



## Clase 3: intro OLS, CLT y Poder

---

Denise Laroze

6 de septiembre de 2018

CESS - Universidad de Santiago  
*denise.laroze@usach.cl*

# Resumen de contenidos

Ordinary Least Squares

Central Limit theorem

Poder estadístico

## Ordinary Least Squares

---

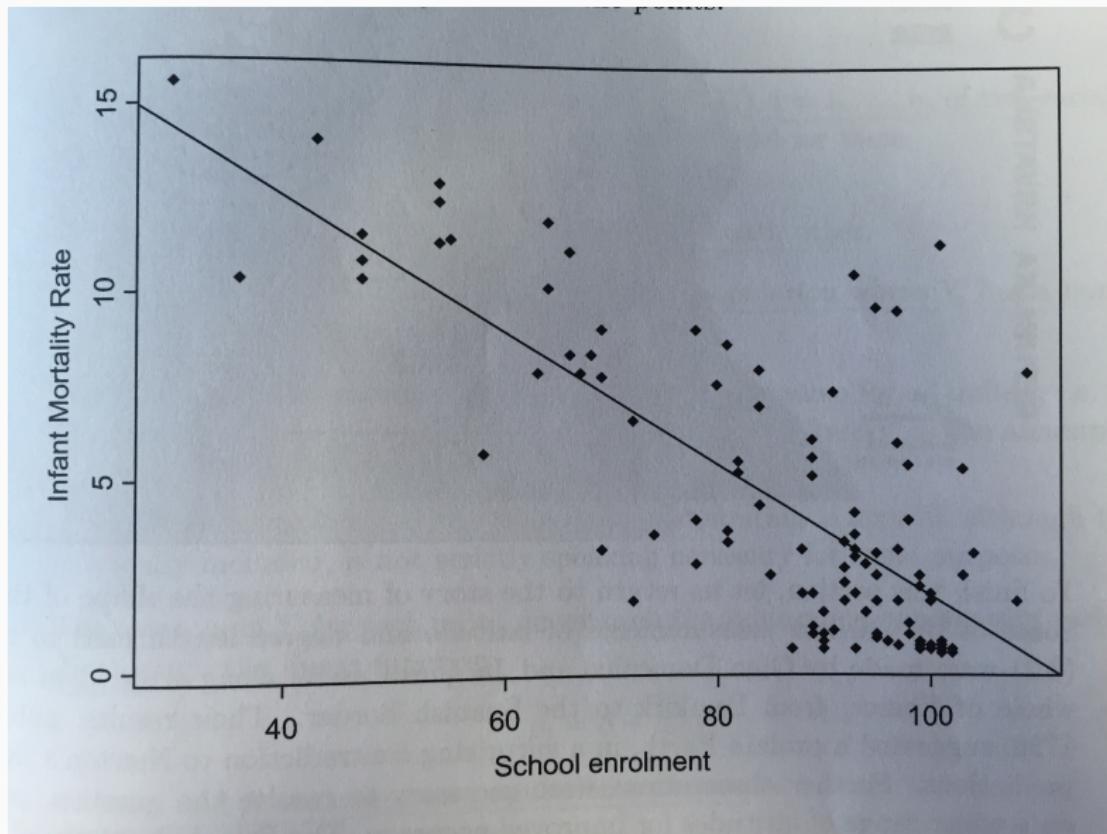
La regresiones lineales simples del tipo OLS (not so – Ordinary Least Squares) estiman una ecuación del tipo:

$$Y = \alpha + \beta X + \epsilon$$

Variable Dependiente = [parte sistemática] + [parte aleatoria]

Con el objetivo de establecer un modelos estadístico que nos permita interpretar lo que ocurre en ( y predecir) un fenómeno de interés.

## OLS - Dispersion



# Definición y presunciones de una regresión lineal simple I

1. Las observaciones  $Y_i$  son estadísticamente independientes unas de otras (iid).
2. Las observaciones  $Y_i$  son una muestra aleatoria de una población donde  $Y_i$  tiene una distribución normal con media  $\mu_i$  y varianza  $\sigma^2$ .
  - 2.1 Tomar en consideración que la varianza  $\sigma^2$  se asume igual para todas las unidades  $i$ . En otras palabras que no depende de  $X_i$ , presunción conocida como *homoskedasticidad*.
  - 2.2 La presunción de que la distribución de la población es normal, aunque regularmente incluida, no es estrictamente necesaria para algunos casos
3. La media de  $\mu_i$  de  $Y_i$  para cada unidad de  $i$  depende del valor de la variable independiente  $X_i$ , a través de una función lineal del tipo  $\mu_i = \alpha + \beta X_i$  donde los parámetros  $\alpha$  y  $\beta$  son desconocidos.

## Definición y presunciones de una regresión lineal simple II

En la estimación hay un elemento aleatorio no observable  $\epsilon_i$  que se asume tiene las siguientes propiedades

4. Todos los  $\epsilon_i$  son estadísticamente independientes unos de otros.
5. La media (valor esperado) de  $\epsilon_i$  es 0 para todos los  $i$ , sin depender de  $X_i$ . Denominado  $E(\epsilon_i) = 0$ .
6. La varianza de  $\epsilon_i$  es  $\sigma^2$  para todos los  $i$ , independiente de  $X_i$ . Denominado  $var(\epsilon_i) = \sigma^2$ .
7. Los  $\epsilon_i$  están distribuidos normalmente. Nuevamente, se suele incluir esta presunción, pero no es estrictamente necesario.

## Interpretación del modelo

Una regresión lineal simple tiene tres parámetros  $\alpha$ ,  $\beta$  y  $\sigma^2$  (este último también denominado  $\epsilon$ ). Los parámetros  $\alpha$  y  $\beta$  son conocidos como los **coeficientes de la regresión** y se interpretan de la siguiente forma:

1.  $\alpha$  es el valor esperado de  $Y$  cuando  $X$  es igual a 0. Se le conoce como el **intercepto** o **constante** de una regresión. Como  $X = 0$  no es un valor interesante en si mismo (normalmente),  $\alpha$  no se suele interpretar.
2.  $\beta$  en cambio, sí es muy interesante y corresponde al cambio en el valor esperado de  $Y$  cuando  $X$  aumenta en 1 unidad. A  $\beta$  se lo conoce como la **pendiente** o el **coeficiente de  $X$**  y es típicamente el único parámetro de interés en una regresión porque describe la asociación entre  $X$  e  $Y$ .
3. El signo de  $\beta$  (+ o -) denota la dirección de la asociación.

## (cont.) interpretación

### Ejemplo

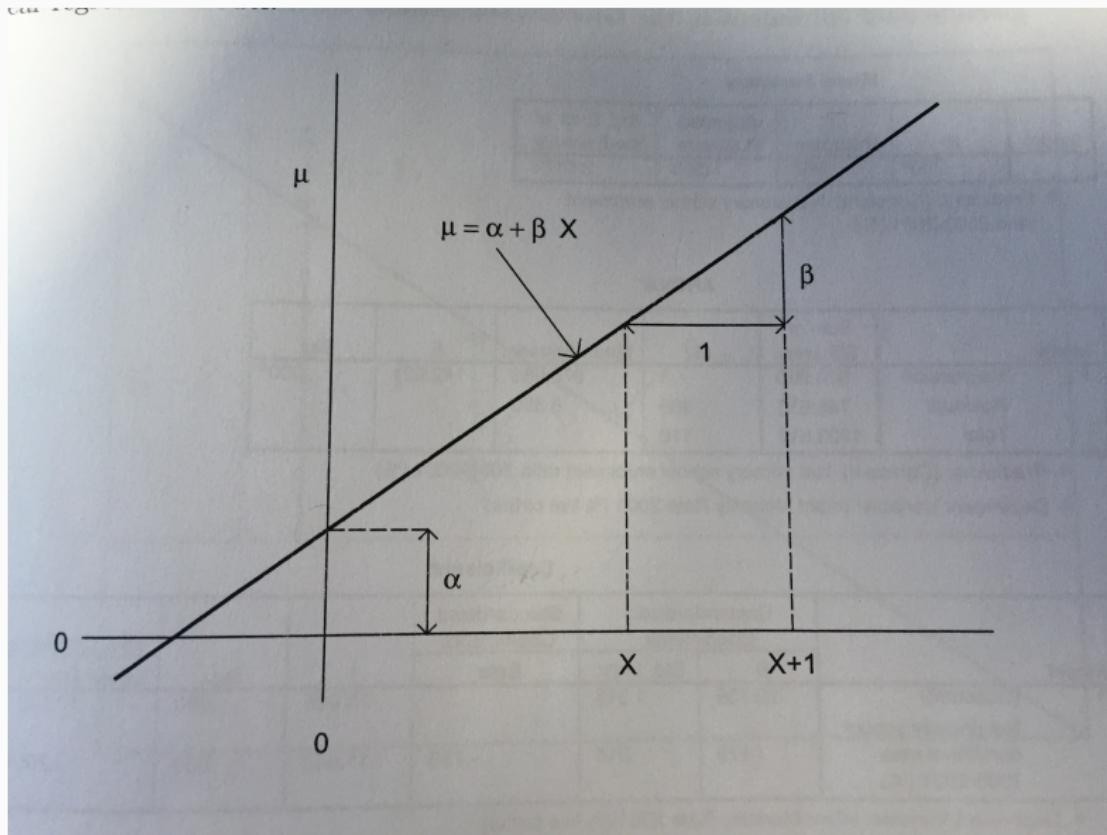
$$\text{con } X+1 : \mu = \alpha + \beta * (X + 1) = \alpha + \beta X + \beta$$

$$\text{con } X : \mu = \alpha + \beta * (X) = \alpha + \beta X$$

---

Diferencia:  $\beta$

## OLS - visualización parte sistemática



"All models are wrong because the world, especially the social world, is an exceedingly complex place, full of local detail. As social scientists we do not want to reproduce that local detail in our models. What we are trying to do is capture the essentials and leave out the inessentials. A model that was one hundred percent correct would be of no value because it would be as complex as reality itself and if we could understand reality in all its complexity we would have no need for models!" (Kuha, 2011:26)

"Essentially, all models are wrong, but some are useful" (Box, 1987: 424)

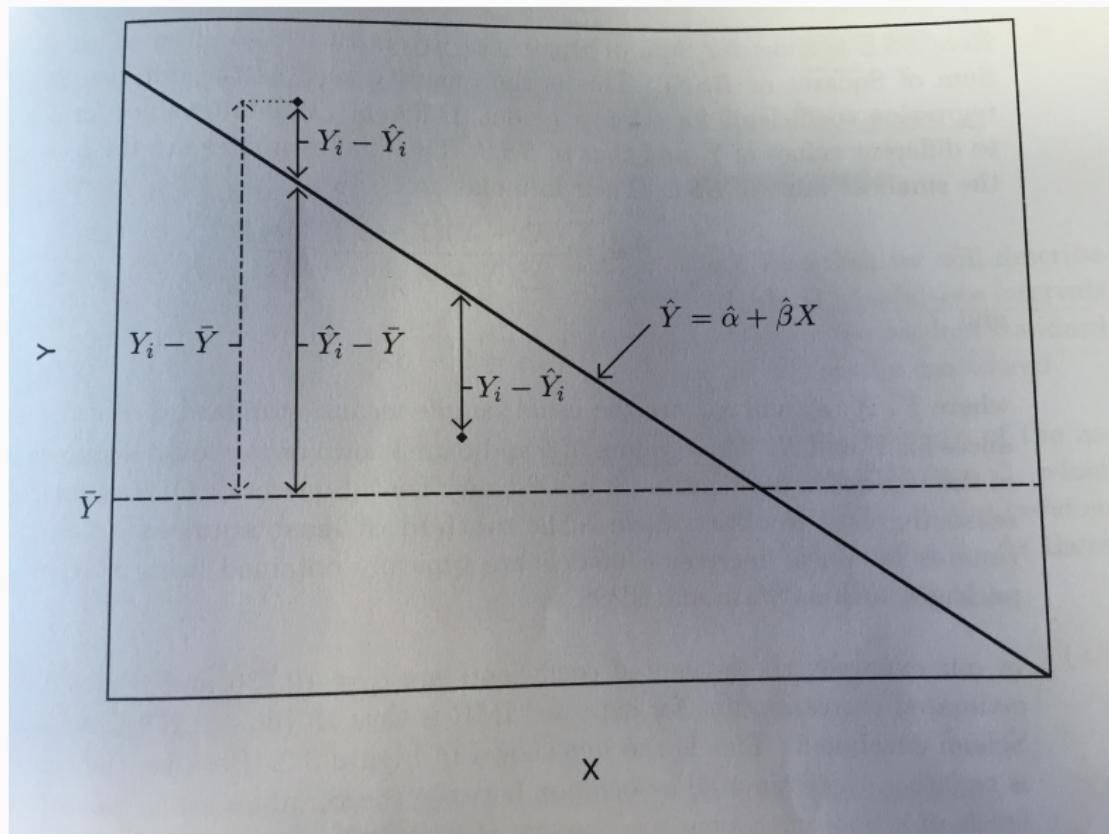
## La varianza condicional $\sigma^2$

El tercer parámetro en una regresión es  $\sigma^2$  que corresponde a la varianza de la distribución de  $Y$  dado  $X$ . También se le conoce como la **varianza de error** o la **varianza residual** y su raíz cuadrada  $\sigma$  se le conoce como el error condicional o **desviación estándar residual**.

Por ejemplo, si miramos la figura de la lámina 7 (inscripción escolar y mortalidad infantil) y tomamos un valor de  $X (=85)$  y vemos todas los valores de  $Y$ , lo que observamos en ese punto con los valores de  $Y$  condicionales a  $X = 85$  ( $Y|X = 85$ ).  $X$  está fija, pero no todos los valores de  $Y$  son iguales. La media de ese  $Y$  es  $\mu = \alpha + \beta * (X)$  y  $\sigma^2$  es la distribución de  $Y$  en ese punto.

De esta forma  $\sigma^2$  y su correspondiente  $\sigma$  indican qué tan concentrados están los valores de  $Y$  alrededor de esa media.

# OLS - Error



## Fitted values

Una vez realizada la regresión y estimados los parámetros  $\alpha$  y  $\beta$  se pueden calcular los valores estimados (*fitted values*).

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X$$

Estos *fitted values* se comparan con los valores reales de  $Y_i$ .

Su diferencia  $Y_1 - \hat{Y}_i$  es conocida como los residuos (de la muestra). La estimación de OLS minimiza esos residuos, entregando la línea que mejor se aproxima a los datos.

## Fitted values

Una vez realizada la regresión y estimados los parámetros  $\alpha$  y  $\beta$  se pueden calcular los valores estimados (*fitted values*).

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X$$

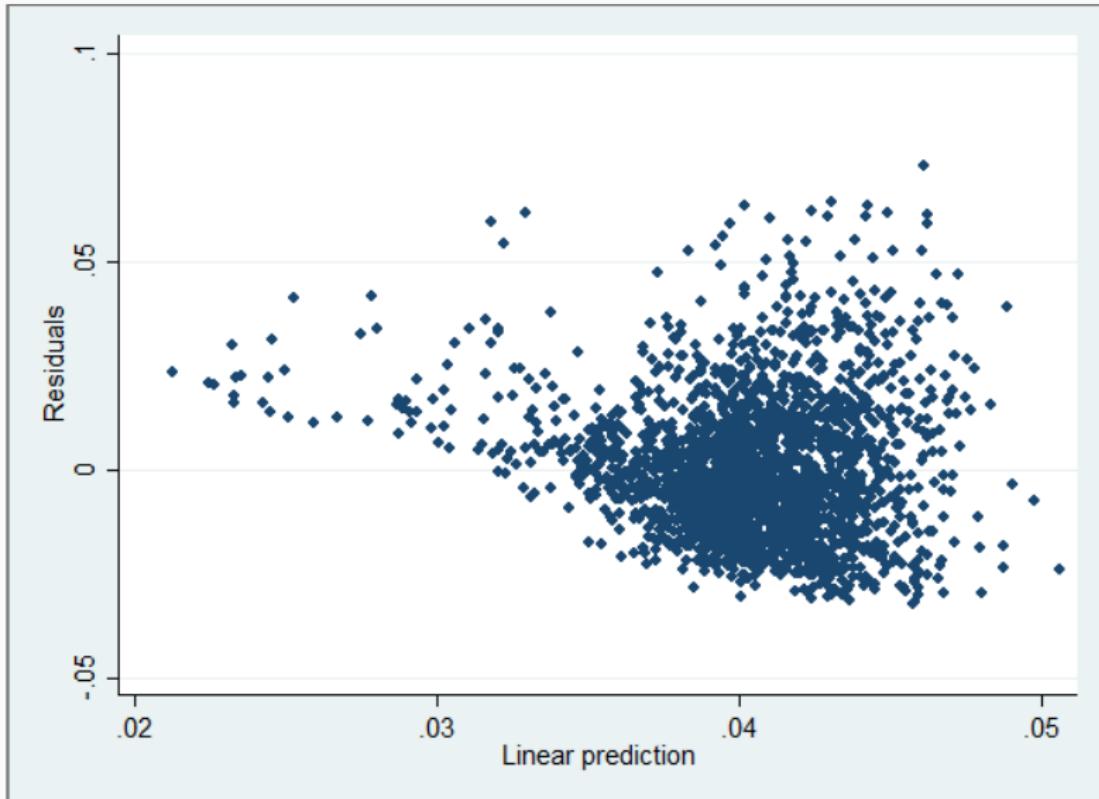
Estos *fitted values*  $\hat{Y}_i$  se comparan con los valores reales de  $Y_i$ .

Su diferencia  $Y_1 - \hat{Y}_i$  son los residuos (de la muestra). La estimación de OLS minimiza la suma de esos residuos al cuadrado (sum of squared residuals - SSR) , entregando la línea que mejor se aproxima a los datos.

Está demostrado que la formula para calcular  $\beta$  que reduce la SSR es:

$$\hat{\beta} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{S_{xy}}{S_x^2} = (X'X)^{-1}X'Y$$

# Heteroskedasticidad



## Central Limit theorem

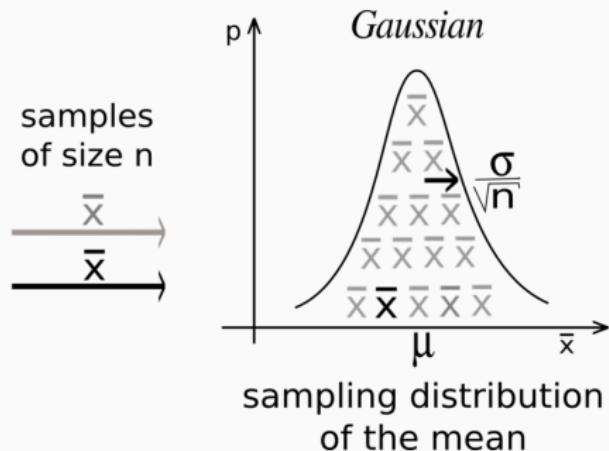
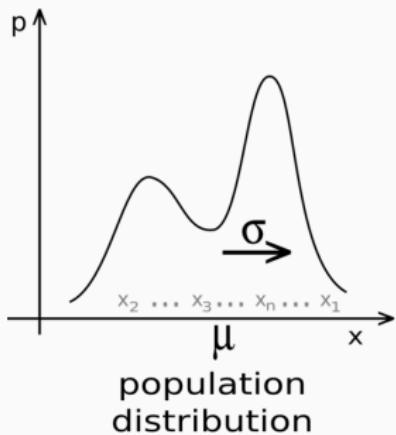
---

## Central Limit theorem

En teoría de la probabilidad el central limit theorem (CLT) establece que, cuando existen variables iid (independent and identically distributed), sus medias (o sumas normalizadas) tienden hacia una distribución normal (en el infinito), incluso si la distribución de la que provienen no es normal.

Este teorema es clave para las estadísticas porque implica que los métodos que aplican para distribuciones normales se pueden aplicar para otras distribuciones.

Supongamos, por ejemplo que se obtiene una muestra con un gran número de observaciones, cada una de las cuales es generada en una forma aleatoria que no depende de los valores de otras observaciones. Supongamos también que se calcula la media aritmética de esa muestra. Si se repite el procedimiento múltiples (infinitas) veces, el CLT indica que los valores estimados de esas medias se distribuirán normalmente, y que la media de las medias corresponderá a la media de la población.



Ahora probemos por simulación

Un muy buen video explicativo se encuentra en

[https://www.khanacademy.org/math/ap-statistics/  
sampling-distribution-ap/sampling-distribution-mean/v/  
central-limit-theorem](https://www.khanacademy.org/math/ap-statistics/sampling-distribution-ap/sampling-distribution-mean/v/central-limit-theorem)

## Poder estadístico

---

# Tamaño de la muestra - Poder estadístico

Poder es fundamental para evitar errores del Tipo II

Type I Error



Type II Error



# Poder estadístico

Error Tipo I : Significancia estadística del tipo  $p < 0,001$

Error tipo II: Beta es la probabilidad de que ocurra un error tipo II ¿Con qué frecuencia podemos rechazar la hipótesis nula con éxito?

$$\beta = \Phi\left(\frac{|\mu_t - \mu_c| \sqrt{N}}{2\sigma} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right),$$

1-Beta es el poder estadístico: Probabilidad de que la prueba rechace  $H_0$

En la medida que **N** aumenta, el poder estadístico de la muestra aumenta. Pero también depende de la magnitud del **impacto** y la **varianza**.

# Poder estadístico

