MINIMOS CUADRADOS NO LINEALES

1 Conceptos Generales

Se entiende por un modelo de regresión **no linea**l aquel para el cual sus primeras derivadas con respecto a los parámetros son funciones no lineales de éstos.

Consideremos el modelo general:

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i \tag{1}$$

El estimador de mínimos cuadrados no lineales es aquel que minimiza la suma de cuadrados residuales:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - h(\mathbf{x}_i, \boldsymbol{\beta}))^2 \tag{2}$$

Como veremos, si εi se distribuye normal, entonces el estimador de mínimos cuadrados no lineales coincide con el estimador de máxima verosimilitud.

Las condiciones de primer y segundo orden, necesarias y suficientes para la minimización (óptimo local), respectivamente, vienen dadas por:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\sum_{i=1}^{n} (y_i - h(\mathbf{x}_i, \boldsymbol{\beta})) \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0} \tag{3}$$

$$\frac{\partial^2 S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = 2\left(\sum_{i=1}^{n} \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} - \sum_{i=1}^{n} (y_i - h(\mathbf{x}_i, \boldsymbol{\beta})) \frac{\partial^2 h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right) \tag{4}$$

donde la matriz en la ecuación (4) debe ser positiva definida.

Ejemplo

Para el modelo $y_i = \beta_1 + \beta_2 \exp(\beta_3 x_i) + \varepsilon_i$

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 e^{\beta_3 x_i}) = 0 \tag{5}$$

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_2} = -2\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 e^{\beta_3 x_i}) e^{\beta_3 x_i} = 0 \tag{6}$$

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_3} = -2\sum_{i=1}^{n}(y_i - \beta_1 - \beta_2 e^{\beta_3 x_i}) \beta_2 x_i e^{\beta_3 x_i} = 0 \tag{7}$$

donde $\varepsilon_i = y_i - \beta_1 - \beta_2 - \exp(\beta_3 x_i)$

$$\frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \beta_1} \\ \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \beta_2} \\ \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \beta_3} \end{pmatrix} = \begin{pmatrix} 1 \\ e^{\beta_3 x_i} \\ \beta_2 x_i e^{\beta_3 x_i} \end{pmatrix} \qquad i=1, 2, ..., n$$

Las ecuaciones (5)-(7) no tienen una solución analítica. Por lo tanto, se requiere de algún método iterativo para encontrar $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ ♦

La distribución asintótica de $\hat{\boldsymbol{\beta}}_{NLS}$, estimador de mínimos cuadrados no lineales, viene dada por:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{NLS} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, \sigma^2 Q^{-1}) \tag{8}$$

donde $\sigma^2 = E(\varepsilon_i^2)$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - h(\mathbf{x}_i, \hat{\boldsymbol{\beta}}))^2 \xrightarrow{p} \sigma^2$

$$\mathbf{Q} = \text{plim}\left(\frac{\tilde{\mathbf{X}}'\tilde{\mathbf{X}}}{n}\right) = \text{plim}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial h(\mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}\frac{\partial h(\mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}'}\right)$$

(ver apéndice para mayores detalles).

Para dar una intuición al resultado anterior, pensemos en la matriz $\tilde{\mathbf{X}}$ como en la matriz de regresores del modelo. Recordemos que para mínimos cuadrados ordinarios, la condición de primer orden (la llamada "**condición de ortogonalidad"**) viene dada por:

$$\frac{\partial S(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{MICO}}) = -2\mathbf{X}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$$

En el caso de mínimos cuadrados no lineales la condición anterior viene dada por:

$$\frac{\partial S(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = -2\tilde{\mathbf{X}}'(\mathbf{y} - \mathbf{h}(\mathbf{x}, \hat{\boldsymbol{\beta}}_{\text{NLS}})) = -2\tilde{\mathbf{X}}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$$

donde $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$, $\mathbf{h}(\mathbf{x}, \hat{\boldsymbol{\beta}}_{\text{NLS}}) = \begin{pmatrix} h(\mathbf{x}_1, \hat{\boldsymbol{\beta}}_{\text{NLS}}) \\ h(\mathbf{x}_2, \hat{\boldsymbol{\beta}}_{\text{NLS}}) \\ \dots \\ h(\mathbf{x}_n, \hat{\boldsymbol{\beta}}_{\text{NLS}}) \end{pmatrix}$, $\hat{\boldsymbol{\varepsilon}} = \begin{pmatrix} y_1 - h(\mathbf{x}_1, \hat{\boldsymbol{\beta}}_{\text{NLS}}) \\ y_2 - h(\mathbf{x}_2, \hat{\boldsymbol{\beta}}_{\text{NLS}}) \\ \dots \\ y_n - h(\mathbf{x}_n, \hat{\boldsymbol{\beta}}_{\text{NLS}}) \end{pmatrix}$

$$\tilde{\mathbf{X}} = \begin{pmatrix} \frac{\partial h(\mathbf{x}_1, \hat{\boldsymbol{\beta}}_{\text{NLS}})}{\partial \beta_1} & \frac{\partial h(\mathbf{x}_1, \hat{\boldsymbol{\beta}}_{\text{NLS}})}{\partial \beta_2} & \dots & \frac{\partial h(\mathbf{x}_1, \hat{\boldsymbol{\beta}}_{\text{NLS}})}{\partial \beta_k} \\ \frac{\partial h(\mathbf{x}_2, \hat{\boldsymbol{\beta}}_{\text{NLS}})}{\partial \beta_1} & \frac{\partial h(\mathbf{x}_2, \hat{\boldsymbol{\beta}}_{\text{NLS}})}{\partial \beta_2} & \dots & \frac{\partial h(\mathbf{x}_2, \hat{\boldsymbol{\beta}}_{\text{NLS}})}{\partial \beta_k} \\ \dots & \dots & \dots & \dots \\ \frac{\partial h(\mathbf{x}_n, \hat{\boldsymbol{\beta}}_{\text{NLS}})}{\partial \beta_1} & \frac{\partial h(\mathbf{x}_n, \hat{\boldsymbol{\beta}}_{\text{NLS}})}{\partial \beta_2} & \dots & \frac{\partial h(\mathbf{x}_n, \hat{\boldsymbol{\beta}}_{\text{NLS}})}{\partial \beta_k} \end{pmatrix}_{n \times k}$$

2 Función de Verosimilitud de la Muestra

Sea $g(y_i, \theta) = h(\mathbf{x}_i, \beta) + \varepsilon_i$ donde $\varepsilon_i \sim N(0, \sigma^2)$ (9)

Entonces la función densidad de probabilidad de y_i está dada por:

$$f(y_i) = \left|\frac{\partial \varepsilon_i}{\partial y_i}\right| (2\pi\sigma^2)^{-1/2} \exp\left(- \frac{(g(y_i, \theta) - h(\mathbf{x}_i, \beta))^2}{2\sigma^2} \right) \qquad (10)$$

El jacobiano de la transformación es:

$$J(y_i, \theta) = \left|\frac{\partial \varepsilon_i}{\partial y_i}\right| = \left|\frac{\partial g(y_i, \theta)}{\partial y_i}\right|$$

El logaritmo de la función de verosimilitud viene dado por:

$$\ln L = \ln\left(\prod_{i=1}^{n} f(y_i) \right) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) + \sum_{i=1}^{n} \ln J(y_i, \theta)$$

$$- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (g(y_i, \theta) - h(\mathbf{x}_i, \beta))^2 \qquad (11)$$

Notemos que si $J(y_i, \theta)$ depende de β y σ^2, entonces el estimador de mínimos cuadrados no lineales NO es el estimador de máxima verosimilitud.

Las **ecuaciones de verosimilitud** vienen dadas, en este caso, por:

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \varepsilon_i \frac{\partial h(\mathbf{x}_i, \beta)}{\partial \beta} = \mathbf{0} \qquad (12)$$

$$\frac{\partial \ln L}{\partial \theta} = \sum_{i=1}^{n} \frac{1}{J_i} \left(\frac{\partial J_i}{\partial \theta} \right) - \frac{1}{\sigma^2} \sum_{i=1}^{n} \varepsilon_i \frac{\partial g(y_i, \theta)}{\partial \theta} = \mathbf{0} \qquad (13)$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} \varepsilon_i^2 = 0 \qquad (14)$$

Usualmente estas ecuaciones son no lineales, por lo cual requerimos de métodos iterativos para resolverlas.

3 Test de Hipótesis

H_0: $\mathbf{R}(\boldsymbol{\beta}_{kx1})=\mathbf{q}_{Jx1}$ conjunto de J restricciones lineales o no lineales

H_1: $\mathbf{R}(\boldsymbol{\beta}_{kx1})\neq\mathbf{q}_{Jx1}$

Ejemplo

Para el modelo genérico $y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i$, bajo H_0 tenemos el siguiente par de restricciones no lineales en los parámetros:

$$\beta_1 + 2\beta_2^3 - \beta_3 = 0 \qquad\qquad \beta_1^2 - \beta_2 = 1$$

En este caso, $\mathbf{R}(\boldsymbol{\beta}) = \begin{pmatrix} \beta_1 + 2\beta_2^3 - \beta_3 \\ \beta_1^2 - \beta_2 \end{pmatrix}$ $\qquad \mathbf{q} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ ♦

Veremos cuatro métodos mediante los cuales podemos contrastar el tipo de hipótesis del ejemplo anterior: test de Wald, razón de verosimilitud, multiplicador de Lagrange y un test F aproximado.

3.1 Test de Wald

$$W = (\mathbf{R}(\hat{\boldsymbol{\beta}}) - \mathbf{q})' (\mathrm{Var}(\mathbf{R}(\hat{\boldsymbol{\beta}}) - \mathbf{q}))^{-1} (\mathbf{R}(\hat{\boldsymbol{\beta}}) - \mathbf{q})$$

$$= (\mathbf{R}(\hat{\boldsymbol{\beta}}) - q)' (\mathbf{C}\hat{\mathbf{V}}\mathbf{C}')^{-1} (\mathbf{R}(\hat{\boldsymbol{\beta}}) - q) \xrightarrow{d} \chi^2(J) \qquad (15)$$

donde $\hat{\mathbf{V}} = \hat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$, estimador de la varianza asintótica de $\hat{\boldsymbol{\beta}}$, estimador de mínimos cuadrados no lineales, $\mathbf{C} = \left(\frac{\partial \mathbf{R}(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}'} \right)_{J \text{ x } k}$.

En el ejemplo anterior, $\mathbf{C}_{2 \text{ x } k} = \begin{pmatrix} 1 & 6\beta_2 & -1 & 0 & \ldots & 0 \\ 2\beta_1 & -1 & 0 & 0 & \ldots & 0 \end{pmatrix}$

3.2 Test de Razón de Verosimilitud

Bajo normalidad de los errores, el logaritmo de la función de verosimilitud del modelo **no** restringido viene dado por:

$$\ln L = -\frac{n}{2}\ln(2\pi\sigma^2) + \sum_{i=1}^{n}\ln J(y_i, \theta) - \frac{\varepsilon'\varepsilon}{2\sigma^2}$$

Sea L* el logaritmo de la función de verosimilitud evaluada en los estimadores restringidos. Entonces se tiene que:

$$-2(\ln L^* - \ln L) \xrightarrow{d} \chi^2(J) \qquad (16)$$

Si el término del jacobiano **no** está presente en la función logaritmo de verosimilitud, entonces lnL evaluada en los estimadores de máxima verosimilitud no restringidos se reduce a :

$$\ln L = -\frac{n}{2}\left(1 + \ln(2\pi) + \ln\left(\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}\right)\right) = -\frac{n}{2}\left(1 + \ln(2\pi) + \ln(\hat{\sigma}^2_{NR})\right)$$

En tanto,

$$\ln L^* = -\frac{n}{2}\left(1 + \ln(2\pi) + \ln\left(\frac{\hat{\varepsilon}_R'\hat{\varepsilon}_R}{n}\right)\right) = -\frac{n}{2}\left(1 + \ln(2\pi) + \ln(\hat{\sigma}^2_R)\right)$$

Con ello, el estadígrafo de razón de verosimilitud se reduce a:

$$n(\ln(\hat{\sigma}^2_R) - \ln(\hat{\sigma}^2_{NR})) = n\ln\left(\frac{\hat{\sigma}^2_R}{\hat{\sigma}^2_{NR}}\right) \xrightarrow{d} \chi^2(J) \qquad (17)$$

Este mismo resultado se puede aplicar al modelo de regresión lineal clásico.

3.3 Test de Multiplicador de Lagrange

$$\text{Sea } \varepsilon_i^* = y_i - h(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_R),\ \tilde{\mathbf{X}}^{*\prime}\tilde{\mathbf{X}}^* = \sum_{i=1}^{n} \frac{\partial h(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_R)}{\partial \boldsymbol{\beta}} \frac{\partial h(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_R)}{\partial \boldsymbol{\beta}'}.$$

Se puede demostrar que el test de multiplicador de Lagrange toma la forma:

$$LM = \frac{\hat{\boldsymbol{\varepsilon}}^{*\prime}\tilde{\mathbf{X}}^*(\tilde{\mathbf{X}}^{*\prime}\tilde{\mathbf{X}}^*)^{-1}\tilde{\mathbf{X}}^{*\prime}\hat{\boldsymbol{\varepsilon}}^*}{(\hat{\boldsymbol{\varepsilon}}^{*\prime}\hat{\boldsymbol{\varepsilon}}^*)/n} \xrightarrow{d} \chi^2(J) \qquad (18)$$

Notemos que sólo necesitamos los estimadores restringidos.

3.4 Test F (aproximado)

Este test es análogo al utilizado en el modelo lineal clásico:

$$F(J, n-k) = \frac{(S(\hat{\boldsymbol{\beta}}_R) - S(\hat{\boldsymbol{\beta}}_{NR}))/J}{S(\hat{\boldsymbol{\beta}}_{NR})/(n-k)} \qquad (19)$$

donde $S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - h(\mathbf{x}_i, \boldsymbol{\beta}))^2$.

Es importante recalcar que la distribución de este estadíagrafo es sólo **aproximada**.

Ejemplo (Greene)

Consideremos el modelo de consumo:

$$C_i = \alpha + \beta Y_i^{\gamma} + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2) \qquad i=1, 2, ..., n$$

donde C representa consumo e Y ingreso agregado. Bajo la hipótesis nula H_0: $\gamma=1$, el modelo es lineal en los parámetros. Por lo tanto, puede ser estimado mediante mínimos cuadrados ordinarios.

Con cifras anuales de la economía americana para el período 1950-1985 se estiman las siguientes funciones de consumo agregado:

Parámetro	Modelo Lineal (Restringido)	Error Estándar	Modelo No Lineal (No Restringido)	Error Estándar
α	11.15	9.64	184.97	39.13
β	0.89	0.0058	0.252	0.081
γ	1.00		1.1535	0.039
$\hat{\varepsilon}'\hat{\varepsilon}$	12068		8421.95	
R^2	0.9956		0.99899	

Número de observaciones, n=36.

Deseamos contrastar la hipótesis H_0: $\gamma=1$ frente a la alternativa H_1: $\gamma\neq1$. Hagamos, entonces, uso de los cuatro tests vistos anteriormente:

i) Test de Razón de Verosimilitud

Dado que en este caso la función de verosimilitud no involucra el término del jacobiano, el test de razón de verosimilitud toma la forma en la ecuación (17):

$$n(\ln(\hat{\sigma}_R^2) - \ln(\hat{\sigma}_{NR}^2)) = n\ln\left(\frac{\hat{\sigma}_R^2}{\hat{\sigma}_{NR}^2}\right) \xrightarrow{d} \chi^2(J)$$

En este caso n=36, J=1, $\hat{\sigma}_R^2 = \frac{12.068}{36} = 335.22$, $\hat{\sigma}_{NR}^2 = \frac{8421.95}{36} = 233.94$

Haciendo los reemplazos correspondientes, se tiene que el test de multiplicador de Lagrange toma el valor de 12.95. Este supera al valor crítico $\chi^2_{95\%}(1)=3.84$. Por lo tanto, rechazamos H_0.

ii) Test de Wald

$$W = \frac{(1.1535-1)^2}{0.0393^2} = 15.29$$

Nuevamente, el estadígrafo calculado supera el valor crítico. Por lo tanto, rechazamos H_0.

iii) Multiplicador de Lagrange

En este caso $\tilde{\mathbf{x}}_i^{*'} = \begin{pmatrix} 1 & Y_i^{\gamma} & \beta Y_i^{\gamma} \ln(Y_i) \end{pmatrix}$

¿Por qué? Dado que $h(\mathbf{x}_i, \boldsymbol{\beta}) = \alpha + \beta Y_i^{\gamma}$, se tiene que $\frac{\partial h}{\partial \alpha} = 1$, $\frac{\partial h}{\partial \beta} = Y_i^{\gamma}$ y $\frac{\partial h}{\partial \gamma} = \beta Y_i^{\gamma} \ln(Y_i)$. Este último resultado se obtiene haciendo uso de $\frac{\partial (Y_i^{\gamma})}{\partial \gamma} = Y_i^{\gamma} \ln(Y_i)$.

Por lo tanto,

$$\tilde{\mathbf{X}} = \begin{pmatrix} \tilde{\mathbf{x}}_1' \\ \tilde{\mathbf{x}}_2' \\ \dots \\ \tilde{\mathbf{x}}_n' \end{pmatrix} = \begin{pmatrix} 1 & Y_1^{\gamma} & \beta Y_1^{\gamma} \ln(Y_1) \\ 1 & Y_2^{\gamma} & \beta Y_2^{\gamma} \ln(Y_2) \\ \dots & \dots & \dots \\ 1 & Y_n^{\gamma} & \beta Y_n^{\gamma} \ln(Y_n) \end{pmatrix}, \quad \hat{\varepsilon} = \begin{pmatrix} C_1 - \hat{\alpha} - \hat{\beta} Y_1^{\hat{\gamma}} \\ C_2 - \hat{\alpha} - \hat{\beta} Y_2^{\hat{\gamma}} \\ \dots \\ C_n - \hat{\alpha} - \hat{\beta} Y_n^{\hat{\gamma}} \end{pmatrix}$$

Para computar el test de multiplicador de Lagrange, se evalúan $\tilde{\mathbf{X}}$ y $\hat{\varepsilon}$ en los estimadores restringidos. Esto es, en aquellos obtenidos con el modelo lineal:

$\hat{\alpha} = 184.97$, $\hat{\beta} = 0.252$ (dado γ=1).

Haciendo los cálculos se obtiene:

$$LM = \frac{3547.3}{12068/36} = 10.582$$

donde $\hat{\boldsymbol{\varepsilon}}^{*\prime}\tilde{\mathbf{X}}^{*}(\tilde{\mathbf{X}}^{*\prime}\tilde{\mathbf{X}}^{*})^{-1}\tilde{\mathbf{X}}^{*\prime}\hat{\boldsymbol{\varepsilon}}^{*}=3547.3$, $\frac{\hat{\boldsymbol{\varepsilon}}^{*\prime}\hat{\boldsymbol{\varepsilon}}^{*}}{n}=\frac{12068}{36}=335.22$. Nótese que $\hat{\boldsymbol{\varepsilon}}^{*\prime}\hat{\boldsymbol{\varepsilon}}^{*}$ es la suma de cuadrados estimados bajo el modelo restringido, esto es, el modelo lineal.

Finalmente,

iv) Test F

$$F=\frac{(12068-8421.95)/1}{8421.95/(36-3)}=14.286$$

El valor crítico $F_{95\%}(1, 33)=4.18$ por lo que, al igual que en los tres casos anteriores, se rechaza H_0 ♦

Apéndice: Propiedades Asintóticas de Mínimos Cuadrados No Lineales

De la condición de primer orden para la minimización de $S(\boldsymbol{\beta})=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(y_i-h(\mathbf{x}_i,\boldsymbol{\beta}))^2$, se tiene:

$$\sum_{i=1}^{n}(y_i-h(\mathbf{x}_i,\hat{\boldsymbol{\beta}}))\frac{\partial h(\mathbf{x}_i,\hat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}}=\mathbf{0} \qquad (1)$$

Si hacemos una expansión de Taylor de primer orden de $h(\mathbf{x}_i,\hat{\boldsymbol{\beta}})$ alrededor del vector de parámetros poblacionales $\boldsymbol{\beta}_0$, tenemos[1]:

$$\sum_{i=1}^{n}(y_i-h(\mathbf{x}_i,\boldsymbol{\beta}_0)-\frac{\partial h(\mathbf{x}_i,\boldsymbol{\beta}^{*})}{\partial\boldsymbol{\beta}'}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0))\frac{\partial h(\mathbf{x}_i,\hat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}}=\mathbf{0} \qquad (2)$$

donde $\boldsymbol{\beta}^{*}$ se ubica entre $\hat{\boldsymbol{\beta}}$ y $\boldsymbol{\beta}_0$.

[1] Recordemos que la aproximación polinomial de f(x) alrededor del punto $x=x_0$ viene dada por $f(x)\approx f(x_0)+\sum_{i=1}^{p}\frac{1}{i!}\frac{d^i f(x_0)}{dx^i}(x-x_0)^i$. Por ejemplo, la aproximación de $f(x)=e^x$ alrededor de $x_0=0$ es $e^x\approx 1+x+\frac{x^2}{2}+...$

La ecuación (2) puede reescribirse como:

$$\sum_{i=1}^{n} \frac{\partial h(\mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \varepsilon_i + \left(\sum_{i=1}^{n} \frac{\partial h(\mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}'} \right) (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \mathbf{0}$$

De lo cual

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \left(\sum_{i=1}^{n} \frac{\partial h(\mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}'} \right)^{-1} \sum_{i=1}^{n} \frac{\partial h(\mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \varepsilon_i \qquad (3)$$

Probar consistencia de $\hat{\boldsymbol{\beta}}$ es técnico. Amemiya (1985) presenta una demostración formal. Probar normalidad asintótica no es difícil una vez que se ha establecido consistencia.

La ecuación (3) se puede reescribir como:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial h(\mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}'} \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial h(\mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \varepsilon_i \quad (4)$$

$$\text{Sea } \tilde{\mathbf{X}}^0 = \begin{pmatrix} \frac{\partial h(\mathbf{x}_1, \boldsymbol{\beta}_0)}{\partial \beta_1} & \frac{\partial h(\mathbf{x}_1, \boldsymbol{\beta}_0)}{\partial \beta_2} & \cdots & \frac{\partial h(\mathbf{x}_1, \boldsymbol{\beta}_0)}{\partial \beta_k} \\ \frac{\partial h(\mathbf{x}_2, \boldsymbol{\beta}_0)}{\partial \beta_1} & \frac{\partial h(\mathbf{x}_2, \boldsymbol{\beta}_0)}{\partial \beta_2} & \cdots & \frac{\partial h(\mathbf{x}_2, \boldsymbol{\beta}_0)}{\partial \beta_k} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial h(\mathbf{x}_n, \boldsymbol{\beta}_0)}{\partial \beta_1} & \frac{\partial h(\mathbf{x}_n, \boldsymbol{\beta}_0)}{\partial \beta_2} & \cdots & \frac{\partial h(\mathbf{x}_n, \boldsymbol{\beta}_0)}{\partial \beta_k} \end{pmatrix}_{n \times k} = \begin{pmatrix} \tilde{\mathbf{x}}_1^{0\prime} \\ \tilde{\mathbf{x}}_2^{0\prime} \\ \cdots \\ \tilde{\mathbf{x}}_n^{0\prime} \end{pmatrix}$$

$$\text{y } \tilde{\mathbf{x}}_i^{0\prime} = \begin{pmatrix} \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}_0)}{\partial \beta_1} & \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}_0)}{\partial \beta_2} & \cdots & \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}_0)}{\partial \beta_k} \end{pmatrix}$$

Entonces $\tilde{\mathbf{X}}^{0\prime}\tilde{\mathbf{X}}^{0} = \sum_{i=1}^{n} \tilde{\mathbf{x}}_i^{0\prime}\tilde{\mathbf{x}}_i^{0}$.

Ahora, supongamos que $\operatorname{plim}\left(\frac{\tilde{\mathbf{X}}^{0\prime}\tilde{\mathbf{X}}^{0}}{n}\right) = \mathbf{Q}$, donde $\mathbf{Q}$ es una matriz invertible.

Entonces $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial h(\mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}'} \xrightarrow{p} \mathbf{Q}$

por consistencia de $\hat{\boldsymbol{\beta}}$.

Por otra parte, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\partial h(\mathbf{x}_i, \hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} \varepsilon_i \xrightarrow{d} N(\mathbf{0}, \sigma^2\mathbf{Q})$, por consistencia de $\hat{\boldsymbol{\beta}}$ y el teorema del límite central de Lindberg-Feller:

Sea $\overline{w} \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \varepsilon_i$. Se tiene, entonces, que $\overline{w}$ es el promedio de n variables independientes $\frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \varepsilon_i$ con esperanza 0 y varianza $\sigma^2 \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}'}$

Entonces, de lo anterior, se deduce que:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \sigma^2\mathbf{Q}^{-1}) \qquad (5)$$

MINIMOS CUADRADOS NO LINEALES: EXTENSIONES

1 Algoritmo Numérico: Gauss-Newton

Consideremos primero el modelo de regresión no lineal:

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i \qquad i=1, 2, ..., n \qquad (1)$$

Linealicemos la función $h(\mathbf{x}_i, \boldsymbol{\beta})$ alrededor del vector de parámetros $\boldsymbol{\beta}_0$ mediante una expansión de Taylor de primer orden:

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx h(\mathbf{x}, \boldsymbol{\beta}_0) + \sum_{k=1}^{K} \frac{\partial h(\mathbf{x}, \boldsymbol{\beta}_0)}{\partial \beta_k} (\beta_k - \beta_k^0) \qquad (2)$$

(después de suprimir el subíndice ‘i’).

Después de agrupar términos obtenemos:

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx \left(h(\mathbf{x}, \boldsymbol{\beta}_0) - \sum_{k=1}^{K} \frac{\partial h(\mathbf{x}, \boldsymbol{\beta}_0)}{\partial \beta_k} \beta_k^0 \right) + \sum_{k=1}^{K} \frac{\partial h(\mathbf{x}, \boldsymbol{\beta}_0)}{\partial \beta_k} \beta_k$$

Denotemos $\frac{\partial h(\mathbf{x}, \boldsymbol{\beta}_0)}{\partial \beta_k}$ como $\mathbf{x}_k^0$. Para un valor dado del vector $\boldsymbol{\beta}_0$, $\mathbf{x}_k^0$ es una función de los datos exclusivamente y NO depende del vector de parámetros desconocidos.

Sea $h(\mathbf{x}_i, \boldsymbol{\beta}_0) \equiv h^0$, entonces:

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx \left(h^0 - \sum_{k=1}^{K} \mathbf{x}_k^0 \beta_k^0 \right) + \sum_{k=1}^{K} \mathbf{x}_k^0 \beta_k$$

De esto, $y \approx \left(h^0 - \mathbf{x}^{0\prime} \boldsymbol{\beta}^0 \right) + \mathbf{x}^{0\prime} \boldsymbol{\beta} + \varepsilon$

donde $\mathbf{x}^{0\prime} = \left(x_1^0 \quad x_2^0 \quad ... \quad x_k^0 \right) = \left(\frac{\partial h(\mathbf{x}, \boldsymbol{\beta}_0)}{\partial \beta_1} \quad \frac{\partial h(\mathbf{x}, \boldsymbol{\beta}_0)}{\partial \beta_2} \quad ... \quad \frac{\partial h(\mathbf{x}, \boldsymbol{\beta}_0)}{\partial \beta_k} \right)$

$$\boldsymbol{\beta}_0 = \begin{pmatrix} \beta_1^0 \\ \beta_2^0 \\ \dots \\ \beta_k^0 \end{pmatrix}.$$

Sea $y_0=y-h^0 + \mathbf{x}^{0\prime}\boldsymbol{\beta}_0$. Entonces tenemos el modelo de regresión (aproximado):

$$y_0 = \mathbf{x}^{0\prime}\boldsymbol{\beta}_0 + \varepsilon$$

Para un valor dado de $\boldsymbol{\beta}_0$ podemos computar y_0 y $\mathbf{x}^0$.

Ejemplo

Consideremos nuevamente el modelo:

$$y = \beta_1 + \beta_2 \exp(\beta_3 x) + \varepsilon \equiv h(\mathbf{x}, \boldsymbol{\beta}) + \varepsilon$$

Recordemos que:

$$\frac{\partial h(\mathbf{x}, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} = \begin{pmatrix} x_1^0 \\ x_2^0 \\ x_3^0 \end{pmatrix} = \begin{pmatrix} 1 \\ \exp(\beta_3^0 x) \\ \beta_2^0 \ x \exp(\beta_3^0 x) \end{pmatrix}, \ \boldsymbol{\beta}_0 = \begin{pmatrix} \beta_1^0 \\ \beta_2^0 \\ \beta_3^0 \end{pmatrix}$$

Entonces, dado el vector $\boldsymbol{\beta}_0$, podemos computar la variable y_0 para cada observación:

$$y_0 = y - h(x, \beta_1^0, \beta_2^0, \beta_3^0) + \beta_1^0 x_1^0 + \beta_2^0 x_2^0 + \beta_3^0 x_3^0$$

y correr una regresión de y_0 en x_1^0, x_2^0 y x_3^0 para así obtener un nuevo vector de parámetros $\hat{\boldsymbol{\beta}}$.

Este proceso continúa hasta que la diferencia entre los sucesivos vectores $\hat{\boldsymbol{\beta}}$ es suficientemente pequeña como para aceptar la convergencia♦

En general, tenemos que:

$$\hat{\boldsymbol{\beta}}_{t+1} = \left(\sum_{i=1}^{n} \mathbf{x}_i^t \mathbf{x}_i^{t\prime} \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{x}_i^t (y_i - h_i^t + \mathbf{x}_i^{t\prime} \hat{\boldsymbol{\beta}}_t) \right)$$

$$= \hat{\boldsymbol{\beta}}_t + \left(\sum_{i=1}^{n} \mathbf{x}_i^t \mathbf{x}_i^{t\prime} \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{x}_i^t (y_i - h_i^t) \right)$$

$$= \hat{\boldsymbol{\beta}}_t + (\mathbf{X}^{t\prime} \mathbf{X}^t)^{-1} \mathbf{X}^{t\prime} \hat{\boldsymbol{\varepsilon}}^t \qquad (3)$$

donde $\mathbf{X}^t = \begin{pmatrix} \mathbf{x}_1^{t\prime} \\ \mathbf{x}_2^{t\prime} \\ \dots \\ \mathbf{x}_n^{t\prime} \end{pmatrix}$ $\qquad \mathbf{x}_i^{t\prime} = \begin{pmatrix} \frac{\partial h(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_t)}{\partial \beta_1} & \frac{\partial h(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_t)}{\partial \beta_2} & \dots & \frac{\partial h(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_t)}{\partial \beta_k} \end{pmatrix}$

$$\hat{\boldsymbol{\varepsilon}}^t = \begin{pmatrix} y_1 - h(\mathbf{x}_1, \hat{\boldsymbol{\beta}}_t) \\ y_2 - h(\mathbf{x}_2, \hat{\boldsymbol{\beta}}_t) \\ \dots \\ y_n - h(\mathbf{x}_n, \hat{\boldsymbol{\beta}}_t) \end{pmatrix}$$

La relación $\hat{\boldsymbol{\beta}}_{t+1} = \hat{\boldsymbol{\beta}}_t + (\mathbf{X}^{t\prime} \mathbf{X}^t)^{-1} \mathbf{X}^{t\prime} \hat{\boldsymbol{\varepsilon}}^t$ indica que en cada **iteración** se actualiza el vector de parámetros estimados en la iteración previa, al correr una regresión de los **residuos** de mínimos cuadrados no lineales en las primeras derivadas de la función h(**x**, $\boldsymbol{\beta}$), siendo todas las expresiones evaluadas en $\hat{\boldsymbol{\beta}}_t$.

El proceso converge una vez que el vector $\mathbf{X}^{t\prime} \hat{\boldsymbol{\varepsilon}}^t \approx \mathbf{0}$. (Notemos que dicho resultado es similar al de las ecuaciones normales del modelo lineal clásico).