Mundo Social **06**, Notas (2026)
Disponible en https://mundo-social.com/

Modelos No Lineales en Econometría

Anjaly Arcos Huaman [1]

Melany Vega Lugo [2]

Curso dictado en 26-1 por Marvin Padilla

[1] *Undergraduate Economist, Universidad Nacional Mayor de San Marcos*
anjaly.arcos@gmail.com
[2] *Undergraduate Economist, Universidad Nacional Mayor de San Marcos*
melany.vega@unmsm.edu.pe
Disponible en: https://mundo-social.com/

A continuación, se presenta una nota académica dedicada al estudio de los modelos no lineales en econometría, desarrollada a partir de los contenidos abordados en el curso de Econometría II, dictado por el profesor Mg. Marvin Padilla. Este material ha sido elaborado con el mayor respeto hacia el profesor y con fines estrictamente educativos, con el propósito de contribuir a la comprensión y difusión de sus enseñanzas, así como de asegurar que su valioso legado académico perdure en el tiempo.

Recomendamos encarecidamente la lectura de los libros y materiales del profesor, los cuales constituyen una fuente rigurosa y profunda para el estudio.

Esta nota incluye el desarrollo de los siguientes subtemas fundamentales:

- Modelos no lineales y su especificación
- Mínimos cuadrados no lineales (NLS)
- Estimación por máxima verosimilitud
- Transformación Box-Cox
- Contraste de restricciones no lineales

Ante cualquier error, comentario o sugerencia, no dude el lector en escribirnos al correo que aparece en la portada del presente documento.

1. Punto de Partida: Del Modelo Lineal al No Lineal

1.1. Los dos modelos fundamentales

En Econometría 1 trabajamos con el modelo lineal general:

$$Y = X\beta + \varepsilon \tag{1}$$

donde la relación entre Y y X es *lineal en los parámetros* β. En Econometría 2 generalizamos esta idea al **modelo no lineal**:

$$y_t = f(x_t, \beta) + \varepsilon_t, \qquad t = 1, 2, \ldots, T \tag{2}$$

$f(x_t, \beta)$ es una función *arbitraria* de las variables explicativas x_t **y** de los parámetros β. El problema central surge cuando f es **no lineal en** β: en ese caso no existe una fórmula cerrada como $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Y$ y debemos recurrir a métodos distintos.

2. Clasificación de Especificaciones No Lineales

A continuación, se presentan cinco ejemplos canónicos. La clave es identificar *dónde* aparece la no linealidad: en las variables o en los parámetros.

2.1. Cinco modelos y su tratamiento

a. Modelo 1: Potencia en la variable endógena

$$y_t = \beta_0 + \beta_1 x_t^{\beta_2} + \varepsilon_t$$

Si $\beta_2 = 1$: $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$ (lineal). El parámetro β_2 controla la *curvatura* de la relación. Si $\beta_2 > 1$ la relación es convexa (crece cada vez más rápido); si $0 < \beta_2 < 1$ es cóncava (crece pero cada vez más despacio). Esta no linealidad afecta a los **parámetros**, por lo que **no puede resolverse** con una transformación simple de datos.

b. Modelo 2: Transformación Box-Cox en la variable independiente

$$y_t^{1-\beta_2} = \beta_0 + \beta_1 x_t + \varepsilon_t$$

Si $\beta_2 = 0$: $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$ (lineal).

c. Modelo 3: Cociente de parámetros

$$y_t = \beta_0 + \beta_1 \frac{x_t^{1-\beta_2}}{\beta_0} + \varepsilon_t$$

Si $\beta_2 = 0$ y $\beta_3 = 1$: se reduce al modelo lineal.

d. Modelo 4: Razón de parámetros como pendiente

$$y_t = \beta_0 + \frac{\beta_1}{\beta_2} x_t + \varepsilon_t$$

Definiendo $\tilde{\beta}_1 = \beta_1/\beta_2$ se obtiene $y_t = \beta_0 + \tilde{\beta}_1 x_t + \varepsilon_t$, que es lineal. El parámetro β_1/β_2 puede recuperarse a partir del coeficiente estimado.

Interpretación:

En este caso la no linealidad es *recuperable*: estimamos el cociente directamente como una sola pendiente lineal y luego hacemos la partición.

e. Modelo 5: No linealidad irreducible en los parámetros

$$C_t = \beta_0 + \beta_1 y_t^{\beta_3} + \varepsilon_t$$

Interpretación: Modelo de consumo no lineal

Este es el caso **problemático**: β_3 está en el exponente de y_t y no puede separarse de β_1 mediante álgebra elemental. No hay transformación de datos que lo linealice. Requiere métodos de estimación no lineales. Económicamente, β_3 captura la *elasticidad ingreso del consumo*: si $\beta_3 = 1$ la propensión marginal es constante (lineal); si $\beta_3 < 1$ el consumo crece pero con rendimientos decrecientes.

3. Transformación Box-Cox

Definición: Transformación Box-Cox

La familia de transformaciones Box-Cox parametriza la forma funcional mediante λ:

$$y_t = \alpha + \beta \cdot \frac{x_t^{\lambda} - 1}{\lambda} + \varepsilon_t \tag{3}$$

3.1. Casos especiales según el valor de λ

λ	Modelo resultante	Expresión
$\lambda = 1$	Lineal	$y_t = \alpha + \beta x_t + \varepsilon_t$
$\lambda = 0$	Logarítmico	$y_t = \alpha + \beta \ln x_t + \varepsilon_t$
$\lambda = -1$	Inverso	$y_t = \alpha + \beta \cdot \frac{1}{x_t} + \varepsilon_t$

Interpretación: Utilidad de Box-Cox

Box-Cox anida en un único parámetro λ las tres especificaciones más comunes. Si λ se estima a partir de los datos, el modelo elige automáticamente la forma funcional más adecuada. Si λ es conocido, el modelo se vuelve **lineal** en α y β y puede estimarse por MCO ordinario.

3.2. Modelo exponencial y la condición de no separabilidad

Considera el modelo:

$$y_t = \beta_0 \, e^{x_t \beta_1} + \varepsilon_t$$

Al tomar logaritmo parecería linealizable: $\ln y_t = \ln \beta_0 + x_t \beta_1 + \varepsilon_t$. Sin embargo, en (3.2) el error ε_t es **aditivo**, mientras que $\ln y_t = \tilde{\beta}_0 + x_t \beta_1 + \varepsilon_t$ requeriría un error **multiplicativo** en el modelo original. Estos dos modelos tienen implicaciones estadísticas distintas y no son equivalentes: esta es la *Condición de No Separabilidad*.

Regla práctica: solo se puede aplicar logaritmo para linealizar si el error original es multiplicativo, es decir, $y_t = \beta_0 e^{x_t \beta_1} \cdot e^{\varepsilon_t}$, lo que da $\ln y_t = \ln \beta_0 + x_t \beta_1 + \varepsilon_t$. Si el error es aditivo como en (3.2), tomar logaritmo introduce una transformación inconsistente.

4. Linealización por Serie de Taylor

4.1. Idea central

Dado el modelo no lineal (2), un primer enfoque consiste en aproximar $f(x_t, \beta)$ localmente alrededor de un estimador inicial $\hat{\beta}$ usando la **serie de Taylor de primer orden**:

$$f(x_t, \beta) \approx f(x_t, \hat{\beta}) + \nabla f(x_t, \hat{\beta})(\beta - \hat{\beta}) \tag{4}$$

Interpretación: Serie de Taylor en $\mathbb{R}^N$

La serie de Taylor dice: cerca de un punto $\hat{\beta}$, cualquier función suave puede aproximarse por un *plano tangente*. $\nabla f(x_t, \hat{\beta})$ es el **vector gradiente**: indica la dirección y magnitud del cambio de f al mover cada parámetro una unidad. La aproximación es mejor cuanto más cerca está β de $\hat{\beta}$.

4.2. Construcción del modelo linealizado

Sustituyendo (4) en el modelo (2):

$$y_t \approx f(x_t, \hat{\beta}) + \nabla f(x_t, \hat{\beta})(\beta - \hat{\beta}) + \varepsilon_t$$

Reordenando - llevando el término $\nabla f(x_t, \hat{\beta})\hat{\beta}$ al lado izquierdo:

$$\underbrace{y_t - f(x_t, \hat{\beta}) + \nabla f(x_t, \hat{\beta})\hat{\beta}}_{y_t^*} \approx \nabla f(x_t, \hat{\beta})\, \beta + \varepsilon_t$$

Se obtiene así el **modelo linealizado**:

$$\boxed{y_t^* \approx \nabla f(x_t, \hat{\beta})\, \beta + \varepsilon_t} \tag{5}$$

donde la variable transformada y_t^* está definida como:

$$y_t^* = y_t - f(x_t, \hat{\beta}) + \nabla f(x_t, \hat{\beta})\,\hat{\beta}$$

Interpretación: y_t^*: la variable endógena ajustada

y_t^* es la variable que efectivamente se regresa en el modelo linealizado. Contiene tres ingredientes:

1. y_t: la observación original.
2. $-f(x_t, \hat{\beta})$: se resta el valor de la función en la estimación inicial.
3. $+\nabla f(x_t, \hat{\beta})\hat{\beta}$: se añade la corrección del gradiente.

Esto permite que el modelo linealizado (5) tenga a $\nabla f(x_t, \hat{\beta})$ jugando el papel de la matriz $\boldsymbol{X}$ y a y_t^* jugando el papel de y_t en el MCO estándar.

4.3. El Jacobiano: la matriz de variables explicativas generalizadas

El **Jacobiano** $\nabla f(\boldsymbol{x}, \beta)$ es la matriz de todas las derivadas parciales de f respecto a cada parámetro, evaluada en cada período t:

$$\nabla f(\boldsymbol{x}, \beta) = \begin{bmatrix} \dfrac{\partial f_1}{\partial \beta_1} & \dfrac{\partial f_1}{\partial \beta_2} & \cdots & \dfrac{\partial f_1}{\partial \beta_k} \\ \dfrac{\partial f_2}{\partial \beta_1} & \dfrac{\partial f_2}{\partial \beta_2} & \cdots & \dfrac{\partial f_2}{\partial \beta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f_T}{\partial \beta_1} & \dfrac{\partial f_T}{\partial \beta_2} & \cdots & \dfrac{\partial f_T}{\partial \beta_k} \end{bmatrix}_{T \times k}$$

Interpretación: Jacobiano como generalización de $\boldsymbol{X}$

- **Dimensión:** $T \times k$ — igual que la matriz de regresores $\boldsymbol{X}$ en el modelo lineal.
- **Cada fila** corresponde a un período t y contiene la sensibilidad de f_t respecto a cada uno de los k parámetros.
- **Cada columna** i contiene $\partial f_t / \partial \beta_i$ para $t = 1, \ldots, T$, es decir, cómo afecta β_i a la función en todos los períodos.
- En el caso lineal $f(x_t, \beta) = x_t'\beta$, el Jacobiano es simplemente $\nabla f = X\beta$, recuperando exactamente la estructura del modelo lineal.

4.4. Ejemplo de linealización: modelo exponencial

Ejemplo: Modelo $y_t = \alpha e^{\beta x_t} + \varepsilon_t$

Sea $f(x_t, \alpha, \beta) = \alpha e^{\beta x_t}$. El vector gradiente (Jacobiano de f respecto a $\nabla f(X_t, \hat{\alpha}, \hat{\beta}) = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$) es:

$$\nabla f(X_t, \hat{\alpha}, \hat{\beta}) = \left[\frac{\partial f_t}{\partial \alpha}, \frac{\partial f_t}{\partial \beta}\right] = \left[e^{\beta x_t}, \alpha x_t e^{\beta x_t}\right]$$

$$\begin{aligned} y_t &= \alpha e^{\beta x_t} + u_t \\ y_t^* &= \nabla f(x_t, \hat{\alpha}, \hat{\beta}) \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + u_t \\ y_t^* &\simeq y_t - \hat{\alpha} e^{\hat{\beta} x_t} + \nabla f(x_t, \hat{\alpha}, \hat{\beta}) \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} \end{aligned}$$

Con valores iniciales $\hat{\alpha} = 1$, $\hat{\beta} = 0$:

$$y_t^* \simeq y_t - \underbrace{(1)e^{(0)x_t}}_{1} + \left[\underbrace{e^{(0)x_t}}_{1}, \underbrace{(1)x_t e^{(0)x_t}}_{1}\right] \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

La variable transformada resulta:

$$y_t^* \simeq y_t - \not{1} + \not{1} \rightarrow y_t^* = y_t$$

Y el modelo linealizado inicial es:

$$y_t^* \simeq \begin{bmatrix} 1 & x_t \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + u_t \Rightarrow y_t^* \approx \alpha + \beta x_t + u_t$$

Este es el modelo de *primera iteración*: se estima por MCO, se obtiene $\hat{\alpha}^{(1)}$ y $\hat{\beta}^{(1)}$, y se repite el proceso hasta convergencia.

5. Mínimos Cuadrados No Lineales (MCNL)

5.1. Planteamiento

El método de Mínimos Cuadrados No Lineales aplica directamente la idea de minimizar la suma de residuos al cuadrado al modelo no lineal, sin necesidad de linealizarlo:

$$\min_{\hat{\beta}} SR(\hat{\beta}) = \sum_{t=1}^{T} \hat{\varepsilon}_t^2 = \sum_{t=1}^{T} \left[y_t - f(x_t, \hat{\beta})\right]^2 \tag{6}$$

Interpretación: MCNL vs. MCO

La lógica es idéntica al MCO: buscar los parámetros que hacen que el modelo se acerque lo más posible a los datos observados. La diferencia es que f es no lineal, por lo que **no hay solución analítica cerrada** y el sistema de ecuaciones normales resultante debe resolverse numéricamente.

5.2. Condición de Primer Orden (C.P.O.)

Derivando (6) respecto a $\hat{\beta}_j$ e igualando a cero:

$$\frac{\partial SR(\hat{\beta})}{\partial \hat{\beta}_j} = 0 \qquad \Longrightarrow \qquad -2\sum_{t=1}^{T}\left[y_t - f(x_t,\hat{\beta})\right]\frac{\partial f(x_t,\hat{\beta})}{\partial \hat{\beta}_j} = 0 \tag{7}$$

$$\sum_{t=1}^{T}\underbrace{[y_t - f(x_t,\hat{\beta})]}_{\hat{\varepsilon}_t}\cdot\frac{\partial f(x_t,\beta)}{\partial \hat{\beta}_j} = 0_k \tag{8}$$

Interpretación: Derivada de la suma residual

La regla de la cadena da: primero derivamos el cuadrado (el 2 baja como coeficiente y el signo negativo viene de la derivada de $-f$) y luego multiplicamos por la derivada interna $\partial f/\partial\hat{\beta}_j$. El factor -2 se cancela al igualar a cero.

Para los k parámetros, se obtiene el **sistema de ecuaciones normales**:

$$\sum_{t=1}^{T}\left[y_t - f(x_t,\hat{\beta})\right]\frac{\partial f(x_t,\hat{\beta})}{\partial \hat{\beta}_j} = 0, \quad j = 1,\ldots,k$$

$$\sum_{t=1}^{T}[y_t - f(x_t,\hat{\beta})]\begin{bmatrix}\frac{\partial f}{\partial \hat{\beta}_1}\\ \frac{\partial f}{\partial \hat{\beta}_2}\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}$$

$\sum_{t=1}^{T}[y_t - f(x_t,\hat{\beta})]\left[\frac{\partial f}{\partial \hat{\beta}}\right] = \mathbf{0}_2$

$\sum_{t=1}^{T}[y_t - f(x_t,\beta)]\frac{\partial f}{\partial \hat{\beta}_{1,2}} = 0 \sum_{t=1}^{T}[y_t - f(x_t,\hat{\beta})]\frac{\partial f}{\partial \hat{\beta}} = 0$

$\sum_{t=1}^{T}[y_t - f(x_t,\hat{\beta})]\begin{bmatrix}\frac{\partial f}{\partial \hat{\beta}_1}\\ \frac{\partial f}{\partial \hat{\beta}_2}\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}$

$\sum_{t=1}^{T}[y_t - f(x_t,\hat{\beta})]\left[\frac{\partial f}{\partial \hat{\beta}}\right] = \mathbf{0}_2$

Descomponiendo:

$$\sum_{t=1}^{T} y_t\frac{\partial f}{\partial \hat{\beta}} - \sum_{t=1}^{T} f(x_t,\hat{\beta})\frac{\partial f}{\partial \hat{\beta}} = 0_k \Rightarrow \sum_{t=1}^{T} y_t\frac{\partial f_t}{\partial \hat{\beta}} = \sum_{t=1}^{T} f(x_t,\hat{\beta})\frac{\partial f_t}{\partial \hat{\beta}}$$

En notación matricial compacta:

$$\boxed{\left(\frac{\partial f(\hat{\beta})}{\partial \beta}\right)' Y = \left(\frac{\partial f(\hat{\beta})}{\partial \beta}\right)' f(\boldsymbol{x}, \hat{\beta})}$$

$$\left(\frac{\partial f(\hat{\beta})}{\partial \beta}\right) y = \left(\frac{\partial f(\hat{\beta})}{\partial \beta}\right) f(x, \hat{\beta})$$

$$\text{Si } f(x, \hat{\beta}) = x\hat{\beta} \quad , \quad \frac{\partial f}{\partial \beta} = x'$$

$$x'y = x'x\hat{\beta} \rightarrow \hat{\beta} = (x'x)^{-1}x'y$$

Interpretación: Condición de ortogonalidad

Esta ecuación se llama **condición de ortogonalidad**: los residuos $\hat{\varepsilon}_t$ deben ser ortogonales (producto punto igual a cero) a las columnas del Jacobiano. En el caso lineal, el Jacobiano es $\boldsymbol{X}$ y la condición se reduce a $\boldsymbol{X'\hat{\varepsilon}} = \mathbf{0}$, que es exactamente la condición de ortogonalidad del MCO estándar. El sistema (5.2) es **no lineal en** $\hat{\beta}$ y debe resolverse mediante algoritmos numéricos iterativos.

$$\cancel{\left(\frac{\partial f(\hat{\beta})}{\partial \beta}\right)' f(x, \hat{\beta})} + \left(\frac{\partial f(\hat{\beta})}{\partial \beta}\right)' \hat{\varepsilon} = \cancel{\left(\frac{\partial f(\hat{\beta})}{\partial \beta}\right)' f(x, \hat{\beta})}$$

$$\left(\frac{\partial f(\hat{\beta})}{\partial \beta}\right)' \hat{\varepsilon} = 0$$

5.3. Varianza del estimador MCNL

Asintóticamente, el estimador MCNL tiene distribución Normal con varianza:

$$Var(\hat{\beta}) = \sigma_\varepsilon^2 \left[\left(\frac{\partial f(\hat{\beta})}{\partial \beta}\right)' \left(\frac{\partial f(\hat{\beta})}{\partial \beta}\right)\right]^{-1}$$

Interpretación: Varianza del MCNL vs. MCO

Comparando con el caso lineal donde $\text{Var}(\hat{\beta}) = \sigma^2(\boldsymbol{X'X})^{-1}$:

- El Jacobiano $\nabla f(\hat{\beta})$ reemplaza a $\boldsymbol{X}$.
- La estructura es idéntica: σ^2 por la inversa de la suma de productos cruzados.
- En el caso lineal, $\nabla f = \boldsymbol{X}$ y (5.3) se reduce exactamente a $\sigma^2(\boldsymbol{X'X})^{-1}$.

6. Estimador de Máxima Verosimilitud (MV)

6.1. Motivación y planteamiento

Dadas n observaciones $Y_1, Y_2, \ldots, Y_n$ con función de densidad $f(y_i, \theta)$, el **estimador de máxima verosimilitud** es el valor de θ que *maximiza la probabilidad de haber observado los datos que vemos*:

$$\hat{\theta}_{MV} = \arg\max_{\theta} L(\theta \mid \boldsymbol{y}) \tag{9}$$

Interpretación: Intuición del MV

Imagina que lanzas una moneda 10 veces y obtienes 7 caras. El estimador MV de la probabilidad de cara es $p = 7/10 = 0{,}7$, porque ese es el valor que hace más "probable" haber observado exactamente 7 caras en 10 lanzamientos. En econometría, el MV elige $\hat{\theta}$ de modo que los datos observados sean lo más "verosímiles" posible bajo el modelo propuesto.

6.2. Función de verosimilitud

La **probabilidad conjunta** de observar $(Y_1, Y_2, \ldots, Y_n)$ es:

$$f(Y_1, Y_2, \ldots, Y_n \mid \theta) = L(\theta \mid Y_1, \ldots, Y_n)$$

Cuando las Y_i son i.i.d. (idénticamente distribuidas e independientes):

$$L(\theta \mid \boldsymbol{y}) = \prod_{t=1}^{T} f(y_t \mid \theta) \tag{10}$$

Interpretación: Por qué el producto

La independencia entre observaciones permite escribir la probabilidad conjunta como el *producto* de las probabilidades individuales: es la misma razón por la que $P(\text{cara y cara}) = P(\text{cara}) \times P(\text{cara})$ al lanzar dos monedas independientes.

6.3. Log-verosimilitud

Para simplificar la optimización, se trabaja con el **logaritmo de la verosimilitud**:

$$\ell(\theta \mid \boldsymbol{y}) = \ln L(\theta \mid \boldsymbol{y}) = \sum_{t=1}^{T} \ln f(y_t \mid \theta) \tag{11}$$

Interpretación: Ventaja del logaritmo

- El logaritmo convierte el *producto* en una *suma*, que es mucho más fácil de derivar.
- Como $\ln(\cdot)$ es una función estrictamente creciente, el argmax de ℓ coincide con el argmax de L: $\hat{\theta}_{MV} = \arg\max_{\theta} \ell(\theta \mid \boldsymbol{y})$.
- Numéricamente, los productos de muchos números pequeños pueden generar underflow computacional; la suma de logaritmos evita este problema.

6.4. C.P.O. para obtener $\hat{\theta}_{MV}$

$$\frac{\partial \ln L(\theta)}{\partial \theta} = 0 \quad \Longrightarrow \quad \hat{\theta}_{MV}$$

Ejemplo: MV para una distribución de Poisson

Sea $f(y \mid x) = \dfrac{\lambda e^{-\lambda}(\lambda y)^x}{x!}$, con $y \geq 0$, $\lambda \geq 0$. Los datos observados son:

y_i	2	5	6	7
x_i	4	10	18	20

La log-verosimilitud conjunta es:

$$\begin{aligned}
\ell(\lambda \mid \text{data}) &= \sum_{i=1}^{4} \ln f(y_i \mid x_i) \\
&= \sum_{i=1}^{4} [\ln \lambda - \lambda y_i + x_i \ln(\lambda y_i) - \ln x_i!] \\
&= 4 \ln \lambda - \lambda \underbrace{\sum y_i}_{20} + \underbrace{\sum x_i \ln(\lambda y_i)}_{\sum x_i \ln y_i + \ln \lambda \sum x_i} - \sum \ln x_i! \\
&= 4 \ln \lambda - 20\lambda + \sum x_i \ln y_i + \ln \lambda \underbrace{\sum x_i}_{52} - \sum \ln x_i!
\end{aligned}$$

C.P.O.:

$$\frac{\partial \ell(\lambda \mid \text{data})}{\partial \lambda} = 0 \qquad \Longrightarrow \qquad \frac{56}{\lambda} - 20 = 0 \qquad \Longrightarrow \qquad \hat{\lambda} = 2{,}8$$

C.S.O. (condición suficiente de máximo):

$$\frac{\partial^2 \ell}{\partial \lambda^2} = -\frac{56}{\lambda^2} < 0 \quad \checkmark$$

La derivada segunda negativa confirma que $\hat{\lambda} = 2{,}8$ es efectivamente un **máximo**.

6.5. Aplicación al Modelo Lineal General (MLG)

a. Función de log-verosimilitud

Para $Y = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}$ con $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, el vector de parámetros es $\theta = [\beta', \sigma^2]'$. La función de densidad de cada ε_i es:

$$f(\varepsilon_i \mid \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right)$$

La log-verosimilitud conjunta resulta:

$$\begin{aligned}\ell(\beta, \sigma^2 \mid \text{data}) &= -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\varepsilon_t^2 \\ &= -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta)\end{aligned}$$

Interpretación: Estructura de la log-verosimilitud

(**??**) tiene tres términos:

1. $-\frac{T}{2}\ln(2\pi)$: constante que no depende de los parámetros.
2. $-\frac{T}{2}\ln(\sigma^2)$: penaliza varianzas grandes.
3. $-\frac{1}{2\sigma^2}\text{SCR}$: penaliza residuos grandes.

Maximizar ℓ respecto a β equivale a **minimizar la SCR**, que es exactamente el principio de MCO. Esto demuestra que bajo normalidad de errores, $\hat{\beta}_{MV} = \hat{\beta}_{MCO} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

b. C.P.O. respecto a β y σ^2

Respecto a β:

$$\begin{aligned}\frac{\partial \ell}{\partial \beta} = 0 \implies -\frac{1}{2\sigma^2}\frac{\partial \text{SCR}}{\partial \beta} = 0 \implies \frac{\partial \text{SCR}}{\partial \beta} = 0 \\ \implies \hat{\beta}_{MV} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\end{aligned}$$

Respecto a σ^2:

$$\begin{aligned}\frac{\partial \ell}{\partial \sigma^2} = 0 \implies -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}\text{SCR} = 0 \\ \implies \hat{\sigma}^2_{MV} = \frac{\text{SCR}}{T}\end{aligned}$$

Interpretación: $\hat{\sigma}^2_{MV}$ vs. $\hat{\sigma}^2_{MCO}$

El estimador MCO divide entre $T - k$ (grados de libertad) para ser insesgado: $\hat{\sigma}^2_{MCO} = \text{SCR}/(T-k)$. El MV divide entre T, por lo que es **sesgado** en muestras finitas, aunque **consistente**: cuando $T \to \infty$, la diferencia entre T y $T - k$ es despreciable.

7. Matriz de Información y Cota de Cramér-Rao

7.1. Derivadas de segundo orden de la log-verosimilitud

Para el MLG con $\hat{\theta} = [\hat{\beta}', \hat{\sigma}^2]'$, la matriz hessiana de la log-verosimilitud es:

$$\frac{\partial^2 \ell}{\partial \hat{\theta}\, \partial \hat{\theta}'} = \begin{bmatrix} \dfrac{\partial^2 \ell}{\partial \beta\, \partial \beta'} & \dfrac{\partial^2 \ell}{\partial \beta\, \partial \sigma^2} \\ \dfrac{\partial^2 \ell}{\partial \sigma^2 \partial \beta'} & \dfrac{\partial^2 \ell}{\partial (\sigma^2)^2} \end{bmatrix}$$

donde:

$$\frac{\partial^2 \ell}{\partial \beta \, \partial \beta'} = -\frac{\boldsymbol{X}'\boldsymbol{X}}{\sigma^2}$$

$$\frac{\partial^2 \ell}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4}$$

$$\frac{\partial^2 \ell}{\partial \beta \, \partial \sigma^2} = -\frac{\boldsymbol{X}'\hat{\boldsymbol{\varepsilon}}}{\sigma^4}$$

7.2. Matriz de Información de Fisher

Definición: Matriz de Información $I(\theta)$

La matriz de información de Fisher se define como la esperanza de la Hessiana negativa de la log-verosimilitud:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta \, \partial \theta'}\right]$$

Para el MLG, usando que $\mathbb{E}[\hat{\boldsymbol{\varepsilon}}] = \mathbf{0}$ (el bloque cruzado se anula en esperanza):

$$I(\hat{\theta})^{-1} = \begin{bmatrix} \dfrac{\boldsymbol{X}'\boldsymbol{X}}{\sigma^2} & \mathbf{0} \\ \mathbf{0}' & \dfrac{n}{2\sigma^4} \end{bmatrix}^{-1} = \begin{bmatrix} \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} & \mathbf{0} \\ \mathbf{0}' & \dfrac{2\sigma^4}{n} \end{bmatrix}$$

Interpretación: Cota de Cramér-Rao

La **cota de Cramér-Rao** establece que para cualquier estimador insesgado de θ, su varianza no puede ser menor que $I(\theta)^{-1}$:

$$Var(\hat{\theta}) \geq I(\theta)^{-1}$$

El estimador MV *alcanza* esta cota asintóticamente: es el estimador de menor varianza posible entre todos los estimadores consistentes y asintóticamente normales. Para $\hat{\beta}_{MV}$, la cota es $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, que coincide exactamente con la varianza del estimador MCO bajo normalidad.

Interpretación: Bloque diagonal de la matriz de información

El hecho de que la matriz de información sea **diagonal por bloques** (ceros en los bloques cruzados entre β y σ^2) indica que las estimaciones de β y σ^2 son **asintóticamente independientes**: conocer mejor uno no te dice nada sobre el otro.

8. Inferencia con $\hat{\theta}_{MV}$

8.1. Contraste de hipótesis vía Razón de Verosimilitudes

Sea el modelo $Y = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}$ y la hipótesis nula $H_0 : \boldsymbol{R}\beta = \boldsymbol{r}$.

- **Modelo no restringido:** $\text{máx}_\theta\, \ell(\theta \mid \text{data}) \Rightarrow \hat{\theta}_{NR}$
- **Modelo restringido:** $\text{máx}_\theta\, \ell(\theta \mid \text{data})$ s.a. $\boldsymbol{R}\beta = \boldsymbol{r} \Rightarrow \hat{\theta}_R \sim \ell(\hat{\theta}_R)$

El **estadístico de razón de verosimilitudes** compara los logaritmos de verosimilitud:

$$\ell(\hat{\theta}_{NR}) - \ell(\hat{\theta}_R) \approx 0 \quad \text{si } H_0 \text{ es verdadera}$$

Interpretación: Lógica del contraste LR

Si H_0 es cierta, imponer la restricción *no debería costar mucho* en términos de verosimilitud: el modelo restringido debería ajustarse casi igual de bien que el no restringido. Si la diferencia $\ell(\hat{\theta}_{NR}) - \ell(\hat{\theta}_R)$ es grande, la restricción es "costosa" y probablemente falsa. Formalmente, el estadístico $-2[\ell(\hat{\theta}_R) - \ell(\hat{\theta}_{NR})] \sim \chi^2_q$, donde q es el número de restricciones impuestas.

9. Resumen Conceptual

Método	Idea central	Cuándo usarlo
MCO lineal	Minimizar $\sum \hat{\varepsilon}^2$ con $f = \boldsymbol{X}\beta$	Siempre que f sea lineal en β
Taylor (linealización)	Aproximar f localmente por un plano tangente e iterar	Primera aproximación cuando f es no lineal
MCNL	Minimizar $\sum[y_t - f(x_t, \hat{\beta})]^2$ directamente	f no lineal; solución numérica iterativa
MV	Maximizar $\ell(\theta \mid \boldsymbol{y}) = \sum \ln f(y_t \mid \theta)$	Cuando se asume distribución de los errores; bajo normalidad = MCNL

Equivalencia fundamental bajo normalidad:

$$\hat{\beta}_{MCO} = \hat{\beta}_{MV} = \hat{\beta}_{MCNL}$$

Los tres métodos coinciden cuando los errores son normales e independientes con varianza constante. Sus diferencias emergen al relajar estos supuestos.

10. Algoritmos de Optimización Numérica

Hasta ahora hemos establecido el problema de minimizar la suma de residuos cuadráticos (MCNL) o maximizar la log-verosimilitud (MV), y hemos visto cómo la linealización por serie de Taylor ofrece una primera aproximación. En la práctica, sin embargo, ninguno de estos problemas tiene solución analítica en forma cerrada cuando $f(x_t, \beta)$ es genuinamente no lineal. Por eso debemos resolverlos de forma *iterativa*: partimos de un punto inicial $\hat{\theta}_0$, aplicamos una regla de actualización y repetimos hasta que la secuencia $\{\hat{\theta}_n\}$ converja.

La elección del algoritmo importa: distintos métodos difieren en velocidad de convergencia, requisitos de memoria y robustez ante funciones objetivo con forma irregular. Los tres algoritmos más utilizados en econometría son Newton–Raphson, Gauss–Newton y el método de Scoring (Berndt–Hall–Hall–Hausman).

10.1. Gradiente y Hessiana: las piezas comunes

Cualquier algoritmo de segundo orden necesita dos ingredientes calculados en el punto actual $\hat{\theta}_n$:

leftmargin=*

- **Gradiente** $\nabla_\theta S(\hat{\theta}_n)$: vector de derivadas parciales de primer orden de la función objetivo. Para MCNL con $S(\theta) = \sum_t (y_t - f(x_t, \theta))^2$ resulta

$$\nabla_\theta S(\theta) = -2 \sum_{t=1}^{T} (y_t - f(x_t, \theta)) \, \frac{\partial f(x_t, \theta)}{\partial \theta} = -2 \sum_{t=1}^{T} \hat{u}_t \, Z_t,$$

  donde $Z_t = \nabla f(x_t, \theta)$ es la fila del Jacobiano correspondiente al período t.

- **Hessiana** $\nabla^2_\theta S(\hat{\theta}_n)$: matriz de derivadas de segundo orden. Para el mismo $S(\theta)$:

$$\nabla^2_\theta S(\theta) = 2 \sum_t Z_t \, Z_t' - 2 \sum_t \hat{u}_t \, H_t,$$

  donde $H_t = \nabla^2_\theta f(x_t, \theta)$ es la Hessiana de f respecto a θ en el período t.

Interpretación:

El gradiente indica la dirección de máximo crecimiento de $S(\theta)$; para minimizar, nos movemos en la dirección opuesta. La Hessiana cuantifica la curvatura local y permite escalar el paso de forma apropiada: si la función es muy "plana" en alguna dirección, la Hessiana lo detecta y el algoritmo da pasos más grandes en esa dirección.

10.2. Newton–Raphson

El método de Newton–Raphson obtiene la regla de actualización expandiendo $\nabla_\theta S(\theta)$ en serie de Taylor de *segundo* orden alrededor de $\hat{\theta}_n$ e igualando a cero:

$$\hat{\theta}_{n+1} = \hat{\theta}_n - \left[\nabla^2_\theta S(\hat{\theta}_n) \right]^{-1} \nabla_\theta S(\hat{\theta}_n). \tag{12}$$

El término $a = \nabla^2_\theta S(\hat{\theta}_n)$ es la Hessiana evaluada en la iteración actual, y el término $b = \nabla_\theta S(\hat{\theta}_n)$ es el gradiente. La actualización consiste entonces en restar al punto actual el producto $a^{-1} b$.

Interpretación:

Newton–Raphson usa información de *curvatura* (Hessiana) para escalar el paso. Si la función objetivo es cuadrática exacta, converge en *una sola iteración*; en casos generales, converge cuadráticamente cerca del óptimo (el error se eleva al cuadrado en cada paso). La desventaja es que requiere calcular e invertir la Hessiana en cada iteración, lo cual puede ser costoso, y que la Hessiana debe ser *definida positiva* (no singular) para que la dirección de actualización sea descendente. Si existen problemas de singularidad, el algoritmo puede fallar.

Para el modelo $y_t = \alpha\beta^{x_t} + u_t$ con $\theta = (\alpha, \beta)'$, el Jacobiano es:

$$Z_t = \nabla f(x_t, \theta) = (\beta^{x_t},\ \alpha x_t \beta^{x_t - 1}),$$

y la Hessiana de f respecto a θ es la matriz simétrica:

$$H_t = \nabla^2_\theta f(x_t, \theta) = \begin{pmatrix} 0 & x_t\beta^{x_t-1} \\ x_t\beta^{x_t-1} & \alpha x_t(x_t-1)\beta^{x_t-2} \end{pmatrix}.$$

Sustituyendo en (12) con valores iniciales $\alpha = 1$, $\beta = 2$, la primera actualización queda:

$$\begin{pmatrix} \hat\alpha_{n+1} \\ \hat\beta_{n+1} \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} - \left[2\sum_t \begin{pmatrix} 2^{2x_t} & x_t 2^{2x_t-1} \\ x_t 2^{2x_t-1} & x_t^2 2^{2x_t-2} \end{pmatrix} - 2\sum_t \hat u_t \begin{pmatrix} 0 & x_t 2^{x_t-1} \\ x_t 2^{x_t-1} & x_t(x_t-1)2^{x_t-2} \end{pmatrix} \right]^{-1}$$
$$\cdot\ 2\sum_t \begin{pmatrix} (y_t - 2^{x_t})\, 2^{x_t} \\ (y_t - 2^{x_t})\, x_t 2^{x_t-1} \end{pmatrix}$$

Regla práctica:

La Hessiana de $S(\theta)$ debe ser **definida positiva** para garantizar que la dirección de Newton sea efectivamente descendente. Si no lo es, el algoritmo puede divergir o quedar atrapado en un máximo local. En la práctica se suelen añadir modificaciones (por ejemplo, el método de Levenberg–Marquardt) para forzar la positividad.

10.3. Gauss–Newton

El método de Gauss–Newton es la versión simplificada de Newton–Raphson que se usa específicamente con MCNL. La clave es que, al expandir la Hessiana de $S(\theta)$:

$$\nabla^2_\theta S(\theta) = 2\sum_t Z_t Z_t' - 2\sum_t \hat u_t H_t,$$

el segundo término $-2\sum_t \hat u_t H_t$ involucra los residuos $\hat u_t$. Cerca del óptimo, cuando el modelo ajusta bien, estos residuos son pequeños y dicho término es despreciable. Gauss–Newton lo ignora deliberadamente, aproximando la Hessiana por $2\sum_t Z_t Z_t'$. La regla de actualización resultante es:

$$\hat\theta_{n+1} = \hat\theta_n - \Big[\sum_t Z_t\, Z_t'\Big]^{-1} \Big[\sum_t Z_t\, \hat u_t\Big]. \tag{13}$$

Interpretación:

Comparando (13) con la fórmula MCO estándar $\hat{\beta} = (X'X)^{-1}X'y$, se aprecia que Gauss–Newton es *exactamente* una regresión MCO de los residuos $\hat{u}_t$ sobre las columnas del Jacobiano Z_t. Esto no es accidental: recuérdese que en la linealización por serie de Taylor, ∇f juega el papel de la matriz de regresores X. Gauss–Newton explota esa analogía en cada iteración.

La ventaja sobre Newton–Raphson es doble: no requiere calcular la Hessiana de f (solo el Jacobiano), y la matriz $\sum Z_t Z_t'$ es siempre semidefinida positiva, lo que garantiza que la dirección de actualización sea siempre descendente.

10.4. Scoring (Berndt–Hall–Hall–Hausman, BHHH): versión para MV

Cuando el objetivo es maximizar la log-verosimilitud $\ell(\theta \mid y)$ en lugar de minimizar $S(\theta)$, la regla de Newton–Raphson se convierte en:

$$\hat{\theta}_{n+1} = \hat{\theta}_n - \Big[\sum_t \nabla^2_\theta \ell_t(\hat{\theta}_n)\Big]^{-1} \Big[\sum_t \nabla_\theta \ell_t(\hat{\theta}_n)\Big],$$

donde ℓ_t es la contribución de la observación t a la log-verosimilitud.

El **método de Scoring** (también conocido como BHHH por sus autores) reemplaza la Hessiana de la log-verosimilitud por su esperanza negativa, es decir, por la *Matriz de Información de Fisher* $\mathcal{I}(\theta)$:

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \Big[\mathcal{I}(\hat{\theta}_n)\Big]^{-1} \Big[\sum_t \nabla_\theta \ell_t(\hat{\theta}_n)\Big], \tag{14}$$

donde $\mathcal{I}(\theta_n) = -\mathbb{E}[\nabla^2_\theta \ell(\theta)]$. Dado que maximizamos (la función *crece* hacia el óptimo), el algoritmo *suma* en lugar de restar.

Interpretación:

En (14), el gradiente $\sum_t \nabla_\theta \ell_t$ es el **score** o función de puntuación, que en el óptimo no restringido vale cero. El término $[\mathcal{I}(\hat{\theta}_n)]^{-1}$ escala el paso según la información disponible en los datos: cuanta más información, menor el paso necesario para alcanzar el máximo. La relación entre los algoritmos y la Hessiana del modelo MV se puede expresar compactamente como:

$$H_{\text{MV}} = \frac{1}{2\hat{\sigma}^2} H_{\text{MCNL}},$$

lo que muestra que el método de Scoring es la contraparte de Gauss–Newton en el mundo de la máxima verosimilitud.

Aplicación al modelo $y_t = \alpha e^{\beta x_t} + u_t$. Para ilustrar cómo se construye el algoritmo en la práctica, consideremos este modelo con $\theta = (\alpha, \beta, \sigma_u^2)'$. La log-verosimilitud bajo normalidad de errores es:

$$\ell(\theta) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma_u^2) - \frac{1}{2\sigma_u^2}\sum_{t=1}^{T}(y_t - \alpha e^{\beta x_t})^2.$$

El gradiente respecto a $(\alpha, \beta, \sigma_u^2)$ es:

$$\frac{\partial \ell}{\partial \theta} = \begin{pmatrix} -\dfrac{1}{2\sigma_u^2}\displaystyle\sum_t (y_t - \alpha e^{\beta x_t})(-e^{\beta x_t}) \\ -\dfrac{1}{2\sigma_u^2}\displaystyle\sum_t (y_t - \alpha e^{\beta x_t})(-\alpha x_t e^{\beta x_t}) \\ -\dfrac{T}{2\sigma_u^2} + \dfrac{1}{2\sigma_u^4}\displaystyle\sum_t \hat{u}_t^2 \end{pmatrix}.$$

Igualando la tercera componente a cero se obtiene directamente:

$$\hat{\sigma}_\mu^2 = \frac{\sum_t \hat{u}_t^2}{T}.$$

Regla práctica:

El estimador MV de la varianza divide entre T (no entre $T-k$), por lo que es **sesgado en muestras finitas**, aunque consistente. La solución asintótica de este estimador requiere que los residuos sean evaluados en el $\hat{\theta}$ que maximiza la verosimilitud, lo que a su vez depende de $\hat{\sigma}^2$; de ahí la naturaleza iterativa del problema.

11. Inferencia con el Estimador de Máxima Verosimilitud

Una vez obtenido el estimador $\hat{\theta}_{\text{MV}}$, la pregunta natural es cómo contrastar hipótesis sobre los parámetros. En el marco de la máxima verosimilitud existen tres estadísticos de contraste clásicos que, bajo la hipótesis nula y condiciones de regularidad, siguen asintóticamente una distribución $\chi^2(q)$, donde q es el número de restricciones impuestas. Los tres son asintóticamente equivalentes bajo H_0, pero pueden diferir en muestras finitas.

Para unificar la exposición, consideremos el modelo general $y = X\beta + \varepsilon$ (o su versión no lineal) con $\hat{\theta} = (\beta', \sigma^2)'$, y la hipótesis nula lineal $H_0 : R\beta = r$, donde R es una matriz $q \times k$ de rango completo y r es un vector de constantes.

11.1. Test de Razón de Verosimilitudes (LR)

La idea central es comparar cuánto "pierde" la verosimilitud al imponer la restricción $H_0 : R\theta = r$. Se estiman dos modelos: el **no restringido** (NR), que maximiza $\ell(\theta \mid y)$ libremente obteniendo $\hat{\theta}^{NR}$, y el **restringido** (R), que maximiza sujeto a $R\theta = r$ obteniendo $\hat{\theta}^{R}$. Por construcción, $\ell^{NR} \geq \ell^{R}$.

El estadístico LR se construye a partir del ratio de verosimilitudes $\lambda = L^R / L^{NR}$, que satisface $0 < \lambda \leq 1$. Tomando logaritmos:

$$\text{LR} = -2\ln\lambda = -2\left[\ell(\hat{\theta}^R) - \ell(\hat{\theta}^{NR})\right] \xrightarrow{d} \chi^2(q). \tag{15}$$

Interpretación:

Si H_0 es verdadera, imponer la restricción no debería costar casi nada en términos de verosimilitud: $\ell^R \approx \ell^{NR}$ y, por tanto, LR ≈ 0. Si LR es grande, la restricción es "costosa" —los datos son muy poco verosímiles bajo H_0— y se rechaza la hipótesis nula. El factor -2 es una normalización que asegura la distribución χ^2 asintótica.

11.2. Test de Wald

El test de Wald no requiere estimar el modelo restringido. Parte del resultado asintótico: bajo condiciones de regularidad, el estimador MV satisface

$$\hat{\beta} \xrightarrow{d} \mathcal{N}(\beta, \mathcal{I}(\beta)^{-1}).$$

Aplicando la restricción lineal $R\hat{\beta} - r$, y usando que una transformación lineal de un vector normal es también normal:

$$R\hat{\beta} - r \xrightarrow{d} \mathcal{N}(0,\, R\mathcal{I}(\beta)^{-1}R').$$

Bajo $H_0 : R\beta = r$ (es decir, $R\beta - r = 0$), el estadístico de Wald es:

$$W = (R\hat{\beta} - r)'\left[R\mathcal{I}(\hat{\beta})^{-1}R'\right]^{-1}(R\hat{\beta} - r) \xrightarrow{d} \chi^2(q), \tag{16}$$

donde $\mathcal{I}(\hat{\beta})^{-1} = \hat{\sigma}^2(X'X)^{-1}$ en el caso del modelo lineal bajo normalidad.

Interpretación:

El estadístico de Wald mide cuán lejos está $R\hat{\beta}$ de r, normalizando por la varianza de esa distancia. Si la restricción es verdadera, $R\hat{\beta}$ debería estar próximo a r y W debería ser pequeño. Una forma útil de interpretar el denominador: $[R\mathcal{I}(\hat{\beta})^{-1}R']$ es la varianza asintótica de $R\hat{\beta}$, de modo que W es esencialmente la distancia al cuadrado medida en unidades de desviación estándar —equivalente a la noción de "cuántos errores estándar" separan al estimador de la restricción.

Ejemplo:

Para $H_0 : \lambda = 5$ en el modelo de Poisson del ejemplo anterior (donde $\hat{\lambda} = 2{,}8$), la varianza de $\hat{\lambda}$ es $\mathcal{I}(\lambda)^{-1} = \lambda^2/56$. El estadístico de Wald resulta:

$$W = (\hat{\lambda} - 5)\,\frac{56}{\hat{\lambda}^2}\,(\hat{\lambda} - 5) = (2{,}8 - 5)\,\frac{56}{2{,}8^2}\,(2{,}8 - 5).$$

11.3. Test del Multiplicador de Lagrange (LM)

El test LM (también llamado *test de Score*) parte del modelo *restringido*: se estima únicamente $\hat{\theta}^R$ y se evalúa el gradiente de la log-verosimilitud en ese punto. En el máximo no restringido, el gradiente es exactamente cero por definición; si la restricción es verdadera, el gradiente en $\hat{\theta}^R$ debería ser *suficientemente cercano* a cero. El vector score evaluado en el modelo restringido es:

$$s(\hat{\theta}^R) = \frac{\partial \ell(\hat{\theta}^R)}{\partial \theta} = \begin{pmatrix} \partial \ell/\partial \beta \\ \partial \ell/\partial \sigma^2 \end{pmatrix}\Bigg|_{\hat{\theta}^R},$$

y el estadístico es:

$$\text{LM} = s(\hat{\theta}^R)' \, \mathcal{I}(\hat{\theta}^R)^{-1} \, s(\hat{\theta}^R) \xrightarrow{d} \chi^2(q). \tag{17}$$

Interpretación:

El LM mide cuán lejos está el gradiente de la log-verosimilitud (evaluado en $\hat{\theta}^R$) de su valor óptimo cero. Si la restricción es verdadera, $\hat{\theta}^R$ debería estar cerca del máximo no restringido y el gradiente debería ser pequeño; si es falsa, el modelo restringido está "lejos" del óptimo y el gradiente es grande.
La gran ventaja práctica del LM es que solo requiere estimar el *modelo restringido*, que a menudo es más simple. Por ejemplo, si H_0 impone $\beta_2 = 0$, basta estimar el modelo sin x_{2t} y evaluar el score.

Los tres tests en perspectiva

La siguiente figura ilustra geométricamente la relación entre los tres estadísticos sobre la curva de log-verosimilitud:

- **LR**: diferencia vertical entre $\ell(\hat{\theta}^{NR})$ y $\ell(\hat{\theta}^R)$.
- **Wald**: distancia horizontal entre $\hat{\theta}^{NR}$ y $\hat{\theta}^R$ (cuán lejos está el estimador no restringido de cumplir la restricción).
- **LM**: pendiente de la log-verosimilitud evaluada en $\hat{\theta}^R$ (cuán "inclinada" está la curva en el punto restringido, lo que indica que aún hay margen de mejora).

En muestras grandes los tres tenderán al mismo valor bajo H_0. En muestras finitas, Wald tiende a sobre-rechazar, LM a sub-rechazar y LR queda en un punto intermedio (Novales, 1993, cap. 12).

Cuadro 1: Comparación de los tres tests clásicos de MV

	LR	**Wald**	**LM**
Modelos estimados	NR y R	Solo NR	Solo R
Insumo principal	$\ell^{NR} - \ell^R$	$R\hat{\beta}^{NR} - r$	$\nabla_\theta \ell(\hat{\theta}^R)$
Distribución asintótica	$\chi^2(q)$	$\chi^2(q)$	$\chi^2(q)$
Ventaja práctica	Intuitivo	No requiere R	No requiere NR

12. Contraste de Restricciones No Lineales

En los modelos no lineales, la hipótesis nula no siempre adopta la forma lineal $R\beta = r$. Frecuentemente las restricciones son *no lineales en los parámetros*, del tipo $H_0 : R(\beta) = r$, donde $R(\cdot)$ es una función vectorial no lineal. Por ejemplo:

$$H_0 : \frac{\beta_2}{1 - \beta_3} = 1 \qquad \text{o} \qquad H_0 : \beta_1 \times \beta_2 = 1.$$

El estadístico de Wald se generaliza a este caso mediante una aproximación de Taylor de $R(\hat{\beta})$ alrededor del verdadero β:

$$W = [R(\hat{\beta}) - r]' \left[\frac{\partial R}{\partial \beta} \widehat{\mathrm{Var}}(\hat{\beta}) \left(\frac{\partial R}{\partial \beta} \right)' \right]^{-1} [R(\hat{\beta}) - r] \xrightarrow{d} \chi^2(q), \tag{18}$$

donde $\partial R/\partial \beta$ es la matriz Jacobiana de R evaluada en $\hat{\beta}$, y $\widehat{\mathrm{Var}}(\hat{\beta})$ es el estimador de la varianza del estimador (por ejemplo, $\hat{\sigma}^2(X'X)^{-1}$ o $\mathcal{I}(\hat{\beta})^{-1}$).

Interpretación:

La lógica es la misma que en el caso lineal: medimos cuán lejos está $R(\hat{\beta})$ de r, escalando por la varianza de esa distancia. La diferencia es que ahora $\widehat{\mathrm{Var}}(R(\hat{\beta}))$ no es simplemente $R\,\mathrm{Var}(\hat{\beta})\,R'$, sino que involucra el Jacobiano $\partial R/\partial \beta$ porque la restricción es no lineal. Esta es la generalización del *método delta*: la varianza de una función no lineal de un estimador se aproxima por la varianza del estimador multiplicada por la derivada al cuadrado.

12.1. Ejemplo: restricción cociente

Considérese el modelo:

$$y_t = \beta_1 + \beta_2\, y_{t-1} + \beta_3\, C_{t-1} + u_t,$$

y la hipótesis $H_0 : \dfrac{\beta_2}{1-\beta_3} = 1$, equivalente a $R(\beta) = \dfrac{\beta_2}{1-\beta_3}$ y $r = 1$.

El Jacobiano de R respecto a $\beta = (\beta_1, \beta_2, \beta_3)'$ es:

$$\frac{\partial R}{\partial \beta} = \begin{pmatrix} \dfrac{\partial R}{\partial \beta_1} \\ \dfrac{\partial R}{\partial \beta_2} \\ \dfrac{\partial R}{\partial \beta_3} \end{pmatrix} = \begin{pmatrix} 0 \\ \dfrac{1}{1-\beta_3} \\ \dfrac{\beta_2}{(1-\beta_3)^2} \end{pmatrix}.$$

Evaluando en $\hat{\beta}$, el estadístico de Wald resulta:

$$W = \left[\frac{\hat{\beta}_2}{1-\hat{\beta}_3} - 1 \right] \left[\left.\frac{\partial R}{\partial \beta}\right|_{\hat{\beta}} \widehat{\mathrm{Var}}(\hat{\beta}) \left(\left.\frac{\partial R}{\partial \beta}\right|_{\hat{\beta}} \right)' \right]^{-1} \left[\frac{\hat{\beta}_2}{1-\hat{\beta}_3} - 1 \right],$$

donde $\widehat{\mathrm{Var}}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$.

12.2. Ejemplo: restricción producto

Para el modelo $y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t$ con $H_0 : \beta_1 \times \beta_2 = 1$, la función de restricción es $R(\beta_1, \beta_2) = \beta_1\beta_2$ y $r = 1$. Su Jacobiano es:

$$\frac{\partial R}{\partial \beta} = \begin{pmatrix} 0 \\ \beta_1 \\ \beta_2 \end{pmatrix},$$

evaluado en $(\hat{\beta}_1, \hat{\beta}_2)$. Esto es un caso típico de *restricción bilineal*, que no puede expresarse como $R\beta = r$ para ninguna matriz R fija. La aplicación directa de (18) permite, no obstante, construir el estadístico de Wald con la fórmula del método delta.

Regla práctica:

Para cualquier restricción no lineal $H_0 : R(\beta) = r$, el procedimiento es siempre el mismo: (i) estimar $\hat{\beta}$ sin restricciones, (ii) evaluar $R(\hat{\beta}) - r$, (iii) calcular el Jacobiano $\partial R/\partial\beta$ en $\hat{\beta}$, y (iv) aplicar la fórmula (18). El estadístico resultante sigue asintóticamente una $\chi^2(q)$ bajo H_0, donde q es el número de restricciones escalares en $R(\beta) = r$.

13. Matriz de Información y Varianza del Estimador MV No Lineal

Para el modelo no lineal general $y_t = f(x_t, \beta) + u_t$ con $u_t \sim \mathcal{N}(0, \sigma_u^2)$, el vector de parámetros completo es $\theta = (\beta', \sigma_u^2)'$. La Matriz de Información de Fisher toma la forma por bloques:

$$\mathcal{I}(\beta, \sigma_u^2) = \begin{pmatrix} \dfrac{1}{\sigma_u^2}\left[\left(\dfrac{\partial f}{\partial \beta}\right)'\left(\dfrac{\partial f}{\partial \beta}\right)\right] & \mathbf{0}_k \\ \mathbf{0}_k' & \dfrac{T}{2\sigma_u^4} \end{pmatrix}, \tag{19}$$

y su inversa (la cota de Cramér–Rao, que coincide con la varianza asintótica del estimador MV) es:

$$\mathcal{I}(\beta, \sigma_u^2)^{-1} = \begin{pmatrix} \sigma_u^2\left[\left(\dfrac{\partial f}{\partial \beta}\right)'\left(\dfrac{\partial f}{\partial \beta}\right)\right]^{-1} & \mathbf{0}_k \\ \mathbf{0}_k' & \dfrac{2\sigma_u^4}{T} \end{pmatrix}. \tag{20}$$

Interpretación:

La estructura por bloques de (19) refleja que, asintóticamente, las estimaciones de β y de σ_u^2 son **independientes**: conocer mejor uno de ellos no aporta información sobre el otro. El bloque superior izquierdo de (20) muestra que la varianza de $\hat{\beta}$ tiene exactamente la misma forma que en el modelo lineal —$\sigma^2(X'X)^{-1}$— pero con el Jacobiano $\partial f/\partial\beta$ reemplazando a X. Esto es consistente con la analogía entre Jacobiano y matriz de regresores que hemos destacado a lo largo de todo el manual.

Ejemplo:

Para $y_t = \alpha e^{\beta x_t} + u_t$, el Jacobiano es $(\partial f/\partial\theta) = (e^{\beta x_t},\ \alpha x_t e^{\beta x_t})'$ en cada t, y el producto externo es:

$$\left(\frac{\partial f_t}{\partial \theta}\right)\left(\frac{\partial f_t}{\partial \theta}\right)' = \begin{pmatrix} e^{2\beta x_t} & \alpha x_t e^{2\beta x_t} \\ \alpha x_t e^{2\beta x_t} & \alpha^2 x_t^2 e^{2\beta x_t} \end{pmatrix}.$$

Sumando sobre $t = 1, \ldots, T$ y sustituyendo en (20) se obtiene la varianza asintótica completa de $(\hat{\alpha}, \hat{\beta})'$.

Resumen de los Algoritmos y Tests

Cuadro 2: Algoritmos numéricos para modelos no lineales

Algoritmo	**Regla de actualización**	**Cuándo usarlo**
Newton–Raphson	$\hat{\theta}_{n+1} = \hat{\theta}_n - [\nabla^2 S]^{-1} \nabla S$	Convergencia rápida cerca del óptimo; requiere Hessiana definida positiva
Gauss–Newton	$\hat{\theta}_{n+1} = \hat{\theta}_n - [\sum Z_t Z_t']^{-1} [\sum Z_t \hat{u}_t]$	MCNL; más robusto; solo necesita Jacobiano
Scoring (BHHH)	$\hat{\theta}_{n+1} = \hat{\theta}_n + [\mathcal{I}(\hat{\theta}_n)]^{-1} \nabla \ell$	Máxima verosimilitud; versión MV de Gauss–Newton

Cuadro 3: Tests clásicos bajo máxima verosimilitud ($H_0 : R(\beta) = r$)

Test	**Estadístico**	**Qué mide**
Razón de Verosimilitud (LR)	$-2[\ell(\hat{\theta}^R) - \ell(\hat{\theta}^{NR})] \sim \chi^2(q)$	Caída en verosimilitud al restringir
Wald (W)	$[R(\hat{\beta}) - r]' [\widehat{\text{Var}}(R(\hat{\beta}))]^{-1} [R(\hat{\beta}) - r] \sim \chi^2(q)$	Distancia del estimador a la restricción
Multiplicador de Lagrange (LM)	$s(\hat{\theta}^R)' \mathcal{I}(\hat{\theta}^R)^{-1} s(\hat{\theta}^R) \sim \chi^2(q)$	Pendiente de la log-verosimilitud en $\hat{\theta}^R$

Referencias

[1] Greene, W. H. (2018). *Econometric Analysis* (8th ed.). Pearson.

[2] Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2nd ed.). MIT Press.

[3] Davidson, R., & MacKinnon, J. G. (2004). *Econometric Theory and Methods*. Oxford University Press.

[4] Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

[5] Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

[6] Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B*, 26(2), 211-252.

[7] Novales, A. (1993). *Econometría*. McGraw-Hill. Capítulos 11 (Modelos no lineales) y 12 (Algoritmos Numéricos de Optimización)