METODO DE MAXIMA VEROSIMILITUD

Supongamos una muestra aleatoria de 10 observaciones de una distribución Poisson:

5, 0, 1, 1, 0, 3, 2, 3, 4, 1.

La densidad de probabilidad para cada observación es:

$$f(x_i, \theta) = \frac{e^{-\theta}\,\theta^{x_i}}{x_i!}, \qquad x_i > 0$$

Con observaciones independientes, la densidad conjunta o verosimilitud de la muestra es:

$$L(\theta) = f(x_1, x_2, \ldots, x_n, \theta) = \prod_{i=1}^{10} f(x_i, \theta) = \frac{e^{-10\theta}\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{10} x_i!} = \frac{e^{-10\theta}\theta^{20}}{207360}$$

Esta última línea da la probabilidad de observar esta muestra en particular, asumiendo que una distribución Poisson con parámetro θ generó los datos. El principio de máxima verosimilitud intenta encontrar aquel θ que maximiza la función de verosimilitud para una muestra dada.

Dado que la función logaritmo natural es monotónicamente creciente, es equivalente maximizar ln L(θ) que L(θ).

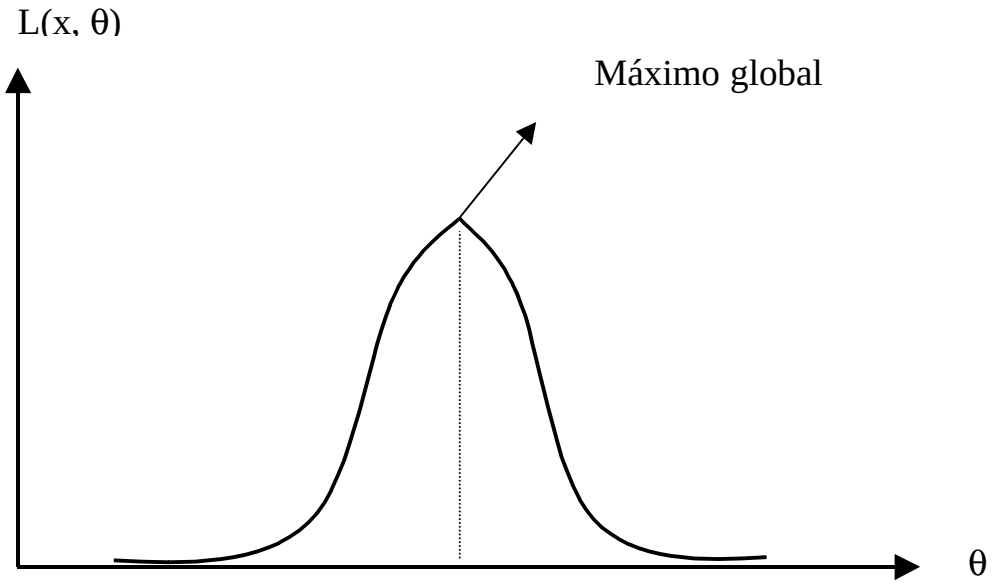
De la ecuación anterior,

$$\ln L(\theta) = -10\,\theta + 20 \ln(\theta) - 12242$$

$$\frac{\partial \ln L(\theta)}{\partial \theta} = -10 + \frac{20}{\theta} = 0 \qquad \Rightarrow \hat{\theta} = 2 \qquad \text{condición de primer orden}$$

$$\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} = -\frac{20}{\theta^2}\Big|_{\theta=\hat{\theta}} < 0$$ condición de segundo orden

Esta última indica que el valor obtenido es un máximo global. Gráficamente la función de verosimilitud luce como:



2

El principio de máxima verosimilitud también se aplica al caso de distribuciones continuas. La densidad conjunta de n observaciones **independientes**, las cuales pueden provenir de una distribución univariada o multivariada, es el producto de las densidades individuales.

Cuando las $\mathbf{x}_i$ son multivariadas, la función de verosimilitud viene dada por:

$$f(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n, \theta) = \prod_{i=1}^{n} f(\mathbf{x}_i, \theta) \equiv L(\theta, \mathbf{X})$$

donde $\boldsymbol{\theta}$ es un vector de parámetros y $\mathbf{X}=(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ es la matriz de datos.

De ello,

$$\ln L(\theta, \mathbf{X}) = \sum_{i=1}^{n} \ln f(\mathbf{x}_i, \theta)$$

Los valores de los parámetros que maximizan esta función son los estimadores de máxima verosimilitud, $\hat{\theta}$.

La condición necesaria para maximizar el logaritmo de la función de verosimilitud es:

$$\frac{\partial \ln L(\theta, \mathbf{X})}{\partial \theta} = \mathbf{0}$$ **Ecuaciones de Verosimilitud**

Ejemplo

Sea $x_i \sim N(\mu, \sigma^2)$. Esto es,

$$f(x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2 \right)$$

Para una muestra de n observaciones independientes:

$$L(\theta, \mathbf{x}) = \prod_{i=1}^{n} f(x_i, \theta) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right)$$

donde $\theta=(\mu\ \sigma^2)'$ y $\mathbf{x}=(x_1\ x_2\ \ldots\ x_n)'$.

Por lo tanto,

$$\ln L(\theta, \mathbf{x}) = \sum_{i=1}^{n} \ln f(x_i, \theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Las ecuaciones de verosimilitud vienen dadas en este caso por:

$$\frac{\partial \ln L(\theta, \mathbf{x})}{\partial \theta} = \begin{pmatrix} \frac{\partial \ln L}{\partial \mu} \\ \frac{\partial \ln L}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Rightarrow \qquad \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} \qquad\qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

A fin de asegurarnos que nos encontramos frente a un máximo global, debemos chequear las condiciones de segundo orden:

El hesiano $\frac{\partial^2 \ln L(\theta, \mathbf{x})}{\partial \theta \partial \theta'}$ viene dado por:

$$H = \begin{pmatrix} \frac{\partial^2 \ln L}{\partial \mu^2} & \frac{\partial^2 \ln L}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 \ln L}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} \end{pmatrix} = \begin{pmatrix} -\frac{n}{\sigma^2} & \\ -\frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^{n} (x_i - \mu)^2 \end{pmatrix}$$

Si evaluamos **H** en los estimadores $\hat{\mu}$ y $\hat{\sigma}^2$, obtenemos que:

$$\mathbf{H} = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & \\ 0 & \frac{n}{2\hat{\sigma}^4} - \frac{n}{\hat{\sigma}^4} \end{pmatrix} = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & \\ 0 & -\frac{n}{\hat{\sigma}^4} \end{pmatrix}$$

el cual es negativo definido. Ello, porque para cualquier vector $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \neq \mathbf{0}$ se tiene que:

$$\mathbf{a}' \mathbf{H} \mathbf{a} = -(a_1^2 \frac{n}{\hat{\sigma}^2} + a_2^2 \frac{n}{2\hat{\sigma}^4}) < 0 \ \blacklozenge$$

Propiedades Asintóticas de los Estimadores de Máxima Verosimilitud

Bajo ciertas condiciones de regularidad (para un tratamiento riguroso, ver Amemiya (1985) o White (1984)), el estimador de máxima verosimilitud tiene las siguientes propiedades asintóticas:

1.- Consistencia: $\text{plim}\ \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$

2.- Normalidad Asintótica: $\hat{\boldsymbol{\theta}} \xrightarrow{d} N(\boldsymbol{\theta}, \mathbf{I}(\boldsymbol{\theta})^{-1})$

donde $\mathbf{I}(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 \ln L}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right) = E\left(\frac{\partial \ln L}{\partial \boldsymbol{\theta}} \frac{\partial \ln L}{\partial \boldsymbol{\theta}'}\right)$, matriz información.

(Véase el apéndice para una demostración de esta igualdad).

3.- Eficiencia Asintótica: el estimador $\hat{\boldsymbol{\theta}}$ es asintóticamente eficiente y alcanza la cota inferior de Cramér-Rao, $\mathbf{I}(\boldsymbol{\theta})^{-1}$, para los estimadores consistentes.

Recordemos que el teorema de Cramér-Rao establece que la varianza de un estimador insesgado de un parámetro (o vector) θ siempre será al menos tan grande como:

$$I(\theta) = -E\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right) = E\left(\frac{\partial \ln L}{\partial \theta}\right)^2$$

donde $I(\theta)$ es el número (matriz) información de la muestra.

Por otra parte, se dice que un estimador es asintóticamente eficiente si es consistente, asintóticamente normal y tiene una matriz varianza-covarianza no más 'grande' que aquella de cualquier estimador consistente y asintóticamente normal.

4.- Invarianza: el estimador máximo verosímil de $\gamma = c(\boldsymbol{\theta})$ es $c(\hat{\boldsymbol{\theta}})$.

Aplicaciones

1.- Estimador de Máxima Verosimilitud en el Modelo Lineal Clásico

Sea $\mathbf{y}_{nx1} = \mathbf{X}_{nxk}\,\boldsymbol{\beta}_{kx1} + \boldsymbol{\varepsilon}_{nx1}$ $\qquad\qquad$ $\boldsymbol{\varepsilon}\sim N(\mathbf{0}, \sigma^2\,\mathbf{I})$

Para una muestra de ε_i, i=1, 2, ...,n, variables i.i.d. normales con media 0 y varianza σ^2, la función de verosimilitud viene dada por:

$$L = (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2\right)$$

con $\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$, $\mathbf{x}_i' = (1\ X_{i2}\ X_{i3}...X_{ik})$.

El jacobiano de cada observación es $\left|\frac{\partial\varepsilon_i}{\partial y_i}\right| = 1$. Por lo tanto, la función de verosimilitud para las y's viene dada por:

$$L = (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\beta)^2\right)$$

$$= (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)\right)$$

Al tomar el logaritmo de L, obtenemos:

$$\ln L = \frac{n}{2}(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

Las ecuaciones de verosimilitud son:

$$\frac{\partial \ln L}{\partial\beta} = \frac{1}{\sigma^2}\mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) = 0$$

De lo cual,

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \qquad \hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n} = \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta})}{n}$$

El hesiano viene dado por:

$$\mathbf{H} = \begin{pmatrix} \frac{\partial^2 \ln L}{\partial\beta\partial\beta'} & \frac{\partial^2 \ln L}{\partial\beta\partial\sigma^2} \\ \frac{\partial^2 \ln L}{\partial\sigma^2\partial\beta'} & \frac{\partial^2 \ln L}{\partial(\sigma^2)^2} \end{pmatrix} = \begin{pmatrix} \frac{\mathbf{X}'\mathbf{X}}{\sigma^2} & -\frac{\mathbf{X}'\varepsilon}{\partial\sigma^4} \\ -\frac{\varepsilon'\mathbf{X}}{\partial\sigma^4} & \frac{n}{2\sigma^4} - \frac{\varepsilon'\varepsilon}{\sigma^6} \end{pmatrix}_{(k+1)x(k+1)}$$

Si tomamos la esperanza de **H**, los términos fuera de la diagonal se hacen cero: $E(\mathbf{X}'\varepsilon) = E(\sum_{i=1}^{n} \varepsilon_i \mathbf{x}_i) = \sum_{i=1}^{n} E(\varepsilon_i)\mathbf{x}_i = \mathbf{0}$, donde $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{pmatrix}$ y $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \cdots \\ \mathbf{x}_n' \end{pmatrix}$.

(La última línea de arriba supone que las X's son no estocásticas. Si ello no es así, hacemos los cálculos de la esperanza condicional en los valores de las X's. Ni el análisis ni los resultados obtenidos cambian).

Por lo tanto,

$$\left(-\mathbf{I}(\beta, \sigma^2)\right)^{-1} = \begin{pmatrix} \sigma^2(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2\sigma^4}{n} \end{pmatrix}_{(k+1)x(k+1)}$$

De ello, es claro que el estimador de mínimos cuadrados ordinarios (MICO) de $\boldsymbol{\beta}$ coincide con el de máxima verosimilitud. Por lo tanto, $\hat{\boldsymbol{\beta}}$ MICO tiene todas las propiedades asintóticas del estimador de máxima verosimilitud.

Recordemos que el estimador MICO de σ^2 es $\hat{\sigma}^2 = \dfrac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$, el cual es insesgado. De ello, resulta evidente que el estimador de máxima verosimilitud de σ^2 es <u>sesgado</u> (hacia cero):

$$E(\hat{\sigma}^2_{MV}) = \frac{(n-k)\,\sigma^2}{n} = \left(1 - \frac{k}{n}\right)\sigma^2 < \sigma^2$$

Sin embargo, el estimador $\hat{\sigma}^2_{MV}$ tiene todas las propiedades asintóticas de los estimadores de máxima verosimilitud. En particular,

$$\sqrt{n}\,(\hat{\sigma}^2_{MV} - \sigma^2) \xrightarrow{d} N(0, 2\sigma^4)$$

<u>Nota: Regresores Estocásticos</u>

Recordemos que $f(\mathbf{y}, \mathbf{X}|\, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$, función de densidad conjunta de la muestra, condicional en los valores de los parámetros poblaciones, puede descomponerse como el producto de las distribuciones de y condicional en X y de X:

$$L = f(\mathbf{y}, \mathbf{X}|\, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = f(\mathbf{y}|\, \mathbf{X}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})\; g(\mathbf{X}|\, \boldsymbol{\theta})$$

Por lo tanto, $\ln L = \ln f(.) + \ln g(.)$

Si $g(\mathbf{X}|\, \boldsymbol{\theta})$ no depende de $\boldsymbol{\beta}$ y σ^2, podemos maximizar las dos partes de el logaritmo de la función de verosimilitud por separado, a fin de obtener el conjunto de parámetros desconocidos $\boldsymbol{\beta}$, σ^2y $\boldsymbol{\theta}$.

2.- Estimador de Máxima Verosimilitud para el Modelo Lineal General

Sea $\mathbf{y}_{nx1} = \mathbf{X}_{nxk}\,\boldsymbol{\beta}_{kx1} + \boldsymbol{\varepsilon}_{nx1}$ $\qquad\qquad$ $\boldsymbol{\varepsilon}\sim N(\mathbf{0}, \sigma^2\,\boldsymbol{\Omega})$

Para una muestra de ε_i, n=1, 2, ..., n, variables aleatorias provenientes de la distribución normal multivariada $\boldsymbol{\varepsilon}\sim N(\mathbf{0}, \sigma^2\,\boldsymbol{\Omega})$, tenemos:

$$L = (2\pi\sigma^2)^{-n/2}\left|\sigma^2\boldsymbol{\Omega}\right|^{-1/2}\exp\left(-\frac{1}{2}\varepsilon'(\sigma^2\boldsymbol{\Omega})^{-1}\varepsilon\right)$$

con $\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}$, $\mathbf{x}_i' = (1\ X_{i2}\ X_{i3}...X_{ik})$.

El jacobiano de cada observación es $\left|\frac{\partial\varepsilon_i}{\partial y_i}\right| = 1$. Por lo tanto, el logaritmo de la función de verosimilitud para las y's viene dado por:

$$\ln L = \frac{n}{2}(2\pi) - \frac{1}{2}\ln\left|\sigma^2\boldsymbol{\Omega}\right| - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\beta)'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\beta)$$

Para una matriz $\boldsymbol{\Omega}$ de constantes conocidas, las ecuaciones de verosimilitud son:

$$\frac{\partial \ln L}{\partial\beta} = \frac{1}{\sigma^2}\mathbf{X}'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\beta) = \mathbf{0}$$

$$\frac{\partial \ln L}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y}-\mathbf{X}\beta)'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\beta) = 0$$

De lo cual,

$$\hat{\beta} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y} \qquad \hat{\sigma}^2 = \frac{\hat{\varepsilon}'\boldsymbol{\Omega}^{-1}\hat{\varepsilon}}{n} = \frac{(\mathbf{y}-\mathbf{X}\hat{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\hat{\beta})}{n}$$

Se puede demostrar de manera análoga que:

$$\begin{pmatrix} \hat{\beta}_{MV} \\ \hat{\sigma}^2_{MV} \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix}, \begin{pmatrix} \sigma^2 (X'\Omega^{-1}X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix} \right)$$

Test de Hipótesis bajo Estimación Vía Máxima Verosimilitud

Veremos tres procedimientos que son asintóticamente equivalentes.

1.- Razón de Verosimilitud

Sea θ un vector de parámetros a ser estimado y sea H_0 alguna hipótesis que queremos contrastar. Asimismo, sea $\hat{\theta}_{NR}$ el estimador de máxima verosimilitud de θ no restringido, y sea $\hat{\theta}_R$ el estimador máximo verosímil restringido. Si $\hat{L}_{NR}$ y $\hat{L}_R$ son las funciones de verosimilitud evaluadas en $\hat{\theta}_{NR}$ y $\hat{\theta}_R$, respectivamente, entonces la razón de verosimilitud es:

$$\lambda = \frac{\hat{L}_R}{\hat{L}_{NR}}$$

donde $0<\lambda<1$, dado que $\hat{L}_{NR}$, $\hat{L}_R > 0$ y $\hat{L}_{NR} > \hat{L}_R$.

Si λ es pequeño, pensaríamos que las restricciones son falsas. Se puede demostrar que:

$$-2\ln(\lambda) \xrightarrow{d} \chi^2(J)$$

donde J es el número de restricciones.

Ejemplo

Recordemos el ejemplo de la distribución Poisson al comienzo de estos apuntes. Bajo H_0: $\theta=1.8$, $\hat{L}_R = 0.0936 \times 10^{-8}$. Por otra parte, $\hat{L}_{NR} = 0.104 \times 10^{-8}$, dado que $\hat{\theta}=2$ (estimador no restringido). Entonces

$$-2\ln(\lambda) = -2\ln\left(\frac{0.0936}{0.104}\right) = 0.21072$$

El valor crítico de una $\chi^2(1)$ al 95 por ciento de confianza es de 3.84. Por lo tanto, no rechazamos H_0 ♦

2.- Test de Wald

Una desventaja práctica del test de razón de verosimilitud es que requiere que estimemos tanto el vector de parámetros restringido como el no restringido. Ello puede ser, en algunos casos, muy intensivo computacionalmente. El test de Wald y el de multiplicador de Lagrange—éste último tratado en el próximo punto—solucionan tal desventaja.

Sea $\hat{\theta}$ el vector de estimadores de máxima verosimilitud no restringidos. Postulamos un conjunto de restricciones del tipo:

H_0: $\mathbf{c}(\theta)=\mathbf{q}$

donde **c** es conjunto de funciones de **θ** no necesariamente lineales.

Si las restricciones son válidas, entonces $\hat{\theta}$ debería satisfacerlas en forma aproximada, dada la variabilidad muestral. El test de Wald toma la forma:

$$W = (\mathbf{c}(\hat{\theta}) - \mathbf{q})'(\mathrm{Var}(\mathbf{c}(\hat{\theta}) - \mathbf{q}))^{-1}(\mathbf{c}(\hat{\theta}) - \mathbf{q}) \xrightarrow{d} \chi^2(J)$$

donde $\mathrm{Var}(\mathbf{c}(\hat{\theta}) - \mathbf{q}) = \mathrm{Var}(\mathbf{c}(\hat{\theta})) = \mathbf{C}\,\mathrm{Var}(\hat{\theta})\,\mathbf{C}'$, $\mathbf{C}_{Jxk} = \left(\frac{\partial \mathbf{c}(\theta)}{\partial \theta'}\right)$, J es el número de restricciones y k es la dimensión del vector θ. (La j-ava fila de C contiene las derivadas de la restricción j-ava con respecto a los k elementos de **θ).**

Para contrastar un conjunto de restricciones lineales, **Rθ=q** el test de Wald se basa en:

H_0: $\mathbf{c}(\theta)-\mathbf{q}=\mathbf{R}\theta-\mathbf{q}=\mathbf{0}$

$$\mathbf{C}_{Jxk} = \left(\frac{\partial \mathbf{c}(\theta)}{\partial \theta'} \right) = \frac{\partial (\mathbf{R}\theta - \mathbf{q})}{\partial \theta'} = \mathbf{R}$$

Entonces, $\text{Var}(\mathbf{c}(\hat{\theta}) - \mathbf{q}) = \mathbf{R}\,\text{Var}(\hat{\theta})\,\mathbf{R}'$. Con ello el test de Wald toma la forma:

$$W = (\mathbf{R}\hat{\theta} - \mathbf{q})'(\mathbf{R}\text{Var}(\hat{\theta})\mathbf{R}')^{-1}(\mathbf{R}\hat{\theta} - \mathbf{q})$$

el cual se distribuye chi-cuadrado con J grados de libertad.

Ejemplo

Si queremos contrastar una restricción simple como:

H_0: $\theta=\theta_0$ versus la alternativa H_1: $\theta\neq\theta_0$

lo podemos hacer con un test de Wald, el cual en este caso toma la forma:

$$W = (\hat{\theta} - \theta_0 - 0)'(\text{Var}(\hat{\theta} - \theta_0 - 0))^{-1}(\hat{\theta} - \theta_0 - 0)$$

$$= \frac{(\hat{\theta} - \theta_0)^2}{\text{Var}(\hat{\theta})} \xrightarrow{d} \chi^2(1)$$

Pero, $z = \dfrac{\hat{\theta} - \theta_0}{\sqrt{\text{Var}(\hat{\theta})}} \xrightarrow{d} N(0,1)$, entonces $z^2 \xrightarrow{d} \chi^2(1)$.

Esto es, cuando se trata de contrastar la hipótesis H_0: $\theta=\theta_0$ versus la alternativa H_1: $\theta\neq\theta_0$, el test de Wald y el test t son asintóticamente equivalentes♦

3.- Test de Multiplicador de Lagrange

Sea λ el vector de multiplicadores de Lagrange. Deseamos maximizar la función de verosimilitud, L(θ), sujeta al conjunto de restricciones $\mathbf{c}(\theta)=\mathbf{q}$:

$$\ln L^*(\theta) = \ln L(\theta) + \lambda'(\mathbf{c}(\theta) - \mathbf{q})$$

Las condiciones de primer orden, necesarias para la maximización de la función de verosimilitud, son las siguientes:

$$\frac{\partial \ln L^{*}(\theta)}{\partial \theta} = \frac{\partial \ln L(\theta)}{\partial \theta} + \frac{\partial(\lambda'(\mathbf{c}(\theta) - \mathbf{q})}{\partial \theta} = \frac{\partial \ln L(\theta)}{\partial \theta} + \frac{\partial \mathbf{c}(\theta)}{\partial \theta}\lambda = \mathbf{0} \qquad (1)$$

$$\frac{\partial \ln L^{*}(\theta)}{\partial \lambda} = \mathbf{c}(\theta) - \mathbf{q} = \mathbf{0} \qquad (2)$$

Si las restricciones son válidas, el valor maximizado de la función de verosimilitud no debiera cambiar sustancialmente. De (1) se tiene que:

$$\frac{\partial \ln L^{*}(\hat{\theta}_R)}{\partial \theta} - \frac{\partial \mathbf{c}(\hat{\theta}_R)}{\partial \theta}\hat{\lambda} = \mathbf{0}$$

donde $\hat{\theta}_R$ es el estimador restringido.

Si las restricciones fueran válidas, entonces $\hat{\lambda} = \mathbf{0}$. (Esto es, las restricciones no serían operativas). En dicho caso, $\frac{\partial \ln L^{*}(\hat{\theta}_R)}{\partial \theta} \approx \mathbf{0}$.

El test de multiplicador de Lagrange se basa en la observación anterior:

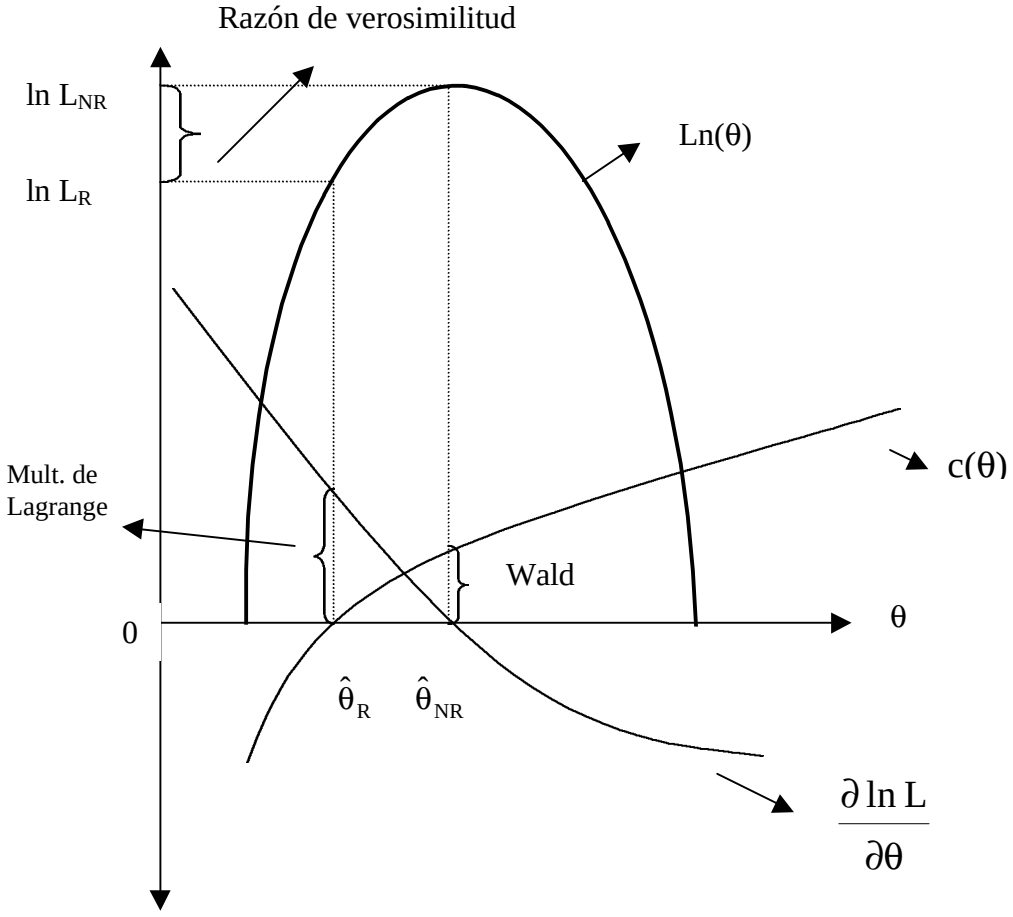
$$LM = \left(\frac{\partial \ln L(\hat{\theta}_R)}{\partial \theta}\right)' \left(I(\hat{\theta}_R)\right)^{-1} \left(\frac{\partial \ln L(\hat{\theta}_R)}{\partial \theta}\right) \xrightarrow{d} \chi^2(J)$$

donde $I(\hat{\theta}_R) = -E\left(\frac{\partial^2 \ln L(\hat{\theta}_R)}{\partial \theta \partial \theta'}\right)$ y J es el número de restricciones.

Es importante notar que todos términos en LM se evalúan en el estimador restringido.

Una característica importante de los test de razón de verosimilitud, Wald y de multiplicador de Lagrange es que son asintóticamente equivalentes. Sin embargo, en muestras pequeñas nuestras conclusiones pueden diferir dependiendo de qué estadígrafo utilicemos.

En términos gráficos los tres estadígrafos pueden ser descritos como sigue, cuando H_0: $c(\theta)=0$:



Analicemos cada test por separado:

- <u>Razón de verosimilitud</u>: Si la restricción $c(\theta)=0$ es válida, el imponerla no debiera reducir notoriamente el valor del logaritmo de la función de verosimilitud. El test se basa en $\ln L(\hat{\theta}_{NR}) - \ln L(\hat{\theta}_R)$, donde $\hat{\theta}_{NR}$ y $\hat{\theta}_R$ son los estimadores de máxima verosimilitud no restringido y restringido, respectivamente. Si esta diferencia es pequeña, en términos estadísticos, entonces no rechazamos H_0.

- <u>Wald</u>: Si la restricción es válida, $c(\hat{\theta}_{NR})$ debería ser cercana a cero. Rechazamos H_0 si $c(\hat{\theta}_{NR})$ es significativamente distinta de cero.

- Multiplicador de Lagrange: Si la restricción es válida, la pendiente (primera derivada) de ln L(θ) debiera ser cercana a cero cuando es evaluada en el estimador restringido.

Ejemplo

Supongamos que el ingreso de una persona proviene de una distribución de probabilidades gama:

$$f(y_i|x_i,\beta,\rho)=\frac{(\beta+x_i)^{-\rho}}{\Gamma(\rho)}y_i^{\rho-1}\exp\left(-\frac{y_i}{\beta+x_i}\right)$$

donde y_i y x_i son los niveles de ingreso y de educación de la persona i , y $\Gamma(\rho)=\int_0^\infty t^{p-1}e^{-t}dt$ es la función gama.

Se desea contrastar la hipótesis H_0: ρ=1, esto es, que el nivel de ingreso de cada persona proviene de una distribución exponencial de parámetro $\lambda_i=\frac{1}{\beta+x_i}$.

El logaritmo de la función de verosimilitud para el modelo no restringido (distribución gama) está dado por:

$$\ln L(\mathbf{y}|\,\mathbf{x},\beta,\rho)=-\rho\sum_{i=1}^{n}\ln(\beta+x_i)-n\ln\Gamma(\rho)-\sum_{i=1}^{n}\frac{y_i}{\beta+x_i}+(\rho-1)\sum_{i=1}^{n}\ln(y_i)$$

Las primeras y segundas derivadas están dadas por:

$$\frac{\partial \ln L}{\partial\beta}=-\rho\sum_{i=1}^{n}\frac{1}{\beta+x_i}+\sum_{i=1}^{n}\frac{y_i}{(\beta+x_i)^2}$$

$$\frac{\partial \ln L}{\partial\rho}=-\sum_{i=1}^{n}\ln(\beta+x_i)-\frac{n\Gamma'(\rho)}{\Gamma(\rho)}+\sum_{i=1}^{n}\ln(y_i)$$

$$\frac{\partial^2 \ln L}{\partial \beta^2} = \rho \sum_{i=1}^{n} \frac{1}{(\beta + x_i)^2} - 2 \sum_{i=1}^{n} \frac{y_i}{(\beta + x_i)^3}$$

$$\frac{\partial^2 \ln L}{\partial \rho^2} = - \frac{n(\Gamma(\rho)\Gamma''(\rho) - \Gamma'(\rho)^2)}{\Gamma(\rho)^2}$$

$$\frac{\partial^2 \ln L}{\partial \beta \partial \rho} = - \sum_{i=1}^{n} \frac{1}{\beta + x_i}$$

donde $\frac{\partial^r \Gamma(\rho)}{\partial \rho^r} = \int_0^\infty \ln(\rho)^r x^{\rho-1} e^{-x} dx$.

Los valores de los estimadores de máxima verosimilitud son obtenidos al igualar las primeras derivadas a cero, sin y con la restricción de que ρ=1. Los resultados obtenidos con los datos del Cuadro 1 son los que se detallan en el Cuadro 2.

Cuadro 1 Ingreso y Años de Educación

Ingreso Anual (Miles US$)	Educación (Años)	Ingreso Anual (Miles US$)	Educación (Años)
20.5	12	55.8	16
31.5	16	25.2	20
47.7	18	29	12
26.2	16	85.5	16
44	12	15.1	10
8.28	12	15.1	10
30.8	16	28.5	18
17.2	12	21.4	16
19.9	10	64.2	12
9.96	12	84.9	16

Fuente: Greene (1997)

A fin de ilustrar el uso de los distintos estadígrafos vistos en estos apuntes, contrastaremos H_0: ρ=1 por cuatro vías distintas:

- Intervalo de confianza para ρ: Un intervalo de confianza (asintótico) al 95 por ciento se basa en el estimador **no restringido**:

$$3.151 \pm 1.96\sqrt{0.6625} = [1.556; 4.746]$$

Dado que este intervalo no contiene el valor de 1, rechazamos la hipótesis nula.

- Test de razón de verosimilitud:

$$-2\lambda = -(-88.436 - (-82.916) = 11.04$$

El valor crítico es $\chi^2(1)_{95\%} = 3.842$, por lo cual rechazamos H_0.

- Test de Wald: En este caso $c(\theta) - q = \rho - 1$. Por ello, $\frac{\partial c(\theta)}{\partial \theta} = 1$ y $\hat{Var}(c(\hat{\theta}) - q) = \hat{Var}(\hat{\rho}) = 0.6625$. Con ello:

$$Wald = (3.151 - 1)(0.6625)^{-1}(3.151 - 1) = 6.984.$$

Como antes, el valor crítico es igual a 3.842. Por lo tanto, rechazamos H_0.

Cuadro 2 Estimadores de Máxima Verosimilitud para el Modelo de Ingreso

	Estimador No Restringido	Estimador Restringido
$\hat{\beta}$	-4.719	15.503
$\hat{\rho}$	3.151	1.00
$\ln L$	-82.916	-88.436
$\frac{\partial \ln L}{\partial \beta}$	0	0
$\frac{\partial \ln L}{\partial \rho}$	0	7.914
$\frac{\partial^2 \ln L}{\partial \beta^2}$	-0.853	-0.0216
$\frac{\partial^2 \ln L}{\partial \rho^2}$	-7.436	-32.894
$\frac{\partial^2 \ln L}{\partial \beta \partial \rho}$	-2.242	-0.669
$\text{Var}(\hat{\beta})$	5.773	46.164
$\text{Var}(\hat{\rho})$	0.663	0
$\text{Cov}(\hat{\beta}, \hat{\rho})$	-1.746	0

- Test de Multiplicador de Lagrange: Las primeras derivadas del logaritmo de la función de verosimilitud evaluadas en el estimador restringido toman el valor de:

$$\frac{\partial \ln L(\hat{\theta}_R)}{\partial \theta} = \begin{pmatrix} \frac{\partial \ln L}{\partial \beta} \\ \frac{\partial \ln L}{\partial \rho} \end{pmatrix} = \begin{pmatrix} 0 \\ 7.914 \end{pmatrix}$$

donde $\theta=(\beta\ \rho)'$.

$$I(\hat{\hat{\theta}}_R) = -\begin{pmatrix} \frac{\partial^2 \ln L(\hat{\beta}_R, \hat{\rho}_R)}{\partial \beta^2} & \frac{\partial^2 \ln L(\hat{\beta}_R, \hat{\rho}_R)}{\partial \beta \partial \rho} \\ \frac{\partial^2 \ln L(\hat{\beta}_R, \hat{\rho}_R)}{\partial \rho \partial \beta} & \frac{\partial^2 \ln L(\hat{\beta}_R, \hat{\rho}_R)}{\partial \rho^2} \end{pmatrix} = \begin{pmatrix} 0.0217 & 0.669 \\ 0.669 & 32.894 \end{pmatrix}$$

Entonces LM toma la forma:

$$LM = \begin{pmatrix} 0 & 7.914 \end{pmatrix} \begin{pmatrix} 0.0217 & 0.669 \\ 0.669 & 32.894 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 7.914 \end{pmatrix} = 5.12$$

Como antes, rechazamos H_0.

Apéndice

Proposición

$$-E\left(\frac{\partial^2 \ln L}{\partial \theta \partial \theta'}\right) = E\left(\frac{\partial \ln L}{\partial \theta}\frac{\partial \ln L}{\partial \theta'}\right)$$

Demostración

Consideremos primero la función densidad de probabilidades de x_i, $f(x_i)$. Supongamos que el rango de x_i no depende del valor de $\boldsymbol{\theta}$. Si $x_i \in$ al intervalo $(-\infty, \infty)$, entonces:

$$\int_{-\infty}^{\infty} f(x_i|\theta)dx_i = 1$$

De lo anterior,

$$\frac{\partial}{\partial \theta}\left(\int_{-\infty}^{\infty} f(x_i|\theta)dx_i\right) = \int_{-\infty}^{\infty} \frac{\partial f(x_i|\theta)}{\partial \theta}dx_i = \mathbf{0}$$

dado que el rango de x_i no depende de θ.

Pero

$$\frac{\partial}{\partial\theta}\left(\int_{-\infty}^{\infty} f(x_i|\theta)dx_i\right)=\int_{-\infty}^{\infty}\frac{\partial f(x_i|\theta)}{\partial\theta}dx_i=\int_{-\infty}^{\infty}\frac{\partial \ln f(x_i|\theta)}{\partial\theta}f(x_i|\theta)dx_i$$

$$=E\left(\frac{\partial \ln f(x_i|\theta)}{\partial\theta}\right)=\mathbf{0}$$

Diferenciemos $\frac{\partial}{\partial\theta}\left(\int_{-\infty}^{\infty} f(x_i|\theta)dx_i\right)$ con respecto a θ nuevamente:

$$\frac{\partial^2}{\partial\theta\partial\theta'}\left(\int_{-\infty}^{\infty} f(x_i|\theta)dx_i\right)=\frac{\partial}{\partial\theta'}\left(\int_{-\infty}^{\infty}\frac{\partial \ln f(x_i|\theta)}{\partial\theta}f(x_i|\theta)dx_i\right)$$

$$\int_{-\infty}^{\infty}\frac{\partial^2 \ln f(x_i|\theta)}{\partial\theta\partial\theta'}f(x_i|\theta)dx_i+\int_{-\infty}^{\infty}\frac{\partial \ln f(x_i|\theta)}{\partial\theta}\frac{\partial f(x_i|\theta)}{\partial\theta'}dx_i=\mathbf{0}$$

Pero $\frac{\partial f(x_i|\theta)}{\partial\theta'}=f(x_i|\theta)\frac{\partial \ln f(x_i|\theta)}{\partial\theta'}$. Entonces la relación anterior se reduce a:

$$\int_{-\infty}^{\infty}\frac{\partial^2 \ln f(x_i|\theta)}{\partial\theta\partial\theta'}f(x_i|\theta)dx_i+\int_{-\infty}^{\infty}\frac{\partial \ln f(x_i|\theta)}{\partial\theta}\frac{\partial \ln f(x_i|\theta)}{\partial\theta'}f(x_i|\theta)dx_i=\mathbf{0}$$

lo cual implica que:

$$-E\left(\frac{\partial^2 \ln f(x_i|\theta)}{\partial\theta\partial\theta'}\right)=E\left(\frac{\partial \ln f(x_i|\theta)}{\partial\theta}\frac{\partial \ln f(x_i|\theta)}{\partial\theta'}\right)$$

Consideremos ahora n observaciones independientes. El logaritmo de la función de verosimilitud de la muestra viene dado por:

$$\ln L = \sum_{i=1}^{n} \ln f(x_i|\theta)$$

con primeras y segundas derivadas:

$$\frac{\partial \ln L}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f(x_i|\theta)}{\partial \theta} = \sum_{i=1}^{n} \mathbf{g}_i \text{ , con } \mathbf{g}_i \equiv \frac{\partial \ln f(x_i|\theta)}{\partial \theta}$$

$$\frac{\partial^2 \ln L}{\partial \theta \partial \theta'} = \sum_{i=1}^{n} \frac{\partial^2 \ln f(x_i|\theta)}{\partial \theta \partial \theta'} = \sum_{i=1}^{n} \mathbf{H}_i \text{ , con } \mathbf{H}_i \equiv \frac{\partial^2 \ln f(x_i|\theta)}{\partial \theta \partial \theta'}$$

Por la demostración anterior sabemos que $E\left(\frac{\partial \ln L}{\partial \theta}\right) = \sum_{i=1}^{n} E(\mathbf{g}_i) = \mathbf{0}$. Sea $\mathbf{g}=(\mathbf{g}_1 \ \mathbf{g}_2 \ ... \ \mathbf{g}_n)'$. Con observaciones independientes se tiene que $E(\mathbf{g}\mathbf{g}') = \sum_{i=1}^{n} E(\mathbf{g}_i \mathbf{g}_i')$, dado que $E(\mathbf{g}_i \mathbf{g}_j')=\mathbf{0} \ \forall \ i \neq j$.

Ahora, $\sum_{i=1}^{n} E(\mathbf{g}_i \mathbf{g}_i') = -\sum_{i=1}^{n} E(\mathbf{H}_i) = -E(\mathbf{H})$, con $\mathbf{H} \equiv E\left(\sum_{i=1}^{n} \mathbf{H}_i\right)$, en virtud de la demostración anterior. De ello deducimos que:

$$-E\left(\frac{\partial^2 \ln L}{\partial \theta \partial \theta'}\right) = E\left(\frac{\partial \ln L}{\partial \theta} \frac{\partial \ln L}{\partial \theta'}\right) \blacklozenge$$

El resultado anterior es el utilizado en el algoritmo de BHHH (Berndt-Hausman-Hall-Hall).

Nota: Dado que $E(\mathbf{g})=\mathbf{0}$, $Var(\mathbf{g})=E(\mathbf{g}\mathbf{g}')= -E(\mathbf{H})$. Vimos que el test de multiplicador de Lagrange se basa en el estadígrafo

$$LM = \mathbf{g}(\hat{\theta}_R)'(\mathbf{H}(\hat{\theta}_R))^{-1}\mathbf{g}(\hat{\theta}_R) \xrightarrow{d} \chi^2(J)$$

Ello, porque bajo la hipótesis nula las restricciones son verdaderas, lo cual implica que $\sqrt{n}\mathbf{g}(\hat{\theta}_R) \xrightarrow{d} N(\mathbf{0}, -E(\mathbf{H}(\theta)))$ y $\frac{1}{n}\mathbf{H}(\hat{\theta}_R) \xrightarrow{p} E(\mathbf{H}(\theta))$.