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.

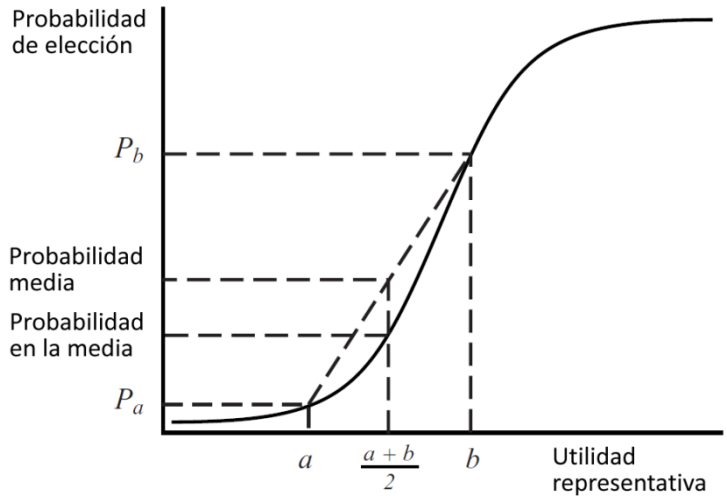

Figura 2.1: Diferencia entre la probabilidad media y la probabilidad calculada en la utilidad representativa media.

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.

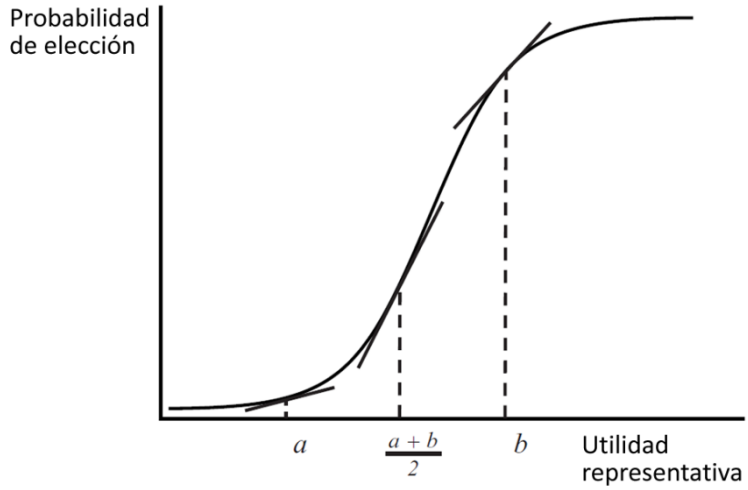

Figura 2.2: Diferencia entre la respuesta media y la respuesta calculada en la utilidad representativa media.

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.