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 $\beta_n, n = 1, \dots, N$. 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.