Métodos de estimación

Randall Romero Aguilar, PhD
randall.romero@ucr.ac.cr

EC4300 - Microeconometría

II Semestre 2023

Última actualización: 13 de agosto de 2023

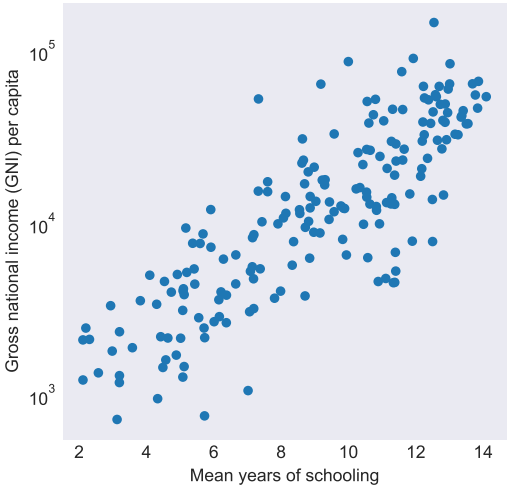
UCR
UNIVERSIDAD DE COSTA RICA

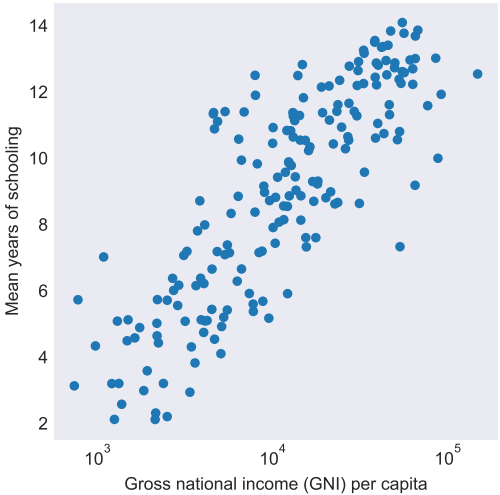
ESCUELA de ECONOMÍA
UNIVERSIDAD DE COSTA RICA

Tabla de contenidos

1. Motivación

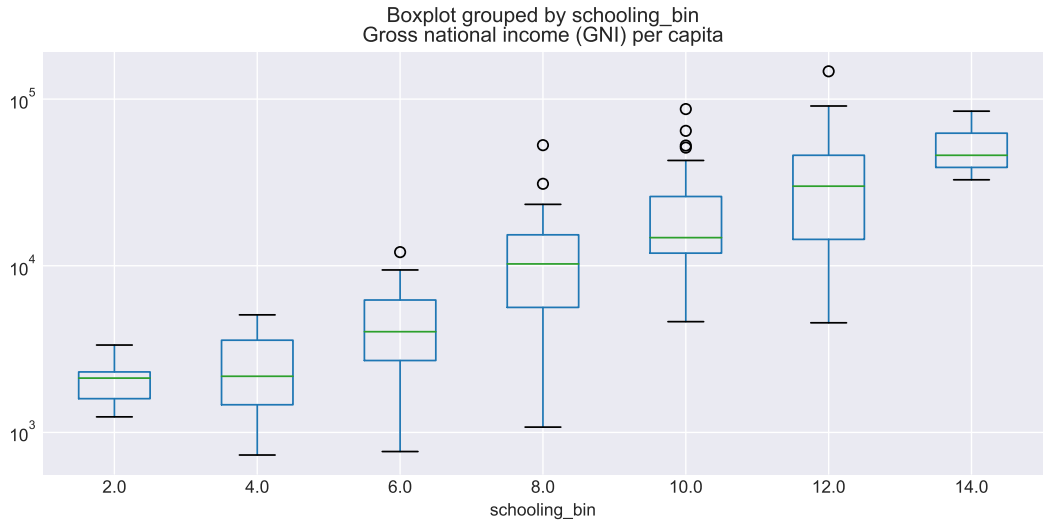
Correlación versus causalidad





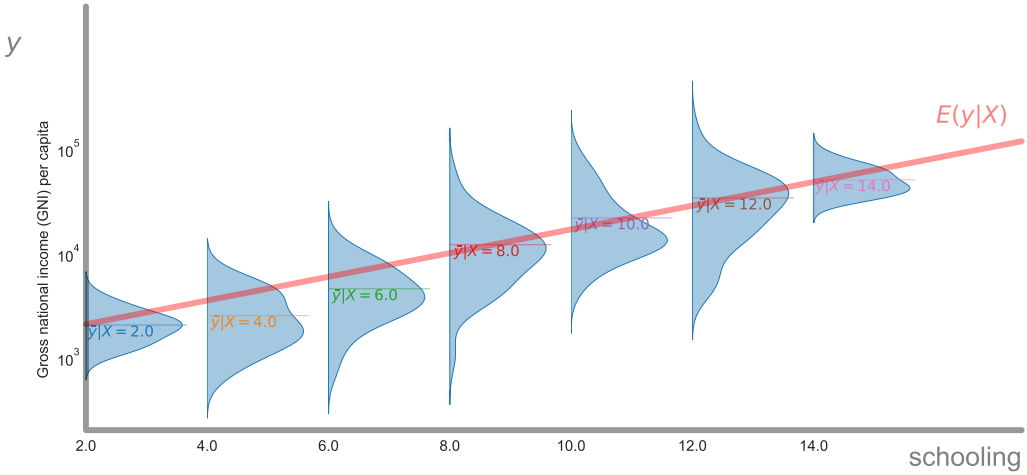
Fuente: Elaborado por el autor con datos de https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

Distribuciones condicionales, intento 1: boxplot



Fuente: Elaborado por el autor con datos de https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

Distribuciones condicionales, intento 2: kernel



Fuente: Elaborado por el autor con datos de https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

2. Repaso del Modelo Clásico de Regresión Lineal

El Modelo Clásico de Regresión Lineal

- El MCRL es la herramienta más útil en econometría.
- A pesar de que en la literatura contemporánea es sólo el punto de partida del análisis completo, sigue siendo una herramienta usada para empezar casi todos los trabajos empíricos.
- Se usa para estudiar la relación entre una variable dependiente y una o más variables independientes.

$$
\begin{aligned}
y_i &= \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + \varepsilon_i, \qquad \forall i = 1, 2, \ldots, n \\
&= \underbrace{\begin{bmatrix} x_{1i} & \ldots & x_{Ki} \end{bmatrix}'}_{x_i'} \underbrace{\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}}_{\beta} + \varepsilon_i \\
&= x_i' \beta + \varepsilon_i
\end{aligned}
$$

Agrupando todos los datos

- Si escribimos todas las observaciones como vectores columna:

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{Y}} = \begin{bmatrix} x_1'\beta + \varepsilon_1 \\ \vdots \\ x_n'\beta + \varepsilon_n \end{bmatrix} = \begin{bmatrix} x_1'\beta \\ \vdots \\ x_n'\beta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \underbrace{\begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}}_{\mathbf{X}} \beta + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}}_{\varepsilon}$$

$$\mathbf{Y} = \mathbf{X}\,\beta + \varepsilon$$

- Tenemos que
  - $\mathbf{Y}$ es un vector columna con las n observaciones de la variable dependiente.
  - $\mathbf{X}$ es una matriz con n filas (observaciones) y K columnas (variables independientes).
  - ε es un vector columna con las n perturbaciones o errores.

Supuestos del MCRL

El modelo es lineal en sus parámetros β:

$$\mathbf{Y} = \mathbf{X}\,\beta + \varepsilon \tag{A1}$$

No hay multicolinealidad: ningún regresor es una combinación lineal de los demás

$$\text{rango}\,(\mathbf{X}_{n\times K}) = K \tag{A2}$$

La esperanza del error, condicional en las variables explicativas, es cero:

$$\mathbb{E}[\varepsilon \,|\, \mathbf{X}] = 0 \tag{A3}$$

No hay autocorrelación ni heteroscedasticidad:

$$\mathbb{V}[\varepsilon \,|\, \mathbf{X}] = \mathbb{E}\left[\varepsilon\,\varepsilon' \,|\, \mathbf{X}\right] = \sigma^2\mathrm{I} \tag{A4}$$

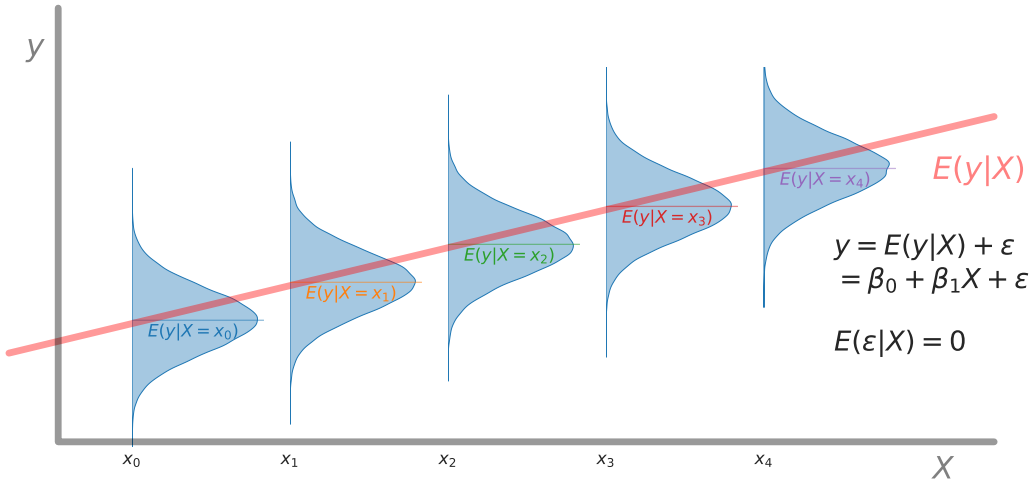
No hay sesgo de simultaneidad:

$$\mathbf{X} \text{ es independiente de } \varepsilon \tag{A5}$$

Condicional en las variables explicativas, el error tiene distribución normal:

$$\varepsilon \,|\, \mathbf{X} \sim \mathrm{N}\left(0, \sigma^2\mathrm{I}\right) \tag{A6}$$

Regresión lineal = modelo de la esperanza condicional



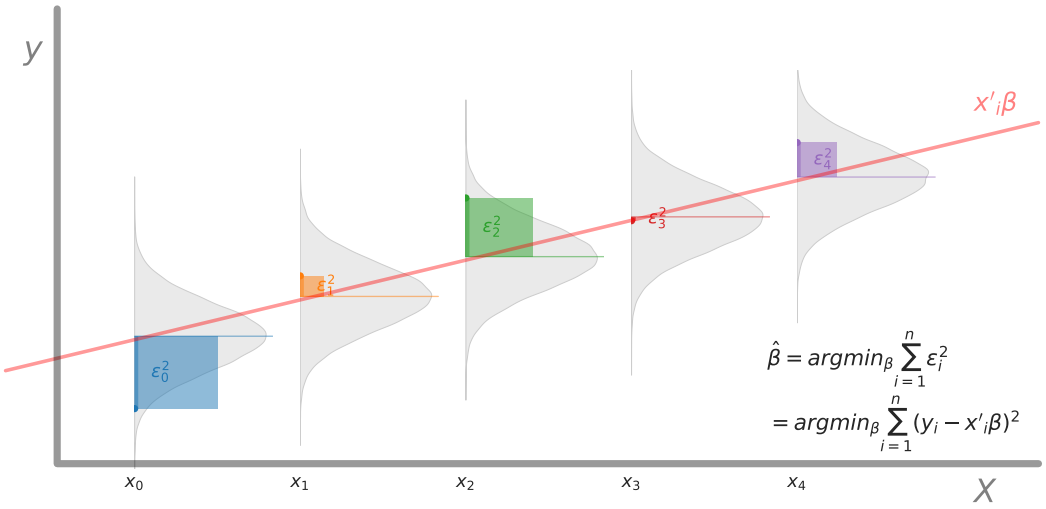
3. Mínimos Cuadrados Ordinarios

Estimación por mínimos cuadrados ordinarios: Definiendo el problema

- **Estrategia**: Se escoge aquel valor β que minimice la suma de los cuadrados de los residuos entre los $\mathbf{Y}$ observados y los ajustados por regresión.

$$
\begin{aligned}
\hat{\beta}^{\text{OLS}} &= \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i' \beta\right)^2 \\
&= \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \varepsilon_i^2 \\
&= \underset{\beta}{\operatorname{argmin}} \left[\varepsilon' \varepsilon\right] \\
&= \underset{\beta}{\operatorname{argmin}} \left[(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)\right] \\
&= \underset{\beta}{\operatorname{argmin}} (\mathbf{Y}'\mathbf{Y} - 2\,\mathbf{Y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta)
\end{aligned}
$$

MCO: ilustración del método



$$\hat{\beta} = argmin_\beta \sum_{i=1}^{n} \varepsilon_i^2$$

$$= argmin_\beta \sum_{i=1}^{n} (y_i - x'_i\beta)^2$$

Nota:

Derivadas con matrices

- Si tenemos una función $f : \mathbb{R}^n \to \mathbb{R}$, entonces la derivada de f con respecto a un vector columna $x \in \mathbb{R}^n$ es simplemente el vector columna de las derivadas parciales de f con respecto a cada elemento de x:

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

- Suponga que tenemos una matriz $A \in \mathbb{R}^{n \times n}$, y un vector $b \in \mathbb{R}^n$. Entonces:
  - La derivada de $b'x$ con respecto a x es

$$\frac{\partial b'x}{\partial x} = b$$

  - La derivada de $x'Ax$ con respecto a x es

$$\frac{\partial x'Ax}{\partial x} = (A + A')x$$

  - Si A es simétrica, entonces la derivada de $x'Ax$ con respecto a x es $2Ax$.

Estimación por mínimos cuadrados ordinarios: Resolviendo el problema

- Tomando condición de primer orden:

$$\begin{aligned}\frac{\partial}{\partial \beta}(\mathbf{Y}'\mathbf{Y} - 2\,\mathbf{Y}'\mathbf{X}\,\beta + \beta'\mathbf{X}'\mathbf{X}\,\beta) &= 0 \\ -2\,\mathbf{X}'\mathbf{Y} + 2\,\mathbf{X}'\mathbf{X}\,\beta &= 0 \\ \mathbf{X}'\mathbf{X}\,\beta &= \mathbf{X}'\mathbf{Y}\end{aligned}$$

- Si se cumple (A2), de manera que $\mathbf{X}'\mathbf{X}$ sea invertible:

$$\hat{\beta}^{\text{OLS}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

Interpretando la fórmula de MCO

- Recordemos que interpretar a $\mathbf{Y}$ y $\mathbf{X}$ como dos vectores de observaciones.
- Entonces

$$\hat{\beta}^{\text{OLS}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} = \left(\begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}\begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}\right)^{-1}\begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$= \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1}\sum_{i=1}^{n} x_i y_i = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right)$$

- Es decir, podemos pensar en el estimador como la correlación de $\mathbf{X}$ con $\mathbf{Y}$ "dividido" entre la covarianza de $\mathbf{X}$.

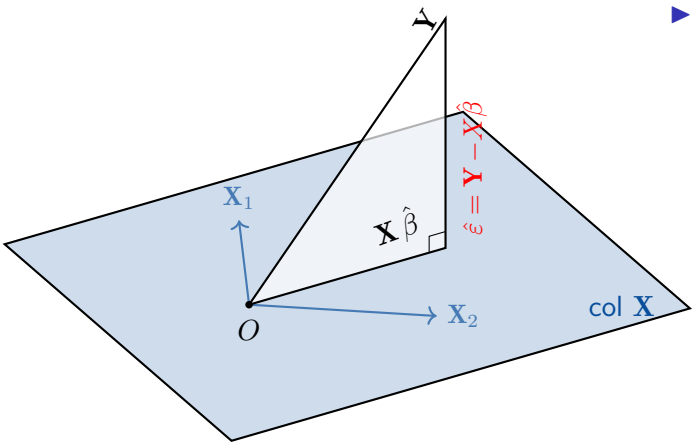
El modelo lineal visto como álgebra vectorial

- Podemos ver el término $\mathbf{X}\,\beta$ como la combinación lineal de las columnas de $\mathbf{X}$:

$$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1K} \\ x_{21} & x_{22} & \dots & x_{2K} \\ x_{31} & x_{32} & \dots & x_{3K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nK} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K \in \mathbb{R}^n$$

- Dado que $K < n$, no podemos generar todo el espacio $\mathbb{R}^n$ a partir de las columnas de $\mathbf{X}$.
- Por ello probablemente será imposible encontrar un β tal que $\mathbf{Y} = \mathbf{X}\,\beta$.
- Busquemos entonces el valor de β que haga $\mathbf{X}\,\beta$ lo más cercano posible a Y.
- Es decir, queremos encontrar el vector $\varepsilon = \mathbf{Y} - \mathbf{X}\,\beta$ más pequeño posible.

Mínimos cuadrados ordinarios, versión 2



- El valor ε más pequeño se obtiene escogiendo β tal que ε sea perpendicular al espacio generado por las columnas de $\mathbf{X}$, es decir cuando:

$$\mathbf{X}'\varepsilon = 0$$
$$\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta) = 0$$
$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y}$$
$$\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Aspectos algebraicos

- Los valores ajustados son

$$\hat{\mathbf{Y}} \equiv \mathbf{X}\,\hat{\beta}^{\text{OLS}} = \mathbf{X}\left(\mathbf{X}'\,\mathbf{X}\right)^{-1}\mathbf{X}'\,\mathbf{Y} = P\,\mathbf{Y}$$

- Los residuos son

$$\hat{\varepsilon} \equiv \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - P\,\mathbf{Y} = (I - P)\,\mathbf{Y} = M\,\mathbf{Y}$$

- Notemos que las matrices P y M son simétricas e idempotentes:

$$P = P^2 = P' \qquad\qquad M = M^2 = M'$$

- El rango de estas matrices es

$$\text{rango}(P) = K \qquad\qquad \text{rango}(M) = n - K$$

- Finalmente, notemos estos resultados

$$P\,\mathbf{X} = \mathbf{X} \qquad\qquad M\,\mathbf{X} = 0 \qquad\qquad PM = 0$$

Un caso particular: una regresión solo con una constante

- Supongamos que tenemos una regresión con una constante y sin ninguna otra variable explicativa: $\mathbf{X} = \mathbf{i}$, una columna de unos.
- Entonces

$$\hat{\mathbf{Y}} = PY = \mathbf{i}\left(\underbrace{\mathbf{i}'\mathbf{i}}_{n}\right)^{-1}\underbrace{\mathbf{i}'\,\mathbf{Y}}_{\sum y_i} = \mathbf{i}\frac{1}{n}\sum_{i=1}^{n} y_i = \bar{\mathbf{Y}}\mathbf{i}$$

  es decir, el valor ajustado es el promedio de los valores de Y.
- Por otro lado, los residuos son

$$M_0\,\mathbf{Y} = \left(I - \mathbf{i}\left(\mathbf{i}'\mathbf{i}\right)^{-1}\mathbf{i}'\right)\mathbf{Y}$$

  simplemente la desviación de cada observación respecto al promedio.
- Así, la suma de los cuadrados de las desviaciones respecto a la media se puede escribir como:

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \mathbf{Y}'\,M_0\,\mathbf{Y}$$

Bondad de ajuste

- Dado que $\mathbf{Y} = X\hat{\beta} + \hat{\varepsilon}$, y que $M_0\hat{\varepsilon} = \hat{\varepsilon}$ (asumiendo que hay intercepto) y $\hat{\varepsilon}'\mathbf{X} = 0$, tenemos que la varianza de Y podemos descomponerla como

$$\begin{aligned}\mathbf{Y}'M_0\mathbf{Y} &= \left(X\hat{\beta} + \hat{\varepsilon}\right)' M_0 \left(X\hat{\beta} + \hat{\varepsilon}\right) \\ &= \underbrace{\left(X\hat{\beta}\right)' M_0 \left(X\hat{\beta}\right)}_{\text{varianza de regresión}} + \underbrace{\hat{\varepsilon}'\hat{\varepsilon}}_{\text{varianza del residuo}}\end{aligned}$$

- La bondad de ajuste de un modelo se mide entonces con el coeficiente de determinación R^2: la proporción de la varianza de Y que es explicada por el modelo:

$$R^2 = \frac{\hat{\beta}'X'M^0X\hat{\beta}}{Y'M^0Y} = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{Y'M^0Y}$$

Estimación de la varianza

- La varianza del error σ^2 es desconocida, pero podemos estimarla con

$$s^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-K} \qquad \text{o bien con} \qquad \sigma^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}$$

- Dado que $\hat{\beta}^{\text{OLS}} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$, su varianza es

$$\begin{aligned}
\mathbb{V}\left[\hat{\beta}^{\text{OLS}}|\,\mathbf{X}\right] &= \mathbb{E}\left[\left(\hat{\beta}^{\text{OLS}} - \beta\right)\left(\hat{\beta}^{\text{OLS}} - \beta\right)'|\,\mathbf{X}\right] \\
&= \mathbb{E}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}|\,\mathbf{X}\right] \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbb{E}\left[\varepsilon\varepsilon'|\,\mathbf{X}\right]\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\sigma^2 I\,\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \qquad = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}
\end{aligned}$$

- Con lo cual podemos estimar la varianza de $\hat{\beta}^{\text{OLS}}$ con

$$\widehat{\mathbb{V}}\left[\hat{\beta}^{\text{OLS}}|\,\mathbf{X}\right] = s^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

Propiedades de muestra finita (1)

Resultados generales

El estimador MCO es insesgado $>>>$ su esperanza es β:

$$E\left[\hat{\beta}^{\text{OLS}} \mid \mathbf{X}\right] = E\left[\hat{\beta}^{\text{OLS}}\right] = \beta \tag{FS1}$$

La varianza del estimador es

$$\mathbb{V}\left[\hat{\beta}^{\text{OLS}} \mid \mathbf{X}\right] = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} \tag{FS2}$$

Es el mejor estimador lineal insesgado:

$$\min\left\{\mathbb{V}\left(w'\tilde{\beta}\right)\right\} = \mathbb{V}\left(w'\hat{\beta}^{\text{OLS}}\right) \qquad \forall \text{ estimador lineal insesgado } \tilde{\beta} \tag{FS3}$$

s^2 es un estimador insesgado de la varianza del error:

$$E\left[s^2 \mid \mathbf{X}\right] = E\left[s^2\right] = \sigma^2 \tag{FS4}$$

No hay correlación entre el estimador y los residuos:

$$\text{Cov}\left[\hat{\beta}^{\text{OLS}}, \hat{\varepsilon} \mid \mathbf{X}\right] = 0 \tag{FS5}$$

Propiedades de muestra finita (2)

Resultados que siguen al supuesto (A6)

Independencia de las estimaciones de la pendiente y la varianza:

$$\hat{\beta}^{\text{OLS}} \text{ y } \hat{\varepsilon} \text{ son independientes } \Rightarrow \hat{\beta}^{\text{OLS}} \text{ y } s^2 \text{ también} \tag{FS6}$$

Distribución del estimador de las pendientes:

$$\hat{\beta}^{\text{OLS}} \mid X \sim N\left[\beta, \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right] \tag{FS7}$$

Distribución del estimador de la varianza:

$$(n-K)\frac{s^2}{\sigma^2} \sim \chi^2(n-K) \tag{FS8}$$

Propiedades de muestra finita (3)

Resultados que siguen al supuesto (A6)

Distribución para prueba de hipótesis de una pendiente:

$$t_{n-K} = \frac{\hat{\beta}_k^{\text{OLS}} - \beta_k}{\sqrt{s^2 \left(\mathbf{X}' \mathbf{X}\right)^{-1}_{kk}}} \sim t[n-K] \tag{FS9}$$

Distribución para prueba de hipótesis de todas las pendientes:

$$F_{K-1,n-K} = \frac{(n-K)R^2}{(K-1)\left(1-R^2\right)} \sim F[K-1, n-K] \tag{FS10}$$

Test de Wald

- Para contrastar las hipótesis

$$H_0 : R\beta = q$$
$$H_1 : R\beta \neq q$$

- utilizamos el estadístico

$$F = \frac{1}{J}(R\hat{\beta} - q)' \left\{ R \left[s^2 \left(\mathbf{X}' \mathbf{X} \right)^{-1} \right] R' \right\}^{-1} (R\hat{\beta} - q)$$
$$= \frac{\frac{R^2 - R_*^2}{J}}{\frac{1-R^2}{n-K}} \sim F[J, n-K]$$

4. Máxima Verosimilitud

Introducción al estimador de máxima verosimilitud

- Sea $f(y|\theta)$ la función de densidad conjunta de la variable $\mathbf{Y} = [Y_1, \ldots, Y_n]$. Entonces, para una muestra observada $\mathbf{y}$ de esta distribución, la función del vector de parámetros θ definida por

$$\mathcal{L}(\theta|\,\mathbf{y}) = f(\mathbf{y}\,|\theta)$$

  se conoce como la función de verosimilitud.

- El estimador de máxima verosimilitud es el valor del vector de parámetros θ que maximiza la función de verosimilitud

$$\hat{\theta}_{\text{ML}} \equiv \underset{\theta}{\operatorname{argmax}} \mathcal{L}(\theta|\,\mathbf{y}) = \underset{\theta}{\operatorname{argmax}} \ln \mathcal{L}(\theta|\,\mathbf{y})$$

- Es decir, $\hat{\theta}_{\text{ML}}$ es el parámetro que hace más plausible haber obtenido la muestra $\mathbf{y}$ si la verdadera distribución era $f(y|\theta)$.

Obteniendo el estimador

- Asumiendo que la función $\mathcal{L}$ es diferenciable respecto a sus parámetros, podemos encontrar su máximo igualando su gradiente a cero:

$$s(\theta) \equiv \frac{\partial \ln \mathcal{L}(\theta|\,\mathbf{y})}{\partial \theta} = \frac{1}{f(\mathbf{y}\,|\theta)} \frac{\partial \mathcal{L}(\theta|\,\mathbf{y})}{\partial \theta} = 0$$

- La función $s(\theta)$ es conocida como el score.
- Notemos que $s(\theta)$ depende de los datos, por lo que (antes de observar la muestra) es una variable aleatoria.
- Si se cumplen ciertas condiciones de regularidad respecto a f, se tiene que

$$\mathbb{E}\left[s(\theta)\right] = 0 \qquad\qquad \mathbb{V}\left[s(\theta)\right] = -\,\mathbb{E}\left(\frac{\partial^2 \ln \mathcal{L}}{\partial \theta \partial \theta'}\right) \equiv \mathcal{I}(\theta)$$

- La matriz $\mathcal{I}(\theta)$ se conoce como información de Fisher.

Nota:

Regla de Leibniz de la integral

- Consideremos la integral

$$\int_{a(t)}^{b(t)} f(x,t)\ \mathrm{d}x$$

- Como la integral es con respecto a x, vemos que el resultado final depende de t.
- Bajo ciertas condiciones,

La regla de Leibniz

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{a(t)}^{b(t)} f(x,t)\ \mathrm{d}x = \int_{a(t)}^{b(t)} \frac{\partial f(x,t)}{\partial t}dx + f[b(t),t]b'(t) - f[a(t),t]a'(t)$$

- Si los límites de integración no dependen de t la fórmula es muy sencilla

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{a(t)}^{b(t)} f(x,t)\ \mathrm{d}x = \int_{a(t)}^{b(t)} \frac{\partial f(x,t)}{\partial t}dx$$

Demostración informal: Asumiendo que $F(x,t)$ es una antiderivada de $f(x,t)$ respecto a x:

$$\int_{a(t)}^{b(t)} f(x,t)\ \mathrm{d}x = F(x,t)|_{x=a(t)}^{x=b(t)} = F[b(t),t] - F[a(t),t]$$

derivamos ambos lados respecto a t

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\int_{a(t)}^{b(t)} f(x,t)\ \mathrm{d}x &= \frac{\mathrm{d}}{\mathrm{d}t}\left(F[b(t),t] - F[a(t),t]\right) \\
&= \frac{\partial F[b(t),t]}{\partial x}b'(t) + \frac{\partial F[b(t),t]}{\partial t} - \frac{\partial F[a(t),t]}{\partial x}a'(t) - \frac{\partial F[a(t),t]}{\partial t} \\
&= \left.\frac{\partial F(x,t)}{\partial t}\right|_{x=a(t)}^{x=b(t)} + \frac{\partial F[b(t),t]}{\partial x}b'(t) - \frac{\partial F[a(t),t]}{\partial x}a'(t) \\
&= \int_{a(t)}^{b(t)} \frac{\partial}{\partial t}f(x,t)\ \mathrm{d}x + f[b(t),t]b'(t) - f[a(t),t]a'(t)
\end{aligned}$$

Demostración de $\mathbb{E}\left[s(\theta)\right] = 0$

$$1 = \int_{a(\theta)}^{b(\theta)} f(y|\theta)\ \mathrm{d}y \qquad \text{(por definión de } f\text{). Derivamos respecto a } \theta$$

$$0 = \int_{a(\theta)}^{b(\theta)} \frac{\partial f(y|\theta)}{\partial \theta}\ \mathrm{d}y + f[b(\theta)|\theta]b'(\theta) - f[a(\theta)|\theta]a'(\theta)$$

Asumiendo que los últimos dos términos son cero:

$$\begin{aligned} 0 &= \int_{a(\theta)}^{b(\theta)} \frac{\partial f(y|\theta)}{\partial \theta}\ \mathrm{d}y \\ &= \int_{a(\theta)}^{b(\theta)} \frac{\partial \ln f(y|\theta)}{\partial \theta} f(y|\theta)\ \mathrm{d}y \\ &= \int_{a(\theta)}^{b(\theta)} s(\theta) f(y|\theta)\ \mathrm{d}y = \mathbb{E}[s(\theta)] \end{aligned}$$

Propiedades asintóticas

Si se cumplen ciertas condiciones de regularidad,

El estimador ML es consistente:

$$\operatorname{plim} \hat{\theta}^{\mathsf{ML}} = \theta_0 \tag{M1}$$

Es asintóticamente normal

$$\hat{\theta}^{\mathsf{ML}} \sim N\left[\theta_0, \{\mathcal{I}(\theta_0)\}^{-1}\right] \tag{M2}$$

Es asintóticamente eficiente, alcanza el límite inferior Cramér-Rao:

$$\mathcal{I}(\theta_0) = -\mathbb{E}\left[\partial^2 \ln \mathcal{L}/\partial\theta\partial\theta'\right] \tag{M3}$$

Es invariante: si g es una función continua y continuamente diferenciable,

$$\text{el estimador ML de } \gamma = g(\theta) \text{ es } g(\hat{\theta}^{\mathsf{ML}}) \tag{M4}$$

Prueba de razón de verosimilitud

- Supongamos que estimamos los parámetros θ con máxima verosimilitud y que deseamos hacer una prueba de hipótesis H_0 respecto a esos parámetros.
- Sea $\hat{\mathcal{L}}_U$ el valor que se obtiene sin restricciones, y $\hat{\mathcal{L}}_R$ cuando se imponen las restricciones de la hipótesis H_0.
- La razón de verosimilitud $\lambda = \frac{\hat{\mathcal{L}}_R}{\hat{\mathcal{L}}_U}$ necesariamente cumple $0 < \lambda < 1$.
- Entonces

Distribución del estadístico de razón de verosimilitud

Bajo la hipótesis H_0 y condiciones de regularidad, la distribución de

$$-2 \ln \lambda = -2 \left(\ln \mathcal{L}_R - \ln \mathcal{L}_U \right) \sim \chi^2(k)$$

donde k es el número de restricciones impuestas por H_0.

Ejemplo 1:

Goles en mundiales de FIFA

worldcup.ipynb

- En este ejemplo, inspirado en uno similar de Chu (2003), usamos datos compilados por Fjelstul (2023) de todos los goles anotados en mundiales de futbol mayor masculinos de la FIFA.
- Trataremos de estimar dos parámetros relacionados:
  - ¿Cuántos goles se anotan, en tiempo regular, en cada partido?
  - ¿Cada cuántos minutos se anota un gol?

- La distribución Poisson es útil para modelar el número de eventos que ocurren por unidad de tiempo, asumiendo que estos eventos suceden de manera independiente.
- Tenemos una muestra de $\{k_1, \ldots, k_N\}$ del número de goles anotados en todos los $N = 964$ partidos disputados en mundiales mayores masculinos de la FIFA.
- Asumamos que k tiene una distribución Poisson con λ goles por partido.
- La función de densidad de una observación (partido) es $f(k_i|\lambda) = \frac{\lambda^{k_i} e^{-\lambda}}{k_i!}$
- La función de verosimilitud es la probabilidad conjunta de observar esta muestra:

$$\mathcal{L}(\lambda|k_1, \ldots, k_N) = \prod_{i=1}^{N} \frac{\lambda^{k_i} e^{-\lambda}}{k_i!}$$

o bien, tomando su logaritmo

$$\ln \mathcal{L}(\lambda|k_1, \ldots, k_N) = \sum_{i=1}^{N} [k_i \ln \lambda - \lambda - \ln k_i!] = \ln \lambda \sum_{i=1}^{N} k_i - N\lambda - \sum_{i=1}^{N} \ln k_i!$$

- Para encontrar el máximo:

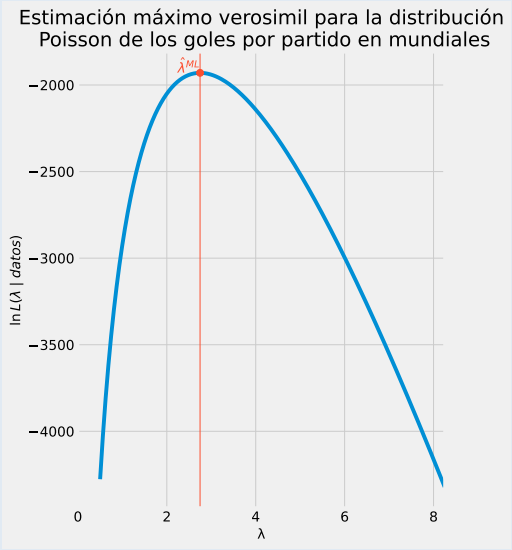
$$\frac{\partial \ln \mathcal{L}(\lambda)}{\partial \lambda} = \frac{1}{\lambda} \sum_{i=1}^{N} k_i - N = 0$$

- Por lo tanto, el estimador de máxima verosimilitud es:

$$\hat{\lambda}^{\mathsf{ML}} = \frac{1}{N} \sum_{i=1}^{N} k_i = \bar{k}$$

donde $\bar{k}$ es el promedio simple de los datos.

- En nuestra muestra particular, $\bar{k} = 2.75$ goles por partido (90 minutos), por lo que el estimador máximo verosimil es $\hat{\lambda}^{\mathsf{ML}} = 2.75$.




Goles por partido en mundiales FIFA masculinos
Observado
Poisson
% de partidos
0
5
10
15
20
25
0
1
2
3
4
5
6
7
8
9
10
11
12
número de goles

- Por su parte, la distribución exponencial es útil para modelar el tiempo que transcurre entre dos eventos, cuando la tasa a la que suceden esos eventos es constante.
- Tenemos una muestra $\{x_1, \ldots, x_N\}$ del tiempo en minutos (enteros, aunque sería mejor reales) entre un gol y el siguiente en los mundiales de la FIFA.
- Supongamos que x_i corresponde a realizaciones independientes de una distribución exponencial con parámetro λ
- La función de densidad de una observación (tiempo entre dos goles) es $f(x|\lambda) = \lambda e^{-\lambda x}$
- La función de verosimilitud es la probabilidad conjunta de observar esta muestra:

$$\mathcal{L}(\lambda|x_1, \ldots, x_N) = \prod_{i=1}^{N} \lambda e^{-\lambda x_i}$$

o bien, tomando su logaritmo

$$\ln \mathcal{L}(\lambda|x_1, \ldots, x_N) = \sum_{i=1}^{N} [\ln \lambda - \lambda x_i] = N \ln \lambda - \lambda \sum_{i=1}^{N} x_i$$

- Para encontrar el máximo:

$$\frac{\partial \ln \mathcal{L}(\lambda)}{\partial \lambda} = \frac{N}{\lambda} - \sum_{i=1}^{N} x_i = 0$$

- Por lo tanto, el estimador de máxima verosimilitud es:

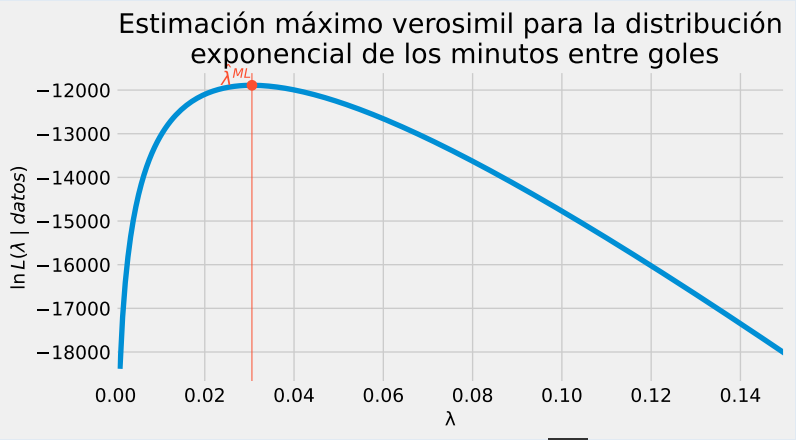
$$\hat{\lambda}_{\mathsf{ML}} = \frac{N}{\sum_{i=1}^{N} x_i} = \frac{1}{\bar{x}}$$

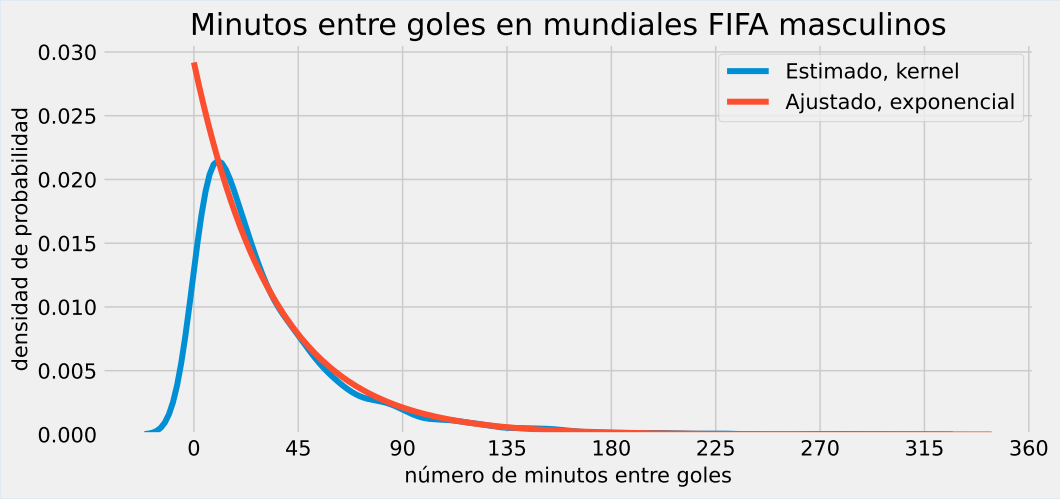
  donde $\bar{x}$ es el promedio simple de los datos.

- En nuestra muestra particular $\bar{x} = 32.75$ minutos entre cada gol, por lo que el estimador máximo verosimil es $\hat{\lambda}^{\mathsf{ML}} = 1/32.75 \approx 0.0305$.

```
# Negativo de la log−verosimilitud de una exponencial: para optimizar con scipy.optimize.minimize
negativelogL2 = lambda λ: −expon.logpdf(goles90['diff'], scale=1/λ).sum()

# Optimización de la log−verosimilitud
MLE_expon = minimize(negativelogL2, 1.0, method='Nelder−Mead')
```




Minutos entre goles en mundiales FIFA masculinos
densidad de probabilidad
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
45
90
135
180
225
270
315
360
número de minutos entre goles
Estimado, kernel
Ajustado, exponencial

La distribución normal multivariada

- Si tenemos $z \in \mathbb{R}^k$ una variable normal multivariada con media μ y matriz de covarianza Σ, entonces su función de densidad de probabilidad es:

$$(2\pi)^{-k/2} \left|\Sigma\right|^{-1/2} \exp\left(-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)\right)$$

- y su logaritmo es

$$\ln \mathcal{L} = -\frac{k}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\Sigma\right| - \left(\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)\right)$$

Estimación por máxima verosimilitud: planteando $\mathcal{L}$

- En el modelo clásico de regresión lineal asumimos que habían n observaciones y que los errores $\varepsilon \,|\, \mathbf{X} \sim \mathrm{N}\left(0, \sigma^2 \mathrm{I}\right)$ (supuesto (A6)).
- Entonces sustituimos $z = \varepsilon$, $\mu = 0$, $\Sigma = \sigma^2 \mathrm{I}$, y $k = n$ en la fórmula de la página anterior para obtener la función de verosimilitud

$$\begin{aligned}
\ln \mathcal{L}(\beta, \sigma^2 | \, \mathbf{Y}, \mathbf{X}) &= -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln \left|\sigma^2 \mathrm{I}\right| - \left(\frac{1}{2} (\varepsilon)' \left(\sigma^2 \mathrm{I}\right)^{-1} (\varepsilon) \right) \\
&= -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln \sigma^{2n} - \frac{1}{2\sigma^2} \varepsilon' \, \varepsilon \\
&= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\, \beta)' (\mathbf{Y} - \mathbf{X}\, \beta)
\end{aligned}$$

- donde en el último paso hemos sustituido $\varepsilon = \mathbf{Y} - \mathbf{X}\, \beta$ por el supuesto (A1).

Estimación por máxima verosimilitud: Definiendo el problema

- **Estrategia**: Buscamos los valores de β y de σ^2 que hagan más plausible o verosimil haber obtenido la muestra $\mathbf{Y}$, $\mathbf{X}$ que tenemos.

$$\hat{\beta}^{\text{ML}} = \underset{\beta}{\operatorname{argmin}} \ln \mathcal{L}(\beta, \sigma^2 | \, \mathbf{Y}, \mathbf{X})$$

$$\hat{\sigma^2}^{\text{ML}} = \underset{\sigma^2}{\operatorname{argmin}} \ln \mathcal{L}(\beta, \sigma^2 | \, \mathbf{Y}, \mathbf{X})$$

donde

$$\ln \mathcal{L}(\beta, \sigma^2 | \, \mathbf{Y}, \mathbf{X}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \left(\mathbf{Y}' \, \mathbf{Y} - 2 \, \mathbf{Y}' \, \mathbf{X} \, \beta + \beta' \, \mathbf{X}' \, \mathbf{X} \, \beta \right)$$

- Vemos que en este caso estamos estimando las pendientes y la varianza del error de forma simultánea.

Estimación por máxima verosimilitud: Resolviendo el problema

- Tomando condiciones de primer orden:

  Para las pendientes:

$$-\frac{1}{2\sigma^2}\left(-2\,\mathbf{X}'\,\mathbf{Y}+2\,\mathbf{X}'\,\mathbf{X}\,\beta\right)=0$$

$$\Rightarrow \hat{\beta}^{\mathsf{ML}}=\left(\mathbf{X}'\,\mathbf{X}\right)^{-1}\mathbf{X}'\,\mathbf{Y}$$

  para la varianza del error:

$$\frac{n}{2\sigma^2}+\frac{\varepsilon'\,\varepsilon}{2(\sigma^2)^2}=0$$

$$\Rightarrow \hat{\sigma^2}^{\mathsf{ML}}=\frac{\varepsilon'\,\varepsilon}{n}$$

- De nuevo, hemos asumido que se cumple (A2), de manera que $\mathbf{X}'\,\mathbf{X}$ sea invertible.

5. El método de momentos (MM)

Momentos

- Sea X una variable aleatoria, y $k = 1, 2, \ldots$. Entonces definimos:
  - $\mathbb{E}\left(X^k\right)$ es el k-ésimo momento (alrededor del origen) de la distribución.
  - $\mathbb{E}\left[(X-\mu)^k\right]$ es el k-ésimo momento central de la distribución.
  - $M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ es el k-ésimo momento muestral.
  - $M_k^* = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^k$ es el k-ésimo momento muestral central.
- Recuerde que según la Ley de los Grandes Números, bajo ciertas condiciones se tiene que

$$\operatorname{plim} M_k = \mathbb{E}\left(X^k\right)$$
$$\operatorname{plim} M_k^* = \mathbb{E}\left[(X-\mu)^k\right]$$

es decir, los momentos muestrales convergen a sus contrapartes poblacionales conforme aumenta el tamaño de muestra n

El método de momentos

- Supongamos que deseamos estimar k parámetros desconocidos $\theta \equiv [\theta_1, \theta_2, \ldots, \theta_k]'$ que caracterizan la distribución de la variable aleatoria X, a partir de una muestra de n realizaciones de esta variable.
- Supongamos que los primeros k momentos de la distribución de X pueden expresarse en términos de los parámetros θ:

$$\begin{aligned} \mathbb{E}[X] &= g_1(\theta_1, \theta_2, \ldots, \theta_k) \\ \mathbb{E}\left[X^2\right] &= g_2(\theta_1, \theta_2, \ldots, \theta_k) \\ &\vdots \\ \mathbb{E}\left[X^k\right] &= g_k(\theta_1, \theta_2, \ldots, \theta_k) \end{aligned}$$

- La idea básica del método de momentos (MM) es:
  1. Sustituir los momentos poblacionales con sus contrapartes muestrales.
  2. Despejar los valores de $\theta_1, \theta_2, \ldots, \theta_k$
- A los valores $\hat{\theta}$ encontrados los denominamos estimador del método de momentos.

Ejemplo 2:

Estimación de parámetros de una distribución uniforme

- Supongamos que tenemos una muestra $x_1, \ldots, x_n$ tomados de una distribución uniforme, $X \sim U(a, b)$.
- Queremos estimar dos parámetros, a y b, por lo que necesitamos los primeros dos momentos de la distribución:

$$\mathbb{E}[X] = \frac{a+b}{2} \qquad \mathbb{E}\left[X^2\right] = \frac{a^2+ab+b^2}{3}$$

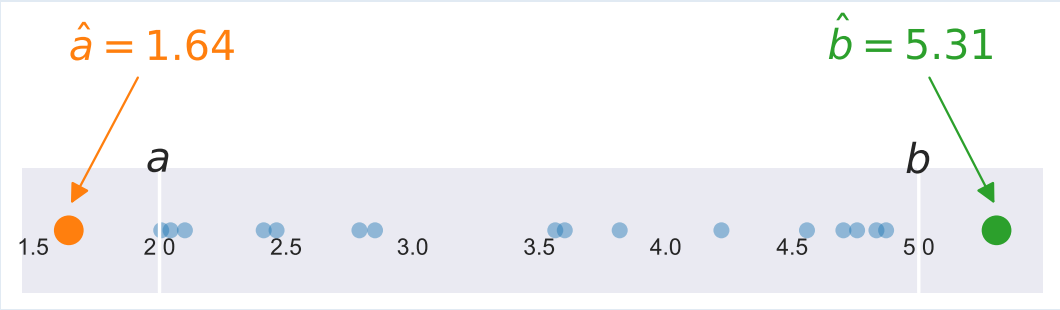
- Sustituyendo los momentos poblacionales con sus contrapartes muestrales:

$$M_1 = \frac{\hat{a}+\hat{b}}{2} \qquad M_2 = \frac{\hat{a}^2+\hat{a}\hat{b}+\hat{b}^2}{3}$$

- Despejando $\hat{a}$ y $\hat{b}$ encontramos

$$\hat{a} = M_1 - \sqrt{3(M_2 - M_1^2)} \qquad \hat{b} = M_1 + \sqrt{3(M_2 - M_1^2)}$$

```
1  import numpy as np
2  a, b, n = 2.0, 5.0, 16
3  np.random.seed(12)
4  x = np.random.uniform(low=a, high=b, size=n)
5  M1, M2 = x.mean(), (x**2).mean()
6  hat_a = M1 - np.sqrt(3*(M2-M1**2))
7  hat_b = M1 + np.sqrt(3*(M2-M1**2))
```



- El método también podemos ejecutarlo en términos de momentos centrales.
- En este caso:

$$\mathbb{E}[X] = \frac{a+b}{2} \qquad \mathbb{E}\left[(X-\mu)^2\right] = \frac{(b-a)^2}{12}$$

- De nuevo, sustituimos momentos poblacionales con muestrales:

$$M_1^* = \frac{\hat{a}+\hat{b}}{2} \qquad M_2^* = \frac{\left(\hat{b}-\hat{a}\right)^2}{12}$$

- Despejando $\hat{a}$ y $\hat{b}$ encontramos

$$\hat{a} = M_1 - \sqrt{3M_2^*} \qquad \hat{b} = M_1 + \sqrt{3M_2^*}$$

Estimador MCO como caso particular del estimador MM

- Como en el modelo clásico de regresión lineal, supongamos que

$$y_i = x_i'\beta + \epsilon_i \qquad \mathbb{E}[\epsilon|x] = 0$$

- Por la teoría de probabilidad, sabemos que $\mathbb{E}[\epsilon|x] = 0 \Rightarrow \mathbb{E}[x\epsilon] = 0$.
- Entonces las condiciones de momentos poblacional son

$$\mathbb{E}[x(y - x'\beta)] = 0$$

- Las condiciones de momentos muestrales entonces son

$$\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - x_i'\beta) = 0 \Rightarrow \qquad 0 = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i x_i'\beta \Rightarrow$$

$$X'X\beta = X'y \Rightarrow \qquad \beta^{\text{MM}} = \beta^{\text{OLS}} = (X'X)^{-1}X'y$$

Estimador VI como caso particular del estimador MM

- De manera similar, el estimador de variables instrumentales también es un caso particular de MM: supongamos que

$$y_i = x_i'\beta + \epsilon_i \qquad \mathbb{E}\left[\epsilon|x\right] \neq 0 \qquad \mathbb{E}\left[\epsilon|z\right] = 0 \Rightarrow \mathbb{E}[z\epsilon] = 0$$

- Entonces las condiciones de momentos poblacional son $\mathbb{E}[z(y - x'\beta)] = 0$
- Las condiciones de momentos muestrales entonces son

$$\frac{1}{n}\sum_{i=1}^{n} z_i(y_i - x_i'\beta) = 0 \Rightarrow \qquad 0 = \sum_{i=1}^{n} z_i y_i - \sum_{i=1}^{n} z_i x_i'\beta \Rightarrow$$

$$Z'X\beta = Z'y \Rightarrow \qquad \beta^{\mathsf{MM}} = \beta^{\mathsf{IV}} = (Z'X)^{-1}Z'y$$

- Notemos que para poder obtener el estimador de variables instrumentales $\beta^{\mathsf{IV}} = (Z'X)^{-1}Z'y$, es necesario que hayan tantos instrumentos (columnas de Z) como variables explicativas (columnas de X, igual al número de parámetros), de manera que $Z'X$ sea cuadrada y potencialmente invertible.

6. El método generalizado de momentos (GMM)

Limitaciones del método de momentos

- El MM solo funciona cuando el número de condiciones de momentos es igual al número de parámetros por estimar.
- Si hay más condiciones de momentos que parámetros, el sistema de ecuaciones se sobreidentifica algebraicamente y no se puede resolver.
- Los estimadores del método generalizado de momentos (GMM) eligen las estimaciones que minimizan una forma cuadrática de las condiciones de momentos.
- GMM se acerca lo más posible a resolver el sistema sobreidentificado.
- GMM es igual a MM cuando el número de parámetros es igual al número de condiciones de momentos.

Generalizando el método de momentos: más instrumentos que parámetros

- Supongamos que hay más instrumentos (q) que parámetros por estimar (k).
- Denotemos las condiciones de momentos por $\mathbb{E}[m_j(w, \theta)] = 0$, donde $j = 1, \ldots, q$.
- Al sustituir por momentos muestrales $\bar{m}_j = \frac{1}{n} \sum_{i=1}^{n} m_j(w_i, \theta) = 0$, tenemos un sistema de q ecuaciones (posiblemente no lineales) en k incógnitas (los parámetros θ).
- Como el sistema está sobreidentificado, podríamos buscar entonces el vector θ que mejor aproxime el sistema, es decir

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ \bar{m}_1^2 + \bar{m}_2^2 + \cdots + \bar{m}_q^2 \right\}$$

Generalizando el método de momentos: más instrumentos que parámetros

- Pero bien podríamos asignarle una importancia (peso $w > 0$) distinto a cada momento:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ w_1 \bar{m}_1^2 + w_2 \bar{m}_2^2 + \cdots + w_q \bar{m}_q^2 \right\}$$

$$= \underset{\theta}{\operatorname{argmin}} \left\{ \underbrace{\begin{bmatrix} \bar{m}_1 & \bar{m}_2 & \dots & \bar{m}_q \end{bmatrix}}_{m(\theta)'} \underbrace{\begin{bmatrix} w_1 & 0 & \dots 0 & \\ 0 & w_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & w_q \end{bmatrix}}_{W} \underbrace{\begin{bmatrix} \bar{m}_1 \\ \bar{m}_2 \\ \vdots \\ \bar{m}_q \end{bmatrix}}_{m(\theta)} \right\}$$

Definición del estimador GMM

- En resumen, en la práctica tendremos un modelo teórico que implica q condiciones de momentos poblacionales

$$\mathbb{E}\left[m(w_i, \theta)\right] = 0$$

  - m es un vector de $q \times 1$ funciones que tienen esperanza cero en la población.
  - w_i son datos de la observación i
  - θ es un vector de $k \times 1$ parámetros, con $k \leq q$.
- Los momentos condicionales correspondientes son

$$\bar{m}(\theta) = \frac{1}{n} \sum_{i=1}^{n} m(w_i, \theta)$$

- El estimador GMM escoge los parámetros que están lo más cerca posible de resolver el sistema de ecuaciones sobreidentificado:

$$\hat{\theta}^{\text{GMM}} \equiv \underset{\theta}{\operatorname{argmin}} \left\{ \bar{m}(\theta)' W \bar{m}(\theta) \right\}$$

donde W es una matriz definida positiva.

Escogiendo la matriz de ponderaciones W

W solo afecta la eficiencia del estimador GMM

- Si fijamos $W = I$ el estimador es consistente, pero ineficiente.
- Hansen (1982) demostró que una condición necesaria (pero no suficiente) para obtener un estimador eficiente es fijar $W = \mathbb{V}\left[m(\theta)\right]^{-1}$.
- Intuición: darle *menos* peso a los momentos que tienen *más* varianza.
- Pero como $W = \mathbb{V}\left[m(\theta)\right]^{-1}$ depende precisamente de los parámetros θ que deseamos estimar, para estimar un modelo por GMM procedemos así:
  1. Fijamos $W = I$ y obtenemos
  $$\hat{\theta}^{\text{GMM,1}} \equiv \underset{\theta}{\operatorname{argmin}} \left\{ \bar{m}(\theta)' \bar{m}(\theta) \right\}$$
  2. Usamos $\hat{\theta}^{\text{GMM,1}}$ para calcular $\tilde{W}$, el cual es un estimador de $\mathbb{V}\left[\bar{m}(\theta)\right]^{-1}$.
  3. Obtenemos
  $$\hat{\theta}^{\text{GMM,2}} \equiv \underset{\theta}{\operatorname{argmin}} \left\{ \bar{m}(\theta)' \hat{W} \bar{m}(\theta) \right\}$$
- Opcionalmente, repetimos los pasos 2 y 3, usando ahora $\hat{\theta}^{\text{GMM,2}}$.
- En condiciones adecuadas, este estimador es consistente, asintóticamente normal, y asintóticamente eficiente.

Ejemplo 3:

Estimación del parámetro de una distribución t de Student

- Este ejemplo está basado en el capítulo 14 de Hamilton (1994).
- Supongamos que tenemos una muestra i.i.d. de T observaciones $y_1, y_2, \ldots, y_T$ de una variable aleatoria que tiene distribución t de Student.
- La función de densidad de esta distribución es

$$f_{Y_1}(y_i; \nu) = \frac{\Gamma[(\nu+1)/2]}{(\pi\nu)^{1/2}\Gamma(\nu/2)} \left[1 + \left(y_t^2/\nu\right)\right]^{-(\nu+1)/2}$$

  donde Γ es la función gamma.
- **Problema**: ¿Cómo estimar ν a partir de los datos?
- A continuación consideramos tres alternativas:
  - el método de máxima verosimilitud (MLE)
  - el método de momentos (MM)
  - el método generalizado de momentos (GMM)

Opción 1: MLE

- Este enfoque consiste en escoger como estimador $\hat{\nu}$ el valor ν que maximiza la función de log-verosimilitud:

$$\begin{aligned}\mathcal{L}(\nu) &= \sum_{i=1}^{T} \log f_{Y_i}\left(y_t;\nu\right) \\ &= T\log\Gamma\left(\frac{\nu+1}{2}\right) - T\log\Gamma\left(\frac{\nu}{2}\right) - \frac{T}{2}\log(\pi\nu) - \frac{\nu+1}{2}\sum_{t=1}^{T}\log\left(1+\frac{y_t^2}{\nu}\right)\end{aligned}$$

- El estimador $\hat{\nu}^{\mathsf{ML}}$ satisface la condición de primer orden

$$\psi\left(\frac{\nu+1}{2}\right) - \psi\left(\frac{\nu}{2}\right) - \frac{1}{\nu} + \frac{1}{T}\sum_{t=1}^{T}\left[\frac{y_t^2+\nu y_t^2}{\nu^2+\nu y_t^2} - \log\left(1+\frac{y_t^2}{\nu}\right)\right] = 0$$

donde $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\ln(\Gamma(x))$ es la función digamma.

- Evidentemente, no es nada sencillo despejar ν en esta ecuación.

Opción 2: MM

- Asumiendo que la distribución tiene más de dos grados de libertad, $\nu > 2$, su media poblacional es cero y su varianza es

$$\mu_2 \equiv E\left(Y_1^2\right) = \frac{\nu}{\nu - 2}$$

- Si calculamos el segundo momento muestral, podemos estimar el parámetro ν a partir de

$$\hat{\mu}_2 \equiv \frac{1}{T}\sum_{t=1}^{T} y_t^2 = \frac{\nu}{\nu - 2} \qquad \Rightarrow \qquad \hat{\nu}^{\text{MM}} = \frac{2\hat{\mu}_2}{\hat{\mu}_2 - 1}$$

- Como $\operatorname{plim} \hat{\mu}_2 = \mu_2$ tenemos que $\operatorname{plim} \hat{\nu} = \nu$, es decir, el método de momentos es consistente.

Este código de Python simula una muestra de 500 datos de una distribución t de Student con 9 grados de libertad.

```
1      from scipy.optimize import fmin
2      from scipy.stats import t
3
4      nu, nu_ini, T = 9, 11, 500
5      np.random.seed(2021)
6      y = t(nu).rvs(T)
7
8      hat_nu2 = (y**2).mean()
9      nu_MM = 2*hat_nu2/(hat_nu2-1)
```

Al ejecutarlo, obtenemos que $\hat{\nu}^{\mathsf{MM}} = 9.1385$.

Opción 3: GMM

- Asumiendo que la distribución tiene más de cuatro grados de libertad, $\nu > 4$, su cuarto momento es

$$\mu_4 \equiv E\left(Y_t^4\right) = \frac{3\nu^2}{(\nu-2)(\nu-4)}$$

- Si calculamos el cuarto momento muestral,

$$\hat{\mu}_4 \equiv \frac{1}{T}\sum_{t=1}^{T} y_t^4$$

podríamos estimar ν a partir de este sistema de ecuaciones:

$$\bar{m}_1(\nu) \equiv \hat{\mu}_2 - \frac{\nu}{\nu-2} = 0 \qquad \bar{m}_2(\nu) \equiv \hat{\mu}_4 - \frac{3\nu^2}{(\nu-2)(\nu-4)} = 0$$

- No podemos escoger un único ν que satisfaga ambas condiciones.
- Por ello, minimizaremos una función de la forma

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ \begin{bmatrix} \bar{m}_1 & \bar{m}_2 \end{bmatrix} \underset{2\times 2}{W} \begin{bmatrix} \bar{m}_1 \\ \bar{m}_2 \end{bmatrix} \right\}$$

Este código de Python estima el parámetro ν con los datos generados anteriormente, usando dos momentos:

```
1     def m(nu):
2     m1 = y**2 - nu/(nu-2)
3     m2 = y**4 - (3*nu**2)/((nu-2)*(nu-4))
4     return np.array([m1,m2])
5
6     def objetivo(nu):
7     mnu = m(nu[0]).mean(axis=1)
8     return mnu.T @ W @ mnu
9
10    W = np.eye(2)
11    hat_nu, = fmin(objetivo, nu_ini)
12
13    W = np.linalg.inv(np.cov(m(hat_nu)))
14    hat_nu, = fmin(objetivo, hat_nu)
```

Al ejecutarlo, obtenemos que $\hat{\nu}^{\text{GMM}} = 7.9062$.

Ejemplo 4:

Estimando una ecuación de Euler

Un modelo de consumo intertemporal:

- Supongamos que tenemos un problema del consumidor, quien desea maximizar

$$\max_{c_{t+i},A_{t+i}} \mathbb{E}_t \sum_{i=0}^{\infty} (1+\delta)^{-i} u\left(c_{t+i}\right)$$

sujeto a las restricciones

$$A_{t+i} = (1+r)A_{t+i-1} + y_{t+i} - c_{t+i}$$
$$\lim_{i\to\infty} E_t A_{t+i}(1+r)^{-i} = 0$$

- Sabemos que la ecuación de Bellman en este caso es

$$V(A) = \max_{A'} \left[u((1+r)A + y - A') + \mathbb{E}\, V(A')\right]$$

- La ecuación de Euler es

$$\mathbb{E}_t \left[\frac{1+r}{1+\delta} u'(c_{t+1})\right] = u'(c_t)$$

- Suponiendo que la función de utilidad es CRRA, la utilidad marginal es $u'(c) = c^{-\gamma}$, la ecuación de Euler es entonces

$$\mathbb{E}_t \left[\underbrace{\frac{1+r}{1+\delta} c_{t+1}^{-\gamma} - c_t^{-\gamma}}_{f_{t+1}} \right] = 0$$

- Los únicos parámetros por estimar son "profundos": δ y γ describen las preferencias del consumidor.
- Esta ecuación no es la función de consumo.
- Pero asumiendo expectativas racionales, esta ecuación nos dice que la única información útil en el periodo t para predecir el consumo futuro c_{t+1} es el consumo actual c_t.
- Esto no significa que otras variables $\mathbf{z}_t$ (como el ingreso o la riqueza) no determinen el nivel de consumo, solo que no son útiles para pronosticarlo.

Estimando los parámetros de preferencias:

- Lo anterior nos sugiere que podemos estimar los parámetros utilizando los momentos

$$\mathbb{E}_t \, f_{t+1} \mathbf{z}_t = 0$$

- o más explícitamente

$$\mathbb{E}_t \, f_{t+1}(c; \delta, \gamma) \mathbf{z}_t = 0$$

- Note que la ecuación de Euler original la obtenemos simplemente fijando $z_t = 1$.

- Planteamos entonces, por ejemplo, estas condiciones de ortogonalidad (momentos):

$$\mathbb{E}_t \left[\frac{1 + r_{t+1}}{1 + \delta} \left(\frac{c_t}{c_{t+1}} \right)^{\gamma} - 1 \right] = 0$$

$$\mathbb{E}_t \left[\left(\frac{1 + r_{t+1}}{1 + \delta} \left(\frac{c_t}{c_{t+1}} \right)^{\gamma} - 1 \right) r_t \right] = 0$$

$$\mathbb{E}_t \left[\left(\frac{1 + r_{t+1}}{1 + \delta} \left(\frac{c_t}{c_{t+1}} \right)^{\gamma} - 1 \right) \left(\frac{c_{t-1}}{c_t} \right) \right] = 0$$

- En este caso, los instrumentos son

$$\mathbf{z}_t' = \left[1,\ r_t,\ \frac{c_{t-1}}{c_t} \right]$$

Algunas limitaciones de este estimador:

Teoréticas: Este procedimiento de estimación tiene sentido únicamente si asumimos que el consumidor no enfrenta restricciones de crédito: vemos los datos como soluciones de equilibrio interior, no como soluciones de esquina.

Empíricas: los parámetros estimados con series de tiempo agregadas usualmente son inestables. Esto contradice su naturaleza de parámetros "profundos".

Por lo anterior, en la práctica estos tipos de modelos usualmente se calibran a partir de estimaciones de estudios microeconómicos con datos desagregados.

Ejemplo 5:

Métodos de estimación

Estimación con scipy minimize.ipynb

- En este ejemplo ilustramos algunos de los métodos de estimación que hemos visto.
- Consideramos el siguiente modelo de regresión lineal

$$\text{esperanza\_de\_vida} = \beta_0 + \beta_1 \text{escolaridad} + \beta_2 \text{ingreso} + \varepsilon$$

el cual estimamos con datos de 191 países, tomados del WDI.

```
1 import numpy as np
2 from scipy.stats import norm
3 from scipy.optimize import minimize
4 import pandas as pd
5
6 # Para guardar los resultados
7 estimaciones = pd.DataFrame(index=['intercepto', 'escolaridad', 'ingreso'])
```

Mínimos cuadrados ordinarios, solución analítica

```
8  # Matrices de datos
9  Y = wdi[['esperanza_de_vida']].values
10 X = np.c_[np.ones_like(Y), wdi[['escolaridad', 'ingreso']].values]
11
12 # estimación de los parámetros: (X'X)^{-1}X'Y
13 XXinv = np.linalg.inv(X.T@X) # (X'X)^{-1}
14 beta_hat = XXinv @ (X.T @ Y)
15
16 estimaciones['OLS_formula'] = beta_hat
```

Mínimos cuadrados ordinarios, solución numérica

```
17 # Función que calcula los residuos a partir de los parámetros β
18 def ε(β):
19     return wdi['esperanza_de_vida'] − β[0] − β[1] * wdi['escolaridad'] − β[2] * wdi['ingreso']
20
21 # Función objetivo: Suma de cuadrados de los residuos
22 def ssr(β):
23     return (ε(β)**2).sum()
24
25 # Optimización de la función objetivo
26 initial_guess = np.ones(3)
27 ols_results = minimize(ssr, initial_guess)
28 estimaciones['OLS_optim'] = ols_results.x
```

Método de máxima verosimilitud

```
# Función objetivo: (negativo de) la log−verosimilitud para distribución normal
def logL(params):
    β, σ2 = params[:−1], params[−1]
    return −norm.logpdf(ε(β), loc=0, scale=np.sqrt(σ2)).sum()

# Optimización de la función de verosimilitud
mle_results = minimize(logL, np.ones(4))
estimaciones['MLE'] = mle_results.x[:3]
```

Método de momentos

```
37 # Función objetivo: los residuos son ortogonales a las variables explicativas
38 def moments(β):
39     m = ε(β) @ X # E[εX] = 0
40     return (m**2).sum()
41
42 # Minimización de la función objetivo
43 mm_results = minimize(moments, np.ones(3), method='SLSQP')
44 estimaciones['MM'] = mm_results.x
```

Resultados

	OLS_formula	OLS_optim	MLE	MM
intercepto	58.535871	58.535870	58.535875	58.536390
schooling	1.069237	1.069237	1.069236	1.069177
income	0.156444	0.156444	0.156444	0.156447

Referencias I

- Cameron, A. Colin y Pravin K. Trivedi (2022a). *Microeconometrics Using Stata. Volumen I: Cross-Sectional and Panel Regression Methods*. 2ª ed. Vol. 1. Stata Press. ISBN: 978-1-59718-361-1.
- — (2022b). *Microeconometrics Using Stata. Nonlinear Models and Causal Infrerence Methods*. 2ª ed. Vol. 2. Stata Press. ISBN: 978-1-59718-362-8.
- Chu, Singfat (2003). "Using Soccer Goals to Motivate the Poisson Process". En: *INFROMS Transactions on Education* 3.2, págs. 62-68.
- Fjelstul, Josjua (2023). *The Fjelstul World Cup Database v.1.0*. URL: https://github.com/jfjelstul/worldcup.
- Greene, William H. (2012). *Econometric Analysis*. 7ª ed. Prentice Hall. ISBN: 978-0-13-139538-1.
- Hamilton, James M. (1994). *Time Series Analysis*. Princeton University Press. ISBN: 0-691-04289-6.
- Hansen, Bruce E. (2022). *Econometrics*. Princeton University Press.