

# Cheat Sheet Econometría

Por Marcelo Moreno - Universidad Rey Juan Carlos

## 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.

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

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

## Supuestos y propiedades

### Supuestos del modelo econométrico

Bajo estos supuestos, los estimadores de los parámetros MCO presentarán buenas propiedades. **Supuestos de Gauss-Markov extendidos:**

1. **Linealidad en parámetros** (+ 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:  $Var(x_j) \neq 0$
  - 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 no incluidas en el modelo:  $Cov(x_j, u) = 0$  for any  $j = 1, \dots, k \rightarrow$  **exogeneidad débil**.
5. **Homocedasticidad**. La variabilidad de los residuos es la misma para todos los niveles de  $x$ :  $Var(u|x) = \sigma^2$
6. **No autocorrelación**. Los residuos no contienen información sobre otros residuos:  $Corr(u_t, u_s|x) = 0$  para cualquier  $t \neq s$ . (SÓLO series temporales)
7. **Normalidad**. Los residuos son independientes e idénticamente distribuidos:  $u \sim N(0, \sigma^2)$
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:

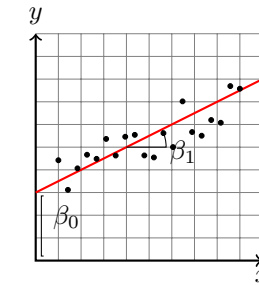
- De (1) a (4a): MCO es **insesgado**.  $E(\hat{\beta}_j) = \beta_j$
- De (1) a (4): MCO es **consistente**.  $plim(\hat{\beta}_j) = \beta_j$  (a (4b) sin (4a), exogeneidad estricta, insesg. y consistente)
- De (1) a (5): **normalidad asintótica** de MCO (entonces, (7) es necesariamente satisfecho):  $u \sim_a N(0, \sigma^2)$ .
- De (1) a (6): **estimador insesgado de  $\sigma^2$** .  $E(\hat{\sigma}^2) = \sigma^2$
- De (1) a (6): MCO es **MELI** (Mejor Estimador Lineal Insesgado) 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):

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

### Modelo de regresión simple



Ecuación:

$$y_i = \beta_0 + \beta_1 x_{1i} + u_i$$

Estimación:

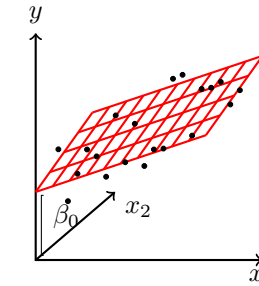
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$$

Donde:

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

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

### Modelo de regresión múltiple



Ecuación:

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + u_i$$

Estimación:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_k x_{ki}$$

Donde:

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

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

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

### Interpretación de coeficientes

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

### Medidas de error

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

Suma Expl. de Cuad.:  $SEC = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

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

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

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

Raíz Cuadrada del Error Cuadrático Medio:  $\sqrt{\frac{\sum_{i=1}^n (\hat{u}_i - \bar{u})^2}{n}}$

Error Medio Absoluto:  $\frac{\sum_{i=1}^n |\hat{u}_i|}{n}$

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

## 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 de las variaciones de  $y$ ) **y 1** (explicación total de las variaciones de  $y$ ).

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

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

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

## Contrastes de hipótesis

### Base de los contrastes de hipótesis

Un contraste de hipótesis 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 la hipótesis nula es rechazada.
- **Estadístico de contraste**: es una variable aleatoria cuya distribución de probabilidad es conocida bajo la hipótesis nula y está tabulada.
- **Nivel de significación ( $\alpha$ )** - es la probabilidad de rechazar la hipótesis nula siendo cierta (error Tipo I). Es elegido por quien conduce el contraste. Usualmente es 0.10, 0.05, 0.01 ó 0.001.
- **Valor crítico** - es el valor contra el cual se compara el estadístico de contraste para determinar si se rechaza o no la hipótesis nula.
- **p-valor** - es el nivel de significación máximo por el cual la hipótesis nula no puede ser rechazada ( $H_0$ ).

**Regla general**: si el p-valor es menor que  $\alpha$ , existe evidencia para rechazar la hipótesis nula a ese determinado  $\alpha$  (existe evidencia para aceptar la hipótesis alternativa).

### Contrastes individuales

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

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

Bajo  $H_0$ :

$$t = \frac{\hat{\beta}_j - \theta}{ee(\hat{\beta}_j)} \sim t_{n-k-1, \alpha/2}$$

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

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

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

Bajo  $H_0$ :

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

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

### 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:

- $\sum_{i=1}^n \hat{u}_{nr}^2$  - es la Suma de Resid. Cuadrados del modelo no restringido ( $SRC_{nr}$ ).
- $\sum_{i=1}^n \hat{u}_r^2$  - es la Suma de Resid. Cuadrados del modelo restringido ( $SRC_r$ ).

Bajo  $H_0$ :

$$F = \frac{SRC_r - SRC_{nr}}{SRC_{nr}} \frac{(n-k_{nr}-1)}{q} \sim F_{q, n-k_{nr}-1}$$

Donde  $k_{nr}$  es el número de parámetros del modelo no restringido y  $q$  es el número de hipótesis lineales a probar.

Si  $F_{q, n-k_{nr}-1} < F$ , existe evidencia para rechazar la hipótesis nula.

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

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0 \dots \text{ and/or } \beta_k \neq 0$$

Si  $F_{k, n-k_{nr}-1} < F$ , existe evidencia para rechazar la hipótesis nula.

## Intervalos de confianza

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

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

## Variables ficticias y cambio estructural

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

- Toman el **valor de 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 posición de las variables ficticias es importante:

- **En el intercepto ( $\beta_0$ )** - representa la diferencia media entre los valores producidos por el cambio estructural.
- **En los parámetros que determinan la pendiente de la línea de regresión ( $\beta_j$ )** - representa la diferencia en el efecto (pendiente) entre los valores producidos por el cambio estructural.

**El contraste de Chow para cambio estructural** - cuando se quiere analizar la existencia de cambio estructural en todos los parámetros del modelo, es común hacer uso de una expresión particular del contraste F, conocido como el contraste de Chow, donde la hipótesis nula es:  $H_0$  : No hay cambio estructural

## Predicciones

Dos tipos de predicciones:

- Del valor medio de  $y$  para un valor específico de  $x$ .
- De un valor individual de  $y$  para un valor específico de  $x$ .

Si los valores de las variables ( $x$ ) se aproximan al valor medio ( $\bar{x}$ ), la amplitud del intervalo de confianza de la predicción será menor.

## 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 (mayores a 0.7) entre variables independientes.
- **Factor de Inflación de la Varianza (FIV o VIF)** - indica el incremento en  $Var(\hat{\beta}_j)$  debido a la multicolinealidad.

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

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 sugieren que es recomendable analizar en mayor profundidad si pueden existir problemas de multicolinealidad.
- Valores por encima de 10 indican que 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 de manera conjunta.

## Heterocedasticidad

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

$$Var(u|x) = Var(y|x) \neq \sigma^2$$

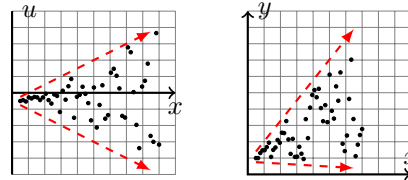
Es el **incumplimiento del quinto (5) supuesto del modelo**.

### 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.  $y$ .
- **Test formales** - White, Bartlett, Breusch-Pagan, etc. Comúnmente, la hipótesis nula:  $H_0 = \text{Homocedasticidad}$



### Corrección

- Usar MCO con un estimador de la matriz de varianzas-covarianzas robusto a la heterocedasticidad, por ejemplo, la propuesta de White.
- Si la estructura de la varianza es conocida, usar Mínimos Cuadrados Ponderados (MCP) o Mínimos Cuadrados Generalizados (MCG).
- 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.
- Hacer suposiciones sobre la posible varianza:
  - Suponiendo que  $\sigma_i^2$  es proporcional a  $x_i$ , dividir las variables del modelo entre la raíz cuadrada de  $x_i$  y aplicar MCO.
  - Suponiendo que  $\sigma_i^2$  es proporcional a  $x_i^2$ , dividir las variables del modelo entre  $x_i$  y aplicar MCO.
- Nueva especificación del modelo, por ejemplo, transformación logarítmica.

## Autocorrelación

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

$$Corr(u_t, u_s|x) \neq 0 \text{ para cualquier } t \neq s$$

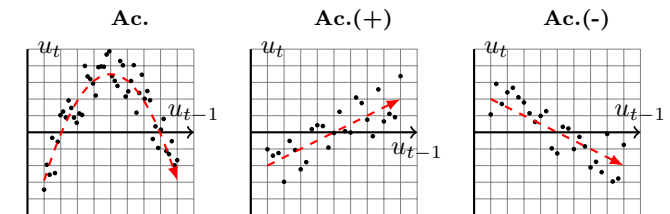
El contexto “natural” de este fenómeno son las series temporales. Es el **incumplimiento del sexto (6) supuesto del modelo**.

### 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.



- **Test formales** - Durbin-Watson, Breusch-Godfrey, etc. Comúnmente, la hipótesis nula:  $H_0 : \text{No autocorrelación}$

### Corrección

- Usar MCO con un estimador de la matriz de varianzas-covarianzas robusto a la autocorrelación, 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.