PUBLICACIONES DE 4er CURSO

Grado: Economía

Asignatura: ECONOMETRÍA III

Suplemento docente nº 2:
Estimación Máxima-Verosímil. Contrastes W, LM, y LR.

Autores: Antonio Aznar y Mª Teresa Aparicio

Profesores: Antonio Aznar

Departamento de ANÁLISIS ECONÓMICO

Curso Académico 2015/16

Facultad de
Economía y Empresa
Universidad Zaragoza

2.1 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

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

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

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

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-verosímil es

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

Notar que la estimación máximo-verosímil 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-verosímil es asintóticamente insesgado y eficiente.

2.2.- CONTRASTES W, LM Y LR

Hemos visto en la Sección anterior que la Región Crítica de un contraste se determinaba a partir del cociente de los valores tomados por la función de

verosimilitud sustituyendo los parámetros por sus estimaciones con y sin restricción. En particular, se definía una distancia entre los valores que toma la función de verosimilitud, y si esta distancia era grande se rechazaba la hipótesis nula.

Pero no hay un procedimiento único para dar cuenta de esta distancia. En la literatura se han desarrollado otras formas de tratar esa distancia y nosotros vamos a prestar atención a tres de ellas asociadas con los llamados contrastes de la Wald (W), Multiplicadores de Lagrange (LM) y Razón de Verosimilitud (LR).

En principio, la argumentación la llevaremos a cabo utilizando un caso muy simple en el que la muestra de tamaño T se ha obtenido a partir de una población con media μ y varianza σ^2 conocida.

En aquellos casos en que sea de interés para el contenido de otras secciones posteriores, los resultados se extenderán a un marco más general en el que la función de verosimilitud depende de un vector de k parámetros , θ , y se pretende contrastar las r restricciones lineales que escribimos como:

$$R\,\theta = q \qquad (1)$$

En donde R es una matriz r.k y q es un vector de r elementos.

La función de verosimilitud para el caso simple viene dada por:

$$L\left(\mu;\sigma^2,y\right) = (2\pi)^{-T/2}(\sigma^2)^{-T/2}\exp\left\{-\frac{1}{2\sigma^2}\Sigma\left(y_t-\mu\right)^2\right\}$$

Su logaritmo puede escribirse como:

$$\ell(\mu) = \log L(\,) = -\frac{T}{2}\log 2\pi - \frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\Sigma\,(y_t-\mu)^2$$

<u>Estimación MV sin restricciones</u> ($\tilde{\mu}$)

El estimador MV sin restricciones es aquél que cumple que:

$$\ell(\tilde{\mu}) = \sup_{\forall \mu} \ell(\mu)$$

La condición necesaria para alcanzar el máximo es que el gradiente se iguale a cero. El gradiente se define como:

$$d(\mu) = \frac{\partial \ell(\mu)}{\partial \mu} = \frac{\sum (y_t - \mu)}{\sigma^2}$$

Por lo tanto, la condición necesaria puede escribirse:

$$d(\tilde{\mu}) = 0 \Rightarrow \tilde{\mu} = \bar{y}$$

El valor que toma el logaritmo de la función de verosimilitud sustituyendo los parámetros por los estimadores MV sin restricciones es:

$$\ell(\tilde{\mu}) = -\frac{T}{2}\log 2\pi - \frac{T}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\Sigma (y_t - \tilde{\mu})^2$$

<u>Estimación MV con restricciones</u> ($\tilde{\mu}_R$)

Supongamos que queremos estimar μ cumpliéndose la restricción: $\mu=\mu_0$, siendo μ_0 un valor conocido.

El estimador MV restringido de μ es aquel que cumple que:

$$\ell(\tilde{\mu}_R) = \sup_{\mu=\mu_0} \ell(\mu)$$

Para obtener la condición necesaria se define la función lagrangiana como:

$$L = \ell(\mu) + \lambda(\mu - \mu_0)$$

Las primeras derivadas de esta función respecto a μ y λ vienen dadas por:

$$\frac{\partial L}{\partial \mu} = d(\mu) + \lambda$$

$$\frac{\partial L}{\partial \lambda} = \mu - \mu_0$$

Las condiciones necesarias vienen dadas, por tanto, por:

$$d(\tilde{\mu}_R) + \tilde{\lambda} = 0 \Rightarrow d(\tilde{\mu}_R) = -\tilde{\lambda}$$

$$\tilde{\mu}_R = \mu_0$$

Podemos resumir los resultados obtenidos en el siguiente Cuadro:

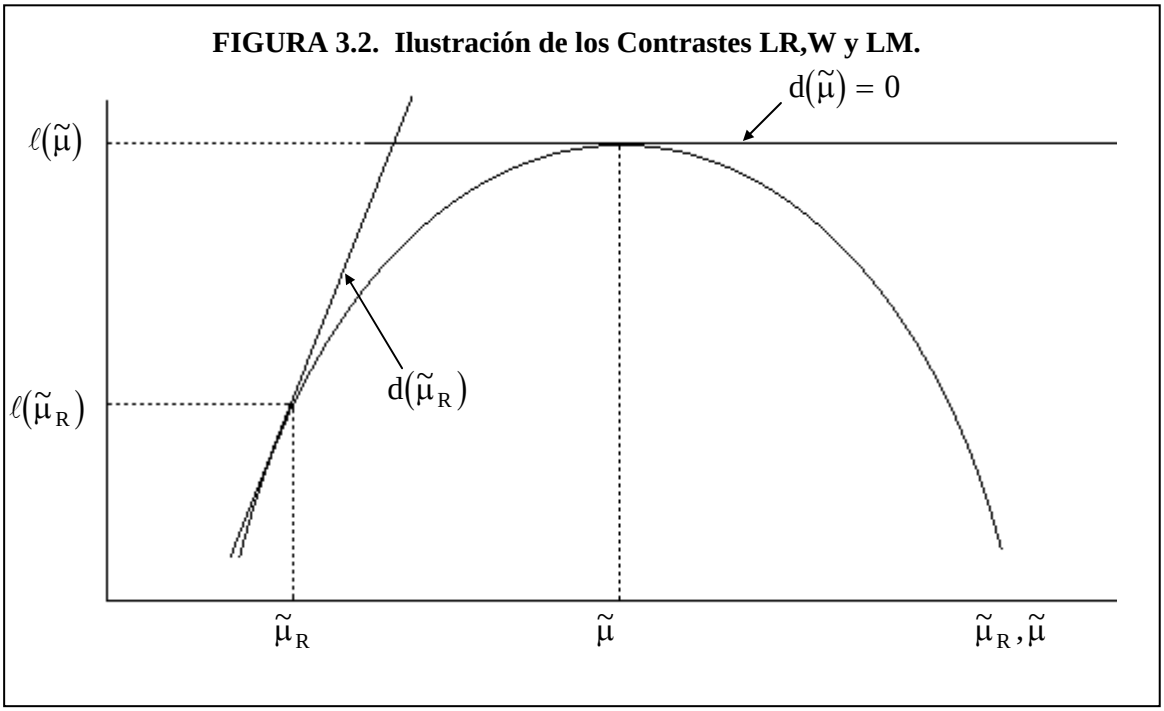
	SIN RESTRICCIONES	CON RESTRICCIONES
Logaritmo de la Func. de Verosimilitud	$\ell(\tilde{\mu})$	$\ell(\tilde{\mu}_R)$
Estimador MV	$\tilde{\mu}$	$\tilde{\mu}_R$
Gradiente	$d(\tilde{\mu}) = 0$	$d(\tilde{\mu}_R)$

Siguiendo a Neyman-Pearson, la Región Crítica se determina prestando atención a la distancia entre los valores tomados por la función de verosimilitud o, equivalentemente, por los tomados por sus logaritmos, es decir, $\ell(\tilde{\mu}) - \ell(\tilde{\mu}_R)$.

Pero a similares resultados puede llegarse si prestamos atención a la distancia entre los dos estimadores o entre los valores tomados por el gradiente.

El siguiente gráfico puede ayudarnos a entender lo que subyace a estos tres planteamientos:

FIGURA 3.2. Ilustración de los Contrastes LR,W y LM.



Puede apreciarse cómo existe una relación monótona entre las tres medidas alternativas de la distancia. Conforme $\ell(\tilde{\mu}_R)$ se aproxima a $\ell(\tilde{\mu}), \tilde{\mu}_R$ se aproxima a $\tilde{\mu}$ y $d(\tilde{\mu}_R)$ se aproxima a $d(\tilde{\mu})$.

El contraste de cualquier hipótesis está basado en el siguiente principio: **"sólo se puede aceptar una restricción cuando su toma en consideración no distorsiona de forma relevante la evidencia contenida en los datos".**

El que una distorsión sea o no relevante puede ponerse en relación con las tres medidas de distancia comentadas.

Podemos decir que la distorsión es relevante cuando:

1) La distancia entre $\ell(\tilde{\mu})$ y $\ell(\tilde{\mu}_R)$ es grande.

2) La distancia entre $\tilde{\mu}$ y $\tilde{\mu}_R$ es grande.

3) La distancia entre $d(\tilde{\mu})$ y $d(\tilde{\mu}_R)$ es grande o, equivalentemente, si la distancia

entre $d(\tilde{\mu}_R)$ y cero es grande.

Ahora bien, la distancia entre estadísticos hay que establecerla prestando atención a la distribución estocástica que siguen dichos estadísticos.

Los dos resultados que se precisan para calibrar esas distancias son:

$$\sqrt{T}\left(\tilde{\mu}-\mu_0\right) \xrightarrow[H_0 \text{ cierta}]{d} \text{N}[0, \lim \text{T I}\left(\mu\right)^{-1}]$$

$$\frac{1}{\sqrt{T}} d\left(\tilde{\mu}_R\right) \xrightarrow[H_0 \text{ cierta}]{d} \text{N}[0, \lim \frac{\text{I}(\mu)}{\text{T}}]$$

(la demostración de estos resultados puede verse en Godfrey (1988))

en donde I(μ) es la matriz de información -en este caso en que μ es un escalar, elemento de información- definida como:

$$\text{I}(\mu) = -\text{E}\,\frac{\partial^2 \ell(\mu)}{\partial \mu^2} = \frac{\text{T}}{\sigma^2}$$

Para llegar a los estadísticos que sirven de base para el contraste necesitamos un resultado adicional que podemos formular como:

<u>Resultado 3.2.1</u>: Sea X un vector de p variables aleatorias distribuidas conjuntamente como: $\text{N}(0, \Sigma_p)$ siendo Σ_p una matriz no singular. Entonces se tiene que $\text{X}' \sum_p^{-1} \text{X} \sim \chi^2(\text{p})$.

Veamos ahora la forma que toman los tres contrastes comentados.

<u>Contraste de Wald</u>: Está basado en la distancia $(\tilde{\mu}-\tilde{\mu}_R)$ o bien en la distancia $(\tilde{\mu}-\mu_0)$. Por los resultados comentados, podemos escribir:

$$\text{W} = \sqrt{\text{T}}(\tilde{\mu}-\mu_0)\,[\hat{\text{Var}}\,\sqrt{\text{T}}(\tilde{\mu}-\mu_0)]^{-1}\,\sqrt{\text{T}}(\tilde{\mu}-\mu_0) =$$

$= \sqrt{T}(\tilde{\mu} - \mu_0)\,[\,I(\tilde{\mu})/T\,]\,\sqrt{T}(\tilde{\mu} - \mu_0) = (\tilde{\mu} - \mu_0)\,I(\tilde{\mu})(\tilde{\mu} - \mu_0)$

Asintóticamente y bajo la hipótesis nula este estadístico se distribuye como una χ^2 con un grado de libertad.

La forma de operar con este estadístico es la siguiente: se fija un tamaño de error tipo 1, $\varepsilon = \varepsilon_0$, se busca en las tablas el correspondiente punto crítico $\chi^2_{\varepsilon_0}(1)$ y la Región Crítica del contraste se determina como:

$$W > \chi^2_{\varepsilon_0}(1)$$

Para el caso general en el que lo que se pretende es contrastar (1) entonces el contraste toma la forma siguiente:

$$W = (R\tilde{\theta} - q)'[RI(\tilde{\theta})^{-1}R']^{-1}(R\tilde{\theta} - q) \qquad (2)$$

y bajo H_0, W sigue una distribución χ^2 con r grados de libertad.

Contraste de Multiplicadores de Lagrange (LM)

Este contraste se basa en la distancia d($\tilde{\mu}_R$) - d($\tilde{\mu}$), y teniendo en cuenta que d($\tilde{\mu}$) = 0, simplemente en lo distante que d($\tilde{\mu}_R$) está de cero. Por los resultados comentados se tiene que:

$$LM = T^{-1/2}d(\tilde{\mu}_R)\left[\hat{Var}\left(T^{-1/2}d(\tilde{\mu}_R)\right)\right]^{-1}T^{-1/2}d(\tilde{\mu}_R) = d(\tilde{\mu}_R)\,I(\tilde{\mu}_R)^{-1}d(\tilde{\mu}_R)$$

Asintóticamente y bajo la hipótesis nula este estadístico se distribuye como una χ^2 con un grado de libertad.

La Región Crítica del contraste es:

$$LM > \chi^2_{\varepsilon_0}(1)$$

Para el caso general asociado con el contraste de (1), el contraste de los Multiplicadores de Lagrange es:

$$LM = d(\tilde{\theta}_R)' I(\tilde{\theta}_R)^{-1} d(\tilde{\theta}_R) \qquad (3)$$

que, bajo H_0, sigue una distribución χ^2 con r grados de libertad.

Contraste de la Razón de Verosimilitud

El contraste está basado en la distancia $(\ell(\tilde{\mu}) - \ell(\tilde{\mu}_R))$. La distribución de este estadístico puede obtenerse fácilmente tras tomar una expansión de Taylor a partir de la de $\sqrt{T}(\tilde{\mu} - \mu_0)$ pudiéndose escribir:

$$LR = 2[\ell(\tilde{\mu}) - \ell(\tilde{\mu}_R)] \sim \chi^2(1) \qquad (4)$$

la Región Crítica del contraste se determina como:

$$LR > \chi^2_{\varepsilon_0}(1)$$

Para el caso general, el contraste de la Razón de Verosimilitud es:

$$LR = 2[\ell(\tilde{\theta}) - \ell(\tilde{\theta}_R)] \qquad (5)$$

que, bajo H_0 sigue una distribución χ^2 con r grados de libertad.

Hemos presentado una justificación intuitiva de los tres contrastes. Pero ahora la pregunta relevante es: ¿son admisibles estos tres contrastes?. La admisibilidad con carácter general se ha demostrado asintóticamente; como puede verse en Godfrey (1988), los tres son consistentes garantizando el mejor uso de la evidencia disponible cuando ésta es grande.