TEMA 10. Regresión y correlación

Alicia Nieto Reyes

BIOESTADÍSTICA

Regresión y Correlación

Estudiaremos la relación entre 2 variables cuantitativas:
I.e.

- tenemos una muestra o población
- medimos dos variables cuantitativas a los individuos de la muestra o población

Ejemplo1: **Regresión**

- muestra de 200 personas que viven en Santander
- tomamos nota de la edad y medimos la tensión de cada una de las 200 personas

La tensión depende de la edad (aparte de depender de otros factores), queremos estudiar en que relación

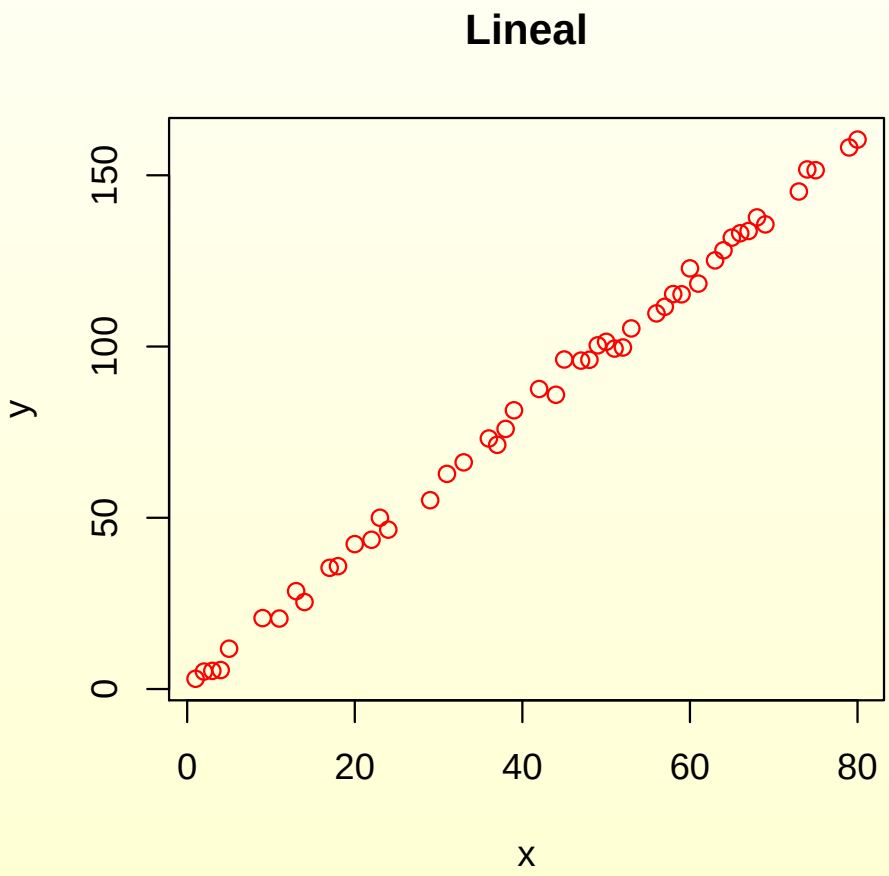
Ejemplo2: **Correlación**

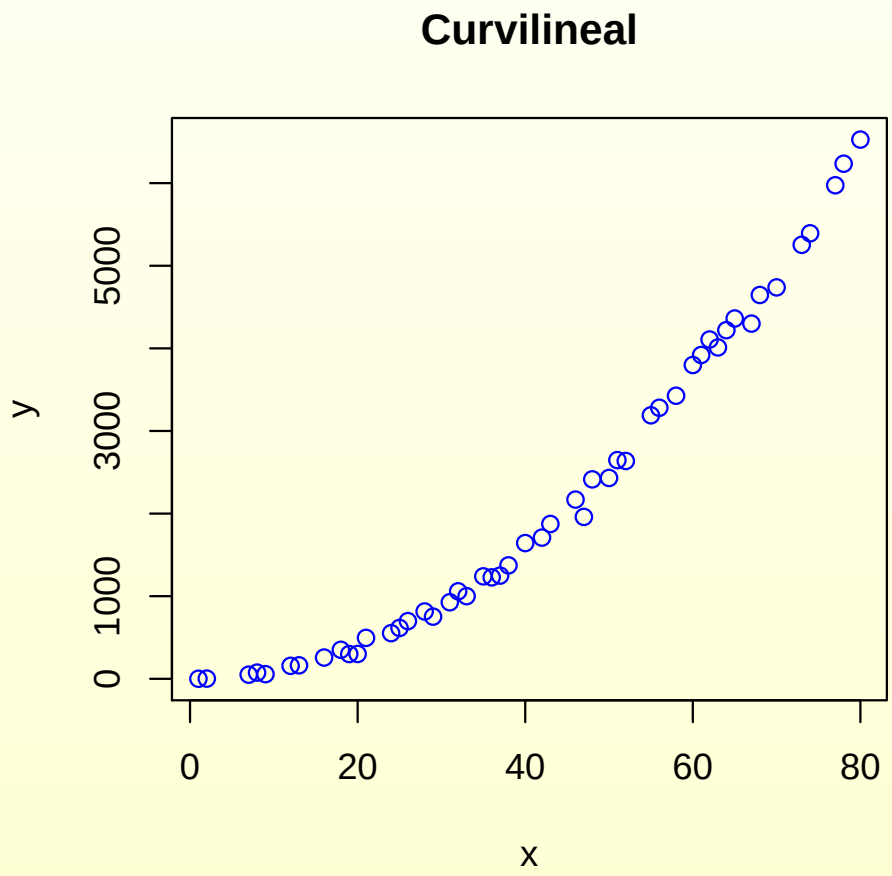
- muestra de 200 niños de 17 años
- medimos la longitud de brazos y la altura de cada uno de los 200 niños

Queremos estudiar la relación entre las variables longitud de brazos y altura (aquí no es una la causa o efecto de la otra)

Regresión

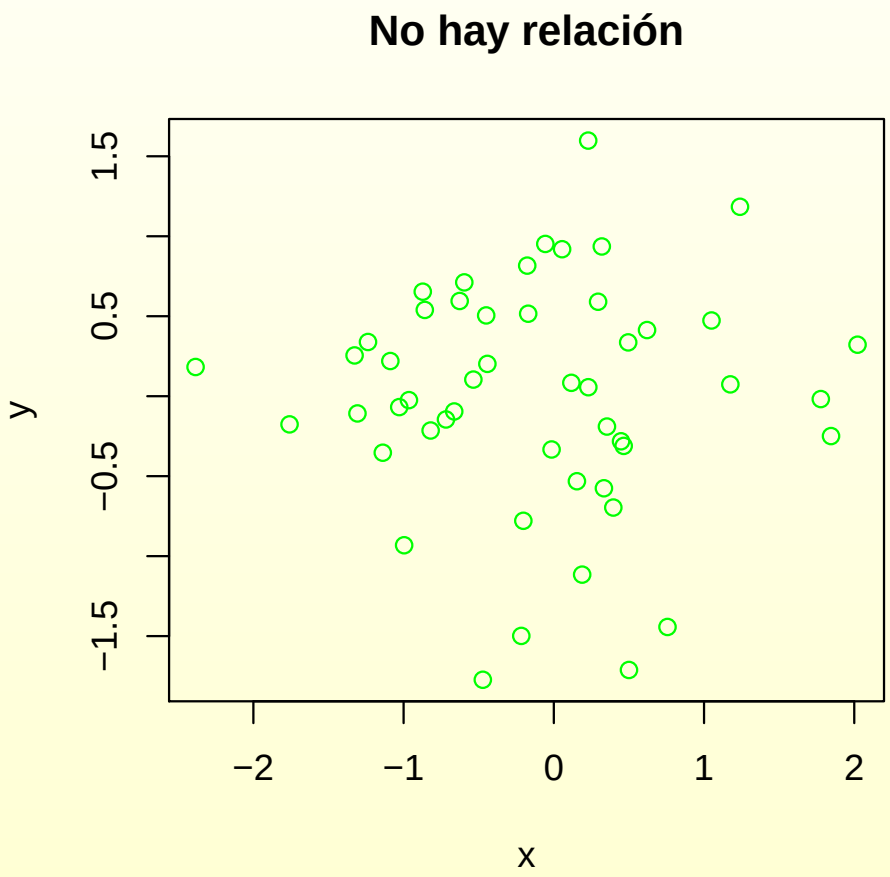
Regresión simple: intervienen dos variables

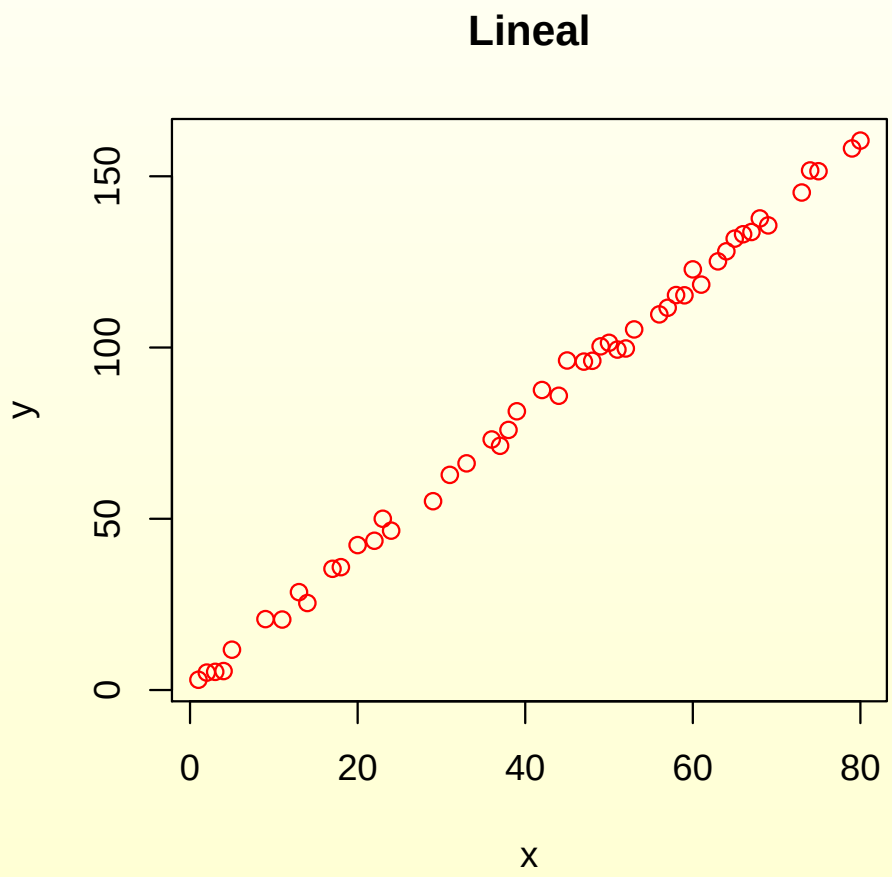




Regresión

Regresión simple: intervienen dos variables





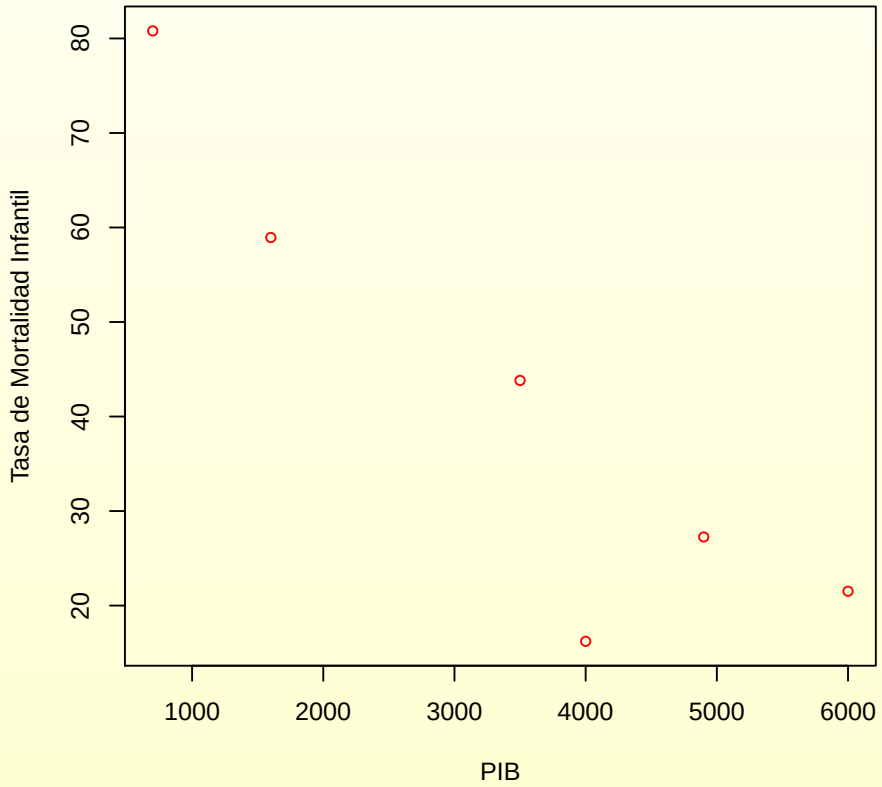
Ejemplo3: Población:Países con un PIB bajo
Variables:

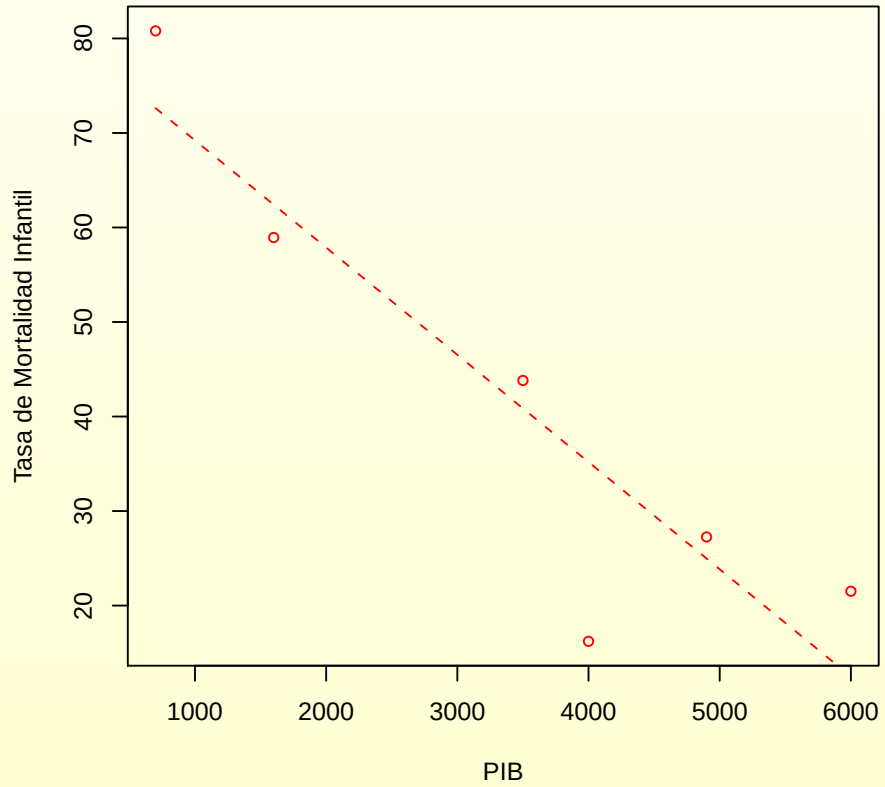
- Producto Interior Bruto
- Tasa de Mortalidad Infantil

La Tasa de Mortalidad Infantil depende del Producto Interior Bruto
Tomando una muestra de la población, obtenemos los siguientes paises:

Paises	Tasa de Mortalidad Infantil	PIB
Etiopía	80.8	700
Senegal	58.94	1600
Irak	43.82	3500
Egipto	27.26	4900
El Salvador	21.52	6000
Georgia	16.22	4000

Graficas de los datos y de los datos con la recta de regresión





Variables:

1. X : Producto Interior Bruto
2. Y : Tasa de Mortalidad Infantil

Recta de regresión:

$$\hat{Y} = aX + b,\ a \text{ y } b \text{ son constantes}$$

Objetivo: Obtener el valor de a y b

Método: mínimos cuadrados

Notación:

- $X_1 = 700, \ldots, X_7 = 4000$
- $Y_1 = 80.8, \ldots, Y_7 = 16.22$
- $\hat{Y}_1 = a \cdot 700 + b, \ldots, \hat{Y}_7 = a \cdot 4000 + b$

Paises	Tasa de Mortalidad Infantil	PIB
Etiopía	80.8	700
⋮	⋮	⋮
Georgia	16.22	4000

Regresión lineal simple

Método de mínimos cuadrados

Recta de regresión:

$$\hat{Y} = aX + b,\ a \text{ y } b \text{ son constantes}$$

Objetivo: Obtener el valor de a (pendiente de la recta) y b (punto de corte)
Queremos:

- que Y y $\hat{Y}$ sean lo más cercanos posibles
- i.e., que $(Y_1 - \hat{Y}_1)^2 + \cdots + (Y_n - \hat{Y}_n)^2$ sea lo más chico posible (en el ejemplo $n = 7$)
- i.e., encontrar a y b tal que $(Y_1 - a \cdot X_1 - b)^2 + \cdots + (Y_n - a \cdot X_n - b)^2$ sea lo más chico posible

Calculando:

$$a := \frac{(X_1 - \bar{X}) \cdot (Y_1 - \bar{Y}) + \cdots + (X_n - \bar{X}) \cdot (Y_n - \bar{Y})}{(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}$$

$$b := \bar{Y} - a \cdot \bar{X}$$

Regresión lineal simple

Método de mínimos cuadrados

Como $a = \dfrac{(X_1 - \bar{X}) \cdot (Y_1 - \bar{Y}) + \cdots + (X_n - \bar{X}) \cdot (Y_n - \bar{Y})}{(X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}$

Si denotamos

$Cov(X, Y) := (X_1 - \bar{X}) \cdot (Y_1 - \bar{Y}) + \cdots + (X_n - \bar{X}) \cdot (Y_n - \bar{Y})$

Entonces $a = \dfrac{Cov(X, Y)}{Var(X)}$

Ejemplo3:

$\bar{X} = 3450 \qquad \bar{Y} = 41.42667$

$$a = \frac{(700 - 3450)(80.8 - 41.43) + \cdots + (4000 - 3450)(16.22 - 41.43)}{(700 - 3450)^2 + \cdots + (4000 - 3450)^2}$$

$$= -45144.8/3979000 = -0.01134577$$

$$b = 41.42667 + 0.01134577 \cdot 3450 = 80.56958$$

Queremos saber como de bien ajusta la recta de regresión a los datos

El coeficiente de determinación evalua la fuerza de la relación lineal existente entre X e Y :

$$r^2 = \frac{(\hat{Y}_1 - \bar{Y})^2 + \cdots + (\hat{Y}_n - \bar{Y})^2}{(Y_1 - \bar{Y})^2 + \cdots + (Y_n - \bar{Y})^2}$$

$r^2 \in [0, 1]$,

- si r^2 está cerca de cero la relación lineal es baja
- si r^2 está cerca de 1 es alta

Resultado

Si para cada X fijo, la variable Y tiene una distribución Normal de media $aX+b$ y desviación típica constante, el estadístico $T := (n-2)\sqrt{Var(X)}asqrt(\hat{Y}_1 - Y_n)^2 + \cdots + (\hat{Y}_n - Y_n)^2$ sigue una distribución t de Student con $n-2$ grados de libertad

Nota: R utiliza el estadístico T^2

Con este resultado se puede contrastar

- H_0 : la pendiente de la recta es cero

contra

- H_a : la pendiente de la recta es distinta de cero

Correlación

Cuando una variable no es la causa-efecto de la otra:

- No parece razonable explicar una variable en términos de la otra
- Pero si: averiguar el grado de relación lineal entre ambas (sin pretender estimar ninguna recta de regresión)

Coeficiente de correlación lineal

$$\rho := \frac{Cov(X, Y)}{Var(X) \cdot Var(Y)}$$

Para estimar ρ se toma una muestra y se calcula

$$r = \sqrt{\frac{(\hat{Y}_1 - \bar{Y})^2 + \cdots + (\hat{Y}_n - \bar{Y})^2}{(Y_1 - \bar{Y})^2 + \cdots + (Y_n - \bar{Y})^2}}$$