PUBLICACIONES DE 4er CURSO

Grado: Economía

Asignatura: ECONOMETRÍA III

Tema 2: Estimación
Apartado 2.2: Enfoque de la Máxima Verosimilitud

Grupos:

Profesores: Antonio Aznar y M. Teresa Aparicio

Departamento de ANÁLISIS ECONÓMICO

Curso Académico 2014/15

Facultad de
Economía y Empresa
Universidad Zaragoza

2.2 Método de la Máxima Verosimilitud

Sea **y** un vector de n observaciones muestrales de la variable y definido como $y' = (y_1, y_2, ...y_n)$. Suponemos que la distribución de probabilidad de cada elemento del vector depende de un vector de k parámetros, θ. La función de probabilidad conjunta de los n elementos de la muestra también dependerá de este vector. Esta probabilidad o densidad conjunta tiene dos interpretaciones. Para un θ dado indica la probabilidad del conjunto de n observaciones. Alternativamente, puede también ser interpretado como una función de θ manteniendo constante el conjunto de observaciones muestrales. En este caso, se llama la función de verosimilitud. La definición formal es

Función de Verosimilitud= $L(\theta; y) = f(\theta; y)$

Los estimadores máximo-verosímiles (MV) son aquellos que maximizan la función de verosimilitud. Sea $\tilde{\theta}$ el vector de estos estimadores. Maximizar la verosimilitud es equivalente a maximizar su logaritmo definido como:

$$l(\theta) = \ln L(\theta; y)$$

Se utiliza el logaritmo porque, en muchas ocasiones, son más fáciles las derivaciones. Condición necesaria para alcanzar el máximo es que la primera derivada con respecto a los parámetros sea igual a cero. Esta primera derivada es el gradiente definido como.

$$d(\theta) = \frac{\partial l(\theta)}{\partial \theta}$$

Notar que el gradiente es un vector con k elementos. Otro concepto relevante es el de la Matriz de Información, que se define como:

$$I(\theta) = -E \frac{\partial^2 l(\theta)}{\partial \theta \partial \theta'}$$

Ejemplo 1. Distribución de Poisson

Supongamos una serie temporal compuesta de T extracciones iid a partir de la siguiente distribución de probabilidad

$$f(y_t;\theta)=\frac{\theta^{y_t}\exp[-\theta]}{y_t!},\qquad y_t=0,1,2,3,....$$

En donde $\theta>0$ es un parámetro desconocido. Notar que se trata de una variable discreta. Por ser independientes las observaciones muestrales, la función de verosimilitud se puede escribir como

$$L(\theta)=\prod_{i=1}^{T} f\left(\theta;y_t\right)=\frac{\theta^{\sum y_t}\exp[-T\theta]}{y_1!y_2!.....y_T!}$$

El logaritmo de la función de verosimilitud es

$$l(\theta)=\ln\theta\sum y_t-T\theta-\ln(y_1!....y_T!)$$

Suponer ahora que tenemos tres observaciones, T=3, con valores 8, 3 y 4. Teniendo en cuenta que $y_1!y_2!y_3!=5806080$ y que $\ln(y_1!y_2!y_3!)=15,57$ los valores de las dos funciones pueden verse en la siguiente tabla.

Valor de θ	$L(\theta)$	$l(\theta)$
1	8,5749E-09	-18,5744
2	1,3989E-05	-11,1772
3	0,0003	-8,0952
4	0,0011	-6,7800
5	**0,0016**	**-6,4328**
6	0,0012	-6,6980
7	0,0006	-7,3857
8	0,0002	-8,3827
9	6,6650E-05	-9,6160

10	1,6116E-05	-11,0356

Como puede verse en esta tabla el máximo valor de ambas funciones corresponde al valor del parámetro igual a 5. Por eso se dice que este valor es la estimación máximo-verosimil. Luego demostraremos analíticamente este resultado.

El gradiente, que es la primera derivada del logaritmo respecto al parámetro, toma la forma siguiente,

$$d(\theta) = \frac{\partial l(\theta)}{\partial \theta} = \frac{1}{\theta}\sum y_t - T$$

La segunda derivada con respecto al parámetro es

$$H(\theta) = \frac{\partial^2 l(\theta)}{\partial \theta^2} = -\frac{\sum y_t}{\theta^2}$$

La matriz de Información (en este caso, el elemento de información porque es un escalar) toma la forma siguiente

$$I(\theta) = -EH(\theta) = \frac{\sum Ey_t}{\theta^2} = \frac{T}{\theta}$$

La cota de Cramer-Rao es igual a la inversa de la matriz de información, es decir

$$\text{Cota de Cramer-Rao} = I\left(\theta\right)^{-1}$$

Este concepto es importante porque establece el nivel mínimo de varianza que puede alcanzar un estimador insesgado. Es la forma de determinar cuando un estimador insesgado es eficiente.

En un marco asintótico decimos que un estimador consistente, $\tilde{\theta}$, es eficiente si la matriz de varianzas y covarianzas de $\sqrt{T}(\tilde{\theta} - \theta)$ es igual a

$$\lim\left[T^{-1} I(\theta)\right]^{-1}$$

Estimadores Máximo-verosímiles(MV)

Los estimadores MV son aquellos que maximizan la función de verosimilitud. La condición necesaria para maximizar esta función es que las primeras derivadas de su logaritmo con respecto a los parámetros sean iguales a cero. O, equivalentemente, son aquellos valores de los parámetros que hacen que los elementos del gradiente sean cero. Es decir, son aquellos que hacen que

$$d(\tilde{\theta}) = 0$$

En nuestra ilustración podemos escribir

$$\frac{1}{\tilde{\theta}} \sum y_t - T = 0$$

De donde el estimador máximo-verosimil es

$$\tilde{\theta} = \frac{\sum y_t}{T} = \bar{y}$$

Notar que la estimación máximoverosimil del parámetro en el ejemplo con tres observaciones es (8+4+3)/3=5, que es el valor que maximizaba tanto la función de verosimilitud como su logaritmo.

Veamos las propiedades de este estimador. Teniendo en cuenta que $Ey_t = \theta$ y que $Var(y_t) = \theta$ entonces se tiene que

$$E\tilde{\theta} = \frac{\sum Ey_t}{T} = \theta$$

Por lo tanto, es un estimador insesgado. En cuanto a la varianza se tiene que, por ser independientes

$$Var(\tilde{\theta}) = \frac{1}{T^2} \sum Var(y_t) = \frac{\theta}{T}$$

Se ve que se trata de un estimador eficiente porque su varianza coincide con la inversa del elemento de información.

Asintóticamente, el estimador es consistente porque el sesgo es cero y la varianza tiende a cero conforme el tamaño muestral crece. Si consideramos la expresión

$$\sqrt{T}(\tilde{\theta}-\theta)$$

La esperanza y varianza son

$$E\sqrt{T}(\tilde{\theta}-\theta)=0$$

$$Var(\sqrt{T}(\tilde{\theta}-\theta))=\theta$$

Este valor de la varianza coincide con la inversa del elemento de información dividido por T. Por lo tanto, el estimador máximo-verosimil es asintóticamente insesgado y eficiente.

Ejemplo 2. Distribución Normal

Sea $y_1, y_2, ...y_T$ una muestra de tamaño T obtenida siguiendo el Muestreo Aleatorio Simple (MAS), a partir de una población cuyos elementos se generan siguiendo el siguiente esquema:

$$y_t=\beta+u_t$$

En donde u_t se distribuye como i.i.d. $N(0,\sigma^2)$, de forma que

$$Ey_t=\beta$$

$$V(y_t)=E(y_t-Ey_t)^2=Eu_t^2=\sigma^2$$

$$Cov(y_t,y_{t'})=0$$

Sea θ el vector de parámetros que define la distribución de probabilidad de $y_t, \theta=(\beta,\sigma^2)'$. La **función de verosimilitud** de cualquier elemento de la muestra, asumiendo la hipótesis de normalidad, puede escribirse como:

$$L(\theta; y_t) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y_t - \beta)^2\right\}$$

Teniendo en cuenta la independencia de los elementos muestrales, la verosimilitud para el conjunto de la muestra viene dada por

$$L(\theta; y_1, ...y_T) = L(\theta; y_1) \times ..... \times L(\theta; y_T)$$

Por lo que,

$$L(\theta; y_1, ..y_T) = (2\pi\sigma^2)^{-\frac{T}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum(y_t - \beta)^2\right\}$$

A partir de ahora denotaremos la verosimilitud de la muestra con $L(\theta)$. **El logaritmo de la función de verosimilitud,** que denotaremos con $l(\theta)$, toma la forma siguiente

$$l(\theta) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum(y_t - \beta)^2$$

El gradiente que denotamos con $d(\theta)$ es el vector de las derivadas parciales del logaritmo de la función de verosimilitud respecto a todos los elementos del vector de parámetros, es decir,

$$d(\theta) = \begin{bmatrix} \frac{\partial l(\theta)}{\partial \beta} \\ \frac{\partial l(\theta)}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{\sum(y_t - \beta)}{\sigma^2} \\ -\frac{T}{2\sigma^2} + \frac{\sum(y_t - \beta)^2}{2\sigma^4} \end{bmatrix}$$

La matriz de segundas derivadas que denotamos con $H(\theta)$ viene dada por

$$H(\theta) = \begin{bmatrix} \frac{\partial^2 l(\theta)}{\partial\beta\partial\beta'} & \frac{\partial^2 l(\theta)}{\partial\beta\partial\sigma^2} \\ \frac{\partial^2 l(\theta)}{\partial\sigma^2\partial\beta} & \frac{\partial^2 l(\theta)}{\partial\sigma^2\partial\sigma^2} \end{bmatrix}$$

Y aplicando las reglas estándar de derivación se obtiene

$$H(\theta)=\begin{bmatrix} -\frac{T}{\sigma^2} & -\frac{\sum(y_t-\beta)}{\sigma^4} \\ -\frac{\sum(y_t-\beta)}{\sigma^4} & \frac{T}{2\sigma^4}-\frac{\sum(y_t-\beta)^2}{\sigma^6} \end{bmatrix}$$

La matriz de información que denotamos con $I(\theta)$ se define como

$$I(\theta)=-EH(\theta)$$

y teniendo en cuenta que

$$E(y_t-\beta)=Eu_t=0$$

$$E\sum(y_t-\beta)^2=E\sum u_t^2=T\sigma^2$$

Se obtiene que

$$I(\theta)=\begin{bmatrix} \frac{T}{\sigma^2} & 0 \\ 0 & \frac{T}{2\sigma^4} \end{bmatrix}$$

La cota de Cramer-Rao es igual a la inversa de la matriz de información, es decir

$$\text{Cota de Cramer-Rao}=I\left(\theta\right)^{-1}$$

Este concepto es importante porque establece el nivel mínimo de varianza que puede alcanzar un estimador insesgado. Es la forma de determinar cuando un estimador insesgado es eficiente.

En un marco asintótico decimos que un estimador consistente, $\tilde{\theta}$, es eficiente si la matriz de varianzas y covarianzas de $\sqrt{T}(\tilde{\theta}-\theta)$ es igual a

$$\lim\left[T^{-1}I(\theta)\right]^{-1}$$

Estimadores Máximo-verosímiles(MV)

Los estimadores MV son aquellos que maximizan la función de verosimilitud. La condición necesaria para maximizar esta función es que las primeras derivadas de

su logaritmo con respecto a los parámetros sean iguales a cero. O, equivalentemente, son aquellos valores de los parámetros que hacen que los elementos del gradiente sean cero. Es decir, son aquellos que hacen que

$$d(\tilde{\theta}) = 0$$

En nuestra ilustración podemos escribir

$$d(\tilde{\theta}) = \begin{bmatrix} \dfrac{\sum (y_t - \tilde{\beta})}{\tilde{\sigma}^2} \\ -\dfrac{T}{2\tilde{\sigma}^2} + \dfrac{\sum (y_t - \tilde{\beta})^2}{2\tilde{\sigma}^4} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

A partir de estas dos ecuaciones podemos definir los estimadores MV

$$\tilde{\theta} = \begin{bmatrix} \tilde{\beta} \\ \tilde{\sigma}^2 \end{bmatrix} = \begin{bmatrix} \dfrac{\sum y_t}{T} \\ \dfrac{\sum (y_t - \tilde{\beta})^2}{T} \end{bmatrix}$$

Derivemos las propiedades de estos estimadores. Por el resultado ya comentado, $Ey_t = \beta$ para todo t, se obtiene que

$$E(\tilde{\beta}) = \frac{1}{T} E(\sum y_t) = \frac{T\beta}{T} = \beta$$

Por lo tanto, es un estimador insesgado. En cuanto a la varianza,

$$Var(\tilde{\beta}) = \frac{1}{T^2} Var(\sum y_t) = \frac{T\sigma^2}{T^2} = \frac{\sigma^2}{T}$$

Es un estimador eficiente porque siendo insesgado alcanza la cota de Cramer-Rao. Es también consistente y asintóticamente eficiente porque la varianza de $\sqrt{T}(\tilde{\beta} - \beta)$, conforme la muestra crece es igual a σ^2.

En cuanto a las propiedades de $\tilde{\sigma}^2$ tener en cuenta que

$$y_t - \tilde{\beta} = \beta + u_t - \tilde{\beta} = u_t - (\tilde{\beta} - \beta)$$

Por lo tanto,

$$E(\tilde{\sigma}^2) = \frac{1}{T} E\left\{\sum u_t^2 + T(\tilde{\beta} - \beta)^2 - 2(\tilde{\beta} - \beta)\sum u_t\right\} =$$

$$\frac{1}{T}\left\{\sum Eu_t^2 + TE(\tilde{\beta} - \beta)^2 - 2E\bar{u}\sum u_t\right\} = \sigma^2 + \frac{\sigma^2}{T} - 2\frac{\sigma^2}{T} = \sigma^2 - \frac{\sigma^2}{T}$$

Se trata de un estimador sesgado pero con un sesgo, $\frac{\sigma^2}{T}$, que tiende a cero. De la misma forma se puede demostrar que

$$Var(\tilde{\sigma}^2) = \frac{2\sigma^4}{T}$$

Por lo tanto, el estimador MV de la varianza es consistente porque tanto el sesgo como la varianza tienden a cero. Además, se tiene que

$$Var\left\{\sqrt{T}(\tilde{\sigma}^2 - \sigma^2)\right\} = 2\sigma^4$$

Lo que nos permite decir que el estimador $\tilde{\sigma}^2$ es consistente y asintóticamente eficiente.