Los estimadores mínimo cuadráticos bajo los supuestos clásicos

Propiedades estadísticas e inferencia

Mariana Marchionni
marchionni.mariana@gmail.com

Econometría I - FCE - UNLP
econometria1unlp.weebly.com

Temario de la clase

1. Repaso
2. El modelo lineal clásico
3. Propiedades estadísticas de los estimadores
4. Inferencia

Estimación por MCO del modelo lineal simple

Tenemos el modelo lineal con 2 variables

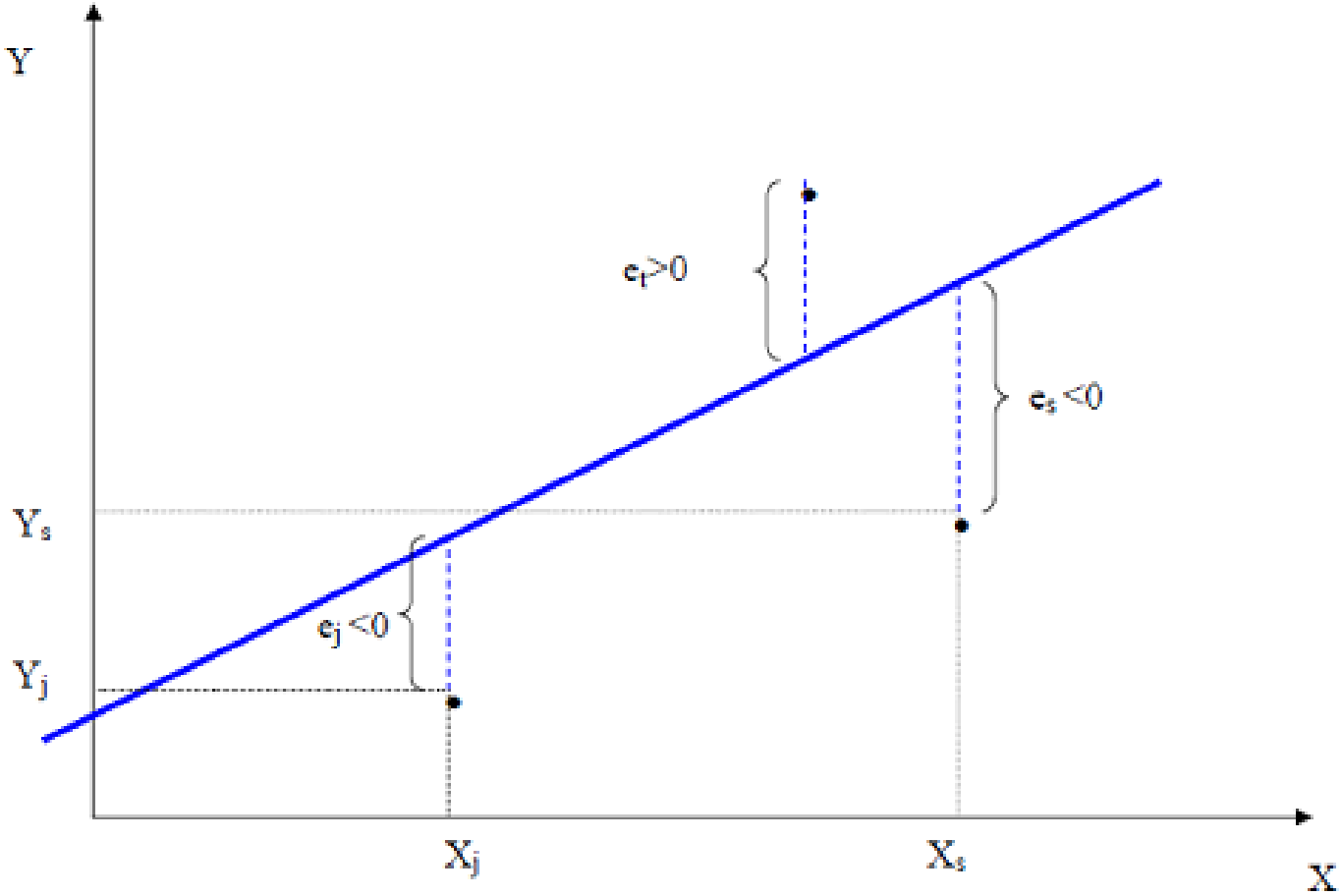
$$Y_i = \alpha + \beta X_i + \mu_i \qquad i = 1, ...n$$

Y el objetivo es estimar α y β a partir de (Y_i, X_i) $i = 1, ...n$.

Recordemos notación y definiciones:

- $\hat{\alpha}$ y $\hat{\beta}$ son los estimadores de α y β
- $\hat{Y}_i \equiv \hat{\alpha} + \hat{\beta} X_i$ es la versión estimada de Y_i
- $e_i \equiv Y_i - \hat{Y}_i = Y_i - (\hat{\alpha} + \hat{\beta} X_i)$ es el error de estimación (residuo)

Cada posible valor que asignemos a $\hat{\alpha}$ y $\hat{\beta}$ define una recta en el plano (Y, X)



Mínimos Cuadrados Ordinarios (MCO)

Criterio de MCO: elegir $\hat{\alpha}$ y $\hat{\beta}$ de manera de minimizar la **S**uma de los **R**esiduos al **C**uadrado (SRC).
Formalmente:

$$Min_{\hat{\alpha},\hat{\beta}}\ SRC(\hat{\alpha},\hat{\beta}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}[Y_i - (\hat{\alpha} + \hat{\beta}X_i)]^2$$

$$CPO: \begin{cases} (1) & \frac{\partial SRC}{\partial \hat{\alpha}} = 0 \\ \\ (2) & \frac{\partial SRC}{\partial \hat{\beta}} = 0 \end{cases}$$

Solución:

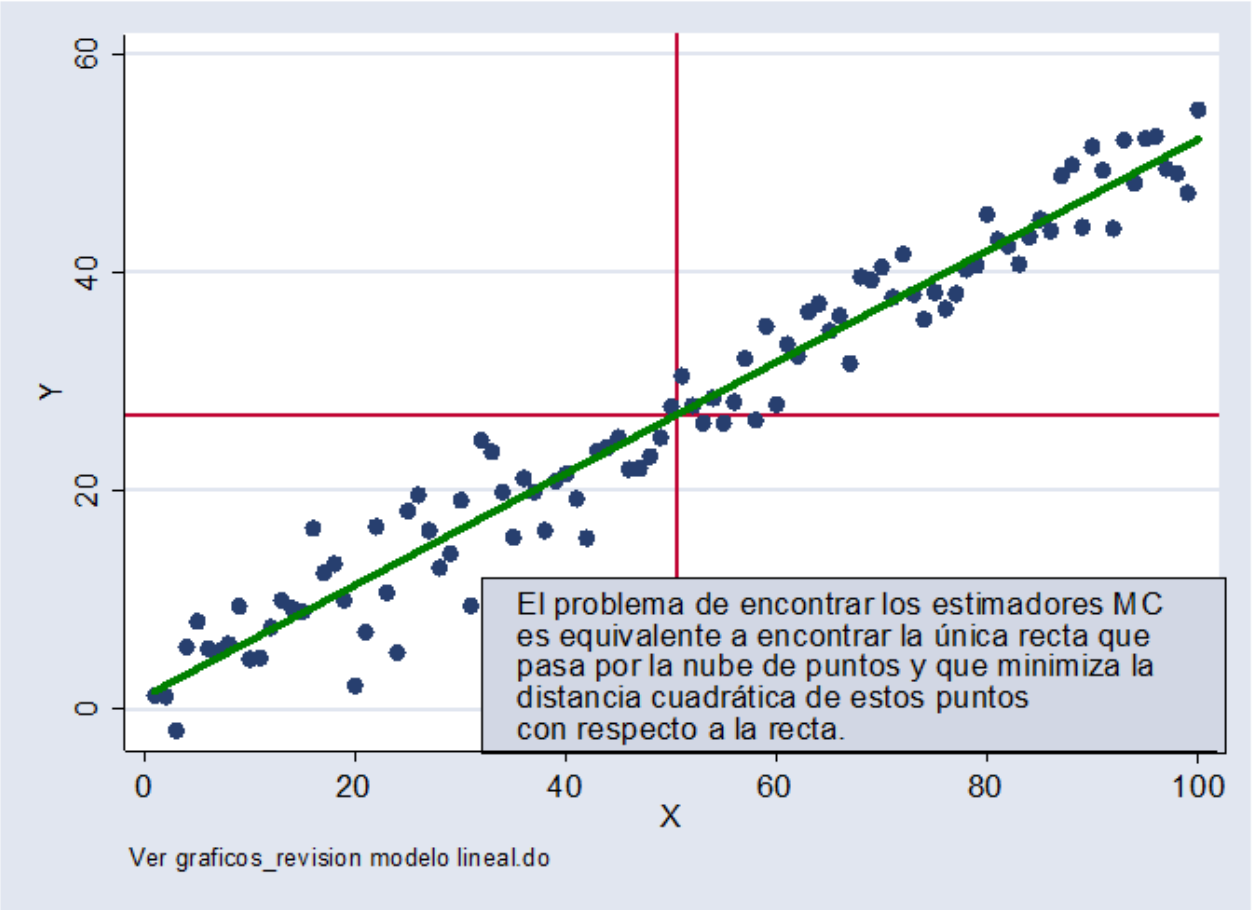
$$(3)\ \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \text{ y } (4)\ \hat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{Y}\bar{X}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} \text{ o } (4')\ \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Propiedades algebraicas

1. $\sum_{i=1}^{n} e_i = 0$
2. $\sum_{i=1}^{n} e_i X_i = 0 \quad (\Rightarrow Cov(X, e) = 0)$
3. $\hat{Y}(\bar{X}) = \bar{Y}$ (la recta de regresión pasa por las medias muestrales)
4. $\hat{\beta} = r_{Y,X} s_Y / s_X$ (relación entre el coeficiente de regresión y el de correlación)
5. $\bar{Y} = \bar{\hat{Y}}$ (la media de las estimaciones de Y coincide con $\bar{Y}$)
6. $\hat{\beta} = \sum \omega_i Y_i$ ($\hat{\beta}$ es una función lineal de las Y_i, donde ω_i dependen únicamente de las X_i)
7. $r_{\hat{Y},e} = 0$ (la correlación entre las estimaciones de Y y los residuos es nula)

La regresión estimada por MCO

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$



Bondad del ajuste

- Vimos que cuando se estima por MCO es válida la siguiente descomposición:

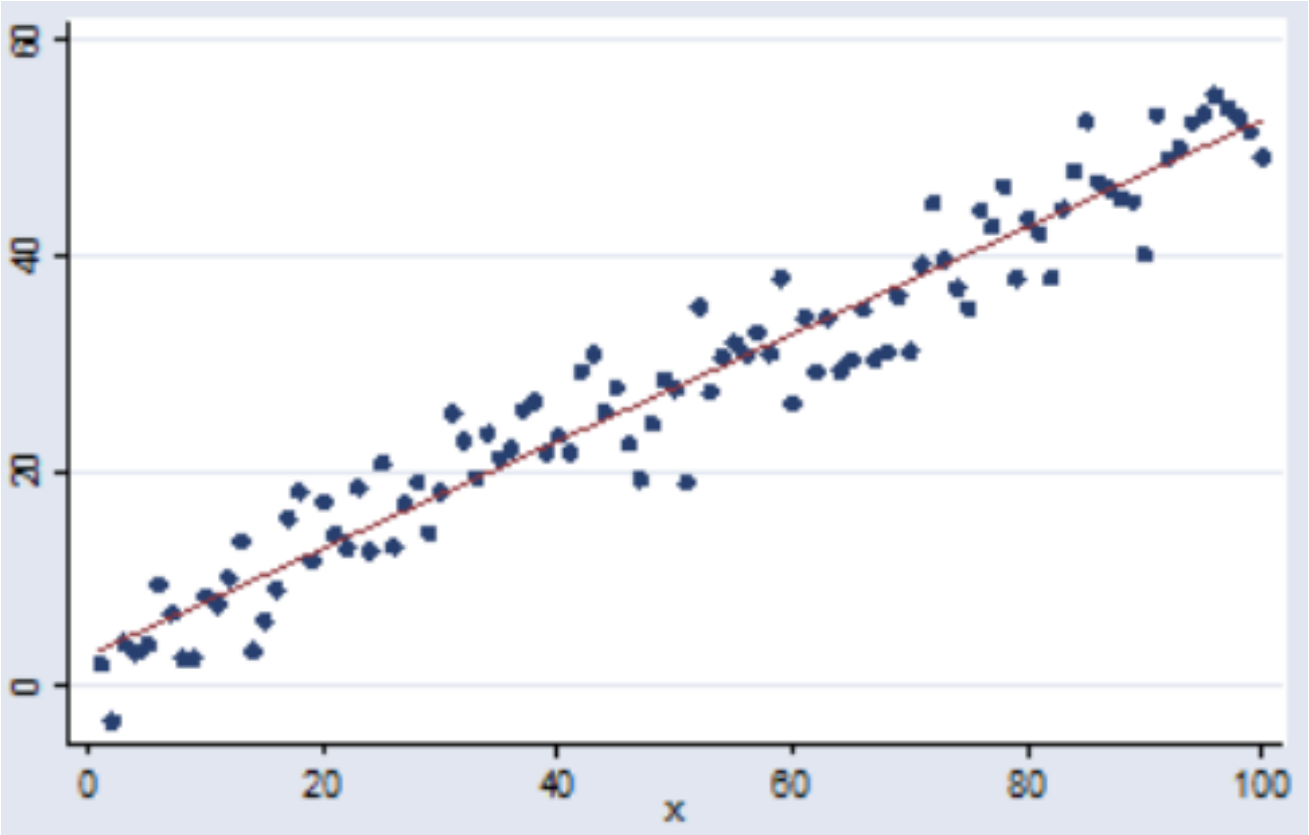
$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{STC} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{SEC} + \underbrace{\sum_{i=1}^{n}e_i^2}_{SRC}$$

STC: Suma Total de Cuadrados, SEC: Suma Explicada de Cuadrados, SRC: Suma de los Residuos al Cuadrado.

- Definimos la medida de bondad del ajuste R^2 como:

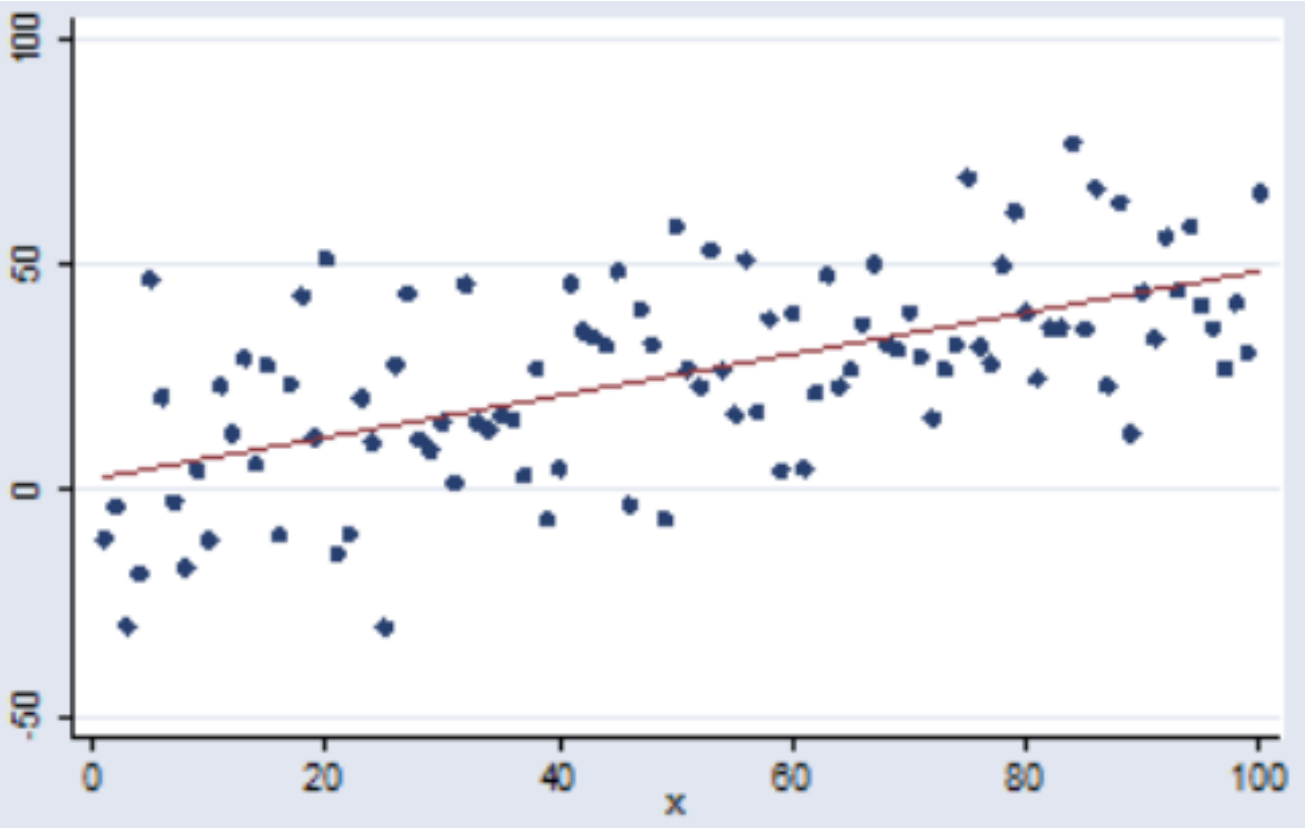
$$R^2 = \frac{SEC}{STC} = 1 - \frac{SRC}{STC}$$

R^2=0.94



94 % de la variabilidad muestral de Y está explicada por el modelo lineal en X

R^2=0.35



35 % de la variabilidad muestral de Y está explicada por el modelo lineal en X

El modelo lineal clásico

Adoptaremos algunos supuestos llamados "supuestos clásicos"

- ***Linealidad***: $Y_i = \alpha + \beta X_i + \mu_i, \ i = 1, ...n$
- ***Esperanza nula*** *de* μ_i: $E[\mu_i] = 0, \ i = 1, ..., n.$
- ***Homocedasticidad*** *de* μ_i: $V[\mu_i] = cte. \equiv \sigma^2, \ i = 1, ..., n.$
- ***No correlación serial*** *de* μ_i: $Cov[\mu_i, \mu_j] = 0, \ \forall i \neq j.$
- X_i ***no aleatorias***: las X_i son determinísticas o "fijas en muestreos repetidos".
- X_i **no pueden ser todas iguales**: los valores de X_i, $i = 1, ...n$ no pueden ser todos iguales (*no multicolinealidad perfecta*).

Algunos de estos supuestos pueden no ser realistas. Levantaremos la mayoría a lo largo del curso.

Linealidad

- Nuestro modelo:

$$Y_i = \alpha + \beta X_i + \mu_i, \ i = 1, ...n$$

- Lo importante es la linealidad en los parámetros α y β. ¿Por qué?

- Vimos algunos modelos no lineales en las variables pero que son lineales en los parámetros:
  - log-log: $lnY_i = \alpha + \beta lnX_i + \mu_i$
  - semi-log: $lnY_i = \alpha + \beta X_i + \mu_i$

Esperanza nula

$$E[\mu_i] = 0 \ , \ \ i = 1, ..., n$$

- Esto implica que, tomando a las X_i como fijas, $E[Y_i] = \alpha + \beta X_i, \ \ i = 1, ..., n$
- En promedio, la relación entre Y y X es lineal y exacta

Homocedasticidad

$$V[\mu_i] = V[\mu_j], \ \ \forall\, i \neq j$$

$$V[\mu_i] = cte. \equiv \sigma^2, \ \ i = 1, ..., n$$

- ¿Qué mide la varianza de una variable con media nula?
- Intuitivamente: para todas las observaciones, la relación entre Y y X está "igual de cerca" de una relación lineal
- Si la varianza no es constante decimos que hay *heterocedasticidad*
- Notar: si $E[\mu_i] = 0$ y $V[\mu_i] = \sigma^2 \Rightarrow E[\mu_i^2] = \sigma^2$

No correlación serial (o no autocorrelación)

$$Cov[\mu_i, \mu_j] = 0, \ \forall i \neq j.$$

- Es una forma débil de independencia entre los términos aleatorios de distintas observaciones
- Notar que si $E[\mu_i] = 0$ y $Cov[\mu_i, \mu_j] = 0 \Rightarrow E[\mu_i \mu_j] = 0$
- ¿Cuánto vale $Cov[\mu_i, \mu_j]$ para $i = j$?

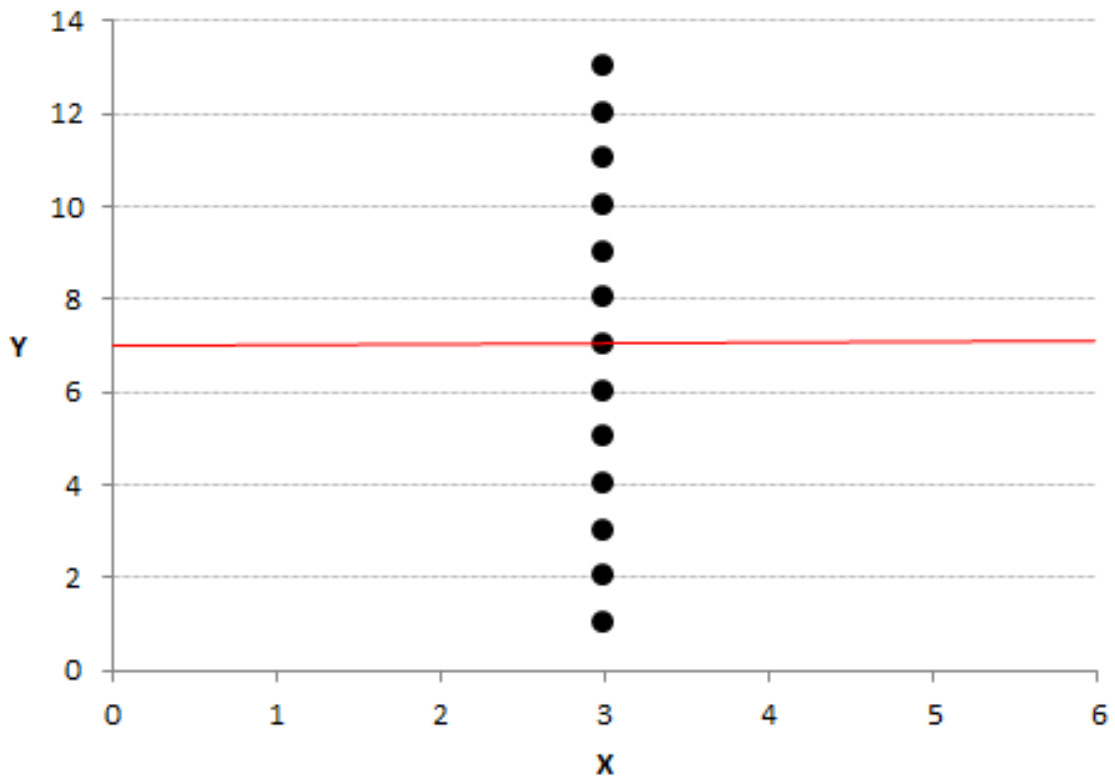
X no aleatoria

- Tratamos a las X_i como números fijos (*fijas en muestreos repetidos*).
  - Tiene sentido en un contexto experimental, donde el investigador decide cuánto valen las X_i y luego mide Y_i.
  - Además, el experimento se puede repetir usando los mismos valores de las X_i.
- No es lo más común en las ciencias sociales.
  - Los datos usualmente no surgen de experimentos sino de la observación del comportamiento de personas.
- Uso de experimentos en Economía
  - Se asigna un "tratamiento" a un grupo y se monitorea a otro grupo llamado "control", que no recibe el tratamiento.
  - Esto **permite estimar el efecto causal de una intervención o tratamiento en algún resultado de interés**.
  - El Premio Nobel de Economía 2019 fue para Banerjee, Duflo y Kremer por sus contribuciones al estudio del desarrollo económico y del combate a la pobreza con estas herramientas. Ver esta entrada en el Blog del CEDLAS.

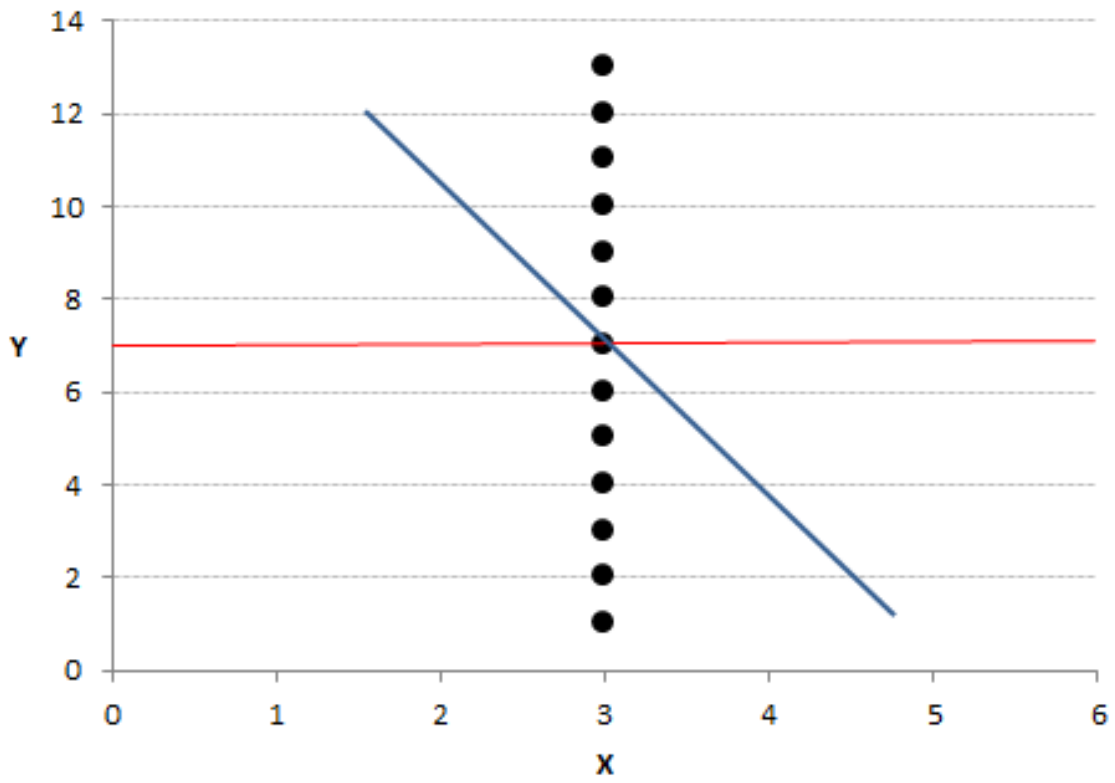
Las X_i no pueden ser todas iguales (no multicolinealidad perfecta)

- En el modelo lineal con 2 variables, este supuesto requiere que la variable X no tenga el mismo valor para todas las observaciones.
- En otras palabras: $X_i, \quad i = 1, ...n$ no son todas iguales
- ¿Por qué?

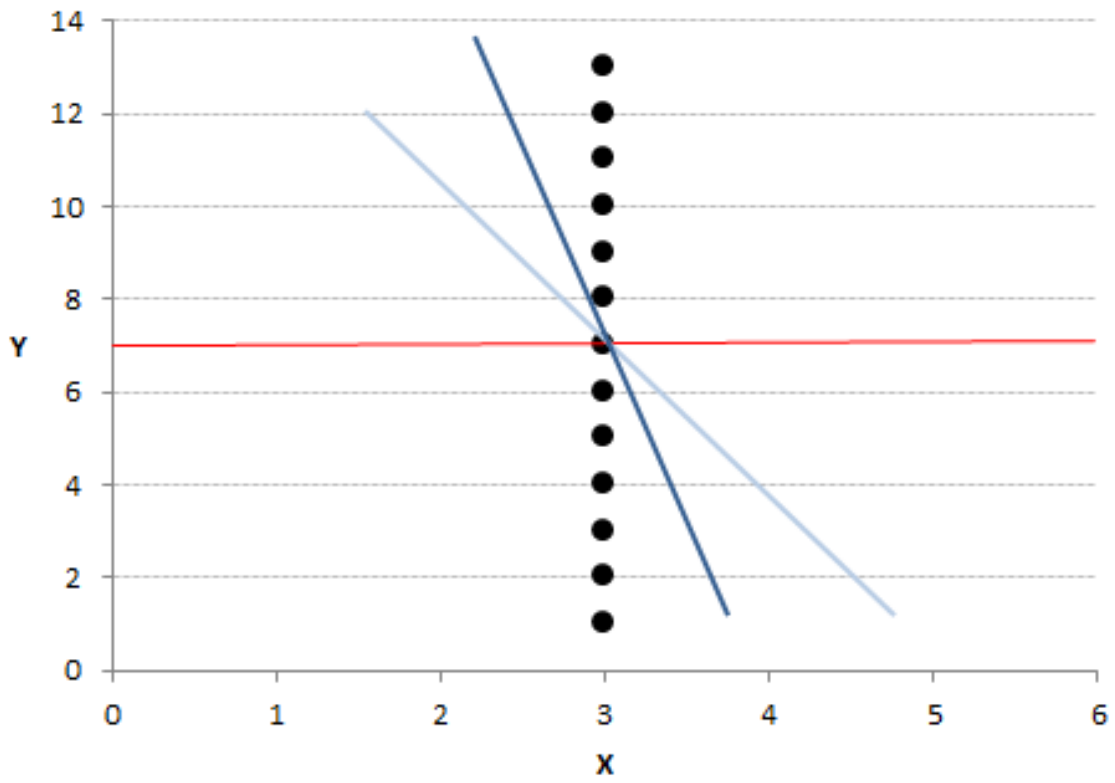
Si todas las X_i son iguales...



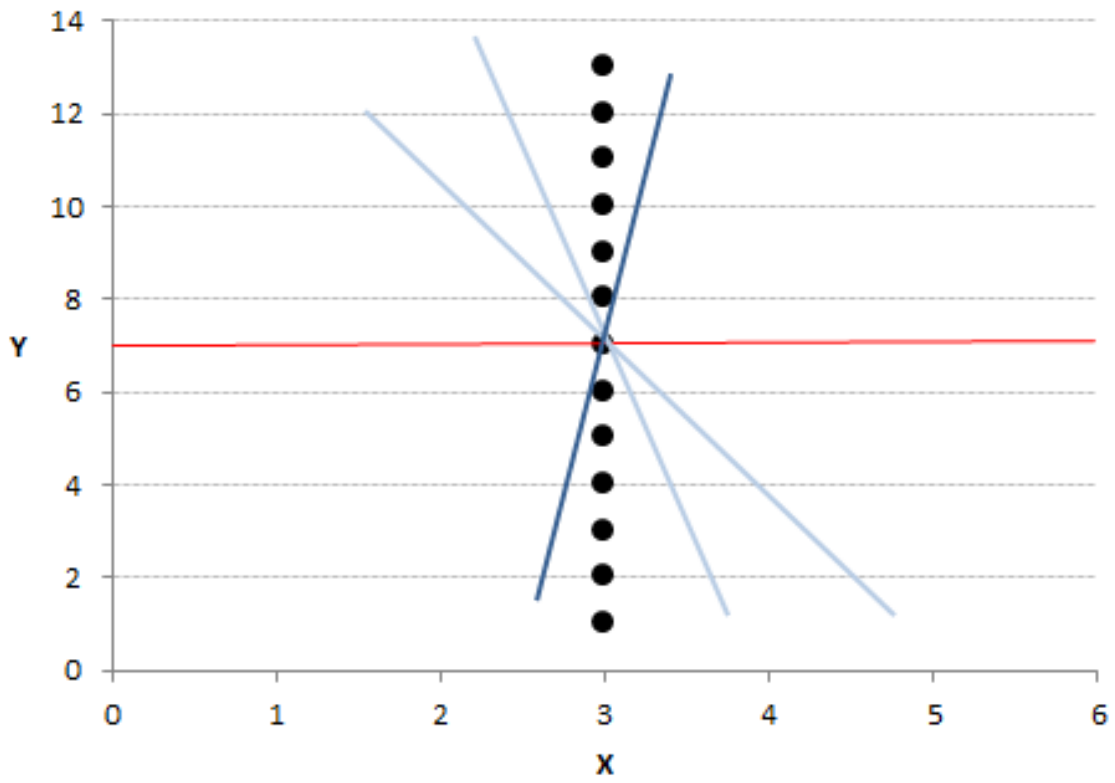
Si todas las X_i son iguales...



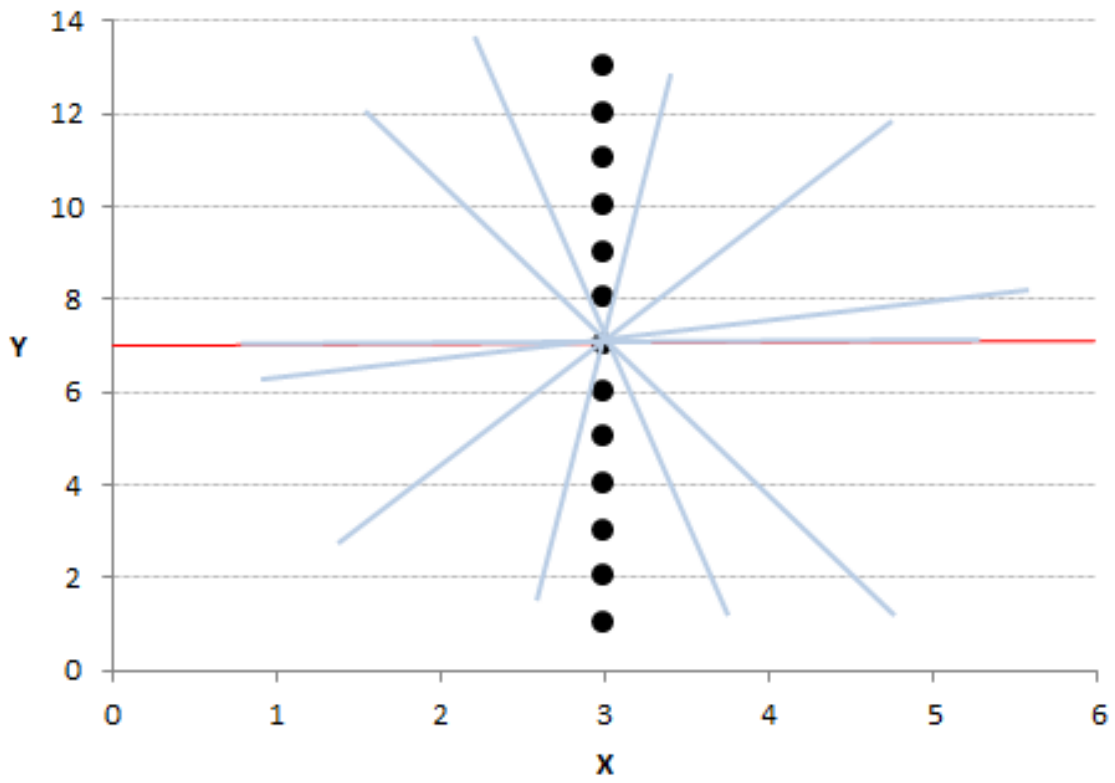
Si todas las X_i son iguales...



Si todas las X_i son iguales...



Si todas las X_i son iguales...



Entonces, ¿por qué las X_i no pueden ser todas iguales?

- Una respuesta más formal: si las X_i son todas iguales, las condiciones de primer orden del problema de MCO son linealmente dependientes $\Rightarrow \nexists$ una única solución. Probarlo.

- ¿Intuición? Estimar β implica que le estamos preguntando a los datos cuánto cambia $E[Y]$ cuando cambia X. Si las X_i no varían en la muestra, los datos no pueden responder esa pregunta.

Recapitulando... el modelo lineal clásico

$Y_i = \alpha + \beta X_i + \mu_i, \ i = 1, ...n$

1. $E[\mu_i] = 0, \ i = 1, ..., n.$
2. $V[\mu_i] = \sigma^2, \ i = 1, ..., n.$
3. $Cov[\mu_i, \mu_j] = 0, \ i \neq j.$
4. Las X_i no son aleatorias y no son todas iguales

Notar: en el modelo lineal clásico (con 2 variables) los parámetros desconocidos son α, β y σ^2

Rol de los supuestos clásicos

- Ya sabemos que los estimadores mínimo cuadráticos $\hat{\alpha}$ y $\hat{\beta}$ son "buenos" en el sentido que minimizan la SRC
- En la derivación de los estimadores MCO de la clase pasada nunca recurrimos a los supuestos clásicos
- Entonces, ¿para qué hacemos los supuestos clásicos?
- Vamos a explorar si los estimadores MCO son buenos en algún otro sentido, ahora sí basándonos en los supuestos clásicos
- Nuevas propiedades: insesgadez, varianza mínima, etc.
- Nos concentramos en $\hat{\beta}$. Los resultados para $\hat{\alpha}$ serán cubiertos en los prácticos.

Primero, recordemos una de las propiedades algebraicas

- En la propiedad algebraica Nro. 6 vimos que $\hat{\beta}$ **es una función lineal de las** Y_i**,** es decir, que puede escribirse

$$\hat{\beta} = \sum \omega_i Y_i$$

  donde ω_i dependen únicamente de las X_i. Además, con el supuesto 4 tenemos que ω_i son números (no aleatorias).
- En este sentido decimos que $\hat{\beta}$ **es un estimador lineal.**
- Repasemos la demostración (ver cap. 1.4 de notas de clase)

$\hat{\beta}$ es un estimador lineal, es decir, $\hat{\beta} = \sum \omega_i Y_i$

Demostración: Comenzaremos escribiendo a $\hat{\beta}$ como sigue:

$$\hat{\beta} = \sum \left(\frac{x_i}{\sum x_i^2} \right) y_i$$

y definamos $w_i = x_i / \sum x_i^2$. Notar que:

$$\sum x_i = \sum (X_i - \bar{X}) = \sum X_i - n\bar{X} = 0$$

lo que implica $\sum w_i = 0$. Del resultado anterior:

$$\begin{aligned} \hat{\beta} &= \sum w_i y_i \\ &= \sum w_i (Y_i - \bar{Y}) \\ &= \sum w_i Y_i - \bar{Y} \sum w_i \\ &= \sum w_i Y_i \end{aligned}$$

- Intuitivamente poco interesante, analíticamente conveniente

Las propiedades estadísticas

1. Insesgadez: $E[\hat{\beta}] = \beta$, es decir, $\hat{\beta}$ es un estimador insesgado del parámetro β

- Ver demostración en cap. 1.6 de notas de clase.

- *$\hat{\beta}$ es un estimador insesgado*, esto es: $E(\hat{\beta}) = \beta$

Demostración:

$$\begin{aligned}
\hat{\beta} &= \sum w_i y_i \\
E(\hat{\beta}) &= \sum w_i E(y_i) \qquad \text{(los } w_i \text{ son no estocásticos)} \\
&= \sum w_i x_i \beta \\
&= \beta \sum w_i x_i \\
&= \beta \sum x_i^2 / (\sum x_i^2) \\
&= \beta
\end{aligned}$$

- ¿Qué supuestos usamos en esta demostración?
- ¿Qué pasa con la propiedad de insesgadez si levantamos el supuesto de homocedasticidad?
- Intuitivamente: ¿qué quiere decir que un estimador sea insesgado?

Insesgadez: intuición

- Supongamos que el verdadero modelo es $Y_i = 2 + 5X_i + \mu_i$
  - $\alpha = 2$ y $\beta = 5$
  - recta de regresión verdadera: $E[Y_i] = 2 + 5X_i$
- Vamos a actuar como si no conociéramos los verdaderos valores de los parámetros:
  - tomamos una muestra de n observaciones
  - estimamos α y β por MCO
- **Primera iteración**. Con la realización de la muestra (datos) obtenemos una estimación: a_1 y b_1
  - Notar que a_1 y b_1 son números, no son aleatorios
  - Seguramente $a_1 \neq 2$ y $b_1 \neq 5$. ¿Por qué?

Insesgadez: intuición

- Repetimos el ejercicio J veces: cada vez obtenemos una nueva muestra de n observaciones de la misma población y volvemos a estimar
  - En la **segunda iteración** obtenemos otra estimación de los parámetros: a_2 y b_2
  - En la **tercera iteración** obtenemos los valores a_3 y b_3
  - ...
  - En la **J-ésima iteración** obtenemos a_J y b_J

Insesgadez: intuición

# iteración	realización de $\hat{\alpha}$	realización de $\hat{\beta}$
1	a_1	b_1
2	a_2	b_2
3	a_3	b_3
$\vdots$	$\vdots$	$\vdots$
J	a_J	b_J
$\sum(...)/J$	$\bar{a}$	$\bar{b}$

- Si J es lo suficientemente grande el promedio de los valores estimados coincide con el valor esperado del estimador (LGN).
- Y como los estimadores son insesgados bajo los supuestos clásicos, el valor esperado del estimador coincide con el verdadero valor del parámetro
  - $\bar{a} = E[\hat{\alpha}] = \alpha = 2$
  - $\bar{b} = E[\hat{\beta}] = \beta = 5$

Insesgadez: intuición

- Interpretación: si usamos un estimador insesgado y pudiésemos repetir el experimento muchas veces obtendríamos el verdadero valor del parámetro.
- Si usamos un estimador insesgado **esperamos** obtener el verdadero valor del parámetro

Las propiedades estadísticas

2. $V[\hat{\beta}] = \sigma^2 / \sum x_i^2$

- Ver demostración en cap. 1.6 de las notas de clase
- ¿Qué supuestos usamos para la prueba?

¿Qué mide $V[\hat{\beta}]$?

- Si $\hat{\beta}$ es insesgado, $V[\hat{\beta}]$ mide la dispersión del estimador alrededor del verdadero β
- Cuanto menor sea $V[\hat{\beta}]$, más eficiente es el estimador $\hat{\beta}$
- ¿De qué depende la magnitud de $V[\hat{\beta}]$?

$$V[\hat{\beta}] = \frac{\sigma^2}{\sum x_i^2} = \frac{\sigma^2}{nV[X]}$$

  - depende de la dispersión del término aleatorio (σ^2)
  - de la cantidad de observaciones (n)
  - de la variabilidad de X (es decir, de $V(X)$)
    - que X tenga poca variabilidad muestral tiene el mismo efecto que tener pocas observaciones

Las propiedades estadísticas

3. Teorema de Gauss-Markov (TGM): si los supuestos clásicos se cumplen, $\hat{\beta}$ tiene la menor varianza entre todos los estimadores lineales e insesgados de β.

- En el modelo lineal clásico, los estimadores MCO son los ***Mejores Estimadores Lineales e Insesgados*** **(MELI)**
- "Mejores" en el sentido de "más eficientes" dentro del grupo de estimadores lineales e insesgados
- Entonces, ¿los estimadores MCO son buenos?
- **Los supuestos clásicos son condiciones necesarias y suficientes del TGM:**
  - **suficientes:** basta que se cumplan para que saber que los estimadores MCO son MELI.
  - **necesarios**: con que solo uno de los supuestos no se cumpla, los estimadores MCO ya no serán los MELI.

Inferencia en el modelo lineal clásico con 2 variables

Consideremos el modelo $Y_i = \alpha + \beta X_i + \mu_i, \quad i = 1, ...n$

- Sabemos cómo producir estimaciones puntuales de α y β
- ¿Qué podemos hacer si quisiéramos evaluar la veracidad de una hipótesis como $\beta = 0$? (test de hipótesis)
- Necesitamos asociar alguna medida de confianza a la estimación puntual

Test de hipótesis

- Hipótesis: una conjetura acerca de un parámetro desconocido
- Por ejemplo: $H_0 : \beta = 0$
- Puede ser cierta o falsa
- Si pudiéramos observar β no hay problema
- ¿Y si no conocemos β? No sabemos realmente
- Pero esperamos poder conocer algo acerca de si H_0 es cierta o falsa

Supongamos que estamos interesados en distinguir entre $H_0 : \beta = 0$ y $H_A : \beta \neq 0$, pero sin observar β.

- Podemos usar $\hat{\beta}$
- $\hat{\beta}$ es una variable aleatoria. ¿Por qué?
- Como las realizaciones de $\hat{\beta}$ pueden tomar cualquier valor tanto bajo H_0 como bajo H_A, no podemos resolver nuestro problema observando los valores que toma $\hat{\beta}$ para una muestra particular.

- Supongamos que $\hat{\beta}$ es un "buen" estimador
- Si $H_0 : \beta = 0$ es cierta, entonces, aunque $\hat{\beta}$ pueda tomar cualquier valor, esperamos que tome valores "cercanos" a cero. ¿Por qué?
- Con esta lógica, podemos sospechar que H_0 es falsa si $\hat{\beta}$ toma valores "lejanos" de cero: *rechazamos* H_0.
- Problema: ¿cómo definimos qué valores están "cerca" y qué valores estan "lejos" del cero?

- A fines de definir qué valores de $\hat{\beta}$ están "cerca" y qué valores están "lejos" del cero en el caso de que H_0 fuera cierta, necesitamos conocer la distribución de $\hat{\beta}$ (*distribución de* $\hat{\beta}$ bajo H_0)
- A partir de las propiedades de $\hat{\beta}$ bajo los supuestos clásicos, sabemos que si $H_0 : \beta = 0$ entonces $E[\hat{\beta}] = 0$. ¿Por qué? También sabemos su varianza.
- Pero no tenemos información suficiente para conocer toda la distribución de $\hat{\beta}$ bajo H_0
- ¿Dónde conseguimos esa información?

Nuevo supuesto: normalidad del término aleatorio

$$\mu_i \sim N(0, \sigma^2)$$

- Con este nuevo supuesto, tanto Y_i, $i = 1, ..., n$, como $\hat{\beta}$ (y también $\hat{\alpha}$) son variables aleatorias con **distribución normal**. ¿Por qué?
  - Recordar que cualquier función lineal de una variable normal es también normal
  - Y que el modelo de Y_i viene dado por: $Y_i = \alpha + \beta X_i + \mu_i$
  - Y que $\hat{\beta}$ se puede escribir como $\hat{\beta} = \sum \omega_i Y_i$

- Entonces, con los supuestos clásicos más el supuesto de normalidad tenemos:

$$\hat{\beta} \sim N(\beta, \sigma^2 / \sum x_i^2)$$

- De manera que cuando es cierta $H_0 : \beta = 0$, se cumple:

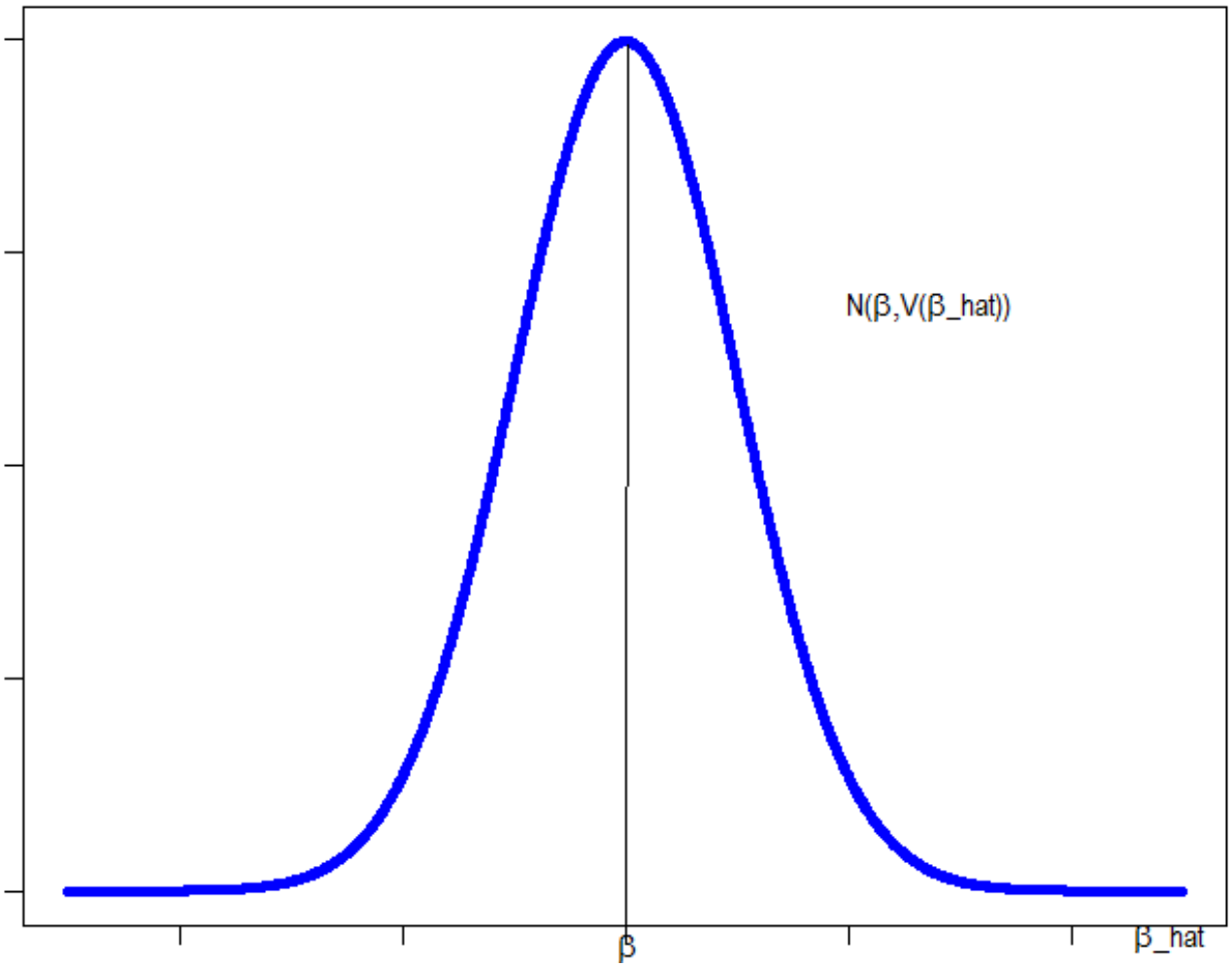
$$\hat{\beta} \sim N(0, \sigma^2 / \sum x_i^2)$$

- También se cumple:

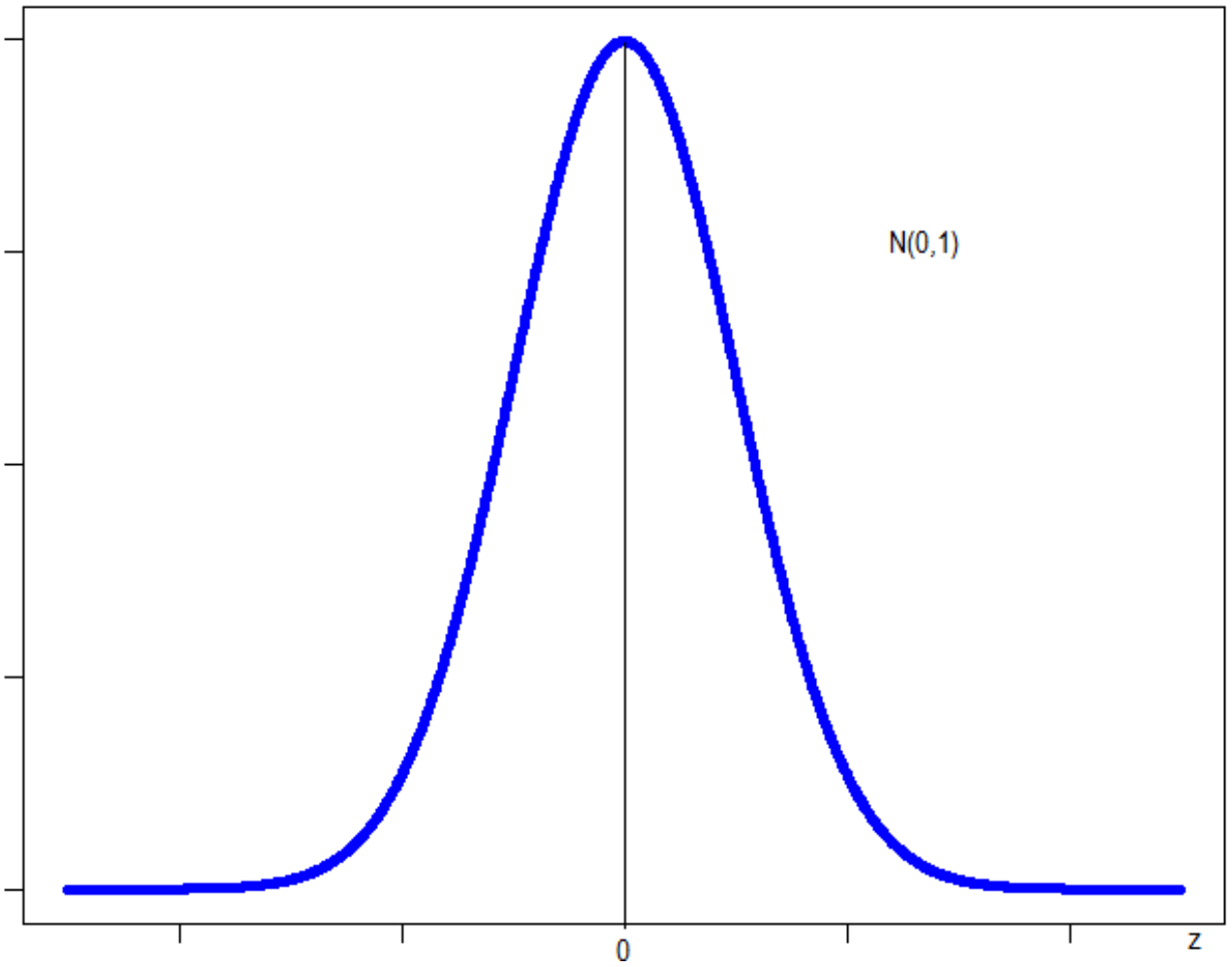
$$z \equiv \frac{\hat{\beta} - 0}{\sqrt{\sigma^2 / \sum x_i^2}} \sim N(0, 1)$$

- Cuando $\hat{\beta}$ es "chico", z es "chico"

Distribución de $\hat{\beta}$



Distribución de z (estandarización de $\hat{\beta}$)



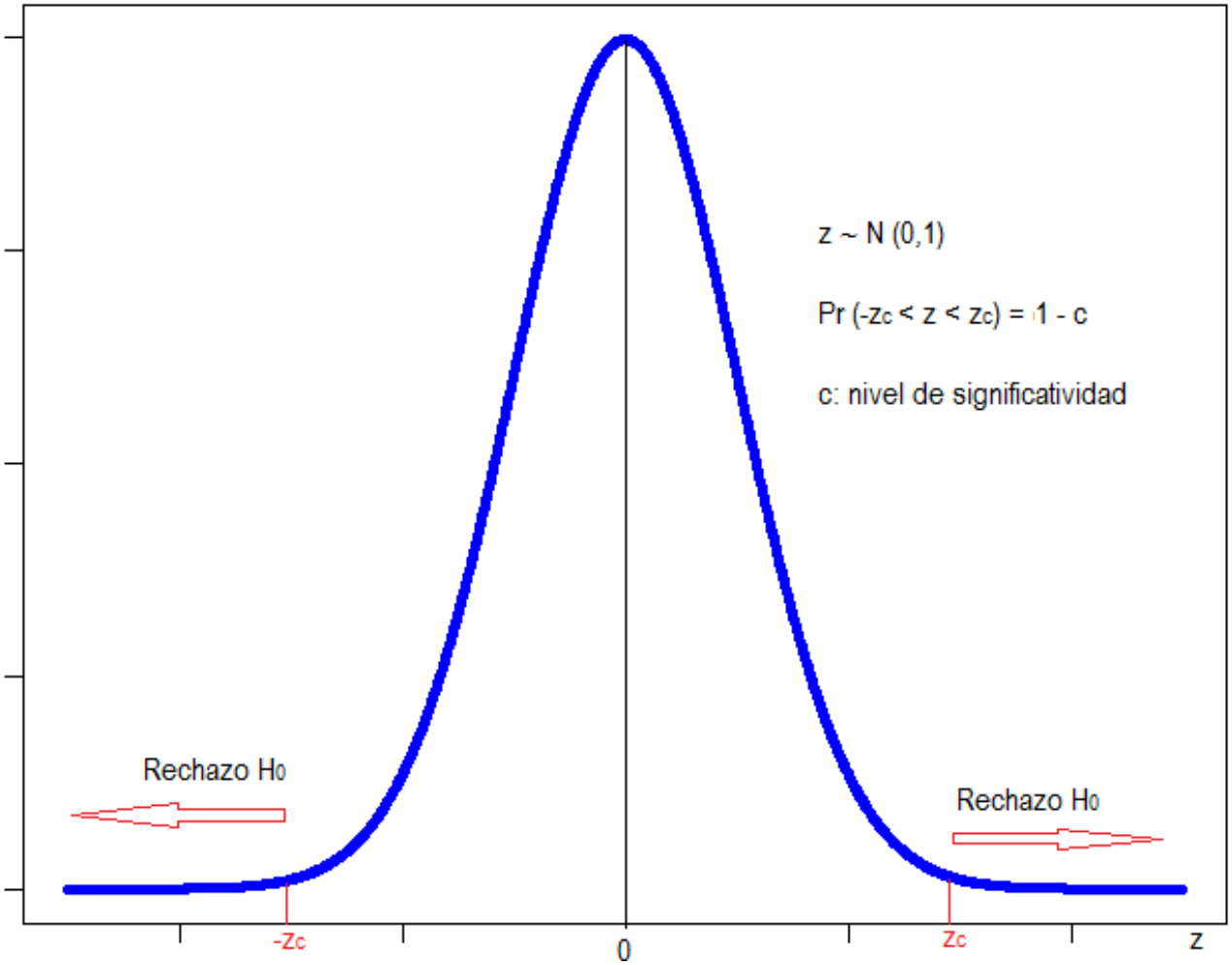
- **Regla de decisión para el test de hipótesis: *aceptamos* H_0 si la realización de $\hat{\beta}$ está cerca del valor que postula esa hipótesis nula.**

- Para definir "cercano" elegimos una constante $c \in (0, 1)$ tal que:

$$\Pr\left[-z_c \leq z \leq z_c\right] = 1 - c$$

  - Notar: la elección de c determina los valores $-z_c$ y z_c.

- Si el valor de z (estandarización de $\hat{\beta}$) está entre los límites $-z_c$ y z_c, concluimos que $\hat{\beta}$ está "cerca" de lo que postula H_0 y aceptamos esa hipótesis.


z ~ N (0,1)
Pr (-zc < z < zc) = 1 - c
c: nivel de significatividad
Rechazo H0
Rechazo H0
-Zc
0
Zc
z

- La región de aceptación de H_0 en términos de z está dada por:

$$\Pr\left[-z_c \leq z \leq z_c\right] = 1 - c$$

- Reemplazando z por su definición:

$$\Pr\left[-z_c\left(\sqrt{\sigma^2/\sum x_i^2}\right) \leq \hat{\beta} \leq z_c\left(\sqrt{\sigma^2/\sum x_i^2}\right)\right] = 1 - c$$

- Entonces la región de aceptación de H_0 viene dada por los valores de $\hat{\beta}$ en el intervalo:

$$0 \pm z_c\left(\sqrt{\sigma^2/\sum x_i^2}\right)$$

- Aceptamos H_0 si $\hat{\beta}$ "cae" en ese intervalo. En caso contrario, rechazamos H_0
- Notar: elegimos c de antemano.
- c es la probabilidad de rechazar una H_0 cuando es cierta (error tipo I) y se conoce como el nivel de **significatividad del test**.
- Al elegir de antemano el nivel de significatividad del test, estamos resignándonos a cometer el error tipo I un $c \times 100\%$ de las veces. ¿Por qué aceptamos cometer errores tipo I?
- Dado c, z_c puede obtenerse de una tabla de percentiles de la distribución normal estándar
- Luego las zonas de aceptación y rechazo se computan usando $0 \pm z_c \left(\sqrt{\sigma^2 / \sum x_i^2}\right)$

Problema: σ^2 no se observa!

- Estimador insesgado de σ^2:

$$S^2 = \frac{\sum e_i^2}{n-2}$$

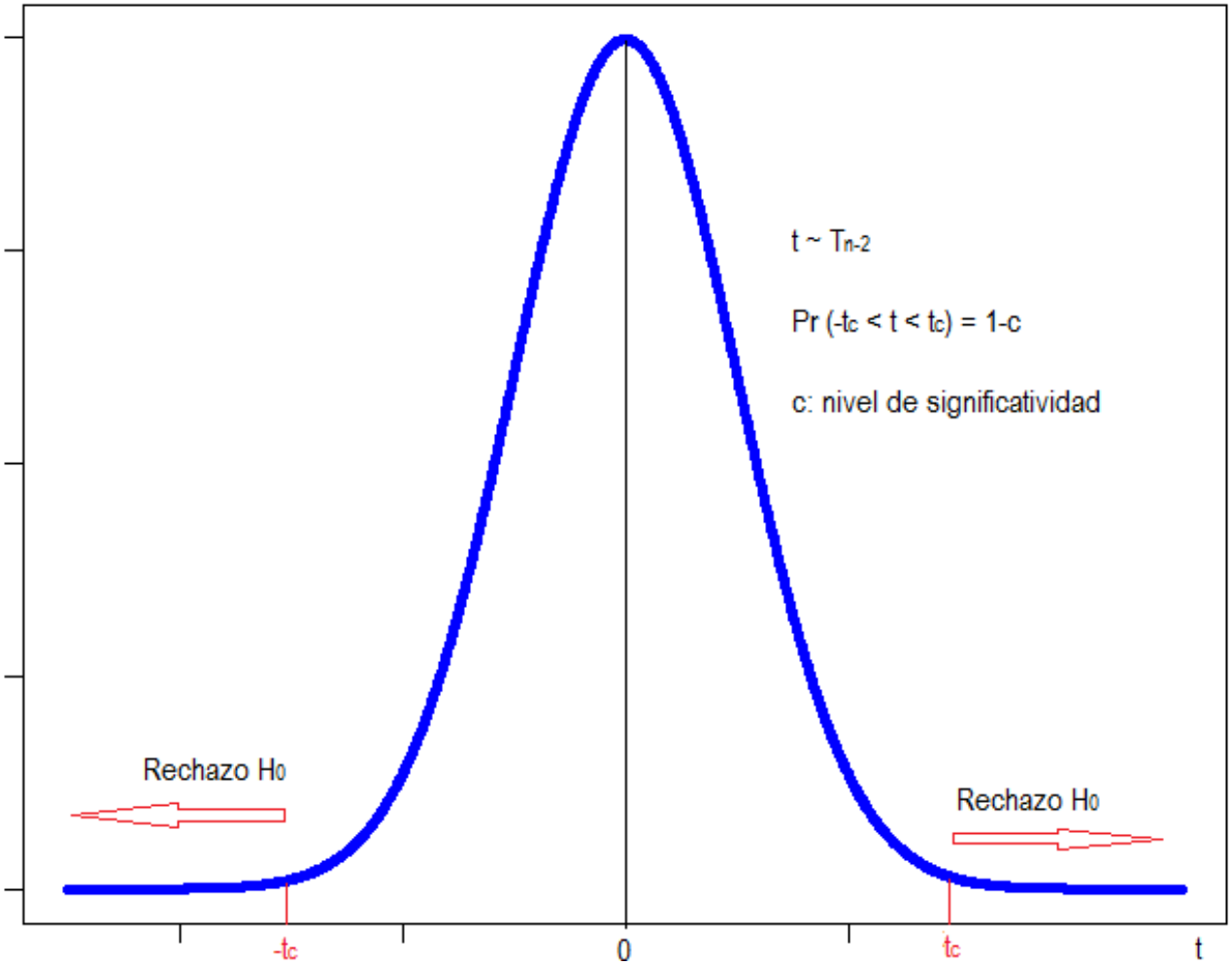
- Resultado: bajo todos los supuestos, si $H_0 : \beta = 0$ es cierta, entonces

$$t \equiv \frac{\hat{\beta}}{\sqrt{S^2 / \sum x_i^2}} \sim T_{n-2}$$

donde T_{n-2} es la distribición T de Student con $n-2$ grados de libertad

$$t \equiv \frac{\hat{\beta}}{\sqrt{S^2 / \sum x_i^2}} \sim T_{n-2}$$

- t es z pero reemplazando σ^2 por S^2
- Ahora todo puede ser computado con los datos disponibles


t ~ Tn-2
Pr (-tc < t < tc) = 1-c
c: nivel de significatividad
Rechazo H0
Rechazo H0
-tc
0
tc
t

- Hasta ahora vimos el caso particular: $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$
- Caso general: $H_0 : \beta = \beta_0$ vs. $H_A : \beta \neq \beta_0$
- Ejemplo: supongamos que en el modelo $Y_i = \alpha + \beta X_i + \mu_i$, donde Y es el consumo agregado y X es el ingreso disponible, queremos evaluar si la propensión marginal a consumir es igual a 1. Para eso postulamos $H_0 : \beta = 1$

- Entonces, bajo todos los supuestos y cuando es cierta $H_0 : \beta = \beta_0$, se cumple:

$$\hat{\beta} \sim N(\beta_0, \sigma^2 / \sum x_i^2)$$

- Que estandarizando resulta:

$$z \equiv \frac{\hat{\beta} - \beta_0}{\sqrt{\sigma^2 / \sum x_i^2}} \sim N(0, 1)$$

- Y si reemplazamos por el estimador de σ^2 tenemos:

$$t \equiv \frac{\hat{\beta} - \beta_0}{\sqrt{S^2 / \sum x_i^2}} \sim T_{n-2}$$

$$t \equiv \frac{\hat{\beta} - \beta_0}{\sqrt{S^2 / \sum x_i^2}} \sim T_{n-2}$$

- En la práctica, reemplazamos β_0 por cualquier valor que nos interese
- En general, el estadístico t es la versión estimada de H_0, dividida por la raíz cuadrada de la varianza estimada.

Bibliografía para esta clase

- Notas de Clase, Capítulo 1.
- Wooldridge, Capítulo 2.