

# Sesión 6: Modelos estadísticos Lineales

Modelos de regresión Simples y múltiples

Cristian E García

[cegarcia@uao.edu.co](mailto:cegarcia@uao.edu.co)

Maestría en Inteligencia Artificial y Ciencias de Datos  
Facultad de Ingeniería  
2025



# Outline

## 1. Modelos Estadísticos

1.1. Introducción

1.2. Definición

1.3. Modelos Simples Y múltiples

## 2. Large Sample en Regresión

## 3. Ejemplos

# Introducción

Los métodos estadísticos para múltiples variables suelen analizar cómo el resultado de una **variable de respuesta** está asociado o puede ser predicho por los valores de las **variables explicativas**. Por ejemplo, un estudio podría analizar cómo la cantidad anual donada a la caridad está asociada con variables explicativas como el ingreso anual de una persona, el número de años de educación, la religiosidad, la edad y el género.

Para la inferencia estadística, los métodos asumen una distribución de probabilidad para la variable de respuesta en cada combinación de valores de las variables explicativas. No se requiere hacer suposiciones sobre las distribuciones de las variables explicativas.

# Introducción

Los métodos estadísticos para múltiples variables suelen analizar cómo el resultado de una **variable de respuesta** está asociado o puede ser predicho por los valores de las **variables explicativas**.

Por ejemplo, un estudio podría analizar cómo la cantidad anual donada a la caridad está relacionada con variables explicativas como el ingreso anual de una persona, el número de años de educación, la religiosidad, la edad y el género. Para la **inferencia estadística**, estos métodos asumen una distribución de probabilidad para la variable de respuesta en cada combinación de valores de las variables explicativas, sin necesidad de asumir nada sobre la distribución de las variables explicativas.

# Introducción

Los métodos estadísticos para múltiples variables suelen analizar cómo el resultado de una **variable de respuesta** está asociado o puede ser predicho por los valores de las **variables explicativas**.

Por ejemplo, un estudio podría analizar cómo la cantidad anual donada a la caridad está relacionada con variables explicativas como el ingreso anual de una persona, el número de años de educación, la religiosidad, la edad y el género. Para la **inferencia estadística**, estos métodos asumen una distribución de probabilidad para la variable de respuesta en cada combinación de valores de las variables explicativas, sin necesidad de asumir nada sobre la distribución de las variables explicativas.

# ¿Qué es un Modelo?

Para comprender que es un modelo, primero debemos pensar en que tipos de modelos existen.

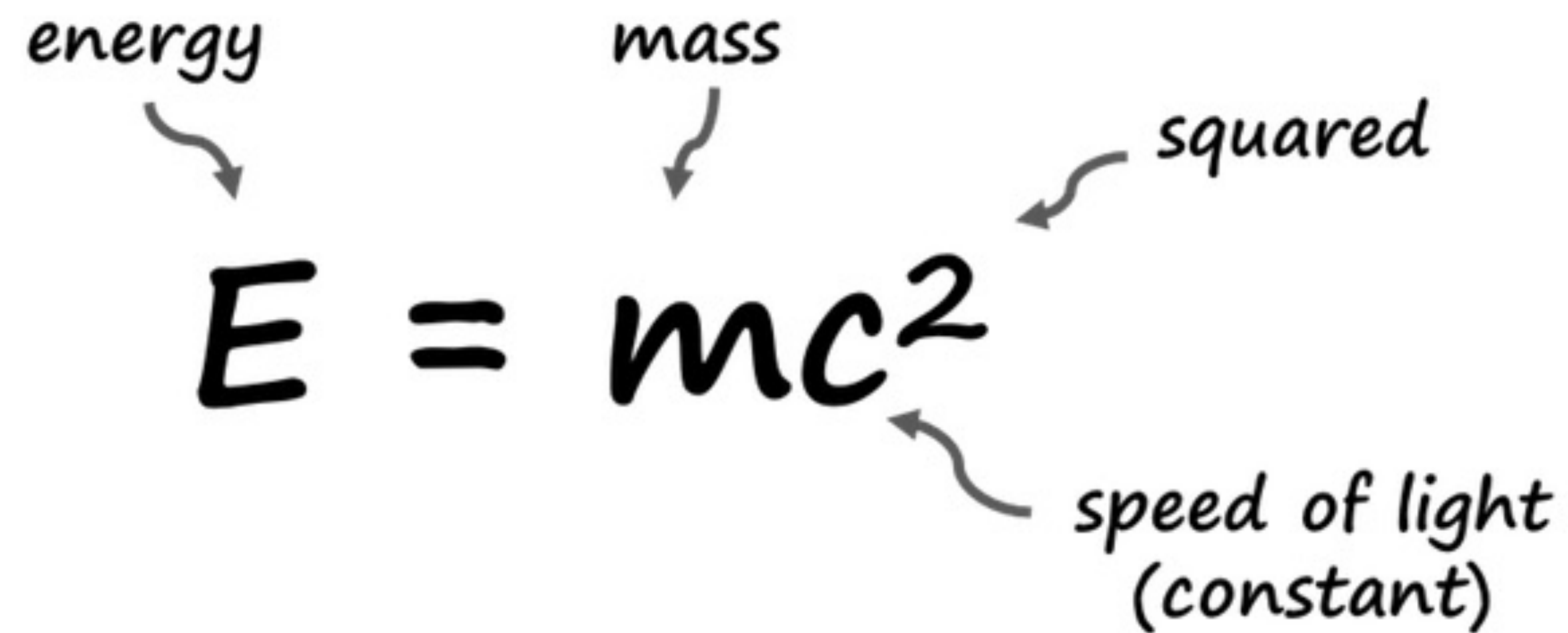
- ✱ Modelo Matemático
- ✱ Modelo Probabilístico
- ✱ Modelo Estadístico

# Modelo Matemático

Podemos describir fenómenos de la naturaleza con exactitud y sin ningún error.

# Modelo Matemático

Podemos describir fenómenos de la naturaleza con exactitud y sin ningún error.



The diagram shows the equation  $E = mc^2$  in a large, bold, black font. Four handwritten labels with arrows point to the components of the equation: 'energy' points to 'E', 'mass' points to 'm', 'squared' points to the '2' in 'c^2', and 'speed of light (constant)' points to 'c'.

$$E = mc^2$$

energy

mass

squared

speed of light  
(constant)



# Modelo Probabilístico

Los modelos probabilísticos definen fenómenos que son gobernados por el azar, sin embargo no conducen a ningún error.



# Proceso de Inferir

Antes de que veamos los **modelos estadísticos** es necesario que entendamos qué es el proceso de inferir.

## Definición de Libro

La palabra inferir deriva del latín inferre que significa llevar dentro, por lo que podríamos entenderlo como sacar una conclusión desde información que ya existe y está disponible, todo dentro del contexto de inferir entendido como un proceso de pensamiento.

# Modelo Estadístico

## Idea 1

Un modelo estadístico es una ecuación matemática que reproduce los fenómenos que observamos de la forma más exacta posible. El modelo es diferente cada vez que se modifica la información.

## Idea 2

Un modelo estadístico se encarga de dar explicación a fenómenos los cuales se pueden representar por medio de una relación matemática las cuales tienen dos componentes asociadas, una de azar y una de error.

# Modelo Estadístico

Un modelo estadístico por lo general se relaciona por medio de una variable de interés y una componente de error.

$$Y = f(x) + \epsilon$$

donde  $Y$  es la característica de interés,  $f(x)$  es la función que relaciona la característica de interés con las mediciones y  $\epsilon$  relaciona la parte del azar y el error o la incertidumbre que generan las mediciones.

# Modelo Estadístico

Algunas situaciones reales:

- \* Relación de la altura con la edad en niños
- \* Relación entre la Zona y el precio de apartamentos o casa
- \* Relación entre el cilindraje de carros y el consumo de gasolina
- \* .....



# Modelo Estadístico

¿Qué forma le daremos a  $f(x)$ ?

Pensemos en el modelo más simple,

$$Y = aX + b$$

# Modelo Estadístico

En este sentido es evidente notar que bajo la estructura del modelo estadístico.

$$Y = f(x) + \epsilon$$

$$Y = 2 + 3X$$

¿Y qué pasa con  $\epsilon$ ?

# Modelo Estadístico

En este sentido es evidente notar que bajo la estructura del modelo estadístico.

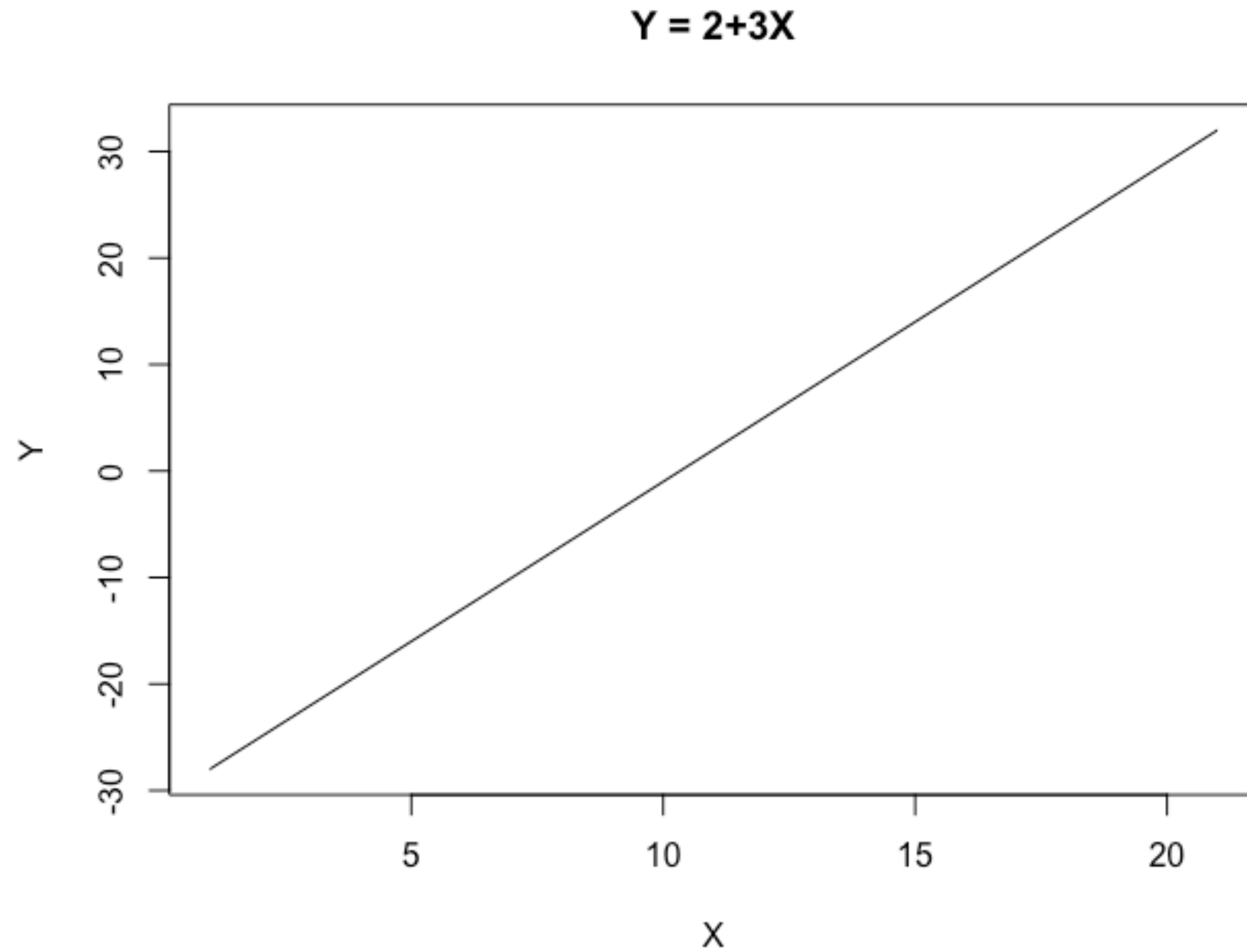
$$Y = f(x) + \epsilon$$

$$Y = 2 + 3X$$

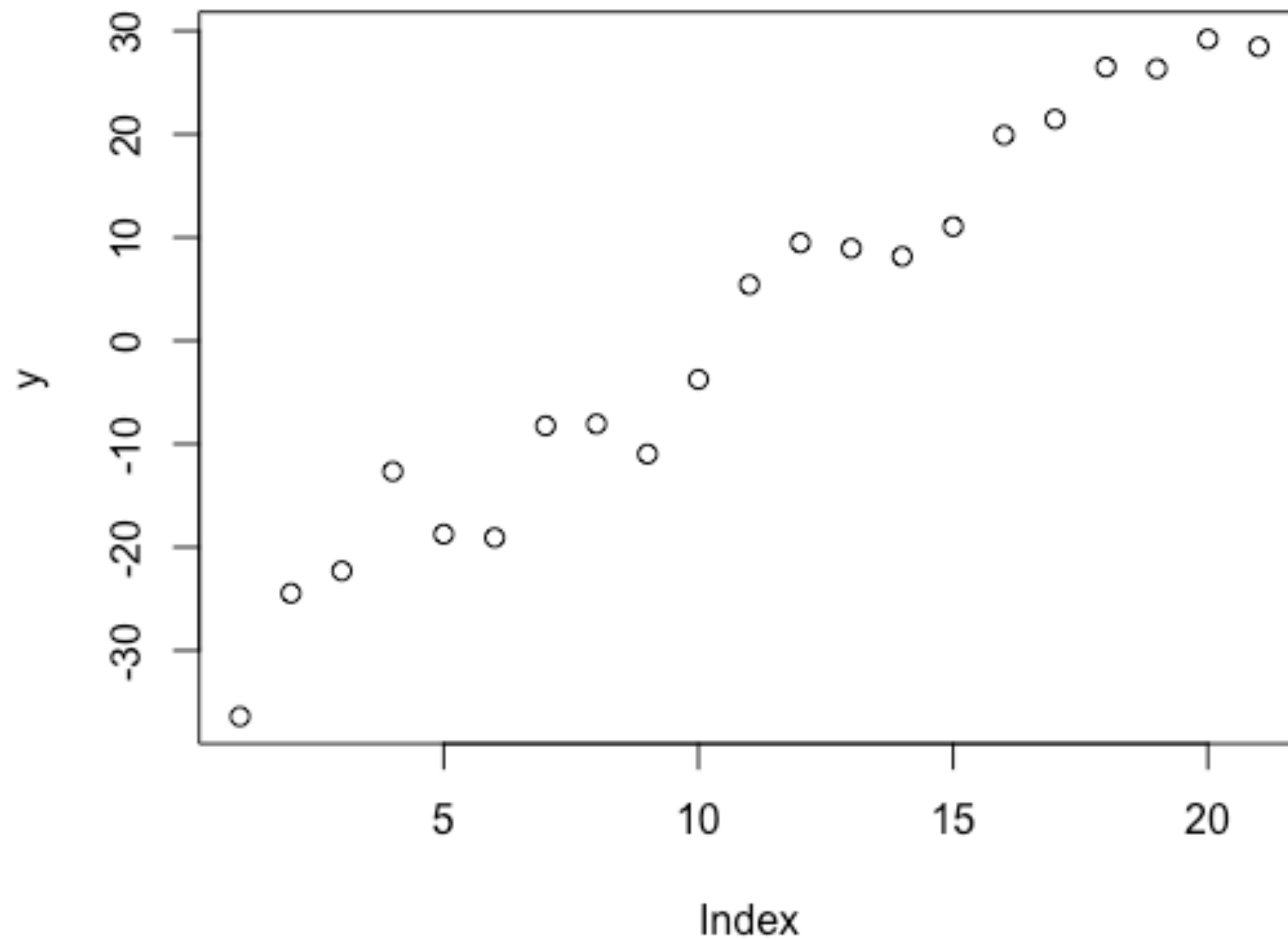
¿Y qué pasa con  $\epsilon$ ? **Ahora lo veremos.**



# Modelo Estadístico



# Modelo Estadístico



# Modelos de Regresión

# Modelos de Regresión Simple

Recordemos que tenemos un modelo de la forma

$$Y = f(x) + \epsilon$$

Ahora consideremos el que una forma razonable para  $f(x) = \beta_0 + \beta_1 x$ , de esta forma el modelo planteado es,

$$Y = \beta_0 + \beta_1 x + \epsilon$$

# Modelos de Regresión Simple

El objetivo es obtener estimaciones razonables de  $Y$  para distintos valores de  $X$  a partir de una muestra de  $n$  pares de valores  $(x_1, y_1), \dots, (x_n, y_n)$ .

Nos interesa obtener estimaciones para describir la relación lineal existente entre  $X$  e  $Y$

En la realidad nunca conocemos los verdaderos valores de  $\beta_0$  y  $\beta_1$  es por eso que debemos estimarlos.

# Modelos de Regresión Simple

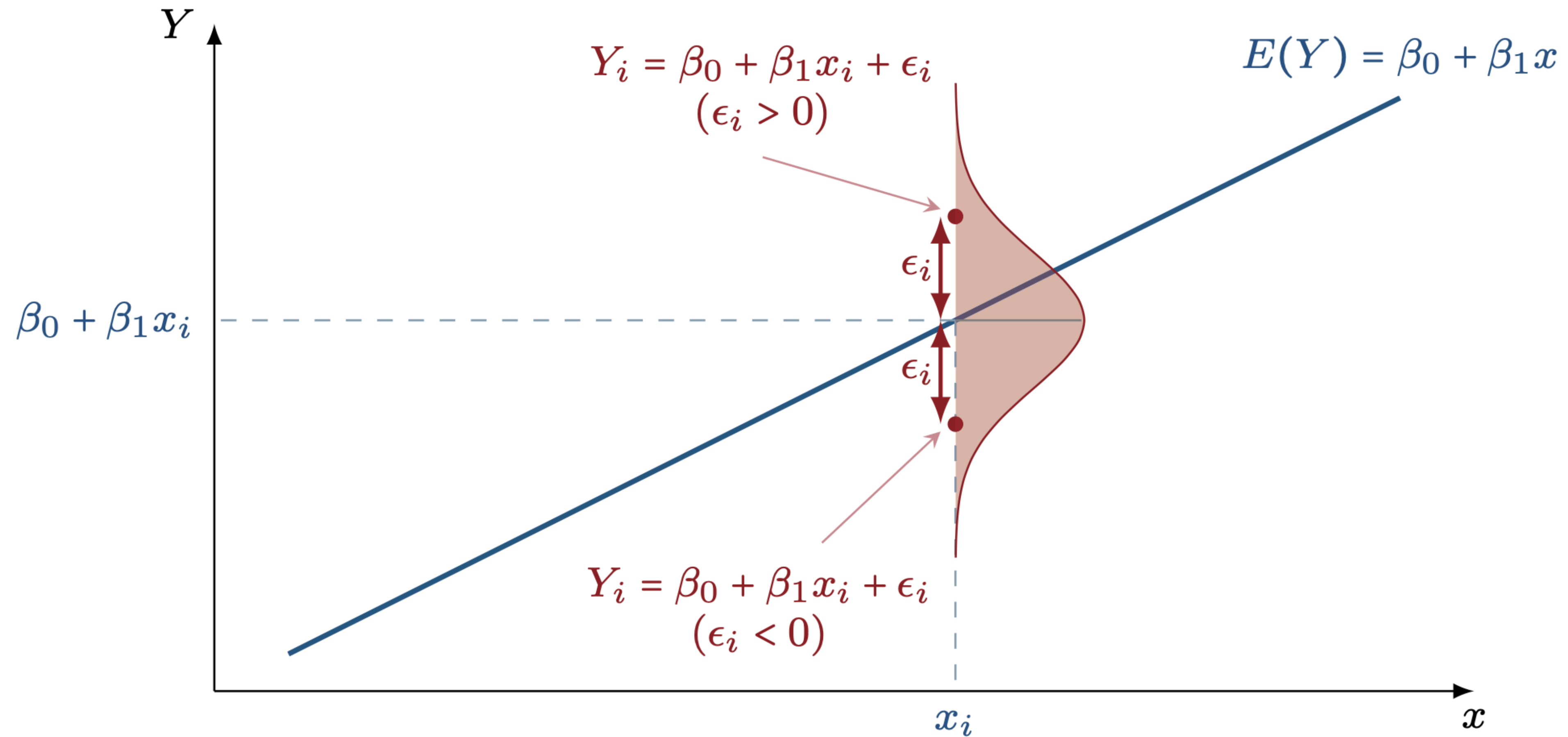
El objetivo general será obtener valores estimados de  $\hat{\beta}_0$  y  $\hat{\beta}_1$ . **(“aquí haremos un acto de fé por ahora”).**  
Las estimaciones ser plantearán de la siguiente manera,

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

y

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \frac{\left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}$$

# Modelos de Regresión Simple



# Modelo Lineal

El modelo lineal está dado por

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

La distribución de  $\mathbf{Y}$  pertenece a la clase

$$\mathcal{P} = \left\{ P_{\theta} : \mathbb{E}_{\theta} \mathbf{Y} = X\boldsymbol{\beta}, \quad \text{Cov}_{\theta} \mathbf{Y} = C, \quad \theta = (\boldsymbol{\beta}, C, \kappa) \in \Theta \subseteq \mathbb{R}^p \times \mathcal{M}^{\geq} \times \mathcal{K} \right\}.$$



# Modelo Lineal

Ahora es valido pensar, que pasará si hay más de una variable de interés. En ese caso pensemos en un modelo de la siguiente forma.

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

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{p1} \\ 1 & x_{12} & \cdots & x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Está es una estructura más adecuada a la realidad, en la practica siempre se cuentan con más variables.

## Mínimos Cuadrados

Para el modelo lineal  $E(Y_i) = \beta_0 + \beta_1 x_i$ , con una muestra de  $n$  observaciones, el **método de mínimos cuadrados** determina los valores de  $\hat{\beta}_0$  y  $\hat{\beta}_1$  que minimizan

$$\sum_{i=1}^n (y_i - \hat{\mu}_i)^2 = \sum_{i=1}^n \left[ y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2 = \sum_{i=1}^n e_i^2,$$

la suma de los residuos al cuadrado.

# Estimación

Como función de los parámetros del modelo  $(\beta_0, \beta_1)$ , la expresión

$$S(\beta_0, \beta_1) = \sum_i (y_i - \mu_i)^2 = \sum_i [y_i - (\beta_0 + \beta_1 x_i)]^2$$

es cuadrática en  $\beta_0$  y  $\beta_1$ . Podemos minimizarla igualando

$$\frac{\partial S}{\partial \beta_0} = - \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)] = 0,$$

y

$$\frac{\partial S}{\partial \beta_1} = - \sum_{i=1}^n x_i [y_i - (\beta_0 + \beta_1 x_i)] = 0.$$

# Estimación

Reescribimos las ecuaciones como

$$\sum_{i=1}^n y_i = n\beta_0 + \beta_1 \sum_{i=1}^n x_i, \quad \sum_{i=1}^n x_i y_i = \beta_0 \left( \sum_{i=1}^n x_i \right) + \beta_1 \sum_{i=1}^n x_i^2.$$

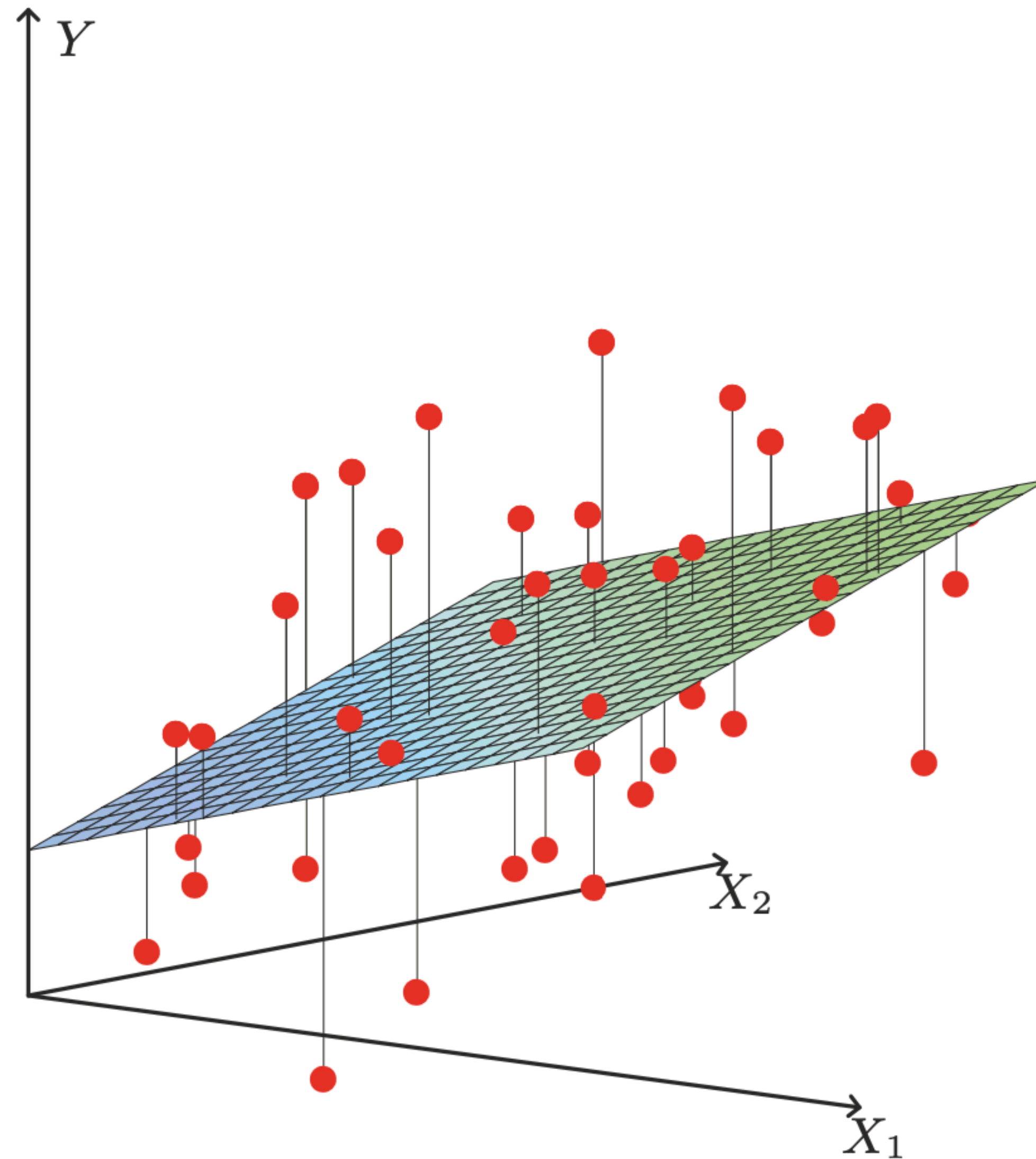
La solución simultánea de estas dos ecuaciones da las estimaciones por mínimos cuadrados, que son únicas. Estas son

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

donde

$$s_x = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n - 1}}, \quad s_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n - 1}.$$

# Estimación



# Estimación Múltiples Parámetros

Consideremos la matriz  $\mathbf{X}$  de tamaño  $N \times (p + 1)$ , donde cada fila representa un vector de entrada (con un 1 en la primera posición), y sea  $\mathbf{y}$  el vector de tamaño  $N$  que contiene las salidas del conjunto de entrenamiento. Entonces, podemos escribir la suma de los cuadrados de los residuos como

$$\text{RSS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Esta es una función cuadrática en los  $p + 1$  parámetros. Al diferenciar con respecto a  $\boldsymbol{\beta}$ , obtenemos

$$\begin{aligned}\frac{\partial \text{RSS}}{\partial \boldsymbol{\beta}} &= -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \\ \frac{\partial^2 \text{RSS}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= 2\mathbf{X}^T\mathbf{X}.\end{aligned}$$

# Estimación Múltiples Parámetros

Si asumimos que  $\mathbf{X}$  tiene rango completo en sus columnas y, por lo tanto,  $\mathbf{X}^T \mathbf{X}$  es definida positiva, igualamos la primera derivada a cero

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

para obtener la solución única

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} .$$

# Estimación Múltiples Parámetros

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ip} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ip} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ip} & \sum_{i=1}^n x_{ip}x_{i1} & \sum_{i=1}^n x_{ip}x_{i2} & \cdots & \sum_{i=1}^n x_{ip}^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \vdots \\ \sum_{i=1}^n x_{ip}y_i \end{pmatrix}$$



# Modelos Lineales

**Teorema** Supongamos el modelo estadístico con  $\text{rank}(\mathbf{X}) = p$ . Entonces, las ecuaciones normales tienen una solución única  $\hat{\boldsymbol{\beta}}$ . Además, el estimador

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

tiene las siguientes propiedades:

- ✻  $\hat{\boldsymbol{\beta}}$  es lineal en  $\mathbf{Y}$ .
- ✻  $\hat{\boldsymbol{\beta}}$  es insesgado, es decir,  $\mathbb{E}_\theta[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$  para todo  $\theta$ .
- ✻ La covarianza está dada por

$$\text{Cov}_\theta \hat{\boldsymbol{\beta}} = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}.$$

# Supuestos Modelo Lineal

Es importante observar que pasa con los errores del modelo, de los cuales se tiene una estimación, que se obtiene de hacer  $\hat{e} = y - \hat{y}$ .

- ✱ Para comprobar si un modelo es un buen modelo recordemos que debe seguir el comportamiento de una distribución normal es decir  $N(0, \sigma^2)$ .
- ✱ El modelo debe tener varianza constante.
- ✱ Los residuos son incorrelacionados.

# Supuestos Modelo Lineal

**Normalidad:** Para el supuesto de normalidad se utilizará la prueba de Shapiro Wilk.

$H_0$ : los datos cumplen con el supuesto de normalidad

$H_1$ : los datos no cumplen con el supuesto de normalidad

# Supuestos Modelo Lineal

**Normalidad:** Para el supuesto de normalidad se utilizará la prueba de Shapiro Wilk.

$H_0$ : los datos cumplen con el supuesto de normalidad

$H_1$ : los datos no cumplen con el supuesto de normalidad

**Homogeneidad de varianza:** Para la homogeneidad de varianzas se hará uso de la prueba de Berusch Pagan

$H_0$ :: Las varianzas son homogéneas

$H_1$ : Las varianzas

# Supuestos Modelo Lineal

**Normalidad:** Para el supuesto de normalidad se utilizará la prueba de Shapiro Wilk.

$H_0$ : los datos cumplen con el supuesto de normalidad

$H_1$ : los datos no cumplen con el supuesto de normalidad

**Homogeneidad de varianza:** Para la homogeneidad de varianzas se hará uso de la prueba de Berusch Pagan

$H_0$ :: Las varianzas son homogéneas

$H_1$ : Las varianzas

**No correlación de los errores:** Para la prueba de la de los errores se utiliza la prueba de Durbin-Watson.

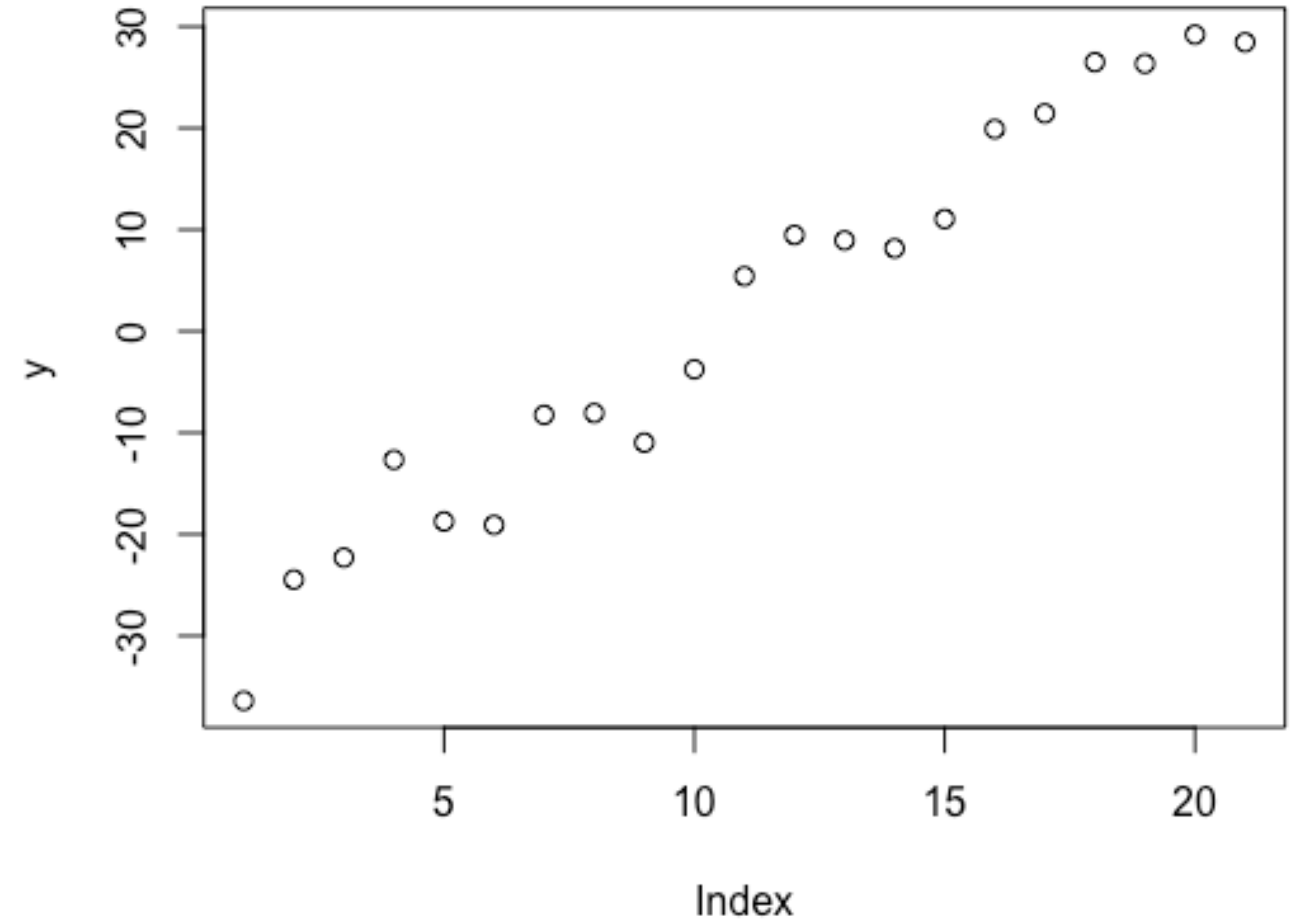
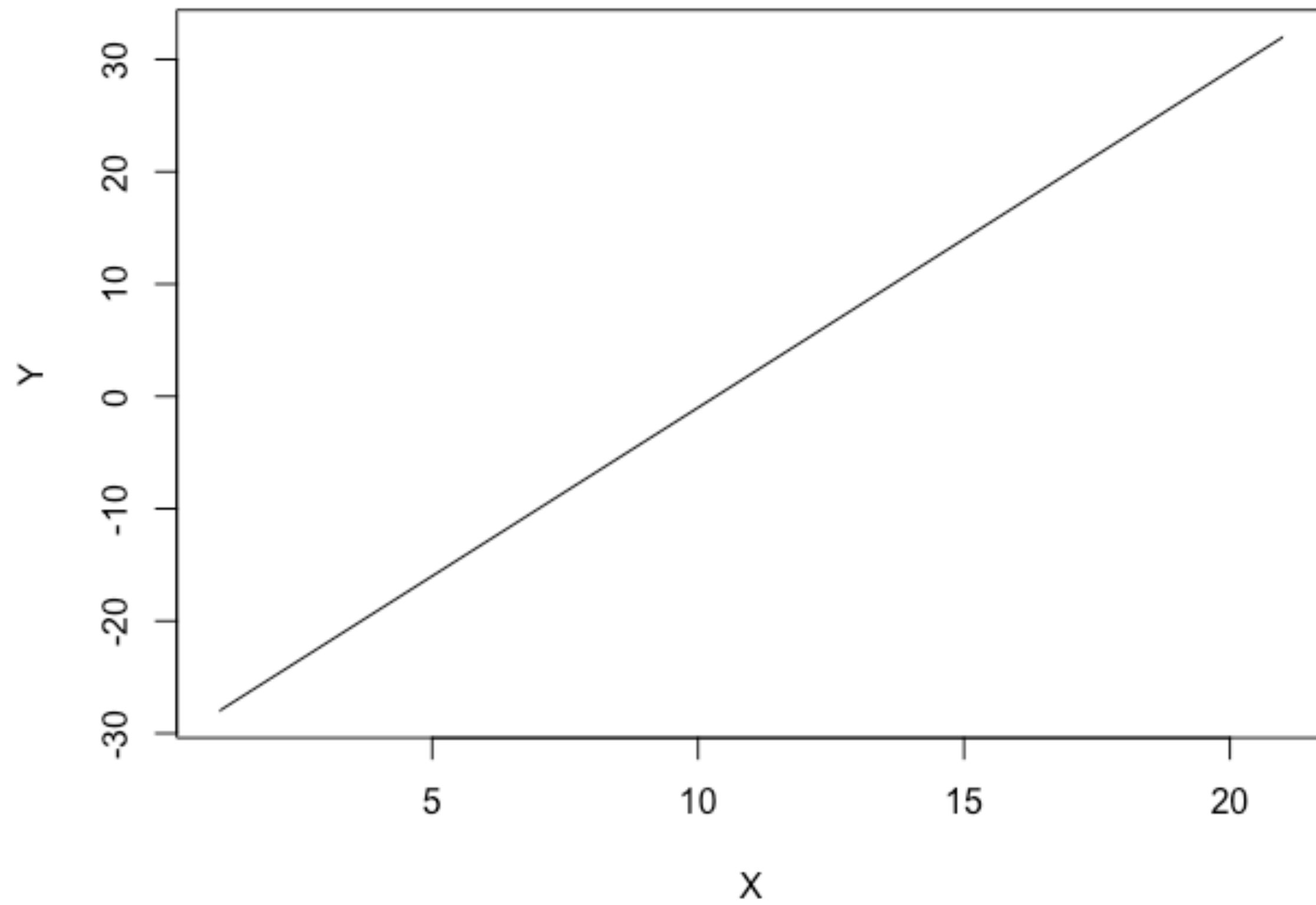
$H_0$ : No hay correlación

$H_1$ : Hay correlación

**¡Veámoslo en R!**  **Studio<sup>®</sup>**

# Ejemplo 1

$$Y = 2 + 3X$$



# Ejemplo 2

Pensemos en una situación “real”. El ejercicio de analizar el comportamiento de en las variables que influyen en el rendimiento de la gasolina en un conjunto de carros.

```
#####  
#   Variables   #  
#####  
  
# mpg:  Miles/(US) gallon  
# cyl:  Number of cylinders  
  
model2 <-lm ( mpg~ cyl , data = mtcars )  
summary ( model2 )
```





¿Preguntas?