

# Cheat Sheet Econometría

Por Marcelo Moreno - Universidad Rey Juan Carlos

The Econometrics Cheat Sheet Project

## Conceptos básicos

### Definiciones

**Econometría** - es una disciplina de las ciencias sociales que tiene como objetivo cuantificar las relaciones entre agentes económicos, contrastar teorías económicas y evaluar e implementar políticas públicas y privadas.

**Modelo econométrico** - es una representación simplificada de la realidad para explicar fenómenos económicos.

**Ceteris paribus** - si todos los demás factores relevantes permanecen constantes.

### Tipos de datos

**Sección cruzada** - datos recogidos en un momento dado en el tiempo, una *foto* estática. El orden no importa.

**Serie temporales** - observación de una/muchas variable/s durante un periodo de tiempo. El orden sí importa.

**Datos de panel** - consiste una una serie temporal por cada observación de una sección cruzada.

**Secciones transversales agrupadas** - combina secciones cruzadas de diferentes periodos temporales.

### Fases de un modelo econométrico

1. Especificación.
2. Estimación.
3. Validación.
4. Utilización.

### Análisis de regresión

Estudiar y predecir el valor medio de una variable (dependiente,  $y$ ) respecto a unos valores fijos de otras variables (variables independientes,  $x$ 's). En econometría es común usar Mínimos Cuadrados Ordinarios (MCO) para análisis de regresión.

### Análisis de correlación

El análisis de correlación no distingue entre variables dependientes e independientes.

- La correlación simple mide el grado de asociación lineal entre dos variables.

$$r = \frac{\text{Cov}(x,y)}{\sigma_x \cdot \sigma_y} = \frac{\sum_{i=1}^n ((x_i - \bar{x}) \cdot (y_i - \bar{y}))}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- La correlación parcial mide el grado de de asociación lineal entre dos variables controlando una tercera.

## Supuestos y propiedades

### Supuestos del modelo econométrico

Bajo estos supuestos, el estimador de MCO presentará buenas propiedades. Supuestos **Gauss-Markov**:

1. **Linealidad en parámetros** (y dependencia débil en series temporales).  $y$  debe ser una función lineal de  $\beta$ 's.
2. **Muestreo aleatorio**. La muestra de la población se ha tomado de forma aleatoria. (Sólo sección cruzada)
3. **No colinealidad perfecta**.
  - No hay variables independientes que sean constantes:  $\text{Var}(x_j) \neq 0, \forall j = 1, \dots, k$
  - No hay una relación lineal exacta entre variables independientes.
4. **Media condicional cero y correlación cero**.
  - a. No hay errores sistemáticos:  $E(u | x_1, \dots, x_k) = E(u) = 0 \rightarrow$  **exogeneidad fuerte** (a implica b).
  - b. No hay variables relevantes fuera del modelo:  $\text{Cov}(x_j, u) = 0, \forall j = 1, \dots, k \rightarrow$  **exogeneidad débil**.
5. **Homocedasticidad**. La variabilidad de los residuos es igual para todos los niveles de  $x$ :  
 $\text{Var}(u | x_1, \dots, x_k) = \sigma_u^2$
6. **No autocorrelación**. Los residuos no contienen información sobre otros residuos:  
 $\text{Corr}(u_t, u_s | x_1, \dots, x_k) = 0, \forall t \neq s$
7. **Normalidad**. Los residuos son independientes e idénticamente distribuidos:  $u \sim \mathcal{N}(0, \sigma_u^2)$
8. **Tamaño de datos**. El número de observaciones disponibles debe ser mayor a  $(k + 1)$  parámetros a estimar. (Ya satisfecho bajo situaciones asintóticas)

### Propiedades asintóticas de MCO

Bajo los supuestos del modelo econométrico y el Teorema Central del Límite (TCL):

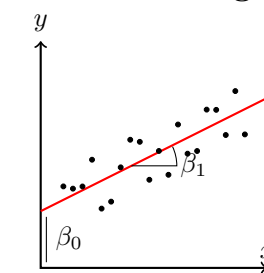
- De 1 a 4a: MCO es **insesgado**.  $E(\hat{\beta}_j) = \beta_j$
- De 1 a 4: MCO es **consistente**.  $\text{plim}(\hat{\beta}_j) = \beta_j$  (a 4b sin 4a, exogeneidad débil, insesgado y consistente).
- De 1 a 5: **normalidad asintótica** de MCO (entonces, 7 es necesariamente satisfecho):  $u \sim_a \mathcal{N}(0, \sigma_u^2)$
- De 1 a 6: **estimador insesgado** de  $\sigma_u^2$ .  $E(\hat{\sigma}_u^2) = \sigma_u^2$
- De 1 a 6: MCO es MELI (Mejor Estimador Lineal Insesgado, **BLUE** en inglés) ó **eficiente**.
- De 1 a 7: contrastes de hipótesis e intervalos de confianza son fiables.

## Mínimos Cuadrados Ordinarios

**Objetivo** - minimizar Suma de Resid. Cuadrados (SRC):

$$\min \sum_{i=1}^n \hat{u}_i^2, \text{ donde } \hat{u}_i = y_i - \hat{y}_i$$

### Modelo de regresión simple



Ecuación:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Estimación:

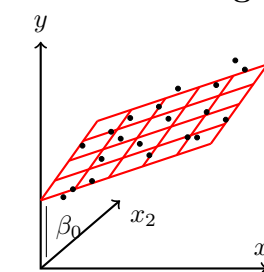
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

donde:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\text{Cov}(y, x)}{\text{Var}(x)}$$

### Modelo de regresión múltiple



Ecuación:

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + u_i$$

Estimación:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik}$$

donde:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \dots - \hat{\beta}_k \bar{x}_k$$

$$\hat{\beta}_j = \frac{\text{Cov}(y, \text{resid } x_j)}{\text{Var}(\text{resid } x_j)}$$

Matriz:  $\hat{\beta} = (X^T X)^{-1} (X^T y)$

### Interpretación de coeficientes

Modelo	Depend.	Independ.	Interpretación $\beta_1$
Nivel-nivel	$y$	$x$	$\Delta y = \beta_1 \Delta x$
Nivel-log	$y$	$\log(x)$	$\Delta y \approx (\beta_1/100)(\% \Delta x)$
Log-nivel	$\log(y)$	$x$	$\% \Delta y \approx (100 \beta_1) \Delta x$
Log-log	$\log(y)$	$\log(x)$	$\% \Delta y \approx \beta_1 (\% \Delta x)$
Cuadrático	$y$	$x + x^2$	$\Delta y = (\beta_1 + 2\beta_2 x) \Delta x$

### Medidas de error

Suma de Resid. Cuad.:  $\text{SRC} = \sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

Suma Explicada de Cuadrados:  $\text{SEC} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

Suma Tot. de Cuad.:  $\text{STC} = \text{SEC} + \text{SRC} = \sum_{i=1}^n (y_i - \bar{y})^2$

Error Estándar de la Regresión:  $\hat{\sigma}_u = \sqrt{\frac{\text{SRC}}{n-k-1}}$

Error Estándar de los  $\hat{\beta}$ 's:  $\text{ee}(\hat{\beta}) = \sqrt{\hat{\sigma}_u^2 \cdot (X^T X)^{-1}}$

Raíz del Error Cuadrático Medio:  $\text{RECM} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$

Error Medio Absoluto:  $\text{EMA} = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$

Porcentaje Medio de Error:  $\text{PME} = \frac{\sum_{i=1}^n |\hat{u}_i / y_i|}{n} \cdot 100$

## R-cuadrado

Es una medida de la **bondad del ajuste**, cómo la regresión se ajusta a los datos:

$$R^2 = \frac{SEC}{STC} = 1 - \frac{SRC}{STC}$$

- Mide el **porcentaje de variación** en  $y$  que es linealmente **explicado** por variaciones de las  $x$ 's.
- Toma valores **entre 0** (no hay explicación lineal) **y 1** (explicación total).

Cuando el número de regresores incrementa, el R-cuadrado también, independientemente de si las nuevas variables son relevantes o no. Para resolver este problema, hay un **R-cuadrado ajustado** por grados de libertad (o corregido):

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \cdot \frac{SRC}{STC} = 1 - \frac{n-1}{n-k-1} \cdot (1 - R^2)$$

Para muestras grandes:  $\bar{R}^2 \approx R^2$

## Contrastes de hipótesis

### Definiciones

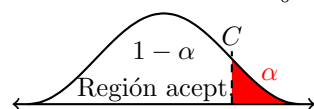
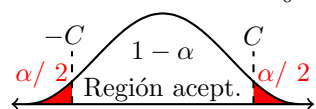
Es una regla diseñada para, a partir de una muestra, explicar si existe **evidencia para rechazar (o no) una hipótesis** sobre uno o más parámetros poblacionales.

Elementos de un contraste de hipótesis:

- **Hipótesis nula** ( $H_0$ ) - es la hipótesis a ser probada.
- **Hipótesis alternativa** ( $H_1$ ) - es la hipótesis que no puede rechazarse si  $H_0$  es rechazada.
- **Estadístico de contraste** - es una variable aleatoria cuya distribución de probabilidad es conocida bajo  $H_0$ .
- **Valor crítico** ( $C$ ) - es el valor contra el cual se compara el estadístico de contraste para determinar si se rechaza o no la hipótesis nula. Determina la frontera entre la región de aceptación y la de rechazo de  $H_0$ .
- **Nivel de significación** ( $\alpha$ ) - es la probabilidad de rechazar la  $H_0$  siendo cierta (Error Tipo I). Es elegido por quien conduce el contraste. Usualmente 10%, 5% ó 1%.
- **p-valor** - es el nivel de significación máximo por el cual  $H_0$  no puede ser rechazada.

Dos colas. Distrib.  $H_0$

Una cola. Distrib.  $H_0$



**Regla general:** si  $p\text{-valor} < \alpha$ , existe evidencia para rechazar  $H_0$ , es decir, existe evidencia para aceptar  $H_1$ .

## Contrastes individuales

Prueba si un parámetro es significativamente diferente de un cierto valor,  $\vartheta$ .

- $H_0 : \beta_j = \vartheta$
- $H_1 : \beta_j \neq \vartheta$

$$\text{Bajo } H_0: \quad t = \frac{\hat{\beta}_j - \vartheta}{\text{ee}(\hat{\beta}_j)} \sim t_{n-k-1, \alpha/2}$$

Si  $|t| > |t_{n-k-1, \alpha/2}|$ , existe evidencia para rechazar  $H_0$ .

**Contraste de significación individual** - prueba si un parámetro es **significativamente distinto de cero**.

- $H_0 : \beta_j = 0$
- $H_1 : \beta_j \neq 0$

$$\text{Bajo } H_0: \quad t = \frac{\hat{\beta}_j}{\text{ee}(\hat{\beta}_j)} \sim t_{n-k-1, \alpha/2}$$

Si  $|t| > |t_{n-k-1, \alpha/2}|$ , existe evidencia para rechazar  $H_0$ .

## Contraste F

Prueba simultáneamente múltiples hipótesis (lineales) sobre los parámetros. Hace uso de un modelo no restringido y uno restringido:

- **Modelo no restringido** - es el modelo donde se quiere probar la hipótesis.
- **Modelo restringido** - es el modelo donde se ha impuesto la hipótesis que se quiere probar.

Entonces, viendo los errores, hay:

- $SRC_{UR}$  - es la SRC del modelo no restringido.
- $SRC_R$  - es la SRC del modelo restringido.

Bajo  $H_0$ :  $F = \frac{SRC_R - SRC_{UR}}{SRC_{UR}} \cdot \frac{n-k-1}{q} \sim F_{q, n-k-1}$   
donde  $k$  es el número de parámetros del modelo no restringido y  $q$  es el número de hipótesis lineales a probar.  
Si  $F > F_{q, n-k-1}$ , existe evidencia para rechazar  $H_0$ .

**Contraste de significación global** - prueba si todos los parámetros asociados a  $x$ 's son **simultáneamente cero**.

- $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$
- $H_1 : \beta_1 \neq 0 \text{ y/o } \beta_2 \neq 0 \dots \text{ y/o } \beta_k \neq 0$

Podemos simplificar la fórmula para el estadístico  $F$ :

$$\text{Bajo } H_0: \quad F = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k} \sim F_{k, n-k-1}$$

Si  $F > F_{k, n-k-1}$ , existe evidencia para rechazar  $H_0$ .

## Intervalos de confianza

Los intervalos de confianza al nivel de confianza  $(1 - \alpha)$ , se pueden calcular:

$$\hat{\beta}_j \mp t_{n-k-1, \alpha/2} \cdot \text{ee}(\hat{\beta}_j)$$

## Variables ficticias

Las variables ficticias (o binarias) son usadas para recoger información cualitativa: sexo, estado civil, país, etc.

- Toman **valor 1** en una categoría dada y **0 en el resto**.
- Se usan para analizar y modelizar **cambios estructurales** en los parámetros del modelo.

Si una variable cualitativa tiene  $m$  categorías, sólo hay que incluir  $(m - 1)$  variables ficticias en el modelo.

## Cambio estructural

El cambio estructural se refiere a los cambios en los valores de los parámetros del modelo producidos por el efecto de diferentes sub-poblaciones. El cambio estructural se puede incluir en el modelo a través de variables ficticias.

La ubicación de las variables ficticias ( $D$ ) es importante:

- **En la constante** (efecto aditivo) - representa la diferencia media entre los valores producidos por el cambio estructural.

$$y = \beta_0 + \delta_1 D + \beta_1 x_1 + u$$

- **En la pendiente** (efecto multiplicativo) - representa la diferencia en el efecto (pendiente) entre los valores producidos por el cambio estructural.

$$y = \beta_0 + \beta_1 x_1 + \delta_1 D \cdot x_1 + u$$

**Contraste de Chow para cambio estructural** - analiza la existencia de cambio estructural en todos los parámetros del modelo, es una expresión particular del contraste F, donde  $H_0$ : No hay cambio estructural (todos  $\delta = 0$ ).

## Cambios de escala

Cambios en las **unidades de medida** de las variables:

- Sobre la variable **endógena**,  $y^* = y \cdot \lambda$  - afecta a todos los parámetros del modelo,  $\beta_j^* = \beta_j \cdot \lambda$ ,  $\forall j = 1, \dots, k$
- Sobre una variable **exógena**,  $x_j^* = x_j \cdot \lambda$  - sólo afecta al parámetro ligado a dicha variable exógena,  $\beta_j^* = \beta_j \cdot \lambda$
- Mismo cambio de escala sobre endógena y exógena - sólo afecta al término constante,  $\beta_0^* = \beta_0 \cdot \lambda$

## Cambios de origen

Cambios en el **origen de medida** de las variables (endógenas o exógenas),  $y^* = y + \lambda$  - sólo afectan al término constante del modelo,  $\beta_0^* = \beta_0 + \lambda$

## Multicolinealidad

- **Multicolinealidad perfecta** - hay variables independientes que son constantes y/o hay una relación lineal exacta entre variables independientes. Es el **incumplimiento del tercer (3) supuesto** del modelo.
- **Multicolinealidad aproximada** - hay variables independientes que son aproximadamente constantes y/o hay una relación lineal aproximada entre variables independientes. **No implica el incumplimiento de algún supuesto** del modelo, pero tiene un efecto en MCO.

### Consecuencias

- **Multicolinealidad perfecta** - el sistema de ecuaciones de MCO no puede resolverse (infinitas soluciones).
- **Multicolinealidad aproximada**
  - Pequeñas variaciones en la muestra producen grandes variaciones en las estimaciones de MCO.
  - La varianza de los estimadores MCO de las  $x$ 's que son colineales incrementa, la inferencia de los parámetros es afectada (intervalo de confianza grande).

### Detección

- **Análisis de correlación** - buscar altas correlaciones entre variables independientes,  $|r| > 0.7$ .
- **Factor de Inflación de la Varianza (FIV o VIF)** - indica el incremento en  $\text{Var}(\hat{\beta}_j)$  debido a la multicolinealidad.

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1-R_j^2}$$

donde  $R_j^2$  denota el R-cuadrado de una regresión entre  $x_j$  y todas las otras  $x$ 's.

- Valores entre 4 y 10 - pueden existir problemas de multicolinealidad.
  - Valores  $> 10$  - existen problemas de multicolinealidad.
- Una característica típica de la multicolinealidad es que los coeficientes de regresión del modelo no son individualmente significativos (por las altas varianzas), pero sí que son conjuntamente significativos.

### Corrección

- Eliminar una de las variables colineales.
- Realizar análisis factorial (u otra técnica de reducción de dimensiones) en las variables colineales.
- Interpretar los coeficientes con multicolinealidad conjuntamente.

## Heterocedasticidad

Los residuos  $u_i$  de la función de regresión poblacional no tienen una varianza constante  $\sigma_u^2$ :

$$\text{Var}(u \mid x_1, \dots, x_k) = \text{Var}(u) \neq \sigma_u^2$$

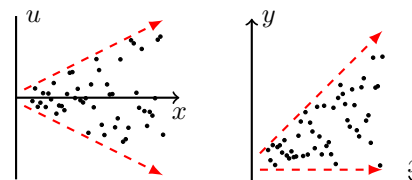
Es el **incumplimiento del quinto (5) supuesto** del modelo.

### Consecuencias

- Estimadores MCO son insesgados.
- Estimadores MCO son consistentes.
- MCO ya **no es eficiente**, pero sigue siendo ELI (Estimador Lineal Insesgado).
- La **estimación de la varianza** de los estimadores es **sesgada**: la construcción de intervalos de confianza y contraste de hipótesis no son fiables.

### Detección

- **Gráficos** - buscar patrones de dispersión en gráficos  $x$  vs.  $u$  ó  $x$  vs.  $y$ .



- **Contrastes** - White, Bartlett, Breusch-Pagan, etc. Generalmente,  $H_0$ : No heterocedasticidad.

### Corrección

- Usar MCO con un estimador de la matriz de varianzas-covarianzas robusto a la heterocedasticidad (HC), por ejemplo, la propuesta de White.
- Si la estructura de la varianza es conocida, usar Mínimos Cuadrados Ponderados (MCP) o Mínimos Cuadrados Generalizados (MCG):
  - Suponiendo que  $\text{Var}(u) = \sigma_u^2 \cdot x_i$ , dividir las variables del modelo entre la raíz cuadrada de  $x_i$  y aplicar MCO.
  - Suponiendo que  $\text{Var}(u) = \sigma_u^2 \cdot x_i^2$ , dividir las variables del modelo entre  $x_i$  (la raíz cuadrada de  $x_i^2$ ) y aplicar MCO.
- Si la estructura de la varianza es desconocida, hacer uso de Mínimos Cuadrados Ponderados Factibles (MCPF), que estima una posible varianza, divide las variables del modelo entre ella y entonces aplica MCO.
- Nueva especificación del modelo, por ejemplo, transformación logarítmica (reduce la varianza).

## Autocorrelación

El residuo de cualquier observación,  $u_t$ , está correlacionado con el residuo de cualquier otra observación. Las observaciones no son independientes.

$$\text{Corr}(u_t, u_s \mid x_1, \dots, x_k) = \text{Corr}(u_t, u_s) \neq 0, \quad \forall t \neq s$$

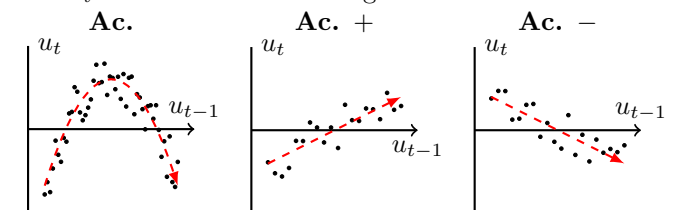
El contexto “natural” de este fenómeno son las series temporales. Es el **incumplimiento del sexto (6) supuesto** del modelo.

### Consecuencias

- Estimadores MCO son insesgados.
- Estimadores MCO son consistentes.
- MCO ya **no es eficiente**, pero sigue siendo ELI (Estimador Lineal Insesgado).
- La **estimación de la varianza** de los estimadores es **sesgada**: la construcción de intervalos de confianza y contraste de hipótesis no son fiables.

### Detección

- **Gráficos** - buscar patrones de dispersión en gráficos  $u_{t-1}$  vs.  $u_t$  o hacer uso del correlograma.



- **Contrastes** - Durbin-Watson, Breusch-Godfrey, etc. Generalmente,  $H_0$ : No autocorrelación.

### Corrección

- Usar MCO con un estimador de la matriz de varianzas-covarianzas robusto a la heterocedasticidad y autocorrelación (HAC), por ejemplo, la propuesta de Newey-West.
- Usar Mínimos Cuadrados Generalizados. Suponiendo  $y_t = \beta_0 + \beta_1 x_t + u_t$ , con  $u_t = \rho u_{t-1} + \varepsilon_t$ , donde  $|\rho| < 1$  y  $\varepsilon_t$  es ruido blanco.
  - Si  $\rho$  es conocido, crear un modelo cuasi-diferenciado donde  $u_t$  es ruido blanco y estimarlo por MCO.
  - Si  $\rho$  es desconocido, estimarlo -por ejemplo- por el método de Cochrane-Orcutt, crear un modelo cuasi-diferenciado donde  $u_t$  es ruido blanco y estimarlo por MCO.