10

Estimación asistida por simulación

10.1 Motivación

Hasta ahora hemos estudiado cómo simular probabilidades de elección pero no hemos estudiado las propiedades de los estimadores de los parámetros que se basan en estas probabilidades simuladas. En los casos que hemos presentado, simplemente hemos insertado las probabilidades simuladas en la función log-verosimilitud y hemos maximizado dicha función, de la misma forma que lo habríamos hecho si las probabilidades hubieran sido exactas. Este procedimiento parece intuitivamente razonable. Sin embargo, no hemos mostrado realmente, al menos hasta ahora, que el estimador resultante tenga propiedades deseables, como consistencia, normalidad asintótica o eficiencia. Tampoco hemos explorado la posibilidad de que otras formas de estimación puedan ser preferibles cuando usamos simulación, en lugar de las probabilidades exactas.

El propósito de este capítulo es examinar varios métodos de estimación en el contexto de la simulación. Derivaremos las propiedades de estos estimadores y mostraremos las condiciones en las que cada estimador es consistente y asintóticamente equivalente al estimador que obtendríamos si usásemos valores exactos en lugar de simulación. Estas condiciones proporcionan una guía al investigador sobre cómo debe llevarse a cabo la simulación para obtener estimadores con propiedades deseables. El análisis también pone en evidencia las ventajas y limitaciones de cada forma de estimación, facilitando así la elección del investigador entre los diferentes métodos.

Consideraremos 3 métodos de estimación:

1. *Máxima verosimilitud simulada (maximum simulated likelihood, MSL)*: Este procedimiento es igual al de máxima verosimilitud (ML) excepto que emplea las probabilidades simuladas en lugar de las probabilidades exactas. Las propiedades del método MSL han sido obtenidas, por ejemplo, por Gourieroux y Monfort, (1993), Lee (1995), y Hajivassiliou y Ruud (1994).
2. *Método de momentos simulados (method of simulated moments, MSM)*: Este procedimiento, sugerido por McFadden (1989), es el análogo simulado del método de momentos tradicional (*method of moments,* MOM). Usando el MOM tradicional en elección discreta, los residuos se definen como la diferencia entre la variable dependiente 0-1 que identifica la alternativa elegida y la probabilidad de dicha alternativa. Se identifican variables exógenas que no estén correlacionadas con los residuos del modelo en la población. Las estimaciones son los valores de los parámetros que hacen que las variables y los residuos no estén correlacionados en la

muestra. La versión simulada de este procedimiento calcula los residuos con las probabilidades simuladas en lugar de las probabilidades exactas.

3. *Método de puntuaciones simuladas (method of simulated scores, MSS)*: Como vimos en el Capítulo 8, el gradiente de la función log-verosimilitud de una observación recibe el nombre de puntuación (*score*) de la observación. El método de puntuaciones encuentra los valores de los parámetros que hacen que la puntuación media sea cero. Cuando se utilizan probabilidades exactas, el método de las puntuaciones es el mismo que el de máxima verosimilitud, ya que la función log-verosimilitud se maximiza cuando la puntuación media es cero. Hajivassiliou y McFadden (1998) sugirieron el uso de puntuaciones simuladas en lugar de puntuaciones exactas. Ellos mostraron que, dependiendo de cómo se simulan las puntuaciones, MSS puede diferir de MSL y, más importante, puede alcanzar consistencia y eficiencia bajo condiciones más relajadas.

En la siguiente sección definimos estos estimadores más formalmente y los relacionamos con sus equivalentes no simulados. A continuación describimos las propiedades de cada estimador en dos etapas. En primer lugar, se obtienen las propiedades del estimador tradicional basado en los valores exactos. En segundo lugar, se muestra cómo cambia la formulación cuando se utilizan valores simulados y no valores exactos. Mostramos que la simulación añade elementos adicionales a la distribución muestral del estimador. El análisis nos permite identificar las condiciones en que estos elementos adicionales desaparecen asintóticamente para que el estimador sea asintóticamente equivalente a su análogo no simulado. También identificamos las condiciones más relajadas en las que el estimador, aunque no sea asintóticamente equivalente a su homólogo no simulado, es sin embargo consistente.

10.2 Definición de estimadores

10.2.1 Máxima Verosimilitud Simulada (maximum simulated likelihood, MSL)

La función de verosimilitud es

$$LL(\theta) = \sum_n ln\, P_n(\theta),$$

donde θ es un vector de parámetros, $P_n(\theta)$ es la probabilidad (exacta) de la elección observada correspondiente a la observación n, y el sumatorio es sobre una muestra de N observaciones independientes. El estimador ML es el valor de θ que maximiza $LL(\theta)$. Dado que el gradiente de $LL(\theta)$ es cero en el máximo, el estimador ML también se puede definir como el valor de θ en el que

$$\sum_n s_n(\theta) = 0,$$

donde $s_n(\theta) = \partial \ln P_n(\theta) / \partial\theta$ es la puntuación de la observación n.

Sea $\check{P}_n(\theta)$ una aproximación simulada de $P_n(\theta)$. La función log-verosimilitud simulada es $SLL(\theta) = \sum_n ln\, \check{P}_n(\theta)$ y el estimador MSL es el valor de θ que maximiza $SLL(\theta)$. Dicho de forma equivalente, el estimador es el valor de θ en el que $\sum_n \check{s}_n(\theta) = 0$, donde $\check{s}_n(\theta) = \partial \ln \check{P}_n(\theta) / \partial\theta$.

Podemos echar ahora un primer vistazo a las propiedades del estimador MSL, reservando una explicación completa para la siguiente sección. El principal problema con el estimador MSL surge debido a la transformación logarítmica. Supongamos que $\check{P}_n(\theta)$ es un simulador no sesgado de $P_n(\theta)$, de manera que $E_r \check{P}_n(\theta) = P_n(\theta)$, donde la esperanza es sobre los valores extraídos al azar utilizados en la simulación. Todos los simuladores que hemos considerado son no sesgados respecto a la

verdadera probabilidad. Sin embargo, dado que el operador logarítmico es una transformación no lineal, $\ln \check{P}_n(\theta)$ es sesgado respecto a $\ln P_n(\theta)$ a pesar de que $\check{P}_n(\theta)$ es no sesgado respecto $P_n(\theta)$. El sesgo en el simulador de $\ln P_n(\theta)$ se traduce en un sesgo en el estimador MSL. Este sesgo disminuye a medida que se utilizan más valores en la simulación.

Para determinar las propiedades asintóticas del estimador MSL, se plantea la cuestión de cómo se comporta el sesgo de simulación cuando el tamaño de la muestra aumenta. La respuesta depende críticamente de la relación entre el número de valores que se utilizan en la simulación, etiquetado como R, y el tamaño de la muestra N. Si R se considera fijo, entonces el estimador MSL no converge a los parámetros reales, debido al sesgo de simulación en $\ln \check{P}_n(\theta)$. Supongamos por el contrario que R se eleva con N; es decir, el número de valores usados en la simulación aumenta con el tamaño de la muestra. En este caso, el sesgo de simulación desaparece a medida que N (y por lo tanto R) se eleva sin límite. MSL es consistente en este caso. Como veremos, si R aumenta más rápidamente que $\sqrt{N}$, MSL no sólo es consistente sino también eficiente, asintóticamente equivalente a la máxima verosimilitud con probabilidades exactas.

En resumen, si R es fijo, entonces MSL es inconsistente. Si R se eleva con N en cualquier proporción, MSL es consistente. Si R se eleva más rápido que $\sqrt{N}$, MSL es asintóticamente equivalente a ML.

La principal limitación de MSL es que es inconsistente para un R fijo. Los otros estimadores que consideraremos están motivados por el deseo de tener un estimador basado en simulación que sea consistente para un R fijo. Tanto MSM como MSS, si se estructuran adecuadamente, logran este objetivo. Este beneficio tiene un precio, sin embargo, como veremos en la siguiente sección.

10.2.2 Método de momentos simulados (method of simulated moments, MSM)

El método de momentos tradicional (*method of moments, MOM*) está motivado por el hecho de que los residuos de un modelo están necesariamente incorrelacionados en la población con factores que son exógenos al comportamiento que está siendo modelado. El estimador MOM es el valor de los parámetros que hace que los residuos en la *muestra* no estén correlacionados con las variables exógenas. Para los modelos de elección discreta, MOM se define como los parámetros que resuelven la ecuación

$$\sum_n \sum_j [d_{nj} - P_{nj}(\theta)] z_{nj} = 0, \qquad (10.1)$$

donde

- d_{nj} es la variable dependiente que identifica la alternativa elegida: $d_{nj} = 1$ si n eligió j, y $d_{nj} = 0$ en caso contrario, y
- z_{nj} es un vector de variables exógenas llamadas instrumentos (*instruments*).

Los residuos son $d_{nj} - P_{nj}(\theta)$, y el estimador MOM es el conjunto de valores de los parámetros para los que los residuos no están correlacionados con los instrumentos en la muestra.

Este estimador MOM es análogo a los estimadores MOM de los modelos de regresión estándar. Un modelo de regresión adopta la forma $y_n = x_n'\beta + \varepsilon_n$. El estimador MOM para esta regresión es la β en la que

$$\sum_n (y_n - x_n'\beta) z_n = 0$$

para un vector de instrumentos exógenos z_n. Cuando las variables explicativas en el modelo son exógenas, entonces éstas sirven como instrumentos. En este caso, el estimador MOM se convierte en el estimador de mínimos cuadrados ordinarios:

$$\sum_n (y_n - x_n'\beta)x_n = 0,$$

$$\sum_n x_n y_n = \sum_n x_n x_n'\beta,$$

$$\hat{\beta} = \left(\sum_n x_n x_n'\right)^{-1}\left(\sum_n x_n y_n\right),$$

que es la fórmula para el estimador de mínimos cuadrados. Cuando los instrumentos se especifican para que sean otras variables distintas a las variables explicativas, el estimador se convierte en el estimador de variables instrumentales estándar:

$$\sum_n (y_n - x_n'\beta)z_n = 0,$$

$$\sum_n z_n y_n = \sum_n z_n x_n'\beta,$$

$$\hat{\beta} = \left(\sum_n z_n x_n'\right)^{-1}\left(\sum_n z_n y_n\right),$$

que es la fórmula para el estimador de variables instrumentales. Este estimador es consistente si los instrumentos son independientes de ε en la población. El estimador es más eficiente cuanto más correlacionados están los instrumentos con las variables explicativas del modelo. Cuando las variables explicativas, x_n, son a su vez exógenas, los instrumentos ideales (es decir, los que dan la eficiencia más alta) son las propias variables explicativas, $z_n = x_n$.

Para los modelos de elección discreta, MOM se define de forma análoga y tiene una relación similar a otros estimadores, especialmente ML. El investigador identifica los instrumentos z_{nj} que son variables exógenas y por lo tanto independientes de los residuos $\left[d_{nj} - P_{nj}(\theta)\right]$ en la población. El estimador MOM es el valor de θ en el que la correlación de la muestra entre los instrumentos y los residuos es cero. A diferencia del caso lineal, la ecuación (10.1) no se puede resolver de forma explícita para $\hat{\theta}$. En lugar de ello, se utilizan procedimientos numéricos para encontrar el valor de θ que resuelve esta ecuación.

Al igual que sucede con la regresión, ML para un modelo de elección discreta es un caso especial de MOM. Hagamos que los instrumentos sean las puntuaciones: $z_{nj} = \partial \ln P_{nj}(\theta)/\partial\theta$. Con estos instrumentos, MOM es el mismo que ML:

$$\sum_n \sum_j [d_{nj} - P_{nj}(\theta)]\, z_{nj} = 0,$$

$$\sum_n \left\{ \left(\sum_j d_{nj} \frac{\partial \ln P_{nj}(\theta)}{\partial \theta} \right) - \left(\sum_j P_{nj}(\theta) \frac{\partial \ln P_{nj}(\theta)}{\partial \theta} \right) \right\} = 0,$$

$$\sum_n \frac{\partial \ln P_{ni}(\theta)}{\partial \theta} - \sum_n \sum_j P_{nj}(\theta) \frac{1}{P_{nj}(\theta)} \frac{\partial P_{nj}(\theta)}{\partial \theta} = 0,$$

$$\sum_n s_n(\theta) - \sum_n \sum_j \frac{\partial P_{nj}(\theta)}{\partial \theta} = 0,$$

$$\sum_n s_n(\theta) = 0,$$

que es la condición que define ML. En la tercera línea, i es la alternativa elegida, reconociendo que $d_{nj} = 0$ para todo j ≠ i. La cuarta línea utiliza el hecho de que la suma de $\partial P_{nj}(\theta)/\partial\theta$ sobre las alternativas es cero, ya que las probabilidades deben sumar 1 antes y después del cambio en θ.

Dado que MOM se convierte en ML y por lo tanto es plenamente eficiente cuando los instrumentos son las puntuaciones, las puntuaciones son llamadas instrumentos ideales. MOM es consistente siempre que los instrumentos sean independientes de los residuos del modelo. Es más eficiente cuanto mayor es la correlación entre los instrumentos y los instrumentos ideales.

Una simplificación interesante surge con el modelo logit estándar. Para el modelo logit estándar, los instrumentos ideales son las propias variables explicativas. Como se muestra en la sección 3.7.1, el estimador ML para logit estándar es el valor de θ que resuelve $\sum_n \sum_j [d_{nj} - P_{nj}(\theta)]\, x_{nj} = 0$, donde x_{nj} son las variables explicativas. Se trata de un estimador MOM con las variables explicativas como instrumentos.

Una versión simulada de MOM, llamado el método de momentos simulados (*method of simulated moments, MSM*), se obtiene mediante la sustitución de las probabilidades exactas $P_{nj}(\theta)$ por las probabilidades simuladas $\check{P}_{nj}(\theta)$. El estimador MSM es el valor de θ que resuelve

$$\sum_n \sum_j [d_{nj} - \check{P}_{nj}(\theta)]\, z_{nj} = 0,$$

para los instrumentos z_{nj}. Al igual que sucede con su analógo no simulado, MSM es consistente si z_{nj} es independiente de $d_{nj} - \check{P}_{nj}(\theta)$.

La característica importante de este estimador es que $\check{P}_{nj}(\theta)$ entra en la ecuación linealmente. Como resultado, si $\check{P}_{nj}(\theta)$ es un simulador no sesgado de $P_{nj}(\theta)$, entonces $[d_{nj} - \check{P}_{nj}(\theta)]\, z_{nj}$ es no sesgado respecto $[d_{nj} - P_{nj}(\theta)]\, z_{nj}$. Puesto que no hay sesgo de simulación en la condición de estimación, el estimador MSM es consistente, incluso cuando el número R de valores extraídos para la simulación es fijo. Por el contrario, MSL contiene sesgo de simulación debido a la transformación logarítmica de las probabilidades simuladas. Al no hacer una transformación no lineal de las probabilidades simuladas, MSM evita el sesgo de simulación.

Aun así, MSM contiene ruido de simulación (la varianza debida a la simulación). Este ruido se reduce a medida que R se eleva y desaparece cuando R aumenta sin límite. Como resultado, MSM es asintóticamente equivalente a MOM si R aumenta con N.

Al igual que su análogo no simulado, MSM es menos eficiente que MSL a no ser que se utilicen los instrumentos ideales. Sin embargo, los instrumentos ideales son funciones de $\ln P_{nj}$. Estos no pueden ser calculados de forma exacta excepto para los modelos más simples y, si son simulados utilizando la probabilidad simulada, se introduce sesgo de simulación debido a la operación logarítmica. MSM se aplica por lo general con pesos no ideales, lo que significa que se produce una pérdida de eficiencia. MSM con pesos ideales simulados sin sesgo se convierte en MSS, algo que veremos en la siguiente sección.

En resumen, MSM tiene la ventaja sobre MSL de ser consistente usando un número fijo de valores extraídos para simulación. Sin embargo, nada es gratuito, y el costo de esta ventaja es una pérdida de eficiencia cuando se utilizan pesos no ideales.

10.2.3 Método de puntuaciones simuladas (method of simulated scores, MSS)

MSS proporciona una posibilidad de lograr consistencia sin pérdida de eficiencia. El costo de esta doble ventaja es numérico: las versiones de MSS que proporcionan eficiencia tienen propiedades numéricas bastante pobres, de manera que el cálculo del estimador puede ser difícil.

El método de puntuaciones se define por la condición

$$\sum_n s_n(\theta) = 0,$$

donde $s_n(\theta) = \partial P_n(\theta)/\partial\theta$ es la puntuación de la observación n. Esta es la misma condición que define ML: cuando se utilizan probabilidades exactas, el método de puntuaciones es simplemente ML.

El método de puntuaciones simuladas reemplaza la puntuación exacta por su análogo simulado. El estimador MSS es el valor de θ que resuelve

$$\sum_n \check{s}_n(\theta) = 0,$$

donde $\check{s}_n(\theta)$ es un simulador de la puntuación. Si $\check{s}_n(\theta)$ se calcula como la derivada del logaritmo de la probabilidad simulada, es decir, $\check{s}_n(\theta) = \partial \check{P}_n(\theta)/\partial\theta$, entonces MSS es igual a MSL. Sin embargo, la puntuación se puede simular de otras maneras. Cuando la puntuación se simula de otras maneras, MSS difiere de MSL y tiene propiedades diferentes.

Supongamos que es posible construir un simulador no sesgado de la puntuación. Con este simulador, la ecuación que define el método, $\sum_n \check{s}_n(\theta) = 0$, no incorpora ningún sesgo de simulación, ya que el simulador entra en la ecuación de forma lineal. Por lo tanto, MSS es consistente con una R fija. El ruido de simulación disminuye a medida que aumenta R, de tal forma que MSS es asintóticamente eficiente, equivalente a MSL, cuando R aumenta con N. En contraste, MSL utiliza el simulador de puntuación sesgado $\check{s}_n(\theta) = \partial \check{P}_n(\theta)/\partial\theta$, que es sesgado debido al uso del operador logarítmico. Por lo tanto, MSS con un simulador de puntuación no sesgado es mejor que MSL con su simulador de puntuación sesgado, en dos aspectos: es consistente en condiciones menos estrictas (para una R fija en lugar de una R creciente con N) y es eficiente en condiciones menos estrictas (R creciente con N en cualquier proporción, en lugar de R creciendo más rápido que $\sqrt{N}$).

La dificultad en el uso de MSS está en encontrar un simulador de puntuación no sesgado. La puntuación puede ser reescrita como

$$s_n(\theta) = \frac{\partial \ln P_{nj}(\theta)}{\partial \theta} = \frac{1}{P_{nj}(\theta)} \frac{\partial P_{nj}}{\partial \theta}.$$

Un simulador no sesgado para el segundo término $\partial P_{nj}(\theta)/\partial\theta$ se obtiene fácilmente tomando la derivada de la probabilidad simulada. Puesto que la diferenciación es una operación lineal, $\partial \check{P}_{nj}(\theta)/\partial\theta$ es no sesgado respecto $\partial P_{nj}(\theta)/\partial\theta$ si $\check{P}_{nj}(\theta)$ es a su vez no sesgado respecto $P_{nj}(\theta)$. Dado que el segundo término de la puntuación puede ser simulado sin sesgo, la dificultad se presenta en la búsqueda de un simulador no sesgado para el primer término $1/P_{nj}(\theta)$. Por supuesto, simplemente tomar la inversa de la probabilidad simulada no proporciona un simulador no sesgado, ya que $\mathrm{E}_r(1/\check{P}_{nj}(\theta)) \neq 1/P_{nj}(\theta)$. Al igual que la operación logarítmica, una inversa introduce sesgo.

Una propuesta para resolver este problema se basa en el hecho de que $1/P_{nj}(\theta)$ es el número esperado de valores extraídos al azar de los términos aleatorios que se necesitan hasta lograr una "aceptación". Para ilustrar esta idea, considere la extracción de bolas de una urna que contiene muchas bolas de diferentes colores. Supongamos que la probabilidad de obtener una bola roja es 0.20. Es decir, una quinta parte de las bolas son de color rojo. ¿Cuántas extracciones se necesitarían, en promedio, para obtener una bola roja? La respuesta es 1/0.2 = 5. La misma idea se puede aplicar a las probabilidades de elección. $P_{nj}(\theta)$ es la probabilidad de que una extracción de los términos aleatorios del modelo resulte en que la alternativa j tenga la mayor utilidad. La inversa $1/P_{nj}(\theta)$ se puede simular como sigue:

1. Extraiga un valor al azar de los términos aleatorios a partir de su densidad.
2. Calcule la utilidad de cada alternativa con este valor.
3. Determine si la alternativa j tiene la mayor utilidad.
4. Si es así, catalogue el valor como una “aceptación”. Si no es así, catalogue el valor como un “rechazo” y repita los pasos 1 a 3 con un nuevo valor. Defina B^r como el número de extracciones que se realizan hasta que se obtiene la primera aceptación.
5. Realice los pasos 1 a 4 R veces, obteniendo B^r para $\mathrm{r} = 1, \ldots, \mathrm{R}$. El simulador de $1/P_{nj}(\theta)$ es $(1/\mathrm{R}) \sum_{r=1}^{R} \mathrm{B}^r$.

Este simulador es no sesgado respecto $1/P_{nj}(\theta)$. El producto de este simulador con el simulador $\partial \check{P}_{nj}(\theta)/\partial\theta$ proporciona un simulador no sesgado de la puntuación. MSS basado en este simulador de puntuación no sesgado es consistente para un R fijo y asintóticamente eficiente cuando R aumenta con N.

Por desgracia, el simulador de $1/P_{nj}(\theta)$ tiene las mismas dificultades que los simuladores de aceptación-rechazo que vimos en la sección 5.6. No hay garantía de que vayamos a obtener una aceptación dentro de un número dado de valores extraídos. Además, el simulador no es continuo en los parámetros. La discontinuidad dificulta los procedimientos numéricos que se utilizan para localizar los parámetros que resuelven la ecuación de MSS.

En resumen, MSS tiene ventajas y desventajas en relación a MSL, al igual que sucede con MSM. La comprensión de las capacidades de cada estimador permite al investigador realizar una elección informada entre ellos.

10.3 El teorema del límite central

Antes de obtener las propiedades de nuestros estimadores, es útil revisar el teorema del límite central. Este teorema proporciona la base de las distribuciones de los estimadores.

Uno de los resultados más básicos en estadísticas es que, si extraemos valores al azar de una distribución con media μ y varianza σ, la media de estos valores se distribuye normalmente con media μ

y varianza σ/N, donde N es un número grande de valores extraídos. Este resultado es el teorema del límite central, expresado de forma intuitiva en lugar de precisa. Vamos a ofrecer un desarrollo más completo y preciso de estas ideas.

Sea $t = (1/N)\sum_n t_n$, donde cada t_n es un valor extraído al azar de una distribución con media μ y varianza σ. Una realización concreta de valores extraídos al azar recibe el nombre de muestra y t es la media de la muestra. Si tomamos una muestra diferente (es decir, obtenemos diferentes valores para las extracciones de cada t_n), entonces obtenemos un valor diferente para el estadístico t. Nuestro objetivo es obtener la distribución muestral de t.

Para la mayoría de estadísticos, no podemos determinar con exactitud la distribución muestral para un tamaño de muestra dado. En su lugar, analizamos cómo se comporta la distribución muestral a medida que el tamaño de la muestra aumenta sin límite. Llegados a este punto, debemos hacer una distinción entre la distribución límite (*limiting distribution*) y la distribución asintótica (*asymptotic distribution*) de un estadístico. Supongamos que, a medida que aumenta el tamaño de la muestra, la distribución muestral del estadístico t converge a una distribución fija. Por ejemplo, la distribución muestral de t podría llegar a estar arbitrariamente cerca de una normal con media t^* y varianza σ. En este caso, decimos que $N(t^*, \sigma)$ es la distribución límite de t y que t converge en distribución a $N(t^*, \sigma)$. Denotamos esta situación como $t \xrightarrow{d} N(t^*, \sigma)$.

En muchos casos, un estadístico no tendrá una distribución límite. A medida que aumenta N, la distribución muestral sigue cambiando. La media de una muestra de valores extraídos es un ejemplo de un estadístico sin una distribución límite. Como se ha indicado anteriormente, si t es la media de una muestra de valores extraídos de una distribución con media μ y varianza σ, entonces t se distribuye normalmente con media μ y varianza σ/N. La varianza disminuye a medida que N se eleva. La distribución cambia a medida que N aumenta, siendo cada vez más y más estrecha alrededor de la media. Si se tuviera que definir una distribución límite para este caso, tendría que ser la distribución degenerada en μ: a medida que N se eleva sin límite, la distribución de t colapsa en μ. Esta distribución límite es inútil para la comprensión de la varianza del estadístico, ya que la varianza de esta distribución límite es cero. ¿Qué hacemos en este caso para comprender las propiedades del estadístico?

Si nuestro estadístico original no tiene una distribución límite, a menudo podemos transformar el estadístico de tal manera que el estadístico transformado sí tenga una distribución límite. Supongamos, como en nuestro ejemplo de una media de la muestra, que el estadístico que nos interesa no tiene una distribución límite porque su varianza disminuye a medida que aumenta N. En ese caso, podemos considerar una transformación del estadístico normalizado respecto al tamaño muestral. En particular, podemos considerar $\sqrt{N}(t-\mu)$. Supongamos que este estadístico sí tiene una distribución límite, por ejemplo, $\sqrt{N}(t-\mu) \xrightarrow{d} N(0, \sigma)$. En este caso, podemos obtener las propiedades de nuestro estadístico original a partir de la distribución límite del estadístico transformado. Recordemos, a partir de principios básicos de probabilidad, que para unos valores a y b dados, si $a(t-b)$ se distribuye normalmente con media cero y varianza σ, entonces t se distribuye normalmente con media b y varianza σ/a^2. Esta relación puede aplicarse a nuestra distribución límite. Para un N suficientemente grande, $\sqrt{N}(t-\mu)$ se distribuye aproximadamente $N(0, \sigma)$. Por lo tanto, para un N suficientemente grande, t se distribuye aproximadamente $N(\mu, \sigma/N)$. Denotamos esto como $t \sim^a N(\mu, \sigma/N)$. Observe que ésta no es la distribución límite de t, ya que t no tiene una distribución límite no degenerada. En su lugar, se denomina distribución asintótica de t, obtenida a partir de la distribución límite de $\sqrt{N}(t-\mu)$.

Ahora podemos re-expresar de forma precisa nuestros conceptos acerca de la distribución muestral de la media de la muestra. El teorema del límite central establece lo siguiente. Supongamos que t es la

media de una muestra de N valores extraídos de una distribución con media μ y varianza σ. Entonces $\sqrt{N}(t-\mu) \xrightarrow{d} N(0,\sigma)$. Con esta distribución límite, podemos decir que $t \sim^{a} N(\mu, \sigma/N)$.

Hay otra versión, más general, del teorema del límite central. En la versión que acabamos de exponer, cada t_n es una extracción de la misma distribución. Supongamos que t_n es una extracción de una distribución con media μ y varianza σ_n, para $n = 1, \ldots, N$. Es decir, cada t_n proviene de una distribución diferente; las distribuciones tienen la misma media pero diferentes varianzas. La versión generalizada del teorema del límite central establece que $\sqrt{N}(t-\mu) \xrightarrow{d} N(0,\sigma)$, donde σ es ahora la varianza media: $\sigma = (1/N)\sum_n \sigma_n$. Dada esta distribución límite, podemos decir que $t \sim^{a} N(\mu, \sigma/N)$. Vamos a utilizar ambas versiones del teorema del límite central al obtener las distribuciones de nuestros estimadores.

10.4 Propiedades de los estimadores tradicionales

En esta sección, revisaremos el procedimiento para obtener las propiedades de los estimadores y aplicaremos este procedimiento para los estimadores tradicionales, no basados en simulación. Esta exposición es el fundamento del análisis de las propiedades de los estimadores basados en simulación que abordaremos en la siguiente sección.

Denotemos el verdadero valor de los parámetros como θ^*. Los estimadores ML y MOM son las raíces de una ecuación que toma la forma

$$\sum_n g_n(\hat{\theta})/N = 0. \qquad (10.2)$$

Es decir, el estimador $\hat{\theta}$ es el valor de los parámetros que resuelve esta ecuación. Dividimos por N, a pesar de que esta división no afecta a la raíz de la ecuación, ya que al hacerlo facilitamos el cálculo de las propiedades de los estimadores. La condición establece que el valor promedio de $g_n(\theta)$ en la muestra es cero en los parámetros estimados. Para ML, $g_n(\theta)$ es la puntuación $\partial \ln P_n(\theta)/\partial\theta$. Para MOM, $g_n(\theta)$ es el conjunto de los primeros momentos de los residuos respecto a un vector de instrumentos, $\sum_j (d_{nj} - P_{nj}) z_{nj}$. La ecuación (10.2) se llama a menudo la condición de momento. En su forma no simulada, el método de puntuaciones es igual a ML y por lo tanto no necesita ser considerado por separado en esta sección. Tenga en cuenta que nosotros llamamos ecuación a (10.2) a pesar de que en realidad es un conjunto de ecuaciones, ya que $g_n(\theta)$ es un vector. Los parámetros que resuelven estas ecuaciones son los estimadores.

En cualquier valor particular de θ pueden calcularse la media y la varianza de $g_n(\theta)$ en la muestra. Etiquete la media como $g(\theta)$ y la varianza como $W(\theta)$. Estamos especialmente interesados en la media muestral y la varianza de $g_n(\theta)$ en los verdaderos parámetros, θ^*, ya que nuestro objetivo es estimar estos parámetros.

La clave para entender las propiedades de un estimador está en darse cuenta de que cada $g_n(\theta^*)$ es una extracción de una distribución de $g_n(\theta^*)$'s en la población. No sabemos los verdaderos parámetros, pero sabemos que cada observación tiene un valor de $g_n(\theta^*)$ en los verdaderos parámetros. El valor de $g_n(\theta^*)$ varía entre personas de la población. Así, extrayendo una persona de nuestra muestra, básicamente estamos extrayendo un valor de $g_n(\theta^*)$ de su distribución en la población.

La distribución de $g_n(\theta^*)$ en la población tiene una media y una varianza. Etiquete la media de $g_n(\theta^*)$ en la población como $\mathbf{g}$ y su varianza en la población como $\mathbf{W}$. La media y la varianza muestral en los verdaderos parámetros, $g(\theta^*)$ y $W(\theta^*)$, son el equivalente en la muestra a la media y varianza en la población, $\mathbf{g}$ y $\mathbf{W}$.

Asumimos que $\mathbf{g} = 0$. Es decir, asumimos que el promedio de $g_n(\theta^*)$ en la población es cero en los parámetros verdaderos. Bajo este supuesto, el estimador proporciona un análogo en la muestra a la

esperanza en la población: $\hat{\theta}$ es el valor de los parámetros en los cuales el promedio de $g_n(\theta)$ en la muestra es igual a cero, como se indica en la condición definitoria (10.2). Para ML, la suposición de que $\mathbf{g} = 0$ simplemente establece que la puntuación media en la población es cero, cuando se evalúa en los verdaderos parámetros. En cierto sentido, esto se puede considerar la definición de parámetros reales, es decir, θ^* son los parámetros en los que la función log-verosimilitud para toda la población obtiene su máximo y por lo tanto tiene pendiente cero. Los parámetros estimados son los valores que hacen que la pendiente de la función de verosimilitud en la muestra sea cero. Para MOM, el supuesto se cumple si los instrumentos son independientes de los residuos. En cierto sentido, la hipótesis con MOM es simplemente una reiteración de que los instrumentos son exógenos. Los parámetros estimados son los valores que hacen que los instrumentos y los residuos no estén correlacionados en la muestra.

Ahora consideraremos la varianza en la población de $g_n(\theta^*)$, lo que hemos denotado como $\mathbf{W}$. Cuando $g_n(\theta)$ es la puntuación, como sucede en ML, esta varianza tiene un significado especial. Como se ha mostrado en la sección 8.7, la identidad de información establece que $\mathbf{V} = -\mathbf{H}$, donde

$$-\boldsymbol{H} = -E\left(\frac{\partial^2 \ln P_n(\theta^*)}{\partial\theta\partial\theta'}\right)$$

es la matriz de información y $\mathbf{V}$ es la varianza de las puntuaciones evaluadas en los verdaderos parámetros: $\mathbf{V} = \mathrm{Var}(\partial \ln P_n(\theta^*)/\partial\theta)$. Cuando $g_n(\theta)$ es la puntuación, $\mathbf{W} = \mathbf{V}$ por definición y, por tanto, $\mathbf{W} = -\mathbf{H}$ por la identidad de información. Es decir, cuando $g_n(\theta)$ es la puntuación, $\mathbf{W}$ es la matriz de información. Para MOM con instrumentos no ideales, $\mathbf{W} \neq -\mathbf{H}$, de modo que $\mathbf{W}$ no es igual a la matriz de información.

¿Por qué es importante esta distinción? Veremos que saber si $\mathbf{W}$ es igual a la matriz de información nos permite determinar si el estimador es eficiente. La menor varianza que un estimador cualquiera puede lograr es $-\mathbf{H}^{-1}/N$. Para obtener una prueba, véase, por ejemplo, Greene (2000) o Ruud (2000). Un estimador es eficiente si su varianza alcanza este límite inferior. Como veremos, este límite inferior se logra cuando $\mathbf{W} = -\mathbf{H}$, pero no cuando $\mathbf{W} \neq -\mathbf{H}$.

Nuestro objetivo es determinar las propiedades de $\hat{\theta}$. Derivamos estas propiedades en un proceso en dos pasos. En primer lugar, se analiza la distribución de $\mathrm{g}(\theta^*)$, que, como se estableció anteriormente, es la media muestral de $g_n(\theta^*)$. En segundo lugar, la distribución de $\hat{\theta}$ se obtiene de la distribución de $\mathrm{g}(\theta^*)$. Este proceso en dos pasos no es necesariamente la forma más directa de examinar estimadores tradicionales. Sin embargo, como veremos en la siguiente sección, proporciona una forma muy conveniente de generalizar el análisis a estimadores basados en simulación.

Paso 1: Distribución de $\mathbf{g}(\boldsymbol{\theta}^*)$

Recuerde que el valor de $g_n(\theta^*)$ varía entre decisores de la población. Al tomar una muestra, el investigador está extrayendo valores $g_n(\theta^*)$ de su distribución en la población. Esta distribución tiene media cero por hipótesis y una varianza denotada por $\mathbf{W}$. El investigador calcula la media de la muestra de estos valores extraídos, $\mathrm{g}(\theta^*)$. Por el teorema del límite central, $\sqrt{\mathrm{N}}(\mathrm{g}(\theta^*) - 0) \xrightarrow{d} N(0, \mathbf{W})$, de tal manera que la media de la muestra tiene una distribución $\mathrm{g}(\theta^*) \sim^a N(0, \mathbf{W}/\mathrm{N})$.

Paso 2: Obtenga la distribución de $\widehat{\boldsymbol{\theta}}$ a partir de la distribución de $\mathbf{g}(\boldsymbol{\theta}^*)$

Podemos relacionar el estimador $\hat{\theta}$ con su término definitorio $\mathrm{g}(\theta)$ de la siguiente manera. Tome una expansión de Taylor de primer orden de $\mathrm{g}(\hat{\theta})$ alrededor $\mathrm{g}(\theta^*)$:

(10.3) $$g(\hat{\theta}) = g(\theta^*) + D[\hat{\theta} - \theta^*],$$

donde $D = \partial g(\theta^*)/\partial\theta'$. Por definición de $\hat{\theta}$ (es decir, mediante la definición de la condición (10.2)), $\mathrm{g}(\hat{\theta}) = 0$, de manera que el lado derecho de esta expansión es 0. Entonces

$$0 = g(\theta^*) + D\left[\hat{\theta} - \theta^*\right],$$

$$\hat{\theta} - \theta^* = -D^{-1}g(\theta^*),$$

$$\sqrt{N}\left(\hat{\theta} - \theta^*\right) = \sqrt{N}(-D^{-1})g(\theta^*). \qquad \textit{(10.4)}$$

Denotemos la media de $\partial g_n(\theta^*)/\partial\theta'$ en la población como $\boldsymbol{D}$. La media de $\partial g_n(\theta^*)/\partial\theta'$ en la muestra es D, tal y como se define por la ecuación (10.3). La media en la muestra D converge a la media poblacional $\boldsymbol{D}$ a medida que el tamaño de la muestra crece. Sabemos del paso 1 que $\sqrt{\mathrm{N}}\mathrm{g}(\theta^*) \xrightarrow{d} N(0, \mathbf{W})$. Usando este hecho en (10.4), tenemos

$$\sqrt{N}\left(\hat{\theta} - \theta^*\right) \xrightarrow{d} N(0, \boldsymbol{D}^{-1}\boldsymbol{W}\boldsymbol{D}^{-1}). \qquad \textit{(10.5)}$$

Esta distribución límite nos dice que $\hat{\theta} \sim^a N(\theta^*, \boldsymbol{D}^{-1}\boldsymbol{W}\boldsymbol{D}^{-1}/N)$.

Ahora podemos observar las propiedades del estimador. La distribución asintótica de $\hat{\theta}$ se centra en el valor verdadero, y su varianza disminuye a medida que el tamaño de la muestra crece. Como resultado, $\hat{\theta}$ converge en probabilidad a θ^* a medida que el tamaño de la muestra se eleva sin límite: $\hat{\theta} \xrightarrow{p} \theta$. Por consiguiente, el estimador es consistente. El estimador es asintóticamente normal. Y su varianza es $\boldsymbol{D}^{-1}\boldsymbol{W}\boldsymbol{D}^{-1}/N$, que puede ser comparada con la varianza más baja posible, $-\boldsymbol{H}^{-1}/N$, para determinar si es eficiente.

Para ML, $g_n(\cdot)$ es la puntuación, de manera que la varianza de $g_n(\theta^*)$ es la varianza de las puntuaciones: $\boldsymbol{W} = \boldsymbol{V}$. Además, la derivada media de $g_n(\theta^*)$ es la derivada media de las puntuaciones: $\boldsymbol{D} = \boldsymbol{H} = E(\partial^2 \ln P_n(\theta^*)\,/\partial\theta\partial\theta')$, donde la esperanza se calcula en la población. Por la identidad de información, $\boldsymbol{V} = -\boldsymbol{H}$. La varianza asintótica de $\hat{\theta}$ se convierte en $\boldsymbol{D}^{-1}\boldsymbol{W}\boldsymbol{D}^{-1}/N = \boldsymbol{H}^{-1}\boldsymbol{V}\boldsymbol{H}^{-1}/N = \boldsymbol{H}^{-1}(-\boldsymbol{H})\boldsymbol{H}^{-1}/N = -\boldsymbol{H}^{-1}/N$, que es la varianza más baja posible de cualquier estimador. Por lo tanto, ML es eficiente. Puesto que $\boldsymbol{V} = -\boldsymbol{H}$, la varianza del estimador ML también puede ser expresada como $\boldsymbol{V}^{-1}/N$, que tiene un significado fácilmente interpretable: la varianza del estimador es igual a la inversa de la varianza de las puntuaciones evaluadas en los verdaderos parámetros, dividida por el tamaño de la muestra.

Para MOM, $g_n(\cdot)$ es un conjunto de momentos. Si se utilizan los instrumentos ideales, entonces MOM se convierte en ML y es eficiente. Si se utilizan otros instrumentos, entonces MOM no es ML . En este caso, $\boldsymbol{W}$ es la varianza en la población de los momentos y $\boldsymbol{D}$ es la derivada media de los momentos, en lugar de la varianza y derivada media de las puntuaciones. La varianza asintótica de $\hat{\theta}$ no es igual $-\boldsymbol{H}^{-1}/N$. Por lo tanto, MOM sin pesos ideales no es eficiente.

10.5 Propiedades de los estimadores basados en simulación

Supongamos que los términos que entran en la ecuación definitoria de un estimador se obtienen por simulación en lugar de calcularse con exactitud. Sea $\check{g}_n(\theta)$ el valor simulado de $g_n(\theta)$, y $\check{g}(\theta)$ la media de estos valores simulados en la muestra, de manera que $\check{g}(\theta)$ es la versión simulada de $g(\theta)$. Llamaremos R al número de valores extraídos al azar que usamos en la simulación para cada n, y asumiremos que para cada n usamos valores extraídos de forma independiente (por ejemplo, usando extracciones separadas para cada n). Supondremos, además, que los mismos valores extraídos al azar se utilizan para cada valor de θ en el cálculo de $\check{g}_n(\theta)$. Este procedimiento evita *vibraciones* (*chatter*) en la

simulación: la diferencia entre $\breve{g}(\theta_1)$ y $\breve{g}(\theta_2)$ para dos valores diferentes de θ no se debe al uso de diferentes valores extraídos al azar.

Estos supuestos sobre los valores extraídos al azar empleados en la simulación son fáciles de implementar para el investigador y simplifican nuestro análisis considerablemente. Para los lectores interesados , Lee (1992) examina el caso en que se usan los mismos valores extraídos al azar para todas las observaciones. Pakes y Pollard (1989) proporcionan una manera de caracterizar una condición de equicontinuidad que, cuando se satisface, facilita el análisis de los estimadores basados en simulación. McFadden (1989) caracteriza esta condición de un modo diferente y muestra que se puede cumplir mediante el uso de los mismos valores extraídos al azar para cada valor de θ, que es la hipótesis que nosotros asumimos. McFadden (1996) ofrece una útil síntesis que incluye un análisis de la necesidad de prevenir la vibración (*chatter*)

El estimador se define por la condición $\breve{g}(\hat{\theta}) = 0$. Derivamos las propiedades de $\hat{\theta}$ mediante los dos mismos pasos que hemos empleado para los estimadores tradicionales.

Paso 1 : Distribución de $\breve{g}(\theta^*)$

Para identificar los distintos componentes de esta distribución, vamos a re-expresar $\breve{g}(\theta^*)$ sumando y restando algunos términos, así como reordenando:

$$\breve{g}(\theta^*) = \breve{g}(\theta^*) + g(\theta^*) - g(\theta^*) + E_r\breve{g}(\theta^*) - E_r\breve{g}(\theta^*)$$

$$= g(\theta^*) + [E_r\breve{g}(\theta^*) - g(\theta^*)] + [\breve{g}(\theta^*) - E_r\breve{g}(\theta^*)],$$

donde $\mathrm{g}(\theta^*)$ es el valor no simulado y $E_r\breve{g}(\theta^*)$ es la esperanza del valor simulado entre los valores al azar utilizados en la simulación. Sumar y restar términos obviamente no cambia $\breve{g}(\theta^*)$. Sin embargo, la posterior reordenación de los términos nos permite identificar los componentes que tienen un significado intuitivo.

El primer término $\mathrm{g}(\theta^*)$ es el mismo que aparece para el estimador tradicional. Los otros dos términos son elementos adicionales que surgen debido a la simulación. El término $E_r\breve{g}(\theta^*) - \mathrm{g}(\theta^*)$ capta el sesgo, si existe, en el simulador de $\mathrm{g}(\theta^*)$. Es la diferencia entre el valor real de $\mathrm{g}(\theta^*)$ y la esperanza del valor simulado. Si el simulador de $\mathrm{g}(\theta^*)$ es no sesgado, entonces $E_r\breve{g}(\theta^*) = \mathrm{g}(\theta^*)$ y este término desaparece. A menudo, sin embargo, el simulador de $\mathrm{g}(\theta^*)$ es sesgado. Por ejemplo, con MSL, $\breve{g}_n(\theta) = \partial \ln \breve{P}_n(\theta)/\partial\theta$, donde $\breve{P}_n(\theta)$ es un simulador no sesgado de $P_n(\theta)$. Dado que $\breve{P}_n(\theta)$ entra de forma no lineal a través del operador logarítmico, $\breve{g}_n(\theta)$ es sesgado. El tercer término, $\breve{g}(\theta^*) - E_r\breve{g}(\theta^*)$, capta el ruido de simulación, es decir, la desviación del simulador para cada valor al azar empleado, respecto a su esperanza calculada sobre todos los posibles valores al azar.

Combinando todo estos conceptos, tenemos

$$\breve{g}(\theta) = A + B + C, \qquad (10.6)$$

donde

A es el mismo que en el estimador tradicional,

B es el sesgo de la simulación,

C es el ruido de simulación.

Para ver cómo los estimadores basados en simulación difieren de sus equivalentes tradicionales, examinaremos el sesgo de simulación B y el ruido C.

Consideremos primero el ruido. Este término puede ser re-expresado como

$$C = \breve{g}(\theta^*) - E_r\breve{g}(\theta^*)$$

$$= \frac{1}{N}\sum_n [\breve{g}_n(\theta^*) - E_r\breve{g}_n(\theta^*)]$$

$$= \sum_n d_n/N,$$

donde d_n es la desviación del valor simulado para la observación n respecto su esperanza. La clave para entender el comportamiento del ruido de simulación está en observar que d_n es simplemente un estadístico para la observación n. La muestra está constituida por N extracciones al azar de este estadístico, uno para cada observación: $d_n, n = 1, \dots, N$. El ruido de simulación C es el promedio de estas N extracciones al azar. Por lo tanto, el teorema del límite central nos da la distribución de C.

En particular, para una observación dada, los valores extraídos al azar que se utilizan en la simulación proporcionan un valor particular de d_n. Si se hubieran extraído valores diferentes, entonces se habría obtenido un valor diferente de d_n. Hay una distribución de los valores de d_n sobre las posibles realizaciones de los valores al azar utilizados en simulación. La distribución tiene media cero, ya que la esperanza de los valores extraídos al azar se resta en el momento de crear d_n. Etiquetemos la varianza de la distribución como S_n/R, donde S_n es la varianza cuando se utiliza un valor extraído al azar en la simulación. Hay dos cosas a tener en cuenta acerca de esta varianza. En primer lugar, S_n/R es inversamente proporcional a R, el número de valores al azar que se utilizan en la simulación. En segundo lugar, la variación es diferente para diferentes n. Dado que$g_n(\theta^*)$ es diferente para diferentes n, la varianza de la desviación de simulación también difiere.

Extraemos un valor al azar de d_n para cada una de las N observaciones; el ruido de simulación global, C, es el promedio de estos N valores de ruido de simulación específico de cada observación. Como acabamos de establecer, cada d_n es un valor extraído de una distribución con media cero y varianza S_n/R. La versión generalizada del teorema del límite central nos permite calcular la distribución de un promedio en la muestra de valores extraídos al azar de distribuciones que tienen la misma media pero diferentes varianzas. En nuestro caso,

$$\sqrt{N}C \xrightarrow{d} N(0, \boldsymbol{S}/R),$$

donde **S** es la media de S_n en la población. Por lo tanto $C \sim^a N(0, \boldsymbol{S}/(\mathrm{N}R))$.

La característica más relevante de la varianza asintótica de C es que disminuye a medida que N se incrementa, incluso cuando R es fija. El ruido de simulación desaparece a medida que aumenta el tamaño de la muestra, incluso sin aumentar el número de valores al azar utilizados en la simulación. Este es un hecho muy importante y de gran alcance. Significa que el aumento del tamaño de la muestra es una forma de disminuir los efectos de la simulación en el estimador. El resultado es intuitivamente lógico. Básicamente, el ruido de simulación se cancela entre observaciones. La simulación de una observación podría, por casualidad, hacer la $\breve{g}_n(\theta)$ de esa observación demasiado grande. Sin embargo, la simulación para otra observación es probable que, por casualidad, sea demasiado pequeña. Promediando las simulaciones entre observaciones, los errores tienden a anularse entre sí. A medida que el tamaño de la muestra aumenta, esta propiedad de cancelación se vuelve más relevante hasta que, con muestras lo suficientemente grandes, el ruido de simulación es insignificante.

Consideremos ahora el sesgo. Si $\breve{g}(\theta)$ es un simulador no sesgado de $g(\theta)$, entonces el término de sesgo B expresado en (10.6) es cero. Si por el contrario el simulador es sesgado, como sucede con MSL, entonces el efecto de este sesgo en la distribución de $\breve{g}(\theta^*)$ debe ser considerado.

Por lo general, el término definitorio $g_n(\theta)$ es una función de un estadístico, l_n, que puede ser simulado sin sesgo. Por ejemplo, en MSL, $g_n(\theta)$ es una función de la probabilidad de elección, que puede ser simulada sin sesgo; en este caso l_n es la probabilidad. Más generalmente, l_n puede ser cualquier estadístico que se simule sin sesgo y que sirve para definir $g_n(\theta)$. Podemos escribir la dependencia en general como $g_n(\theta) = g(l_n(\theta))$ y el simulador no sesgado de $l_n(\theta)$ como $\breve{l}_n(\theta)$ donde $E_r\breve{l}_n(\theta) = l_n(\theta)$.

Ahora podemos re-expresar $\breve{g}_n(\theta)$ mediante una expansión de Taylor alrededor del valor no simulado $g_n(\theta)$:

$$\breve{g}_n(\theta) = g_n(\theta) + \frac{\partial g(l_n(\theta))}{\partial l_n}\left[\breve{l}_n(\theta) - l_n(\theta)\right] + \frac{1}{2}\frac{\partial^2 g(l_n(\theta))}{\partial l_n^2}\left[\breve{l}_n(\theta) - l_n(\theta)\right]^2,$$

$$\breve{g}_n(\theta) - g_n(\theta) = g_n'\left[\breve{l}_n(\theta) - l_n(\theta)\right] + \frac{1}{2}g_n''\left[\breve{l}_n(\theta) - l_n(\theta)\right]^2,$$

donde g_n' y g_n'' son simplemente formas abreviadas de referirse a la primera y la segunda derivada de $g_n(l(\cdot))$ respecto a l. Dado que $\breve{l}_n(\theta)$ no está sesgado respecto a $l_n(\theta)$, sabemos que $E_r g_n'\left[\breve{l}_n(\theta) - l_n(\theta)\right] = g_n'\left[E_r\breve{l}_n(\theta) - l_n(\theta)\right] = 0$. Como resultado de ello, sólo el término de la varianza permanece en la esperanza:

$$E_r\breve{g}_n(\theta) - g_n(\theta) = \frac{1}{2}g_n''E_r\left[\breve{l}_n(\theta) - l_n(\theta)\right]^2$$

$$= \frac{1}{2}g_n''Var_r\breve{l}_n(\theta).$$

Indiquemos $Var_r\breve{l}_n(\theta) = Q_n/R$ para reflejar el hecho de que la varianza es inversamente proporcional al número de valores al azar utilizados en la simulación. El sesgo de simulación es entonces

$$E_r\breve{g}(\theta) - g(\theta) = \frac{1}{N}\sum_n E_r\breve{g}_n(\theta) - g_n(\theta)$$

$$= \frac{1}{N}\sum_n g_n''\frac{Q_n}{2R}$$

$$= \frac{Z}{R},$$

donde Z es el promedio en la muestra de $g_n''Q_n/2$.

Puesto que $B = Z/R$, el valor de este estadístico normalizado para el tamaño de la muestra es

(10.7) $$\sqrt{N}B = \frac{\sqrt{N}}{R}Z.$$

Si R es fijo, entonces B es distinto de cero. Incluso peor, $\sqrt{N}\mathrm{B}$ se eleva con N, de tal manera que no tiene ningún valor límite. Supongamos que consideramos que R se incrementa con N. En este caso, el término de sesgo desaparecería asintóticamente: $B = Z/R \xrightarrow{p} 0$. Sin embargo, el término de sesgo normalizado no necesariamente desaparece. Como $\sqrt{N}$ entra en el numerador de este término, $\sqrt{N}\mathrm{B} = \left(\sqrt{N}/R\right)Z \xrightarrow{p} 0$ sólo si R aumenta más rápidamente que $\sqrt{N}$, de manera que la relación $\sqrt{N}/R$ se aproxime a cero cuando N aumenta. Si R se incrementa más lento que $\sqrt{N}$, el ratio $\sqrt{N}/R$ se eleva, de tal manera que el término de sesgo normalizado no desaparece sino que, de hecho, se hace más y más grande a medida que aumenta el tamaño de la muestra.

Podemos ahora recopilar nuestros resultados para la distribución del término definitorio normalizado por el tamaño de muestra:

$$\sqrt{N}\breve{g}(\theta^*) = \sqrt{N}(A + B + C), \tag{10.8}$$

donde

$\sqrt{N}A \xrightarrow{d} N(0, \boldsymbol{W})$, igual al estimador tradicional,

$\sqrt{N}B = \frac{\sqrt{N}}{R}Z$, captura el sesgo de simulación,

$\sqrt{N}C \xrightarrow{d} N(0, \boldsymbol{S}/R)$, captura el ruido de simulación,

Paso 2: Obtenga la distribución de $\hat{\theta}$ a partir de la distribución de $\breve{\boldsymbol{g}}(\boldsymbol{\theta}^*)$

Al igual que con los estimadores tradicionales, la distribución de $\hat{\theta}$ está directamente relacionada con la distribución de $\breve{g}(\theta^*)$. Usando la misma expansión de Taylor usada en (10.3), tenemos

$$\sqrt{N}\left(\hat{\theta} - \theta^*\right) = -\breve{D}^{-1}\sqrt{N}\breve{g}(\theta^*) = -\breve{D}^{-1}\sqrt{N}(A + B + C), \tag{10.9}$$

donde $\breve{D}$ es la derivada de $\breve{g}(\theta^*)$ respecto a los parámetros, que converge a su esperanza $\mathbf{D}$ a medida que el tamaño de la muestra crece. El estimador mismo se expresa como

$$\hat{\theta} = \theta^* - \breve{D}^{-1}(A + B + C), \tag{10.10}$$

Ahora podemos examinar las propiedades de nuestros estimadores.

10.5.1 Máxima verosimilitud simulada (maximum simulated likelihood, MSL)

Para MSL, $\breve{g}_n(\theta)$ está sesgado respecto $g_n(\theta)$. El término de sesgo en (10.9) es $\sqrt{N}\mathrm{B} = \left(\sqrt{N}/R\right)Z$. Supongamos que R aumenta con N. Si R aumenta más rápido que N, entonces

$$\sqrt{N}B = \left(\sqrt{N}/R\right)Z \xrightarrow{p} 0,$$

ya que el ratio $\sqrt{N}/R$ cae a cero. Consideremos ahora el tercer término en (10.9), que capta el ruido de simulación: $\sqrt{N}C \xrightarrow{d} N(0, \boldsymbol{S}/R)$. Dado que $\boldsymbol{S}/R$ disminuye a medida que R aumenta, tenemos que $\boldsymbol{S}/R \xrightarrow{d} 0$ a medida que $\mathrm{N} \to \infty$ cuando R aumenta con N. El segundo y tercer términos desaparecen, quedando sólo el primer término. Este primer término es el mismo que aparece en el estimador no simulado. Tenemos

$$\sqrt{N}\left(\hat{\theta} - \theta^*\right) = -\boldsymbol{D}^{-1}\sqrt{N}A \xrightarrow{d} N(0, \boldsymbol{D^{-1}WD^{-1}})$$

$$= N(0, \boldsymbol{H^{-1}VH^{-1}})$$

$$= N(0, -\boldsymbol{H^{-1}}),$$

donde la penúltima igualdad se debe a que $g_n(\theta)$ es la puntuación, y la última igualdad se debe a la identidad de información. El estimador se distribuye

$$\hat{\theta} \sim^a N(\theta^*, -\boldsymbol{H^{-1}}/N)$$

Esta es la misma distribución asintótica de ML. Cuando R aumenta más rápidamente que $\sqrt{N}$, MSL es consistente, asintóticamente normal y eficiente, y asintóticamente equivalente a ML.

Supongamos que R crece con N, pero en una proporción menor a $\sqrt{N}$. En este caso, el ratio $\sqrt{N}/R$ se hace más grande a medida que N aumenta. No hay distribución límite para $\sqrt{N}(\hat{\theta} - \theta^*)$, porque el término de sesgo, $(\sqrt{N}/R\ )$Z, crece con N. Sin embargo, el propio estimador converge en el valor verdadero. $\hat{\theta}$ depende de $(1/\mathrm{R})$Z, sin multiplicar por $\sqrt{N}$. Este término de sesgo desaparece cuando R crece a cualquier velocidad (en cualquier proporción respecto N). Por lo tanto, el estimador converge en el valor verdadero, al igual que su equivalente no simulado, lo que significa que $\hat{\theta}$ es consistente. Sin embargo, el estimador no es asintóticamente normal, ya que $\sqrt{N}(\hat{\theta} - \theta^*)$ no tiene distribución límite. Los errores estándar no se pueden calcular, y los intervalos de confianza no se pueden construir.

Cuando R es fijo, el sesgo aumenta a medida que crece N. $\sqrt{N}(\hat{\theta} - \theta^*)$ no tiene distribución límite. Además, el propio estimador, $\hat{\theta}$, contiene un sesgo $\mathrm{B} = (1/R)$Z que no desaparece a medida que el tamaño de la muestra aumenta con un R fijo. El estimador MSL no es ni consistente ni asintóticamente normal cuando R es fijo.

Las propiedades de MSL se pueden resumir de la siguiente manera:

1. Si R es fijo, MSL es inconsistente.
2. Si R se eleva más lentamente que $\sqrt{N}$, MSL es consistente pero no asintóticamente normal.
3. Si R se eleva más rápido que $\sqrt{N}$, MSL es consistente, asintóticamente normal y eficiente, y equivalente a ML.

10.5.2 Método de momentos simulados (method of simulated moments, MSM)

Para MSM con instrumentos fijos, $\check{g}_n(\theta) = \sum_j [d_{nj} - \check{P}_{nj}(\theta)]z_{nj}$, que es no sesgado respecto $g_n(\theta)$, ya que la probabilidad simulada entra linealmente en la expresión. El término de sesgo es cero. La distribución del estimador está determinada sólo por el término A, que es el mismo que en el MOM tradicional sin simulación, y por el término C, el cual refleja el ruido de simulación:

$$\sqrt{N}(\hat{\theta} - \theta^*) = -\check{D}^{-1}\sqrt{N}(A + C).$$

Supongamos que R es fijo. Dado que $\check{D}$ converge a su valor esperado $\boldsymbol{D}$, tenemos $-\sqrt{N}\check{D}^{-1}\mathrm{A} \xrightarrow{d} N(0, \boldsymbol{D^{-1}WD^{-1}})$ y $-\sqrt{N}\check{D}^{-1}\mathrm{C} \xrightarrow{d} N(0, \boldsymbol{D^{-1}}(\boldsymbol{S}/R)\boldsymbol{D^{-1}})$, de manera que

$$\sqrt{N}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, \boldsymbol{D^{-1}}[\boldsymbol{W} + \boldsymbol{S}/R]\boldsymbol{D^{-1}}).$$

La distribución asintótica del estimador es entonces

$$\hat{\theta} \sim^{a} N(\theta^{*}, \boldsymbol{D^{-1}}[\boldsymbol{W} + \boldsymbol{S}/R]\boldsymbol{D^{-1}}/N).$$

El estimador es consistente y asintóticamente normal. Su varianza es mayor que la de su equivalente no simulado en una cantidad $\boldsymbol{D^{-1}SD^{-1}}/(\mathrm{R}N)$, que refleja el ruido de simulación.

Supongamos ahora que R se eleva con N en una proporción cualquiera. La varianza adicional debida al ruido de simulación desaparece, de modo que $\hat{\theta} \sim^{a} N(\theta^{*}, \boldsymbol{D^{-1}WD^{-1}}/N)$, igual a la de su equivalente no simulado. Cuando se utilizan instrumentos no ideales, $\boldsymbol{D^{-1}WD^{-1}} \neq -\boldsymbol{H}^{-1}$ y por lo tanto, el estimador (ya sea en su forma simulada o no simulada) es menos eficiente que ML.

Si se utilizan instrumentos simulados en MSM, entonces las propiedades del estimador dependen de cómo se simulan los instrumentos. Si los instrumentos se simulan sin sesgo y con independencia de la probabilidad que entra en el residuo, entonces este MSM tiene las mismas propiedades que el MSM con pesos fijos. Si se simulan los instrumentos con sesgo y los instrumentos no son ideales, entonces el estimador tiene las mismas propiedades que MSL, excepto que no es asintóticamente eficiente, ya que la identidad de información no aplica. MSM con instrumentos ideales simulados es MSS, que se trata a continuación.

10.5.3 Método de puntuaciones simuladas (method of simulated scores, MSS)

Con MSS utilizando simuladores de puntuación no sesgados, $\breve{g}_n(\theta)$ es no sesgado respecto $g_n(\theta)$, y, por otra parte, $g_n(\theta)$ es la puntuación, de forma que la identidad de información aplica. El análisis es el mismo de MSM excepto que la identidad de información hace que el estimador sea eficiente cuando R aumenta con N. Como en el caso MSM, tenemos

$$\hat{\theta} \sim^{a} N(\theta^{*}, \boldsymbol{D^{-1}}[\boldsymbol{W} + \boldsymbol{S}/R]\boldsymbol{D^{-1}}/N),$$

que, dado que $g_n(\theta)$ es la puntuación, se convierte en

$$\hat{\theta} \sim^{a} N\left(\theta^{*}, \frac{\boldsymbol{H^{-1}}[\boldsymbol{V} + \boldsymbol{S}/R]\boldsymbol{H^{-1}}}{N}\right) = N\left(\theta^{*}, -\frac{\boldsymbol{H^{-1}}}{N} + \frac{\boldsymbol{H^{-1}SH^{-1}}}{RN}\right).$$

Cuando R es fijo, el estimador es consistente y asintóticamente normal, pero su covarianza es más grande que en el caso de ML debido al ruido de la simulación. Si R crece con N en cualquier proporción, entonces tenemos

$$\hat{\theta} \sim^{a} N(\theta^{*}, -\boldsymbol{H^{-1}}/N).$$

MSS con simuladores de puntuación no sesgados es asintóticamente equivalente a ML cuando R crece con N en cualquier proporción.

Este análisis muestra que MSS con simuladores de puntuación no sesgados tiene mejores propiedades que MSL en dos aspectos. En primer lugar, para R fijo, MSS es consistente y asintóticamente normal, mientras que MSL no es ninguna de las dos cosas. En segundo lugar, para R creciendo con N, MSS es equivalente a ML sin importar lo rápido que R esté aumentando, mientras que MSL es equivalente a ML sólo si la proporción en que crece es mayor a $\sqrt{\mathrm{N}}$.

Tal y como vimos en la sección 10.2.3, es difícil encontrar simuladores de puntuación no sesgados con buenas propiedades numéricas. MSS se emplea en ocasiones con simuladores de puntuación sesgados. En este caso, las propiedades del estimador son las mismas que las de MSL: el sesgo en las puntuaciones simuladas se traduce en sesgo en el estimador, que desaparece de la distribución límite sólo si R crece más rápidamente que $\sqrt{\mathrm{N}}$.

10.6 Solución numérica

Los estimadores se definen como el valor de θ que resuelve $\breve{g}(\theta) = 0$, donde $\breve{g}(\theta) = \sum_n \breve{g}_n(\theta)\,/N$ es la media muestral de un estadístico simulado $\breve{g}_n(\theta)$. Dado que $\breve{g}_n(\theta)$ es un vector, tenemos que resolver el conjunto de ecuaciones para los parámetros. La cuestión es: ¿cómo se resuelven numéricamente estas ecuaciones para obtener las estimaciones?

El capítulo 8 describe los métodos numéricos que permiten maximizar una función. Estos procedimientos también se pueden utilizar para resolver un conjunto de ecuaciones. Sea T el negativo del producto interior (*inner product*) del término definitorio de un estimador: $T = -\breve{g}(\theta)'\breve{g}(\theta) = -(\sum_n \breve{g}_n(\theta))'(\sum_n \breve{g}_n(\theta))/N^2$. T es necesariamente menor o igual a cero, ya que es el negativo de una suma de cuadrados. T tiene como valor máximo 0, que se consigue sólo cuando los términos cuadráticos que la componen son todos iguales a 0. Es decir, el máximo de T se alcanza cuando $\breve{g}(\theta) = 0$. Maximizar T es equivalente a resolver la ecuación $\breve{g}(\theta) = 0$. Los enfoques para resolver el problema descritos en el capítulo 8, con la excepción de BHHH, se pueden utilizar para esta maximización. BHHH no se puede utilizar debido a que el método asume que la función que está siendo maximizada es una suma de términos específicos de cada observación, mientras que T contiene el cuadrado de cada suma de términos específicos de cada observación. Los otros métodos, especialmente BFGS y DFP, han demostrado ser muy eficaces en la localización de los parámetros en los que $\breve{g}(\theta) = 0$.

Con MSL, por lo general es más fácil maximizar la función de verosimilitud simulada que maximizar T. BHHH se puede utilizar en este caso, así como el resto de métodos.