serán $\sqrt{1.6}$ veces mayores que los del modelo probit, simplemente debido a la diferencia en la normalización. Si el investigador no tiene en cuenta esta diferencia de escala al comparar los modelos, podría pensar inadvertidamente que el modelo logit implica que las personas se preocupan más por los atributos (ya que los coeficientes son más grandes) que el modelo probit. Por ejemplo, en un modelo de elección del medio de transporte, supongamos que el coeficiente estimado de costo es de -0.55 a partir de un modelo logit y -0.45 a partir de un modelo probit independiente. Es incorrecto decir que el modelo logit asigna una mayor sensibilidad a los costos que el modelo probit. Los coeficientes en uno de los modelos tienen que ajustarse para tener en cuenta la diferencia en la escala. Los coeficientes logit se pueden dividir por $\sqrt{1.6}$, de manera que la varianza del error sea 1, al igual que en el modelo probit. Con este ajuste, los coeficientes comparables pasan a ser -0.43 para el modelo logit y -0.45 para el modelo probit. El modelo logit implica una menor sensibilidad al precio que el probit. Alternativamente, los coeficientes probit podrían ser convertidos a la escala de los coeficientes logit multiplicándolos por $\sqrt{1.6}$, en cuyo caso los coeficientes comparables serían -0.55 para logit y -0.57 para probit.

Un problema de interpretación similar surge cuando el mismo modelo se estima en diferentes conjuntos de datos. La escala relativa de las estimaciones de los dos conjuntos de datos refleja la variación de los factores no observados entre los conjuntos de datos. Supongamos que los modelos de elección del medio de transporte fueron estimados en Chicago y Boston. Para Chicago, el coeficiente de costo estimado es de -0.55 y el coeficiente de tiempo es -1.78. Para Boston, las estimaciones son -0.81 y -2.69 respectivamente. El ratio entre el coeficiente de costo y el coeficiente de tiempo es muy similar en las dos ciudades: 0.309 en Chicago y 0.301 en Boston. Sin embargo, la magnitud de los coeficientes es el cincuenta por ciento más alto para Boston que para Chicago. Esta diferencia de escala significa que la parte no observada de la utilidad tiene menos variación en Boston que en Chicago: dado que los coeficientes se dividen por la desviación estándar de la parte no observada de la utilidad, los coeficientes más bajos significan mayor desviación estándar y por lo tanto mayor varianza. Los modelos están revelando que otros factores distintos de tiempo y costo tienen menos efecto en la gente de Boston que en la de Chicago. Dicho de forma más intuitiva, tiempo y costo tienen más importancia, en relación a factores no observados, en Boston que en Chicago, lo cual es consistente con la mayor escala de los coeficientes para Boston.

Normalización con errores heterocedásticos

En algunas situaciones, la varianza de los términos de error puede ser diferente para diferentes segmentos de la población. El investigador no puede establecer el nivel global de utilidad mediante la

normalización de la varianza de los errores para todos los segmentos, ya que la variación es diferente en los distintos segmentos. En lugar de ello, el investigador establece la escala global de utilidad mediante la normalización de la varianza para un segmento y luego calcula la varianza (y por lo tanto la escala) para cada segmento en relación con este primer segmento.

Por ejemplo, considere la situación descrita en el apartado anterior, donde los factores no observados tienen mayor varianza en Chicago que en Boston. Si se estiman modelos por separado en Chicago y en Boston, la varianza del término de error queda normalizada por separado. La escala de los parámetros de cada modelo refleja la variación de los factores no incluidos en ese área. Supongamos, sin embargo, que el investigador desea estimar un modelo de datos agregado para Chicago y Boston. El investigador no puede normalizar la varianza de los factores no observados a la misma cantidad para todos los viajeros, ya que la varianza es diferente para los viajeros de Boston y de Chicago. En lugar de ello, el investigador establece la escala global de utilidad mediante la normalización de la varianza en una zona (por ejemplo Boston) y luego calcula la varianza en la otra zona respecto a la primera zona (varianza en Chicago relativa a la de Boston).

El modelo en su forma original es

$$U_{nj} = \alpha T_{nj} + \beta M_{nj} + \varepsilon_{nj}^{B} \; \forall n \; en \; Boston$$

$$U_{nj} = \alpha T_{nj} + \beta M_{nj} + \varepsilon_{nj}^{C} \; \forall n \; en \; Chicago,$$

donde ahora la varianza de ε_{nj}^{B} no es igual a la varianza de ε_{nj}^{C}. Etiquetemos el cociente de varianzas como $k = Var(\varepsilon_{nj}^{C})/\mathrm{Var}(\varepsilon_{nj}^{B})$. Ahora podemos dividir la utilidad para los viajeros de Chicago por $\sqrt{\mathrm{k}}$; por supuesto, esta división no afecta a sus elecciones, ya que la escala de la utilidad no importa. Sin embargo, esto nos permite reescribir el modelo como

$$U_{nj} = \alpha T_{nj} + \beta M_{nj} + \varepsilon_{nj} \; \forall n \; en \; Boston$$

$$U_{nj} = (\alpha/\sqrt{k})T_{nj} + (\beta/\sqrt{k})M_{nj} + \varepsilon_{nj} \; \forall n \; en \; Chicago,$$

donde ahora la varianza de ε_{nj} es la misma para todo n en ambas ciudades, ya que $Var(\varepsilon_{nj}^{C}/\sqrt{\mathrm{k}}) = (1/k)Var(\varepsilon_{nj}^{C}) = [Var(\varepsilon_{nj}^{B})/Var(\varepsilon_{nj}^{C})]Var(\varepsilon_{nj}^{C}) = Var(\varepsilon_{nj}^{B})$. La escala de la utilidad queda establecida mediante la normalización de la varianza de ε_{nj}. El parámetro k, que a menudo se llama parámetro de escala, se estima junto con β y α. El valor estimado $\hat{\mathrm{k}}$ de k informa al investigador sobre la varianza de los factores no observados en Chicago respecto a Boston. Por ejemplo, $\hat{\mathrm{k}} = 1.2$ implica que la varianza de los factores no observados es el veinte por ciento mayor en Chicago que en Boston.

La varianza del término de error puede ser diferente en distintas regiones geográficas, conjuntos de datos, tiempo u otros factores. En todos los casos, el investigador establece la escala global de utilidad normalizando una de las varianzas y estimando luego las otras varianzas relativas a la varianza normalizada. Swait y Louviere (1993) han estudiado el papel del parámetro de escala en los modelos de elección discreta, describiendo la variedad de razones por las que la varianza puede diferir entre observaciones. Asimismo, dependiendo de la situación de elección y de la interpretación que hace el investigador de la situación, pueden entrar en juego factores psicológicos del mismo modo que el concepto tradicional de varianza de factores no observados. Por ejemplo, Bradley y Daly (1994) permiten que el parámetro de escala varíe entre experimentos de preferencia declarada con el fin de permitir que entre en el modelo el efecto de la fatiga de los encuestados al responder las preguntas de

la encuesta. Ben-Akiva y Morikawa (1990) permiten que el parámetro de escala difiera entre las intenciones declaradas por los respondientes y sus elecciones reales de mercado.

Normalización con errores correlacionados

En la explicación previa hemos asumido que ε_{nj} es independiente entre alternativas. Cuando los errores están correlacionados entre las alternativas, la normalización de la escala es más compleja. Hasta ahora hemos hablado de fijar la escala de utilidad, sin embargo, dado que sólo las diferencias en utilidad importan, es más apropiado hablar en términos de ajuste de la escala de las *diferencias* de utilidad. Cuando los errores están correlacionados, la normalización de la varianza del error de una alternativa no es suficiente para establecer la escala de las diferencias de utilidad.

El problema se describe más fácilmente a través de un ejemplo con cuatro alternativas La utilidad para las cuatro alternativas es $U_{nj} = V_{nj} + \varepsilon_{nj}$, j $= 1, \dots, 4$. El vector de error $\varepsilon_n = \langle \varepsilon_{n1}, \dots, \varepsilon_{n4} \rangle$ tiene media cero y matriz de covarianza

(2.3)
$$\Omega = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \cdot & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \cdot & \cdot & \sigma_{33} & \sigma_{34} \\ \cdot & \cdot & \cdot & \sigma_{44} \end{pmatrix},$$

donde los puntos se refieren a los elementos correspondientes en la parte superior de la matriz simétrica.

Dado que sólo las diferencias en la utilidad cuentan, este modelo es equivalente a otro en el que todas las utilidades estén expresadas como la diferencia respecto a la primera alternativa, por ejemplo. El modelo equivalente es $\tilde{U}_{nj1} = \tilde{V}_{nj1} + \tilde{\varepsilon}_{nj1}$ para j $= 2,3,4$, donde $\tilde{U}_{nj1} = U_{nj} - U_{n1}, \tilde{V}_{nj1} = V_{nj} - V_{n1}$ y el vector de las diferencias de error es $\tilde{\varepsilon}_{n1} = \langle (\varepsilon_{n2} - \varepsilon_{n1}), (\varepsilon_{n3} - \varepsilon_{n1}), (\varepsilon_{n4} - \varepsilon_{n1}) \rangle$. La varianza de cada diferencia de error depende de las varianzas y covarianzas de los errores originales. Por ejemplo, la varianza de la diferencia entre el primer y el segundo error es $\mathrm{Var}(\tilde{\varepsilon}_{n21}) = \mathrm{Var}(\varepsilon_{n2} - \varepsilon_{n1}) = \mathrm{Var}(\varepsilon_{n1}) + \mathrm{Var}(\varepsilon_{n2}) - 2\mathrm{Cov}(\varepsilon_{n1}, \varepsilon_{n2}) = \sigma_{11} + \sigma_{22} - 2\sigma_{12}$. Podemos calcular de forma similar la covarianza entre $\tilde{\varepsilon}_{n21}$, que es la diferencia entre el primer y el segundo error , y $\tilde{\varepsilon}_{n31}$, que es la diferencia entre el primer y el tercer error: $\mathrm{Cov}(\tilde{\varepsilon}_{n21}, \tilde{\varepsilon}_{n31}) = E(\varepsilon_{n2} - \varepsilon_{n1})(\varepsilon_{n3} - \varepsilon_{n1}) = \mathrm{E}(\varepsilon_{n2}\varepsilon_{n3} - \varepsilon_{n2}\varepsilon_{n1} - \varepsilon_{n3}\varepsilon_{n1} + \varepsilon_{n1}\varepsilon_{n1}) = \sigma_{23} - \sigma_{21} - \sigma_{31} + \sigma_{11}$. La matriz de covarianza para el vector de las diferencias de error se convierte en

$$\tilde{\Omega}_1 = \begin{pmatrix} \sigma_{11} + \sigma_{22} - 2\sigma_{12} & \sigma_{11} + \sigma_{23} - \sigma_{12} - \sigma_{13} & \sigma_{11} + \sigma_{24} - \sigma_{12} - \sigma_{14} \\ . & \sigma_{11} + \sigma_{33} - 2\sigma_{13} & \sigma_{11} + \sigma_{34} - \sigma_{13} - \sigma_{14} \\ . & . & \sigma_{11} + \sigma_{44} - 2\sigma_{14} \end{pmatrix}.$$

Ajustar la varianza de uno de los errores originales no es suficiente para establecer la varianza de las diferencias de error. Por ejemplo, si la varianza de la primera alternativa se establece a un número $\sigma_{11} = k$, la varianza de la diferencia entre los errores de las dos primeras alternativas se convierte en $\mathrm{k} + \sigma_{22} - 2\sigma_{12}$. Un número infinito de valores para σ_{22} y σ_{12} conducen al mismo valor de la diferencia $\sigma_{22} - 2\sigma_{12}$, generando modelos equivalentes.

Una forma habitual para establecer la escala de la utilidad cuando los errores no son iid es normalizar la varianza de una de las diferencias de error a algún número. Ajustar la varianza de una diferencia de error establece la escala de las diferencias de utilidad y por tanto, de la utilidad. Supongamos que normalizamos la varianza de $\tilde{\varepsilon}_{n21}$ a 1 . La matriz de covarianza para las diferencias de error, expresada en términos de las covarianzas de los errores originales, se convierte en

$$(2.4) \qquad \begin{pmatrix} 1 & (\sigma_{11}+\sigma_{23}-\sigma_{12}-\sigma_{13})/m & (\sigma_{11}+\sigma_{24}-\sigma_{12}-\sigma_{14})/m \\ . & (\sigma_{11}+\sigma_{33}-2\sigma_{13})/m & (\sigma_{11}+\sigma_{34}-\sigma_{13}-\sigma_{14})/m \\ . & . & (\sigma_{11}+\sigma_{44}-2\sigma_{14})/m \end{pmatrix},$$

donde $m = \sigma_{11} + \sigma_{22} - 2\sigma_{12}$. De esta forma, la utilidad queda dividida por $\sqrt{\sigma_{11} + \sigma_{22} - 2\sigma_{12}}$ para obtener esta escala.

Nótese que cuando los términos de error son iid, la normalización de la varianza de uno de estos errores normaliza de forma automática la varianza de las diferencias de los errores. Con errores iid, $\sigma_{\mathrm{jj}} = \sigma_{\mathrm{ii}}$ y $\sigma_{\mathrm{ij}} = 0$ para i ≠ j. Por lo tanto, si σ_{11} se normaliza a k, la varianza de la diferencia de error se convierte en $\sigma_{11} + \sigma_{22} - 2\sigma_{12} = k + k - 0 = 2k$. La varianza de la diferencia de error queda efectivamente normalizada, los mismo que sucede con errores no-iid.

La normalización tiene implicaciones en el número de parámetros que pueden estimarse en la matriz de covarianza. La covarianza de los errores originales Ω en la ecuación (2.3), cuenta con diez elementos para nuestro ejemplo con cuatro alternativas. Sin embargo, la matriz de covarianza de las diferencias de error tiene seis elementos, uno de los cuales se normaliza para establecer la escala de las diferencias de utilidad. La matriz de covarianza para las diferencias de error con la varianza de la primera diferencia de error normalizada a k toma la forma

$$(2.5) \qquad \tilde{\Omega}_1^* = \begin{pmatrix} k & \omega_{ab} & \omega_{ac} \\ \cdot & \omega_{bb} & \omega_{bc} \\ \cdot & \cdot & \omega_{cc} \end{pmatrix},$$

que sólo tiene cinco parámetros. Al reconocer que sólo las diferencias de utilidad son importantes y que la escala de utilidad es arbitraria, el número de parámetros de covarianza cae de diez a cinco. Un modelo con J alternativas tiene como máximo $J(J-1)/2 - 1$ parámetros de covarianza después de la normalización.

La interpretación del modelo se ve afectada por la normalización. Supongamos por ejemplo que se han estimado los elementos de la matriz (2.5). El parámetro ω_{bb} es la varianza de la diferencia entre los errores de la primera y la tercera alternativa, en relación a la varianza de la diferencia entre los errores de la primera y la segunda alternativa. Para complicar aún más la interpretación, la varianza de la diferencia entre los errores de dos alternativas refleja las varianzas de ambos así como su covarianza.

Como veremos, la normalización de los modelos logit y logit jerárquicos es automática con los supuestos de distribución que asumen para los términos de error. La interpretación bajo estos supuestos es relativamente sencilla. Para logit mixto y probit asumimos menor número de hipótesis sobre la distribución de los términos de error, por lo que la normalización no es automática. El investigador debe tener en cuenta las cuestiones de normalización al especificar e interpretar un modelo. Volveremos a este tema cuando tratemos cada modelo de elección discreta en los capítulos siguientes

2.6 Agregación

Los modelos de elección discreta operan a nivel de decisores individuales. Sin embargo, el investigador suele estar interesado en alguna medida agregada, como la probabilidad promedio dentro de una población o la respuesta media a un cambio en algunos de los factores.

En los modelos de regresión lineal, las estimaciones de los valores agregados de la variable dependiente se obtienen mediante la inserción en el modelo de los valores agregados de las variables explicativas. Por ejemplo, supongamos que h_n son los gastos en vivienda para una persona n, y_n es el ingreso de esa persona y el modelo que relaciona ambos datos es $h_n = \alpha + \beta y_n$. Puesto que este modelo es lineal, el

gasto medio en vivienda se calcula simplemente como $\alpha + \beta\bar{y}$, donde $\bar{y}$ es el ingreso medio. Del mismo modo, el promedio de respuesta a un cambio en el ingreso de una unidad es simplemente β, ya que β es la respuesta para cada persona.

Los modelos de elección discreta no son lineales en las variables explicativas y en consecuencia, la introducción de los valores agregados de las variables explicativas en los modelos no proporciona una estimación objetiva de la probabilidad media o de la respuesta media. Este hecho se puede constatar de forma muy visual. Considere la figura 2.1, que muestra las probabilidades de elegir una alternativa concreta para dos personas, cuya parte observada de la utilidad (utilidad representativa) es a y b respectivamente. La probabilidad promedio es el promedio de las probabilidades de las dos personas, es decir, $(P_a + P_b)/2$. La utilidad representativa media es $(a + b)/2$ y la probabilidad evaluada en este promedio es el punto de la curva de probabilidad encima de $(a + b)/2$. Como se observa en este caso, la probabilidad media es mayor que la probabilidad evaluada en la utilidad representativa media. En general, la probabilidad evaluada en la utilidad representativa media subestima la probabilidad media cuando las probabilidades de elección de los individuos son bajas y la sobreestima cuando son altas.

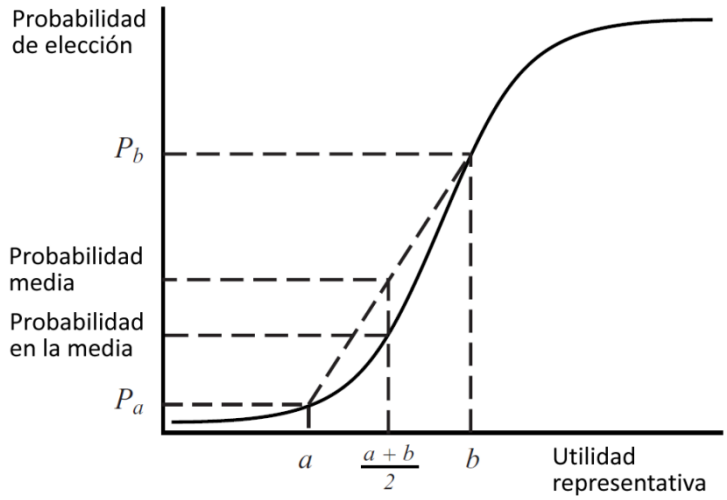


Figura 2.1: Diferencia entre la probabilidad media y la probabilidad calculada en la utilidad representativa media.

Estimar la respuesta promedio mediante el cálculo de derivadas y elasticidades en el promedio de las variables explicativas es igualmente problemático. Considere la figura 2.2, que representa a dos personas con utilidades representativas a y b. La derivada de la probabilidad de elección para un cambio en la utilidad representativa para estas dos personas es pequeña (la pendiente de la curva en a y b). En consecuencia, la derivada promedio también es pequeña. Sin embargo, la derivada en la utilidad representativa media es muy grande (la pendiente en el valor $(a + b)/2$). Estimar la respuesta media de esta manera puede ser tremendamente engañoso. De hecho, Talvitie (1976) encontró, en una situación de elección, que las elasticidades en la utilidad representativa media pueden ser hasta dos o tres veces mayores o menores que el promedio de las elasticidades individuales.

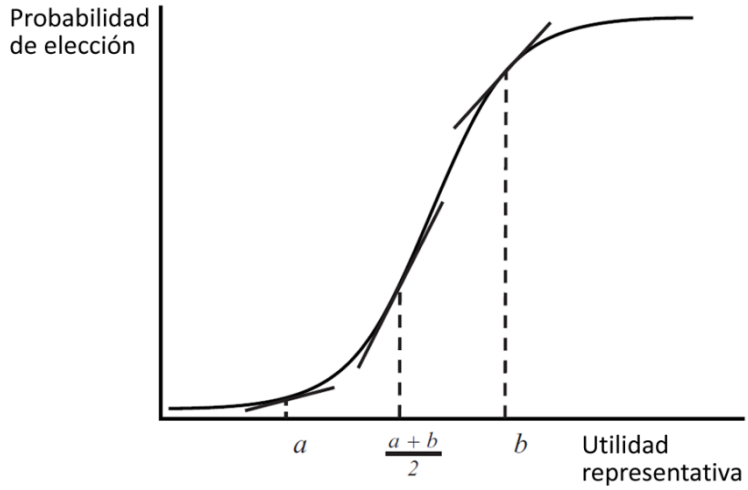


Figura 2.2: Diferencia entre la respuesta media y la respuesta calculada en la utilidad representativa media.

Las variables resultado agregadas de modelos de elección discreta se pueden obtener consistentemente de dos maneras: mediante la enumeración de la muestra o mediante segmentación. Tratamos cada enfoque en las siguientes secciones.

2.6.1 Enumeración de la muestra

La aproximación más directa y la más popular con mucha diferencia, es la enumeración de la muestra, mediante la cual las probabilidades de elección de cada decisor en la muestra se suman, o se promedian, sobre el total de decisores. Considere un modelo de elección discreta que otorga una probabilidad P_{ni} de que el decisor n elegirá la alternativa i entre un conjunto de alternativas. Suponga que una muestra de decisores N, etiquetados $n = 1, \dots, N$, se extrae de la población para la cual se desea calcular estadísticos agregados. (Esta muestra podría ser la misma muestra sobre la que se estimó el modelo. Sin embargo, también podría ser una muestra diferente, recogida en un área diferente o en una fecha posterior a la de la muestra de estimación). Cada decisor de la muestra n tiene un cierto peso asociado w_n que representa el número de decisores similares a él en la población. Para muestras con base a factores exógenos, este peso es el inverso de la probabilidad de que el decisor haya sido seleccionado para la muestra. Si la muestra es puramente aleatoria, w_n es igual para todo n; y si la muestra se estratificó al azar, w_n es el mismo para todos los n dentro de un estrato.

Una estimación consistente del número total de decisores en la población que eligen la alternativa i, etiquetada $\widehat{N}_i$, es simplemente la suma ponderada de las probabilidades individuales:

$$\widehat{N}_i = \sum_n w_n P_{ni}$$

La probabilidad media, que es la cuota de mercado estimada, es $\widehat{N}_i/N$. Las derivadas y elasticidades medias se obtienen de forma similar calculando la derivada y la elasticidad para cada persona muestreada y calculando el promedio ponderado.

2.6.2 Segmentación

Cuando el número de variables explicativas es pequeño y esas variables toman sólo unos pocos valores, es posible estimar los resultados agregados sin utilizar una muestra de decisores. Consideremos, por ejemplo, un modelo con sólo dos variables formando parte de la utilidad representativa de cada alternativa: el nivel educativo y el sexo. Supongamos que la variable educación se compone de cuatro categorías: (1) no completó la escuela secundaria, (2) ha terminado la escuela secundaria pero no asistió a la universidad, (3) ha asistido a la universidad pero no recibió el título y (4) recibió un título universitario. El número total de los diferentes tipos de decisores (llamados segmentos) es ocho: los cuatro niveles de educación por cada uno de los dos sexos. Las probabilidades de elección varían sólo entre estos ocho segmentos y no entre individuos dentro de cada segmento.

Si el investigador tiene datos sobre el número de personas en cada segmento, los resultados agregados se pueden estimar mediante el cálculo de la probabilidad de elección de cada segmento y calculando la suma ponderada de estas probabilidades. El número de personas que se estima que eligen la alternativa i es

$$\widehat{N}_i = \sum_{s=1}^{8} w_s P_{si},$$

donde P_{si} es la probabilidad de que un decisor del segmento s escoja la alternativa i y w_s es el número de decisores en el segmento s.

2.7 Predicción

Para hacer pronósticos en algún instante futuro se aplican los procedimientos descritos anteriormente para las variables agregadas. Sin embargo, las variables exógenas y/o los pesos son ajustados para reflejar los cambios que se anticipan en el tiempo. Si usamos la enumeración de la muestra, la muestra se ajusta de manera que se parezca a cómo será una muestra extraída en el futuro. Por ejemplo, para pronosticar el número de personas que van a elegir una determinada alternativa dentro de cinco años, una muestra extraída en el año en curso se ajusta para reflejar los cambios en los factores socioeconómicos y de otra índole que se espera que ocurran en los próximos cinco años. La muestra se ajusta (1) cambiando el valor de las variables asociadas a cada decisor en la muestra (por ejemplo, aumentando los ingresos de cada decisor para representar el crecimiento de los ingresos reales en el tiempo) y/o (2) cambiando la ponderación asignada a cada decisor para reflejar los cambios en el número de decisores en la población que son similares al decisor de la muestra (por ejemplo, aumentando el peso de los hogares unipersonales y disminuyendo los pesos para familias numerosas para reflejar disminuciones esperadas en el tamaño del hogar con el paso del tiempo).

Para ajustar el enfoque de segmentación, los cambios en el tiempo de las variables explicativas son representados por los cambios en el número de decisores en cada segmento. Lógicamente, las mismas variables explicativas no pueden ser ajustadas, ya que los distintos valores de las variables explicativas definen los segmentos. Cambiar las variables asociadas a un decisor en un segmento simplemente desplaza al decisor a otro segmento.

2.8 Recalibración de constantes

Como se describe en la Sección 2.5.1, a menudo se incluyen constantes específicas de alternativa en un modelo para capturar el efecto promedio de los factores no observados. En la realización de predicciones, suele ser útil ajustar estas constantes para reflejar el hecho de que los factores no observados son diferentes para el área o año pronosticados en comparación con la muestra empleada en la estimación. Los datos de cuota de mercado para el ámbito sobre el que hacemos la previsión se

pueden utilizar para *recalibrar* las constantes adecuadamente. El modelo recalibrado se puede utilizar para predecir cambios en las cuotas de mercado debidos a cambios en los factores explicativos.

Para recalibrar las constantes se utiliza un proceso iterativo. Sea α_j^0 la constante específica de alternativa para la alternativa j . El superíndice 0 se utiliza para indicar que estos son los valores de inicio en el proceso iterativo. Sea S_j la cuota de mercado de decisores en el ámbito de pronóstico que eligen la alternativa j en el año *base* (por lo general, el último año del que se dispone de esos datos). Utilizando el modelo de elección discreta con sus valores originales de $\alpha_j^0\ \forall j$, predecimos la cuota de decisores en el ámbito de pronóstico que elegirá cada alternativa. Etiquetamos estas predicciones como $\hat{S}_j^0\ \forall j$. Comparamos las cuotas de mercado previstas con las cuotas reales. Si el porcentaje real de una alternativa supera la cuota prevista, elevamos la constante de esa alternativa. Si por el contrario la cuota real es inferior a la prevista, bajamos la constante. Un ajuste eficaz es

$$\alpha_j^1 = \alpha_j^0 + ln(S_j/\hat{S}_j{}^0)$$

Con las nuevas constantes, predecimos la cuota de nuevo, comparamos con las cuotas reales y, si es necesario, ajustamos las constantes de nuevo. El proceso se repite hasta que las cuotas previstas estén suficientemente cerca de las cuotas reales. El modelo con estas constantes recalibradas se puede utilizar para predecir los cambios en las cuotas de mercado que se producirán partiendo del año base debido a cambios en los factores observados que afecten a las elecciones de los decisores.

3

Logit

3.1 Probabilidades de elección

Logit es con diferencia el modelo de elección discreta más simple y de uso más extendido. Su popularidad se debe al hecho de que la fórmula para las probabilidades de elección tiene una expresión cerrada y es fácilmente interpretable. Originalmente, la fórmula logit fue obtenida por Luce (1959) a partir de ciertas asunciones sobre las características de las probabilidades de elección, la principal de las cuales era la independencia de alternativas irrelevantes (*independence from irrelevant alternatives*, IIA), tratada en la sección 3.3.2. Marschak (1960) mostró que estos axiomas implicaban que el modelo era consistente con un comportamiento del decisor orientado a la maximización de la utilidad. La relación de la fórmula logit con la distribución de la utilidad no observada (como opuesta a las características de las probabilidades de elección) fue desarrollada por Marley, tal y como citan Luce y Suppes (1965), quienes mostraron que la distribución de valor extremo conduce a la fórmula logit. McFadden (1974) completó el análisis mostrando la relación inversa, es decir, que la fórmula logit para las probabilidades de elección necesariamente implica que la utilidad no observada se distribuye de acuerdo a una distribución de valor extremo. En la ceremonia de entrega de su premio Nobel, McFadden (2001) explicó una historia fascinante sobre el desarrollo de este modelo pionero.

Para obtener el modelo logit, usamos la notación general del Capítulo 2 y añadimos una distribución específica para la utilidad no observada. Un decisor etiquetado como n se enfrenta a J alternativas. La utilidad que el decisor obtiene de la alternativa j se descompone en (1) una parte denominada V_{nj} que es conocida por el investigador a través de algunos parámetros, y (2) una parte ε_{nj} desconocida que es tratada por el investigador como una variable aleatoria: $U_{nj} = V_{nj} + \varepsilon_{nj}\ \forall j$. El modelo logit se obtiene suponiendo que cada ε_{nj} se distribuye independientemente y de forma idénticamente distribuida de acuerdo a una densidad de probabilidad de tipo valor extremo. Esta distribución también se denomina Gumbel y tipo I valor extremo (y en ocasiones, de forma errónea, Weibull). La densidad para cada componente no observado de utilidad es

$$f(\varepsilon_{nj}) = e^{-\varepsilon_{nj}} e^{-e^{-\varepsilon_{nj}}}, \tag{3.1}$$

y la distribución acumulativa es

$$F(\varepsilon_{nj}) = e^{-e^{-\varepsilon_{nj}}}. \tag{3.2}$$

La varianza de esta distribución es $\pi^2/6$. Asumiendo que la varianza es $\pi^2/6$, implícitamente estamos normalizando la escala de la utilidad, como se trata en la sección 2.5. Volveremos sobre este tema y su relevancia para la interpretación del modelo en la siguiente sección. La media de la distribución de valor extremo no es cero, sin embargo, la media es irrelevante ya que sólo las diferencias entre utilidades importan (véase el capítulo 2) y la diferencia entre dos términos aleatorios que tienen la misma media tiene una media igual a cero.

La diferencia entre dos variables de tipo valor extremo se distribuye de forma logística. Es decir, si ε_{nj} y ε_{ni} son de tipo valor extremo iid, entonces $\varepsilon^*_{nji} = \varepsilon_{nj} - \varepsilon_{ni}$ sigue una distribución logística

$$\boldsymbol{F}(\varepsilon^*_{nji}) = \frac{\boldsymbol{e}^{\varepsilon^*_{nji}}}{\mathbf{1}+\boldsymbol{e}^{\varepsilon^*_{nji}}}. \tag{3.3}$$

Esta fórmula se utiliza en ocasiones para describir modelos logit binarios, es decir, modelos con dos alternativas. Usar la distribución de valor extremo para los errores (y por lo tanto la distribución logística para las diferencias entre errores) es casi lo mismo que asumir que los errores se distribuyen normalmente y de forma independiente. La distribución de valor extremo tiene colas ligeramente más gruesas que una distribución normal, lo que implica que permite un comportamiento ligeramente más aberrante que la normal. Por lo general, sin embargo, la diferencia entre errores distribuidos según el valor extremo y según distribuciones normales independientes es indistinguible empíricamente.

El supuesto clave del modelo no es tanto la forma de la distribución como que los errores sean independientes entre sí. Esta independencia significa que la parte no observada de la utilidad de una alternativa no está relacionada con la parte no observada de la utilidad de otra alternativa. Es un supuesto bastante restrictivo, por lo que el desarrollo de otros modelos como los descritos en los capítulos 4-6 ha surgido en gran medida con el fin de evitar este supuesto y permitir la existencia de errores correlacionados.

No obstante, es importante observar que la hipótesis de independencia no es tan restrictiva como podría parecer a primera vista y que, de hecho, puede ser interpretada como un resultado natural de un modelo bien especificado. Recuerde del capítulo 2 que ε_{nj} se define como la diferencia entre la utilidad que el decisor obtiene realmente, U_{nj}, y la representación de la utilidad que el investigador ha desarrollado utilizando las variables observadas, V_{nj}. Como tal, ε_{nj} y su distribución dependen de la especificación que el investigador haga de la utilidad representativa; no está definida por la situación de elección *per se*. En este sentido, la hipótesis de independencia permite una interpretación diferente. Bajo la hipótesis de independencia, el error de una alternativa no proporciona al investigador ninguna información sobre el error de otra alternativa diferente. Dicho de forma equivalente, el investigador ha especificado V_{nj} lo suficiente para que el resto de la utilidad (no observada) sea esencialmente "ruido blanco". En un sentido profundo, el objetivo último del investigador es representar la utilidad tan bien que los únicos aspectos que queden sin representar constituyan simplemente ruido blanco; es decir, el objetivo es especificar la utilidad suficientemente bien como para que un modelo logit sea apropiado. Visto de esta forma, logit es el modelo ideal en lugar de una restricción.

Si el investigador considera que la parte no observada de la utilidad está correlacionada entre alternativas dada su especificación de la utilidad representativa, tiene tres opciones: (1) utilizar un modelo diferente que permita errores correlacionados, tales como los descritos en los capítulos 4-6, (2) especificar de nuevo la utilidad representativa de forma que la fuente de correlación quede capturada explícitamente y por lo tanto los errores restantes sean independientes o (3) utilizar el modelo logit bajo la especificación actual de la utilidad representativa y considerar el modelo como una aproximación. La viabilidad de la última opción depende, por supuesto, de los objetivos de la investigación. Las violaciones de los supuestos del modelo logit parecen tener menos efecto en la estimación de las

preferencias medias que en el pronóstico de patrones de sustitución (cambio en las preferencias futuras al alterar los atributos de las alternativas). Estas cuestiones se tratan en las siguientes secciones.

Derivamos a continuación las probabilidades de elección logit, siguiendo la aproximación de McFadden (1974). La probabilidad de que el decisor n elija la alternativa i es

$$P_{ni} = Prob(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \ \ \forall j \neq i)$$

$$= Prob\left(\varepsilon_{nj} < \varepsilon_{ni} + V_{ni} - V_{nj} \ \ \forall j \neq i\right). \tag{3.4}$$

Si consideramos ε_{ni} como dado, esta expresión es la distribución acumulativa para cada ε_{nj} evaluada en $\varepsilon_{ni} + V_{ni} - V_{nj}$, que de acuerdo con (3.2) es $exp(-exp(-(\varepsilon_{ni} + V_{ni} - V_{nj})))$. Dado que los εs son independientes, esta distribución acumulativa sobre todo $j \neq i$ es el producto de las distribuciones acumulativas individuales:

$$P_{ni}|\varepsilon_{ni} = \prod_{j \neq i} e^{-e^{-\left(\varepsilon_{ni} + V_{ni} - V_{nj}\right)}}.$$

Por supuesto, ε_{ni} no está realmente dado, por lo que la probabilidad de elección es la integral de $\mathrm{P}_{ni}|\varepsilon_{ni}$ sobre todos los valores de ε_{ni} ponderados por su densidad de probabilidad (3.1):

$$P_{ni} = \int \left(\prod_{j \neq i} e^{-e^{-\left(\varepsilon_{ni} + V_{ni} - V_{nj}\right)}}\right) e^{-\varepsilon_{ni}}\, e^{-e^{-\varepsilon_{ni}}} d\varepsilon_{ni}. \tag{3.5}$$

Algunas manipulaciones algebraicas de esta integral resultan en una expresión cerrada y compacta:

$$P_{ni} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}, \tag{3.6}$$

que es la probabilidad de elección logit. El álgebra que obtiene (3.6) de (3.5) se detalla en la última sección de este capítulo.

La utilidad representativa suele especificarse de forma que sea lineal en relación a los parámetros: $V_{nj} = \beta' x_{nj}$, donde x_{nj} es un vector de variables observadas en la alternativa j. Con esta especificación, las probabilidades logit se convierten en

$$P_{ni} = \frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{nj}}}.$$

En condiciones bastante generales, cualquier función puede aproximarse de forma arbitrariamente precisa por una función lineal en parámetros. La asunción es bastante buena por lo tanto. Es importante destacar que McFadden (1974) demostró que la función logaritmo de la verosimilitud (log-verosimilitud o *log-likelihood*) con estas probabilidades de elección, es globalmente cóncava respecto a los parámetros β, lo cual ayuda en los procedimientos de maximización numérica (como se trata en el Capítulo 8). Numerosos paquetes de software contienen rutinas para la estimación de modelos logit con utilidad representativa lineal en parámetros.

Las probabilidades logit exhiben varias propiedades que son deseables. En primer lugar, P_{ni} está necesariamente entre cero y uno, un requisito para que pueda ser una probabilidad. Cuando V_{ni} crece, lo que refleja una mejora en los atributos observados de la alternativa, manteniendo $\mathrm{V}_{nj} \forall j \neq i$ constante, P_{ni} se acerca a uno. Y P_{ni} se acerca a cero cuando V_{ni} disminuye, ya que la exponencial en el

numerador de (3.6) se aproxima a cero a medida que V_{ni} se acerca a $-\infty$. La probabilidad logit para una alternativa nunca es exactamente cero. Si el investigador cree que una alternativa no tiene en realidad ninguna posibilidad de ser elegida por un decisor, puede excluir la alternativa del conjunto de elección. Una probabilidad exactamente igual a 1 se obtiene sólo si el conjunto de elección consiste en una única alternativa.

En segundo lugar, las probabilidades de elección de todas las alternativas suman uno: $\sum_{i=1}^{J} P_{ni} = \sum_i \exp(V_{ni}) / \sum_j \exp(V_{nj}) = 1$. El decisor necesariamente elige una de las alternativas. El denominador de (3.6) es simplemente la suma del numerador sobre todas las alternativas, lo que produce esta propiedad de suma de forma automática. Con logit, así como con algunos modelos un poco más complejos como el logit jerárquico descrito en el Capítulo 4, la interpretación de las probabilidades de elección se ve facilitada al observarse que el denominador sirve para asegurar que las probabilidades sumen uno. En otros modelos, como el logit mixto y el probit, no hay un denominador *per se* que pueda ser interpretado de esta manera.

La relación entre la probabilidad logit y la utilidad representativa es una sigmoidea o función-S, tal y como se muestra en la figura 3.1.

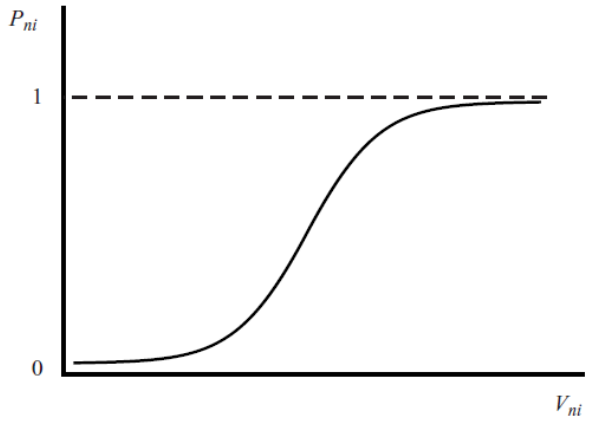


Figura 3.1. Gráfica de la curva logit.

Esta forma tiene implicaciones en el impacto que producen los cambios de las variables explicativas. Si la utilidad representativa de una alternativa es muy baja en comparación con otras alternativas, un pequeño aumento en la utilidad de la alternativa tiene poco efecto sobre la probabilidad de que sea elegida: las otras alternativas son todavía suficientemente mejores por lo que esta pequeña mejora no ayuda mucho. Del mismo modo, si una alternativa es muy superior a las demás en los atributos observados, un aumento adicional en su utilidad representativa tiene poco efecto sobre la probabilidad de elección. El punto en el que el aumento en la utilidad representativa tiene mayor efecto sobre la probabilidad de ser elegida es cuando la probabilidad es próxima a 0.5, lo que significa una probabilidad del 50% de que la alternativa sea elegida. En este caso, una pequeña mejora es decisiva en las elecciones de las personas, induciendo un gran cambio en la probabilidad. La forma sigmoidea de las probabilidades logit se observa en la mayoría de los modelos de elección discreta y tiene implicaciones importantes para los reguladores y responsables de legislar. Por ejemplo, mejorar el servicio de autobuses en zonas en las que el servicio es tan pobre que pocos viajeros viajan en autobús sería menos eficaz, en términos de uso del transporte público, que hacer la misma mejora en áreas donde el servicio de autobús es suficientemente bueno como para que una parte moderada de viajeros ya lo esté usando (pero no tan bueno como para que casi todo el mundo lo esté usando).

La fórmula de probabilidad logit es fácilmente interpretable en el contexto de un ejemplo. Considere en primer lugar una situación de elección binaria: la elección de un hogar entre un sistema de calefacción eléctrica o de gas. Supongamos que la utilidad que obtiene el hogar de cada tipo de sistema depende

sólo del precio de compra, el costo anual de operación, el punto de vista del hogar respecto a la conveniencia y calidad de calentarse con cada tipo de sistema, así como la estética de los sistemas dentro de la casa. Los dos primeros factores pueden ser observados por el investigador, pero no el resto de factores. Si el investigador considera que la parte observada de la utilidad es una función lineal de los factores observados, la utilidad de cada sistema de calefacción se puede expresar como: $\mathrm{U}_g = \beta_1 PP_g + \beta_2 OC_g + \varepsilon_g$ y $\mathrm{U}_e = \beta_1 PP_e + \beta_2 OC_e + \varepsilon_e$, donde el subíndice g y e se refieren a gas y electricidad, PP y OC son el precio de compra (*purchase price*) y los costos de operación (*operating cost*), β_1 y β_2 son parámetros escalares y el subíndice n para el hogar se suprime para simplificar la notación. Dado que costos más altos implican menos dinero para gastar en otros bienes, esperamos que la utilidad disminuya a medida que el precio de compra o el costo de operación suban (con todo lo demás constante): $\beta_1 < 0$ y $\beta_2 < 0$.

El componente no observado de la utilidad de cada alternativa, ε_g y ε_e, varía en los hogares en función de cómo ve cada hogar la calidad, comodidad y estética de cada tipo de sistema. Si estos componentes no observados se distribuyen con una densidad valor extremo iid, entonces la probabilidad de que el hogar elija la calefacción de gas es

$$P_g = \frac{e^{\beta_1 PP_g + \beta_2 OC_g}}{e^{\beta_1 PP_g + \beta_2 OC_g} + e^{\beta_1 PP_e + \beta_2 OC_e}} \tag{3.7}$$

y la probabilidad de que elija la calefacción eléctrica es la misma pero con $\exp(\beta_1 PP_e + \beta_2 OC_e)$ como numerador. La probabilidad de elegir un sistema de gas disminuye cuando su precio de adquisición o costo de operación sube, permaneciendo igual el del sistema eléctrico (suponiendo que β_1 y β_2 son negativos, como es de esperar).

Como en la mayoría de los modelos de elección discreta, el ratio entre los coeficientes de este ejemplo tiene sentido económico. En particular, el ratio β_2/β_1 representa la disposición de la familia a pagar para obtener una reducción de costos de operación. Si β_1 se ha estimado con valor -0.20 y β_2 con valor -1.14, estas estimaciones implicarían que los hogares están dispuestos a pagar hasta (-1,14)/(-0,20)=5,70 dólares más por un sistema cuyo costo de operación anual sea un dólar menos. Esta relación se obtiene de la siguiente manera. Por definición, la disposición de un hogar a pagar por la reducción del costo de operación es el incremento en el precio de compra que mantiene constante la utilidad del hogar dada una reducción en los costos operativos. Tomamos la derivada total de la utilidad respecto al precio de compra y al costo de operación, y fijamos esta derivada igual a cero, por lo que la utilidad no cambia: $\mathrm{dU} = \beta_1 dPP + \beta_2 dOC = 0$. A continuación, resolvemos la ecuación para el cambio en el precio de compra que mantenga la utilidad constante (es decir, que satisfaga esta ecuación) frente a un cambio en los costos de operación: $\partial \mathrm{PP}/\partial \mathrm{OC} = -\beta_2/\beta_1$. El signo negativo indica que los dos cambios están en la dirección opuesta: para mantener constante la utilidad, el precio de compra se eleva cuando el costo de operación disminuye.

En esta situación de elección binaria, las probabilidades de elección se pueden expresar de forma más sucinta. Dividiendo el numerador y el denominador de (3.7) por el numerador, y reconociendo que $\exp(\mathrm{a})/\exp(\mathrm{b}) = \exp(\mathrm{a} - \mathrm{b})$, tenemos

$$P_g = \frac{1}{1 + e^{(\beta_1 PP_e + \beta_2 OC_e) - (\beta_1 PP_g + \beta_2 OC_g)}}.$$

En general, las probabilidades logit binarias con utilidades representativas V_{n1} y V_{n2} pueden ser rescritas como $\mathrm{P}_{n1} = 1/(1 + \exp(\mathrm{V}_{n2} - \mathrm{V}_{n1}))$ y $\mathrm{P}_{n2} = 1/(1 + \exp(\mathrm{V}_{n1} - \mathrm{V}_{n2}))$. Si sólo las características sociodemográficas del decisor, s_n, entran en el modelo y los coeficientes de estas variables sociodemográficas se normalizan a cero para la primera alternativa (como se describe en el Capítulo 2),

la probabilidad de elección de la primera alternativa es $\mathrm{P}_{n1} = 1/(1 + e^{\alpha' s_n})$, que es la forma que se utiliza en la mayoría de los libros de texto y manuales de informática para el logit binario.

La elección multinomial es una extensión simple. Supongamos que hay un tercer tipo de sistema de calefacción, por ejemplo con petróleo (*oil,* indicado con subíndice "o") como combustible. La utilidad del sistema de petróleo se especifica de la misma forma que para los sistemas de electricidad y gas: $\mathrm{U}_o = \beta_1 PP_o + \beta_2 OC_o + \varepsilon_o$. Con esta opción adicional disponible, la probabilidad de que el hogar elija un sistema de gas es

$$P_g = \frac{e^{\beta_1 PP_g + \beta_2 OC_g}}{e^{\beta_1 PP_g + \beta_2 OC_g} + e^{\beta_1 PP_e + \beta_2 OC_e} + e^{\beta_1 PP_o + \beta_2 OC_o}},$$

que es igual a (3.7), excepto un término adicional incluido en el denominador para representar el calentador de petróleo. Dado que el denominador pasa a ser mayor mientras que el numerador permanece igual, la probabilidad de elegir un sistema de gas es más pequeña cuando está disponible un sistema de petróleo que cuando no lo está, como sería de esperar en el mundo real.

3.2 El parámetro de escala

En el apartado anterior hemos obtenido la fórmula logit bajo el supuesto de que los factores no observados se distribuyen con densidad valor extremo y varianza $\pi^2/6$. Fijar la varianza a $\pi^2/6$ es equivalente a normalizar el modelo respecto a la escala de la utilidad, como se trata en la Sección 2.5. Para hacer estos conceptos más explícitos es útil mostrar el papel que la varianza de los factores no observados juega en los modelos logit.

En general, la utilidad se puede expresar como $\mathrm{U}_{nj}^* = V_{nj} + \varepsilon_{nj}^*$, donde la parte no observada de la utilidad tiene varianza $\sigma^2 \times (\pi^2/6)$. Es decir, la varianza es cualquier número re-expresado como un múltiplo de $\pi^2/6$. Como la escala de la utilidad es irrelevante para el comportamiento, la utilidad se puede dividir por σ sin cambiar el comportamiento. La utilidad pasa a ser $\mathrm{U}_{nj} = V_{nj}/\sigma + \varepsilon_{nj}$ donde $\varepsilon_{nj} = \varepsilon_{nj}^*/\sigma$. Ahora, la parte no observada de la utilidad tiene varianza $\pi^2/6$: $\mathrm{Var}(\varepsilon_{nj}) = \mathrm{Var}(\varepsilon_{nj}^*/\sigma) = (1/\sigma^2)\mathrm{Var}(\varepsilon_{nj}^*) = (1/\sigma^2) \times \sigma^2 \times (\pi^2/6) = \pi^2/6$. La probabilidad elección es

$$P_{ni} = \frac{e^{V_{ni}/\sigma}}{\sum_j e^{V_{nj}/\sigma}},$$

que es la misma fórmula de la ecuación (3.6) pero con la utilidad representativa dividida por σ. Si V_{nj} es lineal en los parámetros con coeficientes β^*, las probabilidades de elección pasan a ser

$$P_{ni} = \frac{e^{(\beta^*/\sigma)' x_{ni}}}{\sum_j e^{(\beta^*/\sigma)' x_{nj}}},$$

Cada uno de los coeficientes está escalado por un factor $1/\sigma$. El parámetro σ se denomina el *parámetro de escala*, ya que escala los coeficientes para reflejar la varianza de la parte no observada de la utilidad.

Sólo es posible estimar la relación β^*/σ; β^* y σ no pueden identificarse por separado. Por lo general, el modelo se expresa en su forma reducida, con $\beta = \beta^*/\sigma$, lo que genera la expresión logit estándar

$$P_{ni} = \frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{nj}}}.$$

Los parámetros β son los que se estiman en el modelo, pero para la interpretación es útil reconocer que estos parámetros estimados son en realidad las estimaciones de los coeficientes "originales" β^* divididos por el parámetro de escala σ. Los coeficientes estimados indican el efecto de cada variable observada en relación a la varianza de los factores no observados. Una variación mayor en los factores no observados conduce a estimar coeficientes más pequeños, incluso si los factores observados tienen el mismo efecto en la utilidad (es decir, mayor σ significa β inferiores incluso si β^* es el mismo).

El parámetro de escala no afecta al ratio entre dos coeficientes cualesquiera, dado que queda suprimido en la división; por ejemplo, $\beta_1/\beta_2 = (\beta_1^*/\sigma)/(\beta_2^*/\sigma) = \beta_1^*/\beta_2^*$, donde los subíndices se refieren al primer y segundo coeficientes. La disposición a pagar, valor del tiempo y otras medidas de tasas marginales de sustitución no se ven afectadas por el parámetro de escala. Sólo se ve afectada la interpretación de las magnitudes de todos los coeficientes.

Hasta ahora hemos supuesto que la varianza de los factores no observados es la misma para todos los decisores, ya que la misma σ se utiliza para todo n. Supongamos ahora que los factores no observados tienen mayor varianza para unos decisores que para otros. En la Sección 2.5 se trata una situación en la que la varianza de factores no observados es diferente en Boston y en Chicago. Denotemos la varianza para todos los decisores de Boston como $(\sigma^B)^2(\pi^2/6)$ y para los decisores de Chicago como$(\sigma^C)^2(\pi^2/6)$. El ratio de la varianza en Chicago respecto a la de Boston es $k = (\sigma^C/\sigma^B)^2$. Las probabilidades de elección para las personas de Boston pasan a ser

$$P_{ni} = \frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{nj}}},$$

y para la gente de Chicago

$$P_{ni} = \frac{e^{(\beta/\sqrt{k})' x_{ni}}}{\sum_j e^{(\beta/\sqrt{k})' x_{nj}}},$$

donde $\beta = \beta^*/\sigma^B$. El ratio de las varianzas k se puede estimar junto con los coeficientes β. Los valores de β estimados se interpretan en relación con la varianza de los factores no observados en Boston y la k estimada proporciona información sobre la varianza en Chicago relativa a la de Boston. Es posible obtener relaciones más complejas permitiendo que la varianza para una observación dependa de más factores. Además, a menudo es de esperar que los datos de diferentes conjuntos de datos puedan tener diferente varianza para factores no observados, dando un parámetro de escala diferente para cada conjunto de datos. Ben-Akiva y Morikawa (1990) y Swait y Louviere (1993) tratan estos temas y proporcionan más ejemplos.

3.3 Potencia y limitaciones de logit

Tres cuestiones permiten dilucidar la potencia que los modelos logit tienen para representar el comportamiento de elección, así como delimitar los límites de esta potencia. Estas cuestiones son: variación de preferencias (*taste variation*), patrones de sustitución y elecciones reiteradas a lo largo del tiempo.

La aplicabilidad de los modelos logit se puede resumir de la siguiente manera:

1. Logit puede representar la variación sistemática de la preferencia (es decir, la variación de preferencia que se relaciona con las características observadas del decisor) pero no la variación de preferencia aleatoria (diferencias en las preferencias que no pueden vincularse a las características observadas).

2. El modelo logit implica sustitución proporcional entre alternativas, dada la especificación de la utilidad representativa hecha por el investigador. Para capturar formas más flexibles de sustitución, se necesitan otros modelos.

3. Si en situaciones de elecciones repetidas, los factores no observados son independientes a lo largo del tiempo, los modelos logit pueden capturar la dinámica de la elección repetida, incluyendo la dependencia del estado. Sin embargo, logit no puede manejar situaciones en las que los factores no observados se correlacionan a lo largo del tiempo.

Elaboramos cada una de estas afirmaciones en las próximas tres subsecciones.

3.3.1 Variación de preferencias

El valor o importancia que los decisores dan a cada atributo de las alternativas varía, en general, para los diferentes decisores. Por ejemplo, el tamaño de un automóvil es probablemente más importante para los hogares con muchos miembros que para los hogares más pequeños. Los hogares con ingresos bajos probablemente están más preocupados por el precio de compra de un bien, en relación al resto de sus características, que los hogares con mayores ingresos. En la elección de en qué barrio vivir, los hogares con niños pequeños estarán más preocupados por la calidad de las escuelas que aquellos hogares sin hijos y así sucesivamente. Las preferencias de los decisores también varían por razones que no están vinculadas a características sociodemográficas observadas, simplemente porque personas diferentes son diferentes. Dos personas que tienen el mismo nivel de ingresos, educación, etc., harán diferentes elecciones, lo que refleja sus preferencias individuales y preocupaciones.

Los modelos logit pueden capturar variaciones de preferencia, pero sólo dentro de ciertos límites. En particular, se pueden incorporar en los modelos logit preferencias que varían sistemáticamente respecto a variables observadas, mientras que preferencias que varían con variables no observadas o puramente al azar no pueden ser manejadas. El siguiente ejemplo ilustra la distinción.

Considere la elección que realizan los hogares en el momento de comprar un automóvil entre marcas y modelos disponibles. Supongamos, por simplicidad, que los dos únicos atributos de los automóviles que el investigador observa son el precio de compra, PP_j para la marca/modelo j, y los centímetros de espacio para los hombros, SR_j, que es una medida del tamaño del interior de un automóvil. El valor que las familias dan a estos dos atributos varía entre los diferentes hogares, por lo que la utilidad se escribe como

$$U_{nj} = \alpha_n SR_j + \beta_n PP_j + \varepsilon_{nj} \tag{3.8}$$

donde α_n y β_n son parámetros específicos para el hogar n.

Los parámetros varían entre los hogares reflejando diferencias en preferencias. Supongamos por ejemplo que el valor otorgado al espacio para los hombros varía con el número de miembros de los hogares, $\mathrm{M_n}$, pero nada más:

$$\alpha_n = \rho M_n,$$

de manera que a medida que aumenta $\mathrm{M_n}$ el valor otorgado al espacio para los hombros, α_n, también aumenta. De forma similar, supongamos que la importancia del precio de compra se relaciona inversamente con el nivel de ingresos, I_n, por lo que los hogares de bajos ingresos dan más importancia al precio de compra:

$$\beta_n = \theta / I_n.$$

Sustituyendo estas relaciones en (3.8) resulta

$$U_{nj} = \rho(M_n SR_j) + \theta(PP_j/I_n) + \varepsilon_{nj}.$$

Bajo el supuesto de que cada ε_{nj} se distribuye valor extremo iid, obtenemos un modelo logit estándar con dos variables describiendo la utilidad representativa, siendo ambas una interacción entre un atributo del vehículo y una característica del hogar.

Podríamos emplear otras especificaciones para la variación en las preferencias. Por ejemplo, podríamos asumir que el valor del espacio para los hombros aumenta con el tamaño del hogar, pero con una tasa decreciente, de modo que $\alpha_n = \rho \mathrm{M}_n + \phi M_n^2$, donde se espera que ρ sea positivo y ϕ negativo. Entonces $U_{nj} = \rho\big(M_n SR_j\big) + \phi(M_n^2 SR_j) + \theta(PP_j/I_n) + \varepsilon_{nj}$, lo que resulta en un modelo logit con tres variables entrando en la utilidad representativa.

La limitación del modelo logit surge cuando intentamos permitir variación de preferencias respecto a variables no observadas o puramente al azar. Supongamos, por ejemplo, que el valor del espacio para los hombros varía con el tamaño del hogar y con algunos otros factores (por ejemplo, el tamaño de las propias personas o la frecuencia con la que la familia viaja junta) que pasan desapercibidos para el investigador y, por tanto, son considerados factores aleatorios:

$$\alpha_n = \rho M_n + \mu_n,$$

donde μ_n es una variable aleatoria. Del mismo modo, la importancia del precio de compra consta de sus componentes observados y no observados:

$$\beta_n = \theta/I_n + \eta_n.$$

Sustituyendo en (3.8) resulta en

$$U_{nj} = \rho\big(M_n SR_j\big) + \mu_n SR_j + \theta\big(PP_j/I_n\big) + \eta_n PP_j + \varepsilon_{nj}.$$

Dado que μ_n y η_n no se observan, los términos $\mu_n SR_j$ y $\eta_n PP_j$ pasan a formar parte del componente no observado de la utilidad,

$$U_{nj} = \rho\big(M_n SR_j\big) + \theta\big(PP_j/I_n\big) + \tilde{\varepsilon}_{nj},$$

donde $\tilde{\varepsilon}_{nj} = \mu_n SR_j + \eta_n PP_j + \varepsilon_{nj}$. Los nuevos términos de error $\tilde{\varepsilon}_{nj}$ posiblemente no pueden estar distribuidos idénticamente y de forma independiente, como se requiere para la formulación logit. Desde el momento en que μ_n y η_n entran a formar parte de cada alternativa, $\tilde{\varepsilon}_{nj}$ está necesariamente correlacionada entre las alternativas: $Cov\big(\tilde{\varepsilon}_{nj}, \tilde{\varepsilon}_{nk}\big) = Var(\mu_n) SR_j SR_k + Var(\eta_n) PP_j PP_k \neq 0$ para cualquier pareja de modelos de automóvil j y k. Además, puesto que SR_j y PP_j varían entre alternativas, la varianza de $\tilde{\varepsilon}_{nj}$ varía entre alternativas, violando la asunción de errores idénticamente distribuidos: $Var\big(\tilde{\varepsilon}_{nj}\big) = Var(\mu_n) SR_j{}^2 + Var(\eta_n) PP_j{}^2 + Var(\varepsilon_{nj})$, que es diferente para los diferentes j.

Este ejemplo ilustra la idea general de que cuando las preferencias varían de forma sistemática respecto a las variables observadas, la variación se puede incorporar en los modelos logit. Al contrario, si la variación de preferencia es al menos en parte debida al azar, logit es una mala especificación. Como aproximación, logit podría ser capaz de captar las preferencias promedio bastante bien, incluso cuando las preferencias son al azar, ya que la fórmula logit parece ser bastante robusta frente a malas especificaciones. El investigador podría, por tanto, optar por utilizar logit incluso cuando se sabe que las preferencias tienen un componente aleatorio, en aras de la simplicidad. Sin embargo, no hay garantías de que un modelo logit se vaya a aproximar a las preferencias promedio. E incluso si lo hace, logit no

proporcionará información sobre la distribución de las preferencias alrededor de la media. Esta distribución puede ser importante en muchas situaciones, tales como en la predicción de la penetración de mercado que logrará un nuevo producto que atrae a una minoría de personas en lugar de dirigirse a las preferencias promedio. Para incorporar la variación aleatoria de preferencias de forma apropiada y completa, pueden utilizarse en su lugar un modelo probit o un logit mixto.

3.3.2 Patrones de sustitución

Cuando los atributos de una alternativa mejoran (por ejemplo, su precio baja), la probabilidad de ser elegida aumenta. Algunas de las personas que habrían elegido otras alternativas con sus atributos originales ahora eligen esta alternativa en su lugar. Dado que las probabilidades de todas las alternativas suman uno, un aumento en la probabilidad de una alternativa necesariamente implica una disminución de la probabilidad de que otras alternativas sean escogidas. El patrón de sustitución entre alternativas tiene implicaciones importantes en numerosas situaciones.

Por ejemplo, cuando un fabricante de teléfonos celulares lanza un nuevo producto con prestaciones adicionales, está muy interesada en conocer en qué medida el nuevo producto atraerá a los clientes de sus otros modelos de teléfonos celulares en lugar de los teléfonos de la competencia, ya que la empresa logra más beneficios de lo segundo que de lo primero. Como veremos, el patrón de sustitución también afecta a la demanda de un producto y al cambio en la demanda cuando sus atributos cambian. Por lo tanto, los patrones de sustitución son importantes incluso cuando el investigador está interesado sólo en la cuota de mercado (*market share*) sin preocuparse de donde proviene la cuota.

El modelo logit implica un cierto patrón de sustitución entre alternativas. Si la sustitución se produce realmente de esta manera, dada la especificación que el investigador hace de la utilidad representativa, entonces el modelo logit es apropiado.

Sin embargo, para permitir patrones de sustitución más generales y para investigar qué patrón es más preciso, necesitamos modelos más flexibles. La cuestión se puede ver de dos maneras: como una restricción de los ratios de probabilidades de alternativas y/o como una restricción de las elasticidades cruzadas de las probabilidades. Presentamos cada una de estas maneras de caracterizar el problema a continuación.

Propiedad de independencia de alternativas irrelevantes

Para cualesquiera dos alternativas i y k, el ratio de las probabilidades logit es

$$\frac{P_{ni}}{P_{nk}} = \frac{e^{V_{ni}} / \sum_j e^{V_{nj}}}{e^{V_{nk}} / \sum_j e^{V_{nj}}}$$

$$= \frac{e^{V_{ni}}}{e^{V_{nk}}} = e^{V_{ni} - V_{nk}}$$

Este ratio no depende de las alternativas que no sean i y k. Es decir, las probabilidades relativas de elegir i sobre k son las mismas sin importar qué otras alternativas estén disponibles o cuáles sean los atributos de las otras alternativas. Dado que el ratio es independiente de alternativas distintas de i y k, se dice que es independiente de alternativas irrelevantes. El modelo logit exhibe esta independencia de alternativas irrelevantes (*independence from irrelevant alternatives*, IIA).

En muchas circunstancias, las probabilidades de elección que exhiben IIA proporcionan una representación precisa de la realidad. De hecho, Luce (1959) consideraba que la IIA es una propiedad de las probabilidades de elección que han sido especificadas apropiadamente. Él derivó el modelo logit directamente de la suposición de que las probabilidades de elección debían cumplir con la IIA, en lugar

de (como hemos hecho nosotros) obtener la fórmula logit a partir de un supuesto acerca de la distribución de la utilidad no observada, para luego observar que la IIA es una propiedad resultante del modelo.

Mientras que la propiedad IIA es realista en algunas situaciones de elección, es claramente inadecuada en otras, como señaló por primera vez Chipman (1960) y Debreu (1960). Consideremos el famoso problema del autobús rojo – autobús azul. Un viajero tiene la opción de ir al trabajo en automóvil o tomar un autobús azul. Por simplicidad suponemos que la utilidad representativa de los dos medios de transporte es la misma, de tal manera que las probabilidades de elección son iguales: $P_c = P_{bb} = 1/2$, donde c representa al automóvil (*car*) y bb al autobús azul (*blue bus*). En este caso, el ratio de probabilidades es uno: $P_c/P_{bb} = 1$.

Ahora suponga que se introduce una nueva opción de transporte, un autobús rojo (*red bus*), y que el viajero considera que el autobús rojo es exactamente igual que el autobús azul. La probabilidad de que el viajero elija el autobús rojo es por lo tanto la misma que para el autobús azul, de manera que el ratio de sus probabilidades es uno: $P_{rb}/P_{bb} = 1$. Sin embargo, en el modelo logit el ratio P_c/P_{bb} es el mismo haya o no una nueva alternativa, en este caso, la alternativa del autobús rojo. Por tanto, este ratio se mantiene en uno. Las únicas probabilidades de elección que mantienen ambos ratios constantes, $P_c/P_{bb} = 1$ y $P_{rb}/P_{bb} = 1$, son $P_c = P_{bb} = P_{rb} = 1/3$, que son justamente las probabilidades que el modelo logit predice.

En el mundo real, sin embargo, esperaríamos que la probabilidad de que el viajero eligiese el automóvil se mantuviese igual cuando un nuevo autobús exactamente igual al ya existente estuviese disponible. También esperaríamos que la probabilidad inicial de elegir el autobús se dividiese entre los dos autobuses disponibles, una vez introducido el nuevo autobús. Es decir, podríamos esperar $P_c = 1/2$ y $P_{bb} = P_{rb} = 1/4$. En este caso, el modelo logit, debido a su propiedad IIA, sobreestima la probabilidad de elegir cualquiera de los autobuses y subestima la probabilidad de elegir el automóvil. El ratio de las probabilidades de elección del automóvil y el autobús azul, P_c/P_{bb}, realmente cambia con la introducción del autobús rojo, en lugar de permanecer constante como requiere el modelo logit.

Este ejemplo es bastante rígido y es poco probable que se produzca en la vida real. Sin embargo, el mismo tipo de predicción errónea surge con modelos logit siempre que el ratio de probabilidades para dos alternativas cambie con la introducción o cambio de otra alternativa. Por ejemplo, supongamos que se añade un nuevo medio de transporte que es similar, pero no exactamente igual, a los medios existentes, como un autobús expreso a lo largo de una línea que ya cuenta con servicio de autobús estándar. Podría esperarse que este nuevo medio de transporte redujese la probabilidad del autobús regular en una proporción mayor de lo que se reduciría la probabilidad del automóvil, por lo que el ratio de probabilidades para el automóvil y el autobús regular no permanecería constante. En esta situación, el modelo logit sobrestimaría la predicción de la demanda para los dos tipos de autobuses. Otros ejemplos han sido proporcionados por Ortuzar (1983) y Brownstone y Train (1999).

Sustitución proporcional

El mismo problema puede expresarse en términos de elasticidades cruzadas de las probabilidades logit. Vamos a considerar el cambio de un atributo de la alternativa j. Queremos conocer qué efecto tiene este cambio en las probabilidades de elección de todas las alternativas restante. En la sección 3.6 se obtiene la fórmula de la elasticidad de P_{ni} respecto a una variable que entra en la utilidad representativa de la alternativa j:

$$E_{iz_{nj}} = -\beta_z z_{nj} P_{nj},$$

donde z_{nj} es el atributo de la alternativa j percibido por la persona n y β_z es su coeficiente (o si la variable entra en la utilidad representativa de forma no lineal, β_z sería la derivada de V_{nj} con respecto a z_{nj}) .

Esta elasticidad cruzada es la misma para todas las i: i no entra en la fórmula. Una mejora en los atributos de una alternativa reduce las probabilidades de elección de todas las otras alternativas en el mismo porcentaje. Si la probabilidad de elección de una alternativa se reduce un 10%, entonces todas las otras probabilidades de elección de las restantes alternativas también caen un 10% (excepto, claro está, la alternativa cuyo atributo ha cambiado; su probabilidad se eleva debido a la mejora). Una forma concisa de expresar este fenómeno es indicar que una mejora en una alternativa se extrae proporcionalmente de todas las otras alternativas restantes. Del mismo modo, para una disminución de la utilidad representativa de una alternativa, las probabilidades de todas las restante alternativas se incrementan en el mismo porcentaje.

Este patrón de sustitución, que puede ser llamado *desplazamiento proporcional (proportionate shifting)*, es una manifestación de la propiedad IIA. El ratio de probabilidades de elección de las alternativas i y k se mantiene constante cuando un atributo de la alternativa j cambia sólo si las dos probabilidades varían en la misma proporción. Si denotamos con el superíndice 0 las probabilidades antes del cambio y con 1 después del cambio, la propiedad IIA exige que

$$\frac{P_{ni}^1}{P_{nk}^1} = \frac{P_{ni}^0}{P_{nk}^0}$$

cuando un atributo de la alternativa j cambia. Esta igualdad sólo puede mantenerse si cada probabilidad cambia en la misma proporción: $P_{ni}^1 = \lambda P_{ni}^0$ y $P_{nk}^1 = \lambda P_{nk}^0$, donde en ambos casos λ es la misma.

La sustitución proporcional puede ser realista para algunas situaciones, en cuyo caso el modelo logit es apropiado. Sin embargo, en muchos casos podemos esperar encontrar otros patrones de sustitución por lo que imponer la sustitución proporcional a través de un modelo logit puede conducir a previsiones poco realistas. Considere una situación que es importante para la *California Energy Comission* (CEC), entidad que tiene la responsabilidad de investigar las políticas reguladoras para promover vehículos de energía eficiente en California, así como reducir la dependencia del estado respecto a la gasolina para los automóviles. Supongamos en aras de la simplicidad que hay tres clases de vehículos: automóviles grandes de gasolina, automóviles pequeños de gasolina y automóviles pequeños eléctricos. Supongamos también que en las condiciones actuales las probabilidades de que un hogar elija cada uno de estos vehículos son 0.66, 0.33 y 0.01 respectivamente. La CEC está interesada en conocer el impacto de subvencionar los automóviles eléctricos. Supongamos que el subsidio es suficiente para elevar la probabilidad de que el automóvil eléctrico sea elegido de 0.01 a 0.10. A través del modelo logit, la probabilidad de elección de cada uno de los automóviles de gasolina se prevé que caiga en el mismo porcentaje. La probabilidad de elección de un automóvil grande de gasolina se reduciría en un diez por ciento, de 0.66 a 0.60, y lo mismo sucedería para el automóvil pequeño de gasolina, reducción del diez por ciento, de 0.33 hasta 0.30. En términos de números absolutos, la probabilidad incrementada del automóvil pequeño eléctrico (0.09) se prevé, según el modelo logit, que provenga dos veces más de los automóviles grandes de gasolina (0.06) que de los automóviles pequeños de gasolina (0.03).

Este patrón de sustitución es claramente poco realista. Dado que el automóvil eléctrico es pequeño, es de esperar que subsidiarlo atraiga más a usuarios de automóviles pequeños de gasolina que a usuarios de automóviles grandes de gasolina. En términos de elasticidades cruzadas, esperaríamos que la elasticidad cruzada de los automóviles pequeños de gasolina respecto a una mejora en los automóviles pequeños eléctricos vaya a ser más alta que la de los automóviles grandes de gasolina. Esta diferencia es importante en el análisis de políticas de la CEC. El modelo logit sobrestimará la predicción de ahorro de gasolina que resultará de la subvención, dado que sobrestimará la sustitución de los automóviles

grandes de gasolina (los que más gastan) y subestimará la sustitución de los automóviles pequeños de gasolina. Desde una perspectiva reguladora, esta predicción errónea puede ser crítica, causando que un programa de subsidios parezca más beneficioso de lo que realmente es. Por esta razón la CEC utiliza modelos que son más generales que logit para representar la sustitución entre vehículos. Los modelos logit jerárquico, probit y logit mixto explicados en los capítulos 4-6 ofrecen opciones viables para el investigador.

Ventajas de la IIA

Como acabamos de mencionar, la propiedad de IIA del modelo logit puede ser poco realista en muchos entornos. Sin embargo, cuando la IIA refleja la realidad (o una aproximación adecuada de la realidad), se obtienen considerables ventajas en su uso. En primer lugar, debido a la IIA, es posible estimar los parámetros del modelo de forma consistente en un subconjunto de alternativas para cada decisor estudiado. Por ejemplo, en una situación con 100 alternativas, el investigador podría, con el fin de reducir el tiempo de cálculo, estimar sobre un subconjunto de 10 alternativas para cada persona de la muestra, con la elección de la persona incluida así como 9 alternativas adicionales seleccionadas aleatoriamente entre las restantes 99 . Dado que las probabilidades relativas dentro de un subconjunto de alternativas no se ven afectadas por los atributos o por la existencia de alternativas fuera del subconjunto, la exclusión de alternativas en la estimación no afecta a la consistencia del estimador. Los detalles de este tipo de estimación se ofrecen en la sección 3.7.1. Este hecho tiene gran importancia práctica. En el análisis de situaciones de elección para las que el número de alternativas es grande, la estimación en un subconjunto de alternativas puede ahorrar cantidades sustanciales de tiempo de computación. En casos extremos, el número de alternativas podría ser tan grande como para impedir por completo la estimación si no fuese posible utilizar un subconjunto de las alternativas.

Otro uso práctico de la propiedad IIA surge cuando el investigador sólo está interesado en examinar elecciones entre un subconjunto de alternativas y no entre todas las alternativas. Por ejemplo, considere un investigador que está interesado en la comprensión de los factores que afectan a la elección que realizan los trabajadores entre el automóvil y el autobús como medios de transporte para ir al trabajo. El conjunto completo de alternativas incluye medios como caminar, andar en bicicleta, motocicletas, patines, etc. Si el investigador pensó que la propiedad IIA se cumplía adecuadamente en este caso, podría estimar un modelo de elección con sólo automóvil y autobús como alternativas y excluir de la muestra analizada a los trabajadores que utilizaron otros medios de transporte. Esta estrategia le ahorraría al investigador un tiempo y costo considerables en el desarrollo de los datos de los otros medios, sin limitar su capacidad de examinar los factores relacionados con el automóvil y el autobús.

Pruebas de IIA

Si la IIA se cumple en un entorno particular es una cuestión empírica, susceptible de investigación estadística. Las primeras pruebas para verificar la IIA fueron desarrolladas por McFadden et al . (1978). Se sugieren dos tipos de pruebas. En primer lugar, el modelo se puede re-estimar en un subconjunto de alternativas. Según la IIA, el ratio de probabilidades para cualesquiera dos alternativas debería ser el mismo estén o no estén presentes otras alternativas. Como resultado, si la IIA se observa realmente, las estimaciones obtenidas de los parámetros en el subconjunto de alternativas no deberían ser significativamente diferentes a las obtenidas usando el total de las alternativas. Un test de la hipótesis de que los parámetros en el subgrupo son los mismos que los parámetros para el conjunto completo de alternativas constituye una prueba de la IIA. Hausman y McFadden (1984) proporcionan un estadístico adecuado para este tipo de prueba. En segundo lugar, el modelo se puede re-estimar con nuevas variables intercambiadas entre alternativas, es decir, con las variables de una alternativa formando parte de la utilidad de otra alternativa. Si el ratio de probabilidades de las alternativas i y k en realidad depende de los atributos y de la existencia de una tercera alternativa j (en violación de la IIA), los

atributos de la alternativa j entrarán significativamente en la utilidad de las alternativas i o k dentro de una especificación logit. Por lo tanto, una prueba de si las variables cruzadas entre alternativas entran en el modelo constituye una prueba de la IIA. McFadden (1987) desarrolló un procedimiento para realizar este tipo de prueba con regresiones: para ello usaba los residuos del modelo logit original como variable dependiente y las variables cruzadas entre alternativas, apropiadamente especificadas, como variables explicativas. Train et al. (1989) muestran cómo este procedimiento se puede realizar convenientemente dentro del modelo logit mismo.

La llegada de modelos que no presentan la IIA y especialmente el desarrollo de software para la estimación de estos modelos, hace que las pruebas de IIA sean más fáciles que antes. Para especificaciones más flexibles, tales como GEV y logit mixto, el modelo logit simple con IIA es un caso especial que surge bajo ciertas restricciones sobre los parámetros del modelo más flexible. En estos casos, la IIA puede ser probada mediante pruebas de estas limitaciones. Por ejemplo, un modelo logit mixto se convierte en un logit simple si la distribución de mezcla tiene varianza cero. La IIA se puede probar mediante la estimación de un logit mixto, probando posteriormente si la varianza de la distribución mixta es cero en la práctica.

Una prueba de la IIA como una restricción a un modelo más general necesariamente opera bajo la suposición de que el modelo más general es en sí mismo una especificación apropiada. Las pruebas sobre subconjuntos de alternativas (Hausman y McFadden, 1984) y sobre variables cruzadas entre alternativas (McFadden, 1987; Train et al, 1989), aunque son más difíciles de realizar, operan bajo hipótesis menos restrictivas. El contrapunto a esta ventaja, por supuesto, es que cuando la IIA falla, estas pruebas no proporcionan tanta orientación sobre la especificación correcta a usar en lugar del modelo logit.

3.3.3 Datos de panel

En muchas ocasiones el investigador puede observar numerosas elecciones realizadas por cada decisor. Por ejemplo, en los estudios de ocupación, se observa si las personas en la muestra trabajan o no trabajan en cada mes durante varios años. Un investigador interesado en las dinámicas de elección de compra de un automóvil podría obtener datos sobre las compras actuales y pasadas de vehículos en una muestra de hogares. En las encuestas usadas en investigación de mercados, a los encuestados a menudo se les pide responder una serie de preguntas de elección hipotética, llamadas experimentos de "preferencia declarada". Para cada uno de estos experimentos, se describe al encuestado un conjunto de productos alternativos con diferentes atributos y se le pide que indique cuál es el producto que elegiría. Al encuestado se le administra una batería de este tipo de preguntas, variando cada vez los atributos de los productos con el fin de determinar cómo la elección del entrevistado cambia cuando cambian los atributos. Por lo tanto, el investigador puede observar la secuencia de opciones elegidas por cada encuestado. Los datos que representan repeticiones de elecciones como éstas se llaman datos de panel.

Si los factores no observados que afectan a los decisores son independientes entre las elecciones repetidas, un modelo logit puede utilizarse para examinar los datos de panel de la misma forma como se usaría para examinar datos obtenidos de una sola vez. Cualesquiera dinámicas relacionadas con factores observados que entren en el proceso de decisión, como la dependencia del estado del decisor (por la cual elecciones pasadas de la persona influyen en sus elecciones presentes) o la respuesta diferida a cambios en los atributos, se puede acomodar al modelo. Sin embargo, dinámicas asociadas a factores no observados no se pueden manejar, dado que se supone que los factores no observados no guardan relación entre elecciones.

La utilidad que el decisor n obtiene de la alternativa j en el período o situación de elección t es

$$U_{njt} = V_{njt} + \varepsilon_{njt} \; \forall j, t.$$

Si ε_{njt} se distribuye con densidad valor extremo, independiente respecto a n, j y sobre todo, t, entonces, usando la misma prueba descrita en (3.6), las probabilidades de elección son

$$P_{nit} = \frac{e^{V_{nit}}}{\sum_j e^{V_{njt}}}. \tag{3.9}$$

Cada situación de elección de cada decisor se convierte en una observación independiente. Si se especifica que la utilidad representativa de cada período dependa sólo de las variables correspondientes a ese período, por ejemplo, $V_{njt} = \beta' x_{njt}$, donde x_{njt} es un vector de variables que describen la alternativa j tal y como se presenta al decisor n en el período t, entonces no hay ninguna diferencia esencial entre el modelo logit con datos de panel y con datos puntuales.

Podemos capturar aspectos dinámicos del comportamiento especificando que la utilidad representativa en cada período dependa de variables observadas en otros períodos. Por ejemplo, una respuesta diferida al precio puede representarse mediante la introducción del precio en el período $t-1$ como variable explicativa en la utilidad del período t. La introducción de los precios en períodos futuros puede realizarse, tal y como hace Adamowicz (1994), para capturar la anticipación de los consumidores a futuros cambios de precios. Bajo los supuestos del modelo logit, la variable dependiente en períodos anteriores también se puede introducir como variable explicativa. Supongamos por ejemplo que existe una inercia o formación de hábitos en las elecciones de las personas de tal manera que tienden a quedarse con la alternativa que hayan elegido previamente a menos que otra alternativa proporcione una utilidad suficientemente mayor para justificar un cambio. Este comportamiento se puede capturar como $V_{njt} = \alpha y_{nj(t-1)} + \beta x_{njt}$, donde $y_{nj(t-1)} = 1$ si n eligió j en el período t y 0 en caso contrario. Con $\alpha > 0$, la utilidad de la alternativa j en el período actual es mayor si ya se consumía la alternativa j en el período anterior. La misma especificación también puede capturar una especie de búsqueda de la variedad. Si α es negativo, el consumidor obtiene mayor utilidad de no elegir la misma alternativa que eligió en el último período. Son posibles numerosas variaciones sobre estos conceptos. Adamowicz (1994) introduce el número de veces que la alternativa ha sido elegida previamente en lugar de usar un simple indicador para la elección inmediatamente anterior. Erdem (1996) introduce los atributos de las alternativas elegidas previamente, con la utilidad de cada alternativa en el período actual dependiendo de la similitud de sus atributos con los atributos previamente experimentados.

La inclusión de la variable dependiente diferida no induce inconsistencia en la estimación, ya que para un modelo logit se supone que los errores deben ser independientes en el tiempo. La variable dependiente diferida $y_{nj(t-1)}$ no está correlacionada con el error actual ε_{njt} debido a esta independencia. La situación es análoga a los modelos de regresión lineal, en los que una variable dependiente diferida puede añadirse sin inducir sesgo, siempre y cuando los errores sean independientes a lo largo del tiempo.

Por supuesto, la asunción de errores independientes en el tiempo es severa. Por lo general, es de esperar que existan algunos factores que no están siendo observados por el investigador que afecten a cada una de las elecciones de los decisores. En particular, si hay dinámicas en los factores observados, el investigador también podría esperar que haya dinámicas en los factores no observados. En estas situaciones, el investigador puede utilizar un modelo como el probit o el logit mixto que permiten que los factores no observados estén correlacionados en el tiempo, o re-especificar la utilidad representativa para incorporar de forma explícita las fuentes de las dinámicas no observadas en el modelo, de tal manera que los errores restantes sean independientes en el tiempo.

3.4 Utilidad representativa no lineal

En algunos contextos, el investigador encontrará útil permitir que los parámetros entren en la utilidad representativa de forma no lineal. La estimación pasa a ser más difícil, ya que la función log-

verosimilitud (*log-likelihood*) puede que no sea globalmente cóncava y los procedimientos informáticos capaces de tratar esta situación no están tan ampliamente disponibles como para modelos logit con utilidad lineal respecto a los parámetros. Sin embargo, los aspectos del comportamiento que el investigador está estudiando pueden incluir parámetros que sólo son interpretables cuando entran en la utilidad de forma no lineal. En estos casos, el esfuerzo necesario para desarrollar un código propio puede estar justificado. Dos ejemplos ilustran este punto.

Ejemplo 1: El balance entre bienes y ocio

Considere la posibilidad de elección de los trabajadores respecto al medio de transporte a su lugar de trabajo (automóvil o autobús). Supongamos que los trabajadores también eligen el número de horas que quieren trabajar basándose en el equilibrio estándar entre bienes y ocio. Train y McFadden (1978) desarrollaron un procedimiento para examinar estas decisiones interrelacionadas. Como veremos a continuación, los parámetros de la función de utilidad de los trabajadores respecto a los bienes y al ocio entran de forma no lineal en la utilidad de los medios de transporte.

Supongamos que las preferencias de los trabajadores con respecto a los bienes G (*goods*) y al ocio L (*leisure*) están representados por una función de utilidad Cobb-Douglas de la forma

$$U = (1-\beta)lnG + \beta lnL.$$

El parámetro β refleja la preferencia relativa de los trabajadores entre obtener bienes y disponer tiempo de ocio, donde una β mayor implica una mayor preferencia por el ocio en relación con los bienes. Cada trabajador tiene una cantidad fija de tiempo (24 horas al día) y se enfrenta a un salario fijo por hora trabajada, w. En el modelo estándar de bienes-ocio, el trabajador elige el número de horas a trabajar que maximiza U, sujeto a las restricciones de que (1) el número de horas trabajadas más el número de horas de ocio es igual al número de horas disponibles y (2) el valor de los bienes consumidos es igual al salario por hora *w* por el número de horas trabajadas.

Cuando se añade el medio de transporte al modelo, las restricciones sobre el tiempo y el dinero cambian. Cada medio de transporte resta una cierta cantidad de tiempo y de dinero. Condicionado a la elección del automóvil como medio de transporte, el trabajador maximiza su utilidad U sujeta a la restricción de que (1) el número de horas trabajadas más el número de horas de ocio es igual al número de horas disponibles después de *restar el tiempo pasado en el automóvil para ir al trabajo* y (2) el valor de los bienes consumidos es igual al salario por hora *w* por el número de horas trabajadas *menos el costo de conducir hasta el trabajo*. La utilidad asociada con la elección de viajar en automóvil es el mayor valor posible de U que se puede alcanzar bajo estas restricciones. Del mismo modo, la utilidad de viajar en autobús al trabajo es el valor máximo de U que se puede obtener teniendo en cuenta el tiempo y el dinero que quedan después de restar el tiempo y el dinero consumidos en el viaje en autobús. Train y McFadden obtuvieron los valores de maximización de U condicionados a cada medio de transporte. Para la U dada anteriormente, estos valores son

$$U_j = -\alpha\left(c_j/w^{\beta} + w^{1-\beta}t_j\right) para\ j = automóvil\ y\ autobús.$$

El costo del viaje se divide por w^{β} y el tiempo de viaje se multiplica por $w^{1-\beta}$. El parámetro β, que denota la preferencia relativa de los trabajadores por los bienes o por el ocio, entra en la expresión de la utilidad de la elección de cada medio de transporte de forma no lineal. Dado que este parámetro tiene sentido, el investigador desearía estimarlo dentro de esta utilidad no lineal en lugar de utilizar una aproximación mediante un modelo lineal en parámetros.

Ejemplo 2: Agregación geográfica

Se han desarrollado modelos ampliamente utilizados para la elección que los viajeros hacen de sus destinos para diferentes tipos de viajes, tales como viajes dentro de un área metropolitana para ir de compras. Habitualmente, el área metropolitana se divide en zonas y los modelos dan la probabilidad de que una persona elija viajar a una zona en particular. La utilidad representativa para cada zona depende del tiempo y el costo de los viajes a la zona, además de otro tipo de variables, tales como cantidad de población residente y número de personas empleadas en comercios, que reflejan motivos por los que la gente puede desear visitar la zona. Estas últimas variables se llaman variables de *atracción*; las etiquetaremos con el vector a_j para la zona j. Puesto que son estas variables de atracción las que dan lugar a los parámetros que entran en la utilidad de forma no lineal, asumimos por simplicidad que la utilidad representativa depende únicamente de estas variables.

La dificultad en la especificación de la utilidad representativa proviene del hecho de que la decisión del investigador en relación a qué tamaño asignar a cada zona es bastante arbitraria. Sería útil tener un modelo que no fuese sensible al nivel de agregación usado en la definición de las zonas. Si dos zonas se combinan, sería bueno para el modelo dar una probabilidad de viajar a la zona combinada que sea la misma que la suma de las probabilidades de viajar a las dos zonas originales. Esta consideración impone restricciones a la forma de la utilidad representativa.

Considere las zonas j y k que, cuando se combinan, se etiquetan como zona c .La población y el empleo en la zona combinada son necesariamente las sumas de la población y el empleo de las dos zonas originales: $a_j + a_k = a_c$. Con el fin de que los modelos den la misma probabilidad de elección de estas zonas antes y después de su fusión, el modelo debe satisfacer

$$P_{nj} + P_{nk} = P_{nc},$$

que para los modelos logit toma la forma

$$\frac{e^{V_{nj}} + e^{V_{nk}}}{e^{V_{nj}} + e^{V_{nk}} + \sum_{l \neq j,k} e^{V_{nl}}} = \frac{e^{V_{nc}}}{e^{V_{nc}} + \sum_{l \neq j,k} e^{V_{nl}}}$$

Esta igualdad se cumple sólo cuando $exp(V_{nj}) + exp(V_{nk}) = exp(V_{nc})$. Si la utilidad representativa se especifica como $V_{nl} = \ln(\beta' a_l)$ para todas las zonas l, entonces la igualdad se cumple: $exp(\ln(\beta' a_j)) + exp(\ln(\beta' a_k)) = \beta' a_j + \beta' a_k = \beta' a_c = exp(\ln(\beta' a_c))$. Por lo tanto, para especificar un modelo de elección de destino que no sea sensible al nivel de agregación zonal, la utilidad representativa tiene que ser especificada con los parámetros dentro de una operación logarítmica.

3.5 Excedente del consumidor

Para el análisis de políticas reguladoras, el investigador a menudo está interesado en medir el cambio en el excedente que obtiene el consumidor (*consumer surplus*) por efecto de una política en particular. Por ejemplo, si se está considerando una nueva alternativa como la construcción de un sistema de tren ligero en la ciudad, es importante medir los beneficios del proyecto para ver si se justifican los costos. Del mismo modo, un cambio en los atributos de una alternativa puede tener un impacto en el excedente que recibe el consumidor cuya evaluación es importante. La degradación de la calidad del agua de los ríos daña a los pescadores que ya no pueden pescar de forma efectiva en los sitios dañados. La medición de este daño en términos monetarios es un elemento central de la acción legal contra el contaminador. A menudo, es importante evaluar los efectos distributivos de una política, por ejemplo, cómo la carga de un impuesto es soportado por diferentes grupos de la población.

Bajo los supuestos logit, el excedente percibido por el consumidor asociado a un conjunto de alternativas toma una forma cerrada que es fácil de calcular. Por definición, el excedente del consumidor de una persona concreta es la utilidad, expresada en dólares, que la persona recibe en la

situación de elección. El decisor escoge la alternativa que ofrece la mayor utilidad. El excedente del consumidor es por lo tanto $CS_n = (1/\alpha_n)max_j(U_{nj})$, donde α_n es la utilidad marginal de los ingresos: $dU_n/dY_n = \alpha_n$, siendo Y_n los ingresos de la persona n. La división por α_n traduce utilidad a dólares, ya que $1/\alpha_n = dY_n/dU_n$. El investigador no observa U_{nj} y por lo tanto no se puede utilizar esta expresión para calcular el excedente del consumidor del decisor. En lugar de ello, el investigador observa V_{nj} y conoce la distribución de la porción restante de la utilidad. Con esta información, el investigador es capaz de calcular el excedente del consumidor esperado:

$$E(CS_n) = \frac{1}{\alpha_n} E\big[max_j\big(V_{nj} + \varepsilon_{nj}\big)\big],$$

donde la esperanza se calcula sobre todos los valores posibles de ε_{nj}. Williams (1977) y Small y Rosen (1981) muestran que si cada ε_{nj} se disitribuye como valor extremo iid y la utilidad es lineal respecto al ingreso (de modo que α_n es constante respecto a ingresos), entonces esta expectativa se convierte en

(3.10) $$E(CS_n) = \frac{1}{\alpha_n} ln\big(\sum_{j=1}^{J} e^{V_{nj}}\big) + C,$$

donde C es una constante desconocida que representa el hecho de que el nivel absoluto de utilidad no puede ser medido. Como veremos a continuación, esta constante es irrelevante para definir una política y puede ser ignorada.

Observe que el argumento entre paréntesis de esta expresión es el denominador de la probabilidad de elección logit (3.6). Excepto por la división y la adición de constantes, el excedente del consumidor esperado en un modelo logit es simplemente el logaritmo del denominador de la probabilidad de elección. A menudo esta expresión es llamada el término log-suma (*log-sum*). Esta semejanza entre las dos fórmulas no tiene significado económico, en el sentido de que no hay nada en el denominador de una probabilidad de elección que se relacione necesariamente con el excedente del consumidor. Es simplemente el resultado de la fórmula matemática de la distribución de valor extremo. Sin embargo, la relación hace el cálculo del excedente del consumidor esperado extremadamente fácil, lo que representa otra de las muchas comodidades de usar el modelo logit.

Bajo la interpretación estándar de la distribución de los errores, tal y como se describe en el último párrafo del punto 2.3, $E(CS_n)$ es el promedio del excedente del consumidor en la sub-población de personas que tienen las mismas utilidades representativas que la persona n. El excedente total del consumidor en la población se calcula como la suma ponderada de $E(CS_n)$ sobre una muestra de decisores, con las ponderaciones reflejando el número de personas en la población que afrontan la misma utilidad representativa que la persona de la muestra.

El cambio en el excedente del consumidor que resulta de un cambio en las alternativas y/o en el conjunto de elección se calcula a partir de (3.10). En particular, $E(CS_n)$ se calcula dos veces: primero en las condiciones previas al cambio y de nuevo en las condiciones posteriores al cambio. La diferencia entre los dos resultados es el cambio en el excedente del consumidor:

$$\Delta E(CS_n) = \frac{1}{\alpha_n}\left[ln\left(\sum_{j=1}^{J^1} e^{V_{nj}^1}\right) - ln\left(\sum_{j=1}^{J^0} e^{V_{nj}^0}\right)\right],$$

donde los superíndices 0 y 1 se refieren a antes y después del cambio. El número de alternativas puede cambiar (por ejemplo, se puede añadir una nueva alternativa) así como los atributos de las alternativas. Dado que la constante desconocida C entra en el excedente del consumidor esperado tanto antes como

después del cambio, desaparece de la diferencia y por lo tanto puede ser ignorada cuando se calculan los cambios en el excedente del consumidor.

Para calcular el cambio en el excedente del consumidor, el investigador debe conocer o haber estimado la utilidad marginal del ingreso, α_n. Por lo general, un precio o un costo variable forma parte de la utilidad representativa, en cuyo caso el negativo de su coeficiente es α_n por definición. (Un coeficiente asociado a un precio o a un costo es negativo; el negativo de un coeficiente negativo da una α_n positiva). Por ejemplo, en la elección entre el automóvil y el autobús, la utilidad es $U_{nj} = \beta_1 t_{nj} + \beta_2 c_{nj}$, donde t es el tiempo, c es el costo y tanto β_1 como β_2 son negativos, lo que indica que la utilidad disminuye a medida que el tiempo o el costo de un viaje aumenta. El negativo del coeficiente de costo $-\beta_2$ es la cantidad en que la utilidad se incrementa debido a una disminución de un dólar en los costos. Una reducción de un dólar en costos es equivalente a un incremento de un dólar en ingresos, ya que la persona puede gastar el dólar que ahorra en costos de viaje de la misma manera que si hubiese obtenido un dólar adicional de ingresos. Por consiguiente, la cantidad $-\beta_2$ es el aumento en la utilidad generada por un incremento de un dólar en los ingresos: la utilidad marginal del ingreso. En este caso, la cantidad es la misma para todo n. Si c_{nj} ha entrado en la utilidad representativa interactuando con características de la persona que no sean el ingreso, como en el producto $c_{nj}H_n$ donde H_n es el tamaño del hogar, entonces la utilidad marginal del ingreso sería $-\beta_2 H_n$, cantidad que varía para diferentes n.

A lo largo de esta explicación hemos asumido que α_n está fijada para cada persona con independencia de sus ingresos. La fórmula (3.10) para la esperanza del excedente del consumidor depende de manera crítica del supuesto de que la utilidad marginal de los ingresos es independiente de los ingresos. Si la utilidad marginal de los ingresos cambia con los ingresos, necesitamos una fórmula más complicada, dado que α_n mismo se convierte en una función de los cambios en los atributos. McFadden (1999) y Karlstrom (2000) proporcionan procedimientos para el cálculo de los cambios en el excedente del consumidor en estas condiciones.

Las condiciones para el uso de la expresión (3.10) en realidad son menos estrictas de lo que hemos indicado. Dado que sólo los cambios en el excedente del consumidor son relevantes para el análisis de políticas reguladoras, la fórmula (3.10) se puede utilizar si la utilidad marginal del ingreso es constante en el rango de cambios en el ingreso que implícitamente se están considerando en la política reguladora. Por lo tanto, para cambios en políticas reguladoras que cambian el excedente del consumidor en pequeñas cantidades por persona en relación con los ingresos, la fórmula se puede utilizar a pesar de que la utilidad marginal del ingreso en realidad varíe con el ingreso.

La suposición de que α_n no depende de los ingresos tiene implicaciones en la especificación de la utilidad representativa. Como ya se ha mencionado, α_n por lo general se toma como el valor absoluto del coeficiente de precio o de costo. Por lo tanto, si el investigador tiene previsto utilizar su modelo para estimar los cambios en el excedente del consumidor y quiere aplicar la fórmula (3.10), no se puede especificar que este coeficiente dependa de los ingresos. En el ejemplo de la elección del medio de transporte, el costo puede aparecer multiplicado por el tamaño del hogar, de manera que el coeficiente de costo y, por lo tanto, la utilidad marginal del ingreso, varíe entre hogares de diferente tamaño. Sin embargo, si el costo aparece dividido por el ingreso del hogar, el coeficiente de costo pasa a depender de los ingresos, algo que viola el supuesto necesario para la expresión (3.10). Esta violación puede no ser importante para los pequeños cambios en el excedente del consumidor, pero sin duda pasa a ser importante para los grandes cambios.

3.6 Derivadas y elasticidades

Dado que las probabilidades de elección son una función de las variables observadas, a menudo es útil conocer en qué medida cambian dichas probabilidades en respuesta a un cambio en algún factor observado. Por ejemplo, en el caso de un hogar eligiendo marca y modelo de automóvil a comprar, una

pregunta natural es: ¿en qué medida aumenta la probabilidad de escoger un automóvil dado si se mejora su eficiencia en el uso de combustible? Desde el punto de vista de los demás fabricantes de automóviles, una cuestión relacionada es: ¿hasta qué punto descenderá la probabilidad de que los hogares elijan, por ejemplo, un Toyota, si mejora la eficiencia de uso de combustible de un Honda?

Para abordar estas cuestiones, calculamos las derivadas de las probabilidades de elección. El cambio en la probabilidad de que el decisor n elija la alternativa i dado un cambio en un factor observado z_{ni} que forma parte de la utilidad representativa de esa alternativa (manteniendo constante la utilidad representativa de las otras alternativas) es

$$\frac{\partial P_{ni}}{\partial z_{ni}} = \frac{\partial\left(e^{V_{ni}}/\sum_j e^{V_{nj}}\right)}{\partial z_{ni}}$$

$$= \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}}\frac{\partial V_{ni}}{\partial z_{ni}} - \frac{e^{V_{ni}}}{\left(\sum_j e^{V_{nj}}\right)^2} e^{V_{ni}}\frac{\partial V_{ni}}{\partial z_{ni}}$$

$$= \frac{\partial V_{ni}}{\partial z_{ni}}\left(P_{ni} - {P_{ni}}^2\right)$$

$$= \frac{\partial V_{ni}}{\partial z_{ni}} P_{ni}(1 - P_{ni})$$

Si la utilidad representativa es lineal en z_{ni} con coeficiente β_z, la derivada se convierte en $\beta_z P_{ni}(1 - P_{ni})$. Esta derivada es máxima cuando $P_{ni} = 1 - P_{ni}$, algo que sucede cuando $P_{ni} = 0.5$. Inversamente, disminuye a medida P_{ni} se aproxima a cero o a uno. La curva de probabilidad sigmoidea de la figura 3.1 es consistente con estos hechos. Dicho de forma intuitiva, el efecto de un cambio en una variable observada es mayor cuando las probabilidades de elección indican un alto grado de incertidumbre en cuanto a la elección. A medida que la elección se hace más cierta (es decir, las probabilidades se acercan a cero o a uno), el efecto de un cambio en una variable observada disminuye.

También se puede determinar en qué medida cambia la probabilidad de elegir una alternativa particular cuando cambia una variable observada relacionada con *otra* alternativa. Supongamos que z_{nj} se refiere a un atributo de la alternativa j ¿Cómo cambia la probabilidad de elegir la alternativa i cuando z_{nj} aumenta? tenemos

$$\frac{\partial P_{ni}}{\partial z_{nj}} = \frac{\partial(e^{V_{ni}}/\sum_k e^{V_{nk}})}{\partial z_{nj}}$$

$$= -\frac{e^{V_{ni}}}{(\sum_k e^{V_{nk}})^2} e^{V_{nj}}\frac{\partial V_{nj}}{\partial z_{nj}}$$

$$= -\frac{\partial V_{nj}}{\partial z_{nj}} P_{ni}P_{nj}.$$

Cuando V_{nj} es lineal en z_{nj} con coeficiente β_z, entonces esta derivada cruzada se convierte en $-\beta_z P_{ni}P_{nj}$. Si z_{nj} es un atributo deseable, de manera que β_z es positivo, incrementar z_{nj} disminuye la probabilidad de elegir cada alternativa que no sea j. Además, la disminución de la probabilidad es proporcional al valor de la probabilidad antes de que z_{nj} cambiase.

Un aspecto lógicamente necesario de las derivadas de las probabilidades de elección es que, cuando una variable observada cambia, los cambios en las probabilidades de elección sumen cero. Esto es una consecuencia del hecho de que las probabilidades deben sumar a uno antes y después del cambio; la demostración de este hecho para modelos logit es la siguiente:

$$\sum_{i=1}^{J} \frac{\partial P_{ni}}{\partial z_{nj}} = \frac{\partial V_{nj}}{\partial z_{nj}} P_{nj}\left(1 - P_{nj}\right) + \sum_{i \neq j} \left(-\frac{\partial V_{nj}}{\partial z_{nj}}\right) P_{nj} P_{ni}$$

$$= \frac{\partial V_{nj}}{\partial z_{nj}} P_{nj} \left[\left(1 - P_{nj}\right) - \sum_{i \neq j} P_{ni}\right]$$

$$= \frac{\partial V_{nj}}{\partial z_{nj}} P_{nj}\left[\left(1 - P_{nj}\right) - \left(1 - P_{nj}\right)\right]$$

$$= 0.$$

En términos prácticos, si una alternativa se mejora de manera que su probabilidad de ser elegida aumenta, la probabilidad adicional necesariamente debe extraerse de otras alternativas. Para aumentar la probabilidad de una alternativa es necesario disminuir la probabilidad de otra alternativa. Aunque es obvio, este hecho es a menudo olvidado por planificadores que quieren mejorar la demanda de una alternativa sin reducir la demanda de otras alternativas.

Los economistas suelen medir la respuesta a los cambios mediante elasticidades en lugar de derivadas, dado que las elasticidades están normalizadas por las unidades de las variables. Una elasticidad es el cambio porcentual en una variable asociado a un cambio del uno por ciento en otra variable. La elasticidad de P_{ni} respecto a z_{ni}, una variable que entra en la utilidad de la alternativa i, es

$$E_{iz_{ni}} = \frac{\partial P_{ni}}{\partial z_{ni}} \frac{z_{ni}}{P_{ni}}$$

$$= \frac{\partial V_{ni}}{\partial z_{ni}} P_{ni}(1 - P_{ni}) \frac{z_{ni}}{P_{ni}}$$

$$= \frac{\partial V_{ni}}{\partial z_{ni}} z_{ni}(1 - P_{ni}).$$

Si la utilidad representativa es lineal en z_{ni} con coeficiente β_z, entonces $E_{iz_{ni}} = \beta_z z_{ni}(1 - P_{ni})$.

La elasticidad cruzada de P_{ni} respecto a una variable que entra en la especificación de la alternativa j es

$$E_{iz_{nj}} = \frac{\partial P_{ni}}{\partial z_{nj}} \frac{z_{nj}}{P_{ni}}$$

$$= -\frac{\partial V_{nj}}{\partial z_{nj}} z_{nj} P_{nj}.$$

que en el caso de que la utilidad sea lineal se reduce a $E_{iz_{nj}} = -\beta_z z_{nj} P_{nj}$. Tal y como se ha analizado en la sección 3.3.2, esta elasticidad cruzada es igual para todas las alternativas i: un cambio en un atributo de la alternativa j cambia las probabilidades de todas las otras alternativas en el mismo porcentaje. Esta propiedad de las elasticidades cruzadas de logit es una manifestación, o re-expresión, de la propiedad de IIA de las probabilidades de elección logit.

3.7 Estimación

Manski y McFadden (1981) y Cosslett (1981) describen métodos de estimación para varios procedimientos de muestreo. Nosotros trataremos en esta sección la estimación para el esquema de muestreo más habitual. En primer lugar, describimos la estimación cuando la muestra es exógena y todas las alternativas se utilizan en la estimación. Luego tratamos la estimación en un subconjunto de alternativas y con ciertos tipos de muestras basadas en la propia elección (es decir, no exógenas).

3.7.1 Muestra exógena

Consideremos en primer lugar la situación en que la muestra se ha seleccionado exógenamente, es decir, la muestra se ha seleccionado aleatoriamente o aleatoriamente por estratos, con los estratos definidos sobre factores exógenos a la elección que se desea analizar. Si el procedimiento de muestreo se relaciona con la elección que se desea analizar (por ejemplo, si se está examinando la elección del medio de transporte y la muestra se obtiene mediante la selección de gente en los autobuses y se une a una selección de personas reclutadas en peajes) entonces necesitaremos procedimientos de estimación más complejos en general, como se trata en la próxima sección. También asumimos que las variables explicativas son exógenas a la situación de elección. Es decir, las variables que entran en la utilidad representativa son independientes del componente no observado de utilidad.

Con el propósito de hacer la estimación, obtenemos una muestra de N decisores. Dado que las probabilidades de elección logit tienen una expresión cerrada, podemos aplicar los procedimientos habituales de máxima verosimilitud (*maximum-likelihood*). La probabilidad de que la persona n elija la alternativa que realmente hemos observado que eligió se puede expresar como

$$\prod_i (P_{ni})^{y_{ni}}$$

donde $y_{ni} = 1$ si la persona n eligió i y cero en caso contrario. Observe que dado que $y_{ni} = 0$ para todas las alternativas no elegidas y P_{ni} elevado a la potencia cero es 1, este término es simplemente la probabilidad de la alternativa elegida.

Asumiendo que la elección de cada decisor es independiente de las elecciones del resto de decisores, la probabilidad de que cada persona de la muestra haya elegido la alternativa que realmente hemos observado que eligió es

$$L(\beta) = \prod_{n=1}^{N} \prod_i (P_{ni})^{y_{ni}},$$

donde β es un vector que contiene los parámetros del modelo. La función logaritmo de verosimilitud (*log-likelihood*) es por lo tanto

(3.11) $$LL(\beta) = \sum_{n=1}^{N} \sum_i y_{ni} ln P_{ni}$$

y el estimador es el valor de β que maximiza esta función. McFadden (1974) muestra que $LL(\beta)$ es globalmente cóncava para una especificación de la utilidad lineal en parámetros y numerosos paquetes de software estadístico disponibles en el mercado permiten estimar estos modelos. Cuando los parámetros entran en la utilidad representativa de forma no lineal, el investigador puede tener que escribir su propio código para hacer la estimación utilizando los procedimientos descritos en el Capítulo 8.

En este caso, la estimación de la máxima verosimilitud puede ser rescrita y reinterpretada de una manera que ayuda a comprender la naturaleza de las estimaciones. En el máximo de la función de verosimilitud, su derivada con respecto a cada uno de los parámetros es cero:

(3.12) $$\frac{LL(\beta)}{d\beta} = 0$$

La estimación de máxima verosimilitud son por lo tanto los valores de β que satisfacen esta condición de primer orden. Por conveniencia, hagamos que la utilidad representativa sea lineal en parámetros: $V_{nj} = \beta' x_{nj}$. Esta restricción no es necesaria, pero hace que la notación y el análisis sean más concisos. Usando (3.11) y la fórmula para las probabilidades logit, se muestra al final de este apartado que la condición de primer orden (3.12) se convierte en

(3.13) $$\sum_n \sum_i (y_{ni} - P_{ni}) x_{ni} = 0$$

Reorganizando y dividiendo ambos lados por N, obtenemos

(3.14) $$\frac{1}{N}\sum_n \sum_i y_{ni} x_{ni} = \frac{1}{N}\sum_n \sum_i P_{ni} x_{ni}$$

Esta expresión es fácilmente interpretable. Sea $\bar{x}$ el promedio de x entre las alternativas elegidas por los individuos muestreados: $\bar{x} = (1/\mathrm{N}) \sum_n \sum_i y_{ni} x_{ni}$. Sea $\hat{x}$ la media de x sobre las elecciones previstas de los decisores en la muestra: $\hat{x} = (1/\mathrm{N}) \sum_n \sum_i P_{ni} x_{ni}$. El promedio observado de x en la muestra es $\bar{x}$, mientras que $\hat{x}$ es la media predicha. Según (3.14), estas dos medias son iguales cuando el estimador de verosimilitud es máximo. Es decir, las estimaciones de máxima verosimilitud de β son aquellas que hacen que el promedio pronosticado de cada variable explicativa sea igual al promedio observado en la muestra. En este sentido, las estimaciones inducen al modelo a reproducir los promedios observados en la muestra.

Esta propiedad del estimador de máxima verosimilitud para los modelos logit adquiere un significado especial para constantes específicas de alternativa. Una constante específica de alternativa es el coeficiente de una variable indicador (*dummy variable*) que identifica a una alternativa. Una variable indicador para la alternativa j es una variable cuyo valor en la utilidad representativa de la alternativa i es $d_i^j = 1$ para $i = j$ y cero en caso contrario. Según (3.14) la constante estimada es la que cumple

$$\frac{1}{N}\sum_n \sum_i y_{ni} d_i^j = \frac{1}{N}\sum_n \sum_i P_{ni} d_i^j$$

$$S_j = \widehat{S_j},$$

donde S_j es la proporción o cuota de personas de la muestra que eligieron la alternativa j y $\widehat{S_j}$ es la proporción prevista para la alternativa j. Con constantes específicas de alternativa, las proporciones previstas para la muestra son iguales a las proporciones observadas. Por lo tanto, el modelo estimado es correcto en promedio dentro de la muestra. Esta característica es similar a la función que cumple una

constante en un modelo de regresión lineal, donde la constante asegura que la media del valor predicho de la variable dependiente sea igual a su promedio observado en la muestra.

La condición de primer orden (3.13) proporciona otra interpretación importante. La diferencia entre la elección real de una persona, y_{ni}, y la probabilidad de esta elección, P_{ni} , es un error del modelo o residuo. El lado izquierdo de (3.13) es la covarianza muestral de los residuos con las variables explicativas. Las estimaciones de máxima verosimilitud son, por lo tanto, los valores de las βs que hacen que esta covarianza sea cero, es decir, que hacen que los residuos no estén correlacionados con las variables explicativas. Esta condición para las estimaciones logit es la misma que se aplica en modelos de regresión lineal. Para un modelo de regresión $y_n = \beta' x_n + \varepsilon_n$, la estimación ordinaria de mínimos cuadrados son los valores de β que hacen que $\sum_n (y_n - \beta' x_n) x_n = 0$. Este hecho se comprueba resolviendo para β: $\beta = (\sum_n x_n x_n')^{-1} (\sum_n x_n y_n')$, que es la fórmula para el estimador ordinario de mínimos cuadrados. Dado que $y_n - \beta' x_n$ es el residuo en el modelo de regresión, los estimadores hacen que los residuos no estén correlacionados con las variables explicativas.

Bajo esta interpretación, los estimadores pueden verse como una forma de proporcionar una muestra análoga a las características de la población. Hemos supuesto que las variables explicativas son exógenas, lo que significa que no están correlacionadas dentro de la población con los errores del modelo. Dado que las variables y los errores no están correlacionados en la población, tiene sentido elegir estimadores que hagan las variables y los residuos no correlacionados en la muestra. Los estimadores hacen justamente eso: proporcionan un modelo que reproduce en la muestra la covarianza nula que se produce en la población.

Los estimadores que resuelven las ecuaciones de la forma (3.13) se dice que emplean el método de los momentos, ya que para definir el estimador utilizan condiciones sobre los momentos (correlaciones en este caso) entre los residuos y las variables. Volveremos a estos estimadores cuando hablemos de la estimación asistida por simulación en el capítulo 10.

Hemos afirmado sin pruebas que (3.13) es la condición de primer orden para el estimador de máxima verosimilitud del modelo logit. Vamos a mostrar esta prueba ahora. La función de logaritmo de la verosimilitud (3.11) puede ser re-expresada como

$$LL(\beta) = \sum_n \sum_i y_{ni} ln P_{ni}$$

$$= \sum_n \sum_i y_{ni} ln \left(\frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{nj}}} \right)$$

$$= \sum_n \sum_i y_{ni} (\beta' x_{ni}) - \sum_n \sum_i y_{ni} ln \left(\sum_j e^{\beta' x_{nj}} \right)$$

La derivada de la función log-verosimilitud se convierte en

$$\frac{dLL(\beta)}{d\beta} = \frac{\sum_n \sum_i y_{ni} (\beta' x_{ni})}{d\beta} - \frac{\sum_n \sum_i y_{ni} ln\left(\sum_j e^{\beta' x_{nj}}\right)}{d\beta}$$

$$= \sum_n \sum_i y_{ni} x_{ni} - \sum_n \sum_i y_{ni} \sum_j P_{nj} x_{nj}$$

$$= \sum_n \sum_i y_{ni} x_{ni} - \sum_n \left(\sum_j P_{nj} x_{nj} \right) \sum_i y_{ni}$$

$$= \sum_n \sum_i y_{ni} x_{ni} - \sum_n \left(\sum_j P_{nj} x_{nj} \right)$$

$$= \sum_n \sum_i (y_{ni} - P_{nj}) x_{ni}$$

Estableciendo este derivada igual a cero obtenemos la condición de primer orden (3.13).

Estimación en un subconjunto de alternativas

En algunas situaciones, el número de alternativas que enfrenta el decisor es tan grande que la estimación de los parámetros del modelo es muy costosa o incluso imposible. Con un modelo logit, la estimación puede realizarse en un subconjunto de las alternativas sin producir inconsistencia. Por ejemplo, un investigador que examina una situación de elección que involucra 100 alternativas puede estimar sobre un subconjunto de 10 alternativas para cada decisor de la muestra, incluyendo siempre la alternativa elegida por cada persona así como 9 alternativas más, seleccionadas al azar entre las 99 restantes. Si todas las alternativas tienen las mismas oportunidades de ser seleccionadas dentro del subconjunto, entonces la estimación puede realizarse en el subconjunto de alternativas como si fuera el conjunto completo. Si las alternativas tienen desigual probabilidad de ser seleccionadas, se requieren procedimientos de estimación más complejos. El procedimiento se describe como sigue.

Supongamos que el investigador ha utilizado algún método específico para la selección aleatoria de alternativas que forman el subconjunto que se utiliza en la estimación para cada decisor de la muestra. Denotamos el conjunto completo de alternativas como F y un subconjunto de alternativas como K. Sea $q(K|i)$ la probabilidad, bajo el método empleado por el investigador para seleccionar la muestra, de que el subconjunto K sea seleccionado teniendo en cuenta que el decisor eligió la alternativa i. Asumiendo que el subconjunto incluye necesariamente la alternativa elegida, tenemos que $q(K|i) = 0$ para cualquier K que no incluya i. La probabilidad de que la persona n elija la alternativa i del conjunto completo es P_{ni}. Nuestro objetivo es obtener una fórmula para la probabilidad de que la persona elija la alternativa i condicionada a que el investigador haya seleccionado el subconjunto K para él. Esta probabilidad condicionada se denota $P_n(i|K)$.

Esta probabilidad condicionada se obtiene de la siguiente manera. La probabilidad conjunta de que el investigador seleccione el subconjunto K y el decisor elija la alternativa i es $Prob(K, i) = q(K|i)Pni$. La probabilidad conjunta también se puede expresar con la condicionada opuesta como $Prob(K, i) = Pni(i|K)Q(K)$ donde $Q(K) = \sum_{j \in F} P_{nj} q(K|j)$ es la probabilidad marginal de que el investigador seleccione el subconjunto K sobre todas las alternativas que la persona podría elegir. Igualando estas dos expresiones y despejando $P_n(i|K)$, tenemos

$$P_n(i|K) = \frac{P_{ni}\, q(K|i)}{\sum_{j \in F} P_{nj} q(K|j)}$$

$$= \frac{e^{V_{ni}}\, q(K|i)}{\sum_{j \in F} e^{V_{nj}} q(K|j)}$$

(3.15)
$$\frac{e^{V_{ni}}\, q(K|i)}{\sum_{j\in K} e^{V_{nj}} q(K|j)}$$

donde en la segunda línea hemos cancelado los denominadores de P_{ni} y P_{nj} $\forall j$, y en la tercera, la igualdad utiliza el hecho de que $q(K|j) = 0$ para cualquier j que no esté en K.

Supongamos que el investigador ha diseñado el proceso de selección de manera que $q(K|j)$ es la misma para todos los $j \in K$. Esta propiedad se produce, por ejemplo, si el investigador asigna la misma probabilidad de selección a todas las alternativas no escogidas, de modo que la probabilidad de escoger j en el subconjunto cuando i es la opción escogida por el decisor es la misma probabilidad de escoger i en el subconjunto cuando j es la opción elegida. McFadden (1978) llama a esto la "propiedad de condicionamiento uniforme", ya que el subconjunto de alternativas tiene una probabilidad uniforme (igual) de ser seleccionado, condicionada a que cualquiera de sus miembros haya sido escogido por el decisor. Cuando esta propiedad se cumple, $q(K|j)$ desaparece de la expresión anterior y la probabilidad se convierte en

$$P_n(i|K) = \frac{e^{V_{ni}}}{\sum_{j\in K} e^{V_{nj}}},$$

que es simplemente la fórmula logit para una persona que se enfrenta a las alternativas disponibles en el subconjunto K.

La función log-verosimilitud condicionada en virtud de la propiedad de condicionamiento uniforme es

$$CLL(\beta) = \sum_n \sum_{i\in K_n} y_{ni} ln \frac{e^{V_{ni}}}{\sum_{j\in K_n} e^{V_{nj}}},$$

donde K_n es el subconjunto seleccionado para la persona n . Esta función es la misma que la función log-verosimilitud dada en (3.11) excepto que el subconjunto de alternativas K_n sustituye, para cada persona de la muestra, al conjunto completo. La maximización de CLL proporciona un estimador consistente de β. Sin embargo, dado que hay información excluida de CLL que sí está incorporada en LL (es decir, información sobre alternativas que no están en cada subconjunto) el estimador basado en CLL no es eficiente.

Supongamos que el investigador diseña un proceso de selección que no presenta la propiedad de condicionamiento uniforme. En este caso, la probabilidad $q(K|i)$ se puede incorporar en el modelo como variable separada. La expresión en (3.15) se puede reescribir como

$$P_n(i|K) = \frac{e^{V_{ni} + ln\, q(K|i)}}{\sum_{j\in K} e^{V_{nj} + ln\, q(K|j)}}$$

Una variable z_{nj} calculada como $\ln q(K_n|j)$ se añade a la utilidad representativa de cada alternativa. El coeficiente de esta variable está limitado a 1 en la estimación.

Nos podemos plantear la siguiente cuestión: ¿por qué un investigador va a querer diseñar un procedimiento de selección que no satisfaga la propiedad de condicionamiento uniforme, cuando satisfacer esta propiedad hace que la estimación sea tan sencilla? Un ejemplo de los beneficios potenciales que tiene el condicionamiento no uniforme la proporcionan Train et al. (1987a) en su estudio de la demanda de telecomunicaciones. La situación de elección en este caso incluye una enorme cantidad de alternativas que representan las diferentes posibilidades de llamadas por hora del día, distancia y duración de las mismas. La gran mayoría de alternativas casi nunca fueron elegidas por

los decisores dentro de la población. Si se hubiesen seleccionado las alternativas con igual probabilidad para cada alternativa, habría sido muy probable que los subconjuntos resultantes hubiesen consistido en su totalidad en alternativas que casi nunca fueron elegidas, junto con la alternativa realmente elegida por la persona. Comparar la alternativa elegida de una persona con un grupo de alternativas altamente indeseables por el individuo proporciona poca información sobre las razones que han motivado la elección de la persona. Para evitar este problema, se seleccionaron las alternativas en proporción a sus cuotas de mercado en la población (o para ser más precisos, estimaciones de las cuotas de mercado de la población). Este procedimiento incrementó la probabilidad de que las alternativas relativamente deseables se incluyesen en cada subconjunto de alternativas usado en la estimación.

3.7.2 Muestras basadas en la elección

En algunas situaciones, una muestra tomada sobre la base de factores exógenos incluiría pocas personas que han optado por alternativas concretas. Por ejemplo, en una elección de calentadores de agua, una muestra aleatoria de hogares en la mayoría de las áreas incluiría sólo un pequeño número de hogares que hayan elegido sistemas de calefacción de agua solares. Si el investigador está particularmente interesado en los factores que afectan a la penetración en el mercado de los dispositivos solares, una muestra al azar tendría que ser muy grande para asegurar un número razonable de hogares con calor solar.

En situaciones como éstas, el investigador podría optar por seleccionar la muestra, o parte de la muestra, sobre la base de la elección que se analiza. Por ejemplo, el investigador que desea estudiar los calentadores de agua podría complementar una muestra aleatoria de hogares con hogares que se sabe (quizá a través de registros de ventas en las tiendas si el investigador tiene acceso a los mismos) que han instalado recientemente calentadores de agua solares.

Las muestras seleccionadas sobre la base de las elecciones de los decisores pueden ser puramente basadas en la elección o un híbrido entre selección basada en la elección y en factores exógenos. En una muestra totalmente basada en la elección, la población se divide en grupos por cada alternativa elegida y los decisores se extraen al azar dentro de cada grupo, aunque en proporciones diferentes. Por ejemplo, un investigador que esté estudiando cómo eligen las personas la ubicación de su lugar de residencia y esté interesado en identificar los factores que contribuyen a que la gente elija una ubicación en particular, podría extraer al azar dentro de esa ubicación concreta uno de cada L hogares y extraer aleatoriamente del resto de ubicaciones uno de cada M hogares, donde M es mayor que L. Este procedimiento asegura que el investigador dispone de un número adecuado de personas de la zona de interés en la muestra. Una muestra híbrida es como la elaborada por el investigador interesado en la calefacción solar de agua del ejemplo anterior, en la que una muestra exógena se complementa con una muestra extraída sobre la base de las decisiones de los hogares.

La estimación de los parámetros del modelo con muestras extraídas al menos parcialmente sobre la base de la elección del decisor es bastante compleja en general y varía con el método exacto del procedimiento de muestreo. Para los lectores interesados, Ben-Akiva y Lerman (1985, pp 234-244) proporcionan un estudio útil. Uno de los resultados es particularmente relevante, ya que permite a los investigadores estimar modelos logit sobre muestras basadas en la elección sin necesidad de emplear procedimientos de estimación complejos. Este resultado, debido a Manski y Lerman (1977), se puede describir de la siguiente manera. Si el investigador está utilizando una muestra basada totalmente en la elección de los decisores e incluye una constante específica de alternativa en la utilidad representativa para cada alternativa, estimar un modelo logit como si la muestra fueses exógena produce estimaciones consistentes para todos los parámetros del modelo a excepción de las constantes específicas de alternativa. Además, estas constantes resultan sesgadas por un factor conocido y por lo tanto se pueden ajustar de manera que las constantes ajustadas sean consistentes. En particular, la esperanza de la constante estimada para la alternativa j, denominada $\widehat{\alpha_j}$, está relacionada con la verdadera constante $\alpha_j{}^*$ por

$$E(\widehat{\alpha_j}) = \alpha_j^{*} - ln(A_j/S_j),$$

donde A_j es la proporción de decisores en la población que eligió la alternativa j y S_j es la proporción en la muestra basada en la elección que eligió la alternativa j . En consecuencia, si A_j es conocido (es decir, si las cuotas de mercado de la población son conocidas para cada alternativa) entonces $\widehat{\alpha_j}$ es un estimador consistente de la constante específica de alternativa, el cual se estima en la muestra basada en la elección y al que se le suma el logaritmo del ratio entre la proporción de esa alternativa en la población y en la muestra.

3.8 Bondad de ajuste y pruebas de hipótesis

Trataremos a continuación la bondad de ajuste y los test de hipótesis en el contexto de los modelos logit, en los cuales la función logaritmo de verosimilitud (log-verosimilitud) se puede calcular con exactitud. Los conceptos aplican a otros modelos, con los debidos ajustes por la varianza de la simulación, cuando la función log-verosimilitud se simula en lugar de calcularse con exactitud.

3.8.1 Bondad de ajuste

Un estadístico denominado *índice de ratio de verosimilitud* (*likelihood ratio index*) se utiliza a menudo con modelos de elección discreta para medir lo bien que se ajustan a los datos. Dicho de forma más precisa, el estadístico mide lo bien que el modelo, con sus parámetros estimados, se comporta en comparación con un modelo en el que todos los parámetros son iguales a cero (que generalmente es equivalente a no tener ningún modelo en absoluto). Esta comparación se realiza sobre la base de la función log-verosimilitud, evaluada tanto para los parámetros estimados como para todos los parámetros iguales a cero.

El índice de ratio de verosimilitud se define como

$$\rho = 1 - \frac{LL(\hat{\beta})}{LL(0)},$$

donde $LL(\hat{\beta})$ es el valor de la función log-verosimilitud en los parámetros estimados y $LL(0)$ es su valor cuando todos los parámetros se igualan a cero. Si los parámetros estimados no lo hacen mejor, en términos de la función de verosimilitud, que los parámetros nulos (es decir, si el modelo estimado no es mejor que no tener modelo) entonces $LL(\hat{\beta}) = LL(0)$ y por lo tanto $\rho = 0$. Este es el valor más bajo que ρ puede tomar (dado que si $LL(\hat{\beta})$ fuese inferior a $LL(0)$, entonces $\hat{\beta}$ no sería la estimación de máxima verosimilitud).

En el extremo opuesto, supongamos que el modelo estimado ha resultado ser tan bueno que podríamos predecir perfectamente la elección de cada decisor de la muestra. En este caso, la función de verosimilitud en los parámetros estimados sería igual a uno, ya que la probabilidad de observar las elecciones que realmente se hicieron es uno. Y dado que el logaritmo de uno es cero, la función log-verosimilitud sería cero en los parámetros estimados. Si $LL(\hat{\beta}) = 0$ entonces $\rho = 1$. Este es el valor más alto que ρ puede tomar. En resumen, el índice de ratio de verosimilitud va desde cero, cuando los parámetros estimados no son mejores que los parámetros nulos, hasta uno, cuando los parámetros estimados predicen perfectamente las decisiones de los decisores incluidos en la muestra.

Es importante destacar que el índice de ratio de verosimilitud no es en absoluto similar en su interpretación a la R^2 utilizada en regresiones, a pesar de que los dos estadísticos tienen el mismo rango. R^2 indica el porcentaje de la variación de la variable dependiente que se "explica" por el modelo estimado. El coeficiente de verosimilitud no tiene un sentido intuitivamente interpretable para los valores que se encuentran entre los extremos (cero y uno). Es el porcentaje de incremento en la función

log-verosimilitud por encima del valor resultante cuando fijamos a cero los parámetros (ya que $\rho = 1 - LL(\hat{\beta})/LL(0) = (LL(0) - LL(\hat{\beta}))/LL(0))$. Sin embargo, el significado de un aumento en dicho porcentaje no está claro. En la comparación de dos modelos estimados con los mismos datos y con el mismo conjunto de alternativas (tal que $LL(0)$ es igual para ambos modelos), por lo general es válido decir que el modelo con el ρ más alto se ajusta mejor a los datos. Pero eso no es más que decir que es preferible incrementar el valor de la función log-verosimilitud. Dos modelos estimados en muestras que no sean idénticas o con un conjunto diferente de alternativas para cualquier decisor en la muestra no se pueden comparar a través de sus valores de índice de ratio de verosimilitud.

Otro estadístico de bondad de ajuste que se utiliza ocasionalmente, pero que realmente debería ser evitado, es el "porcentaje correctamente predicho". Este estadístico se calcula identificando para cada decisor en la muestra la alternativa con la probabilidad más alta de ser elegida, basándose en el modelo estimado, y determinando si esta alternativa fue o no la que realmente el decisor escogió. El porcentaje de decisores en la muestra para los cuales la alternativa más probable y la alternativa elegida es la misma se denomina "porcentaje correctamente predicho".

Este estadístico incorpora una noción opuesta al significado de las probabilidades y al propósito de especificar probabilidades de elección. El estadístico está basado en la idea de que el investigador predice que el decisor elegirá la alternativa para la que el modelo ha estimado una probabilidad de elección más alta. Sin embargo tal y como vimos en la formulación de las probabilidades de elección en el capítulo 2, el investigador no tiene suficiente información para predecir la elección del decisor. El investigador sólo tiene información suficiente como para indicar la probabilidad de que el decisor elija cada alternativa. Cuando calcula probabilidades de elección, el investigador está afirmando que si la situación de elección se repitiese varias veces (o fuese afrontada por muchas personas con los mismos atributos) cada alternativa sería elegida una determinada proporción de las veces. Esto es bastante diferente a decir que la alternativa con la probabilidad más alta será la elegida cada vez.

Un ejemplo puede ser útil para comprender la diferencia. Supongamos que un modelo estimado predice probabilidades de elección de 0.75 y 0.25 en una situación de elección con dos alternativas. Estas probabilidades significan que si 100 personas se enfrentan a las utilidades representativas que dieron lugar a estas probabilidades (o una persona enfrenta estas utilidades representativas 100 veces) la mejor predicción que el investigador puede hacer sobre cuántas personas podrían elegir cada alternativa son 75 y 25. Sin embargo, el estadístico "porcentaje correctamente predicho" está basado en la idea de que la mejor predicción para cada persona es la alternativa con mayor probabilidad. Esta noción podría predecir que una alternativa sería elegida por las 100 personas, mientras que la otra alternativa nunca sería elegida. El procedimiento olvida el sentido de las probabilidades, obviamente da cuotas de mercado inexactas y parece implicar que el investigador tiene información perfecta.

3.8.2 Test de hipótesis

Como sucede con las regresiones, para probar hipótesis sobre los parámetros individuales en los modelos de elección discreta – por ejemplo probar si el parámetro es cero - se usan estadísticos-t estándar (*t-statistics*). Para hipótesis más complejas, casi siempre puede usarse un test de ratio de verosimilitud de la siguiente manera. Considere la hipótesis nula H que puede ser expresada como restricciones sobre los valores de los parámetros. Dos de las hipótesis más comunes son (1) varios parámetros son iguales a cero y (2) dos o más parámetros son iguales entre ellos. La estimación de máxima verosimilitud restringida de los parámetros (etiquetados $\hat{\beta}^H$) es el valor de β que da el mayor valor de LL sin violar las restricciones de la hipótesis nula H. Definamos el ratio de verosimilitudes $R = L(\hat{\beta}^H)/L(\hat{\beta})$, donde $\hat{\beta}^H$ es el valor máximo (restringido) de la función de verosimilitud (sin logaritmo) bajo la hipótesis nula H y $\hat{\beta}$ es el máximo sin restricciones de la función de verosimilitud. Al igual que en los test de ratios de verosimilitud para modelos diferentes a los de elección discreta, el estadístico de test definido como $-2logR$ se distribuye chi-cuadrado con un número de grados de libertad igual al número de restricciones que implica la hipótesis

nula. Por lo tanto, el estadístico de test de hipótesis es $-2(LL(\hat{\beta}^H) - LL(\hat{\beta}))$. Dado que el logaritmo de la verosimilitud es siempre negativo, esto es simplemente dos veces la (magnitud de la) diferencia entre los máximos restringidos y no restringidos de la función log-verosimilitud. Si este valor supera al valor crítico de chi-cuadrado con los grados de libertad apropiados, entonces la hipótesis nula es rechazada.

Hipótesis nula I: Los coeficientes de varias variables explicativas son cero

Para probar esta hipótesis, estime el modelo dos veces: una vez con estas variables explicativas incluidas y una segunda vez sin ellas (ya que excluyendo las variables obliga a que sus coeficientes sean cero). Observe el valor máximo de la función log-verosimilitud para cada estimación; dos veces la diferencia entre estos valores máximos es el valor del estadístico de test. Compare el estadístico de test con el valor crítico de chi-cuadrado con un número de grados de libertad igual al número de variables explicativas excluidas de la segunda estimación.

Hipótesis nula II: Los coeficientes de las dos primeras variables don las mismas

Para probar esta hipótesis, estime el modelo dos veces: una vez con cada una de las variables explicativas entrando por separado en el modelo, incluyendo las dos primeras; luego con las dos primeras variables reemplazadas por una única variable que es la suma de las dos variables (dado que sumar las variables obliga a sus coeficientes a ser iguales). Observe el valor máximo de la función log-verosimilitud para cada una de las estimaciones. Multiplique la diferencia de estos valores máximos por dos y compare esta cifra con el valor crítico de chi-cuadrado con un grado de libertad.

3.9 Estudio de un caso: predicción para un nuevo sistema de tráfico

Una de las primeras aplicaciones de los modelos logit que constituye una prueba importante de sus capacidades, surgió a mediados de la década de 1970 en el área de la Bahía de San Francisco. Un nuevo sistema de tren, llamado *Bay Area Rapid Transit (BART)* había sido construido. Daniel McFadden obtuvo una subvención de la *National Science Foundation* para aplicar modelos logit a la elección del medio de transporte de los viajeros en el área de la bahía y usar los modelos para predecir el número de pasajeros del sistema BART. Tuve la suerte de trabajar como su asistente de investigación en este proyecto. Una muestra de los pasajeros fue seleccionada antes de que el BART fuese abierto al público. Sobre esta muestra se estimaron modelos de elección sobre el medio de transporte. Estas estimaciones proporcionaron información importante sobre los factores que entran en juego en las decisiones de los viajeros, incluyendo el valor otorgado al ahorro de tiempo. Posteriormente, se utilizaron los modelos para pronosticar las decisiones que los pasajeros incluidos en la muestra harían una vez que el BART estuviese disponible. Cuando el BART fue abierto al público, los pasajeros fueron contactados de nuevo y se observaron sus elecciones reales de medios de transporte. La proporción prevista de usuarios del BART se comparó con la proporción realmente observada. Los modelos lograron predecir bastante bien las proporciones, de forma mucho más precisa que los procedimientos empleados por los consultores del BART, que no habían utilizado los modelos de elección discreta.

El equipo del proyecto recolectó datos de 771 viajeros antes de la apertura del BART. Se consideraron cuatro medios de transporte como opciones disponibles para viajar al trabajo: (1) la conducción de un automóvil por uno mismo, (2) tomar el autobús y caminar hasta la parada de autobús, (3) tomar el autobús y conducir hasta la parada de autobús y (4) compartir automóvil entre viajeros. El tiempo y el costo de viajar en cada medio de transporte se determinaron para cada viajero, basándose en la ubicación del domicilio y del trabajo de cada persona. El tiempo de viaje se diferenció entre tiempo caminando a pie (para el modo autobús – a pie), tiempo de espera (para los dos modos de autobús) y tiempo en vehículo (para todos los modos). También se registraron las características de los viajeros, incluyendo ingresos, tamaño del hogar, número de vehículos y número de conductores en el hogar, y si el viajero era el cabeza de familia. Con toda esta información se estimó un modelo logit con utilidad lineal en los parámetros.

El modelo estimado se muestra en la tabla 3.1, extraído de Train (1978).

Tabla 3.1. Modelo Logit de elección de medio de transporte para viajar al trabajo

Variable explicativa[a]		Coeficiente	Estadístico-t
Costo dividido por salario después de impuestos, minutos (1-4)		−0.0284	4.31
Tiempo dentro del automóvil, minutos (1, 3, 4)		−0.0644	5.65
Tiempo dentro del transporte público, minutos (2, 3)		−0.0259	2.94
Tiempo a pie, minutos (2, 3)		−0.0689	5.28
Tiempo de espera en transbordos, minutos (2, 3)		−0.0538	2.30
Número de transbordos (2, 3)		−0.1050	0.78
Tiempo de paso del primer autobús, minutos (2, 3)		−0.0318	3.18
Ingresos del hogar con techo de $7500 (1)		0.00000454	0.05
Ingresos del hogar – $7500 con suelo 0, techo $3000 (1)		−0.0000572	0.43
Ingresos del hogar – $10500 con suelo 0, techo $5000 (1)		−0.0000543	0.91
Número de conductores en el hogar (1)		1.02	4.81
Número de conductores en el hogar (3)		0.990	3.29
Número de conductores en el hogar (4)		0.872	4.25
Indicador de si el trabajador es cabeza de familia (1)		0.627	3.37
Densidad de empleo en la ubicación del trabajo (1)		−0.0016	2.27
Ubicación del hogar en o cerca del distrito principal de negocios (1)		−0.502	4.18
Autos por conductor con techo 1 (1)		5.00	9.65
Autos por conductor con techo 1 (3)		2.33	2.74
Autos por conductor con techo 1 (4)		2.38	5.28
Indicador de sólo automóvil (1)		−5.26	5.93
Indicador de autobús con acceso en automóvil (3)		−5.49	5.33
Indicador de compartición de automóvil (4)		−3.84	6.36
Índice de ratio de verosimilitud	0.4426		
Log-verosimilitud en convergencia	−595.8		
Número de observaciones	771		
Valor del tiempo ahorrado como % del salario:			
Tiempo en automóvil	227		3.20
Tiempo en transporte público	91		2.43
Tiempo a pie	243		3.10
Tiempo de espera en transbordo	190		2.01

a La variable entra en los medios de transporte indicados en paréntesis y es cero en el resto de medios.
Medios de transporte: 1. Sólo Automóvil. 2. Autobús con acceso a pies. 3. Autobús con acceso en automóvil. 4. Automóvil compartido.

El costo del viaje se dividió por el salario del viajero para reflejar la expectativa de que los trabajadores con salarios más bajos van a estar más preocupados por el costo que los trabajadores mejor remunerados. El tiempo de viaje pasado en el vehículo entra en el modelo por separado para automóvil y autobús, con el fin de indicar que los viajeros podrían encontrar el tiempo empleado en el autobús más o menos molesto que el tiempo durante el que conducen un automóvil. Los viajes en autobús a menudo implican transbordos y estos transbordos pueden ser pesados para los viajeros. Es por ello que el modelo incluye el número de transbordos y el tiempo estimado de espera en los transbordos. El tiempo transcurrido entre el paso de dos autobuses para la primera línea de autobús que el pasajero tomaría se incluye como una medida de la cantidad máxima de tiempo que la persona tendría que esperar usando este medio.

Los coeficientes estimados de costo y de los diversos componentes de tiempo proporcionan información sobre el valor del tiempo. Por definición, el valor del tiempo es el costo extra en que una persona estaría dispuesta a incurrir para ahorrar tiempo. La utilidad toma la forma $U_{nj} = \alpha c_{nj}/w_n + \beta t_{nj} + \cdots$, donde c es el costo y t es el tiempo. La derivada total con respecto a los cambios en el tiempo y el costo es $dU_{nj} = (\alpha/w_n)dc_{nj} + \beta dt_{nj}$, que igualamos a cero y resolvemos para dc/dt para poder encontrar el

cambio en el costo que mantiene inalterada la utilidad de un cambio en el tiempo: $dc/dt = -(\beta/\alpha)w_n$. Por tanto, el valor del tiempo es una proporción β/α del salario de la persona. Los valores estimados del tiempo se presentan en la parte inferior de la tabla 3.1. El tiempo ahorrado al viajar en autobús es valorado como un 91% del salario ((-0.0259/-0.0284) × 100), mientras que el ahorro de tiempo de viaje en automóvil vale más del doble: 227% del salario. Esta diferencia sugiere que los pasajeros consideran la conducción considerablemente más pesada que ir en autobús, cuando se evalúa por minuto invertido en el viaje. Los viajeros parece que eligen el automóvil no porque les guste la conducción en sí, sino porque conducir es generalmente más rápido. Caminar se considera más molesto que esperar un autobús (243% del salario frente a 190%) y esperar un autobús es más molesto que viajar en él.

Los ingresos entran en la utilidad representativa de la alternativa "sólo automóvil". Entran de una forma lineal por tramos para permitir la posibilidad de que un ingreso adicional tenga un impacto diferente dependiendo del nivel general de ingresos. Ninguna de las variables de ingreso contribuye significativamente. Al parecer, dividir el costo del viaje por los salarios suprime cualquier efecto que el ingreso pudiera tener en la elección del medio de transporte de un viajero. Es decir, los salarios más altos inducen al viajero a preocuparse menos por los gastos del viaje, pero no inducen una predilección por la conducción más allá del impacto que tiene a través de los costos. El número de personas y el número de vehículos por conductor en el hogar tienen un efecto significativo en la elección del medio de transporte, tal y como se esperaba. Se han incluido también constantes específicas de alternativa, con la constante para la alternativa "autobús - a pie" normalizada a cero.

El modelo en la tabla 3.1 se utilizó para predecir la elección de medio de transporte de los pasajeros después de la inauguración del BART. Se consideró como conjunto de elección los cuatro medios de transporte enumerados anteriormente más dos modalidades del BART que se diferenciaban en función de si la persona necesitaba usar el autobús o conducir para llegar a la estación del BART. La tabla 3.2 presenta las cuotas de mercado previstas y reales para cada medio de transporte. La demanda del BART se estimó que sería de un 6.3%, en comparación con el 6.2% que realmente obtuvo. La estrecha correspondencia entre predicción y valor real es notable.

Tabla 3.2. Predicciones para post-apertura del BART

	Cuota real	Cuota prevista
Sólo automóvil	59.90	55.84
Autobús con acceso a pie	10.78	12.51
Autobús con acceso en automóvil	1.426	2.411
BART con acceso a pie	0.951	1.053
BART con acceso en automóvil	5.230	5.286
Automóvil compartido	21.71	22.89

Las cifras del cuadro 3.2 tienden a enmascarar varias complicaciones que surgieron en la predicción. Por ejemplo, caminar hasta la estación del BART fue incluido originalmente como un medio de transporte separado. El modelo predijo esta opción muy pobremente, sobre-estimando el número de personas que iban a caminar hasta el BART por un factor de doce. El problema fue investigado y se encontró que se debía principalmente a las diferencia entre la experiencia de caminar a las estaciones del BART y la de caminar hacia el autobús, debido a los barrios en los que se ubicaban las estaciones del BART. Estos problemas se analizan con mayor detenimiento por McFadden et al. (1977)

3.10 Obtención de las probabilidades logit

Se afirmó sin pruebas en la sección 3.1 que si el componente no observado de utilidad se distribuye con una densidad iid valor extremo para cada alternativa, las probabilidades de elección toman la forma de la ecuación (3.6). Nos proponemos a continuación demostrar este resultado. De (3.5) tenemos

$$P_{ni} = \int_{s=-\infty}^{\infty} \left(\prod_{j \neq i} e^{-e^{-(s+V_{ni}-V_{nj})}} \right) e^{-s} e^{-e^{-s}} ds$$

donde s es ε_{ni}. Nuestra tarea consiste en evaluar esta integral. Observando que $V_{ni} - V_{ni} = 0$ y agrupando términos en el exponente de e, tenemos

$$P_{ni} = \int_{s=-\infty}^{\infty} \left(\prod_{j} e^{-e^{-(s+V_{ni}-V_{nj})}} \right) e^{-s} ds$$

$$= \int_{s=-\infty}^{\infty} exp\left(-\sum_{j} e^{-(s+V_{ni}-V_{nj})} \right) e^{-s} ds$$

$$= \int_{s=-\infty}^{\infty} exp\left(-e^{-s} \sum_{j} e^{-(V_{ni}-V_{nj})} \right) e^{-s} ds$$

Definimos $t = \exp(-s)$ tal que $-\exp(-s)\, ds = dt$. Tenga en cuenta que a medida que s tiende a infinito, t se aproxima a cero, y cuando s se acerca a menos infinito, t se convierte infinitamente grande. Usando este nuevo término llegamos a

$$P_{ni} = \int_{\infty}^{0} exp\left(-t \sum_{j} e^{-(V_{ni}-V_{nj})} \right) (-dt)$$

$$= \int_{0}^{\infty} exp\left(-t \sum_{j} e^{-(V_{ni}-V_{nj})} \right) dt$$

$$= \left. \frac{exp\left(-t \sum_j e^{-(V_{ni}-V_{nj})} \right)}{-\sum_j e^{-(V_{ni}-V_{nj})}} \right|_{0}^{\infty}$$

$$= \frac{1}{\sum_j e^{-(V_{ni}-V_{nj})}} = \frac{e^{V_{ni}}}{\sum_j e^{V_{nj}}},$$

según se pretendía demostrar.

4

GEV

4.1 Introducción

El modelo logit estándar exhibe independencia de alternativas irrelevantes (IIA), lo que implica la sustitución proporcional entre alternativas. Como ya comentamos en el capítulo 3, esta propiedad puede ser vista como una restricción impuesta por el modelo o como el resultado natural de un modelo bien especificado, capaz de capturar en la utilidad representativa todas las fuentes de correlación entre alternativas, por lo que sólo queda ruido blanco como inobservado. A menudo, el investigador es incapaz de capturar todas las fuentes de correlación de forma explícita, por lo que las partes no observadas de utilidad están correlacionadas y la IIA no se sostiene. En estos casos se necesita un modelo más general que el logit estándar.

Los modelos generalizados de valor extremo (GEV) constituyen una amplia clase de modelos que exhiben patrones de sustitución variados. El atributo unificador de estos modelos es que las partes no observadas de la utilidad de todas las alternativas se distribuyen conjuntamente de acuerdo a una distribución generalizada de valor extremo. Esta distribución permite correlaciones entre alternativas y, como su propio nombre indica, es una generalización de la distribución de valor extremo univariante que se utiliza en los modelos logit estándar. Cuando todas las correlaciones son cero, la distribución GEV se convierte en el producto de distribuciones de valor extremo independientes y el modelo GEV se convierte en el logit estándar. Por consiguiente, esta clase de modelos incluye el logit simple pero también una variedad de otros modelos. Los test de hipótesis sobre las correlaciones dentro de un modelo GEV se pueden emplear para examinar si las correlaciones son cero, lo que es equivalente a probar si logit estándar proporciona una representación precisa de los patrones de sustitución.

El miembro más ampliamente utilizado de la familia GEV es el logit jerárquico o anidado. Este modelo ha sido aplicado por muchos investigadores en una gran variedad de situaciones, incluyendo la energía, el transporte, la vivienda, las telecomunicaciones y una multitud de otros campos; véase por ejemplo, Ben- Akiva (1973), Train (1986, capítulo 8), Train et al. (1987a), Forinash y Koppelman (1993) y Lee (1999). Su forma funcional es simple en comparación con otros tipos de modelos GEV y proporciona un conjunto flexible de posibles patrones de sustitución. Las secciones 4.2 y 4.3 describen la especificación y estimación de modelos logit jerárquicos. Esta descripción es útil en sí misma, dada la prominencia adquirida por los modelos logit jerárquicos, pero también como base para la comprensión de modelos GEV más complejos. En la Sección 4.4

pasamos a estudiar otros modelos GEV puestos en práctica por algunos investigadores, con especial énfasis en dos de los modelos más prometedores, a saber, el logit combinacional emparejado (*paired combinatorial logit*, PCL) y el logit jerárquico generalizado (*generalized nested logit*, GNL). La sección final del capítulo describe la clase completa de modelos GEV y cómo se generan las nuevas especificaciones de esta clase.

Sólo una pequeña parte de los posibles modelos GEV se han implementado alguna vez. Esto significa que aún no han sido plenamente explotadas todas las capacidades de esta clase de modelos y que nuevas investigaciones en esta área tienen el potencial de encontrar modelos aún más poderosos que los que ya se utilizan actualmente. Un ejemplo de este potencial lo proporciona Karlstrom (2001), que especificó un modelo GEV de una forma nunca empleada anteriormente y que encontró que se ajustaba a sus datos mejor que otros tipos de modelos GEV especificados previamente. Los modelos GEV tienen la ventaja de que las probabilidades de elección suelen tener una forma cerrada, de modo que pueden ser estimadas sin recurrir a la simulación. Sólo por esta razón, los modelos GEV seguirán siendo fuente de nuevas y potentes especificaciones para satisfacer las necesidades de los investigadores.

4.2 Logit jerárquico

4.2.1 Patrones de sustitución

Un modelo logit jerárquico es apropiado cuando el conjunto de alternativas a las que se enfrenta un decisor puede dividirse en subconjuntos, llamados nidos (*nests*), de tal manera que las siguientes propiedades se cumplen:

1. Para cualesquiera dos alternativas que están en el *mismo* nido, el ratio de probabilidades es independiente de los atributos o de la existencia de cualesquiera otras alternativas. Es decir, se verifica la IIA dentro de cada nido.
2. Para cualesquiera dos alternativas en *diferentes* nidos, el ratio de probabilidades puede depender de los atributos de otras alternativas en los dos nidos. La IIA no se sostiene, en general, entre alternativas en diferentes nidos.
3.

Un ejemplo puede explicar mejor si un conjunto de alternativas puede ser objeto de una partición de este tipo. Supongamos que el conjunto de alternativas disponibles para un trabajador para ir a su lugar de trabajo consisten en conducir un automóvil en solitario, compartir un automóvil, usar el autobús y usar el tren. Si se elimina cualquiera de estas alternativas, las probabilidades de elección del resto de alternativas se incrementaría (por ejemplo, si el automóvil del trabajador está siendo reparado, por lo que no puede conducir al trabajo por sí mismo, las probabilidades de compartir automóvil, ir en autobús e ir en ferrocarril aumentarían). La pregunta relevante al dividir estas alternativas en grupos es: si quitase una alternativa, ¿en qué proporción aumentaría la probabilidad de cada una de las restantes alternativas? Supongamos que los cambios en las probabilidades se producen como se indica en la tabla 4.1.

Tabla 4.1. Ejemplo de cumplimiento de la IIA dentro de los nidos de alternativas: cambio en las probabilidades cuando se suprime una alternativa

	Probabilidad				
		Con alternativa suprimida			
Alternativa	Original	Automóvil	Automóvil compartido	Autobús	Tren
Automóvil	.40	—	.45 (+12.5%)	.52 (+30%)	.48 (+20%)
Automóvil compartido	.10	.20 (+100%)	—	.13 (+30%)	.12 (+20%)
Autobús	.30	.48 (+60%)	.33 (+10%)	—	.40 (+33%)
Tren	.20	.32 (+60%)	.22 (+10%)	.35 (+70%)	—

Obsérvese que las probabilidades del autobús y el tren siempre aumentan en la misma proporción cada vez que se elimina una de las otras alternativas. Por lo tanto, la IIA se mantiene entre estas dos alternativas. Pongamos estas dos alternativas en un nido y llamémosle "transporte público". Del mismo modo, la probabilidad del automóvil (en solitario) y el automóvil compartido crecen en la misma proporción cada vez que una de las otras alternativas se elimina. La IIA se mantiene entre estas dos alternativas, por lo que las ponemos en un nido llamado "automóviles". La IIA no se mantiene entre una alternativa del nido “automóviles” y una del nido “transporte público”. Por ejemplo, cuando se elimina la alternativa automóvil, la probabilidad de compartir automóvil se eleva proporcionalmente más que la probabilidad del autobús o del tren. Con nuestros dos nidos, podemos enunciar las pautas de sustitución de forma resumida como: la IIA se mantiene dentro de cada nido, pero no entre nidos. Un modelo logit jerárquico con las dos alternativas de automóvil en un nido y las dos alternativas de transporte público en otro nido, es apropiado para representar esta situación.

Una manera conveniente de representar los patrones de sustitución es con un diagrama en árbol. En el árbol, cada rama representa un subconjunto de alternativas dentro de las cuales la IIA se mantiene y cada hoja de cada rama indica una alternativa. Por ejemplo, el diagrama en árbol de la elección que el trabajador hace del medio de transporte que acabamos de describir se muestra en la figura 4.1. El árbol (de arriba a abajo) se compone de dos ramas, denominadas "automóviles" y "transporte público", para los dos subconjuntos de alternativas, y cada una de las ramas contiene dos sub-ramas para las dos alternativas dentro del subconjunto. Existe sustitución proporcional entre sub-ramas pero no entre ramas.

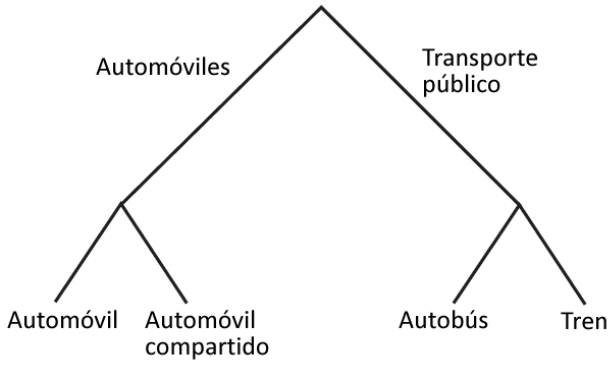


Figura 4.1. Diagrama en árbol para la elección de medio de transporte.

4.2.2 Probabilidades de elección

Daly y Zachary (1978), McFadden (1978) y Williams (1977) mostraron, de forma independiente y utilizando diferentes pruebas, que el modelo logit jerárquico es consistente con la maximización de la utilidad. Supongamos que dividimos el conjunto de alternativas j en K subconjuntos no solapados $B_1, B_2, \ldots, B_K$ y los llamamos nidos. La utilidad que una persona n obtiene de la alternativa j en el nido

B_K se denota, como de costumbre, como $U_{nj} = V_{nj} + \varepsilon_{nj}$, donde V_{nj} es observada por el investigador y ε_{nj} es una variable aleatoria cuyo valor no es observado por el investigador. El modelo logit jerárquico se obtiene asumiendo que el vector de utilidad no observada $\varepsilon_n = \langle \varepsilon_{n1}, \dots, \varepsilon_{nJ} \rangle$, tiene distribución acumulativa

$$exp\left(-\sum_{k=1}^{K}\left(\sum_{j\in B_k} e^{-\varepsilon_{nj}/\lambda_k}\right)^{\lambda_k}\right). \qquad (4.1)$$

Esta distribución es un tipo de distribución GEV. Es una generalización de la distribución que da lugar al modelo logit. Para logit, cada ε_{nj} es independiente con una distribución univariante valor extremo. Para esta GEV, la distribución marginal de cada ε_{nj} es de tipo valor extremo univariante. Sin embargo, los ε_{nj}s se correlacionan dentro de los nidos. Para cualesquiera dos alternativas j y m en el nido B_K, ε_{nj} se correlaciona con ε_{nm}. Para cualesquiera dos alternativas en diferentes nidos, la parte no observada de la utilidad sigue estando no correlacionada: $Cov(\varepsilon_{nj}, \varepsilon_{nm}) = 0$ para cualquier $j \in B_K$ y $m \in B_l$ con $l \neq k$.

El parámetro λ_k es una medida del grado de independencia de la utilidad no observada entre las alternativas dentro del nido k. Un valor más alto de λ_k significa mayor independencia y menor correlación. El estadístico $1 - \lambda_k$ es una medida de correlación, en el sentido de que a medida que λ_k aumenta, indicando una menor correlación, este estadístico cae. Como señala McFadden (1978), la correlación en realidad es más compleja que $1 - \lambda_k$, pero $1 - \lambda_k$ se puede utilizar como una indicación de la correlación. Un valor de $\lambda_k = 1$ indica independencia completa dentro del nido k, es decir, que no hay correlación. Cuándo $\lambda_k = 1$ para todo k, representando independencia entre todas las alternativas en todos los nidos, la distribución GEV se convierte en el producto de términos tipo valor extremo independientes, cuya distribución está expresada en (3.2). En este caso, el modelo logit jerárquico se reduce al modelo logit estándar.

Como han mostrado los autores citados anteriormente, la distribución de los componentes no observados de la utilidad da lugar a la siguiente probabilidad de elección para la alternativa $i \in B_k$:

$$P_{ni} = \frac{e^{V_{ni}/\lambda_k}\left(\sum_{j\in B_k} e^{V_{nj}/\lambda_k}\right)^{\lambda_k - 1}}{\sum_{l=1}^{K}\left(\sum_{j\in B_l} e^{V_{nj}/\lambda_l}\right)^{\lambda_l}}. \qquad (4.2)$$

Podemos utilizar esta fórmula para demostrar que la IIA se mantiene dentro de cada subconjunto de alternativas, pero no entre subconjuntos. Considere las alternativas $i \in B_k$ y $m \in B_l$. Dado que el denominador de (4.2) es el mismo para todas las alternativas, el ratio de probabilidades es el ratio de numeradores:

$$\frac{P_{ni}}{P_{nm}} = \frac{e^{V_{ni}/\lambda_k}\left(\sum_{j\in B_k} e^{V_{nj}/\lambda_k}\right)^{\lambda_k - 1}}{e^{V_{nm}/\lambda_l}\left(\sum_{j\in B_l} e^{V_{nj}/\lambda_l}\right)^{\lambda_l - 1}}.$$

Si $k = l$ (es decir, i y m están en el mismo nido) entonces los factores que aparecen entre paréntesis se anulan y obtenemos

$$\frac{P_{ni}}{P_{nm}} = \frac{e^{V_{ni}/\lambda_k}}{e^{V_{nm}/\lambda_l}}.$$

Este ratio es independiente de todas las otras alternativas. Para $k \neq l$ (es decir, i y m están en diferentes nidos), los factores que aparecen entre paréntesis no se anulan. La relación de probabilidades

depende de los atributos de todas las alternativas presentes en los nidos que contienen i y m. Obsérvese, sin embargo, que la relación no depende de los atributos de las alternativas de los nidos que no contienen i y m. Por lo tanto, un tipo de IIA se mantiene incluso para alternativas en diferentes nidos. Este tipo de IIA puede describirse en términos generales como "independencia de los nidos irrelevantes" o IIN. Con un modelo logit jerárquico, se cumple la IIA sobre alternativas dentro de cada nido y la IIN sobre alternativas de diferentes nidos. Esta propiedad de los modelos logit jerárquicos se refuerza en la siguiente sección en la que descomponemos la probabilidad de elección del logit jerárquico en dos probabilidades logit estándar.

Cuando $\lambda_k = 1$ para todo k (y por tanto $1 - \lambda_k = 0$), indicando que no hay correlación entre los componentes no observados de la utilidad de alternativas dentro de un mismo nido, las probabilidades de elección se convierten simplemente en logit. El modelo logit jerárquico es una generalización del logit que permite un patrón particular de correlación en la utilidad no observada.

El parámetro λ_k puede variar para diferentes nidos, lo que refleja distintas correlaciones entre los factores no observados dentro de cada nido. El investigador puede restringir los λ_ks para que sean el mismo para todos (o algunos) nidos, indicando que la correlación es la misma en cada uno de estos nidos. Es posible usar un test de hipótesis para determinar si aplicar restricciones sobre los valores de las λ_ks es razonable. Así, por ejemplo, hacer un test sobre la restricción $\lambda_k = 1 \; \forall k$ es equivalente a probar si el modelo logit estándar es una especificación razonable para el problema de elección frente al modelo logit jerárquico más general. Estos test se llevan a cabo más fácilmente con el estadístico de ratio de verosimilitud descrito en la Sección 3.8.2.

El valor de λ_k debe estar dentro de un rango determinado para que el modelo sea consistente con el comportamiento de maximización de la utilidad. Si $\lambda_k \; \forall k$ está entre cero y uno, el modelo es consistente con la maximización de la utilidad para todos los posibles valores de las variables explicativas. Para λ_k mayor que uno, el modelo es coherente con el comportamiento maximizador de la utilidad de algún rango de las variables explicativas, pero no para todos los posibles valores. Kling y Herriges (1995) y Herriges y Kling (1996) proporcionan pruebas de la consistencia del logit jerárquico con la maximización de la utilidad cuando $\lambda_k > 1$; y Train et al. (1987a) y Lee (1999) ofrecen ejemplos de modelos para los cuales $\lambda_k > 1$. Un valor negativo de λ_k es inconsistente con la maximización de la utilidad e implica que la mejora de los atributos de una alternativa (como la reducción de su precio) podría disminuir la probabilidad de que la alternativa fuese elegida. Con λ_k positivo, el modelo logit jerárquico se aproxima al modelo de "eliminación por aspectos" (*elimination by aspects*) de Tversky (1972) a medida que $\lambda_k \to 0$.

En la notación que hemos estado utilizando, cada λ_k es un parámetro fijo, lo que implica que todos los decisores tienen las mismas correlaciones entre los factores no observados. En realidad, las correlaciones podrían diferir entre decisores en base a sus características observadas. Para dar cabida a esta posibilidad, cada λ_k puede ser especificada como una función paramétrica de datos demográficos observados u otras variables, siempre y cuando la función mantenga un valor positivo. Por ejemplo, Bhat (1997a) especifica $\lambda = \exp(\alpha z_n)$, donde z_n es un vector de características del decisor n y α es un vector de parámetros a estimar, junto con los parámetros que entran en la utilidad representativa. La transformación exponencial asegura que λ es positivo.

4.2.3 La descomposición en dos logits

La expresión (4.2) no es muy esclarecedora como fórmula. Sin embargo, las probabilidades de elección pueden ser expresadas de una forma alternativa bastante simple y fácilmente interpretable. Sin pérdida de generalidad, el componente observado de utilidad se puede descomponer en dos partes: (1) una parte etiquetada como W constante para todas las alternativas dentro de un nido y (2) una parte etiquetada como Y que varía entre alternativas dentro de un nido. La utilidad se rescribe como

(4.3) $$U_{nj} = W_{nk} + Y_{nj} + \varepsilon_{nj},$$

para $j \in B_k$, donde:

- W_{nk} sólo depende de variables que describen el nido k. Estas variables difieren entre los nidos pero no entre las alternativas de cada nido.
- Y_{nj} depende de variables que describen la alternativa j. Estas variables varían entre alternativas dentro del nido k.

Observe que esta descomposición es totalmente general, dado que para cualquier W_{nk}, Y_{nj} se define como $V_{nj} - W_{nk}$.

Con esta descomposición de la utilidad, la probabilidad logit jerárquica puede escribirse como el producto de dos probabilidades logit estándar. Expresemos la probabilidad de elegir la alternativa $i \in B_k$ como el producto de dos probabilidades: la probabilidad de elegir una alternativa dentro del nido B_k y la probabilidad de elegir concretamente la alternativa i, condicionada a que una alternativa de B_k ha sido escogida:

$$P_{ni} = P_{ni|B_k} P_{nB_k},$$

donde $P_{ni|B_k}$es la probabilidad condicionada de elegir la alternativa i dado que se ha elegido una alternativa del nido B_k y P_{nB_k}es la probabilidad marginal de elegir una alternativa en el nido B_k (con la marginalidad aplicada sobre todas las alternativas en B_k). Esta igualdad es exacta, ya que cualquier probabilidad se puede escribir como el producto de una probabilidad marginal y una probabilidad condicionada.

La razón para descomponer P_{ni} en una probabilidad marginal y una condicionada es que, con la fórmula logit jerárquica descrita para P_{ni}, las probabilidades marginales y condicionadas toman la forma de logits. En particular, estas probabilidades se pueden expresar como

(4.4) $$P_{nB_k} = \frac{e^{W_{nk} + \lambda_k I_{nk}}}{\sum_{l=1}^{K} e^{W_{nl} + \lambda_l I_{nl}}},$$

(4.5) $$P_{ni|B_k} = \frac{e^{Y_{ni}/\lambda_k}}{\sum_{j \in B_k} e^{Y_{nj}/\lambda_k}},$$

donde

$$I_{nk} = ln \sum_{j \in B_k} e^{Y_{nj}/\lambda_k}.$$

La obtención de estas expresiones a partir de la probabilidad de elección (4.2) es una simple cuestión de reordenamiento algebraico. Para los lectores interesados, se proporciona en la sección 4.2.5.

Expresado en palabras, la probabilidad de elegir una alternativa en B_k toma la forma de la fórmula logit, como si fuese el resultado de un modelo para una elección entre nidos. Esta probabilidad incluye variables W_{nk} que varían entre nidos pero no entre alternativas dentro de cada nido. También incluye una cantidad llamada I_{nk}, cuyo significado dilucidaremos posteriormente. La probabilidad de elegir i condicionada al hecho de que una alternativa de B_k ha sido seleccionada, también está dada por una fórmula logit, como si fuese el resultado de un modelo de elección entre las alternativas dentro del nido. Esta probabilidad condicionada incluye las variables Y_{nj} que varían entre alternativas dentro del nido.

Tenga en cuenta que estas variables se dividen por λ_k, de modo que cuando Y_{nj} es lineal en los parámetros, los coeficientes que entran en esta probabilidad condicionada son los coeficientes originales divididos por λ_k. Es habitual referirse a la probabilidad marginal (elección del nido) como al *modelo superior* (*upper model*) y a la probabilidad condicionada (elección de alternativa dentro del nido) como el modelo inferior (*lower model*), lo que refleja sus posiciones relativas en la figura 4.1.

La cantidad I_{nk} enlaza los modelos superior e inferior al traer información desde el modelo inferior hacia el modelo superior. Ben-Akiva (1973) identificó por primera vez la fórmula correcta para este enlace. En particular, I_{nk} es el logaritmo del denominador del modelo inferior. Esta fórmula tiene un significado importante. Recuerde de la explicación sobre el excedente del consumidor en un modelo logit (sección 3.5) que el logaritmo del denominador del modelo logit es la utilidad esperada que el decisor obtiene de la situación de elección, como muestran Williams (1977) y Small y Rosen (1981). La misma interpretación aplica aquí: $\lambda_k I_{nk}$ es la utilidad esperada que recibe el decisor n de la elección entre las alternativas del nido B_k. La fórmula para la utilidad esperada es igual a la de un modelo logit porque, condicionada a un nido, la elección de las alternativas dentro del nido es de hecho un logit, tal y como se muestra en la ecuación (4.5). I_{nk} es a menudo llamado *valor inclusivo* o *utilidad inclusiva* del nido B_k. También se le llama "término log-suma", porque es el logaritmo de una suma (de utilidades representativas exponenciadas). El término "precio inclusivo" también se utiliza en ocasiones; sin embargo, en todo caso sería el negativo de I_{nk} lo que se asemejaría más a un precio.

El coeficiente λ_k de I_{nk} en el modelo superior es a menudo llamado el coeficiente log-suma. Como se ha mencionado, λ_k refleja el grado de independencia entre las partes no observadas de la utilidad de las alternativas del nido B_k, con una λ_k menor indicando menor independencia (mayor correlación).

Es apropiado que el valor inclusivo entre como variable explicativa en el modelo superior. Dicho de forma general, la probabilidad de elegir el nido B_k depende de la utilidad esperada que la persona recibe de ese nido. Esta utilidad esperada incluye la utilidad que recibe sin importar qué alternativa elige dentro del nido, que es W_{nk}, además de la utilidad adicional esperada que recibe por ser capaz de elegir la mejor alternativa dentro del nido, que es $\lambda_k I_{nk}$.

Recordemos que los coeficientes que entran en el modelo inferior se dividen por λ_k, como se indica en la ecuación (4.5). Estos mismos modelos han sido especificados y estimados sin dividir por λ_k en el modelo inferior. Daly (1987) y Greene (2000) describen un modelo de este tipo y el paquete de software STATA lo incluye en su modelo de logit jerárquico en el comando *nlogit*. El paquete NLOGIT permite cualquier especificación. Si los coeficientes en el modelo inferior no se dividen por λ_k, las probabilidades de elección no son iguales a las expresadas en la ecuación (4.2). Como se muestra en la obtención de estas fórmulas en la sección 4.2.5, se necesita la división por λ_k para que el producto de las probabilidades condicionadas y marginales sean iguales a las probabilidades logit jerárquicas dadas por la ecuación (4.2). Sin embargo, el hecho de que el modelo no dé las probabilidades de la ecuación (4.2) no necesariamente significa que sea inadecuado. Koppelman y Wen (1998) y Hensher y Greene (2002) comparan los dos enfoques (dividir por λ_k frente a no dividir) y muestran que este último modelo no es consistente con la maximización de la utilidad cuando algún coeficiente es común entre nidos (como un coeficiente de costo que sea el mismo para los medios de transporte autobús y automóvil). Heiss (2002) señala la relación inversa: si no hay coeficientes comunes entre nidos, el segundo modelo es consistente con la maximización de la utilidad, ya que la división necesaria por λ_k en cada nido se lleva a cabo de manera implícita (en lugar de explícitamente) al permitir coeficientes separados en cada uno de los nidos, de tal manera que la escala de los coeficientes difiera entre nidos. Por el contrario, cuando los coeficientes son comunes entre los nidos, Heiss encontró que no dividir por λ_k conlleva implicaciones contrarias a la intuición.

4.2.4 Estimación

Los parámetros de un modelo jerárquico se pueden estimar mediante técnicas estándar de máxima verosimilitud. Sustituyendo las probabilidades de elección de la expresión (4.2) en la función de log - verosimilitud da una función explícita de los parámetros de este modelo. Los valores de los parámetros que maximizan esta función son, bajo condiciones bastante generales, consistentes y eficientes (Brownstone y Small, 1989).

Existen rutinas en paquetes de software comercial para la estimación de modelos jerárquicos por máxima verosimilitud. Hensher y Greene (2002) ofrecen una guía para estimar logits jerárquicos utilizando software disponible. La maximización numérica a veces es difícil, ya que la función log-verosimilitud no es globalmente cóncava e incluso en áreas cóncavas dista mucho de ser cuadrática. El investigador puede necesitar ayudar a las rutinas probando diferentes algoritmos y/o distintos valores de inicio, como se trata en el Capítulo 8.

En lugar de utilizar máxima verosimilitud, los modelos logit jerárquicos pueden ser estimados consistentemente (pero no de manera eficiente) de una manera secuencial, explotando el hecho de que las probabilidades de elección se pueden descomponer en probabilidades marginales y condicionadas que son logit. Esta estimación secuencial se realiza "de abajo hacia arriba". En primer lugar se estiman los modelos inferiores (para la elección de la alternativa dentro de un nido). Usando los coeficientes estimados, el valor inclusivo se calcula para cada modelo inferior. A continuación, se estima el modelo superior (para la elección del nido) con el valor inclusivo entrando como variable explicativa.

La estimación secuencial presenta dos dificultades que desaconsejan su uso. En primer lugar, los errores estándar de los parámetros de los modelos superiores están sesgados a la baja, como Amemiya (1978) señaló por primera vez. Este sesgo surge debido a que la varianza de la estimación del valor inclusivo que entra en el modelo superior no se incorpora en el cálculo de los errores estándar. Con errores estándar sesgados a la baja, los intervalos de confianza se estiman inferiores a la realidad y los estadísticos-t mayores, por lo que el modelo superior parecerá ser mejor de lo que realmente es. Ben-Akiva y Lerman (1985, p. 298) proporcionan un procedimiento para ajustar los errores estándar para eliminar el sesgo.

En segundo lugar, por lo general es habitual que algunos parámetros aparezcan en varios sub-modelos. La estimación de los diversos modelos superiores e inferiores por separado proporciona estimaciones independientes de cualquier parámetro común que aparezca en el modelo. La estimación simultánea por máxima verosimilitud asegura que los parámetros comunes son forzados a ser los mismos dondequiera que aparezcan en el modelo.

Estas dos complicaciones son síntomas de una circunstancia más general, a saber, que la estimación secuencial de modelos logit jerárquicos, aunque coherente, no es tan eficiente como la estimación simultánea por máxima verosimilitud. Con estimación simultánea, toda la información se utiliza en la estimación de cada parámetro y los parámetros que son comunes entre componentes necesariamente son obligados a ser iguales. Dado que hay software comercial disponible para la estimación simultánea, existen pocas razones para estimar secuencialmente un logit jerárquico. Si surgen problemas en la estimación simultánea, el investigador podría considerar útil estimar el modelo secuencialmente y luego usar las estimaciones secuenciales como valores iniciales en la estimación simultánea. El principal valor de la descomposición del logit jerárquico en sus componentes superior e inferior no proviene de su uso como una herramienta de estimación sino más bien como un dispositivo heurístico: la descomposición ayuda en gran medida a la comprensión del significado y la estructura del modelo logit jerárquico.

4.2.5 Equivalencia de las fórmulas del logit jerárquico

Hemos afirmado en la sección 4.2.3 que el producto de las probabilidades marginales y condicionadas en (4.4) y (4.5) es igual a la probabilidad conjunta de (4.2). Vamos a verificar esta afirmación:

$$P_{ni} = \frac{e^{V_{ni}/\lambda_k}\left(\sum_{j\in B_k} e^{V_{nj}/\lambda_k}\right)^{\lambda_k - 1}}{\sum_{l=1}^{K}\left(\sum_{j\in B_l} e^{V_{nj}/\lambda_l}\right)^{\lambda_l}} \; por\ (4.2)$$

$$= \frac{e^{V_{ni}/\lambda_k}}{\sum_{j\in B_k} e^{V_{nj}/\lambda_k}} \frac{\left(\sum_{j\in B_k} e^{V_{nj}/\lambda_k}\right)^{\lambda_k}}{\sum_{l=1}^{K}\left(\sum_{j\in B_l} e^{V_{nj}/\lambda_l}\right)^{\lambda_l}}$$

$$= \frac{e^{(W_{nk}+Y_{ni})/\lambda_k}}{\sum_{j\in B_k} e^{(W_{nk}+Y_{nj})/\lambda_k}} \frac{\left(\sum_{j\in B_k} e^{(W_{nk}+Y_{nj})/\lambda_k}\right)^{\lambda_k}}{\sum_{l=1}^{K}\left(\sum_{j\in B_l} e^{(W_{nl}+Y_{nj})/\lambda_l}\right)^{\lambda_l}}$$

$$= \frac{e^{W_{nk}/\lambda_k} e^{Y_{ni}/\lambda_k}}{e^{W_{nk}/\lambda_k}\sum_{j\in B_k} e^{Y_{nj}/\lambda_k}} \frac{e^{W_{nk}}\left(\sum_{j\in B_k} e^{Y_{nj}/\lambda_k}\right)^{\lambda_k}}{\sum_{l=1}^{K} e^{W_{nl}}\left(\sum_{j\in B_l} e^{Y_{nj}/\lambda_l}\right)^{\lambda_l}}$$

$$= \frac{e^{Y_{ni}/\lambda_k}}{\sum_{j\in B_k} e^{Y_{nj}/\lambda_k}} \frac{e^{W_{nk}+\lambda_k I_{nk}}}{\sum_{l=1}^{K} e^{W_{nl}+\lambda_l I_{nl}}}$$

$$= P_{ni|B_k} P_{nB_k},$$

donde la penúltima igualdad se debe a que $I_{nk} = \ln \sum_{j\in B_k} e^{Y_{nj}/\lambda_k}$, reconociendo que $e^x b^c = e^{x + c \ln b}$.

4.3 Logit jerárquico de tres niveles

El modelo logit jerárquico que hemos visto hasta este punto se denomina modelo logit jerárquico de dos niveles, ya que hay dos niveles de modelado: las probabilidades marginales (modelo superior) y las probabilidades condicionadas (modelos inferiores). En el caso de la elección del medio de transporte, los dos niveles son: el modelo marginal de automóviles frente a transporte público y los modelos condicionados de tipo de automóvil o tipo de transporte público (automóvil en solitario o compartido condicionado a la elección de automóvil, y autobús o tren condicionado a la elección de transporte público).

En algunas situaciones es apropiado emplear modelos logit jerárquicos de tres o más niveles. Los modelos de tres niveles se obtienen al dividir el conjunto de alternativas en nidos y luego dividir cada nido en sub-nidos. La fórmula de la probabilidad de elección en este caso es una generalización de (4.2) con las sumas adicionales para los sub-nidos dentro de las sumas de los nidos. Ver McFadden (1978) o Ben-Akiva y Lerman (1985) para la fórmula.

Al igual que con un logit jerárquico de dos niveles, las probabilidades de elección para un modelo de tres niveles se pueden expresar como una serie de logits. El modelo superior describe la elección del nido; las modelos intermedios describen la elección de sub-nidos dentro de cada nido; y los modelos inferiores describen la elección de alternativas dentro de cada sub-nido. El modelo superior incluye un valor inclusivo para cada nido. Este valor representa la utilidad esperada que el decisor puede obtener de los sub-nidos dentro del nido. Se calcula como el logaritmo del denominador del modelo intermedio para ese nido. Del mismo modo, los modelos intermedios incluyen un valor inclusivo para cada sub-nido, que representa la utilidad esperada que el decisor puede obtener de las alternativas dentro del sub-nido. Se calcula como el logaritmo del denominador del modelo inferior para cada sub-nido.

Como ejemplo, considere la elección que una familia (hogar) hace de su vivienda dentro de un área metropolitana. La familia elige una entre todas las viviendas disponibles en la ciudad. Las viviendas están disponibles en diferentes barrios de la ciudad y con diferente número de habitaciones. Es razonable asumir que hay factores no observados que son comunes a todas las viviendas en el mismo barrio, tales como la proximidad a tiendas y a lugares de entretenimiento. Por lo tanto, se espera que la parte no observada de la utilidad esté correlacionada entre todas las viviendas dentro de un barrio determinado. También hay factores no observados que son comunes a todas las viviendas con el mismo número de habitaciones, como por ejemplo la comodidad para trabajar en casa. Por lo tanto, esperamos que la utilidad no observada esté aún más correlacionada entre viviendas del mismo tamaño en el mismo barrio que entre viviendas de diferente tamaño en el mismo barrio. Este patrón de correlación puede ser representado por la anidación de las viviendas por barrio y posteriormente sub-anidando por número de dormitorios. Un diagrama en árbol que representa esta situación se puede ver en la figura 4.2 para San Francisco. Hay tres niveles de sub-modelos: la probabilidad de elección de barrio, la probabilidad de elección de número de dormitorios, dado el barrio, y la elección de la vivienda dada la elección de vecindad y número de dormitorios.

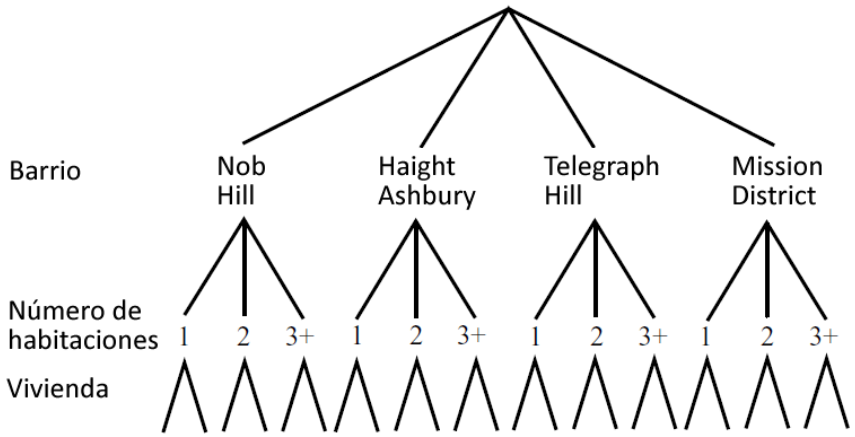


Figura 4.2. Logit jerárquico de tres niveles.

Un modelo logit jerárquico con esta estructura de anidación soporta la IIA de las siguientes formas:

1. El ratio de probabilidades de dos viviendas en el mismo barrio y con el mismo número de dormitorios es independiente de las características de todas las demás viviendas. Por ejemplo, la reducción del precio de un apartamento de dos dormitorios en Haight Ashbury extrae probabilidad de elección proporcionalmente de todas las viviendas de un dormitorio en Telegraph Hill.
2. El ratio de probabilidades de dos viviendas en el mismo barrio, pero con diferente número de habitaciones es independiente de las características de las viviendas en otros barrios, pero depende de las características de las viviendas en el mismo barrio que tienen el mismo número de dormitorios que cualquiera de estas viviendas. Bajar el precio de un apartamento de dos dormitorios en Haight Ashbury extrae probabilidad de elección proporcionalmente de viviendas de uno y dos dormitorios en Telegraph Hill, pero extrae de manera desproporcional de viviendas de dos dormitorios en Haight Ashbury en comparación a viviendas de un dormitorio en Haight Ashbury.

3. El ratio de probabilidades de dos viviendas en distintos barrios depende de las características de todas las demás viviendas en esos barrios, pero no de las características de viviendas en otros barrios. Bajar el precio de un apartamento de dos dormitorios en Haight Ashbury extrae probabilidad de elección proporcionalmente de todas las viviendas fuera de Haight Ashbury pero extrae de manera desproporcional de viviendas en Haight Ashbury en relación a las viviendas fuera de Haight Ashbury.

Cada nivel de agrupamiento en un logit jerárquico introduce parámetros que representan el grado de correlación entre alternativas dentro de los nidos. Con el conjunto completo de alternativas dividido en nidos, el parámetro λ_k se introduce para el nido k, como se ha descrito para los modelos de dos niveles. Si los nidos se dividen además en sub-nidos, entonces un parámetro σ_{mk} se introduce para el sub-nido m del nido k. Al descomponer la probabilidad en una serie de modelos logit, σ_{mk} es el coeficiente del valor inclusivo en el modelo intermedio y $\lambda_k\sigma_{mk}$ es el codeficiente del valor inclusivo en el modelo superior. Del mismo modo que para un logit jerárquico de dos niveles, los valores de estos parámetros deben encontrarse dentro de ciertos rangos para ser consistentes con la maximización de la utilidad. Si $0 < \lambda_k < 1$ y $0 < \sigma_{mk} < 1$, el modelo es consistente con la maximización de la utilidad para todos los niveles de las variables explicativas. Un valor negativo para los parámetros es incompatible con la maximización de la utilidad. Y valores superiores a uno son consistentes sólo para un cierto rango de valores de las variables explicativas.

4.4 Solapamiento de nidos

Para los modelos logit jerárquicos que hemos considerado, cada alternativa forma parte de un único nido (y para los modelos de tres niveles, un único sub-nido). Este aspecto de los modelos logit jerárquicos es una restricción en ocasiones inadecuada. Por ejemplo, en nuestro caso de elección de medio de transporte, ponemos los medios de transporte de automóvil en solitario y compartido en un mismo nido ya que tienen algunos atributos no observados similares. Sin embargo, compartir automóvil también tiene algunos atributos no observados que son similares a los del autobús y el tren, como la falta de flexibilidad en la planificación (el trabajador no puede ir a trabajar en cualquier momento del día, sino que tiene que ir en el momento en que el viaje compartido ha sido pactado, de manera similar a tomar una línea de autobús o tren con salidas fijadas). Sería útil tener un modelo en el que la utilidad no observada para la alternativa automóvil compartido pudiese estar correlacionada con la de viajar solo en automóvil y al mismo tiempo estar correlacionada, aunque en un grado diferente, con las alternativas autobús y tren. Dicho de manera equivalente, sería útil poder incluir la alternativa vehículo compartido en dos nidos: en un nido junto a la opción de viajar en automóvil en solitario y en otro nido junto a las opciones de viajar en autobús y tren.

Varios tipos de modelos GEV se han especificado con nidos solapados, de manera que una alternativa pueda ser miembro de más de un nido. Vovsha (1997), Bierlaire (1998) y Ben-Akiva y Bierlaire (1999) han propuesto diversos modelos llamados logits jerárquicos cruzados (*cross-nested logits*, CNLs) que contienen múltiples nidos solapados. Small (1987) consideró una situación en la que las alternativas tienen un orden natural, como por ejemplo el número de automóviles que posee un hogar (0, 1, 2, 3...) o la destinación para ir de compras, con las zonas de tiendas ordenadas por la distancia desde la residencia del decisor. Small especificó un modelo, llamado modelo ordenado generalizado de valor extremo (*ordered generalized extreme value*, OGEV) en el que la correlación en utilidad no observada entre dos alternativas depende de su proximidad en la ordenación. Este modelo tiene nidos solapados como en el CNL, pero cada nido consta de dos alternativas y se impone un patrón en las correlaciones (mayor correlación para los pares más cercanos). Small (1994) y Bhat (1998b) describieron una versión anidada del OGEV, que es similar a un logit jerárquico excepto en que los modelos inferiores (modelos de elección de alternativas dada la elección de un nido) son OGEV en lugar de logit estándar. Chu (1981,

1989) propuso un modelo llamado logit combinacional emparejado (*paired combinatorial logit* , PCL) en el que cada par de alternativas constituye un nido con su propia correlación. Con J alternativas, cada alternativa es miembro de $J-1$ nidos, y la correlación de su utilidad no observada con cada una de las otras alternativas puede ser estimada. Wen y Koppelman (2001) han desarrollado un modelo logit jerárquico generalizado (*generalized nested logit*, GNL) que incluye cruzados como casos especiales el PCL y otros modelos jerárquicos. En los apartados siguientes describo los modelos PCL y GNL, el primero debido a su simplicidad y el segundo debido a su generalidad.

4.4.1 Logit combinacional emparejado (PCL)

Cada par de alternativas se considera que es un nido. Dado que cada alternativa está emparejada con cada una de las otras alternativas, cada alternativa es miembro de $J-1$ nidos. Un parámetro etiquetado como λ_{ij} indica el grado de independencia entre las alternativas i y j. Dicho de forma equivalente: $1-\lambda_{ij}$ es una medida de la correlación existente entre la utilidad no observada de la alternativa i y de la alternativa j. Este parámetro es análogo a λ_k en un modelo logit jerárquico, donde λ_k indica el grado de independencia entre alternativas dentro del nido y $1-\lambda_k$ es una medida de correlación dentro del nido. Y como sucede con el logit jerárquico, el modelo PCL se convierte en un logit estándar cuando $\lambda_{ij}=1$ para todos los pares de alternativas.

Las probabilidades de elección para el modelo PCL son

$$P_{ni}=\frac{\sum_{j\neq i}e^{V_{ni}/\lambda_{ij}}(e^{V_{ni}/\lambda_{ij}}+e^{V_{nj}/\lambda_{ij}})^{\lambda_{ij}-1}}{\sum_{k=1}^{J-1}\sum_{l=k+1}^{J}(e^{V_{nk}/\lambda_{kl}}+e^{V_{nl}/\lambda_{kl}})^{\lambda_{kl}}}. \tag{4.6}$$

La suma en el numerador es sobre el total de $J-1$ nidos en los que la alternativa i se encuentra. Para cada uno de estos nidos, el término que se añade a la suma es el mismo que el numerador de la probabilidad logit jerárquica (4.2). Así, el modelo PCL es como el logit jerárquico salvo que permite que i esté en más de un nido. El denominador en la probabilidad PCL también tiene la misma forma de un logit jerárquico: es la suma sobre todos los nidos de la suma de los $exp(V/\lambda)$s dentro del nido, elevado a la potencia adecuada λ. Si λ_{ij} está entre cero y uno para todos los pares ij, el modelo es consistente con la maximización de la utilidad para todos los niveles de los datos. Es fácil verificar que P_{ni} se convierte en la fórmula logit estándar cuando $\lambda_{ij}=1\ \forall i,j$. En su aplicación práctica, Koppelman y Wen (2000) comprobaron que el modelo PCL obtenía mejores resultados que el logit jerárquico o el logit estándar.

El investigador puede hacer un test de hipótesis sobre si $\lambda_{ij}=1$ para algunos o todos los pares, mediante el test de ratio de verosimilitudes de la Sección 3.8.2. La aceptación de la hipótesis para un par de alternativas implica que no hay una correlación significativa en la utilidad no observada para ese par. El investigador también puede añadir cierta estructura al patrón de correlación. Por ejemplo, es posible asumir que las correlaciones son iguales entre un grupo de alternativas; esta asunción se impone fijando $\lambda_{ij}=\lambda_{kl}$ para todo i,j,k y l dentro del grupo. El modelo OGEV de Small es un modelo PCL en el que se especifica que λ_{ij} sea una función de la proximidad entre i y j. Con un número grande de alternativas, el investigador tendrá probablemente necesidad de imponer alguna forma de estructura en las λ_{ij}s, simplemente para evitar la proliferación de parámetros que afloran con un valor de J grande. Esta proliferación de parámetros, uno para cada par de alternativas, es lo que hace al modelo PCL tan flexible. El objetivo del investigador es aplicar esta flexibilidad con sentido en su situación particular.

Como se ha visto casi al final de la sección 2.5, ya que la escala y el nivel de utilidad son irrelevantes, como máximo pueden estimarse $J(J-1)/2-1$ parámetros de covarianza en un modelo de elección discreta. Un modelo PCL contiene $J(J-1)/2$ λs: una λ por cada alternativa emparejada con cada una del resto de alternativas, teniendo en cuenta que i emparejado con j es lo mismo que j emparejado con

i. El número de λs excede el número máximo de parámetros de covarianza identificables exactamente en uno. El investigador, por lo tanto, debe fijar al menos una restricción sobre las λs. Esto puede lograrse mediante la normalización de una de las λs a 1. Si se impone cierta estructura sobre el patrón de correlación, como se describe en el párrafo anterior, en general esta estructura impondrá la normalización de forma automática.

4.4.2 Logit jerárquico generalizado (GNL)

Los nidos de alternativas son etiquetados $B_1, B_2, \dots, B_K$. Cada alternativa puede ser miembro de más de un nido. Es importante destacar que una alternativa puede estar en un nido en diversos grados. Dicho de otra manera, una alternativa se distribuye entre los nidos, estando más presente en unos nidos que en otros. Un parámetro de *asignación* α_{jk}indica en qué medida la alternativa j es un miembro del nido k. Este parámetro debe ser no negativo: $\alpha_{jk} \geq 0 \; \forall j,k$. Un valor nulo significa que la alternativa no está presente en el nido en absoluto. La interpretación se facilita al imponer que los parámetros de asignación sumen uno entre los nidos para cualquier alternativa: $\sum_k \alpha_{jk} = 1 \; \forall j$. Bajo esta condición, α_{jk} refleja la porción de la alternativa que se asigna a cada nido.

Se define un parámetro λ_k para cada nido que realiza la misma función que en los modelos logit jerárquicos, es decir, indica el grado de independencia entre alternativas dentro del nido: un valor mayor de λ_k se traduce en una mayor independencia y menor correlación.

La probabilidad de que la persona n elija la alternativa i es

$$P_{ni} = \frac{\sum_k \left(\alpha_{ik} e^{V_{ni}}\right)^{1/\lambda_k} \left(\sum_{j \in B_k} \left(\alpha_{jk} e^{V_{nj}}\right)^{1/\lambda_k}\right)^{\lambda_k - 1}}{\sum_{l=1}^{K} \left(\sum_{j \in B_l} \left(\alpha_{jl} e^{V_{nj}}\right)^{1/\lambda_l}\right)^{\lambda_l}}. \tag{4.7}$$

Esta fórmula es similar a la probabilidad logit jerárquica dada en la ecuación (4.2), excepto que el numerador es una suma sobre todos los nidos que contienen la alternativa i, con ponderaciones aplicadas a estos nidos. Si cada alternativa entra en un solo nido, con $\alpha_{jk} = 1$ para $j \in B_k$ y cero en caso contrario, el modelo se convierte en un modelo logit jerárquico. Y si además, $\lambda_k = 1$ para todos los nidos, entonces el modelo se convierte en el logit estándar. Wen y Koppelman (2001) obtienen varios modelos jerárquicos cruzados como casos especiales del modelo GNL.

Para facilitar la interpretación, la probabilidad GNL se puede descomponer como

$$P_{ni} = \sum_k P_{ni|B_k} P_{nk},$$

donde la probabilidad del nido k es

$$P_{nk} = \frac{\left(\sum_{j \in B_k} \left(\alpha_{jk} e^{V_{nj}}\right)^{1/\lambda_k}\right)^{\lambda_k}}{\sum_{l=1}^{K} \left(\sum_{j \in B_l} \left(\alpha_{jl} e^{V_{nj}}\right)^{1/\lambda_l}\right)^{\lambda_l}}.$$

y la probabilidad de la alternativa i dado el nido k es

$$P_{ni|B_k} = \frac{\left(\alpha_{ik} e^{V_{ni}}\right)^{1/\lambda_k}}{\sum_{j \in B_k} \left(\alpha_{jk} e^{V_{nj}}\right)^{1/\lambda_k}}.$$

4.5 Logit heterocedástico

En lugar de capturar las correlaciones entre alternativas, el investigador puede desear simplemente permitir que la varianza de factores no observados difiera entre alternativas. Steckel y Vanhonacker (1988), Bhat (1995) y Recker (1995) describen un tipo de modelo GEV, llamado valor extremo heterocedástico (*heteroskedastic extreme value*, HEV) que es igual al logit excepto por permitir una varianza diferente para cada alternativa. La utilidad se especifica como $U_{nj} = V_{nj} + \varepsilon_{nj}$, donde ε_{nj} se distribuye independientemente como valor extremo con varianza $(\theta_j \pi)^2/6$. No hay correlación en factores no observados entre las alternativas, sin embargo, la varianza de los factores no observados es diferente para alternativas diferentes. Para ajustar la escala global de la utilidad, la varianza de una alternativa se normaliza a $\pi^2/6$, que es la varianza de la distribución normalizada de valor extremo. Las varianzas de las otras alternativas se estiman posteriormente en relación a la varianza normalizada.

Las probabilidades de elección para este logit heterocedástico son (Bhat, 1995)

$$P_{ni} = \int \left[\prod_{j \neq i} e^{-e^{-(V_{ni} - V_{nj} + \theta_i w)/\theta_j}} \right] e^{-e^{-w}} e^{-w} dw.$$

donde $w = \varepsilon_{ni}/\theta_i$. La integral no tiene una forma cerrada, sin embargo, se puede aproximar por simulación . Tenga en cuenta que $exp(-exp(-w))exp(-w)$ es la densidad de la distribución valor extremo, facilitada en la Sección 3.1. Por lo tanto, P_{ni} es la integral del factor entre corchetes sobre la densidad de valor extremo. Puede ser simulada de la siguiente manera: (1) Extraiga un valor de la distribución de valor extremo, usando el procedimiento descrito en la Sección 9.2.3. (2) Para esta extracción de w, calcule el factor entre paréntesis, es decir $\prod_{j \neq i} \exp(-\exp(-(V_{ni} - V_{nj} + \theta_i w)/\theta_j))$. (3) Repita los pasos 1 y 2 múltiples veces y promedie los resultados. Este promedio es una aproximación de P_{ni}. Bhat (1995) muestra que, puesto que la integral es sólo unidimensional, las probabilidades logit heterocedásticas se pueden calcular eficazmente mediante cuadratura numérica en lugar de simulación.

4.6 La familia GEV

Pasamos ahora a describir el proceso que McFadden (1978) desarrolló para generar modelos GEV. Utilizando este proceso, el investigador puede desarrollar nuevos modelos GEV que se adapten mejor a las circunstancias específicas de su problema de elección. Como ejemplo, se muestra cómo se utiliza el procedimiento para generar modelos que ya hemos comentado, es decir, logit, logit jerárquico y logit combinacional emparejado. El mismo procedimiento puede ser aplicado por un investigador para generar nuevos modelos con propiedades que cumplan con sus necesidades de investigación.

Para simplificar la notación, vamos a omitir el subíndice n que denota al decisor. También, ya que vamos a utilizar $\exp(V_j)$ varias veces, lo referiremos de forma más compacta como Y_j. Es decir, $Y_j = \exp(V_j)$. Tenga presente que Y_j es necesariamente positivo.

Considere una función G que depende de Y_j para todo j. Denotamos esta función $G = G(Y_1, \ldots, Y_J)$. Sea G_i la derivada de G con respecto a Y_i: $G_i = \partial G/\partial Y_i$. Si esta función reúne ciertas condiciones, entonces es posible basar un modelo de elección discreta en ella. En particular, si G satisface las condiciones que se enumeran en el párrafo siguiente, entonces

$$P_i = \frac{Y_i G_i}{G} \tag{4.8}$$

es la probabilidad de elección de un modelo de elección discreta que es consistente con la maximización de la utilidad. Cualquier modelo que pueda ser formulado de esta manera es un modelo GEV. Por tanto, esta fórmula define la familia de modelos GEV.

Las propiedades que la función debe cumplir son las siguientes:

1. $\mathrm{G} \geq 0$ para todos los valores positivos de Y_j $\forall j$.
2. G es una función homogénea de grado uno. Es decir si cada Y_j se incrementa en una proporción determinada ρ, G se incrementa también en esa misma proporción ρ: $G(\rho Y_1, \ldots, \rho Y_j) = \rho G(Y_1, \ldots, Y_j)$. En realidad, Ben-Akiva y Francois (1983) mostraron que esta condición podía relajarse para permitir cualquier grado de homogeneidad. Mantenemos el uso de grado uno, ya que al hacerlo la condición es más fácil de interpretar y es consistente con la descripción original de McFadden.
3. $\mathrm{G} \to \infty$ cuando $Y_j \to \infty$ para cualquier j.
4. Las derivadas parciales cruzadas de G cambian de signo de una manera particular. Es decir, $\mathrm{G}_i \geq 0$ para todo i, $\mathrm{G}_{ij} = \partial \mathrm{G}_i / \partial \mathrm{Y}_j \leq 0$ para todo $j \neq i$, $\mathrm{G}_{ijk} = \partial \mathrm{G}_{ij} / \partial \mathrm{Y}_k \geq 0$ para cualquier i, j, k distintos, y así sucesivamente para derivadas parciales cruzadas de mayor orden.

Hay poca intuición económica que motive estas propiedades, particularmente la última. Sin embargo, es fácil verificar si una función las cumple. La falta de intuición detrás de las propiedades es al mismo tiempo una bendición y una maldición. La desventaja es que el investigador tiene poca orientación sobre cómo especificar una función G que proporcione un modelo que satisfaga las necesidades de su investigación. La ventaja es que el enfoque puramente matemático permite al investigador generar modelos que él no podría haber desarrollado confiando solamente en su intuición económica. Karlstrom (2001) ofrece un ejemplo: arbitrariamente especificó una G (en el sentido de que no se basaba en conceptos de comportamiento) y se encontró que la fórmula de probabilidad resultante se ajustaba mejor a sus datos que logit, logit jerárquico y PCL.

Para ilustrar el proceso de generación podemos mostrar cómo logit, logit jerárquico y los modelos PCL se obtienen bajo las especificaciones apropiadas de G.

Logit

Sea $\mathrm{G} = \sum_{j=1}^{J} Y_j$. Esta G exhibe las cuatro propiedades requeridas: (1) La suma de Y_js positivas es positiva. (2) Si todas las Y_js son incrementadas por un factor ρ, G se incrementa en ese mismo factor. (3) Si alguna Y_j se incrementa sin límite, entonces G también lo hace. (4) La primera derivada parcial de G es $\mathrm{G}_i = \partial \mathrm{G} / \partial \mathrm{Y}_j = 1$, que cumple el criterio de que $\mathrm{G}_i > 0$. Y las derivadas de orden superior son todas cero, que cumple claramente el criterio expuesto, ya que son ≥ 0 o ≤ 0 como se requiere.

Si insertamos esta función G y su primera derivada G_i en (4.8), la probabilidad de elección resultante es

$$P_i = \frac{Y_i G_i}{G}$$

$$= \frac{Y_i}{\sum_{j=1}^{J} Y_i}$$

$$= \frac{e^{V_i}}{\sum_{j=1}^{J} e^{V_i}},$$

que es la fórmula logit.

Logit jerárquico

Las J alternativas se dividen en K nidos etiquetados $B_1, \ldots, B_K$. Sea

$$G = \sum_{l=1}^{K} \left(\sum_{j \in B_l} Y_j^{1/\lambda_l} \right)^{\lambda_l},$$

con cada λ_k entre cero y uno. Las tres primeras propiedades son fáciles de verificar. Para la cuarta propiedad, calculamos la primera derivada parcial

$$G_i = \lambda_k \left(\sum_{j \in B_k} Y_j^{1/\lambda_k} \right)^{\lambda_k - 1} \frac{1}{\lambda_k} Y_i^{(1/\lambda_k)-1}$$

$$= Y_i^{(1/\lambda_k)-1} \left(\sum_{j \in B_k} Y_j^{1/\lambda_k} \right)^{\lambda_k - 1}$$

para $i \in B_k$. Dado que $Y_j \geq 0 \; \forall j$, tenemos que $\mathrm{G}_i \geq 0$, según se requería. La segunda derivada parcial cruzada es

$$G_{im} = \frac{\partial G_i}{\partial Y_m}$$

$$= (\lambda_k - 1) Y_i^{(1/\lambda_k)-1} \left(\sum_{j \in B_k} Y_j^{1/\lambda_k} \right)^{\lambda_k - 2} \frac{1}{\lambda_k} Y_m{}^{(1/\lambda_k)-1}$$

$$= \frac{\lambda_k - 1}{\lambda_k} (Y_i Y_m)^{(1/\lambda_k)-1} \left(\sum_{j \in B_k} Y_j^{1/\lambda_k} \right)^{\lambda_k - 2}$$

para $m \in B_k$ y $m \neq i$. Con $\lambda_k \leq 1$, $G_{ij} \leq 0$, según se requería. Para j en un nido diferente a i, $G_{ij} = 0$, que también cumple el criterio. Derivadas parciales cruzadas de orden superior se calculan de manera similar; exhiben la propiedad requerida si $0 < \lambda_k \leq 1$.

La probabilidad de elección se convierte en

$$P_i = \frac{Y_i G_i}{G}$$

$$= \frac{Y_i Y_i^{(1/\lambda_k)-1} \left(\sum_{j \in B_k} Y_j^{1/\lambda_l} \right)^{\lambda_k - 1}}{\sum_{l=1}^{K} \left(\sum_{j \in B_l} Y_j^{1/\lambda_l} \right)^{\lambda_l}}$$

$$= \frac{Y_i^{1/\lambda_k}\left(\sum_{j\in B_k} Y_j^{1/\lambda_l}\right)^{\lambda_k-1}}{\sum_{l=1}^{K}\left(\sum_{j\in B_l} Y_j^{1/\lambda_l}\right)^{\lambda_l}}$$

$$= \frac{(e^{V_i})^{1/\lambda_k}\left(\sum_{j\in B_k}(e^{V_j})^{1/\lambda_l}\right)^{\lambda_k-1}}{\sum_{l=1}^{K}\left(\sum_{j\in B_l}(e^{V_j})^{1/\lambda_l}\right)^{\lambda_l}}$$

$$= \frac{e^{V_i/\lambda_k}\left(\sum_{j\in B_k} e^{V_j/\lambda_l}\right)^{\lambda_k-1}}{\sum_{l=1}^{K}\left(\sum_{j\in B_l} e^{V_j/\lambda_l}\right)^{\lambda_l}}$$

Que es la fórmula logit jerárquica (4.2).

Logit combinacional emparejado

Sea

$$G = \sum_{k=1}^{J-1}\sum_{l=k+1}^{J}(Y_k^{1/\lambda_{kl}} + Y_l^{1/\lambda_{kl}})^{\lambda_{kl}}.$$

Las propiedades requeridas se verifican de la misma forma que para el logit jerárquico. Tenemos

$$G_i = \sum_{j\neq i}\lambda_{ji}(Y_i^{1/\lambda_{ij}} + Y_j^{1/\lambda_{ij}})^{\lambda_{ij}-1}\frac{1}{\lambda_{ij}}Y_i^{(1/\lambda_{ij})-1}$$

$$= \sum_{j\neq i} Y_i^{(1/\lambda_{ij})-1}(Y_i^{1/\lambda_{ij}} + Y_j^{1/\lambda_{ij}})^{\lambda_{ij}-1}.$$

De esta forma, la probabilidad de elección es

$$\mathrm{P_i} = \frac{\mathrm{Y_i G_i}}{\mathrm{G}}$$

$$= \frac{Y_i\sum_{j\neq i} Y_i^{(1/\lambda_{ij})-1}(Y_i^{1/\lambda_{ij}} + Y_j^{1/\lambda_{ij}})^{\lambda_{ij}-1}}{\sum_{k=1}^{J-1}\sum_{l=k+1}^{J}(Y_k^{1/\lambda_{kl}} + Y_l^{1/\lambda_{kl}})^{\lambda_{kl}}}$$

$$= \frac{\sum_{j\neq i} Y_i^{(1/\lambda_{ij})}(Y_i^{1/\lambda_{ij}} + Y_j^{1/\lambda_{ij}})^{\lambda_{ij}-1}}{\sum_{k=1}^{J-1}\sum_{l=k+1}^{J}(Y_k^{1/\lambda_{kl}} + Y_l^{1/\lambda_{kl}})^{\lambda_{kl}}}$$

$$= \frac{\sum_{j\neq i} e^{V_i/\lambda_{ij}}(e^{V_i/\lambda_{ij}} + e^{V_j/\lambda_{ij}})^{\lambda_{ij}-1}}{\sum_{k=1}^{J-1}\sum_{l=k+1}^{J}(e^{V_k/\lambda_{kl}} + e^{V_l/\lambda_{kl}})^{\lambda_{kl}}}$$

que es la fórmula del modelo PCL (4.6).

Logit jerárquico generalizado

El lector puede verificar que las probabilidades GNL en la ecuación (4.7) se obtienen de

$$G = \sum_{k=1}^{K} \left(\sum_{j \in B_k} (\alpha_{jk} Y_j)^{1/\lambda_k} \right)^{\lambda_k} .$$

Usando el mismo proceso, los investigadores pueden generar otros modelos GEV.

5

Probit

5.1 Probabilidades de elección

El modelo logit tiene tres limitaciones importantes. (1) No puede representar la variación aleatoria de preferencias. (2) Presenta patrones de sustitución restrictivos debido a la propiedad de IIA. Y (3) no puede utilizarse con datos de panel cuando los factores no observados están correlacionados en el tiempo para cada decisor. Los modelos GEV relajan la segunda de estas restricciones, pero no las otros dos. Los modelos probit abordan las tres limitaciones. Pueden manejar variación de preferencias aleatoria, permiten cualquier patrón de sustitución y son aplicables a datos de panel con errores correlacionados temporalmente.

La única limitación de los modelos probit es que requieren distribuciones normales para todos los componentes no observados de utilidad. En muchos casos, quizá en la mayoría de las situaciones, las distribuciones normales proporcionan una representación adecuada de los componentes aleatorios. Sin embargo, en algunas situaciones las distribuciones normales son inadecuadas y pueden conducir a predicciones perversas. Un ejemplo destacado es el de los coeficientes de precios. Para un modelo probit con variación de preferencias aleatoria, el coeficiente de precio se supone distribuido normalmente en la población. Dado que la distribución normal tiene densidad en ambos lados del cero, el modelo implica necesariamente que algunas personas tienen un coeficiente de precio positivo (preferencia por precios más caros). El uso de una distribución con densidad en un solo lado del cero, como la distribución logarítmica normal (log-normal) es más apropiado y, sin embargo, no puede ser acomodado dentro de un modelo probit. Exceptuando esta restricción, probit es bastante general.

El modelo probit se obtiene bajo el supuesto de que las utilidades no observadas siguen una distribución normal conjunta. La primera formulación de un probit binario a cargo de Thurstone (1927) utilizaba la terminología de estímulos psicológicos, terminología que Marschak (1960) tradujo a términos económicos como utilidad. Hausman y Wise (1978) y Daganzo (1979) dilucidaron la generalidad de la especificación para representar diversos aspectos del comportamiento de elección. Empecemos descomponiendo la utilidad en su parte observada y su parte no observada: $U_{nj} = V_{nj} + \varepsilon_{nj}\ \forall j$. Considere el vector compuesto por cada ε_{nj}, etiquetado como $\varepsilon_n' = \langle \varepsilon_{n1}, \ldots, \varepsilon_{nJ} \rangle$. Supondremos que ε_n se distribuye de acuerdo a una distribución normal con un vector de medias cero y una matriz de covarianza Ω. La densidad de ε_n es

$$\phi(\varepsilon_n) = \frac{1}{(2\pi)^{J/2}|\Omega|^2} e^{-\frac{1}{2}\varepsilon_n' \Omega^{-1} \varepsilon_n}$$

La covarianza Ω puede depender de variables percibidas por el decisor n, por lo que Ω_n sería la notación más apropiada; sin embargo, omitimos el subíndice en aras de la simplicidad.

La probabilidad elección es

$$P_{ni} = Prob(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \; \forall j \neq i)$$

$$= \int I\big(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \; \forall j \neq i\big)\phi(\varepsilon_n) d\varepsilon_n, \qquad (5.1)$$

donde $I(\cdot)$ es un indicador de si la expresión entre paréntesis es verdadera y la integral es sobre todos los valores de ε_n. Esta integral no tiene una forma cerrada. Debe ser evaluada numéricamente mediante simulación.

Las probabilidades de elección pueden expresarse de otras dos maneras que son útiles para la simulación de la integral. Sea B_{ni} el conjunto de valores de los términos de error que producen que la elección del decisor sea la alternativa i: $B_{ni} = \big\{\varepsilon_n \text{ s.t. } V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj} \; \forall \text{j} \neq \text{i}\big\}$. Entonces

$$P_{ni} = \int_{\varepsilon_n \in B_{ni}} \phi(\varepsilon_n) d\varepsilon_n, \qquad (5.2)$$

que es una integral sobre algunos de los valores de ε_n en lugar de sobre todos los valores posibles, es decir, es sobre los valores de B_{ni}.

Las expresiones (5.1) y (5.2) son integrales J-dimensionales sobre los J errores $\varepsilon_{nj}, j = 1, \dots, J$. Dado que sólo las diferencias de utilidad importan, las probabilidades de elección pueden expresarse de manera equivalente como integrales $(\text{J} - 1)$-dimensionales sobre las diferencias entre errores. Definamos las diferencias respecto a la alternativa i, la alternativa para la que estamos calculando la probabilidad de elección. Definimos $\widetilde{U}_{nji} = U_{nj} - U_{ni}$, $\tilde{V}_{nji} = V_{nj} - V_{ni}$ y $\tilde{\varepsilon}_{nji} = \varepsilon_{nj} - \varepsilon_{ni}$. Entonces $P_{ni} = Prob(\widetilde{U}_{nji} < 0 \; \forall j \neq i)$. Es decir, la probabilidad de elegir la alternativa i es la probabilidad de que todas las diferencias de utilidad, cuando se refieren a la alternativa i, sean negativas. Definimos el vector $\tilde{\varepsilon}_{ni} = \langle \tilde{\varepsilon}_{n1i}, \dots, \tilde{\varepsilon}_{nji} \rangle$ donde los puntos "..." representan todas las alternativas excepto i, de manera que $\tilde{\varepsilon}_{ni}$ tiene dimensión $J - 1$. Dado que la diferencia entre dos variables normales es normal, la densidad de las diferencias de error es

$$\phi(\tilde{\varepsilon}_{ni}) = \frac{1}{(2\pi)^{\frac{1}{2}(J-1)}\left|\widetilde{\Omega}_i\right|^{1/2}} e^{-\frac{1}{2}\tilde{\varepsilon}_{ni}' \widetilde{\Omega}_i^{-1} \tilde{\varepsilon}_{ni}},$$

donde $\widetilde{\Omega}_i$ es la covarianza de $\tilde{\varepsilon}_{ni}$, obtenida a partir de Ω. La probabilidad de elección expresada en diferencias de utilidad es por tanto

$$P_{ni} = \int I\big(\tilde{V}_{nji} + \tilde{\varepsilon}_{nji} < 0 \; \forall j \neq i\big)\phi(\tilde{\varepsilon}_{ni}) d\tilde{\varepsilon}_{ni}, \qquad (5.3)$$

que es una integral $(\text{J} - 1)$-dimensional sobre todos los valores posibles de las diferencias de error. Una expresión equivalente es

$$P_{ni} = \int_{\tilde{\varepsilon}_{ni} \in \tilde{B}_{ni}} \phi(\tilde{\varepsilon}_{ni}) d\tilde{\varepsilon}_{ni}, \qquad (5.4)$$

donde $\tilde{B}_{ni} = \{\tilde{\varepsilon}_{ni} \text{ s.t. } \tilde{V}_{nji} + \tilde{\varepsilon}_{nji} < 0 \; \forall j \neq i\}$, que es una integral $(\mathrm{J}-1)$-dimensional sobre las diferencias de error en $\tilde{B}_{ni}$.

Las expresiones (5.3) y (5.4) utilizan la matriz de covarianza $\widetilde{\Omega}_i$ de las diferencias de error. Hay una manera directa de obtener $\widetilde{\Omega}_i$ a partir de la covarianza de los errores, Ω. Sea M_i la matriz identidad de dimensión $(\mathrm{J}-1)$ con una columna adicional de -1s agregada como columna i-ésima. La columna adicional hace que la matriz tenga dimensiones $(\mathrm{J}-1) \times J$. Por ejemplo, con $J = 4$ alternativas e $i = 3$,

$$M_i = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}.$$

Esta matriz puede ser usada para transformar la matriz de covarianza de los errores en la matriz de covarianza de las diferencias entre errores: $\widetilde{\Omega}_i = M_i \Omega M_i'$. Observe que $\widetilde{\Omega}_i$ es de dimensión $(\mathrm{J}-1) \times (J-1)$, mientras que Ω es de dimensión $J \times J$, ya que M_i es $(J-1) \times J$. A modo ilustrativo, considere una situación de tres alternativas con errores $\langle \varepsilon_{n1}, \varepsilon_{n2}, \varepsilon_{n3} \rangle$ que tiene covarianza

$$\Omega = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix}.$$

Supongamos que calculamos las diferencias de error respecto la alternativa 2. Sabemos por postulación que las diferencias de error $\langle \tilde{\varepsilon}_{n12}, \tilde{\varepsilon}_{n32} \rangle$ tienen covarianza

$$\widetilde{\Omega}_2 = Cov \begin{pmatrix} \varepsilon_{n1} - \varepsilon_{n2} \\ \varepsilon_{n3} - \varepsilon_{n2} \end{pmatrix}$$

$$\begin{pmatrix} \sigma_{11} + \sigma_{22} - 2\sigma_{12} & \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} \\ \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} & \sigma_{33} + \sigma_{22} - 2\sigma_{23} \end{pmatrix}.$$

Esta matriz de covarianza también se puede obtener a partir de la transformación $\widetilde{\Omega}_2 = M_2 \Omega M_2'$:

$$\widetilde{\Omega}_n = \begin{pmatrix} 1 & -1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & -1 \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_{11} - \sigma_{12} & \sigma_{12} - \sigma_{22} & \sigma_{13} - \sigma_{23} \\ -\sigma_{12} + \sigma_{13} & -\sigma_{22} + \sigma_{23} & -\sigma_{23} + \sigma_{33} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & -1 \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_{11} - \sigma_{12} - \sigma_{12} + \sigma_{22} & -\sigma_{12} + \sigma_{22} + \sigma_{13} - \sigma_{23} \\ -\sigma_{12} + \sigma_{13} + \sigma_{22} - \sigma_{23} & \sigma_{22} - \sigma_{23} - \sigma_{23} + \sigma_{33} \end{pmatrix}$$

$$\begin{pmatrix} \sigma_{11} + \sigma_{22} - 2\sigma_{12} & \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} \\ \sigma_{13} + \sigma_{22} - \sigma_{12} - \sigma_{23} & \sigma_{33} + \sigma_{22} - 2\sigma_{23} \end{pmatrix}.$$

Como veremos, esta transformación mediante M_i es muy práctica cuando se simulan probabilidades probit.

5.2 Identificación

Tal y como se describió en la sección 2.5, cualquier modelo de elección discreta debe estar normalizado para tener en cuenta el hecho de que el nivel y la escala de la utilidad son irrelevantes. El nivel de utilidad es irrelevante porque se puede añadir una constante a la utilidad de todas las alternativas sin cambiar la alternativa que tiene mayor utilidad: la alternativa con la utilidad más alta antes de añadir la constante sigue siendo la de mayor utilidad después de la adición. Del mismo modo, la escala de la utilidad no importa porque la utilidad de cada alternativa puede ser multiplicada por una constante (positiva) sin cambiar la alternativa que tiene mayor utilidad. En los modelos logit y logit jerárquico la normalización de escala y de nivel se produce de forma automática bajo los supuestos que se realizan relativos a la distribución de los términos de error. Como resultado, para estos modelos no es necesario considerar de forma explícita la normalización. Con modelos probit, sin embargo, la normalización de escala y de nivel no se produce automáticamente. El investigador debe normalizar el modelo directamente.

La normalización del modelo está relacionada con la identificación de parámetros. Un parámetro es *identificado* si se puede estimar, y es *no identificado* si no puede ser estimado. Un ejemplo de un parámetro no identificado es k en la especificación de la utilidad: $U_{nj} = V_{nj} + \mathrm{k} + \varepsilon_{nj}$. Aunque el investigador podría escribir la utilidad de esta forma y podría intentar estimar k para obtener una medida del nivel general de utilidad, eso es imposible. El comportamiento del decisor no se ve afectado por k, por lo que el investigador no puede deducir su valor a partir de las elecciones que los decisores han hecho. Dicho de forma directa, los parámetros que no afectan el comportamiento de los decisores no pueden ser estimados. En un modelo no normalizado, pueden aparecer parámetros no identificados; estos parámetros se refieren a la escala y al nivel de la utilidad, algo que no afecta al comportamiento. Una vez que el modelo se normaliza, estos parámetros desaparecen. La dificultad reside en que no siempre es evidente qué parámetros se refieren a la escala y al nivel de utilidad. En el ejemplo anterior, el hecho de que k es un parámetro no identificado es bastante obvio. En muchos casos, no es en absoluto evidente qué parámetros son identificados. Bunch y Kitamura (1989) han demostrado que los modelos probit que aparecen en varios artículos publicados no están normalizados y contienen parámetros no identificados. El hecho de que ni los autores ni los revisores de estos artículos pudiesen detectar que los modelos no estaban normalizados es un buen testimonio de la complejidad de la cuestión.

A continuación proporciono un procedimiento que siempre puede ser usado para normalizar un modelo probit y asegurar que todos los parámetros son identificados. No es el único procedimiento que puede usarse; véase, por ejemplo, Bunch (1991). En algunos casos, un investigador puede encontrar otros procedimientos de normalización más convenientes. Sin embargo, el procedimiento que facilito siempre se puede utilizar, ya sea por sí mismo o como un control sobre otro procedimiento.

Describo el procedimiento para un modelo de cuatro alternativas. La generalización a más alternativas es obvia. Como de costumbre, la utilidad se expresa como $U_{nj} = V_{nj} + \varepsilon_{nj},\ j = 1, \ldots, 4$. El vector de errores es $\varepsilon_n' = \langle \varepsilon_{n1}, \ldots, \varepsilon_{n4} \rangle$. Este vector se distribuye normalmente con media cero y una matriz de covarianza que se puede expresar de forma explícita como

$$\Omega = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \cdot & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \cdot & \cdot & \sigma_{33} & \sigma_{34} \\ \cdot & \cdot & \cdot & \sigma_{44} \end{pmatrix}, \tag{5.5}$$

donde los puntos se refieren a los elementos correspondientes en la parte superior de la matriz. Tenga en cuenta que hay diez elementos en esta matriz, es decir, diez σs distintas que representan las

varianzas y covarianzas entre los cuatro errores. En general, un modelo con J alternativas tiene $J(J+1)/2$ elementos distintos en la matriz de covarianza de los errores.

Para tener en cuenta el hecho de que el nivel de utilidad es irrelevante, usamos diferencias de utilidad. En mi procedimiento, siempre uso las diferencias respecto a la primera alternativa, ya que simplifica el análisis tal y como veremos. Definimos las diferencias de error como $\tilde{\varepsilon}_{nj1} = \varepsilon_{nj} - \varepsilon_{n1}$ para $j = 2,3,4$ y definimos el vector de las diferencias de error como $\tilde{\varepsilon}_{n1} = \langle \tilde{\varepsilon}_{n21}, \tilde{\varepsilon}_{n31}, \tilde{\varepsilon}_{n41} \rangle$. Observe que el subíndice 1 en $\tilde{\varepsilon}_{n1}$ significa que las diferencias de error son respecto a la primera alternativa, en lugar de indicar que los errores son de la primera alternativa.

La matriz de covarianza para el vector de las diferencias de error toma la siguiente forma

$$\widetilde{\Omega}_1 = \begin{pmatrix} \theta_{22} & \theta_{23} & \theta_{24} \\ \cdot & \theta_{33} & \theta_{34} \\ \cdot & \cdot & \theta_{44} \end{pmatrix},$$

donde las θs se relacionan con las σs originales así:

$$\theta_{22} = \sigma_{22} + \sigma_{11} - 2\sigma_{12},$$

$$\theta_{33} = \sigma_{33} + \sigma_{11} - 2\sigma_{13},$$

$$\theta_{44} = \sigma_{44} + \sigma_{11} - 2\sigma_{14},$$

$$\theta_{23} = \sigma_{23} + \sigma_{11} - \sigma_{12} - \sigma_{13},$$

$$\theta_{24} = \sigma_{24} + \sigma_{11} - \sigma_{12} - \sigma_{14},$$

$$\theta_{34} = \sigma_{34} + \sigma_{11} - \sigma_{13} - \sigma_{14}.$$

Computacionalmente, esta matriz se puede obtener utilizando la matriz de transformación M_i definida en la Sección 5.1 como $\widetilde{\Omega}_1 = M_1 \Omega M_1'$.

Para ajustar la escala de la utilidad, uno de los elementos de la diagonal se normaliza. Yo fijo el elemento de la parte superior izquierda de $\widetilde{\Omega}_1$, que es la varianza de $\tilde{\varepsilon}_{n21}$, a 1. Esta normalización de la escala nos da la siguiente matriz de covarianza:

(5.6)
$$\widetilde{\Omega}_1^* = \begin{pmatrix} 1 & \theta_{23}^* & \theta_{24}^* \\ \cdot & \theta_{33}^* & \theta_{34}^* \\ \cdot & \cdot & \theta_{44}^* \end{pmatrix}$$

Las θ^*s se relacionan con las σs originales como sigue:

$$\theta_{33}^* = \frac{\sigma_{33} + \sigma_{11} - 2\sigma_{13}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}},$$

$$\theta_{44}^* = \frac{\sigma_{44} + \sigma_{11} - 2\sigma_{14}}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}},$$

$$\theta_{23}^{*} = \frac{\sigma_{23} + \sigma_{11} - \sigma_{12} - \sigma_{13},}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}},$$

$$\theta_{24}^{*} = \frac{\sigma_{24} + \sigma_{11} - \sigma_{12} - \sigma_{14},}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}},$$

$$\theta_{34}^{*} = \frac{\sigma_{34} + \sigma_{11} - \sigma_{13} - \sigma_{14},}{\sigma_{22} + \sigma_{11} - 2\sigma_{12}},$$

Hay cinco elementos en $\tilde{\Omega}_1^*$. Estos son los únicos parámetros identificados del modelo. Este número es menor a los diez elementos que entran en Ω. Cada θ^* es una función de las σs. Puesto que hay cinco θ^*s y diez σs, no es posible resolver para todas las σs a partir de los valores estimados de las θ^*s. Por tanto, no es posible obtener estimaciones para todas las σs.

En general, un modelo con J alternativas y una matriz de covarianza sin restricciones tendrá $[(\mathrm{J}-1)\mathrm{J}/2]-1$ parámetros de covarianza al ser normalizado, en comparación con los $\mathrm{J}(\mathrm{J}+1)/2$ parámetros cuando no se normaliza. Sólo $[(\mathrm{J}-1)\mathrm{J}/2]-1$ parámetros son identificados. Esta reducción en el número de parámetros *no* es una restricción. La reducción en el número de parámetros es una normalización que simplemente elimina los aspectos irrelevantes de la matriz de covarianza original, que son la escala y el nivel de utilidad. Los diez elementos en Ω permiten una varianza y una covarianza debida simplemente a la escala y al nivel de utilidad, que no tiene relevancia en el comportamiento de los decisores. Sólo los cinco elementos de $\tilde{\Omega}_1^*$ contendrán información acerca de la varianza y covarianza de los errores con independencia de la escala y el nivel de utilidad. En este sentido, sólo los cinco parámetros tienen contenido económico y sólo esos cinco parámetros se pueden estimar.

Supongamos ahora que el investigador impone una estructura sobre la matriz de covarianza. Es decir, en lugar de permitir una matriz de covarianza completa para los errores, el investigador considera que los errores siguen un patrón que implica ciertos valores particulares para - o ciertas relaciones entre - los elementos de la matriz de covarianza. El investigador restringe la matriz de covarianza para incorporar este patrón.

La estructura puede adoptar diversas formas, dependiendo de la aplicación. Yai et al . (1997) estiman un modelo probit de elección de ruta donde la covarianza entre dos rutas cualesquiera sólo depende de la longitud de los segmentos de ruta compartidos; esta estructura reduce el número de parámetros de covarianza a uno solo, que captura la relación de la covarianza con la longitud compartida. Bolduc et al. (1996) estiman un modelo de elección de ubicación de médicos donde la covarianza entre ubicaciones es una función de la proximidad entre las propias ubicaciones, usando lo que Bolduc (1992) ha denominado una estructura "generalizada auto-regresiva". Haaijer et al . (1998) imponen una estructura factor-analítica (*factor-analytic structure*) que surge de coeficientes aleatorios de las variables explicativas; este tipo de estructura se describe en detalle en la Sección 5.3. Elrod and Keane (1995) también imponen una estructura factor-analítica, pero que surge a partir de componentes de error en lugar de coeficientes aleatorios per se.

A menudo la estructura impuesta será suficiente para normalizar el modelo. Es decir, las restricciones que el investigador impone a la matriz de covarianza para ajustar sus expectativas acerca de la forma en que los errores se relacionan entre sí, servirán también para normalizar el modelo. Sin embargo, esto no siempre sucede. Los ejemplos citados por Bunch y Kitamura (1989) son casos en que las restricciones que el investigador aplicaba a la matriz de covarianza parecían suficientes para normalizar el modelo, pero en realidad no lo eran.

El procedimiento que he facilitado en el texto anterior se puede utilizar para determinar si las restricciones aplicadas a la matriz de covarianza son suficientes para normalizar el modelo. El investigador especifica Ω con sus restricciones sobre los elementos de la matriz. A continuación, el procedimiento indicado se utiliza para obtener $\tilde{\Omega}_1^*$, que está normalizada para la escala y el nivel. Sabemos que cada elemento de $\tilde{\Omega}_1^*$ es identificado. Si cada uno de los elementos restringidos de Ω puede ser calculado a partir de los elementos de $\tilde{\Omega}_1^*$, entonces las restricciones son suficientes para normalizar el modelo. En este caso, cada parámetro de la Ω restringida es identificado. Por otro lado, si los elementos de Ω no se pueden calcular a partir de los elementos de $\tilde{\Omega}_1^*$, las restricciones no son suficientes para normalizar el modelo y los parámetros de Ω no son identificados.

Para ilustrar este enfoque, supongamos que el investigador está estimando un modelo con cuatro alternativas y asume que la matriz de covarianza de los errores tiene la siguiente forma:

$$\Omega = \begin{pmatrix} 1+\rho & \rho & 0 & 0 \\ \cdot & 1+\rho & 0 & 0 \\ \cdot & \cdot & 1+\rho & \rho \\ \cdot & \cdot & \cdot & 1+\rho \end{pmatrix}.$$

Esta matriz de covarianza permite que el primer y el segundo error estén correlacionados, al igual que el error de la tercera y la cuarta alternativas, pero no permite ninguna otra correlación. La correlación entre los pares apropiados es $\rho/(1+\rho)$. Observe que mediante la especificación de los elementos de la diagonal como $1+\rho$, el investigador asegura que la correlación está entre -1 y 1 para cualquier valor de $\rho \geq -\frac{1}{2}$, como se requiere para una correlación. ¿Está este modelo, tal y como se especifica, normalizado para la escala y el nivel? Para responder a esta pregunta, aplicamos el procedimiento descrito. En primer lugar, tomamos las diferencias respecto a la primera alternativa. La matriz de covarianza de las diferencias de error es

$$\widetilde{\Omega}_1 = \begin{pmatrix} \theta_{22} & \theta_{23} & \theta_{24} \\ \cdot & \theta_{33} & \theta_{34} \\ \cdot & \cdot & \theta_{44} \end{pmatrix}$$

donde las θs se refieren a las σs originales según las siguientes relaciones:

$$\theta_{22} = 2$$

$$\theta_{33} = 2 + 2\rho,$$

$$\theta_{44} = 2 + 2\rho,$$

$$\theta_{23} = 1,$$

$$\theta_{24} = 1,$$

$$\theta_{34} = 1 + 2\rho.$$

A continuación normalizamos la escala estableciendo el elemento superior izquierdo a 1. La matriz de covarianza normalizada es

$$\widetilde{\Omega}_1^* = \begin{pmatrix} 1 & \theta_{23}^* & \theta_{24}^* \\ \cdot & \theta_{33}^* & \theta_{34}^* \\ \cdot & \cdot & \theta_{44}^* \end{pmatrix},$$

donde las θ^*s se relacionan con las σs originales por las siguientes fórmulas:

$$\theta_{33}^* = 1 + \rho,$$

$$\theta_{44}^* = 1 + \rho,$$

$$\theta_{23}^* = \frac{1}{2},$$

$$\theta_{24}^* = \frac{1}{2},$$

$$\theta_{34}^* = \frac{1}{2} + \rho.$$

Observe que $\theta_{33}^* = \theta_{44}^* = \theta_{34}^* + \frac{1}{2}$ y que el resto de θ^*s tienen valores fijos. Sólo hay un parámetro en $\tilde{\Omega}_1^*$, tal y como sucedía en Ω. Definimos $\theta = 1 + \rho$. La matriz $\tilde{\Omega}_1^*$ resulta

$$\widetilde{\Omega}_1^* = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \cdot & \theta & \theta - \frac{1}{2} \\ \cdot & \cdot & \theta \end{pmatrix},$$

El parámetro ρ original puede ser calculado directamente a partir de θ. Por ejemplo, si θ se estima que es 2.4, entonces la estimación de ρ es $\theta - 1 = 1.4$ y la correlación es $1.4/2.4 = 0.58$. El hecho de que los parámetros que entran en Ω puedan calcularse a partir de los parámetros que entran en la matriz de covarianza normalizada $\tilde{\Omega}_1^*$ significa que el modelo original ya está normalizado para la escala y el nivel de utilidad. Es decir, las restricciones que el investigador ha colocado en Ω también han proporcionado al mismo tiempo la normalización necesaria.

A veces, las restricciones en la matriz de covarianza original pueden parecer suficientes para normalizar el modelo, pero realmente no es así. Aplicar nuestro procedimiento determinará si realmente es el caso. Considere el modelo del ejemplo anterior, pero ahora supongamos que el investigador permite una correlación diferente entre el primer y segundo error, a la permitida entre el tercer y cuarto error. La matriz de covarianza de errores se especifica como

$$\Omega = \begin{pmatrix} 1 + \rho_1 & \rho_1 & 0 & 0 \\ \cdot & 1 + \rho_1 & 0 & 0 \\ \cdot & \cdot & 1 + \rho_2 & \rho_2 \\ \cdot & \cdot & \cdot & 1 + \rho_2 \end{pmatrix}$$

La correlación entre el primer y el segundo error es $\rho_1/(1+\rho_1)$, y la correlación entre el tercer y el cuarto error es $\rho_2/(1+\rho_2)$. Podemos obtener $\tilde{\Omega}_1$ para las diferencias de error y luego obtener $\tilde{\Omega}_1^*$ estableciendo el elemento superior izquierdo de $\tilde{\Omega}_1$ a 1. La matriz resultante es

$$\widetilde{\Omega}_1^* = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \cdot & \theta & \theta - \frac{1}{2} \\ \cdot & \cdot & \theta \end{pmatrix},$$

donde ahora $\theta = 1 + (\rho_1 + \rho_2)/2$. Los valores de ρ_1 y ρ_2 no se pueden calcular a partir de un valor de θ. Por lo tanto, el modelo original no está normalizado para la escala y el nivel, y los parámetros ρ_1 y ρ_2 no son identificados. Este hecho es algo sorprendente, ya que sólo dos parámetros entran en la matriz de covarianza original Ω. Parecería, a menos que el investigador explícitamente haga la prueba que acabamos de hacer, que restringir la matriz de covarianza para que conste de sólo dos elementos debería ser suficiente para normalizar el modelo. En este caso, sin embargo, no es así.

En el modelo normalizado, sólo aparece el promedio de la ρs: $(\rho_1 + \rho_2)/2$. Es posible calcular la ρ promedio a partir de θ, simplemente como $\theta - 1$. Esto significa que la ρ promedio es identificada, pero no los valores individuales. Cuando $\rho_1 = \rho_2$, como en el ejemplo anterior, el modelo queda normalizado porque cada ρ es igual a la ρ promedio. Sin embargo, como vemos ahora, cualquier modelo con el mismo promedio de ρ es equivalente, después de normalizar la escala y el nivel. Por lo tanto, asumir que $\rho_1 = \rho_2$ es lo mismo que asumir que $\rho_1 = 3\rho_2$ o cualquier otra relación. Lo único que importa para el comportamiento es el promedio de estos parámetros, no sus valores relativos entre ellos. Este hecho es bastante sorprendente y sería difícil darse cuenta del mismo sin el uso de nuestro procedimiento para la normalización.

Ahora que sabemos cómo asegurar que un modelo probit está normalizado para el nivel y la escala, y que por lo tanto contiene únicamente información económicamente relevante, podemos examinar cómo se utiliza el modelo probit para representar distintos tipos de situaciones de elección. Nos fijaremos en tres situaciones en las que los modelos logit presentan limitaciones y mostraremos cómo estas limitaciones se superan con probit. Estas situaciones son las variaciones de preferencia, los patrones de sustitución y las elecciones repetidas a lo largo del tiempo.

5.3 Variaciones de preferencia

Probit se adapta particularmente bien a la incorporación de coeficientes aleatorios, a condición de que los coeficientes se distribuyan normalmente. Hausman y Wise (1978) fueron los primeros, que yo tenga conocimiento, en desarrollar esta formulación. Haaijer et al. (1998) proporcionan una aplicación convincente de su uso. Suponga que la utilidad representativa es lineal en los parámetros y que los coeficientes varían aleatoriamente entre los decisores en lugar de estar fijados como se ha supuesto hasta ahora en este libro. La utilidad es $U_{nj} = \beta_n' x_{nj} + \varepsilon_{nj}$, donde β_n es el vector de coeficientes para el decisor n que representa las preferencias de esa persona. Supongamos que β_n presenta una distribución normal en la población con media b y covarianza W: $\beta_n \sim N(b, W)$. El objetivo de la investigación es estimar los parámetros b y W.

La utilidad puede ser reescrita descomponiendo β_n entre su media y las desviaciones de su media: $U_{nj} = b' x_{nj} + \tilde{\beta}_n' x_{nj} + \varepsilon_{nj}$, donde $\tilde{\beta}_n = \beta_n - b$. Los dos últimos términos de la utilidad son aleatorios; denominemos η_{nj} la suma de ambos términos aleatorios para obtener $U_{nj} = b' x_{nj} + \eta_{nj}$. La covarianza de los η_{nj}s depende de W así como de las x_{nj}s, por lo que la covarianza difiere entre decisores.

La covarianza de los términos η_{nj}s se puede describir fácilmente para un modelo con dos alternativas y una variable explicativa. En este caso, la utilidad es

$$U_{n1} = \beta_n x_{n1} + \varepsilon_{n1},$$

$$U_{n2} = \beta_n x_{n2} + \varepsilon_{n2}.$$

Supongamos que β_n se distribuye normalmente con media b y varianza σ_β. Supongamos que ε_{n1} y ε_{n2} se distribuyen de forma independiente e idéntica con varianza σ_ε. El supuesto de independencia es para este ejemplo y no es necesario en general. La utilidad se reescribe entonces como

$$U_{n1} = bx_{n1} + \eta_{n1},$$

$$U_{n2} = bx_{n2} + \eta_{n2},$$

donde η_{n1} y η_{n2} se distribuyen de acuerdo a una distribución normal conjunta. Cada una tiene una media de cero: $\mathrm{E}(\eta_{nj}) = \mathrm{E}(\tilde{\beta}_n x_{nj} + \varepsilon_{nj}) = 0$. La covarianza se determina como sigue. La varianza de cada una es $\mathrm{V}(\eta_{nj}) = \mathrm{V}(\tilde{\beta}_n x_{nj} + \varepsilon_{nj}) = x_{nj}^2\sigma_\beta + \sigma_\varepsilon$. Su covarianza es

$$Cov(\eta_{n1}, \eta_{n2}) = E\left[(\tilde{\beta}_n x_{n1} + \varepsilon_{n1})(\tilde{\beta}_n x_{n2} + \varepsilon_{n2})\right] =$$

$$E\left(\tilde{\beta}_n^2 x_{n1}x_{n2} + \varepsilon_{n1}\varepsilon_{n2} + \varepsilon_{n1}\tilde{\beta}_n x_{n2} + \varepsilon_{n2}\tilde{\beta}_n x_{n1}\right) =$$

$$x_{n1}x_{n2}\sigma_\beta.$$

La matriz de covarianza es

$$\Omega = \begin{pmatrix} x_{n1}^2\sigma_\beta + \sigma_\varepsilon & x_{n1}x_{n2}\sigma_\beta \\ x_{n1}x_{n2}\sigma_\beta & x_{n2}^2\sigma_\beta + \sigma_\varepsilon \end{pmatrix}$$

$$= \sigma_\beta \begin{pmatrix} x_{n1}^2 & x_{n1}x_{n2} \\ x_{n1}x_{n2} & x_{n2}^2 \end{pmatrix} + \sigma_\varepsilon \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Se requiere un último paso para la estimación. Recordemos que el comportamiento de los decisores no se ve afectado por una transformación multiplicativa de la utilidad. Por lo tanto, necesitamos establecer la escala de la utilidad. Una normalización conveniente para este caso es $\sigma_\varepsilon = 1$. Bajo esta normalización

$$\Omega = \sigma_\beta \begin{pmatrix} x_{n1}^2 & x_{n1}x_{n2} \\ x_{n1}x_{n2} & x_{n2}^2 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Los valores de x_{n1} y x_{n2} son observados por el investigador y los parámetros b y σ_β son estimados. De esta forma, el investigador obtiene tanto la media como la varianza del coeficiente aleatorio en la población. La generalización de este caso a más de una variable explicativa y más de dos alternativas es directa.

5.4 Patrones de sustitución y fallo de la IIA

Probit puede representar cualquier patrón de sustitución. Las probabilidades probit no exhiben la propiedad de IIA que da lugar a la sustitución proporcional de logit. Diferentes matrices de covarianza Ω proporcionan diferentes patrones de sustitución y, mediante la estimación de la matriz de covarianza, el investigador determina el patrón de sustitución que es más adecuado para sus datos.

Es posible estimar una matriz de covarianza completa o, alternativamente, el investigador puede imponer cierta estructura en la matriz de covarianza para representar fuentes particulares de no independencia. Esta estructura suele reducir el número de parámetros y facilita su interpretación. Consideraremos en primer lugar la situación en la que el investigador estima una matriz de covarianza completa y posteriormente una situación en la que el investigador impone una estructura sobre la matriz de covarianza.

Covarianza Completa: patrones de sustitución no restringidos

Para simplificar la notación, considere un modelo probit con cuatro alternativas. Una matriz de covarianza completa para los componentes no observados de utilidad toma la forma de Ω en (5.5). Cuando normalizamos la escala y el nivel, la matriz de covarianza se convierte en $\widetilde{\Omega}_1^*$ en (5.6). Los elementos de $\widetilde{\Omega}_1^*$ son estimados. Los valores estimados pueden representar cualquier tipo de patrón de sustitución; es importante destacar que la normalización de la escala y el nivel no restringe los patrones de sustitución. La normalización sólo elimina los aspectos de Ω que son irrelevantes para el comportamiento.

Observe, sin embargo, que los valores estimados de las θ^*s no proporcionan esencialmente ninguna información interpretable en ellas mismas (Horowitz, 1991). Por ejemplo, supongamos que estimamos que θ_{33}^* es mayor que θ_{44}^*. Puede ser tentador interpretar este resultado como una indicación de que la varianza en la utilidad no observada de la tercera alternativa es mayor que la de la cuarta alternativa, es decir, que $\sigma_{33} > \sigma_{44}$. Sin embargo, esta interpretación no es correcta. Es perfectamente posible que $\theta_{33}^* > \theta_{44}^*$ y sin embargo $\sigma_{44} > \sigma_{33}$, si la covarianza σ_{14} es suficientemente mayor a σ_{13}. Del mismo modo, supongamos que se estima que el valor θ_{23}^* es negativo. Esto no significa que la utilidad no observada de la segunda alternativa se correlacione negativamente con la utilidad no observada de la tercera alternativa (es decir, $\sigma_{23} < 0$). Es posible que σ_{23} sea positiva y sin embargo σ_{12} y σ_{13} sean suficientemente grandes para hacer θ_{23}^* sea negativa. El punto a destacar aquí es que la estimación de una matriz de covarianza completa permite que el modelo represente cualquier patrón de sustitución, pero hace que los parámetros estimados sean esencialmente imposibles de interpretar.

Covarianza estructurada: patrones de sustitución restringidos

Al imponer estructura en la matriz de covarianza, los parámetros estimados por lo general se vuelven más interpretables. La estructura es una restricción en la matriz de covarianza y, como tal, reduce la capacidad del modelo de representar diversos patrones de sustitución. Sin embargo, si la estructura es correcta (es decir, representa realmente el comportamiento de los decisores), entonces el verdadero patrón de sustitución podrá ser representado por la matriz de covarianza restringida.

La estructura es necesariamente dependiente de la situación: una estructura adecuada para una matriz de covarianza depende de las características específicas de la situación de elección que se está modelando. En la sección 5.2 se describen varios estudios que utilizan diferentes tipos de estructura. Como ejemplo de cómo se puede imponer estructura a la matriz de covarianza y por lo tanto a los patrones de sustitución, considere la elección que un comprador de vivienda hace entre diferentes tipos de hipotecas. Supongamos que el comprador puede escoger entre cuatro hipotecas de cuatro entidades financieras diferentes: una con un tipo de interés fijo y tres con tipos variables. Supongamos que la parte no observada de la utilidad de esta decisión se compone de dos partes: la preocupación del

comprador de vivienda por el riesgo de incremento de los tipos de interés, etiquetada r_n, y que es común a todos los préstamos de tipo variable, y todos los demás factores no observados, etiquetados colectivamente η_{nj}. El componente no observado de utilidad es por lo tanto

$$\varepsilon_{nj} = -r_n d_j + \eta_{nj},$$

donde $d_j = 1$ para los préstamos a tipo variable y 0 para el préstamo a tipo fijo, y donde el signo negativo indica que la utilidad disminuye a medida que la preocupación por el riesgo aumenta. Supongamos que r_n se distribuye normalmente sobre los compradores de vivienda con varianza σ y que η_{nj} $\forall j$ es normal iid con media cero y varianza ω. La matriz de covarianza para $\varepsilon_n = \langle \varepsilon_{n1}, \dots, \varepsilon_{n4} \rangle$ es

$$\Omega = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \cdot & \sigma & \sigma & \sigma \\ \cdot & \cdot & \sigma & \sigma \\ \cdot & \cdot & \cdot & \sigma \end{pmatrix} + \omega \begin{pmatrix} 1 & 0 & 0 & 0 \\ \cdot & 1 & 0 & 0 \\ \cdot & \cdot & 1 & 0 \\ \cdot & \cdot & \cdot & 1 \end{pmatrix}.$$

El modelo necesita ser normalizado para la escala pero, como veremos, ya está normalizado para el nivel. La covarianza de las diferencias de error es

$$\Omega = \begin{pmatrix} \sigma & \sigma & \sigma \\ \cdot & \sigma & \sigma \\ \cdot & \cdot & \sigma \end{pmatrix} + \omega \begin{pmatrix} 2 & 1 & 1 \\ \cdot & 2 & 1 \\ \cdot & \cdot & 2 \end{pmatrix}.$$

Esta matriz no tiene menos parámetros que Ω. Es decir, el modelo ya estaba normalizado para el nivel. Para normalizar la escala, fijamos $\sigma + 2\omega = 1$. La matriz de covarianza se convierte en

$$\widetilde{\Omega}_1^* = \begin{pmatrix} 1 & \theta & \theta \\ \cdot & 1 & \theta \\ \cdot & \cdot & 1 \end{pmatrix},$$

donde $\theta = (\sigma + \omega)/(\sigma + 2\omega)$. Los valores de σ y ω no se pueden calcular a partir de θ. Sin embargo, el parámetro θ proporciona información sobre la varianza en la utilidad debida a la preocupación por el riesgo respecto a la varianza debida a todos los demás factores no observados. Por ejemplo, supongamos que se estima θ en 0.75. Esta estimación puede ser interpretada como una indicación de que la varianza en la utilidad atribuible a la preocupación sobre el riesgo es dos veces la varianza en la utilidad atribuible a todos los demás factores:

$$\theta = 0.75,$$

$$\frac{\sigma + \omega}{\sigma + 2\omega} = 0.75,$$

$$\sigma + \omega = 0.75\sigma + 1.5\omega,$$

$$0.25\sigma = 0.5\omega,$$

$$\sigma = 2\omega,$$

Dicho de forma equivalente, $\hat{\theta} = 0.75$ significa que la preocupación por el riesgo representa dos tercios de la varianza en el componente no observado de la utilidad.

Dado que el modelo original ya estaba normalizado para el nivel, el modelo podría ser estimado sin reformular la matriz de covarianza en términos de las diferencias de error. La normalización de la escala podría lograrse simplemente estableciendo $\omega = 1$ en la Ω original. Usando este procedimiento, el parámetro σ es estimado directamente. Su valor en relación a 1 indica la varianza debida a la preocupación por el riesgo en relación a la varianza debida a la percepción sobre la facilidad de tratar con cada entidad financiera. Una estimación de $\hat{\theta} = 0.75$ corresponde a una estimación de $\hat{\sigma} = 2$.

5.5 Datos de panel

El modelo probit para elecciones repetidas es similar al probit para una elección por decisor. La única diferencia es que la dimensión de la matriz de covarianza de los errores se ve expandida. Considere un decisor que se enfrenta a una elección entre J alternativas en cada uno de los T períodos de tiempo o situaciones de elección. Las alternativas pueden cambiar a lo largo del tiempo, y J y T pueden diferir para diferentes decisores; sin embargo, suprimimos la notación para estas posibilidades. La utilidad que el decisor n obtiene de la alternativa j en el período T es $U_{njt} = V_{njt} + \varepsilon_{njt}$. En general, sería de esperar que ε_{njt} estuviese correlacionado en el tiempo así como respecto a otras alternativas, dado que los factores no observados por el investigador pueden persistir a lo largo del tiempo. Denotemos el vector de errores para todas las alternativas en todos los períodos de tiempo como $\varepsilon_n' = \langle \varepsilon_{n11}, \dots, \varepsilon_{nJ1}, \varepsilon_{n12}, \dots, \varepsilon_{nJ2}, \dots, \varepsilon_{n1T}, \dots, \varepsilon_{nJT} \rangle$. La matriz de covarianza para este vector se denomina Ω, que tiene dimensiones JT × JT.

Considere una secuencia de alternativas concreta, una alternativa para cada período de tiempo, $\mathbf{i} = \{i_1, \dots, i_T\}$. La probabilidad de que el decisor haga esta secuencia de elecciones es

$$P_{n\boldsymbol{i}} = Prob(U_{ni_t t} > U_{njt} \; \forall j \neq i_t, \forall t)$$

$$= Prob(V_{ni_t t} + \varepsilon_{ni_t t} > V_{njt} + \varepsilon_{njt} \; \forall j \neq i_t, \forall t)$$

$$= \int_{\varepsilon_n \in B_n} \phi(\varepsilon_n) d\varepsilon_n,$$

donde $B_n = \{\varepsilon_n \; s.t. V_{ni_t t} + \varepsilon_{ni_t t} > V_{njt} + \varepsilon_{njt} \; \forall j \neq i_t, \forall t\}$ y $\phi(\varepsilon_n)$ es la densidad normal conjunta con media cero y covarianza Ω. En comparación con la probabilidad probit para una situación de una única elección, la integral simplemente se ha ampliado hasta JT dimensiones en lugar de J.

A menudo es más conveniente trabajar con diferencias de utilidad. La probabilidad de la secuencia $\mathbf{i}$ es la probabilidad de que las diferencias de utilidad sean negativas para cada alternativa en cada período de tiempo, cuando las diferencias se calculan respecto a la alternativa identificada por $\mathbf{i}$ para ese período de tiempo:

$$P_{ni} = Prob(\widetilde{U}_{nji_t t} < 0 \; \forall j \neq i_t, \forall t)$$

$$= \int_{\tilde{\varepsilon}_n \in \tilde{B}_n} \phi(\tilde{\varepsilon}_n) d\tilde{\varepsilon}_n,$$

donde $\widetilde{U}_{nji_t t} = U_{njt} - U_{ni_t t}$; $\tilde{\varepsilon}_n' = \langle (\varepsilon_{n11} - \varepsilon_{ni_1 1}), \dots, (\varepsilon_{nJ1} - \varepsilon_{ni_1 1}), \dots, (\varepsilon_{n1T} - \varepsilon_{ni_T T}), \dots, (\varepsilon_{nJT} - \varepsilon_{ni_T T}) \rangle$ con cada "…" refiriéndose a todas las alternativas excepto i, y $\tilde{B}_n$ es el conjunto de $\tilde{\varepsilon}_n$s para las que $\widetilde{U}_{nji_t t} < 0 \; \forall j \neq i_t, \forall t$. Esta es una integral $(J-1)$T-dimensional. La densidad $\phi(\tilde{\varepsilon}_n)$ se distribuye

normalmente con matriz de covarianza obtenida a partir de Ω. La simulación de la probabilidad de elección es la misma que para situaciones con una elección por decisor, descrita en la Sección 5.6, pero con una dimensión mayor tanto en la matriz de covarianza como en la integral. Borsch-Supan et al. (1991) proporcionan un ejemplo de un probit multinomial sobre datos de panel que permite covarianza en el tiempo y entre alternativas.

Para elecciones binarias, tales como si una persona compra un producto en particular en cada período de tiempo o si trabaja en un puesto de trabajo remunerado cada mes, el modelo probit se simplifica considerablemente (Gourieroux y Monfort, 1993). La utilidad neta de tomar la acción (por ejemplo, trabajar) en el período t es $U_{nt} = V_{nt}+\varepsilon_{nt}$, y la persona realiza la acción si $U_{nt} > 0$. Esta utilidad se llama utilidad neta, ya que es la diferencia entre la utilidad de tomar la acción y no tomarla. Como tal, ya está expresada en términos de diferencias. Los errores están correlacionados en el tiempo, y la matriz de covarianza para $\varepsilon_{n1}, \dots, \varepsilon_{nT}$ es Ω, que es de dimensiones $T \times T$.

Una secuencia de elecciones binarias se representa más fácilmente por un conjunto de T variables indicadoras (*dummy*): $d_{nt} = 1$ si la persona n ha tomado la acción en el período t y $d_{nt} = -1$ en caso contrario. La probabilidad de que se produzca la secuencia de elecciones $d_n = d_{n1}, \dots, d_{nT}$ es

$$P_{nd_n} = Prob(U_{nt}d_{nt} > 0 \; \forall t)$$

$$= Prob(V_{nt}d_{nt} + \varepsilon_{nt}d_{nt} > 0 \; \forall t)$$

$$= \int_{\varepsilon_n \in B_n} \phi(\varepsilon_n) d\varepsilon_n,$$

donde B_n es el conjunto de ε_ns para las que $V_{nt}d_{nt} + \varepsilon_{nt}d_{nt} > 0 \; \forall t$ y $\phi(\varepsilon_n)$ es la densidad normal conjunta con covarianza Ω.

Es posible aplicar cierta estructura a la covarianza de los errores a lo largo del tiempo. Supongamos que en el caso binario, por ejemplo, el error consiste en una parte que fija específica del decisor, reflejando en qué medida es proclive a tomar la acción, y una parte que varía en el tiempo para cada decisor: $\varepsilon_{nt} = \eta_n+\mu_{nt}$, donde μ_{nt} es iid a lo largo del tiempo y entre personas, con una densidad normal estándar, y η_n es iid entre personas con una densidad de probabilidad normal con media cero y varianza σ. La varianza del error en cada período es $V(\varepsilon_{nt}) = V(\eta_n+\mu_{nt}) = \sigma + 1$. La covarianza entre los errores en dos períodos diferentes t y s es $\mathrm{Cov}(\varepsilon_{nt}, \varepsilon_{ns}) = \mathrm{E}(\eta_n+\mu_{nt})(\eta_n+\mu_{ns}) = \sigma$. Por tanto, la matriz de covarianza toma la forma

$$\Omega = \begin{pmatrix} \sigma+1 & \sigma & \dots & \dots & \sigma \\ \sigma & \sigma+1 & \sigma & \dots & \sigma \\ \dots & \dots & \dots & \dots & \dots \\ \sigma & \dots & \dots & \sigma & \sigma+1 \end{pmatrix}.$$

Sólo un parámetro, σ, entra en la matriz de covarianza. Su valor indica la varianza en la utilidad no observada entre individuos (la varianza de η_n) en relación con la varianza a lo largo del tiempo para cada individuo (la varianza de μ_{nt}). A menudo, este parámetro recibe el nombre de *varianza entre-sujetos relativa a la varianza intra-sujetos* (*cross-subject variance relative to the within-subject variance*).

Bajo esta estructura aplicada a los errores, las probabilidades de elección se pueden simular fácilmente utilizando los conceptos de partición conveniente del error, tratados en la sección 1.2. Condicionada a η_n, la probabilidad de no tomar la acción en el período t es $\mathrm{Prob}(V_{nt} + \eta_n + \mu_{nt} < 0) = \mathrm{Prob}(\mu_{nt} <$

$-(V_{nt}+\eta_n)\big) = \Phi(-(V_{nt}+\eta_n))$, donde $\Phi(\cdot)$ es la distribución normal acumulativa. La mayoría de los paquetes de software estadístico comerciales incluyen rutinas para calcular esta función. La probabilidad de tomar la acción condicionada a η_n, es $1-\Phi\big(-(V_{nt}+\eta_n)\big) = \Phi(V_{nt}+\eta_n)$. La probabilidad de la secuencia de elecciones d_n, condicionada a η_n, es por lo tanto $\prod_t \Phi((V_{nt}+\eta_n)\mathrm{d}_{nt})$, que podemos etiquetar como $\mathrm{H}_{nd_n}(\eta_n)$.

Hasta ahora hemos condicionado a η_n, cuando en realidad η_n es aleatoria. La probabilidad *no condicionada* es la integral de la probabilidad condicionada $\mathrm{H}_{nd_n}(\eta_n)$ sobre todos los valores posibles de η_n:

$$P_{nd_n} = \int H_{nd_n}(\eta_n)\, \phi(\eta_n) d\eta_n$$

donde $\phi(\eta_n)$ es la densidad normal con media cero y varianza σ. Esta probabilidad se puede simular muy fácilmente de la siguiente manera:

1. Extraiga un valor al azar de una densidad normal estándar utilizando un generador de números aleatorios. Multiplique el valor extraído por $\sqrt{\sigma}$, de manera que se convierta en una extracción de η_n, de una densidad normal con varianza σ.
2. Para esta extracción de η_n, calcule $\mathrm{H}_{nd_n}(\eta_n)$.
3. Repita los pasos 1-2 numerosas veces y promedie los resultados. Este promedio es una aproximación simulada de P_{nd_n}.

Este simulador es mucho más fácil de calcular que los simuladores probit generales que se describen en la siguiente sección. La posibilidad de utilizarlo surge de la estructura que impusimos al modelo, es decir, de imponer que la dependencia temporal de los factores no observados podía ser capturada en su totalidad por un componente aleatorio η_n que se mantiene constante en el tiempo para cada persona. Gourieroux y Monfort (1993) proporcionan un ejemplo de la utilización de este simulador con un modelo probit de este tipo.

La utilidad representativa en un período de tiempo puede incluir variables exógenas para otros períodos de tiempo, tal y como ya hemos comentado respecto a los modelos logit sobre datos de panel (sección 3.3.3). Es decir, V_{nt} puede incluir variables exógenas que se refieren a períodos distintos de t. Por ejemplo, una respuesta diferida a cambios de precios se puede representar mediante la inclusión de los precios en períodos anteriores en la V del período actual. Una conducta anticipatoria (por la cual, por ejemplo, una persona compra un producto ahora porque anticipa correctamente que el precio se incrementará en el futuro) puede ser representada incluyendo los precios previstos en períodos futuros en la V del período actual.

Introducir una variable dependiente diferida es posible, pero introduce dos dificultades que el investigador debe abordar. En primer lugar, dado que los errores están correlacionados en el tiempo, la elección en un período está correlacionada con los errores en períodos posteriores. Como resultado, la inclusión de una variable dependiente diferida sin ajustar convenientemente el procedimiento de estimación da como resultado estimaciones inconsistentes. Este problema es análogo al del análisis de regresión, donde el estimador de mínimos cuadrados ordinarios es inconsistente cuando se incluye una variable dependiente diferida y los errores están correlacionados en serie. Para estimar un probit consistentemente en esta situación, el investigador debe determinar la distribución de cada ε_{nt} condicionada al valor de las variables dependientes diferidas. La probabilidad de elección se basa en esta distribución condicionada en lugar de basarse en la distribución no condicionada $\phi(\cdot)$ que se utilizó anteriormente. En segundo lugar, el investigador a menudo no puede observar las elecciones de los decisores desde la primera elección que estos tuvieron a su disposición. Por ejemplo, un investigador que estudia los patrones de empleo tal vez

observe la situación laboral de una persona durante un período de tiempo (por ejemplo, 1998-2001), pero por lo general no va a observar la situación laboral de la persona desde la primera vez que esa persona podría haber tenido un trabajo (algo que podría ser muy anterior a 1998). En este caso, la probabilidad para el primer período que el investigador observa depende de las decisiones de la persona en los períodos anteriores que el investigador no observa. El investigador debe determinar una forma de representar la primera probabilidad de elección que permita una estimación consistente teniendo en cuenta la información perdida de las elecciones anteriores. Esto se conoce como el *problema de las condiciones iniciales* (*initial conditions problem*) de los modelos de elección dinámicos. Ambas cuestiones, así como los posibles enfoques para tratar con ellas, han sido abordadas por Heckman (1981b, 1981a) y Heckman y Singer (1986). Debido a su complejidad, no describo los procedimientos aquí y remito al lector interesado - y valiente - a que lea estos artículos.

Papatla y Krishnamurthi (1992) evitan estos problemas en su modelo probit con variables dependientes diferidas, al asumir que los factores no observados son independientes en el tiempo. Como ya comentamos en relación con el modelo logit para datos de panel (Sección 3.3.3), las variables dependientes diferidas no se correlacionan con los errores actuales cuando los errores son independientes en el tiempo y por lo tanto se pueden introducir sin inducir inconsistencia. Por supuesto, este procedimiento sólo es apropiada si el supuesto de errores independientes en el tiempo es realmente cierto, en lugar de ser simplemente un supuesto.

5.6 Simulación de las probabilidades de elección

Las probabilidades probit no tienen una expresión cerrada por lo que deben aproximarse numéricamente. Para ello, se han empleado varios procedimientos sin simulación que pueden ser efectivos en ciertas circunstancias.

Los métodos de cuadratura numérica aproximan la integral mediante una función ponderada de puntos de evaluación especialmente elegidos. Una buena explicación de estos procedimientos la proporciona Geweke (1996). Ejemplos de su uso para probit los podemos encontrar en Butler y Moffitt (1982) y Guilkey y Murphy (1993). La cuadratura numérica opera de forma efectiva cuando la dimensión de la integral es pequeña, pero no sucede así con dimensiones altas. Se puede utilizar para modelos probit si el número de alternativas (o, con datos de panel, el número de alternativas por el número de períodos de tiempo) no es mayor a cuatro o cinco. También se puede utilizar si el investigador ha especificado una estructura de componentes de error con no más de cuatro o cinco términos. Sin embargo, no es eficaz para modelos probit generales. E incluso para integración con pocas dimensiones, la simulación es a menudo más fácil

Otro procedimiento sin simulación que ha sido sugerido es el algoritmo de Clark, introducido por Daganzo et al. (1977). Este algoritmo utiliza el hecho, mostrado por Clark (1961), de que el máximo de varias variables distribuidas normalmente es en sí mismo aproximadamente una distribución normal. Desafortunadamente, la aproximación puede ser muy imprecisa en algunas circunstancias (como se muestra por Horowitz et al., 1982) y el grado de precisión es difícil de evaluar para un contexto determinado.

Por último, la simulación ha demostrado ser muy general y útil para aproximar probabilidades probit. Se han propuesto numerosos simuladores para los modelos probit, un resumen lo proporciona Hajivassiliou et al. (1996). En la sección anterior he descrito un simulador que es apropiado para un modelo probit que tiene una estructura particularmente conveniente: un probit binario sobre datos de panel donde la dependencia del tiempo es capturada por un factor aleatorio. En esta sección, describo tres simuladores que son aplicables para probits de todo tipo: simulador aceptación-rechazo, simulador aceptación-rechazo suavizado y GHK. El simulador GHK es, con mucha diferencia, el simulador probit más utilizado, por razones que se tratan a continuación. Los otros dos métodos son valiosos

pedagógicamente. También tienen relevancia más allá de los modelos probit y pueden aplicarse en prácticamente cualquier situación. Pueden ser muy útiles cuando el investigador está desarrollando sus propios modelos en lugar de emplear modelos probit o cualquier otro modelo descrito en este libro.

5.6.1 Simulador por aceptación-rechazo

El método de aceptación-rechazo (*accept-reject,* AR) es el simulador más directo. Considere la simulación de P_{ni}. Se extraen valores al azar de los términos aleatorios a partir de sus distribuciones. Para cada valor extraído, el investigador determina si esos valores de los errores, al ser combinados con las variables observadas que afronta la persona n, darían lugar a que la alternativa i fuese la elegida. Si es así, el valor extraído se califica como una *aceptación*. Si el valor extraído resultaría en la elección de otra alternativa, el valor es un *rechazo*. La probabilidad simulada es la proporción de valores extraídos que son aceptados. Este procedimiento se puede aplicar a cualquier modelo de elección con cualquier distribución de los términos aleatorios. Se propuso originalmente para probits (Manski y Lerman, 1981), y por ello facilitamos los detalles del enfoque adoptado por este método en términos del modelo probit. Su uso para otros modelos es obvio.

Utilizamos la expresión (5.1) para las probabilidades probit:

$$P_{ni} = \int I\left(V_{ni} + \varepsilon_{ni} > V_{nj} + \varepsilon_{nj}\ \forall j \neq i\right)\phi(\varepsilon_n)d\varepsilon_n,$$

donde $\mathrm{I}(\cdot)$ es una función indicadora de si la expresión entre paréntesis es verdadera y $\phi(\varepsilon_n)$ es la densidad normal conjunta con media cero y covarianza Ω. El simulador AR de esta integral se calcula como sigue:

1. Haga una extracción de valores al azar para el vector J-dimensional de errores ε_n, a partir de una densidad normal con media cero y covarianza Ω. Etiquete el vector de valores extraído como ε_n^r con $\mathrm{r} = 1$ y los elementos de la extracción como $\varepsilon_{n1}^r, \ldots, \varepsilon_{nJ}^r$.
2. Utilizando estos valores para los errores, calcule la utilidad que cada alternativa obtiene con estos errores. Es decir, calcule $\mathrm{U}_{nj}^r = \mathrm{V}_{nj} + \varepsilon_{nj}^r\ \forall j$.
3. Determine si la utilidad de la alternativa i es mayor a la de todas las otras alternativas. Es decir, calcule $\mathrm{I}^r = 1$ si $\mathrm{U}_{ni}^r > \mathrm{U}_{nj}^r \forall j \neq i$, lo que indicaría una aceptación, y $\mathrm{I}^r = 0$ en cualquier otro caso, indicando un rechazo.
4. Repita los pasos 1-3 múltiples veces. Etiquete el número de repeticiones (incluyendo la primera) como R, de modo que r toma valores de 1 a R.
5. La probabilidad simulada es la proporción de extracciones que son aceptaciones: $\breve{\mathrm{P}}_{ni} = \frac{1}{R}\sum_{r=1}^{R} \mathrm{I}^r$.

La integral $\int \mathrm{I}(\cdot)\phi(\varepsilon_n)d\varepsilon$ se aproxima por el promedio $\frac{1}{R}\sum_{r=1}^{R} \mathrm{I}^r(\cdot)$ para extracciones de valores de $\phi(\cdot)$. Obviamente, $\breve{\mathrm{P}}_{ni}$ es un estimador no sesgado de P_{ni}: $\mathrm{E}\left(\breve{\mathrm{P}}_{ni}\right) = \frac{1}{R}\sum E[\mathrm{I}^r(\cdot)] = \frac{1}{R}\sum \mathrm{P}_{ni} = \mathrm{P}_{ni}$, donde la expectativa es entre diferentes conjuntos de R extracciones. La varianza de $\breve{\mathrm{P}}_{ni}$ entre diferentes conjuntos de extracciones disminuye a medida que el número de extracciones se eleva. El simulador es a menudo llamado el "simulador de frecuencia cruda" ("*crude frequency simulator"),* ya que es la frecuencia de veces que la extracción de valores de los errores produce que la alternativa especificada sea la elegida. La palabra "crudo" distingue este simulador del simulador de frecuencia suavizado que describimos en la siguiente sección.

El primer paso del simulador AR para un modelo probit es extraer un valor al azar de una densidad normal conjunta. La siguiente pregunta surge de inmediato: ¿cómo se obtienen estas extracciones? El

procedimiento más sencillo es el descrito en la Sección 9.2.5, que utiliza el factor Choleski. La matriz de covarianza de los errores es Ω. Un factor Choleski de Ω es una matriz triangular inferior L tal que $\mathrm{LL}' = \Omega$. Este factor es llamado en ocasiones la raíz cuadrada generalizada de Ω. La mayoría de los paquetes de software estadístico contienen rutinas para calcular el factor Choleski de cualquier matriz simétrica. Supongamos ahora que η es un vector de J normales estándar iid, tal que $\eta \sim \mathrm{N}(0, \mathrm{I})$, donde I es la matriz identidad. Este vector se puede obtener extrayendo J valores de un generador de números aleatorios de una distribución normal estándar, agrupándolos luego en un vector. Podemos construir un vector ε que se distribuya $\mathrm{N}(0, \Omega)$ usando el factor Choleski para transformar η. Concretamente, calculamos $\varepsilon = \mathrm{L}\eta$. Dado que la suma de normales es normal, ε se distribuye normalmente. Y como η tiene media cero, también ε tiene media cero. La covarianza de ε es $Cov(\varepsilon) = \mathrm{E}(\,\varepsilon\,\varepsilon') = \mathrm{E}(\mathrm{L}\eta(\mathrm{L}\eta)') = \mathrm{E}(\mathrm{L}\eta\eta'\mathrm{L}') = \mathrm{LE}(\eta\eta')\mathrm{L}' = \mathrm{LIL}' = \mathrm{LL}' = \Omega$.

Utilizando el factor Choleski L de Ω, el primer paso del simulador AR se descompone en dos sub-etapas:

1A. Extraiga J valores al azar de una densidad normal estándar, utilizando un generador de números aleatorios. Agrupe estos valores en un vector y etiquete el vector como η^r.

1B. Calcule $\varepsilon_n^r = L\eta^r$.

Posteriormente, utilizando ε_n^r, calcule la utilidad de cada alternativa y verifique si la alternativa i es la que tiene mayor utilidad.

El procedimiento que hemos descrito funciona empleando utilidades y la expresión (5.1), que es una integral J-dimensional. El procedimiento se puede aplicar de forma análoga a las diferencias de utilidad, lo que reduce la dimensión de la integral a $\mathrm{J} - 1$. Como se ha indicado en (5.3), las probabilidades de elección se pueden expresar en términos de diferencias de utilidad:

$$P_{ni} = \int I\big(\tilde{V}_{nji} + \tilde{\varepsilon}_{nji} < 0 \;\forall j \neq i\big)\boldsymbol{\phi}(\tilde{\varepsilon}_{ni})d\tilde{\varepsilon}_{ni},$$

donde $\phi(\tilde{\varepsilon}_{ni})$ es la densidad normal conjunta con media cero y covarianza $\widetilde{\Omega}_i = M_i \Omega M_i'$. Esta integral puede ser simulada con métodos AR a través de los siguientes pasos:

1. Extraiga un valor $\tilde{\varepsilon}_{ni}^r = L_i\eta^r$ de la siguiente manera:
   a. Extraiga $\mathrm{J} - 1$ valores de una densidad normal estándar utilizando un generador de números aleatorios. Agrupe estos valores en un vector y etiquete el vector como η^r.
   b. Calcule $\tilde{\varepsilon}_{ni}^r = L_i\eta^r$, donde L_i es el factor Choleski de $\widetilde{\Omega}_i$.
2. Utilizando estos valores de los errores, calcule la diferencia de utilidad para cada alternativa, respecto a la utilidad de la alternativa i. Es decir, calcule $\widetilde{\mathrm{U}}_{nji}^r = V_{nj} - V_{ni} + \tilde{\varepsilon}_{nji}^r \;\forall j \neq i$.
3. Determine si cada diferencia de utilidad es negativa. Es decir, calcule $I^r = 1$ si $\widetilde{\mathrm{U}}_{nji}^r < 0 \;\forall j \neq i$, lo que indicaría una aceptación, y $I^r = 0$ en caso contrario, lo que indicaría un rechazo.
4. Repita los pasos 1 a 3 R veces.
5. La probabilidad simulada es el número de aceptaciones dividido por el número de repeticiones: $\widetilde{\mathrm{P}}_{ni} = \frac{1}{R}\sum_{r=1}^{R} \mathrm{I}^r$.

Usar diferencias de utilidad es ligeramente más rápido computacionalmente que usar las utilidades en sí mismas, ya que reducimos una dimensión. Sin embargo, a menudo es más fácil conceptualmente trabajar con utilidades.

Como acabamos de indicar, el simulador AR es muy general. Puede ser aplicado a cualquier modelo para el cual podamos extraer valores al azar de sus términos aleatorios y para el que sea posible determinar el comportamiento que el decisor adoptaría con esos valores. Asimismo, es un simulador muy intuitivo,

lo que representa una ventaja desde el punto de vista de programación, ya que la depuración se convierte en una tarea relativamente fácil. Sin embargo, el simulador AR tiene varias desventajas, especialmente cuando se utiliza en el contexto de la estimación de máxima verosimilitud.

Recordemos que la función log-verosimilitud es $LL = \sum_n \sum_j d_{nj} log P_{nj}$, donde $d_{nj} = 1$ si n eligió j y 0 en caso contrario. Cuando las probabilidades no se pueden calcular con exactitud, como en el caso del modelo probit, se utiliza la función log-verosimilitud simulada en lugar de la log-verosimilitud real, reemplazando las verdaderas probabilidades por las probabilidades simuladas: $SLL = \sum_n \sum_j d_{nj} log \check{P}_{nj}$. El valor de los parámetros que maximiza la SLL recibe el nombre de estimador de máxima verosimilitud simulada (*máximum simulated likelihood estimator*, MSLE). Es, con diferencia, el procedimiento de estimación basada en simulación más utilizado. Sus propiedades se describen en el capítulo 8. Desafortunadamente, usar el simulador AR en la SLL puede ser problemático.

Los problemas son dos. En primer lugar, $\check{P}_{ni}$ puede ser cero para cualquier número finito de extracciones R. Es decir, es posible que cada uno de los R valores extraídos de los términos de error dé como resultado un rechazo, de manera que la probabilidad simulada resulte cero. Los valores cero para $\check{P}_{ni}$ son problemáticos debido a que calculamos el logaritmo de $\check{P}_{ni}$ cuando entra en la función de verosimilitud y el logaritmo de cero es indeterminado. SLL no puede calcularse si la probabilidad simulada es cero para algún decisor de la muestra.

La ocurrencia de una probabilidad simulada igual a cero es particularmente probable cuando la verdadera probabilidad es baja. A menudo, al menos un decisor de la muestra habrá hecho una elección que tenga baja probabilidad. Por ejemplo, cuando los decisores se enfrentan a numerosas alternativas (como miles de marcas y modelos en la elección de un automóvil), cada alternativa tiene baja probabilidad. Con elecciones repetidas, la probabilidad de cualquier secuencia concreta de elecciones puede ser extremadamente pequeña; por ejemplo, si la probabilidad de elegir una alternativa concreta es 0.25 en cada uno de los 10 períodos de tiempo en los que se repite una elección, la probabilidad de la secuencia consistente en repetir 10 veces la misma elección es $(0.25)^{10}$, que es menor a 0.000001.

Además de este problema, la SLL se debe calcular en cada paso del proceso de búsqueda de su máximo. Algunos de los valores de los parámetros para los que necesitaremos calcular la SLL durante el proceso de maximización pueden estar muy lejos de los verdaderos valores. Durante el proceso, pueden aparecer probabilidades bajas incluso aunque éstas no existan en los valores que maximizan la SLL.

Siempre podemos obtener probabilidades simuladas no nulas al realizar suficientes extracciones. Sin embargo, si el investigador continúa extrayendo valores hasta obtener al menos una aceptación para cada decisor, el número de extracciones se convierte en una función de las probabilidades. El proceso de simulación deja de ser independiente del proceso de elección que se está modelando y las propiedades del estimador pasan a ser más complejas.

Existe una segunda dificultad en el uso del simulador AR para obtener el MSLE. Las probabilidades simuladas no son funciones suaves en relación a los parámetros, es decir, no son dos veces diferenciables. Como se explica en el capítulo 8, los procedimientos numéricos que se utilizan para localizar el máximo de la función log-verosimilitud se basan en las primeras derivadas y, en ocasiones, en las segundas derivadas, de las probabilidades de elección. Si no existen estas derivadas, o no se dirigen hacia el máximo, el procedimiento numérico no será efectivo.

La probabilidad AR simulada es una función escalonada, tal como se representa en la figura 5.1. $\check{P}_{ni}$ es la proporción de valores extraídos para los que la alternativa i tiene la utilidad mayor. Un cambio infinitesimalmente pequeño en un parámetro por lo general no va a producir que un valor extraído pase de ser una aceptación a un rechazo, y viceversa. Si U_{ni}^r está por debajo de U_{nj}^r para algunas alternativas j en un nivel determinado de los parámetros, también lo estará para un cambio infinitesimalmente

pequeño en cualquier parámetro. Así que, por lo general, $\check{P}_{ni}$ es constante respecto a pequeños cambios en los parámetros. Sus derivadas respecto a los parámetros son cero en este rango. Si los parámetros cambian de tal manera que un rechazo se convierte en una aceptación, entonces $\check{P}_{ni}$ se incrementa en una cantidad discreta, que va desde M/R a $(M+1)/R$, donde M es el número de aceptaciones contabilizadas en los valores originales de los parámetros. $\check{P}_{ni}$ es constante (pendiente cero) hasta que una aceptación se convierta en un rechazo, o viceversa, momento en el que $\check{P}_{ni}$ salta en una cantidad $1/R$. Su pendiente en ese momento no está definida. Por lo tanto, la primera derivada de $\check{P}_{ni}$ con respecto a los parámetros o es cero o es indefinida.

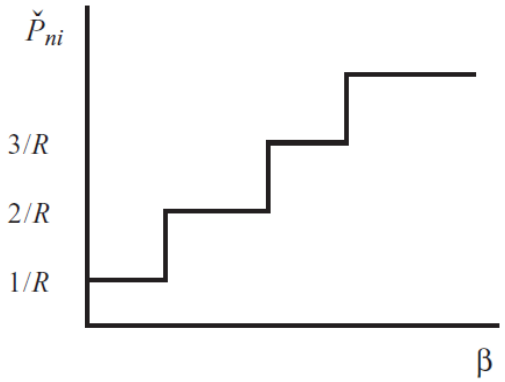


Figura 5.1. El simulador AR es una función escalonada en los parámetros.

Este hecho dificulta los procedimientos numéricos que se utilizan para localizar el máximo de la SLL. Como se trata en el Capítulo 8, los procedimientos de maximización utilizan el gradiente calculado en unos valores de prueba de los parámetros para determinar la dirección en la cual moverse para encontrar los parámetros con mayor SLL. Siendo la pendiente $\check{P}_{ni}$ para cada n cero o indefinida, el gradiente de la SLL es cero o indefinido. Este gradiente no proporciona ninguna ayuda en la búsqueda del máximo.

Este problema realmente no es tan drástico como parece. El gradiente de la SLL se puede aproximar como el cambio producido en la SLL para un cambio no infinitesimalmente pequeño en los parámetros. De esta forma, los parámetros se cambian en una cantidad que es lo suficientemente grande como para producir cambios entre aceptaciones y rechazos para, por lo menos, algunas de las observaciones. El gradiente aproximado, que puede ser llamado un gradiente de arco (*arc gradient*), se calcula como la cantidad en que ha cambiado la SLL dividida por el cambio en los parámetros. Siendo precisos: para un vector de parámetros β de longitud K, la derivada de la SLL respecto al parámetro k-ésimo se calcula como $(SLL^1 - SLL^0)/(\beta_k^1 - \beta_k^0)$, donde SLL^0 se calcula con los parámetros β originales con el elemento k-ésimo igual a β_k^0 y SLL^1 se calcula en β_k^1 con todos los demás parámetros iguales a sus valores originales. El gradiente de arco calculado de esta manera no es cero o indefinido, y proporciona información sobre la dirección de incremento. Sin embargo, la experiencia indica que aun así, la probabilidad simulada AR es difícil de usar.

5.6.2 Simuladores AR suavizados

Una forma de mitigar las dificultades del simulador AR es reemplazar el indicador AR 0-1 por una función suave y estrictamente positiva. La simulación comienza del mismo modo que con un simulador AR, extrayendo valores al azar de los términos aleatorios y calculando la utilidad de cada alternativa para cada valor extraído: U_{nj}^r. Pero en lugar de determinar si la alternativa i tiene la mayor utilidad (es decir, en lugar de calcular la función indicadora I^r), las utilidades simuladas $U_{nj}^r \forall j$ se introducen en una función. Puede utilizarse cualquier función para simular P_{ni} siempre y cuando se incremente cuando U_{ni}^r

se incremente y disminuya cuando el resto de utilidades U_{nj}^r se incrementen, sea estrictamente positiva y tenga definidas la primera y segunda derivadas respecto a $U_{nj}^r \forall j$. Una función particularmente adecuada es la función logit, como sugirió McFadden (1989). El uso de esta función da lugar al simulador AR suavizado-logit (*logit-smoothed AR simulator*).

El simulador se implementa mediante los pasos siguientes, que son los mismos del simulador AR excepto el paso 3:

1. Haga una extracción de valores al azar para el vector J-dimensional de errores ε_n, a partir de una densidad normal con media cero y covarianza Ω. Etiquete el vector de valores extraído como ε_n^r con $\mathrm{r} = 1$ y los elementos de la extracción como $\varepsilon_{n1}^r, \ldots, \varepsilon_{nJ}^r$.
2. Utilizando estos valores para los errores, calcule la utilidad que cada alternativa obtiene con estos errores. Es decir, calcule $\mathrm{U}_{nj}^r = \mathrm{V}_{nj} + \varepsilon_{nj}^r \; \forall j$.
3. Introduzca estas utilidades en la fórmula logit. Es decir, calcule

$$S^r = \frac{e^{U_{ni}^r/\lambda}}{\sum_j e^{U_{nj}^r/\lambda}},$$

   donde $\lambda > 0$ es un factor de escala especificado por el investigador y que se trata en el siguiente texto.
4. Repita los pasos 1-3 muchas veces. Etiquete el número de repeticiones (incluyendo la primera) como R, de modo que r toma valores de 1 a R.
5. La probabilidad simulada es el promedio de los valores de la fórmula logit: $\breve{\mathrm{P}}_{ni} = \frac{1}{R}\sum_{r=1}^{R} \mathrm{S}^r$.

Dado que $S^r > 0$ para cualquier valor finito de U_{nj}^r, la probabilidad simulada es estrictamente positiva para cualquier extracción de valores de error. Se incrementa cuando U_{ni}^r se incrementa y disminuye cuando $\mathrm{U}_{nj}^r \; j \neq i$ se incrementa. Es suave (dos veces diferenciable), dado que la propia fórmula logit es suave.

El simulador AR suavizado-logit se puede aplicar a cualquier modelo de elección, simplemente simulando las utilidades bajo la correspondiente hipótesis relativa a las distribuciones de los errores e insertando posteriormente las utilidades en la fórmula logit. Cuando este simulador se aplica al modelo probit, Ben-Akiva y Bolduc (1996) lo han denominado "probit logit-kernel" (*logit-kernel probit*).

El factor de escala λ determina el grado de suavizado. A medida que $\lambda \to 0$, S^r se acerca a la función indicadora I^r. La figura 5.2 ilustra esta circunstancia para un caso con dos alternativas. Para una extracción dada de ε_n^r, se calcula la utilidad de las dos alternativas. Considere la probabilidad simulada para la alternativa 1. Empleando AR, la función indicadora 0-1 es cero si U_{n1}^r está por debajo de U_{n2}^r y uno si U_{n1}^r excede U_{n2}^r. Empleando el suavizado-logit, la función escalonada se sustituye por una curva sigmoidea suave. El factor λ determina el grado de proximidad de la sigmoidea con la función indicadora 0-1. Bajar λ aumenta la escala de las utilidades cuando entran en la función logit (ya que las utilidades se dividen por λ). Incrementar la escala de utilidad aumenta la diferencia absoluta entre las dos utilidades. La fórmula logit da probabilidades que están más cerca de cero o uno cuando la diferencia de utilidades es mayor. Por lo tanto, la función logit suavizada S^r se vuelve más cercana a la función escalonada a medida que λ se aproxima más a cero.

El investigador necesita establecer el valor de λ. Un valor de λ menor hace del logit suave una mejor aproximación de la función indicadora. Sin embargo, este hecho es un arma de doble filo: si el logit suave se aproxima a la función indicadora demasiado bien, las dificultades numéricas del uso del simulador AR no suavizado simplemente se reproducirán en el simulador logit suavizado. Lo que quiere el investigador es establecer una λ lo suficientemente baja como para obtener una buena aproximación,

pero no tan baja como para reintroducir dificultades numéricas. Existen pocas recomendaciones a dar sobre el nivel apropiado de λ. Tal vez el mejor enfoque posible para el investigador es experimentar con diferentes λs. Para experimentar, los mismos valores extraídos de ε_n deberían ser usados con cada posible λ, de manera que aseguremos que las diferencias en los resultados son debidos al cambio en la λ y no a diferencias en los propios valores extraídos.

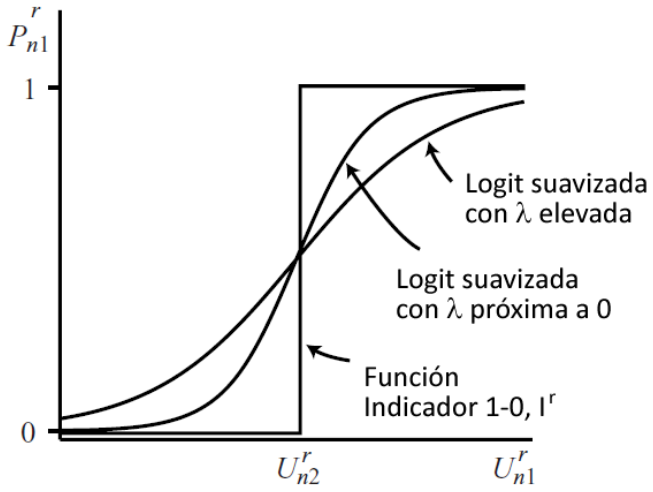


Figura 5.2. AR suavizado

McFadden (1989) describe otras funciones de suavizado. Para todas ellas, el investigador debe especificar el grado de suavizado. Una ventaja del logit suavizado es su simplicidad. Además, veremos en el capítulo 6 que el logit suavizado aplicado a un modelo probit o a cualquier otro modelo, constituye un tipo de especificación logit mixta. Es decir, en lugar de ver el logit suavizado como una aproximación que no tiene relación alguna con el modelo de comportamiento (sólo tiene un objetivo numérico), podemos verlo como el resultado de un tipo particular de estructura de error dentro del propio modelo de comportamiento. Según esta interpretación, la fórmula logit aplicada a las utilidades simuladas no es una aproximación, sino que en realidad representa el verdadero modelo.

5.6.3 Simulador GHK

El simulador probit más utilizado se denomina GHK, en referencia a los autores Geweke (1989, 1991), Hajivassiliou (tal y como se informa en Hajivassiliou y McFadden, 1998) y Keane (1990, 1994), quien desarrolló el procedimiento. En una comparación entre numerosos simuladores probit, Hajivassiliou et al. (1996) encontraron que el simulador GHK era el más preciso en las situaciones de elección que se examinaron. Geweke et al. (1994) encontraron que el simulador GHK funciona mejor que el AR suavizado. La experiencia ha confirmado su utilidad y exactitud relativa (por ejemplo, Borsch-Supan y Hajivassiliou, 1993).

El simulador GHK opera con diferencias de utilidades. La simulación de la probabilidad P_{ni} comienza restando la utilidad de la alternativa i de la utilidad de cada una del resto de las alternativas. Es importante destacar que se resta la utilidad de una alternativa diferente dependiendo de qué probabilidad se está simulando: para P_{ni}, U_{ni} es la utilidad restada de las otras utilidades, mientras que para P_{nj}, se resta U_{nj}. Este hecho es crítico para la aplicación del procedimiento.

Voy a explicar el procedimiento GHK en primer lugar para un caso de tres alternativas, ya que esta situación se puede representar gráficamente en dos dimensiones para las diferencias de utilidad. A continuación describiré el procedimiento en general, para cualquier número de alternativas. Bolduc

(1993, 1999) proporciona una excelente descripción alternativa del procedimiento, junto con los métodos para simular las derivadas analíticas de las probabilidades probit. Keane (1994) proporciona una descripción de la utilización de GHK para probabilidades de transición.

Tres alternativas

Empezamos con una especificación del modelo de comportamiento en las utilidades: $\mathrm{U}_{nj} = \mathrm{V}_{nj} + \varepsilon_{nj}, j = 1,2,3$. Suponemos el vector $\varepsilon'_n = \langle \varepsilon_{n1}, \varepsilon_{n2}, \varepsilon_{n3} \rangle \sim N(0, \Omega)$. Asumimos que el investigador ha normalizado el modelo para la escala y el nivel, por lo que los parámetros que entran en Ω están identificados. Asimismo, Ω puede ser una función paramétrica de los datos, así como incluir variación aleatoria de las preferencias, aunque no mostramos esta dependencia en nuestra notación.

Supongamos que queremos simular la probabilidad de la primera alternativa, P_{n1}. Podemos reformular el modelo en diferencias de utilidad substrayendo la utilidad de la alternativa 1:

$$U_{nj} - U_{n1} = \left(V_{nj} - V_{n1}\right) + \left(\varepsilon_{nj} - \varepsilon_{n1}\right),$$

$$\widetilde{U}_{nj1} = \tilde{V}_{nj1} + \tilde{\varepsilon}_{nj1},$$

para $j = 2,3$. El vector $\varepsilon'_{n1} = \langle \tilde{\varepsilon}_{n21}, \tilde{\varepsilon}_{n31} \rangle$ se distribuye $N(0, \widetilde{\Omega}_1)$,donde $\widetilde{\Omega}_1$ se obtiene de Ω.

Hacemos una transformación más para hacer el modelo más conveniente para la simulación. Para ello, sea L_1 el factor Choleski de $\widetilde{\Omega}_1$. Dado que $\widetilde{\Omega}_1$ es 2×2 en la situación de nuestro ejemplo, L_1 es una matriz triangular inferior que toma la forma

$$L_1 = \begin{pmatrix} C_{aa} & 0 \\ C_{ab} & C_{bb} \end{pmatrix}.$$

Usando este factor Choleski, las diferencias de error originales, que están correlacionadas, pueden reescribirse como funciones lineales de normales estándar *no correlacionadas*:

$$\tilde{\varepsilon}_{n21} = c_{aa}\eta_1,$$

$$\tilde{\varepsilon}_{n31} = c_{ab}\eta_1 + c_{bb}\eta_2,$$

donde η_1 y η_2 son $N(0,1)$ iid. Las diferencias de error $\tilde{\varepsilon}_{n21}$ y $\tilde{\varepsilon}_{n31}$ están correlacionadas porque ambas diferencias dependen de η_1. Expresando las diferencias de error de esta forma, las diferencias de utilidad pueden ser escritas como

$$\widetilde{U}_{n21} = \tilde{V}_{n21} + c_{aa}\eta_1,$$

$$\widetilde{U}_{n31} = \tilde{V}_{n31} + c_{ab}\eta_1 + c_{bb}\eta_2,$$

La probabilidad de la alternativa 1 es $\mathrm{P}_{\mathrm{n1}} = \mathrm{Prob}\left(\widetilde{\mathrm{U}}_{\mathrm{n21}} < 0 \; y \; \widetilde{\mathrm{U}}_{\mathrm{n31}} < 0\right) = \mathrm{Prob}(\widetilde{\mathrm{V}}_{\mathrm{n21}} + \tilde{\varepsilon}_{\mathrm{n21}} < 0 \; y \; \widetilde{\mathrm{V}}_{\mathrm{n31}} + \tilde{\varepsilon}_{\mathrm{n31}} < 0)$. Esta probabilidad es difícil de evaluar numéricamente en términos de los $\tilde{\varepsilon}$s, porque están correlacionados. Sin embargo, utilizando la transformación basada en el factor Choleski, la probabilidad se puede escribir de forma que involucre términos aleatorios independientes. La

probabilidad se convierte en una función de la distribución normal acumulativa estándar unidimensional:

$$P_{n1} = Prob(\tilde{V}_{n21} + c_{aa}\eta_1 < 0 \; y \; \tilde{V}_{n31} + c_{ab}\eta_1 + c_{bb}\eta_2 < 0)$$

$$= Prob\left(\tilde{V}_{n21} + c_{aa}\eta_1 < 0\right) \times Prob\left(\tilde{V}_{n31} + c_{ab}\eta_1 + c_{bb}\eta_2 < 0 \,\middle|\, \tilde{V}_{n21} + c_{aa}\eta_1 < 0\right)$$

$$= Prob\left(\eta_1 < -\tilde{V}_{n21}/c_{aa}\right) \times Prob\left(\eta_2 < -(\tilde{V}_{n31} + c_{ab}\eta_1)/c_{bb} \,\middle|\, \eta_1 < -\tilde{V}_{n21}/c_{aa}\right)$$

$$= \Phi\left(\frac{-\tilde{V}_{n21}}{c_{aa}}\right) \times \int_{\eta_1=-\infty}^{-\tilde{V}_{n21}/c_{aa}} \Phi\left(\frac{-\left(\tilde{V}_{n31} + c_{ab}\eta_1\right)}{c_{bb}}\right) \bar{\phi}(\eta_1) d\eta_1$$

donde $\Phi(\cdot)$ es la distribución normal estándar acumulativa evaluada en el punto indicado entre paréntesis y $\bar{\phi}(\cdot)$ es la densidad normal truncada[i]. El primer factor $\Phi\left(-\tilde{V}_{n21}/c_{aa}\right)$ es fácil de calcular: simplemente es la distribución normal acumulativa estándar evaluada en $-\tilde{V}_{n21}/c_{aa}$. Los paquetes informáticos de estadística contienen rutinas rápidas para la distribución normal acumulativa. El segundo factor es una integral. Como sabemos, las computadoras no pueden integrar, por lo que utilizamos la simulación para aproximar las integrales. Este es el corazón del proceso GHK: usar la simulación para aproximar la integral en P_{n1}.

Examinemos esta integral más de cerca. Es una integral sobre una normal truncada, es decir, sobre η_1 hasta $-\tilde{V}_{n21}/c_{aa}$. La simulación se realiza como sigue. Extraiga un valor al azar de η_1 de una densidad normal estándar truncada por encima de $-\tilde{V}_{n21}/c_{aa}$. Para este valor, calcule el factor $\Phi(-(\tilde{V}_{n31} + c_{ab}\eta_1)/c_{bb})$. Repita este proceso para muchas extracciones y promedie los resultados. Este promedio será una aproximación simulada de $\int_{\eta_1=-\infty}^{-\tilde{V}_{n21}/c_{aa}} \Phi\left(-(\tilde{V}_{n31} + c_{ab}\eta_1)/c_{bb}\right)\bar{\phi}(\eta_1)d\eta_1$. La probabilidad simulada se obtiene entonces multiplicando esta media por el valor de $\Phi(-\tilde{V}_{n21}/c_{aa})$, que se calcula exactamente. ¡Bastante simple!

No obstante, se nos plantea la siguiente cuestión: ¿cómo extraemos un valor al azar de una distribución normal truncada? Describiremos cómo extraer valores al azar de distribuciones univariadas truncadas en la Sección 9.2.4. Llegados a este punto, si el lector lo desea, puede consultar esta sección antes de continuar. Pero básicamente, el proceso consiste en extraer un valor al azar de una distribución uniforme estándar y etiquetarlo μ. Posteriormente se calcula $\eta = \Phi^{-1}(\mu\Phi(-\tilde{V}_{n21}/c_{aa}))$. El η resultante es una extracción de un valor al azar de una densidad normal truncada por encima de $-\tilde{V}_{n21}/c_{aa}$.

Ahora podemos poner todo esto junto para mostrar explícitamente los pasos concretos que se utilizan para el simulador GHK en nuestro caso de tres alternativas. La probabilidad de la alternativa 1 es

$$P_{n1} = \Phi\left(\frac{-\tilde{V}_{n21}}{c_{aa}}\right) \times \int_{\eta_1=-\infty}^{-\tilde{V}_{n21}/c_{aa}} \Phi\left(\frac{-\left(\tilde{V}_{n31} + c_{ab}\eta_1\right)}{c_{bb}}\right) \bar{\phi}(\eta_1) d\eta_1$$

Esta probabilidad se simula de la siguiente manera:

1. Calcule $k = \Phi(-\tilde{V}_{n21}/c_{aa})$.
2. Extraiga un valor de η_1, etiquetado como η_1^r, de una distribución normal estándar truncada en $-\tilde{V}_{n21}/c_{aa}$. Esto se logra de la siguiente manera:
    a. Extraiga un valor de una distribución uniforme estándar μ^r.

[i] Para ser precisos $\bar{\phi}(\eta_1) = \phi(\eta_1)/\Phi\left(-\tilde{V}_{n21}/c_{aa}\right)$ para $-\infty < \eta_1 < -V_{n21}/c_{aa}$, y =0 en otro caso.

b. Calcule $\eta_1^r = \Phi^{-1}(\mu^r \Phi(-\tilde{V}_{n21}/c_{aa}))$.

3. Calcule $g^r = \Phi(-(\tilde{V}_{n31} + c_{ab}\eta_1^r)/c_{bb})$.
4. La probabilidad simulada de este valor es $\breve{P}_{n1}^r = k \times g^r$.
5. Repita los pasos 1- 4 R veces y promedie los resultados. Este promedio es la probabilidad simulada: $\breve{P}_{n1} = (1/R)\sum P_{n1}^r$.

Una representación gráfica puede resultar útil. La figura 5.3 muestra la probabilidad de la alternativa 1 en el espacio de los errores independientes η_1 y η_2. El eje x es el valor de η_1 y el eje y es el valor de η_2. La línea etiquetada como A indica la zona en la que η_1 es igual a $-\tilde{V}_{n21}/c_{aa}$. La condición de que η_1 esté por debajo de $-\tilde{V}_{n21}/c_{aa}$ se cumple en la zona de rayas a la izquierda de la línea A. La línea etiquetada como B indica donde $\eta_2 = -(\tilde{V}_{n31} + c_{ab}\eta_1)/c_{bb}$. Tenga en cuenta que la intersección en el eje y se produce donde $\eta_1 = 0$, de modo que $\eta_2 = -\tilde{V}_{n31}/c_{bb}$ en este punto. La pendiente de la línea es $-c_{ab}/c_{bb}$. La condición de que $\eta_2 < -(\tilde{V}_{n31} + c_{ab}\eta_1)/c_{bb}$ se satisface debajo de la línea B. El área sombreada es donde η_1 está a la izquierda de la línea A y η_2 está por debajo de la línea B. Por tanto, la probabilidad de que se escoja la alternativa 1 es la masa de densidad de probabilidad en el área sombreada.

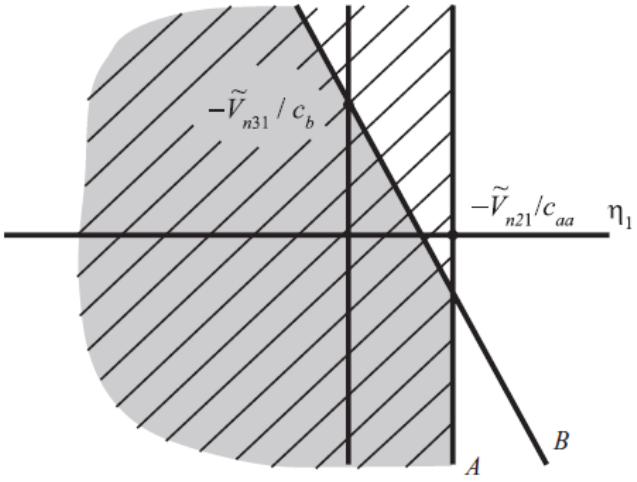


Figura 5.3. Probabilidad de la alternativa 1.

La probabilidad (es decir, la masa sombreada) es el producto de la masa de densidad de la zona rayada por la proporción de esta masa rayada que está por debajo de la línea B. El área rayada tiene masa $\Phi(-\tilde{V}_{n21}/c_{aa})$. Esto es fácil de calcular. Para cualquier valor dado de η_1, la porción de la masa rayada que está por debajo de la línea B también es fácil de calcular. Por ejemplo, en la figura 5.4, cuando η_1 toma el valor η_1^r, la probabilidad de que η_2 esté por debajo de la línea B es la proporción de la masa de la línea C que está por debajo de la línea B. Esta proporción es simplemente $\Phi(-(\tilde{V}_{n31} + c_{ab}\eta_1^r)/c_{bb})$. Por tanto, la proporción de la masa rayada que está por debajo de la línea B es el promedio de $\Phi(-(\tilde{V}_{n31} + c_{ab}\eta_1^r)/c_{bb})$ sobre todos los valores de η_1 que están a la izquierda de la línea A. Este promedio se simula mediante la extracción de valores de η_1 a la izquierda de la línea A, calculando $\Phi(-(\tilde{V}_{n31} + c_{ab}\eta_1^r)/c_{bb})$ para cada valor extraído y promediando los resultados. La probabilidad es el resultado de este promedio por la masa de la zona rayada, $\Phi(-\tilde{V}_{n21}/c_{aa})$.

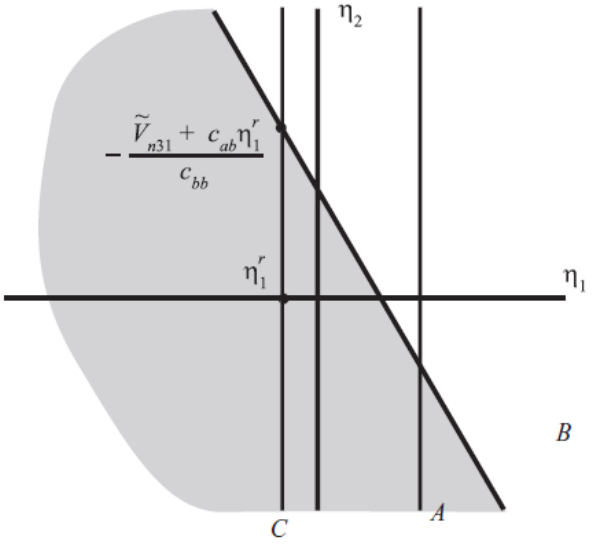


Figura 5.4. Probabilidad de que η_2 esté en el rango correcto, dado η_1^r.

Modelo general

Ahora podemos describir el simulador GHK en términos generales de forma rápida, ya que la lógica básica detrás del modelo ya ha sido expuesta. Esta expresión sucinta sirve para reforzar la idea de que el simulador GHK es realmente más simple de lo que puede parecer a primera vista

La utilidad se expresa como

$$U_{nj} = V_{nj} + \varepsilon_{nj}, \qquad j = 1, \dots, J,$$

$$\varepsilon_n' = \langle \varepsilon_{n1}, \dots, \varepsilon_{nJ} \rangle, \qquad \varepsilon_n : J \times 1,$$

$$\varepsilon_n \sim N(0, \Omega).$$

Transformamos a diferencias de utilidad respecto a la alternativa i:

$$\tilde{U}_{nji} = \tilde{V}_{nji} + \tilde{\varepsilon}_{nji}, \qquad j \neq i,$$

$$\varepsilon_{ni}' = \langle \tilde{\varepsilon}_{n1}, \dots, \tilde{\varepsilon}_{nJ} \rangle, \qquad donde \dots es\ sobre\ toda\ alternativa\ excepto\ i,$$

$$\tilde{\varepsilon}_{ni} : (J-1) \times 1,$$

$$\tilde{\varepsilon}_{ni} \sim N\left(0, \tilde{\Omega}_i\right),$$

donde $\tilde{\Omega}_i$ se obtiene de Ω.

Re-expresamos los errores como una transformación Choleski de normales estándar iid:

$$L_i \ s.t. \ L_i L_i' = \tilde{\Omega}_i,$$

$$L_i = \begin{pmatrix} c_{11} & 0 & \dots & \dots & \dots & 0 \\ c_{21} & c_{22} & 0 & \dots & \dots & 0 \\ c_{31} & c_{32} & c_{33} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}.$$

A continuación, agrupando utilidades $\widetilde{\mathrm{U}}'_{ni} = \left(\widetilde{U}_{n1i}, \dots, \widetilde{U}_{nJi}\right)$, obtenemos la forma vectorial del modelo,

$$\widetilde{U}_{ni} = \tilde{V}_{ni} + L_i \eta_n,$$

donde $\eta'_n = \langle \eta_{1n}, \dots, \eta_{J-1,n} \rangle$, es un vector de normales estándar iid: $\eta_{nj} \sim N(0,1)\ \forall j$. Escrito de forma explícita, el modelo es

$$\widetilde{U}_{n1i} = \tilde{V}_{n1i} + c_{11}\eta_1,$$

$$\widetilde{U}_{n2i} = \tilde{V}_{n2i} + c_{21}\eta_1 + c_{22}\eta_2,$$

$$\widetilde{U}_{n3i} = \tilde{V}_{n3i} + c_{31}\eta_1 + c_{32}\eta_2 + c_{33}\eta_3,$$

y así sucesivamente. Las probabilidades de elección son

$$\mathrm{P_{ni}} = Prob(\widetilde{\mathrm{U}}_{\mathrm{nji}} < 0\ \forall j \neq i)$$

$$= Prob\left(\eta_1 < \frac{-\tilde{V}_{n1i}}{c_{11}}\right)$$

$$\times Prob\left(\eta_2 < \frac{-\left(\tilde{V}_{n2i} + c_{21}\eta_1\right)}{c_{22}} \middle| \eta_1 < \frac{-\tilde{V}_{n1i}}{c_{11}}\right)$$

$$\times Prob\left(\eta_3 < \frac{-\left(\tilde{V}_{n3i} + c_{31}\eta_1 + c_{32}\eta_2\right)}{c_{33}} \middle| \eta_1 < \frac{-\tilde{V}_{n1i}}{c_{11}}\ y\ \eta_2 < \frac{-\left(\tilde{V}_{n2i} + c_{21}\eta_1\right)}{c_{22}}\right)$$

$$\times \cdots.$$

El simulador GHK se calcula como sigue:

1. Calculamos

$$Prob\left(\eta_1 < \frac{-\tilde{V}_{n1i}}{c_{11}}\right) = \Phi\left(\frac{-\tilde{V}_{n1i}}{c_{11}}\right).$$

2. Extraemos un valor al azar de η_1 , etiquetado como η_1^r, de una distribución normal estándar truncada en $-\widetilde{\mathrm{V}}_{\mathrm{n1i}}/\mathrm{c}_{11}$. Este valor se obtiene de la siguiente manera:
    a. Extraemos un valor al azar de una distribución uniforme estándar μ_1^r.
    b. Calculamos $\eta_1^{\mathrm{r}} = \Phi^{-1}(\mu_1^{\mathrm{r}}\Phi(-\widetilde{\mathrm{V}}_{\mathrm{n1i}}/\mathrm{c}_{11}))$.
3. Calculamos

$$Prob\left(\eta_2 < \frac{-\left(\tilde{V}_{n2i} + c_{21}\eta_1\right)}{c_{22}}\middle|\, \eta_1 = \eta_1^r\right) = \Phi\left(\frac{-\left(\tilde{V}_{n2i} + c_{21}\eta_1^r\right)}{c_{22}}\right)$$

4. Extraemos un valor al azar de η_2, etiquetado como η_2^r , de una distribución normal estándar truncada en $-\left(\tilde{V}_{n2i} + c_{21}\eta_1^r\right)/c_{22}$. Este valor se obtiene de la siguiente manera:
    a. Extraemos un valor al azar de una distribución uniforme estándar μ_2^r.
    b. Calculamos $\eta_2^r = \Phi^{-1}(\mu_2^r \Phi(-\left(\tilde{V}_{n2i} + c_{21}\eta_1^r\right)/c_{22}))$.
5. Calculamos

$$Prob\left(\eta_3 < \frac{-\left(\tilde{V}_{n3i} + c_{31}\eta_1 + c_{32}\eta_2\right)}{c_{33}}\middle|\, \eta_1 = \eta_1^r, \eta_2 = \eta_2^r\right)$$
$$= \Phi\left(\frac{-\left(\tilde{V}_{n3i} + c_{31}\eta_1^r + c_{32}\eta_2^r\right)}{c_{33}}\right)$$

6. Y así sucesivamente para todas las alternativas exceptuando i.
7. La probabilidad simulada para esta r-ésima extracción de valores de η_1, η_2,… se calcula como

$$\check{P}_{ni}^r = \Phi\left(\frac{-\tilde{V}_{n1i}}{c_{11}}\right)$$

$$\times \Phi\left(\frac{-\left(\tilde{V}_{n2i} + c_{21}\eta_1^r\right)}{c_{22}}\right)$$

$$\times \Phi\left(\frac{-\left(\tilde{V}_{n3i} + c_{31}\eta_1^r + c_{32}\eta_2^r\right)}{c_{33}}\right)$$

$$\times \cdots.$$

8. Repetimos los pasos 1-7 muchas veces, para $r = 1, \ldots, R$.
9. La probabilidad simulada es

$$\check{P}_{in} = \frac{1}{R}\sum_r \check{P}_{in}^r$$

Simulador GHK con estimación de máxima verosimilitud

Hay varias cuestiones que deben abordarse al utilizar el simulador GHK en una estimación de máxima verosimilitud. En primer lugar, en la función log-verosimilitud utilizamos la probabilidad de la alternativa elegida por el decisor. Dado que diferentes decisores eligen diferentes alternativas, P_{ni} debe calcularse para diferentes is. El simulador GHK usa diferencias de utilidad respecto a la alternativa para la que se calcula la probabilidad y, por lo tanto, es necesario considerar diferentes diferencias de utilidad para los decisores que eligieron distintas alternativas. En segundo lugar, para una persona que eligió la alternativa i, el simulador GHK utiliza la matriz de covarianza $\widetilde{\Omega}_i$, mientras que para una persona que

eligió la alternativa j, se utiliza la matriz $\widetilde{\Omega}_j$. Ambas matrices se obtienen de la misma matriz de covarianza Ω de los errores originales. Debemos asegurar que los parámetros en $\widetilde{\Omega}_i$ son consistentes con los de $\widetilde{\Omega}_j$, en el sentido de que ambas matrices se obtienen de una Ω común. En tercer lugar, tenemos que asegurar que los parámetros que se estiman por máxima verosimilitud implican el uso de matrices de covarianza $\widetilde{\Omega}_j$ $\forall j$ que son definidas positivas, como debe ser cualquier matriz de covarianza. En cuarto lugar, como siempre, debemos asegurarnos de que el modelo está normalizado para la escala y el nivel de utilidad, por lo que los parámetros son identificados.

Los investigadores utilizan diversos procedimientos para abordar estas cuestiones. Voy a describir el procedimiento que yo uso.

Para asegurar que el modelo es identificado, parto de la matriz de covarianza de las diferencias de utilidad escaladas, con las diferencias calculadas respecto a la primera alternativa. Esta es la matriz $\widetilde{\Omega}_1$, que es $(J-1)\times(J-1)$. Para asegurar que la matriz de covarianza es definida positiva, parametrizo el modelo en términos del factor Choleski de $\widetilde{\Omega}_1$. Es decir, empiezo con una matriz triangular inferior que es $(J-1)\times(J-1)$ y cuyo elemento superior izquierdo es 1:

$$L_1 = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ c_{21} & c_{22} & 0 & \dots & \dots & 0 \\ c_{31} & c_{32} & c_{33} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Los elementos c_{kl} de este factor Choleski son los parámetros que se estiman en el modelo. Cualquier matriz que sea resultado del producto de una matriz de rango completo triangular inferior por sí misma es definida positiva. De esta forma, usando los elementos de L_1 como parámetros, puedo estar seguro de que $\widetilde{\Omega}_1$ es definida positiva para cualquier valor estimado de estos parámetros.

La matriz Ω para los J errores no diferenciados se crea a partir de L_1. Yo creo un factor Cholesky $J\times J$ para Ω mediante la adición de una fila de ceros en la parte superior de L_1 y una columna de ceros a la izquierda. La matriz resultante es

$$L = \begin{pmatrix} 0 & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & c_{21} & c_{22} & 0 & \dots & \dots & 0 \\ 0 & c_{31} & c_{32} & c_{33} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

A continuación Ω se calcula como LL'. Teniendo esta Ω, puedo obtener $\widetilde{\Omega}_j$ para cualquier j. Observe que la matriz Ω construida de esta manera es totalmente general (es decir , permite cualquier patrón de sustitución), ya que utiliza todos los parámetros de la matriz $\widetilde{\Omega}_1$ normalizada.

La utilidad se expresa en forma de vector agrupado por alternativas: $U_n = V_n + \varepsilon_n$, $\varepsilon_n \sim N(0,\Omega)$. Considere una persona que ha elegido la alternativa i. Para la función log-verosimilitud, queremos calcular P_{ni}. Recordemos la matriz M_i que introdujimos en la sección 5.1. Las diferencias de utilidad se calculan usando esta matriz: $\widetilde{U}_{ni} = M_i U_n$, $\tilde{V}_{ni} = M_i V_n$ y $\tilde{\varepsilon}_{ni} = M_i \varepsilon_n$. La covarianza de las diferencias de error $\tilde{\varepsilon}_{ni}$ se calcula como $\widetilde{\Omega}_i = M_i \Omega M_i'$. Se toma el factor Choleski de $\widetilde{\Omega}_i$ y se etiqueta como $\mathrm{L_i}$. (Observe que la matriz L_1 obtenida aquí será necesariamente la misma L_1 que se utilizó al principio para parametrizar el modelo). La utilidad de la persona se expresa como: $\widetilde{U}_{ni} = \tilde{V}_{ni} + \mathrm{L_i}\eta_n$, donde η_n es un $(J-1)$-vector de normales estándar iid. El simulador GHK se aplica a esta expresión.

Este procedimiento satisface todos nuestros requerimientos. El modelo está necesariamente normalizado para la escala y el nivel, ya que lo parametrizamos en términos del factor Choleski L_1 de la covarianza de las *diferencias* de error *escaladas*, $\widetilde{\Omega}_1$. Cada $\widetilde{\Omega}_i$ es consistente con cada $\widetilde{\Omega}_j$ para $j \neq i$, porque ambos se obtienen de la misma Ω (que está construida a partir de L_1) . Cada $\widetilde{\Omega}_i$ es definida positiva para cualquier valor de los parámetros, ya que los parámetros son los elementos de L_1. Como se dijo anteriormente, cualquier matriz que sea el resultado del producto de una matriz triangular inferior multiplicada por sí misma es definida positiva, por lo que $\widetilde{\Omega}_1 = LL'$ es definida positiva. Y cada una de las otras matrices $\widetilde{\Omega}_j$s , para $j = 2, \ldots, J$, también es definida positiva, ya que se construyen para ser consistentes con Ω_1, que es definida positiva.

GHK como muestreo por importancia

Como describí en el caso de tres alternativas, el simulador GHK proporciona una aproximación simulada de la integral

$$\int_{\eta_1=-\infty}^{-\widetilde{\mathrm{V}}_{\mathrm{n21}}/\mathrm{c}_{\mathrm{aa}}} \Phi\left(\frac{-\left(\widetilde{\mathrm{V}}_{\mathrm{n31}} + \mathrm{c}_{\mathrm{ab}}\eta_1\right)}{\mathrm{c}_{\mathrm{bb}}}\right)\phi(\eta_1)\mathrm{d}\eta_1$$

El simulador GHK puede interpretarse de una forma alternativa que a menudo es útil. El muestreo por importancia (*importance sampling*) es una manera de transformar una integral para que sea más conveniente para la simulación. El procedimiento se describe en la Sección 9.2.7, por lo que el lector puede encontrar útil avanzar hasta ese apartado para leer la descripción. El muestreo por importancia puede resumirse de la siguiente manera. Considere cualquier integral $\bar{t} = \int t(\varepsilon)g(\varepsilon)d\varepsilon$ sobre una densidad g. Supongamos que existe otra densidad de la que es fácil extraer valores al azar. Etiquetamos esta otra densidad $f(\varepsilon)$. La densidad g se denomina densidad objetivo y f densidad generadora. La integral puede rescribirse como $\bar{t} = \int [t(\varepsilon)g(\varepsilon)/f(\varepsilon)]f(\varepsilon)d\varepsilon$. Esta integral se simula extrayendo valores de f, calculando $t(\varepsilon)g(\varepsilon)/f(\varepsilon)$ para cada valor y promediando los resultados. Este procedimiento se denomina muestreo por importancia porque cada extracción de f se pondera por g/f cuando se calcula el promedio de t; el peso g/f es la "importancia" del valor extraído de f. Este procedimiento es ventajoso si (1) resulta más fácil extraer valores al azar de f que de g , y/o (2) el simulador basado en $t(\varepsilon)g(\varepsilon)/f(\varepsilon)$ tiene mejores propiedades (por ejemplo , suavidad) que el simulador basado en $t(\varepsilon)$.

El simulador GHK puede considerarse que hace este tipo de transformación y, por lo tanto, puede ser visto como un tipo de muestreo por importancia. Sea ser un vector η de $\mathrm{J}-1$ normales estándar iid. La probabilidad de elección se puede expresar como

$$P_{ni} = \int I(\eta \in B)g(\eta)d\eta, \qquad (5.7)$$

donde $\mathrm{B} = \left\{\eta \text{ s.t. } \widetilde{\mathrm{U}}_{\mathrm{nji}} < 0 \ \forall j \neq i\right\}$ es el conjunto de ηs que producen que la alternativa i sea la elegida; $\mathrm{g}(\eta) = \phi(\eta_1)\cdots\phi(\eta_{J-1})$ es la densidad , donde ϕ es la densidad normal estándar; y las utilidades son

$$\widetilde{\mathrm{U}}_{\mathrm{n1i}} = \widetilde{\mathrm{V}}_{\mathrm{n1i}} + \mathrm{c}_{11}\eta_1,$$
$$\widetilde{\mathrm{U}}_{\mathrm{n2i}} = \widetilde{\mathrm{V}}_{\mathrm{n2i}} + \mathrm{c}_{21}\eta_1 + \mathrm{c}_{22}\eta_2,$$
$$\widetilde{\mathrm{U}}_{\mathrm{n3i}} = \widetilde{\mathrm{V}}_{\mathrm{n3i}} + \mathrm{c}_{31}\eta_1 + \mathrm{c}_{32}\eta_2 + \mathrm{c}_{33}\eta_3,$$

y así sucesivamente.

La forma más directa para simular esta probabilidad es extraer valores de η, calcular $I(\eta \in B)$ para cada valor, y promediar los resultados. Este es el simulador AR. Este simulador tiene las desafortunadas propiedades de que puede ser cero y no es suave.

Para el simulador GHK extraemos η de una densidad diferente, no de g(η). Recordemos que para el simulador GHK extraemos el valor η_1 de una densidad normal estándar truncada en $-\tilde{V}_{n1i}/c_{11}$. La densidad de esta normal truncada es $\phi(\eta_1)/\Phi(-\tilde{V}_{n1i}/c_{11})$, es decir, la densidad normal estándar normalizada por la probabilidad total bajo el punto de truncamiento. Se obtienen extracciones de η_2, η_3 y así sucesivamente, de densidades truncadas pero con diferentes puntos de truncamiento. Cada una de estas densidades truncadas toma la forma $\phi(\eta_j)/\Phi(\cdot)$ para algún punto de truncamiento en el denominador. Por tanto, la densidad de la que extraemos valores para el simulador GHK es

$$f(\eta) = \begin{cases} \frac{\phi(\eta_1)}{\Phi(-\tilde{V}_{n1i}/c_{11})} \times \frac{\phi(\eta_2)}{\Phi(-(\tilde{V}_{n2i}+c_{21}\eta_1)/c_{22})} \times \cdots & para\ \eta \in B \\ 0 & para\ \eta \notin B \end{cases} \tag{5.8}$$

Tenga en cuenta que sólo extraemos valores que sean consistentes con la persona que elige la alternativa i (dado que extraemos valores de las distribuciones correctamente truncadas). Por lo tanto, $f(\eta) = 0$ para $\eta \notin B$.

Recuerde que para una extracción de η dentro del simulador GHK, calculamos:

$$\check{P}_{in}(\eta) = \Phi\left(\frac{-\tilde{V}_{n1i}}{c_{11}}\right)$$

$$\times \Phi\left(\frac{-\left(\tilde{V}_{n2i} + c_{21}\eta_1\right)}{c_{22}}\right)$$

$$\times \cdots. \tag{5.9}$$

Observe que esta expresión es el denominador de f(η) para $\eta \in B$, dada en la ecuación (5.8). Usando este hecho, podemos reescribir la densidad f(η) como

$$f(\eta) = \begin{cases} g(\eta)/\check{P}_{ni}(\eta) & para\ \eta \in B \\ 0 & para\ \eta \notin B \end{cases}$$

Con esta expresión para f(η), podemos probar que para el simulador GHK, $\check{P}_{in}(\eta)$ es un estimador no sesgado de $P_{ni}(\eta)$.

$$E(\check{P}_{in}(\eta)) = \int \check{P}_{in}(\eta) f(\eta) d\eta$$

$$= \int_{\eta \in B} \check{P}_{in}(\eta) \frac{g(\eta)}{\check{P}_{in}(\eta)} d\eta \quad por\ (5.6.3)$$

$$= \int_{\eta \in B} g(\eta) d\eta$$

$$= \int I(\eta \in B) g(\eta) d\eta$$

$$= P_{in}.$$

La interpretación del simulador GHK como un muestreo por importancia también se obtiene a partir de esta expresión:

$$P_{in} = \int I(\eta \in B) g(\eta) d\eta$$

$$= \int I(\eta \in B) g(\eta) \frac{f(\eta)}{f(\eta)} d\eta$$

$$= \int I(\eta \in B) \frac{g(\eta)}{g(\eta)/\check{P}_{in}(\eta)} f(\eta) d\eta \quad por\ (5.6.3)$$

$$= \int I(\eta \in B) \check{P}_{in}(\eta) f(\eta) d\eta$$

$$= \int \check{P}_{in}(\eta) f(\eta) d\eta$$

donde la última igualdad se debe a que $f(\eta) > 0$ sólo cuando $\eta \in B$. El procedimiento GHK extrae valores de $f(\eta)$, calcula $\check{P}_{in}(\eta)$ para cada valor extraído y promedia los resultados. Básicamente, GHK reemplaza la función indicadora $0 - 1\ I(\eta \in B)$ por una $\check{P}_{in}(\eta)$ suave, y hace el cambio correspondiente en la densidad de $g(\eta)$ a $f(\eta)$.

6

Logit mixto

6.1 Probabilidades de elección

Logit mixto es un modelo muy flexible que puede aproximar cualquier modelo de utilidad aleatoria (McFadden y Train, 2000). Este modelo elude las tres limitaciones del modelo logit estándar, permitiendo variación aleatoria de preferencias, patrones de sustitución no restringidos y correlación entre factores no observados a lo largo del tiempo. A diferencia de probit, no está limitado a distribuciones normales. Su formulación es sencilla y la simulación de sus probabilidades de elección es computacionalmente simple.

Al igual que probit, el modelo logit mixto se conoce desde hace muchos años, pero sólo se ha convertido en un modelo plenamente aplicable con la llegada de la simulación. Según parece, las primeras aplicaciones reales del logit mixto fueron los modelos de demanda de automóviles creados conjuntamente por Boyd y Mellman (1980) y Cardell y Dunbar (1980). En estos estudios, las variables explicativas no variaban entre decisores y la variable dependiente observada era la cuota de mercado y no las elecciones individuales de los clientes. Como resultado, la integración computacionalmente intensiva inherente al modelo logit mixto (como se explica más adelante) sólo fue necesario llevarla a cabo una única vez para el mercado como un todo, en lugar de realizarla para cada decisor de la muestra. Las primeras aplicaciones sobre datos a nivel de consumidor individual, como Train et al. (1987a) y Ben-Akiva et al. (1993), incluían sólo una o dos dimensiones de integración, las cuales podían calcularse por cuadratura numérica. Las mejoras en la velocidad de las computadoras y en el conocimiento de los métodos de simulación han permitido poder utilizar toda la potencia del modelo logit mixto. Entre los estudios que evidencian esta potencia están los de Bhat (1998a) y Brownstone y Train (1999) sobre datos transversales (elecciones en un único período de tiempo), y Erdem (1996), Revelt & Train (1998) y Bhat (2000), sobre datos de panel. La descripción que se ofrece en el presente capítulo se basa en gran medida en Train (1999).

Los modelos logit mixtos pueden formularse bajo diversas especificaciones de comportamiento y cada formulación proporciona una interpretación particular. El modelo logit mixto se *define* sobre la base de la forma funcional de sus probabilidades de elección. Cualquier especificación de comportamiento cuya formulación de las probabilidades de elección adopte esta forma particular se denomina un modelo logit mixto.

Las probabilidades del logit mixto son las integrales de las probabilidades logit estándar sobre una densidad de probabilidad de los parámetros. Dicho de manera más explícita, un modelo logit mixto es cualquier modelo cuyas probabilidades de elección se puedan expresar en la forma

$$P_{ni} = \int L_{ni}(\beta) f(\beta) d\beta,$$

donde $L_{ni}(\beta)$ es la probabilidad logit evaluada en los parámetros ß:

$$L_{ni}(\beta) = \frac{e^{V_{ni}(\beta)}}{\sum_{j=1}^{J} e^{V_{nj}(\beta)}}$$

y $f(\beta)$ es una función de densidad de probabilidad. $V_{ni}(\beta)$ es la parte observada de la utilidad, que depende de los parámetros β. Si la utilidad es lineal en β, entonces $V_{ni}(\beta) = \beta x_{ni}$. En este caso, la probabilidad del modelo logit mixto toma su forma habitual:

(6.1)
$$P_{ni} = \int \left(\frac{e^{\beta' x_{ni}}}{\sum_{j=1}^{J} e^{\beta' x_{nj}}} \right) f(\beta) d\beta,$$

La probabilidad del modelo logit mixto es un promedio ponderado de la fórmula logit evaluada en diferentes valores de β, con los pesos dados por la densidad $f(\beta)$. En la literatura estadística, la media ponderada de varias funciones se llama una función mixta (*mixed function* o *mixture function*) y la densidad que proporciona los pesos se llama la distribución de mezcla o mixtura (*mixing distribution*). El modelo logit mixto es una mezcla de la función logit evaluada en diferentes βs con $f(\beta)$ como distribución de mezcla.

Logit estándar es un caso especial de logit mixto en el que la distribución de mezcla $f(\beta)$ degenera en unos parámetros fijos b: $f(\beta) = 1$ para $\beta = b$ y 0 para $\beta \neq b$. La probabilidad elección (6.1) se convierte en la fórmula logit simple

$$P_{ni} = \frac{e^{\beta' x_{ni}}}{\sum_{j=1}^{J} e^{\beta' x_{nj}}}$$

La distribución de mezcla $f(\beta)$ puede ser discreta, con β tomando un conjunto finito de posibles valores. Supongamos que β toma M valores posibles etiquetados como $b_1, \ldots, b_M$ con una probabilidad s_m de que $\beta = b_m$. En este caso, el modelo logit mixto se convierte en el *modelo de clases latentes* (*latent class model*), que ha sido popular durante mucho tiempo en psicología y marketing; algunos ejemplos los proporcionan Kamakura y Russell (1989) y Chintagunta et al. (1991). La probabilidad de elección en este caso resulta

$$P_{ni} = \sum_{m=1}^{M} s_m \left(\frac{e^{b_m' x_{ni}}}{\sum_{j=1}^{J} e^{b_m' x_{nj}}} \right).$$

Esta especificación resulta útil si hay M segmentos de población, cada uno de los cuales tiene su propio comportamiento de elección o preferencias. La proporción de la población que pertenece al segmento m es s_m, proporción que el investigador puede estimar en el modelo junto con las bs para cada segmento.

En la mayoría de aplicaciones que realmente han sido denominadas logits mixtos (tales como las citadas en los párrafos introductorios de este capítulo), se especifica que $f(\beta)$ sea continua. Por ejemplo, la densidad de β se puede especificar que sea una distribución normal con media b y covarianza W. La probabilidad de elección bajo esta densidad se convierte en

$$P_{ni} = \int \left(\frac{e^{\beta' x_{ni}}}{\sum_{j=1}^{J} e^{\beta' x_{nj}}} \right) \phi(\beta|b, W) d\beta,$$

donde $\phi(\beta|b, W)$ es la densidad normal con media b y covarianza W.El investigador estima b y W. Distribuciones como la log-normal, la uniforme, la triangular, la gamma o cualquier otra pueden ser utilizadas del mismo modo. Como se verá en la sección 6.5, especificando apropiadamente las variables explicativas y la densidad, el investigador puede representar a través de un modelo logit mixto cualquier comportamiento orientado a maximizar la utilidad, así como muchas formas de comportamiento no relacionadas con la maximización de la utilidad.

McFadden y Train (2000) y Chesher & Santos - Silva (2002) han desarrollado varios test que muestran la necesidad de usar una distribución de mezcla no degenerada en parámetros fijos, así como la adecuación de cualquier distribución específica dada. Asimismo, varios estudios han comparado las distribuciones de mezcla discretas y continuas en el contexto de modelos logit mixtos; véase, por ejemplo, Wedel y Kamakura (2000) y Andrews et al. (2002).

Una cuestión terminológica surge en relación a los modelos logit mixtos. Hay dos conjuntos de parámetros en un modelo logit mixto. En primer lugar, tenemos los parámetros β, que entran en la fórmula logit. Estos parámetros tienen densidad $f(\beta)$. El segundo conjunto son parámetros que describen esta densidad. Por ejemplo, si β se distribuye normalmente con media b y covarianza W, entonces b y W son parámetros que describen la densidad $f(\beta)$. Por lo general (aunque no siempre, como señalaremos a continuación) el investigador está interesado en la estimación de los parámetros de f.

Denotemos los parámetros que describen la densidad de β como θ. La forma más adecuada para referirse a esta densidad es $f(\beta|\theta)$. Las probabilidades de elección del modelo logit mixto no dependen de los valores de β. Estas probabilidades son $P_{ni} = \int L_{ni}(\beta) f(\beta|\theta) d\beta$, que son funciones de θ. Los parámetros β son las variables de integración y desaparecen del resultado. Por lo tanto, las βs son similares a la ε_{nj}s, en el sentido en que ambos son términos aleatorios que se integran para obtener la probabilidad de elección.

Bajo ciertas formulaciones del modelo logit mixto, los valores de β tienen un significado interpretable como una representación de las preferencias individuales de los decisores. En estos casos, el investigador desearía obtener información acerca de las βs para cada decisor de la muestra, así como los parámetros θ que describen la distribución de las βs entre decisores. En el capítulo 11, se describe la forma en que el investigador puede obtener esta información a partir de estimaciones de θ y de las elecciones observadas de cada decisor. En el presente capítulo se describe la estimación e interpretación de θ, usando procedimientos clásicos de estimación. En el capítulo 12 se describen los procedimientos bayesianos que proporcionan información sobre θ y sobre la β de cada decisor simultáneamente.

6.2 Coeficientes Aleatorios

La probabilidad del modelo logit mixto puede obtenerse bajo la hipótesis de un comportamiento orientado a la maximización de la utilidad, de varias formas que son formalmente equivalentes pero que ofrecen diferentes interpretaciones. La formulación más directa y más ampliamente utilizada en

estudios recientes, se basa en coeficientes aleatorios. El decisor se enfrenta a una elección entre J alternativas. La utilidad que obtiene la persona n de la alternativa j se especifica como

$$U_{nj} = \beta_n' x_{nj} + \varepsilon_{nj},$$

donde x_{nj} son variables observadas que se relacionan con la alternativa y el decisor, β_n es un vector de coeficientes de estas variables para la persona n que representa las preferencias de esa persona y ε_{nj} es un término aleatorio de tipo valor extremo iid. Los coeficientes varían entre decisores de la población con densidad $f(\beta)$. Esta densidad es una función de los parámetros θ que representan, por ejemplo, la media y la covarianza de las βs en la población. Esta especificación es igual a la del logit estándar, excepto que las βs varían entre decisores en lugar de ser fijas.

El decisor conoce el valor de su propia β_n y de las ε_{nj}s para toda alternativa j, y elige la alternativa i si y sólo si $U_{ni} > U_{nj}\ \forall j \neq i$. El investigador observa las x_{nj} pero no las β_ns o los ε_{nj}s. Si el investigador observase las β_ns, entonces la probabilidad de elección sería logit estándar, ya que los ε_{nj}s son valor extremo iid. Es decir, la probabilidad *condicionada* sobre β_n es

$$L_{ni}(\beta_n) = \frac{e^{\beta_n' x_{ni}}}{\sum_{j=1}^{J} e^{\beta_n' x_{nj}}}$$

Sin embargo, el investigador no conoce las β_ns y por lo tanto no puede condicionar sobre β. Por eso, la probabilidad de elección no condicionada es la integral de $L_{ni}(\beta_n)$ sobre todos los posibles valores de β_n:

$$P_{ni} = \int \left(\frac{e^{\beta' x_{ni}}}{\sum_j e^{\beta' x_{nj}}} \right) f(\beta) d\beta,$$

que es la probabilidad del modelo logit mixto (6.1).

El investigador especifica una distribución para los coeficientes y estima los parámetros de esa distribución. En la mayoría de los usos de este modelo, como Revelt & Train (1998), Mehndiratta (1996) y Ben-Akiva y Bolduc (1996), $f(\beta)$ se ha especificado como normal o log-normal : $\beta \sim N(b, W)$ o $\ln \beta \sim N(b, W)$ con parámetros b y W estimados. La distribución logarítmica normal es útil cuando se sabe que el coeficiente tiene el mismo signo para todos los decisores, como por ejemplo un coeficiente de precio que se sabe que es negativo para todo el mundo. Revelt & Train (2000), Hensher & Greene (2003), y Train (2001) han utilizado distribuciones triangulares y uniformes. Con la densidad uniforme, β se distribuye uniformemente entre $b - s$ y $b + s$, donde la media b y la extensión s deben ser estimadas. La distribución triangular tiene una densidad positiva que comienza en $b - s$, aumenta linealmente hasta b y luego desciende linealmente hasta $b + s$, tomando la forma de una tienda de campaña o triángulo. La media b y la extensión s se estiman, como en el caso de la distribución uniforme, pero la densidad es puntiaguda en lugar de plana. Estas densidades tienen la ventaja de estar limitadas por ambos lados, lo que evita el problema que puede surgir con las distribuciones normales y log-normales, las cuales pueden generar coeficientes anormalmente grandes para algún grupo de decisores. Al limitar s = b, el investigador puede asegurar que los coeficientes tienen el mismo signo para todos los decisores. Siikamaki (2001) y Siikamaki y Layton (2001) utilizan la distribución de Rayleigh (Johnson et al., 1994), que se encuentra en un solo lado del cero como la distribución log-normal pero, tal y como hallaron estos investigadores, puede resultar más simple a efectos de estimación que la log-normal. Revelt (1999) utilizó distribuciones normales truncadas. Como indican estos ejemplos, el

investigador es libre de especificar una distribución que satisfaga sus expectativas sobre el comportamiento de su propia situación de elección.

Las variaciones en las preferencias que están relacionadas con los atributos observados de los decisores se capturan a través de la especificación de variables explicativas y/o la distribución de mezcla. Por ejemplo, el costo puede dividirse por los ingresos del decisor para permitir que el valor o la importancia relativa de los costos disminuya a medida que aumenta el ingreso. El coeficiente de esta variable aleatoria pasa a representar la variación del valor que las personas con el mismo nivel de ingresos otorgan al precio. La valoración media del costo se reduce con el aumento del ingreso mientras que la varianza entorno a la media es fija. Los atributos observados del decisor también pueden entrar en $f(\beta)$, de manera que los momentos de orden superior de la variación de las preferencias también puedan depender de los atributos del decisor. Por ejemplo, Bhat (1998a, 2000) especifica una $f(\beta)$ log-normal con media y varianza en función de las características del decisor.

6.3 Componentes de error

Un modelo logit mixto se puede utilizar sin una interpretación de coeficientes aleatorios, simplemente como una representación de los componentes de error que crea correlaciones entre las utilidades de diferentes alternativas. La utilidad se especifica como

$$U_{nj} = \alpha' x_{nj} + \mu_n' z_{nj} + \varepsilon_{nj},$$

donde x_{nj} y z_{nj} son vectores que contienen variables observadas en relación a la alternativa j, α es un vector de coeficientes fijos, μ es un vector de términos aleatorios con media cero y ε_{nj} es de tipo valor extremo iid. Los términos en z_{nj} son componentes de error que, junto con ε_{nj}, definen la parte estocástica de la utilidad. Es decir, la parte no observada (aleatoria) de utilidad es $\eta_{nj} = \mu_n' z_{nj} + \varepsilon_{nj}$, que puede estar correlacionada entre alternativas en función de la especificación de z_{nj}. Para el modelo logit estándar, z_{nj} es idénticamente igual a cero, de modo que no hay correlación en la utilidad entre alternativas. Esta falta de correlación da lugar a la propiedad IIA y sus patrones de sustitución restrictivos. Con componentes de error distintos de cero, la utilidad está correlacionada entre alternativas: $\mathrm{Cov}(\eta_{ni}, \eta_{nj}) = \mathrm{E}(\mu_n' z_{ni} + \varepsilon_{ni})(\mu_n' z_{nj} + \varepsilon_{nj}) = z_{ni} W z_{nj}$, donde W es la covarianza de μ_n. La utilidad está correlacionada entre alternativas incluso cuando los componentes de error son independientes (como sucede en la mayoría de las especificaciones), de forma que W es diagonal.

Varios patrones de correlación, y por lo tanto varios patrones de sustitución, se pueden obtener mediante la elección apropiada de las variables que entran como componentes de error. Por ejemplo, un modelo análogo al logit jerárquico se obtiene mediante la especificación de una variable indicadora (*dummy*) para cada nido que sea igual a 1 para cada alternativa en el nido y cero para las alternativas fuera del nido. Con K nidos no solapados, los componentes de error son $\mu_n' z_{nj} = \sum_{k=1}^{K} \mu_{nk} d_{jk}$, donde $d_{jk} = 1$ si j está en el nido k y cero en caso contrario. Es conveniente en esta situación especificar que los componentes de error se distribuyan como normales independientes: $\mu_{nk}\ iid\ N(0, \sigma_k)$. La cantidad aleatoria μ_{nk} entra en la utilidad de cada alternativa del nido k, induciendo correlación entre estas alternativas. No entra en ninguna de las alternativas de otros nidos, con lo cual no induce correlación entre alternativas del nido con alternativas fuera del nido. La varianza σ_k capta la magnitud de la correlación. Desempeña un papel análogo al coeficiente de valor inclusivo de los modelos logit jerárquicos.

Para ser más precisos, la covarianza entre dos alternativas en el nido k es $\mathrm{Cov}(\eta_{ni}, \eta_{nj}) = \mathrm{E}(\mu_k + \varepsilon_{ni})(\mu_k + \varepsilon_{nj}) = \sigma_k$. La varianza para cada una de las alternativas en el nido k es $\mathrm{Var}(\eta_{ni}) = \mathrm{E}(\mu_k + \varepsilon_{ni})^2 = \sigma_k + \pi^2/6$, ya que la varianza del término de valor extremo, ε_{ni}, es $\pi^2/6$ (véase la

Sección 3.1). La correlación entre dos alternativas cualesquiera dentro del nido de k, es por lo tanto $\sigma_k/(\sigma_k + \pi^2/6)$. Restringir la varianza de cada componente de error de cada nido para que sea igual en todos los nidos (es decir, forzar que $\sigma_k = \sigma, k = 1, \dots, K$) es análogo a restringir el coeficiente de log-suma para que sea igual para todos los nidos en un modelo logit jerárquico. Esta restricción también asegura que el modelo logit mixto esté normalizado para la escala y el nivel.

Permitir diferentes varianzas de las cantidades aleatorias de nidos diferentes es análogo a permitir que el coeficiente de valor inclusivo difiera entre los nidos en un logit jerárquico. Un efecto análogo al de los nidos solapados se captura con variables indicadoras que identifiquen conjuntos solapados de alternativas, como en Bhat (1998a). Un modelo análogo al logit heterocedástico (visto en la sección 4.5) se obtiene mediante la introducción de un componente de error para cada alternativa. Walker et al. (2007) proporcionan orientación sobre cómo especificar estas variables de manera apropiada.

Las especificaciones basadas en componentes de error y en coeficientes aleatorios son formalmente equivalentes. Basándonos en coeficientes aleatorios, la utilidad se especifica como $\mathrm{U_{nj}} = \beta_n' \mathrm{x_{nj}} + \varepsilon_{nj}$ con β_n aleatorio. Los coeficientes β_n se pueden descomponer entre su media α y las desviaciones μ_n, de manera que $\mathrm{U_{nj}} = \alpha_n' \mathrm{x_{nj}} + \mu_n' \mathrm{x_{nj}} + \varepsilon_{nj}$, que tiene componentes de error definidos por $\mathrm{z_{nj}} = \mathrm{x_{nj}}$. En el sentido contrario, basándonos en componentes de error, la utilidad es $\mathrm{U_{nj}} = \alpha' \mathrm{x_{nj}} + \mu_n' \mathrm{z_{nj}} + \varepsilon_{nj}$, lo que equivale a un modelo de parámetros aleatorios con coeficientes fijos para las variables $\mathrm{x_{nj}}$ y coeficientes aleatorios con media cero para las variables $\mathrm{z_{nj}}$. Si $\mathrm{x_{nj}}$ y $\mathrm{z_{nj}}$ se solapan (en el sentido de que algunas variables entran tanto en $\mathrm{x_{nj}}$ como en $\mathrm{z_{nj}}$), los coeficientes de estas variables pueden considerarse que varían de forma aleatoria con media α y la misma distribución de μ_n alrededor de sus medias.

Aunque coeficientes aleatorios y componentes de error son formalmente equivalentes, la forma en que un investigador piensa en el modelo afecta a la especificación del logit mixto. Por ejemplo, cuando se piensa en términos de parámetros aleatorios, es natural permitir que el coeficiente de cada variable pueda variar e incluso permitir correlaciones entre coeficientes. Este es el enfoque adoptado por Revelt & Train (1998). Sin embargo, cuando el objetivo principal es representar patrones de sustitución apropiadamente a través de la utilización de componentes de error, el énfasis se pone en especificar variables que puedan inducir correlación entre alternativas de una manera lo más simple posible como para proporcionar patrones de sustitución suficientemente realistas. Este es el enfoque adoptado por Brownstone y Train (1999). Los objetivos eran diferentes en estos estudios: Revelt y Train estaban interesados en el patrón de preferencias, mientras que Brownstone y Train estaban más preocupados por la predicción. El número de variables explicativas también difirió en ambos casos, con Revelt y Train examinando 6 variables, lo que permitía que la estimación de la distribución conjunta de sus coeficientes fuese una meta razonable, mientras que Brownstone y Train incluían 26 variables en su análisis. Aspirar a estimar la distribución de 26 coeficientes no es razonable, y sin embargo, pensar en términos de parámetros aleatorios en lugar de componentes de error puede llevar al investigador a tales expectativas. Es importante recordar que la distribución de mezcla, ya sea motivada por parámetros aleatorios o por componentes de error, captura la varianza y las correlaciones de los factores no observados. Hay un límite natural en cuánto puede aprenderse sobre las cosas que no son observadas.

6.4 Patrones de sustitución

Logit mixto no presenta independencia de alternativas irrelevantes (IIA) o los patrones de sustitución restrictivos de logit. El ratio de las probabilidades de elección en los modelos logit mixtos, $\mathrm{P_{ni}/P_{nj}}$, depende de todos los datos, incluidos los atributos de las alternativas que no son i o j. Los denominadores de la fórmula logit se encuentran dentro de las integrales, por lo que no se anulan. El

porcentaje de cambio en la probabilidad de una de las alternativas dado un cambio porcentual en el atributo m-ésimo de otra alternativa, es

$$E_{nix_{nj}^m} = -\frac{x_{nj}^m}{P_{ni}} \int \beta^m L_{ni}(\beta) L_{nj}(\beta) f(\beta) d\beta,$$

$$= -x_{nj}^m \int \beta^m L_{nj}(\beta) \left[\frac{L_{ni}(\beta)}{P_{ni}}\right] f(\beta) d\beta,$$

donde β^m es el elemento m-ésimo de β. Esta elasticidad es diferente para cada alternativa i. Una reducción del diez por ciento en una alternativa no implica necesariamente (como en logit) una reducción del diez por ciento en cada una de las otras alternativas. En este caso, el patrón de sustitución depende de la especificación de las variables y de la distribución de mezcla, y ambas puede determinarse empíricamente.

Observe que el porcentaje de cambio en la probabilidad depende de la correlación entre $L_{ni}(\beta)$ y $L_{nj}(\beta)$ a través de diferentes valores de β, la cual está determinada por la especificación que el investigador hace de las variables y la distribución de mezcla. Por ejemplo, para representar una situación en la que una mejora en la alternativa j reduce proporcionalmente más la alternativa i que la alternativa k, el investigador puede especificar un elemento de x que correlacione positivamente entre i y j, pero que no correlacione o correlacione negativamente entre k y j, con una distribución de mezcla que permita variar al coeficiente de esta variable.

6.5 Aproximación de cualquier modelo de utilidad aleatoria

McFadden y Train (2000) muestran que cualquier modelo de utilidad aleatoria (*random utility model*, RUM) puede ser aproximado con cualquier grado de precisión por un modelo logit mixto, con la elección apropiada de las variables y la distribución de mezcla. Esta demostración es análoga a las aproximaciones consistentes con RUM proporcionadas por Dagsvik (1994). Es posible proporcionar una explicación intuitiva fácilmente. Supongamos que el verdadero modelo es $U_{nj} = \alpha_n' z_{nj}$, donde z_{nj} son variables relacionadas con la alternativa j y α sigue una distribución $f(\alpha)$ cualquiera. Cualquier RUM puede expresarse de esta forma. (La notación más tradicional $U_{nj} = \beta_n' x_{nj} + \varepsilon_{nj}$ se obtiene definiendo $z_{nj}' = \langle x_{nj}', d_j \rangle$, $\alpha' = \langle \beta_n', \varepsilon_{nj} \rangle$ y $f(\alpha)$ como la densidad conjunta de β_n y ε_{nj} $\forall j$). Condicionada a α, la elección de la persona está totalmente determinada, dado que U_{nj} se conocería entonces para cada j. Por tanto, la probabilidad condicionada es

$$q_{ni}(\alpha) = I\left(\alpha_n' z_{ni} > \alpha_n' z_{nj} \; \forall j \neq i\right),$$

donde $I(\cdot)$ es la función indicadora 1-0 de si se produce el evento entre paréntesis. Esta probabilidad condicionada es determinista en el sentido de que la probabilidad o es cero o es uno: condicionada a todos los términos aleatorios desconocidos, la elección del decisor está completamente determinada. La probabilidad de elección no condicionada es la integral de $q_{ni}(\alpha)$ sobre α:

$$Q_{ni} = \int I\left(\alpha_n' z_{ni} > \alpha_n' z_{nj} \; \forall j \neq i\right) f(\alpha) d\alpha,$$

Podemos aproximar esta probabilidad con un modelo logit mixto. Escalemos la utilidad por un factor λ, de modo que $U_{nj}^* = (\alpha/\lambda)' z_{nj}$. Este escalado no cambia el modelo, ya que el comportamiento no se ve afectado por la escala de la utilidad. A continuación, agregamos un término tipo valor extremo iid: ε_{nj}.

La adición del término valor extremo no cambia el modelo, ya que cambia la utilidad de cada alternativa. Lo agregamos porque al hacerlo resulta un logit mixto. Y, como veremos (este es el propósito de la demostración), añadir el término valor extremo es inocuo. La probabilidad del modelo logit mixto basado en esta utilidad es

$$P_{ni} = \int \left(\frac{e^{(\alpha/\lambda)' z_{ni}}}{\sum_j e^{(\alpha/\lambda)' z_{nj}}} \right) f(\alpha) d\alpha,$$

A medida que λ se aproxima a cero, los coeficientes α/λ en la fórmula logit se hacen mayores y P_{ni} se parece cada vez más a un indicador 1-0 de la alternativa con mayor utilidad. Es decir, la probabilidad del modelo logit mixto P_{ni} se aproxima a la verdadera probabilidad Q_{ni} a medida que λ se aproxima a cero. Escalando los coeficientes al alza suficientemente, el logit mixto basado en estos coeficientes escalados se aproxima arbitrariamente al modelo verdadero. Srinivasan y Mahmassani (2005) utilizan este concepto de aumentar la escala de los coeficientes para demostrar que un modelo logit mixto puede aproximar un modelo probit; el concepto aplica en general a la aproximación de cualquier RUM.

Recuerde que hemos añadido un término valor extremo iid a la verdadera utilidad de cada alternativa. Estos términos cambian el modelo, porque la alternativa con mayor utilidad antes de que los términos fuesen añadidos puede no tener la mayor utilidad después (ya que se añade una cantidad diferente a cada utilidad). Sin embargo, al aumentar la escala de utilidad suficientemente, podemos estar totalmente seguros de que la adición de los términos valor extremo no tiene ningún efecto. Consideremos un ejemplo con dos alternativas. Supongamos, utilizando el verdadero modelo con su escala original, que la utilidad de la alternativa 1 es 0.5 unidades mayor que la utilidad de la alternativa 2, de modo que la alternativa 1 es la elegida. Supongamos que añadimos un término valor extremo para cada alternativa. Hay una probabilidad considerable, dada la varianza de estos términos aleatorios, de que el valor obtenido para la alternativa 2 exceda el de la alternativa 1 por lo menos en media unidad, de manera que la alternativa 2 sea ahora la que obtiene mayor utilidad en lugar de la alternativa 1. Por lo tanto, la adición de los términos valor extremo cambia el modelo, ya que cambia la alternativa que tiene mayor utilidad. Supongamos, sin embargo, que podemos aumentar la escala de la utilidad original por un factor 10 (es decir, $\lambda = 0.10$). La utilidad de la alternativa 1 supera ahora la utilidad de la alternativa 2 de 5 unidades en lugar de 0.5 unidades. Es muy poco probable que la adición de términos valor extremo a estas utilidades invierta esta diferencia. Es decir, es muy poco probable, de hecho casi imposible, que el valor de $\varepsilon_{\mathrm{n2}}$ que se agrega a la utilidad de la alternativa 2 sea mayor en 5 unidades al término $\varepsilon_{\mathrm{n1}}$ que se agrega a la utilidad de la alternativa 1. Si el re-escalado en un factor 10 no es suficiente para asegurar que la adición del término valor extremo no tenga ningún efecto, entonces las utilidades originales podrían re-escalarse en un factor 100 o 1000. En algún momento, encontraremos una escala para la que la suma de los términos valor extremo no tenga ningún efecto. Dicho de manera sucinta, la adición de un término valor extremo a la verdadera utilidad, que convierte el modelo en un logit mixto, no cambia la utilidad de manera significativa cuando la escala de la utilidad es lo suficientemente grande. Un logit mixto puede aproximar cualquier RUM simplemente ampliando suficientemente la escala de la utilidad.

Esta demostración no pretende sugerir que aumentar la escala de la utilidad es la forma en que el investigador procederá realmente cuando especifique un logit mixto como una aproximación al verdadero modelo. Más bien, la demostración simplemente indica que si no se pueden encontrar otros medios para especificar un modelo logit mixto que aproxime el verdadero modelo, entonces este procedimiento de cambio de escala puede ser usado para lograr la aproximación. Por lo general, un logit mixto se puede especificar de manera que refleje adecuadamente el verdadero modelo sin necesidad de recurrir a una escala aumentada de utilidad. Por ejemplo, el verdadero modelo por lo general contendrá algún término iid que se agrega a la utilidad de cada alternativa. Suponiendo una distribución tipo valor

extremo para este término tal vez está lo suficientemente cerca de la realidad como para ser empíricamente indistinguible de otros supuestos de distribución para el término iid. En este caso, la escala de la utilidad se determina naturalmente por la varianza de este término iid. La tarea del investigador es simplemente encontrar las variables y una distribución de mezcla que capten las otras partes de la utilidad, es decir, las partes que están correlacionadas entre alternativas o heterocedásticas.

6.6 Simulación

El modelo logit mixto se adapta bien a los métodos de simulación para la estimación. La utilidad es $U_{nj} = \beta_n' x_{nj} + \varepsilon_{nj}$, donde los coeficientes β_n se distribuyen con densidad $f(\beta|\theta)$, donde θ se refiere colectivamente a los parámetros de esta distribución (tales como la media y la covarianza de β). El investigador especifica la forma funcional $f(\cdot)$ y desea estimar los parámetros θ. Las probabilidades de elección son

$$P_{ni} = \int L_{ni}(\beta) f(\beta|\theta) d\beta,$$

donde

$$L_{ni}(\beta) = \frac{e^{\beta' x_{ni}}}{\sum_{j=1}^{J} e^{\beta' x_{nj}}}.$$

Las probabilidades se aproximan mediante simulación para cualquier valor dado de θ: (1) Extraiga al azar un valor β de $f(\beta|\theta)$, y etiquételo como β^r, con el superíndice $r = 1$ en referencia al primer valor extraído. (2) Calcule la fórmula logit $L_{ni}(\beta^r)$ con este valor. (3) Repita los pasos 1 y 2 múltiples veces y promedie los resultados. Este promedio es la probabilidad simulada:

$$\breve{P}_{ni} = \frac{1}{R} \sum_{r=1}^{R} L_{ni}(\beta^r)$$

donde R es el número de valores extraídos al azar usados en la simulación. $\breve{P}_{ni}$ es un estimador no sesgado de P_{ni} por la forma en que se construye. Su varianza disminuye a medida que aumenta R. Es estrictamente positivo, de modo que $\ln \breve{P}_{ni}$ está definido, lo que es útil para aproximar a continuación la función log-verosimilitud. $\breve{P}_{ni}$ es suave (dos veces diferenciable) en los parámetros θ y en las variables x, lo que facilita la búsqueda numérica de la máxima verosimilitud y el cálculo de elasticidades. Y la suma de $\breve{P}_{ni}$ para todas las alternativas es uno, algo útil para hacer pronósticos.

Las probabilidades simuladas se insertan en la función log -verosimilitud para calcular una log-verosimilitud simulada:

$$SLL = \sum_{n=1}^{N} \sum_{j=1}^{J} d_{nj} \, ln \, \breve{P}_{nj},$$

donde $d_{nj} = 1$ si n eligió j y cero en caso contrario. El estimador de máxima verosimilitud simulada (MSLE) es el valor de θ que maximiza SLL. Las propiedades de este estimador se tratan en el Capítulo 10. Generalmente, se extraen al azar valores diferentes para cada observación. Este procedimiento mantiene la independencia entre decisores de las probabilidades simuladas que entran en la SLL. Lee

(1992) describe las propiedades del MSLE cuando se utilizan para todas las observaciones los mismos valores extraídos al azar.

La probabilidad simulada de un logit mixto puede relacionarse con métodos de simulación de tipo aceptación-rechazo (AR). La simulación AR se describe en la Sección 5.6 para los modelos probit, pero es aplicable de forma más general. Para cualquier modelo de utilidad aleatoria, el simulador AR se construye como sigue: (1) Se extrae al azar un valor de los términos aleatorios. (2) Se calcula la utilidad de cada alternativa a partir de ese valor y se identifica la alternativa con mayor utilidad. (3) Los pasos 1 y 2 se repiten múltiples veces. (4) La probabilidad de elección simulada para una alternativa concreta se calcula como la proporción de valores extraídos para los que esa alternativa ha sido la de mayor utilidad. El simulador AR es no sesgado por construcción. Sin embargo, no es estrictamente positivo para cualquier número finito de valores extraídos. Tampoco es una función suave, sino una función escalonada: constante dentro de los rangos de parámetros para los que la identidad de la alternativa con mayor utilidad no cambia para cualquier valor extraído, y con saltos donde los cambios en los parámetros cambian la identidad de la alternativa de mayor utilidad. Los métodos numéricos de maximización basados en el simulador AR se ven perjudicados por estas características. Para hacer frente a estos problemas numéricos, el simulador AR puede ser suavizado reemplazando la función indicadora 0-1 por la fórmula logit. Tal y como vimos en la Sección 5.6.2, el simulador AR suavizado-logit puede aproximar el simulador AR hasta un nivel de similitud arbitrario mediante un re-escalado de la utilidad apropiado.

El simulador logit mixto puede verse como un simulador AR suavizado-logit de cualquier RUM: se extraen valores al azar de los términos aleatorios, se calculan las utilidades para estos valores, las utilidades calculadas se insertan en la fórmula logit y se promedian los resultados. El teorema que afirma que un logit mixto puede aproximar cualquier modelo de utilidad aleatoria (sección 6.5) puede ser visto desde esta perspectiva. Sabemos por la Sección 5.6.2 que el simulador AR suavizado-logit puede aproximarse arbitrariamente al simulador AR de cualquier modelo, siempre y cuando se escale suficientemente la utilidad. Dado que el simulador logit mixto es equivalente a un simulador AR suavizado-logit, el modelo logit mixto simulado puede estar arbitrariamente cerca del simulador AR de cualquier modelo.

6.7 Datos de panel

La especificación del modelo logit mixto se puede generalizar fácilmente para permitir elecciones repetidas de cada decisor de la muestra. La especificación más simple trata los coeficientes que entran en la utilidad como parámetros que varían entre personas pero que son constantes en situaciones entre situaciones de elección de una misma persona. La utilidad de la alternativa j en una situación de elección t por parte de una persona n es $\mathrm{U}_{njt} = \beta_n \mathrm{x}_{njt} + \varepsilon_{njt}$, con ε_{njt} siendo valor extremo iid a lo largo del tiempo, de las personas y de las alternativas. Considere una secuencia de alternativas, una para cada período de tiempo, $\mathbf{i} = \{i_1, \dots, i_T\}$. Condicionada a β, la probabilidad de que la persona haga esta secuencia de elecciones es el producto de fórmulas logit:

$$\boldsymbol{L}_{ni}(\beta) = \prod_{t=1}^{T} \left[\frac{e^{\beta_n' x_{ni_t t}}}{\sum_j e^{\beta_n' x_{njt}}} \right], \tag{6.2}$$

dado que los ε_{njt}s son independientes en el tiempo. La probabilidad no condicionada es la integral de este producto sobre todos los valores de β:

$$P_{ni} = \int \boldsymbol{L}_{ni}(\beta) f(\beta) d\beta. \tag{6.3}$$

La única diferencia entre un logit mixto con elecciones repetidas y uno con una sola elección por decisor es que el integrando implica un producto de fórmulas logit, una para cada período de tiempo, en lugar de sólo una fórmula logit. La probabilidad se simula de manera similar a la probabilidad con un único período de elección. Se extrae al azar un valor β de su distribución de probabilidad. Se calcula la fórmula logit para cada período y se calcula el producto de estas fórmulas logit. Este proceso se repite para múltiples valores extraídos y se promedian los resultados.

Es posible agregar a la utilidad variables exógenas pasadas y futuras en un período determinado para representar una respuesta diferida o un comportamiento anticipatorio, como se describe en la sección 5.5 en relación a un modelo probit con datos de panel. Sin embargo, a diferencia de probit, en un modelo logit mixto es posible añadir variables dependientes diferidas sin cambiar el procedimiento de estimación. Si condicionamos a β_n, los únicos términos aleatorios que quedan en el logit mixto son los términos ε_{nj}s, que son independientes en el tiempo. Una variable dependiente diferida que entre en U_{njt} no está correlacionada con estos términos de error restantes para el período t, ya que estos términos son independientes a lo largo del tiempo. Las probabilidades condicionadas (condicionadas a β) son, por lo tanto, las mismas de la ecuación (6.2), pero con x incluyendo variables dependientes diferidas. La probabilidad no condicionada es la integral de esta probabilidad condicionada sobre todos los valores de ß, que es justamente la ecuación (6.3). En este sentido, logit mixto es un modelo más conveniente que probit para la representación de la dependencia del estado, ya que las variables dependientes diferidas se pueden añadir al logit mixto sin ajustar la fórmula de probabilidad o el método de simulación. Erdem (1996) y Johannesson y Lundin (2000) explotan esta ventaja para examinar la formación de hábitos y la búsqueda de la variedad dentro de un logit mixto que también captura la variación aleatoria de preferencias.

Si las elecciones y los datos no se observan desde el inicio del proceso (es decir, desde la primera situación de elección que la persona afronta), es necesario resolver la cuestión de las condiciones iniciales, al igual que con probit. El investigador debe representar de alguna manera la probabilidad de la primera elección observada, que depende de las elecciones previas no observadas. Heckman y Singer (1986) proporcionan formas de manejar este problema. Sin embargo, cuando el investigador observa el proceso de elección desde el principio, el problema de las condiciones iniciales no se plantea. En este caso, el uso de variables dependientes diferidas para capturar la inercia en la elección u otros tipos de dependencia del estado, es algo sencillo para el modelo logit mixto. Los datos de preferencia declarada (es decir, las respuestas a una serie de hipotéticas situaciones de elección planteadas a los participantes en una encuesta) son un ejemplo claro de datos en los que el investigador observa toda la secuencia de elecciones.

En la especificación desarrollada hasta el momento, así como en casi todas las aplicaciones prácticas, se asume que los coeficientes β_n son constantes entre diferentes situaciones de elección para un mismo decisor. Este supuesto es apropiado si las preferencias del decisor son estables a lo largo del período de tiempo que comprende las elecciones repetidas. Sin embargo, es posible especificar que los coeficientes asociados a cada persona puedan variar con el tiempo de diversas maneras. Por ejemplo, las preferencias de cada persona pueden estar correlacionadas en serie entre situaciones de elección, por lo que la utilidad sería

$$U_{njt} = \beta_{nt} x_{njt} + \varepsilon_{njt},$$

$$\beta_{nt} = b + \tilde{\beta}_{nt},$$

$$\tilde{\beta}_{nt} = \rho \tilde{\beta}_{nt-1} + \mu_{nt},$$

donde b es fijo y μ_{nt} es iid sobre n y t. La simulación de la probabilidad de que la secuencia de elecciones se produzca se realiza como sigue:

1. Extraiga un valor al azar para μ_{n1}^r para el período inicial y calcule la fórmula logit para este período utilizando $\beta_{n1}^r = b + \mu_{n1}^r$.
2. Extraiga un valor al azar para μ_{n2}^r para el segundo período, calcule $\beta_{n2}^r = b + \rho\mu_{n1}^r + \mu_{n2}^r$ y luego calcule la fórmula logit en base a esta β_{n2}^r.
3. Continúe para todos los períodos de tiempo T.
4. Tome el producto de los T logits.
5. Repita los pasos 1-4 para numerosas extracciones de valores.
6. Promedie los resultados.

La carga que colocamos en la tarea de simulación es mayor que con coeficientes constantes en el tiempo para cada persona, requiriendo la realización de extracciones T veces.

6.8 Estudio de un caso

A modo ilustrativo, considere un modelo logit mixto relativo a las elecciones que los pescadores realizan sobre los sitios a los que ir a pescar (Train, 1999). La especificación emplea el uso de coeficientes aleatorios. La utilidad es $U_{njt} = \beta_n x_{njt} + \varepsilon_{njt}$, con coeficientes β_n que varían entre pescadores pero no entre las decisiones (los viajes realizados) de cada pescador. La probabilidad de la secuencia de los sitios elegidos por cada pescador viene dada por la ecuación (6.3).

La muestra consta de 962 viajes a ríos de Montana realizados por un total de 258 pescadores durante el período comprendido entre julio de 1992 y agosto de 1993. Se definieron un total de 59 posibles sitios en los ríos, con base a la situación geográfica y a otros factores relevantes. Cada sitio contiene uno o más tramos de ríos definidos en el sistema de información fluvial de Montana. Las siguientes variables entran como elementos de x para cada sitio:

1. Disponibilidad de peces, medida en unidades de 100 peces por cada 1000 pies de río.
2. Valoración estética, medida en una escala de 0 a 3, siendo 3 la valoración más alta.
3. Costo del viaje: Costo de viajar desde la residencia del pescador hasta el sitio de pesca, incluyendo la variable de costo de conducir (combustible, mantenimiento, neumáticos, aceite) y el valor del tiempo consumido en la conducción (con el tiempo valorado como un tercio del salario del pescador).
4. Indicador de que la "Guía del pescador en Montana" menciona el sitio como sitio de pesca mayor.
5. Número de campings por bloque en el sitio, tal y como se define bloque en el "U.S. Geological Survey (USGS)".
6. Número de zonas recreativas estatales por bloque USGS.
7. Número de especies de pesca restringida (especies restringidas) en el sitio.
8. Logaritmo del tamaño del sitio, en bloques USGS.

Lógicamente, los coeficientes de las variables 4-7 pueden tener cualquier signo; por ejemplo, a algunos pescadores les puede gustar tener campings cerca mientras que otros pueden preferir la privacidad que proporciona no tener campings cercanos. A cada uno de estos coeficientes se le asigna una distribución normal independiente con media y desviación estándar a estimar. Se espera que los coeficientes de costo del viaje, disponibilidad de peces y valoración estética del sitio tengan el mismo signo para todos los pescadores, difiriendo entre pescadores sólo en sus magnitudes. A estos coeficientes se asignan distribuciones log-normales independientes. Se estiman la media y la desviación estándar del logaritmo del coeficiente y, a partir de estas estimaciones, se calculan la media y la desviación estándar del propio coeficiente. Dado que la distribución log-normal se define en el rango positivo y se espera que el costo del viaje tenga un coeficiente negativo para todos los pescadores, introducimos en el modelo el costo del viaje con el signo invertido (negativa del costo). Se asume que el coeficiente del logaritmo del tamaño del sitio sea fijo. Esta variable permite contemplar el hecho de que la probabilidad de visitar un sitio más grande sea mayor que la de un sitio pequeño, en igualdad de todos los demás parámetros. Permitir que el coeficiente de esta variable variase entre personas, aunque sería posible, no sería particularmente significativo. Una versión del modelo con coeficientes correlacionados la proporciona Train (1998). El modelo de elección del sitio forma parte de un modelo global, dado por Desvousges et al. (1996), de la elección conjunta de la frecuencia de viaje y la elección del sitio.

La simulación se llevó a cabo utilizando mil extracciones de valores al azar para cada pescador de la muestra. Los resultados se facilitan en la tabla 6.1. La desviación estándar de cada coeficiente aleatorio es altamente significativa, lo que indica que esos coeficientes en efecto varían en la población.

Tabla 6.1. Modelo logit mixto de la elección de sitios de pesca en ríos

Variable	Parámetro	Valor	Error estándar
Disponibilidad de peces	Media del ln(coeficiente)	−2.876	0.6066
	Desv. est. del ln(coeficiente)	1.016	0.2469
Valoración estética	Media del ln(coeficiente)	−0.794	0.2287
	Desv. est. del ln(coeficiente)	0.849	0.1382
Costo total (negativo)	Media del ln(coeficiente)	−2.402	0.0631
	Desv. est. del ln(coeficiente)	0.801	0.0781
Listado en la guía como sitio mayor	Media del coeficiente	1018	0.2887
	Desv. est. del coeficiente	2.195	0.3518
Campings	Media del coeficiente	0.116	0.3233
	Desv. est. del coeficiente	1.655	0.4350
Áreas de acceso	Media del coeficiente	−0.950	0.3610
	Desv. est. del coeficiente	1.888	0.3511
Especies restringidas	Media del coeficiente	−0.499	0.1310
	Desv. est. del coeficiente	0.899	0.1640
Log(tamaño)	Media del coeficiente	0.984	0.1077
Índice del coeficiente de verosimilitud		0.5018	
SLL en convergencia		−1932.33	

Consideremos en primer lugar los coeficientes distribuidos normalmente. Las medias estimadas y las desviaciones estándar de estos coeficientes proporcionan información sobre la proporción de la

población que valora positivamente un atributo del sitio y la que lo valora negativamente. La distribución del coeficiente relativo a la variable que indica si el sitio es mencionado como importante en la "Guía del pescador en Montana" obtiene una media estimada de 1.018 y una desviación estándar estimada de 2.195, de tal manera que el 68 por ciento de la distribución está por encima de cero y el 32 por ciento por debajo. Esto implica que estar catalogado como un sitio importante en la "Guía del pescador en Montana" es un incentivo positivo para cerca de dos tercios de los pescadores y un factor negativo para el otro tercio, que aparentemente prefiere la soledad. La presencia de campings es preferida por aproximadamente la mitad (53 por ciento) de los pescadores y es evitada por la otra mitad. Y se estima que cerca de un tercio de los pescadores (31 por ciento) prefieren tener numerosas áreas recreativas, mientras que las otras dos terceras partes prefieren tener menos áreas.

Consideremos ahora los coeficientes definidos con distribuciones log-normales. Un coeficiente β^k sigue una distribución log-normal si el logaritmo de β^k se distribuye normalmente. Parametrizamos la distribución log-normal en términos de la distribución normal subyacente. Es decir, estimamos los parámetros m y s que representan la media y la varianza del logaritmo del coeficiente: $\ln \beta^k \sim N(m, s)$. La media y la varianza de β^k se obtienen acto seguido a partir de las estimaciones de m y s. La mediana es $\exp(\mathrm{m})$, la media es $\exp(\mathrm{m} + \mathrm{s}/2)$ y la varianza es $\exp(2\mathrm{m} + \mathrm{s})\,[\exp(\mathrm{s}) - 1]$. Las estimaciones puntuales (*point estimates*) implican que los coeficientes relativos a la disponibilidad de peces, valoración estética y costo del viaje tienen las siguientes medianas, medias y desviaciones estándar:

Variable	Mediana	Media	Desv. Estándar
Disponibilidad de peces	0.0563	0.0944	0.1270
Valoración estética	0.4519	0.6482	0.6665
Costo del viaje	0.0906	0.1249	0.1185

El ratio para un pescador entre sus coeficiente de disponibilidad de peces y costo del viaje es una medida de la cantidad económica que el pescador está dispuesto a pagar para tener más peces en el río. Dado que el ratio entre dos términos distribuidos log-normal independientemente también se distribuye log-normal, podemos calcular los momentos estadísticos de la distribución de la predisposición a pagar. El logaritmo del ratio entre el coeficiente de disponibilidad de peces y de costo del viaje tiene una media estimada de -0.474 y una desviación estándar de 1.29. La relación en sí, por lo tanto, tiene una mediana de 0.62, media de 1.44 y desviación estándar de 2.96. Es decir, la predisposición media a pagar para que se incremente la disponibilidad de peces en 100 peces por cada 1000 pies de río se estima en 1.44$, y la variación en la predisposición de los pescadores a pagar por un disponibilidad de peces adicional es muy amplia. Del mismo modo, la predisposición media estimada a pagar por un sitio que tiene una valoración estética superior a 1 es de 9.87$, y de nuevo la variación es bastante grande.

Como ilustra este caso, el modelo logit mixto proporciona más información que un logit estándar, ya que el logit mixto estima hasta qué punto los pescadores difieren en sus preferencias por los atributos de los sitios. Las desviaciones estándar de los coeficientes son significativas, lo que indica que un logit mixto proporciona una mejor representación de la situación de elección que un logit estándar, que supone que los coeficientes son los mismos para todos los pescadores. El modelo logit mixto permite también contemplar varios viajes de cada pescador en la muestra y que las preferencias de cada pescador se apliquen a cada uno sus viajes.

7

Variaciones sobre un mismo tema

7.1 Introducción

La simulación da al investigador la libertad de especificar modelos que representen adecuadamente las situaciones de elección objeto de estudio, sin ser obstaculizado por consideraciones puramente matemáticas. Esta perspectiva ha sido el tema principal de nuestro libro. Los modelos de elección discreta que hemos estudiado - es decir, logit, logit jerárquico, probit y logit mixto – están siendo utilizados en la gran mayoría de los casos estudiados. Sin embargo, los lectores no deben sentirse obligados a utilizar estos modelos. En el presente capítulo, se describen varios modelos que se obtienen en condiciones un tanto diferentes de comportamiento. Estos modelos son variaciones de los ya expuestos, dirigidos hacia temas y datos específicos. El objetivo no es simplemente describir modelos adicionales, sino facilitar una explicación que ilustre cómo el investigador puede examinar una situación de elección y desarrollar un modelo y un procedimiento de estimación adecuados para esa situación particular, usando elementos (y adaptándolos) del conjunto estándar de modelos y herramientas.

Cada sección de este capítulo está motivada por un tipo de datos que representan el resultado de un proceso de elección en particular. El escenario en el que podrían surgir estos datos se describe y se identifican las limitaciones que los modelos principales presentan en el tratamiento de estos datos. En cada caso se describe un nuevo modelo que representa mejor la situación de elección. A menudo, este nuevo modelo es sólo un ligero cambio de uno de los modelos principales. Sin embargo, el ligero cambio a menudo hará inservible el software estándar, por lo que el investigador tendrá que desarrollar su propio software, tal vez mediante la modificación de los códigos que están disponibles para los modelos estándar. La capacidad de revisar el código para representar nuevas especificaciones permite al investigador aprovechar la libertad que ofrece este campo.

7.2 Datos de preferencia declarada y de preferencia revelada

Decimos que unos datos son de *preferencia revelada* (*revealed-preference data*) si se refieren a las elecciones que las personas realizan en situaciones del mundo real. Estos datos se denominan así porque en ellos la gente revela sus gustos o preferencias a través de las decisiones que toman en su vida. Asimismo, decimos que unos datos son de *preferencia declarada* (*stated-preference data*) cuando

se recogen a través de experimentos o encuestas en los que se presentan a los participantes situaciones de elección hipotéticas. El término se refiere al hecho de que los respondientes declaran cuáles serían sus elecciones en las situaciones hipotéticas. Por ejemplo, en una encuesta, podríamos presentar a una persona tres automóviles con diferentes precios y otros atributos. Preguntaríamos cuál de los tres automóviles compraría si sólo pudiese elegir entre estos tres modelos en el mundo real. La respuesta que facilitase sería la elección declarada por esa persona. Un dato revelado similar de ese respondiente lo obtendríamos preguntando qué automóvil compró la última vez que compró uno.

Cada tipo de dato tiene sus ventajas y sus limitaciones. Los datos de preferencia revelada tienen la ventaja de que reflejan las elecciones reales. Esto, por supuesto, es una gran ventaja. Sin embargo, este tipo de datos se limitan a las situaciones de elección y a los atributos de las alternativas que ya existen actualmente o que han existido en algún momento. A menudo, un investigador querrá examinar las respuestas de las personas a situaciones que no existen en la actualidad, tales como la demanda de un nuevo producto. Los datos de preferencia revelada simplemente no están disponibles para estas nuevas situaciones. Incluso para situaciones de elección que sí existen actualmente, puede haber una variación insuficiente de los factores relevantes como para permitir la estimación con datos de preferencia revelada. Por ejemplo, supongamos que el investigador quiere examinar los factores que afectan la elección del proveedor de energía en los hogares de California. Aunque los clientes residenciales han podido elegir entre diferentes proveedores durante muchos años, apenas ha habido diferencias apreciables de precio entre las ofertas disponibles. La respuesta de los clientes al precio no se puede estimar partiendo de unos datos que contienen poca o ninguna variación de precios. En este sentido, se plantea una paradoja interesante. Si los clientes fuesen muy sensibles al precio, entonces los proveedores, sabiendo esto, ofrecerían precios que igualasen los precios de sus competidores; en estas condiciones se acostumbra a producir una situación de equilibrio muy conocida, en la que todas las empresas ofrecen su servicio (esencialmente) al mismo precio. Si se utilizasen los datos de este mercado en un modelo de elección, el coeficiente de precio resultaría ser insignificante, ya que existe poca variación de precios en los datos. El investigador podría concluir erróneamente de esta poca significación que el precio no es importante para los consumidores. Esta paradoja es inherente a los datos de preferencia revelada. Los factores más importantes para los consumidores a menudo exhiben la menor variación debido a las fuerzas naturales de equilibrio del mercado. Su importancia, por tanto, podría ser difícil de detectar con datos de preferencia revelada.

Los datos de preferencia declarada complementan los datos de preferencia revelada. Para ello, se diseña un cuestionario en el que se presenta al entrevistado uno o más experimentos de elección. En cada experimento, se describen dos o más opciones y se pregunta al encuestado qué opción elegiría si se le presentase esa situación en la vida real. Por ejemplo, en los datos que se examinan en el capítulo 11, se presentan a cada encuestado 12 experimentos. En cada experimento, se describieron cuatro proveedores de energía hipotéticos, incluyendo el precio, los términos del contrato y otros atributos dados para cada proveedor. Se pidió al encuestado que indicase cuál de los cuatro proveedores elegiría.

La ventaja de los datos de preferencia declarada es que los experimentos pueden diseñarse para que contengan tanta variación en cada atributo como el investigador considere apropiado. Aunque pueda existir muy poca variación de precios entre proveedores en el mundo real, los proveedores descritos en los experimentos pueden mostrar suficiente diferencia de precios como para permitir una estimación precisa. Es posible hacer que los atributos varíen entre encuestados y entre experimentos para cada encuestado. Este grado de variación contrasta con los datos de mercado, donde a menudo los mismos productos están disponibles para todos los consumidores de modo que no hay variación en los atributos de los productos entre clientes. Es importante destacar que para productos que no han sido ofrecidos con anterioridad, o para nuevos atributos de productos existentes, los datos de preferencia declarada permiten la estimación

de modelos de elección cuando no existen datos de preferencia revelada. Louviere et al. (2000) describen la forma apropiada de recolectar y analizar datos de preferencia declarada.

Las limitaciones de los datos de preferencia declarada son obvias: lo que las personas declaran que van a hacer a menudo no es lo mismo que lo que realmente hacen. Las personas pueden no saber lo que harían si una situación hipotética fuera real. O pueden no estar dispuestos a decir lo que harían. De hecho, la opinión de los respondientes sobre qué harían frente a una situación hipotética podría estar influenciada por factores que no surgirían en situaciones de elección reales, como por ejemplo su percepción de lo que entrevistador espera o desea que respondan.

Al combinar los datos de preferencias reveladas y declaradas, es posible aprovechar las ventajas de cada método y mitigar al mismo tiempo sus limitaciones. Los datos de preferencia declarada proporcionan la variación necesaria de los atributos, mientras que los datos de preferencia revelada acercan las predicciones de cuota de mercado a la realidad. Para aprovechar las fortalezas de cada tipo de dato se necesita un procedimiento de estimación que (1) permita que los ratios entre coeficientes (que representan la importancia relativa de los diversos atributos) se estimen principalmente a partir de datos de preferencia declarada (o, más generalmente, de cualquier variación en los atributos existente, que por lo general proviene de datos de preferencia declarada), al mismo tiempo que (2) permita que las constantes específicas de alternativa y la escala global de los parámetros se determinen a través de datos de preferencia revelada (ya que las constantes y la escala determinan cuotas de mercado promedio en las condiciones de partida).

Diversos procedimientos para la estimación de modelos de elección discreta que combinan datos de preferencia declarada y observada han sido descritos por Ben-Akiva y Morikawa (1990), Hensher y Bradley (1993) y Hensher et al. (1999) en el contexto de modelos logit, y por Bhat y Castelar (2002) y Brownstone et al. (2000) para modelos logit mixtos. Estos procedimientos constituyen variaciones de los métodos que ya hemos examinado. El problema más frecuente que surge al combinar datos de preferencia revelada y declarada es que los factores no observados acostumbran a diferir entre los dos tipos de datos. Describimos en los siguientes párrafos cómo afrontar fácilmente esta cuestión.

Especificamos la utilidad que una persona n obtiene de la alternativa j en una situación t como $U_{njt} = \beta' x_{njt} + e_{njt}$, donde x_{njt} no incluye constantes específicas de alternativa y e_{njt}representa el efecto de factores no observados por el investigador. Estos factores tienen una media para cada alternativa (que representa el efecto promedio que todos los factores excluidos tienen sobre la utilidad de esa alternativa) y una distribución en torno a esta media. La media es capturada por una constante específica de alternativa denominada c_j, y la distribución alrededor de esta media, para un modelo logit estándar, es de tipo valor extremo con varianza $\lambda^2\pi^2/6$. Como se describe en los capítulos 2 y 3, la escala de la utilidad se establece por la normalización de la varianza de la parte no observada de la utilidad. La función de utilidad se convierte $U_{njt} = (\beta/\lambda)' x_{njt} + c_j/\lambda + \varepsilon_{njt}$, donde el error normalizado $\varepsilon_{njt} = (e_{njt} - c_j)/\lambda$ es ahora valor extremo iid con varianza $\pi^2/6$. La probabilidad de elección viene dada por la fórmula logit basada en $(\beta/\lambda)' x_{njt} + c_j/\lambda$. Los parámetros a estimar son los parámetros originales divididos por un factor de escala λ.

Esta especificación es razonable para muchos tipos de datos y situaciones de elección. Sin embargo, no hay ninguna razón que nos haga esperar que las constantes específicas de alternativa y el factor de escala sean iguales tanto para los datos de preferencia declarada como para los datos de preferencia revelada. Estos parámetros reflejan los efectos de factores no observados, que son necesariamente diferentes en situaciones de elección reales y en situaciones hipotéticas planteadas en una encuesta. En elecciones reales, entran en juego múltiples factores que afectan a la persona pero que no son observados por el investigador. En un experimento de preferencias declaradas, generalmente se solicita al respondiente que asuma que todas las alternativas son iguales en relación a cualquier factor que no

sea mencionado explícitamente en el experimento. Si el entrevistado sigue estas instrucciones de forma precisa, no habría, por definición, factores no observados en las elecciones de preferencia declarada. Por supuesto, los encuestados inevitablemente incorporan algunos conceptos externos a los experimentos en el momento de participar, de forma que sí entran factores no observados. Sin embargo, no hay razón para esperar que estos factores sean los mismos, en términos de media o varianza, a los factores que operan en las elecciones de la vida real.

Para tener en cuenta estas diferencias, se especifican constantes y parámetros de escala diferenciados para situaciones de elección de preferencia declarada y revelada. Definamos c_j^s y c_j^r como los parámetros que representan la media del efecto de factores no observados de la alternativa j en experimentos de preferencia declarada y en elecciones de preferencia revelada, respectivamente. Del mismo modo, sean λ^s y λ^r los parámetros que representan las escalas (proporcionales a las desviaciones estándar) de las distribuciones de los factores no observados en torno a estas medias en situaciones de preferencia declarada y revelada, respectivamente. Para ajustar la escala global de utilidad, normalizamos cualquiera de los dos parámetros de escala a 1, lo que hace que el otro parámetro de escala pase a ser el ratio de los dos parámetros de escala original. Vamos a normalizar λ^r, por lo que λ^s refleja la varianza de los factores no observados en situaciones de preferencia declarada respecto a la varianza en situaciones de preferencia revelada. La utilidad se convierte de este modo en

$$U_{njt} = (\beta/\lambda^s)' x_{njt} + c_j^s/\lambda^s + \varepsilon_{njt},$$

para cada t que sea una situación de preferencia declarada, y

$$U_{njt} = \beta' x_{njt} + c_j^r + \varepsilon_{njt},$$

para cada t que sea una situación de preferencia revelada.

El modelo se estima usando los datos disponibles, tanto los de elecciones de preferencia revelada como declarada. Ambos grupos de observaciones se "apilan" conjuntamente para ser introducidos en una rutina de estimación logit. Un conjunto separado de constantes específicas de alternativa se estima para los datos de preferencia declarada y preferencia revelada. Es importante destacar que los coeficientes del modelo se dividen por un parámetro $1/\lambda^s$ sólo para las observaciones de preferencia declarada. Este escalado diferenciado no es factible en la mayoría de los paquetes de software de estimación logit estándar. Sin embargo, el investigador puede modificar fácilmente el código disponible (o su propio código) para permitir este parámetro extra. Hensher y Bradley (1993) muestran cómo estimar este modelo usando software para logit jerárquico.

Observe que con esta configuración, los elementos de β se estiman usando ambos tipos de datos. Las estimaciones necesariamente reflejarán la cantidad de variación que cada tipo de dato contiene para los atributos (es decir, los elementos de x). Si hay poca varianza en los datos de preferencia revelada, reflejando las condiciones de los mercados del mundo real, entonces las βs se determinarán fundamentalmente a través de los datos de preferencia declarada, que contienen cualquier variación que se haya definido en los experimentos. En la medida en que los datos de preferencia revelada contengan variación utilizable, esta información será incorporada en las estimaciones.

Las constantes específicas de alternativa se calculan por separado para los dos tipos de datos. Esta distinción permite al investigador evitar muchos de los sesgos que los datos de preferencia declarada podrían arrojar. Por ejemplo, los encuestados a menudo declaran que van a comprar un producto mucho más de lo que en realidad terminan haciendo. La probabilidad promedio de comprar el producto se captura en la constante específica de alternativa del producto. Si este sesgo se está produciendo, entonces la constante estimada para los datos de preferencia declarada será mayor que la de los datos

de preferencia revelada. En el momento de hacer predicciones, el investigador puede utilizar la constante de los datos de preferencia revelada, vinculando así la predicción a una realidad basada en el mercado actual. Del mismo modo, la escala de los datos de preferencia revelada (que se normaliza a 1) se puede utilizar en la predicción en lugar de la escala de los datos de preferencia declarada, incorporando así correctamente la varianza del mundo real en factores no observados.

7.3 Datos de ordenación

En experimentos de preferencia declarada, puede requerirse a los encuestados que ordenen por preferencia las alternativas disponibles en lugar de identificar únicamente la alternativa que escogerían. Esta ordenación puede realizarse de diferentes maneras. Puede pedirse a los encuestados que indiquen qué alternativa elegirían y, posteriormente, después de haber hecho esta primera elección, se les puede pedir cuál de las alternativas restantes elegirían, continuando así con el total de las alternativas. Otra forma de obtener la misma información sería pedir a los encuestados que simplemente ordenen las alternativas de mejor a peor. En cualquier caso, los datos que el investigador obtiene constituyen una ordenación de las alternativas que, presumiblemente, refleja la utilidad que el encuestado obtiene de cada una de ellas.

Los datos de ordenación se pueden manejar a través de un modelo logit estándar o logit mixto, utilizando software disponible en la actualidad sin necesidad de modificaciones. Lo único que se requiere es que los datos de entrada del modelo puedan construirse de una manera particular, que se describe en el texto siguiente. Para un modelo probit, el software disponible necesitaría modificarse ligeramente para poder manejar los datos de ordenación. Sin embargo, la modificación es simple. En primer lugar, consideraremos el caso del logit estándar y mixto.

7.3.1 Logit estándar y mixto

Bajo los supuestos del modelo logit estándar, la probabilidad de cualquier ordenación posible de las alternativas de mejor a peor puede expresarse como el producto de fórmulas logit. Consideremos, por ejemplo, un encuestado al que se le presentan cuatro alternativas denominadas A, B, C y D. Supongamos que la persona ordenó las alternativas de la siguiente manera: C, B, D, A, donde C es la primera elección. Si la utilidad de cada alternativa se distribuye valor extremo iid (como se requiere para un modelo logit), entonces la probabilidad de esta ordenación se puede expresar como la probabilidad logit de elegir la alternativa C del conjunto A, B, C, D, *por* la probabilidad logit de elegir la alternativa B entre las alternativas restantes A, B, D, *por* la probabilidad de elegir la alternativa D entre las alternativas restantes A y D.

Dicho de forma más explícita, si $U_{nj} = \beta' x_{nj} + \varepsilon_{nj}$ para $\mathrm{j} = \mathrm{A}, \ldots, \mathrm{D}$ con ε_{nj} tipo valor extremo iid, entonces

$$Prob(orden\ C, B, D, A) =$$

$$= \frac{e^{\beta' x_{nC}}}{\sum_{j=A,B,C,D} e^{\beta' x_{nj}}} \frac{e^{\beta' x_{nB}}}{\sum_{j=A,B,D} e^{\beta' x_{nj}}} \frac{e^{\beta' x_{nD}}}{\sum_{j=A,D} e^{\beta' x_{nj}}} \qquad (7.1)$$

Esta simple expresión para la probabilidad de una ordenación es el resultado de la forma particular que tiene la distribución de valor extremo, mostrada por primera vez por Luce y Suppes (1965). No puede ser aplicada en general; por ejemplo, no aplica a modelos probit.

La ecuación (7.1) implica que la ordenación de las cuatro alternativas se puede representar como tres elecciones independientes del respondiente. Estas tres elecciones se denominan pseudo-observaciones, ya que la ordenación completa de cada encuestado, lo que constituye una observación real, se re-escribe como si se tratase de varias observaciones. En general, una ordenación de J alternativas ofrece

$J-1$ pseudo-observaciones en un modelo logit estándar. Para la primera pseudo-observación se considera que todas las alternativas están disponibles y la variable dependiente identifica la alternativa ordenada en primera posición. Para la segunda pseudo-observación, se descarta la alternativa que ocupó el primer lugar de la ordenación. Las alternativas restantes constituyen el nuevo conjunto de elección, y la variable dependiente identifica la segunda mejor alternativa, y así sucesivamente. Al crear el archivo de entrada para la estimación logit, las variables explicativas para cada alternativa se repiten $J-1$ veces, lo que hace proliferar muchas pseudo-observaciones. La variable dependiente para estas pseudo-observaciones identifica, respectivamente, la alternativas clasificada en primera posición, en segunda posición y así sucesivamente. Para cada pseudo-observación, las alternativas que han sido ordenadas por delante de la variable dependiente para esa pseudo-observación se omiten. Una vez los datos se han construido de esta manera, podemos hacer la estimación logit como de costumbre.

Un modelo logit sobre alternativas ordenadas se llama a menudo un *logit expandido o explotado* (*exploded logit*), ya que cada observación explota en varias pseudo-observaciones para facilitar la estimación. Aplicaciones destacadas de este modelo incluyen Beggs et al. (1981), Chapman y Staelin (1982), y Hausman y Ruud (1987).

Un modelo logit mixto puede ser estimado con datos ordenados usando la misma explosión. Supongamos ahora que β es aleatoria con densidad $g(\beta|\theta)$, donde θ son los parámetros de esta distribución. Condicionada a β, la probabilidad de la ordenación de la persona es un producto de logits, como se indica en la ecuación (7.1). La probabilidad no condicionada es la integral de este producto sobre la densidad de β:

$$Prob(orden\ C,B,D,A) =$$

$$= \int \left(\frac{e^{\beta' x_{nC}}}{\sum_{j=A,B,C,D} e^{\beta' x_{nj}}} \frac{e^{\beta' x_{nB}}}{\sum_{j=A,B,D} e^{\beta' x_{nj}}} \frac{e^{\beta' x_{nD}}}{\sum_{j=A,D} e^{\beta' x_{nj}}} \right) \times g(\beta|\theta) d\beta \qquad (7.2)$$

El modelo logit mixto sobre alternativas ordenadas se calcula con las rutinas convencionales para logit mixto con datos de panel, utilizando la configuración de datos de entrada tal y como se ha descrito anteriormente para logit, donde las $J-1$ pseudo-observaciones por cada ordenación se tratan como $J-1$ elecciones en un panel. El modelo logit mixto incorpora el hecho de que cada respondiente tiene sus propios coeficientes y, sobre todo, que los coeficientes del respondiente afectan a toda su ordenación, de manera que las pseudo-observaciones están correlacionadas. Un modelo logit estándar sobre datos ordenados no permite esta correlación.

7.3.2 Probit

Los datos de ordenación también pueden utilizarse de forma efectiva en un modelo probit. Sea la utilidad de cuatro alternativas tal y como se ha definido para logit, excepto que los términos de error siguen una distribución normal conjunta: $U_{nj} = \beta' x_{nj} + \varepsilon_{nj}$ para $j = A, B, C, D$, donde $\varepsilon_n = \langle \varepsilon_{nA}, \ldots, \varepsilon_{nD} \rangle'$se distribuye $N(0, \Omega)$. Como antes, la probabilidad de la ordenación de la persona es $\text{Prob}\big(\text{ordenación}(C,B,D,A)\big) = \text{Prob}(U_{nC} > U_{nB} > U_{nD} > U_{nA})$. Descomponiendo esta probabilidad conjunta en varias probabilidades condicionadas y una marginal no ayuda en el caso del modelo probit de la misma forma que lo hacía en logit, ya que las probabilidades condicionadas no colapsan en probabilidades no condicionadas como sí lo hacen bajo la hipótesis de errores independientes. Debemos usar una estrategia diferente. Recordemos que para modelos probit vimos que era muy conveniente trabajar con diferencias de utilidad en lugar de trabajar con las utilidades directamente. Denotemos $\tilde{U}_{njk} = U_{nj} - U_{nk}$, $\tilde{x}_{njk} = x_{nj} - x_{nk}$ y $\tilde{\varepsilon}_{njk} = \varepsilon_{nj} - \varepsilon_{nk}$. La probabilidad de la ordenación puede

expresarse ahora como $\text{Prob}\big(\text{ordenación}(C,B,D,A)\big) = \text{Prob}(U_{nC} > U_{nB} > U_{nD} > U_{nA}) = \text{Prob}(\widetilde{U}_{nBC} < 0, \widetilde{U}_{nDB} < 0, \widetilde{U}_{nAD} < 0)$.

Para expresar esta probabilidad, definimos una matriz de transformación M que calcula las diferencias apropiadas. El lector puede querer revisar en este punto la sección 5.6.3 en relación a la simulación de probabilidades probit para una alternativa elegida, que utiliza una matriz de transformación similar. El mismo procedimiento se utiliza para datos de ordenación, pero con una matriz de transformación diferente.

Apilamos las alternativas de la A a la D, de manera que la utilidad queda expresada en forma vectorial como $U_n = V_n + \varepsilon_n$, donde $\varepsilon_n \sim N(0, \Omega)$. Definimos la matriz 3 × 4

$$M = \begin{pmatrix} 0 & 1 & -1 & 0 \\ 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & -1 \end{pmatrix}$$

Esta matriz tiene una fila por cada desigualdad en el argumento de la probabilidad $\text{Prob}(\widetilde{U}_{nBC} < 0, \widetilde{U}_{nDB} < 0, \widetilde{U}_{nAD} < 0)$. Cada fila contiene un 1 y un -1, junto con ceros, donde el 1 y el -1 identifican las alternativas que están siendo diferenciadas para la desigualdad. Con esta matriz, la probabilidad de las alternativas ordenadas se convierte en

$$Prob\big(ordenación(C,B,D,A)\big) = Prob(\widetilde{U}_{nBC} < 0, \widetilde{U}_{nDB} < 0, \widetilde{U}_{nAD} < 0)$$

$$= Prob(MU_n < 0)$$

$$= Prob(MV_n + M\varepsilon_n < 0)$$

$$= Prob(M\varepsilon_n < -MV_n).$$

Las diferencias de error definidas por $M\varepsilon_n$ se distribuyen con densidad normal conjunta con media cero y covarianza $M\Omega M'$. La probabilidad de que estas diferencias de error correlacionadas caigan por debajo de $-MV_n$ se simula mediante GHK de acuerdo al método dado en la sección 5.6.3. Este procedimiento ha sido aplicado por Hajivassiliou y Ruud (1994) y Schechter (2001).

7.4 Escalas de respuesta ordenadas

En las encuestas, a menudo se pide a los encuestados que proporcionen clasificaciones de las alternativas de diversa índole. Algunos ejemplos:

¿En qué medida crees que el presidente está haciendo un buen trabajo? Marca una opción:

1. Muy buen trabajo
2. Buen trabajo
3. Ni bueno ni malo
4. Mal trabajo
5. Muy mal trabajo

¿En qué medida te gusta este libro? Valora el libro de 1 a 7, donde 1 es lo peor que jamás hayas leído (aparte de *Los puentes de Madison*, por supuesto) y 7 es lo mejor

1 2 3 4 5 6 7

¿Qué probabilidad existe de que compres un ordenador nuevo este año?

1. Nada probable

2. Algo probable
3. Muy probable

La principal característica común de estas preguntas, desde la perspectiva del modelo de datos, es que las posibles respuestas están ordenadas. Una calificación de un libro de 6 es superior a 5, que es superior a 4, y una calificación del presidente de "muy mal trabajo" es peor que "mal trabajo", que es peor que "ni bueno ni malo". Podría especificarse un modelo logit estándar con cada respuesta potencial como una alternativa. Sin embargo, la hipótesis del modelo logit de errores independientes para cada alternativa es incompatible con el hecho de que las alternativas tengan un orden: con alternativas ordenadas, una alternativa es más similar a las alternativas próximas en la escala y menos similar a las alternativas más alejadas. La naturaleza ordenada podría ser manejada mediante la especificación de un logit jerárquico, un logit mixto o un modelo probit que represente el patrón de similitud y disimilitud entre las alternativas. Por ejemplo, podría estimarse un modelo probit con correlación entre alternativas, siendo la correlación existente entre 2 y 3 mayor que la existente entre 1 y 3, y la correlación entre 1 y 2 mayor que la existente entre 1 y 3. Sin embargo, tal especificación, aunque pueda proporcionar buenos resultados, no se ajusta realmente a la estructura de los datos. Recordemos que la formulación tradicional para estos modelos se inicia con una especificación de la utilidad asociada con cada alternativa. Para la pregunta de valoración sobre el trabajo del presidente, la formulación asumiría que hay cinco utilidades, una para cada posible respuesta, y que la persona está eligiendo entre las alternativas de 1 a 5 aquella que tiene la mayor utilidad. Si bien es posible pensar en el proceso de decisión de esta manera (y el modelo resultante probablemente proporcionará resultados útiles), no es una forma muy natural de pensar en la decisión del respondiente.

Una representación más natural del proceso de decisión es pensar que el respondiente tiene un cierto nivel de utilidad u opinión asociado con el objeto de la pregunta, y que está respondiendo la pregunta en base a lo grande que es esta utilidad. Por ejemplo, sobre la pregunta relativa al presidente, la siguiente formulación parece representar mejor el proceso de decisión. Supongamos que el encuestado tiene una opinión sobre lo bien que lo está haciendo el presidente. Esta opinión está representada en una variable (no observable) que etiquetamos U, donde los niveles superiores de U significan que la persona piensa que el presidente está haciendo un buen trabajo y los niveles más bajos significan que piensa que el presidente está haciendo un mal trabajo. Al responder a la pregunta, a la persona se le pide que exprese esta opinión seleccionando una de cinco categorías posibles: "muy buen trabajo", "buen trabajo" y así sucesivamente. Es decir, a pesar de que la opinión de la persona U puede tomar muchos niveles diferentes que representan diferentes niveles de agrado o desagrado con el trabajo que el presidente está haciendo, la pregunta sólo permite cinco posibles respuestas. La persona elige una respuesta con base al nivel de su U. Si U está por encima de cierto límite, que denominamos k_1, el entrevistado elige la respuesta "muy buen trabajo". Si U está por debajo de k_1 pero por encima de otro umbral k_2, entonces responderá "buen trabajo." Y así sucesivamente. La decisión se representa como

- "Muy buen trabajo" si $U > k_1$
- "Buen trabajo" si $k_1 > U > k_2$
- "Ni bueno ni malo" si $k_2 > U > k_3$
- "Mal trabajo" si $k_3 > U > k_4$
- "Muy mal trabajo" si $k_4 > U$.

El investigador observa algunos de los factores que se relacionan con la opinión del entrevistado, como la afiliación política de la persona, los ingresos, etc. Sin embargo, otros factores que afectan a la opinión de la persona no pueden ser observados. Descomponemos U en componentes observados y no observados: $U = \beta' x + \varepsilon$. Como de costumbre, los factores no observados ε se consideran aleatorios. Su distribución determina la probabilidad de las cinco respuestas posibles a la pregunta.

La figura 7.1 ilustra la situación. U se distribuye alrededor de $\beta' x$ con una forma de distribución que sigue la distribución de ε. Hay unos puntos de corte para las posibles respuestas: $k_1, \ldots, k_4$. La probabilidad de que la persona responda "muy mal trabajo" es la probabilidad de que U sea menor a k_4, que es la zona correspondiente a la cola izquierda de la distribución. La probabilidad de que la persona diga "mal trabajo" es la probabilidad de que U esté por encima de k_4, indicando que no piensa que el trabajo sea muy malo, pero está por debajo k_3. Esta probabilidad es el área entre k_4 y k_3.

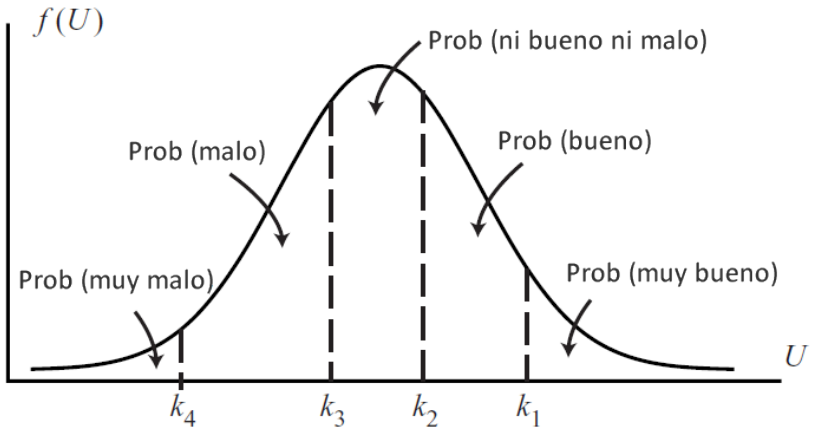


Figura 7.1 Distribución de la opinión acerca del trabajo del presidente

Una vez se especifica una distribución para ε, las probabilidades se pueden calcular exactamente. Por simplicidad, supongamos que ε se distribuye con densidad logística, lo que significa que la distribución acumulativa de ε es $F(\varepsilon) = \exp(\varepsilon)/(1 + \exp(\varepsilon))$. La probabilidad de la respuesta "muy mal trabajo" es entonces

$$Prob(\text{"muy mal trabajo"}) = Prob(U < k_4)$$

$$= Prob(\beta' x + \varepsilon < k_4)$$

$$= Prob(\varepsilon < k_4 - \beta' x)$$

$$= \frac{e^{k_4 - \beta' x}}{1 + e^{k_4 - \beta' x}}.$$

La probabilidad de "mal trabajo" es

$$Prob(\text{"mal trabajo"}) = Prob(k_4 < U < k_3)$$

$$= Prob(k_4 < \beta' x + \varepsilon < k_3)$$

$$= Prob(k_4 - \beta' x < \varepsilon < k_3 - \beta' x)$$

$$= Prob(\varepsilon < k_3 - \beta' x) - Prob(\varepsilon < k_4 - \beta' x)$$

$$= \frac{e^{k_3 - \beta' x}}{1 + e^{k_3 - \beta' x}} - \frac{e^{k_4 - \beta' x}}{1 + e^{k_4 - \beta' x}}.$$

Las probabilidades para el resto de respuestas se obtienen de forma análoga. Las probabilidades entran en la función log-verosimilitud como de costumbre, y la maximización de dicha función de verosimilitud proporciona las estimaciones de los parámetros. Observe que los parámetros estimados son β, que informa del efecto de las variables explicativas sobre la opinión que tiene la gente sobre el presidente, así como los puntos de corte $k_1, \ldots, k_4$.

Este modelo se denomina logit ordenado (*ordered logit*), ya que utiliza la distribución logística sobre alternativas ordenadas. Desafortunadamente, los modelos logit jerárquicos en ocasiones han sido llamados logits ordenados; esta nomenclatura causa confusión y esperamos que se evite en el futuro.

Observe que las probabilidades del modelo logit ordenado incorporan la fórmula logit binaria. Esta similitud con el logit binario es sólo incidental: la formulación tradicional de un logit binario especifica dos alternativas con una utilidad para cada una, mientras que el modelo logit ordenado tiene una única utilidad con múltiples alternativas que representan el nivel de esa utilidad. La similitud en la fórmula surge del hecho de que si dos variables aleatorias son tipo valor extremo iid, su diferencia sigue una distribución logística. Por lo tanto, asumir que en un logit binario ambas utilidades son de valor extremo iid es equivalente a asumir que la diferencia en las utilidades se distribuye de forma logística, la misma distribución de la utilidad del modelo logit ordenado.

Un modelo similar se obtiene bajo el supuesto de que ε se distribuye de forma normal estándar en lugar de logística (Zavoina y McKelvey, 1975). La única diferencia surge por el hecho de la fórmula logit binaria se sustituye por la distribución normal estándar acumulativa. Es decir,

$$Prob(\text{"}muy\ mal\ trabajo\text{"}) = Prob(\varepsilon < k_4 - \beta' x)$$

$$= \Phi(k_4 - \beta' x)$$

y

$$Prob(\text{"}mal\ trabajo\text{"}) = Prob(\varepsilon < k_3 - \beta' x) - Prob(\varepsilon < k_4 - \beta' x)$$

$$= \Phi(k_3 - \beta' x) - \Phi(k_4 - \beta' x),$$

donde Φ es la función normal acumulativa estándar. Este modelo se denomina probit ordenado (*ordered probit*). Existe software para manejar logits y probits ordenados en muchos paquetes comerciales.

El investigador puede pensar que los parámetros varían al azar en la población. En ese caso, se puede especificar una versión mixta del modelo, como hace Bhat (1999). Sea $g(\beta|\theta)$ la densidad de β. En este caso, las probabilidades del modelo logit ordenado mixto son simplemente las probabilidades del logit ordenado integradas sobre la densidad $g(\cdot)$. Por ejemplo

$$Prob(\text{"}muy\ mal\ trabajo\text{"}) = \int \left(\frac{e^{k_4 - \beta' x}}{1 + e^{k_4 - \beta' x}} \right) g(\beta|\theta) d\beta$$

y

$$Prob(\text{"}mal\ trabajo\text{"}) = \int \left(\frac{e^{k_3 - \beta' x}}{1 + e^{k_3 - \beta' x}} - \frac{e^{k_4 - \beta' x}}{1 + e^{k_4 - \beta' x}} \right) g(\beta|\theta) d\beta,$$

y así sucesivamente. Estas probabilidades se simulan de la misma manera como se hace para logits mixtos, mediante la extracción de valores β al azar de $g(\cdot)$, calculando la probabilidad del logit ordenado para cada valor y promediando los resultados. El probit ordenado mixto se obtiene de manera similar.

7.4.1 Escalas de respuesta ordenadas múltiples

Las respuestas de los encuestados a diferentes preguntas suelen estar relacionadas. Por ejemplo, la calificación que una persona da sobre lo bien que el presidente lo está haciendo está probablemente relacionada con la calificación que la persona da sobre lo bien que la economía está yendo. El investigador desearía incorporar en el análisis el hecho de que las respuestas están relacionadas. Para ser concretos, supongamos que se pide a los encuestados que califiquen tanto al presidente como a la economía en una escala de cinco puntos. Sea U la opinión del encuestado sobre la tarea que el presidente está haciendo, y sea W la evaluación del entrevistado sobre la economía. Cada una de estas evaluaciones se puede descomponer en factores observados y no observados: $\mathrm{U} = \beta'\mathrm{x} + \varepsilon$ y $\mathrm{W} = \alpha'\mathrm{z} + \mu$. En la medida en que las evaluaciones estén relacionadas debido a factores observados, las mismas variables pueden incluirse en x y z. Para contemplar la posibilidad de que las evaluaciones estén relacionadas debido a factores no observados, especificamos ε y μ para que se distribuyan conjuntamente normales con correlación ρ (y con varianzas unitarias a efectos de normalización). Denotemos los puntos de corte de U como $k_1, \dots, k_4$ como antes, y los puntos de corte de W como $c_1, \dots, c_4$. Queremos obtener la probabilidad de cada posible combinación de respuestas a las dos preguntas.

La probabilidad de que una persona diga que el presidente está haciendo un “muy mal trabajo” y al mismo tiempo también diga que la economía está yendo “muy mal” se obtiene de la siguiente manera:

$$Prob(Presidente \text{ "muy mal" y Economía "}muy\ mal\text{"})$$

$$= Prob(U < k_4\ y\ W < c_4)$$

$$= Prob(\varepsilon < k_4 - \beta' x \quad y \quad \mu < c_4 - \alpha' z)$$

$$= Prob(\varepsilon < k_4 - \beta' x\ ) \times Prob(\mu < c_4 - \alpha' z | \varepsilon < k_4 - \beta' x)$$

Del mismo modo, la probabilidad de una calificación "muy mala" para el presidente y “buena” para la economía es

$$Prob(Presidente \text{ "muy mal" y Economía "}bien\text{"})$$

$$= Prob(U < k_4\ y\ c_2 < W < c_1)$$

$$= Prob(\varepsilon < k_4 - \beta' x\ y\ c_2 - \alpha' z < \mu < c_1 - \alpha' z)$$

$$= Prob(\varepsilon < k_4 - \beta' x\ ) \times Prob(c_2 - \alpha' z < \mu < c_1 - \alpha' z \mid \varepsilon < k_4 - \beta' x)$$

Las probabilidades para las demás combinaciones se obtienen del mismo modo, y la generalización a más de dos preguntas relacionadas es directa. Este modelo se denomina probit ordenado multivariado o multirrespuesta (*multivariate or multiresponse ordered probit*). Las probabilidades pueden simularse mediante GHK de manera similar a como se describe en el capítulo 5. La explicación en el capítulo 5 supone que el truncamiento de la distribución normal conjunta se produce sólo por un lado (ya que para

un probit estándar la probabilidad que se está calculando es la probabilidad de que todas las diferencias de utilidad estén por debajo de cero, que es un truncamiento de la parte superior), mientras que las probabilidades para un probit ordenado multivariado se truncan por ambos lados (como es el caso de la segunda probabilidad indicada anteriormente). Sin embargo, la lógica es la misma. Los lectores interesados pueden referirse a Hajivassiliou y. Ruud (1994) para un tratamiento explícito de GHK con truncamiento por los dos lados.

7.5 Valoración contingente

En algunas encuestas, se pide a los entrevistados que expresen sus opiniones en relación a un número específico que el entrevistador indica. Por ejemplo, el entrevistador podría preguntar: "Considera un proyecto destinado a proteger los peces de determinados ríos de Montana. ¿Estarías dispuesto a gastar $50 en saber que los peces de estos ríos están a salvo?". Esta tipo de pregunta en ocasiones se acompaña de una segunda cuestión que depende de la respuesta del encuestado a la primera pregunta. Por ejemplo, si la persona dijo "sí" a la pregunta anterior, el entrevistador podría seguir preguntando: "¿Y qué tal $75? ¿Estarías dispuesto a pagar $75?". Si la persona contestó "no" a la primera pregunta, indicando que no estaba dispuesto a pagar $50, el entrevistador podría seguir con "¿Estarías dispuesto a pagar $25?".

Este tipo de preguntas se utilizan a menudo en estudios ambientales, en los que la inexistencia de mercados relativos a la calidad del medio ambiente impide la valoración de los recursos naturales por medio de datos revelados (reales); los artículos editados por Hausman (1993) proporcionan una revisión y una crítica de este procedimiento, que a menudo se denomina "valoración contingente" ("*contingent valuation"*). Cuando únicamente se pregunta una cuestión, como por ejemplo si la persona está dispuesta a pagar $50, el método se llama de límite simple (*single-bounded*), dado que la respuesta de la persona informa sobre un límite de su predisposición a pagar. Si la persona responde "sí", el investigador sabe que su verdadera predisposición a pagar es por lo menos de $50, pero no sabe cuánto más puede ser. Si la persona responde "no", el investigador sabe que la predisposición a pagar de la persona es menor a $50. Ejemplos de estudios que utilizan métodos de límite simple los proporcionan Cameron y James (1987) y Cameron (1988).

Cuando se realiza una segunda pregunta, el método se denomina *límite doble ("double-bounded"*). Si la persona dice que está dispuesta a pagar $50, pero no $75, el investigador averigua que su verdadera disposición a pagar está entre $50 y $75, es decir, está delimitada por ambos lados. Si la persona afirma no estar dispuesta a pagar $50, pero sí está dispuesta a pagar $25, su disposición a pagar se sabe que está entre $25 y $50. Por supuesto, incluso con un método de límite doble, la disposición a pagar de algunos de los encuestados sólo queda limitada por un lado, como la de una persona que diga que está dispuesta a pagar $50 y también $75. Ejemplos de este enfoque los proporcionan Hanemann et al. (1991), Cameron y Quiggin (1994), y Cai et al. (1998).

La cifra que se utiliza como referencia (es decir, los $50 en nuestro ejemplo) se modifica para diferentes encuestados. Para estimar la distribución de la predisposición a pagar se utilizan las respuestas de una muestra de personas. El procedimiento de estimación está estrechamente relacionado con el que acabamos de describir para logits y probits ordenados, a excepción de que los puntos de corte vienen dados por el diseño del cuestionario y no se estiman como parámetros. Describimos el procedimiento a continuación.

Sea W_n un parámetro que representa la verdadera predisposición a pagar de la persona n. W_n varía entre personas con una distribución $f(W|\theta)$, donde θ son los parámetros de la distribución, tales como la media y la varianza. El objetivo del investigador es estimar estos parámetros poblacionales. Supongamos que el investigador diseña un cuestionario con un enfoque de límite simple, facilitando un valor de referencia diferente para diferentes encuestados. Denominemos el valor de referencia que se le da a la persona n como k_n. La persona responde a la pregunta con un "sí" si $\mathrm{W}_n > k_n$ y con un "no"

en caso contrario. El investigador asume que $W_n$se distribuye normalmente en la población con media $\overline{W}$ y varianza σ^2.

La probabilidad de "sí" es $\text{Prob}(W_n > k_n) = 1 - \text{Prob}(W_n < k_n) = 1 - \Phi((k_n - \overline{W})/\sigma)$, y la probabilidad de "no" es $\Phi((k_n - \overline{W})/\sigma)$, donde $\Phi(\cdot)$ es la función normal acumulativa estándar. La función log-verosimilitud resulta en este caso $\sum_n (y_n \ln(1 - \Phi((k_n - \overline{W})/\sigma)) + (1 - y_n)\ln(\Phi((k_n - \overline{W})/\sigma)))$, donde $y_n = 1$ si la persona n ha dicho "sí" y 0 en caso contrario. Maximizar esta función proporciona estimaciones para $\overline{W}$ y σ.

Un procedimiento similar se usa si el investigador diseña un cuestionario con límite doble. Denominemos k_{nu} (u en referencia a "*upper*") al valor de referencia de la segunda pregunta en caso de que el encuestado haya respondido "sí" a la primera pregunta, donde $k_{nu} > k_n$, y denominemos k_{nl} (l en referencia a "*lower*") al segundo valor de referencia si la persona inicialmente ha respondido "no", donde $k_{nl} < k_n$. Hay cuatro posibles secuencias de respuestas a las dos preguntas. Las probabilidades para estas secuencias se ilustran en la figura 7.2 y vienen dadas por

- Primero "no", luego "no": $P = \text{Prob}(W_n < k_{nl}) = \Phi((k_{nl} - \overline{W})/\sigma)$.
- Primero "no", luego "sí": $P = \text{Prob}(k_{nl} < W_n < k_n) = \Phi((k_n - \overline{W})/\sigma) - \Phi((k_{nl} - \overline{W})/\sigma)$.
- Primero "sí", luego "no": $P = \text{Prob}(k_n < W_n < k_{nu}) = \Phi((k_{nu} - \overline{W})/\sigma) - \Phi((k_n - \overline{W})/\sigma)$.
- Primero "sí", luego "sí": $P = \text{Prob}(W_n > k_{nu}) = 1 - \Phi((k_{nu} - \overline{W})/\sigma)$.

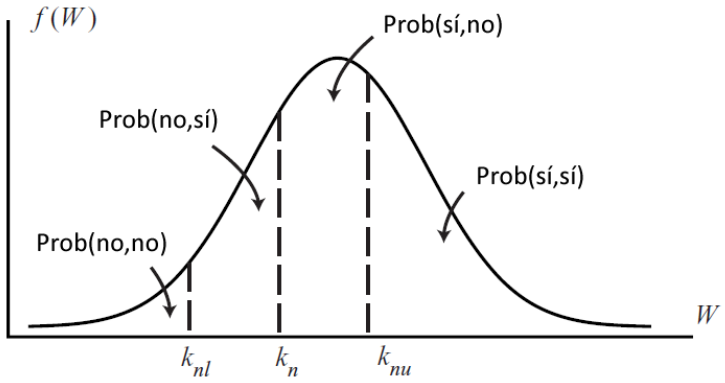


Figura 7.2. Distribución de la predisposición a pagar.

Estas probabilidades entran en la función log-verosimilitud que maximizamos para obtener estimaciones de $\overline{W}$ y σ. Por supuesto, otras distribuciones se podrían utilizar en lugar de la normal. Log-normal es una opción interesante si el investigador supone que todas las personas tienen una predisposición positiva a pagar. Asimismo, el investigador podría especificar una distribución que tuviese una masa en cero para representar la proporción de personas que no están predispuestas a pagar nada y una log-normal para la parte restante. La generalización a múltiples dimensiones es directa, para reflejar, por ejemplo, que la predisposición de la gente a pagar por un paquete ambiental también podría estar relacionada con su predisposición a pagar por otro. Al igual que con el probit ordenado multivariado, el simulador GHK es muy útil cuando se supone que los múltiples valores se distribuirán de forma normal conjunta.

7.6 Modelos mixtos

Hemos visto los modelos logit mixto y logit ordenado mixto. Por supuesto, se pueden desarrollar modelos mixtos de todo tipo utilizando la misma lógica. Cualquier modelo cuyas probabilidades se puedan escribir como una función de unos parámetros también se pueden hacer mixtos al permitir que los parámetros sean aleatorios e integrando la función resultante sobre la distribución de probabilidad

de los parámetros (Greene, 2001). Las probabilidades se simulan mediante la extracción de valores al azar de la distribución, calculando la función para cada valor y promediando el resultado. En la siguiente sección facilitamos dos ejemplos, pero los investigadores inevitablemente necesitarán desarrollar otros que satisfagan las necesidades de sus proyectos particulares, como el uso que hace Bhat (1999) del logit ordenado mixto.

7.6.1 Logit jerárquico mixto

El modelo logit mixto no muestra la propiedad de independencia de alternativas irrelevantes como sí lo hace logit, y puede aproximar cualquier patrón de sustitución mediante una especificación adecuada de las variables y de la distribución de mezcla. Este hecho ha llevado a algunos a pensar que el desarrollo de este modelo suprimía la necesidad de usar modelos logit jerárquicos. Es posible estimar un modelo logit mixto que proporcione patrones de sustitución y de correlación análogos a los de un logit jerárquico. Por ejemplo, considere un logit jerárquico con dos nidos de alternativas etiquetados como A y B. Siempre y cuando los coeficientes log-suma estén entre 0 y 1, la sustitución dentro de cada nido será mayor que la sustitución entre nidos. Este patrón de sustitución puede ser representado en un modelo logit mixto mediante la especificación de una variable indicadora para cada nido y permitiendo que los coeficientes de las variables indicadoras sean aleatorios (restringiendo, a efectos de identificación, que las medias sean cero si se incluye un conjunto completo de constantes específicas de alternativa, y que las dos varianzas sean iguales).

Aunque es posible especificar un modelo logit mixto de esta manera, al hacerlo perdemos la perspectiva del uso de la simulación. Como ya vimos en el Capítulo 1, la simulación se utiliza como una forma de aproximar las integrales cuando no existe una forma cerrada que permita el cálculo analítico. La integración analítica siempre es más precisa que la simulación y se debe utilizar siempre que sea posible, a menos que haya una razón de peso para hacer lo contrario. El uso de un logit mixto para representar patrones de sustitución propios de un logit jerárquico, si bien es posible, reemplaza la forma cerrada de la integral del modelo logit jerárquico por una integral que necesita ser simulada. Desde una perspectiva numérica, esta sustitución sólo puede reducir la precisión. Las únicas posibles ventajas del logit mixto en este contexto son que (1) puede ser más fácil para el investigador probar numerosas estructuras de anidación, incluyendo nidos superpuestos, dentro de un modelo logit mixto que en un logit jerárquico, y (2) el investigador podría haber especificado otros coeficientes como aleatorios, por lo que ya se está utilizando un modelo logit mixto.

La segunda razón sugiere la posibilidad de definir un logit jerárquico mixto. Supongamos que el investigador cree que algunos de los coeficientes del modelo son aleatorios, y también que, condicionados a estos coeficientes, los factores no observados se correlacionan entre alternativas de una forma que puede ser representada por un logit jerárquico. Para representar esta situación podemos especificar un modelo logit jerárquico mixto. Condicionadas a los coeficientes que entran en la utilidad, las probabilidades de elección serían las propias de un logit jerárquico, que tienen una forma cerrada y pueden ser calculadas exactamente. La probabilidad no condicionada pasaría a ser la fórmula logit jerárquica integrada sobre la distribución de los coeficientes aleatorios. Es posible modificar el software existente destinado a estimar un modelo logit mixto, simplemente localizando la fórmula logit dentro del código y reemplazándola por la fórmula logit jerárquica apropiada. La experiencia indica que la maximización de la función de verosimilitud para logits jerárquicos no mixtos es a menudo difícil numéricamente, por lo que hacer el modelo mixto agravará esta dificultad. La estimación bayesiana jerárquica (Capítulo 12) podría resultar particularmente útil en esta situación, ya que no implica la maximización de la función de verosimilitud.

7.6.2 Probit mixto

Una limitación de los modelos probit y, de hecho, la característica que los define, es que todos los términos aleatorios entran en la utilidad linealmente y se distribuyen aleatoriamente de tal manera que la utilidad misma se distribuye normalmente. Esta limitación se puede eliminar mediante la especificación de un probit mixto. Supongamos que algunos términos aleatorios entran de forma no lineal o no se distribuyen aleatoriamente, pero que la utilidad condicionada a estos términos sí se distribuye normalmente. Por ejemplo, un coeficiente de precio podría ser log-normal para asegurar que es negativo para todo el mundo, y sin embargo, todos los demás coeficientes podrían ser fijos o normales, y los términos de error finales conjuntamente normales. Un modelo probit mixto es apropiado para una especificación así. Las probabilidades de elección condicionadas al coeficiente de precio seguirían la fórmula probit estándar. Las probabilidades no condicionadas serían la integral de esta fórmula probit sobre la distribución del coeficiente de precio. Para aproximar estas probabilidades necesitaríamos un proceso de simulación de dos niveles: (1) se extrae un valor al azar del coeficiente de precio y (2) para este valor, el simulador GHK o cualquier otro simulador probit se utiliza para aproximar la probabilidad de elección condicionada. Este proceso se repite muchas veces y se promedian los resultados.

Es de esperar que el modelo probit mixto requiera tiempos de ejecución largos, ya que el simulador GHK debe ser calculado para cada valor extraído al azar del coeficiente de precio. Sin embargo, es posible reducir el número de extracciones al azar en el simulador GHK, ya que promediar entre extracciones al azar del coeficiente de precio reduce la varianza generada por el simulador GHK. En principio, el simulador GHK puede basarse en sólo un valor extraído al azar por cada valor extraído al azar del coeficiente de precio. En la práctica, puede que sea aconsejable utilizar más de un valor extraído, pero muchos menos de los que se usarían en un probit no mixto.

El modelo probit mixto proporciona al investigador una forma de evitar algunas de las dificultades prácticas que pueden surgir con un modelo logit mixto. Por ejemplo, para representar heterocedasticidad pura (es decir, una varianza diferente para la utilidad de cada alternativa) o un patrón de correlación fija entre alternativas (es decir, una matriz de covarianza que no depende de las variables), a menudo puede ser más fácil estimar un probit en lugar de especificar numerosos componentes de error dentro de un modelo logit mixto. Como destacó Ben-Akiva et al. (2001), la especificación de la covarianza y la heterocedasticidad puede ser más compleja en un modelo logit mixto que en un probit, porque en el primer caso necesariamente deben añadirse los términos valor extremo iid a cualesquiera otros elementos aleatorios que especifique el investigador. Probit es una especificación más natural en estas situaciones. Sin embargo, si el investigador quiere incluir algunos términos aleatorios no normales, no es posible usar un probit no mixto. El probit mixto permite al investigador incluir términos no normales, manteniendo la simplicidad de la representación que hace probit de la covarianza fija para errores aditivos. Conceptualmente, la especificación y el procedimiento de estimación son sencillos. El único inconveniente es el tiempo de cálculo adicional, algo que se vuelve menos relevante a medida que las computadoras se vuelven más rápidas.

7.7 Optimización dinámica

En los capítulos anteriores hemos examinado ciertos tipos de dinámicas, por las que las elecciones en un período afectan a las elecciones en otro período. Por ejemplo, hemos descrito cómo una variable dependiente diferida puede ser incluida para capturar la inercia del comportamiento o la búsqueda de variedad. Estos casos de uso sugieren un ámbito mucho más amplio en relación a estas dinámicas de lo que habíamos realmente considerado. En particular: si elecciones pasadas afectan elecciones presentes, entonces elecciones presentes afectan elecciones futuras, y un decisor que sea consciente de este hecho tomará estos efectos futuros en consideración. Un vínculo desde el pasado hasta el presente necesariamente implica un vínculo desde el presente hasta el futuro.

En muchas situaciones, las elecciones que hace una persona en un momento de su vida tienen una profunda influencia en las opciones que estarán a su disposición en el futuro. Ir a la universidad, aunque caro y a veces irritante, mejora las posibilidades futuras de empleo. Ahorrar dinero ahora permite a una persona comprar cosas más tarde que de otra manera no sería capaz de pagar. Ir al gimnasio hoy significa nos lo podemos saltar mañana. La mayoría de nosotros tenemos en cuenta los efectos futuros cuando elegimos entre alternativas presentes.

La pregunta que se plantea es: ¿cómo puede representarse un comportamiento como éste en modelos de elección discreta? En general, la situación se puede describir de la siguiente manera. Una persona hace una serie de elecciones a lo largo del tiempo. La alternativa elegida en un período afecta a los atributos y a la disponibilidad de alternativas en el futuro. A veces los efectos futuros no se conocen completamente, o dependen de factores que aún no han ocurrido (como la situación futura de la economía). Sin embargo, la persona sabe que en el futuro maximizará la utilidad entre las alternativas que estén disponibles en ese momento, de acuerdo a las condiciones que prevalezcan en ese momento. Este conocimiento le permite elegir la alternativa en el período actual que maximiza su utilidad esperada en los períodos actuales y futuros. El investigador reconoce que el decisor actúa de esta manera, pero no observa todo lo que el decisor está teniendo en consideración en los períodos actuales y futuros. Como de costumbre, la probabilidad de elección es una integral del comportamiento del decisor sobre todos los posibles valores de la factores que el investigador no observa.

En esta sección especificaremos modelos que incorporan las consecuencias futuras de decisiones actuales. Para estos modelos, vamos a suponer que el decisor es totalmente racional en el sentido de que optimiza sus decisiones a la perfección en cada período de tiempo, dada la información que está disponible para él en ese momento y dado que sabe que va a actuar de manera óptima en el futuro, cuando la información futura le sea revelada. Los procedimientos para modelar estas decisiones fueron desarrollados en primer lugar para diversas aplicaciones prácticas, por ejemplo, por Wolpin (1984) sobre la fertilidad femenina, Pakes (1986) sobre las opciones de patentes, Wolpin (1987) sobre la búsqueda de empleo, Rust (1987) sobre la sustitución de motores, Berkovec y Stern (1991) sobre la jubilación, y otros. Eckstein y Wolpin (1989) proporcionan una excelente visión de estas primeras contribuciones. Los esfuerzos de los trabajos más recientes se han dirigido principalmente hacia la solución de algunos de los problemas computacionales que pueden surgir en estos modelos, como veremos a continuación.

Antes de embarcarse en esta tarea, es importante mantener en perspectiva el concepto de racionalidad. Un modelo de toma de decisiones racionales a lo largo del tiempo no necesariamente representa el comportamiento con mayor precisión que un modelo de comportamiento miope, en el que el decisor no tiene en cuenta las consecuencias futuras. De hecho, la realidad en una situación concreta puede estar entre estos dos extremos: los decisores podrían estar actuando de una forma que no sea ni completamente miope ni completamente racional. Como veremos, el comportamiento verdaderamente optimizador es muy complejo. La personas pueden emprender un comportamiento que sólo es aproximadamente óptimo simplemente porque ellas (nosotros) no pueden encontrar la auténtica forma óptima de proceder. Visto desde otro punto de vista, se podría incluso argumentar que las personas siempre optimizan cuando el ámbito de la optimización se amplía suficientemente. Por ejemplo, algunas reglas de oro u otros comportamientos que parece que sólo aproximan un comportamiento óptimo, podrían resultar realmente óptimos si consideramos los costos de la propia optimización.

Los conceptos y procedimientos que se han desarrollado para examinar el comportamiento optimizador nos llevan, de una forma diferente, a otros tipos de comportamiento que reconocen los efectos futuros de las decisiones actuales. Es más, el investigador a menudo puede poner a prueba representaciones alternativas del comportamiento. El comportamiento miope casi siempre aparece como una restricción comprobable en un modelo totalmente racional, concretamente, un coeficiente cero para la variable que captura los efectos futuros. A veces, el modelo racional estándar es una restricción en un modelo

supuestamente no racional. Por ejemplo, O'Donoghue y Rabin (1999), entre otros, argumentan que las personas son incoherentes a lo largo del tiempo: cuando es lunes, sopesamos los beneficios y costos que vendrán, por ejemplo, el miércoles, sólo ligeramente más que los que llegarán el jueves, y sin embargo, cuando el miércoles llega, sopesamos los beneficios y costos del miércoles (hoy) mucho más que los del jueves. Básicamente, tenemos un sesgo hacia el presente. El modelo racional estándar, donde se utiliza la misma tasa de descuento entre dos períodos independientemente de si la persona se encuentra en uno de esos períodos, constituye una restricción en el modelo incoherente en el tiempo.

Los conceptos en esta área de análisis son más sencillos que la notación empleada. Para desarrollar los conceptos con un mínimo de notación, vamos a empezar con un modelo de dos períodos de tiempo en el que el decisor conoce el efecto exacto que sus elecciones en el primer período tienen en las alternativas y utilidades del segundo período. Ampliaremos posteriormente el modelo a más períodos y a situaciones en las que el decisor se enfrenta a la incertidumbre de los futuros efectos.

7.7.1 Dos períodos, sin incertidumbre sobre efectos futuros

Para facilitar una explicación lo más concreta posible, considere la elección que un estudiante de secundaria hace sobre si debe ir o no a la universidad. La elección puede ser examinada en un contexto de dos períodos: los años universitarios y los años post-universidad. En el primer período, el estudiante va a la universidad o no va. A pesar de que este período lo llamamos los años universitarios, el estudiante no tiene por qué haber ido a la universidad, sino que puede haber elegido trabajar en lugar de estudiar. En el segundo período, el estudiante escoge entre los trabajos que están a su disposición en ese momento. Ir a la universidad durante los años universitarios implica menos ingresos durante ese período, pero mejores opciones de trabajo en los años posteriores a la universidad. U_{1C} (c en referencia a *"college"*) es la utilidad que el estudiante obtiene en el período 1 al ir a la universidad y U_{1W} (w en referencia a *"work"*) es la utilidad que obtiene en el primer período si trabaja en lugar de ir a la universidad. Si el estudiante fuera miope, elegiría la universidad sólo si $U_{1C} > U_{1W}$. Sin embargo, suponemos que no es miope. Para el segundo período, J denota el conjunto de todos los puestos de trabajo posibles. La utilidad del trabajo j en el período 2 es U_{2j}^{C} si el estudiante fue a la universidad y U_{2j}^{W} si trabajó en el primer período. La utilidad de un trabajo depende del salario que la persona recibe así como de otros factores. Para muchos puestos de trabajo, las personas con un título universitario reciben salarios más altos y gozan de mayor autonomía y responsabilidad. Para estos trabajos, $\mathrm{U}_{2j}^{C} > \mathrm{U}_{2j}^{W}$. Sin embargo, trabajar durante el primer período proporciona experiencia laboral que da acceso a mayores salarios y responsabilidades que un título universitario para algunos trabajos; para estos trabajos, $\mathrm{U}_{2j}^{W} > \mathrm{U}_{2j}^{C}$. Un trabajo que no está disponible se representa teniendo una utilidad infinitamente negativa. Por ejemplo, si el trabajo j está disponible sólo para los titulados universitarios, entonces $\mathrm{U}_{2j}^{W} = -\infty$.

¿Cómo decidirá el estudiante de secundaria si va a la universidad o no? Asumimos por el momento que el estudiante conoce U_{2j}^{C} y U_{2j}^{W} para todos los trabajos j ∈ J en el momento de decidir si va a ir a la universidad en el primer período. Esto es, el estudiante tiene un conocimiento perfecto de sus futuras opciones sea cual sea su elección en el primer período. Posteriormente consideraremos cómo cambia el proceso de decisión cuando el estudiante desconoce las utilidades futuras. El estudiante sabe que cuando el segundo período llegue, escogerá el trabajo que proporcione mayor utilidad. Es decir, él sabe en el primer período que la utilidad que va a obtener en el segundo período, si elige la universidad en el primer período, es el máximo de las utilidades U_{2j}^{C} todos los posibles trabajos. Denominamos esta utilidad como $\mathrm{U}_{2}^{C} = \max_j(\mathrm{U}_{2j}^{C})$. El estudiante, por lo tanto, se da cuenta de que si elige la universidad en el primer período, la utilidad total sobre ambos períodos será

$$TU_C = U_{1C} + \lambda U_2^C$$

$$= U_{1C} + \lambda\, max_j\left(U_{2j}^{C}\right),$$

Donde λ refleja el peso relativo que las utilidades en los dos períodos tienen en el proceso de decisión del estudiante. Dada la forma en que hemos definido los períodos de tiempo, λ incorpora la duración relativa de los intervalos de tiempo de cada período, así como la tendencia a subestimar la utilidad futura en relación a la utilidad presente. Es por ello que λ puede ser superior a uno, incluso con el descuento que se aplica a la utilidad futura, si el segundo período representa por ejemplo cuarenta años, mientras que el primer período es de únicamente cuatro años. El comportamiento miope se representa como $\lambda = 0$ (es decir, no contabilizar la utilidad futura).

Aplicamos la misma lógica a la opción de trabajar en el primer período en vez de ir a la universidad. El estudiante sabe que va a elegir el trabajo que ofrezca mayor utilidad, por lo que $\mathrm{U}_2^W = \max_j(\mathrm{U}_{2j}^W)$, y la utilidad total, durante ambos períodos de tiempo, que proporciona la opción de trabajar en el primer período es

$$TU_W = U_{1W} + \lambda U_2^W$$

$$= U_{1W} + \lambda\, max_j\left(U_{2j}^{W}\right).$$

El estudiante elige ir a la universidad si $TU_{\mathrm{C}} > TU_W$ y en caso contrario, opta por trabajar en el primer período.

Esto completa la descripción del comportamiento del decisor. Ahora vamos a por el investigador. Como siempre, el investigador observa sólo algunos de los factores que afectan a la utilidad que percibe el estudiante. Cada utilidad en cada período de tiempo se descompone en una parte observada y otra parte no observada:

$$U_{1C} = V_{1C} + \varepsilon_{1C},$$

$$U_{1W} = V_{1W} + \varepsilon_{1W}$$

y

$$U_{2j}^{C} = V_{2j}^{C} + \varepsilon_{2j}^{C},$$

$$U_{2j}^{W} = V_{2j}^{W} + \varepsilon_{2j}^{W},$$

para todos los $j \in J$. Agrupamos todos los componentes no observados en un vector $\varepsilon = \langle \varepsilon_{1C}, \varepsilon_{1W}, \varepsilon_{2j}^{C}, \varepsilon_{2j}^{W}, \forall j \rangle$, y llamamos a la densidad de estos términos $f(\varepsilon)$. La probabilidad de que el estudiante escoja la universidad es

$$P_c = Prob(TU_C > TU_W)$$

$$= Prob\left(U_{1C} + \lambda\, max_j\left(U_{2j}^{C}\right) > U_{1W} + \lambda\, max_j\left(U_{2j}^{W}\right)\right)$$

$$= Prob\left(V_{1C} + \varepsilon_{1C} + \lambda\, max_j\left(V_{2j}^{C} + \varepsilon_{2j}^{C}\right) > V_{1W} + \varepsilon_{1W} + \lambda\, max_j\left(V_{2j}^{W} + \varepsilon_{2j}^{W}\right)\right)$$

$$= \int I\left[V_{1C} + \varepsilon_{1C} + \lambda\, max_j\left(V_{2j}^C + \varepsilon_{2j}^C\right) > V_{1W} + \varepsilon_{1W} + \lambda\, max_j\left(V_{2j}^W + \varepsilon_{2j}^W\right)\right] f(\varepsilon) d\varepsilon$$

Donde $I[\cdot]$ es una función indicadora de si la declaración entre paréntesis es verdadera.

La integral puede aproximarse a través de simulación. Para un simulador tipo aceptación-rechazo:

1. Extraiga un valor al azar de $f(\varepsilon)$, con sus componentes etiquetados como $\varepsilon_{1C}^r, \varepsilon_{2j}^{Cr}, \ldots$
2. Calcule $U_{2j}^C = V_{2j}^C + \varepsilon_{2j}^{Cr}$ para todo j, determine la utilidad más alta y etiquétela como U_2^{Cr}. De forma análoga calcule U_2^{Wr}.
3. Calcule las utilidades totales como $TU_C^r = \mathrm{V}_{1C}^r + \varepsilon_{1C}^r + \lambda \mathrm{U}_2^{Cr}$ y lo mismo para TU_{W}^r.
4. Determine si $TU_C^r > TU_{\mathrm{W}}^r$. Si es así, establezca que $\mathrm{I}^r = 1$. De lo contrario, $\mathrm{I}^r = 0$.
5. Repita los pasos 1-4 R veces. La probabilidad de elección simulada de escoger la universidad es $\widetilde{\mathrm{P}}_c = \sum_r \mathrm{I}^r / R$.

Podemos utilizar la participación conveniente del error (como se explica en la sección 1.2) para obtener un simulador suave y más preciso que el simulador de aceptación-rechazo, siempre y cuando la integral sobre los errores del primer período tenga una forma cerrada al condicionar respecto a los errores del segundo período. Supongamos por ejemplo que ε_{1C} y ε_{1W} son de tipo valor extremo iid. Denominemos a los errores del segundo período colectivamente como ε_2 con cualquier densidad $\mathrm{g}(\varepsilon_2)$. Condicionada a los errores del segundo período, la probabilidad de que el estudiante vaya a la universidad está dada por un modelo logit estándar con una variable explicativa adicional que captura el efecto futuro de la elección actual. Es decir

$$P_c = \frac{e^{V_{1C} + \lambda U_2^C(\varepsilon_2)}}{e^{V_{1C} + \lambda U_2^C(\varepsilon_2)} + e^{V_{1W} + \lambda U_2^W(\varepsilon_2)}},$$

donde $\mathrm{U}_2^C(\varepsilon_2)$ se calcula a partir de los errores del segundo período como $\mathrm{U}_2^C(\varepsilon_2) = \max_j(V_{2j}^C + \varepsilon_{2j}^C)$, y de forma similar para $\mathrm{U}_2^W(\varepsilon_2)$. La probabilidad no condicionada es la integral de esta fórmula logit sobre todos los valores posibles de los errores del segundo período:

$$P_c = \int P_c(\varepsilon_2) g(\varepsilon_2) d\, \varepsilon_2.$$

La probabilidad se simula de la siguiente manera: (1) Extraiga un valor al azar de la densidad $\mathrm{g}(\cdot)$ y etiquételo como ε_2^r. (2) Usando este valor de los errores del segundo período, calcule la utilidad que se obtendría de cada posible puesto de trabajo si la persona fue a la universidad. Es decir, calcule $U_{2j}^{Cr} = V_{2j}^C + \varepsilon_{2j}^{Cr}$ para todo j. (3) Determine el máximo de estas utilidades y etiquételo como U_2^{Cr}. Esta es la utilidad que la persona obtendría en el segundo período si fue a la universidad en el primer período, en base a este valor extraído al azar de los errores del segundo período. (4)-(5) Análogamente, calcule $U_{2j}^{Wr}\ \forall j$, y determine luego el máximo U_2^{Wr}. (6) Calcule la probabilidad de elección condicionada para este valor extraído al azar como

$$P_C^r = \frac{e^{V_{1C} + \lambda U_2^{Cr}}}{e^{V_{1C} + \lambda U_2^{Cr}} + e^{V_{1W} + \lambda U_2^{Wr}}}.$$

(7) Repita los pasos 1-6 múltiples veces, etiquetando $r = 1, \ldots, R$. (8) La probabilidad simulada es $\widetilde{\mathrm{P}}_c = \sum_r \mathrm{P}_C^r / R$.

Si los errores del segundo período son también valor extremo iid, entonces la probabilidad de aceptar un trabajo particular en el segundo período es logit estándar. La probabilidad de ir a la universidad y elegir el trabajo j es

$$P_{Cj} = \left(\int \left[\frac{e^{V_{1C}+\lambda U_2^C(\varepsilon_2)}}{e^{V_{1C}+\lambda U_2^C(\varepsilon_2)} + e^{V_{1W}+\lambda U_2^W(\varepsilon_2)}} \right] g(\varepsilon_2) d\varepsilon_2 \right) \left(\frac{e^{V_{2j}^C}}{\sum_k e^{V_{2k}^C}} \right).$$

Las probabilidades de elección para el primer período se simulan mediante la extracción de valores al azar de los errores del segundo período, como se acaba de describir, siendo g(·) la distribución de valor extremo. Sin embargo, las probabilidades para el segundo período se calculan de forma exacta. Las extracciones de valores al azar de los errores del segundo período sólo se utilizan en el cálculo de las probabilidades del primer período, donde no se integran en forma cerrada. Los errores del segundo período se integran (y se cancelan) en las probabilidades de forma cerrada del segundo período, lo que permite calcular las probabilidades del segundo período de forma exacta. La aplicación a otras distribuciones que permitan correlación entre alternativas, como GEV o normal, es sencilla. Permitir que los errores se correlacionen a lo largo del tiempo se puede lograr con una distribución normal conjunta y la simulación de las probabilidades de ambos períodos.

7.7.2 Múltiples períodos

En primer lugar expandiremos el modelo anterior a tres períodos y posteriormente generalizaremos a cualquier número de períodos. El modelo de elección de la universidad se puede extender al considerar las opciones de jubilación. Cuando una persona llega a la edad de jubilación, por lo general hay varias opciones disponibles. La persona puede seguir trabajando a tiempo completo, puede trabajar a tiempo parcial y gastar parte de sus fondos de retiro, o retirarse totalmente y cobrar de la seguridad social y tal vez una pensión. Los ingresos de la persona para estas alternativas dependen en gran medida del trabajo que la persona haya realizado y del plan de jubilación que su empleo le haya proporcionado. Tres períodos son suficientes para capturar el proceso de decisión. La persona va a la universidad o no en el primer período, elige un empleo en el segundo período y elige entre las diferentes opciones de jubilación disponibles en el tercer período. El estudiante de secundaria sabe, en el momento de decidir si va a la universidad, que esta decisión afectará a sus oportunidades de trabajo, que a su vez afectarán a sus opciones de jubilación. (Esta capacidad de predicción está empezando a parecer una carga pesada para un estudiante de secundaria).

El conjunto de alternativas disponibles en la edad de jubilación se etiqueta como S y sus elementos se indexan mediante s. En el tercer período, la utilidad que la persona obtiene de la alternativa s si fue a la universidad en el primer período y tenía un trabajo j en el segundo período es U_{3s}^{Cj}. Condicionado a estas elecciones previas, la persona elige la opción s if $U_{3s}^{Cj} > U_{3t}^{Cj}$ para todo $s \neq t$ y $s, t \in S$. Una notación y un comportamiento similar se aplican condicionando a otras elecciones en el primer y segundo períodos.

En el segundo período, la persona reconoce que su elección de trabajo afectará a sus opciones en edad de jubilación. Él sabe que va a maximizar entre las opciones disponibles cuando llegue la edad de jubilación. Supongamos que eligió la universidad en el primer período. En el segundo período, él sabe que la utilidad que obtendrá en el tercer período si opta por un trabajo j es $max_s U_{3s}^{Cj}$. La utilidad total de elegir el trabajo j en el segundo período, dado que él eligió la universidad en el primer período, es por lo tanto $TU_j^C = U_{2j}^C + \theta\, max_s U_{3s}^{Cj}$, donde θ pondera el peso del tercer período en relación al segundo período. La persona escoge el trabajo j si $TU_j^C > TU_k^C$ para todo k ≠ j y j, k ∈ J. Una notación y un comportamiento similar se producen si decide trabajar en el primer período.

Consideremos ahora el primer período. Él sabe que si escoge ir a la universidad, posteriormente escogerá el trabajo que maximice su utilidad condicionada a haber ido a la universidad, y luego elegirá la opción en la edad de jubilación que maximice su utilidad condicionada al trabajo elegido. El total de utilidad que proporciona la elección de ir a la universidad es

$$TU_C = U_{1C} + \lambda\, max_j\, TU_j^C$$

$$= U_{1C} + \lambda\, max_j\big(U_{2j}^C + \theta\, max_s\, U_{3s}^{Cj}\big).$$

Esta expresión es similar a la del modelo de dos períodos excepto que incluye un nivel adicional de maximización: la maximización para el tercer período está contenida en cada maximización del segundo período. Una expresión similar da la utilidad total de trabajar en el primer período, TU_W. La persona elige la universidad si $TU_C > TU_W$.

Esto completa la descripción del comportamiento de la persona. El investigador observa una parte de cada función de utilidad: U_{1C}, U_{1W}, U_{2j}^C y U_{2j}^W $\forall j$, y U_{3s}^{Cj} y U_{3s}^{Wj} $\forall s \in S, j \in J$. Las partes no observadas están etiquetadas colectivamente por el vector ε con densidad $\mathrm{f}(\varepsilon)$. La probabilidad de que la persona elija la universidad es

$$P_c = \int I(\varepsilon) f(\varepsilon) d\varepsilon,$$

donde

$$I(\varepsilon) = 1$$

si

$$V_{1C} + \varepsilon_{1C} + \lambda\, max_j\big(V_{2j}^C + \varepsilon_{2j}^C + \theta\, max_s(V_{3s}^{Cj} + \varepsilon_{3s}^{Cj})\big)$$
$$> V_{1W} + \varepsilon_{1W} + \lambda\, max_j\big(V_{2j}^W + \varepsilon_{2j}^W + \theta\, max_s(V_{3s}^{Wj} + \varepsilon_{3s}^{Wj})\big).$$

Esta expresión es la misma que en el modelo de dos períodos, excepto que ahora el término dentro de la función indicadora tiene un nivel extra de maximización. Podemos obtener un simulador aceptación-rechazo de la siguiente manera: (1) Extraiga un valor al azar de $\mathrm{f}(\varepsilon)$, (2) calcule la utilidad del tercer período U_{3s}^{Cj} para cada s, (3) identifique el máximo sobre s, (4) calcule $T\mathrm{U}_{2j}^C$ con este máximo, (5) repita los pasos (2)-(5) para cada j, e identifique el máximo de $T\mathrm{U}_{2j}^C$ sobre j; (6) calcule TU_C usando este máximo, (7) repita los pasos (2)-(6) para TU_W; (8) determine si $\mathrm{TU}_C > \mathrm{TU}_W$ y establezca $\mathrm{I} = 1$ si es así; (9) Repita los pasos (1)-(8) numerosas veces y promedie los resultados. También es posible usar partición conveniente del error. Por ejemplo, si todos los errores son valor extremo iid, entonces las probabilidades de elección del primer período, condicionadas a valores extraídos al azar de los errores del segundo y tercer períodos, son logit; las probabilidades del segundo período, condicionadas a los errores del tercer período, son logit; y las probabilidades del tercer período son logit.

Ahora podemos generalizar estos conceptos e introducir algo de terminología ampliamente utilizada. Tenga en cuenta que el análisis del comportamiento de la persona y de la simulación de las probabilidades de elección por parte del investigador empieza por el último período y se extiende hacia atrás en el tiempo hasta el primer período. Este proceso se llama recursión hacia atrás (*backwards recursion*). Supongamos que hay J alternativas en cada uno de los T períodos de tiempo de igual

longitud. Denotemos una secuencia de elecciones hasta el período t como $\{\mathrm{i}_1, \mathrm{i}_2, \dots, \mathrm{i}_t\}$. La utilidad que la persona obtiene en el período t de la alternativa j es $\mathrm{U}_{tj}(\mathrm{i}_1, \mathrm{i}_2, \dots, \mathrm{i}_{t-1})$, que depende de todas las elecciones anteriores. Si la persona elige la alternativa j en el período t, obtendrá esta utilidad más la utilidad futura de sus elecciones, condicionada a esta elección. La utilidad total (actual y futura) que la persona obtiene de elegir la alternativa j en el período t es $\mathrm{TU}_{tj}(\mathrm{i}_1, \mathrm{i}_2, \dots, \mathrm{i}_{t-1})$. La persona elige la alternativa en el período actual que proporciona la mayor utilidad total. Por lo tanto, la utilidad total que recibe de su elección óptima en el período t es $\mathrm{TU}_t(\mathrm{i}_1, \mathrm{i}_2, \dots, \mathrm{i}_{t-1}) = max_j \mathrm{TU}_{tj}(\mathrm{i}_1, \mathrm{i}_2, \dots, \mathrm{i}_{t-1})$. Esta utilidad total de la opción óptima en el período t, TU_t, se denomina la función de valoración (*valuation function*) en el periodo t.

La persona elige de manera óptima en el período actual con el conocimiento de que elegirá óptimamente en el futuro. Este hecho establece una relación conveniente entre la función de valoración en períodos sucesivos. En particular,

$$TU_t(i_1, \dots, i_{t-1}) = max_j\big[U_{jt}(i_1, \dots, i_{t-1}) + \delta TU_{t+1}(i_1, \dots, i_t = j)\big],$$

donde δ es un parámetro que descuenta el valor de la utilidad futura. En el lado derecho, TU_{t+1} es la utilidad total que la persona va a obtener del período $\mathrm{t}+1$ en adelante si opta por la alternativa j en el período t (es decir, si $\mathrm{i}_t = j$). La ecuación establece que la utilidad total que la persona obtiene de optimizar su comportamiento del período t en adelante, dadas sus elecciones anteriores, es el máximo sobre j de la utilidad de j en el período t, más la utilidad total descontada de optimizar el comportamiento del período $t+1$ en adelante, condicionada a la elección de j en el período t. Esta relación es la ecuación de Bellman (1957) aplicada a la elección discreta con información perfecta.

$\mathrm{TU}_{tj}(\mathrm{i}_1, \dots, \mathrm{i}_{t-1})$ a veces es llamada la función de valoración condicionada, condicionada a la elección de la alternativa j en el período t. Una ecuación de Bellman también opera para este término:

$$TU_{tj}(i_1, \dots, i_{t-1}) = U_{jt}(i_1, \dots, i_{t-1}) + \delta max_k\big[TU_{t+1,k}(i_1, \dots, i_t = j)\big].$$

Dado que, por definición, $\mathrm{TU}_t(\mathrm{i}_1, \dots, \mathrm{i}_{t-1}) = \max_j[\mathrm{TU}_{tj}(\mathrm{i}_1, \dots, \mathrm{i}_{t-1})]$, la ecuación de Bellman en términos de la función de valoración condicionada es equivalente a la ecuación en términos de la función de valoración no condicionada.

Si T es finito, la ecuación de Bellman se puede aplicar con recursividad hacia atrás para calcular TU_{tj} para cada período de tiempo. En $t = T$ no hay período de tiempo futuro, así que $\mathrm{TU}_{Tj}(\mathrm{i}_1, \dots, \mathrm{i}_{T-1}) = \mathrm{U}_{Tj}(\mathrm{i}_1, \dots, \mathrm{i}_{T-1})$. Entonces $\mathrm{TU}_{T-1,j}(\mathrm{i}_1, \dots, \mathrm{i}_{T-2})$ se calcula a partir de $\mathrm{TU}_{Tj}(\mathrm{i}_1, \dots, \mathrm{i}_{T-1})$ usando la ecuación de Bellman, y así sucesivamente hasta $\mathrm{t} = 1$. Tenga en cuenta que $\mathrm{U}_{tj}(\mathrm{i}_1, \dots, \mathrm{i}_{t-1})$ debe ser calculada para cada t, cada j, y, muy importante, para cada posible secuencia de elecciones pasadas, $\mathrm{i}_1, \dots, \mathrm{i}_{t-1}$.Con J alternativas en períodos T de tiempo, la recursión requiere el cálculo de $(J^T)T$ utilidades (es decir, J^T posibles secuencias de elecciones, con cada secuencia conteniendo T utilidades de un único período). Para simular las probabilidades, el investigador debe calcular estas utilidades para cada valor extraído al azar de factores no observados. Y estas probabilidades deben ser simuladas para cada valor de los parámetros en la búsqueda numérica para las estimaciones. Esta enorme carga computacional ha recibido el nombre de la maldición de la dimensionalidad (*curse of dimensionality*) y es el principal obstáculo para la aplicación de estos procedimientos para más de unos pocos períodos de tiempo y/o alternativas. Veremos en las próximas secciones procedimientos que se han sugerido para evitar o mitigar esta maldición, después de mostrar que la maldición es aún mayor cuando se considera la incertidumbre.

7.7.3 Incertidumbre sobre efectos futuros

Hasta ahora hemos asumido en el análisis que el decisor conoce la utilidad de cada alternativa en cada período de tiempo futuro y cómo esta utilidad se ve afectada por decisiones anteriores. Por lo general, el decisor no posee tal conocimiento previo. Un grado de incertidumbre envuelve los efectos futuros de las elecciones actuales. Es posible adaptar el modelo de comportamiento para incorporar la incertidumbre. Para simplificar, volvemos al modelo de dos períodos relativo a nuestro estudiante de secundaria. Supongamos que en el primer período el estudiante no sabe a ciencia cierta las utilidades del segundo período, U_{2j}^{C} y U_{2j}^{W} $\forall j$. Por ejemplo, el estudiante no sabe antes de ir a la universidad cómo irá la economía, y por lo tanto cuáles serán sus posibilidades de empleo cuando se gradúe. Estas utilidades pueden ser expresadas como funciones de factores desconocidos $U_{2j}^{C}(e)$, donde e se refiere colectivamente a todos los factores del segundo período que son desconocidos en el período uno. Estos factores desconocidos se convertirán en conocidos (es decir, se revelarán) cuando el estudiante alcance el segundo período, pero son desconocidos para la persona en el primer período. El estudiante tiene una distribución subjetiva sobre e que refleja la probabilidad que él atribuye a los factores desconocidos de que tomen unos valores particulares en el segundo período. Esta densidad la denominaremos $g(e)$. Él sabe que, independientemente de qué posibilidades de e se concreten en la realidad, en el segundo período seleccionará el trabajo que le dé la máxima utilidad. Es decir, él sabe que va a recibir la utilidad $\max_j U_{2j}^{C}(e)$ en el segundo período, si elige la universidad en el primer período y los factores desconocidos terminan siendo e. En el primer período, cuando evalúa la posibilidad de ir a la universidad, él toma la expectativa de esta futura utilidad sobre todas las posibles realizaciones de los factores desconocidos, utilizando su distribución subjetiva sobre estas realizaciones. Por tanto, la utilidad esperada que va a obtener en el segundo período si elige la universidad en el primer período es $\int \left[\max_j U_{2j}^{C}(e)\right] g(e)de$. La utilidad total esperada de elegir la universidad en el primer período es por lo tanto

$$TEU_c = U_{1c} + \lambda \int \left[max_j\, U_{2j}^{C}(e)\right] g(e)de.$$

TEU_W se define de manera similar. La persona elige la universidad si $TEU_C > TEU_W$. En el segundo período, los factores hasta ese momento desconocidos se dan a conocer, y la persona elige el trabajo j si ha elegido la universidad sólo si $U_{2j}^{C}(e^*) > U_{2k}^{C}(e^*)$ para todo $k \neq j$, donde e^* es la realización de los factores que en realidad ocurrieron.

En cuanto al investigador, se le presenta una complicación adicional introducida por $g(e)$, la distribución subjetiva del decisor sobre los factores desconocidos. Además de no conocer las utilidades en su totalidad, el investigador tiene sólo un conocimiento parcial de la probabilidad subjetiva $g(e)$ del decisor. Esta falta de información se suele representar mediante parametrización. El investigador especifica una densidad, $h(e|\theta)$, que depende de parámetros desconocidos θ. El investigador asume entonces que la densidad subjetiva de la persona es la densidad especificada, evaluada en los parámetros verdaderos θ^*. Es decir, el investigador asume que $h(e|\theta^*) = g(e)$. Dicho de forma más convincente y precisa: los verdaderos parámetros son, por definición, los parámetros para los que la densidad especificada por el investigador $h(e|\theta)$ se convierte en la densidad $g(e)$ que la persona realmente usó. Con una función h suficientemente flexible, cualquier g puede representarse como h evaluada en algunos parámetros, que se llaman los parámetros verdaderos. Estos parámetros se estiman junto con los parámetros que forman parte de la utilidad. (Otras formas de representar la falta de conocimiento que tiene el investigador acerca de $g(e)$ son posibles; sin embargo, son generalmente más complejas).

Las utilidades se descomponen en sus partes observadas y no observadas, con las partes no observadas llamadas colectivamente ε con densidad $f(\varepsilon)$. La probabilidad de que la persona vaya a la universidad es

$$P_c = Prob(TEU_c > TEU_W)$$

$$= \int I(TEU_c > TEU_W) f(\varepsilon) d\varepsilon$$

donde TEU_C y TEU_W, según las definiciones anteriores, incluyen una integral cada una sobre e con densidad $h(e|\theta)$. La probabilidad se puede aproximar mediante la simulación de las integrales en TEU_C y TEU_W, dentro de la simulación de la integral sobre $I(TEU_C > TEU_W)$. Para ello se deben seguir los siguientes pasos. (1) Extraiga un valor al azar de ε. (2a) Extraiga un valor al azar de $h(e|\theta)$. (2b) Usando este valor, calcule los integrandos en TEU_Cy TEU_W. (2c) Repita los pasos 2a-b numerosas veces y promedie los resultados. (3) Usando el valor de 2c, calcule $I(TEU_C > TEU_W)$. (4) Repita los pasos 1-3 múltiples veces y promedie los resultados. Como el lector puede ver, la maldición de la dimensionalidad empeora.

Varios autores han sugerido formas de reducir la carga computacional. Keane y Wolpin (1994) calculan la función de valoración en una selección de realizaciones de los factores desconocidos y de las decisiones pasadas; luego aproximan la función de valoración en otras realizaciones y decisiones pasadas mediante interpolación a partir de las valoraciones calculadas. Rust (1997) sugiere simular futuros caminos y usar el promedio de estos caminos simulados como una aproximación de la función de valoración. Hotz y Miller (1993) y Hotz et al. (1993) demuestran que existe una correspondencia entre la función de valoración en cada período de tiempo y las probabilidades de elección en períodos futuros. Esta correspondencia permite que las funciones de valoración sean calculadas con estas probabilidades en lugar de usar recursividad hacia atrás.

Cada uno de estos procedimientos tiene limitaciones y es aplicable sólo en determinadas situaciones, que los propios autores describen. Como Rust (1994) ha observado, es poco probable que surja un gran avance de aplicación general que haga simple la estimación para todas las tipologías de modelos de optimización dinámica. Inevitablemente, el investigador tendrá que hacer concesiones en el momento de especificar el modelo para asegurar la viabilidad, y el método de especificación y estimación más adecuado dependerá de las particularidades del proceso de elección y los objetivos de la investigación. En este sentido, he encontrado que dos simplificaciones son muy poderosas en cuanto a que frecuentemente proporcionan una gran ganancia en la viabilidad computacional del modelo a cambio de una pérdida relativamente pequeña (y a veces, una ganancia) en el contenido.

La primera sugerencia es que el investigador considere formas de captar la naturaleza de la situación de elección con el menor número de períodos de tiempo posible. A veces, de hecho por norma general, los períodos de tiempo no deben definirse usando intervalos estándar, como el año o el mes, sino más bien en unidades de distancia que sean más estructurales respecto al proceso de decisión. Por ejemplo, para el estudiante de secundaria y su decisión de ir o no a la universidad, podría parecer natural decir que hace una elección cada año entre los posibles puestos de trabajo y opciones de enseñanza que están disponibles en ese año, teniendo en cuenta sus decisiones pasadas. De hecho, esta afirmación es cierta: el estudiante, en efecto, hace elecciones anuales (o incluso mensuales, semanales, diarias). Sin embargo, este modelo se enfrentaría claramente a la maldición de la dimensionalidad. Por el contrario, la especificación que hemos comentado anteriormente implica sólo dos períodos de tiempo, o tres si se considera la jubilación. La estimación es bastante factible para esta especificación. De hecho, el modelo de dos períodos puede ser más preciso que un modelo anual: los estudiantes que deciden sobre la universidad probablemente piensan en términos de los años universitarios y sus opciones después de la universidad, en lugar de tratar de anticipar sus decisiones futuras para cada año futuro. McFadden y Train (1996) proporcionan un ejemplo de cómo un modelo de optimización dinámica, con sólo unos pocos períodos bien pensados, puede capturar con precisión la naturaleza de una situación de elección.

Una segunda simplificación de gran alcance fue observada por primera vez por Rust (1987). Supongamos que los factores que el decisor no observa de antemano también son los factores que el investigador no observa (ya sea antes o después), y que el decisor piensa que esos factores son valor extremo iid. Bajo este supuesto ciertamente restrictivo, las probabilidades de elección toman una forma cerrada que es fácil de calcular. El resultado se puede obtener fácilmente para nuestro modelo de elección de la universidad. Supongamos que el estudiante, cuando está en el primer período, descompone la utilidad del segundo período en una parte conocida y otra desconocida, por ejemplo, $U_{2j}^{C}(e) = V_{2j}^{C} + e_{2j}^{C}$, y que asume que e_{2j}^{C} sigue una distribución de valor extremo independiente de todo lo demás. Este factor desconocido pasa a ser conocido para el estudiante en el segundo período, de modo que la elección del segundo período implica la maximización sobre una $U_{2j}^{C} \forall j$ conocida. Sin embargo, en el primer período es desconocida. Recuerde de la sección 3.5 que el máximo esperado de utilidades tipo valor extremo iid toma la conocida fórmula log-suma. En nuestro contexto, este resultado significa que

$$E\left(max_j\left(V_{2j}^{C} + \varepsilon_{2j}^{C}\right)\right) = \alpha\, ln\left(\sum_j e^{V_{2j}^{C}}\right),$$

que podemos etiquetar LS_2^C. LS_2^W se obtiene de manera similar. Por tanto, la persona elige la universidad si

$$TEU_c > TEU_W,$$

$$U_{1c} + \lambda LS_2^C > U_{1W} + \lambda LS_2^W$$

Observe que esta regla de decisión tiene una forma cerrada: la integral sobre los factores futuros desconocidos se convierte en la fórmula log-suma. Consideremos ahora al investigador. Cada utilidad del primer período se descompone en una parte observada y otra no observada ($U_{1C} = V_{1C} + \varepsilon_{1C}$, $U_{1W} = V_{1\mathrm{W}} + \varepsilon_{1\mathrm{W}}$), y suponemos que las partes no observadas son valor extremo iid. Para las utilidades del segundo período, hacemos una suposición bastante restrictiva. Suponemos que la parte de la utilidad que el investigador no observa es la misma parte que el estudiante no sabe de antemano en el momento de decidir. Es decir, asumimos $\mathrm{U}_{2j}^{C} = \mathrm{V}_{2j}^{C} + \varepsilon_{2j}^{C}\ \forall j$, donde el término ε_{2j}^{C} del investigador es el mismo término e_{2j}^{C} del estudiante. Bajo esta hipótesis, el investigador puede calcular los términos log-suma de la utilidad futura, LS_2^C y LS_2^W, de forma exacta, ya que estos términos sólo dependen de la parte observada de la utilidad en el segundo período, $\mathrm{V}_{2j}^{C}\ \forall j$, que es observada por el investigador y conocida de antemano por el decisor. La probabilidad de que el estudiante escoja la universidad es ahora

$$P_c = Prob(TEU_c > TEU_W)$$

$$= Prob(U_{1c} + \lambda LS_2^C > U_{1W} + \lambda LS_2^W)$$

$$= Prob(V_{1c} + \varepsilon_{1c} + \lambda LS_2^C > V_{1W} + \varepsilon_{1W} + \lambda LS_2^W)$$

$$= \frac{e^{V_{1C} + \lambda LS_2^C}}{e^{V_{1C} + \lambda LS_2^C} + e^{V_{1W} + \lambda LS_2^W}}$$

El modelo toma la misma forma que la parte superior de un modelo logit jerárquico: la probabilidad de elección del primer período es la fórmula logit con un término log-suma incluido como una variable explicativa adicional. Múltiples períodos se manejan de la misma manera que logits jerárquicos de varios niveles.

En realidad, es dudoso que el investigador observe todo lo que el decisor conoce de antemano cuando elige. Sin embargo, la simplificación que se plantea a partir de esta suposición es tan grande, y la maldición de la dimensionalidad que surgiría de otra forma es tan severa, que proceder como si esta hipótesis fuese cierta vale la pena en muchas situaciones.

PARTE II

Estimación

8

Maximización numérica

8.1 Motivación

Muchos procesos de estimación implican la maximización de alguna función, como la función de verosimilitud, la de verosimilitud simulada o la de condiciones de momentos cuadráticos (*squared moment conditions*). Este capítulo describe los procedimientos numéricos que se utilizan para maximizar una función de verosimilitud. Procedimientos análogos se utilizan para maximizar otras funciones.

Conocer y ser capaz de aplicar estos procedimientos es fundamental en esta nueva era de los modelos de elección discreta. En el pasado, los investigadores adaptaron sus especificaciones a los pocos modelos prácticos que estaban a su disposición. Estos modelos se incluyeron en los paquetes de software de estimación disponibles en el mercado, de modo que el investigador podía estimar los modelos sin conocer los detalles de cómo se llevaba a cabo realmente la estimación desde una perspectiva numérica. La potencia de esta nueva ola de métodos de elección discreta es liberar al investigador para que pueda especificar modelos pensados a medida de su situación y de sus problemas. Utilizar esta libertad significa que el investigador a menudo se encontrará especificando un modelo que no es exactamente igual a uno de los modelos disponibles en el software comercial. En estos casos, tendrá que escribir código especial para su modelo especial.

El propósito de este capítulo es ayudar a hacer posible este ejercicio. Aunque por lo general no se enseñan en cursos de econometría, los procedimientos de maximización son bastante sencillos y fáciles de implementar. Una vez aprendidos, la libertad que ofrecen tiene un valor incalculable.

8.2 Notación

La función logaritmo de la verosimilitud (log-verosimilitud) tiene la forma $LL(\beta) = \sum_{n=1}^{N} \ln P_n(\beta)/N$, donde $P_n(\beta)$ es la probabilidad del resultado observado para el decisor n, N es el tamaño de la muestra y β es un vector de $K \times 1$ parámetros. En este capítulo, dividimos la función de verosimilitud por N, de modo que LL es la verosimilitud promedio en la muestra. Hacer esta división no altera la posición del máximo (dado que N es un valor fijo para una muestra dada) y sin embargo facilita la interpretación de alguno de los procedimientos. Todos los procedimientos funcionan igual dividamos o no la función log-verosimilitud por N. El lector puede verificar este hecho a medida que avancemos al observar que la N se cancela y queda excluida de las fórmulas relevantes.

El objetivo es encontrar el valor de β que maximice $LL(\beta)$. En términos de la figura 8.1, el objetivo es localizar $\hat{\beta}$. Observe que en esta figura LL es siempre negativa, ya que la verosimilitud es una probabilidad entre 0 y 1, y el logaritmo de cualquier número entre 0 y 1 es negativo. Numéricamente, el máximo se puede encontrar "subiendo" por la función de verosimilitud hasta que no podamos lograr ningún incremento adicional. El investigador especifica unos valores iniciales β_0. Cada iteración o paso, nos movemos a un nuevo valor de los parámetros en los que $LL(\beta)$ sea mayor que en el valor anterior. Llamemos al valor actual de β como β_t, valor que se alcanza después de t pasos desde los valores iniciales. La pregunta es: ¿cuál es el mejor paso que podemos tomar a continuación?, es decir, ¿cuál es el mejor valor para β_{t+1}?

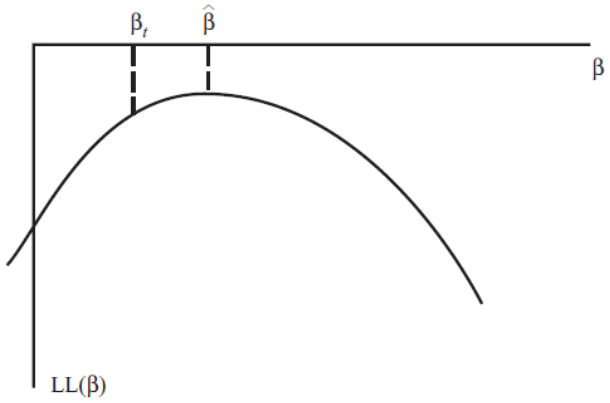


Figura 8.1. Estimación de máxima verosimilitud

El gradiente en β_t es el vector de las primeras derivadas de $LL(\beta)$ evaluadas en β_t:

$$g_t = \left(\frac{\partial LL(\beta)}{\partial \beta}\right)_{\beta_t}.$$

Este vector nos dice en qué dirección dar el siguiente paso con el fin de desplazarnos a un valor mayor de la función de verosimilitud. La matriz hessiana (o hessiano) es la matriz de segundas derivadas:

$$H_t = \left(\frac{\partial g_t}{\partial \beta'}\right)_{\beta_t} = \left(\frac{\partial^2 LL(\beta)}{\partial \beta \partial \beta'}\right)_{\beta_t}.$$

El gradiente tiene dimensiones $K \times 1$ y el hessiano $K \times K$. Como veremos, el hessiano nos puede ayudar a saber qué tan largo debemos dar el paso, mientras que el gradiente nos dice en qué dirección dar el paso.

8.3 Algoritmos

De los numerosos algoritmos de maximización que se han desarrollado a lo largo de los años, describiré sólo los más destacados, con un énfasis en el valor pedagógico de los procedimientos así como en su uso práctico. Los lectores que se sientan atraídos a estudiar más a fondo esta cuestión pueden encontrar gratificante el tratamiento dado a este tema por Judge et al. (1985, Apéndice B) y Ruud (2000).

8.3.1 Newton-Raphson

Para determinar el mejor valor de β_{t+1}, tome una aproximación de Taylor de segundo orden de la función $LL(\beta_{t+1})$ en torno a $LL(\beta_t)$:

$$(8.1) \qquad LL(\beta_{t+1}) = LL(\beta_t) + (\beta_{t+1} - \beta_t)' g_t + \frac{1}{2}(\beta_{t+1} - \beta_t)' H_t (\beta_{t+1} - \beta_t)$$

Ahora encuentre el valor de β_{t+1} que maximice esta aproximación de $LL(\beta_{t+1})$:

$$\frac{\partial LL(\beta_{t+1})}{\partial \beta_{t+1}} = g_t + H_t(\beta_{t+1} - \beta_t) = 0,$$

$$H_t(\beta_{t+1} - \beta_t) = -g_t,$$

$$\beta_{t+1} - \beta_t = -{H_t}^{-1} g_t,$$

$$\beta_{t+1} = \beta_t + \left(-{H_t}^{-1}\right) g_t.$$

El procedimiento Newton-Raphson (NR) utiliza esta fórmula. El paso a dar desde el valor actual de β al nuevo valor es $\left(-{\mathrm{H}_t}^{-1}\right) g_t$, es decir, el vector del gradiente multiplicado por el negativo de la inversa del hessiano.

Esta fórmula es intuitivamente comprensible. Considere $k = 1$, como se ilustra en la figura 8.2. La pendiente de la función de verosimilitud es g_t. La segunda derivada es el hessiano H_t, que es negativo en este gráfico, ya que la curva se ha dibujado cóncava. El negativo de este hessiano negativo es positivo y representa el grado de curvatura. Es decir, $-\mathrm{H}_t$ es la curvatura positiva. Cada paso de β es la pendiente de la función log-verosimilitud dividida por su curvatura. Si la pendiente es positiva, β se incrementa como se muestra en el primer dibujo, y si la pendiente es negativa, β se reduce como en el segundo dibujo. La curvatura determina cómo de grande se da un paso. Si la curvatura es grande, lo que significa que la pendiente cambia rápidamente tal y como se muestra en el primer dibujo de la figura 8.3, entonces es probable que estemos cerca del máximo, así que se da un paso pequeño. (Dividir el gradiente por un número grande da un número pequeño). Por el contrario, si la curvatura es pequeña, lo que significa que la pendiente no está cambiando mucho, entonces el máximo probablemente esté más lejos y por lo tanto se da un paso más grande.

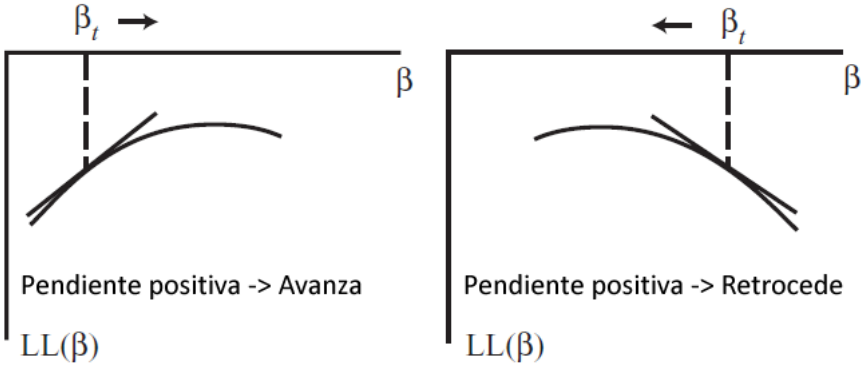


Figura 8.2. La dirección del paso a dar sigue la pendiente.

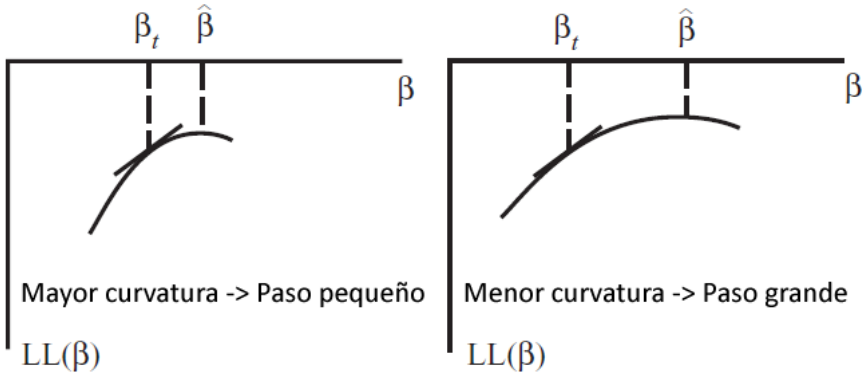


Figura 8.3. El tamaño del paso se relaciona inversamente con la curvatura.

Debemos tener en cuenta tres cuestiones relevantes en el procedimiento NR.

Cuadrático

Si $LL(\beta)$ fuese exactamente una función cuadrática de β, entonces el procedimiento NR alcanzaría el máximo en un solo paso desde cualquier valor inicial. Este hecho se puede demostrar fácilmente con $K = 1$. Si $LL(\beta)$ es cuadrática, entonces puede ser escrita como

$$LL(\beta) = a + b\beta + c\beta^2$$

El máximo es

$$\frac{\partial LL(\beta)}{\partial \beta} = b + 2c\beta = 0,$$

$$\hat{\beta} = -\frac{b}{2c}.$$

El gradiente y el hessiano son $g_t = b + 2c\beta_t$ y $H_t = 2c$, así que el procedimiento NR nos da

$$\beta_{t+1} = \beta_t - {H_t}^{-1} g_t$$

$$= \beta_t - \frac{1}{2c}(b + 2c\beta_t)$$

$$= \beta_t - \frac{b}{2c} - \beta_t$$

$$= -\frac{b}{2c} = \hat{\beta}.$$

La mayoría de las funciones log-verosimilitud no son cuadráticas, por lo que el procedimiento NR necesitará más de un paso para alcanzar el máximo. Sin embargo, saber cómo se comporta el procedimiento NR en el caso cuadrático ayuda a comprender su comportamiento con LL no cuadráticas, como veremos en la siguiente sección.

Tamaño del paso

Es posible que el procedimiento NR, al igual que otros procedimientos que se explican posteriormente, dé un paso que vaya más allá del máximo, llegando a un valor de los parámetros β en el que la $LL(\beta)$

sea inferior a la de partida. La figura 8.4 representa la situación. La función LL real está representada por la línea continua. La línea discontinua es una función cuadrática que tiene la pendiente y la curvatura que tiene LL en el punto β_t. El procedimiento NR se mueve hasta la parte superior de la función cuadrática, hasta β_{t+1}. Sin embargo, en este caso $LL(\beta_{t+1})$ es inferior a $LL(\beta_t)$.

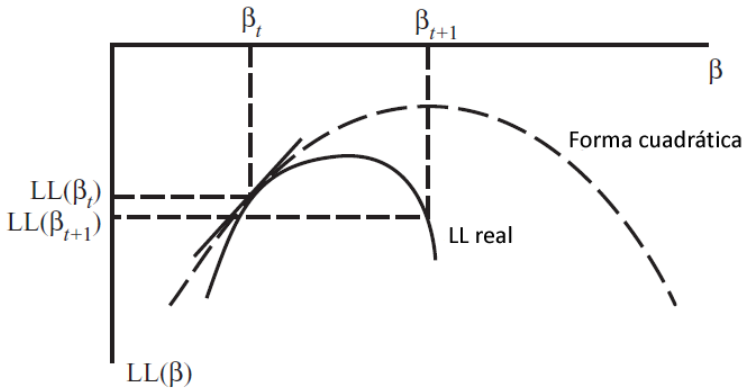


Figura 8.4. El paso podría ir más allá del máximo, hasta un valor inferior de LL

Para contemplar esta posibilidad, el paso se puede multiplicar por un escalar λ en la fórmula NR:

$$\beta_{t+1} = \beta_t + \lambda\left(-{H_t}^{-1}\right)g_t.$$

El vector $\left(-{H_t}^{-1}\right)g_t$ se denomina dirección y λ recibe el nombre de tamaño del paso. (Esta terminología es estándar, aunque $\left(-{H_t}^{-1}\right)g_t$ también contiene información sobre el tamaño del paso a través de H_t, como ya se ha explicado en relación a la figura 8.3). El tamaño de paso λ se reduce para asegurar que en cada paso del procedimiento NR logremos un aumento en $LL(\beta)$. El ajuste se realiza por separado en cada iteración, de la siguiente manera.

Empezamos con $\lambda = 1$. Si $LL(\beta_{t+1}) > LL(\beta_t)$, nos movemos a β_{t+1} y comenzamos una nueva iteración. Si $LL(\beta_{t+1}) < LL(\beta_t)$, establecemos $\lambda = 1/2$ y volvemos a intentarlo. Si, con $\lambda = 1/2$, $LL(\beta_{t+1})$ sigue siendo inferior a $LL(\beta_t)$, establecemos $\lambda = 1/4$ y volvemos a intentarlo. Continuamos este proceso hasta encontrar una λ para la que $LL(\beta_{t+1}) > LL(\beta_t)$. Si este proceso acaba generando una λ pequeña, entonces se ha avanzado poco en la búsqueda del máximo. Esto puede ser interpretado como una señal para el investigador de que puede ser necesario un procedimiento iterativo diferente.

Es posible hacer un ajuste análogo del tamaño de paso en la dirección contraria, es decir, mediante el aumento de λ cuando sea apropiado. Un caso se muestra en la figura 8.5. La parte superior (el máximo) de la función cuadrática se obtiene con un tamaño de paso de $\lambda = 1$. Sin embargo, la $LL(\beta)$ no es cuadrática, y su máximo está más lejos. El tamaño del paso se puede ajustar hacia arriba, siempre y cuando $LL(\beta)$ siga creciendo. Es decir, calculamos β_{t+1} con $\lambda = 1$ en β_{t+1}. Si $LL(\beta_{t+1}) > LL(\beta_t)$, intentamos $\lambda = 2$. Si la β_{t+1} basada en $\lambda = 2$ da un valor mayor de la función log-verosimilitud que con $\lambda = 1$, entonces probamos $\lambda = 4$, y así sucesivamente, doblando λ siempre y cuando al hacerlo elevemos aún más la función de verosimilitud. Cada vez, $LL(\beta_{t+1})$ con el valor de λ doblado se compara con el valor de λ probado anteriormente, en lugar de compararlo con el valor para $\lambda = 1$, con el fin de asegurar que cada vez que doblamos λ aumentamos la función de verosimilitud más de lo que previamente se había incrementado con una λ más pequeña. En la figura 8.5, se utiliza un tamaño de paso final de 2, ya que la función de verosimilitud con $\lambda = 4$ es menor que con $\lambda = 2$, a pesar de que es mayor que con $\lambda = 1$.

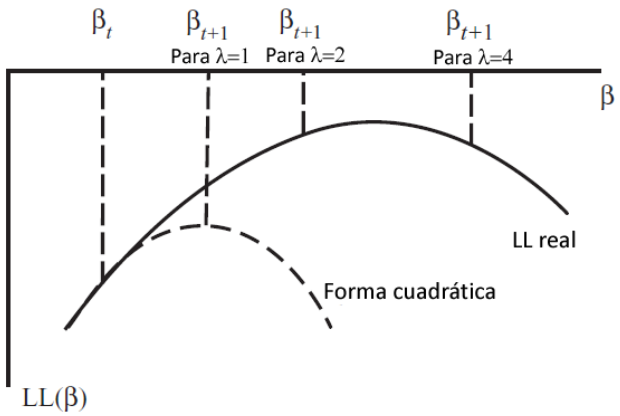


Figura 5.8. Doblamos λ mientras LL siga creciendo

La ventaja de este enfoque (incrementar λ) es que por lo general reduce el número de iteraciones necesarias para alcanzar el máximo. Podemos probar nuevos valores de λ sin necesidad de volver a calcular g_t y H_t, mientras que cada nueva iteración del procedimiento NR requiere el cálculo de estos términos. Por lo tanto, ajustar λ puede acelerar la búsqueda del máximo.

Concavidad

Si la función log-verosimilitud es globalmente cóncava, el procedimiento NR garantiza un aumento de la función de verosimilitud en cada iteración. Este hecho se demuestra de la siguiente manera. Que $LL(\beta)$ sea cóncava significa que su matriz hessiana es definida negativa en todos los valores de β. (En una dimensión, la pendiente de $LL(\beta)$ está disminuyendo, de modo que la segunda derivada es negativa). Si H es definida negativa, entonces H^{-1} también es definida negativa y $-H^{-1}$ es definida positiva. Por definición, una matriz simétrica M se dice definida positiva si $x'Mx > 0$ para cualquier $x \neq 0$. Considere la aproximación de Taylor de primer orden de $LL(\beta_{t+1})$ alrededor de $LL(\beta_t)$:

$$LL(\beta_{t+1}) = LL(\beta_t) + (\beta_{t+1} - \beta_t)' g_t$$

En el marco del procedimiento NR, $\beta_{t+1} - \beta_t = \lambda\left(-{H_t}^{-1}\right)g_t$. Sustituyendo obtenemos

$$LL(\beta_{t+1}) = LL(\beta_t) + (\lambda\left(-{H_t}^{-1}\right)g_t)' g_t$$

$$= LL(\beta_t) + \lambda g_t'\left(-{H_t}^{-1}\right)g_t.$$

Dado que $-H^{-1}$ es definida positiva, tenemos que $g_t'\left(-{H_t}^{-1}\right)g_t > 0$ y $LL(\beta_{t+1}) > LL(\beta_t)$. Tenga en cuenta que dado que esta comparación se basa en una aproximación de primer orden, un aumento en $LL(\beta)$ podría obtenerse sólo en una pequeña zona de β_t. Es decir, el valor de λ que proporciona un aumento podría ser pequeño. Sin embargo, un incremento está garantizado en cada iteración si $LL(\beta)$ es globalmente cóncava.

Supongamos que la función log-verosimilitud tiene regiones que no son cóncavas. En estas áreas, el procedimiento NR puede no encontrar un incremento de la verosimilitud. Si la función es convexa en β_t, el procedimiento NR se mueve en la dirección opuesta a la pendiente de la función log-verosimilitud. La situación se ilustra en la figura 8.6 para $K = 1$. El paso que da el procedimiento NR con un único parámetro es $LL'(\beta)/(-LL''(\beta))$, donde el símbolo ' hace referencia a la derivada. La segunda derivada es positiva en β_t, ya que la pendiente está aumentando. Por lo tanto, $-LL''(\beta)$ es negativo, y el paso queda definido en la dirección opuesta a la pendiente. Con $K > 1$, si la matriz hessiana es definida

positiva en β_t, entonces $-{H_t}^{-1}$ es definida negativa, y los pasos del procedimiento NR quedan definidos en la dirección opuesta a g_t.

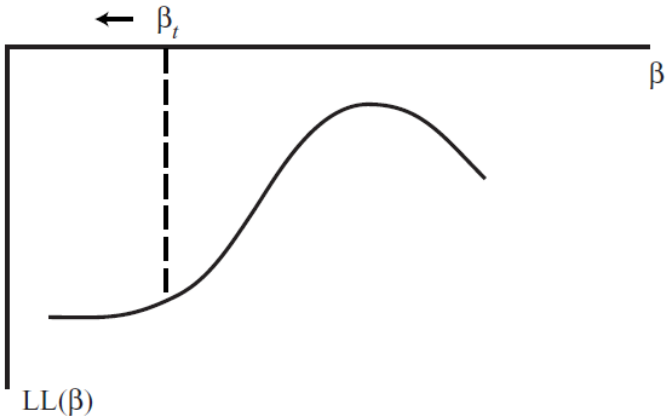


Figura 8.6. NR en la porción convexa de la función LL.

El signo del hessiano puede invertirse en estas situaciones. Sin embargo, no hay ninguna razón para el uso del hessiano allí donde la función no es cóncava, dado que el hessiano en las regiones convexas no proporciona ninguna información útil sobre dónde puede estar el máximo. Hay maneras más fáciles para lograr un incremento de LL en estas situaciones que calcular el hessiano e invertir su signo. Este tema es una de las razones que motiva el uso de otros procedimientos.

El procedimiento NR tiene dos inconvenientes. En primer lugar, el cálculo del hessiano es por lo general computacionalmente costoso. Procedimientos que eviten el cálculo del hessiano en cada iteración pueden ser mucho más rápidos. En segundo lugar, como acabamos de mostrar, el procedimiento NR no garantiza un incremento en cada paso si la función log-verosimilitud no es globalmente cóncava. Cuando $-{H_t}^{-1}$ no es definida positiva, no se puede garantizar un aumento.

Otros enfoques utilizan aproximaciones del hessiano que resuelven estas dos cuestiones. Los métodos difieren en la forma de la aproximación. Cada procedimiento define un paso como

$$\beta_{t+1} = \beta_t + \lambda M_t g_t,$$

donde M_t es una matriz $\mathrm{K} \times \mathrm{K}$. Para el procedimiento NR, $M_t = -{H_t}^{-1}$. Otros procedimientos utilizan M_ts que son más fáciles de calcular que el hessiano y que son necesariamente definidas positivas, a fin de garantizar un aumento en cada iteración, incluso en regiones convexas de la función log-verosimilitud.

8.3.2 BHHH

El procedimiento NR no utiliza el hecho de que la función que se está maximizado en realidad es la suma de logaritmos de verosimilitudes sobre una muestra de observaciones. El gradiente y el hessiano se calculan tal y como se haría en la maximización de cualquier función. Esta característica del procedimiento NR le proporciona generalidad, en el sentido de que se puede utilizar para maximizar cualquier función, no sólo un logaritmo de verosimilitud. Sin embargo, como veremos, la maximización puede ser más rápida si utilizamos el hecho de que la función que se está maximizando es una suma de términos en una muestra.

Necesitamos un poco de notación adicional para reflejar el hecho de que la función log-verosimilitud es una suma sobre observaciones. Definimos la *puntuación* (*score*) de una observación como la derivada del logaritmo de verosimilitud de esa observación respecto a los parámetros: $s_n(\beta_t) = \partial ln P_n(\beta)/\partial \beta$

evaluada en β_t. El gradiente, que hemos definido anteriormente y que se utiliza en el procedimiento NR, es la puntuación media: $g_t = \sum_n s_n(\beta_t)/N$. El producto exterior (*outer product*) de la puntuación correspondiente a la observación n es la matriz K × K

$$s_n(\beta_t)s_n(\beta_t)' = \begin{pmatrix} s_n^1 s_n^1 & s_n^1 s_n^2 & \cdots & s_n^1 s_n^K \\ s_n^1 s_n^2 & s_n^2 s_n^2 & \cdots & s_n^2 s_n^K \\ \vdots & \vdots & & \vdots \\ s_n^1 s_n^K & s_n^2 s_n^K & \cdots & s_n^K s_n^K \end{pmatrix},$$

donde s_n^k es el elemento k-ésimo de $s_n(\beta_t)$ con la dependencia de β_t omitida por conveniencia. El producto exterior medio en la muestra es $B_t = \sum_n s_n(\beta_t)s_n(\beta_t)'$ /N. Esta media está relacionada con la matriz de covarianza: si la puntuación media fuera cero, entonces B sería la matriz de covarianza de las puntuaciones de la muestra. A menudo B_t se denomina como el "producto exterior del gradiente". Este término puede ser confuso, ya que B_t no es el producto exterior de g_t. Sin embargo, refleja el hecho de que la puntuación es el gradiente de una observación específica y B_t es el producto exterior medio de estos gradientes de observación específica.

En los parámetros que maximizan la función de verosimilitud, la puntuación media es nula. El máximo se produce donde la pendiente es cero, lo que significa que el gradiente, es decir, la puntuación media, es cero. Dado que la puntuación media es cero, el producto exterior de las puntuaciones, B_t, se convierte en la varianza de las puntuaciones. Es decir, en los valores de maximización de los parámetros, B_t es la varianza de las puntuaciones en la muestra.

La varianza de las puntuaciones proporciona información importante para localizar el máximo de la función de verosimilitud. En concreto, esta variación proporciona una medida de la curvatura de la función log-verosimilitud, similar al hessiano. Supongamos que todas las personas de la muestra tienen puntuaciones similares. Si esto sucede, la muestra contiene muy poca información. La función log-verosimilitud será bastante plana en esta situación, lo que refleja el hecho de que diferentes valores de los parámetros se ajustan a los datos de forma similar. El primer dibujo de la figura 8.7 ilustra esta situación: con una log-verosimilitud casi plana, diferentes valores de β dan valores similares de $LL(\beta)$. La curvatura es pequeña cuando la varianza de las puntuaciones es pequeña. Por el contrario, puntuaciones que difieren considerablemente entre observaciones indican que las observaciones son muy diferentes y que la muestra proporciona una cantidad considerable de información. La función log-verosimilitud es muy puntiaguda, reflejando el hecho de que la muestra ofrece buena información sobre los valores de β. Alejarse de los valores maximización de β provoca una gran pérdida de ajuste. El segundo panel de la figura 8.7 ilustra esta situación. La curvatura es grande cuando la varianza de las puntuaciones es alta.

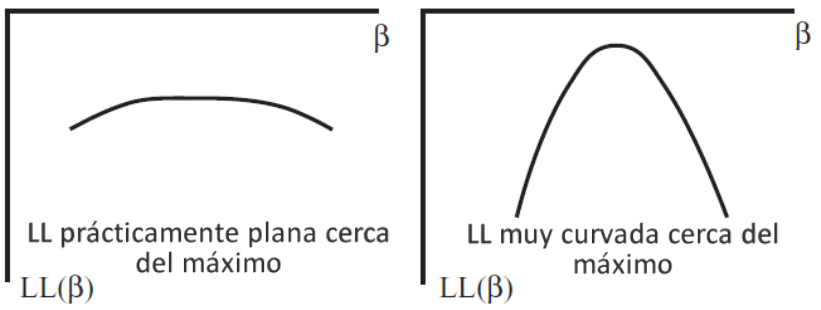


Figura 8.7. Forma de la función log-verosimilitud cerca del máximo.

Estas ideas sobre la varianza de las puntuaciones y su relación con la curvatura de la función log-verosimilitud se formalizan en la famosa *identidad de información* (*information identity*). Esta igualdad afirma que la covarianza de las puntuaciones en los verdaderos parámetros es igual a la negativa del hessiano. Demostraremos esta igualdad en la última sección de este capítulo; Theil (1971) y Ruud (2000) también proporcionan pruebas útiles y heurísticas. Sin embargo, incluso sin pruebas, intuitivamente tiene sentido que la varianza de las puntuaciones proporcione información sobre la curvatura de la función log-verosimilitud.

Berndt, Hall, Hall y Hausman (1974), en lo sucesivo BHHH (y comúnmente pronunciado triple-B H) propusieron el uso de esta relación en la búsqueda numérica del máximo de la función log-verosimilitud. En concreto, el procedimiento BHHH utiliza B_t en la rutina de optimización en lugar de $-H_t$. Cada iteración se define por

$$\beta_{t+1} = \beta_t + \lambda B_t^{-1} g_t.$$

Cada paso de este procedimiento se define igual que en el procedimiento NR, excepto que B_t se utiliza en lugar de $-H_t$. Teniendo en cuenta la explicación anterior acerca de cómo la varianza de las puntuaciones indican qué curvatura tiene la función de verosimilitud, reemplazar $-H_t$ por B_t tiene sentido.

El procedimiento BBBH tiene dos ventajas respecto al procedimiento NR:

1. B_t es mucho más rápido de calcular que H_t. En el procedimiento NR, las puntuaciones deben ser calculadas de todos modos para obtener el gradiente, por lo que el cálculo de B_t como el producto exterior medio de las puntuaciones apenas requiere tiempo adicional de computación. Por el contrario, el cálculo de H_t requiere el cálculo de las segundas derivadas de la función log-verosimilitud.
2. B_t necesariamente es definida positiva. Por consiguiente, el procedimiento BHHH garantiza un incremento de $LL(\beta)$ en cada iteración, incluso en partes convexas de la función. Utilizando la prueba dada anteriormente para el procedimiento NR cuando $-H_t$ es definida positiva, el tamaño del paso $\lambda B_t^{-1} g_t$ dado en cada iteración por el procedimiento BHHH incrementa $LL(\beta)$ para una λ suficientemente pequeña.

Nuestra exposición sobre la relación de la varianza de las puntuaciones con la curvatura de la función log-verosimilitud se puede establecer de forma un poco más precisa. Para un modelo correctamente especificado en los parámetros verdaderos, $B \to -H$ a medida que $N \to \infty$. Esta relación entre las dos matrices es una implicación de la identidad de información, expuesta con mayor detalle en la última sección. Esta convergencia sugiere que B_t puede considerarse como una aproximación a $-H_t$. A medida que el tamaño de la muestra aumenta, se espera que la aproximación sea mejor. Y la aproximación se puede esperar que sea mejor cerca de los verdaderos parámetros, donde la esperanza de la puntuación es cero y la identidad de información se cumple, que para valores de β que están lejos de los valores verdaderos. Es decir, se puede esperar que B_t sea una aproximación mejor cerca del máximo de $LL(\beta)$ que lejos del máximo.

BHHH tiene algunos inconvenientes. El procedimiento puede dar pasos pequeños que incrementan $LL(\beta)$ muy poco, especialmente cuando el proceso iterativo está lejos del máximo. Este comportamiento puede surgir porque B_t no es una buena aproximación de $-H_t$ lejos del valor verdadero, o si $LL(\beta)$ es altamente no cuadrática en el área donde está ocurriendo el problema. Si la función es altamente no cuadrática, el procedimiento NR no funciona bien, tal y como se explicó anteriormente; puesto que BHHH es una aproximación de NR, BHHH tampoco funcionará bien en estas circunstancias incluso aunque B_t sea una buena aproximación de $-H_t$.

8.3.3 BHHH-2

El procedimiento BHHH se basa en la matriz B_t, que como hemos descrito, captura la covarianza de las puntuaciones cuando la puntuación media es igual a cero (es decir, en el valor de maximización de β). Cuando el proceso iterativo no está en el máximo, la puntuación media no es cero y B_t no representa la covarianza de las puntuaciones.

Una variante del procedimiento BHHH se obtiene restando la puntuación media antes de calcular el producto exterior. Para cualquier nivel de la puntuación media, la covarianza de las puntuaciones entre decisores de la muestra es

$$W_t = \sum_n \frac{(s_n(\beta_t) - g_t)(s_n(\beta_t) - g_t)'}{N}$$

donde el gradiente g_t es la puntuación media. W_t es la covarianza de las puntuaciones alrededor de su media y B_t es el producto externo promedio de las puntuaciones. W_t y B_t son iguales cuando el gradiente medio es cero (es decir, en el valor de maximización de β), pero difieren en caso contrario.

El procedimiento de maximización puede utilizar W_t en lugar de B_t:

$$\beta_{t+1} = \beta_t + \lambda W_t^{-1} g_t.$$

Este procedimiento, que denomino BHHH-2, tiene las mismas virtudes de BHHH. W_t es necesariamente definida positiva, ya que es una matriz de covarianza, por lo que el procedimiento garantiza un aumento de $LL(\beta)$ en cada iteración. A su vez, para un modelo especificado correctamente en los parámetros verdaderos, $W \to -H$ a medida que $N \to \infty$, de modo que W_t puede considerarse como una aproximación a $-H_t$. La identidad de información establece esta equivalencia, como lo hace para B.

Para βs cercanas al valor de maximización, BHHH y BHHH-2 dan casi los mismos resultados. Pero pueden diferir en gran medida en valores lejanos al máximo. La experiencia indica, sin embargo, que los dos métodos son bastante similares en el sentido de que, o bien ambos trabajan con eficacia para una función de probabilidad dada, o ninguno de los dos lo hace. El valor principal de BHHH-2 es pedagógico y permite dilucidar la relación entre la covarianza de las puntuaciones y el producto exterior medio de las puntuaciones. Esta relación es crítica en el análisis de la identidad de información que se hace en la sección 8.7.

8.3.4 Ascenso más rápido (steepest ascent)

Este procedimiento se define por la fórmula iterativa siguiente

$$\beta_{t+1} = \beta_t + \lambda g_t.$$

La matriz que caracteriza este procedimiento es la matriz identidad I. Dado que I es definida positiva, el método garantiza un incremento en la función de verosimilitud en cada iteración. Este procedimiento recibe el nombre de "ascenso más rápido" (*"steepest ascent"*) debido a que proporciona el mayor incremento posible de $LL(\beta)$ para la distancia existente entre β_t y β_{t+1}, por lo menos para una distancia suficientemente pequeña. Cualquier otro paso de igual distancia proporciona menor incremento. Este hecho se demuestra de la siguiente manera. Considere una expansión en series de Taylor de primer orden de $LL(\beta_{t+1})$ alrededor de $LL(\beta_t)$: $LL(\beta_{t+1}) = LL(\beta_t) + (\beta_{t+1} - \beta_t)g_t$. Maximice esta expresión para $LL(\beta_{t+1})$ sujeta a que la distancia euclidiana entre $LL(\beta_{t+1})$ y $LL(\beta_{t+1})$ sea $\sqrt{k}$. Es decir, maximice con la restricción $(\beta_{t+1} - \beta_t)'(\beta_{t+1} - \beta_t) = k$. El lagrangiano es

$$L = LL(\beta_t) + (\beta_{t+1} - \beta_t)g_t - \frac{1}{2\lambda}\left[(\beta_{t+1} - \beta_t)'(\beta_{t+1} - \beta_t) - k\right],$$

y tenemos

$$\frac{\partial L}{\partial \beta_{t+1}} = g_t - \frac{1}{\lambda}(\beta_{t+1} - \beta_t) = 0,$$

$$\beta_{t+1} - \beta_t = \lambda g_t,$$

$$\beta_{t+1} = \beta_t + \lambda g_t,$$

que es la fórmula que define el procedimiento de ascenso más rápido.

A primera vista, uno podría pensar que el método de ascenso más rápido es el mejor procedimiento alcanzable, ya que da el mayor incremento posible de la función log-verosimilitud en cada paso. Sin embargo, esta propiedad del método es en realidad menos fuerte de lo que parece. Tenga en cuenta que el método se basa en una aproximación de primer orden que sólo es exacta en las cercanías de β_t. La consecuencia estrictamente correcta basada en el resultado anterior sería afirmar que existe una cierta distancia suficientemente pequeña para la cual el método de ascenso más rápido da el mayor incremento posible. Esta distinción es fundamental. La experiencia indica que los tamaños de paso a menudo resultan ser muy pequeños con este método. El hecho de que el ascenso sea mayor que para cualquier otro método para una determinada distancia no es especialmente útil cuando los pasos dados son tan pequeños. Por lo general, BHHH y BHHH-2 convergen más rápidamente que el método de ascenso más rápido.

8.3.5 DFP y BFGS

Los métodos Davidon-Fletcher-Powell (DFP) y Broyden-Fletcher-Goldfarb-Shanno (BFGS) calculan una aproximación del hessiano utilizando para ello información de más de un punto de la función de verosimilitud. Recordemos que el método NR utiliza el hessiano en β_t para determinar el paso a dar para llegar a β_{t+1}, mientras que los métodos BHHH y BHHH-2 utilizan las puntuaciones en β_t para aproximar el hessiano. En estos procedimientos sólo se utiliza información en β_t para determinar el paso. Si la función es cuadrática, la información en un punto de la función proporciona toda la información necesaria sobre la forma de la función. Por lo tanto, estos métodos funcionan bien cuando la función log-verosimilitud tiene una forma aproximadamente cuadrática. Por el contrario, los procedimientos DFP y BFGS utilizan información en varios puntos para obtener una descripción sobre la curvatura de la función log-verosimilitud.

El hessiano es la matriz de las segundas derivadas. Como tal, informa sobre cuanto cambia la curva a medida que uno se mueve a lo largo de la misma. El hessiano se define para movimientos infinitesimales. Dado que estamos interesados en dar grandes pasos en cada iteración de los algoritmos de búsqueda, comprender cómo cambia la pendiente para movimientos no infinitesimales es útil. Podemos definir un arco-hessiano sobre la base de cómo cambia el gradiente de un punto a otro. Por ejemplo, supongamos que para la función $f(x)$ la pendiente en $x = 3$ es 25 y en $x = 4$ es 19. El cambio en la pendiente para un cambio en x de una unidad es -6. En este caso, el arco-hessiano es -6, representando el cambio producido en la pendiente al dar un paso de $x = 3$ a $x = 4$.

Los procedimientos DFP y BFGS utilizan estos conceptos para aproximar el hessiano. En cada iteración del proceso se calcula el gradiente. La diferencia en el gradiente entre los diversos puntos que se han alcanzado se utiliza para calcular el arco-hessiano sobre estos puntos. Este arco-hessiano refleja el cambio que se produce en el gradiente para el movimiento real sobre la curva, en oposición al hessiano, que simplemente refleja el cambio en la pendiente para pasos infinitesimalmente pequeños alrededor de ese punto. Cuando la función log-verosimilitud no es cuadrática, el hessiano en cualquier punto proporciona muy poca información sobre la forma de la función. El arco-hessiano proporciona mejor información en estos casos.

En cada iteración, los procedimientos DFP y BFGS actualizan el arco-hessiano utilizando la información que se obtiene en el nuevo punto, es decir, utilizando el nuevo gradiente. Los dos procedimientos difieren en cómo se realiza la actualización; véase Greene (2000) para más detalles. Ambos métodos son extremadamente efectivos - por lo general mucho más eficientes que NR, BHHH, BHHH-2 o el método de ascenso más rápido. BFGS es una versión refinada de DFP. Mi experiencia indica que casi siempre funciona mejor. BFGS es el algoritmo predeterminado en las rutinas de optimización de muchos paquetes de software comercial.

8.4 Criterio de Convergencia

En teoría, el máximo de $LL(\beta)$ se produce cuando el vector gradiente es cero. En la práctica, el vector gradiente calculado nunca es exactamente igual a cero: se puede estar muy cerca, pero una serie de cálculos en una computadora no pueden producir un resultado exactamente igual a cero (a no ser, claro está, que el resultado se fije a cero a través de un operador booleano o mediante la multiplicación por cero, algo que no sucede en el cálculo del gradiente). Por ello surge la siguiente pregunta: ¿cuándo estamos suficientemente cerca del máximo para justificar la detención del proceso iterativo?

El estadístico $m_t = g_t'(-H_t^{-1})g_t$ se utiliza a menudo para evaluar la convergencia. El investigador especifica un valor pequeño para m, como $\breve{m} = 0.00001$, y determina en cada iteración si $g_t'(-H_t^{-1})g_t < \breve{m}$. Si esta desigualdad se cumple, el proceso iterativo se detiene y los parámetros en esa iteración se consideran los valores convergentes, es decir, las estimaciones. Para procedimientos distintos al método NR que utilizan una aproximación del hessiano en el proceso iterativo, dicha aproximación se utiliza en el estadístico de convergencia a fin de evitar el cálculo del hessiano real. Cerca del máximo, que es donde el criterio se vuelve relevante, se espera que todas las aproximaciones del hessiano que hemos visto sean similares al hessiano real.

m_t es el estadístico para la hipótesis de que todos los elementos del vector gradiente son cero. El estadístico se distribuye chi-cuadrado con K grados de libertad. Sin embargo, el criterio de convergencia $\breve{m}$ por lo general se suele fijar mucho más restrictivo (es decir, menor) que el valor crítico de una distribución chi-cuadrada con niveles estándar de significación, a fin de asegurar que los parámetros estimados son muy próximos a los valores de maximización. Por lo general, la hipótesis de que los elementos del gradiente son cero no puede ser rechazada para una zona bastante amplia alrededor del máximo. Esta distinción puede ilustrarse para un coeficiente estimado que tenga un estadístico-t de 1.96. En este caso, la hipótesis no podría ser rechazada si este coeficiente tiene un valor entre cero y el doble de su valor estimado. Sin embargo, no deseamos que la convergencia se defina por el hecho de haber llegado a un valor del parámetro dentro de este rango tan amplio.

Es tentador ver los pequeños cambios en β_t entre una iteración y la siguiente, y en consecuencia los pequeños aumentos en $\mathrm{LL}(\beta_t)$, como una prueba de que se ha obtenido la convergencia. Sin embargo, como se dijo anteriormente, los procedimientos iterativos pueden producir pasos pequeños debido a que la función de verosimilitud no está cerca de ser cuadrática y no porque estemos cerca del máximo. Cambios pequeños en β_t y en $\mathrm{LL}(\beta_t)$ acompañados de un vector gradiente que no esté cerca de cero indican que la rutina numérica no es eficaz en la búsqueda del máximo.

La convergencia a veces se evalúa sobre la base del propio vector gradiente en lugar de a través del estadístico de prueba m_t. Existen dos procedimientos: (1) determinar si cada elemento del vector gradiente es menor en magnitud a un cierto valor especificado por el investigador y (2) dividir cada elemento del vector gradiente por el elemento correspondiente de β, y determinar si cada uno de estos cocientes es menor en magnitud a un valor especificado por el investigador. El segundo enfoque

normaliza las unidades de los parámetros, que están determinadas por las unidades de las variables que entran en el modelo.

8.5 Máximo local y máximo global

Todos los métodos que hemos visto son susceptibles de converger en un máximo local que no es el máximo global que realmente buscamos, como se muestra en la figura 8.8. Cuando la función log-verosimilitud es globalmente cóncava, como sucede en un modelo logit con utilidad lineal en los parámetros, sólo existe un máximo y la cuestión no se plantea. Sin embargo, la mayoría de los modelos de elección discreta no son globalmente cóncavos.

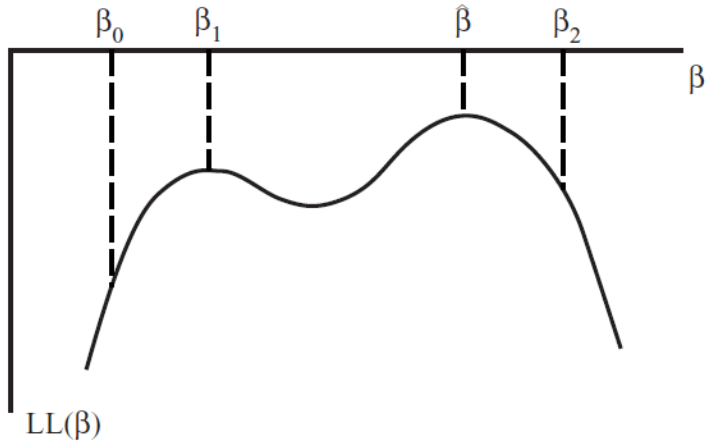


Figura 8.8. Máximo local comparado con el máximo global

Una manera de investigar el problema es utilizar varios valores de inicio en los algoritmos y observar si la convergencia se produce en los mismos valores de los parámetros. Por ejemplo, en la figura 8.8, empezar en β_0 conducirá a la convergencia en β_1. Si el investigador no probase otros valores de inicio, podría creer erróneamente que ya ha alcanzado el máximo de $\mathrm{LL}(\beta)$. Sin embargo, empezando en β_2, la convergencia se logra en $\hat{\beta}$. Mediante la comparación entre $\mathrm{LL}(\beta_1)$ y $\mathrm{LL}(\hat{\beta})$, el investigador observa que β_1 no es el valor que maximiza la verosimilitud. Liu y Mahmassani (2000) proponen una manera de seleccionar los valores de inicio que obliga al investigador a establecer límites superiores e inferiores de cada parámetro, y posteriormente seleccionar al azar los valores de inicio dentro de esos límites.

8.6 Varianza de las estimaciones

En cursos estándar de econometría, se muestra que para un modelo correctamente especificado,

$$\sqrt{N}(\hat{\beta} - \beta^*) \xrightarrow{d} N(0, (-\boldsymbol{H})^{-1})$$

a medida que $N \to \infty$, donde β^* es el vector de los parámetros verdaderos, $\hat{\beta}$ es el estimador de máxima verosimilitud y $\boldsymbol{H}$ es el hessiano esperado en la población. El negativo del hessiano esperado, $-\boldsymbol{H}$, a menudo se llama matriz de información. Expresado en palabras, la distribución en la muestra de la diferencia entre el estimador y el valor verdadero, normalizado para el tamaño de muestra, converge asintóticamente a una distribución normal centrada en cero y con covarianza igual a la inversa de la matriz de información, $-\boldsymbol{H}^{-1}$. Dado que la covarianza asintótica de $\sqrt{N}(\hat{\beta} - \beta^*)$ es $-\boldsymbol{H}^{-1}$, la covarianza asintótica de $\hat{\beta}$ es $-\boldsymbol{H}^{-1}/N$.

El tipo de letra negrita en estas expresiones indica que $\boldsymbol{H}$ es la media en la población, en contraposición a H, que es el hessiano promedio en la muestra. El investigador calcula la covarianza asintótica usando

H como una estimación de $\boldsymbol{H}$. Es decir, la covarianza asintótica de $\hat{\beta}$ se calcula como $-H^{-1}/\mathrm{N}$, donde H se evalúa en $\hat{\beta}$.

Recordemos que W es la covarianza de las puntuaciones en la muestra. En los valores de maximización de β, B también es la covarianza de las puntuaciones. Debido a la identidad de información que acabamos de exponer, y que se explica en la última sección, $-H$, que es el (negativo de la) matriz hessiana promedio en la muestra, converge a la covarianza de las puntuaciones de un modelo correctamente especificado en los parámetros verdaderos. En el cálculo de la covarianza asintótica de las estimaciones $\hat{\beta}$, cualquiera de estas tres matrices se puede utilizar como una estimación de $-\boldsymbol{H}$. La varianza asintótica de $\hat{\beta}$ se calcula como W^{-1}/N, B^{-1}/N o $-H^{-1}/N$, donde cada una de estas matrices se evalúa en $\hat{\beta}$.

Si no se especifica correctamente el modelo, entonces la covarianza asintótica de $\hat{\beta}$ es más compleja. En concreto, para cualquier modelo para el cual la puntuación esperada sea cero en los parámetros verdaderos,

$$\sqrt{N}\left(\hat{\beta} - \beta^*\right) \xrightarrow{d} N(0, \boldsymbol{H}^{-1}\boldsymbol{V}\boldsymbol{H}^{-1}),$$

donde $\boldsymbol{V}$ es la varianza de las puntuaciones en la población. Cuando el modelo está correctamente especificado, la matriz $-\boldsymbol{H} = \boldsymbol{V}$ de acuerdo a lo establecido por la identidad de información, de tal manera que $\boldsymbol{H}^{-1}\mathbf{V}\boldsymbol{H}^{-1} = -\boldsymbol{H}^{-1}$ y tenemos la fórmula para un modelo especificado correctamente. Sin embargo, si no se especifica correctamente el modelo, no se produce esta simplificación. La varianza asintótica de $\hat{\beta}$ es $\boldsymbol{H}^{-1}\mathbf{V}\boldsymbol{H}^{-1}/N$. Esta matriz se denomina *matriz de covarianza robusta (robust covariance matrix)*, ya que es válida tanto si el modelo se ha especificado correctamente como si no.

Para estimar la matriz de covarianza robusta, el investigador debe calcular el hessiano H. Si para alcanzar la convergencia se utiliza un procedimiento diferente al NR, no es necesario calcular el hessiano en cada iteración; sin embargo, éste debe ser calculado en la iteración final. Acto seguido, la covarianza asintótica se calcula como $H^{-1}\mathrm{W}H^{-1}$, o usando B en lugar de W. Esta fórmula se denomina en ocasiones el estimador "sándwich" de la covarianza, debido a que la inversa del hessiano aparece en ambos lados.

Una forma alternativa de estimar la matriz de covarianza es a través de *bootstrapping*, tal y como sugirió Efron (1979). De acuerdo a este procedimiento, el modelo se re-estima numerosas ocasiones usando diferentes muestras tomadas de la muestra original. Denominemos la muestra original como A, formada por los decisores indexados por $n = 1, \dots, N$. Es decir, la muestra original consta de N observaciones. El estimador que se obtiene de esta muestra es $\hat{\beta}$. El procedimiento de bootstrapping consiste en los siguiente pasos:

1. Seleccione aleatoriamente una muestra de N observaciones *con reemplazo* a partir de la muestra original A. Dado que el muestreo es con reemplazo, algunos decisores pueden estar representados más de una vez en esta nueva muestra y otros pueden no estar representados ninguna. Esta nueva muestra es del mismo tamaño que la muestra original, pero tiene una composición diferente a la original, ya que algunos decisores están repetidos y otros no están incluidos.
2. Re-estime el modelo usando esta nueva muestra y etiquete la estimación como β_{r}, con $\mathrm{r} = 1$ para esta primera nueva muestra.
3. Repita los pasos 1 y 2 varias veces, obteniendo estimaciones β_{r} para $\mathrm{r} = 1, \dots, ,\mathrm{R}$ donde R es el número de veces que la estimación se repite con una nueva muestra.
4. Calcule la covarianza de las estimaciones resultantes en torno a la estimación original: $\mathrm{V} = \frac{1}{\mathrm{R}}\sum_{\mathrm{r}}(\beta_{\mathrm{r}} - \hat{\beta})(\beta_{\mathrm{r}} - \hat{\beta})'$.

Esta V es una estimación de la matriz de covarianza asintótica. La varianza de muestreo de cualquier estadístico que esté basado en los parámetros se calcula de manera similar. Para estadísticos escalares $\mathrm{t}(\beta)$, la varianza muestral es $\sum_{\mathrm{r}}(\mathrm{t}(\beta_{\mathrm{r}}) - \mathrm{t}(\hat{\beta}))^2/\mathrm{R}$.

La lógica del procedimiento es la siguiente. La covarianza muestral de un estimador es, por definición, una medida de la variación de las estimaciones cuando se obtienen diferentes muestras de la población. Nuestra muestra original es una muestra de la población. Sin embargo, si esta muestra es lo suficientemente grande, es probable que sea similar a la población, de tal manera que extraer valores al azar de ella sea similar a extraer valores al azar de la población en sí misma. El método de *bootstrap* hace justamente eso: extrae valores al azar de la muestra original, con reemplazo, como un proxy de extraer valores al azar de la propia población. Las estimaciones obtenidas en las muestras creadas mediante *bootstrap* proporcionan información sobre la distribución de las estimaciones que se obtendría si realmente hubiésemos extraído muestras alternativas de la población.

La ventaja del *bootstrap* es que es conceptualmente sencillo y no se basa en fórmulas que son válidas asintóticamente pero que podrían no ser especialmente precisas para un tamaño de muestra dado. Su desventaja es que es un método computacionalmente costoso, ya que requiere la estimación del modelo numerosas veces. Efron y Tibshirant (1993) y Vinod (1993) proporcionan un estudio sobre el tema, junto con aplicaciones reales.

8.7 Identidad de información

La identidad de información establece que, para un modelo correctamente especificado en los parámetros verdaderos, $\boldsymbol{V} = -\boldsymbol{H}$, donde $\boldsymbol{V}$ es la matriz de covarianza de las puntuaciones en la población y $\boldsymbol{H}$ es el hessiano promedio de la población. La puntuación de una persona es el vector de las primeras derivadas de $ln\ P(\beta)$ de esa persona respecto a los parámetros, y el hessiano es la matriz de segundas derivadas. La identidad de información establece que, en la población, la matriz de covarianza de las primeras derivadas es igual a la matriz promedio de las segundas derivadas (en realidad, el negativo de esta matriz). Esto es un hecho sorprendente, no es algo que podríamos esperar o incluso creer si no tuviésemos una prueba. Tiene implicaciones en toda la econometría. Las implicaciones que hemos utilizado en las secciones anteriores de este capítulo se demuestran fácilmente a partir de esta igualdad. En particular:

(1) En el valor maximización de β, $W \to -H$ a medida que $N \to \infty$, donde W es la covarianza de las puntuaciones en la muestra y H es el promedio en la muestra del hessiano de cada observación. A medida que el tamaño de la muestra aumenta, la covarianza de la muestra se aproxima a la covarianza de la población: $W \to \mathbf{V}$. De forma similar, el promedio en la muestra del hessiano se acerca al promedio de la población: $H \to \mathbf{H}$. Dado que $\mathbf{V} \to -\mathbf{H}$ de acuerdo a la identidad de información, W se aproxima a la misma matriz a la que se aproxima $-H$, de modo que se aproximan entre sí.

(2) En el valor maximización de β, $B \to -H$ a medida que $N \to \infty$, donde B es el promedio en la muestra del producto exterior de las puntuaciones. En $\hat{\beta}$, la puntuación media en la muestra es igual a cero, por lo que B es igual a W. El resultado para W también aplica a B.

A continuación, vamos a demostrar la identidad de información. Necesitamos para ello ampliar nuestra notación con el fin de abarcar la población en lugar de limitarnos a la muestra. Sea $P_i(x, \beta)$ la probabilidad de que una persona que se enfrenta a unas variables explicativas x escoja la alternativa i dados los parámetros β. De las personas de la población que se enfrentan a las variables x, la proporción que elige la alternativa i es esta probabilidad calculada en los parámetros verdaderos: $S_i(x) = P_i(x, \beta^*)$ donde β^* son los parámetros verdaderos. Consideremos ahora el gradiente de $\ln P_i(x, \beta)$ respecto a β. El gradiente medio en la población es

(8.2) $$g = \int \sum_i \frac{\partial \ln P_i(x,\beta)}{\partial \beta} S_i(x) f(x) dx,$$

donde $f(x)$ es la densidad de las variables explicativas en la población. Esta expresión se puede explicar de la siguiente manera. El gradiente para las personas que se enfrentan a x y eligen i es $\partial \ln \mathrm{P_i(x,\beta)} / \partial \beta$. El gradiente medio es el promedio de este término entre todos los valores de x y todas las alternativas i. La proporción de personas que se enfrentan a un valor dado de x está dada por $f(x)$ y la proporción de personas que se enfrentan a esta x que eligen i es $S_i(x)$. Así que $S_i(x)f(x)$ es la proporción de la población que se enfrenta a x y elige i, por lo que tienen gradiente $\partial \ln \mathrm{P_i(x,\beta)} / \partial \beta$ Sumando este término sobre todos los valores de i e integrando sobre todos los valores de x (suponiendo que las xs son continuas) nos da el gradiente medio, tal y como se expresa en (8.2).

El gradiente medio en la población es igual a cero en los parámetros verdaderos. Este hecho puede ser considerado como la definición de parámetros verdaderos o el resultado de un modelo correctamente especificado. Además, sabemos que $S_i(x) = P_i(x, \beta^*)$. Sustituyendo estos hechos en (8.2), tenemos

$$0 = \int \sum_i \frac{\partial \ln P_i(x,\beta)}{\partial \beta} P_i(x,\beta) f(x) dx,$$

donde todas las funciones se evalúan en β^*. Derivamos esta ecuación respecto a los parámetros:

$$0 = \int \sum_i \left(\frac{\partial^2 \ln P_i(x,\beta)}{\partial \beta \partial \beta'} P_i(x,\beta) + \frac{\partial \ln P_i(x,\beta)}{\partial \beta} \frac{\partial P_i(x,\beta)}{\partial \beta'} \right) f(x) dx.$$

Dado que $\partial \ln \mathrm{P} / \partial \beta = (1/\mathrm{P})\ \partial \mathrm{P} / \partial \beta$ por las reglas de derivación, podemos reemplazar $[\partial \ln \mathrm{P_i(x,\beta)} / \partial \beta']\mathrm{P_i(x,\beta)}$ por $\partial \mathrm{P_i(x,\beta)} / \partial \beta'$en el último término entre paréntesis:

$$0 = \int \sum_i \left(\frac{\partial^2 \ln P_i(x,\beta)}{\partial \beta \partial \beta'} P_i(x,\beta) + \frac{\partial \ln P_i(x,\beta)}{\partial \beta} \frac{\partial \ln P_i(x,\beta)}{\partial \beta'} P_i(x,\beta) \right) f(x) dx.$$

Reorganizando

$$-\int \sum_i \frac{\partial^2 \ln P_i(x,\beta)}{\partial \beta \partial \beta'} P_i(x,\beta) f(x) dx = \int \sum_i \frac{\partial \ln P_i(x,\beta)}{\partial \beta} \frac{\partial \ln P_i(x,\beta)}{\partial \beta'} P_i(x,\beta) f(x) dx$$

Dado que todos los términos se evalúan en los parámetros verdaderos, podemos sustituir $P_i(x,\beta)$ con $S_i(x)$ para obtener

$$-\int \sum_i \frac{\partial^2 \ln P_i(x,\beta)}{\partial \beta \partial \beta'} S_i(x) f(x) dx = \int \sum_i \frac{\partial \ln P_i(x,\beta)}{\partial \beta} \frac{\partial \ln P_i(x,\beta)}{\partial \beta'} S_i(x) f(x) dx$$

El lado izquierdo es el negativo del hessiano promedio de la población, $-\mathbf{H}$. El lado derecho es el promedio del producto exterior del gradiente, que es la covarianza del gradiente, $\mathbf{V}$, ya que el gradiente medio es cero. Por lo tanto, $-\mathbf{H} = \mathbf{V}$, la identidad de información. Como se ha indicado, la matriz $-\mathbf{H}$ a menudo es llamada matriz de información.

9

Extrayendo valores de densidades

9.1 Introducción

La simulación de un estadístico consiste en extraer valores al azar de una densidad de probabilidad, calcular el estadístico para cada valor extraído y promediar los resultados. En todos los casos, el investigador quiere calcular un promedio de tipo $\bar{t} = \int t(\varepsilon) f(\varepsilon) d\varepsilon$, donde $t(\cdot)$ es el estadístico de interés y $f(\cdot)$ es una densidad de probabilidad. Para aproximar esta media a través de simulación, el investigador debe poder extraer valores de la densidad $f(\cdot)$. Para algunas densidades, esta tarea es simple. Sin embargo, en muchas situaciones, el cómo se extraen valores de la densidad correspondiente no es algo inmediato. Incluso con densidades simples, puede haber formas de extraer valores que proporcionen una mejor aproximación a la integral que una secuencia puramente aleatoria de extracciones.

En este capítulo exploramos estos temas. En las primeras secciones, se describen los métodos más importantes que han sido desarrollados para extraer valores puramente aleatorios de varias clases de densidades. Estos métodos se presentan de forma progresiva, empezando por procedimientos simples que funcionan para unas pocas densidades especialmente convenientes, pasando posteriormente a tratar métodos cada vez más complejos que trabajan con densidades menos convenientes. La exposición culmina con el algoritmo Metropolis-Hastings, que se puede utilizar con (prácticamente) cualquier densidad. El capítulo finaliza regresando a la cuestión de si es posible obtener - y cómo - una secuencia de valores que proporcionen una mejor aproximación a la integral correspondiente que una secuencia puramente aleatoria. Expondremos los métodos de antitéticos, el muestreo sistemático y las secuencias de Halton, y mostraremos el beneficio que este tipo de extracción de valores proporciona en la estimación de los parámetros de un modelo.

9.2 Extracción de valores aleatorios

9.2.1 Distribuciones normales y uniformes estándar

Si el investigador quiere extraer un valor aleatorio de una densidad normal estándar (es decir, una normal con media cero y varianza unitaria) o de una densidad uniforme estándar (uniforme entre 0 y 1), el proceso es muy simple desde una perspectiva de programación. La mayoría de los paquetes de software estadístico contienen generadores de números aleatorios para estas densidades. El investigador simplemente debe llamar a estas rutinas para obtener una secuencia de valores al azar. En

las siguientes secciones nos referiremos a un valor extraído al azar de una densidad normal estándar como η y a una extracción de una densidad uniforme estándar como μ.

Los valores generados por estas rutinas son en realidad números *pseudo-aleatorios*, porque nada de lo que una computadora puede hacer es verdaderamente aleatorio. Hay muchas cuestiones que intervienen en el diseño de estas rutinas. El objetivo de su diseño es producir números que exhiban las propiedades de extracciones al azar. En qué medida se logra este objetivo depende, por supuesto, de cómo se definen las propiedades de una extracción "al azar". Estas propiedades son difíciles de definir con precisión, ya que el azar es un concepto teórico que no tiene un reflejo en el mundo real. Desde una perspectiva práctica, mi consejo es el siguiente: a menos que uno esté dispuesto a pasar un tiempo considerable investigando y resolviendo (de hecho, diría re-resolviendo) estos temas, probablemente la mejor idea es usar las rutinas disponibles en lugar de programar una nueva.

9.2.2 Transformaciones de la normal estándar

Algunas variables aleatorias son transformaciones de una normal estándar. Por ejemplo, un valor al azar de una densidad normal con media b y varianza s^2 se obtiene como $\varepsilon = b + s\eta$. Un valor al azar de una densidad log-normal se obtiene mediante una exponencial de una extracción de una densidad normal: $\varepsilon = e^{b+s\eta}$. Los momentos estadísticos de la log-normal son funciones de la media y la varianza de la normal exponenciada. En concreto, la media de ε es $\exp(b + (s^2/2))$ y su varianza es $\exp(2b + s^2)(\exp(s^2) - 1)$. Dados los valores de la media y la varianza de la log-normal, es posible calcular los valores correctos de b y s a utilizar en la transformación. Sin embargo, es más común tratar b y s como los parámetros de la log-normal, y calcular su media y su varianza a partir de estos parámetros.

9.2.3 Densidades acumulativas inversas para densidades univariadas

Considere una variable aleatoria con densidad $f(\varepsilon)$ y la correspondiente distribución acumulativa $F(\varepsilon)$. Si F es invertible (es decir, si se puede calcular F^{-1}) entonces es posible extraer valores al azar de ε a partir de valores al azar de una distribución uniforme estándar. Por definición, $F(\varepsilon) = k$ significa que la probabilidad de extraer un valor al azar igual o menor a ε es k, donde k está entre cero y uno. Un valor μ extraído al azar de una distribución uniforme estándar proporciona un número entre cero y uno. Podemos establecer $F(\varepsilon) = \mu$ y resolver para el correspondiente ε: $\varepsilon = F^{-1}(\mu)$. Cuando ε se obtiene de esta manera, la distribución acumulativa de los valores extraídos es igual a F, de tal manera que los valores son equivalentes a valores extraídos directamente de F. En la figura 9.1 puede observarse un ejemplo. Un valor μ^1 extraído de una distribución uniforme estándar se transforma en el valor de ε etiquetado como ε^1, valor en el que $F(\varepsilon^1) = \mu^1$.

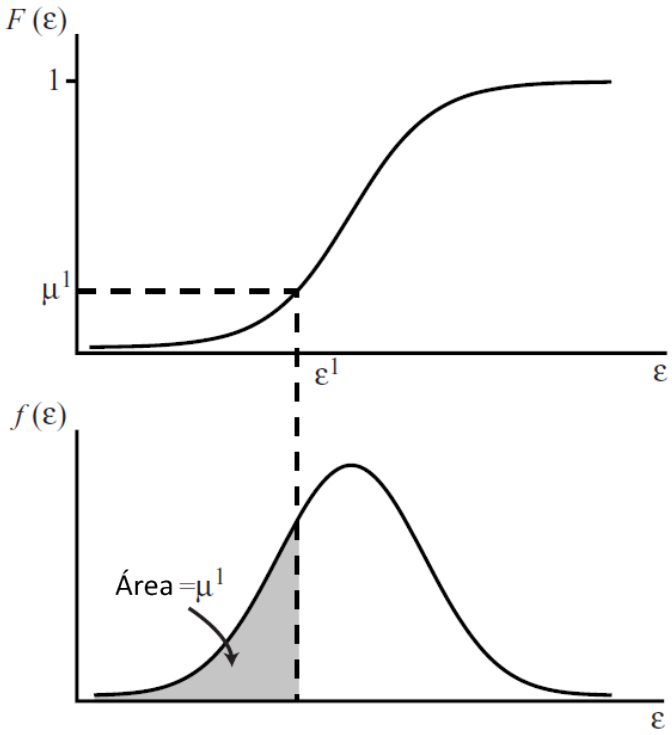


Figura 9.1 Extraemos un valor μ^1 de una uniforme y creamos $\varepsilon^1 = F^{-1}(\mu)$

La distribución de valor extremo, que es la base de los modelos logit multinomiales, nos proporciona un ejemplo de este método. La densidad es $f(\varepsilon) = \exp(-\varepsilon) \cdot \exp(-\exp(-\varepsilon))$ con una distribución acumulativa $F(\varepsilon) = \exp(-\exp(-\varepsilon))$. Podemos extraer un valor al azar de esta densidad de probabilidad como $\varepsilon = -\ln(-\ln \mu)$.

Tenga en cuenta que este procedimiento sólo funciona para distribuciones univariadas. Si ε tiene dos o más elementos, entonces $F^{-1}(\mu)$ no es único, ya que diversas combinaciones de los elementos de ε tienen la misma probabilidad acumulativa.

9.2.4 Densidades univariadas truncadas

Considere una variable aleatoria que va de a a b con densidad de probabilidad proporcional a $f(\varepsilon)$ dentro de este rango. Es decir, la densidad es $(1/k)f(\varepsilon)$ para $a \leq \varepsilon \leq b$ y 0 en caso contrario, donde k es la constante de normalización que asegura que la integral de la densidad resulta 1: $k = \int_a^b f(\varepsilon)d\varepsilon = F(b) - F(a)$. Podemos extraer un valor de esta densidad mediante la aplicación del procedimiento descrito en la Sección 9.2.3, asegurando que el valor está dentro del rango apropiado.

Para ello, extraiga μ de una densidad uniforme estándar. Calcule el promedio ponderado de $F(a)$ y $F(b)$ como $\bar{\mu} = (1-\mu)F(a) + \mu F(b)$. Luego calcule $\varepsilon = F^{-1}(\bar{\mu})$. Dado que $\bar{\mu}$ está entre $F(a)$ y $F(b)$, ε necesariamente está entre a y b. Básicamente, el valor extraído de μ determina hasta dónde llegar entre entre a y b. Tenga en cuenta que la constante de normalización k no se utiliza en los cálculos y por lo tanto no necesita ser calculada. La figura 9.2 ilustra el proceso.

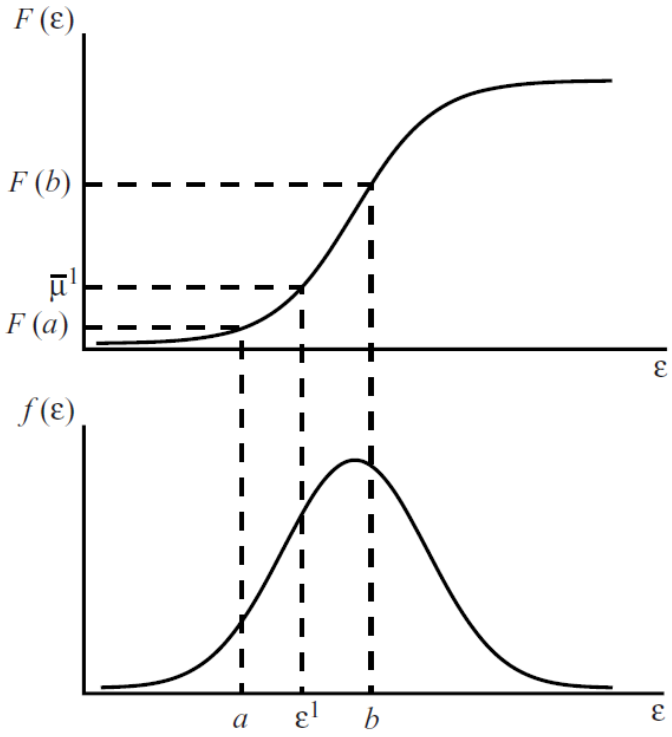


Figura 9.2. Un valor $\bar{\mu}^1$ entre $F(a)$ y $F(b)$ genera un valor ε^1 de $f(\varepsilon)$ entre a y b.

9.2.5 Transformación Choleski de normales multivariadas

Como se describe en la Sección 9.2.2, una normal univariada con media b y varianza s^2 se obtiene como $\varepsilon = b + s\eta$, donde η es una normal estándar. Es posible utilizar un procedimiento análogo para extraer valores al azar de una normal multivariada. Sea ε un vector con K elementos distribuidos $N(b, \Omega)$. Un factor Choleski de Ω se define como una matriz triangular inferior L tal que $LL' = \Omega$. A menudo recibe el nombre de raíz cuadrada generalizada de Ω o desviación estándar generalizada de ε. Con $K = 1$ y varianza s^2, el factor Choleski es s, que simplemente es la desviación estándar de ε. La mayoría de los paquetes de software estadístico y de manipulación de matrices tienen rutinas para calcular un factor Choleski de cualquier matriz simétrica definida positiva.

Es posible extraer un valor al azar ε de $N(b, \Omega)$ de la siguiente manera. Extraiga K valores de una normal estándar y etiquete el vector que contiene estos valores como $\eta = \langle \eta_1, \ldots, \eta_K \rangle'$. Calcule $\varepsilon = b + L\eta$. Podemos verificar las propiedades del ε resultante: se distribuye normalmente ya que la suma de normales es normal. Su media es b: $E(\varepsilon) = b + LE(\eta) = b$. Y su covarianza es Ω: $Var(\varepsilon) = E(L\eta(L\eta)') = LE(\eta\eta')L' = LVar(\eta)L' = LIL' = LL' = \Omega$.

Para ser concretos, considere un caso de ε tridimensional con media cero. Un valor de ε puede calcularse como

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} = \begin{pmatrix} s_{11} & 0 & 0 \\ s_{21} & s_{22} & 0 \\ s_{31} & s_{32} & s_{33} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix}$$

o

$$\varepsilon_1 = s_{11}\eta_1,$$

$$\varepsilon_2 = s_{21}\eta_1 + s_{22}\eta_2,$$

$$\varepsilon_3 = s_{31}\eta_1 + s_{32}\eta_2 + s_{33}\eta_3,$$

A través de estas relaciones vemos que $Var(\varepsilon_1) = s_{11}^2$, $Var(\varepsilon_2) = s_{21}^2 + s_{22}^2$ y $Var(\varepsilon_3) = s_{31}^2 + s_{32}^2 + s_{33}^2$. También que $\mathrm{Cov}(\varepsilon_1, \varepsilon_2) = \mathrm{s}_{11}\mathrm{s}_{21}$ y así sucesivamente. Los elementos ε_1 y ε_2 están correlacionados debido a la influencia común de η_1 en ambos. No están perfectamente correlacionados porque η_2 entra en ε_2 sin afectar ε_1. Un análisis similar aplica a ε_1 y ε_3, y a ε_2 y ε_3. En esencia, el factor Choleski expresa K términos correlacionados a partir de K componentes independientes, de manera que cada componente *carga* (afecta) de manera diferente cada término. Para cualquier patrón de covarianza, existe un conjunto de cargas de componentes independientes que reproduce esa covarianza.

9.2.6 Aceptación-rechazo para densidades multivariadas truncadas

El procedimiento descrito en la Sección 9.2.4 para la extracción de valores al azar de densidades truncadas sólo aplica a distribuciones univariadas. Con densidades multivariadas, extraer valores de distribuciones truncadas es más difícil. A continuación describimos un procedimiento de aceptación-rechazo que siempre es aplicable. Sin embargo, como veremos más adelante, este enfoque tiene desventajas que pueden hacer elegir al investigador otros métodos siempre que estén disponibles.

Supongamos que queremos extraer un valor al azar de una densidad multivariada $\mathrm{g}(\varepsilon)$ dentro del rango $a \leq \varepsilon \leq b$, donde a y b son vectores con la misma longitud de ε. Es decir, queremos extraer valores de $f(\varepsilon) = \frac{1}{k} g(\varepsilon)$ si $a \leq \varepsilon \leq b$, e iguales a cero en caso contrario, donde k es la constante de normalización. Podemos extraer valores de f simplemente extrayendo valores de g y reteniendo ("aceptando") los valores que se encuentran dentro del área de la distribución pertinente y descartando ("rechazando") los valores que se encuentran fuera de ese área. La ventaja de este procedimiento es que se puede aplicar siempre que sea posible extraer valores de la densidad no truncada. Es importante destacar que no es necesario conocer la constante de normalización k para la densidad truncada. Este hecho es útil debido a que suele ser difícil calcular esta constante de normalización.

La desventaja de este procedimiento es que el número de valores extraídos que son aceptados (es decir, el número de valores de f que obtenemos) no es fijo, sino que es a su vez un número aleatorio. Si se extraen R valores al azar de g, el número esperado de aceptaciones es kR. Esta esperanza no puede conocerse sin conocer k, que, como ya se ha mencionado, es por lo general difícil de calcular. Por tanto, es difícil determinar un número apropiado de valores a extraer de g. Más importante aún, el número real de aceptaciones generalmente diferirá de la cantidad esperada. De hecho, hay una probabilidad positiva de no obtener ninguna aceptación a partir de un número fijo de valores extraídos. Cuando el espacio de truncamiento es pequeño (o, más precisamente, cuando k es pequeño) no obtener ninguna aceptación, y por lo tanto no obtener ningún valor de la densidad truncada, es un evento probable.

Esta dificultad se puede sortear extrayendo valores de g hasta obtener un cierto número de valores aceptados. Es decir, en lugar de establecer de antemano el número de valores a extraer de g, el investigador puede establecer el número de valores extraídos de f que quiere obtener. Por supuesto, el investigador no sabe cuánto tiempo tardará en alcanzar el número establecido. En la mayoría de situaciones, es posible aplicar más fácilmente otros procedimientos para extraer valores de una densidad truncada multivariada. Sin embargo, es importante recordar que cuando ningún otro método parece posible con una distribución truncada, el procedimiento de aceptación-rechazo puede ser aplicado.

9.2.7 Muestreo por importancia

Supongamos que ε tiene una densidad $f(\varepsilon)$ de la cual no pueden extraerse valores fácilmente por otros procedimientos. Supongamos además que existe otra densidad $g(\varepsilon)$ de la que sí pueden extraerse valores al azar fácilmente. Podemos extraer valores de $f(\varepsilon)$ de la siguiente manera. Extraiga un valor de $g(\varepsilon)$ y etiquételo como ε^1. Pondere el valor por el factor $f(\varepsilon^1)/g(\varepsilon^1)$. Repita este proceso muchas veces. El conjunto de valores ponderados es equivalente a un conjunto de valores extraídos de f.

Para verificar este hecho, mostraremos que la distribución acumulativa de los valores ponderados extraídos de g es la misma que la distribución acumulativa de valores extraídos de f. Considere la proporción de valores extraídos de g que están por debajo cierto valor m, con cada valor ponderado por el factor f/g. Esta proporción es

$$\int \frac{f(\varepsilon)}{g(\varepsilon)} I(\varepsilon < m) g(\varepsilon) d\varepsilon = \int_{-\infty}^{m} \frac{f(\varepsilon)}{g(\varepsilon)} g(\varepsilon) d\varepsilon = \int_{-\infty}^{m} f(\varepsilon)\, d\varepsilon = F(m).$$

En simulación, los valores extraídos de una densidad se usan para calcular el promedio de un estadístico sobre dicha densidad. El muestreo por importancia puede ser visto como un cambio en el estadístico acompañado del correspondiente cambio en la densidad que haga que extraer valores de esa densidad sea fácil. Supongamos que queremos calcular $\int t(\varepsilon) f(\varepsilon) d\varepsilon$, pero resulta difícil extraer valores de f. Podemos multiplicar el integrando por $g \div g$ sin cambiar su valor, por lo que la integral es $\int t(\varepsilon)[f(\varepsilon)/g(\varepsilon)] g(\varepsilon) d\varepsilon$. Para simular la integral, extraemos valores de g, calculamos $t(\varepsilon)[f(\varepsilon)/g(\varepsilon)]$ para cada valor y promediamos los resultados. Simplemente hemos transformado la integral de modo que resulte más fácil para la simulación.

La densidad f recibe el nombre de densidad objetivo y g de densidad propuesta. Los requisitos para poder aplicar el muestreo por importancia son que (1) el ámbito de $g(\varepsilon)$ cubra el ámbito de f, de modo que cualquier ε que podría surgir con f también pueda surgir con g, y (2) el ratio $f(\varepsilon)/g(\varepsilon)$ debe ser finito para todos los valores posibles de ε, de modo que este ratio siempre pueda ser calculado.

Un ejemplo útil de muestreo por importancia lo encontramos en densidades normales truncadas multivariadas. Supongamos que queremos extraer valores al azar de $N(0, \Omega)$, pero de forma que cada elemento sea positivo (es decir, truncando la densidad de probabilidad por debajo de cero). La densidad es

$$f(\varepsilon) = \frac{1}{k(2\pi)^{\frac{1}{2}K} |\Omega|^{1/2}} e^{-\frac{1}{2}\varepsilon' \Omega^{-1} \varepsilon},$$

para $\varepsilon > 0$, y 0 en caso contrario, donde K es la dimensión de ε y k es la constante de normalización. (Asumimos para los fines de este ejemplo que k es conocido. En realidad, el cálculo de k puede a su vez requerir simulación). Extraer valores al azar de esta densidad es difícil debido a que los elementos de ε están correlacionados así como truncados. Sin embargo, podemos utilizar el procedimiento de la sección 9.2.4 para extraer valores al azar de normales independientes truncadas y luego aplicar muestreo por importancia para crear la correlación. Para ello, extraiga valores de K normales univariadas truncadas por debajo de cero, utilizando el procedimiento de la Sección 9.2.4. Estos valores constituyen colectivamente una extracción de un vector k-dimensional de ε desde el cuadrante positivo con densidad

$$g(\varepsilon) = \frac{1}{m(2\pi)^{\frac{1}{2}K}} e^{-\frac{1}{2}\varepsilon' \varepsilon},$$

donde $m = 1/2^K$. Para cada valor extraído, asigne el peso

$$\frac{f(\varepsilon)}{g(\varepsilon)} = \frac{m}{k}|\Omega|^{-1/2} e^{\varepsilon'(\Omega^{-1}-I)\varepsilon}.$$

Los valores ponderados son equivalentes a valores extraídos a partir de $N(0,\Omega)$ truncada por debajo de cero.

Una última reflexión: observe que el procedimiento de aceptación-rechazo descrito en la sección 9.2.6 es un tipo de muestreo por importancia. La distribución truncada es la densidad objetivo y la distribución no truncada es la densidad propuesta. Cada valor extraído de la densidad no truncada es ponderado por una constante si el valor está dentro del espacio de truncamiento y es ponderado por cero si el valor está fuera del espacio de truncamiento. Ponderar por una constante o por cero es equivalente a ponderar por uno (aceptar) o cero (rechazar).

*9.2.8 Muestreo de Gibbs (*Gibbs Sampling*)*

Para distribuciones multinomiales, a veces es difícil extraer valores al azar directamente de la densidad conjunta y sin embargo resulta sencillo hacerlo de la densidad condicionada de cada elemento dados los valores del resto de elementos. El muestro de Gibbs (el término aparentemente fue introducido por Geman y Geman, 1984) se puede utilizar en estas situaciones. Una explicación general del método lo proporcionan Casella y George (1992), explicación que el lector puede utilizar como complemento a la descripción más concisa que facilito a continuación.

Considere dos variables aleatorias ε_1 y ε_2. La generalización a una dimensión mayor es obvia. La densidad conjunta es $\mathrm{f}(\varepsilon_1,\varepsilon_2)$ y las densidades condicionadas son $\mathrm{f}(\varepsilon_1|\varepsilon_2)$ y $\mathrm{f}(\varepsilon_2|\varepsilon_1)$. El muestreo de Gibbs actúa extrayendo valores al azar iterativamente de las densidades condicionadas: extrae ε_1 condicionado a un valor de ε_2, extrae ε_2 condicionado a ese valor de ε_1, extrae de nuevo ε_1 condicionado al nuevo valor de ε_2 y así sucesivamente. Este proceso converge a valores extraídos de la densidad conjunta.

Para ser más precisos, el método consiste en: (1) Elija un valor inicial para ε_1, llamado ε_1^0. Cualquier valor con densidad distinta de cero puede ser elegido. (2) Extraiga un valor de ε_2, llamado ε_2^0, de la densidad $f(\varepsilon_2|\varepsilon_1^0)$. (3) Extraiga un valor de ε_1, llamado ε_1^1, a partir de $f(\varepsilon_1|\varepsilon_2^0)$. (4) Extraiga ε_2^1 de $f(\varepsilon_2|\varepsilon_1^1)$ y así sucesivamente. Los valores ε_1^t de $f(\varepsilon_1|\varepsilon_2^{t-1})$ y los valores ε_2^t de $f(\varepsilon_2|\varepsilon_1^t)$ constituyen una secuencia en t. Para una t suficientemente grande (es decir, para un número suficientemente grande de iteraciones) la secuencia converge a valores extraídos de la densidad conjunta $f(\varepsilon_1,\varepsilon_2)$.

Como ejemplo, considere dos normales con desviación estándar, independientes salvo por el hecho de que se truncan en base a su suma: $\varepsilon_1+\varepsilon_2 \le m$. La figura 9.3 representa la densidad truncada. Los círculos son contornos de la densidad no truncada y el área sombreada representa la densidad truncada. Para obtener las densidades condicionadas, considere primero las normales no truncadas. Puesto que las dos variables son independientes, la densidad condicionada de cada una es igual a su densidad no condicionada. Es decir, haciendo caso omiso del truncamiento, $\varepsilon_1|\varepsilon_2 \sim N(0,1)$. La regla de truncamiento que hemos definido es $\varepsilon_1+\varepsilon_2 \le m$ que puede ser reexpresada como $\varepsilon_1 \le m-\varepsilon_2$. Por lo tanto, $\varepsilon_1|\varepsilon_2$ se distribuye como una normal estándar univariada, truncada por encima de $m-\varepsilon_2$. Dado ε_2, un valor de ε_1 se obtiene usando el procedimiento de la sección 9.2.4: $\varepsilon_1 = \Phi^{-1}(\mu\Phi(m-\varepsilon_2))$, donde μ es un valor extraído de una uniforme estándar y $\Phi(\cdot)$ es la distribución normal estándar acumulativa. Valores de ε_2 condicionados a ε_1 se obtienen de forma análoga. Obtener valores secuencialmente de estas densidades condicionadas finalmente proporciona valores de la densidad truncada conjunta.

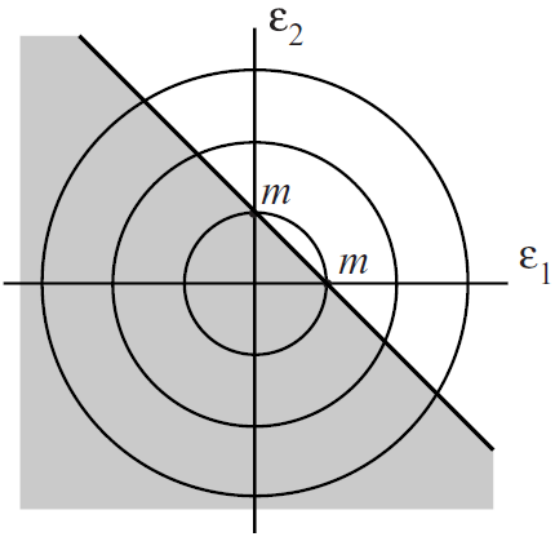


Figura 9.3 Densidad normal truncada

9.2.9 Algoritmo Metropolis-Hastings

Si todo lo demás falla, puede usarse el algoritmo Metropolis-Hastings (MH) para extraer valores de una densidad de probabilidad. El algoritmo fue inicialmente desarrollado por Metropolis et al. (1953) y generalizado posteriormente por Hastings (1970). Funciona como se indica a continuación. El objetivo es extraer valores al azar de $f(\varepsilon)$:

1. Empiece con un valor del vector ε, etiquetado ε^0.
2. Elija un valor de prueba de ε^1 como $\tilde{\varepsilon}^1 = \varepsilon^0 + \eta$, donde η se extrae de una distribución $g(\eta)$ que tiene media cero. Por lo general, se suele especificar una distribución normal para $g(\eta)$.
3. Calcule la densidad en el valor de prueba $\tilde{\varepsilon}^1$ y compárelo con la densidad en el valor original ε^0. Es decir, compare $f(\tilde{\varepsilon}^1)$ con $f(\varepsilon^0)$. Si $f(\tilde{\varepsilon}^1) > f(\varepsilon^0)$, entonces acepte $\tilde{\varepsilon}^1$, etiquételo como ε^1 y vaya al paso 4. Si $f(\tilde{\varepsilon}^1) \leq f(\varepsilon^0)$, entonces acepte $\tilde{\varepsilon}^1$ con probabilidad $f(\tilde{\varepsilon}^1)/f(\varepsilon^0)$ y rechácelo con probabilidad $1 - f(\tilde{\varepsilon}^1)/f(\varepsilon^0)$. Para determinar si aceptamos o rechazamos $\tilde{\varepsilon}^1$, en este caso, extraiga un valor al azar de una distribución uniforme estándar μ. Si $\mu \leq f(\tilde{\varepsilon}^1)/f(\varepsilon^0)$, entonces mantenga $\tilde{\varepsilon}^1$. De lo contrario, rechace $\tilde{\varepsilon}^1$. Si se acepta $\tilde{\varepsilon}^1$, etiquételo como ε^1. Si $\tilde{\varepsilon}^1$ es rechazado, entonces utilice ε^0 como ε^1.
4. Elija un valor de prueba de ε^2 como $\tilde{\varepsilon}^2 = \varepsilon^1 + \eta$, donde η es un nuevo valor extraído al azar de $g(\eta)$.
5. Aplique la regla descrita en el paso 3 para aceptar $\tilde{\varepsilon}^2$ como ε^2 o rechace $\tilde{\varepsilon}^2$ y utilice ε^1 como valor de ε^2.
6. Continúe este proceso durante múltiples iteraciones. La secuencia ε^t acaba siendo equivalente a extraer valores al azar de $f(\varepsilon)$ para una t suficientemente grande.

Los valores generados se correlacionan en serie, ya que cada valor depende del valor anterior. De hecho, cuando un valor de prueba es rechazado, el valor empleado en la serie es igual al valor previo. Esta correlación en serie debe tenerse en cuenta cuando se utilizan los valores generados.

El algoritmo MH se puede aplicar a cualquier densidad que pueda ser calculada. Este algoritmo es particularmente útil cuando la constante de normalización de una densidad no se conoce o no se puede calcular fácilmente. Supongamos que sabemos que ε se distribuye proporcional a $f^*(\varepsilon)$. Esto significa

que la densidad de ε es $f(\varepsilon) = \frac{1}{k} f^*(\varepsilon)$, donde la constante de normalización $k = \int f^*(\varepsilon) d\varepsilon$ asegura que la integral de f es 1. Por lo general, k no se puede calcular analíticamente, por la misma razón por la que tenemos que simular integrales en otras circunstancias. Por suerte, el algoritmo MH no utiliza k. Un valor de prueba de ε^t se prueba en primer lugar determinando si $f(\tilde{\varepsilon}^t) > f(\varepsilon^{t-1})$. Esta comparación no se ve afectada por la constante de normalización, ya que la constante entra en el denominador en ambos lados. Por lo tanto, si $f(\tilde{\varepsilon}^t) \leq f(\varepsilon^{t-1})$, aceptamos el valor de prueba con probabilidad $f(\tilde{\varepsilon}^t)/f(\varepsilon^{t-1})$. La constante de normalización desaparece de este ratio.

El algoritmo MH es en realidad más general de lo que describo en esta explicación, aunque en la práctica suele aplicarse como lo describo. Chib y Greenberg (1995) proporcionan una excelente descripción del algoritmo más general, así como una explicación de por qué funciona. Bajo la definición más general, el muestreo de Gibbs es un caso particular del algoritmo MH, tal y como Gelman (1992) señaló. El algoritmo MH y el muestreo de Gibbs a menudo se llaman métodos de Monte Carlo – Cadena de Markov (*Markov chain Monte Carlo* ,MCMC o MC al cuadrado); Chib y Greenberg (1996) proporcionan una descripción de su uso en econometría. Los valores extraídos son cadenas de Markov porque cada valor depende sólo del valor extraído inmediatamente antes, y los métodos son Monte Carlo porque los valores se extraen de forma aleatoria. Exploraremos más cuestiones sobre el algoritmo MH, tales como la forma de elegir $g(\varepsilon)$, en el contexto de su uso con los procedimientos bayesianos jerárquicos (en el capítulo 12).

9.3 Reducción de la varianza

El uso de valores independientes extraídos al azar para la simulación es atractivo por ser conceptualmente simple y porque las propiedades estadísticas del simulador resultante son fáciles de obtener. Sin embargo, hay otras formas de extraer valores de una densidad que pueden proporcionar mayor precisión para un número dado de extracciones. Examinaremos estos métodos alternativos en las siguientes secciones.

Recordemos que el objetivo es aproximar una integral de la forma $\int t(\varepsilon) f(\varepsilon) d\varepsilon$. Al extraer una secuencia de valores de la densidad $f(\varepsilon)$, dos cuestiones están en juego: la cobertura y la covarianza. Considere en primer lugar la cobertura. La integral es sobre toda la densidad f. Parece razonable pensar que obtendríamos una aproximación más precisa si evaluásemos $t(\varepsilon)$ en valores de ε que se extendiesen en todo el dominio de f. Con valores independientes extraídos al azar, es posible que los valores se agrupen, sin que obtengamos valores en grandes áreas del dominio. Es de esperar que procedimientos que garanticen una mejor cobertura del dominio proporcionen una mejor aproximación.

La covarianza es otro tema relevante. Con valores independientes, la covarianza entre valores es cero.

Por tanto, la varianza de un simulador basado en R valores independientes es la varianza basada en un valor dividida por R. Si los valores extraídos se correlacionan negativamente en lugar de ser independientes, entonces la varianza del simulador es menor. Considere $R = 2$. La varianza de $\check{t} = [t(\varepsilon_1) + t(\varepsilon_2)]/2$ es $\left[V\left(t(\varepsilon_1)\right) + V\left(t(\varepsilon_2)\right) + 2Cov(t(\varepsilon_1), t(\varepsilon_2))\right]/4$. Si los valores son independientes, la varianza es $V\left(t(\varepsilon_r)\right)/2$. Si los dos valores se correlacionan negativamente entre sí, el término de covarianza es negativo y la varianza es inferior a $V\left(t(\varepsilon_r)\right)/2$. Básicamente, cuando los valores están correlacionados negativamente dentro de un simulador no sesgado, un valor extraído que esté por encima de $\bar{t} = E_r(t(\varepsilon))$ tiende a estar asociado a un valor para la siguiente extracción que está por debajo de $\bar{t} = E_r(t(\varepsilon))$, de tal manera que su media está más cerca del verdadero valor $\bar{t}$.

El mismo concepto aplica cuando los simuladores se suman entre observaciones. Por ejemplo, la función log-verosimilitud simulada es una suma sobre observaciones del logaritmo de probabilidades simuladas. Si los valores extraídos para la simulación de cada observación son independientes de los valores

extraídos para otras observaciones, la varianza de la suma es simplemente la suma de las varianzas. Si los valores extraídos se toman de manera que se cree una correlación negativa entre observaciones, la varianza de la suma es inferior.

Para una observación dada, la cuestión de la covarianza está relacionada con la de la cobertura. Al inducir una correlación negativa entre valores, usualmente estamos asegurando una mejor cobertura. Con $R = 2$, si los dos valores se extraen de forma independiente, entonces ambos pueden estar en la parte baja de la distribución. Por el contrario, si se induce correlación negativa, el segundo valor tenderá a ser alto si el primer valor fue bajo, lo que proporciona una mejor cobertura.

A continuación se describen métodos para lograr una mejor cobertura para la integral de cada observación y para inducir correlación negativa entre valores extraídos para cada observación, así como entre observaciones. Para facilitar la explicación asumimos que la integral es una probabilidad de elección y que la suma sobre las observaciones es la función log-verosimilitud simulada. Sin embargo, los conceptos aplican a otras integrales, tales como puntuaciones (*scores*), así como para otras sumas, como las condiciones de momentos (*moment conditions*) y cuotas de mercado (*market shares*). Además, a menos que se indique lo contrario, ilustraremos los métodos con sólo dos términos aleatorios de manera que los valores extraídos puedan ser representados gráficamente. Los términos aleatorios se etiquetan como ε^a y ε^b, y colectivamente como $\varepsilon = \langle \varepsilon^a, \varepsilon^b \rangle'$. Un valor de ε extraído de su densidad $f(\varepsilon)$ se denota como $\varepsilon_r = \langle \varepsilon_r^a, \varepsilon_r^b \rangle'$ para $r = 1, \ldots, R$. Por lo tanto, ε_3^a, por ejemplo, es el tercer valor extraído del primer término aleatorio.

*9.3.1 Antitéticos (*antithetics*)*

La extracción de valores antitéticos, sugerida por Hammersley y Morton (1956), se obtiene mediante la creación de diferentes tipos de imágenes espejo de un mismo valor. Para una densidad simétrica centrada en cero, la variable aleatoria antitética más simple se crea invirtiendo el signo de todos los elementos de un valor extraído. La figura 9.4 ilustra este método. Supongamos que extraemos de $f(\varepsilon)$ un valor al azar $\varepsilon_1 = \langle \varepsilon_1^a, \varepsilon_1^b \rangle'$. El segundo "valor extraído", al que llamamos antítesis del primer valor, se crea como $\varepsilon_2 = \langle -\varepsilon_1^a, -\varepsilon_1^b \rangle'$. Cada valor extraído de f crea un par de "extracciones", el valor original y su imagen espejo (reflejado en relación al origen). Para obtener un total de R valores, extraemos de f un total de $R/2$ valores independientemente y los otros $R/2$ restantes se crean como los negativos de los valores originales.

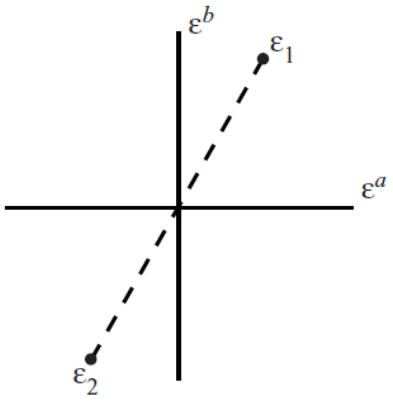


Figura 9.4 Signo invertido de los dos elementos.

Cuando la densidad no está centrada en cero, es posible aplicar el mismo concepto pero a través de un proceso diferente. Por ejemplo, la densidad uniforme estándar se extiende entre 0 y 1, con 0.5 como valor central. Si extraemos de esta densidad un valor μ_1, podemos crear su valor antitético como $\mu_2 = 1 - \mu_1$.

Este valor dista igual de 0.5 que la primera extracción realizada, pero en el otro lado de 0.5. En general, para cualquier densidad univariada con función acumulativa $F(\varepsilon)$, la antítesis de un valor se crea como $F^{-1}(1-F(\varepsilon))$. En el caso de una densidad simétrica centrada en cero, esta fórmula general es equivalente a invertir directamente el signo. En lo que queda de exposición supondremos que la densidad es simétrica y está centrada en cero, lo que hace los conceptos más fáciles de expresar y visualizar.

La correlación entre un valor extraído y su antítesis es exactamente - 1, de manera que la varianza de su suma es igual a cero: $V(\varepsilon_1+\varepsilon_2)=V(\varepsilon_1)+V(\varepsilon_2)+2Cov(\varepsilon_1,\varepsilon_2)=0$. Este hecho no significa que no haya varianza en la probabilidad simulada que se basa en estos valores. La probabilidad simulada es una función no lineal de los términos aleatorios y, por lo tanto, la correlación entre $P(\varepsilon_1)$ y $P(\varepsilon_2)$ es menor que uno. La varianza de la probabilidad simulada $\check{P}=\frac{1}{2}[P(\varepsilon_1)+P(\varepsilon_2)\,]$ es mayor que cero. Sin embargo, la varianza de las probabilidades simuladas es menor que $\frac{1}{2}V_r(P(\varepsilon_r))$, que es la varianza que tendríamos con dos valores extraídos de forma independiente.

Como se muestra en la figura 9.4, invertir el signo de un valor extraído da puntos de evaluación en cuadrantes opuestos. El concepto se puede ampliar para obtener valores en cada cuadrante. Para ello, se extrae un valor y posteriormente se crean valores antitéticos invirtiendo cada elemento por separado (dejando el signo de los otros elementos sin alterar), invirtiendo el signo de cada pareja de elementos, de cada triplete de elementos, y así sucesivamente. Para un ε con dos elementos, este proceso crea tres valores antitéticos por cada valor extraído independiente. Para $\varepsilon_1=\langle\varepsilon_1^a,\varepsilon_1^b\rangle'$, los valores antitéticos son

$$\varepsilon_2=\langle-\varepsilon_1^a,\varepsilon_1^b\rangle',$$

$$\varepsilon_3=\langle\varepsilon_1^a,-\varepsilon_1^b\rangle',$$

$$\varepsilon_4=\langle-\varepsilon_1^a,-\varepsilon_1^b\rangle'.$$

Estos valores se muestran en la figura 9.5. Cada cuadrante contiene un valor.

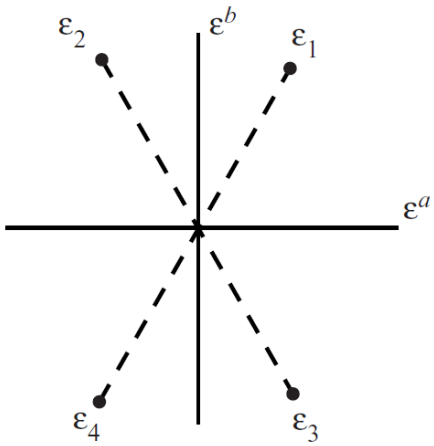


Figura 9.5. Invierta el signo de cada elemento, luego el de ambos.

Moviendo la posición de cada elemento e invirtiendo sus signos, se puede obtener mejor cobertura y mayor correlación negativa. En la figura 9.5, ε_1 y ε_2 están bastante próximos entre sí, al igual que ε_3y ε_4. Esta disposición deja grandes áreas descubiertas entre ε_1 y ε_3 y entre ε_2 y ε_4. Es posible obtener valores ortogonales con disposición equilibrada intercambiando el elemento ε_1^a con el elemento ε_1^b, e invirtiendo los signos al mismo tiempo. Los valores antitéticos resultantes son

$$\varepsilon_2 = \langle -\varepsilon_1^b, \varepsilon_1^a \rangle',$$

$$\varepsilon_3 = \langle \varepsilon_1^b, -\varepsilon_1^a \rangle',$$

los cuales se ilustran en la figura 9.6. Por supuesto, estos conceptos pueden extenderse a cualquier número de dimensiones. Para un ε M-dimensional, cada valor extraído crea 2^M valores antitéticos (incluyendo el original), con un valor en cada cuadrante.

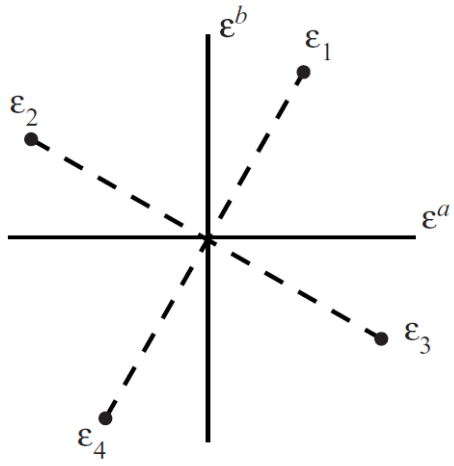


Figura 9.6. Intercambie posiciones e invierta signos.

Las comparaciones realizadas por Vijverberg (1997) y Sándor y András (2001) muestran que los antitéticos mejoran sustancialmente la estimación de modelos probit. Del mismo modo, Geweke (1988) ha demostrado el valor de su uso al calcular estadísticos basados en distribuciones posteriores bayesianas.

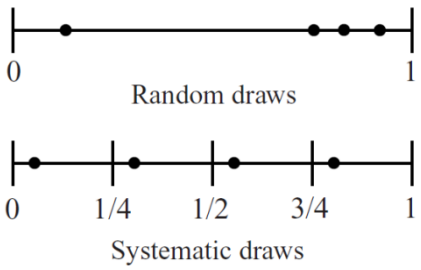


Figura 9.7. Extracción de valores de una uniforme estándar.

9.3.2 Muestreo sistemático

La cobertura también puede mejorarse mediante muestreo sistemático (McGrath, 1970), un método que crea una malla de puntos sobre el dominio de la densidad y que aleatoriamente desplaza la malla de puntos al completo. Considere la extracción de valores de una distribución uniforme entre 0 y 1. Si se extraen cuatro valores de forma independiente, los puntos podrían distribuirse como se muestra en la parte superior de la figura 9.7, obteniendo una cobertura bastante pobre. En lugar de extraerlos de forma totalmente independiente, podemos dividir el intervalo unitario en cuatro segmentos y extraer valores al azar de forma que aseguremos que hay un valor en cada segmento, con idéntica distancia entre valores. Para ello, extraiga un valor de una uniforme entre 0 y 0.25 (mediante una extracción de

un valor de una densidad uniforme estándar, dividiendo el resultado por 4). Etiquete el valor como ε_1. Los otros tres valores se crean como

$$\varepsilon_2 = 0.25 + \varepsilon_1,$$
$$\varepsilon_3 = 0.50 + \varepsilon_1,$$
$$\varepsilon_4 = 0.75 + \varepsilon_1.$$

Estos valores se distribuyen tal y como se muestra en la parte inferior de la figura 9.7, de forma que proporcionan una mejora cobertura que la extracción de valores independientes.

La cuestión es, ¿qué nivel de segmentación debemos aplicar sobre el intervalo? Por ejemplo, para extraer un total de 100 valores, el intervalo unitario puede dividirse en 100 segmentos. Se extrae un valor entre 0 y 0.01, y posteriormente definimos los restantes 99 valores a partir de este primer valor. En lugar de hacer esto, el intervalo unitario puede dividirse en menos de 100 valores y extraer más valores independientes. Si el intervalo se divide en cuatro segmentos, podemos extraer 25 valores independientes entre 0 y 0.25, y posteriormente calcular tres valores en los otros segmentos para cada uno de los valores extraídos de forma independiente. El investigador debe buscar un término medio al decidir el nivel de segmentación de la malla empleada en el muestro sistemático. Un número mayor de segmentos proporcionan una cobertura más uniforme para un número total dado de valores. Sin embargo, un número menor de segmentos proporcionan más aleatoriedad al proceso. En nuestro ejemplo con R = 100, sólo hay un valor aleatorio si utilizamos 100 segmentos, mientras que hay 25 valores aleatorios si usamos cuatro segmentos.

La aleatoriedad de los valores extraídos para una simulación es una propiedad necesaria en la obtención de las propiedades asintóticas de los estimadores basados en simulación, como se describe en el capítulo 10. Muchas de las propiedades asintóticas se basan en el concepto de que el número de valores extraídos al azar aumenta sin límites con el tamaño de la muestra. Las distribuciones asintóticas se vuelven relativamente exactas sólo cuando se han tomado suficientes valores al azar. Por lo tanto, para un número total dado de valores, el objetivo de lograr una mejor cobertura, que se logra con una segmentación definida con más precisión, debe ser equilibrado con el objetivo de tener suficiente aleatoriedad en las fórmulas asintóticas como para poder aplicarlas, lo que se logra con un menor nivel de segmentación (segmentos mayores). El mismo problema aplica a los antitéticos vistos anteriormente.

El muestreo sistemático puede realizarse en múltiples dimensiones. Considere una distribución uniforme bidimensional en una región cuadrada unitaria. Podemos crear una malla dividiendo cada dimensión en segmentos. Tal y como se muestra en la figura 9.8, cuando cada dimensión se divide en cuatro segmentos, el cuadrado unitario queda dividido en 16 zonas. Se extrae al azar un valor entre 0 y 0.25 para cada elemento de ε, resultando $\varepsilon_1 = \langle \varepsilon_1^a, \varepsilon_1^b \rangle'$, donde $0 < \varepsilon_1^a < 0.25$ y $0 < \varepsilon_1^b < 0.25$. Este valor extraído cae en algún lugar de la zona inferior izquierda en la figura 9.8. Para obtener otros quince valores, sumamos al "origen" de cada zona el valor $\varepsilon_1 = \langle \varepsilon_1^a, \varepsilon_1^b \rangle'$. Por ejemplo, el punto que se crea para la zona inferior derecha es $\varepsilon_4 = \langle (0.75 + \varepsilon_1^a), (0 + \varepsilon_1^b) \rangle'$.

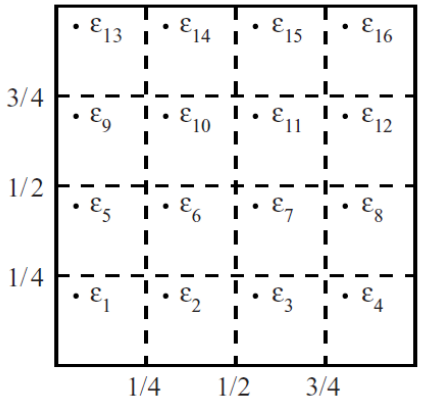


Figura 9.8. Valores extraídos sistemáticamente en dos dimensiones.

Estos valores se definen para una distribución uniforme. Cuando f representa otra densidad, los puntos se transforman utilizando el método descrito en la sección 9.2.3. En particular, sea F la distribución acumulativa asociada a la densidad f univariada. Es posible hacer un muestreo sistemático de f transformando cada extracción sistemática de una uniforme mediante F^{-1}. Por ejemplo, para una normal estándar, se divide la densidad en cuatro segmentos de igual tamaño con puntos de corte: $\Phi^{-1}(0.25) = -0.67$, $\Phi^{-1}(0.5) = 0$ y $\Phi^{-1}(0.75) = 0.67$. Tal y como se muestra en la figura 9.9, estos segmentos son de igual tamaño en el sentido de que cada uno contiene la misma masa de densidad. Las extracciones de la normal estándar se obtienen mediante la extracción de un valor de una uniforme entre 0 y 0.25, etiquetado μ_1. El punto correspondiente en la normal es $\varepsilon_1 = \Phi^{-1}(\mu_1)$, que se ubica en el primer segmento. Los puntos para los otros tres segmentos restantes se crean como $\varepsilon_2 = \Phi^{-1}(0.25 + \mu_1)$, $\varepsilon_3 = \Phi^{-1}(0.5 + \mu_1)$ y $\varepsilon_4 = \Phi^{-1}(0.75 + \mu_1)$.

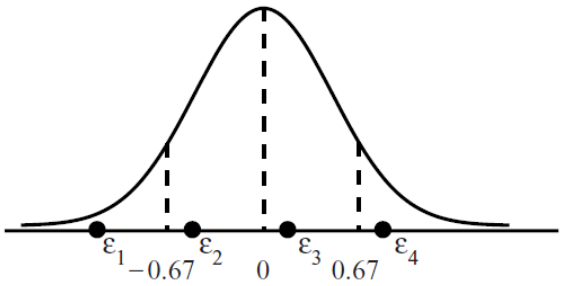


Figura 9.9 Valores sistemáticos para una normal univariada.

Se pueden extraer valores de términos aleatorios multidimensionales de forma similar, con la condición de que los elementos sean independientes. Por ejemplo, si ε se compone de dos elementos cada uno de los cuales se distribuye de forma normal estándar, entonces se pueden extraer valores análogos a los de la figura 9.8 de la siguiente manera: Extraiga los valores μ_1^a y μ_1^b de una uniforme entre 0 y 0.25. Calcule ε_1 como $\langle \Phi^{-1}(\mu_1^a), \Phi^{-1}(\mu_1^b) \rangle'$. Calcule los otros 15 puntos restantes ε_r como $\langle \Phi^{-1}(x_r + \mu_1^a), \Phi^{-1}(y_r + \mu_1^b) \rangle'$, donde $\langle x_r, y_r \rangle'$ es el origen del área r en el cuadrado unitario.

El requisito de que los elementos de ε sean independientes no es restrictivo. Se pueden crear elementos aleatorios correlacionados a través de transformaciones de elementos independientes, como la transformación de Choleski. Los elementos independientes se extraen de su densidad y posteriormente se crea la correlación dentro del modelo.

Obviamente, se pueden obtener varios conjuntos de valores muestreados sistemáticamente para lograr más aleatorización. En dos dimensiones con cuatro segmentos en cada dimensión, se pueden obtener

64 valores extrayendo 4 valores independientes en el cuadrado que va de 0 a 1/4 y creando 15 valores adicionales para cada valor independiente. Este procedimiento proporciona mayor aleatorización pero una cobertura menos uniforme que la obtenida al generar los valores usando 8 segmentos en cada dimensión, de forma que cada extracción aleatoria en el cuadrante que va de 0 a 1/8 se traduce en 64 valores sistemáticos.

Para una distribución normal, los valores que se extraen como se acaba de describir no son simétricos alrededor de cero. Se puede utilizar una aproximación alternativa a este problema para asegurar dicha simetría. Para una normal unidimensional, se pueden obtener 4 valores que son simétricos alrededor de cero como sigue. Extraiga un valor al azar de una uniforme entre 0 y 0.25, y etiquételo como μ_1. Cree el valor correspondiente para la normal como $\varepsilon_1 = \Phi^{-1}(\mu_1)$. Calcule el valor para el segundo segmento como $\varepsilon_2 = \Phi^{-1}(0.25 + \mu_1)$. A continuación, cree los valores para el tercer y cuarto segmento como los negativos de estos valores: $\varepsilon_3 = -\varepsilon_2$ y $\varepsilon_4 = -\varepsilon_1$. La figura 9.10 ilustra los valores generados usando el mismo μ_1 que en la figura 9.9. Este procedimiento combina el muestreo sistemático con los antitéticos. Se puede extender a múltiples dimensiones mediante la creación de valores extraídos sistemáticamente para el cuadrante positivo y la posterior creación de valores antitéticos para los otros cuadrantes.

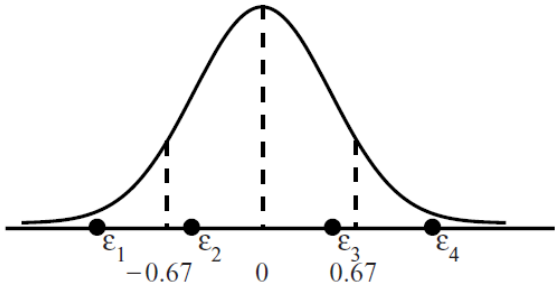


Figura 9.10 Valores simétricos extraídos sistemáticamente.

9.3.3 Secuencias de Halton

Las secuencias de Halton (Halton, 1960) proporcionan cobertura y, a diferencia de los otros métodos que hemos expuesto aquí, inducen correlación negativa entre observaciones. Una secuencia de Halton se define en función de un número dado, por lo general un número primo. El concepto detrás de la secuencia se entiende más fácilmente a través de un ejemplo. Considere el número primo 3. La secuencia de Halton para el 3 se crea dividiendo el intervalo unitario en tres partes con puntos de corte en $\frac{1}{3}$ y $\frac{2}{3}$, como se muestra en la parte superior de la figura 9.11. Los primeros términos de la secuencia son estos puntos de corte: $\frac{1}{3}, \frac{2}{3}$. A continuación, cada uno de los tres segmentos se divide en tres partes, y los puntos de corte de estos nuevos segmentos se añaden a la secuencia pero de una forma particular. La secuencia se convierte en $\frac{1}{3}, \frac{2}{3}, \frac{1}{9}, \frac{4}{9}, \frac{7}{9}, \frac{2}{9}, \frac{5}{9}, \frac{8}{9}$. Observe que los puntos de corte más bajos de los tres segmentos $(\frac{1}{9}, \frac{4}{9}, \frac{7}{9})$ se introducen en la secuencia antes que los puntos de corte más altos $(\frac{2}{9}, \frac{5}{9}, \frac{8}{9})$. Posteriormente, cada uno de los nueve segmentos se divide en tres partes y se añaden nuevamente los puntos de corte a la secuencias. La secuencia se convierte así en $\frac{1}{3}, \frac{2}{3}, \frac{1}{9}, \frac{4}{9}, \frac{7}{9}, \frac{2}{9}, \frac{5}{9}, \frac{8}{9}, \frac{1}{27}, \frac{10}{27}, \frac{19}{27}, \frac{4}{27}, \frac{13}{27}$, y así sucesivamente. Este proceso continúa hasta obtener tantos puntos como el investigador desee.

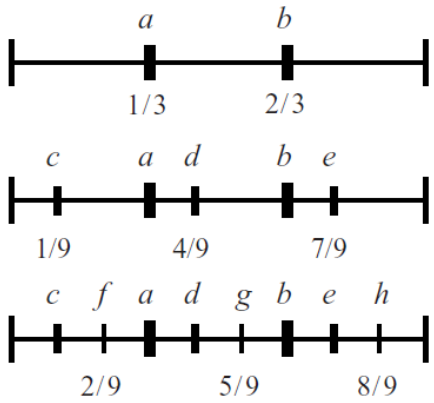


Figura 9.11. Secuencia de Halton para el número primo 3.

Desde el punto de vista de programación, es fácil crear una secuencia de Halton. La secuencia se crea de forma iterativa. En cada iteración t, la secuencia se denota s_t, que es una serie de números. La secuencia se amplía en cada iteración, siendo la nueva secuencia $s_{t+1} = \{s_t, s_t + 1/3^t, s_t + 2/3^t\}$. El proceso sería el siguiente: comience con 0 como secuencia inicial: $s_0 = \{0\}$. El número cero en realidad no es parte de una secuencia de Halton, pero considerarlo como primer elemento facilita la creación de la secuencia, como veremos más adelante. Se puede suprimir una vez se ha creado la secuencia completa. En la primera iteración, agregue $1/3^1 = \frac{1}{3}$ y luego $2/3^1 = \frac{2}{3}$ al elemento inicial, obteniendo $\left\{0, \frac{1}{3}, \frac{2}{3}\right\}$. La secuencia tiene tres elementos. En la segunda iteración, añada $1/3^2 = \frac{1}{9}$ y luego $2/3^2 = \frac{2}{9}$ a cada elemento de la secuencia, y añada los resultados:

$0 = 0,$

$1/3 = 1/3,$

$2/3 = 2/3,$

$0 + 1/9 = 1/9,$

$1/3 + 1/9 = 4/9,$

$2/3 + 1/9 = 7/9,$

$0 + 2/9 = 2/9,$

$1/3 + 2/9 = 5/9,$

$2/3 + 2/9 = 8/9.$

La nueva secuencia consta de nueve elementos. En la tercera iteración, añada $1/3^3 = \frac{1}{27}$ y luego $2/3^3 = \frac{2}{27}$ a cada elemento de esta secuencia y añada los resultados:

$0 = 0,$

$1/3 = 1/3,$

$2/3 = 2/3,$

$1/9 = 1/9,$

$4/9 = 4/9,$

$7/9 = 7/9,$

$2/9 = 2/9,$

$5/9 = 5/9,$

$8/9 = 8/9,$

$0 + 1/27 = 1/27,$

$1/3 + 1/27 = 10/27,$

$2/3 + 1/27 = 19/27,$

$1/9 + 1/27 = 4\ /\ 27,$

$4/9 + 1/\ 27 = 13/27,$

$7/9 + 1/\ 27 = 22/27,$

$2/9 + 1/\ 27 = 7/27,$

$5/9 + 1/\ 27 = 16/27,$

$8/9 + 1/27 = 25/27,$

$0 + 2/27 = 2/27,$

$1/3 + 2/27 = 11/27,$

$2/3 + 2/27 = 20/27,$

$1/9 + 2/27 = 5\ /27,$

$4/9 + 2/27 = 14/27,$

$7/9 + 2/27 = 23\ /27,$

$2/9 + 2/27 = 8\ /27,$

$5/9 + 2\ /\ 27 = 17/27,$

$8/9 + 2/27 = 26\ /27.$

La secuencia se compone ahora de 27 elementos. En la cuarta iteración, añada $1/3^4 = \frac{1}{81}$ y luego $2/3^4 = \frac{2}{81}$ a cada elemento de la secuencia y añada los resultados, y así sucesivamente.

Observe que la secuencia se compone así de ciclos sobre el intervalo unitario cada tres números:

0	$1/3$	$2/3$
$1/9$	$4/9$	$7/9$
$2/9$	$5/9$	$8/9$
$1/27$	$10/27$	$19/27$
$4/27$	$13/27$	$22/27$
$7/27$	$16/27$	$25/27$
$2/27$	$11/27$	$20/27$
$5/27$	$14/27$	$23/27$
$8/27$	$17/27$	$26/27$

Dentro de cada ciclo, los números son ascendentes.

Las secuencias de Halton para otros números primos se crean de manera similar. La secuencia para 2 es $\left\{\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{5}{8}, \frac{3}{8}, \frac{7}{8}, \frac{1}{16}, \frac{9}{16}, \ldots\right\}$. En general, la secuencia para el número primo k se crea iterativamente, siendo la secuencia en la iteración $t + 1$ $s_{t+1} = \left\{s_t, s_t + \frac{1}{k^t}, s_t + \frac{2}{k^t}, \ldots, s_t + \frac{k-1}{k^t}\right\}$. La secuencia contiene ciclos de longitud k, donde cada ciclo consiste en k puntos ascendentes, equidistantes entre sí, dentro del intervalo unitario.

Dado que una secuencia de Halton está definida en el intervalo unitario, sus elementos pueden ser considerados como una extracción de valores “bien colocados” de una densidad uniforme estándar. La extracción de valores de Halton proporciona, en promedio, una mejor cobertura que la extracción de

valores puramente al azar, debido a que se crean para llenar progresivamente el intervalo unitario de forma uniforme y cada vez más densa. Los elementos de cada ciclo son equidistantes entre sí, y cada ciclo abarca las zonas del intervalo unitario no cubiertas por los ciclos anteriores.

Cuando se utilizan secuencias de Halton para generar una muestra de observaciones, normalmente se crea una secuencia larga y posteriormente se utiliza una parte de la secuencia para cada observación. Los elementos iniciales de la secuencia se descartan por razones que detallaremos posteriormente. Los elementos restantes se utilizan en grupos, con cada grupo de elementos constituyendo los valores extraídos para una observación. Por ejemplo, supongamos que hay dos observaciones, y el investigador quiere $R = 5$ valores extraídos para cada una. Si se utiliza el número primo 3, y el investigador decide descartar los primeros 10 elementos, se debe crear una secuencia de longitud 20. Esta secuencia es

$$
\begin{array}{lll}
0 & 1/3 & 2/3 \\
1/9 & 4/9 & 7/9 \\
2/9 & 5/9 & 8/9 \\
1/27 & 10/27 & 19/27 \\
4/27 & 13/27 & 22/27 \\
7/27 & 16/27 & 25/27 \\
2/27 & 11/27. &
\end{array}
$$

Después de eliminar los 10 primeros elementos, los valores de Halton para la primera observación son $\left\{\frac{10}{27}, \frac{19}{27}, \frac{4}{27}, \frac{13}{27}, \frac{22}{27}\right\}$ y los valores de Halton para la segunda observación son $\left\{\frac{7}{27}, \frac{16}{27}, \frac{25}{27}, \frac{2}{27}, \frac{11}{27}\right\}$. Estos valores se ilustran en la figura 9.12. Observe que las deficiencias en la cobertura de la primera observación se compensan por los valores de la segunda observación. Por ejemplo, la gran brecha que existe entre $\frac{4}{27}$ y $\frac{10}{27}$ para la primera observación es rellenada por el punto medio de este vacío, $\frac{7}{27}$, en la segunda observación. La brecha entre $\frac{13}{27}$ y $\frac{19}{27}$ es rellenada por su punto medio, $\frac{16}{27}$, en la segunda observación, y así sucesivamente. El patrón por el cual se crean secuencias de Halton las hace de tal manera que cada sub-secuencia llena los vacíos de las sub-secuencias anteriores.

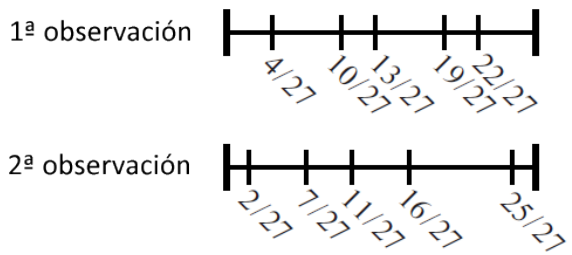


Figura 9.12. Valores de Halton para dos observaciones.

Debido a esta propiedad de llenado de espacios, las probabilidades simuladas basadas en la extracción de valores de Halton tienden a ser auto-correctoras entre observaciones. Los valores de una observación tienden a estar correlacionados negativamente con los de la observación anterior. En nuestro ejemplo, el promedio de los valores correspondiente a la primera observación está por encima de 0.5, mientras que el promedio de los valores para la segunda observación es inferior a 0.5. Esta correlación negativa reduce el error en la función log-verosimilitud simulada.

Cuando el número de valores utilizados para cada observación se eleva, la cobertura para cada observación mejora. La covarianza negativa entre observaciones disminuye, ya que hay menos espacios en la cobertura de cada observación para ser rellenados por la siguiente observación. Esta característica

de auto-corrección entre observaciones de las secuencias de Halton es mayor cuando se extraen pocos valores para cada observación, de manera que la corrección es más necesaria. Sin embargo, la precisión mejora cuantos más valores de Halton se utilizan, ya que la cobertura es mejor para cada observación.

Como se ha descrito hasta ahora, la extracción de valores de Halton se define para una densidad uniforme. Para obtener una secuencia de puntos para otras densidades univariadas, la distribución acumulativa inversa se evalúa en cada elemento de la secuencia de Halton. Por ejemplo, supongamos que el investigador quiere extraer valores de una densidad normal estándar. Se crea una secuencia de Halton para, por ejemplo, el número primo 3, y se evalúa la inversa de la normal acumulativa para cada elemento. La secuencia resultante es

$$\Phi^{-1}\left(\frac{1}{3}\right) = -0.43,$$

$$\Phi^{-1}\left(\frac{2}{3}\right) = 0.43,$$

$$\Phi^{-1}\left(\frac{1}{9}\right) = -1.2,$$

$$\Phi^{-1}\left(\frac{4}{9}\right) = -0.14,$$

$$\Phi^{-1}\left(\frac{7}{9}\right) = 0.76,$$

$$\vdots$$

Esta secuencia se representa en la figura 9.13. Se puede considerar igual a la del intervalo unitario, en el sentido de que divide la densidad en tres segmentos de igual masa, con puntos de corte en -0.43 y 0.43, y luego divide cada segmento en tres sub-segmentos de igual masa, y así sucesivamente.

Figura 9.13. Valores de Halton para una normal estándar.

Para obtener secuencias de Halton en múltiples dimensiones se crea una secuencia de Halton para cada dimensión, usando un número primo diferente para cada una. Por ejemplo, una secuencia en dos dimensiones se obtiene mediante la creación de pares de valores de las secuencias de Halton para los números primos 2 y 3. Los puntos son

$$\varepsilon_1 = \langle\frac{1}{2}, \frac{1}{3}\rangle,$$

$$\varepsilon_2 = \langle \frac{1}{4}, \frac{2}{3} \rangle,$$

$$\varepsilon_3 = \langle \frac{3}{4}, \frac{1}{9} \rangle,$$

$$\varepsilon_4 = \langle \frac{1}{8}, \frac{4}{9} \rangle,$$

$$\varepsilon_5 = \langle \frac{5}{8}, \frac{7}{9} \rangle,$$

$$\varepsilon_6 = \langle \frac{3}{8}, \frac{2}{9} \rangle,$$

$$\vdots$$

Esta secuencia se representa en la figura 9.14. Para extraer valores de una normal estándar de dos dimensiones independientes, se calcula la normal acumulativa inversa de cada elemento de estos pares. Los valores son

$$\varepsilon_1 = \langle 0, -0.43 \rangle,$$

$$\varepsilon_2 = \langle -0.67, 0.43 \rangle,$$

$$\varepsilon_3 = \langle 0.67, -1.2 \rangle,$$

$$\varepsilon_4 = \langle -1.15, -1.2 \rangle,$$

$$\varepsilon_5 = \langle 0.32, 0.76 \rangle,$$

$$\varepsilon_6 = \langle -0.32, -0.76 \rangle,$$

$$\vdots$$

que se muestran en la figura 9.15.

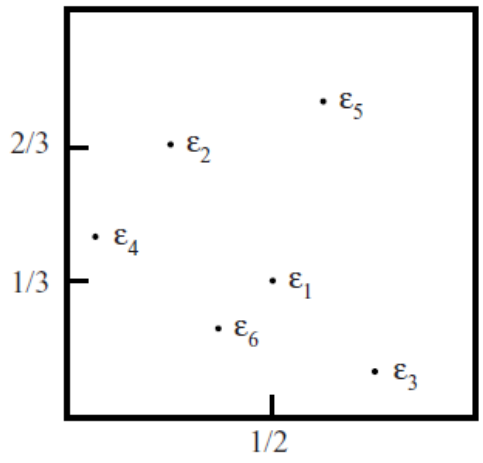


Figura 9.14. Secuencia de Halton en dos dimensiones para los números primeros 2 y 3.

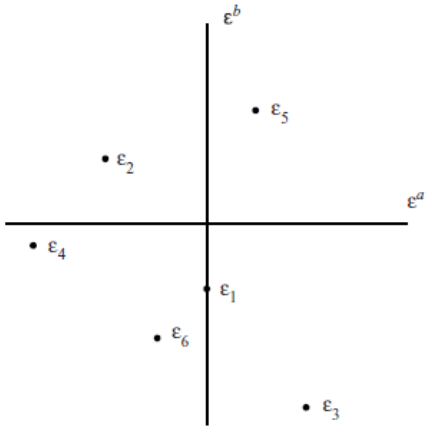


Figura 9.15. Secuencia de Halton para una normal estándar de dos dimensiones.

Al crear secuencias en varias dimensiones, es costumbre eliminar la parte inicial de las series. Los términos iniciales de dos secuencias de Halton están altamente correlacionados, por lo menos durante el primer ciclo de cada secuencia. Por ejemplo, las secuencias para 7 y 11 comienzan con $\left\{\frac{1}{7},\frac{2}{7},\frac{3}{7},\frac{4}{7},\frac{5}{7},\frac{6}{7}\right\}$ y $\left\{\frac{1}{11},\frac{2}{11},\frac{3}{11},\frac{4}{11},\frac{5}{11},\frac{6}{11}\right\}$. Estos primeros elementos caen en una línea en dos dimensiones, como se muestra en la figura 9.16. La correlación se disipa después de que cada secuencia haya completado un ciclo a través del intervalo unitario, puesto que secuencias con diferentes números primos completan un ciclo a diferentes velocidades. El descarte de la parte inicial de la secuencia elimina la correlación. El número de elementos iniciales a descartar necesita ser al menos tan grande como el mayor primo empleado en la creación de las secuencias.

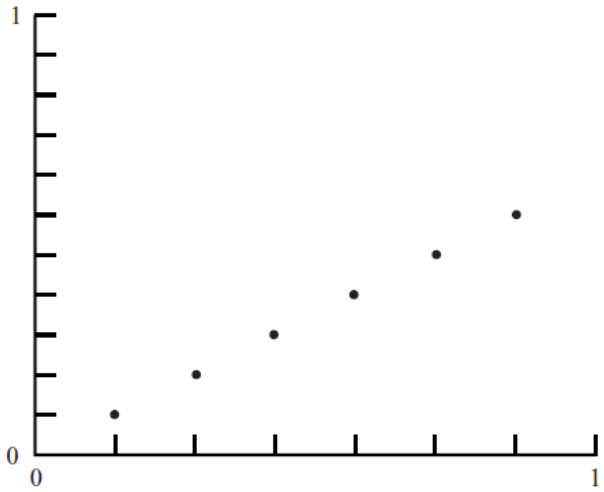


Figura 9.16. Primeros 6 elementos de una secuencia de Halton para los primos 7 y 11.

La potencial correlación es la razón por la que se utilizan números primos para crear las secuencias de Halton en lugar de no primos. Si se utiliza un número compuesto (no primo), entonces hay una posibilidad de que los ciclos coincidan a lo largo de toda la secuencia, en lugar de sólo para los elementos iniciales. Por ejemplo, si se crean secuencias de Halton con los números 3 y 6, la secuencia de 3 completa un ciclo dos veces por cada ciclo completado de la secuencia de 6. Puesto que los elementos dentro de un ciclo son ascendentes, los elementos en cada ciclo de la secuencia de 3

estarán correlacionados con los elementos en el ciclo de la secuencia de 6. Usando sólo números primos evitamos esta superposición de ciclos.

La mayor cobertura y la correlación negativa entre observaciones que se obtiene al usar secuencias de Halton hacen que este método sea mucho más eficaz que la extracción de valores al azar a efectos de simulación. Spanier y Maize (1991) han demostrado que un pequeño número de valores de Halton proporciona una integración relativamente buena. En el contexto de los modelos de elección discreta, Bhat (2001) encontró que 100 valores de Halton proporcionaban resultados más precisos para su logit mixto que 1.000 valores extraídos al azar. De hecho, el error de simulación con 125 valores de Halton era la mitad del error obtenido con 1.000 valores extraídos de forma aleatoria y algo inferior al de 2.000 valores al azar. Train (2000), Munizaga y Alvarez-Daziano (2001), y Hensher (2001) confirman estos resultados en otros conjuntos de datos.

Como ilustración, considere el modelo logit mixto que se describe ampliamente en el capítulo 11. De forma resumida, el modelo describe la elección de proveedor de electricidad de los hogares. En una encuesta de preferencias declaradas, a los encuestados se les presentó una serie de situaciones de elección hipotéticas. En cada situación, se describieron cuatro proveedores de energía y se preguntó al entrevistado qué empresa elegiría. Los proveedores se diferenciaban en función de su precio, de si la empresa requería que el cliente firmase un contrato a largo plazo, de si el proveedor era el suministrador local de energía, de si el proveedor era una empresa bien conocida y de si el proveedor ofrecía tarifas diferenciadas por horario (*time-of-day, TOD*) o tarifas estacionales. Se estimó un modelo logit mixto con estas seis características como variables explicativas. Se supuso que el coeficiente de cada variable seguía una distribución normal, excepto para el coeficiente de precio, que se supuso fijo. Por tanto el modelo contenía cinco términos aleatorios para la simulación. Una descripción completa de los datos, del modelo estimado y de sus implicaciones, se facilitan en el capítulo 11, en el cual el contenido del modelo es relevante para el tema tratado en el capítulo. Por ahora, sólo nos interesa la comparación entre el uso de secuencias de Halton y el uso de valores al azar.

Para investigar esta cuestión, el modelo se estimó con 1.000 valores extraídos al azar y luego con 100 valores de Halton. Más concretamente, el modelo se estimó en cinco ocasiones utilizando cinco grupos diferentes de 1.000 valores al azar. Se calculó la media y la desviación estándar de los parámetros estimados en estas cinco simulaciones. Posteriormente, el modelo se estimó en cinco ocasiones usando secuencias de Halton. El primer modelo utilizó los números primos 2, 3, 5, 7 y 11 para las cinco dimensiones de la simulación. Para el resto de modelos se cambió el orden de los números primos, de modo que la dimensión para la que se utilizó cada número primo cambió en cada una de las cinco estimaciones. A continuación, se calculó el promedio y la desviación estándar de los cinco conjuntos de estimaciones.

Las medias de las estimaciones de los parámetros de las cinco simulaciones se pueden ver en la tabla 9.1. La media de las simulaciones basadas en valores extraídos al azar se dan en la primera columna, y las medios para las simulaciones basadas en secuencias de Halton se dan en la segunda columna. Los dos conjuntos de medias son muy similares. Este resultado indica que las secuencias de Halton proporcionan las mismas estimaciones, *en promedio*, que los valores extraídos al azar.

Tabla 9.1. Medias de las estimaciones de parámetros

	1.000 valores al azar	100 valores de Halton
Precio:	−0.8607	−0.8588
Duración del contrato:		
Media	−0.1955	−0.1965

Desv. Estándar	0.3092	0.3158
Proveedor local:		
Media	2.0967	2.1142
Desv. Estándar	1.0535	1.0236
Compañía conocida:		
Media	1.431	1.4419
Desv. Estándar	0.8208	0.6894
Tarifas por franjas horarias:		
Media	−8.3760	−8.4149
Desv. Estándar	2.4647	2.5466
Tarifas estacionales:		
Media	−8.6286	−8.6381
Desv. Estándar	1.8492	1.8977

Las desviaciones estándar de las estimaciones de los parámetros se pueden ver en la tabla 9.2. Para los 11 parámetros excepto uno, las desviaciones estándar son inferiores al emplear 100 valores de Halton respecto a 1.000 valores extraídos al azar. Para ocho de los parámetros, las desviaciones estándar son la mitad de grandes. Teniendo en cuenta que ambos conjuntos de valores dan esencialmente las mismos medias, las desviaciones estándar más bajas obtenidas con los valores de Halton indican que el investigador puede esperar estar más cerca de los valores esperados de las estimaciones si usa 100 valores de Halton en lugar de usar 1.000 valores extraídos al azar.

Tabla 9.2. Desviaciones estándar de las estimaciones de parámetros

	1.000 valores al azar	100 valores de Halton
Precio:	0.0310	0.0169
Duración del contrato:		
Media	0.0093	0.0045
Desv. Estándar	0.0222	0.0108
Proveedor local:		
Media	0.0844	0.0361
Desv. Estándar	0.1584	0.1180
Compañía conocida:		
Media	0.0580	0.0242
Desv. Estándar	0.0738	0.1753
Tarifas por franjas horarias:		
Media	0.3372	0.1650
Desv. Estándar	0.1578	0.0696
Tarifas estacionales:		
Media	0.4134	0.1789
Desv. Estándar	0.2418	0.0679

Estos resultados demuestran el beneficio proporcionado por las secuencias de Halton. El tiempo de computación requerido en simulación se puede reducir en un factor diez mediante el uso de secuencias de Halton en lugar de valores extraídos al azar, sin reducir, y de hecho incrementando, la precisión.

Sin embargo, estos resultados deben ser tomados con precaución. El uso de secuencias de Halton y otros números cuasi-aleatorios en la estimación basada en simulación es algo bastante nuevo y no completamente comprendido. Por ejemplo, durante el análisis surgió una anomalía que sirve como advertencia. El modelo se volvió a estimar con 125 valores de Halton en lugar de 100. Se estimó cinco veces con cada una de las cinco posibles ordenaciones de los números primos, como se ha descrito anteriormente. Cuatro de las cinco simulaciones proporcionaron estimaciones muy similares. Sin embargo, la quinta simulación dio estimaciones que eran notablemente diferentes de las demás. Por ejemplo, el coeficiente de precio estimado para las primeras cuatro simulaciones fue -0.862, -0.865, -0.863 y -0.864 respectivamente, mientras que la quinta resultó -0.911. Las desviaciones estándar entre los cinco conjuntos de estimaciones fueron menores a las obtenidas con 1.000 valores al azar, lo que confirma el valor de las secuencias de Halton. Sin embargo, las desviaciones estándar fueron mayores con 125 valores de Halton que con 100 valores de Halton, debido a que la última simulación con 125 valores proporcionaba resultados muy diferentes. La razón de esta anomalía no ha sido determinada aún. Su presencia indica la necesidad de una mayor investigación de las propiedades de las secuencias de Halton en la estimación basada en simulación.

9.3.4 Secuencias de Halton aleatorizadas

Las secuencias Halton son sistemáticas y no aleatorias. Sin embargo, las propiedades asintóticas de los estimadores basados en simulación se obtienen bajo el supuesto de que los valores empleados son aleatorios. Hay dos maneras de abordar esta cuestión. En primer lugar, uno puede darse cuenta de que cuando emplea un generador de números aleatorios, estos tampoco son realmente al azar. Son sistemáticos, como cualquier cosa hecha por una computadora. Un generador de números aleatorios crea valores que tienen muchas de las características de los valores realmente generados al azar, pero en realidad son sólo pseudo-aleatorios. En este sentido, por lo tanto, los valores de Halton pueden ser vistos como una forma sistemática de aproximar la integración que es más precisa que el uso de valores pseudo-aleatorios, que también son sistemáticos. Ninguno de los dos métodos coincide con el concepto teórico de aleatoriedad y, de hecho, no está claro que el concepto teórico en realidad tenga una representación en el mundo real. Ambos métodos cumplen con el objetivo básico subyacente de aproximar una integral sobre una densidad.

En segundo lugar, las secuencias de Halton pueden ser transformadas de una manera que pasan a ser aleatorias, al menos en la misma medida en que los números pseudo-aleatorios son aleatorios. Bhat (2003) describe el proceso, basado en los procedimientos introducidos por Tuffin (1996):

1. Extraiga un valor de una densidad uniforme estándar. Etiquete este valor como μ.
2. Añada μ a cada elemento de la secuencia de Halton. Si el elemento resultante es superior a 1, se resta 1 del mismo. De lo contrario, mantenga el elemento resultante como está (sin restar 1).

La fórmula que describe esta transformación es $s_n = mod(s_0 + \mu)$, donde s_0 es el elemento original de la secuencia de Halton, s_n es el elemento transformado y mod es la operación consistente en tomar la parte decimal del argumento entre paréntesis.

La transformación se representa en la figura 9.17. Supongamos que el valor μ extraído de la densidad uniforme es 0.40. El número 0.33 es el primer elemento de la secuencia de Halton para el número primo 3. Transformamos este elemento, como se muestra en la parte superior de la figura 9.17, a $0.33 + 0.40 = 0.73$, que es sólo un desplazamiento de 0.40 en la línea. El número 0.67 es el segundo elemento de la

secuencia. Se transforma mediante la adición de 0.40 y, a continuación, ya que el resultado es superior a 1, restamos 1 para obtener 0.07 $(0.67 + 0.40 - 1 = 0.07)$. Como se muestra en la parte inferior de la figura 9.17, esta transformación se visualiza como un movimiento del punto original hacia arriba a una distancia de 0.40, pero volviendo al inicio cuando se alcanza el final del intervalo unitario. El punto se mueve hacia arriba 0.33 hasta donde termina la línea y luego se lleva al inicio de la línea y continúa moviéndose hacia arriba otros 0.07, para completar así un movimiento total de 0.40.

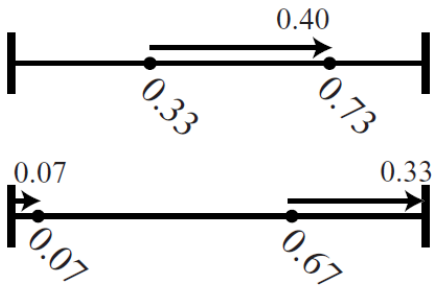


Figura 9.17. Transformación aleatoria de secuencias de Halton con μ=0.40.

La figura 9.18 representa la transformación para los cinco primeros elementos de la secuencia. La posición relativa entre los puntos y el grado de cobertura es el mismo antes y después de la transformación. Sin embargo, dado que la transformación se basa en el valor aleatorio μ, los valores numéricos de la secuencia transformada son aleatorios. La secuencia resultante se denomina secuencia de Halton aleatorizada. Tiene las mismas propiedades de cobertura y correlación negativa entre observaciones de la secuencia de Halton original, ya que la colocación relativa de los elementos es la misma; sin embargo, ahora la secuencia es aleatoria.

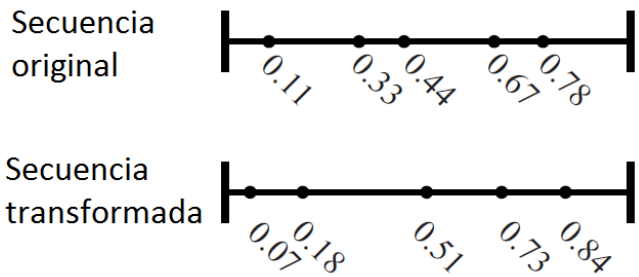


Figura 9.18. Aleatorización de una secuencia de Halton en una dimensión.

Con múltiples dimensiones, la secuencia utilizada para cada dimensión se transforma por separado en función de su propio valor extraído al azar de la densidad uniforme estándar. La figura 9.19 representa una transformación de una secuencia de dos dimensiones de longitud 3 definida para números primos 2 y 3. La secuencia para el número primo 3 está representada en el eje x y obtiene un valor aleatorio de 0.40. La secuencia para el número primo 2 obtiene un valor aleatorio de 0.35. Cada punto en la secuencia bidimensional original se mueve a la derecha 0.40 y hacia arriba 0.35, volviendo al origen cuando es necesario. La posición relativa entre los puntos en cada dimensión se mantiene, y sin embargo la secuencia es ahora aleatoria.

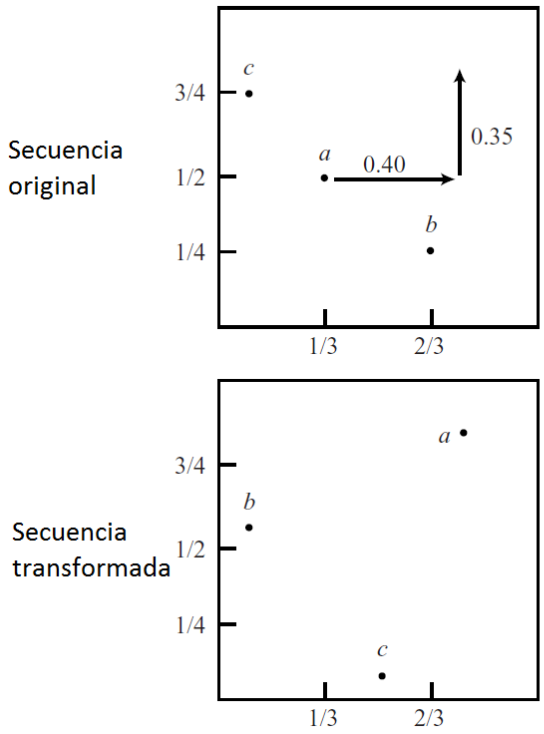


Figura 9.19. Aleatorización de una secuencia de Halton en dos dimensiones.

9.3.5 Secuencias de Halton mezcladas

Otro problema con las secuencias de Halton surge cuando se utilizan con muchas dimensiones. Para la simulación de integrales de alta dimensionalidad, se necesitan secuencias de Halton basadas en números primos grandes. Por ejemplo, con 15 dimensiones, se necesitan los números primos existentes hasta 47. Sin embargo, las secuencias de Halton definidas por números primos grandes pueden estar altamente correlacionadas entre sí a través de extensas porciones de la secuencia. La correlación no se limita a los elementos iniciales como se ha descrito anteriormente, por lo que no podemos eliminar la correlación descartando estos elementos. Dos secuencias definidas por números primos grandes y similares se sincronizan periódicamente entre sí y permanecen así durante muchos ciclos.

Bhat (2003) describe el problema y proporciona una solución eficaz. La figura 9.20 reproduce un gráfico incluido en su artículo que representa la secuencia de Halton para los números primos 43 y 47. Claramente, estas secuencias están altamente correlacionadas.

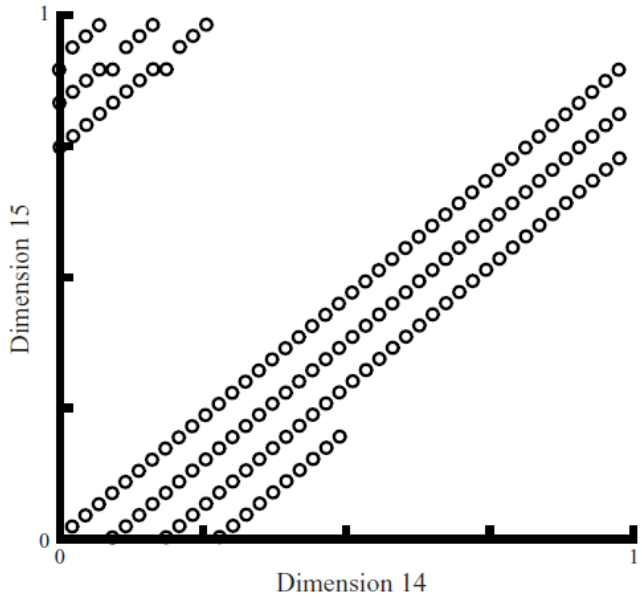


Figura 9.20. Secuencia de Halton estándar.

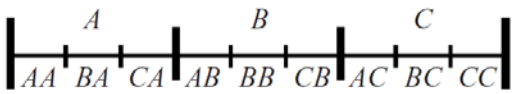


Figura 9.21. Segmentos definidos para mezclar secuencias de Halton.

Esta correlación se puede suprimir, preservando al mismo tiempo la cobertura deseable de las secuencias de Halton, mediante la mezcla de los dígitos de cada elemento de la secuencia. La mezcla se puede hacer de varias maneras. Braatan y Weller (1979) describen un procedimiento que se explica más fácilmente a través de un ejemplo. Considere la secuencia de Halton para el número primo 3:

$$\frac{1}{3},\frac{2}{3},\frac{1}{9},\frac{4}{9},\frac{7}{9},\frac{2}{9},\frac{5}{9},\frac{8}{9},\ldots$$

Recordemos que la secuencia se crea dividiendo el intervalo unitario en tres segmentos, que denominamos A, B y C en la figura 9.21. Cada segmento se divide en tres sub-segmentos, etiquetados como AA (indicando sub-segmento A del segmento A), BA (sub-segmento B del segmento A), CA, AB, BB, CB, AC, BC y CC. La secuencia de Halton no mezclada es el punto de inicio de cada segmento en orden alfabético e ignorando A (es decir, ignoramos A, $\frac{1}{3}$ para B, $\frac{2}{3}$ para C), seguido por el punto de inicio de cada sub-segmento en orden alfabético e ignorando A (es decir, ignoramos AA, AB y AC, $\frac{1}{9}$ para BA, $\frac{4}{9}$ para BB, $\frac{7}{9}$ para BC, $\frac{2}{9}$ para CA, $\frac{5}{9}$ para CB y $\frac{8}{9}$ para CC). Tenga en cuenta que los segmentos y sub-segmentos que empiezan con A se ignoran debido a que sus puntos de inicio o bien son 0 (para el segmento A) o ya están incluidos en la secuencia (por ejemplo, el punto de inicio del sub-segmento AB es el mismo que el punto de inicio del segmento B).

La secuencia mezclada se obtiene mediante la inversión de B y C, es decir, considerando que C está antes que B en el alfabeto. La lista alfabética pasa a ser: segmentos A C B, sub-segmentos AA AC AB CA

CC CB BA BC BB. La secuencia se crea así de la misma manera que antes pero con este nuevo orden alfabético: ignoramos A, $\frac{2}{3}$ para C, $\frac{1}{3}$ para B; ignoramos AA, AC y AB, $\frac{2}{9}$ para CA, $\frac{8}{9}$ para CC, $\frac{5}{9}$ para CB, $\frac{1}{9}$ para BA, $\frac{7}{9}$ para BC, $\frac{4}{9}$ para BB. Las secuencia original y la mezclada son:

original	mezclado
1/3	2/3
2/3	1/3
1/9	2/9
4/9	8/9
7/9	5/9
2/9	1/9
5/9	7/9
8/9	4/9

Diferentes permutaciones de las letras se utilizan para diferentes números primos. La figura 9.22, hecha por Bhat (2003), muestra la secuencia mezclada para los números primos 43 y 47. Los puntos no están correlacionados como sí están en la secuencia original. Bhat demuestra que las secuencias mezcladas funcionan bien para integrales de alta dimensionalidad de la misma manera que las secuencias no mezcladas lo hacen para integrales de baja dimensionalidad.

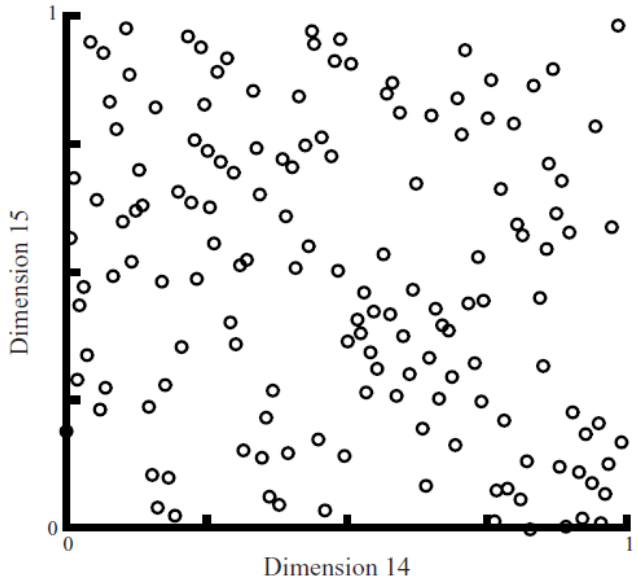


Figura 9.22. Secuencia de Halton mezclada.

9.3.6 Otros procedimientos

Hemos descrito sólo unos pocos de los procedimientos antitéticos y cuasi-aleatorios más importantes y directos. Procedimientos más complejos, con propiedades teóricas deseables, son descritos por Niederreiter (1978, 1988), Morokoff y Caflisch (1995), Joe y Sloan (1993), y Sloan y Wozniakowski (1998), por nombrar sólo unos pocos en esta creciente área de investigación. Como hemos visto con las secuencias de Halton, procedimientos bastante simples pueden proporcionar grandes mejoras respecto a la extracción de valores al azar. Las comparaciones realizadas por Sándor y András (2001) sobre probit y Sándor y Train (2004) sobre logit mixtos, indican que la precisión de la estimación basada en la

simulación de modelos de elección discreta se puede mejorar aún más con procedimientos más complejos. Es importante recordar, sin embargo, pese al entusiasmo que despiertan los resultados de estos métodos, que la precisión siempre se puede mejorar mediante el simple uso de más valores. El investigador debe decidir si le conviene aprender y programar nuevos métodos de extracción de valores, dadas sus limitaciones de tiempo, en lugar de simplemente estimar su modelo con más valores.

10

Estimación asistida por simulación

10.1 Motivación

Hasta ahora hemos estudiado cómo simular probabilidades de elección pero no hemos estudiado las propiedades de los estimadores de los parámetros que se basan en estas probabilidades simuladas. En los casos que hemos presentado, simplemente hemos insertado las probabilidades simuladas en la función log-verosimilitud y hemos maximizado dicha función, de la misma forma que lo habríamos hecho si las probabilidades hubieran sido exactas. Este procedimiento parece intuitivamente razonable. Sin embargo, no hemos mostrado realmente, al menos hasta ahora, que el estimador resultante tenga propiedades deseables, como consistencia, normalidad asintótica o eficiencia. Tampoco hemos explorado la posibilidad de que otras formas de estimación puedan ser preferibles cuando usamos simulación, en lugar de las probabilidades exactas.

El propósito de este capítulo es examinar varios métodos de estimación en el contexto de la simulación. Derivaremos las propiedades de estos estimadores y mostraremos las condiciones en las que cada estimador es consistente y asintóticamente equivalente al estimador que obtendríamos si usásemos valores exactos en lugar de simulación. Estas condiciones proporcionan una guía al investigador sobre cómo debe llevarse a cabo la simulación para obtener estimadores con propiedades deseables. El análisis también pone en evidencia las ventajas y limitaciones de cada forma de estimación, facilitando así la elección del investigador entre los diferentes métodos.

Consideraremos 3 métodos de estimación:

1. *Máxima verosimilitud simulada (maximum simulated likelihood, MSL)*: Este procedimiento es igual al de máxima verosimilitud (ML) excepto que emplea las probabilidades simuladas en lugar de las probabilidades exactas. Las propiedades del método MSL han sido obtenidas, por ejemplo, por Gourieroux y Monfort, (1993), Lee (1995), y Hajivassiliou y Ruud (1994).
2. *Método de momentos simulados (method of simulated moments, MSM)*: Este procedimiento, sugerido por McFadden (1989), es el análogo simulado del método de momentos tradicional (*method of moments,* MOM). Usando el MOM tradicional en elección discreta, los residuos se definen como la diferencia entre la variable dependiente 0-1 que identifica la alternativa elegida y la probabilidad de dicha alternativa. Se identifican variables exógenas que no estén correlacionadas con los residuos del modelo en la población. Las estimaciones son los valores de los parámetros que hacen que las variables y los residuos no estén correlacionados en la

muestra. La versión simulada de este procedimiento calcula los residuos con las probabilidades simuladas en lugar de las probabilidades exactas.

3. *Método de puntuaciones simuladas (method of simulated scores, MSS)*: Como vimos en el Capítulo 8, el gradiente de la función log-verosimilitud de una observación recibe el nombre de puntuación (*score*) de la observación. El método de puntuaciones encuentra los valores de los parámetros que hacen que la puntuación media sea cero. Cuando se utilizan probabilidades exactas, el método de las puntuaciones es el mismo que el de máxima verosimilitud, ya que la función log-verosimilitud se maximiza cuando la puntuación media es cero. Hajivassiliou y McFadden (1998) sugirieron el uso de puntuaciones simuladas en lugar de puntuaciones exactas. Ellos mostraron que, dependiendo de cómo se simulan las puntuaciones, MSS puede diferir de MSL y, más importante, puede alcanzar consistencia y eficiencia bajo condiciones más relajadas.

En la siguiente sección definimos estos estimadores más formalmente y los relacionamos con sus equivalentes no simulados. A continuación describimos las propiedades de cada estimador en dos etapas. En primer lugar, se obtienen las propiedades del estimador tradicional basado en los valores exactos. En segundo lugar, se muestra cómo cambia la formulación cuando se utilizan valores simulados y no valores exactos. Mostramos que la simulación añade elementos adicionales a la distribución muestral del estimador. El análisis nos permite identificar las condiciones en que estos elementos adicionales desaparecen asintóticamente para que el estimador sea asintóticamente equivalente a su análogo no simulado. También identificamos las condiciones más relajadas en las que el estimador, aunque no sea asintóticamente equivalente a su homólogo no simulado, es sin embargo consistente.

10.2 Definición de estimadores

10.2.1 Máxima Verosimilitud Simulada (maximum simulated likelihood, MSL)

La función de verosimilitud es

$$LL(\theta) = \sum_n ln\, P_n(\theta),$$

donde θ es un vector de parámetros, $P_n(\theta)$ es la probabilidad (exacta) de la elección observada correspondiente a la observación n, y el sumatorio es sobre una muestra de N observaciones independientes. El estimador ML es el valor de θ que maximiza $LL(\theta)$. Dado que el gradiente de $LL(\theta)$ es cero en el máximo, el estimador ML también se puede definir como el valor de θ en el que

$$\sum_n s_n(\theta) = 0,$$

donde $s_n(\theta) = \partial \ln P_n(\theta) / \partial\theta$ es la puntuación de la observación n.

Sea $\check{P}_n(\theta)$ una aproximación simulada de $P_n(\theta)$. La función log-verosimilitud simulada es $SLL(\theta) = \sum_n ln\, \check{P}_n(\theta)$ y el estimador MSL es el valor de θ que maximiza $SLL(\theta)$. Dicho de forma equivalente, el estimador es el valor de θ en el que $\sum_n \check{s}_n(\theta) = 0$, donde $\check{s}_n(\theta) = \partial \ln \check{P}_n(\theta) / \partial\theta$.

Podemos echar ahora un primer vistazo a las propiedades del estimador MSL, reservando una explicación completa para la siguiente sección. El principal problema con el estimador MSL surge debido a la transformación logarítmica. Supongamos que $\check{P}_n(\theta)$ es un simulador no sesgado de $P_n(\theta)$, de manera que $E_r \check{P}_n(\theta) = P_n(\theta)$, donde la esperanza es sobre los valores extraídos al azar utilizados en la simulación. Todos los simuladores que hemos considerado son no sesgados respecto a la

verdadera probabilidad. Sin embargo, dado que el operador logarítmico es una transformación no lineal, $\ln \check{P}_n(\theta)$ es sesgado respecto a $\ln P_n(\theta)$ a pesar de que $\check{P}_n(\theta)$ es no sesgado respecto $P_n(\theta)$. El sesgo en el simulador de $\ln P_n(\theta)$ se traduce en un sesgo en el estimador MSL. Este sesgo disminuye a medida que se utilizan más valores en la simulación.

Para determinar las propiedades asintóticas del estimador MSL, se plantea la cuestión de cómo se comporta el sesgo de simulación cuando el tamaño de la muestra aumenta. La respuesta depende críticamente de la relación entre el número de valores que se utilizan en la simulación, etiquetado como R, y el tamaño de la muestra N. Si R se considera fijo, entonces el estimador MSL no converge a los parámetros reales, debido al sesgo de simulación en $\ln \check{P}_n(\theta)$. Supongamos por el contrario que R se eleva con N; es decir, el número de valores usados en la simulación aumenta con el tamaño de la muestra. En este caso, el sesgo de simulación desaparece a medida que N (y por lo tanto R) se eleva sin límite. MSL es consistente en este caso. Como veremos, si R aumenta más rápidamente que $\sqrt{N}$, MSL no sólo es consistente sino también eficiente, asintóticamente equivalente a la máxima verosimilitud con probabilidades exactas.

En resumen, si R es fijo, entonces MSL es inconsistente. Si R se eleva con N en cualquier proporción, MSL es consistente. Si R se eleva más rápido que $\sqrt{N}$, MSL es asintóticamente equivalente a ML.

La principal limitación de MSL es que es inconsistente para un R fijo. Los otros estimadores que consideraremos están motivados por el deseo de tener un estimador basado en simulación que sea consistente para un R fijo. Tanto MSM como MSS, si se estructuran adecuadamente, logran este objetivo. Este beneficio tiene un precio, sin embargo, como veremos en la siguiente sección.

10.2.2 Método de momentos simulados (method of simulated moments, MSM)

El método de momentos tradicional (*method of moments, MOM*) está motivado por el hecho de que los residuos de un modelo están necesariamente incorrelacionados en la población con factores que son exógenos al comportamiento que está siendo modelado. El estimador MOM es el valor de los parámetros que hace que los residuos en la *muestra* no estén correlacionados con las variables exógenas. Para los modelos de elección discreta, MOM se define como los parámetros que resuelven la ecuación

$$\sum_n \sum_j [d_{nj} - P_{nj}(\theta)] z_{nj} = 0, \qquad (10.1)$$

donde

- d_{nj} es la variable dependiente que identifica la alternativa elegida: $d_{nj} = 1$ si n eligió j, y $d_{nj} = 0$ en caso contrario, y
- z_{nj} es un vector de variables exógenas llamadas instrumentos (*instruments*).

Los residuos son $d_{nj} - P_{nj}(\theta)$, y el estimador MOM es el conjunto de valores de los parámetros para los que los residuos no están correlacionados con los instrumentos en la muestra.

Este estimador MOM es análogo a los estimadores MOM de los modelos de regresión estándar. Un modelo de regresión adopta la forma $y_n = x_n'\beta + \varepsilon_n$. El estimador MOM para esta regresión es la β en la que

$$\sum_n (y_n - x_n'\beta) z_n = 0$$

para un vector de instrumentos exógenos z_n. Cuando las variables explicativas en el modelo son exógenas, entonces éstas sirven como instrumentos. En este caso, el estimador MOM se convierte en el estimador de mínimos cuadrados ordinarios:

$$\sum_n (y_n - x_n'\beta)x_n = 0,$$

$$\sum_n x_n y_n = \sum_n x_n x_n'\beta,$$

$$\hat{\beta} = \left(\sum_n x_n x_n'\right)^{-1}\left(\sum_n x_n y_n\right),$$

que es la fórmula para el estimador de mínimos cuadrados. Cuando los instrumentos se especifican para que sean otras variables distintas a las variables explicativas, el estimador se convierte en el estimador de variables instrumentales estándar:

$$\sum_n (y_n - x_n'\beta)z_n = 0,$$

$$\sum_n z_n y_n = \sum_n z_n x_n'\beta,$$

$$\hat{\beta} = \left(\sum_n z_n x_n'\right)^{-1}\left(\sum_n z_n y_n\right),$$

que es la fórmula para el estimador de variables instrumentales. Este estimador es consistente si los instrumentos son independientes de ε en la población. El estimador es más eficiente cuanto más correlacionados están los instrumentos con las variables explicativas del modelo. Cuando las variables explicativas, x_n, son a su vez exógenas, los instrumentos ideales (es decir, los que dan la eficiencia más alta) son las propias variables explicativas, $z_n = x_n$.

Para los modelos de elección discreta, MOM se define de forma análoga y tiene una relación similar a otros estimadores, especialmente ML. El investigador identifica los instrumentos z_{nj} que son variables exógenas y por lo tanto independientes de los residuos $\left[d_{nj} - P_{nj}(\theta)\right]$ en la población. El estimador MOM es el valor de θ en el que la correlación de la muestra entre los instrumentos y los residuos es cero. A diferencia del caso lineal, la ecuación (10.1) no se puede resolver de forma explícita para $\hat{\theta}$. En lugar de ello, se utilizan procedimientos numéricos para encontrar el valor de θ que resuelve esta ecuación.

Al igual que sucede con la regresión, ML para un modelo de elección discreta es un caso especial de MOM. Hagamos que los instrumentos sean las puntuaciones: $z_{nj} = \partial \ln P_{nj}(\theta)/\partial\theta$. Con estos instrumentos, MOM es el mismo que ML:

$$\sum_n \sum_j [d_{nj} - P_{nj}(\theta)]\, z_{nj} = 0,$$

$$\sum_n \left\{ \left(\sum_j d_{nj} \frac{\partial \ln P_{nj}(\theta)}{\partial \theta} \right) - \left(\sum_j P_{nj}(\theta) \frac{\partial \ln P_{nj}(\theta)}{\partial \theta} \right) \right\} = 0,$$

$$\sum_n \frac{\partial \ln P_{ni}(\theta)}{\partial \theta} - \sum_n \sum_j P_{nj}(\theta) \frac{1}{P_{nj}(\theta)} \frac{\partial P_{nj}(\theta)}{\partial \theta} = 0,$$

$$\sum_n s_n(\theta) - \sum_n \sum_j \frac{\partial P_{nj}(\theta)}{\partial \theta} = 0,$$

$$\sum_n s_n(\theta) = 0,$$

que es la condición que define ML. En la tercera línea, i es la alternativa elegida, reconociendo que $d_{nj} = 0$ para todo j ≠ i. La cuarta línea utiliza el hecho de que la suma de $\partial P_{nj}(\theta)/\partial\theta$ sobre las alternativas es cero, ya que las probabilidades deben sumar 1 antes y después del cambio en θ.

Dado que MOM se convierte en ML y por lo tanto es plenamente eficiente cuando los instrumentos son las puntuaciones, las puntuaciones son llamadas instrumentos ideales. MOM es consistente siempre que los instrumentos sean independientes de los residuos del modelo. Es más eficiente cuanto mayor es la correlación entre los instrumentos y los instrumentos ideales.

Una simplificación interesante surge con el modelo logit estándar. Para el modelo logit estándar, los instrumentos ideales son las propias variables explicativas. Como se muestra en la sección 3.7.1, el estimador ML para logit estándar es el valor de θ que resuelve $\sum_n \sum_j [d_{nj} - P_{nj}(\theta)]\, x_{nj} = 0$, donde x_{nj} son las variables explicativas. Se trata de un estimador MOM con las variables explicativas como instrumentos.

Una versión simulada de MOM, llamado el método de momentos simulados (*method of simulated moments, MSM*), se obtiene mediante la sustitución de las probabilidades exactas $P_{nj}(\theta)$ por las probabilidades simuladas $\check{P}_{nj}(\theta)$. El estimador MSM es el valor de θ que resuelve

$$\sum_n \sum_j [d_{nj} - \check{P}_{nj}(\theta)]\, z_{nj} = 0,$$

para los instrumentos z_{nj}. Al igual que sucede con su analógo no simulado, MSM es consistente si z_{nj} es independiente de $d_{nj} - \check{P}_{nj}(\theta)$.

La característica importante de este estimador es que $\check{P}_{nj}(\theta)$ entra en la ecuación linealmente. Como resultado, si $\check{P}_{nj}(\theta)$ es un simulador no sesgado de $P_{nj}(\theta)$, entonces $[d_{nj} - \check{P}_{nj}(\theta)]\, z_{nj}$ es no sesgado respecto $[d_{nj} - P_{nj}(\theta)]\, z_{nj}$. Puesto que no hay sesgo de simulación en la condición de estimación, el estimador MSM es consistente, incluso cuando el número R de valores extraídos para la simulación es fijo. Por el contrario, MSL contiene sesgo de simulación debido a la transformación logarítmica de las probabilidades simuladas. Al no hacer una transformación no lineal de las probabilidades simuladas, MSM evita el sesgo de simulación.

Aun así, MSM contiene ruido de simulación (la varianza debida a la simulación). Este ruido se reduce a medida que R se eleva y desaparece cuando R aumenta sin límite. Como resultado, MSM es asintóticamente equivalente a MOM si R aumenta con N.

Al igual que su análogo no simulado, MSM es menos eficiente que MSL a no ser que se utilicen los instrumentos ideales. Sin embargo, los instrumentos ideales son funciones de $\ln P_{nj}$. Estos no pueden ser calculados de forma exacta excepto para los modelos más simples y, si son simulados utilizando la probabilidad simulada, se introduce sesgo de simulación debido a la operación logarítmica. MSM se aplica por lo general con pesos no ideales, lo que significa que se produce una pérdida de eficiencia. MSM con pesos ideales simulados sin sesgo se convierte en MSS, algo que veremos en la siguiente sección.

En resumen, MSM tiene la ventaja sobre MSL de ser consistente usando un número fijo de valores extraídos para simulación. Sin embargo, nada es gratuito, y el costo de esta ventaja es una pérdida de eficiencia cuando se utilizan pesos no ideales.

10.2.3 Método de puntuaciones simuladas (method of simulated scores, MSS)

MSS proporciona una posibilidad de lograr consistencia sin pérdida de eficiencia. El costo de esta doble ventaja es numérico: las versiones de MSS que proporcionan eficiencia tienen propiedades numéricas bastante pobres, de manera que el cálculo del estimador puede ser difícil.

El método de puntuaciones se define por la condición

$$\sum_n s_n(\theta) = 0,$$

donde $s_n(\theta) = \partial P_n(\theta)/\partial\theta$ es la puntuación de la observación n. Esta es la misma condición que define ML: cuando se utilizan probabilidades exactas, el método de puntuaciones es simplemente ML.

El método de puntuaciones simuladas reemplaza la puntuación exacta por su análogo simulado. El estimador MSS es el valor de θ que resuelve

$$\sum_n \check{s}_n(\theta) = 0,$$

donde $\check{s}_n(\theta)$ es un simulador de la puntuación. Si $\check{s}_n(\theta)$ se calcula como la derivada del logaritmo de la probabilidad simulada, es decir, $\check{s}_n(\theta) = \partial \check{P}_n(\theta)/\partial\theta$, entonces MSS es igual a MSL. Sin embargo, la puntuación se puede simular de otras maneras. Cuando la puntuación se simula de otras maneras, MSS difiere de MSL y tiene propiedades diferentes.

Supongamos que es posible construir un simulador no sesgado de la puntuación. Con este simulador, la ecuación que define el método, $\sum_n \check{s}_n(\theta) = 0$, no incorpora ningún sesgo de simulación, ya que el simulador entra en la ecuación de forma lineal. Por lo tanto, MSS es consistente con una R fija. El ruido de simulación disminuye a medida que aumenta R, de tal forma que MSS es asintóticamente eficiente, equivalente a MSL, cuando R aumenta con N. En contraste, MSL utiliza el simulador de puntuación sesgado $\check{s}_n(\theta) = \partial \check{P}_n(\theta)/\partial\theta$, que es sesgado debido al uso del operador logarítmico. Por lo tanto, MSS con un simulador de puntuación no sesgado es mejor que MSL con su simulador de puntuación sesgado, en dos aspectos: es consistente en condiciones menos estrictas (para una R fija en lugar de una R creciente con N) y es eficiente en condiciones menos estrictas (R creciente con N en cualquier proporción, en lugar de R creciendo más rápido que $\sqrt{N}$).

La dificultad en el uso de MSS está en encontrar un simulador de puntuación no sesgado. La puntuación puede ser reescrita como

$$s_n(\theta) = \frac{\partial \ln P_{nj}(\theta)}{\partial \theta} = \frac{1}{P_{nj}(\theta)} \frac{\partial P_{nj}}{\partial \theta}.$$

Un simulador no sesgado para el segundo término $\partial P_{nj}(\theta)/\partial\theta$ se obtiene fácilmente tomando la derivada de la probabilidad simulada. Puesto que la diferenciación es una operación lineal, $\partial \check{P}_{nj}(\theta)/\partial\theta$ es no sesgado respecto $\partial P_{nj}(\theta)/\partial\theta$ si $\check{P}_{nj}(\theta)$ es a su vez no sesgado respecto $P_{nj}(\theta)$. Dado que el segundo término de la puntuación puede ser simulado sin sesgo, la dificultad se presenta en la búsqueda de un simulador no sesgado para el primer término $1/P_{nj}(\theta)$. Por supuesto, simplemente tomar la inversa de la probabilidad simulada no proporciona un simulador no sesgado, ya que $\mathrm{E}_r(1/\check{P}_{nj}(\theta)) \neq 1/P_{nj}(\theta)$. Al igual que la operación logarítmica, una inversa introduce sesgo.

Una propuesta para resolver este problema se basa en el hecho de que $1/P_{nj}(\theta)$ es el número esperado de valores extraídos al azar de los términos aleatorios que se necesitan hasta lograr una "aceptación". Para ilustrar esta idea, considere la extracción de bolas de una urna que contiene muchas bolas de diferentes colores. Supongamos que la probabilidad de obtener una bola roja es 0.20. Es decir, una quinta parte de las bolas son de color rojo. ¿Cuántas extracciones se necesitarían, en promedio, para obtener una bola roja? La respuesta es 1/0.2 = 5. La misma idea se puede aplicar a las probabilidades de elección. $P_{nj}(\theta)$ es la probabilidad de que una extracción de los términos aleatorios del modelo resulte en que la alternativa j tenga la mayor utilidad. La inversa $1/P_{nj}(\theta)$ se puede simular como sigue:

1. Extraiga un valor al azar de los términos aleatorios a partir de su densidad.
2. Calcule la utilidad de cada alternativa con este valor.
3. Determine si la alternativa j tiene la mayor utilidad.
4. Si es así, catalogue el valor como una “aceptación”. Si no es así, catalogue el valor como un “rechazo” y repita los pasos 1 a 3 con un nuevo valor. Defina B^r como el número de extracciones que se realizan hasta que se obtiene la primera aceptación.
5. Realice los pasos 1 a 4 R veces, obteniendo B^r para $\mathrm{r} = 1, \ldots, \mathrm{R}$. El simulador de $1/P_{nj}(\theta)$ es $(1/\mathrm{R}) \sum_{r=1}^{R} \mathrm{B}^r$.

Este simulador es no sesgado respecto $1/P_{nj}(\theta)$. El producto de este simulador con el simulador $\partial \check{P}_{nj}(\theta)/\partial\theta$ proporciona un simulador no sesgado de la puntuación. MSS basado en este simulador de puntuación no sesgado es consistente para un R fijo y asintóticamente eficiente cuando R aumenta con N.

Por desgracia, el simulador de $1/P_{nj}(\theta)$ tiene las mismas dificultades que los simuladores de aceptación-rechazo que vimos en la sección 5.6. No hay garantía de que vayamos a obtener una aceptación dentro de un número dado de valores extraídos. Además, el simulador no es continuo en los parámetros. La discontinuidad dificulta los procedimientos numéricos que se utilizan para localizar los parámetros que resuelven la ecuación de MSS.

En resumen, MSS tiene ventajas y desventajas en relación a MSL, al igual que sucede con MSM. La comprensión de las capacidades de cada estimador permite al investigador realizar una elección informada entre ellos.

10.3 El teorema del límite central

Antes de obtener las propiedades de nuestros estimadores, es útil revisar el teorema del límite central. Este teorema proporciona la base de las distribuciones de los estimadores.

Uno de los resultados más básicos en estadísticas es que, si extraemos valores al azar de una distribución con media μ y varianza σ, la media de estos valores se distribuye normalmente con media μ

y varianza σ/N, donde N es un número grande de valores extraídos. Este resultado es el teorema del límite central, expresado de forma intuitiva en lugar de precisa. Vamos a ofrecer un desarrollo más completo y preciso de estas ideas.

Sea $t = (1/N)\sum_n t_n$, donde cada t_n es un valor extraído al azar de una distribución con media μ y varianza σ. Una realización concreta de valores extraídos al azar recibe el nombre de muestra y t es la media de la muestra. Si tomamos una muestra diferente (es decir, obtenemos diferentes valores para las extracciones de cada t_n), entonces obtenemos un valor diferente para el estadístico t. Nuestro objetivo es obtener la distribución muestral de t.

Para la mayoría de estadísticos, no podemos determinar con exactitud la distribución muestral para un tamaño de muestra dado. En su lugar, analizamos cómo se comporta la distribución muestral a medida que el tamaño de la muestra aumenta sin límite. Llegados a este punto, debemos hacer una distinción entre la distribución límite (*limiting distribution*) y la distribución asintótica (*asymptotic distribution*) de un estadístico. Supongamos que, a medida que aumenta el tamaño de la muestra, la distribución muestral del estadístico t converge a una distribución fija. Por ejemplo, la distribución muestral de t podría llegar a estar arbitrariamente cerca de una normal con media t^* y varianza σ. En este caso, decimos que $N(t^*, \sigma)$ es la distribución límite de t y que t converge en distribución a $N(t^*, \sigma)$. Denotamos esta situación como $t \xrightarrow{d} N(t^*, \sigma)$.

En muchos casos, un estadístico no tendrá una distribución límite. A medida que aumenta N, la distribución muestral sigue cambiando. La media de una muestra de valores extraídos es un ejemplo de un estadístico sin una distribución límite. Como se ha indicado anteriormente, si t es la media de una muestra de valores extraídos de una distribución con media μ y varianza σ, entonces t se distribuye normalmente con media μ y varianza σ/N. La varianza disminuye a medida que N se eleva. La distribución cambia a medida que N aumenta, siendo cada vez más y más estrecha alrededor de la media. Si se tuviera que definir una distribución límite para este caso, tendría que ser la distribución degenerada en μ: a medida que N se eleva sin límite, la distribución de t colapsa en μ. Esta distribución límite es inútil para la comprensión de la varianza del estadístico, ya que la varianza de esta distribución límite es cero. ¿Qué hacemos en este caso para comprender las propiedades del estadístico?

Si nuestro estadístico original no tiene una distribución límite, a menudo podemos transformar el estadístico de tal manera que el estadístico transformado sí tenga una distribución límite. Supongamos, como en nuestro ejemplo de una media de la muestra, que el estadístico que nos interesa no tiene una distribución límite porque su varianza disminuye a medida que aumenta N. En ese caso, podemos considerar una transformación del estadístico normalizado respecto al tamaño muestral. En particular, podemos considerar $\sqrt{N}(t-\mu)$. Supongamos que este estadístico sí tiene una distribución límite, por ejemplo, $\sqrt{N}(t-\mu) \xrightarrow{d} N(0, \sigma)$. En este caso, podemos obtener las propiedades de nuestro estadístico original a partir de la distribución límite del estadístico transformado. Recordemos, a partir de principios básicos de probabilidad, que para unos valores a y b dados, si $a(t-b)$ se distribuye normalmente con media cero y varianza σ, entonces t se distribuye normalmente con media b y varianza σ/a^2. Esta relación puede aplicarse a nuestra distribución límite. Para un N suficientemente grande, $\sqrt{N}(t-\mu)$ se distribuye aproximadamente $N(0, \sigma)$. Por lo tanto, para un N suficientemente grande, t se distribuye aproximadamente $N(\mu, \sigma/N)$. Denotamos esto como $t \sim^a N(\mu, \sigma/N)$. Observe que ésta no es la distribución límite de t, ya que t no tiene una distribución límite no degenerada. En su lugar, se denomina distribución asintótica de t, obtenida a partir de la distribución límite de $\sqrt{N}(t-\mu)$.

Ahora podemos re-expresar de forma precisa nuestros conceptos acerca de la distribución muestral de la media de la muestra. El teorema del límite central establece lo siguiente. Supongamos que t es la

media de una muestra de N valores extraídos de una distribución con media μ y varianza σ. Entonces $\sqrt{N}(t-\mu) \xrightarrow{d} N(0,\sigma)$. Con esta distribución límite, podemos decir que $t \sim^{a} N(\mu, \sigma/N)$.

Hay otra versión, más general, del teorema del límite central. En la versión que acabamos de exponer, cada t_n es una extracción de la misma distribución. Supongamos que t_n es una extracción de una distribución con media μ y varianza σ_n, para $n = 1, \ldots, N$. Es decir, cada t_n proviene de una distribución diferente; las distribuciones tienen la misma media pero diferentes varianzas. La versión generalizada del teorema del límite central establece que $\sqrt{N}(t-\mu) \xrightarrow{d} N(0,\sigma)$, donde σ es ahora la varianza media: $\sigma = (1/N)\sum_n \sigma_n$. Dada esta distribución límite, podemos decir que $t \sim^{a} N(\mu, \sigma/N)$. Vamos a utilizar ambas versiones del teorema del límite central al obtener las distribuciones de nuestros estimadores.

10.4 Propiedades de los estimadores tradicionales

En esta sección, revisaremos el procedimiento para obtener las propiedades de los estimadores y aplicaremos este procedimiento para los estimadores tradicionales, no basados en simulación. Esta exposición es el fundamento del análisis de las propiedades de los estimadores basados en simulación que abordaremos en la siguiente sección.

Denotemos el verdadero valor de los parámetros como θ^*. Los estimadores ML y MOM son las raíces de una ecuación que toma la forma

$$\sum_n g_n(\hat{\theta})/N = 0. \qquad (10.2)$$

Es decir, el estimador $\hat{\theta}$ es el valor de los parámetros que resuelve esta ecuación. Dividimos por N, a pesar de que esta división no afecta a la raíz de la ecuación, ya que al hacerlo facilitamos el cálculo de las propiedades de los estimadores. La condición establece que el valor promedio de $g_n(\theta)$ en la muestra es cero en los parámetros estimados. Para ML, $g_n(\theta)$ es la puntuación $\partial \ln P_n(\theta)/\partial\theta$. Para MOM, $g_n(\theta)$ es el conjunto de los primeros momentos de los residuos respecto a un vector de instrumentos, $\sum_j (d_{nj} - P_{nj}) z_{nj}$. La ecuación (10.2) se llama a menudo la condición de momento. En su forma no simulada, el método de puntuaciones es igual a ML y por lo tanto no necesita ser considerado por separado en esta sección. Tenga en cuenta que nosotros llamamos ecuación a (10.2) a pesar de que en realidad es un conjunto de ecuaciones, ya que $g_n(\theta)$ es un vector. Los parámetros que resuelven estas ecuaciones son los estimadores.

En cualquier valor particular de θ pueden calcularse la media y la varianza de $g_n(\theta)$ en la muestra. Etiquete la media como $g(\theta)$ y la varianza como $W(\theta)$. Estamos especialmente interesados en la media muestral y la varianza de $g_n(\theta)$ en los verdaderos parámetros, θ^*, ya que nuestro objetivo es estimar estos parámetros.

La clave para entender las propiedades de un estimador está en darse cuenta de que cada $g_n(\theta^*)$ es una extracción de una distribución de $g_n(\theta^*)$'s en la población. No sabemos los verdaderos parámetros, pero sabemos que cada observación tiene un valor de $g_n(\theta^*)$ en los verdaderos parámetros. El valor de $g_n(\theta^*)$ varía entre personas de la población. Así, extrayendo una persona de nuestra muestra, básicamente estamos extrayendo un valor de $g_n(\theta^*)$ de su distribución en la población.

La distribución de $g_n(\theta^*)$ en la población tiene una media y una varianza. Etiquete la media de $g_n(\theta^*)$ en la población como $\mathbf{g}$ y su varianza en la población como $\mathbf{W}$. La media y la varianza muestral en los verdaderos parámetros, $g(\theta^*)$ y $W(\theta^*)$, son el equivalente en la muestra a la media y varianza en la población, $\mathbf{g}$ y $\mathbf{W}$.

Asumimos que $\mathbf{g} = 0$. Es decir, asumimos que el promedio de $g_n(\theta^*)$ en la población es cero en los parámetros verdaderos. Bajo este supuesto, el estimador proporciona un análogo en la muestra a la

esperanza en la población: $\hat{\theta}$ es el valor de los parámetros en los cuales el promedio de $g_n(\theta)$ en la muestra es igual a cero, como se indica en la condición definitoria (10.2). Para ML, la suposición de que $\mathbf{g} = 0$ simplemente establece que la puntuación media en la población es cero, cuando se evalúa en los verdaderos parámetros. En cierto sentido, esto se puede considerar la definición de parámetros reales, es decir, θ^* son los parámetros en los que la función log-verosimilitud para toda la población obtiene su máximo y por lo tanto tiene pendiente cero. Los parámetros estimados son los valores que hacen que la pendiente de la función de verosimilitud en la muestra sea cero. Para MOM, el supuesto se cumple si los instrumentos son independientes de los residuos. En cierto sentido, la hipótesis con MOM es simplemente una reiteración de que los instrumentos son exógenos. Los parámetros estimados son los valores que hacen que los instrumentos y los residuos no estén correlacionados en la muestra.

Ahora consideraremos la varianza en la población de $g_n(\theta^*)$, lo que hemos denotado como $\mathbf{W}$. Cuando $g_n(\theta)$ es la puntuación, como sucede en ML, esta varianza tiene un significado especial. Como se ha mostrado en la sección 8.7, la identidad de información establece que $\mathbf{V} = -\mathbf{H}$, donde

$$-\boldsymbol{H} = -E\left(\frac{\partial^2 \ln P_n(\theta^*)}{\partial\theta\partial\theta'}\right)$$

es la matriz de información y $\mathbf{V}$ es la varianza de las puntuaciones evaluadas en los verdaderos parámetros: $\mathbf{V} = \mathrm{Var}(\partial \ln P_n(\theta^*)/\partial\theta)$. Cuando $g_n(\theta)$ es la puntuación, $\mathbf{W} = \mathbf{V}$ por definición y, por tanto, $\mathbf{W} = -\mathbf{H}$ por la identidad de información. Es decir, cuando $g_n(\theta)$ es la puntuación, $\mathbf{W}$ es la matriz de información. Para MOM con instrumentos no ideales, $\mathbf{W} \neq -\mathbf{H}$, de modo que $\mathbf{W}$ no es igual a la matriz de información.

¿Por qué es importante esta distinción? Veremos que saber si $\mathbf{W}$ es igual a la matriz de información nos permite determinar si el estimador es eficiente. La menor varianza que un estimador cualquiera puede lograr es $-\mathbf{H}^{-1}/N$. Para obtener una prueba, véase, por ejemplo, Greene (2000) o Ruud (2000). Un estimador es eficiente si su varianza alcanza este límite inferior. Como veremos, este límite inferior se logra cuando $\mathbf{W} = -\mathbf{H}$, pero no cuando $\mathbf{W} \neq -\mathbf{H}$.

Nuestro objetivo es determinar las propiedades de $\hat{\theta}$. Derivamos estas propiedades en un proceso en dos pasos. En primer lugar, se analiza la distribución de $\mathrm{g}(\theta^*)$, que, como se estableció anteriormente, es la media muestral de $g_n(\theta^*)$. En segundo lugar, la distribución de $\hat{\theta}$ se obtiene de la distribución de $\mathrm{g}(\theta^*)$. Este proceso en dos pasos no es necesariamente la forma más directa de examinar estimadores tradicionales. Sin embargo, como veremos en la siguiente sección, proporciona una forma muy conveniente de generalizar el análisis a estimadores basados en simulación.

Paso 1: Distribución de $\mathbf{g}(\boldsymbol{\theta}^*)$

Recuerde que el valor de $g_n(\theta^*)$ varía entre decisores de la población. Al tomar una muestra, el investigador está extrayendo valores $g_n(\theta^*)$ de su distribución en la población. Esta distribución tiene media cero por hipótesis y una varianza denotada por $\mathbf{W}$. El investigador calcula la media de la muestra de estos valores extraídos, $\mathrm{g}(\theta^*)$. Por el teorema del límite central, $\sqrt{\mathrm{N}}(\mathrm{g}(\theta^*) - 0) \xrightarrow{d} N(0, \mathbf{W})$, de tal manera que la media de la muestra tiene una distribución $\mathrm{g}(\theta^*) \sim^a N(0, \mathbf{W}/\mathrm{N})$.

Paso 2: Obtenga la distribución de $\widehat{\boldsymbol{\theta}}$ a partir de la distribución de $\mathbf{g}(\boldsymbol{\theta}^*)$

Podemos relacionar el estimador $\hat{\theta}$ con su término definitorio $\mathrm{g}(\theta)$ de la siguiente manera. Tome una expansión de Taylor de primer orden de $\mathrm{g}(\hat{\theta})$ alrededor $\mathrm{g}(\theta^*)$:

(10.3) $$g(\hat{\theta}) = g(\theta^*) + D[\hat{\theta} - \theta^*],$$

donde $D = \partial g(\theta^*)/\partial\theta'$. Por definición de $\hat{\theta}$ (es decir, mediante la definición de la condición (10.2)), $\mathrm{g}(\hat{\theta}) = 0$, de manera que el lado derecho de esta expansión es 0. Entonces

$$0 = g(\theta^*) + D\left[\hat{\theta} - \theta^*\right],$$

$$\hat{\theta} - \theta^* = -D^{-1}g(\theta^*),$$

(10.4) $$\sqrt{N}\left(\hat{\theta} - \theta^*\right) = \sqrt{N}(-D^{-1})g(\theta^*).$$

Denotemos la media de $\partial g_n(\theta^*)/\partial\theta'$ en la población como $\boldsymbol{D}$. La media de $\partial g_n(\theta^*)/\partial\theta'$ en la muestra es D, tal y como se define por la ecuación (10.3). La media en la muestra D converge a la media poblacional $\boldsymbol{D}$ a medida que el tamaño de la muestra crece. Sabemos del paso 1 que $\sqrt{\mathrm{N}}\mathrm{g}(\theta^*) \xrightarrow{d} N(0, \mathbf{W})$. Usando este hecho en (10.4), tenemos

(10.5) $$\sqrt{N}\left(\hat{\theta} - \theta^*\right) \xrightarrow{d} N(0, \boldsymbol{D}^{-1}\boldsymbol{W}\boldsymbol{D}^{-1}).$$

Esta distribución límite nos dice que $\hat{\theta} \sim^a N(\theta^*, \boldsymbol{D}^{-1}\boldsymbol{W}\boldsymbol{D}^{-1}/N)$.

Ahora podemos observar las propiedades del estimador. La distribución asintótica de $\hat{\theta}$ se centra en el valor verdadero, y su varianza disminuye a medida que el tamaño de la muestra crece. Como resultado, $\hat{\theta}$ converge en probabilidad a θ^* a medida que el tamaño de la muestra se eleva sin límite: $\hat{\theta} \xrightarrow{p} \theta$. Por consiguiente, el estimador es consistente. El estimador es asintóticamente normal. Y su varianza es $\boldsymbol{D}^{-1}\boldsymbol{W}\boldsymbol{D}^{-1}/N$, que puede ser comparada con la varianza más baja posible, $-\boldsymbol{H}^{-1}/N$, para determinar si es eficiente.

Para ML, $g_n(\cdot)$ es la puntuación, de manera que la varianza de $g_n(\theta^*)$ es la varianza de las puntuaciones: $\boldsymbol{W} = \boldsymbol{V}$. Además, la derivada media de $g_n(\theta^*)$ es la derivada media de las puntuaciones: $\boldsymbol{D} = \boldsymbol{H} = E(\partial^2 \ln P_n(\theta^*)\,/\partial\theta\partial\theta')$, donde la esperanza se calcula en la población. Por la identidad de información, $\boldsymbol{V} = -\boldsymbol{H}$. La varianza asintótica de $\hat{\theta}$ se convierte en $\boldsymbol{D}^{-1}\boldsymbol{W}\boldsymbol{D}^{-1}/N = \boldsymbol{H}^{-1}\boldsymbol{V}\boldsymbol{H}^{-1}/N = \boldsymbol{H}^{-1}(-\boldsymbol{H})\boldsymbol{H}^{-1}/N = -\boldsymbol{H}^{-1}/N$, que es la varianza más baja posible de cualquier estimador. Por lo tanto, ML es eficiente. Puesto que $\boldsymbol{V} = -\boldsymbol{H}$, la varianza del estimador ML también puede ser expresada como $\boldsymbol{V}^{-1}/N$, que tiene un significado fácilmente interpretable: la varianza del estimador es igual a la inversa de la varianza de las puntuaciones evaluadas en los verdaderos parámetros, dividida por el tamaño de la muestra.

Para MOM, $g_n(\cdot)$ es un conjunto de momentos. Si se utilizan los instrumentos ideales, entonces MOM se convierte en ML y es eficiente. Si se utilizan otros instrumentos, entonces MOM no es ML . En este caso, $\boldsymbol{W}$ es la varianza en la población de los momentos y $\boldsymbol{D}$ es la derivada media de los momentos, en lugar de la varianza y derivada media de las puntuaciones. La varianza asintótica de $\hat{\theta}$ no es igual $-\boldsymbol{H}^{-1}/N$. Por lo tanto, MOM sin pesos ideales no es eficiente.

10.5 Propiedades de los estimadores basados en simulación

Supongamos que los términos que entran en la ecuación definitoria de un estimador se obtienen por simulación en lugar de calcularse con exactitud. Sea $\check{g}_n(\theta)$ el valor simulado de $g_n(\theta)$, y $\check{g}(\theta)$ la media de estos valores simulados en la muestra, de manera que $\check{g}(\theta)$ es la versión simulada de $g(\theta)$. Llamaremos R al número de valores extraídos al azar que usamos en la simulación para cada n, y asumiremos que para cada n usamos valores extraídos de forma independiente (por ejemplo, usando extracciones separadas para cada n). Supondremos, además, que los mismos valores extraídos al azar se utilizan para cada valor de θ en el cálculo de $\check{g}_n(\theta)$. Este procedimiento evita *vibraciones* (*chatter*) en la

simulación: la diferencia entre $\breve{g}(\theta_1)$ y $\breve{g}(\theta_2)$ para dos valores diferentes de θ no se debe al uso de diferentes valores extraídos al azar.

Estos supuestos sobre los valores extraídos al azar empleados en la simulación son fáciles de implementar para el investigador y simplifican nuestro análisis considerablemente. Para los lectores interesados , Lee (1992) examina el caso en que se usan los mismos valores extraídos al azar para todas las observaciones. Pakes y Pollard (1989) proporcionan una manera de caracterizar una condición de equicontinuidad que, cuando se satisface, facilita el análisis de los estimadores basados en simulación. McFadden (1989) caracteriza esta condición de un modo diferente y muestra que se puede cumplir mediante el uso de los mismos valores extraídos al azar para cada valor de θ, que es la hipótesis que nosotros asumimos. McFadden (1996) ofrece una útil síntesis que incluye un análisis de la necesidad de prevenir la vibración (*chatter*)

El estimador se define por la condición $\breve{g}(\hat{\theta}) = 0$. Derivamos las propiedades de $\hat{\theta}$ mediante los dos mismos pasos que hemos empleado para los estimadores tradicionales.

Paso 1 : Distribución de $\breve{g}(\theta^*)$

Para identificar los distintos componentes de esta distribución, vamos a re-expresar $\breve{g}(\theta^*)$ sumando y restando algunos términos, así como reordenando:

$$\breve{g}(\theta^*) = \breve{g}(\theta^*) + g(\theta^*) - g(\theta^*) + E_r\breve{g}(\theta^*) - E_r\breve{g}(\theta^*)$$

$$= g(\theta^*) + [E_r\breve{g}(\theta^*) - g(\theta^*)] + [\breve{g}(\theta^*) - E_r\breve{g}(\theta^*)],$$

donde $\mathrm{g}(\theta^*)$ es el valor no simulado y $E_r\breve{g}(\theta^*)$ es la esperanza del valor simulado entre los valores al azar utilizados en la simulación. Sumar y restar términos obviamente no cambia $\breve{g}(\theta^*)$. Sin embargo, la posterior reordenación de los términos nos permite identificar los componentes que tienen un significado intuitivo.

El primer término $\mathrm{g}(\theta^*)$ es el mismo que aparece para el estimador tradicional. Los otros dos términos son elementos adicionales que surgen debido a la simulación. El término $E_r\breve{g}(\theta^*) - \mathrm{g}(\theta^*)$ capta el sesgo, si existe, en el simulador de $\mathrm{g}(\theta^*)$. Es la diferencia entre el valor real de $\mathrm{g}(\theta^*)$ y la esperanza del valor simulado. Si el simulador de $\mathrm{g}(\theta^*)$ es no sesgado, entonces $E_r\breve{g}(\theta^*) = \mathrm{g}(\theta^*)$ y este término desaparece. A menudo, sin embargo, el simulador de $\mathrm{g}(\theta^*)$ es sesgado. Por ejemplo, con MSL, $\breve{g}_n(\theta) = \partial \ln \breve{P}_n(\theta)/\partial\theta$, donde $\breve{P}_n(\theta)$ es un simulador no sesgado de $P_n(\theta)$. Dado que $\breve{P}_n(\theta)$ entra de forma no lineal a través del operador logarítmico, $\breve{g}_n(\theta)$ es sesgado. El tercer término, $\breve{g}(\theta^*) - E_r\breve{g}(\theta^*)$, capta el ruido de simulación, es decir, la desviación del simulador para cada valor al azar empleado, respecto a su esperanza calculada sobre todos los posibles valores al azar.

Combinando todo estos conceptos, tenemos

$$\breve{g}(\theta) = A + B + C, \qquad (10.6)$$

donde

A es el mismo que en el estimador tradicional,

B es el sesgo de la simulación,

C es el ruido de simulación.

Para ver cómo los estimadores basados en simulación difieren de sus equivalentes tradicionales, examinaremos el sesgo de simulación B y el ruido C.

Consideremos primero el ruido. Este término puede ser re-expresado como

$$C = \breve{g}(\theta^*) - E_r\breve{g}(\theta^*)$$

$$= \frac{1}{N}\sum_n [\breve{g}_n(\theta^*) - E_r\breve{g}_n(\theta^*)]$$

$$= \sum_n d_n/N,$$

donde d_n es la desviación del valor simulado para la observación n respecto su esperanza. La clave para entender el comportamiento del ruido de simulación está en observar que d_n es simplemente un estadístico para la observación n. La muestra está constituida por N extracciones al azar de este estadístico, uno para cada observación: $d_n, n = 1, \dots, N$. El ruido de simulación C es el promedio de estas N extracciones al azar. Por lo tanto, el teorema del límite central nos da la distribución de C.

En particular, para una observación dada, los valores extraídos al azar que se utilizan en la simulación proporcionan un valor particular de d_n. Si se hubieran extraído valores diferentes, entonces se habría obtenido un valor diferente de d_n. Hay una distribución de los valores de d_n sobre las posibles realizaciones de los valores al azar utilizados en simulación. La distribución tiene media cero, ya que la esperanza de los valores extraídos al azar se resta en el momento de crear d_n. Etiquetemos la varianza de la distribución como S_n/R, donde S_n es la varianza cuando se utiliza un valor extraído al azar en la simulación. Hay dos cosas a tener en cuenta acerca de esta varianza. En primer lugar, S_n/R es inversamente proporcional a R, el número de valores al azar que se utilizan en la simulación. En segundo lugar, la variación es diferente para diferentes n. Dado que$g_n(\theta^*)$ es diferente para diferentes n, la varianza de la desviación de simulación también difiere.

Extraemos un valor al azar de d_n para cada una de las N observaciones; el ruido de simulación global, C, es el promedio de estos N valores de ruido de simulación específico de cada observación. Como acabamos de establecer, cada d_n es un valor extraído de una distribución con media cero y varianza S_n/R. La versión generalizada del teorema del límite central nos permite calcular la distribución de un promedio en la muestra de valores extraídos al azar de distribuciones que tienen la misma media pero diferentes varianzas. En nuestro caso,

$$\sqrt{N}C \xrightarrow{d} N(0, \boldsymbol{S}/R),$$

donde **S** es la media de S_n en la población. Por lo tanto $C \sim^a N(0, \boldsymbol{S}/(\mathrm{N}R))$.

La característica más relevante de la varianza asintótica de C es que disminuye a medida que N se incrementa, incluso cuando R es fija. El ruido de simulación desaparece a medida que aumenta el tamaño de la muestra, incluso sin aumentar el número de valores al azar utilizados en la simulación. Este es un hecho muy importante y de gran alcance. Significa que el aumento del tamaño de la muestra es una forma de disminuir los efectos de la simulación en el estimador. El resultado es intuitivamente lógico. Básicamente, el ruido de simulación se cancela entre observaciones. La simulación de una observación podría, por casualidad, hacer la $\breve{g}_n(\theta)$ de esa observación demasiado grande. Sin embargo, la simulación para otra observación es probable que, por casualidad, sea demasiado pequeña. Promediando las simulaciones entre observaciones, los errores tienden a anularse entre sí. A medida que el tamaño de la muestra aumenta, esta propiedad de cancelación se vuelve más relevante hasta que, con muestras lo suficientemente grandes, el ruido de simulación es insignificante.

Consideremos ahora el sesgo. Si $\check{g}(\theta)$ es un simulador no sesgado de $g(\theta)$, entonces el término de sesgo B expresado en (10.6) es cero. Si por el contrario el simulador es sesgado, como sucede con MSL, entonces el efecto de este sesgo en la distribución de $\check{g}(\theta^*)$ debe ser considerado.

Por lo general, el término definitorio $g_n(\theta)$ es una función de un estadístico, l_n, que puede ser simulado sin sesgo. Por ejemplo, en MSL, $g_n(\theta)$ es una función de la probabilidad de elección, que puede ser simulada sin sesgo; en este caso l_n es la probabilidad. Más generalmente, l_n puede ser cualquier estadístico que se simule sin sesgo y que sirve para definir $g_n(\theta)$. Podemos escribir la dependencia en general como $g_n(\theta) = g(l_n(\theta))$ y el simulador no sesgado de $l_n(\theta)$ como $\check{l}_n(\theta)$ donde $E_r\check{l}_n(\theta) = l_n(\theta)$.

Ahora podemos re-expresar $\check{g}_n(\theta)$ mediante una expansión de Taylor alrededor del valor no simulado $g_n(\theta)$:

$$\check{g}_n(\theta) = g_n(\theta) + \frac{\partial g(l_n(\theta))}{\partial l_n}\left[\check{l}_n(\theta) - l_n(\theta)\right] + \frac{1}{2}\frac{\partial^2 g(l_n(\theta))}{\partial l_n^2}\left[\check{l}_n(\theta) - l_n(\theta)\right]^2,$$

$$\check{g}_n(\theta) - g_n(\theta) = g_n'\left[\check{l}_n(\theta) - l_n(\theta)\right] + \frac{1}{2}g_n''\left[\check{l}_n(\theta) - l_n(\theta)\right]^2,$$

donde g_n' y g_n'' son simplemente formas abreviadas de referirse a la primera y la segunda derivada de $g_n(l(\cdot))$ respecto a l. Dado que $\check{l}_n(\theta)$ no está sesgado respecto a $l_n(\theta)$, sabemos que $E_r g_n'\left[\check{l}_n(\theta) - l_n(\theta)\right] = g_n'\left[E_r\check{l}_n(\theta) - l_n(\theta)\right] = 0$. Como resultado de ello, sólo el término de la varianza permanece en la esperanza:

$$E_r\check{g}_n(\theta) - g_n(\theta) = \frac{1}{2}g_n''E_r\left[\check{l}_n(\theta) - l_n(\theta)\right]^2$$

$$= \frac{1}{2}g_n''Var_r\check{l}_n(\theta).$$

Indiquemos $Var_r\check{l}_n(\theta) = Q_n/R$ para reflejar el hecho de que la varianza es inversamente proporcional al número de valores al azar utilizados en la simulación. El sesgo de simulación es entonces

$$E_r\check{g}(\theta) - g(\theta) = \frac{1}{N}\sum_n E_r\check{g}_n(\theta) - g_n(\theta)$$

$$= \frac{1}{N}\sum_n g_n''\frac{Q_n}{2R}$$

$$= \frac{Z}{R},$$

donde Z es el promedio en la muestra de $g_n''Q_n/2$.

Puesto que $B = Z/R$, el valor de este estadístico normalizado para el tamaño de la muestra es

(10.7) $$\sqrt{N}B = \frac{\sqrt{N}}{R}Z.$$

Si R es fijo, entonces B es distinto de cero. Incluso peor, $\sqrt{N}\mathrm{B}$ se eleva con N, de tal manera que no tiene ningún valor límite. Supongamos que consideramos que R se incrementa con N. En este caso, el término de sesgo desaparecería asintóticamente: $B = Z/R \xrightarrow{p} 0$. Sin embargo, el término de sesgo normalizado no necesariamente desaparece. Como $\sqrt{N}$ entra en el numerador de este término, $\sqrt{N}\mathrm{B} = \left(\sqrt{N}/R\right)Z \xrightarrow{p} 0$ sólo si R aumenta más rápidamente que $\sqrt{N}$, de manera que la relación $\sqrt{N}/R$ se aproxime a cero cuando N aumenta. Si R se incrementa más lento que $\sqrt{N}$, el ratio $\sqrt{N}/R$ se eleva, de tal manera que el término de sesgo normalizado no desaparece sino que, de hecho, se hace más y más grande a medida que aumenta el tamaño de la muestra.

Podemos ahora recopilar nuestros resultados para la distribución del término definitorio normalizado por el tamaño de muestra:

$$\sqrt{N}\breve{g}(\theta^*) = \sqrt{N}(A + B + C), \tag{10.8}$$

donde

$\sqrt{N}A \xrightarrow{d} N(0, \boldsymbol{W})$, igual al estimador tradicional,

$\sqrt{N}B = \frac{\sqrt{N}}{R}Z$, captura el sesgo de simulación,

$\sqrt{N}C \xrightarrow{d} N(0, \boldsymbol{S}/R)$, captura el ruido de simulación,

Paso 2: Obtenga la distribución de $\hat{\theta}$ a partir de la distribución de $\breve{g}(\theta^*)$

Al igual que con los estimadores tradicionales, la distribución de $\hat{\theta}$ está directamente relacionada con la distribución de $\breve{g}(\theta^*)$. Usando la misma expansión de Taylor usada en (10.3), tenemos

$$\sqrt{N}\left(\hat{\theta} - \theta^*\right) = -\breve{D}^{-1}\sqrt{N}\breve{g}(\theta^*) = -\breve{D}^{-1}\sqrt{N}(A + B + C), \tag{10.9}$$

donde $\breve{D}$ es la derivada de $\breve{g}(\theta^*)$ respecto a los parámetros, que converge a su esperanza $\mathbf{D}$ a medida que el tamaño de la muestra crece. El estimador mismo se expresa como

$$\hat{\theta} = \theta^* - \breve{D}^{-1}(A + B + C), \tag{10.10}$$

Ahora podemos examinar las propiedades de nuestros estimadores.

10.5.1 Máxima verosimilitud simulada (maximum simulated likelihood, MSL)

Para MSL, $\breve{g}_n(\theta)$ está sesgado respecto $g_n(\theta)$. El término de sesgo en (10.9) es $\sqrt{N}\mathrm{B} = \left(\sqrt{N}/R\right)Z$. Supongamos que R aumenta con N. Si R aumenta más rápido que N, entonces

$$\sqrt{N}B = \left(\sqrt{N}/R\right)Z \xrightarrow{p} 0,$$

ya que el ratio $\sqrt{N}/R$ cae a cero. Consideremos ahora el tercer término en (10.9), que capta el ruido de simulación: $\sqrt{N}C \xrightarrow{d} N(0, \boldsymbol{S}/R)$. Dado que $\boldsymbol{S}/R$ disminuye a medida que R aumenta, tenemos que $\boldsymbol{S}/R \xrightarrow{d} 0$ a medida que $\mathrm{N} \to \infty$ cuando R aumenta con N. El segundo y tercer términos desaparecen, quedando sólo el primer término. Este primer término es el mismo que aparece en el estimador no simulado. Tenemos

$$\sqrt{N}\left(\hat{\theta} - \theta^*\right) = -\boldsymbol{D}^{-1}\sqrt{N}A \xrightarrow{d} N(0, \boldsymbol{D^{-1}WD^{-1}})$$

$$= N(0, \boldsymbol{H^{-1}VH^{-1}})$$

$$= N(0, -\boldsymbol{H^{-1}}),$$

donde la penúltima igualdad se debe a que $g_n(\theta)$ es la puntuación, y la última igualdad se debe a la identidad de información. El estimador se distribuye

$$\hat{\theta} \sim^a N(\theta^*, -\boldsymbol{H^{-1}}/N)$$

Esta es la misma distribución asintótica de ML. Cuando R aumenta más rápidamente que $\sqrt{N}$, MSL es consistente, asintóticamente normal y eficiente, y asintóticamente equivalente a ML.

Supongamos que R crece con N, pero en una proporción menor a $\sqrt{N}$. En este caso, el ratio $\sqrt{N}/R$ se hace más grande a medida que N aumenta. No hay distribución límite para $\sqrt{N}(\hat{\theta} - \theta^*)$, porque el término de sesgo, $(\sqrt{N}/R\,)$Z, crece con N. Sin embargo, el propio estimador converge en el valor verdadero. $\hat{\theta}$ depende de $(1/\mathrm{R})$Z, sin multiplicar por $\sqrt{N}$. Este término de sesgo desaparece cuando R crece a cualquier velocidad (en cualquier proporción respecto N). Por lo tanto, el estimador converge en el valor verdadero, al igual que su equivalente no simulado, lo que significa que $\hat{\theta}$ es consistente. Sin embargo, el estimador no es asintóticamente normal, ya que $\sqrt{N}(\hat{\theta} - \theta^*)$ no tiene distribución límite. Los errores estándar no se pueden calcular, y los intervalos de confianza no se pueden construir.

Cuando R es fijo, el sesgo aumenta a medida que crece N. $\sqrt{N}(\hat{\theta} - \theta^*)$ no tiene distribución límite. Además, el propio estimador, $\hat{\theta}$, contiene un sesgo $\mathrm{B} = (1/R)$Z que no desaparece a medida que el tamaño de la muestra aumenta con un R fijo. El estimador MSL no es ni consistente ni asintóticamente normal cuando R es fijo.

Las propiedades de MSL se pueden resumir de la siguiente manera:

1. Si R es fijo, MSL es inconsistente.
2. Si R se eleva más lentamente que $\sqrt{N}$, MSL es consistente pero no asintóticamente normal.
3. Si R se eleva más rápido que $\sqrt{N}$, MSL es consistente, asintóticamente normal y eficiente, y equivalente a ML.

10.5.2 Método de momentos simulados (method of simulated moments, MSM)

Para MSM con instrumentos fijos, $\check{g}_n(\theta) = \sum_j [d_{nj} - \check{P}_{nj}(\theta)]z_{nj}$, que es no sesgado respecto $g_n(\theta)$, ya que la probabilidad simulada entra linealmente en la expresión. El término de sesgo es cero. La distribución del estimador está determinada sólo por el término A, que es el mismo que en el MOM tradicional sin simulación, y por el término C, el cual refleja el ruido de simulación:

$$\sqrt{N}(\hat{\theta} - \theta^*) = -\check{D}^{-1}\sqrt{N}(A + C).$$

Supongamos que R es fijo. Dado que $\check{D}$ converge a su valor esperado $\boldsymbol{D}$, tenemos $-\sqrt{N}\check{D}^{-1}\mathrm{A} \xrightarrow{d} N(0, \boldsymbol{D^{-1}WD^{-1}})$ y $-\sqrt{N}\check{D}^{-1}\mathrm{C} \xrightarrow{d} N(0, \boldsymbol{D^{-1}}(\boldsymbol{S}/R)\boldsymbol{D^{-1}})$, de manera que

$$\sqrt{N}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, \boldsymbol{D^{-1}}[\boldsymbol{W} + \boldsymbol{S}/R]\boldsymbol{D^{-1}}).$$

La distribución asintótica del estimador es entonces

$$\hat{\theta} \sim^{a} N(\theta^{*}, \boldsymbol{D^{-1}}[\boldsymbol{W} + \boldsymbol{S}/R]\boldsymbol{D^{-1}}/N).$$

El estimador es consistente y asintóticamente normal. Su varianza es mayor que la de su equivalente no simulado en una cantidad $\boldsymbol{D^{-1}SD^{-1}}/(\mathrm{R}N)$, que refleja el ruido de simulación.

Supongamos ahora que R se eleva con N en una proporción cualquiera. La varianza adicional debida al ruido de simulación desaparece, de modo que $\hat{\theta} \sim^{a} N(\theta^{*}, \boldsymbol{D^{-1}WD^{-1}}/N)$, igual a la de su equivalente no simulado. Cuando se utilizan instrumentos no ideales, $\boldsymbol{D^{-1}WD^{-1}} \neq -\boldsymbol{H}^{-1}$ y por lo tanto, el estimador (ya sea en su forma simulada o no simulada) es menos eficiente que ML.

Si se utilizan instrumentos simulados en MSM, entonces las propiedades del estimador dependen de cómo se simulan los instrumentos. Si los instrumentos se simulan sin sesgo y con independencia de la probabilidad que entra en el residuo, entonces este MSM tiene las mismas propiedades que el MSM con pesos fijos. Si se simulan los instrumentos con sesgo y los instrumentos no son ideales, entonces el estimador tiene las mismas propiedades que MSL, excepto que no es asintóticamente eficiente, ya que la identidad de información no aplica. MSM con instrumentos ideales simulados es MSS, que se trata a continuación.

10.5.3 Método de puntuaciones simuladas (method of simulated scores, MSS)

Con MSS utilizando simuladores de puntuación no sesgados, $\breve{g}_n(\theta)$ es no sesgado respecto $g_n(\theta)$, y, por otra parte, $g_n(\theta)$ es la puntuación, de forma que la identidad de información aplica. El análisis es el mismo de MSM excepto que la identidad de información hace que el estimador sea eficiente cuando R aumenta con N. Como en el caso MSM, tenemos

$$\hat{\theta} \sim^{a} N(\theta^{*}, \boldsymbol{D^{-1}}[\boldsymbol{W} + \boldsymbol{S}/R]\boldsymbol{D^{-1}}/N),$$

que, dado que $g_n(\theta)$ es la puntuación, se convierte en

$$\hat{\theta} \sim^{a} N\left(\theta^{*}, \frac{\boldsymbol{H^{-1}}[\boldsymbol{V} + \boldsymbol{S}/R]\boldsymbol{H^{-1}}}{N}\right) = N\left(\theta^{*}, -\frac{\boldsymbol{H^{-1}}}{N} + \frac{\boldsymbol{H^{-1}SH^{-1}}}{RN}\right).$$

Cuando R es fijo, el estimador es consistente y asintóticamente normal, pero su covarianza es más grande que en el caso de ML debido al ruido de la simulación. Si R crece con N en cualquier proporción, entonces tenemos

$$\hat{\theta} \sim^{a} N(\theta^{*}, -\boldsymbol{H^{-1}}/N).$$

MSS con simuladores de puntuación no sesgados es asintóticamente equivalente a ML cuando R crece con N en cualquier proporción.

Este análisis muestra que MSS con simuladores de puntuación no sesgados tiene mejores propiedades que MSL en dos aspectos. En primer lugar, para R fijo, MSS es consistente y asintóticamente normal, mientras que MSL no es ninguna de las dos cosas. En segundo lugar, para R creciendo con N, MSS es equivalente a ML sin importar lo rápido que R esté aumentando, mientras que MSL es equivalente a ML sólo si la proporción en que crece es mayor a $\sqrt{\mathrm{N}}$.

Tal y como vimos en la sección 10.2.3, es difícil encontrar simuladores de puntuación no sesgados con buenas propiedades numéricas. MSS se emplea en ocasiones con simuladores de puntuación sesgados. En este caso, las propiedades del estimador son las mismas que las de MSL: el sesgo en las puntuaciones simuladas se traduce en sesgo en el estimador, que desaparece de la distribución límite sólo si R crece más rápidamente que $\sqrt{\mathrm{N}}$.

10.6 Solución numérica

Los estimadores se definen como el valor de θ que resuelve $\breve{g}(\theta) = 0$, donde $\breve{g}(\theta) = \sum_n \breve{g}_n(\theta)\,/N$ es la media muestral de un estadístico simulado $\breve{g}_n(\theta)$. Dado que $\breve{g}_n(\theta)$ es un vector, tenemos que resolver el conjunto de ecuaciones para los parámetros. La cuestión es: ¿cómo se resuelven numéricamente estas ecuaciones para obtener las estimaciones?

El capítulo 8 describe los métodos numéricos que permiten maximizar una función. Estos procedimientos también se pueden utilizar para resolver un conjunto de ecuaciones. Sea T el negativo del producto interior (*inner product*) del término definitorio de un estimador: $T = -\breve{g}(\theta)'\breve{g}(\theta) = -(\sum_n \breve{g}_n(\theta))'(\sum_n \breve{g}_n(\theta))/N^2$. T es necesariamente menor o igual a cero, ya que es el negativo de una suma de cuadrados. T tiene como valor máximo 0, que se consigue sólo cuando los términos cuadráticos que la componen son todos iguales a 0. Es decir, el máximo de T se alcanza cuando $\breve{g}(\theta) = 0$. Maximizar T es equivalente a resolver la ecuación $\breve{g}(\theta) = 0$. Los enfoques para resolver el problema descritos en el capítulo 8, con la excepción de BHHH, se pueden utilizar para esta maximización. BHHH no se puede utilizar debido a que el método asume que la función que está siendo maximizada es una suma de términos específicos de cada observación, mientras que T contiene el cuadrado de cada suma de términos específicos de cada observación. Los otros métodos, especialmente BFGS y DFP, han demostrado ser muy eficaces en la localización de los parámetros en los que $\breve{g}(\theta) = 0$.

Con MSL, por lo general es más fácil maximizar la función de verosimilitud simulada que maximizar T. BHHH se puede utilizar en este caso, así como el resto de métodos.

11

Parámetros a nivel individual

11.1 Introducción

Los modelos logit mixto y probit permiten coeficientes aleatorios cuya distribución en la población se estima. Consideremos, por ejemplo, el modelo descrito el capítulo 6, relativo a la elección hecha por los pescadores entre distintos sitios de pesca disponibles. Los sitios se diferencian por el hecho de tener o no tener campings. A algunos pescadores les gusta tener campings en los sitios de pesca, ya que pueden utilizarlos para acampar y pasar la noche. A otros pescadores no les gustan las multitudes y el ruido que se asocian a los lugares para acampar, y prefieren la pesca en lugares más aislados. Para capturar estas diferencias en las preferencias, se especificó un modelo logit mixto que incluía coeficientes aleatorios tanto para la variable relativa a la presencia de camping como para otros atributos del sitio de pesca. Una vez especificado el modelo, se estimó la distribución de los coeficientes en la población. La figura 11.1 muestra la distribución estimada para el coeficiente de camping. Se especificó una distribución normal para dicho coeficiente. La media estimada fue de 0.116 y la desviación estándar de 1.655. Esta distribución proporciona información útil sobre la población. Por ejemplo, las estimaciones resultantes implican que al 47% de la población no le gusta tener lugares para acampar en sus sitios de pesca, mientras que al otro 53% por ciento sí le gusta.

La cuestión que se plantea es: ¿en qué parte de la distribución de preferencias está un pescador en particular? ¿Hay una manera de determinar si una persona dada tiende a preferir (o no), tener lugares para acampar en los sitios de pesca?

Las elecciones de una persona revelan algo acerca de sus gustos y preferencias, algo que el investigador puede, en principio, descubrir. Si el investigador observa que un pescador concreto escoge consistentemente sitios sin campamentos, aun cuando el costo de conducir hasta estos sitios es mayor, el investigador puede inferir razonablemente que a este pescador no le gusta acampar. Existe una manera precisa de llevar a cabo este tipo de inferencia, dada por Revelt y Train (2000).

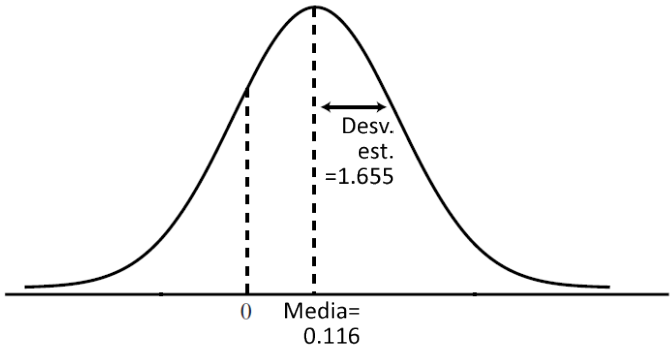


Figura 11.1. Distribución en el total de la población de pescadores del coeficiente relativo a la presencia de campings.

Explicamos el procedimiento en el contexto de un modelo logit mixto; sin embargo, el procedimiento puede usarse con cualquier modelo de comportamiento que incorpore coeficientes aleatorios, incluyendo probit. El concepto central es una distinción entre dos distribuciones: la distribución de las preferencias en la población y la distribución de las preferencias en la subpoblación de personas que toman unas decisiones concretas. Nos referiremos a los coeficientes aleatorios como al vector β. La distribución de β en el total de personas de la población se denota como $g(\beta|\theta)$, donde θ son los parámetros de esta distribución, tales como la media y la varianza.

Una situación de elección consiste en varias alternativas descritas colectivamente por las variables x. Consideremos el siguiente experimento mental. Supongamos que todos los miembros de la población se enfrentan a la misma situación de elección descrita por las mismas variables x. Una parte de la población elegirá cada alternativa. Considere las personas que optan por la alternativa i. Las preferencias de estas personas no son todas iguales: existe una distribución de los coeficientes entre estas personas. Sea $h(\beta|i,x,\theta)$ la distribución de β en la subpoblación de personas que, ante la situación de elección descrita por las variables x, elegiría la alternativa i. De esta forma, $g(\beta|\theta)$ es la distribución de β en toda la población y $h(\beta|i,x,\theta)$ la distribución de β en la subpoblación de personas que elegirían la alternativa i al enfrentarse a una situación de elección descrita por x.

Podemos generalizar la notación para permitir elecciones repetidas. Sea y la secuencia de elecciones en una serie de situaciones de elección que se describen colectivamente por las variables x. La distribución de los coeficientes en la subpoblación de personas que harían la secuencia de elecciones y al enfrentarse a las situaciones de elección descritas por x se denota como $h(\beta|y,x,\theta)$.

Observe que $h(\cdot)$ está condicionada a y, mientras que $g(\cdot)$ no lo está. A veces es útil referirse a h como la distribución condicionada y a g como la distribución no condicionada. Dos de estas distribuciones se muestran en la figura 11.2. Si no supiéramos nada sobre las elecciones pasadas de una persona, entonces lo mejor que podemos hacer para describir sus preferencias es decir que sus coeficientes se encuentran en algún lugar de $g(\beta|\theta)$. Sin embargo, si se ha observado que la persona hizo las elecciones y cuando se enfrentó a las situaciones de elección descritas por x, entonces sabemos que los coeficientes de esa persona están en la distribución $h(\beta|y,x,\theta)$. Dado que h es más estrecha que g, tenemos una mejor información sobre las preferencias de la persona al condicionar sobre sus elecciones pasadas.

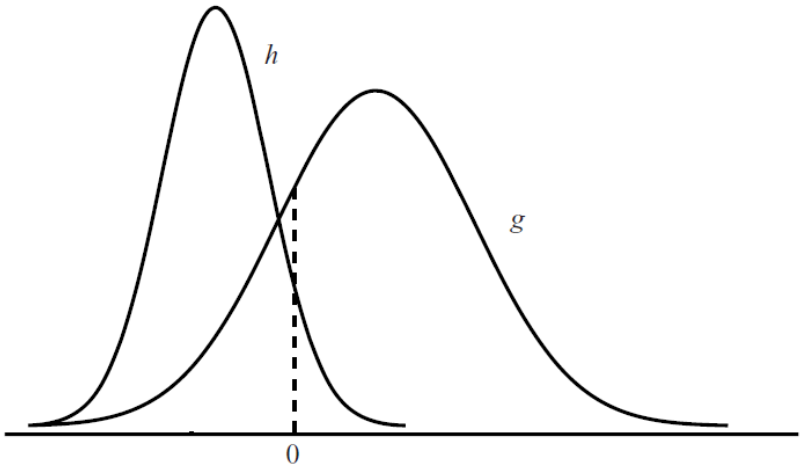


Figura 11.2. Distribución g (no condicionada) de la población y distribución h (condicionada) de la subpoblación de pescadores que eligen sitios de pesca sin campings.

Inferencias de este tipo hace mucho tiempo que se llevan a cabo en modelos de regresión lineal, donde tanto la variable dependiente y como la distribución de los coeficientes son continuas (Griffiths, 1972; Juez et al, 1988). Modelos de cambio de régimen (*regime-switching models*, series temporales que ocasionalmente muestran un cambio brusco en su comportamiento), sobre todo en macroeconomía, han utilizado un procedimiento análogo para evaluar la probabilidad de que una observación se encuentre dentro de un régimen determinado (Hamilton y Susmel, 1994 ; Hamilton, 1996). En estos modelos, la variable dependiente es continua y la distribución de los coeficientes es discreta (representando un conjunto de coeficientes para cada régimen). A diferencia de estos dos casos, nuestros modelos tienen variables dependientes discretas. Kamakura y Russell (1989) y DeSarbo et al. (1995) desarrollaron un enfoque del problema en el contexto de un modelo de elección discreta con una distribución discreta de los coeficientes (es decir, un modelo de clase latente). Utilizaron para ello procedimientos de máxima verosimilitud con el fin de estimar los coeficientes de cada segmento, y luego calcularon la probabilidad de que una observación estuviese dentro de cada segmento, basándose en las elecciones registradas para esa observación. El enfoque que se describe en este capítulo aplica a modelos de elección discreta con coeficientes distribuidos de forma continua o discreta, y utiliza máxima verosimilitud (u otros métodos clásicos) para la estimación. Los modelos de Kamakura y Russell (1989) y DeSarbo et al. (1995) son un caso especial de este método más general. También se han desarrollado procedimientos bayesianos para llevar a cabo esta inferencia en modelos de elección discreta (Rossi et al, 1996;. Allenby y Rossi, 1999). Los métodos bayesianos se describen en el capítulo 12.

11.2 Derivación de la distribución condicionada

La relación entre h y g se puede establecer con precisión. Considere una elección entre alternativas $j = 1, \dots, J$ en diferentes situaciones de elección $t = 1, \dots, T$. La utilidad que la persona n obtiene al elegir la alternativa j en una situación t es

$$U_{njt} = \beta_n' x_{njt} + \varepsilon_{njt},$$

donde $\varepsilon_{njt} \sim iid$ valor extremo y $\beta_n \sim g(\beta|\theta)$ en la población. Podemos referirnos a las variables x_{njt} de forma colectiva, para todas las alternativas y todas las situaciones de elección, como x_n. Sea $y_n = \langle y_{n1}, \dots, y_{nT} \rangle$ la secuencia de alternativas escogidas por la persona. Si conociésemos β_n, entonces la probabilidad de la secuencia de elecciones de la persona sería un producto de logits:

$$P(y_n|x_n,\beta) = \prod_{t=1}^{T} L_{nt}(y_{nt}|\beta),$$

donde

$$L_{nt}(y_{nt}|\beta) = \frac{e^{\beta' x_{ny_{nt}t}}}{\sum_j e^{\beta' x_{njt}}}.$$

Dado que no conocemos β_n, la probabilidad de la secuencia de elecciones de la persona es la integral de $P(y_n|x_n,\beta)$ sobre la distribución de β:

(11.1) $$P(y_n|x_n,\theta) = \int P(y_n|x_n,\beta)g(\beta|\theta)d\beta.$$

Esta es la probabilidad de elección de un modelo logit mixto que ya comentamos en el capítulo 6.

Ahora podemos obtener $h(\beta|y_n,x_n,\theta)$. Por el teorema de Bayes,

$$h(\beta|y_n,x_n,\theta) \times P(y_n|x_n,\theta) = P(y_n|x_n,\beta) \times g(\beta|\theta).$$

Esta ecuación simplemente establece que la densidad conjunta de β e y_n se puede expresar como la probabilidad de y_n por la probabilidad de β condicionada a y_n (que es el lado izquierdo de la expresión), o en la otra dirección del condicionamiento, como la probabilidad de β por la probabilidad de y_n condicionada a β (que es el lado derecho). Reordenando,

(11.2) $$h(\beta|y_n,x_n,\theta) = \frac{P(y_n|x_n,\beta)\times g(\beta|\theta)}{P(y_n|x_n,\theta)}.$$

Sabemos todas las cantidades del lado derecho. A partir de estas cantidades, podemos calcular h.

La ecuación (11.2) también proporciona una forma de interpretar h intuitivamente. Observe que el denominador $P(y_n|x_n,\theta)$ es la integral del numerador, tal y como se indica en la definición (11.1). Como tal, el denominador es una constante que hace que la integral de h sea 1, como se requiere para cualquier densidad. Dado que el denominador es una constante, h es proporcional al numerador, $P(y_n|x_n,\beta) \times g(\beta|\theta)$. Esta relación hace que la interpretación de h sea relativamente fácil. Dicho en palabras, la densidad de β en la subpoblación de personas que elegirían la secuencia y_n al enfrentarse a las situaciones de elección x_n es proporcional al producto de la densidad de β en la totalidad de la población por la probabilidad de que y_n sea elegida si los coeficientes de la persona fuesen β.

Usando (11.2), se pueden obtener varios estadísticos condicionados a y_n. La β media en la subpoblación de personas que elegirían y_n cuando se enfrentan a x_n es

$$\bar{\beta}_n = \int \beta \cdot h(\beta|y_n,x_n,\theta)d\beta.$$

Esta media generalmente difiere de la β media en toda la población. Sustituyendo h en la fórmula,

$$\bar{\beta}_n = \frac{\int \beta \cdot P(y_n|x_n,\beta)g(\beta|\theta)d\beta}{P(y_n|x_n,\theta)}$$

(11.3)
$$= \frac{\int \beta \cdot P(y_n|x_n,\beta) g(\beta|\theta) d\beta}{\int P(y_n|x_n,\beta) g(\beta|\theta) d\beta}.$$

Las integrales en esta ecuación no tienen una forma cerrada; sin embargo, se pueden simular fácilmente. Para ello, extraiga valores al azar de β a partir de la densidad poblacional $g(\beta|\theta)$. Calcule el promedio ponderado de estos valores, siendo el peso del valor β^r proporcional a $P(y_n|x_n,\beta^r)$. La media en la subpoblación simulada es

$$\check{\beta}_n = \sum_r \omega^r \beta^r,$$

donde los pesos son

(11.4)
$$\omega^r = \frac{P(y_n|x_n,\beta^r)}{\sum_r P(y_n|x_n,\beta^r)}.$$

Es posible calcular otros estadísticos. Supongamos que la persona se enfrenta a una nueva situación de elección descrita por variables $x_{njT+1} \forall j$. Si no tuviéramos información sobre las decisiones pasadas de la persona, podríamos asignar la siguiente probabilidad de que eligiera la alternativa i:

(11.5)
$$P(i|x_{nT+1},\theta) = \int L_{nT+1}(i|\beta) g(\beta|\theta) d\,\beta,$$

donde

$$L_{nT+1}(i|\beta) = \frac{e^{\beta' x_{niT+1}}}{\sum_j e^{\beta' x_{njT+1}}}.$$

Esto simplemente es la probabilidad de un modelo logit mixto utilizando la distribución de β en la población. Sin embargo, si hemos observado las últimas elecciones de la persona, entonces la probabilidad puede condicionarse a estas elecciones. La probabilidad se convierte en

(11.6)
$$P(i|x_{nT+1},y_n,x_n,\theta) = \int L_{nT+1}(i|\beta) h(\beta|y_n,x_n,\theta) d\,\beta.$$

Esta es también es la probabilidad del modelo logit mixto, pero utilizando la distribución condicionada h en lugar de la distribución no condicionada g. Cuando no conocemos las elecciones previas de la persona, combinamos la fórmula logit sobre la densidad de β en toda la población. Por el contrario, cuando sí conocemos las elecciones anteriores de la persona, podemos mejorar nuestra predicción mediante la combinación sobre la densidad de β en la subpoblación que habría hecho las mismas elecciones que esta persona.

Para calcular esta probabilidad, sustituimos h en la fórmula a partir de (11.2):

$$P(i|x_{nT+1},y_n,x_n,\theta) = \frac{\int L_{nT+1}(i|\beta) P(y_n|x_n,\beta) g(\beta|\theta) d\beta}{\int P(y_n|x_n,\beta) g(\beta|\theta) d\beta}.$$

La probabilidad se simula extrayendo valores al azar β de la distribución g en la población, calculando la fórmula logit para cada valor, y tomando un promedio ponderado de los resultados:

$$\check{P}_{niT+1}(y_n, x_n, \theta) = \sum_r \omega^r L_{nT+1}(i|\beta^r),$$

donde los pesos están dados por (11.4).

11.3 Implicaciones de la estimación de θ

Los parámetros poblacionales θ se estiman usando cualquiera de los métodos que se describen en el capítulo 10. El enfoque más común es el de máxima verosimilitud simulada, usando el valor simulado de $P(y_n|x_n, \theta)$ en la función log-verosimilitud. Obtenemos así una estimación de θ, etiquetada $\hat{\theta}$. Sabemos que el estimador tiene varianza muestral. La covarianza asintótica del estimador también se estima, la cual denominamos $\widehat{W}$. Por consiguiente, la distribución asintótica se estima como $N(\hat{\theta}, \widehat{W})$.

El parámetro θ describe la distribución de β en la población, dando, por ejemplo, la media y la varianza de β sobre todos los decisores. Para cualquier valor de θ, la ecuación (11.2) nos da la distribución condicionada de β en la subpoblación de personas que harían las elecciones y_n al enfrentarse a situaciones de elección descritas por x_n. Esta relación es exacta en el sentido de que no hay ninguna varianza, muestral o de cualquier otro tipo, asociada a ella. De forma similar, cualquier estadístico basado en h es exacto dado un valor de θ. Por ejemplo, la media de la distribución condicionada, $\bar{\beta}_n$, es exactamente la ecuación (11.3) para un valor dado de θ.

Teniendo en cuenta esta correspondencia entre θ y h, el hecho de que θ sea un valor estimado se puede manejar de dos maneras diferentes. El primer enfoque consiste en usar la estimación puntual (*point estimate*) de θ para calcular los estadísticos asociados con la distribución condicionada h. Bajo este enfoque, la media de la distribución condicionada, $\bar{\beta}_n$, se calcula insertando $\hat{\theta}$ en (11.3). La probabilidad de una nueva situación de elección se calcula insertando $\hat{\theta}$ en (11.6). Si el estimador de θ es consistente, entonces este enfoque es consistente para estadísticos basados en θ.

El segundo enfoque consiste en considerar la distribución muestral de $\hat{\theta}$. Cada posible valor de θ implica un valor de h, y por lo tanto un valor de cualquier estadístico asociado con h, como $\bar{\beta}_n$. La varianza muestral en el estimador de θ induce varianza muestral en los estadísticos que son calculados sobre la base de θ. Esta varianza muestral se puede calcular mediante simulación, a través de la extracción de valores al azar de θ a partir de su distribución muestral estimada y posterior cálculo del estadístico correspondiente para cada uno de estos valores.

Por ejemplo, para representar la distribución muestral de $\hat{\theta}$ en el cálculo de $\bar{\beta}_n$, se siguen los siguientes pasos:

1. Extraiga un valor al azar de $N(\hat{\theta}, \widehat{\mathrm{W}})$, que es la distribución muestral estimada de $\hat{\theta}$. Este paso se lleva a cabo de la siguiente manera. Extraiga K valores al azar de una densidad normal estándar, y etiquete el vector de estos valores como η^r, donde K es la longitud de θ. A continuación, cree $\theta^r = \hat{\theta} + L\eta^r$, donde L es el factor Choleski de $\widehat{\mathrm{W}}$.
2. Calcule $\bar{\beta}_n^r$ basándose en este θ^r. Dado que la fórmula para $\bar{\beta}_n$ implica integración, la simulamos usando la fórmula (11.3).
3. Repita los pasos 1 y 2 múltiples veces, etiquetando el número de veces como R.

Los valores resultantes son extracciones de la distribución muestral de $\bar{\beta}_n$ inducida por la distribución muestral de $\hat{\theta}$. La media de $\bar{\beta}_n^r$ sobre los R sorteos de θ^r es la media de la distribución muestral de $\bar{\beta}_n$. La desviación estándar de los valores extraídos al azar da el error estándar asintótico de $\bar{\beta}_n$ que es inducido por la varianza muestral de $\hat{\theta}$.

Tenga en cuenta que este proceso implica simulación dentro de una simulación. Para cada valor al azar de θ^r, el estadístico $\bar{\beta}_n^r$ es simulado con múltiples sorteos de β extraídos de la densidad poblacional $g(\beta|\theta^r)$.

Supongamos que usamos cualquiera de estos enfoques para estimar $\bar{\beta}_n$. A continuación surge la siguiente pregunta: ¿puede considerarse la estimación de $\bar{\beta}_n$ una estimación de β_n? Es decir: ¿es la media estimada de la distribución condicionada $h(\beta|y_n, x_n, \theta)$, que está condicionada a decisiones pasadas de la persona n, una estimación de los coeficientes de la persona n?

Hay dos respuestas posibles, dependiendo de cómo ve el investigador el proceso de generación de datos. Si el número de situaciones de elección que el investigador puede observar para cada decisor es fijo, entonces la estimación de $\bar{\beta}_n$ no es una estimación consistente de β_n. Cuando T es fijo, la consistencia exige que la estimación converja al verdadero valor cuando el tamaño de la muestra aumenta sin límite. Si el tamaño de la muestra aumenta, pero las situaciones de elección que enfrenta la persona n son fijas, entonces la distribución condicionada y su media no cambian. En la medida en que los coeficientes de la persona n no coinciden con la media de la distribución condicionada (una circunstancia esencialmente imposible), la media de la distribución condicionada nunca será igual a los coeficientes de la persona, sin importar cómo de grande sea la muestra. Elevar el tamaño de la muestra mejora la estimación de θ y por lo tanto proporciona una mejor estimación de la media de la distribución condicionada, ya que esta media sólo depende de θ. Sin embargo, aumentar el tamaño de la muestra no hace que la media condicionada sea igual a los coeficientes de la persona.

Cuando se fija el número de situaciones de elección, la media condicionada tiene la misma interpretación que la media de la población, pero para un grupo de gente diferente y menos diversa. Cuando predecimos el comportamiento futuro de la persona, es de esperar que obtengamos mejores predicciones usando la distribución condicionada, como en (11.6), que la distribución en la población. En el estudio de un caso que se presenta en la siguiente sección, se muestra que la mejora puede ser sustancial.

Si se puede considerar que el número de situaciones de elección a las que una persona se enfrenta crece, entonces la estimación de $\bar{\beta}_n$ puede ser considerada como una estimación de β_n. Sea T el número de situaciones de elección que la persona n afronta. Si observamos más elecciones de la persona (es decir, T se eleva), entonces estamos en mejores condiciones para identificar los coeficientes de la persona. La figura 11.3 muestra la distribución condicionada $h(\beta|y_n, x_n, \theta)$ para tres valores diferentes de T. La distribución condicionada tiende a moverse hacia la β_n de la propia persona a medida que aumenta T, y tiende a ser más concentrada. A medida que T aumenta sin límite, la distribución condicionada colapsa en β_n. La media de la distribución condicionada converge al verdadero valor de β_n a medida que el número de situaciones de elección aumenta sin límite. Por tanto, la estimación de $\bar{\beta}_n$ es consistente para β_n.

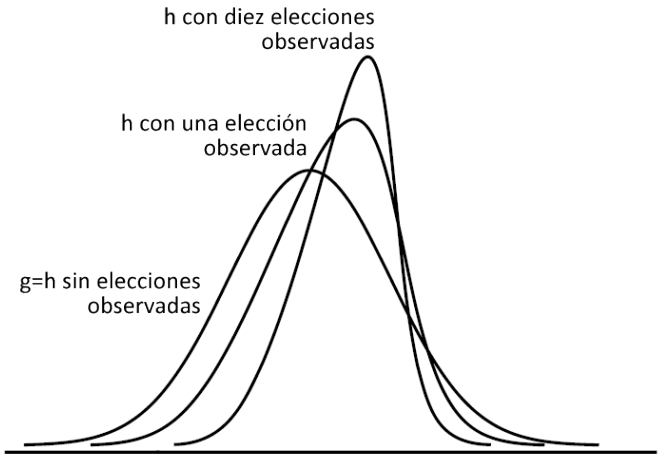


Figura 11.3. Distribución condicionada con T=0, 1 y 10.

En el Capítulo 12 se describe el teorema de Bernstein-von Mises. Este teorema indica que, en condiciones bastante laxas, la media de una distribución posterior de un parámetro es asintóticamente equivalente al máximo de la función de verosimilitud. La distribución condicionada h es una distribución posterior: por (11.2) h es proporcional a una densidad g, que puede ser interpretada como una distribución a priori de β_n, multiplicada por la verosimilitud de las T elecciones hechas por la persona n dadas las β_n, que es $P(y_n|x_n, \beta_n)$. Por el teorema de Bernstein-von Mises, la media de h es por lo tanto un estimador de β_n que es asintóticamente equivalente al estimador de máxima verosimilitud de β_n, donde el comportamiento asintótico se define como un aumento de T. Estos conceptos se describen más detalladamente en el capítulo 12; los mencionamos ahora simplemente para proporcionar otra interpretación de la media de la distribución condicionada.

11.4 Ilustración de Monte Carlo

Para ilustrar los conceptos expuestos, generé un conjunto de datos hipotéticos donde los verdaderos parámetros de la población θ son conocidos, así como el verdadero β_n de cada decisor. Estos datos nos permiten comparar la media de la distribución condicionada de las elecciones de cada decisor, $\bar{\beta}_n$, con la β_n de cada decisor. También nos permiten investigar el impacto de un aumento del número de situaciones de elección en la distribución condicionada. Para este experimento, construí conjuntos de datos que constaban de 300 "clientes", cada uno enfrentando $T = 1{,}10{,}20\ y\ 50$ situaciones de elección. Cada conjunto de datos consta de tres alternativas y cuatro variables. Los coeficientes para las dos primeras variables se mantienen fijos para toda la población, con valor 1.0, y los coeficientes de las dos últimas variables se distribuyen normalmente con media y varianza 1.0. La utilidad se especifica de forma que incluye estas variables más un término final iid que se distribuye valor extremo, de manera que el modelo es logit mixto. La variable dependiente para cada cliente se ha creado extrayendo un valor al azar de las densidades de los términos aleatorios, calculando la utilidad de cada alternativa con este valor, y determinando cuál de las alternativas tenía la utilidad más alta. Para minimizar el efecto del ruido de simulación en la creación de los datos, construí 50 conjuntos de datos para cada nivel de T. Los resultados que se reportan son el promedio entre estos 50 conjuntos de datos.

Calculamos la media de la distribución condicionada para cada cliente, $\bar{\beta}_n$. También calculamos la desviación estándar de $\bar{\beta}_n$ entre los 300 clientes, así como la desviación media absoluta de $\bar{\beta}_n$ respecto a las β_n de los clientes (es decir, la media sobre n de $|\bar{\beta}_n - \beta_n|$). La tabla 11.1 presenta estos estadísticos. Consideremos en primer lugar la desviación estándar. Si no hubiera situaciones de elección observadas sobre las que condicionar ($T = 0$), entonces la distribución condicionada para cada cliente sería igual a la distribución no condicionada (población). Cada cliente tendría la misma $\bar{\beta}_n$, igual a la media de β_n en la población. En este caso, la desviación estándar de $\bar{\beta}_n$ sería cero, ya que todos los clientes tendrían la misma $\bar{\beta}_n$. En el otro extremo, si se observó un número ilimitadamente grande de situaciones de elección ($T \to \infty$), entonces la distribución condicionada para cada cliente colapsaría en su propia β_n. En este caso, la desviación estándar de $\bar{\beta}_n$ sería igual a la desviación estándar de la distribución de β_n en la población, que es 1.0 en este experimento. Para T entre 0 y ∞, la desviación estándar de $\bar{\beta}_n$ está entre 0 y la desviación estándar de β_n en la población.

Tabla 11.1. Ilustración de Monte Carlo

	Coeficiente 1º	Coeficiente 2º
1 situación de elección:		
Desv. Estándar de $\bar{\beta}_n$	0.413	0.416

Diferencia absoluta entre $\bar{\beta}_n$ y β_n	0.726	0.718
10 situación de elección:		
Desv. Estándar de $\bar{\beta}_n$	0.826	0.826
Diferencia absoluta entre $\bar{\beta}_n$ y β_n	0.422	0.448
20 situación de elección:		
Desv. Estándar de $\bar{\beta}_n$	0.894	0.886
Diferencia absoluta entre $\bar{\beta}_n$ y β_n	0.354	0.350
50 situación de elección:		
Desv. Estándar de $\bar{\beta}_n$	0.951	0.953
Diferencia absoluta entre $\bar{\beta}_n$ y β_n	0.243	0.243

En la tabla 11.1 vemos que condicionar sobre pocas situaciones de elección captura una gran parte de la variación de β entre clientes. Con sólo una situación de elección, la desviación estándar de $\bar{\beta}_n$ está en torno a 0.4. Puesto que la desviación estándar de β_n en la población es de 1.0 en este experimento, esto significa que condicionar respecto a una situación de elección captura más del 40 por ciento de la variación en β_n. Con 10 situaciones de elección, capturamos más del 80 de la variación. El rendimiento de observar más situaciones de elección decrece fuertemente. Doblar T de $T = 10$ a $T = 20$ sólo aumenta la proporción de variación capturada desde aproximadamente 0.83 a aproximadamente 0.89. El aumento a $T = 50$ aumenta la proporción a 0.95.

Consideremos ahora la diferencia absoluta entre la media de la distribución condicionada a las elecciones del cliente, $\bar{\beta}_n$, y la β_n real del cliente. Sin condicionar ($T = 0$), la diferencia absoluta media sería de 0.8, que es la diferencia absoluta esperada para variables que siguen un distribución normal estándar como las que tenemos en nuestro experimento. Con un condicionamiento perfecto ($T \to \infty$), $\bar{\beta}_n = \beta_n$ para cada cliente, por lo que la diferencia absoluta sería de 0. Condicionando a una única situación de elección, la desviación absoluta media cae de 0.8 (sin condicionar) a aproximadamente 0.72, es decir, mejora un 10%. La desviación absoluta va decreciendo a medida que el número de situaciones de elección se eleva.

Observe que la caída de la desviación absoluta es menor que el aumento en la desviación estándar. Por ejemplo, con una única situación de elección, la desviación absoluta se mueve un 10% desde su valor en ausencia de condicionamiento hacia el valor con un conocimiento perfecto (de 0.80 con $T = 0$ a 0.72 con $T = 1$, que es un 10% del recorrido entre no condicionar, $T = 0$, y tener un conocimiento perfecto, $T \to \infty$). Sin embargo, la desviación estándar se mueve alrededor de un 40% desde su valor en ausencia de condicionamiento y su valor con conocimiento perfecto (0.4 con $T = 1$ es el 40% de la distancia desde 0 con $T = 0$ a 1 con $T \to \infty$). Esta diferencia se debe al hecho de que la desviación estándar incorpora tanto un movimiento de $\bar{\beta}_n$ en dirección contraria a β_n (alejamiento) como un movimiento hacia β_n (acercamiento). Es importante tener en cuenta este hecho cuando se evalúa la desviación estándar de $\bar{\beta}_n$ en aplicaciones empíricas, en las que la diferencia absoluta no se puede calcular al desconocerse la β_n. Es decir, la desviación estándar de $\bar{\beta}_n$ expresada como un porcentaje de la desviación estándar estimada en la población es una sobreestimación de la cantidad de información contenida en las $\bar{\beta}_n$'s. Con diez situaciones de elección, la desviación estándar promedio en $\bar{\beta}_n$ es más del 80% del valor que tendría con un conocimiento perfecto, y sin embargo la desviación absoluta es menos de la mitad de la que sería sin condicionar.

11.5 Distribución condicionada promedio

Para un modelo correctamente especificado en los verdaderos parámetros de la población, la distribución condicionada de las preferencias, agregada para todos los clientes, es igual a la distribución poblacional de las preferencias. Dada una serie de situaciones de elección descritas por x_n, hay un

conjunto de posibles secuencias de elección. Etiquetemos estas posibles secuencias como y_s para $s = 1, \ldots, S$. Llamemos a la verdadera frecuencia de y_s como $m(y_s|x_n, \theta^*)$, explicitando su dependencia de los verdaderos parámetros θ^*. Si se ha especificado correctamente el modelo y se ha estimado consistentemente, entonces $P(y_s|x_n, \theta^*)$ aproxima $m(y_s|x_n, \theta^*)$ asintóticamente. Condicionando respecto a las variables explicativas, el valor esperado de $h(\beta|y_s, x_n, \hat{\theta})$ es

$$E_y h\left(\beta\middle|y, x_n, \hat{\theta}\right) = \sum_s \frac{P(y_s|x_n, \beta) g\left(\beta\middle|x_n, \hat{\theta}\right)}{P\left(y_s\middle|x_n, \hat{\theta}\right)} m(y_s|x_n, \theta^*)$$

$$\rightarrow \sum_s P(y_s|x_n, \beta) g\left(\beta\middle|x_n, \hat{\theta}\right) = g\left(\beta\middle|x_n, \hat{\theta}\right).$$

Esta relación proporciona una herramienta de diagnóstico del modelo (Allenby y Rossi, 1999). Si la media de las distribuciones condicionadas de las preferencias en los clientes muestreados es similar a la distribución estimada en la población, el modelo está correctamente especificado y calculado con precisión. Si no son similares, la diferencia podría ser debida a (1) un error de especificación, (2) un número insuficiente de valores aleatorios empleados en la simulación, (3) un tamaño de muestra inadecuado y/o (4) la rutina de máxima verosimilitud ha convergido en un máximo local en lugar de un máximo global.

11.6 Caso de estudio: elección de proveedor de energía

11.6.1 Distribución en la población

Se obtuvieron datos de preferencia declarada relativa a la elección del proveedor de electricidad de clientes residenciales. A los clientes encuestados se les presentaron 8-12 situaciones de elección hipotéticas llamadas experimentos. En cada experimento, al cliente se le presentaron cuatro proveedores alternativos con diferentes precios y otras características. En cuanto a precios, los proveedores ofrecían tres tipos de tarifas: (a)precio fijo en centavos de dólar por kilovatio-hora, c/kWh, (b) precios "*time-of-day*" (TOD), variables en función de tramos horarios o (c) precios estacionales con precios diferentes para cada estación del año. El resto de características eran la duración del contrato (tiempo durante el cual el proveedor debe proporcionar el servicio al precio contratado y el cliente debe permanecer con el proveedor si no desea pagar una sanción por haberlo abandonado antes), así como el hecho de si el proveedor es local, una empresa conocida aunque no sea local o una empresa poco conocida. Los datos fueron recogidos por *Research Triangle Institute* (1997) para el *Electric Power Research Institute* y fueron utilizados por Goett (1998) para estimar modelos logits mixtos. Utilizaremos una especificación similar a la de Goett, pero hemos eliminado o combinado variables que resultaron ser insignificantes.

Se estimaron dos modelos logit mixto con estos datos, basados en diferentes especificaciones para la distribución de los coeficientes aleatorios. Todas las elecciones excepto la última situación de elección de cada cliente se utilizaron para estimar los parámetros de la distribución en la población, y la última situación de elección se reservó para comparar la capacidad predictiva de los diferentes modelos y métodos.

La tabla 11.2 muestra los parámetros estimados en la población. El coeficiente de precio en los dos modelos se fija en toda la población de tal manera que la distribución de la predisposición a pagar por cada atributo distinto del precio (que es el ratio entre el coeficiente del atributo en cuestión y el coeficiente de precio) tiene la misma distribución que el coeficiente del propio atributo. Para el modelo 1, todos los coeficientes distintos del precio se especifican mediante una distribución normal en la

población. Se estima la media m y la desviación estándar s de cada coeficiente. Para el modelo 2, los tres primeros coeficientes distintos del precio se especifican para que sean normales, y el cuarto y quinto para que sean log-normales. Las variables cuarta y quinta se refieren a las tarifas TOD y estacionales, y sus coeficientes lógicamente deberían ser negativos para todos los clientes. La distribución log-normal (con los signos de las variables invertidos) proporcionan este comportamiento. El logaritmo de estos coeficientes se distribuye normalmente con media m y desviación estándar s, que son los parámetros que realmente se estiman. Los coeficientes resultantes tienen media $\exp(m + (s^2/2))$ y desviación estándar igual a la media por $\sqrt{\exp(s^2) - 1}$.

Tabla 11.2. Modelo logit mixto de elección de proveedor de energía

	Modelo 1	Modelo 2
Precio, kWh	−0.8574	−0.8827
	(0.0488)	(0.0497)
Duración del contrato, años		
m	−0.1833	−0.2125
	(0.0289)	(0.0261)
s	0.3786	0.3865
	(0.0291)	(0.0278)
Compañía local		
m	2.0977	2.2297
	(0.1370)	(0.1266)
s	1.5585	1.7514
	(0.1264)	(0.1371)
Compañía conocida		
m	1.5247	1.5906
	(0.1018)	(0.0999)
s	0.9520	0.9621
	(0.0998)	(0.0977)
Tarifas TOD(a)		
m	−8.2857	2.1328
	(0.4577)	(0.0543)
s	2.5742	0.4113
	(0.1676)	(0.0397)
Tarifas estacionales(b)		
m	−8.5303	2.1577
	(0.4468)	(0.0509)
s	2.1259	0.2812
	(0.1604)	(0.0217)
Log-verosimilitud en convergencia	−3646.51	−3618.92

Errores estándar entre paréntesis

(a) Tarifas TOD:11c/kWh, 8 a.m.–8 p.m., 5c/kWh, 8 p.m.–8 a.m.

(b) Tarifas estacionales: 10c/kWh en verano, 8c/kWh en invierno, 6c/kWh en primavera y otoño.

Las estimaciones proporcionan los siguientes resultados cualitativos:

- El cliente medio está dispuesto a pagar alrededor de 1/5 a 1/4 c/kWh más, dependiendo del modelo, con el fin de tener un contrato que sea un año más corto. Dicho a la inversa, un

proveedor que requiera a los clientes firmar un contrato de cuatro a cinco años debe descontar de su precio 1 c/kWh para atraer al cliente medio.

- Existe una considerable variación en las actitudes de los clientes hacia la duración del contrato, con una parte relevante de clientes que prefieren un contrato más largo que corto. Un contrato a largo plazo constituye un seguro para el cliente frente a los aumentos de precios, ya que el proveedor está obligado a respetar el precio establecido durante la duración del contrato. Este tipo de contratos, sin embargo, impiden al cliente beneficiarse de precios más bajos que puedan surgir durante la vigencia del mismo. Al parecer, muchos clientes valoran más la seguridad frente a un incremento de los precios que el riesgo de perder una oportunidad de disfrutar precios más bajos. El grado de heterogeneidad de los clientes implica que el mercado puede soportar contratos de diferentes duraciones, con proveedores obteniendo beneficios mediante la oferta de diferentes tipos de contratos dirigidos a diferentes segmentos de población.
- El cliente medio está dispuesto a pagar la friolera de 2.5 c/kWh más por su proveedor local que por un proveedor desconocido. Sólo una pequeña parte de los clientes prefieren un proveedor desconocido a su compañía eléctrica local. Este hallazgo tiene implicaciones importantes para la competencia. Implica que la entrada en el mercado residencial de proveedores previamente desconocidos será muy difícil, sobre todo porque los descuentos en los precios que los competidores pueden ofrecer en la mayoría de los mercados son bastante reducidos. La experiencia en California, donde sólo el 1 por ciento de los clientes residenciales han abandonado a su proveedor local después de varios años de libre competencia, es consistente con este hallazgo.
- El cliente medio está dispuesto a pagar 1.8 c/kWh más por un proveedor conocido que por uno desconocido. Los valores estimados de s implican que una parte considerable de los clientes estarían dispuestos a pagar más por un proveedor conocido que por su proveedor local, presumiblemente a causa de una mala experiencia o una actitud negativa hacia la compañía eléctrica local. Estos resultados implican que las empresas que son conocidas por los clientes , tales como los operadores de larga distancia, los operadores de telecomunicaciones locales, las empresas locales de cable e incluso los minoristas como Sears y Home Depot, pueden ser más exitosas atrayendo clientes para el suministro de electricidad que empresas desconocidas antes de su entrada en el mercado como proveedores de energía.
- El cliente promedio valora las tarifas TOD de una forma bastante consistente con los patrones de uso de este tipo de tarifa. En el modelo 1, el coeficiente medio de la variable indicadora de tarifa TOD implica que el cliente promedio considera que estas tarifas son equivalentes a un precio fijo de 9.7 c/kWh. En el modelo 2, la media estimada y la desviación estándar del logaritmo del coeficiente implican que la mediana de la predisposición a pagar es de 8.4 c/kWh y la media de 10.4 c/kWh, que abarca la media del modelo 1. Aquí 9.5 c/kWh es el precio medio que un cliente pagaría en un sistema tarifario TOD si el 75 por ciento de su consumo se produjese durante el día (entre las 8 de la mañana y las 8 de la tarde) y el otro 25 por ciento se produjese durante la noche. Estas proporciones, aunque tal vez un poco altas para el día, son razonables. Los valores estimados de s son altamente significativos, reflejando heterogeneidad en los patrones de uso y tal vez capacidad de los clientes para desplazar el horario de su consumo en respuesta a las tarifas TOD. Estos valores son mayores de lo que razonablemente cabe esperar, lo que implica que una parte no despreciable de los clientes tratan las tarifas TOD como equivalentes a un precio fijo más alto que el precio de la tarifa TOD más alto, o menor que el precio de la tarifa TOD más bajo.

- El cliente medio parece evitar tarifas estacionales por razones que van más allá de los propios precios. El cliente medio trata las tarifas estacionales como equivalentes a un precio fijo de 10 c/kWh, que es el precio más alto dentro de la tarifa estacional. Una posible explicación de este resultado se relaciona con la variación estacional de las facturas de los clientes. En muchas regiones, el consumo de electricidad es más alto en verano, cuando se usa el aire acondicionado, y las facturas son, por tanto, mayores en verano que en otras épocas del año, incluso en sistemas tarifarios fijos. La variación en las facturas entre diferentes meses, sin variación proporcional en los ingresos, hace que sea más difícil para los clientes pagar sus facturas en verano. De hecho, la falta de pago de la mayoría de los servicios de energía es más frecuente en verano. Las tarifas estacionales, que aplican un precio más alto en verano, incrementan la variación del importe de las facturas entre diferentes estaciones. Los clientes estarían evitando racionalmente un sistema tarifario que exacerba una dificultad ya existente. Si esta interpretación es correcta, entonces las tarifas estacionales combinadas con una facturación suavizada (en la cual el proveedor carga una parte de la factura de verano en la factura de invierno) podrían proporcionar una solución atractiva para los clientes y proveedores por igual.

El modelo 2 alcanza un valor de log-verosimilitud mayor que el modelo 1, presumiblemente debido a que la distribución log-normal asegura coeficientes negativos para las tarifas TOD y las tarifas estacionales.

11.6.2 Distribuciones condicionadas

A continuación usaremos los modelos estimados para calcular las distribuciones condicionadas de los clientes y las medias de estas distribuciones. Calcularemos $\bar{\beta}_n$ para cada cliente de dos maneras. En primer lugar, calculamos $\bar{\beta}_n$ utilizando la ecuación (11.3) con las estimaciones puntuales de los parámetros de la población. En segundo lugar, utilizamos el procedimiento descrito en la sección 11.3 para integrar sobre la distribución muestral de los parámetros estimados de la población.

Las medias y las desviaciones estándar de $\bar{\beta}_n$ entre clientes de la muestra, calculadas mediante estos dos métodos, se muestran en las tablas 11.3 y 11.4, respectivamente. El coeficiente de precio no se muestra en la tabla 11.3, pues es fijo para toda la población. La tabla 11.4 incorpora la distribución muestral de los parámetros en la población, que incluye la varianza en el coeficiente de precio.

Tabla 11.3. $\bar{\beta}_n$ medio usando estimaciones puntuales de $\hat{\theta}$.

	Modelo 1	Modelo 2
Duración del contrato		
Media	−0.2028	−0.2149
Desviación estándar	0.3175	0.3262
Proveedor local		
Media	2.1205	2.2146
Desviación estándar	1.2472	1.3836
Compañía conocida		
Media	1.5360	1.5997
Desviación estándar	0.6676	0.6818
Tarifa TOD		
Media	−8.3194	−9.2584
Desviación estándar	2.2725	3.1051
Tarifa estacional		
Media	−8.6394	−9.1344
Desviación estándar	1.7072	2.0560

Tabla 11.4 $\bar{\beta}_n$. medio usando la distribución muestral de $\hat{\theta}$

	Modelo 1	Modelo 2
Precio		
Media	−0.8753	−0.8836
Desviación estándar	0.5461	0.0922
Duración del contrato		
Media	−0.2004	−0.2111
Desviación estándar	0.3655	0.3720
Proveedor local		
Media	2.1121	2.1921
Desviación estándar	1.5312	1.6815
Compañía conocida		
Media	1.5413	1.5832
Desviación estándar	0.9364	0.9527
Tarifa TOD		
Media	−9.1615	−9.0216
Desviación estándar	2.4309	3.8785
Tarifa estacional		
Media	−9.4528	−8.9408
Desviación estándar	1.9222	2.5615

Observe en primer lugar los resultados de la tabla 11.3. La media de $\bar{\beta}_n$ está muy cerca de la media estimada de la población que aparece en la tabla 11.2. Esta similitud es la esperada para un modelo correctamente especificado y consistentemente estimado. La desviación estándar de $\bar{\beta}_n$ sería cero si no hubiera condicionamiento y sería igual a la desviación estándar de la población, si el coeficiente de cada cliente se conociese con exactitud. Las desviaciones estándar de la tabla 11.3 están considerablemente por encima de cero y están bastante cerca de las desviaciones estándar estimadas para la población, mostradas en la tabla 11.2. Por ejemplo, en el modelo 1, la media condicionada del coeficiente de duración de contrato tiene una desviación estándar entre clientes de 0.318, y la estimación puntual de la desviación estándar en la población es 0.379. Por lo tanto, la variación en $\bar{\beta}_n$ captura más del 70 por ciento de la variación total estimada en este coeficiente. Se obtienen resultados similares para otros coeficientes. Este resultado implica que la media de la distribución condicionada de un cliente captura una proporción bastante grande de la variación en los coeficientes entre clientes y tiene el potencial de ser útil para diferenciar clientes.

Tal y como vimos en la sección 11.5, es posible hacer un diagnóstico relativo a la bondad de la especificación y la estimación del modelo, mediante la comparación de la media muestral de las distribuciones condicionadas con la distribución poblacional estimada. Las medias de la tabla 11.3 representan las medias del promedio muestral de las distribuciones condicionadas. La desviación estándar de la distribución condicionada promediada en la muestra depende de la desviación estándar de $\bar{\beta}_n$, que se muestra en la tabla 11.3, más la desviación estándar de $\beta_n - \bar{\beta}_n$. Cuando se añade este último componente, la desviación estándar de cada coeficiente coincide muy aproximadamente con la desviación estándar estimada en la población. Esta equivalencia sugiere que no hay un error significativo de especificación y que los parámetros estimados en la población son bastante exactos. Esta sugerencia se ve matizada, sin embargo, por los resultados mostrados en la tabla 11.4.

La tabla 11.4 muestra la media muestral y la desviación estándar de la media de la distribución muestral de $\bar{\beta}_n$, que se induce por la distribución muestral de $\hat{\theta}$. Las medias de la tabla 11.4 son las medias del promedio en la muestra de $h(\beta|y_n, x_n, \hat{\theta})$ integradas respecto a la distribución muestral de $\hat{\theta}$. Para el

modelo 1, se produce una discrepancia que podría indicar un posible error de especificación. En particular, las medias de los coeficientes relativos a las tarifas TOD y las tarifas estacionales de la tabla 11.4 exceden su media poblacional estimada, mostrada en la tabla 11.2. Curiosamente, las medias para estos coeficientes en la tabla 11.4 para el modelo 1 están más cerca de las medias análogas del modelo 2 que de las medias poblacionales estimadas para el modelo 1, mostradas en la tabla 11.2. El modelo 2 usa la distribución log-normal, cuya forma es más razonable para estos coeficientes, y obtiene un ajuste considerablemente mejor que el del modelo 1. El condicionamiento en el modelo 1 parece que desplaza los coeficientes hacia los valores obtenidos por el modelo 2 - mejor especificado – y los aleja de sus propias distribuciones poblacionales, deficientemente especificadas. Éste es un ejemplo de cómo comparar la distribución estimada de la población con el promedio en la muestra de la distribución condicionada puede revelar información acerca de la especificación y la estimación.

Las desviaciones estándar mostradas en la tabla 11.4 son más grandes que las de la tabla 11.3. Esta diferencia se debe al hecho de que la varianza muestral de los parámetros estimados en la población está incluida en los cálculos hechos para la tabla 11.4, pero no para la tabla 11.3. Las desviaciones estándar más grandes no significan que la porción de la varianza total de β_n que se captura por la variación en $\bar{\beta}_n$ sea mayor cuando se tiene en cuenta la distribución en la muestra respecto a cuando no se tiene en cuenta.

Es posible obtener información útil para acciones de marketing mediante el examen de la $\bar{\beta}_n$ de cada cliente. El valor de esta información para la comercialización segmentada de productos/servicios se ha destacado por Rossi et al. (1996). La tabla 11.5 muestra la $\bar{\beta}_n$ calculada para los tres primeros clientes del conjunto de datos, junto con la media de β_n en la población.

Tabla 11.5. Medias condicionadas para tres clientes

	Población	Cliente 1	Cliente 2	Cliente 3
Duración del contrato	−0.213	0.198	−0.208	−0.401
Proveedor local	2.23	2.91	2.17	0.677
Compañía conocida	1.59	1.79	2.15	1.24
Tarifa TOD	−9.19	−5.59	−8.92	−12.8
Tarifa estacional	−9.02	−5.86	−11.1	−10.9

El primer cliente prefiere un contrato a largo plazo, a diferencia de la gran mayoría de los clientes, a los cuales no les gustan los contratos a largo plazo. Él está dispuesto a pagar un precio superior por la energía si este precio está garantizado a través de un contrato de larga duración. Este cliente evalúa las tarifas TOD y las tarifas estacionales de forma generosa, como si todo su consumo se produjese en el período de menor precio (tenga en cuenta que el precio más bajo con tarifa TOD es de 5 c/kWh y el precio más bajo con tarifa estacional es de 6 c/kWh). Es decir, el primer cliente probablemente está dispuesto a pagar por tener asignada una tarifa TOD o una tarifa estacional, más de lo que las tarifas realmente valen en términos de reducción de la factura. Por último, este cliente está dispuesto a pagar más que el promedio de los clientes por permanecer con la compañía eléctrica local. Desde una perspectiva de marketing, la compañía eléctrica local fácilmente puede retener y obtener beneficios adicionales de este cliente si le ofrece un contrato a largo plazo con tarifas TOD o tarifas estacionales.

Al tercer cliente no le gustan las tarifas estacionales y las tarifas TOD, evaluando ambas como si todo su consumo se produjese en los periodos con precios más caros. A este cliente le desagradan los contratos de larga duración mucho más que a la media de los clientes y, sin embargo, a diferencia de la mayoría de los clientes, prefiere recibir el servicio de una empresa conocida que no sea su compañía local. Este

cliente es un objetivo prioritario para una compañía bien conocida si esta compañía le ofrece un precio fijo sin requerirle un compromiso de permanencia.

El segundo cliente no es una oportunidad de marketing tan clara. Una empresa bien conocida está aproximadamente en igualdad de condiciones con la empresa local de energía para atraer a este cliente. Esto en sí mismo puede hacer que el cliente sea un objetivo para proveedores reconocidos, ya que está menos ligado a la empresa local de energía que la mayoría de los clientes. Sin embargo, más allá de esta información, hay poco más destacable en las preferencias de este cliente, salvo los precios bajos (que son valorados por todos los clientes). Su evaluación de las tarifas TOD y las tarifas estacionales es suficientemente negativa como para hacer poco probable que un proveedor pueda atraerlo y obtener un beneficio de él ofreciéndole estas tarifas. Asimismo, el cliente está dispuesto a pagar para evitar un contrato a largo plazo, por lo que un proveedor podría atraer a este cliente al no exigirle un contrato si otros proveedores sí se lo están exigiendo. Sin embargo, si otros proveedores tampoco están exigiendo contratos, parece que hay pocas más cosas que un proveedor pueda usar para tener ventaja sobre sus competidores. Aparentemente, este cliente será conquistado por la compañía que le ofrezca el precio fijo más bajo.

El análisis expuesto relativo a estos tres clientes ilustra el tipo de información que se puede obtener mediante el condicionamiento respecto a la elección de los clientes, y cómo la información puede traducirse fácilmente en una segmentación del mercado y en la identificación de oportunidades de marketing rentables.

11.6.3 Probabilidad condicionada para la última elección

Recordemos que la última situación de elección que afronta cada cliente no se incluyó en la estimación. Por lo tanto, se puede considerar una nueva situación de elección y se puede usar para evaluar el efecto de condicionar sobre decisiones pasadas. Para ello, identificamos qué alternativa escogió cada cliente en la nueva situación de elección y calculamos la probabilidad de elección de esa alternativa. La probabilidad se calculó en primer lugar sin condicionar en las elecciones anteriores. Este cálculo utiliza la fórmula del modelo logit mixto (11.5) empleando la distribución en la población de β_n y las estimaciones puntuales de los parámetros de la población. El promedio de esta probabilidad no condicionada entre los clientes es 0.353. A continuación se calculó la probabilidad condicionada a las elecciones anteriores. Se utilizaron cuatro maneras diferentes de calcular esta probabilidad:

1. Basada en la fórmula (11.6), utilizando las estimaciones puntuales de los parámetros de la población.
2. Basada en la fórmula (11.6), junto con el procedimiento descrito en la sección 11.3 que tiene en cuenta la varianza muestral de las estimaciones de los parámetros de la población.
3. (y 4). Empleando la fórmula logit

$$\frac{e^{\beta_n' x_{niT+1}}}{\sum_j e^{\beta_n' x_{njT+1}}},$$

usando la media condicionada $\bar{\beta}_n$ como β_n. Este método es equivalente a utilizar la $\bar{\beta}_n$ del cliente como si fuera una estimación de los verdaderos coeficientes del cliente, β_n. Las dos versiones (3 y 4) difieren en si $\bar{\beta}_n$ se calcula sobre la base de la estimación puntual de los parámetros de la población (método 3) o se tiene en cuenta la distribución en la muestra de estos parámetros (método 4).

Los resultados se dan en la tabla 11.6 para el modelo 2. El resultado más destacado es que el condicionamiento sobre las elecciones anteriores de cada cliente mejora las previsiones para la última situación de elección de forma considerable. La probabilidad promedio de la alternativa elegida se incrementa de 0.35 sin condicionar a más de 0.50 al condicionar. Para cerca de tres cuartas partes de los 361 clientes en la muestra, la predicción de su última situación de elección es mejor al condicionar que al no hacerlo, incrementándose la probabilidad promedio de elección en más de 0.25. Para el resto de clientes, el condicionamiento hace que la predicción en las últimas situaciones de elección sea menos precisa, disminuyendo la probabilidad media de elección para estos clientes.

Tabla 11.6. Probabilidad de la alternativa elegida en la última situación de elección

	Método 1	Método 2	Método 3	Método 4
Probabilidad promedio	0.5213	0.5041	0.5565	0.5487
Número de clientes cuya probabilidad aumenta al condicionar	266	260	268	264
Incremento promedio en la probabilidad para clientes con incremento	0.2725	0.2576	0.3240	0.3204
Número de clientes cuya probabilidad disminuye al condicionar	95	101	93	97
Disminución promedio en la probabilidad para clientes con disminución	0.1235	0.1182	0.1436	0.1391

Hay varias razones por las que la probabilidad predicha después de condicionar no siempre es mayor. En primer lugar, los experimentos de elección se construyeron de manera que cada situación era bastante diferente del resto de situaciones, con el fin de obtener la mayor variación posible. Si la última situación implica nuevas comparaciones, las opciones anteriores no serán de utilidad y de hecho pueden ser perjudiciales para la predicción de la última elección. Una prueba más apropiada podría ser el diseño de una serie de situaciones de elección que obtenga información sobre comparaciones relevantes y luego diseñar una situación extra que esté dentro de la gama de comparaciones empleadas.

En segundo lugar, no hemos incluido en nuestro modelo todos los atributos de las alternativas que se presentaron a los clientes. En particular, hemos omitido los atributos que no entraron de manera significativa en la estimación de los parámetros de la población. Algunos clientes podrían responder a estos atributos omitidos, a pesar de que son insignificantes para la población en su conjunto. En la medida en que la última situación de elección implica comparaciones relativas a estos atributos, las distribuciones de preferencias condicionadas serían engañosas, dado que las preferencias relevantes quedarían excluidas. Esta explicación sugiere que si un modelo logit mixto va a ser usado para la obtención de densidades condicionadas para cada cliente, el investigador quizá debería incluir atributos que podrían ser importantes para algunos individuos a pesar de que son insignificantes para la población como un todo.

En tercer lugar, independientemente de cómo se ha diseñado la encuesta y el modelo, algunos clientes podrían responder a situaciones de elección de forma quijotesca, de manera que las preferencias evidenciadas en las elecciones anteriores no sean aplicadas por el cliente en la última situación de elección.

Por último, factores aleatorios pueden hacer que la probabilidad de algunos clientes caiga al condicionar, incluso cuando las tres primeras razones no lo hagan.

Aunque al menos una de estas razones pudiera estar contribuyendo a reducir la probabilidad de elección para algunos de los clientes de nuestra muestra, la ganancia en la exactitud de predicción para los clientes con un aumento de la probabilidad después del condicionamiento, es más del doble que la

pérdida de precisión para aquellos con una disminución de la probabilidad, y el número de clientes con una ganancia es casi tres veces mayor que el número de cliente con una pérdida.

El tercer método (el más fácil), que se limita a calcular la fórmula logit estándar utilizando la $\bar{\beta}_n$ de los clientes sobre la base de la estimación puntual de los parámetros de la población, da la mayor probabilidad. Este procedimiento no permite contemplar la distribución de β_n en torno a $\bar{\beta}_n$ o la distribución muestral de $\hat{\theta}$. Permitir cualquiera de estas varianzas reduce la probabilidad media: usar la distribución condicionada de $\beta_n$en lugar de usar sólo la media $\bar{\beta}_n$ (métodos 1 y 2 en comparación con los métodos 3 y 4, respectivamente) reduce la probabilidad media, y permitir la distribución muestral de $\hat{\theta}$ en lugar de la estimación puntual (métodos 2 y 4 en comparación con los métodos 1 y 3, respectivamente) también reduce la probabilidad media. Este resultado no significa que el método 3, que incorpora la menor varianza, sea superior a los otros. Los métodos 3 y 4 son consistentes sólo si el número de situaciones de elección es capaz de incrementarse sin límite, de manera que $\bar{\beta}_n$ pueda ser considerado como un estimador de β_n. Con T fijo, los métodos 1 y 2 son más adecuados, ya que incorporan toda la densidad condicionada.

11.7 Exposición

En este capítulo se muestra cómo la distribución de los coeficientes condicionada a las elecciones observadas del cliente se obtiene a partir de la distribución de los coeficientes de la población. Si bien estas distribuciones condicionadas pueden ser útiles de varias formas, es importante reconocer las limitaciones del concepto. En primer lugar, el uso de distribuciones condicionadas para hacer predicciones se limita a aquellos clientes cuyas elecciones anteriores han sido observadas. En segundo lugar, mientras que la distribución condicionada de cada cliente puede utilizarse en el análisis de conglomerados (*cluster analysis*) así como para otros fines de identificación, el investigador a menudo quiere relacionar preferencias con características demográficas observables de los clientes. Sin embargo, estos datos demográficos observables de los clientes podrían ser introducidos directamente en el modelo, de modo que los parámetros de la población varíen con las características observadas de los clientes en la población. De hecho, introducir demográficos en el modelo es más directo y más accesible para hacer pruebas de hipótesis que estimar un modelo sin esas características, calcular la distribución condicionada para cada cliente y luego hacer el análisis de conglomerados y otros análisis usando los momentos estadísticos de las distribuciones condicionadas.

Teniendo en cuenta estas cuestiones, hay tres razones principales por las que un investigador podría beneficiarse del cálculo de las distribuciones condicionadas de los clientes. En primer lugar, cada vez resulta más simple disponer de información sobre elecciones pasadas. Algunos ejemplos de esta disponibilidad incluyen los datos que provienen de un lector de códigos de barras de clientes que compran en tiendas de comestibles usando una tarjeta de fidelización, los programas para viajeros frecuentes que ofrecen las aerolíneas y las compras de particulares a través de Internet. En estas situaciones, condicionar sobre elecciones pasadas permite una comercialización específica eficaz y el desarrollo de nuevos productos y servicios que respondan a las preferencias reveladas de los subgrupos de clientes.

En segundo lugar, las características demográficas que diferencian a los clientes con diferentes preferencias podrían ser más evidentes a través de un análisis de conglomerados aplicado sobre distribuciones condicionadas, que a través de pruebas de especificación en el modelo mismo. El análisis de conglomerados tiene su propia forma de identificar patrones, que puede ser en algunos casos más eficaz que las pruebas de especificación dentro de un modelo de elección discreta.

En tercer lugar, el examen de las distribuciones condicionadas de los clientes a menudo puede identificar patrones que no pueden relacionarse con características observadas de esos clientes, pero que sin

embargo es útil conocerlos. Por ejemplo, saber que un producto o una campaña de marketing serán de interés para una parte de la población debido a sus preferencias particulares suele ser suficiente, sin necesidad de identificar a ese grupo de personas sobre la base de sus características demográficas. Las densidades condicionadas pueden facilitar enormemente los análisis que tienen estos objetivos.

12

Procedimientos bayesianos

12.1 Introducción

Un potente conjunto de procedimientos para la estimación de modelos de elección discreta ha sido desarrollado dentro de la tradición bayesiana. Los conceptos clave fueron introducidos por Albert y Chib (1993) y McCulloch y Rossi (1994) en el contexto de los modelos probit, y por Allenby y Lenk (1994) y Allenby (1997) para logits mixtos con coeficientes distribuidos normalmente. Estos autores mostraron cómo se pueden estimar los parámetros de los modelos sin necesidad de calcular las probabilidades de elección. Sus procedimientos proporcionan una alternativa a los métodos de estimación clásicos descritos en el capítulo 10. Rossi et al . (1996), Allenby (1997), y Allenby y Rossi (1999) también mostraron cómo pueden utilizarse estos procedimientos para obtener información sobre los parámetros a nivel individual dentro de un modelo con variación de preferencias aleatorias. Por esta vía, los autores proporcionan un análogo bayesiano a los procedimientos clásicos que hemos descrito en el capítulo 11. Las variaciones realizadas sobre estos procedimientos para acomodar otros aspectos del comportamiento han sido numerosas. Por ejemplo, Arora et al. (1998) generalizó el procedimiento logit mixto para tener en cuenta la cantidad de compras, así como la elección de la marca en cada elección de compra. Bradlow y Fader (2001) mostraron cómo pueden usarse métodos similares para examinar datos de ordenación (*ranking data*) a nivel agregado en lugar de datos de elección a nivel individual. Chib y Greenberg (1998) y Wang et al. (2002) desarrollaron métodos para estudiar respuestas discretas interrelacionadas. Chiang et al. (1999) estudiaron las situaciones en las que el conjunto de opciones de elección disponibles para el decisor es desconocido para el investigador. Train (2001) amplió el procedimiento bayesiano para logit mixto con el fin de poder usarlo con distribuciones de coeficientes no normales, incluyendo distribuciones log-normales, uniformes y triangulares.

Los procedimientos bayesianos evitan dos de las dificultades más importantes asociadas a los procedimientos clásicos. En primer lugar, los procedimientos bayesianos no requieren la maximización de una función. Con probit y algunos modelos logit mixtos (especialmente los que usan distribuciones log-normales), la maximización de la función de verosimilitud simulada puede ser numéricamente difícil. A menudo, el algoritmo no converge por diversas razones. La elección de los valores de inicio del algoritmo suele ser crítica, lo que puede producir que el algoritmo converja a partir de valores iniciales cercanos al máximo, pero no desde otros valores de partida. La cuestión de los máximos locales frente a los globales complica la maximización aún más, ya que la convergencia no garantiza que se haya alcanzado el máximo global. En segundo lugar, propiedades de estimación deseables, tales como la

consistencia y la eficiencia, se pueden conseguir en condiciones más relajadas con procedimientos bayesianos que con procedimientos clásicos. Como se muestra en el Capítulo 10, la máxima verosimilitud simulada es consistente sólo si se considera que el número de valores al azar utilizados en la simulación aumenta con el tamaño de la muestra, y la eficiencia se alcanza sólo si el número de valores al azar crece más rápido que la raíz cuadrada del tamaño de la muestra. En contraste, los estimadores bayesianos que describiremos son consistentes para un número fijo de valores al azar utilizados en la simulación y son eficientes si el número de valores al azar crece con cualquier proporción respecto al tamaño de la muestra.

Estas ventajas tienen un precio, por supuesto. Para los investigadores que están acostumbrados a trabajar desde una perspectiva clásica, la curva de aprendizaje puede ser difícil. El investigador debe asimilar numerosas técnicas y conceptos relacionados entre sí antes de poder apreciar la potencia de estos métodos. Sin embargo, puedo asegurar al lector que el esfuerzo vale la pena. Otro costo de los procedimientos bayesianos es más fundamental. Para simular los estadísticos pertinentes que se definen para una distribución, los procedimientos bayesianos utilizan un proceso iterativo que converge, con un número suficiente de iteraciones, hacia valores al azar extraídos de esa distribución. Esta convergencia es diferente de la convergencia a un valor máximo que se necesita para los procedimientos clásicos e implica su propio conjunto de dificultades. El investigador no puede determinar fácilmente si la convergencia realmente se ha logrado. Por lo tanto, los procedimientos bayesianos cambian las dificultades de convergencia a un máximo por las dificultades asociadas con este otro tipo de convergencia. El investigador tendrá que decidir, en un entorno en particular, qué tipo de convergencia es menos pesada.

Para algunos modelos de comportamiento y algunas especificaciones de distribución, los procedimientos bayesianos son mucho más rápidos y, una vez el investigador clásico haya afrontado el aprendizaje inicial necesario, son más sencillos desde una perspectiva de programación que los procedimientos clásicos. Para otros modelos, los procedimientos clásicos son más fáciles. Exploraremos la velocidad relativa entre procedimientos bayesianos y clásicos en las secciones que siguen. Las diferencias se pueden clasificar fácilmente, a través de una comprensión de cómo operan los dos conjuntos de procedimientos. El investigador puede utilizar este conocimiento para decidir qué procedimiento utilizar en un entorno particular.

Antes de proceder, dos factores deben tenerse en cuenta. En primer lugar, los procedimientos bayesianos, y el término "jerárquico bayesiano" que se utiliza a menudo en el contexto de los modelos de elección discreta, se refieren a un método de estimación, no a un modelo de comportamiento. Probit, logit mixto o cualquier otro modelo que el investigador especifique, en principio puede ser estimado por procedimientos clásicos o bayesianos. En segundo lugar, la perspectiva bayesiana a partir de la cual surgen estos procedimientos proporciona un paradigma rico e intelectualmente satisfactorio para la inferencia y la toma de decisiones. Sin embargo, aunque un investigador no esté interesado en la perspectiva bayesiana, igualmente puede beneficiarse de este tipo de procedimientos: el uso de procedimientos bayesianos no requiere que el investigador adopte una perspectiva bayesiana de los estadísticos. Como se verá, los procedimientos bayesianos proporcionan un estimador cuyas propiedades pueden ser examinadas e interpretadas de forma totalmente clásica. Bajo ciertas condiciones, el estimador que resulta de los procedimientos bayesianos es asintóticamente equivalente al estimador de máxima verosimilitud. Por ello, el investigador puede utilizar procedimientos bayesianos para obtener estimaciones de los parámetros y luego interpretarlos como si fueran estimadores de máxima verosimilitud. Un hecho destacable de los procedimientos bayesianos es que los resultados pueden ser interpretados desde ambas perspectivas simultáneamente, usando las ideas fundamentales de cada tradición. Esta doble interpretación se puede aplicar a los procedimientos clásicos, cuyos resultados pueden ser transformados para ser interpretados desde un punto de vista bayesiano, tal y

como describe Geweke (1989). En resumen, la perspectiva estadística del investigador no tiene que dictar la elección del procedimiento.

En las secciones siguientes proporcionaremos un resumen de los principios bayesianos en general, introduciendo el concepto de distribuciones a priori y a posteriori. A continuación, mostraremos cómo la media de la distribución posterior puede ser interpretada desde una perspectiva clásica como asintóticamente equivalente al máximo de la función de verosimilitud. Luego abordaremos el problema numérico relativo a cómo calcular la media de la distribución posterior. El muestreo de Gibbs y, más en general, el algoritmo Metropolis-Hastings pueden utilizarse para extraer valores al azar de prácticamente cualquier distribución posterior, sin importar su complejidad. La media de estos valores al azar simula la media de la distribución posterior y es por tanto la estimación de los parámetros. La desviación estándar de los valores extraídos al azar proporciona los errores estándar clásicos de las estimaciones. Aplicaremos el método a un modelo logit mixto y compararemos la dificultad numérica y la velocidad del procedimiento bayesiano con las de los procedimientos clásicos en diversas especificaciones.

12.2 Introducción a los conceptos bayesianos

Considere un modelo con parámetros θ. El investigador tiene algunas ideas iniciales sobre el valor de estos parámetros y recoge datos para mejorar dichas ideas. Bajo la perspectiva del análisis bayesiano, las ideas que el investigador tiene acerca de los parámetros se representan mediante una distribución de probabilidad sobre todos los posibles valores que pueden tomar los parámetros, donde la probabilidad representa las opciones que el investigador otorga a que los parámetros tengan un determinado valor. Antes de recoger datos, las ideas del investigador se basan en la lógica, la intuición o en análisis anteriores. Estas ideas están representadas por una densidad en θ, denominada distribución a priori, que se denota como $k(\theta)$. El investigador recopila datos con el fin de mejorar sus ideas previas sobre el valor de θ. Supongamos que el investigador observa una muestra de N decisores independientes. Denominemos y_n a la elección (o elecciones) observada(s) del decisor n, y denominemos colectivamente $Y = \{y_1, \dots, y_N\}$ al conjunto de elecciones observadas de toda la muestra. En base a esta información de la muestra, el investigador cambia o actualiza sus ideas acerca de θ. Las ideas actualizadas están representadas por una nueva densidad de probabilidad de θ, etiquetada como $K(\theta|Y)$ y a la que nos referimos como distribución posterior. Esta distribución posterior depende de Y, ya que incorpora la información que está contenida en la muestra observada.

La cuestión que surge es: ¿cómo cambian exactamente las ideas del investigador acerca de θ al observar Y? Es decir, ¿cómo difiere la distribución posterior $K(\theta|Y)$ de la distribución a priori $k(\theta)$? Existe una relación precisa entre la distribución a priori y a posteriori, establecida por la regla de Bayes. Sea $P(y_n|\theta)$ la probabilidad de observar los resultados $y_n$por parte del decisor n. Esta probabilidad es el modelo de comportamiento que relaciona las variables explicativas y los parámetros con los resultados, aunque hemos omitido la notación relativa a las variables explicativas por razones de simplicidad. La probabilidad de observar los resultados Y en la muestra es

$$L(Y|\theta) = \prod_{n=1}^{N} P(y_n|\theta).$$

Esta es la función de verosimilitud (sin el logaritmo) de las elecciones observadas. Tenga presente que es una función de los parámetros θ.

La regla de Bayes proporciona el mecanismo por el cual el investigador mejora sus ideas acerca de θ. Por las reglas sobre probabilidades condicionadas

(12.1) $$K(\theta|Y)L(Y) = L(Y|\theta)k(\theta),$$

donde $L(Y)$ es la probabilidad marginal de Y, marginal respecto a θ.

$$L(Y) = \int L(Y|\theta)k(\theta)d\theta.$$

Ambos lados de la ecuación (12.1) representan la probabilidad conjunta de Y y θ, aplicando el condicionamiento en direcciones opuestas. El lado izquierdo es el producto de la probabilidad de Y por la probabilidad de θ dado Y, mientras que el lado derecho es el producto de la probabilidad de θ por la probabilidad de Y dado θ. Reordenando, tenemos

(12.2) $$K(\theta|Y) = \frac{L(Y|\theta)k(\theta)}{L(Y)}.$$

Esta ecuación es la regla de Bayes aplicada a las distribuciones previas y posteriores. En general, la regla de Bayes vincula probabilidades condicionadas e incondicionadas en cualquier situación, y no implica una perspectiva bayesiana de la estadística. La estadística bayesiana surge cuando la probabilidad no condicionada es la distribución a priori (que refleja las ideas que el investigador tiene acerca de θ sin condicionar a la información proporcionada por la muestra) y la probabilidad condicionada es la distribución posterior (que refleja las ideas del investigador acerca de θ, condicionadas a la información proporcionada por la muestra).

Podemos expresar la ecuación (12.2) de forma más compacta y conveniente. La probabilidad marginal de Y, $L(Y)$, es constante respecto a θ y, más específicamente, es la integral del numerador de (12.2). Como tal, $L(Y)$ es simplemente la constante de normalización que asegura que la integral de la distribución posterior suma 1, como se requiere para cualquier densidad bien definida. Usando este hecho, la ecuación (12.2) se puede expresar más sucintamente diciendo simplemente que la distribución posterior es proporcional al producto de la distribución a priori por la función de verosimilitud:

$$K(\theta|Y) \propto L(Y|\theta)k(\theta).$$

Intuitivamente, la probabilidad que el investigador atribuye a un determinado valor de los parámetros después de ver la muestra es la probabilidad que le atribuía antes de ver la muestra por la probabilidad (es decir, la verosimilitud) de que esos valores produzcan las elecciones observadas.

La media de la distribución posterior es

(12.3) $$\bar{\theta} = \int \theta K(\theta|Y)d\theta.$$

Esta media tiene importancia, tanto desde el punto de vista bayesiano como desde la perspectiva clásica. Desde una perspectiva bayesiana, $\bar{\theta}$ es el valor de θ que minimiza el costo esperado que tiene para el investigador equivocarse acerca θ, si el costo del error es una función cuadrática del tamaño del error. Desde una perspectiva clásica, $\bar{\theta}$ es un estimador que tiene la misma distribución muestral asintótica que el estimador de máxima verosimilitud. Explicamos estos dos conceptos en los apartados siguientes.

12.2.1 Propiedades bayesianas de $\bar{\theta}$

La visión del investigador acerca de θ está representada por la distribución posterior $K(\theta|Y)$ después de observar la muestra. Supongamos que pedimos al investigador que adivine el verdadero valor de θ y le decimos que recibirá una penalización en función del grado en que su conjetura difiera del valor real. De forma más realista, suponga que debemos tomar una decisión que depende del valor de θ, como por ejemplo un fabricante que deba fijar el precio de un producto cuando los ingresos a cualquier nivel de precio dependen de la elasticidad de la demanda. Tomar una mala decisión tiene un costo, como fijar el precio basándose en la creencia de que la elasticidad del precio es -0.2 cuando realmente es -0.3. La pregunta que surge es: ¿qué valor de θ debería usar el investigador en estas decisiones con el objetivo de minimizar el costo esperado de estar equivocado, dadas sus creencias acerca de θ, representadas en la distribución posterior?

Si el costo de equivocarse es cuadrático en relación a la distancia entre el valor θ que utiliza en la decisión y la valor verdadero θ^*, entonces el valor óptimo de θ a utilizar en la decisión es $\bar{\theta}$. Este hecho se puede demostrar de la siguiente manera. Si el investigador utiliza θ_0 en sus decisiones cuando el valor real es θ^*, el costo de equivocarse es

$$C(\theta_0, \theta^*) = (\theta_0 - \theta^*)'B(\theta_0 - \theta^*),$$

donde B es una matriz de constantes. El investigador no conoce el verdadero valor de θ, pero tiene creencias acerca de su valor representadas en $K(\theta|Y)$. Por ello, el investigador puede calcular el costo esperado de equivocarse al usar el valor θ_0. Este costo previsto es

$$EC(\theta_0) = \int C(\theta_0, \theta^*)K(\theta|Y)d\theta =$$

$$= \int (\theta_0 - \theta^*)'B(\theta_0 - \theta^*)K(\theta|Y)d\theta.$$

El valor de θ_0 que minimiza este costo esperado se determina derivando $EC(\theta_0)$ e igualando a cero, y resolviendo para θ_0. La derivada es

$$\frac{\partial EC(\theta_0)}{\partial \theta_0} = \int \frac{\partial[(\theta_0 - \theta)'B(\theta_0 - \theta)]}{\partial \theta_0} K(\theta|Y)d\theta$$

$$= \int 2(\theta_0 - \theta)'BK(\theta|Y)d\theta$$

$$= 2\theta_0'B \int K(\theta|Y)d\theta - 2\left(\int \theta K(\theta|Y)d\theta\right)' B$$

$$= 2\theta_0'B - 2\bar{\theta}'B.$$

Al igualar esta expresión a cero y despejar para θ_0, tenemos que

$$2\theta_0'B - 2\bar{\theta}'B = 0,$$

$$\theta_0'B = \bar{\theta}'B,$$

$$\theta_0 = \bar{\theta}.$$

La media de la distribución posterior, $\bar{\theta}$, es el valor de θ que un investigador bayesiano elegiría como óptimo si el costo de equivocarse acerca de θ creciese de forma cuadrática con la distancia al verdadero valor de θ.

Zellner (1971) describe el estimador bayesiano óptimo considerando otras funciones de costo (o de pérdida). Aunque la función de costo se asume generalmente como simétrica y sin límites, como la función cuadrática, no tiene por qué ser así; véase, por ejemplo, Wen y Levy (2001). Por su parte, Bickel y Doksum (2000) muestran que la correspondencia que se describe en la siguiente sección entre la media de la distribución posterior y el estimador de máxima verosimilitud aplica también a estimadores bayesianos que son óptimos considerando muchas otras funciones de costo.

12.2.2 Propiedades clásicas de $\bar{\theta}$: El teorema de Bernstein-von Mises

La estadística clásica no se preocupa de las creencias del investigador y no contempla la noción de distribución a priori y a posteriori. La preocupación de la estadística clásica es determinar la distribución muestral de un estimador. Esta distribución refleja el hecho de que una muestra diferente produciría una estimación puntual diferente. La distribución muestral es la distribución de las estimaciones puntuales que se obtendrían si se tomasen muchas muestras diferentes. Por lo general, la distribución muestral de un estimador no puede calcularse para muestras pequeñas. Sin embargo, la distribución muestral asintótica sí que suele ser posible calcularla, la cual aproxima la distribución muestral real cuando el tamaño de la muestra es lo suficientemente grande. En la estadística clásica, la distribución muestral asintótica determina las propiedades del estimador, tales como su consistencia, normalidad asintótica y eficiencia. La varianza de la distribución asintótica proporciona los errores estándar de las estimaciones y permite el test de hipótesis, cuya precisión aumenta con el tamaño de la muestra.

Desde una perspectiva clásica, $\bar{\theta}$ es simplemente un estadístico como cualquier otro. Su fórmula, dada en (12.3), existe y se puede aplicar incluso si el investigador no interpreta la fórmula como la media de una distribución posterior. El investigador puede considerar $K(\theta|Y)$ como una función definida por la ecuación (12.2) para cualquier $k(\theta)$ definida arbitrariamente que cumpla los requisitos de una densidad de probabilidad. La pregunta relevante para el investigador clásico es la misma que se haría para cualquier estadístico: ¿cuál es la distribución muestral de $\bar{\theta}$?

La respuesta a esta pregunta viene dada por el teorema de Bernstein-von Mises. Este teorema tiene una larga historia y toma múltiples formas. En el siglo XIX, Laplace (1820) observó que las distribuciones posteriores empezaban a parecerse cada vez más a la distribución normal a medida que el tamaño de la muestra aumentaba. Con los años, numerosas versiones de esta observación se han demostrado en diversas condiciones, y sus consecuencias han sido explicadas con más profundidad. Rao (1987), Le Cam y Yang (1990), Lehmann y Casella (1998), y Bickel y Doksum (2000) proporcionan enfoques modernos sobre esta cuestión, con notas históricas. El teorema lleva el nombre de Bernstein (1917) y von Mises (1931) ya que al parecer fueron ellos los primeros en proporcionar una prueba formal de la observación de Laplace, aunque bajo supuestos restrictivos que otros más tarde han relajado.

Describiré a continuación el teorema a través de tres declaraciones relacionadas. En estas declaraciones, la matriz de información, que hemos utilizado ampliamente en los capítulos 8 y 10, juega un papel importante. Recordemos que la puntuación de una observación es la pendiente del logaritmo de la verosimilitud de esa observación respecto a los parámetros:$s_n = \partial \ln P(y_n|\theta) / \partial\theta$, donde $P(y_n|\theta)$ es la probabilidad de las elecciones observadas del decisor n. La matriz de información, $-H$, es el negativo de la esperanza de la derivada de la puntuación, evaluada en los valores verdaderos de los parámetros:

$$-\boldsymbol{H} = -E\left(\frac{\partial^2 ln\, P(y_n|\theta^*)}{\partial\theta\partial\theta'}\right),$$

donde la esperanza es relativa a la población. (Se toma el negativo de modo que la matriz de información pueda ser definida positiva, como una matriz de covarianza). Recordemos también que el estimador de máxima verosimilitud tiene una varianza asintótica igual a $(-\boldsymbol{H})^{-1}/N$. Es decir, $\sqrt{N}(\theta^* - \hat{\theta}) \xrightarrow{d} N(0, (-\boldsymbol{H})^{-1})$, de modo que $\hat{\theta} \sim^{a} N(\theta^*, (-\boldsymbol{H})^{-1}/N)$, donde $\hat{\theta}$ es el estimador de máxima verosimilitud.

Ahora podemos facilitar las tres declaraciones que, en conjunto, constituyen el teorema de Bernstein-von Mises:

1. $\sqrt{N}(\theta - \bar{\theta}) \xrightarrow{d} N(0, (-\boldsymbol{H})^{-1})$.

   Dicho de forma intuitiva, la distribución posterior de θ converge a una distribución normal con varianza $(-\boldsymbol{H})^{-1}/N$ a medida que el tamaño de la muestra crece. Respecto al uso de la expresión $\xrightarrow{d}$ en este contexto, es importante tener en cuenta que la distribución que está convergiendo es la distribución posterior de $\sqrt{N}(\theta - \bar{\theta})$ en lugar de la distribución muestral. En el análisis clásico de estimadores, como se observa en el capítulo 10, la notación $\xrightarrow{d}$ se usa para indicar que la distribución muestral está convergiendo. El análisis bayesiano examina la distribución posterior en lugar de la distribución muestral, y la notación indica que la distribución posterior está convergiendo.

   Los puntos relevantes a destacar de esta primera declaración son que, a medida que el tamaño de la muestra crece, (i) la distribución posterior se vuelve normal y (ii) la varianza de la distribución posterior se convierte en la misma varianza del estimador de máxima verosimilitud. Estos dos puntos son relevantes para las próximas dos declaraciones.

2. $\sqrt{N}(\bar{\theta} - \hat{\theta}) \xrightarrow{p} 0$.

   La media de la distribución posterior converge al máximo de la función de verosimilitud. Estamos haciendo una declaración aún más fuerte. La diferencia entre la media de la distribución posterior y el máximo de la función de verosimilitud desaparece asintóticamente, *incluso cuando* la diferencia se escala por un factor $\sqrt{N}$.

   Intuitivamente este resultado tiene sentido, dado el primer resultado. Dado que la distribución posterior converge a una normal, y la media y el valor máximo son los mismos para una distribución normal, la media de la distribución posterior se convierte en el máximo de la distribución posterior. Además, el efecto de la distribución a priori en la distribución posterior desaparece a medida que crece el tamaño de la muestra (siempre y cuando la distribución a priori no sea cero en los alrededores del valor verdadero, claro está). Por tanto, la distribución posterior es proporcional a la función de verosimilitud para tamaños de muestra suficientemente grandes. El máximo de la función de verosimilitud se convierte en el mismo máximo de la distribución posterior, que, como se ha dicho, es también la media. Dicho sucintamente: dado que la distribución posterior es asintóticamente normal, de modo que su media es igual a su valor máximo, y la distribución posterior es proporcional a la función de verosimilitud asintóticamente, la diferencia entre $\bar{\theta}$ y $\hat{\theta}$ eventualmente desaparece.

3. $\sqrt{N}(\bar{\theta} - \theta^*) \xrightarrow{d} N(0, (-\boldsymbol{H})^{-1})$.

   La media de la distribución posterior, considerado como un estimador clásico, es asintóticamente equivalente al estimador de máxima verosimilitud. Es decir, $\bar{\theta} \sim^{a} N(\theta^*, (-\boldsymbol{H})^{-1}/N)$, al igual que el estimador de máxima verosimilitud. Tenga en cuenta que, dado que ahora estamos hablando en términos clásicos, la notación se refiere a la distribución muestral de $\bar{\theta}$, igual que lo haríamos para cualquier estimador.

Esta tercera declaración es una implicación de las dos primeras. El estadístico $\sqrt{N}(\bar{\theta} - \theta^*)$ puede re-escribirse como

$$\sqrt{N}(\bar{\theta} - \theta^*) = \sqrt{N}\left(\hat{\theta} - \theta^*\right) + \sqrt{N}(\bar{\theta} - \hat{\theta})$$

Gracias a la declaración 2, sabemos que $\sqrt{N}(\bar{\theta} - \hat{\theta}) \xrightarrow{p} 0$, por lo que el segundo término desaparece asintóticamente. Sólo el primer término afecta a la distribución asintótica. Este primer término es el estadístico que define el estimador de máxima verosimilitud $\hat{\theta}$. Hemos demostrado en el capítulo 10 que $\sqrt{N}\left(\hat{\theta} - \theta^*\right) \xrightarrow{d} N(0, (-\boldsymbol{H})^{-1})$. Por lo tanto, el estadístico $\sqrt{N}(\bar{\theta} - \theta^*)$ sigue la misma distribución asintótica. Básicamente, dado que $\bar{\theta}$ y $\hat{\theta}$ convergen, sus distribuciones muestrales asintóticas son las mismas.

El teorema de Bernstein-von Mises establece que $\bar{\theta}$ sigue las mismas pautas, en términos clásicos, de $\hat{\theta}$. En lugar de maximizar la función de verosimilitud, el investigador puede calcular la media de la distribución posterior sabiendo que el estimador resultante es tan bueno en términos clásicos como la máxima verosimilitud.

El teorema también proporciona un procedimiento para la obtención de los errores estándar de las estimaciones. La declaración 1 afirma que, asintóticamente, la varianza de la distribución posterior es $(-\boldsymbol{H})^{-1}/N$, la cual, según la declaración 3, es la varianza muestral asintótica del estimador $\bar{\theta}$. La varianza de la distribución posterior es la varianza asintótica de las estimaciones. El investigador puede realizar la estimación íntegramente mediante el uso de momentos estadísticos de la distribución posterior: la media de la distribución posterior proporciona las estimaciones puntuales, y la desviación estándar de la distribución posterior proporciona los errores estándar.

En la aplicación a casos reales, la media posterior y el máximo de la función de verosimilitud pueden diferir cuando el tamaño de la muestra es insuficiente para lograr la convergencia asintótica. Huber y Train (2001) encontraron que ambos son muy similares en su caso práctico, mientras que Ainslie et al. (2001) encontraron que son lo suficientemente diferentes como para justificar su consideración. Cuando las dos estimaciones no son similares, es necesario usar otros criterios para elegir entre ellas (si elegir es necesario), ya que sus propiedades asintóticas son las mismas.

12.3 Simulación de la media posterior

Para calcular la media de la distribución posterior, generalmente se requieren procedimientos de simulación. Como se mencionó anteriormente, la media es

$$\bar{\theta} = \int \theta K(\theta|Y) d\theta.$$

Una aproximación simulada de esta integral se obtiene extrayendo valores al azar de θ a partir de la distribución posterior y promediando los resultados. La media simulada es

$$\breve{\theta} = \frac{1}{R}\sum_{r=1}^{R} \theta^r,$$

donde θ^r es la extracción al azar r-ésima de $K(\theta|Y)$. La desviación estándar de la distribución posterior, que sirve como error estándar de las estimaciones, se simula calculando la desviación estándar de los R valores extraídos.

Como se ha dicho anteriormente, $\bar{\theta}$ tiene las mismas propiedades asintóticas que el estimador de máxima verosimilitud $\hat{\theta}$. ¿Cómo afecta el uso de la simulación en la aproximación de $\bar{\theta}$ a sus propiedades como estimador? Para la máxima verosimilitud simulada (MSL), vimos que el número de valores al azar utilizados en la simulación debe aumentar más rápidamente que la raíz cuadrada del tamaño de la muestra para que el estimador sea asintóticamente equivalente a la máxima verosimilitud. Con un número fijo de valores al azar, el estimador MSL es inconsistente. Si el número de valores al azar aumenta con el tamaño de la muestra, pero a un ritmo más lento que la raíz cuadrada del tamaño de la muestra, MSL es consistente pero no asintóticamente normal o eficiente. Como veremos, las propiedades que nos gustaría que tuviese como estimador la media simulada de la distribución posterior (*simulated mean of the posterior*, SMP) se alcanzan con unas condiciones más laxas relativas al número de valores al azar empleados. En particular, el estimador SMP es consistente y asintóticamente normal para un número fijo de valores al azar, y se convierte en eficiente y equivalente a la máxima verosimilitud si el número de valores crece en cualquier proporción con el tamaño de la muestra.

Para demostrar estas propiedades, examinaremos el estadístico normalizado $\sqrt{N}(\check{\theta} - \theta^*)$. Este estadístico puede reescribirse como

$$\sqrt{N}(\check{\theta} - \theta^*) = \sqrt{N}(\bar{\theta} - \theta^*) + \sqrt{N}(\check{\theta} - \bar{\theta}).$$

Gracias a la declaración 3 del teorema de Bernstein-von Mises, sabemos que la distribución límite del primer término: $\sqrt{N}(\bar{\theta} - \theta^*) \xrightarrow{d} N(0, (-\boldsymbol{H})^{-1}/N)$. El teorema del límite central nos da la distribución límite del segundo término. $\check{\theta}$ es el promedio de R extracciones al azar de una distribución con media $\bar{\theta}$ y varianza $(-\boldsymbol{H})^{-1}/N$. Suponiendo que los valores extraídos al azar son independientes, el teorema del límite central establece que el promedio de estos R valores se distribuye con media $\bar{\theta}$ y varianza $(-\boldsymbol{H})^{-1}/(RN)$. Al introducir esta información en el segundo término, tenemos $\sqrt{N}(\check{\theta} - \bar{\theta}) \xrightarrow{d} N(0, (-\boldsymbol{H})^{-1}/R)$. Los dos términos son independientes por cómo se construyen, por lo que

$$\sqrt{N}(\check{\theta} - \theta^*) \xrightarrow{d} N\left(0, \left(1 + \frac{1}{R}\right)(-\boldsymbol{H})^{-1}\right).$$

La media simulada de la distribución posterior es consistente y asintóticamente normal para un R fijo. La covarianza se infla por un factor $1/R$ debido a la simulación; sin embargo, la matriz de covarianza se puede calcular, y por lo tanto los errores estándar. Asimismo, es posible lleva a cabo test de hipótesis teniendo en cuenta el ruido de simulación.

Si R crece con N en cualquier proporción, el segundo término desaparece asintóticamente. Tenemos

$$\sqrt{N}(\check{\theta} - \theta^*) \xrightarrow{d} N(0, (-\boldsymbol{H})^{-1}),$$

que es igual a la media $\bar{\theta}$ real (no simulada) y al estimador de máxima verosimilitud $\hat{\theta}$. Cuando R crece con N, $\check{\theta}$ es asintóticamente eficiente y equivalente a la máxima verosimilitud.

Es necesario aclarar dos cosas respecto a los resultados anteriores. En primer lugar, se ha supuesto que los valores extraídos al azar para la simulación de la distribución posterior son independientes. En las secciones siguientes se describen métodos para la extracción de valores de la distribución posterior que se traducen en valores que exhiben un tipo de correlación en serie. Cuando se utilizan valores de este tipo, la varianza de la media simulada se infla en más de un factor $1/R$. El estimador sigue siendo consistente y asintóticamente normal con un número fijo de valores no independientes; simplemente su covarianza es mayor. Y si R se eleva con N, la covarianza adicional debida a la simulación desaparece

asintóticamente incluso con valores no independientes, de tal manera que la media simulada es asintóticamente equivalente a la máxima verosimilitud.

En segundo lugar, hemos supuesto que los valores extraídos al azar de la distribución posterior se pueden obtener sin necesidad de simular las probabilidades de elección. Para algunos modelos, extraer un valor al azar de la distribución posterior requiere simular las probabilidades de elección en que se basa la distribución posterior. En este caso, la media simulada de la distribución posterior implica hacer simulación dentro de la simulación, y la fórmula de su distribución asintótica es más compleja. Sin embargo, veremos que para la mayoría de modelos, incluyendo todos los modelos que consideramos en este libro, es posible extraer valores al azar de la distribución posterior sin simular las probabilidades de elección. Una de las ventajas de los procedimientos bayesianos es que normalmente evitan la necesidad de simular probabilidades de elección.

12.4 Extracción de valores al azar de la distribución posterior

Por lo general, la distribución posterior no tiene una forma muy conveniente para poder extraer valores al azar de ella. Por ejemplo, sabemos cómo extraer valores al azar fácilmente de una distribución normal conjunta no truncada, sin embargo, es raro que la distribución posterior tome esta forma para todo el vector de parámetros. El muestreo por importancia (*importance sampling*), que se describe en la sección 9.2.7 en relación a cualquier densidad, puede ser útil para la simulación de estadísticos sobre la distribución posterior. Geweke (1992, 1997) describe cómo afrontar el problema respecto a distribuciones posteriores y proporciona una guía práctica sobre la selección apropiada de una densidad propuesta dentro del procedimiento definido por el muestreo por importancia. Otros dos métodos que hemos descrito en el capítulo 9 son particularmente útiles para la extracción de valores al azar de una distribución posterior: el muestreo de Gibbs y el algoritmo de Metropolis-Hasting. Estos métodos a menudo son llamados métodos de Monte Carlo – Cadena de Markov (*Markov Chain Monte Carlo methods* , MCMC). Formalmente, el muestreo de Gibbs es un tipo especial de algoritmo Metropolis-Hasting (Gelman, 1992). Sin embargo, el caso es tan especial, y por lo tanto conceptualmente sencillo, que el término Metropolis-Hasting (MH) generalmente se reserva para versiones más complejas que el muestreo de Gibbs. Es decir, cuando el algoritmo MH es el muestreo de Gibbs, suele ser referido como muestreo de Gibbs, y cuando es más complejo que el muestreo de Gibbs, se suele referir como algoritmo MH. Mantengo esta convención de aquí en adelante.

Será de gran utilidad para el lector revisar las secciones 9.2.8 y 9.2.9, que describen el muestreo de Gibbs y el algoritmo MH, ya que vamos a utilizar estos procedimientos ampliamente en el resto de este capítulo. Como hemos visto anteriormente, la media de la distribución posterior se simula extrayendo valores al azar de dicha distribución posterior y promediando los valores. En lugar de extraer valores de la distribución posterior multidimensional para todos los parámetros, el muestreo de Gibbs permite al investigador extraer valores de un parámetro cada vez (o un subconjunto de parámetros), condicionando a los valores del resto de parámetros (Casella y George, 1992). Extraer valores al azar de la distribución posterior para un parámetro condicionado al resto de parámetros suele ser mucho más fácil que extraer valores de todos los parámetros simultáneamente.

En algunos casos, se necesita usar el algoritmo MH en conjunción con el muestreo de Gibbs. Supongamos, por ejemplo, que la distribución posterior de un parámetro condicionado al resto de parámetros no toma una forma simple. En este caso, puede usarse el algoritmo MH, ya que es aplicable a (prácticamente) cualquier distribución (Chib y Greenberg, 1995).

El algoritmo MH es particularmente útil en relación a las distribuciones posteriores porque no es necesario calcular la constante de normalización de dicha distribución. Recordemos que la distribución posterior es el producto de la distribución a priori por la función de verosimilitud, dividido por una constante de normalización que asegura que la integral de la distribución posterior es igual a uno:

$$K(\theta|Y) = \frac{L(Y|\theta)k(\theta)}{L(Y)},$$

donde $L(Y)$ es la constante de normalización

$$L(Y) = \int L(Y|\theta)k(\theta)d\theta.$$

Esta constante puede ser difícil de calcular, ya que implica integración. Como se describe en la sección 9.2.9, el algoritmo MH se puede aplicar sin necesidad de conocer o calcular la constante de normalización de la distribución posterior.

En resumen, el muestreo de Gibbs, combinado si es necesario con el algoritmo MH, permite extraer valores al azar de un vector de parámetros a partir de la distribución posterior para prácticamente cualquier modelo. Estos procedimientos se aplican a un modelo logit mixto en la Sección 12.6. Sin embargo, en primer lugar vamos a obtener la distribución posterior de algunos modelos muy simples. Como veremos, estos resultados se aplican a menudo en modelos más complejos para un subconjunto de los parámetros. Este hecho facilita el muestreo de Gibbs sobre estos parámetros.

12.5 Distribuciones posteriores de la media y la varianza de una distribución normal

La distribución posterior toma una forma muy conveniente para algunos procesos de inferencia simples. Describiremos dos de estas situaciones que, como veremos, a menudo surgen dentro de modelos más complejos para un subconjunto de los parámetros. Ambos resultados se refieren a la distribución normal. Consideremos en primer lugar la situación en la que se conoce la varianza de una distribución normal, pero no su media. Pasamos luego a considerar la media como el parámetro conocido, pero no la varianza. Por último, combinando estas dos situaciones con el muestreo de Gibbs, consideraremos la situación en que tanto la media como la varianza son desconocidas.

12.5.1 Resultado A: Media desconocida, varianza conocida

Expondremos en primer lugar el caso para una sola dimensión, y luego generalizamos a múltiples dimensiones. Considere una variable aleatoria β que se distribuye normalmente con media desconocida b y varianza conocida σ. El investigador observa una muestra de N realizaciones de la variable aleatoria, etiquetadas $\beta_n, n = 1, \ldots, N$. La media de la muestra es $\bar{\beta} = (1/N)\sum_n \beta_n$. Supongamos que la distribución a priori del investigador acerca de b es $N(b_0, s_0)$, es decir, las creencias a priori del investigador están representadas por una distribución normal con media b_0 y varianza s_0. Tenga en cuenta que ahora tenemos dos distribuciones normales: la distribución de β, que tiene media b, y la distribución a priori sobre esta media desconocida, que tiene media b_0. La distribución a priori indica que el investigador cree que lo más probable es que $b = b_0$ y que también piensa que hay una probabilidad del 95 por ciento de que b se encuentre en algún lugar entre $b_0 - 1.96\sqrt{s_0}$ y $b_0 + 1.96\sqrt{s_0}$. Considerando esta distribución a priori, la distribución posterior de b es $N(b_1, s_1)$, donde

$$b_1 = \frac{\frac{1}{s_0}b_0 + \frac{N}{\sigma}\bar{\beta}}{\frac{1}{s_0} + \frac{N}{\sigma}}$$

Y

$$s_1 = \frac{1}{\frac{1}{s_0} + \frac{N}{\sigma}}.$$

La media posterior b_1 es la media ponderada de la media de la muestra y la media a priori.

Prueba: La distribución a priori es

$$k(b) = \frac{1}{\sqrt{2\pi s_0}} e^{-(b-b_0)^2/2s_0}.$$

La probabilidad de extraer al azar el valor β_n de $N(b, \sigma)$ es

$$\frac{1}{\sqrt{2\pi\sigma}} e^{-(b-\beta_n)^2/2\sigma},$$

por lo que la verosimilitud de los N valores extraídos es

$$L(\beta_n \ \forall n \mid b) = \prod_n \frac{1}{\sqrt{2\pi\sigma}} e^{-(b-\beta_n)^2/2\sigma}$$

$$= \frac{1}{(2\pi\sigma)^{N/2}} e^{-\sum(b-\beta_n)^2/2\sigma}$$

$$= \frac{1}{(2\pi\sigma)^{N/2}} e^{(-N\bar{s}-N(b-\bar{\beta})^2)/2\sigma}$$

$$= \frac{1}{(2\pi\sigma)^{N/2}} e^{-N\bar{s}/2\sigma} e^{-N(b-\bar{\beta})^2/2\sigma},$$

donde $\bar{s} = (1/N)\sum(\beta_n - \bar{\beta})^2$ es la varianza muestral de las β_n. Por tanto, la distribución posterior es

$$K(b|\beta_n \ \forall n) \propto L(\beta_n \ \forall n \mid b) k(b)$$

$$= \frac{1}{(2\pi\sigma)^{N/2}} e^{-N\bar{s}/2\sigma} e^{-N(b-\bar{\beta})^2/2\sigma} \times \frac{1}{\sqrt{2\pi s_0}} e^{-(b-b_0)^2/2s_0}$$

$$= m_1 e^{-[N(b-\bar{\beta})^2/2\sigma]-[(b-b_0)^2/2s_0]},$$

donde m_1 es una constante que contiene todos los términos multiplicativos que no dependen de b. Con un poco de manipulación algebraica, tenemos

$$K(b|\beta_n \ \forall n) \propto e^{-[N(b-\bar{\beta})^2/2\sigma]-[(b-b_0)^2/2s_0]}$$

$$\propto e^{(b^2-2b_1b)/2s_1}$$

$$\propto e^{(b-b_1)^2/2s_1}.$$

El segundo $\propto$ suprime $\bar{\beta}^2$ y b_0^2 de la exponencial, ya que no dependen de b y por lo tanto sólo afectan a la constante de normalización. (Recordemos que $\exp(a+b)=\exp(a)\exp(b)$, por lo que añadir y eliminar términos de la exponencial tiene un efecto multiplicativo sobre $K(b|\beta_n\ \forall n)$. La tercera $\propto$ añade $b_1\bar{\beta}$ al exponente, que tampoco depende de b. Por tanto, la distribución posterior es

$$K(b|\beta_n\ \forall n) = me^{(b-b_1)^2/2s_1},$$

donde m es la constante de normalización. Esta fórmula es la densidad normal con media b_1 y varianza s_1.

Como se ha indicado anteriormente, la media de la distribución posterior es un promedio ponderado de la media de la muestra y la media a priori. El peso aplicado a la media de la muestra se eleva a medida que aumenta el tamaño de la muestra, de modo que para un N suficientemente grande, la media a priori pasa a ser irrelevante.

A menudo, el investigador querrá especificar una distribución a priori que contenga muy poca información acerca de los parámetros antes de observar la muestra. En general, la incertidumbre del investigador se refleja en la varianza de la distribución a priori. Una varianza grande significa que el investigador tiene una idea muy vaga sobre el valor del parámetro. Dicho de forma equivalente, una distribución a priori casi plana significa que el investigador considera que todos los valores posibles de los parámetros son igualmente probables. Una distribución a priori que contiene poca información se llama *difusa*.

Podemos examinar el efecto de una distribución a priori difusa en la distribución posterior de b. Al incrementar la varianza de la distribución a priori, s_0, la normal a priori se hace más extendida y plana. A medida que $s_0 \to \infty$, representando una distribución a priori cada vez más difusa, la distribución posterior se acerca $N(\bar{\beta}, \sigma/N)$.

Las versiones multivariadas de este resultado son similares. Considere un vector K-dimensional aleatorio $\beta \sim N(b, W)$ con W conocida y b desconocida. El investigador observa una muestra $\beta_n, n = 1, \ldots, N$, cuya media muestral es $\bar{\beta}$. Si la distribución a priori del investigador sobre b es difusa (normal con una varianza ilimitadamente grande), entonces la distribución posterior es $N(\bar{\beta}, W/N)$.

Extraer valores al azar de esta distribución posterior es fácil. Sea L el factor Choleski de W/N. Extraiga K valores al azar de variables aleatorias normales estándar iid, $\eta_i, i = 1, \ldots, K$, y agrúpelos en un vector $\eta = \langle \eta_1, \ldots, \eta_K \rangle'$. Calcule $\tilde{b} = \bar{\beta} + L\eta$. El vector resultante $\tilde{b}$ es un valor al azar de $N(\bar{\beta}, W/N)$.

12.5.2 Resultado B: Varianza desconocida, media conocida

Considere una variable aleatoria (unidimensional) que se distribuye normalmente con media conocida b y varianza desconocida σ. El investigador observa una muestra de N realizaciones, etiquetadas $\beta_n, n = 1, \ldots, N$. La varianza muestral alrededor de la media conocida es $\bar{s} = (1/N)\sum_n (\beta_n - b)^2$. Supongamos que la distribución a priori sobre σ del investigador es una gamma invertida con v_0 grados de libertad y escala s_0. Esta distribución a priori se denota como $IG(v_0, s_0)$. La densidad es igual a cero para cualquier valor negativo de σ, lo que refleja el hecho de que una varianza debe ser positiva. La moda de la distribución a priori tipo gamma invertida es $s_0 v_0/(1+v_0)$. Usando una gamma invertida como distribución a priori, la distribución posterior de σ es también una gamma invertida $IG(v_1, s_1)$, donde

$$v_1 = v_0 + N,$$

$$s_1 = \frac{v_0 s_0 + N\bar{s}}{v_0 + N}.$$

Prueba: Una gamma invertida con v_0 grados de libertad y escala s_0 tiene una densidad

$$k(\sigma) = \frac{1}{m_0 \sigma^{(v_0/2)+1}} e^{-v_0 s_0/2\sigma},$$

donde m_0 es la constante de normalización. La verosimilitud de la muestra, tratada como una función de σ, es

$$L(\beta_n \ \forall n \mid \sigma) = \frac{1}{(2\pi\sigma)^{N/2}} e^{-\sum(b-\beta_n)^2/2\sigma} = \frac{1}{(2\pi\sigma)^{N/2}} e^{-N\bar{s}/2\sigma}.$$

La distribución posterior es entonces

$$K(\sigma|\beta_n \ \forall n) \propto L(\beta_n \ \forall n \mid \sigma) k(\sigma)$$

$$\propto \frac{1}{\sigma^{N/2}} e^{-N\bar{s}/2\sigma} \times \frac{1}{\sigma^{(v_0/2)+1}} e^{-v_0 s_0/2\sigma}$$

$$= \frac{1}{\sigma^{((N+v_0)/2)+1}} e^{-(N\bar{s}+v_0 s_0)/2\sigma}$$

$$= \frac{1}{\sigma^{(v_1/2)+1}} e^{-v_1 s_1/2\sigma},$$

que es la densidad gamma invertida con v_1 grados de libertad y escala s_1.

La distribución gamma invertida a priori se vuelve más difusa con una v_0 menor. Para que la integral de la densidad sea uno y tenga una media, v_0 debe ser mayor que 1. Es habitual establecer $s_0 = 1$ cuando se especifica $v_0 \to \infty$. En virtud de esta distribución a priori difusa, la distribución posterior se convierte en $IG(1+N, (1+N\bar{s})/(1+N))$. La moda de esta distribución posterior es $(1+N\bar{s})/(2+N)$, que es aproximadamente la varianza de la muestra $\bar{s}$ para N grandes.

El caso multivariado es similar. La generalización multivariado de una distribución gamma invertida es la distribución Wishart invertida. El resultado en el caso multivariado es el mismo que con una única variable aleatoria, excepto que la gamma invertida se sustituye por la Wishart invertida.

Un vector aleatorio K-dimensional $\beta \sim N(b, W)$ tiene una b conocida pero una W desconocida. Una muestra de tamaño N de esta distribución tiene una varianza alrededor de la media conocida de $\bar{S} = (1/N)\sum_n (\beta_n - b)(\beta_n - b)'$. Si la distribución a priori sobre W del investigador es una Wishart invertida con v_0 grados de libertad y matriz de escala S_0, etiquetada $IW(v_0, S_0)$, entonces la distribución posterior de W es $IW(v_1, S_1)$ donde

$$v_1 = v_0 + N,$$

$$S_1 = \frac{v_0 S_0 + N\bar{S}}{v_0 + N},$$

La distribución a priori se vuelve más difusa con una v_0 menor, aunque v_0 debe ser mayor que K para que la distribución a priori integre a uno y tenga media. Con $S_0 = I$, donde I es la matriz identidad K-dimensional, la distribución posterior bajo una distribución a priori difusa se convierte en $IW(K + N, (KI + N\bar{S})/(K + N))$. Conceptualmente, la distribución a priori es equivalente a que el investigador tenga una muestra previa de K observaciones cuya varianza muestral fuese I. A medida que N aumenta sin límite, la influencia de la distribución a priori en la distribución posterior va desapareciendo.

Es fácil extraer valores al azar de una distribución gamma invertida y de una distribución Wishart invertida. Consideremos en primer lugar una gamma invertida $IG(v_1, s_1)$. Para extraer valores aleatorios de la misma procederíamos como sigue:

1. Extraiga v_1 valores al azar de una normal estándar y etiquete los valores como $\eta_i, i = 1, \ldots, v_1$.
2. Divida cada valor por $\sqrt{s_1}$, eleve al cuadrado el resultado y tome la media. Es decir, calcule $r = (1/v_1)\sum_i(\sqrt{1/s_1}\eta_i)^2$, que es la varianza muestral de v_1 valores extraídos al azar de una distribución normal cuya varianza es $1/s_1$.
3. Calcule la inversa de r: $\tilde{s} = 1/r$ es un valor al azar extraído de la gamma invertida.

Puede extraer valores al azar de una Wishart K-dimensional $IW(v_1, S_1)$ de la siguiente manera:

1. Extraiga al azar v_1 vectores K-dimensionales cuyos elementos sean variables normales estándar independientes. Etiquete los valores como $\eta_i, i = 1, \ldots, v_1$.
2. Calcule el factor Choleski de la inversa de S_1, etiquétela como L, donde $LL' = S_1^{-1}$.
3. Calcule $R = (1/v_1)\sum_i(L\eta_i)(L\eta_i)'$. Observe que R es la varianza de los valores extraídos al azar de una distribución con varianza S_1^{-1}.
4. Calcule la inversa de R: $\tilde{S} = R^{-1}$ es un valor al azar extraído de $IW(v_1, S_1)$.

12.5.3 Media y varianza desconocidas

Suponga que tanto la media b como la varianza W son desconocidas. Para ninguno de estos parámetros la distribución posterior tiene una forma conveniente. Sin embargo, pueden extraerse valores al azar fácilmente utilizando el muestreo de Gibbs y los resultados A y B anteriormente obtenidos. Un valor al azar de b se extraer condicionado a W, y luego un valor al azar de W se extrae condicionando a b. El resultado A dice que la distribución posterior de b condicionada a W es normal, de la cuál es fácil extraer valores al azar. El resultado B dice que la distribución posterior de W condicionada a b es una Wishart invertida, de la cual también es fácil extraer valores. Iterando numerosas veces a través de las distribuciones posteriores condicionadas nos proporciona, al final, valores de la distribución posterior conjunta.

12.6 Procedimiento bayesiano jerárquico para logit mixto

En esta sección se muestra cómo se pueden utilizar los procedimientos bayesianos para estimar los parámetros de un modelo logit mixto. Utilizaremos para ello el enfoque desarrollado por Allenby (1997), implementando por Sawtooth Software (1999), y generalizado en Train (2001). Sea la utilidad que una persona n obtiene de la alternativa j en el período de tiempo t

$$U_{njt} = \beta_n' x_{njt} + \varepsilon_{njt},$$

donde ε_{njt} es iid valor extremo y $\beta_n \sim N(b, W)$. Si asignamos a β_n una distribución normal podemos usar los resultados A y B anteriores, lo que acelera considerablemente la estimación. En la siguiente sección, se abordará el uso de distribuciones no normales.

El investigador tiene distribuciones a priori sobre b y W. Supongamos que la distribución a priori de b es normal con una varianza ilimitadamente grande. Supongamos que la distribución a priori de W se distribuye como una Wishart invertida con K grados de libertad y matriz de escala I, la matriz identidad K-dimensional. Observe que estas distribuciones a priori son las utilizadas en los resultados A y B. Se pueden especificar distribuciones a priori más flexibles para W, utilizando los procedimientos proporcionados por McCulloch y Rossi (2000), por ejemplo, aunque hacerlo hace que el muestreo de Gibbs resulte más complejo.

Observamos una muestra de N personas. Las alternativas elegidas en cada período de tiempo por la persona n se indican como $y_n' = \langle y_{n1}, \dots, y_{nT} \rangle$, y las elecciones de toda la muestra se etiquetan como $Y = \langle y_1, \dots, y_N \rangle$. La probabilidad de las elecciones observadas de la persona n, condicionadas a β, es

$$L(y_n|\beta) = \prod_t \left(\frac{e^{\beta' x_{n y_{nt} t}}}{\sum_j e^{\beta' x_{njt}}} \right).$$

La probabilidad *no* condicionada a β es la integral de $L(y_n|\beta)$ sobre todo β:

$$L(y_n|b, W) = \int L(y_n|\beta)\, \phi(\beta|b, W) d\beta,$$

donde $\phi(\beta|b, W)$ es la densidad normal con media b y varianza W. Esta $L(y_n|b, W)$ es la probabilidad de elección de un modelo logit mixto.

La distribución posterior de b y W es, por definición,

$$K(b, W|Y) \propto \prod_n L(y_n|b, W) k(b, W), \qquad (12.4)$$

donde $k(b, W)$ es la distribución a priori de b y W descrita anteriormente (es decir, el producto de una distribución normal para b y una Wishart invertida para W).

Sería *posible* extraer valores directamente de $K(b, W|Y)$ con el algoritmo MH. Sin embargo, hacerlo sería computacionalmente muy lento. Para cada iteración del algoritmo MH, sería necesario calcular el lado derecho de (12.4). Sin embargo, la probabilidad de elección $L(y_n|b, W)$ es una integral sin una forma cerrada y debe ser aproximada a través de la simulación. Por tanto, cada iteración del algoritmo MH requeriría simulación de $L(y_n|b, W)$ para cada n. Eso consumiría mucho tiempo, y las propiedades del estimador resultante se verían afectadas. Recordemos que las propiedades de la media simulada de la distribución posterior se obtuvieron bajo el supuesto de que los valores al azar se pueden extraer sin necesidad de simular las probabilidades de elección. Aplicar el algoritmo MH a (12.4) viola este supuesto.

Extraer valores al azar de $K(b, W|Y)$ se convierte en algo rápido y sencillo si cada β_n se considera como un parámetro junto a b y W, y usamos el muestreo de Gibbs para los tres conjuntos de parámetros b, W y $\beta_n \forall n$. La distribución posterior para b, W y $\beta_n \forall n$ es

$$K(b, W, \beta_n \forall n|Y) \propto \prod_n L(y_n|\beta_n) \phi(\beta_n|b, W) k(b, W).$$

Extraemos valores al azar de esta distribución posterior mediante el muestreo de Gibbs. Se extrae un valor al azar de cada parámetro, condicionando a los otros parámetros: (1) Extraiga un valor de b

condicionando a los valores de W y $\beta_n \forall n$. (2) Extraiga un valor de W condicionando a los valores de b y $\beta_n \forall n$. (3) Extraiga un valor de $\beta_n \forall n$ condicionando a los valores de b y W. Cada uno de estos pasos es fácil, como veremos más adelante. El paso 1 utiliza resultado A, que da la distribución posterior de la media dada la varianza. El paso 2 utiliza el resultado B, que da la distribución posterior de la varianza dada la media. El paso 3 utiliza el algoritmo MH, pero de una manera que no implica el uso de simulación dentro del algoritmo. Cada paso se describe a continuación:

1. $\mathrm{b}|\mathrm{W}, \beta_n \forall n$.

   En este paso condicionamos respecto a W y a la β_n de cada persona, lo que significa que tratamos estos parámetros como si se conocieran. El resultado A nos da la distribución posterior de b en estas condiciones. Las β_ns constituyen una muestra de N realizaciones de una distribución normal con media desconocida b y varianza W conocida. Dada nuestra distribución a priori difusa de b, la distribución posterior de b es $\mathrm{N}(\bar{\beta}, \mathrm{W/N})$, donde $\bar{\beta}$ es la media muestral de las β_ns. Se extrae un valor al azar de esta distribución posterior tal y como se describe en la sección 12.5.1.

2. $\mathrm{W}|\mathrm{b}, \beta_n \forall n$.

   El resultado B nos da la distribución posterior de W condicionada a b y a las β_ns. Las β_ns constituyen una muestra de una distribución normal con media b conocida y varianza W desconocida. Usando nuestra distribución a priori de W, la distribución posterior de W es una Wishart invertida con $\mathrm{K} + \mathrm{N}$ grados de libertad y matriz de escala $(\mathrm{KI} + \mathrm{NS}_1)/(\mathrm{K} + \mathrm{N})$, donde $\mathrm{S}_1 = (1/N)\sum_n(\beta_n - b)(\beta_n - b)'$ es la varianza muestral de las β_ns alrededor de la media conocida b. Se extrae un valor al azar de una distribución Wishart invertida como se describe en la sección 12.5.2.

3. $\beta_n|\mathrm{b}, \mathrm{W}$.

   La distribución posterior de la β_n de cada persona, condicionada respecto a sus elecciones y respecto a la media y varianza de β_n en la población, es

$$K(\beta_n|b, W, y_n) \propto L(y_n|\beta_n)\phi(\beta_n|b, W). \tag{12.5}$$

   No existe una manera simple de extraer valores al azar de esta distribución posterior, por lo que se utiliza el algoritmo MH. Tenga en cuenta que el lado derecho de (12.5) es fácil de calcular: $\mathrm{L}(y_n|\beta_n)$ es un producto de logits y $\phi(\beta_n|\mathrm{b}, \mathrm{W})$ es la densidad normal. El algoritmo MH funciona como sigue:

   (a) Comience con un valor inicial β_n^0.

   (b) Extraiga K valores independientes de una densidad normal estándar, y agrupe los valores en un vector etiquetado como η^1.

   (c) Cree un valor de prueba de β_n^1 como $\tilde{\beta}_n^1 = \beta_n^0 + \rho L \eta^1$, donde ρ es un escalar especificado por el investigador y L es el factor Choleski de W. Tenga en cuenta que la distribución propuesta del algoritmo MH (etiquetada $g(\cdot)$ en la sección 9.2.9) se especifica como normal con media cero y varianza $\rho^2 W$.

   (d) Extraiga un valor de una variable uniforme estándar μ^1.

   (e) Calcule el ratio

$$F = \frac{L(y_n|\tilde{\beta}_n^1)\phi(\tilde{\beta}_n^1|b, W)}{L(y_n|\beta_n^0)\phi(\beta_n^0|b, W)}$$

   (f) Si $\mu^1 \leq F$, acepte $\tilde{\beta}_n^1$ y defina $\beta_n^1 = \tilde{\beta}_n^1$. Si $\mu^1 > F$, rechace $\tilde{\beta}_n^1$ y defina dejar $\beta_n^1 = \beta_n^0$.

(g) Repita el proceso varias veces. Para un t suficientemente alto, β_n^t es un valor extraído al azar de la distribución posterior.

Ahora sabemos cómo extraer valores al azar de la distribución posterior para cada parámetro, condicionando sobre el resto de parámetros. Combinamos estos procedimientos en un muestreador de Gibbs para los tres conjuntos de parámetros. Comience con cualquier valor inicial de b^0, W^0 y $\beta_n^0 \forall n$. La iteración t–ésima del muestreador de Gibbs consta de los siguientes pasos:

1. Extraiga un valor b^t de $\mathrm{N}(\bar{\beta}^{t-1}, \mathrm{W}^{t-1}/\mathrm{N})$, donde $\bar{\beta}^{t-1}$ es la media de las β_n^{t-1}s.
2. Extraiga W^t de $\mathrm{IW}(\mathrm{K}+\mathrm{N}, (\mathrm{KI}+\mathrm{NS}^{t-1})/(\mathrm{K}+\mathrm{N}))$, donde $\mathrm{S}^{t-1} = \sum_n (\beta_n^{t-1} - \mathrm{b}^t)(\beta_n^{t-1} - \mathrm{b}^t)' / N$.
3. Para cada n, extraiga β_n^tusando una iteración del algoritmo MH descrito anteriormente, empezando por β_n^{t-1} y usando la densidad normal $\phi(\beta_n|\mathrm{b}^t, \mathrm{W}^t)$.

Estos tres pasos se repiten para muchas iteraciones. Los valores resultantes convergen a valores extraídos de la distribución posterior conjunta de b, W y $\beta_n \forall n$. Una vez se obtienen los valores convergentes de la distribución posterior, se puede calcular la media y la desviación estándar de los valores extraídos para obtener estimaciones y errores estándar de los parámetros. Tenga en cuenta que este procedimiento proporciona información acerca de las β_n para cada n, de forma similar al procedimiento descrito en el Capítulo 11 usando la estimación clásica.

Como se ha mencionado, el muestreador de Gibbs converge, usando suficientes iteraciones, a valores extraídos de la distribución posterior conjunta de todos los parámetros. Las iteraciones previas a la convergencia a menudo se llaman *burn-in* (quemado). Por desgracia, no siempre es fácil determinar cuándo se ha logrado la convergencia, como subraya Kass et al. (1998). Cowles y Carlin (1996) proporcionan una descripción de las diferentes pruebas y diagnósticos que se han propuesto. Por ejemplo, Gelman y Rubin (1992) sugieren comenzar el muestreo de Gibbs desde varios puntos diferentes y probar la hipótesis de que el estadístico de interés (en nuestro caso, la media posterior) es el mismo cuando se calcula a partir de cada una de las secuencias presumiblemente convergentes. A veces, la convergencia es bastante obvia, por lo que la prueba formal es innecesaria. Durante la fase de *burn-in*, el investigador normalmente podrá ver la tendencia de los valores, es decir, podrá ver cómo avanzan en dirección a la masa principal de la distribución posterior. Una vez se ha logrado la convergencia, los valores extraídos tienden a moverse alrededor de la distribución posterior.

Los valores extraídos mediante el muestreo de Gibbs está correlacionados entre iteraciones incluso cuando se ha logrado la convergencia, ya que cada iteración se basa en la anterior. Esta correlación no impide que los valores puedan ser utilizados para el cálculo de la media posterior y la desviación estándar, o cualquier otro estadístico. Sin embargo, el investigador puede reducir la cantidad de correlación entre valores mediante el uso de sólo una parte de los valores obtenidos después de la convergencia. Por ejemplo, el investigador podría retener uno de cada diez valores y descartar los otros, lo que reduce la correlación entre los valores retenidos en un factor 10. Por tanto, un investigador puede especificar un total de 20.000 iteraciones para obtener 1.000 valores: 10.000 para la fase de *burn-in* y 10.000 posteriores a la convergencia, de los cuales conserva uno de cada diez.

Queda pendiente una cuestión. En el algoritmo MH, el escalar ρ es especificado por el investigador. Este escalar determina el tamaño de cada salto dentro de la distribución. Por lo general, saltos más pequeños se traducen en más aceptaciones y saltos más grandes en menos. Sin embargo, usar saltos pequeños implica que el algoritmo MH necesitará más iteraciones para converger e implica más correlación serial entre valores una vez se alcanza la convergencia. Gelman et al. (1995. P 335) han estudiado la tasa de aceptación óptima para el algoritmo MH. Encontraron que la tasa óptima es de aproximadamente 0.44 cuando $K = 1$ y cae hasta 0.23 a medida que aumenta K. El investigador puede establecer el valor de ρ

para lograr una tasa de aceptación en torno a estos valores, bajando ρ para obtener una tasa de aceptación mayor y elevándolo para obtener una tasa de aceptación menor.

De hecho, ρ se puede ajustar como parte del proceso iterativo. El investigador establece el valor inicial de ρ. En cada iteración, un valor de prueba de β_n es aceptado o rechazado para cada n en la muestra. Si en una iteración, la tasa de aceptación entre las N observaciones está por encima de un valor determinado (por ejemplo, 0.33), entonces ρ se eleva. Si la tasa de aceptación es inferior a este valor, ρ se baja. Por lo tanto, el valor de ρ se altera durante el proceso de iteración para alcanzar el nivel de aceptación especificado.

12.6.1 Reformulación resumida

Una vez hemos descrito por completo los procedimientos bayesianos, el modelo y el muestreo de Gibbs se pueden expresar de manera sucinta, en la forma en que se utiliza en la mayoría de las publicaciones. El modelo es como sigue.

Utilidad:

$U_{njt} = \beta_n' x_{njt} + \varepsilon_{njt}$,

ε_{njt} iid valor extremo

$\beta_n \sim N(b, W)$.

Elección observada:

$y_{nt} = i$ si y solo si $U_{nit} > U_{njt}\ \forall j \neq i$.

Distribuciones a priori:

$k(b, w) = k(b)k(W)$,

donde

$k(b)$ es $N(b_0, W_0)$ con varianza extremadamente grande,

$k(W)$ es $IW(K, I)$.

Distribuciones posteriores condicionadas:

$$K(\beta_n|b, W, y_n) \propto \prod_t \frac{e^{\beta' x_{ny_{nt}t}}}{\sum_j e^{\beta' x_{njt}}} \phi(\beta_n|b, W)\ \forall n,$$

$K(b|W, \beta_n \forall n)$ es $N(\bar{\beta}, W/N)$, donde $\bar{\beta} = \sum_n \beta_n / N$,

$K(W|b, \beta_n \forall n)$ es $IW\left(K + N, \frac{KI + N\bar{S}}{K+N}\right)$, donde $\bar{S} = \sum_n (\beta_n - b)(\beta_n - b)' / N$,

Las tres distribuciones posteriores condicionadas se denominan capas (o niveles) del muestreo de Gibbs. La primera capa para cada n depende sólo de datos de esa persona y no de toda la muestra. Las segunda y tercera capas no dependen de los datos directamente, sólo de los valores extraídos de β_n, que a su vez dependen de los datos.

El muestreo de Gibbs para este modelo es rápido por dos razones. En primer lugar, ninguna de las capas requiere integración. En particular, la primera capa utiliza un producto de fórmulas logit para un valor dado de β_n. El procedimiento bayesiano evita la necesidad de calcular la probabilidad logit mixto, utilizando en su lugar logits simples condicionadas a un valor de β_n. En segundo lugar, las capas 2 y 3 no utilizan los datos en absoluto, ya que dependen sólo de los valores extraídos para $\beta_n \forall n$. En estas capas sólo es necesario calcular la media y la varianza de las β_ns.

El procedimiento a menudo se llama jerárquico bayesiano (*hierarchical Bayes*, HB), porque en él existe una jerarquía de parámetros. β_n son los *parámetros a nivel individual* para la persona n, los cuales

describen las preferencias de esa persona. Las β_ns se distribuyen en la población con media b y varianza W. Los parámetros b y W son llamados a menudo los *parámetros a nivel de población* o *hiper-parámetros*. También existe una jerarquía de distribuciones a priori. La distribución a priori de la β_n de cada persona es la densidad de β_n en la población. Esta distribución a priori tiene parámetros (hiper-parámetros), su media b y su varianza W, que a su vez tienen distribuciones a priori.

12.7 Caso de estudio: elección del proveedor de energía

Aplicamos los procedimientos bayesianos a los datos que se describen en el capítulo 11 sobre la elección que los clientes realizan entre proveedores de energía. Las estimaciones bayesianas son comparadas con las estimaciones obtenidas a través de máxima verosimilitud simulada (MSL).

A cada uno de los 361 clientes se le presentó un máximo de 12 situaciones de elección hipotéticas. En cada situación de elección se describían cuatro proveedores de energía y se requería al encuestado indicar cuál elegiría si se enfrentase a esas opciones de elección en el mundo real. Los proveedores se diferenciaban sobre la base de seis factores: (1) si el proveedor cobraba un precio fijo, y si era así, la tarifa en centavos de dólar por kilovatio-hora, (2) la duración del contrato en años, durante el cual se garantizaba la tarifa y el cliente debería pagar una penalización en caso de querer abandonar la compañía, (3) si el proveedor era la compañía eléctrica local, (4) si la compañía era una empresa conocida sin ser la compañía eléctrica local, (5) si el proveedor cobraba mediante una tarifa TOD (*time-of-day*, precios especificados en cada franja horaria del día) y (6) si el proveedor cobraba tarifas estacionales (precios especificados para cada estación del año). En el diseño experimental, las tarifas fijas variaban entre situaciones de elección, pero cada vez que se indicaba que un proveedor ofrecía tarifas TOD o estacionales, se especificaban los mismos precios en todos los experimentos. El coeficiente de las variables indicadoras para las tarifas TOD y estacionales, por lo tanto, refleja el valor de estas tarifas a los precios indicados. El coeficiente del precio fijo indica el valor de cada centavo por kilovatio-hora.

12.7.1 Coeficientes normales independientes

Se estimó un modelo logit mixto bajo la hipótesis inicial de que los coeficientes son independientes y se distribuyen normalmente en la población. Es decir, $\beta_n \sim N(b, W)$ con una W diagonal. Los parámetros de la población son la media y la desviación estándar de cada coeficiente. La tabla 12.1 muestra la media simulada de la distribución posterior (SMP) para estos parámetros, junto con las estimaciones MSL. Para el procedimiento bayesiano, se usaron 20.000 iteraciones de un muestreo de Gibbs. Las primeras 10.000 iteraciones se consideraron *burn-in* y uno de cada 10 valores se retuvo después de lograr la convergencia, alcanzando así un total de 1.000 valores extraídos de la distribución posterior. La media y la desviación estándar de estos valores constituyen las estimaciones y los errores estándar. Para MSL, la probabilidad de elección del modelo logit mixto se simuló con 200 valores aleatorios de Halton por cada observación.

Tabla 12.1. Modelo logit mixto de elección entre proveedores de energía

Estimadores (a)		MSL	SMP	MSL escalado
Coeficiente de precio	Media	−0.976	−1.04	−1.04
		(.0370)	(.0374)	(.0396)
	Desv.estándar	0.230	0.253	0.246
		(.0195)	(.0169)	(.0209)
Coeficiente de contrato	Media	−0.194	−0.240	−0.208
		(.0224)	(.0269)	(.0240)
	Desv.estándar	0.405	0.426	0.434
		(.0238)	(.0245)	(.0255)

Coeficiente empresa local	Media	2.24	2.41	2.40
		(.118)	(.140)	(.127)
	Desv.estándar	1.72	1.93	1.85
		(.122)	(.123)	(.131)
Coeficiente empresa conocida	Media	1.62	1.71	1.74
		(.0865)	(.100)	(.0927)
	Desv.estándar	1.05	1.28	1.12
		(.0849)	(.0940)	(.0910)
Coeficiente TOD	Media	−9.28	−10.0	−9.94
		(.314)	(.315)	(.337)
	Desv.estándar	2.00	2.51	2.14
		(.147)	(.193)	(.157)
Coeficiente estacional	Media	−9.50	−10.2	−10.2
		(.312)	(.310)	(.333)
	Desv.estándar	1.24	1.66	1.33
		(.188)	(.182)	(.201)

(a) Errores estándar entre paréntesis

Los dos procedimientos proporcionan resultados similares en este caso. La escala de las estimaciones del procedimiento bayesiano es algo mayor que la de MSL. Esta diferencia indica que la distribución posterior es asimétrica, con la media superando la moda. Cuando las estimaciones MSL se escalan para que tengan la misma media estimada para el coeficiente de precio, los dos conjuntos de estimaciones son notablemente parecidos, tanto en errores estándar como en estimaciones puntuales. El tiempo de ejecución fue prácticamente el mismo en cada enfoque.

En otros casos, por ejemplo, Ainslie y otros (2001), las estimaciones de MSL y SMP han dado resultados diferentes. En general, la magnitud de las diferencias depende del número de observaciones en relación con el número de parámetros, así como de la cantidad de variación contenida en las observaciones. Cuando los dos conjuntos de estimaciones difieren, significa que las hipótesis asintóticas aún no están operando completamente (es decir, el tamaño de la muestra es insuficiente para que las propiedades asintóticas sean totalmente observables). El investigador podría querer aplicar una perspectiva bayesiana en este caso (si no lo está haciendo ya) con el fin de hacer una inferencia basada en una muestra pequeña. La distribución posterior contiene la información relevante para el análisis bayesiano con cualquier tamaño de la muestra, mientras que la perspectiva clásica requiere que el investigador confíe en fórmulas asintóticas para la distribución muestral que no tienen por qué ser significativas con muestras pequeñas. Allenby y Rossi (1999) proporcionan ejemplos de las diferencias observables y el valor de los enfoques bayesianos y su perspectiva.

Hemos vuelto a estimar el modelo para varios supuestos de distribución. En las siguientes secciones, se describe cómo cada método se implementa bajo estos supuestos alternativos. Por razones que son inherentes a las metodologías, los procedimientos bayesianos son más fáciles y rápidos de aplicar sobre algunas especificaciones, mientras que los procedimientos clásicos son más fáciles y rápidos para otras. Comprender en qué tipo de situaciones es más conveniente usar un enfoque u otro puede ayudar al investigador a decidir qué método utilizar para un modelo en particular.

12.7.2 Coeficientes normales multivariados

A continuación hemos permitido que los coeficientes estén correlacionados entre sí. Es decir, W es una matriz completa en lugar de una matriz diagonal. El procedimiento clásico es el mismo, salvo que la extracción de valores al azar de $\phi(\beta_n|b,W)$ para la simulación de la probabilidad logit mixto exige la creación de una correlación entre valores extraídos de forma independiente a partir de un generador de números aleatorios. El modelo está parametrizado en términos del factor Choleski de W, etiquetado

como L. Los valores se calculan como $\tilde{\beta}_n = b + L\eta$, donde η es un valor extraído al azar de un vector K-dimensional de variables normales estándar independientes. En términos de tiempo de cálculo del MSL, la principal diferencia es que el modelo tiene muchos más parámetros al usar una W plena respecto a una W diagonal: $K + K(K+1)/2$ en lugar de los $2K$ parámetros que teníamos con coeficientes independientes. En nuestro caso con $K = 6$, el número de parámetros se eleva de 12 a 27. El gradiente respecto a cada uno de los nuevos parámetros toma tiempo de cálculo y el modelo requiere más iteraciones para localizar el máximo de una función log-verosimilitud con mayor dimensionalidad. Como se muestra en la segunda línea de la tabla 12.2, el tiempo de ejecución del modelo con coeficientes correlacionados casi triplica el del modelo con coeficientes independientes.

Tabla 12.2. Tiempos de ejecución

	Tiempo de ejecución (min)	
Especificación	MSL	SMP
Todos normales, sin correlación	48	53
Todos normales, covarianza plena	139	55
1 fijo, otros normales, sin correlación	42	112
3 log-normales, 3 normales, sin correlación	69	54
Todos triangulares, sin correlación	56	206

Con el procedimiento bayesiano, los coeficientes correlacionados no son más difíciles de manejar que los no correlacionados. Para una matriz W completa, la distribución gamma es reemplazada por su generalización multivariada, la Wishart invertida. Los valores al azar de esta distribución se extraen por el procedimiento descrito en la sección 12.5.2. El único tiempo de cálculo adicional respecto al modelo con coeficientes independientes surge por la necesidad de cálculo de la matriz de covarianza de las β_ns y su factor Choleski, en lugar de las desviaciones estándar de las β_ns. Esta diferencia es trivial para una cantidad de parámetros típica. Como se muestra en la tabla 12.2, el tiempo de ejecución para el modelo con covarianza plena entre los coeficientes aleatorios es esencialmente el mismo que con coeficientes independientes.

12.7.3 Coeficientes fijos para algunas variables

Hay varias razones por las que el investigador puede optar por especificar como fijos algunos de los coeficientes:

1. Ruud (1996) argumenta que un modelo logit mixto con todos los coeficientes aleatorios es casi inidentificable empíricamente, ya que sólo los ratios de los coeficientes son económicamente significativos. Él recomienda fijar al menos un coeficiente, sobre todo cuando los datos contienen sólo una situación de elección para cada decisor.
2. En un modelo con constantes específicas de alternativa, los términos finales iid tipo valor extremo constituyen la parte aleatoria de estas constantes. Permitir que los coeficientes de las variables ficticias específicas de alternativa sean aleatorios, adicionalmente a tener términos finales iid de tipo valor extremo, es equivalente a suponer que las constantes siguen una distribución que es una mezcla de la distribución valor extremo y la distribución que se haya asumido para esos coeficientes. Si las dos distribuciones son similares, como la distribución valor extremo y la normal, la mezcla puede ser empíricamente no identificable. En este caso, el analista puede optar por mantener fijos los coeficientes de las constantes específicas de alternativa.
3. El objetivo del análisis puede ser predecir correctamente patrones de sustitución en lugar de comprender la distribución de los coeficientes. En este caso, los componentes de error se

pueden especificar para que capturen los patrones de sustitución correctos mientras se mantienen fijos los coeficientes de las variables explicativas originales (como en Brownstone y Train, 1999).

4. La predisposición a pagar (*willingness to pay*, wtp) por un atributo es el ratio entre el coeficiente de dicho atributo y el coeficiente de precio. Si el coeficiente de precio se mantiene fijo, la distribución de la wtp es simplemente la distribución escalada del coeficiente del atributo. La distribución de la wtp resulta más compleja cuando el coeficiente de precio también varía. Además, si para el coeficiente de precio se emplean las distribuciones habituales, como la normal o la log-normal, se plantea la cuestión de cómo manejar coeficientes de precio positivos, coeficientes de precio que están cerca de cero de modo que el wtp es extremadamente alto, y coeficientes de precio que son extremadamente negativos. El primero de estos problemas se evita con log-normales, pero no los otros dos. El analista puede fijar el coeficiente de precio para evitar estos problemas.

En el enfoque clásico, fijar uno o más coeficientes es muy fácil. Los elementos correspondientes de W y L simplemente se fijan a cero, en lugar de tratarse como parámetros. El tiempo de ejecución se reduce, ya que hay menos parámetros. Como se indica en la tercera línea de la tabla 12.2, el tiempo de ejecución se redujo en un 12 por ciento con un coeficiente fijo y el resto normales independientes, en relación al modelo con todos los coeficientes normales independientes. Con las normales correlacionadas, se produciría una reducción porcentual más grande, ya que el número de parámetros cae más que proporcionalmente.

En el procedimiento bayesiano, tener en cuenta coeficientes fijos requiere la adición de una nueva capa en el muestreo de Gibbs. El coeficiente fijo no se puede extraer como parte del algoritmo MH para los coeficientes aleatorios de cada persona. Recordemos que bajo MH, en cada iteración se aceptan o rechazan valores extraídos. Si el valor de prueba extraído que contiene un nuevo valor de un coeficiente fijo junto con los nuevos valores de los coeficientes aleatorios es aceptado para una persona, pero dicho valor de prueba no es aceptado para otra persona, entonces ambas personas tendrán diferentes valores del coeficiente fijo, lo que contradice el hecho mismo de que sea fijo. En lugar de esto, los coeficientes aleatorios y los parámetros de la población de estos coeficientes, deben ser extraídos condicionados a un valor de los coeficientes fijados; y los coeficientes fijados son extraídos condicionados a los valores de los coeficientes aleatorios. Extraer valores al azar de la distribución posterior para los coeficientes fijados requiere el uso de un algoritmo MH, además del que ya se utiliza para extraer valores de los coeficientes aleatorios.

Para ser explícitos, reescriba la función de utilidad como

$$U_{njt} = \alpha' z_{njt} + \beta_n' x_{njt} + \varepsilon_{njt}, \tag{12.6}$$

donde α es un vector de coeficientes fijos y β_n es aleatorio como antes, con media b y varianza W. La probabilidad de la secuencia de elección de la persona dado α y β_n es

$$L(y_n|\alpha, \beta_n) = \prod_t \left(\frac{e^{\alpha' z_{ny_{nt}t} + \beta_n' x_{ny_{nt}t}}}{\sum_j e^{\alpha' z_{njt} + \beta_n' x_{njt}}} \right). \tag{12.7}$$

Las distribuciones posteriores condicionadas para el muestreo de Gibbs son:

1. $K(\beta_n|\alpha, b, W) \propto L(y_n|\alpha, \beta_n)\phi(\beta_n|b, W)$. MH se utiliza para extraer estos valores de la misma manera que se haría con todos coeficientes normales, excepto que ahora $\alpha' z_{njt}$ forma parte de las fórmulas logit.
2. $K(b|W, \beta_n \forall n)$ es $N(\sum_n \beta_n /N, W/N)$. Observe que α no entra en esta distribución posterior; su efecto está incorporado en los valores de β_n de la capa 1.
3. $K(W|b, \beta_n \forall n)$ es $IW(K+N, (KI + N\bar{S})/(K+N))$, donde $\bar{S} = \sum_n (\beta_n - b)(\beta_n - b)'/N$. De nuevo, α no entra directamente.
4. $K(\alpha|\beta_n) \propto \prod_n L(y_n|\alpha, \beta_n)$, si la distribución a priori de α es esencialmente plana (por ejemplo, normal con una varianza suficientemente grande). Se extraen valores al azar con MH sobre los datos agrupados.

La capa 4 requiere tanto tiempo de cálculo como la capa 1, dado que ambas requieren el cálculo de una fórmula logit para cada observación. Por tanto, el procedimiento bayesiano con coeficientes fijos y normales es de esperar que use aproximadamente el doble de tiempo que con todos los coeficientes normales. Como se indica en la tercera línea de la tabla 12.2, esta expectativa se confirma en nuestro caso práctico.

12.7.4 Log-normales

Las distribuciones log-normales se especifican a menudo cuando el analista quiere asegurar que el coeficiente tiene el mismo signo para todas las personas. Se producen pocos cambios en cualquiera de los procedimientos cuando algunos o todos los coeficientes se distribuyen log-normales en lugar de normales. Básicamente, se extraen coeficientes distribuidos normalmente y luego, aquellos que se distribuyen log-normal, son exponenciados cuando entran en la utilidad. Para todos los coeficientes log-normales, la utilidad se especifica como

$$U_{njt} = \left(e^{\beta_n}\right)' x_{njt} + \varepsilon_{njt}, \tag{12.8}$$

con β_n distribuido normalmente como antes, con media b y varianza W. La probabilidad de la secuencia de elecciones de la persona dada $\beta_n$es

$$L(y_n|\alpha, \beta_n) = \prod_t \left(\frac{e^{\left(e^{\beta_n}\right)' x_{ny_{nt}t}}}{\sum_j e^{\left(e^{\beta_n}\right)' x_{njt}}} \right). \tag{12.9}$$

Con este cambio, el resto de pasos son los mismos en ambos procedimientos. En el enfoque clásico, sin embargo, localizar el máximo de la función de probabilidad es considerablemente más difícil con coeficientes log-normales que con los normales. A menudo, los procedimientos numéricos de maximización no logran encontrar un aumento después de un número de iteraciones. O se encuentra un "máximo" y sin embargo el Hessiano es singular en ese punto. Frecuentemente es necesario especificar valores de inicio para los procedimientos de maximización que están cerca del máximo. Y el hecho de que las iteraciones puedan fallar en la mayoría de los valores de inicio hace que sea difícil determinar si un máximo es local o global. El procedimiento bayesiano no encuentra estos problemas, ya que no busca el máximo. El muestreo de Gibbs parece converger un poco más lentamente, pero no de manera apreciable. Como se indica en la tabla 12.2, el tiempo de ejecución para el enfoque clásico subió casi un 50 por ciento para los coeficientes log-normales con relación a los normales (debido a que se requieren más iteraciones), mientras que el procedimiento bayesiano tomó aproximadamente la misma cantidad de tiempo en cada caso. Esta comparación es generosa con el enfoque clásico, dado que en este caso se ha logrado la convergencia en un máximo, mientras que en muchos otros casos prácticos no hemos sido

capaces de obtener la convergencia con log-normales o la hemos obtenido después de invertir un tiempo considerable encontrando valores de inicio exitosos.

12.7.5 Triangulares

Las distribuciones normales y log-normales permiten coeficientes de magnitud ilimitada. En algunas situaciones, el analista podría querer asegurarse de que los coeficientes de todas las personas se mantienen dentro de un rango razonable de valores. Este objetivo se logra mediante la especificación de distribuciones que tengan un ámbito limitado, tales como uniformes, normales truncadas y distribuciones triangulares. En el enfoque clásico, estas distribuciones son fáciles de manejar. El único cambio en el procedimiento se produce en la línea de código del programa que extraer valores al azar de las distribuciones. Por ejemplo, la densidad de una distribución triangular con media b y extensión s es cero fuera del rango $(\mathrm{b}-\mathrm{s},\mathrm{b}+\mathrm{s})$, se eleva linealmente desde $\mathrm{b}-s$ hasta b y cae linealmente hasta $\mathrm{b}+\mathrm{s}$. Un valor al azar se extrae como $\beta_n = b + s(\sqrt{2\mu}-1)$ si $\mu < 0.5$ y $\beta_n = b + s(1-\sqrt{2(1-\mu)})$ en caso contrario, donde μ es un valor extraído al azar de una uniforme estándar. Dados los valores de β_n, el cálculo de la probabilidad simulada y la maximización de la función de verosimilitud son equivalentes a las que se harían con valores extraídos de una normal. La experiencia indica que la estimación de los parámetros de una distribución uniforme, una normal truncada y una distribución triangular tarda aproximadamente el mismo número de iteraciones que en el caso de distribuciones normales. La última línea de la tabla 12.2 refleja esta experiencia.

Con el enfoque bayesiano, el cambio a distribuciones no normales es mucho más complicado. Con coeficientes distribuidos normalmente, las distribuciones posteriores condicionadas para los momentos estadísticos poblaciones son muy convenientes: distribución normal para la media y distribución Wishart invertida para la varianza. La mayoría del resto de distribuciones no dan posteriores tan convenientes. Por lo general, se necesita un algoritmo MH para los parámetros de la población, además del algoritmo MH para los parámetros β_ns a nivel de cliente. Esta adición aumenta considerablemente el tiempo de cálculo. El problema se agrava para distribuciones con ámbito acotado, ya que, como veremos a continuación, es de esperar que el algoritmo MH converja lentamente para estas distribuciones.

Con distribuciones triangulares independientes para todos los coeficientes con vectores de media y extensión b y s respectivamente, y distribuciones a priori planas en cada caso, las distribuciones posteriores condicionadas son:

1. $K(\beta_n|b,\mathrm{s}) \propto L(y_n|\beta_n)h(\beta_n|b,s)$, donde h es la densidad triangular. Se extraen valores al azar a través de MH, de forma separada para cada persona. Este paso es el mismo que con normales independientes excepto el cambio en la densidad de β_n.
2. $K(b,s|\beta_n) \propto \prod_n h(\beta_n|b,s)$ cuando las distribuciones a priori de b y s son prácticamente planas. Se extraen valores al azar a través de MH sobre cada β_n para todas las personas.

Debido al ámbito acotado de la distribución, el algoritmo es extremadamente lento en converger. Considere, por ejemplo, la extensión de la distribución. En la primera capa, valores extraídos al azar de β_n que están fuera del rango $(\mathrm{b}-\mathrm{s},\mathrm{b}+\mathrm{s})$ de la segunda capa son necesariamente rechazados. Y en la segunda capa, valores de b y s que crean una rango $(\mathrm{b}-\mathrm{s},\mathrm{b}+\mathrm{s})$ que no cubre todas las β_ns de la primera capa son necesariamente rechazados.

Por lo tanto, es difícil que el rango crezca de una iteración a la siguiente. Por ejemplo, si el rango es de 2 a 4 en una iteración de la primera capa, la siguiente iteración generará valores de β_n entre 2 y 4, y por lo general cubrirá la mayor parte del rango si el tamaño de la muestra es suficientemente grande. En la siguiente extracción de valores de b y s, cualquier valor que no cubra el rango de las β_ns (que es aproximadamente de 2 a 4) será rechazado. En efecto, existe un cierto margen para jugar, dado que las β_ns no cubrirán todo el rango de 2 a 4. El algoritmo converge, pero en nuestro caso se encontró que

eran necesarias muchas más iteraciones para lograr algo similar a la convergencia, en comparación con las distribuciones normales. En consecuencia, el tiempo de ejecución aumentó en un factor cuatro.

12.7.6 Resumen de los resultados

Para distribuciones normales con matrices de covarianza completas y para las transformaciones de distribuciones normales que pueden expresarse en la función de utilidad, tales como la exponenciación para representar la distribución log-normal, el enfoque bayesiano parece ser muy atractivo computacionalmente hablando. Usar coeficientes fijos añade una capa de condicionamiento al enfoque bayesiano que duplica su tiempo de ejecución. En contraste, el enfoque clásico se vuelve más rápido por cada coeficiente que se define como fijo en lugar de aleatorio, debido a que se reduce el número de parámetros a estimar. Para distribuciones con ámbito acotado, como las triangulares, el enfoque bayesiano es muy lento, mientras que el enfoque clásico maneja estas distribuciones tan rápidamente como las normales.

Estas comparaciones se refieren sólo a logits mixtos. Es de esperar que otros modelos de comportamiento tengan diferentes tiempos de cálculo para cada uno de los dos enfoques. La comparación realizada con el modelo logit mixto dilucida las cuestiones que se plantean en la aplicación de cada método. La comprensión de estas cuestiones ayuda al investigador a especificar el modelo y el método que son más apropiados y convenientes para la situación de elección.

12.8 Procedimientos bayesianos para modelos probit

Los procedimientos bayesianos pueden aplicarse a modelos probit. De hecho, los métodos son aún más rápidos para modelos probit que para logits mixtos. El procedimiento es descrito por Albert y Chib (1993), McCulloch y Rossi (1994), Allenby y Rossi (1999) y McCulloch y Rossi (2000). El método difiere en un punto crítico del procedimiento para modelos logits mixtos. En particular, para un modelo probit la probabilidad de las elecciones de cada persona condicionada a los coeficientes de las variables, que es el análogo de $\mathrm{L}(\mathrm{y}_n|\beta_n)$ para un modelo logit, no tiene una forma cerrada. Los procedimientos que utilizan esta probabilidad, como sucede en la primera capa del muestreo de Gibbs para un logit mixto, no se pueden aplicar fácilmente a un modelo probit. En lugar de usar esta probabilidad, el muestreo de Gibbs para probits se realiza considerando que las utilidades de las alternativas, U_{njt}, son parámetros en sí mismas. La distribución posterior condicionada de cada U_{njt} es una normal truncada, de la cual es fácil extraer valores al azar. Los distintos niveles del muestreo de Gibbs son los siguientes:

1. Extraiga al azar un valor de b condicionado a W y a $\beta_n \forall n$.
2. Extraiga W condicionado a b y a $\beta_n \forall n$. Estas dos capas son las mismas que hemos definido para el modelo logit mixto.
3. Para cada n, extraiga β_n condicionada a $\mathrm{U}_{njt}\ \forall j, t$. Estos valores se extraen reconociendo que, dado el valor de utilidad, la función $\mathrm{U}_{njt} = \beta_n \mathrm{x}_{njt} + \varepsilon_{njt}$ es una regresión de x_{njt} respecto a U_{njt}. Se han obtenido distribuciones posteriores bayesianas para coeficientes de regresión y errores normalmente distribuidos (de forma similar a nuestros resultados A y B) y resulta simple extraer valores al azar de las mismas.
4. Para cada $\mathrm{n}, \mathrm{i}, \mathrm{t}$, extraiga U_{nit} condicionada a β_n y el valor de U_{njt} para cada $\mathrm{j} \neq \mathrm{i}$. Como se mostró anteriormente, la distribución posterior condicionada para cada U_{nit} es una normal truncada univariante, de la que es fácil extraer valores al azar con el procedimiento indicado en la sección 9.2.4.

Los detalles figuran en los artículos citados.

Bolduc et al. (1997) compararon el método bayesiano con MSL y encontraron que el procedimiento bayesiano requería aproximadamente la mitad de tiempo de cálculo que MSL con valores aleatorios. Si se hubiesen utilizado valores al azar de Halton en la simulación, parece ser que MSL habría sido más rápido para el mismo nivel de exactitud, debido a que se habrían necesitado menos de la mitad de los valores. El procedimiento bayesiano para probit se basa en que todos los términos aleatorios se distribuyen normalmente. Sin embargo, la idea de tratar las utilidades como parámetros se puede generalizar para otras distribuciones, lo que definiría un procedimiento bayesiano para probits mixtos.

Se pueden desarrollar procedimientos bayesianos de una forma u otra para prácticamente cualquier modelo de comportamiento. En muchos casos, proporcionan grandes ventajas computacionales respecto a los procedimientos clásicos. Algunos ejemplos son los modelos de elección discreta dinámicos de Imai et al. (2001), los modelos conjuntos relativos al momento y a la cantidad de las compras de Boatwright et al. (2003) y la combinación de distintos modelos de elección discreta a cargo de Brownstone (2001). El poder de estos procedimientos y sobre todo la posibilidad de combinarlos con métodos clásicos, crean una perspectiva brillante para este campo del conocimiento.

13

Endogeneidad

13.1 Descripción general

Hasta ahora hemos supuesto que las variables explicativas que entran en un modelo de elección discreta son independientes de los factores no observados. Sin embargo, en muchas situaciones las variables explicativas son endógenas, es decir, están correlacionadas o en cualquier caso no son independientes de los factores no observados. Algunos ejemplos son los siguientes:

1. *Atributos no observados de un producto pueden afectar a su precio*.

   Al modelar las elecciones de un consumidor entre diferentes productos, podría ser imposible medir todos los atributos relevantes de los diversos productos. En el caso de los automóviles, por ejemplo, el investigador puede obtener información acerca de la eficiencia del combustible, la longitud, la anchura, la potencia, el peso y muchos otros atributos de cada vehículo ofrecido por los fabricantes, pero atributos tales como la comodidad, belleza del diseño, suavidad de marcha, manejo en curvas, valor esperado de reventa y prestigio no se pueden medir directamente. Sin embargo, es de esperar que el precio del producto refleje estos atributos no observados (es decir, no medidos). Hay dos razones por las cuales el precio se ve afectado. En primer lugar, en la medida en que los atributos no observados supongan un costo para el fabricante, es de esperar que el precio del producto refleje estos costos. En segundo lugar, si los atributos no observados afectan a la demanda del producto, un precio determinado por la interacción entre la oferta y la demanda podría reflejar estas diferencias en la demanda. El resultado final es que el precio está correlacionado con atributos no observados, en lugar de ser independiente como se ha supuesto hasta ahora en este libro.

2. *Los esfuerzos en actividades de marketing pueden estar relacionados con los precios*.

   La publicidad y la promoción de las ventas, a través de acciones como cupones y descuentos, están en todas partes. A menudo, las prácticas habituales de comercialización de las empresas crean una correlación entre el precio de los productos, que observa el investigador, y las actividades de promoción no vinculadas al precio, que por lo general el investigador no puede medir directamente. La correlación puede ir en cualquier dirección. Un producto puede ser promovido a través de una campaña publicitaria junto con descuentos. La publicidad en este

caso estaría correlacionada negativamente con el precio: mayor publicidad se produce simultáneamente con precios más bajos. Por el contrario, las empresas pueden aumentar el precio de sus productos para pagar la publicidad, lo que crearía una correlación positiva. En cualquier caso, el precio del producto ya no es independiente de los factores no observados que afectan a las elecciones de los consumidores.

3. *Decisiones interrelacionadas de los decisores*.

   En muchas situaciones, los factores observados que afectan a la elección realizada por una persona están determinados por otra elección de esa persona. Medio de transporte y ubicación de vivienda son un ejemplo destacado. En un modelo de elección del medio habitual de transporte para ir al trabajo, las variables explicativas observadas son por lo general el costo y el tiempo de desplazamiento desde el hogar hasta el lugar de trabajo para cada medio (automóvil, autobús y tren). Sin embargo, las personas que tienden a preferir el transporte público (o que les disgusta menos que a la persona promedio) también podrían tender a comprar o alquilar casas que están cerca del transporte público. Por lo tanto, el tiempo de viaje en transporte es menor para estas personas que para las personas que residen más lejos de la zonas de tránsito. Expresado en términos de factores observados y no observados del modelo de elección del medio de transporte, las actitudes observadas hacia el transporte público, que afectan a la elección del medio de transporte pero que no pueden ser medidas por completo por el investigador, están (negativamente) correlacionadas con el tiempo observado de los viajes en transporte público.

En situaciones como éstas, estimar un modelo sin tener en cuenta la correlación entre factores observados y no observados es inconsistente. La dirección del sesgo a menudo se puede determinar por pura lógica. Por ejemplo, si los atributos no observados deseables están correlacionados positivamente con el precio, la estimación sin tener en cuenta esta correlación se traducirá en un coeficiente de precio estimado que estará sesgado a la baja en magnitud. La razón es obvia: dado que los precios más altos se asocian con atributos deseables, los consumidores evitan los productos más caros *menos* de lo que lo harían si los precios más altos se produjesen sin ningún cambio en compensación en los atributos no observados. Esencialmente, el coeficiente de precio estimado recoge tanto el efecto del precio (que es negativo) como el efecto de los atributos no observados deseables (que son positivos), con estos últimos ocultando el efecto del primero. Un sesgo similar, aunque en la dirección opuesta, se produce si las acciones de marketing consisten en publicidad acompañada de descuentos en los precios. El aumento de la demanda de los productos comercializados proviene tanto del precio más bajo y como de la publicidad (ajena a los precios). El coeficiente de precio estimado recogería los dos efectos, que sumados son mayores al impacto de los precios más bajos por sí mismos.

Varios métodos han sido desarrollados para estimar modelos de elección en presencia de variables explicativas endógenas. En este capítulo se describen estos métodos, delimitando las ventajas y las limitaciones de cada aproximación. En primer lugar describimos el enfoque BLP, desarrollado por Berry, Levinsohn y Pakes (de ahí las iniciales) a través de una serie de publicaciones. Berry (1994) señaló que se pueden incluir constantes en el modelo de elección para capturar el efecto promedio de los atributos del producto (tanto observados y como no observados). Es posible entonces hacer una regresión lineal de las constantes estimadas respecto a los atributos observados, donde la endogeneidad se maneja de la manera habitual, a través de estimación mediante variables instrumentales. En esencia, demostró que la endogeneidad podía extraerse del modelo de elección, que es intrínsecamente no lineal, e insertarlo en un modelo de regresión lineal, donde la endogeneidad puede ser manejada a través de estimación de variables instrumentales estándar. Para aplicar este método, a menudo es necesario estimar un número muy grande de constantes dentro del modelo de elección, lo cual puede ser difícil utilizando métodos de maximización estándar basados en el gradiente. Para abordar esta cuestión, BLP (1995)

proporcionó un procedimiento, llamado "la contracción" (*the contraction*), que facilitaba la estimación de estas constantes. Estos dos artículos iniciales estaban basados en modelos agregados, es decir, modelos estimados con datos agregados de cuotas de mercado. Sin embargo, los conceptos son aplicables a datos de elección a nivel individual o a una combinación de datos a nivel agregado e individual, tal y como se utilizan en un artículo posterior por parte de BLP (2004). Casos de uso del enfoque BLP incluyen a Nevo (2001), Petrin (2002), Goolsbee y Petrin (2004), Chintagunta, Dubé y Goh (2005), y Train y Winston (2007), por nombrar sólo unos pocos. Una versión bayesiana del procedimiento ha sido desarrollada por Yang, Chen, y Allenby (2003) y Jiang, Manchanda y Rossi (2007).

El segundo procedimiento que describimos es el enfoque llamado de función de control. Los conceptos que motivan este enfoque datan de los trabajos de Heckman (1978) y Hausman (1978), aunque el primer uso del término "función de control" parece haber sido a cargo de Heckman y Robb (1985). La endogeneidad surge cuando las variables observadas están correlacionadas con factores no observados. Esta correlación implica que los factores no observados, condicionados a las variables observadas, no tienen media cero, como se requiere por lo general para una estimación estándar. Una función de control es una variable que captura la media condicionada, esencialmente "controlando" la correlación. Ríos y Voung (1988) adaptaron estas ideas para manejar la endogeneidad a un modelo probit binario con coeficientes fijos, y Petrin y Train (2009) generalizaron este enfoque a los modelos de elección multinomiales con coeficientes aleatorios. El procedimiento se lleva a cabo en dos pasos. En primer lugar, la variable explicativa endógena (como el precio) es objeto de una regresión respecto variables exógenas. La regresión estimada se utiliza para crear una nueva variable (la función de control) que se introduce en el modelo de elección. El modelo de elección es estimado a continuación, con las variables originales más la nueva variable, lo que permite representar adecuadamente la distribución de los factores no observados condicionados tanto a esta nueva variable como a las originales. Algunos ejemplos de este método los proporcionan Ferreira (2004) y Guervara y Ben-Akiva (2006).

El tercer procedimiento es un enfoque completo de máxima verosimilitud, tal y como lo aplican Villas-Boas y Winer (1999) a un modelo logit multinomial con coeficientes fijos, generalizado por Park y Gupta (de próxima publicación) a modelos de elección con coeficientes aleatorios. El procedimiento está estrechamente relacionado con el enfoque de la función de control, en el sentido de que tiene en cuenta la media condicionada distinta de cero de los factores no observados. Sin embargo, en lugar de implementar secuencialmente los dos pasos (es decir, estimar el modelo de regresión para crear la función de control y a continuación estimar el modelo de elección con esta función de control), los dos pasos se combinan en un criterio de estimación conjunta. Se requieren supuestos adicionales para poder hacer la estimación de forma simultánea, sin embargo, el procedimiento es más eficiente cuando se cumplen estos supuestos.

En las secciones siguientes, se trata cada uno de estos procedimientos, y se acompañan de un caso de estudio utilizando el enfoque BLP.

13.2 El Enfoque BLP

Este procedimiento se describe con mayor facilidad para elecciones entre productos en los que el precio es endógeno. Un conjunto de productos se vende en varios mercados y en cada mercado tiene numerosos consumidores. Los atributos de los productos varían entre mercados, pero no entre consumidores de cada mercado (es decir, todos los consumidores dentro de un mercado determinado se enfrentan a los mismos productos con los mismos atributos). La definición de lo que es un mercado depende del caso práctico. El mercado podría ser una zona geográfica, como en el análisis de Goolsbee y Petrin (2004) relativo a la elección hecha por los hogares entre la oferta televisiva. En este caso concreto, el precio de la televisión por cable y las características ofrecidas (como el número de canales) varían entre ciudades, ya que las franquicias de cable se conceden por los gobiernos locales. Alternativamente, el mercado podría ser

definido temporalmente, como en el análisis BLP de la demanda de nuevos vehículos (1995, 2004), donde cada año-modelo consiste en un conjunto de marcas y modelos de vehículos nuevos junto a sus precios y otros atributos. Cada año constituye un mercado en este caso práctico. Si el análisis es entre productos cuyos atributos son los mismos para todos los consumidores, entonces sólo hay un mercado y la diferenciación por mercado es innecesaria.

13.2.1 Especificación

Sea M el número de mercados y J_m el número de opciones disponibles para cada uno de los consumidores en el mercado m. J_m es el número de productos disponibles en el mercado m más, tal vez, dependiendo del análisis, la opción de no comprar ninguno de los productos, que a veces se denomina "el bien externo" [ii] (*the outside good*). El precio del producto j en el mercado m se denota como p_{jm}. Algunos de los atributos de los productos distintos del precio son observados por el investigador y otros no. Los atributos distintos del precio del producto j en el mercado m que sí son observados se denotan por el vector x_{jm}. Los atributos no observados se designan colectivamente como ξ_{jm}, cuyo significado preciso se trata en mayor detalle más adelante.

La utilidad que el consumidor n en el mercado m obtiene del producto j depende de los atributos observados y no observados del producto. Suponga que la utilidad toma la forma

$$U_{njm} = V\left(p_{jm}, x_{jm}, s_n, \beta_n\right) + \xi_{jm} + \varepsilon_{njm},$$

donde s_n es un vector de características demográficas de los consumidores, $V(\cdot)$ es una función de las variables observadas y las preferencias del consumidor, representadas por el β_n, y ε_{njm} es de tipo valor extremo iid. Observe que ξ_{jm} entra en la utilidad de la misma manera para todos los consumidores; en esta configuración, por lo tanto, ξ_{jm} representa el promedio, o la parte común, de la utilidad que los consumidores obtienen de los atributos no observados del producto j en el mercado m.

El problema fundamental que motiva el enfoque de esta estimación es la endogeneidad de los precios. En particular, el precio de cada producto depende en general de todos sus atributos, tanto los que son observados por el investigador como los que no pueden ser medidos por el investigador, pero sin embargo afectan la demanda y/o los costos del producto. Como resultado, el precio p_{jm} depende de ξ_{jm}.

Supongamos ahora que íbamos a estimar este modelo sin tener en cuenta esta endogeneidad. El modelo de elección incluiría el precio p_{jm} y los atributos observados x_{jm} como variables explicativas. La parte no observada de la utilidad, condicionada a β_n, sería $\varepsilon^*_{njm} = \xi_{jm} + \varepsilon_{njm}$, que incluye la utilidad media de atributos no observados. Sin embargo, dado que p_{jm} depende de ξ_{jm}, este componente no observado, ε^*_{njm}, no es independiente de p_{jm}. Por el contrario, se podría esperar una correlación positiva, estando los atributos no observados más deseables asociados a precios más altos.

La aproximación BLP a este problema consiste en mover ξ_{jm} a la parte observada de la utilidad. Esto se logra mediante la introducción de una constante para cada producto en cada mercado. Sea $\bar{V}$ la porción de $V(\cdot)$ que varía entre productos y mercados, pero que es igual para todos los consumidores. Sea $\tilde{V}$ sea la parte que varía entre consumidores así como entre mercados y productos. Entonces $V(\cdot) =$

[ii] Si el bien externo está incluido el modelo se puede utilizar para predecir la demanda total de productos bajo condiciones alteradas. Si el bien externo no está incluido, el análisis examina la elección de los consumidores entre los productos condicionada a la compra de uno de los productos. El modelo puede ser utilizado para predecir la evolución de las cuotas entre los consumidores que originalmente compraron los productos, pero no se puede usar para predecir los cambios en la demanda total, ya que no incluye los cambios en el número de consumidores que decidieron no comprar cualquiera de los productos. Si el bien externo está incluido, generalmente se considera que su precio es cero.

$\bar{V}(p_{jm}, x_{jm}, \bar{\beta}) + \tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n)$, donde $\bar{\beta}$ son parámetros que son iguales para todos los consumidores y $\tilde{\beta}_n$ son parámetros que varían entre consumidores. Observe que $\bar{V}$ no depende de s_n ya que es constante entre consumidores[iii]. Es más natural pensar en $\bar{V}$ como la representación de la media de V en la población; sin embargo, no tiene por qué ser así. Todo lo que se requiere es que $\bar{V}$ sea constante entre consumidores. La variación en la utilidad de los atributos observados alrededor de esta constante es capturada por $\tilde{V}$, que puede depender de datos demográficos observados y de coeficientes que varían aleatoriamente. La utilidad es entonces

$$U_{njm} = \bar{V}(p_{jm}, x_{jm}, \bar{\beta}) + \tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n) + \xi_{jm} + \varepsilon_{njm}.$$

Reordenando los términos, tenemos

$$U_{njm} = \left[\bar{V}(p_{jm}, x_{jm}, \bar{\beta}) + \xi_{jm}\right] + \tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n) + \varepsilon_{njm}.$$

Observe que el término entre paréntesis no varía entre consumidores. Es constante para cada producto en cada mercado. Denotemos esta constante como

(13.1) $$\delta_{jm} = \bar{V}(p_{jm}, x_{jm}, \bar{\beta}) + \xi_{jm}$$

y sustituyámosla en la utilidad

(13.2) $$U_{njm} = \delta_{jm} + \tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n) + \varepsilon_{njm}.$$

Un modelo de elección basado en esta especificación de la utilidad no implica ninguna endogeneidad. Una constante se incluye para cada producto en cada mercado, lo que absorbe ξ_{jm}. La porción restante de utilidad no observada, ε_{njm}, es independiente de las variables explicativas. Las constantes se calculan junto con los otros parámetros del modelo. En esencia, el término que causó el carácter endógeno, es decir, ξ_{jm}, se ha subsumido en la constante de producto-mercado de tal manera que ya no es parte del componente no observado de la utilidad.

El modelo de elección se completa especificando cómo $\tilde{\beta}_n$ varía entre consumidores. Denotemos la densidad de $\tilde{\beta}_n$ como $f(\tilde{\beta}_n|\theta)$, donde θ son parámetros de esta distribución que representan, por ejemplo, la varianza de los coeficientes alrededor de los valores comunes. Teniendo en cuenta que ε_{njm} es de tipo valor extremo iid, la probabilidad de elección es la de un modelo logit mixto:

(13.3) $$P_{nim} = \int \left[\frac{e^{\delta_{im} + \tilde{V}(p_{im}, x_{im}, s_n, \tilde{\beta}_n)}}{\sum_j e^{\delta_{jm} + \tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n)}}\right] f(\tilde{\beta}_n|\theta) d\tilde{\beta}_n.$$

Por lo general, $\tilde{V}$ es lineal con coeficientes $\tilde{\beta}_n$ y variables explicativas que son los atributos observados, p_{jm} y x_{jm}, interactuando quizá con los demográficos, s_n.Pueden especificarse otras distribuciones para ε_{njm}; por ejemplo, Goolsbee y Petrin (2004) asumen que ε_{njm} es una normal conjunta entre productos, de manera que la probabilidad de elección es probit. Además, si disponemos de información sobre el ranking de preferencias del consumidor respecto a las alternativas, total o parcialmente, entonces la probabilidad del ranking se especifica de manera análoga; por ejemplo, Berry, Levinsohn y Pakes (2004) y Train y Winston (2007) disponían de datos sobre el vehículo comprado por cada consumidor así como

[iii] Es posible que $\bar{V}$ incluya datos demográficos agregados, como la renta media del mercado, que varía entre mercados, pero no entre consumidores de cada mercado. Sin embargo, nos abstraemos de esta posibilidad en nuestra notación

sobre su segunda elección del vehículo, y representaron este ranking parcial mediante la inserción de la fórmula logit expandida de la Sección 7.3 dentro de los corchetes de la ecuación (13.3) en lugar de la fórmula logit estándar.

La estimación del modelo de elección descrito en (13.3) proporciona estimaciones de las constantes y de la distribución de preferencias. Sin embargo, no proporciona estimaciones de los parámetros que entran en la parte de la utilidad que es constante entre consumidores; es decir, no proporciona estimaciones de $\bar{\beta}$ en $\bar{V}$. Estos parámetros entran en la definición de las constantes de la ecuación (13.1), lo que constituye un modelo de regresión que puede ser usado para estimar el promedio de las preferencias. Es habitual expresar $\bar{V}$ como una función lineal respecto a los parámetros, de tal manera que (13.1) se convierte en

$$\delta_{jm} = \bar{\beta}'\bar{V}(p_{jm}, x_{jm}) + \xi_{jm} \tag{13.4}$$

donde $\bar{V}(\cdot)$ es una función vectorial de los atributos observados. Es posible estimar una regresión en la que la variable dependiente es la constante para cada producto en cada mercado, y en la que las variables explicativas son el precio y otros atributos observados del producto. El término de error para esta regresión es ξ_{jm}, que se correlaciona con el precio. Sin embargo, los procedimientos para manejar la endogeneidad en los modelos de regresión lineal están bien desarrollados y se describen en cualquier libro de texto de econometría estándar. En particular, la regresión (13.4) se estima a través de variables instrumentales en lugar de mínimos cuadrados ordinarios. Todo lo que se requiere para esta estimación es que el investigador tenga, o pueda calcular, algunas variables exógenas adicionales que se utilizan como instrumentos en lugar del precio endógeno. La selección de instrumentos se trata más tarde como parte del procedimiento de estimación; sin embargo, primero tenemos que afrontar un tema importante que ha estado implícito en nuestra exposición hasta ahora. Nos referimos a la forma de manejar el hecho de que podrían haber (y por lo general hay) un gran número de constantes a estimar, una para cada producto en cada mercado.

13.2.2 La contracción

Como se describió anteriormente, las constantes $\delta_{jm}\ \forall j, m$ se estiman junto con los otros parámetros del modelo de elección. Cuando existen numerosos productos y/o mercados, la estimación de este gran número de constantes puede ser difícil o inviable numéricamente, si uno trata de estimarlas de manera estándar. Por ejemplo, para la elección del vehículo, hay más de 200 marcas y modelos de vehículos nuevos cada año, lo que requiere la estimación de más de 200 constantes para la información correspondiente a cada año. Para 5 años de datos, más de un millar de constantes tendrían que ser estimadas. Si se utilizaran los procedimientos del capítulo 8 para un modelo así, cada iteración implicaría el cálculo de la pendiente respecto a, por ejemplo, más de 1000 parámetros, e invertir una matriz Hessiana de más de 1000 por 1000; también se necesitarían numerosas iteraciones ya que la búsqueda es en un espacio de parámetros de más de 1000 dimensiones.

Por suerte, no es necesario estimar las constantes de la forma estándar. BLP proporciona un algoritmo para estimarlas de forma rápida, dentro del proceso iterativo del resto de parámetros. Este procedimiento se basa en la constatación de que las constantes determinan las cuotas de mercado previstas para cada producto, y que por lo tanto se pueden fijar de tal manera que las cuotas de mercado previstas sean iguales a los porcentajes reales. Para ser precisos, sea S_{jm} sea la cuota o proporción de los consumidores en el mercado de m que eligen el producto j. Para un modelo correctamente especificado, las cuotas *predichas* en cada mercado deben ser iguales a estas cuotas reales (al menos asintóticamente). Podemos encontrar las constantes que hacen cumplir esta igualdad, es decir, que hacen que el modelo prediga las cuotas que realmente se observan. Representemos las

constantes a través de un vector $\delta = \langle \delta_{jm} \, \forall j, m \rangle$. Las cuotas previstas son $\hat{S}_{jm}(\delta) = \sum_n P_{njm}/N_m$, donde la suma se extiende sobre los N_m consumidores muestreados en el mercado m. Estas cuotas previstas se expresan como una función de las constantes δ debido a que las constantes afectan a las probabilidades de elección que a su vez afectan a las cuotas previstas.

Recordemos que en la sección 2.8 se describe un procedimiento iterativo para re-calibrar las constantes en un modelo de manera que las cuotas previstas sean iguales a las cuotas reales. Empezando con cualesquiera valores de las constantes, etiquetados $\delta_{jm}^t \, \forall j, m$, las constantes se ajustan iterativamente mediante la fórmula

$$\delta_{jm}^{t+1} = \delta_{jm}^t + ln\left(\frac{S_{jm}}{\hat{S}_{jm}(\delta^t)}\right).$$

Este proceso de ajuste mueve cada constante en la dirección "correcta", en el sentido siguiente. Si, con el valor actual de la constante, la cuota real de un producto excede la cuota predicha, entonces el ratio entre cuota real y predicha (es decir, $S_{jm}/\hat{S}_{jm}(\delta^t)$) es mayor que 1 y $\ln(\cdot)$ es positivo. En este caso, la constante se ajusta al alza, para elevar la proporción predicha. Cuando la cuota real está por debajo de la predicha, el ratio está por debajo de 1 y $\ln(\cdot) < 0$, de manera que la constante se ajusta a la baja. El ajuste se repite iterativamente hasta que las cuotas previstas son iguales a las cuotas reales (dentro de un margen de tolerancia) para todos los productos en todos los mercados.

Este algoritmo se puede utilizar para estimar las constantes en lugar de estimarlas a través de los métodos habituales basados en el gradiente. Los otros parámetros del modelo sí se estiman a través de métodos basados en el gradiente, y en cada valor de prueba de estos otros parámetros (es decir, cada iteración en la búsqueda de la optimización de los valores de los otros parámetros), las constantes se ajustan de tal manera que las cuotas previstas son iguales a las cuotas reales para este valor de prueba. Básicamente, el procedimiento que había sido utilizado durante muchos años para la recalibración post-estimación de constantes se utiliza *durante* la estimación, en cada iteración para los otros parámetros.

Berry (1994) demostró que para cualesquiera valores de los otros parámetros en el modelo de elección (es decir, de θ), existe un conjunto único de constantes para las que cuotas previstas son iguales a las cuotas reales. Posteriormente BLP (1995) mostró que el proceso de ajuste iterativo es una contracción, de tal manera que la convergencia a un conjunto único de constantes está garantizada. Cuando se utiliza en el contexto de la estimación en lugar del contexto la calibración post-estimación, el algoritmo ha llegado a ser conocido como "la contracción".

Algunas notas adicionales son útiles en relación a la contracción. En primer lugar, anteriormente hemos definido las cuotas S_{jm} como las cuotas "reales". En la práctica, se pueden utilizar tanto las cuotas de mercado agregadas como las cuotas de la muestra. En algunas situaciones, los datos sobre las cuotas agregadas no están disponibles o no son fiables. Las cuotas de la muestra son consistentes con las cuotas de mercado, siempre que el muestreo sea exógeno. En segundo lugar, el procedimiento impone una restricción o condición sobre la estimación: que las cuotas predichas sean iguales a las cuotas reales. Como vimos en la sección 3.7.1, la estimación de máxima verosimilitud de un modelo logit estándar con constantes específicas de alternativa para cada producto en cada mercado, da necesariamente cuotas previstas iguales a las cuotas de la muestra. Por tanto, la condición de que las cuotas predichas sean iguales a las cuotas de la es consistente con (o más precisamente, es una característica de) la máxima verosimilitud en un modelo logit estándar. Sin embargo, para otros modelos, incluyendo probit y logit mixto, el estimador de máxima verosimilitud no equipara las cuotas previstas con las cuotas de la muestra, incluso cuando se incluyen un conjunto completo de constantes. Las constantes estimadas que se obtienen a través de la contracción, no son por tanto las estimaciones

de máxima verosimilitud. Sin embargo, ya que la condición se cumple asintóticamente para un modelo correctamente especificado, imponerlo parece razonable.

13.2.3 Estimación por máxima verosimilitud simulada y variables instrumentales

Hay varias maneras de que los otros parámetros del modelo (es decir, θ y $\bar{\beta}$) puedan ser estimados. El procedimiento más fácil de conceptualizar es el utilizado por Goolsbee y Petrin (2004) y Train y Winston (2007). En estos estudios, el modelo de elección de la ecuación (13.3) se calcula en primer lugar, utilizando máxima verosimilitud simulada (MSL) con la contracción. Este paso proporciona estimaciones de los parámetros que entran en la ecuación (13.2), a saber, las constantes δ_{jm} $\forall j, m$ y los parámetros θ de la distribución de preferencias alrededor de estas constantes. La contracción se utiliza para las constantes, de manera que la maximización de la función log -verosimilitud es sólo sobre θ.

Para ser más precisos, ya que las probabilidades de elección dependen tanto de δ como de θ, esta dependencia se puede denotar funcionalmente como $P_{njm}(\delta, \theta)$. Sin embargo, para cualquier valor dado de θ, las constantes δ están completamente determinadas: son los valores que iguala las cuotas predichas y las cuotas reales cuando este valor de θ se utiliza en el modelo. Por tanto, las constantes calibradas se pueden considerar una función de θ, denotada $\delta(\theta)$. Sustituyendo en la probabilidad de elección, la probabilidad se convierte en una función únicamente de θ: $P_{njm}(\theta) = P_{njm}(\delta(\theta), \theta)$. La función log-verosimilitud se define también como una función de θ: con i_n denotando la alternativa elegida por n, la función log-verosimilitud es $LL(\theta) = \sum_n \ln P_{ni_nm}(\theta)$, donde δ se re-calcula adecuadamente para cualquier θ. Como tal, el estimador es MSL sujeto a la restricción de que las cuotas predichas sean iguales a las cuotas reales (ya sean cuotas de mercado o de la muestra, la que sea que se esté utilizando)[iv].

Una vez que el modelo de elección es estimado, las constantes estimadas se utilizan en la regresión lineal (13.4), que repetimos aquí por conveniencia:

$$\delta_{jm} = \bar{\beta}' \bar{v}(p_{jm}, x_{jm}) + \xi_{jm}$$

Las constantes estimadas del modelo de elección son las variables dependientes en esta regresión, y el precio y otros atributos observados de los productos son las variables explicativas. Puesto que el precio es endógeno en esta regresión, se estima a través de variables instrumentales en lugar de mínimos cuadrados ordinarios. Los instrumentos incluyen los atributos diferentes al precio observados de los productos, x_{jm}, además de al menos un instrumento adicional en lugar del precio. Si nos referimos a todos los instrumentos a través del vector z_{jm}, el estimador de variables instrumentales es el valor de $\bar{\beta}$ que satisface

$$\sum_j \sum_m [\hat{\delta}_{jm} - \bar{\beta}' \bar{v}(p_{jm}, x_{jm})] z_{jm} = 0,$$

donde $\hat{\delta}_{jm}$ es la constante estimada a partir del modelo de elección. Podemos reorganizar esta expresión para expresar el estimador de forma cerrada, como suele mostrarse por lo general en los libros de texto sobre regresión, y como se indica en la Sección 10.2.2:

[iv] Desde un punto de vista de programación, la maximización implica iteraciones dentro de iteraciones. El procedimiento de optimización itera sobre valores de θ en la búsqueda del máximo de la función log-verosimilitud. En cada valor de prueba de θ, la contracción itera sobre los valores de las constantes, ajustándolas hasta que las cuotas previstas son iguales a las cuotas reales en ese valor de prueba de θ.

$$\hat{\hat{\beta}} = \left(\sum_j \sum_m z_{jm} \bar{v}(p_{jm}, x_{jm})' \right)^{-1} \left(\sum_j \sum_m z_{jm} \hat{\delta}_{jm} \right).$$

Si el investigador lo desea, la eficiencia puede ser mejorada teniendo en cuenta la covarianza entre las constantes estimadas, a través de mínimos cuadrados generalizados (GLS); véase, por ejemplo, Greene (2000), en relación a la estimación GLS de modelos de regresión lineal.

El problema surge necesariamente en relación a que variables utilizar como instrumentos. Es frecuente, como ya se mencionó, utilizar como instrumentos los atributos observados diferentes del precio bajo el supuesto de que son exógenos[v]. BLP (1994) sugirieron el uso de instrumentos basados en los conceptos de precio. En particular, cada fabricante fijará el precio de cada uno de sus productos de una manera que tenga en consideración la sustitución con sus otros productos, así como la sustitución con productos de otras empresas. Por ejemplo, cuando una empresa está considerando la posibilidad de un aumento de precio para uno de sus productos, los consumidores que abandonarán el consumo de este producto para consumir otro de los productos de la misma empresa no representan la parte más importante de la pérdida (y de hecho hasta podrían representar una ganancia, en función de los márgenes de beneficio) como sí lo son los consumidores que pasarán a consumir productos de otras empresas. Sobre la base de estas ideas, BLP propuso dos instrumentos: los atributos medios distintos del precio de otros productos del mismo fabricante y los atributos medios distintos del precio de los productos de otras empresas. Por ejemplo, en el contexto de la elección de un vehículo en la que, por ejemplo, el peso del vehículo es un atributo observado, los dos instrumentos para el Toyota Camry para un año determinado son (1) el peso medio de todas los modelos de vehículos Toyota en ese año y (2) el peso promedio de todos los vehículos que no son Toyota en ese año[vi]. Train y Winston (2007) utilizaron una extensión de estos instrumentos que refleja hasta qué punto cada producto se diferencia de otros productos del mismo fabricante y de otros fabricantes. En particular, emplearon la suma de las diferencias al cuadrado entre el producto y cada uno de los otros productos del mismo fabricante y de otros fabricantes. Estos dos instrumentos para el producto j en el mercado m son $\mathrm{z}^1_{jm} = \sum_{k \in K_{jm}} (x_{jm} - x_{km})^2$, donde K_{jm} es el conjunto de productos ofrecidos en el mercado m por la empresa que produjo el producto j, y $\mathrm{z}^2_{jm} = \sum_{k \in S_{jm}} (x_{jm} - x_{km})^2$, donde S_{jm} es el conjunto de productos ofrecidos en el mercado m por todas las empresas, excepto la empresa que produjo el producto j.

En otros contextos, otros instrumentos son apropiados. Goolsebee y Petrin (2004) examinaron las elecciones de los hogares entre las opciones disponibles de canales de televisión, siendo cada ciudad un mercado distinto con diferentes precios y diferentes características distintas del precio, tanto para TV por cable y como por difusión aérea. Siguiendo la práctica sugerida por Hausman (1997), utilizaron los precios ofrecidos en otras ciudades por la misma compañía como instrumentos para cada ciudad. Este instrumento para la ciudad m es $\mathrm{z}_{jm} = \sum_{m' \in K_{jm}} p_{jm'}$, donde K_m es el conjunto de otras ciudades que son atendidas por el operador franquiciado en la ciudad m. El concepto que origina este planteamiento es que los atributos no observados de la televisión por cable en una ciudad determinada (como la

[v] Este supuesto es en gran parte algo conveniente, ya que en general se podría esperar que los atributos no observados de un producto estuviesen relacionados no sólo con el precio sino también con los atributos observados distintos del precio. Sin embargo, un modelo en el que todos los atributos observados son tratados como endógenos dejan poco margen para utilizar instrumentos

[vi] En vez de los promedios, también pueden usarse las sumas, con el número de productos de la misma marca y de otras marcas (es decir, los denominadores de los promedios) entrando también como instrumentos.

calidad de la programación para la franquicia en esa ciudad) están correlacionados con el precio de la televisión por cable en esa ciudad, pero no están correlacionados con el precio de la televisión por cable en otras ciudades. Ellos también incluyeron la tasa de la franquicia de cable impuesta por la ciudad (es decir, impuestos) y la densidad de población de la ciudad.

13.2.4 Estimación por GMM

Como se dijo anteriormente, es posible usar varios métodos para la estimación, no sólo máxima verosimilitud. BLP (1995,2004), Nevo (2001) y Petrin (2002) utilizaron un estimador basado en un método generalizado de momentos (GMM). Este procedimiento es una versión generalizada del método de momentos simulados (MSM) que se describe en el capítulo 10, aumentado por los momentos de la ecuación de regresión. Las condiciones de momentos se crean a partir de las probabilidades de elección como

(13.5) $$\sum_n \sum_j \left(d_{njm} - P_{njm}\right) z_{njm} = 0,$$

donde d_{njm} es la variable dependiente, que es 1 si el consumidor n en el mercado m escoge la alternativa j y 0 en caso contrario, y z_{njm} es un vector de instrumentos que varía entre productos y mercados (como los atributos observados distintos del precio y las funciones de los mismos), así como entre consumidores en cada mercado (tales como las variables demográficas interactuando con atributos distintos del precio). En la estimación, las probabilidades exactas se sustituyen por su versión simulada. Tenga en cuenta que esta es la misma fórmula que la mostrada en la sección 10.2.2, en relación a la estimación MSM de modelos de elección. Estas condiciones de momentos, cuando se satisfacen, implican que la media observada de los instrumentos para las alternativas elegidas, $\sum_{\mathrm{n}} \sum_{\mathrm{j}} d_{njm} z_{njm} / N_m$, es igual a la media predicha por el modelo, $\sum_{\mathrm{n}} \sum_{\mathrm{j}} P_{njm} z_{njm} / N_m$. Observe que, tal como se describe en la sección 3.7, para un modelo logit estándar, estas condiciones de momentos son la condición de primer orden para la estimación de máxima verosimilitud, donde los instrumentos son las variables explicativas del modelo. En otros modelos, esta condición de momentos no es igual a la condición de primer orden de la máxima verosimilitud, de tal manera que se produce una pérdida de eficiencia. Sin embargo, tal como se describe en el capítulo 10 en la comparación entre MSL y MSM, la simulación de estos momentos es no sesgada dado un simulador no sesgado de la probabilidad de elección, de manera que MSM es consistente para un número fijo de valores extraídos al azar en la simulación. En contraste, MSL es consistente sólo cuando se considera que el número de valores al azar aumenta con el tamaño de la muestra

La condición de momentos para la ecuación de regresión se crean como

$$\sum_j \sum_m \xi_{jm} z_{jm} = 0,$$

donde z_{jm} son instrumentos que varían entre productos y mercados, pero no entre las personas en cada mercado (como los atributos observados distintos del precio y las funciones de los mismos). Estos momentos pueden ser re-escritos para incluir los parámetros del modelo de forma explícita, suponiendo una especificación lineal para $\bar{V}$:

(13.6) $$\sum_j \sum_m \left[\delta_{jm} - \bar{\beta}' \bar{v}\left(p_{jm}, x_{jm}\right)\right] z_{jm} = 0.$$

Como vimos en la sección previa, estos momentos definen el estimador de variables instrumentales estándar de los coeficientes de regresión.

Los parámetros del sistema son $\bar{\beta}$, que capturan elementos de preferencias que son iguales para todos los consumidores, y θ, que representa la variación en las preferencias entre consumidores. Las constantes δ_{jm} también se pueden considerar parámetros, ya que se estiman junto con los otros parámetros. Alternativamente, se pueden considerar funciones de los otros parámetros, como vimos en la subsección anterior, calculadas para igualar las cuotas predichas y reales para cualesquiera valores dados de los otros parámetros. Para las distinciones en los párrafos siguientes, consideramos que los parámetros son $\bar{\beta}$ y θ, sin las constantes. Bajo esta terminología, la estimación se realiza de la siguiente manera.

Si el número de condiciones de momentos en (13.5) y (13.6) combinados es igual al número de parámetros (y tiene una solución), entonces el estimador se define como el valor de los parámetros que satisface todas las condiciones de momentos. El estimador es el MSM descrito en el capítulo 10 aumentado con momentos adicionales para la regresión. Si el número de condiciones de momentos supera el número de parámetros, entonces no hay ningún conjunto de valores de los parámetros que pueda satisfacer todas las condiciones. En este caso, el número de condiciones independientes se reduce mediante el uso de un método generalizado de momentos (GMM)[vii]. El estimador GMM se describe más fácilmente definiendo los momentos específicos de observación:

$$g^1_{njm} = \left(d_{njm} - P_{njm}\right)z_{njm},$$

que son los términos en (13.5), y

$$g^2_{njm} = \left[\delta_{jm} - \bar{\beta}'\bar{v}\left(p_{jm}, x_{jm}\right)\right]z_{jm},$$

que son los términos en (13.6). Agrupe estos dos vectores en uno solo, $g_{njm} = \langle g^1_{njm}, g^2_{njm}\rangle$, señalando que el segundo conjunto de momentos se repite para cada consumidor en el mercado m. Las condiciones de momentos se pueden escribir de manera sucinta como $\mathrm{g} = \sum_n \sum_j g_{njm} = 0$. El estimador GMM es el valor del parámetro que minimiza la forma cuadrática $g'\Theta^{-1}g$, donde Θ es una matriz de ponderación definida positiva. La covarianza asintótica de g es la matriz de ponderación óptima, calculada como $\sum_n \sum_j g_{njm}g'_{njm}$. Ruud (2000, capítulo 21) proporciona un estudio útil sobre estimadores GMM.

13.3 Lado de la oferta

El lado de la oferta de un mercado puede ser importante por varias razones. En primer lugar, en la medida en que los precios del mercado están determinados por la interacción entre la oferta y la demanda, la predicción de los resultados del mercado debido a un cambio en las condiciones del mismo requiere una comprensión tanto de la oferta como de la demanda. Un ejemplo importante se plantea en el contexto del análisis antimonopolio de las fusiones de empresas. Antes de la fusión, cada empresa fija los precios de sus propios productos con el fin de maximizar sus propios beneficios. Cuando dos empresas se fusionan, establecen los precios de sus productos para maximizar su beneficio conjunto, lo que puede producir, y por lo general así sucede, precios diferentes a los existentes cuando las empresas competían entre sí. Una tarea central del Departamento de Justicia (*Department of Justice*) y la Comisión Federal de Comercio (*Federal Trade Commission*) en el momento de decidir si aprueba o no

[vii] Cuando se utilizan probabilidades simuladas en las condiciones, el procedimiento puede ser denominado quizá de forma más exacta como estimador mediante el método generalizado de momentos simulados (GMSM). Sin embargo, nunca he visto usar este término; en su lugar, GMM se utiliza para denotar tanto los momentos simulados y como no simulado

una fusión es pronosticar el impacto de la misma sobre los precios. Esta tarea conlleva generalmente modelar el lado de la oferta, es decir, los costos marginales y la política de precios de las empresas, así como la demanda de cada producto como una función del precio.

En segundo lugar, es de esperar en muchos contextos que las empresas fijen sus precios, no sólo en base al costo marginal o a algún margen aplicado sobre el costo marginal, sino también sobre la base de la demanda de su producto y el impacto de los cambios de precios en su demanda. En estos contextos, los precios observados contienen información acerca de la demanda de los productos y la elasticidad respecto al precio. Esta información, si se extrae correctamente, puede ser utilizada en la estimación de los parámetros de la demanda. El investigador podría, por lo tanto, optar por examinar la oferta como un aspecto de la estimación de la demanda, aun cuando entre los objetivos del investigador no esté la predicción de la oferta *per se*.

En los párrafos siguientes, se describen varios tipos de comportamiento de los precios que pueden observarse en los mercados y se muestra cómo se puede combinar la especificación de este comportamiento con la estimación de la demanda vista anteriormente. En cada caso, se requieren algunas suposiciones sobre el comportamiento de las empresas. El investigador debe decidir, por lo tanto, si incorpora la oferta en el análisis bajo estas suposiciones o si por el contrario estima la demanda sin considerar la oferta, utilizando los métodos descritos en la sección anterior. El investigador se enfrenta a una disyuntiva inherente. Incorporar la oferta en el análisis tiene el potencial de mejorar las estimaciones de la demanda y ampliar el uso del modelo. Sin embargo, implica hacer suposiciones sobre el comportamiento de los precios que podrían no ser reales, de forma que estimar la demanda sin considerar la oferta podría ser más seguro.

13.3.1 Costo Marginal

Un principio básico de la teoría económica es que los precios dependen del costo marginal (*marginal cost*, MC). El costo marginal de un producto depende de sus atributos, incluyendo tanto los atributos que son observados por parte del investigador, x_{jm}, como los que no son observador, ξ_{jm}. El costo marginal también depende de los precios de los medios de producción, como alquileres y salarios, y otras "palancas de cambio de los costos" (*cost shifters*). El investigador observa algunas de estas palancas de cambio de los costos, etiquetadas como c_{jm} y no observa otras. A menudo se asume que el costo marginal es separable en términos no observadas, de tal manera que toma la forma de una regresión:

$$MC_{jm} = W(x_{jm}, c_{jm}, \gamma) + \mu_{jm}, \tag{13.7}$$

donde W(·) es una función con parámetros γ. El término de error en esta ecuación, μ_{jm}, depende en general de los atributos no observados ξ_{jm} y de otras palancas de cambio de los costos no observadas.

Los costos marginales se relacionan con los precios de manera diferente, dependiendo del mecanismo de fijación de precios que esté operando en el mercado. Examinamos a continuación las posibilidades más habituales.

13.3.2 Precios MC

En los mercados con competencia perfecta, el precio de cada producto es presionado a la baja, en equilibrio, hacia su costo marginal. Por lo tanto, los precios en un mercado competitivo con costos marginales separables son

$$p_{jm} = W(x_{jm}, c_{jm}, \gamma) + \mu_{jm}, \tag{13.8}$$

Esta ecuación de precios puede ser utilizada conjuntamente con cualquiera de las dos formas descritas anteriormente para la estimación de la demanda, cada uno de los cuales se abordan a continuación.

MSL e IV, con precios MC

Supongamos que se utiliza el método descrito en la sección 13.2.3; es decir, el modelo de elección se estima por máxima verosimilitud y luego se hace una regresión de las constantes estimadas respecto a los atributos del producto, usando variables instrumentales para tener en cuenta la endogeneidad de los precios. Con esta estimación, las variables que entran en la ecuación de precios MC son los instrumentos adecuados para usar en las variables instrumentales. Cualquieras palancas de cambio de costos observadas c_{jm} sirven como instrumentos, junto con los atributos observados distintos del precio x_{jm}. En esta configuración, la ecuación de precios simplemente proporciona información acerca de los instrumentos adecuados. Como la ecuación de precios no depende de parámetros de la demanda, no proporciona ninguna información, más allá de los instrumentos, que pueda mejorar la estimación de los parámetros de la demanda.

El investigador podría querer estimar la ecuación de precios a pesar de que no incluya parámetros de la demanda. Por ejemplo, podría querer predecir los precios y la demanda en condiciones de cambio en las palancas de costos. En este caso, la ecuación (13.8) se estima mediante mínimos cuadrados ordinarios. La estimación de la demanda y la oferta se realiza tres pasos: (1) se estima el modelo de elección mediante MSL, (2) se estima la regresión de las constantes de atributos observados del producto mediante variables instrumentales y (3) se estima la ecuación de precios mediante mínimos cuadrados ordinarios. De hecho, si las variables que entran en la ecuación de fijación de precios son los únicos instrumentos, entonces la ecuación de precios se calcula implícitamente como parte del paso 2, incluso si el investigador no estima la ecuación de precios de manera explícita en el paso 3. Recuerde de los textos estándar relativos a regresión que la estimación de variables instrumentales es equivalente la de mínimos cuadrados en dos etapas. En nuestro contexto, las variables instrumentales en el paso 2 se pueden obtener en dos etapas: (i) estimando la ecuación de precios por mínimos cuadrados ordinarios y utilizando los coeficientes estimados para predecir los precios, y luego (ii) haciendo una regresión de las constantes de los atributos del producto por mínimos cuadrados ordinarios utilizando el precio pronosticado como variable explicativa en lugar del precio observado. Por lo tanto, el proceso conjunto de estimación es el siguiente : (1) estimamos el modelo de elección, (2) estimamos la ecuación de precios por mínimos cuadrados ordinarios y (3) estimamos la ecuación para las constantes por mínimos cuadrados ordinarios utilizando los precios predichos en lugar de los precios reales.

GMM con precios MC

Si se utiliza la estimación GMM, las variables explicativas en la ecuación de precios se convierten en instrumentos $z_{jm} = \langle x_{jm}, c_{jm} \rangle$, que entran en los momentos indicados en las ecuaciones (13.5) y (13.6). La ecuación de precios proporciona condiciones de momentos adicionales basados en estos instrumentos:

$$\sum_j \sum_m \left[p_{jm} - W\left(x_{jm}, c_{jm}, \gamma\right) \right] z_{jm} = 0.$$

Los parámetros de la demanda pueden ser estimados sin estos momentos adicionales. O estos momentos adicionales pueden ser incluidos en la estimación GMM, junto con los definidos en (13.5) y (13.6). El estimador es el valor de los parámetros del modelo de demanda y de la ecuación de precios que minimiza $g' \Theta^{-1} g$, donde g incluye los momentos de la ecuación de precios y Θ incluye su covarianza.

13.3.3 Margen fijo sobre el costo marginal

Las empresas podrían fijar sus precios como un margen fijo sobre el costo marginal, siendo este margen independiente de la demanda. Esta forma de fijación de precios es considerada por la gente de negocios como una práctica omnipresente. Por ejemplo, Shim y Sudit (1995) informan de que esta forma de fijación de precios es utilizada por más del 80 por ciento de los directivos de las empresas fabricantes. Por supuesto, es difícil saber cómo interpretar las declaraciones de los directivos acerca de su política de precios, ya que cada gerente puede pensar que ellos fijan sus precios con un margen fijo por encima del costo marginal y sin embargo, el tamaño de este margen " fijo" puede ser que varíe entre directivos en relación con la demanda de los productos de los distintos fabricantes.

Esta forma de fijación de precios tiene las mismas implicaciones para la estimación de la demanda que los precios MC. El precio es $p_{jm} = \mathrm{k}MC_{jm}$ por alguna constante k. La ecuación de fijación de precios es la misma de antes, la ecuación (13.8), pero con los coeficientes incorporando ahora el factor k. Todos los otros aspectos de la estimación, ya sea usando MSL y IV o GMM, son los mismas que en el caso de precios MC.

13.3.4 Precios de monopolio y equilibrio de Nash para empresas con un solo producto

Considere una situación en la que cada producto es ofrecido por una empresa independiente. Si sólo hay un producto en el mercado, entonces la empresa es monopolista. Se supone que el monopolista fija el precio para maximizar sus beneficios, manteniendo fijos todos los demás precios de la economía. Si hay múltiples productos, entonces las empresas son oligopolistas. Asumimos que cada oligopolista maximiza sus beneficios, dados los precios del resto de empresas. Este supuesto implica que cada oligopolista utiliza la misma condición de maximización de beneficios de un monopolista, manteniendo fijos todos los demás precios (incluidos los precios de sus rivales). El equilibrio de Nash se produce cuando ninguna empresa es inducida a cambiar su precio, dados los precios del resto de empresas.

Sean p_j y q_j el precio y la cantidad del producto j, donde el subíndice que se refiere a los mercados se omite por conveniencia, ya que estamos examinando el comportamiento de las empresas en un mercado determinado. Las ganancias de la empresa son $\pi_j = p_j q_j - \mathrm{TC}(q_j)$, donde TC es el costo total. Los beneficios son máximos cuando

$$\frac{d\pi_j}{dp_j} = 0$$

$$\frac{d(p_j q_j)}{dp_j} - MC_j \frac{dq_j}{dp_j} = 0$$

$$q_j - p_j \frac{dq_j}{dp_j} = MC_j \frac{dq_j}{dp_j}$$

$$p_j + q_j \left(\frac{dq_j}{dp_j}\right)^{-1} = MC_j \tag{13.9}$$

$$p_j + p_j \frac{q_j}{p_j} \left(\frac{dq_j}{dp_j}\right)^{-1} = MC_j$$

$$p_j + \left(p_j / e_j\right) = MC_j,$$

donde MC_jes el costo marginal del producto j y e_j es la elasticidad de la demanda para el producto j respecto a su precio. Esta elasticidad depende de todos los precios, debido a que la elasticidad para un producto es diferente para diferentes niveles de precio (y por lo tanto, diferentes cantidades) para ese producto así como los otros productos. Tenga en cuenta que la elasticidad es negativa, lo que implica que $p_j + \left(p_j / e_j\right)$ es menor que p_j de tal manera que el precio está por encima del costo marginal, como es de esperar[viii]. Sustituyendo la especificación en (13.7) para MC y añadiendo subíndices para los diferentes mercados, tenemos

(13.10)
$$p_{jm} + \left(p_{jm} / e_{jm}\right) = W\left(x_{jm}, c_{jm}, \gamma\right) + \mu_{jm},$$

que es igual a la ecuación de precios bajo precios MC, (13.8), excepto que el lado izquierdo añade $\left(p_{jm} / e_{jm}\right)$ al precio.

Los parámetros de costos marginales pueden estimarse después de los parámetros de demanda. En el contexto de MSL y IV, el proceso consiste en los mismos tres pasos que usamos con precios MC: (1) estimar el modelo de elección mediante MSL, (2) estimar la regresión de las constantes respecto al precio y a otros atributos mediante variables instrumentales, y (3) estimar la ecuación de precios mediante mínimos cuadrados ordinarios. El único cambio es que ahora la variable dependiente en la ecuación de precios no es el precio en sí, sino $p_{jm} + \left(p_{jm} / e_{jm}\right)$. Este término incluye la elasticidad de la demanda, que no es observado directamente. Sin embargo, dados unos parámetros estimados de la demanda en los pasos 1 y 2, la elasticidad de la demanda puede ser calculada y utilizada en el paso 3.

El hecho de que la ecuación de fijación de precios incluya la elasticidad de la demanda implica que los precios observados contienen información acerca de los parámetros de demanda. El procedimiento de estimación secuencial que acabamos de describir no utiliza esta información. En el contexto de GMM, utilizar esta información adicional es sencillo, al menos conceptualmente. Se definen condiciones de momentos adicionales a partir de la ecuación de precios como

$$\sum_j \sum_m \left(p_{jm} + \left(p_{jm} / e_{jm}\right) - W\left(x_{jm}, c_{jm}, \gamma\right)\right) z_{jm} = 0.$$

La única diferencia con el procedimiento utilizado con precios MC es que ahora las condiciones de momentos incluyen el término adicional $\left(p_{jm} / e_{jm}\right)$. En cada valor de prueba de los parámetros en la estimación GMM, la elasticidad se calcula y se inserta en este momento.

13.3.5 Precios de monopolio y equilibrio de Nash para empresas multiproducto

Ahora generalizaremos el análisis de la sección anterior para permitir que cada empresa pueda vender más de un producto en un mercado. Si sólo hay una empresa que ofrece todos los productos en el mercado, esta empresa es un monopolista multiproducto. De lo contrario, el mercado es un oligopolio con empresas multiproducto. El mercado de los vehículos nuevos es un ejemplo representativo, donde cada fabricante, como Toyota, ofrece numerosas marcas y modelos de vehículos. La regla de fijación de precios se diferencia de la situación de un único producto por empresa en que ahora cada empresa

[viii] La condición puede ser reorganizada para que tome la forma que utilizada a menudo para monopolistas y oligopolistas de un único producto: $\left(p_j - MC_j\right)/p_j = -1/e_j$, donde el margen expresado como un porcentaje del precio está inversamente relacionado con la magnitud de la elasticidad.

debe tener en cuenta el impacto de su precio de un producto en la demanda de sus otros productos. Emplearemos el supuesto estándar, como antes, de que cada empresa fija precios con el fin de maximizar los beneficios, dados los precios de los productos de las otras empresas.

Considere una empresa que ofrece un conjunto de k productos. Los beneficios de la empresa son $\pi = \sum_{\mathrm{j}\in\mathrm{K}} p_j q_j - \mathrm{TC}(q_j\ \forall \mathrm{j} \in \mathrm{K})$. La empresa elige el precio del producto $\mathrm{j} \in \mathrm{K}$ que maximiza sus beneficios:

$$d\pi/dp_j = 0$$

$$q_j + \sum_{k\in K} (p_k - MC_k)\left(dq_k/dp_j\right) = 0.$$

Condiciones análogas aplican simultáneamente a todos los productos de la empresa. Las condiciones para todos los productos de la empresa se pueden expresar de forma sencilla mediante el uso de notación matricial. Agrupemos los precios de los K productos en el vector p, las cantidades en el vector q y los costos marginales en el vector MC. Definimos una matriz $\mathrm{K} \times \mathrm{K}$ con las derivadas de la demanda respecto al precio, D, donde el elemento (i,j)-ésimo es $\left(dq_j/dp_i\right)$. Siendo así, los precios que maximizan los beneficios de los productos de la empresa satisfacen

$$q + D(p - MC) = 0$$

$$D^{-1}q + p - MC = 0$$

$$p + D^{-1}q = MC.$$

Tenga en cuenta que esta última ecuación simplemente es una versión generalizada de la regla del monopolista mono-producto dada en (13.9). Se maneja de la misma manera en la estimación. Con estimación MSL y IV, se estima después del modelo de la demanda, usando los parámetros de la demanda para calcular D^{-1}. Con GMM, la ecuación de precios se incluye como una condición de momentos adicional, calculando D^{-1} para cada valor de prueba de los parámetros.

13.4 Funciones de control

El enfoque BLP no siempre es aplicable. Si las cuotas de mercado observadas para algunos productos en algunos mercados son cero, el enfoque BLP no puede ser implementado, ya que las constantes para estos mercados de productos no están identificadas. (Cualquier constante finita genera una cuota de mercado prevista estrictamente positiva, que superaría la cuota real de cero). Un ejemplo es el estudio a cargo de Martin (2008) sobre la elección que hacen los consumidores entre bombillas incandescentes y bombillas fluorescentes compactas (*compact fluorescent lightbulbs*, CFLs), en el que la publicidad y las promociones se produjeron con una frecuencia semanal y con variación entre diferentes tiendas y, sin embargo, era común que una tienda no vendiese ningún CFL en una semana determinada. La endogeneidad también puede surgir entre los propios decisores y no sólo entre mercados (es decir, grupos de decisiones), de tal manera que la endogeneidad puede no quedar absorbida por las constantes de producto-mercado. Por ejemplo, supongamos que las personas a las que les gusta el transporte público optan por vivir cerca de zonas de tráfico de tal manera que el tiempo de tránsito en su elección del medio de transporte es endógeno. No es posible estimar constantes para cada decisor, ya que las constantes serían infinitas (para la alternativa elegida) e negativamente infinitas (para las alternativas no elegidas), prediciendo perfectamente las elecciones, sin dejar información para la estimación de parámetros. Aún cuando la

aproximación BLP pudiese implementarse, el investigador podría querer evitar la complicación de aplicar la contracción que habitualmente se requiere en el enfoque BLP.

Algunas de las alternativas que se han implementado para resolver este problema incluyen el enfoque basado en la función de control, que se aborda en esta sección, y una versión específicamente diseñada de la aproximación mediante función de control que utiliza máxima verosimilitud con información completa, que se trata en la siguiente sección.

La configuración para el enfoque basado en la función de control sigue muy de cerca la especificación de modelos de regresión de ecuaciones simultáneas. Sin embargo, dado que los modelos de elección son no lineales, es necesario tratar una capa adicional de complejidad, y esta complejidad adicional puede restringir la aplicabilidad del método más de lo que podríamos esperar de entrada dada la analogía con los modelos lineales.

Denotemos la variable explicativa endógena para el decisor n y la alternativa j como y_{nj}. No diferenciamos los mercados, como en el enfoque BLP, porque permitimos la posibilidad de que las variables endógenas varíen entre decisores y no sólo entre mercados. La variable endógena podría ser el precio, el tiempo de tránsito o lo que sea relevante en un caso concreto dado. En las siguientes secciones, se abordan los problemas que pueden surgir cuando la variable endógena es el precio.

La utilidad que el consumidor n obtiene del producto j se expresa como

$$U_{nj} = V(y_{nj}, x_{nj}, \beta_n) + \varepsilon_{nj}, \qquad (13.11)$$

donde x_{nj} son variables exógenas observadas relativas a la persona n y el producto j (incluyendo datos demográficos observados). El término no observado ε_{nj} no es independiente de y_{nj} como se requiere para la estimación estándar. Dejemos que la variable explicativa endógena se exprese como una función de los instrumentos observados y de los factores no observados:

$$y_{nj} = W(z_{nj}, \gamma) + \mu_{nj}, \qquad (13.12)$$

donde ε_{nj} y μ_{nj} son independientes de z_{nj}, pero μ_{nj} y ε_{nj} están correlacionados. La correlación entre μ_{nj} y ε_{nj} implica que y_{nj} y ε_{nj} están correlacionados, que es lo que nos concierne. Asumimos para nuestra explicación inicial que μ_{nj} y ε_{nj} son independientes para todo los $k \neq j$.

Consideremos ahora la distribución de ε_{nj} condicionada a μ_{nj}. Si esta distribución condicionada toma una forma conveniente, entonces podemos utilizar un enfoque de función de control para estimar el modelo. Descompongamos ε_{nj} en su media condicionada a μ_{nj} y en desviaciones alrededor de esta media: $\varepsilon_{nj} = E(\varepsilon_{nj}|\mu_{nj}) + \tilde{\varepsilon}_{nj}$. Por construcción, las desviaciones no se correlacionan con μ_{nj} y por lo tanto no se correlacionan con y_{nj}. La esperanza condicionada es una función de μ_{nj} (y quizás de otras variables); se denomina la función de control y se denotada como $CF(\mu_{nj}, \lambda)$, donde λ son los parámetros de esta función. El caso más simple es aquel en el que $E(\varepsilon_{nj}|\mu_{nj}) = \lambda\mu_{nj}$, de manera que la función de control es simplemente μ_{nj} por un coeficiente a estimar. Expondremos las diferentes motivaciones para usar diversas funciones de control más adelante. Sustituyendo la media condicionada y las desviaciones en la ecuación de utilidad, tenemos

$$U_{nj} = V(y_{nj}, x_{nj}, \beta_n) + CF(\mu_{nj}, \lambda) + \tilde{\varepsilon}_{nj}. \qquad (13.13)$$

Las probabilidades de elección se obtienen de la distribución condicionada de las desviaciones $\tilde{\varepsilon}_{nj}$. Definamos $\tilde{\varepsilon}_n = \langle \tilde{\varepsilon}_{nj} \forall j \rangle$ y $\mu_n = \langle \mu_{nj} \forall j \rangle$. La distribución condicionada de $\tilde{\varepsilon}_n$ se denota como $g(\tilde{\varepsilon}_n|\mu_n)$ y la distribución de β_n es $f(\beta_n|\theta)$. La probabilidad de elección es entonces

$$P_{nj} = Prob\big(U_{nj} > U_{nk}\ \forall k \neq j\big)$$

$$(13.14) \quad = \int\int I(V_{nj} + CF_{nj} + \tilde{\varepsilon}_{nj} > V_{nk} + CF_{nk} + \tilde{\varepsilon}_{nk}\ \forall k \neq j) g(\tilde{\varepsilon}_n|\mu_n) f(\beta_n|\theta) d\tilde{\varepsilon} d\beta_n,$$

donde se utilizan las siguientes abreviaturas:

$$V_{nj} = V(p_{nj}, x_{nj}, \beta_n)$$

$$CF_{nj} = CF\big(\mu_{nj}, \lambda\big).$$

Este es un modelo de elección como cualquier otro, con la función de control formando parte como variable explicativa adicional. Tenga en cuenta que la integral es sobre la distribución condicionada de $\tilde{\varepsilon}$ en lugar del ε original. Por construcción, $\tilde{\varepsilon}$ no está correlacionado con la variable endógena, mientras que el ε original sí lo estaba. Básicamente, la parte de ε que está correlacionada con y_{nj} se introduce explícitamente como variable explicativa adicional, concretamente, la función de control, de tal manera que la parte restante no está correlacionada.

El modelo se estima en dos etapas. En primer lugar, se estima la ecuación (13.12). Es una regresión con la variable endógena como variable dependiente y con instrumentos exógenos como variables explicativas. Los residuos de esta regresión proporcionan estimaciones de las μ_{nj}s. Estos residuos se calculan como $\hat{\mu}_{nj} = y_{nj} - W(z_{nj}, \hat{\gamma})$ usando los parámetros estimados $\hat{\gamma}$. En segundo lugar, el modelo de elección se estima con $\hat{\mu}_{nj}$ entrando en la función de control. Es decir, las probabilidades de elección en (13.14) se estiman por máxima verosimilitud, con $\hat{\mu}_{nj}$ y/o una función paramétrica de $\hat{\mu}_{nj}$ entrando como variables explicativas adicionales.

La cuestión central con el enfoque de la función de control es la especificación de la propia función de control y de la distribución condicionada de $\tilde{\varepsilon}_n$. En algunas situaciones, hay formas naturales de especificar estos elementos del modelo. En otras situaciones, es difícil o incluso imposible, especificarlos de una manera que represente la realidad de forma significativa. La aplicabilidad del enfoque depende de que el investigador sea capaz de especificar estos términos de manera significativa.

Algunos ejemplos de especificaciones de la función de control son las siguientes:

1. Sea

   $U_{nj} = V\big(y_{nj}, x_{nj}, \beta_n\big) + \varepsilon_{nj}$

   $y_{nj} = W\big(z_{nj}, \gamma\big) + \mu_{nj}$

   y asumamos que ε_{nj} y μ_{nj} siguen una distribución normal conjunta con media cero y matriz de covarianza constante para todo j. Debido a las propiedades de las distribuciones normales, la esperanza de ε_{nj} condicionada a μ_{nj} es $\lambda\mu_{nj}$, donde λ refleja la covarianza, y las desviaciones alrededor de la media, $\tilde{\varepsilon}_{nj}$, son normales con varianza constante. En este caso, la función de control es $\mathrm{CF}\big(\mu_{nj}, \lambda\big) = \lambda\mu_{nj}$. La utilidad es

   $U_{nj} = V\big(y_{nj}, x_{nj}, \beta_n\big) + \lambda\mu_{nj} + \varepsilon_{nj}$

   El modelo de elección es probit, con el residuo de la ecuación de precios como una variable adicional. Tenga en cuenta sin embargo que la varianza de $\tilde{\varepsilon}_{nj}$ difiere de la varianza de ε_{nj}, de manera que la escala en el modelo probit estimado es diferente de la escala original. Si $\beta_n$es aleatorio, entonces el modelo es un probit mixto.

2. Supongamos que ε_{nj} consta de una parte distribuida normalmente que se correlaciona con y_{nj} y una parte que se distribuye valor extremo iid. En concreto,

   $$U_{nj} = V(y_{nj}, x_{nj}, \beta_n) + \varepsilon_{nj}^1 + \varepsilon_{nj}^2$$

   $$y_{nj} = W(z_{nj}, \gamma) + \mu_{nj}$$

   donde ε_{nj}^1 y μ_{nj} son conjuntamente normales y ε_{nj}^2 es valor extremo iid. La distribución condicionada de ε_{nj}^1 es, como en el ejemplo anterior, normal con media $\lambda\mu_{nj}$ y varianza constante. Sin embargo, la distribución condicionada de ε_{nj}^2 es igual a su distribución no condicionada dado que ε_{nj}^2 y μ_{nj} son independientes. La utilidad se convierte en

   $$U_{nj} = V(y_{nj}, x_{nj}, \beta_n) + \lambda\mu_{nj} + \tilde{\varepsilon}_{nj}^1 + \varepsilon_{nj}^2$$

   donde $\tilde{\varepsilon}_{nj}^1$ es normal con media cero y varianza constante. Este componente de error puede expresarse como $\tilde{\varepsilon}_{nj}^1 = \sigma\eta_{nj}$, donde η_{nj}, es una normal estándar. La utilidad se convierte así en

   $$U_{nj} = V(y_{nj}, x_{nj}, \beta_n) + \lambda\mu_{nj} + \sigma\eta_{nj} + \varepsilon_{nj}^2.$$

   La probabilidad de elección es la de un modelo logit mixto, mezclado respecto a los componentes de error normales $\sigma\eta_{nj}\ \forall j$ así como a los elementos aleatorios de β_n. La desviación estándar de los errores condicionados, σ, es estimada, a diferencia del ejemplo anterior.

3. La notación se puede generalizar para permitir correlación entre ε_{nj} y μ_{nk} para $\mathrm{k} \neq \mathrm{j}$. Definamos $\varepsilon_{nj} = \varepsilon_{nj}^1 + \varepsilon_{nj}^2$ como en el ejemplo anterior, excepto que ahora supondremos que los vectores agrupados ε_n^1 y μ_n son conjuntamente normales. En este caso, ε_n^1 condicionada a μ_n es normal con media $\mathrm{M}\mu_n$y varianza Ω, donde M y Ω son matrices de parámetros. Agrupando utilidades y funciones explicativas, tenemos

   $$U_n = V(y_n, x_n, \beta_n) + \mathrm{M}\mu_n + L\eta_n + \varepsilon_n^2,$$

   donde L es el factor Choleski triangular inferior de Ω y η_n es un vector variables normales estándar iid. Dado que los elementos de $\tilde{\varepsilon}_{nj}^2$ son iid valor extremo, el modelo es logit mixto, mezclado en relación a los componentes de error η_n y a los elementos aleatorios de β_n. Los residuos de todos los productos forman parte de la utilidad de cada producto.

13.4.1 Relación con el comportamiento de los precios

Como se estableció con anterioridad, la principal limitación del enfoque de función de control es la necesidad de especificar dicha función de control así como la distribución condicionada del nuevo término no observado $\tilde{\varepsilon}$. En algunas situaciones, la verdadera distribución condicionada es tan compleja que no se puede obtener, de manera que cualquier intento de especificación que use distribuciones estándar, como la normal, es necesariamente incorrecta. Estas cuestiones se explican más fácilmente usando el ejemplo de la política de precios de las empresas, donde el precio es la variable endógena, pero pueden surgir bajo cualquier tipo de variable endógena, dependiendo de la forma en que se determine la misma.

Supongamos que la elección es entre productos y que la variable endógena y_{nj} es el precio p_{nj}, por lo que podemos hablar de distintos comportamientos de los precios y de si estos comportamientos pueden tener cabida dentro del enfoque basado en la función de control. Consideremos en primer lugar una situación en la que la función de control se puede aplicar fácilmente: precios MC. La utilidad que el consumidor n obtiene del producto j es

$$U_{nj} = V(p_{nj}, x_{nj}, \beta_n) + \varepsilon_{nj}^1 + \varepsilon_{nj}^2,$$

donde ε_{nj}^1 está correlacionada con el precio y ε_{nj}^2 se distribuye valor extremo iid. Por ejemplo, ε_{nj}^1 podría representar atributos no observados del producto. Los precios varían entre personas porque diferentes personas están en diferentes mercados o los precios se establecen por separado para personas o grupos de personas (por ejemplo, para tener en cuenta los costos de transporte de cada cliente). Supongamos, además, que las empresas fijan los precios iguales al costo marginal, que se especifica como $MC_{nj} = W(z_{nj}, \gamma) + \mu_{nj}$, donde z_{nj} son variables exógenas que afectan el costo marginal (incluyendo los atributos observados del producto) y μ_{nj} captura el efecto de palancas de cambio de los costos no observadas (incluyendo los atributos no observados del producto). La ecuación de precios es entonces

$$p_{nj} = W(z_{nj}, \gamma) + \mu_{nj}.$$

Supongamos que ε_{nj}^1 y μ_{nj} son conjuntamente normales, con la misma matriz de covarianza para todo j. Podría surgir correlación, por ejemplo, si los atributos no observados afectan a la utilidad, así como a los costos, entrando así tanto en ε_{nj}^1 como μ_{nj}. Al igual que en el segundo ejemplo dado anteriormente, la utilidad se convierte en

$$U_{nj} = V(p_{nj}, x_{nj}, \beta_n) + \lambda\mu_{nj} + \sigma\eta_{nj} + \varepsilon_{nj}^2,$$

donde η_{nj} es una normal estándar iid. El modelo se estima en dos pasos: en primer lugar, se estima la ecuación de precios y se retienen sus residuos, $\hat{\mu}_{nj}$. En segundo lugar, se estima el modelo de elección usando estos residuos como variables explicativas. El modelo es un logit mixto, mezclado respecto a los nuevos componentes de error η_{nj}. La misma especificación es aplicable si las empresas fijan precios con un margen constante sobre el costo marginal. Con esta política de precios, la ecuación de precios es lo suficientemente simple como para que suposiciones razonables sobre los términos observados en el modelo den una distribución condicionada para los términos no observados de la utilidad que pueda ser obtenida y que sea conveniente.

Consideremos ahora los precios en situación de monopolio o en equilibrio de Nash, donde el precio depende de la elasticidad de la demanda así como del costo marginal. Como se mostró anteriormente en relación con el enfoque BLP, la ecuación de precios para un monopolista mono-producto o un oligopolista de Nash es

$$p_{nj} + (p_{nj}/e_{nj}) = MC$$

$$p_{nj} = -(p_{nj}/e_{nj}) + W(z_{nj}, \gamma) + \mu_{nj}.$$

Supongamos ahora que ε_{nj}^1 y μ_{nj} presentan una distribución normal conjunta. La distribución de ε_{nj}^1 condicionada a μ_{nj} sigue siendo normal. Sin embargo, μ_{nj} ya no es el único error en la ecuación de precios y por lo tanto no podemos obtener una estimación de μ_{nj} sobre la que condicionar. A diferencia de los precios MC, el componente no observado de la demanda, ε_{nj}^1, entra en la ecuación de precios a través de la elasticidad. La ecuación de precios incluye dos términos no observados: μ_{nj} y una transformación altamente no lineal de ε_{nj}^1 entrando a través de e_{nj}.

Si reescribimos la ecuación de precios de forma que pueda ser estimada, con una parte observada y un error separable, tendríamos

$$p_{nj} = Z_j(z_{nj}, \gamma) + \mu_{nj}^*,$$

donde z_n es un vector con todas las variables exógenas observadas que afectan el costo marginal y a la elasticidad, $Z_j(\cdot)$ es una función paramétrica de estas variables y μ_{nj}^* son las desviaciones no observadas del precio en torno a esta función. Podemos estimar esta ecuación y conservar su residuo, que es una estimación de μ_{nj}^*. Sin embargo, μ_{nj}^* no es μ_{nj}; por el contrario, μ_{nj}^*incorpora tanto μ_{nj} como los componentes no observados del margen basado en la elasticidad p_{nj}/e_{nj}. La distribución de ε_{nj}^1 condicionada a este μ_{nj}^* no se ha obtenido, y podría no ser obtenible. Dada la forma en que ε_{nj}^1 entra en la ecuación de precios a través de la elasticidad, su distribución condicionada desde luego no es normal aunque su distribución no condicionada sea normal. De hecho, su distribución condicionada ni siquiera es independiente de las variables exógenas.

Villas- Boas (2007) ha propuesto una dirección alternativa para gestionar esta situación. Señala que la dificultad que nos hemos encontrado anteriormente en la especificación de una función de control apropiada surge del supuesto de que los costos marginales son separables en términos no observados, más que del supuesto de que los precios están relacionados con las elasticidades. Supongamos por el contrario que el costo marginal es una función general de los términos observados y no observados: $\mathrm{MC} = W^*\left(z_{nj}, \gamma, \mu_{nj}\right)$. En muchos sentidos, ese supuesto de términos no observados y no separables es más realista, dado que es de esperar que las palancas de cambio de costo no observadas, en general, interactúen con palancas de costo observadas. En virtud de esta función general de costo, Villas-Boas muestra que para cualquier especificación de la función de control y de la distribución de ε_{nj}^1 condicionada a esta función de control, existe una función de costo marginal $W^*(\cdot)$ y una distribución de los términos no observados μ_{nj}y ε_{nj}^1 que es consistente con ellas. Este resultado implica que el investigador puede aplicar el enfoque basado en la función de control, incluso cuando los precios dependen de la elasticidad, a sabiendas de que existe cierta función de costo marginal y unas distribuciones de los términos no observables que hacen que el enfoque sea consistente. Por supuesto, esta prueba de la existencia no proporciona ninguna orientación sobre qué función de control y qué distribución condicionada son más razonables, lo que debe seguir siendo una cuestión importante para el investigador.

13.5 Enfoque de máxima verosimilitud

El enfoque de máxima verosimilitud es similar al de la función de control, exceptuando el hecho de que los parámetros del modelo se estiman de forma simultánea en lugar de secuencialmente. Al igual que con el enfoque de función de control, la utilidad viene dada por la ecuación (13.13) y la variable explicativa endógena se especifica en (13.12). Para permitir que la notación sea más compacta, agrupamos cada uno de los términos entre alternativas, de tal manera que las dos ecuaciones resultan

(13.15) $$U_n = V(y_n, x_n, \beta_n) + \varepsilon_n$$

(13.16) $$y_n = W(z_n, \gamma) + \mu_n.$$

En lugar de especificar la distribución condicionada de ε_n dado μ_n, el investigador especifica su distribución conjunta, la cual denotamos como $\mathrm{g}(\varepsilon_n, \mu_n)$. Usando la ecuación (13.16), μ_n se puede expresar como una función de los datos y de los parámetros: $\mu_n = y_n - W(z_n, \gamma)$. De esta forma, la distribución conjunta de ε_ny y_n es $\mathrm{g}(\varepsilon_n, y_n - W(z_n, \gamma))$. Denominemos a la alternativa elegida como i. La probabilidad de los datos observados para la persona n es la probabilidad de que la variable explicativa endógena tome el valor y_n y de que la alternativa escogida sea la i. Condicionada respecto a β_n, esta probabilidad es

$$P_n(\beta_n) = \int I(U_{ni} > U_{nj}\ \forall j \neq i) g(\varepsilon_n, y_n - W(z_n, \gamma)) d\varepsilon_n.$$

Si β_n es aleatorio, entonces $P_n(\beta_n)$ se mezcla respecto a su distribución. La probabilidad resultante P_n se inserta en la función log-verosimilitud: $LL = \sum_n \ln(P_n)$. Esta LL se maximiza respecto a los parámetros del modelo. En lugar de estimar (13,16) primero y usar luego los residuos en la probabilidad de elección, los parámetros de (13.16) y el modelo de elección se estiman de forma simultánea.

El tercer ejemplo del enfoque de función de control (que es el ejemplo más general) se puede adaptar a este procedimiento de máxima verosimilitud. La utilidad agrupada es

$$U_n = V(y_n, x_n, \beta_n) + \varepsilon_n^1 + \varepsilon_n^2$$

Siendo la variable endógena agrupada:

$$y_n = W(z_n, \gamma) + \mu_n,$$

donde $W(\cdot)$ tiene ahora valores vectoriales. Cada elemento de ε_n^2 se asume distribuido tipo valor extremo iid. Asumamos que ε_n^1 y μ_n presentan una distribución conjuntamente normal con media cero y covarianza Ω. Su densidad se denota como $\phi(\varepsilon_n^1, \mu_n|\Omega)$. La probabilidad de que forma parte de la función log-verosimilitud para la persona n que eligió la alternativa i es

$$P_n = \int \int \frac{e^{V(y_{ni}, x_{ni}, \beta_n) + \varepsilon_{ni}^1}}{\sum e^{V(y_{nj}, x_{nj}, \beta_n) + \varepsilon_{nj}^1}} \phi(\varepsilon_n, \left(y_n - W(z_n, \gamma)\right)|\Omega) f(\beta_n|\theta) d\varepsilon_n d\beta_n.$$

Esta probabilidad se inserta en la función log-verosimilitud y se maximiza respecto a γ (los parámetros que relacionan la variable explicativa endógena con los instrumentos), θ (que describe la distribución de las preferencias que afectan a la utilidad) y Ω (la covarianza de los términos no observados correlacionados ε_n^1 y μ_n). Por supuesto, en cualquier caso real concreto, es posible aplicar restricciones a Ω para reducir el número de parámetros. Por ejemplo, Park y Gupta (en un trabajo que será publicado próximamente) asumen que ε_{nj}^1 y μ_{nk} no están correlacionados para $k \neq j$.

Los enfoques basados en función de control y en máxima verosimilitud proporcionan una solución de compromiso que es habitual en econometría: generalidad frente a eficiencia. El enfoque de máxima verosimilitud requiere una especificación de la distribución conjunta de ε_n y μ_n, mientras que la función de control requiere una especificación de la distribución condicionada de ε_n dado un μ_n. Cualquier distribución conjunta implica una distribución condicionada particular, pero cualquier distribución condicionada dada no implica necesariamente una distribución conjunta particular. Puede haber numerosas distribuciones conjuntas que generen una misma distribución condicionada. Por tanto, el enfoque de función de control es más general que el enfoque de máxima verosimilitud: es aplicable para cualquier distribución conjunta que sea coherente con la distribución condicionada especificada. Sin embargo, si la distribución conjunta puede especificarse correctamente, el enfoque de máxima verosimilitud es más eficiente que el de función de control, simplemente por el hecho de que es la solución de máxima verosimilitud para todos los parámetros.

13.6 Caso de estudio: elección de consumidores entre vehículos nuevos

Un caso ilustrativo del enfoque BLP lo proporcionan Train y Winston (2007). Su estudio examinó la cuestión de por qué los tres grandes fabricantes (estadounidenses) de automóviles han estado perdiendo cuota de mercado. Como parte del estudio, estimaron un modelo de elección de los compradores de

vehículos nuevos. Dado que muchos de los atributos de los vehículos nuevos no son observados por el investigador, y sin embargo afectan a los precios, es de esperar que el precio sea endógeno.

La estimación se realizó sobre una muestra de consumidores que compraron o alquilaron vehículos nuevos en el año 2000, junto con los datos relativos a las cuotas de mercado para cada marca y modelo de ese año. El análisis no incluyó un bien externo, y como tal, muestra la demanda condicionada a la compra de un vehículo nuevo. Para cada comprador muestreado, la información de la encuesta incluyó la marca y el modelo del vehículo que la persona que compró, más una lista de los vehículos que la persona declaró que tuvo en consideración en el momento de comprar. El vehículo elegido y los vehículos tenidos en cuenta fueron tratados como un ranking, con el vehículo elegido en primera posición y los vehículos considerados clasificados en el orden en que la persona los mencionó. El modelo de elección se especificó como un "logit explotado" (*exploded logit*) para la probabilidad del ranking de la persona, mezclado respecto a una distribución de coeficientes aleatorios. Vea la Sección 7.3 para una extensa explicación relativa a esta especificación para datos de ordenación o ranking. El modelo de elección incluyó constantes para cada marca y modelo de vehículo, así como variables explicativas y coeficientes aleatorios para capturar las variaciones en las preferencias entre consumidores. El análisis distinguía 200 marcas y modelos, y utilizaba la contracción para calcular las 199 constantes (con un normalizada a cero) dentro de la estimación de máxima verosimilitud de los otros parámetros. Las constantes estimadas fueron posteriormente objeto de una regresión respecto a los atributos observados de los vehículos, incluyendo el precio. Puesto que el precio fue considerado endógeno, esta regresión se estimó mediante variables instrumentales en lugar de mínimos cuadrados ordinarios.

Tabla 13.1. Modelo logit explotado para la elección de nuevo vehículo

	Parámetro	Error estándar
Precio dividido por el ingreso de los hogares		
Coeficiente medio	−1.6025	0.4260
Desviación estándar del coeficiente	0.8602	0.4143
Índice de Reparaciones Reportado por el Consumidor (Consumer Report's repair index), para mujeres de 30 o más años	0.3949	0.0588
Vehículos lujosos o deportivos, para personas que alquilan	0.6778	0.4803
Furgoneta, para hogares con adolescentes	3.2337	0.5018
Todoterrenos o familiares, para hogares con adolescentes	2.0420	0.4765
(1+número de concesionarios en un radio de 50 millas del hogar del comprador)	1.4307	0.2714
Potencia: desviación estándar del coeficiente	0.0045	0.0072
Consumo de combustible (l/mpg): desviación estándar del coeficiente	−102.15	20.181
Camioneta ligera, furgoneta o pickup: desviación estándar del coeficiente	6.8505	2.5572
Número de compras previas consecutivas de un vehículo GM	0.3724	0.1471
Número de compras previas consecutivas de un vehículo GM, si el hogar es rural	0.3304	0.2221
Número de compras previas consecutivas de un vehículo Ford	1.1822	0.1498
Número de compras previas consecutivas de un vehículo Chrysler	0.9652	0.2010
Número de compras previas consecutivas de un vehículo japonés	0.7560	0.2255
Número de compras previas consecutivas de un vehículo europeo	1.7252	0.4657
Lealtad al fabricante, componente de error: desviación estándar	0.3453	0.1712
SLL en convergencia	−1994.93	

La tabla 13.1 muestra las estimaciones de los parámetros que se refieren a la variación de las preferencias entre consumidores. Estos son los parámetros que se estimaron por máxima verosimilitud respecto a la

probabilidad del ranking de marcas y modelos de cada comprador, usando la contracción de las constantes. Los coeficientes estimados tienen las siguientes implicaciones:

- El precio dividido por los ingresos entra como variable explicativa, con el fin de captar la idea de que los hogares con ingresos más altos dan menos importancia al precio que los hogares con menores ingresos. Se asigna a la variable un coeficiente aleatorio distribuido normalmente, cuya media y desviaciones estándar se calcularon. La media estimada es negativa, como se esperaba, y la desviación estándar estimada es considerablemente grande y significativa (con un estadístico-*t* en torno a 2), lo que indica una variación considerable en respuesta al precio.
- El Índice de Reparaciones Reportado por el Consumidor (*Consumer Report's repair index*) entra para las mujeres de 30 o más años, y no para los hombres y las mujeres más jóvenes. Esta distinción fue descubierta a través de pruebas de variables demográficas alternativas interactuado con el índice de reparaciones.
- Las personas que alquilan su vehículo tienen mayor preferencia por el lujo y por los vehículos deportivos de la gente que compra su vehículo.
- Se observa que los hogares con adolescentes tienen una mayor preferencia por furgonetas, todoterrenos y coches familiares que otros hogares.
- Se encontró que las ubicaciones de los concesionarios afectan a las elecciones de los hogares. Uno de los propósitos del análisis de Train y Winston fue investigar el impacto de las ubicaciones de los concesionarios en la elección del vehículo, para ver si los cambios en las ubicaciones podían explicar parte de la pérdida de cuota de mercado de los 3 grandes fabricantes. El modelo de la demanda indica que, como era de esperar, la probabilidad de comprar una marca y un modelo determinado se eleva cuando hay más concesionarios que venden esa marca y modelo dentro de un radio de 50 millas del hogar.
- La potencia, el consumo de combustible y una variable indicadora referida a si el vehículo es una camioneta, entran con coeficientes aleatorios. Estas variables no varían entre consumidores, sólo entre marcas y modelos. Como resultado, el producto del coeficiente medio por la variable es absorbido por las constantes de marca/modelo, de forma que sólo la desviación estándar es calculada en el modelo de elección. Para ser precisos, cada una de estas variables entra a formar parte de la utilidad como $\beta_n x_j$, con β_n aleatorio. Descomponemos β_n en su media $\bar{\beta}$ y sus desviaciones $\tilde{\beta}_n$. El impacto medio $\bar{\beta} x_j$ no varía entre consumidores y se convierte en parte de la constante del vehículo j. Las desviaciones $\tilde{\beta}_n x_j$ entran en el modelo de elección por separado de las constantes, y la desviación estándar de $\tilde{\beta}_n$ es estimada. Las estimaciones indican una considerable variación en las preferencias de los consumidores respecto a la potencia, la eficiencia de combustible y al hecho de que los vehículos sean camionetas.
- El último conjunto de variables captura la lealtad de los consumidores hacia los fabricantes. Cada variable se especifica como el número de compras pasadas consecutivas de un vehículo de un fabricante determinado (o grupo de fabricantes). Los coeficientes estimados indican que la lealtad a los fabricantes europeos es mayor, seguida por la lealtad a Ford. Curiosamente, los hogares rurales resultan ser más leales a GM que los hogares urbanos. Estas variables de fidelidad son un tipo de variable dependiente diferida, que introducen aspectos econométricos debido a la posibilidad de que existan factores no observados correlacionados en serie. Winston y Train tratan estos temas y tienen en cuenta la correlación en serie, al menos en parte, en su procedimiento de estimación.

El modelo de elección de la tabla 13.1 incluye constantes para cada marca y modelo de vehículo. Las constantes estimadas fueron objeto de una regresión respecto a los atributos del vehículo para estimar

los aspectos de la demanda que son comunes a todos los consumidores. La tabla 13.2 muestra los coeficientes estimados de esta regresión, utilizando variables instrumentales para tener en cuenta la endogeneidad del precio.

Tabla 13.2. Regresión de constantes respecto a atributos de los vehículos

	Parámetro	Error estándar
Precio (precio de venta recomendado por el fabricante, en miles de dólares)	−0.0733	0.0192
Potencia dividida por peso (en toneladas)	0.0328	0.0117
Cambio automático	0.6523	0.2807
Ancho (en pulgadas)	0.0516	0.0127
Largo menos ancho (en pulgadas)	0.0278	0.0069
Consumo de combustible (en galones por milla)	−31.641	23.288
Vehículo de lujo o deportivo	−0.0686	0.2711
Todoterreno o familiar	0.7535	0.4253
Furgoneta	−1.1230	0.3748
Camioneta pickup	0.0747	0.4745
Chrysler	0.0228	0.2794
Ford	0.1941	0.2808
General Motors	0.3169	0.2292
Europeo	2.4643	0.3424
Coreano	0.7340	0.3910
Constante	−7.0318	1.4884
R-cuadrado	0.394	

Como era de esperar, el precio entra en el modelo con un coeficiente negativo, lo que indica que a los consumidores no les gusta el precio más alto, manteniendo todo lo demás igual (es decir, suponiendo que no hay cambios en los atributos de un vehículo para compensar el precio más alto). Curiosamente, cuando la regresión se estimó por mínimos cuadrados ordinarios, haciendo caso omiso de la endogeneidad, el coeficiente de precio estimado fue considerablemente menor: -0.0434 en comparación con la estimación de -0.0733 obtenida utilizando variables instrumentales. La dirección de esta diferencia es la esperada, ya que una correlación positiva entre el precio y los atributos no observados crea un sesgo a la baja en la magnitud del coeficiente de precio. El tamaño de esta diferencia indica la importancia de tener en cuenta la endogeneidad.

Los otros coeficientes estimados tienen los signos esperados. La estimación indica que los consumidores valoran la potencia adicional, como se evidencia por el coeficiente positivo de la relación potencia-peso. Los consumidores también prefieren tener transmisión automática de serie (manteniendo el precio del vehículo constante). El tamaño del vehículo se mide tanto por su distancia entre ejes como por su longitud respecto a la distancia entre ejes. Ambas medidas entran positivamente en el modelo, y la distancia entre ejes obtiene un coeficiente mayor que la longitud respecto a la distancia entre ejes. Este resultado es de esperar, ya que la distancia entre ejes es generalmente un mejor indicador del espacio interior del vehículo que la longitud. El coeficiente negativo de consumo de combustible implica que los consumidores prefieren una mayor eficiencia de combustible, lo que reduce el consumo por milla recorrida.

Tenga en cuenta que el precio entra en ambas partes del modelo: en la regresión de la tabla 13.2, que captura los impactos que son constantes entre consumidores, y en el modelo de elección de la tabla 13.1, que captura impactos que difieren entre consumidores. Tomando ambas partes conjuntamente, el precio forma parte de la utilidad con un coeficiente: $-0{,}0773 - 1.602/Ingresos + 0{,}860 * \eta/Ingresos$, donde η es un término aleatorio normal estándar. El coeficiente de precio tiene un

componente constante, una parte que varía con el ingreso del hogar y una parte que varía aleatoriamente entre hogares con el mismo nivel de ingresos.

La endogeneidad de los precios ha sido manejada adecuadamente en este caso de estudio mediante la inclusión de constantes en el modelo de elección para absorber los atributos no observados y usando posteriormente variables instrumentales al hacer la regresión de las constantes respecto al precio y a otros atributos observados, es decir, a través del enfoque BLP. Un caso de estudio sobre el enfoque de función de control lo proporcionan Petrin y Train (2009), y del enfoque de máxima verosimilitud lo proporcionan Park y Gupta (de próxima publicación).

14

Algoritmos EM

14.1 Introducción

En el capítulo 8, hemos hablado de métodos para maximizar la función log-verosimilitud (LL). A medida que los modelos se vuelven más complejos, la maximización a través de estos métodos se hace más difícil. Varias cuestiones contribuyen a esta dificultad. En primer lugar, la obtención de modelos más flexibles y realistas suele lograrse aumentando el número de parámetros. Sin embargo, los procedimientos descritos en el capítulo 8 utilizan el cálculo del gradiente respecto a cada parámetro, por lo que a medida que el número de parámetros se eleva, requieren más tiempo de cálculo. El hessiano, o el hessiano aproximado, deben ser calculados e invertidos; con un gran número de parámetros, la inversión puede ser numéricamente difícil. Asimismo, a medida que el número de parámetros crece, la búsqueda de los valores de maximización debe realizarse sobre un espacio de mayor dimensionalidad, de tal manera que la localización del máximo necesita más iteraciones. En definitiva, cada iteración usa más tiempo y se requieren más iteraciones.

En segundo lugar, la función LL para los modelos simples a menudo es aproximadamente cuadrática, de manera que los procedimientos del capítulo 8 operan de manera efectiva. Sin embargo, cuando el modelo se vuelve más complejo, la función LL generalmente se vuelve menos cuadrática, al menos en algunas regiones del espacio de parámetros. Este problema puede manifestarse de dos formas. El procedimiento iterativo puede quedarse "atrapado" en las áreas no cuadráticas de la función LL, dando pasos pequeños que no representan apenas mejoría en la LL. O el procedimiento puede "pasar de largo" el máximo repetidas veces, dando grandes pasos en cada iteración, pero sin poder localizar el máximo.

Por último, existe otro problema más fundamental que el número de parámetros y la forma de la función LL. Por lo general, cuando un investigador especifica un modelo más general, y por lo tanto complejo, lo hace porque quiere depender menos de los supuestos y, por el contrario, obtener más información de los datos. Sin embargo, el objetivo de obtener más información de los datos es intrínsecamente contrario a la simplicidad de la estimación.

Los algoritmos de maximización de la esperanza (*expectation-maximization, EM*) son procedimientos que permiten la maximización de una función LL cuando los procedimientos estándar son numéricamente difíciles o inviables. El procedimiento fue introducido por Dempster, Laird y Rubin (1977), como un mecanismo para manejar información perdida o ausente. Sin embargo, es aplicable de forma mucho más general y se ha utilizado con éxito en muchos campos de la estadística. McLachlan y

Krishnan (1997) ofrecen una revisión de casos prácticos. En el campo de los modelos de elección discreta, los algoritmos EM han sido utilizados por Bhat (1997a) y Train (2008a, b).

El procedimiento consiste en definir una esperanza o expectativa en particular, y luego maximizarla (de ahí el nombre). Esta expectativa está relacionada con la función LL de una forma que vamos a describir, pero difiere de tal manera que facilita la maximización. El procedimiento es iterativo, iniciándose en un cierto valor inicial de los parámetros y actualizando los valores en cada iteración. Los parámetros actualizados en cada iteración son los valores que maximizan la expectativa en esa iteración particular. Como se verá, la maximización repetida de esta función converge al máximo de la propia función LL.

En este capítulo se describe el algoritmo EM en general y se desarrollan algoritmos específicos para modelos de elección discreta con coeficientes aleatorios. Mostraremos que el algoritmo EM se puede utilizar para estimar distribuciones de preferencias muy flexibles, incluidas especificaciones no paramétricas que pueden aproximar asintóticamente cualquier distribución verdadera subyacente. Aplicaremos estos métodos en un caso de estudio relativo a la elección que los consumidores hacen entre vehículos de hidrógeno y vehículos de gas.

14.2 Procedimiento general

En esta sección describiremos el procedimiento EM de una manera muy general, a fin de elucidar sus características. En las secciones siguientes, aplicaremos el procedimiento general a modelos específicos. Denotemos colectivamente las variables dependientes observadas como y, representando las elecciones o la secuencia de elecciones de una muestra completa de decisores. Las elecciones dependen de variables explicativas observadas que, por conveniencia de notación, no indicamos de forma explícita. Las elecciones también dependen de datos que faltan (o datos ausentes, *missing data*), designados colectivamente como z. Dado que los valores de estos datos ausentes no son observados, el investigador especifica una distribución que representa los valores que estos datos ausentes podrían tomar. Por ejemplo, si no tenemos el nivel de ingresos de algunos individuos de la muestra, la distribución de los ingresos en la población puede ser una especificación útil para la distribución de los valores de los ingresos que no tenemos. La densidad de los datos ausentes se denota como $f(z|\theta)$, que en general depende de parámetros θ a estimar.

El modelo de comportamiento relaciona datos observados y ausentes con elecciones de decisores. Este modelo predice las elecciones que se producirían si los datos ausentes fueran realmente observados en lugar de estar ausentes. Este modelo de comportamiento se denota como $P(y|z,\theta)$, donde θ son parámetros que pueden solaparse o ampliar los de f. (Para mantener la notación compacta, utilizaremos θ para referirnos a todos los parámetros a estimar, incluyendo los que entran en f y los que entran en P). Sin embargo, dado que realmente los datos ausentes no están, la probabilidad de las elecciones observadas, utilizando la información que el investigador observa, es la integral de la probabilidad condicionada sobre la densidad de los datos ausentes[ix]:

$$P(y|\theta) = \int P(y|z,\theta) f(z|\theta) dz.$$

La densidad de los datos ausentes, $f(z|\theta)$, se utiliza para predecir las elecciones observadas, y por lo tanto no depende de y. Sin embargo, podemos obtener alguna información acerca de los datos ausentes mediante la observación de las elecciones que se hicieron. Por ejemplo, en la elección del vehículo, si los ingresos de una persona no están disponibles pero se observa que la persona ha

[ix] Asumimos en esta expresión que z es continua, de manera que la probabilidad no condicionada es una integral. Si z es discreta, o una mezcla de variables continuas y discretas, entonces la integración se sustituye por una suma sobre los valores discretos, o una combinación de integrales y sumas.

comprado un Mercedes, se puede inferir que es probable que los ingresos de esta persona estén por encima de la media. Definamos $h(z|y,\theta)$ como la densidad de los datos ausentes condicionada a las elecciones observadas en la muestra. Esta densidad condicionada está relacionada con la densidad no condicionada través de la identidad de Bayes:

$$h(z|y,\theta) = \frac{P(y|z,\theta)f(z|\theta)}{P(y|\theta)}.$$

Dicho sucintamente, la densidad de z condicionada a las elecciones observadas es proporcional al producto de la densidad no condicionada de z por la probabilidad de las elecciones observadas dada esta z. El denominador es simplemente la constante de normalización, igual a la integral del numerador. Este concepto de distribución condicionada debería resultar familiar a los lectores del capítulo 11.

Ahora consideremos la estimación. La función LL se basa en la información que el investigador tiene, que no incluye los datos ausentes. La función LL es por tanto

$$LL(\theta) = log\, P(y|\theta) = log\left(\int P(y|z,\theta)f(z|\theta)dz\right).$$

En principio, esta función se puede maximizar mediante el uso de los procedimientos descritos en el capítulo 8. Sin embargo, como vamos a ver, a menudo es mucho más fácil maximizar LL de forma diferente.

El procedimiento alternativo es iterativo, comenzando con un valor inicial de los parámetros y actualizándolos de una manera que se describirá a continuación. Denotemos el valor de prueba de los parámetros en una iteración dada como θ^t. Definamos una nueva función en θ^t que se relacione con LL pero que utilice la distribución condicionada h. Esta nueva función es

$$\mathcal{E}(\theta|\theta^t) = \int h(z|y,\theta^t)\, log(P(y|z,\theta)f(z|\theta))\, dz,$$

donde la densidad condicionada h se calcula utilizando el valor de prueba de los parámetros actual, θ^t. Esta función tiene un significado específico. Tenga en cuenta que la parte más a la derecha de esta expresión, $P(y|z,\theta)f(z|\theta)$, es la probabilidad conjunta de las elecciones observadas y de los datos ausentes. El logaritmo de esta probabilidad conjunta es la LL de las elecciones observadas *y* de los datos ausentes combinados. Esta LL conjunta está integrada sobre una densidad, concretamente, $h(z|y,\theta^t)$. Por tanto, nuestra función $\mathcal{E}$ es una esperanza de la LL conjunta de los datos ausentes y de las elecciones observadas. Es una esperanza específica, en concreto, la esperanza sobre la densidad de los datos ausentes condicionada a las elecciones observadas. Puesto que la densidad condicionada de z depende de los parámetros, esta densidad se calcula utilizando los valores θ^t. Dicho de manera equivalente, $\mathcal{E}$ es el promedio ponderado de la LL conjunta, utilizando $h(z|y,\theta^t)$ como pesos.

El procedimiento EM consiste en maximizar $\mathcal{E}$ repetidamente. Empezando con un cierto valor inicial, los parámetros se actualizan en cada iteración a través de la siguiente fórmula:

$$\theta^{t+1} = argmax_\theta \mathcal{E}(\theta|\theta^t). \qquad (14.1)$$

En cada iteración, los valores actuales de los parámetros, θ^t, se utilizan para calcular los pesos h, y a continuación se maximiza la LL conjunta ponderada. El nombre EM proviene del hecho de que el procedimiento utiliza una esperanza que es maximizada.

Es importante reconocer la doble función de los parámetros en $\mathcal{E}$. En primer lugar, los parámetros entran en la función log-verosimilitud conjunta de las elecciones observadas y de los datos ausentes, $log(P(y|z,\theta)f(z|\theta))$. En segundo lugar, los parámetros entran en la densidad condicionada de los datos ausentes, $h(z|y,\theta)$. La función $\mathcal{E}$ se maximiza respecto a la primera manteniendo constante la segunda. Es decir, $\mathcal{E}$ se maximiza sobre la θ que entra en $log(P(y|z,\theta)f(z|\theta))$, manteniendo el valor de θ que entra en los pesos $h(z|y,\theta)$ en su valor actual θ^t. Para indicar este doble rol, $\mathcal{E}(\theta|\theta^t)$ se expresa como una función de θ (el argumento sobre el que se realiza la maximización) dado θ^t(el valor utilizado en los pesos que se mantiene fijo durante la maximización).

En condiciones muy generales, las iteraciones definidas por la ecuación (14.1) convergen al máximo de LL. Bolyes (1983) y Wu (1983) proporcionan pruebas formales. En la siguiente sección ofrezco una explicación intuitiva. Sin embargo, los lectores que estén interesados en ver en primer lugar ejemplos del algoritmo pueden consultar directamente la sección 14.3.

14.2.1 ¿Por qué el algoritmo EM funciona?

La relación entre el algoritmo EM y la función LL se puede explicar en tres pasos. Cada paso es un poco opaco, pero los tres combinados proporcionan una comprensión sorprendentemente intuitiva.

Paso 1: Ajustamos $\mathcal{E}$ para igualarla a LL en θ^t

$\mathcal{E}(\theta|\theta^t)$ no es igual a $\mathrm{LL}(\theta)$. Para facilitar la comparación entre ambas funciones, vamos a añadir una constante a $\mathcal{E}(\theta|\theta^t)$ igual a la diferencia entre las dos funciones en θ^t:

$$\mathcal{E}^*(\theta|\theta^t) = \mathcal{E}(\theta|\theta^t) + [LL(\theta^t) - \mathcal{E}(\theta^t|\theta^t)]$$

El término entre corchetes es constante respecto a θ, así que maximizar $\mathcal{E}^*$ es equivalente a maximizar la propia $\mathcal{E}$. Observe sin embargo que, por construcción, $\mathcal{E}^*(\theta|\theta^t) = LL(\theta)$ en $\theta = \theta^t$.

Paso 2: Observe que la derivada respecto a θ es la misma para $\mathcal{E}^*$ y para LL evaluada en $\theta = \theta^t$.

Considere la derivada de $\mathcal{E}^*(\theta|\theta^t)$ respecto a su argumento θ:

$$\frac{d\mathcal{E}^*(\theta|\theta^t)}{d\theta} = \frac{d\mathcal{E}(\theta|\theta^t)}{d\theta}$$

$$= \int h(z|y,\theta^t)\left(\frac{d\, log\, P(y|z,\theta)f(z|\theta)}{d\theta}\right)dz$$

$$= \int h(z|y,\theta^t)\frac{1}{P(y|z,\theta)f(z|\theta)}\frac{dP(y|z,\theta)f(z|\theta)}{d\theta}dz.$$

Ahora evaluemos esta derivada en $\theta = \theta^t$:

$$\left.\frac{d\mathcal{E}^*(\theta|\theta^t)}{d\theta}\right|_{\theta^t}$$

$$= \int h(z|y,\theta^t)\frac{1}{P(y|z,\theta^t)f(z|\theta^t)}\left(\frac{dP(y|z,\theta)f(z|\theta)}{d\theta}\right)_{\theta^t}dz$$

$$= \int \frac{P(y|z,\theta^t)f(z|\theta^t)}{P(y|\theta^t)} \frac{1}{P(y|z,\theta^t)f(z|\theta^t)} \left(\frac{dP(y|z,\theta)f(z|\theta)}{d\theta}\right)_{\theta^t} dz$$

$$= \int \frac{1}{P(y|\theta^t)} \left(\frac{dP(y|z,\theta)f(z|\theta)}{d\theta}\right)_{\theta^t} dz$$

$$= \frac{1}{P(y|\theta^t)} \int \left(\frac{dP(y|z,\theta)f(z|\theta)}{d\theta}\right)_{\theta^t} dz$$

$$= \left(\frac{d\, log\, P(y|\theta)}{d\theta}\right)_{\theta^t}$$

$$= \left(\frac{dLL(\theta)}{d\theta}\right)_{\theta^t}.$$

En $\theta = \theta^t$, las dos funciones, $\mathcal{E}^*$ y LL, tienen la misma pendiente.

Paso 3: Observe que $\mathcal{E}^* \leq LL$ para todo θ.

Esta relación se puede demostrar de la siguiente manera:

(14.2) $$LL(\theta) = logP(y|\theta)$$

$$= log \int P(y|z,\theta)f(z|\theta)dz$$

$$= log \int \frac{P(y|z,\theta)f(z|\theta)}{h(z|y,\theta^t)} h(z|y,\theta^t)dz$$

(14.3) $$\geq \int h(z|y,\theta^t)\, log \frac{P(y|z,\theta)f(z|\theta)}{h(z|y,\theta^t)} dz$$

$$= \int h(z|y,\theta^t)\, log\big(P(y|z,\theta)f(z|\theta)\big)\, dz - \int h(z|y,\theta^t)\, log(h(z|y,\theta^t))\, dz$$

$$= \mathcal{E}(\theta|\theta^t) - \int h(z|y,\theta^t)\, log(h(z|y,\theta^t))\, dz$$

$$= \mathcal{E}(\theta|\theta^t) - \int h(z|y,\theta^t)\, log\left(h(z|y,\theta^t)\frac{P(y|\theta^t)}{P(y|\theta^t)}\right) dz$$

$$= \mathcal{E}(\theta|\theta^t) + \int h(z|y,\theta^t)\, log(P(y|\theta^t))\, dz - \int h(z|y,\theta^t)\, log(h(z|y,\theta^t)P(y|\theta^t))\, dz$$

$$= \mathcal{E}(\theta|\theta^t) + log(P(y|\theta^t)) \int h(z|y,\theta^t)dz - \int h(z|y,\theta^t)\, log(h(z|y,\theta^t)P(y|\theta^t))\, dz$$

$$= \mathcal{E}(\theta|\theta^t) + log(P(y|\theta^t)) - \int h(z|y,\theta^t)\, log(h(z|y,\theta^t)P(y|\theta^t))\, dz \tag{14.4}$$

$$= \mathcal{E}(\theta|\theta^t) + LL(\theta^t) - \int h(z|y,\theta^t)\, log\big(P(y|z,\theta^t)f(z|\theta^t)\big)\, dz \tag{14.5}$$

$$= \mathcal{E}(\theta|\theta^t) + LL(\theta^t) - \mathcal{E}(\theta^t|\theta^t)$$

$$= \mathcal{E}^*(\theta|\theta^t).$$

La desigualdad mostrada en la ecuación (14.3) se debe a la desigualdad de Jensen, que establece que $log(E(x)) > E(log(x))$. En nuestro caso, x es el estadístico $\frac{P(y|z,\theta)f(z|\theta)}{h(z|y,\theta^t)}$ y la esperanza es respecto a la densidad $h(z|y,\theta^t)$. Un ejemplo de esta desigualdad se muestra en la figura 14.1, donde los promedios son respecto a dos valores etiquetados como a y b. El promedio de $log(a)$ y $log(b)$ es el punto medio de la línea discontinua que conecta estos dos puntos de la curva logarítmica. El logaritmo evaluado en el promedio de a y b es $log((a+b)/2)$, que está por encima del punto medio de la línea discontinua. La desigualdad de Jensen es simplemente una consecuencia de la forma cóncava de la función logarítmica.

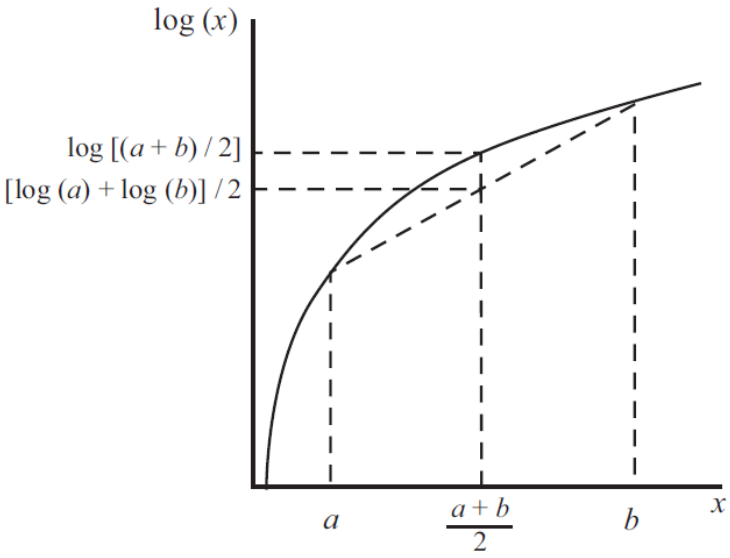


Figura 14.1. Ejemplo de la desigualdad de Jensen

La ecuación (14.4) se obtiene gracias a que la integral de la densidad h es 1. La ecuación (14.5) se obtiene sustituyendo $h(z|y,\theta^t) = P(y|z,\theta^t)f(z|\theta^t)/P(y|\theta^t)$ dentro del logaritmo y cancelando luego los términos $P(y|\theta^t)$.

Combinamos resultados para comparar $\mathcal{E}^*$ y LL.

La figura 14.2 muestra la relación entre $\mathcal{E}^*(\theta|\theta^t)$ y $LL(\theta)$. Como hemos demostrado, en $\theta = \theta^t$ estas dos funciones son iguales y tienen la misma pendiente. Estos resultados implican que las dos funciones son tangentes entre sí en $\theta = \theta^t$. También hemos demostrado que $\mathcal{E}^*(\theta|\theta^t) \leq LL(\theta)$ para todo θ. De acuerdo con esta relación, $\mathcal{E}^*$ se dibuja en el gráfico por debajo de $\mathrm{LL}(\theta)$ en todos los puntos, excepto en θ^t, donde son iguales.

El algoritmo EM maximiza $\mathcal{E}^*(\theta|\theta^t)$ en cada iteración para encontrar el siguiente valor de prueba de θ. El valor de maximización se muestra como θ^{t+1}. Como indica el gráfico, la función LL es necesariamente mayor en el valor del nuevo parámetro θ^{t+1}, que en el valor original θ^t. Mientras la derivada de la

función LL no sea cero en θ^t, maximizar $\mathcal{E}^*(\theta|\theta^t)$ incrementa $\mathrm{LL}(\theta)$[x]. Cada iteración del algoritmo EM incrementa la función LL hasta que el algoritmo converge en el máximo de la función LL.

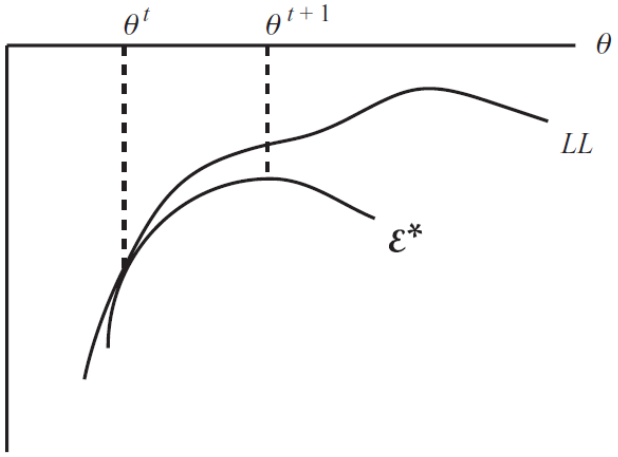


Figura 14.2. Relación entre $\mathcal{E}^*$ y LL.

14.2.2 Convergencia

La convergencia del algoritmo EM se define generalmente como un cambio suficientemente pequeño en los parámetros (por ejemplo, Levine y Casella, 2001) o en la función LL (por ejemplo, Weeks y Lange, 1989; Aitkin y Aitkin, 1996). Estos criterios deben ser utilizados con cuidado, ya que el algoritmo EM se puede mover lentamente cerca de la convergencia. Ruud (1991) muestra que el estadístico de convergencia visto en la sección 8.4 se puede utilizar con el gradiente y con el hessiano de $\mathcal{E}$, en lugar del LL. Sin embargo, el cálculo de este estadístico puede ser computacionalmente más exigente que la iteración del algoritmo EM en sí, llegando a ser inviable en algunos casos.

14.2.3 Errores Estándar

Hay 3 formas en que podemos calcular los errores estándar. En primer lugar, una vez se ha encontrado el máximo de $LL(\theta)$ con el algoritmo EM, los errores estándar se pueden calcular a partir de LL de la misma forma en que se haría si hubiésemos maximizado directamente la función LL. Los procedimientos de la sección 8.6 son aplicables: los errores estándar asintóticos se pueden calcular a partir del hessiano o a partir de la varianza de los gradientes específicos de cada observación (es decir, las puntuaciones), calculados a partir de LL evaluada en θ^t.

Una segunda opción surge del resultado que obtuvimos en el paso 2. Vimos que $\mathcal{E}$ y LL tienen el mismo gradiente en $\theta = \theta^t$. En el punto de convergencia, el valor de θ no cambia de una iteración a la siguiente, de tal manera que $\hat{\theta} = \theta^{t+1} = \theta^t$. Por lo tanto, en $\hat{\theta}$, las derivadas de estas dos funciones son iguales. Este hecho implica que las puntuaciones se pueden calcular a partir de $\mathcal{E}$ en lugar de LL. Si $\mathcal{E}$ toma una forma más conveniente que LL, como suele ser el caso cuando empleamos un algoritmo EM, esta forma de cálculo alternativo puede ser atractiva.

Una tercera opción es el *bootstrapping*, ya visto en la sección 8.6. En este caso, el algoritmo EM se aplica en numerosas ocasiones, usando una muestra diferente de las observaciones en cada ocasión. En muchos de los contextos en los que se aplican los algoritmos EM, el cálculo de los errores estándar a través de *bootstrapping* es más factible y útil que el uso de fórmulas asintóticas. El caso de estudio mostrado en la última sección proporciona un ejemplo.

[x] De hecho, cualquier incremento de $\mathcal{E}^*(\theta|\theta^t)$ conduce a un incremento de $LL(\theta)$

14.3 Ejemplos de algoritmos EM

Describiremos en esta sección varios tipos de modelos de elección discreta cuyas funciones LL son difíciles de maximizar directamente, pero son fáciles de estimar con algoritmos EM. El objetivo de la exposición es proporcionar ejemplos concretos de cómo se especifican los algoritmos EM e ilustrar al mismo tiempo el valor de este enfoque.

14.3.1 Distribución de mezcla discreta con puntos fijos

Una de las cuestiones que se plantean con los modelos logit mixtos (de hecho, con cualquier modelo mixto) es la especificación adecuada de la distribución de mezcla. Es habitual usar una distribución conveniente, como una normal o una log-normal. Sin embargo, es cuestionable que la verdadera distribución de los coeficientes tome una forma matemáticamente conveniente. Usar distribuciones más flexibles puede ser útil, donde flexibilidad significa que la distribución especificada puede tomar una variedad amplia de formas, dependiendo de los valores de sus parámetros.

Por lo general, podemos lograr mayor flexibilidad mediante la inclusión de más parámetros. En la estimación no paramétrica, se especifica una familia de distribuciones que tienen la propiedad de que la distribución se hace más flexible a medida que el número de parámetros se eleva. Al permitir que el número de parámetros aumente con el tamaño de la muestra, el estimador no paramétrico es consistente con cualquier distribución verdadera. El término "no paramétrico" es un nombre poco apropiado en este contexto: "superparamétrico" sería tal vez más apropiado, ya que el número de parámetros empleados es generalmente mayor al de especificaciones estándar, y aumenta con el tamaño de la muestra para obtener cada vez mayor flexibilidad.

El gran número de parámetros en la estimación no paramétrica hace que la maximización directa de la función LL resulte difícil. En muchos casos, sin embargo, es posible desarrollar un algoritmo EM que facilita esta estimación considerablemente. La presente sección muestra uno de estos casos.

Considere un modelo logit mixto con una distribución desconocida de coeficientes. Cualquier distribución se puede aproximar de forma arbitrariamente precisa a través de una distribución discreta con un número suficientemente grande de puntos. Podemos utilizar este hecho para desarrollar un estimador no paramétrico de la distribución de mezcla, utilizando un algoritmo EM para la estimación.

Representemos la densidad de los coeficientes mediante C puntos, siendo β_c el punto de c-ésimo. Supondremos que la ubicación de estos puntos (para este procedimiento en concreto) es fija y la masa en cada punto (es decir, la proporción o cuota de la población en cada punto) es el parámetro a estimar. Una forma de seleccionar los puntos es especificar un máximo y un mínimo de cada coeficiente, y crear una red de puntos uniformemente espaciados entre máximos y mínimos. Por ejemplo, supongamos que hay cinco coeficientes y que el rango entre el mínimo y el máximo de cada coeficiente está representado por 10 puntos uniformemente espaciados. Los 10 puntos en cada dimensión crean una cuadrícula de $10^5 = 100.000$ puntos en el espacio de cinco dimensiones. Los parámetros del modelo son la proporción o cuota de la población en cada uno de los 100.000 puntos. Como veremos, la estimación de un gran número de parámetros de este tipo es bastante asequible con un algoritmo EM. Al aumentar el número de puntos, la red se hace cada vez más fina, de tal manera que la estimación de las cuotas de población en los puntos permite aproximar cualquier distribución subyacente.

La utilidad que el agente n obtiene de la alternativa j es

$$U_{nj} = \beta_n x_{nj} + \varepsilon_{nj}$$

donde ε se distribuye valor extremo iid. Los coeficientes aleatorios tienen la distribución discreta que se ha descrito anteriormente, siendo s_c la cuota de la población en el punto β_c. La distribución se expresa a través de la siguiente función

$$f(\beta_n) = \begin{cases} s_1 & si\ \beta_n = \beta_1 \\ s_2 & si\ \beta_n = \beta_2 \\ \vdots & \\ s_C & si\ \beta_n = \beta_C \\ 0 & en\ otro\ caso, \end{cases}$$

donde las cuotas suman 1: $\sum_c s_c = 1$. Para mayor comodidad, nos referiremos al conjunto de todas las cuotas a través del vector $s = \langle s_1, \dots, s_c \rangle$[xi].

Condicionando a $\beta_n = \beta_c$ para unos valores c determinados, el modelo de elección es un logit estándar:

$$L_{ni}(\beta_c) = \frac{e^{\beta_c x_{ni}}}{\sum_j e^{\beta_c x_{ni}}}.$$

Dado que β_n no se conoce para cada persona, la probabilidad de elección es la de un modelo logit mixto, mezclando respecto a la distribución discreta de β_n:

$$P_{ni}(s) = \sum_c s_c L_{ni}(\beta_c).$$

La función LL es $LL(s) = \sum_n log\ P_{ni_n}(s)$, donde i_n es la alternativa elegida por el agente n.

Para estimar las cuotas s podemos maximizar directamente esta función LL. Con un gran número de clases, como se requiere normalmente para representar de forma flexible la distribución real, esta maximización directa puede ser difícil. Sin embargo, es posible utilizar un algoritmo EM increíblemente sencillo para este modelo, incluso con cientos de miles de puntos.

Los "datos ausentes" en este modelo son los coeficientes de cada agente. La distribución f indica la cuota de la población con cada valor del coeficiente. Sin embargo, como vimos en el capítulo 11, las elecciones que hace una persona revelan información sobre sus coeficientes. Condicionando a que la persona n elige la alternativa i_n, la probabilidad de que la persona tenga coeficientes β_c está dada por la identidad de Bayes:

$$h(\beta_c | i_n, s) = \frac{s_c L_{ni_n}(\beta_c)}{P_{ni_n}(s)}.$$

El algoritmo EM utiliza esta distribución condicionada. En particular, la esperanza para el algoritmo EM es

$$\mathcal{E}(s|s^t) = \sum_n \sum_c h(\beta_c | i_n, s^t)\ log\left(s_c L_{ni_n}(\beta_c)\right).$$

[xi] Esta especificación se puede considerar un tipo de modelo de clases latentes, donde hay C clases, los coeficientes de las personas de la clase c son β_c, y s_c es la proporción de la población en la clase c. Sin embargo, el término "modelo de clases latentes " por lo general se refiere a un modelo en el que la ubicación de los puntos son parámetros, así como las proporciones. Consideraremos esta forma más tradicional en nuestro siguiente ejemplo.

Dado que $log\left(s_c L_{ni_n}(\beta_c)\right) = log(s_c) + log\left(L_{ni_n}(\beta_c)\right)$, esta esperanza puede ser reescrita en dos partes:

$$\mathcal{E}(s|s^t) = \sum_n \sum_c h(\beta_c|i_n, s^t)\, log(s_c) + \sum_n \sum_c h(\beta_c|i_n, s^t)\, log\left(L_{ni_n}(\beta_c)\right).$$

Esta esperanza es la que debe maximizarse respecto a los parámetros s. Sin embargo, tenga en cuenta que el segundo término de esta expresión no depende de s: sólo depende de los coeficientes β_c, que en este procedimiento no paramétrico son puntos fijos. Por lo tanto, maximizar la fórmula anterior es equivalente a maximizar sólo la primera parte:

$$\mathcal{E}(s|s^t) = \sum_n \sum_c h(\beta_c|i_n, s^t)\, log(s_c).$$

Esta función es muy fácil de maximizar. En particular, el valor de maximización de s_c, teniendo en cuenta la restricción de que la suma de las cuotas debe ser 1, es

$$s_c^{t+1} = \frac{\sum_n h(\beta_c|i_n, s^t)}{\sum_n \sum_{c'} h(\beta_{c'}|i_n, s^t)}.$$

Usando la nomenclatura que hemos definido en la descripción general de los algoritmos EM, $h(\beta_c|i_n, s^t)$ son los pesos, calculados en el valor actual de las cuotas s^t. La cuota actualizada para la clase c es el porcentaje que la suma de los pesos en el punto c representa respecto a la suma de los pesos en todos los puntos.

Este algoritmo EM se implementa mediante los siguientes pasos:

1. Definimos los puntos β_c para c $= 1, \ldots,$ C.
2. Calculamos la fórmula logit para cada persona en cada punto: $L_{ni}(\beta_c)\ \forall n, c$.
3. Especificamos los valores iniciales de las cuotas en cada punto, etiquetadas colectivamente como s^0. Es conveniente que las cuotas iniciales sean $s_c = 1/C\ \forall c$.
4. Para cada persona y cada punto, se calcula la probabilidad de que la persona tenga esos coeficientes condicionando a las elecciones que ha realizado, usando las cuotas iniciales s^0 como las probabilidades no condicionadas: $h(\beta_c|i_n, s^0) = s_c^0 L_{ni}(\beta_c)/P_{ni}(s^0)$. Observe que el denominador es la suma sobre todos los puntos del numerador. Para mayor comodidad, etiquetamos este valor calculado como h_{nc}^0.
5. Actualizamos la cuota de la población en el punto c como $s_c^1 = \frac{\sum_n h_{nc}^0}{\sum_n \sum_{c'} h_{nc'}^0}$.
6. Repita los pasos 4 y 5 utilizando las cuotas actualizadas s en lugar de los valores iniciales originales. Repita estos pasos hasta lograr la convergencia.

Este procedimiento no requiere el cálculo de ningún gradiente ni la inversión de ningún hessiano, algo que los procedimientos descritos en el capítulo 8 sí utilizan para la maximización directa de LL. Por otra parte, las probabilidades logit se calculan sólo una vez (en el paso 2), y no en cada iteración. Las iteraciones consisten en volver a calibrar las cuotas en cada punto, algo que es pura aritmética. Dado que se necesita tan poco cálculo para cada punto, el procedimiento puede implementarse para un gran número de puntos. Por ejemplo, el caso real de Train (2008a) incluía más de 200.000 puntos, y sin embargo la estimación se llevó a cabo en sólo unos 30 minutos. Por el contrario, difícilmente la

maximización directa descrita en los métodos del capítulo 8 habría sido siquiera factible, puesto que implicaría una inversión de un hessiano de 200.000×200.000 valores.

14.3.2 Distribución de mezcla discreta con puntos como parámetros

Podemos modificar el modelo anterior tratando los coeficientes β_c para cada c, como parámetros a estimar en lugar de considerarlos puntos fijos. Los parámetros del modelo son, por lo tanto, la ubicación y el porcentaje o cuota de la población en cada punto. Etiquetamos estos parámetros colectivamente como $\theta = \langle s_c, \beta_c, c = 1, \dots, C \rangle$. Esta especificación a menudo recibe el nombre de *modelo de clases latentes* (*latent class model*): la población está compuesta por C clases distintas, de manera que todas las personas dentro de una clase tienen los mismos coeficientes, siendo los coeficientes diferentes para clases diferentes. Los parámetros del modelo son las proporciones o cuotas de la población en cada clase y los coeficientes de cada clase.

Los datos ausentes de este modelo son la pertenencia de las personas a cada clase. La esperanza empleada por el algoritmo EM es la misma de la especificación anterior, exceptuando que ahora que los coeficientes β_c son tratados como parámetros:

$$\mathcal{E}(\theta|\theta^t) = \sum_n \sum_c h(\beta_c|i_n, s^t)\, log\left(s_c L_{ni_n}(\beta_c)\right).$$

Observe que cada conjunto de parámetros entra solamente en un término dentro del logaritmo: el vector de cuotas s no entra en ninguno de los términos $L_{ni_n}(\beta_c)$s y cada β_c sólo entra a formar parte de $L_{ni_n}(\beta_c)$ para la clase c. Por lo tanto, la maximización de esta función es equivalente a la maximización por separado de cada una de las siguientes funciones:

$$\mathcal{E}(s|\theta^t) = \textstyle\sum_n \sum_c h(\beta_c|i_n, s^t)\, log(s_c), \qquad (14.6)$$

y para cada c:

$$\mathcal{E}(\beta_c|\theta^t) = \textstyle\sum_n h(\beta_c|i_n, s^t)\, log\, L_{ni_n}(\beta_c). \qquad (14.7)$$

El máximo de (14.6) se alcanza, como antes, en

$$s_c^{t+1} = \frac{\sum_n h(\beta_c|i_n, s^t)}{\sum_n \sum_{c'} h(\beta_{c'}|i_n, s^t)}.$$

Para la actualización de los coeficientes β_c, tenga en cuenta que (14.7) es la función LL para un modelo logit estándar, con cada observación ponderada por $h(\beta_c|i_n, s^t)$. Los valores actualizados de β_c se obtienen mediante la estimación de un modelo logit estándar, donde cada persona proporciona una observación que es ponderada adecuadamente. La misma estimación logit se lleva a cabo para cada clase c, pero con diferentes pesos para cada clase.

El algoritmo EM se implementa a través de los siguientes pasos:

1. Especificamos los valores iniciales de la cuota y de los coeficientes de cada clase, etiquetados como s_c y β_c $\forall c$. Es conveniente que las cuotas iniciales sean $1/C$. He observado que es posible obtener fácilmente valores de inicio para los coeficientes mediante la partición de la muestra en C grupos y la posterior ejecución de un logit en cada grupo[xii].

[xii] Tenga en cuenta que estos grupos no representan una división de la muestra en clases. Las clases son latentes por lo que dicha partición no es posible. En realidad, el objetivo es obtener C conjuntos de

2. Para cada persona y cada clase, calculamos la probabilidad de formar parte de cada clase condicionando a la elección de la persona, usando las cuotas iniciales s^0 como las probabilidades no condicionadas: $h(\beta_c^0|i_n, s^0) = s_c^0 L_{ni}(\beta_c^0)/P_{ni_n}(s^0)$. Observe que el denominador es la suma del numerador para todas las clases. Para mayor comodidad, etiquetamos este valor calculado como h_{nc}^0.
3. Actualizamos la cuota de la clase c como $s_c^1 = \frac{\sum_n h_{nc}^0}{\sum_n \sum_{c'} h_{nc'}^0}$.
4. Actualizamos los coeficientes para cada clase c mediante la estimación de un modelo logit, ponderando cada persona n por h_{nc}^0. Se estiman un total de C modelos logit, usando las mismas observaciones pero con diferentes pesos en cada una.
5. Repita los pasos 2-4 utilizando las cuotas actualizadas s y los coeficientes β_c $\forall c$ en lugar de los valores iniciales originales. Continúe repitiendo estos pasos hasta lograr la convergencia.

Una ventaja de este enfoque es que el investigador puede aplicar la estimación no paramétrica, con las cuotas y los coeficientes de cada clase tratados como parámetros, usando cualquier paquete de software estadístico que incluya una rutina de estimación logit. Para un número dado de clases, este procedimiento requiere mucho más tiempo que el anterior, ya que se deben estimar C modelos logit en cada iteración. Sin embargo, probablemente se necesiten muchas menos clases con este enfoque que con el anterior para representar adecuadamente la verdadera distribución, ya que este procedimiento estima la "mejor" ubicación de los puntos (es decir, los coeficientes), mientras que el anterior usa puntos fijos.

Bhat (1997a) desarrolló un algoritmo EM para un modelo similar a éste, salvo que las cuotas s_c de las clases no son parámetros en sí mismos sino que se especifican para que dependan de las características demográficas de la persona. El algoritmo EM que usó reemplaza nuestro paso 3 con un modelo logit de pertenencia de clase, con la probabilidad condicionada de pertenencia a una clase actuando como la variable dependiente. Su caso práctico muestra el punto fundamental de este capítulo: que los algoritmos EM se pueden desarrollar fácilmente para muchos tipos de modelos de elección complejos.

14.3.3 Distribución de mezcla normal con covarianza completa

En el capítulo 12 vimos que un logit mixto con covarianza plena entre coeficientes puede ser difícil de estimar mediante la maximización estándar de la función LL, debido tanto al gran número de parámetros de covarianza como al hecho de que la LL es altamente no cuadrática. Train (2008b) desarrolló un algoritmo EM muy simple y rápido para logits mixtos con covarianza plena. El algoritmo toma la siguiente forma, muy simple:

1. Especificamos valores iniciales para la media y la covarianza de los coeficientes en la población.
2. Para cada persona, extraemos valores al azar de la distribución de la población utilizando esta media y covarianza iniciales.
3. Ponderamos los valores extraídos de cada persona por la densidad condicionada de las extracciones de esa persona.
4. Calculamos la media y la covarianza de los valores extraídos ponderados de todas las personas. Éstos se convierten en la media y covarianza actualizadas de los coeficientes en la población.

valores iniciales para los coeficientes de las C clases. Estos valores iniciales no deben ser los mismos para todas las clases, ya que si fuesen iguales, el algoritmo realizaría los mismos cálculos para cada clase y devolvería la misma cuota y las mismas estimaciones actualizadas para todas las clases en cada iteración. Una manera fácil de obtener C conjuntos diferentes de valores iniciales es a dividir la muestra en C grupos y estimar un logit en cada grupo.

5. Repetimos los pasos 2-4 usando la media y covarianza actualizadas, y continuamos repitiendo estos pasos hasta que no haya ningún cambio adicional (con una cierta tolerancia) en la media y la covarianza.

Los valores convergentes son las estimaciones de la media y covarianza en la población. No se requieren gradientes. Todo lo que se necesita para la estimación de un modelo con coeficientes totalmente correlacionados es extraer repetidamente valores al azar utilizando la media y covarianza calculadas previamente, ponderar los valores apropiadamente, y calcular la media y covarianza de los valores ponderados.

El procedimiento es fácilmente aplicable cuando los coeficientes se distribuyen normalmente o cuando son transformaciones de términos conjuntamente normales. Siguiendo la notación empleada por Train y Sonnier (2005), la utilidad que obtiene el agente n de la alternativa j es

$$U_{nj} = \alpha_n x_{nj} + \varepsilon_{nj},$$

donde los coeficientes aleatorios son transformaciones de términos distribuidos normalmente: $\alpha_n = T(\beta_n)$, con β_n distribuido normalmente con media b y covarianza W. La transformación permite una flexibilidad considerable en la elección de la distribución. Por ejemplo, una distribución log-normal se obtiene especificando una transformación $\alpha_n = exp(\beta_n)$. Una distribución S_b que tenga un límite superior e inferior, se obtiene mediante la especificación de $\alpha_n = exp(\beta_n)/(1 + exp(\beta_n))$. Por supuesto, si el coeficiente es en sí mismo normal, entonces $\alpha_n = \beta_n$. La densidad normal se denota como $\phi(\beta_n|b, W)$.

Condicionadas a β_n, las probabilidades de elección son logit:

$$L_{ni}(\beta_n) = \frac{e^{T(\beta_n)x_{ni}}}{\sum_j e^{T(\beta_n)x_{nj}}}.$$

Dado que β_n no es conocido, la probabilidad de elección es un logit mixto, mezclado respecto a la distribución de β_n:

$$P_{ni}(b, W) = \int L_{ni}(\beta)\phi(\beta|b, W)d\beta.$$

La función LL es $LL(b, W) = \sum_n \log P_{ni_n}(b, W)$, donde i_n es la alternativa elegida por el agente n.

Como se trata en la sección 12.7, la estimación clásica de este modelo mediante la maximización estándar de LL es difícil, y esta dificultad es una de las razones para usar procedimientos bayesianos. Sin embargo, es posible aplicar un algoritmo EM que es considerablemente más fácil que la maximización estándar y que hace que la estimación clásica sea prácticamente tan conveniente para este modelo como la estimación bayesiana.

Los datos ausentes para el algoritmo EM son los parámetros β_n de cada persona. La densidad $\phi(\beta|b, W)$ es la distribución de β en la población. Para el algoritmo EM, utilizamos la distribución condicionada para cada persona. De acuerdo con la identidad de Bayes, la densidad de β condicionada a la alternativa i escogida por la persona n es $h(\beta|i, b, W) = L_{ni}(\beta)\phi(\beta|b, W)/P_{ni}(b, W)$. La esperanza para el algoritmo EM es

$$\mathcal{E}(b, W|b^t, W^t) = \sum_n \int h(\beta|i_n, b^t, W^t)\, log\big(L_{ni}(\beta)\phi(\beta|b, W)\big)\, d\beta.$$

Observe que $L_{ni}(\beta)$ no depende de los parámetros b y W. Por lo tanto, maximizar esta expectativa respecto a los parámetros es equivalente a maximizar

$$\mathcal{E}(b, W|b^t, W^t) = \sum_n \int h(\beta|i_n, b^t, W^t)\, log\big(\phi(\beta|b, W)\big)\, d\beta.$$

La integral dentro de esta esperanza no tiene una forma cerrada. Sin embargo podemos aproximar dicha integral a través de simulación. Sustituyendo la definición de $h(\cdot)$ y reordenando, tenemos

$$\mathcal{E}(b, W|b^t, W^t) = \sum_n \int \frac{L_{ni_n}(\beta)}{P_{ni_n}(b^t, W^t)} log\big(\phi(\beta|b, W)\big)\phi(\beta|b^t, W^t)\, d\beta.$$

La esperanza respecto a ϕ se simula mediante la extracción de R valores al azar de $\phi(\beta|b^t, W^t)$ para cada persona, etiquetando como β_{nr} el r–ésimo sorteo de la persona n. La esperanza simulada es

$$\tilde{\mathcal{E}}(b, W|b^t, W^t) = \sum_n \sum_r w_{nr}^t\, log\big(\phi(\beta_{nr}|b, W)\big)\,/R.$$

donde los pesos son

$$w_{nr}^t = \frac{L_{ni_n}(\beta_{nr})}{\frac{1}{R}\sum_{r'} L_{ni_n}(\beta_{nr'})}$$

Esta esperanza simulada tiene una forma familiar: es la función LL para una muestra de valores extraídos de una distribución normal, con cada valor ponderado por w_{nr}^t[xiii]. El estimador de máxima verosimilitud de la media y la covarianza de una distribución normal, dada una muestra ponderada de valores extraídos de esa distribución, no es más que la media y la covarianza ponderadas de los valores de la muestra. La media actualizada es

$$b^{t+1} = \frac{1}{NR}\sum_n \sum_r w_{nr}^t \beta_{nr}$$

y la matriz de covarianza actualizada es

$$W^{t+1} = \frac{1}{NR}\sum_n \sum_r w_{nr}^t (\beta_{nr} - b^{t+1})(\beta_{nr} - b^{t+1})'.$$

Observe que W^{t+1} es necesariamente definida positiva, como se requiere para una matriz de covarianza, dado que se construye como la covarianza de los valores extraídos al azar.

El algoritmo EM se implementa de la siguiente manera:

1. Especificamos los valores iniciales de la media y la covarianza, etiquetados b^0 y W^0.
2. Extraemos R valores al azar para cada una de las N personas en la muestra como $\beta_{nr}^0 = b^0 + chol(W^0)\eta_{nr}$, donde $chol(W^0)$ es el factor Choleski triangular inferior de W^0 y η_{nr} es un vector ajustado de valores normales estándar iid.

[xiii] La división por R puede ser ignorada dado que no afecta a la maximización, de la misma forma en que la división por el tamaño de la muestra N se omite en $\mathcal{E}$.

3. Para cada valor extraído de cada persona, calculamos la probabilidad logit de la elección observada de dicha persona: $L_{ni_n}(\beta_{nr}^0)$.
4. Para cada valor extraído de cada persona, calculamos el peso

$$w_{nr}^0 = \frac{L_{ni_n}(\beta_{nr}^0)}{\sum_{r'} L_{ni_n}\left(\beta_{nr'}^0\right)/R}.$$

5. Calculamos la media y la covarianza ponderadas de los $N * R$ valores extraídos $\beta_{nr}^0, r = 1, \dots, R, n = 1, \dots, N$, usando los pesos w_{nr}^0. La media y covarianza ponderada son los parámetros actualizados b^1 y W^1.
6. Repetimos los pasos 2-5 utilizando la media b y la varianza W actualizadas en lugar de los valores iniciales originales. Continuamos repitiendo estos pasos hasta lograr la convergencia.

Este procedimiento puede llevarse a cabo sin necesitad de usar un software de estimación, simplemente extrayendo valores al azar, calculando fórmulas logit para construir los pesos y calculando la media y covarianza ponderadas de los valores. Un investigador puede estimar un modelo logit mixto con covarianza plena y con coeficientes que posiblemente son transformaciones de normales a través de estos sencillos pasos.

Train (2008a) muestra que este procedimiento se puede generalizar para una mezcla finita de normales, donde β se extrae de cualquiera de entre C normales con diferentes medias y covarianzas. La probabilidad de extraer β de cada normal (es decir, la cuota de la población cuyos coeficientes son descritos por cada distribución normal) es un parámetro, junto con las medias y covarianzas. Cualquier distribución se puede aproximar por una mezcla finita de normales, con un número suficiente de normales subyacentes. Al permitir que el número de normales crezca con el tamaño de la muestra, la aproximación se convierte en una forma de estimación no paramétrica de la verdadera distribución de mezcla. El algoritmo EM para este tipo de estimaciones no paramétricas combina los conceptos mostrados en la presente sección para distribuciones normales con plena covarianza y los conceptos que vimos en la sección inmediatamente anterior sobre distribuciones discretas.

Una última observación útil: en los valores de convergencia, las derivadas de $\tilde{\mathcal{E}}(b, W|\hat{b}, \widehat{W})$ proporcionan puntuaciones que pueden ser usadas para estimar los errores estándar asintóticos de las estimaciones. En particular, el gradiente respecto a b y W es

$$\frac{\tilde{\mathcal{E}}(b, W|\hat{b}, \widehat{W})}{db} = \sum_n \left[\frac{1}{R}\sum_r -w_{nr}W^{-1}(\beta_{nr} - b)\right]$$

y

$$\frac{\tilde{\mathcal{E}}(b, W|\hat{b}, \widehat{W})}{dW} = \sum_n \left[\frac{1}{R}\sum_r w_{nr}\left(-\frac{1}{2}W^{-1} + \frac{1}{2}W^{-1}(\beta_{nr} - b)(\beta_{nr} - b)'W^{-1}\right)\right],$$

donde los pesos w_{nr} se calculan en los valores estimados $\hat{b}$ y $\widehat{W}$. Los términos entre corchetes son las puntuaciones de cada persona, las cuales podemos agrupar en un vector etiquetado como s_n. La varianza de las puntuaciones es $V = \sum_n s_n s_n'/N$. La covarianza asintótica del estimador se calcula por lo tanto como V^{-1}/N, como vimos en la sección 8.6.

14.4 Caso de estudio: demanda de coches impulsados por hidrógeno

Train (2008a) estudió las preferencias de los compradores de vehículos impulsados por hidrógeno, utilizando varios de los algoritmos EM que hemos visto. Vamos a describir uno de sus modelos estimados como una ilustración del procedimiento EM. Se realizó una encuesta a compradores de automóviles nuevos en el sur de California para evaluar la importancia que estos compradores otorgaban a varios aspectos relevantes en relación a los vehículos de hidrógeno, tales como la disponibilidad de estaciones de servicio. A cada encuestado se le presentó una serie de 10 experimentos de preferencia declarada. En cada experimento, se solicitó al encuestado que eligiese entre tres alternativas posibles: el vehículo de combustible convencional (*conventional-fuel vehicle*, CV) que el encuestado había comprado recientemente y dos vehículos de combustible alternativo (*alternative-fuel vehicle*, AVs) con los atributos especificados. Se pidió al encuestado que evaluase las tres opciones, indicando cuál consideraba mejor y cuál peor. Los atributos representan las características relevantes de los vehículos de hidrógeno, pero a los encuestados no se les dijo que el combustible alternativo era el hidrógeno a fin de evitar cualquier idea preconcebida que pudiesen haber desarrollado respecto a este tipo de vehículos. Los atributos que se incluyeron en los experimentos son los siguientes:

- Costo del combustible (*fuel cost*, FC), expresado como diferencia porcentual respecto a los vehículos de combustible convencional, CV. En la estimación, el atributo se definió con una escala porcentual, de tal manera que unos costos de combustible un 50 por ciento menores que los del vehículo de combustible convencional entraban en el modelo como -0.5 y los costos un 50 por ciento mayores entraban como 0.5.
- Precio de compra (*purchase price*, PP), expresado como diferencia porcentual respecto al CV, escalado al formar parte del modelo de manera análoga al costo del combustible.
- Radio de conducción (*driving radius*, DR): la distancia más lejana a la que uno puede viajar desde el hogar y posteriormente regresar, partiendo con un depósito lleno de combustible. Según la definición, el radio de conducción es la mitad de la autonomía del vehículo. En los modelos estimados, el DR fue escalado en cientos de millas.
- Destinos convenientes de media distancia (*convenient medium-distance destinations*, CMDD): porcentaje de destinos dentro del radio de conducción que "no requieren una planificación anticipada, ya que es posible abastecerse de combustible durante el camino o en el destino", en contraposición a los destinos que "requieren de abastecimiento de combustible (o al menos estimar si se dispone de suficiente combustible) antes de salir para asegurar que es posible hacer el viaje de ida y vuelta". Este atributo refleja la distribución de los posibles destinos y estaciones de servicio dentro del radio de conducción, reconociendo el hecho de que el depósito de combustible no estará siempre lleno en el momento de arrancar. En los modelos estimados, esta variable se introduce como una cuota o proporción, de tal manera que, por ejemplo, 50 por ciento entra como 0.50.
- Posibles destinos de larga distancia (p*ossible long-distance destinations*, PLDD) : el porcentaje de destinos más allá del radio de conducción que es posible alcanzar gracias a que el reabastecimiento es posible, en oposición a destinos que no se pueden alcanzar debido a la cobertura limitada de estaciones de repostaje. Este atributo refleja la disponibilidad de estaciones de servicio fuera del radio de conducción y su proximidad a potenciales destinos de conducción. Se introdujo en los modelos usando una escala análoga al CMDD.
- Tiempo adicional respecto a las estaciones locales (*extra time to local stations*, ETLS): tiempo de viaje adicional de ida, más allá del tiempo requerido normalmente para encontrar una estación de combustible convencional, que se requiere para llegar a una estación de combustible alternativo en el área local. El ETLS se definió para tener valores de 0, 3 y 10 minutos en los

experimentos; sin embargo, en el análisis preliminar, se encontró que los encuestados consideraron que 3 minutos no representaba ningún inconveniente (es decir, era equivalente a 0 minutos). En los modelos estimados, por lo tanto, se introdujo una variable indicadora para ETLS igual a 10 o diferente de 10, en lugar del ETLS en sí mismo.

En los experimentos, el CV comprado por el encuestado fue descrito como un vehículo con un radio de conducción de 200 millas, CMDD y PLDD igual al 100 por cien y, por definición, ETLS, FC y PP iguales a 0.

Como se indicó anteriormente, se pidió al entrevistado que identificase la mejor y la peor de las tres alternativas, proporcionando así un ranking de los tres vehículos. Condicionadas a los coeficientes de los entrevistados, las probabilidades del ranking se especificaron con la fórmula de "logit expandido" (tal como se describe en la sección 7.3.1). Mediante esta formulación, la probabilidad del ranking es la probabilidad logit de la primera elección entre las tres alternativas posibles del experimento, por la probabilidad logit de la segunda elección entre las dos alternativas restantes. Esta probabilidad se combina con la distribución de probabilidad de los coeficientes, cuyos parámetros han sido estimados.

Se aplicaron los tres métodos que hemos descrito anteriormente. Nos concentramos en el método de la sección 14.3.2, ya que proporciona una ilustración sucinta de la potencia del algoritmo EM. Para este método, hay C clases de compradores, y los coeficientes β_n y las cuotas s_c de la población en cada clase son tratados como parámetros.

Train (2008a) estimó el modelo con diferente número de clases, que van desde una clase (que es un logit estándar) hasta 30 clases. La tabla 14.1 muestra el valor de la LL para estos modelos. Aumentar el número de clases mejora la LL considerablemente, desde -7.884.6 con una clase hasta -5.953.4 con 30 clases. Por supuesto, un mayor número de clases implica más parámetros, lo que plantea la cuestión de si el ajuste mejorado justifica el esfuerzo de tratar parámetros adicionales. En situaciones como ésta, es habitual evaluar los modelos por el criterio de información de Akaike (a*kaike information criterion*, AIC) o mediante el criterio de información bayesiano (*bayesian information criterion*, BIC)[xiv]. Los valores de estos estadísticos también se muestran en la tabla 14.1. El AIC es más bajo (mejor) con 25 clases y el BIC, que penaliza en mayor medida el uso de parámetros adicionales que el AIC, es más bajo con 8 clases.

Tabla 14.1. Modelos logit mixtos con distribuciones discretas de coeficientes y diferente número de clases

Clases	Log-verosimilitud (LL)	Parámetros	AIC	BIC
1	−7,884.6	7	15,783.2	15,812.8
5	−6,411.5	39	12,901.0	13,066.0
6	−6,335.3	47	12,764.6	12,963.4
7	−6,294.4	55	12,698.8	12,931.5
8	−6,253.9	63	12,633.8	12,900.3
9	−6,230.4	71	12,602.8	12,903.2
10	−6,211.4	79	12,580.8	12,915.0
15	−6,124.5	119	12,487.0	12,990.4
20	−6,045.1	159	12,408.2	13,080.8
25	−5,990.7	199	12,379.4	13,221.3
30	−5,953.4	239	12,384.8	13,395.9

[xiv] 6 Véase, por ejemplo, Mittelhammer et. al. (2000, sección 18.5) para una exposición relativa a los criterios de información. El AIC (Akaike, 1974) es - 2LL + 2K, donde LL es el valor del logaritmo de la verosimilitud y K es el número de parámetros. El BIC, también llamado criterio de Schwarz (1978), es - 2LL + log (N) K, donde N es el tamaño de la muestra.

A efectos de evaluar el algoritmo EM, es útil tener en cuenta que la estimación de estos modelos requirió un tiempo de ejecución en torno a 1.5 minutos por clase, pariendo de los valores iniciales hasta lograr la convergencia. Esto significa que el modelo con 30 clases, que tiene 239 parámetros[xv], se estimó en tan sólo 45 minutos.

La tabla 14.2 presenta las estimaciones para el modelo con 8 clases, la mejor opción de acuerdo al criterio BIC. El modelo con 25 clases, el mejor modelo según el criterio AIC, proporciona aún mayor detalle pero no se proporciona en aras de la brevedad. Como se muestra en la tabla 14.2, la mayor de las 8 clases es la última, con el 25 por ciento. Esta clase tiene un coeficiente positivo grande para CV, a diferencia de todas las otras clases. Por lo tanto, esta clase aparentemente se compone de personas que prefieren su CV frente a los AV, incluso cuando los AV tienen los mismos atributos, tal vez a causa de la incertidumbre asociada a las nuevas tecnologías de combustible. Otras características distintivas de las clases son evidentes. Por ejemplo, la clase 3 se preocupa más por el PP (precio de compra) que las otras clases, mientras que la clase 1 da más importancia al FC (costo de combustible) que las otras clases.

Tabla 14.2. Modelo con 8 clases

Clase	1	2	3	4
Cuotas	0.107	0.179	0.115	0.0699
Coeficientes				
FC	−3.546	−2.576	−1.893	−1.665
PP	−2.389	−5.318	−12.13	0.480
DR	0.718	0.952	0.199	0.472
CMDD	0.662	1.156	0.327	1.332
PLPP	0.952	2.869	0.910	3.136
ETLS=10 (variable indicadora)	−1.469	−0.206	−0.113	−0.278
CV (variable indicadora)	−1.136	−0.553	−0.693	−2.961
Clase	5	6	7	8
Cuotas	0.117	0.077	0.083	0.252
Coeficientes				
FC	−1.547	−0.560	−0.309	−0.889
PP	−2.741	−1.237	−1.397	−2.385
DR	0.878	0.853	0.637	0.369
CMDD	0.514	3.400	−0.022	0.611
PLPP	0.409	3.473	0.104	1.244
ETLS=10 (variable indicadora)	0.086	−0.379	−0.298	−0.265
CV (variable indicadora)	−3.916	−2.181	−0.007	2.656

La tabla 14.3 muestra la media y la desviación estándar entre coeficientes de las 8 clases. Como Train (2008a) señala, estas medias y desviaciones estándar son similares a las obtenidas con un modelo logit mixto más estándar, con coeficientes distribuidos normalmente (coeficientes que se pueden consultar en su artículo publicado, pero que no se repiten aquí). Este resultado indica que el uso en este caso práctico de numerosas clases, algo que el algoritmo EM hace posible, proporciona mayor detalle en la explicación de las diferencias en las preferencias, manteniendo al mismo tiempo estadísticos resumidas muy similares.

[xv] Siete coeficientes y una cuota de mercado por cada una de las 30 clases, con una cuota de clase determinada por la restricción de que las cuotas deben sumar uno.

Tabla 14.3. Estadísticos resumidos de los coeficientes

	Medias		Desviaciones estándar	
	Est.	EE	Est.	EE
Coeficientes				
FC	−1.648	0.141	0.966	0.200
PP	−3.698	0.487	3388	0.568
DR	0.617	0.078	0.270	0.092
CMDD	0.882	0.140	0.811	0.126
PLPP	1575	0.240	1098	0.178
ETLS=10 (variable indicadora)	−0.338	0.102	0.411	0.089
CV (variable indicadora)	−0.463	1181	2142	0.216

Est=Estimación, EE=Error Estándar

Sería difícil calcular los errores estándar de las fórmulas asintóticas para este modelo (es decir, calcularlos mediante la inversa del hessiano estimado), debido a la gran cantidad de parámetros existentes. Además, estamos interesados en los estadísticos resumidas, como la media y la desviación estándar de los coeficientes entre todas las clases, dadas en la tabla 14.4. Obtener los errores estándar de estos estadísticos resumidos a partir de fórmulas asintóticas de la covarianza de los parámetros mismos sería computacionalmente difícil. En cambio, es posible calcular fácilmente los errores estándar mediante *bootstrapping*. Dada la velocidad del algoritmo EM en este caso práctico, usar *bootstrapping* es factible. Asimismo, *bootstrapping* proporciona automáticamente los errores estándar de nuestros estadísticos resumidos (mediante el cálculo de los estadísticos resumidos para cada estimación *bootstrap* y tomando sus desviaciones estándar).

Tabla 14.4. Errores estándar para la clase 1

	Est.	EE
Coeficientes		
FC	−3.546	2.473
PP	−2.389	6.974
DR	0.718	0.404
CMDD	0.662	1.713
PLPP	0.952	1.701
ETLS=10 (variable indicadora)	−1.469	0.956
CV (variable indicadora)	−1.136	3.294

Est=Estimación, EE=Error Estándar

Los errores estándar para los estadísticos resumidos se facilitan en la tabla 14.3, basados en 20 muestras de *bootstrapping*. Los errores estándar no se proporcionan en la tabla 14.2 para los parámetros de cada clase. En lugar de ello, la tabla 14.4 da los errores estándar para la clase 1, como un ejemplo ilustrativo de todas las clases. Tal y como se muestra en dicha tabla, los errores estándar de los parámetros de la clase 1 son elevados. Estos errores estándar elevados eran de esperar, y surgen por el hecho de que el etiquetado de clases en este modelo es arbitrario. Supongamos, como ejemplo extremo pero ilustrativo, que las dos muestras diferentes de *bootstrapping* dan las mismas estimaciones para dos clases pero con su orden cambiado (es decir, las estimaciones para la clase 1 convirtiéndose en las estimaciones para la clase 2 y viceversa). En este caso, los errores estándar obtenidos por *bootstrapping* para los parámetros de ambas clases aumentan a pesar de que el modelo para estas dos clases juntas es exactamente el mismo. Los estadísticos resumidos evitan este problema. Todas las medias excepto una son

estadísticamente significativas, siendo la variable indicadora de CV la única que obtiene una media no significativa. Todas las desviaciones estándar son significativamente diferentes de cero.

Train (2008a) también estimó otros dos modelos a partir de estos mismos datos utilizando algoritmos EM: (1) un modelo con una distribución discreta de coeficientes, donde los puntos son fijos y las cuotas de población en cada punto se estiman usando el procedimiento de la Sección 14.3.1 y (2) un modelo con una distribución de mezcla discreta, formada por dos distribuciones normales con covarianza plena, utilizando una generalización del procedimiento de la sección 14.3.3. La flexibilidad de los algoritmos EM para dar cabida a una amplia variedad de modelos complejos es la razón por la que vale la pena aprender su uso. Mejoran la capacidad del investigador para construir modelos diseñados a medida, que se ajustan estrechamente a la realidad de la situación y de los objetivos de la investigación, algo que ha sido el objetivo primordial de este libro.

15

Bibiografía

Adamowicz, W. (1994), 'Habit formation and variety seeking in a discrete choice model of recreation demand', Journal of Agricultural and Resource Economics 19,19–31.

Aitkin, M. and I. Aitkin (1996), 'A hybrid EM/Gauss-Newton algorithm for maximum likelihood in mixture distributions', Statistics and Computing 6, 127–130.

Akaike, H. (1974), 'Anewlook at the statistical identification model', IEEE Transactions on Automatic Control 19(6), 716–723.

Albert, J. and S. Chib 1993), 'Bayesian analysis of binary and polychotomous response data', Journal of the American Statistical Association 88, 669–679.

Allenby,G. (1997), 'An introduction to hierarchical Bayesian modeling', Tutorial Notes, Advanced Research Techniques Forum, American Marketing Association.

Allenby, G. and P. Lenk (1994), 'Modeling household purchase behavior with logistic normal regression', Journal of the American Statistical Association 89, 1218–1231.

Allenby,G. and P. Rossi (1999), 'Marketing models of consumer heterogeneity', Journal of Econometrics 89, 57–78.

Amemiya, T. (1978), 'On two-step estimation of multivariate logit models', Journal of Econometrics 8, 13–21.

Andrews, R., A. Ainslie, and I. Currim (2002), 'An empirical comparison of logit choice models with discrete vs. continuous representation of heterogeneity', Journal of Marketing Research 39, 479–487.

Arora, N., G. Allenby, and J. Ginter (1998), 'A hierachical Bayes model of primary and secondary demand', Marketing Science 17, 29–44.

Beggs, S., S. Cardell, and J. Hausman (1981), 'Assessing the potential demand for electric cars', Journal of Econometrics 16, 1–19.

Bellman, R. (1957), Dynamic Programming, Princeton University Press, Princeton, NJ.

Ben-Akiva, M. (1973), 'The structure of travel demand models', PhD Thesis, MIT.

Ben-Akiva,M. and M. Bierlaire (1999), 'Discrete choice methods and their applications in short term travel decisions', in R. Hall, ed., The Handbook of Transportation Science, Kluwer, Dordrecht, pp. 5–33.

Ben-Akiva, M. and D. Bolduc (1996), 'Multinomial probit with a logit kernel and a general parametric specification of the covariance structure', Working Paper, Department of Civil Engineering, MIT.

Ben-Akiva, M. and B. Francois (1983), 'Mu-homogenous generalized extreme value model', Working Paper, Department of Civil Engineering, MIT.

Ben-Akiva,M. and S. Lerman (1985), Discrete Choice Analysis: Theory and Application to Travel Demand, MIT Press, Cambridge, MA.

Ben-Akiva,M. and T. Morikawa (1990), 'Estimation of switching models from revealed preferences and stated intentions', Transportation Research A 24, 485–495.

Ben-Akiva, M., D. Bolduc, and M. Bradley (1993), 'Estimation of travel model choice models with randomly distributed values of time', Transportation Research Record 1413, 88–97.

Berkovec, J. and S. Stern (1991), 'Job exit behavior of older men', Econometrica 59, 189–210.

Berndt, E., B. Hall, R. Hall, and J. Hausman (1974), 'Estimation and inference in nonlinear structural models', Annals of Economic and Social Measurement 3/4, 653–665.

Bernstein, S. (1917), Calcul des probabilit´es.

Berry, S. (1994), 'Estimating discrete choice models of product differentiation', RAND Journal of Economics 25, 242–262.

Berry, S., J. Levinsohn, and A. Pakes (1995), 'Automobile prices in market equilibrium', Econometrica 63, 841–889.

Berry, S., J. Levinsohn, and A. Pakes (2004), 'Differentiated products demand system from a combination of micro and macro data: The new car market', Journal of Political Economy 112, 68–105.

Bhat, C. (1995), 'A heteroscedastic extreme value model of intercity mode choice', Transportation Research B 29, 471–483.

Bhat, C. (1997a), 'An endogenous segmentation mode choice model with an application to intercity travel', Transportation Science 31, 34–48.

Bhat, C. (1997b), 'Covariance heterogeneity in nested logit models: Econometric structure and application to intercity travel', Transportation Research B 31, 11–21.

Bhat, C. (1998a), 'Accommodating variations in responsiveness to level-of-service variables in travel mode choice models', Transportation Research A 32, 455–507.

Bhat, C. (1998b), 'An analysis of travel mode and departure time choice for urban shopping trips', Transportation Research B 32, 361–371.

Bhat, C. (1999), 'An analysis of evening commute stop-making behavior using repeated choice observation from a multi-day survey', Transportation Research B 33, 495–510.

Bhat, C. (2000), 'Incorporating observed and unobserved heterogeneity in urban work mode choice modeling', Transportation Science 34, 228–238.

Bhat, C. (2001), 'Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model', Transportation Research B 35, 677–693.

Bhat, C. (2003), 'Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences', Transportation Research B 37, 837–855.

Bhat, C. and S. Castelar (2002), 'A unified mixed logit framework for modeling revealed and stated preferences: Formulation and application to congestion pricing analysis in the San Francisco Bay area', Transportation Research 36, 577–669.

Bickel, P. and K. Doksum (2000), Mathematical Statistics: Basic Ideas and Selected Topics, Vol. 1, Prentice Hall, Upper Saddle River, NJ.

Bierlaire, M. (1998), Discrete choice models, in M. Labbe, G. Laporte, K. Tanczos, and P. Toint, eds., Operations Research and Decision Aid Methodologies in Traffic and Transportation Management, Springer, Heidelberg, pp. 203–227.

Boatwright, P., S. Borle, and J. Kadane (2003), 'A model of the joint distribution of purchase quantity and timing', Journal of the American Statistical Association 98, 564–572.

Bolduc, D. (1992), 'Generalized autoregressive errors: The multinomial probit model', Transportation Research B 26, 155–170.

Bolduc, D. (1993), 'Maximum simulated likelihood estimation of MNP models using the GHK probability simulation with analytic derivatives', Working Paper, D´epartement d'Economique, Universit´e Laval, Quebec.

Bolduc, D. (1999), 'A practical technique to estimate multinomial probit models in transportation', Transportation Research B 33, 63–79.

Bolduc, D., B. Fortin, and M. Fournier (1996), 'The impact of incentive policies on the practice location of doctors: A multinomial probit analysis', Journal of Labor Economics 14, 703–732.

Bolduc, D., B. Fortin, and S. Gordon (1997), 'Multinomial probit estimation of spatially interdependent choices: An empirical comparison of two newtechniques', International Regional Science Review 20, 77–101.

Borsch-Supan, A. and V. Hajivassiliou (1993), 'Smooth unbiased multivariate probability simulation for maximum likelihood estimation of limited dependent variable models', Journal of Econometrics 58, 347–368.

Borsch-Supan, A.,V. Hajivassiliou, L.Kotlikoff, and J. Morris (1991), 'Health, children, and elderly living arrangements: A multiperiod multinomial probit model with unobserved heterogeneity and autocorrelated errors', in D.Wise, ed., Topics in the Economics of Aging, University of Chicago Press, Chicago.

Boyd, J. and J. Mellman (1980), 'The effect of fuel economy standards on the U.S. automotive market: A hedonic demand analysis', Transportation Research A 14, 367–378.

Boyles, R. (1983), 'On the convergence of the EM algorithm', Journal of the Royal Statistical Society B 45, 47–50.

Braatan, E. and G. Weller (1979), 'An improved low-discrepancy sequence for multidimensional quasi-Monte Carlo integration', Journal of Computational Physics 33, 249–258.

Bradley, M. and A. Daly (1994), 'Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data', Transportation 21, 167–184.

Bradlow, E. and P. Fader (2001), 'A Bayesian lifetime model for the "hot 100" billboard songs', Journal of the American Statistical Association 96, 368–381.

Brownstone, D. (2001), 'Discrete choice modeling for transportation', in D. Hensher, ed., Travel Behavior Research: The Leading Edge, Elsevier, Oxford, pp. 97–124.

Brownstone, D. and K. Small (1989), 'Efficient estimation of nested logit model', Journal of Business and Economic Statistics 7, 67–74.

Brownstone, D. and K. Train (1999), 'Forecasting new product penetration with flexible substitution patterns', Journal of Econometrics 89, 109–129.

Brownstone, D., D. Bunch, and K. Train (2000), 'Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles', Transportation Research B 34, 315–338.

Bunch, D. (1991), 'Estimability in the multinomial probit model', Transportation Research B 25, 1–12.

Bunch, D. and R. Kitamura (1989), 'Multinomial probit estimation revisited: Testing new algorithms and evaluation of alternative model specification of household car ownership', Transportation Research Group Report UCD-TRG-RR-4, University of California, Davis.

Butler, J. and R. Moffitt (1982), 'A computationally efficient quadrature procedure for the one factor multinomial probit model', Econometrica 50, 761–764.

Cai, Y., I. Deilami, and K. Train (1998), 'Customer retention in a competitive power market: Analysis of a "double-bounded plus follow-ups" questionnaire', The Energy Journal 19, 191–215.

Cameron, T. (1988), 'A new paradigm for valuing non-market goods using referendum data: Maximum likelihood estimation by censored logistic regression', Journal of Environmental Economics and Management 15, 355–379.

Cameron, T. and M. James (1987), 'Efficient estimation methods for closed-ended contingent valuation survey data', Review of Economics and Statistics 69, 269–276.

Cameron, T. and J. Quiggin (1994), 'Estimation using contingent valuation data from a "dichotomous choice with follow-up" questionnaire', Journal of Environmental Economics and Management 27, 218 234.

Cardell, S. and F. Dunbar (1980), 'Measuring the societal impacts of automobile downsizing', Tranportation Research A 14, 423–434.

Casella, G. and E. George (1992), 'Explaining the Gibbs sampler', American Statistician 46, 167–174.

Chapman, R. and R. Staelin (1982), 'Exploiting rank ordered choice set data within the stochastic utility model', Journal of Marketing Research 14, 288–301.

Chesher, A. and J. Santos-Silva (2002), 'Taste variation in discrete choice models', Review of Economic Studies 69, 62–78.

Chiang, J., S. Chib, and C. Narasimhan (1999), 'Markov chain Monte Carlo and models of consideration set and parameter heterogeneity', Journal of Econometrics 89, 223–248.

Chib, S. and E. Greenberg (1995), 'Understanding the Metropolis–Hastings algorithm', American Statistician 49, 327–335.

Chib, S. and E. Greenberg (1996), 'Markov chain Monte Carlo simulation methods in econometrics', Econometric Theory 12, 409–431.

Chib, S. and E. Greenberg (1998), 'Analysis of multivariate probit models', Biometrika 85, 347–361.

Chintagunta, P., J. Dub´e, and K. Goh (2005), 'Beyond the endogeneity bias: The effect of unmeasured brand characteristics on household-level brand choice models', Management Science 52, 832–849.

Chintagunta, P., D. Jain, and N. Vilcassim (1991), 'Investigating heterogeneity in brand preference in logit models for panel data', Journal of Marketing Research 28, 417–428.

Chipman, J. (1960), 'The foundations of utility', Econometrica 28, 193–224.

Chu, C. (1981), 'Structural issues and sources of bias in residential location and travel choice models', PhD Thesis, Northwestern University.

Chu, C. (1989), 'A paired combinational logit model for travel demand analysis', Proceedings of Fifth World Conference on Transportation Research 4, 295–309.

Clark, C. (1961), 'The greatest of a finite set of random variables', Operations Research 9, 145–162.

Cosslett, S. (1981), 'Efficient estimation of discrete choice models', in C. Manski and D. McFadden, eds., Structural Analysis of Discrete Data with Econometric Applications, MIT Press, Cambridge, MA.

Cowles, M. and B. Carlin (1996), 'Markov chain Monte Carlo convergence diagnostics: A comparative review', Journal of the American Statistical Association 91, 883–904.

Daganzo, C. (1979), Multinomial Probit: The Theory and Its Application to Demand Forecasting, Academic Press, New York.

Daganzo, C., F. Bouthelier, and Y. Sheffi (1977), 'Multinomial probit and qualitative choice: A computationally efficient algorithm', Transportation Science 11, 338–358.

Dagsvik, J. (1994), 'Discrete and continuous choice max-stable processes and independence from irrelevant alternatives', Econometrica 62, 1179–1205.

Daly, A. (1987), 'Estimating "tree" logit models', Transportation Research B 21, 251–267.

Daly, A. and S. Zachary (1978), Improved multiple choice models, in D. Hensher and M. Dalvi, eds., Determinants of Travel Choice, Saxon House, Sussex.

Debreu, G. (1960), 'Review of R.D. Luce individual choice behavior', American Economic Review 50, 186–188.

Dempster, A., N. Laird, and D. Rubin (1977), 'Maximum likelihood from incomplete data via the EM algorithm', Journal of the Royal Statistical Society B 39, 1–38.

DeSarbo,W., V. Ramaswamy, and S. Cohen (1995), 'Market segmentation with choicebased conjoint analysis', Marketing Letters 6, 137–147.

Desvousges, W., S. Waters, and K. Train (1996), 'Potential economic losses associated with recreational services in the Upper Clark Fork River basin', Report, Triangle Economic Research, Durham, NC.

Eckstein, Z. and K.Wolpin (1989), 'The specification and estimation of dynamic stochastic discrete choice models: A survey', Journal of Human Resources 24, 562–598.

Efron, B. (1979), 'Bootstrapping methods: Another look at the jackknife', Annals of Statistics 7, 1–26.

Efron, B. and R. Tibshirani (1993), An Introduction to the Bootstrap, Chapman and Hall, New York.

Elrod, T. and M. Keane (1995), 'A factor analytic probit model for representing the market structure in panel data', Journal of Marketing Research 32, 1–16.

Erdem, T. (1996), 'A dynamic analysis of market structure based on panel data', Marketing Science 15, 359–378.

Ferreira, F. (2004), 'You can take it with you: Transferability of proposition 13 tax benefits, residential mobility, and willingness to pay for housing amenities', Center for Labor Economics Working Paper No. 72, at http://ssrn.com/abstract=661421.

Forinash, C. and F. Koppelman (1993), 'Application and interpretation of nested logit models of intercity mode choice', Transportation Research Record 1413, 98–106.

Gelman, A. (1992), 'Iterative and non-iterative simulation algorithms', Computing Science and Statistics (Interface Proceedings) 24, 433–438.

Gelman, A. and D. Rubin (1992), 'Inference from iterative simulation using multiple sequences', Statistical Sciences 7, 457–511.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (1995), Bayesian Data Analysis, Chapman and Hall, Suffolk.

Geman, S. and D. Geman (1984), 'Stochastic relaxation Gibbs distributions and the Bayesian restoration of images', IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721–741.

Geweke, J. (1988), 'Antithetic acceleration of Monte Carlo integration in Bayesian inference', Journal of Econometrics 38, 73–89.

Geweke, J. (1989), 'Bayesian inference in econometric models using Monte Carlo integration', Econometrica 57, 1317–1339.

Geweke, J. (1991), 'Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints', in E. M. Keramidas, ed., Computer Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface, Interface Foundation of North America, Inc., Fairfax, pp. 571–578.

Geweke, J. (1992), 'Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments', in J. Bernardo, J. Berger, A. Dawid, and F. Smith, eds., Bayesian Statistics, Oxford University Press, New York, pp. 169–193.

Geweke, J. (1996), 'Monte Carlo simulation and numerical integration', in D. Kendrick and J. Rust, eds., Handbook of Computational Economics, Elsevier Science, Amsterdam, pp. 731–800.

Geweke, J. (1997), 'Posterior simulators in econometrics', in D. Kreps and K. Wallis, eds., Advance Economics and Econometric Theory and Applications, Cambridge University Press, New York.

Geweke, J., M. Keane, and D. Runkle (1994), 'Alternative computational approaches to inference in the multinomial probit model', Review of Economics and Statistics 76, 609–632.

Goett, A. (1998), 'Estimating customer preferences for new pricing products', Electric Power Research Institute Report TR-111483, Palo Alto, CA.

Goolsbee, A. and A. Petrin (2004), 'The consumer gains from direct broadcast satellites and the competition with cable TV', Econometrica 72, 351–382.

Gourieroux, C. and A. Monfort (1993), 'Simulation-based inference: A survey with special reference to panel data models', Journal of Econometrics 59, 5–33.

Greene, W. (2000), Econometric Analysis, 4th edn, Prentice Hall, Upper Saddle River, NJ.

Greene, W. (2001), 'Fixed and random effects in nonlinear models', Working Paper, Stern School of Business, New York University.

Griffiths, W. (1972), 'Estimation of actual response coefficients in the Hildreth–Horck random coefficient model', Journal of the American Statistical Association 67, 663–635.

Guevara, C. and M. Ben-Akiva (2006), 'Endogeneity in residential location choice models', Transportation Research Record 1977, 60–66.

Guilkey, D. and J. Murphy (1993), 'Estimation and testing in the random effects probit model', Journal of Econometrics 59, 301–317.

Haaijer, M., M. Wedel, M. Vriens, and T. Wansbeek (1998), 'Utility covariances and context effects in conjoint MNP models', Marketing Science 17, 236–252.

Hajivassiliou, V. and D. McFadden (1998), 'The method of simulated scores for the estimation of LDV models', Econometrica 66, 863–896.

Hajivassiliou, V. and P. Ruud (1994), 'Classical estimation methods for LDV models using simulation', in R. Engle and D. McFadden, eds., Handbook of Econometrics, North-Holland, Amsterdam, pp. 2383–2441.

Hajivassiliou, V., D. McFadden, and P. Ruud (1996), 'Simulation of multivariate normal rectangle probabilities and their derivatives: Theoretical and computational results', Journal of Econometrics 72, 85–134.

Halton, J. (1960), 'On the efficiency of evaluating certain quasi-random sequences of points in evaluating multi-dimensional integrals', Numerische Mathematik 2, 84–90.

Hamilton, J. (1996), 'Specification testing in Markov-switching time-series models', Journal of Econometrics 70, 127–157.

Hamilton, J. and R. Susmel (1994), 'Autoregressive conditional heteroskedasticity and changes in regime', Journal of Econometrics 64, 307–333.

Hammersley, J. and K. Morton (1956), 'A new Monte Carlo technique: Antithetic variates', Proceedings of the Cambridge Philosophical Society 52, 449–474.

Hanemann, M., J. Loomis, and B. Kanninen (1991), 'Statistical efficiency of doublebounded dichotomous choice contengent valuation', American Journal of Agricultural Economics 73, 1255–1263.

Hastings, W. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', Biometrika 57, 97–109.

Hausman, J. (1978), 'Specification tests in econometrics', Econometrica 46, 1251–1272.

Hausman, J., ed. (1993), Contingent Valuation: A Critical Assessment, North-Holland, New York.

Hausman, J. (1997), 'Valuation of new goods under perfect and imperfect competition', in R. Gordon and T. Bresnahan, eds., The Economics of New Goods, University of Chicago Press, Chicago.

Hausman, J. and D. McFadden (1984), 'Specification tests for the multinomial logit model', Econometrica 52, 1219–1240.

Hausman, J. and P. Ruud (1987), 'Specifying and testing econometric models for rankordered data', Journal of Econometrics 34, 83–103.

Hausman, J. and D. Wise (1978), 'A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences', Econometrica 48, 403–429.

Heckman, J. (1978), 'Dummy endogenous variables in a simultaneous equation system', Econometrica 46, 931–959.

Heckman, J. (1981a), 'The incidental parameters problem and the problem of initial condition in estimating a discrete time–discrete data stochastic process', in C. Manski and D. McFadden, eds., Structural Analysis of Discrete Data with Econometric Applications, MIT Press, Cambridge, MA, pp. 179–185.

Heckman, J. (1981b), 'Statistical models for the analysis of discrete panel data', in C. Manski and D. McFadden, eds., Structural Analysis of Discrete Data with Econometric Applications, MIT Press, Cambridge, MA, pp. 114–178.

Heckman, J. and R. Robb (1985), 'Alternative methods for evaluating the impacts of interventions: An overview', Journal of Econometrics 30, 239–267.

Heckman, J. and B. Singer (1986), 'Econometric analysis of longitudinal data', in Z. Griliches and M. Intriligator, eds., Handbook of Econometrics, North-Holland, Amsterdam, pp. 1689–1763.

Heiss, F. (2002), 'Structural choice analysis with nested logit models', Stata Journal 2 227–252.

Hensher, D. (2001), 'The valuation of commuter travel time savings for car drivers in New Zealand: Evaluating alternative model specifications', Transportation 28, 101–118.

Hensher, D. and M. Bradley (1993), 'Using stated response data to enrich revealed preference discrete choice models', Marketing Letters 4, 39–152.

Hensher, D. and W. Greene (2003), 'The mixed logit model: The state of practice and warnings for the unwary', Transportation 30, 133–176.

Hensher, D. andW. Greene (2002), 'Specification and estimation of nested logit model', Transportation Research B, 36, 1–17.

Hensher, D., J. Louviere, and J. Swait (1999), 'Combining sources of preference data', Journal of Econometrics 89, 197–221.

Herriges, J. and C. Kling (1996), 'Testing the consistency of nested logit models with utility maximization', Economic Letters 50, 33–39.

Horowitz, J. (1991), 'Reconsidering the multinomial probit model', Transportation Research B 25, 433–438.

Horowitz, J., J. Sparmann, and C. Daganzo (1982), 'An investigation of the accuracy of the Clark approximation for the multinomial probit model', Transportation Science 16, 382–401.

Hotz, V. and R. Miller (1993), 'Conditional choice probabilities and the estimation of dynamic models', Review of Economic Studies 60, 497–529.

Hotz, V., R. Miller, S. Sanders, and J. Smith (1993), 'A simulation estimator for dynamic models of discrete choice', Review of Economic Studies 61, 265–289.

Huber, J. and K. Train (2001), 'On the similarity of classical and Bayesian estimates of individual mean partworths', Marketing Letters 12, 259–269.

Imai, S., N. Jain, and A. Ching (2001), 'Bayesian estimation of dynamics discrete choice models', paper presented at Bayesian Applications and Methods in Marketing Conference,

Ohio State University; and Working Paper, Department of Economics, Pennsylvania State University.

Jiang, R., P. Manchanda, and P. Rossi (2007), 'Bayesian analysis of random coefficient logit models using aggregate data', Working Paper, Graduate School of Business, University of Chicago, Chicago.

Joe, S. and I. Sloan (1993), 'Implementation of a lattice method for numerical multiple integration', ACM Transactions in Mathematical Software 19, 523–545.

Johannesson, M. and D. Lundin (2000), 'The impact of physical preferences and patient habits on the diffusion of new drugs', Working Paper, Department of Economics, Stockholm School of Economics.

Johnson, N., S. Kotz, and N. Balakrishnan (1994), Continuous Multivariate Distributions, 2nd edn, John Wiley and Sons, New York.

Judge, G., R. Hill, W. Griffiths, and T. Lee (1985), The Theory and Practice of Econometrics, 2nd edn, John Wiley and Sons, New York.

Judge, G., R. Hill, W. Griffiths, H. Lutkepohl, and T. Lee (1988), Introduction to the Theory and Practice Econometrics, 2nd edn, John Wiley and Sons, New York.

Kamakura, W. A. and G. Russell (1989), 'A probabilistic choice model for market segmentation and elasticity structure', Journal of Marketing Research 26, 379–390.

Karlstrom, A. (2000), 'Non-linear value functions in random utility econometrics', Conference Presentation, 9th IATBR Travel Behavior Conference, Australia; and Working Paper, Infrastructure and Planning, Royal Institute of Technology, Stockholm.

Karlstrom, A. (2001), 'Developing generalized extreme value models using the Piekands representation theorem', Working Paper, Infrastructure and Planning, Royal Institute of Technology, Stockholm.

Kass, R., B. Carlin, A. Gelman, and R. Neal (1998), 'Markov chain Monte Carlo in practice: A roundtable discussion', American Statistician 52, 93–100.

Keane, M. (1990), 'Four essays in empirical macro and labor economics', PhD Thesis, Brown University.

Keane, M. (1994), 'A computationally practical simulation estimator for panel data', Econometrica 62, 95–116.

Keane, M. and K. Wolpin (1994), 'The solutions and estimation of discrete choice dynamic programming models by simulation and interpretation: Monte Carlo evidence', Review of Economics and Statistics 76, 648–672.

Kling, C. and J. Herriges (1995), 'An empirical investigation of the consistency of nested logit models with utility maximization', American Journal of Agricultural Economics 77, 875–884.

Koppelman, F. and C.Wen (1998), 'Alternative nested logit models: Structure, properties and estimation', Transportation Research B 32, 289–298.

Koppelman, F. and C. Wen (2000), 'The paired combination logit model: Properties, estimation and application', Transportation Research B 34, 75–89.

Laplace, P. (1820), Th´eorie Analytique des Probabilit´es, 3rd edn, Paris.

Le Cam, L. and G. Yang (1990), Asymptotics in Statistics, Springer, New York.

Lee, B. (1999), 'Calling patterns and usage of residential toll service under self-selecting tariffs', Journal of Regulatory Economics 16, 45–82.

Lee, L. (1992), 'On the efficiency of methods of simulated moments and simulated likelihood estimation of discrete choice models', Econometric Theory 8, 518–552.

Lee, L. (1995), 'Asymptotic bias in simulated maximum likelihood estimation of discrete choice models', Econometric Theory 11, 437–483.

Lehmann, E. and G. Casella (1998), Theory of Point Estimation, 2nd edn, Springer, New York.

Levine, R. and G. Casella (2001), 'Implementation of the Monte Carlo EM algorithm', Journal of Computational and Graphical Statistics 10, 422–439.

Liu, Y. and H. Mahmassani (2000), 'Global maximum likelihood estimation procedures for multinomial probit(MND)model parameters', Transportation ResearchB 34, 419–444.

Louviere, J., D. Hensher, and J. Swait (2000), Stated Choice Methods: Analysis and Applications, Cambridge University Press, New York.

Luce, D. (1959), Individual Choice Behavior, John Wiley and Sons, New York.

Luce, D. and P. Suppes (1965), 'Preferences, utility and subjective probability', in R. Luce, R. Bush, and E. Galanter, eds., Handbook of Mathematical Psychology, John Wiley and Sons, New York, pp. 249–410.

Manski, C. and S. Lerman (1977), 'The estimation of choice probabilities from choice based samples', Econometrica 45, 1977–1988.

Manski, C. and S. Lerman (1981), 'On the use of simulated frequencies to approximate choice probabilities', in C. Manski and D. McFadden, eds., Structural Analysis of Discrete Data with Econometric Applications, MIT Press, Cambridge, MA, pp. 305–319.

Manski, C. and D. McFadden (1981), 'Alternative estimators and sample designs for discrete choice analysis', in C. Manski and D. McFadden, eds., Structural Analysis of Discrete Data with Econometric Applications, MIT Press, Cambridge, MA, pp. 2–50.

Marschak, J. (1960), 'Binary choice constraints on random utility indications', in K. Arrow, ed., Stanford Symposium on Mathematical Methods in the Social Sciences, Stanford University Press, Stanford, CA, pp. 312–329.

Martin, L. (2008), 'Consumer demand for compact flourescent light bulbs', Working Paper, Department of Agricultural and Resource Economics, University of California, Berkeley.

McCulloch, R. and P. Rossi (1994), 'An exact likelihood analysis of the multinomial probit model', Journal of Econometrics 64, 207–240.

McCulloch, R. and P. Rossi (2000), 'Bayesian analysis of the multinomial probit model', in R. Mariano, T. Schuermann, and M.Weeks, eds., Simulation-Based Inference in Econometrics, Cambridge University Press, New York.

McFadden, D. (1974), 'Conditional logit analysis of qualitative choice behavior', in P. Zarembka, ed., Frontiers in Econometrics, Academic Press, New York, pp. 105–142.

McFadden, D. (1978), 'Modeling the choice of residential location', in A. Karlqvist, L. Lundqvist, F. Snickars, and J. Weibull, eds., Spatial Interaction Theory and Planning Models, North-Holland, Amsterdam, pp. 75–96.

McFadden, D. (1987), 'Regression-based specification tests for the multinomial logit model', Journal of Econometrics 34, 63–82.

McFadden, D. (1989), 'A method of simulated moments for estimation of discrete response models without numerical integration', Econometrica 57, 995–1026.

McFadden, D. (1996), 'Lectures on simulation-assisted statistical inference', Conference Presentation, EC-squared Conference, Florence, Italy; and Working Paper, Department of Economics, University of California, Berkeley.

McFadden, D. (1999), 'Computing willingness-to-pay in random utility models', in J. Moore, R. Riezman, and J. Melvin, eds., Trade, Theory and Econometrics: Essays in Honour of John S. Chipman, Routledge, London, pp. 253–274.

McFadden, D. (2001), 'Economic choices', American Economic Review 91, 351–378.

McFadden, D. and K. Train (1996), 'Consumers' evaluation of new products: Learning from self and others', Journal of Political Economy 104, 683–703.

McFadden, D. and K. Train (2000), 'Mixed MNL models of discrete response', Journal of Applied Econometrics 15, 447–470.

McFadden, D., A. Talvitie, S. Cosslett, I. Hasan, M. Johnson, F. Reid, and K. Train (1977), 'Demand model estimation and validation', Final Report, Volume V, Urban Travel Demand Forecasting Project, Institute of Transportation Studies, University of California, Berkeley.

McFadden, D., K. Train, and W. Tye (1978), 'An application of diagnostic tests for the independence from irrelevant alternatives property of the multinomial logit model', Transportation Research Record 637, 39–46.

McGrath, E. (1970), Fundamentals of Operations Research, West Coast University Press, San Francisco.

McLachlan, G. and T. Krishnan (1997), The EM Algorithm and Extensions, JohnWiley and Sons, New York.

Mehndiratta, S. (1996), 'Time-of-day effects in inter-city business travel', PhD Thesis, University of California, Berkeley.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A.Teller, and E.Teller (1953), 'Equations of state calculations by fast computing machines', Journal of Chemical Physics 21, 1087–1092.

Mittelhammer, R., G. Judge, and D. Miller (2000), EconometricFoundations, Cambridge University Press, New York.

Morokoff, W. and R. Caflisch (1995), 'Quasi-Monte Carlo integration', Journal of Computational Physics 122, 218–230.

Munizaga, M. and R. Alvarez-Daziano (2001), 'Mixed logit versus nested logit and probit', Working Paper, Departmento de Ingeniera Civil, Universidad de Chile.

Nevo, A. (2001), 'Measuring market power in the ready-to-eat cereal industry', Econometrica 69, 307–342.

Niederreiter, H. (1978), 'Quasi-Monte Carlo methods and pseudo-random numbers', Bulletin of the American Mathematical Society 84, 957–1041.

Niederreiter, H. (1988), 'Low-discrepancy and low dispersion sequences', Journal of Number Theory 30, 51–70.

O'Donoghue, T. and M. Rabin (1999), 'Doing it now or later', American Economic Review 89, 103–124.

Ortuzar, J. (1983), 'Nested logit models for mixed-mode travel in urban corridors', Transportation Research A 17, 283–299.

Pakes, A. (1986), 'Patents as options: Some estimates of the value of holding European patent stocks', Econometrica 54, 755–785.

Pakes, A. and D. Pollard (1989), 'Simulation and asymptotics of optimization estimators', Econometrica 57, 1027–1057.

Papatla, P. and L. Krishnamurthi (1992), 'Aprobit model of choice dynamics', Marketing Science 11, 189–206.

Park, S. and S. Gupta (forthcoming), 'A simulated maximum likelihood estimator for the random coefficient logit model using aggregate data', Journal of Marketing Research.

Petrin, A. (2002), 'Quantifying the benefits of new products; the case of the Minivan', Journal of Political Economy 110, 705–729.

Petrin, A. and K. Train (2009), 'A control function approach to endogeneity in consumer choice models', Journal of Marketing Research, forthcoming.

Rao, B. (1987), Asymptotic Theory of Statistical Inference, John Wiley and Sons, New York.

Recker, W. (1995), 'Discrete choice with an oddball atlernative', Transportation Research B 29, 207–211.

Research Triangle Institute (1997), 'Predicting retail customer choices among electricity pricing alternatives', Electric Power Research Institute Report, Palo Alto, CA.

Revelt, D. (1999), 'Three discrete choice random coefficients papers and one police crime study', PhD Thesis, University of California, Berkeley.

Revelt, D. and K.Train (1998), 'Mixed logit with repeated choices', Reviewof Economics and Statistics 80, 647–657.

Revelt, D. and K. Train (2000), 'Customer-specific taste parameters and mixed logit', Working Paper No. E00-274, Department of Economics, University of California, Berkeley.

Rivers, D. and Q. Vuong (1988), 'Limited information estimators and exogeneity tests for simultaneous probit models', Journal of Econometrics 39, 347–366.

Rossi, P., R. McCulloch, and G. Allenby (1996), 'The value of household information in target marketing', Marketing Science 15, 321–340.

Rust, J. (1987), 'Optimal replacement of GMC bus engines: An empirical model of Harold Zurchner', Econometrica 55, 993–1033.

Rust, J. (1994), 'Estimation of dynamic structural models, problems and prospects: Discrete decision processes', in C. Sims, ed., Advances in Econometrics: Sixth World Congress, Vol. II, Cambridge University Press, New York, pp. 5–33.

Rust, J. (1997), 'Using randomization to break the curse of dimensionality', Econometrica 65, 487–516.

Ruud, P. (1991), 'Extensions of estimation methods using the EM algorithm', Journal of Econometrics 49, 305–341.

Ruud, P. (1996), 'Simulation of the multinomial probit model: An analysis of covariance matrix estimation', Working Paper, Department of Economics, University of California, Berkeley.

Ruud, P. (2000), An Introduction to Classical Econometric Theory, Oxford University Press, New York.

S´andor, Z. and P. Andr´as (2001), 'Alternative sampling methods for estimating multivariate normal probabilities',Working Paper, Department of Economics, University of Groningen, The Netherlands.

S´andor, Z. and K. Train (2004), 'Quasi-random simulation of discrete choice models', Journal of Econometrics 120, 207–234.

SawtoothSoftware (1999), 'The CBC/HB module for hierarchical Bayes', at www.sawtoothsoftware.com.

Schechter, L. (2001), 'The apple and your eye: Visual and taste rankordered probit analysis', Working Paper, Department of Agricultural and Resource Economics, University of California, Berkeley.

Schwarz, G. (1978), 'Estimating the dimension of a model', The Annals of Statistics 6, 461–464.

Shim, E. and E. Sudit (1995), 'How manufacturers price products', Management Accounting 76, 37–39.

Siikamaki, J. (2001), 'Discrete choice experiments valuing biodiversity conservation in Finland', PhD Thesis, University of California, Davis.

Siikamaki, J. and D. Layton (2001), 'Pooled models for contingent valuation and contingent ranking data:Valuing benefits from biodiversity conservation',Working Paper, Department of Agricultural and Resource Economics, University of California, Davis.

Sloan, I. and H.Wozniakowski (1998), 'When are quasi-Monte Carlo algorithms efficient for high dimensional integrals?', Journal of Complexity 14, 1–33.

Small, K. (1987), 'A discrete choice model for ordered alternatives', Econometrica 55, 409–424.

Small, K. (1994), 'Approximate generalized extreme value models of discrete choice', Journal of Econometrics 62, 351–382.

Small, K. and H. Rosen (1981), 'Applied welfare economics of discrete choice models', Econometrica 49, 105–130.

Spanier, J. and E. Maize (1991), 'Quasi-random methods for estimating integrals using relatively small samples', SIAM Review 36, 18–44.

Srinivasan, K. and H. Mahmassani (2005), 'Dynamic kernel logit model for the analysis of longtitude discrete choice data: Properties and computational assessment', Transportation Science 39, 160–181.

Steckel, J. and W. Vanhonacker (1988), 'A heterogeneous conditional logit model of choice', Journal of Business and Economic Statistics 6, 391–398.

Swait, J. and J. Louviere (1993), 'The role of the scale parameter in the estimation and use of multinomial logit models', Journal of Marketing Research 30, 305–314.

Talvitie, A. (1976), 'Disaggregate travel demand models with disaggregate data, not aggregate data, and why',Working Paper No. 7615, UrbanTravelDemandForecasting

Project, Institute of Transportation Studies, University of California, Berkeley.

Theil, H. (1971), Principles of Econometrics, John Wiley and Sons, New York.

Thurstone, L. (1927), 'A law of comparative judgement', Psychological Review 34, 273–286.

Train, K. (1978), 'A validation test of a diaggregate mode choice model', Transportation Research 12, 167–174.

Train, K. (1986), Qualitative Choice Analysis, MIT Press, Cambridge, MA.

Train, K. (1995), 'Simulation methods for probit and related models based on convenient error partitioning', Working Paper No. 95-237, Department of Economics, University of California, Berkeley.

Train, K. (1998), 'Recreation demand models with taste variation', Land Economics 74, 230–239.

Train, K. (1999), 'Mixed logit models for recreation demand', in J. Herriges and C. Kling, eds., Valuing Recreation and the Environment, Edward Elgar, Northampton, MA.

Train, K. (2000), 'Halton sequences for mixed logit', Working Paper No. E00-278, Department of Economics, University of California, Berkeley.

Train, K. (2001), 'A comparison of hierarchical Bayes and maximum simulated likelihood for mixed logit', Working Paper, Department of Economics, University of California, Berkeley.

Train, K. (2008a),'EMalgorithms for nonparametric estimation of mixing distributions', Journal of Choice Modelling 1, 40–69.

Train, K. (2008b), 'Arecursive estimator for random coefficient models',Working Paper, Department of Economics, University of California, Berkeley.

Train, K. and D. McFadden (1978), 'The goods–leisure tradeoff and disaggregate work trip mode choice models', Transportation Research 12, 349–353.

Train, K., D. McFadden, and M. Ben-Akiva (1987a), 'The demand for local telephone service: A fully discrete model of residential calling patterns and service choice', Rand Journal of Economics 18, 109–123.

Train, K., D. McFadden, and A. Goett (1987b), 'Consumer attitudes and voluntary rate schedules for public utilities', Review of Economics and Statistics LXIX, 383–391.

Train, K., M. Ben-Akiva, and T. Atherton (1989), 'Consumption patterns and selfselecting tariffs', Review of Economics and Statistics 71, 62–73.

Train, K. and C. Winston (2007), 'Vehicle choice behavior and the declining market share of U.S. automakers', International Economic Review 48, 1469–1496.

Train, K. and G. Sonnier (2005), 'Mixed logit with bounded distributions of correlated partworths', in R. Scarpa and A. Alberini, eds., Applications of Simulation Methods in Environmental and Resource Economics, Springer, Dordrecht, pp. 117–134.

Tuffin, B. (1996), 'On the use of low-discrepancy sequences in Monte Carlo methods', Monte Carlo Methods and Applications 2, 295–320.

Tversky, A. (1972), 'Elimination by aspects: A theory of choice', Psychological Review 79, 281–299.

Vijverberg, W. (1997), 'Monte Carlo evaluation of multivariate normal probabilities', Journal of Econometrics 76, 281–307.

Villas-Boas, M. (2007), 'A note on limited versus full information estimation in nonlinear models',Working Paper, Haas School of Business, University of California, Berkeley.

Villas-Boas, M. and R. Winer (1999), 'Endogeneity in brand choice models', Management Science 45, 1324–1338.

Vinod, H. (1993), 'Bootstrap, jackknife, resampling and simulation: Applications in econometrics', in G. Maddala, C. Rao, and H. Vinod, eds., Handbook of Statistics: Econometrics, Vol. II, North-Holland, Amsterdam, chapter 11. von Mises, R. (1931), Wahrscheinlichkeitsrechnung, Springer, Berlin.

Vovsha, P. (1997), 'The cross-nested logit model: Application to mode choice in the Tel Aviv metropolitan area', Conference Presentation, 76th Transportation Research Board Meetings, Washington, DC.

Walker, J., M. Ben-Akiva, and D. Bolduc (2007), 'Identification of parameters in normal error component logit-mixture (NECLM) models', Journal of Applied Econometrics 22, 1095–1125.

Wang, X., E. Bradlow, and H. Wainer (2002), 'A general Bayesian model for testlets: Theory and application', Applied Psychological Measurement 26, 1090–1128.

Wedel, M. and W. Kamakura (2000), Market Segmentation: Conceptual and Methodological Foundations, 2nd edn, Kluwer Academic Publishers, Boston.

Weeks, D. and K. Lange (1989), 'Trials, tribulations, and triumphs of the EM algorithm in pegigree analysis', Journal of Mathematics Applied in Medicine and Biology 6, 209–232.

Wen, C.-H. and F.Koppelman (2001), 'The generalized nested logit model', Transportation Research B 35, 627–641.

Wen, D. and M. Levy (2001), 'Blindex: A bounded asymmetric loss function with application to Bayesian estimation', Communications in Statistics—Theory and Methods 30, 147–153.

Williams, H. (1977), 'On the formation of travel demand models and economic evaluation measures of user benefits', Environment and Planning A 9, 285–344.

Wolpin, K. (1984), 'An estimable dynamic stochastic model of fertility and child mortality', Journal of Political Economy 92, 852–874.

Wolpin, K. (1987), 'Estimating a structural search model: The transition from school to work', Econometrica 55, 801–818.

Wu, C. (1983), 'On the convergence properties of the EM algorithm', Annals of Statistics 11, 95–103.

Yai, T., S. Iwakura, and S. Morichi (1997), 'Multinomial probit with structured covariance for route choice behavior', Transporation Research B 31, 195–207.

Yang, S., Y. Chen, and G. Allenby (2003), 'Bayesian analysis of simultaneous demand and supply', Quantitative Marketing and Economics 1, 251–275.

Zavoina, R. and W. McKelvey (1975), 'A statistical model for the analysis of ordinal level dependent variables', Journal of Mathematical Sociology Summer, 103–120.

Zellner, A. (1971), An Introduction to Bayesian Inference in Econometrics, John Wiley and Sons, New York.