PARTE II

Estimación

8

Maximización numérica

8.1 Motivación

Muchos procesos de estimación implican la maximización de alguna función, como la función de verosimilitud, la de verosimilitud simulada o la de condiciones de momentos cuadráticos (*squared moment conditions*). Este capítulo describe los procedimientos numéricos que se utilizan para maximizar una función de verosimilitud. Procedimientos análogos se utilizan para maximizar otras funciones.

Conocer y ser capaz de aplicar estos procedimientos es fundamental en esta nueva era de los modelos de elección discreta. En el pasado, los investigadores adaptaron sus especificaciones a los pocos modelos prácticos que estaban a su disposición. Estos modelos se incluyeron en los paquetes de software de estimación disponibles en el mercado, de modo que el investigador podía estimar los modelos sin conocer los detalles de cómo se llevaba a cabo realmente la estimación desde una perspectiva numérica. La potencia de esta nueva ola de métodos de elección discreta es liberar al investigador para que pueda especificar modelos pensados a medida de su situación y de sus problemas. Utilizar esta libertad significa que el investigador a menudo se encontrará especificando un modelo que no es exactamente igual a uno de los modelos disponibles en el software comercial. En estos casos, tendrá que escribir código especial para su modelo especial.

El propósito de este capítulo es ayudar a hacer posible este ejercicio. Aunque por lo general no se enseñan en cursos de econometría, los procedimientos de maximización son bastante sencillos y fáciles de implementar. Una vez aprendidos, la libertad que ofrecen tiene un valor incalculable.

8.2 Notación

La función logaritmo de la verosimilitud (log-verosimilitud) tiene la forma $LL(\beta) = \sum_{n=1}^{N} \ln P_n(\beta)/N$, donde $P_n(\beta)$ es la probabilidad del resultado observado para el decisor n, N es el tamaño de la muestra y β es un vector de $K \times 1$ parámetros. En este capítulo, dividimos la función de verosimilitud por N, de modo que LL es la verosimilitud promedio en la muestra. Hacer esta división no altera la posición del máximo (dado que N es un valor fijo para una muestra dada) y sin embargo facilita la interpretación de alguno de los procedimientos. Todos los procedimientos funcionan igual dividamos o no la función log-verosimilitud por N. El lector puede verificar este hecho a medida que avancemos al observar que la N se cancela y queda excluida de las fórmulas relevantes.

El objetivo es encontrar el valor de β que maximice $LL(\beta)$. En términos de la figura 8.1, el objetivo es localizar $\hat{\beta}$. Observe que en esta figura LL es siempre negativa, ya que la verosimilitud es una probabilidad entre 0 y 1, y el logaritmo de cualquier número entre 0 y 1 es negativo. Numéricamente, el máximo se puede encontrar "subiendo" por la función de verosimilitud hasta que no podamos lograr ningún incremento adicional. El investigador especifica unos valores iniciales β_0. Cada iteración o paso, nos movemos a un nuevo valor de los parámetros en los que $LL(\beta)$ sea mayor que en el valor anterior. Llamemos al valor actual de β como β_t, valor que se alcanza después de t pasos desde los valores iniciales. La pregunta es: ¿cuál es el mejor paso que podemos tomar a continuación?, es decir, ¿cuál es el mejor valor para β_{t+1}?

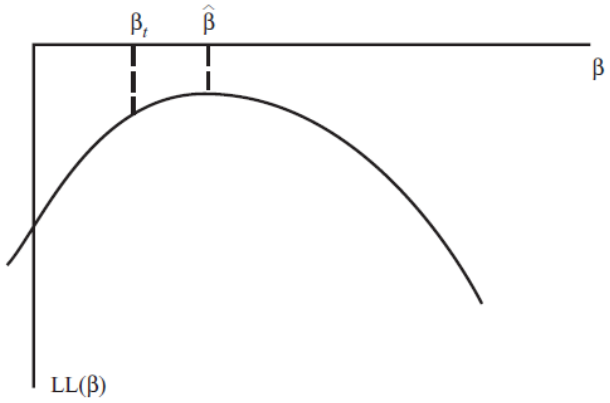

Figura 8.1. Estimación de máxima verosimilitud

El gradiente en β_t es el vector de las primeras derivadas de $LL(\beta)$ evaluadas en β_t:

$$g_t = \left(\frac{\partial LL(\beta)}{\partial \beta}\right)_{\beta_t}.$$

Este vector nos dice en qué dirección dar el siguiente paso con el fin de desplazarnos a un valor mayor de la función de verosimilitud. La matriz hessiana (o hessiano) es la matriz de segundas derivadas:

$$H_t = \left(\frac{\partial g_t}{\partial \beta'}\right)_{\beta_t} = \left(\frac{\partial^2 LL(\beta)}{\partial \beta \partial \beta'}\right)_{\beta_t}.$$

El gradiente tiene dimensiones $K \times 1$ y el hessiano $K \times K$. Como veremos, el hessiano nos puede ayudar a saber qué tan largo debemos dar el paso, mientras que el gradiente nos dice en qué dirección dar el paso.

8.3 Algoritmos

De los numerosos algoritmos de maximización que se han desarrollado a lo largo de los años, describiré sólo los más destacados, con un énfasis en el valor pedagógico de los procedimientos así como en su uso práctico. Los lectores que se sientan atraídos a estudiar más a fondo esta cuestión pueden encontrar gratificante el tratamiento dado a este tema por Judge et al. (1985, Apéndice B) y Ruud (2000).

8.3.1 Newton-Raphson

Para determinar el mejor valor de β_{t+1}, tome una aproximación de Taylor de segundo orden de la función $LL(\beta_{t+1})$ en torno a $LL(\beta_t)$:

$$(8.1) \qquad LL(\beta_{t+1}) = LL(\beta_t) + (\beta_{t+1} - \beta_t)' g_t + \frac{1}{2}(\beta_{t+1} - \beta_t)' H_t (\beta_{t+1} - \beta_t)$$

Ahora encuentre el valor de β_{t+1} que maximice esta aproximación de $LL(\beta_{t+1})$:

$$\frac{\partial LL(\beta_{t+1})}{\partial \beta_{t+1}} = g_t + H_t(\beta_{t+1} - \beta_t) = 0,$$

$$H_t(\beta_{t+1} - \beta_t) = -g_t,$$

$$\beta_{t+1} - \beta_t = -{H_t}^{-1} g_t,$$

$$\beta_{t+1} = \beta_t + \left(-{H_t}^{-1}\right) g_t.$$

El procedimiento Newton-Raphson (NR) utiliza esta fórmula. El paso a dar desde el valor actual de β al nuevo valor es $\left(-{\mathrm{H}_t}^{-1}\right) g_t$, es decir, el vector del gradiente multiplicado por el negativo de la inversa del hessiano.

Esta fórmula es intuitivamente comprensible. Considere $k = 1$, como se ilustra en la figura 8.2. La pendiente de la función de verosimilitud es g_t. La segunda derivada es el hessiano H_t, que es negativo en este gráfico, ya que la curva se ha dibujado cóncava. El negativo de este hessiano negativo es positivo y representa el grado de curvatura. Es decir, $-\mathrm{H}_t$ es la curvatura positiva. Cada paso de β es la pendiente de la función log-verosimilitud dividida por su curvatura. Si la pendiente es positiva, β se incrementa como se muestra en el primer dibujo, y si la pendiente es negativa, β se reduce como en el segundo dibujo. La curvatura determina cómo de grande se da un paso. Si la curvatura es grande, lo que significa que la pendiente cambia rápidamente tal y como se muestra en el primer dibujo de la figura 8.3, entonces es probable que estemos cerca del máximo, así que se da un paso pequeño. (Dividir el gradiente por un número grande da un número pequeño). Por el contrario, si la curvatura es pequeña, lo que significa que la pendiente no está cambiando mucho, entonces el máximo probablemente esté más lejos y por lo tanto se da un paso más grande.

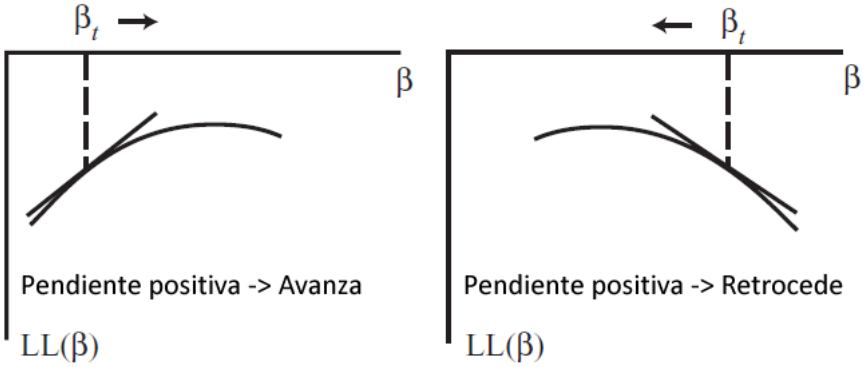

Figura 8.2. La dirección del paso a dar sigue la pendiente.

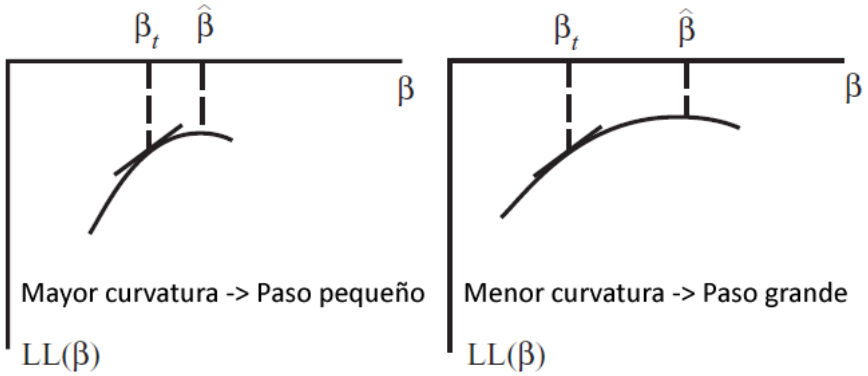

Figura 8.3. El tamaño del paso se relaciona inversamente con la curvatura.

Debemos tener en cuenta tres cuestiones relevantes en el procedimiento NR.

Cuadrático

Si $LL(\beta)$ fuese exactamente una función cuadrática de β, entonces el procedimiento NR alcanzaría el máximo en un solo paso desde cualquier valor inicial. Este hecho se puede demostrar fácilmente con $K = 1$. Si $LL(\beta)$ es cuadrática, entonces puede ser escrita como

$$LL(\beta) = a + b\beta + c\beta^2$$

El máximo es

$$\frac{\partial LL(\beta)}{\partial \beta} = b + 2c\beta = 0,$$

$$\hat{\beta} = -\frac{b}{2c}.$$

El gradiente y el hessiano son $g_t = b + 2c\beta_t$ y $H_t = 2c$, así que el procedimiento NR nos da

$$\beta_{t+1} = \beta_t - {H_t}^{-1} g_t$$

$$= \beta_t - \frac{1}{2c}(b + 2c\beta_t)$$

$$= \beta_t - \frac{b}{2c} - \beta_t$$

$$= -\frac{b}{2c} = \hat{\beta}.$$

La mayoría de las funciones log-verosimilitud no son cuadráticas, por lo que el procedimiento NR necesitará más de un paso para alcanzar el máximo. Sin embargo, saber cómo se comporta el procedimiento NR en el caso cuadrático ayuda a comprender su comportamiento con LL no cuadráticas, como veremos en la siguiente sección.

Tamaño del paso

Es posible que el procedimiento NR, al igual que otros procedimientos que se explican posteriormente, dé un paso que vaya más allá del máximo, llegando a un valor de los parámetros β en el que la $LL(\beta)$

sea inferior a la de partida. La figura 8.4 representa la situación. La función LL real está representada por la línea continua. La línea discontinua es una función cuadrática que tiene la pendiente y la curvatura que tiene LL en el punto β_t. El procedimiento NR se mueve hasta la parte superior de la función cuadrática, hasta β_{t+1}. Sin embargo, en este caso $LL(\beta_{t+1})$ es inferior a $LL(\beta_t)$.

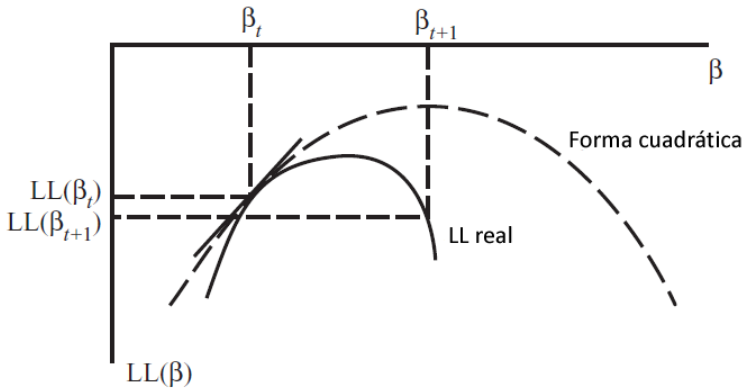

Figura 8.4. El paso podría ir más allá del máximo, hasta un valor inferior de LL

Para contemplar esta posibilidad, el paso se puede multiplicar por un escalar λ en la fórmula NR:

$$\beta_{t+1} = \beta_t + \lambda\left(-{H_t}^{-1}\right)g_t.$$

El vector $\left(-{H_t}^{-1}\right)g_t$ se denomina dirección y λ recibe el nombre de tamaño del paso. (Esta terminología es estándar, aunque $\left(-{H_t}^{-1}\right)g_t$ también contiene información sobre el tamaño del paso a través de H_t, como ya se ha explicado en relación a la figura 8.3). El tamaño de paso λ se reduce para asegurar que en cada paso del procedimiento NR logremos un aumento en $LL(\beta)$. El ajuste se realiza por separado en cada iteración, de la siguiente manera.

Empezamos con $\lambda = 1$. Si $LL(\beta_{t+1}) > LL(\beta_t)$, nos movemos a β_{t+1} y comenzamos una nueva iteración. Si $LL(\beta_{t+1}) < LL(\beta_t)$, establecemos $\lambda = 1/2$ y volvemos a intentarlo. Si, con $\lambda = 1/2$, $LL(\beta_{t+1})$ sigue siendo inferior a $LL(\beta_t)$, establecemos $\lambda = 1/4$ y volvemos a intentarlo. Continuamos este proceso hasta encontrar una λ para la que $LL(\beta_{t+1}) > LL(\beta_t)$. Si este proceso acaba generando una λ pequeña, entonces se ha avanzado poco en la búsqueda del máximo. Esto puede ser interpretado como una señal para el investigador de que puede ser necesario un procedimiento iterativo diferente.

Es posible hacer un ajuste análogo del tamaño de paso en la dirección contraria, es decir, mediante el aumento de λ cuando sea apropiado. Un caso se muestra en la figura 8.5. La parte superior (el máximo) de la función cuadrática se obtiene con un tamaño de paso de $\lambda = 1$. Sin embargo, la $LL(\beta)$ no es cuadrática, y su máximo está más lejos. El tamaño del paso se puede ajustar hacia arriba, siempre y cuando $LL(\beta)$ siga creciendo. Es decir, calculamos β_{t+1} con $\lambda = 1$ en β_{t+1}. Si $LL(\beta_{t+1}) > LL(\beta_t)$, intentamos $\lambda = 2$. Si la β_{t+1} basada en $\lambda = 2$ da un valor mayor de la función log-verosimilitud que con $\lambda = 1$, entonces probamos $\lambda = 4$, y así sucesivamente, doblando λ siempre y cuando al hacerlo elevemos aún más la función de verosimilitud. Cada vez, $LL(\beta_{t+1})$ con el valor de λ doblado se compara con el valor de λ probado anteriormente, en lugar de compararlo con el valor para $\lambda = 1$, con el fin de asegurar que cada vez que doblamos λ aumentamos la función de verosimilitud más de lo que previamente se había incrementado con una λ más pequeña. En la figura 8.5, se utiliza un tamaño de paso final de 2, ya que la función de verosimilitud con $\lambda = 4$ es menor que con $\lambda = 2$, a pesar de que es mayor que con $\lambda = 1$.

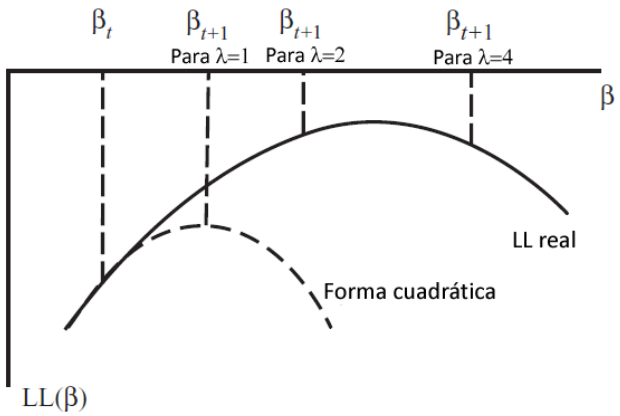

Figura 5.8. Doblamos λ mientras LL siga creciendo

La ventaja de este enfoque (incrementar λ) es que por lo general reduce el número de iteraciones necesarias para alcanzar el máximo. Podemos probar nuevos valores de λ sin necesidad de volver a calcular g_t y H_t, mientras que cada nueva iteración del procedimiento NR requiere el cálculo de estos términos. Por lo tanto, ajustar λ puede acelerar la búsqueda del máximo.

Concavidad

Si la función log-verosimilitud es globalmente cóncava, el procedimiento NR garantiza un aumento de la función de verosimilitud en cada iteración. Este hecho se demuestra de la siguiente manera. Que $LL(\beta)$ sea cóncava significa que su matriz hessiana es definida negativa en todos los valores de β. (En una dimensión, la pendiente de $LL(\beta)$ está disminuyendo, de modo que la segunda derivada es negativa). Si H es definida negativa, entonces H^{-1} también es definida negativa y $-H^{-1}$ es definida positiva. Por definición, una matriz simétrica M se dice definida positiva si $x'Mx > 0$ para cualquier $x \neq 0$. Considere la aproximación de Taylor de primer orden de $LL(\beta_{t+1})$ alrededor de $LL(\beta_t)$:

$$LL(\beta_{t+1}) = LL(\beta_t) + (\beta_{t+1} - \beta_t)' g_t$$

En el marco del procedimiento NR, $\beta_{t+1} - \beta_t = \lambda\left(-{H_t}^{-1}\right)g_t$. Sustituyendo obtenemos

$$LL(\beta_{t+1}) = LL(\beta_t) + (\lambda\left(-{H_t}^{-1}\right)g_t)' g_t$$

$$= LL(\beta_t) + \lambda g_t'\left(-{H_t}^{-1}\right)g_t.$$

Dado que $-H^{-1}$ es definida positiva, tenemos que $g_t'\left(-{H_t}^{-1}\right)g_t > 0$ y $LL(\beta_{t+1}) > LL(\beta_t)$. Tenga en cuenta que dado que esta comparación se basa en una aproximación de primer orden, un aumento en $LL(\beta)$ podría obtenerse sólo en una pequeña zona de β_t. Es decir, el valor de λ que proporciona un aumento podría ser pequeño. Sin embargo, un incremento está garantizado en cada iteración si $LL(\beta)$ es globalmente cóncava.

Supongamos que la función log-verosimilitud tiene regiones que no son cóncavas. En estas áreas, el procedimiento NR puede no encontrar un incremento de la verosimilitud. Si la función es convexa en β_t, el procedimiento NR se mueve en la dirección opuesta a la pendiente de la función log-verosimilitud. La situación se ilustra en la figura 8.6 para $K = 1$. El paso que da el procedimiento NR con un único parámetro es $LL'(\beta)/(-LL''(\beta))$, donde el símbolo ' hace referencia a la derivada. La segunda derivada es positiva en β_t, ya que la pendiente está aumentando. Por lo tanto, $-LL''(\beta)$ es negativo, y el paso queda definido en la dirección opuesta a la pendiente. Con $K > 1$, si la matriz hessiana es definida

positiva en β_t, entonces $-{H_t}^{-1}$ es definida negativa, y los pasos del procedimiento NR quedan definidos en la dirección opuesta a g_t.

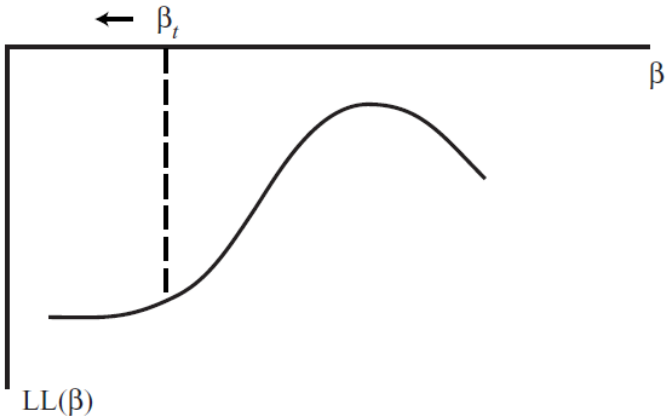

Figura 8.6. NR en la porción convexa de la función LL.

El signo del hessiano puede invertirse en estas situaciones. Sin embargo, no hay ninguna razón para el uso del hessiano allí donde la función no es cóncava, dado que el hessiano en las regiones convexas no proporciona ninguna información útil sobre dónde puede estar el máximo. Hay maneras más fáciles para lograr un incremento de LL en estas situaciones que calcular el hessiano e invertir su signo. Este tema es una de las razones que motiva el uso de otros procedimientos.

El procedimiento NR tiene dos inconvenientes. En primer lugar, el cálculo del hessiano es por lo general computacionalmente costoso. Procedimientos que eviten el cálculo del hessiano en cada iteración pueden ser mucho más rápidos. En segundo lugar, como acabamos de mostrar, el procedimiento NR no garantiza un incremento en cada paso si la función log-verosimilitud no es globalmente cóncava. Cuando $-{H_t}^{-1}$ no es definida positiva, no se puede garantizar un aumento.

Otros enfoques utilizan aproximaciones del hessiano que resuelven estas dos cuestiones. Los métodos difieren en la forma de la aproximación. Cada procedimiento define un paso como

$$\beta_{t+1} = \beta_t + \lambda M_t g_t,$$

donde M_t es una matriz $\mathrm{K} \times \mathrm{K}$. Para el procedimiento NR, $M_t = -{H_t}^{-1}$. Otros procedimientos utilizan M_ts que son más fáciles de calcular que el hessiano y que son necesariamente definidas positivas, a fin de garantizar un aumento en cada iteración, incluso en regiones convexas de la función log-verosimilitud.

8.3.2 BHHH

El procedimiento NR no utiliza el hecho de que la función que se está maximizado en realidad es la suma de logaritmos de verosimilitudes sobre una muestra de observaciones. El gradiente y el hessiano se calculan tal y como se haría en la maximización de cualquier función. Esta característica del procedimiento NR le proporciona generalidad, en el sentido de que se puede utilizar para maximizar cualquier función, no sólo un logaritmo de verosimilitud. Sin embargo, como veremos, la maximización puede ser más rápida si utilizamos el hecho de que la función que se está maximizando es una suma de términos en una muestra.

Necesitamos un poco de notación adicional para reflejar el hecho de que la función log-verosimilitud es una suma sobre observaciones. Definimos la *puntuación* (*score*) de una observación como la derivada del logaritmo de verosimilitud de esa observación respecto a los parámetros: $s_n(\beta_t) = \partial ln P_n(\beta)/\partial \beta$

evaluada en β_t. El gradiente, que hemos definido anteriormente y que se utiliza en el procedimiento NR, es la puntuación media: $g_t = \sum_n s_n(\beta_t)/N$. El producto exterior (*outer product*) de la puntuación correspondiente a la observación n es la matriz K × K

$$s_n(\beta_t)s_n(\beta_t)' = \begin{pmatrix} s_n^1 s_n^1 & s_n^1 s_n^2 & \cdots & s_n^1 s_n^K \\ s_n^1 s_n^2 & s_n^2 s_n^2 & \cdots & s_n^2 s_n^K \\ \vdots & \vdots & & \vdots \\ s_n^1 s_n^K & s_n^2 s_n^K & \cdots & s_n^K s_n^K \end{pmatrix},$$

donde s_n^k es el elemento k-ésimo de $s_n(\beta_t)$ con la dependencia de β_t omitida por conveniencia. El producto exterior medio en la muestra es $B_t = \sum_n s_n(\beta_t)s_n(\beta_t)'$ /N. Esta media está relacionada con la matriz de covarianza: si la puntuación media fuera cero, entonces B sería la matriz de covarianza de las puntuaciones de la muestra. A menudo B_t se denomina como el "producto exterior del gradiente". Este término puede ser confuso, ya que B_t no es el producto exterior de g_t. Sin embargo, refleja el hecho de que la puntuación es el gradiente de una observación específica y B_t es el producto exterior medio de estos gradientes de observación específica.

En los parámetros que maximizan la función de verosimilitud, la puntuación media es nula. El máximo se produce donde la pendiente es cero, lo que significa que el gradiente, es decir, la puntuación media, es cero. Dado que la puntuación media es cero, el producto exterior de las puntuaciones, B_t, se convierte en la varianza de las puntuaciones. Es decir, en los valores de maximización de los parámetros, B_t es la varianza de las puntuaciones en la muestra.

La varianza de las puntuaciones proporciona información importante para localizar el máximo de la función de verosimilitud. En concreto, esta variación proporciona una medida de la curvatura de la función log-verosimilitud, similar al hessiano. Supongamos que todas las personas de la muestra tienen puntuaciones similares. Si esto sucede, la muestra contiene muy poca información. La función log-verosimilitud será bastante plana en esta situación, lo que refleja el hecho de que diferentes valores de los parámetros se ajustan a los datos de forma similar. El primer dibujo de la figura 8.7 ilustra esta situación: con una log-verosimilitud casi plana, diferentes valores de β dan valores similares de $LL(\beta)$. La curvatura es pequeña cuando la varianza de las puntuaciones es pequeña. Por el contrario, puntuaciones que difieren considerablemente entre observaciones indican que las observaciones son muy diferentes y que la muestra proporciona una cantidad considerable de información. La función log-verosimilitud es muy puntiaguda, reflejando el hecho de que la muestra ofrece buena información sobre los valores de β. Alejarse de los valores maximización de β provoca una gran pérdida de ajuste. El segundo panel de la figura 8.7 ilustra esta situación. La curvatura es grande cuando la varianza de las puntuaciones es alta.

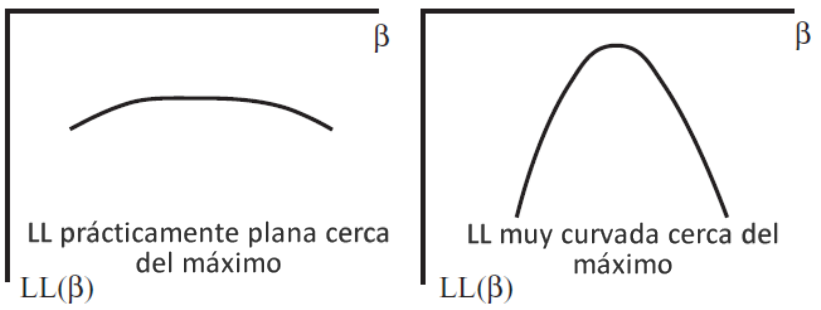

Figura 8.7. Forma de la función log-verosimilitud cerca del máximo.

Estas ideas sobre la varianza de las puntuaciones y su relación con la curvatura de la función log-verosimilitud se formalizan en la famosa *identidad de información* (*information identity*). Esta igualdad afirma que la covarianza de las puntuaciones en los verdaderos parámetros es igual a la negativa del hessiano. Demostraremos esta igualdad en la última sección de este capítulo; Theil (1971) y Ruud (2000) también proporcionan pruebas útiles y heurísticas. Sin embargo, incluso sin pruebas, intuitivamente tiene sentido que la varianza de las puntuaciones proporcione información sobre la curvatura de la función log-verosimilitud.

Berndt, Hall, Hall y Hausman (1974), en lo sucesivo BHHH (y comúnmente pronunciado triple-B H) propusieron el uso de esta relación en la búsqueda numérica del máximo de la función log-verosimilitud. En concreto, el procedimiento BHHH utiliza B_t en la rutina de optimización en lugar de $-H_t$. Cada iteración se define por

$$\beta_{t+1} = \beta_t + \lambda B_t^{-1} g_t.$$

Cada paso de este procedimiento se define igual que en el procedimiento NR, excepto que B_t se utiliza en lugar de $-H_t$. Teniendo en cuenta la explicación anterior acerca de cómo la varianza de las puntuaciones indican qué curvatura tiene la función de verosimilitud, reemplazar $-H_t$ por B_t tiene sentido.

El procedimiento BBBH tiene dos ventajas respecto al procedimiento NR:

1. B_t es mucho más rápido de calcular que H_t. En el procedimiento NR, las puntuaciones deben ser calculadas de todos modos para obtener el gradiente, por lo que el cálculo de B_t como el producto exterior medio de las puntuaciones apenas requiere tiempo adicional de computación. Por el contrario, el cálculo de H_t requiere el cálculo de las segundas derivadas de la función log-verosimilitud.
2. B_t necesariamente es definida positiva. Por consiguiente, el procedimiento BHHH garantiza un incremento de $LL(\beta)$ en cada iteración, incluso en partes convexas de la función. Utilizando la prueba dada anteriormente para el procedimiento NR cuando $-H_t$ es definida positiva, el tamaño del paso $\lambda B_t^{-1} g_t$ dado en cada iteración por el procedimiento BHHH incrementa $LL(\beta)$ para una λ suficientemente pequeña.

Nuestra exposición sobre la relación de la varianza de las puntuaciones con la curvatura de la función log-verosimilitud se puede establecer de forma un poco más precisa. Para un modelo correctamente especificado en los parámetros verdaderos, $B \to -H$ a medida que $N \to \infty$. Esta relación entre las dos matrices es una implicación de la identidad de información, expuesta con mayor detalle en la última sección. Esta convergencia sugiere que B_t puede considerarse como una aproximación a $-H_t$. A medida que el tamaño de la muestra aumenta, se espera que la aproximación sea mejor. Y la aproximación se puede esperar que sea mejor cerca de los verdaderos parámetros, donde la esperanza de la puntuación es cero y la identidad de información se cumple, que para valores de β que están lejos de los valores verdaderos. Es decir, se puede esperar que B_t sea una aproximación mejor cerca del máximo de $LL(\beta)$ que lejos del máximo.

BHHH tiene algunos inconvenientes. El procedimiento puede dar pasos pequeños que incrementan $LL(\beta)$ muy poco, especialmente cuando el proceso iterativo está lejos del máximo. Este comportamiento puede surgir porque B_t no es una buena aproximación de $-H_t$ lejos del valor verdadero, o si $LL(\beta)$ es altamente no cuadrática en el área donde está ocurriendo el problema. Si la función es altamente no cuadrática, el procedimiento NR no funciona bien, tal y como se explicó anteriormente; puesto que BHHH es una aproximación de NR, BHHH tampoco funcionará bien en estas circunstancias incluso aunque B_t sea una buena aproximación de $-H_t$.

8.3.3 BHHH-2

El procedimiento BHHH se basa en la matriz B_t, que como hemos descrito, captura la covarianza de las puntuaciones cuando la puntuación media es igual a cero (es decir, en el valor de maximización de β). Cuando el proceso iterativo no está en el máximo, la puntuación media no es cero y B_t no representa la covarianza de las puntuaciones.

Una variante del procedimiento BHHH se obtiene restando la puntuación media antes de calcular el producto exterior. Para cualquier nivel de la puntuación media, la covarianza de las puntuaciones entre decisores de la muestra es

$$W_t = \sum_n \frac{(s_n(\beta_t) - g_t)(s_n(\beta_t) - g_t)'}{N}$$

donde el gradiente g_t es la puntuación media. W_t es la covarianza de las puntuaciones alrededor de su media y B_t es el producto externo promedio de las puntuaciones. W_t y B_t son iguales cuando el gradiente medio es cero (es decir, en el valor de maximización de β), pero difieren en caso contrario.

El procedimiento de maximización puede utilizar W_t en lugar de B_t:

$$\beta_{t+1} = \beta_t + \lambda W_t^{-1} g_t.$$

Este procedimiento, que denomino BHHH-2, tiene las mismas virtudes de BHHH. W_t es necesariamente definida positiva, ya que es una matriz de covarianza, por lo que el procedimiento garantiza un aumento de $LL(\beta)$ en cada iteración. A su vez, para un modelo especificado correctamente en los parámetros verdaderos, $W \to -H$ a medida que $N \to \infty$, de modo que W_t puede considerarse como una aproximación a $-H_t$. La identidad de información establece esta equivalencia, como lo hace para B.

Para βs cercanas al valor de maximización, BHHH y BHHH-2 dan casi los mismos resultados. Pero pueden diferir en gran medida en valores lejanos al máximo. La experiencia indica, sin embargo, que los dos métodos son bastante similares en el sentido de que, o bien ambos trabajan con eficacia para una función de probabilidad dada, o ninguno de los dos lo hace. El valor principal de BHHH-2 es pedagógico y permite dilucidar la relación entre la covarianza de las puntuaciones y el producto exterior medio de las puntuaciones. Esta relación es crítica en el análisis de la identidad de información que se hace en la sección 8.7.

8.3.4 Ascenso más rápido (steepest ascent)

Este procedimiento se define por la fórmula iterativa siguiente

$$\beta_{t+1} = \beta_t + \lambda g_t.$$

La matriz que caracteriza este procedimiento es la matriz identidad I. Dado que I es definida positiva, el método garantiza un incremento en la función de verosimilitud en cada iteración. Este procedimiento recibe el nombre de "ascenso más rápido" (*"steepest ascent"*) debido a que proporciona el mayor incremento posible de $LL(\beta)$ para la distancia existente entre β_t y β_{t+1}, por lo menos para una distancia suficientemente pequeña. Cualquier otro paso de igual distancia proporciona menor incremento. Este hecho se demuestra de la siguiente manera. Considere una expansión en series de Taylor de primer orden de $LL(\beta_{t+1})$ alrededor de $LL(\beta_t)$: $LL(\beta_{t+1}) = LL(\beta_t) + (\beta_{t+1} - \beta_t)g_t$. Maximice esta expresión para $LL(\beta_{t+1})$ sujeta a que la distancia euclidiana entre $LL(\beta_{t+1})$ y $LL(\beta_{t+1})$ sea $\sqrt{k}$. Es decir, maximice con la restricción $(\beta_{t+1} - \beta_t)'(\beta_{t+1} - \beta_t) = k$. El lagrangiano es

$$L = LL(\beta_t) + (\beta_{t+1} - \beta_t)g_t - \frac{1}{2\lambda}\left[(\beta_{t+1} - \beta_t)'(\beta_{t+1} - \beta_t) - k\right],$$

y tenemos

$$\frac{\partial L}{\partial \beta_{t+1}} = g_t - \frac{1}{\lambda}(\beta_{t+1} - \beta_t) = 0,$$

$$\beta_{t+1} - \beta_t = \lambda g_t,$$

$$\beta_{t+1} = \beta_t + \lambda g_t,$$

que es la fórmula que define el procedimiento de ascenso más rápido.

A primera vista, uno podría pensar que el método de ascenso más rápido es el mejor procedimiento alcanzable, ya que da el mayor incremento posible de la función log-verosimilitud en cada paso. Sin embargo, esta propiedad del método es en realidad menos fuerte de lo que parece. Tenga en cuenta que el método se basa en una aproximación de primer orden que sólo es exacta en las cercanías de β_t. La consecuencia estrictamente correcta basada en el resultado anterior sería afirmar que existe una cierta distancia suficientemente pequeña para la cual el método de ascenso más rápido da el mayor incremento posible. Esta distinción es fundamental. La experiencia indica que los tamaños de paso a menudo resultan ser muy pequeños con este método. El hecho de que el ascenso sea mayor que para cualquier otro método para una determinada distancia no es especialmente útil cuando los pasos dados son tan pequeños. Por lo general, BHHH y BHHH-2 convergen más rápidamente que el método de ascenso más rápido.

8.3.5 DFP y BFGS

Los métodos Davidon-Fletcher-Powell (DFP) y Broyden-Fletcher-Goldfarb-Shanno (BFGS) calculan una aproximación del hessiano utilizando para ello información de más de un punto de la función de verosimilitud. Recordemos que el método NR utiliza el hessiano en β_t para determinar el paso a dar para llegar a β_{t+1}, mientras que los métodos BHHH y BHHH-2 utilizan las puntuaciones en β_t para aproximar el hessiano. En estos procedimientos sólo se utiliza información en β_t para determinar el paso. Si la función es cuadrática, la información en un punto de la función proporciona toda la información necesaria sobre la forma de la función. Por lo tanto, estos métodos funcionan bien cuando la función log-verosimilitud tiene una forma aproximadamente cuadrática. Por el contrario, los procedimientos DFP y BFGS utilizan información en varios puntos para obtener una descripción sobre la curvatura de la función log-verosimilitud.

El hessiano es la matriz de las segundas derivadas. Como tal, informa sobre cuanto cambia la curva a medida que uno se mueve a lo largo de la misma. El hessiano se define para movimientos infinitesimales. Dado que estamos interesados en dar grandes pasos en cada iteración de los algoritmos de búsqueda, comprender cómo cambia la pendiente para movimientos no infinitesimales es útil. Podemos definir un arco-hessiano sobre la base de cómo cambia el gradiente de un punto a otro. Por ejemplo, supongamos que para la función $f(x)$ la pendiente en $x = 3$ es 25 y en $x = 4$ es 19. El cambio en la pendiente para un cambio en x de una unidad es -6. En este caso, el arco-hessiano es -6, representando el cambio producido en la pendiente al dar un paso de $x = 3$ a $x = 4$.

Los procedimientos DFP y BFGS utilizan estos conceptos para aproximar el hessiano. En cada iteración del proceso se calcula el gradiente. La diferencia en el gradiente entre los diversos puntos que se han alcanzado se utiliza para calcular el arco-hessiano sobre estos puntos. Este arco-hessiano refleja el cambio que se produce en el gradiente para el movimiento real sobre la curva, en oposición al hessiano, que simplemente refleja el cambio en la pendiente para pasos infinitesimalmente pequeños alrededor de ese punto. Cuando la función log-verosimilitud no es cuadrática, el hessiano en cualquier punto proporciona muy poca información sobre la forma de la función. El arco-hessiano proporciona mejor información en estos casos.

En cada iteración, los procedimientos DFP y BFGS actualizan el arco-hessiano utilizando la información que se obtiene en el nuevo punto, es decir, utilizando el nuevo gradiente. Los dos procedimientos difieren en cómo se realiza la actualización; véase Greene (2000) para más detalles. Ambos métodos son extremadamente efectivos - por lo general mucho más eficientes que NR, BHHH, BHHH-2 o el método de ascenso más rápido. BFGS es una versión refinada de DFP. Mi experiencia indica que casi siempre funciona mejor. BFGS es el algoritmo predeterminado en las rutinas de optimización de muchos paquetes de software comercial.

8.4 Criterio de Convergencia

En teoría, el máximo de $LL(\beta)$ se produce cuando el vector gradiente es cero. En la práctica, el vector gradiente calculado nunca es exactamente igual a cero: se puede estar muy cerca, pero una serie de cálculos en una computadora no pueden producir un resultado exactamente igual a cero (a no ser, claro está, que el resultado se fije a cero a través de un operador booleano o mediante la multiplicación por cero, algo que no sucede en el cálculo del gradiente). Por ello surge la siguiente pregunta: ¿cuándo estamos suficientemente cerca del máximo para justificar la detención del proceso iterativo?

El estadístico $m_t = g_t'(-H_t^{-1})g_t$ se utiliza a menudo para evaluar la convergencia. El investigador especifica un valor pequeño para m, como $\breve{m} = 0.00001$, y determina en cada iteración si $g_t'(-H_t^{-1})g_t < \breve{m}$. Si esta desigualdad se cumple, el proceso iterativo se detiene y los parámetros en esa iteración se consideran los valores convergentes, es decir, las estimaciones. Para procedimientos distintos al método NR que utilizan una aproximación del hessiano en el proceso iterativo, dicha aproximación se utiliza en el estadístico de convergencia a fin de evitar el cálculo del hessiano real. Cerca del máximo, que es donde el criterio se vuelve relevante, se espera que todas las aproximaciones del hessiano que hemos visto sean similares al hessiano real.

m_t es el estadístico para la hipótesis de que todos los elementos del vector gradiente son cero. El estadístico se distribuye chi-cuadrado con K grados de libertad. Sin embargo, el criterio de convergencia $\breve{m}$ por lo general se suele fijar mucho más restrictivo (es decir, menor) que el valor crítico de una distribución chi-cuadrada con niveles estándar de significación, a fin de asegurar que los parámetros estimados son muy próximos a los valores de maximización. Por lo general, la hipótesis de que los elementos del gradiente son cero no puede ser rechazada para una zona bastante amplia alrededor del máximo. Esta distinción puede ilustrarse para un coeficiente estimado que tenga un estadístico-t de 1.96. En este caso, la hipótesis no podría ser rechazada si este coeficiente tiene un valor entre cero y el doble de su valor estimado. Sin embargo, no deseamos que la convergencia se defina por el hecho de haber llegado a un valor del parámetro dentro de este rango tan amplio.

Es tentador ver los pequeños cambios en β_t entre una iteración y la siguiente, y en consecuencia los pequeños aumentos en $\mathrm{LL}(\beta_t)$, como una prueba de que se ha obtenido la convergencia. Sin embargo, como se dijo anteriormente, los procedimientos iterativos pueden producir pasos pequeños debido a que la función de verosimilitud no está cerca de ser cuadrática y no porque estemos cerca del máximo. Cambios pequeños en β_t y en $\mathrm{LL}(\beta_t)$ acompañados de un vector gradiente que no esté cerca de cero indican que la rutina numérica no es eficaz en la búsqueda del máximo.

La convergencia a veces se evalúa sobre la base del propio vector gradiente en lugar de a través del estadístico de prueba m_t. Existen dos procedimientos: (1) determinar si cada elemento del vector gradiente es menor en magnitud a un cierto valor especificado por el investigador y (2) dividir cada elemento del vector gradiente por el elemento correspondiente de β, y determinar si cada uno de estos cocientes es menor en magnitud a un valor especificado por el investigador. El segundo enfoque

normaliza las unidades de los parámetros, que están determinadas por las unidades de las variables que entran en el modelo.

8.5 Máximo local y máximo global

Todos los métodos que hemos visto son susceptibles de converger en un máximo local que no es el máximo global que realmente buscamos, como se muestra en la figura 8.8. Cuando la función log-verosimilitud es globalmente cóncava, como sucede en un modelo logit con utilidad lineal en los parámetros, sólo existe un máximo y la cuestión no se plantea. Sin embargo, la mayoría de los modelos de elección discreta no son globalmente cóncavos.

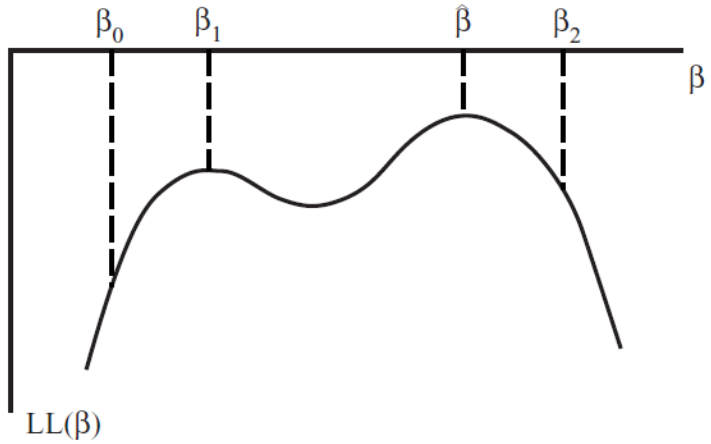

Figura 8.8. Máximo local comparado con el máximo global

Una manera de investigar el problema es utilizar varios valores de inicio en los algoritmos y observar si la convergencia se produce en los mismos valores de los parámetros. Por ejemplo, en la figura 8.8, empezar en β_0 conducirá a la convergencia en β_1. Si el investigador no probase otros valores de inicio, podría creer erróneamente que ya ha alcanzado el máximo de $\mathrm{LL}(\beta)$. Sin embargo, empezando en β_2, la convergencia se logra en $\hat{\beta}$. Mediante la comparación entre $\mathrm{LL}(\beta_1)$ y $\mathrm{LL}(\hat{\beta})$, el investigador observa que β_1 no es el valor que maximiza la verosimilitud. Liu y Mahmassani (2000) proponen una manera de seleccionar los valores de inicio que obliga al investigador a establecer límites superiores e inferiores de cada parámetro, y posteriormente seleccionar al azar los valores de inicio dentro de esos límites.

8.6 Varianza de las estimaciones

En cursos estándar de econometría, se muestra que para un modelo correctamente especificado,

$$\sqrt{N}(\hat{\beta} - \beta^*) \xrightarrow{d} N(0, (-\boldsymbol{H})^{-1})$$

a medida que $N \to \infty$, donde β^* es el vector de los parámetros verdaderos, $\hat{\beta}$ es el estimador de máxima verosimilitud y $\boldsymbol{H}$ es el hessiano esperado en la población. El negativo del hessiano esperado, $-\boldsymbol{H}$, a menudo se llama matriz de información. Expresado en palabras, la distribución en la muestra de la diferencia entre el estimador y el valor verdadero, normalizado para el tamaño de muestra, converge asintóticamente a una distribución normal centrada en cero y con covarianza igual a la inversa de la matriz de información, $-\boldsymbol{H}^{-1}$. Dado que la covarianza asintótica de $\sqrt{N}(\hat{\beta} - \beta^*)$ es $-\boldsymbol{H}^{-1}$, la covarianza asintótica de $\hat{\beta}$ es $-\boldsymbol{H}^{-1}/N$.

El tipo de letra negrita en estas expresiones indica que $\boldsymbol{H}$ es la media en la población, en contraposición a H, que es el hessiano promedio en la muestra. El investigador calcula la covarianza asintótica usando

H como una estimación de $\boldsymbol{H}$. Es decir, la covarianza asintótica de $\hat{\beta}$ se calcula como $-H^{-1}/\mathrm{N}$, donde H se evalúa en $\hat{\beta}$.

Recordemos que W es la covarianza de las puntuaciones en la muestra. En los valores de maximización de β, B también es la covarianza de las puntuaciones. Debido a la identidad de información que acabamos de exponer, y que se explica en la última sección, $-H$, que es el (negativo de la) matriz hessiana promedio en la muestra, converge a la covarianza de las puntuaciones de un modelo correctamente especificado en los parámetros verdaderos. En el cálculo de la covarianza asintótica de las estimaciones $\hat{\beta}$, cualquiera de estas tres matrices se puede utilizar como una estimación de $-\boldsymbol{H}$. La varianza asintótica de $\hat{\beta}$ se calcula como W^{-1}/N, B^{-1}/N o $-H^{-1}/N$, donde cada una de estas matrices se evalúa en $\hat{\beta}$.

Si no se especifica correctamente el modelo, entonces la covarianza asintótica de $\hat{\beta}$ es más compleja. En concreto, para cualquier modelo para el cual la puntuación esperada sea cero en los parámetros verdaderos,

$$\sqrt{N}\left(\hat{\beta} - \beta^*\right) \xrightarrow{d} N(0, \boldsymbol{H}^{-1}\boldsymbol{V}\boldsymbol{H}^{-1}),$$

donde $\boldsymbol{V}$ es la varianza de las puntuaciones en la población. Cuando el modelo está correctamente especificado, la matriz $-\boldsymbol{H} = \boldsymbol{V}$ de acuerdo a lo establecido por la identidad de información, de tal manera que $\boldsymbol{H}^{-1}\mathbf{V}\boldsymbol{H}^{-1} = -\boldsymbol{H}^{-1}$ y tenemos la fórmula para un modelo especificado correctamente. Sin embargo, si no se especifica correctamente el modelo, no se produce esta simplificación. La varianza asintótica de $\hat{\beta}$ es $\boldsymbol{H}^{-1}\mathbf{V}\boldsymbol{H}^{-1}/N$. Esta matriz se denomina *matriz de covarianza robusta (robust covariance matrix)*, ya que es válida tanto si el modelo se ha especificado correctamente como si no.

Para estimar la matriz de covarianza robusta, el investigador debe calcular el hessiano H. Si para alcanzar la convergencia se utiliza un procedimiento diferente al NR, no es necesario calcular el hessiano en cada iteración; sin embargo, éste debe ser calculado en la iteración final. Acto seguido, la covarianza asintótica se calcula como $H^{-1}\mathrm{W}H^{-1}$, o usando B en lugar de W. Esta fórmula se denomina en ocasiones el estimador "sándwich" de la covarianza, debido a que la inversa del hessiano aparece en ambos lados.

Una forma alternativa de estimar la matriz de covarianza es a través de *bootstrapping*, tal y como sugirió Efron (1979). De acuerdo a este procedimiento, el modelo se re-estima numerosas ocasiones usando diferentes muestras tomadas de la muestra original. Denominemos la muestra original como A, formada por los decisores indexados por $n = 1, \dots, N$. Es decir, la muestra original consta de N observaciones. El estimador que se obtiene de esta muestra es $\hat{\beta}$. El procedimiento de bootstrapping consiste en los siguiente pasos:

1. Seleccione aleatoriamente una muestra de N observaciones *con reemplazo* a partir de la muestra original A. Dado que el muestreo es con reemplazo, algunos decisores pueden estar representados más de una vez en esta nueva muestra y otros pueden no estar representados ninguna. Esta nueva muestra es del mismo tamaño que la muestra original, pero tiene una composición diferente a la original, ya que algunos decisores están repetidos y otros no están incluidos.
2. Re-estime el modelo usando esta nueva muestra y etiquete la estimación como β_{r}, con $\mathrm{r} = 1$ para esta primera nueva muestra.
3. Repita los pasos 1 y 2 varias veces, obteniendo estimaciones β_{r} para $\mathrm{r} = 1, \dots, ,\mathrm{R}$ donde R es el número de veces que la estimación se repite con una nueva muestra.
4. Calcule la covarianza de las estimaciones resultantes en torno a la estimación original: $\mathrm{V} = \frac{1}{\mathrm{R}}\sum_{\mathrm{r}}(\beta_{\mathrm{r}} - \hat{\beta})(\beta_{\mathrm{r}} - \hat{\beta})'$.

Esta V es una estimación de la matriz de covarianza asintótica. La varianza de muestreo de cualquier estadístico que esté basado en los parámetros se calcula de manera similar. Para estadísticos escalares $\mathrm{t}(\beta)$, la varianza muestral es $\sum_{\mathrm{r}}(\mathrm{t}(\beta_{\mathrm{r}}) - \mathrm{t}(\hat{\beta}))^2/\mathrm{R}$.

La lógica del procedimiento es la siguiente. La covarianza muestral de un estimador es, por definición, una medida de la variación de las estimaciones cuando se obtienen diferentes muestras de la población. Nuestra muestra original es una muestra de la población. Sin embargo, si esta muestra es lo suficientemente grande, es probable que sea similar a la población, de tal manera que extraer valores al azar de ella sea similar a extraer valores al azar de la población en sí misma. El método de *bootstrap* hace justamente eso: extrae valores al azar de la muestra original, con reemplazo, como un proxy de extraer valores al azar de la propia población. Las estimaciones obtenidas en las muestras creadas mediante *bootstrap* proporcionan información sobre la distribución de las estimaciones que se obtendría si realmente hubiésemos extraído muestras alternativas de la población.

La ventaja del *bootstrap* es que es conceptualmente sencillo y no se basa en fórmulas que son válidas asintóticamente pero que podrían no ser especialmente precisas para un tamaño de muestra dado. Su desventaja es que es un método computacionalmente costoso, ya que requiere la estimación del modelo numerosas veces. Efron y Tibshirant (1993) y Vinod (1993) proporcionan un estudio sobre el tema, junto con aplicaciones reales.

8.7 Identidad de información

La identidad de información establece que, para un modelo correctamente especificado en los parámetros verdaderos, $\boldsymbol{V} = -\boldsymbol{H}$, donde $\boldsymbol{V}$ es la matriz de covarianza de las puntuaciones en la población y $\boldsymbol{H}$ es el hessiano promedio de la población. La puntuación de una persona es el vector de las primeras derivadas de $ln\ P(\beta)$ de esa persona respecto a los parámetros, y el hessiano es la matriz de segundas derivadas. La identidad de información establece que, en la población, la matriz de covarianza de las primeras derivadas es igual a la matriz promedio de las segundas derivadas (en realidad, el negativo de esta matriz). Esto es un hecho sorprendente, no es algo que podríamos esperar o incluso creer si no tuviésemos una prueba. Tiene implicaciones en toda la econometría. Las implicaciones que hemos utilizado en las secciones anteriores de este capítulo se demuestran fácilmente a partir de esta igualdad. En particular:

(1) En el valor maximización de β, $W \to -H$ a medida que $N \to \infty$, donde W es la covarianza de las puntuaciones en la muestra y H es el promedio en la muestra del hessiano de cada observación. A medida que el tamaño de la muestra aumenta, la covarianza de la muestra se aproxima a la covarianza de la población: $W \to \mathbf{V}$. De forma similar, el promedio en la muestra del hessiano se acerca al promedio de la población: $H \to \mathbf{H}$. Dado que $\mathbf{V} \to -\mathbf{H}$ de acuerdo a la identidad de información, W se aproxima a la misma matriz a la que se aproxima $-H$, de modo que se aproximan entre sí.

(2) En el valor maximización de β, $B \to -H$ a medida que $N \to \infty$, donde B es el promedio en la muestra del producto exterior de las puntuaciones. En $\hat{\beta}$, la puntuación media en la muestra es igual a cero, por lo que B es igual a W. El resultado para W también aplica a B.

A continuación, vamos a demostrar la identidad de información. Necesitamos para ello ampliar nuestra notación con el fin de abarcar la población en lugar de limitarnos a la muestra. Sea $P_i(x, \beta)$ la probabilidad de que una persona que se enfrenta a unas variables explicativas x escoja la alternativa i dados los parámetros β. De las personas de la población que se enfrentan a las variables x, la proporción que elige la alternativa i es esta probabilidad calculada en los parámetros verdaderos: $S_i(x) = P_i(x, \beta^*)$ donde β^* son los parámetros verdaderos. Consideremos ahora el gradiente de $\ln P_i(x, \beta)$ respecto a β. El gradiente medio en la población es

(8.2) $$g = \int \sum_i \frac{\partial \ln P_i(x,\beta)}{\partial \beta} S_i(x) f(x) dx,$$

donde $f(x)$ es la densidad de las variables explicativas en la población. Esta expresión se puede explicar de la siguiente manera. El gradiente para las personas que se enfrentan a x y eligen i es $\partial \ln \mathrm{P_i(x,\beta)} / \partial \beta$. El gradiente medio es el promedio de este término entre todos los valores de x y todas las alternativas i. La proporción de personas que se enfrentan a un valor dado de x está dada por $f(x)$ y la proporción de personas que se enfrentan a esta x que eligen i es $S_i(x)$. Así que $S_i(x)f(x)$ es la proporción de la población que se enfrenta a x y elige i, por lo que tienen gradiente $\partial \ln \mathrm{P_i(x,\beta)} / \partial \beta$ Sumando este término sobre todos los valores de i e integrando sobre todos los valores de x (suponiendo que las xs son continuas) nos da el gradiente medio, tal y como se expresa en (8.2).

El gradiente medio en la población es igual a cero en los parámetros verdaderos. Este hecho puede ser considerado como la definición de parámetros verdaderos o el resultado de un modelo correctamente especificado. Además, sabemos que $S_i(x) = P_i(x, \beta^*)$. Sustituyendo estos hechos en (8.2), tenemos

$$0 = \int \sum_i \frac{\partial \ln P_i(x,\beta)}{\partial \beta} P_i(x,\beta) f(x) dx,$$

donde todas las funciones se evalúan en β^*. Derivamos esta ecuación respecto a los parámetros:

$$0 = \int \sum_i \left(\frac{\partial^2 \ln P_i(x,\beta)}{\partial \beta \partial \beta'} P_i(x,\beta) + \frac{\partial \ln P_i(x,\beta)}{\partial \beta} \frac{\partial P_i(x,\beta)}{\partial \beta'} \right) f(x) dx.$$

Dado que $\partial \ln \mathrm{P} / \partial \beta = (1/\mathrm{P})\ \partial \mathrm{P} / \partial \beta$ por las reglas de derivación, podemos reemplazar $[\partial \ln \mathrm{P_i(x,\beta)} / \partial \beta']\mathrm{P_i(x,\beta)}$ por $\partial \mathrm{P_i(x,\beta)} / \partial \beta'$en el último término entre paréntesis:

$$0 = \int \sum_i \left(\frac{\partial^2 \ln P_i(x,\beta)}{\partial \beta \partial \beta'} P_i(x,\beta) + \frac{\partial \ln P_i(x,\beta)}{\partial \beta} \frac{\partial \ln P_i(x,\beta)}{\partial \beta'} P_i(x,\beta) \right) f(x) dx.$$

Reorganizando

$$-\int \sum_i \frac{\partial^2 \ln P_i(x,\beta)}{\partial \beta \partial \beta'} P_i(x,\beta) f(x) dx = \int \sum_i \frac{\partial \ln P_i(x,\beta)}{\partial \beta} \frac{\partial \ln P_i(x,\beta)}{\partial \beta'} P_i(x,\beta) f(x) dx$$

Dado que todos los términos se evalúan en los parámetros verdaderos, podemos sustituir $P_i(x,\beta)$ con $S_i(x)$ para obtener

$$-\int \sum_i \frac{\partial^2 \ln P_i(x,\beta)}{\partial \beta \partial \beta'} S_i(x) f(x) dx = \int \sum_i \frac{\partial \ln P_i(x,\beta)}{\partial \beta} \frac{\partial \ln P_i(x,\beta)}{\partial \beta'} S_i(x) f(x) dx$$

El lado izquierdo es el negativo del hessiano promedio de la población, $-\mathbf{H}$. El lado derecho es el promedio del producto exterior del gradiente, que es la covarianza del gradiente, $\mathbf{V}$, ya que el gradiente medio es cero. Por lo tanto, $-\mathbf{H} = \mathbf{V}$, la identidad de información. Como se ha indicado, la matriz $-\mathbf{H}$ a menudo es llamada matriz de información.