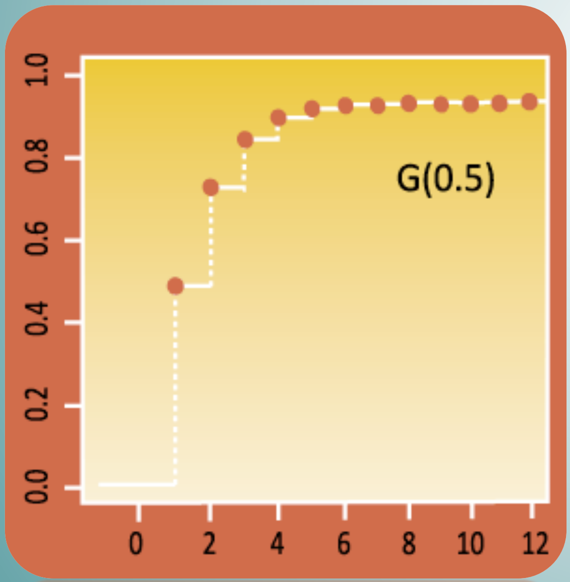
# Estadística y Métodos Numéricos

## Tema 6. Modelos de Regresión





**Ángel Barón Caldera**
**Ángel Cobo Ortega**
**María Dolores Frías Domínguez**
**Jesús Fernández Fernández**
**Francisco Javier González Ortiz**
**Carmen María Sordo García**

DEPARTAMENTO DE MATEMÁTICA APLICADA Y CIENCIAS DE LA COMPUTACIÓN

**UNIVERSIDAD DE CANTABRIA**

# TEMA 6: Modelos de regresión

Variables bidimensionales

Regresión lineal

Medidas de ajuste

# Variables bidimensionales

Variables que surgen cuando se estudian dos o más características asociadas a la observación de un fenómeno (e.g. peso y altura de una muestra de personas).

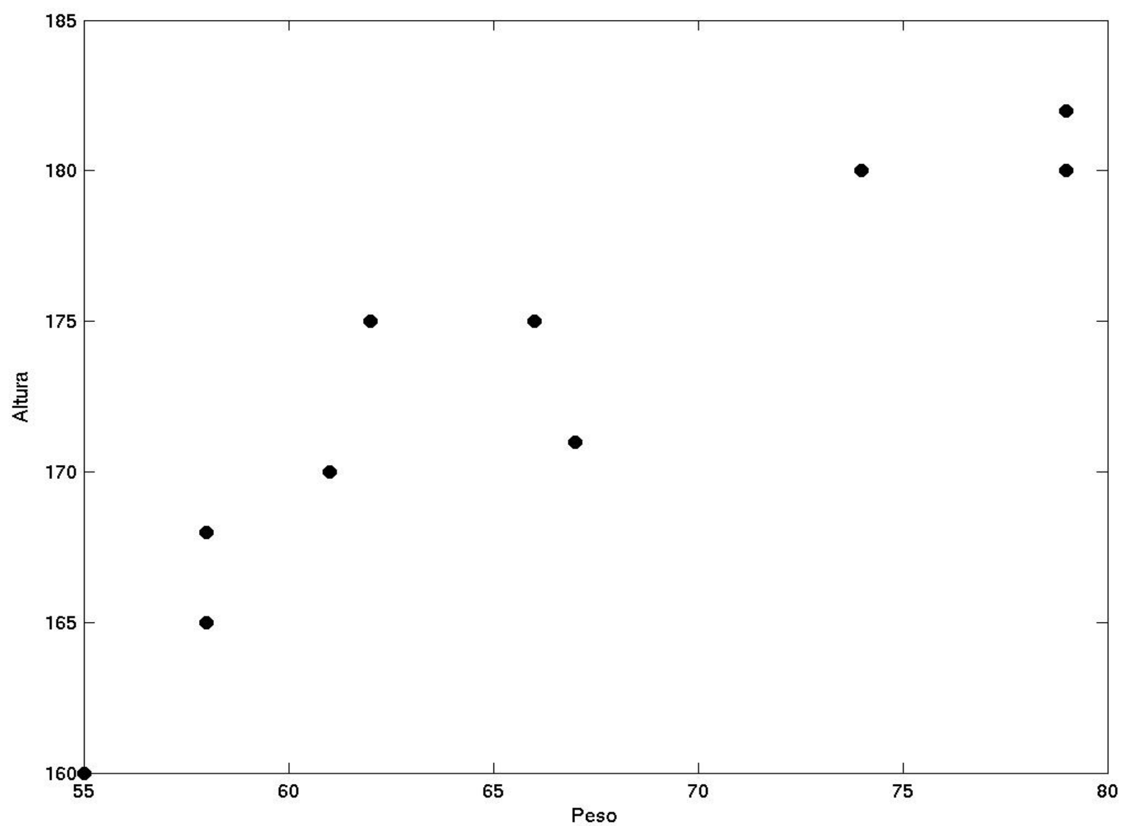
variable X: $x_1$, $x_2$..... $x_n$

variable Y: $y_1$, $y_2$..... $y_n$

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Altura (cm) | 160 | 165 | 168 | 170 | 171 | 175 | 175 | 180 | 180 | 182 |
| Peso (kg) | 55 | 58 | 58 | 61 | 67 | 62 | 66 | 74 | 79 | 79 |

# Variables bidimensionales

La forma más sencilla de representar gráficamente datos bidimensionales es mediante los denominados diagramas de dispersión (dibujo en los ejes cartesianos de la muestra).

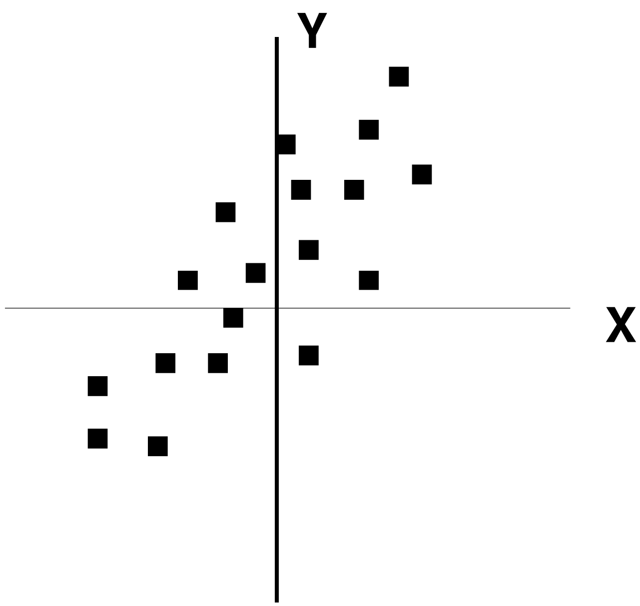



Se observa que cuando la altura aumenta el peso aumenta.

Existe una relación lineal directa entre las variables.

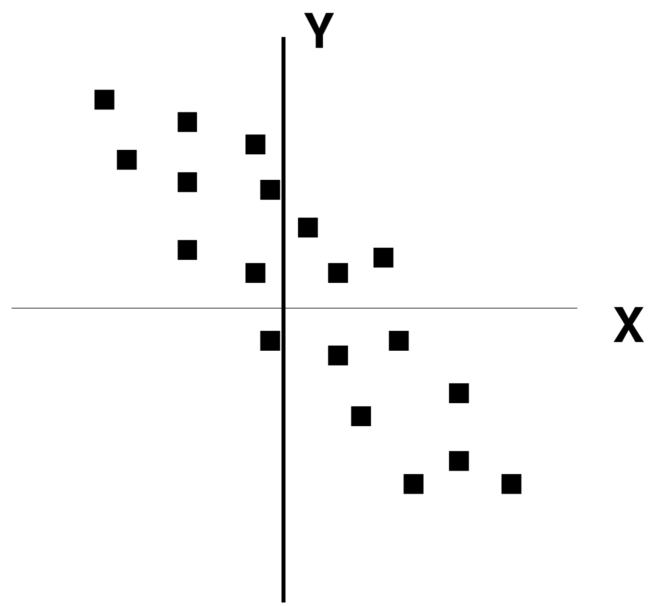
# Variables bidimensionales

La forma más sencilla de representar gráficamente datos bidimensionales es mediante los denominados diagramas de dispersión (dibujo en los ejes cartesianos de la muestra).




Cuando X crece Y crece: relación lineal directa.

Casi todos los puntos pertenecen al primer y tercer cuadrante




Cuando X crece Y decrece: relación lineal inversa.

Casi todos los puntos pertenecen al segundo y cuarto cuadrante.

# Variables bidimensionales

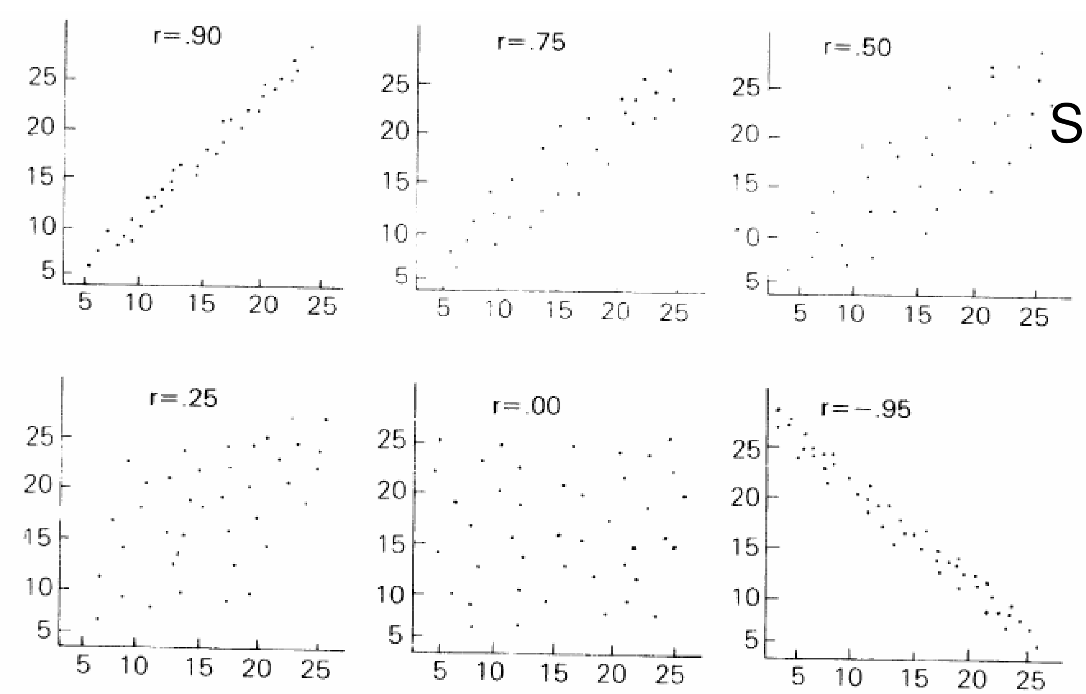
Es posible estimar la relación lineal entre dos variables mediante el coeficiente de correlación:

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{S_n(x)S_n(y)} = \frac{S_n(x,y)}{S_n(x)S_n(y)}$$

donde $S_n(x,y)$ es la covarianza muestral.

Toma valores entre 1 (dependencia directa) y -1 (dependencia inversa).

Si se acerca a 0 la dependencia lineal es débil.



# Variables bidimensionales

**Ejemplo:**

En una ciudad se quiere hacer un estudio sobre la utilización de una determinada línea urbana de autobús. Para ello se midieron en una misma parada 31 valores de intervalos de tiempo, en minutos, que transcurren entre las sucesivas llegadas de autobuses de dicha línea (X) y el número de viajeros que suben a él (Y), resultando los siguientes valores:

$\sum_i x_i = 290$ $\quad \sum_i x_i^2 = 2848$ $\quad \sum_i x_i y_i = 2995$

$\sum_i y_i = 315$ $\quad \sum_i y_i^2 = 3981$

$c)$ Estudie la relación entre X e Y mediante el coeficiente de correlación.

# Regresión

En la práctica surge con frecuencia la necesidad de tener que relacionar un conjunto de variables a través de una ecuación (ej, el peso de unas personas con su altura).

La **regresión** es una técnica estadística que permite construir modelos que representan la dependencia entre variables o hacer predicciones de una variable Y en función de las observaciones de otras ($X_1$, ..., $X_p$).

$$y = f(x_1, \cdots, x_p) + \epsilon$$

Y es la variable respuesta o dependiente

$X_1$, ..., $X_p$ son las variables predictoras, dependientes o covariables

$\epsilon$ es el término de error que se supone con media cero y varianza constante.

# Regresión

Las ecuaciones más comunes que se utilizan para expresar estas relaciones son:

Lineal

$$Y = a + bX + \epsilon$$

Cuadrática

$$Y = a + bX + cX^2 + \epsilon$$

Polinómica

$$Y = a_0 + a_1 X + \cdots + a_p X^p + \epsilon$$

Logarítmica

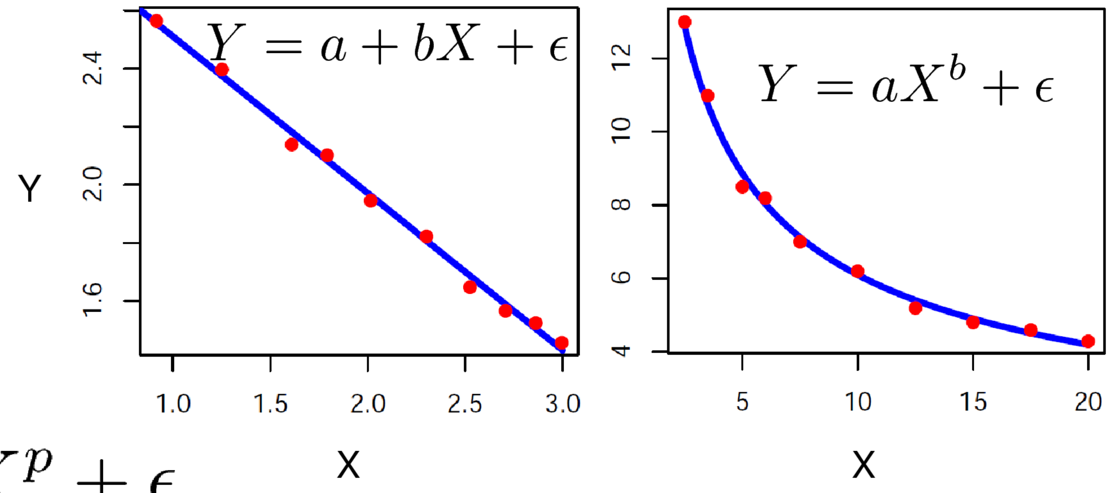
$$Y = a + b\, ln X + \epsilon$$

Exponencial

$$Y = a\, e^{bX} + \epsilon$$

Potencial

$$Y = aX^b + \epsilon$$



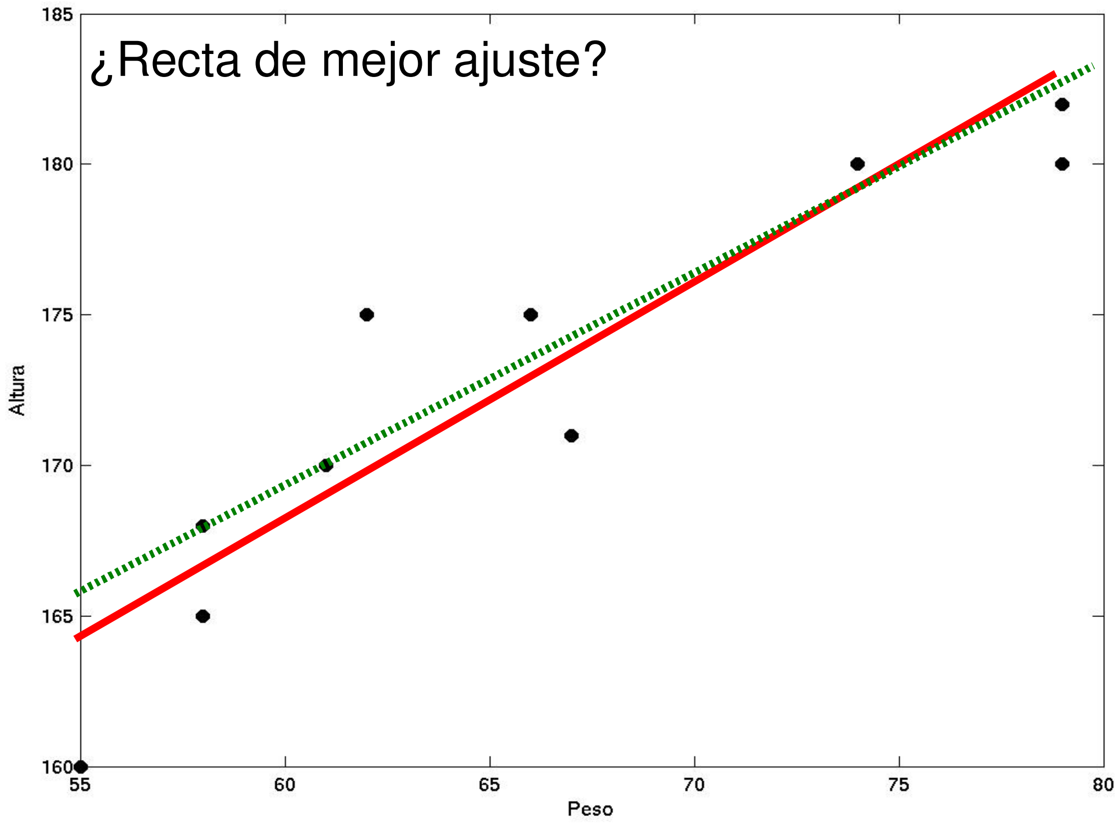

El diagrama de dispersión puede servir de gran ayuda a la hora de determinar la relación entre las variables.

Nos centraremos en los modelos de regresión lineales.

# Regresión lineal

Una vez seleccionado el modelo (lineal en nuestro caso) a ajustar a partir de las observaciones de una muestra se está interesado en estimar los parámetros de dicho modelo ($\beta_i$).

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$




Uno de los métodos más comunes es el de mínimos cuadrados que consiste en ajustar los parámetros del modelo de manera que la suma de los cuadrados de los errores sea mínima.

# Regresión lineal

En el caso más sencillo, regresión lineal simple, la ecuación

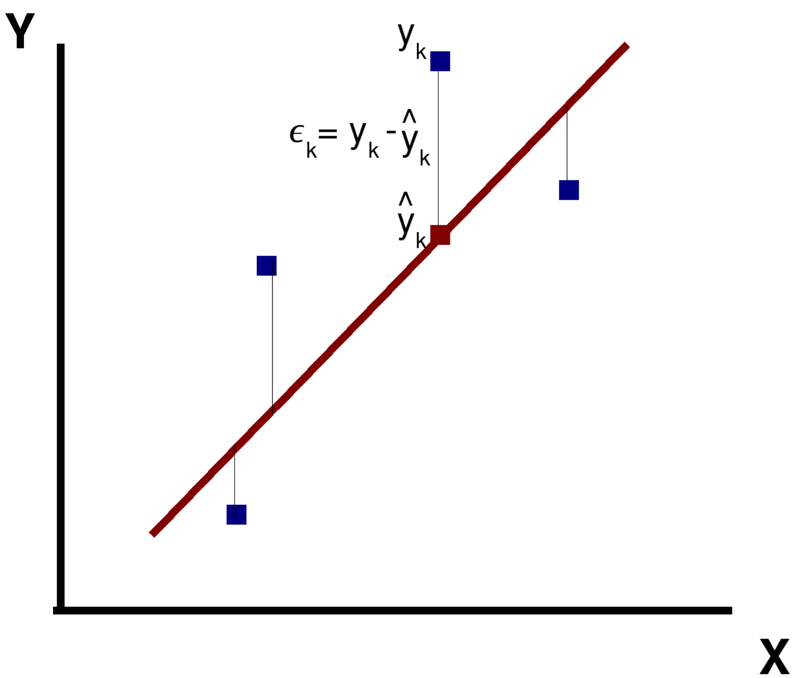
$$\hat{y} = a + bx$$

nos da una estimación de y, siendo el error que se comete,

$$\epsilon_k = y_k - \hat{y}_k$$

En este caso a y b se eligen de manera que,

$$E^2 = \sum_{k=1}^{n} (\varepsilon_k)^2 = \sum_{k=1}^{n} (y_k - \widehat{y}_k)^2 = \sum_{k=1}^{n} (y_k - a - bx_k)^2$$




sea mínimo $$\frac{\partial E^2}{\partial a} = \frac{\partial E^2}{\partial b} = 0$$

$$a = \frac{\sum_{k=1}^{n} y_k}{n} - b\frac{\sum_{k=1}^{n} x_k}{n} = \bar{y} - b\,\bar{x}$$

$$b = \frac{n \sum_{k=1}^{n} x_k\, y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} y_k}{n \sum_{k=1}^{n} x_k^2 - \left(\sum_{k=1}^{n} x_k\right)^2} = \frac{S_n(x,y)}{S_n^2(x)}$$

# Regresión lineal

## Ejemplo:

En una ciudad se quiere hacer un estudio sobre la utilización de una determinada línea urbana de autobús. Para ello se midieron en una misma parada 31 valores de intervalos de tiempo, en minutos, que transcurren entre las sucesivas llegadas de autobuses de dicha línea (X) y el número de viajeros que suben a él (Y), resultando los siguientes valores:

$\sum_i x_i = 290$ $\quad \sum_i x_i^2 = 2848$ $\quad \sum_i x_i y_i = 2995$

$\sum_i y_i = 315$ $\quad \sum_i y_i^2 = 3981$

*d*) Calcule la recta de regresión lineal del número de viajeros respecto al tiempo.

*e*) Estime, con la recta hallada, el número de viajeros que suben al autobús cuando el intervalo de tiempo entre autobuses es de 5 minutos.

# Regresión lineal

Esta formulación se extiende al caso de la regresión lineal múltiple $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$

en la que se observa una muestra ($y_k$, $x_{1k}$,...,$x_{pk}$) con k=1,...,n y se está interesado en estimar los parámetros del modelo.

Ej, estudios sobre el efecto de diversas condiciones climáticas (temperatura, humedad, radiación...) sobre la resistencia de un metal a la corrosión.

El modelo lineal se puede expresar en forma matricial de la forma:

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

$$\underset{n \times 1}{Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}} \quad \underset{n \times (p+1)}{X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{p1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{pmatrix}} \quad \underset{(p+1) \times 1}{\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}} \quad \underset{n \times 1}{\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}$$

# Regresión lineal

Aplicando el método de mínimos cuadrados para obtener los parámetros del modelo debemos minimizar:

$$E^2 = \sum_{k=1}^{n} \epsilon_k^2 = \sum_{k=1}^{n} [y_k - (\beta_0 + \beta_1 x_{1k} + \cdots + \beta_p x_{pk})]^2 = \epsilon^T \epsilon$$

$$\begin{aligned} \epsilon^T \epsilon &= (Y - X\beta)^T (Y - X\beta) \\ &= Y^T Y - \beta^T X^T Y - Y^T X\beta + \beta^T X^T X\beta \\ &= Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta \end{aligned}$$

derivando con respecto a $\beta$ e igualando la expresión resultante a cero se obtienen las ecuaciones normales:

$$(X^T X)\beta = X^T Y$$

que se reducirían a las ecuaciones normales obtenidas antes para el caso de la regresión lineal simple.

# Regresión lineal

**Ejemplo:**

Utilice la regresión lineal múltiple para ajustar los siguientes datos:

| Y | $X_1$ | $X_2$ |
|---|---|---|
| 5 | 0 | 0 |
| 10 | 2 | 1 |
| 9 | 2.5 | 2 |
| 0 | 1 | 3 |
| 3 | 4 | 6 |
| 27 | 7 | 2 |

# Regresión lineal

**Ejemplo:**

Utilice la regresión lineal múltiple para ajustar los siguientes datos:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Aplicando mínimos cuadrados:

$$E^2 = \sum_{k=1}^{n} \epsilon^2 = \sum_{k=1}^{n} (y_k - \beta_0 - \beta_1 x_{1k} - \beta_2 x_{2k})^2$$

Y derivando respecto a los parámetros:

$$\frac{\partial E^2}{\partial \beta_0} = -2 \sum_{k=1}^{n} (y_k - \beta_0 - \beta_1 x_{1k} - \beta_2 x_{2k}) = 0$$

$$\frac{\partial E^2}{\partial \beta_1} = -2 \sum_{k=1}^{n} x_{1k}(y_k - \beta_0 - \beta_1 x_{1k} - \beta_2 x_{2k}) = 0$$

$$\frac{\partial E^2}{\partial \beta_2} = -2 \sum_{k=1}^{n} x_{2k}(y_k - \beta_0 - \beta_1 x_{1k} - \beta_2 x_{2k}) = 0$$

| y | $x_1$ | $x_2$ | $x_1^2$ | $x_2^2$ | $x_1x_2$ | $x_1y$ | $x_2y$ |
|---|---|---|---|---|---|---|---|
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 2 | 1 | 4 | 1 | 2 | 20 | 10 |
| 9 | 2.5 | 2 | 6.25 | 4 | 5 | 22.5 | 18 |
| 0 | 1 | 3 | 1 | 9 | 3 | 0 | 0 |
| 3 | 4 | 6 | 16 | 36 | 24 | 12 | 18 |
| 27 | 7 | 2 | 49 | 4 | 14 | 189 | 54 |
| **54** | **16.5** | **14** | **76.25** | **54** | **48** | **243.5** | **100** |

En forma matricial:

$$\begin{pmatrix} n & \sum x_{1k} & \sum x_{2k} \\ \sum x_{1k} & \sum x_{1k}^2 & \sum x_{1k}x_{2k} \\ \sum x_{2k} & \sum x_{1k}x_{2k} & \sum x_{2k}^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \sum y_k \\ \sum x_{1k}y_k \\ \sum x_{2k}y_k \end{pmatrix}$$

$$\begin{pmatrix} 6 & 16.5 & 14 \\ 16.5 & 76.25 & 48 \\ 14 & 48 & 54 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 54 \\ 243.5 \\ 100 \end{pmatrix}$$

$$\beta_0 = 5 \quad \beta_1 = 4 \quad \beta_2 = -3$$

# Regresión lineal

## Ejercicio:

En un estudio se han recogido los siguientes valores de los ingresos totales de 40 familias frente a los gastos fijos por mes en euros:

| Gastos \ Ingresos | 300-600 | 600-840 | 840-1100 | 1100-1350 | 1350-1700 |
|---|---|---|---|---|---|
| 35-70 | 2 | | | | |
| 70-110 | 1 | 3 | 5 | | |
| 110-150 | | | 8 | 10 | |
| 150-180 | | | | 6 | 2 |
| 180-300 | | | | | 3 |

Calcular:

*a*) la recta de regresión mínimo cuadrática de los gastos fijos sobre los ingresos.

*b*) el coeficiente de correlación lineal.

*c*) con la recta calculada estime, los gastos fijos de una familia cuyos ingresos son de 200 euros.

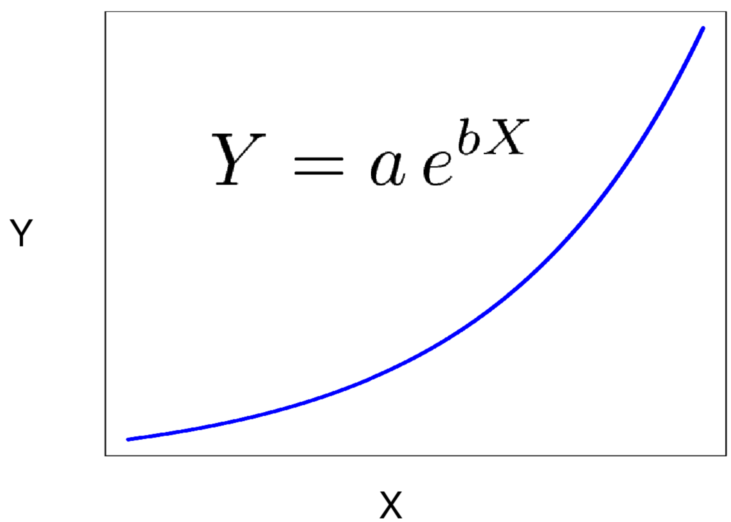
# Modelos no lineales

El método de mínimos cuadrados permite obtener la mejor recta de ajuste a los datos en el caso de la regresión lineal.

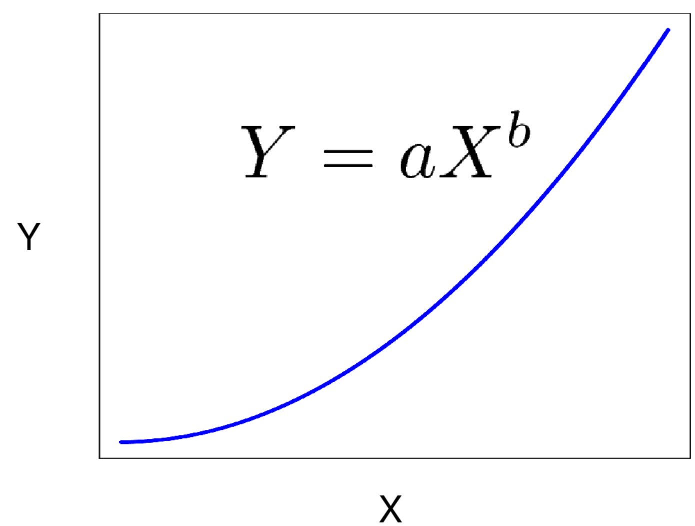
Sin embargo, no siempre existe una relación lineal entre la variable dependiente e independiente.

En algunos casos es posible aplicar transformaciones para expresar los datos en una forma compatible con la regresión lineal. Este es el caso del modelo exponencial y de potencias.

Exponencial

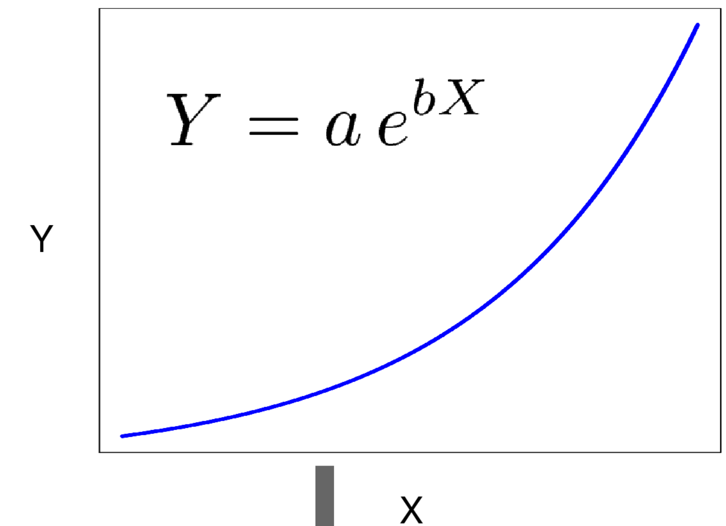



Potencial

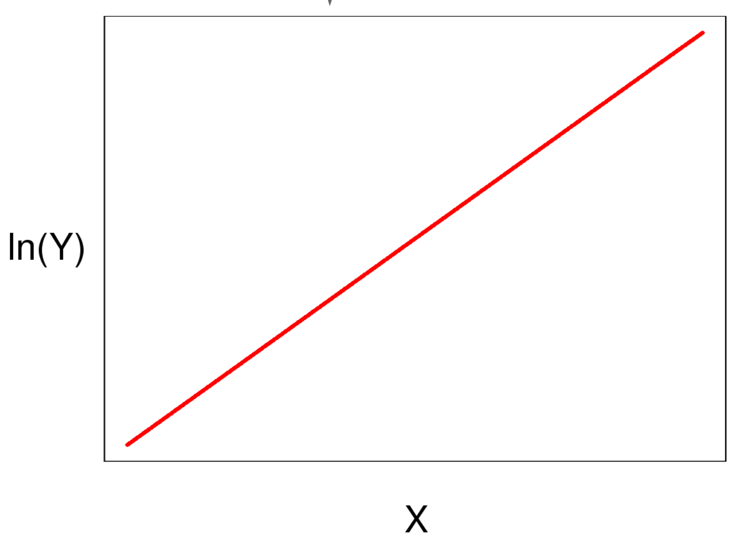


# Modelos no lineales

Exponencial




**Linealización**

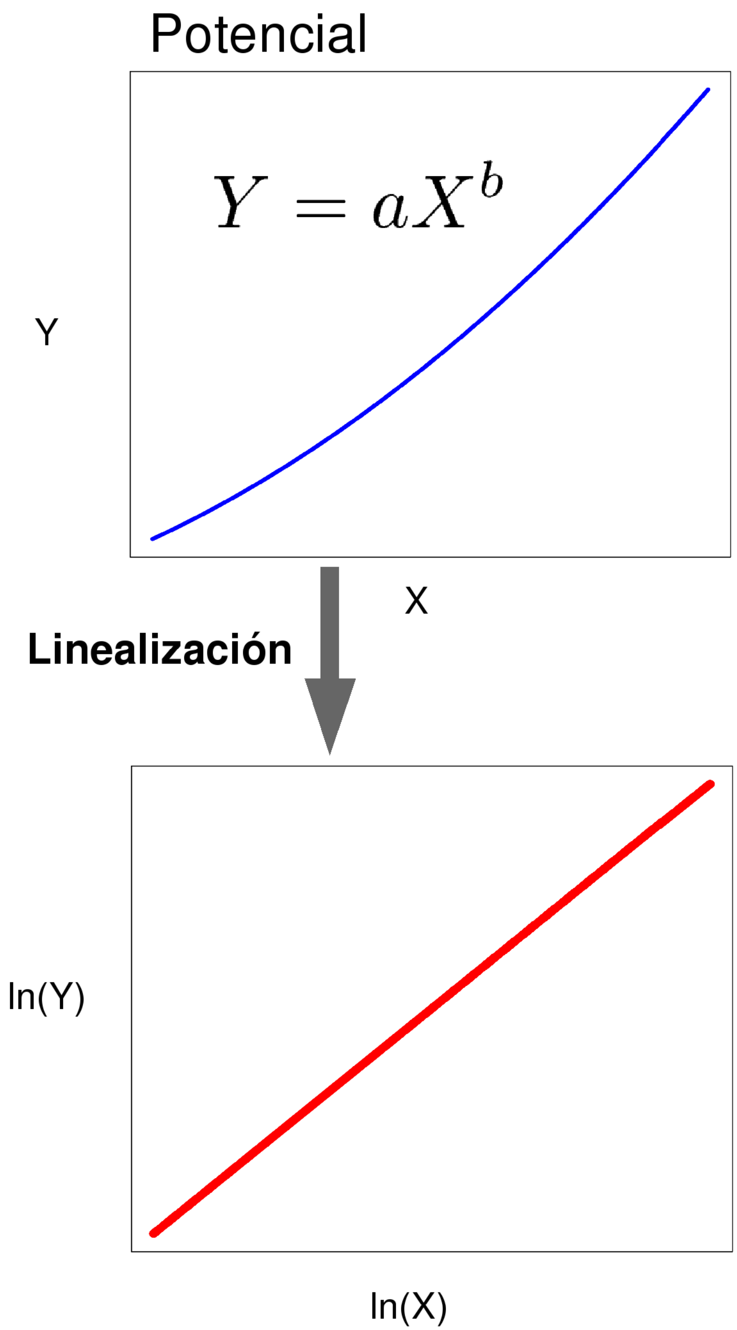



El modelo exponencial se linealiza al aplicar el logaritmo natural:

$$
\begin{aligned}
ln(Y) &= ln(a\, e^{bX}) \\
&= ln(a) + ln(e^{bX}) \\
&= ln(a) + bX
\end{aligned}
$$

donde si representamos el ln(Y) frente a X obtendremos una recta con pendiente b y corte con el eje de ordenadas ln(a).

# Modelos no lineales




El modelo potencial se linealiza al aplicar el logaritmo natural:

$$
\begin{aligned}
ln(Y) &= ln(a\,X^b) \\
&= ln(a) + ln(X^b) \\
&= ln(a) + b\ ln(X)
\end{aligned}
$$

donde si representamos el ln(Y) frente a ln(X) obtendremos una recta con pendiente b y corte con el eje de ordenadas ln(a).

# Modelos no lineales

**Ejemplo:**

Ajuste los datos siguientes con el modelo de potencias y aplique una transformación logarítmica para estimar los parámetros de dicho modelo. Use la ecuación resultante para hacer el pronóstico para x=9

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| x | 2.5 | 3.5 | 5 | 6 | 7.5 | 10 | 12.5 | 15 | 17.5 | 20 |
| y | 13 | 11 | 8.5 | 8.2 | 7 | 6.2 | 5.2 | 4.8 | 4.6 | 4.3 |

# Modelos no lineales

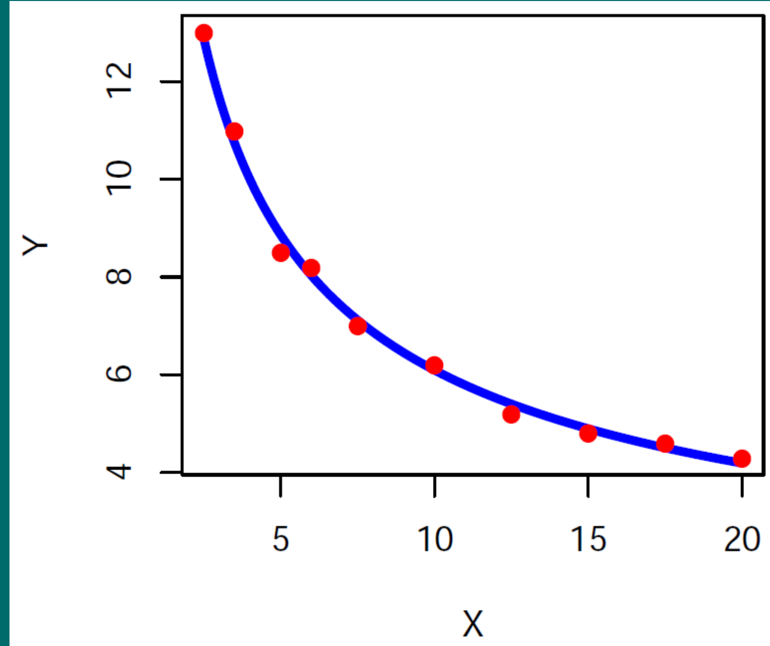
**Resolución con R:**

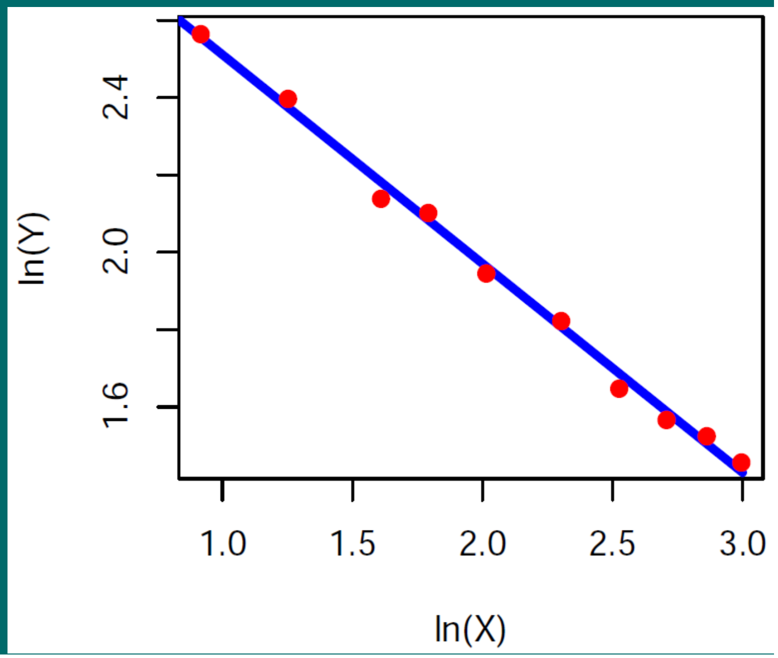
```
# Definición de variables
v1 <- c(2.5, 3.5, 5, 6, 7.5, 10, 12.5, 15, 17.5, 20)
v2 <- c(13, 11, 8.5, 8.2, 7, 6.2, 5.2, 4.8, 4.6, 4.3)
x <- log(v1)
y <- log(v2)

# Calculo regresión lineal
fit <- lm(y~x)
a <- fit$coeff[1]
b <- fit$coeff[2]

# Funcion potencial y recta
fx <- function(x,a,b) exp(a) * x^b
fxrecta <- function(x,a,b) a+b*x

# Plots
pdf("figura.pdf", width=7, height=3)
  par(mfrow=c(1,2), mar=c(4,4,1,1))
  plot(v1,v2, xlab="v1", ylab="v2", type="n")
    curve(fx(x,a,b), col="blue",lwd=4,add=TRUE)
    points(v1,v2, pch=19, col="red")
  plot(x,y, xlab="log(v1)", ylab="log(v2)", type="n")
    curve(fxrecta(x,a,b), col="blue",lwd=4,add=TRUE)
    points(x,y , pch=19, col="red")
dev.off()
```



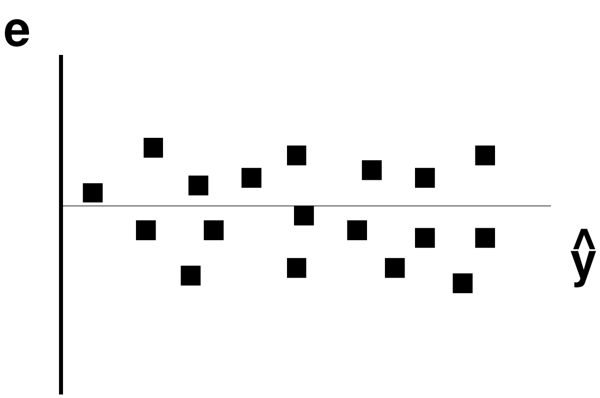



# Medidas de la idoneidad del modelo

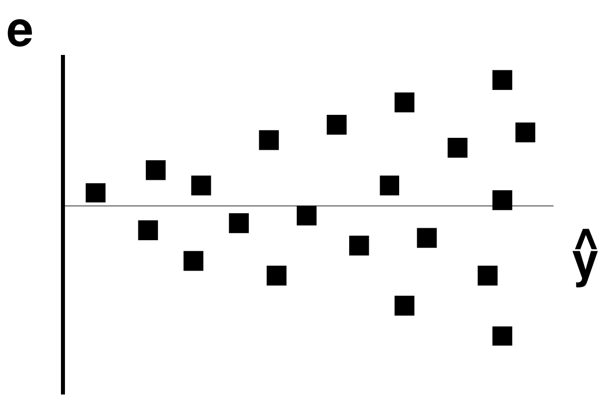
Toda la información sobre la falta de ajuste del modelo está contenida en los residuos.

Un diagrama de los residuos frente a los valores predichos nos sirve para detectar posibles desviaciones de las hipótesis de partida: valor medio cero y varianza constante.

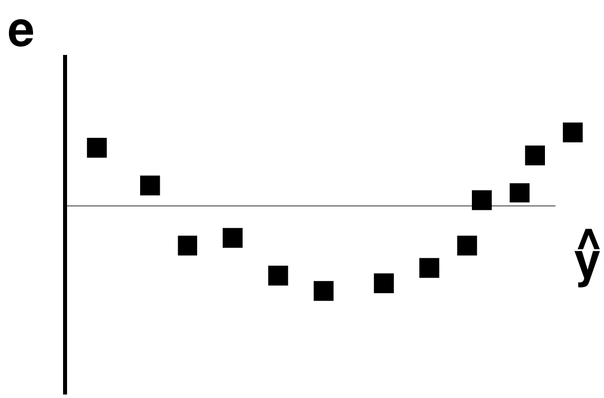
Errores típicos cuando el modelo no es el adecuado:




Caso ideal: media cero y varianza constante




Varianza no constante




Dependencia sistemática

También se recomienda pintar los residuos frente a la variable independiente para detectar posibles tendencias.

# Medidas de la calidad de ajuste

Es posible cuantificar la bondad del ajuste realizado en la regresión lineal simple al aplicar el método de mínimos cuadrados mediante las siguientes magnitudes:

**Error estandar de la estimación**, $S_e$:

$$S_e = \sqrt{\frac{E^2}{n-2}}$$

Cuantifica la dispersión de los datos alrededor de la línea de regresión.

Se divide entre n-2 ya que se usaron dos datos estimados ($\beta_0$ y $\beta_1$) para calcular $S_r$.

**Coeficiente de correlación**, r:

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{S_n(x)S_n(y)} = \frac{S_n(x,y)}{S_n(x)S_n(y)}$$

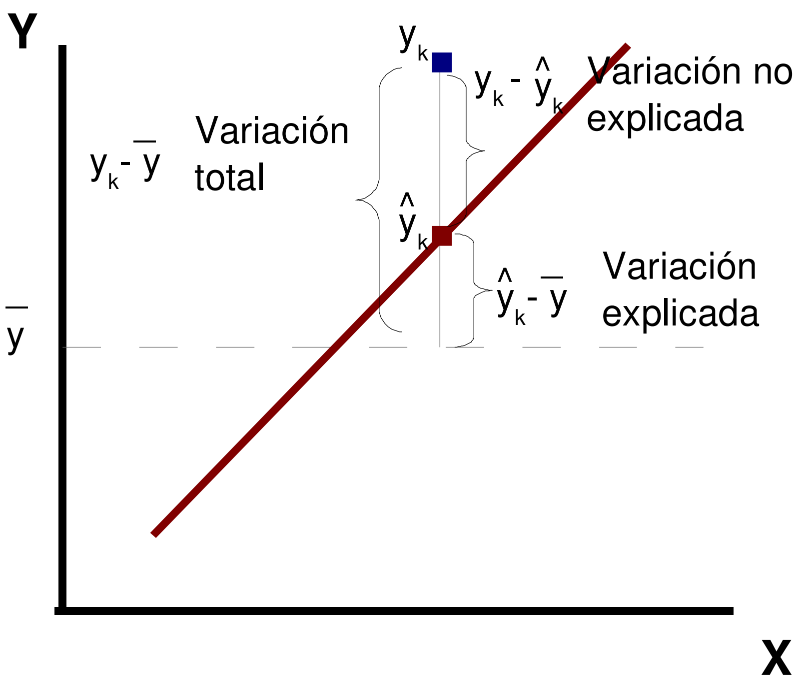
Cuantifica la relación lineal entre dos variables.

# Medidas de la calidad de ajuste

Es posible cuantificar la bondad del ajuste realizado en la regresión lineal simple al aplicar el método de mínimos cuadrados mediante las siguientes magnitudes:

Coeficiente de determinación, $r^2$:

Medida de la bondad del ajuste lineal. Indica la fracción de variación explicada por la recta de regresión respecto a la variación total.




$$r^2 = \frac{\sum_{k=1}^{n} (\hat{y}_k - \overline{y})^2}{\sum_{k=1}^{n} (y_k - \overline{y})^2} = \frac{S_{\hat{y}}^2}{S_y^2} = \frac{S_y^2 - \frac{E^2}{n}}{S_y^2}$$

Toma valores entre 0 y 1.

Cuanto más próximo a 1 mejor será el ajuste lineal y cuanto más proximo a 0 peor.

Coincide con el cuadrado del coeficiente de correlación

## Ejemplo:

Un ciclista se desplaza en línea recta con un movimiento uniforme para el cual según las leyes de la mecánica su posición x en un instante t vendrá dada por la ecuación $x = x_0 + vt$ donde $x_0$ es la posición inicial y v la velocidad.
Se han tomado los siguientes valores de su posición x en metros y el tiempo t en segundos:

| x (metros) | 14 | 26.2 | 37.7 | 51 | 61.8 | 76 | 84.2 |
|---|---|---|---|---|---|---|---|
| t (segundos) | 2 | 4 | 6 | 8 | 10 | 12 | 14 |

A partir de estos datos estimar:

*a*) el coeficiente de correlación

*b*) los valores de la posición inicial y la velocidad del ciclista por medio de una regresión lineal.

*c*) el espacio que recorrido por el ciclista transcurridos 9 segundos.

*d*) el error estandar de la estimación y la fracción de varianza explicada por el modelo.