# Ajuste por mínimos cuadrados

Mathieu Kessler

Departamento de Matemática Aplicada y Estadística
Universidad Politécnica de Cartagena

Cartagena, Enero 2010

# Guión

# Guión

1. Planteamiento
2. Criterio de mínimos cuadrados
3. Casos concretos: regresión lineal
   - La recta $y = ax + b$
   - Algunas transformaciones útiles

## Conjuntos de datos reales

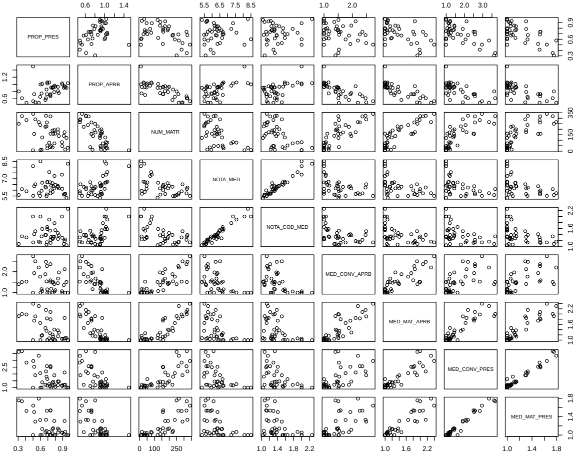
- A menudo, más de una variable asociada a cada individuo.
- Nos interesa las posibles relaciones
- Empezamos por una matriz de gráficas por pares...

# Matriz de gráficas, ejemplo

Para cada asignatura de la UPCT y por curso académico, hemos calculado:

- Proporción de presentados resp. a matriculados
- Proporción de aprobados respecto a los presentados.
- Número de matriculados.
- Nota media entre los alumnos que superan la asignatura durante ese académico.
- Entre los alumnos que han superado la asignatura durante ese curso, el número medio de convocatorias que han agotado para aprobar.
- Entre los alumnos que han superado la asignatura durante ese curso, el número medio de matriculas que han agotado para aprobar.
- Entre los alumnos que se han presentado por primera vez durante el curso, el número medio de convocatorias que han necesitado para ello.
- Entre los alumnos que se han presentado por primera vez durante el curso, el número medio de matriculas que han necesitado para ello

# Datos académicos, UPCT, Arquitectura Técnica, Curso 2006-07



### Nos centramos a partir de ahora

- Una variable respuesta $Y$
- Una (o más) variable(s) explicativas, $X$, (ó $X_1$, $X_2, \ldots$)

### Buscamos explicar la evolución de la respuesta en función de las explicativas

Construimos un modelo basado en los datos

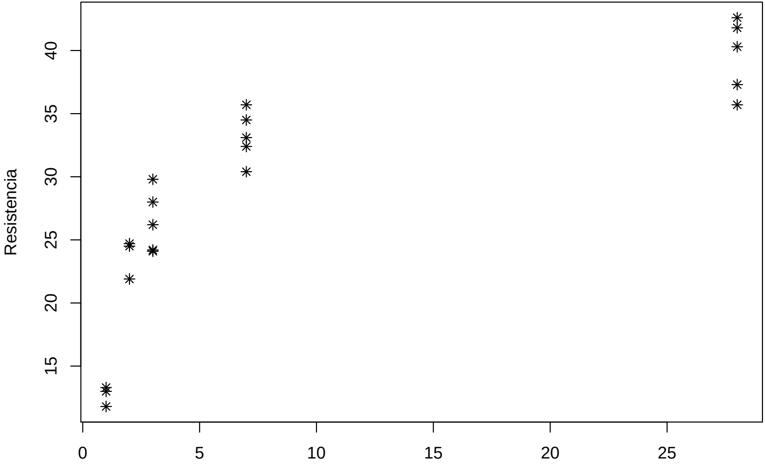
- Para entender
- Para predecir

### Ejemplo: evolución media en agosto en San Javier

Modelo: $Temperatura = -582.5 + 0.3 año,$

# Ejemplos de cuatro conjuntos de datos reales

## Resistencia del cemento en función del tiempo de fraguado en días (Hald 1952)

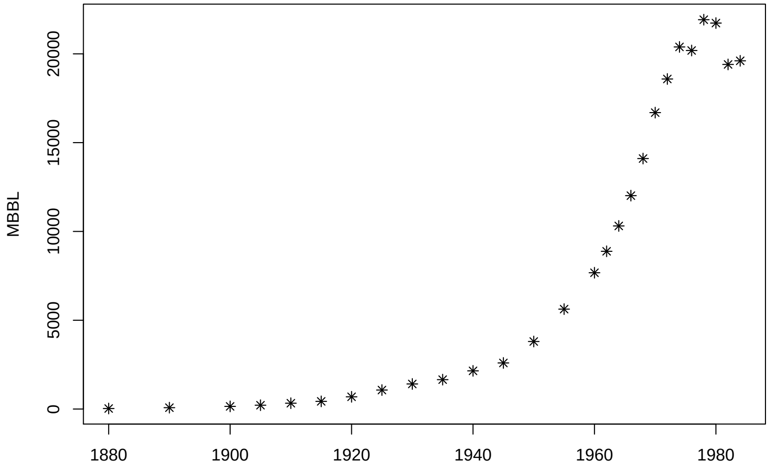


# Ejemplos de cuatro conjuntos de datos reales

## Nivel máximo anual del mar en Venecia

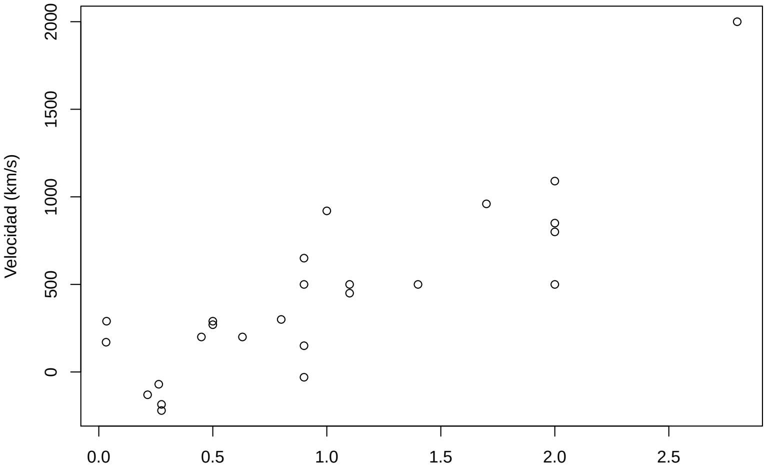
# Ejemplos de cuatro conjuntos de datos reales

## Producción mundial de petróleo



# Ejemplos de cuatro conjuntos de datos reales

## Velocidad de Recesión de 24 nebulosas



# Guión

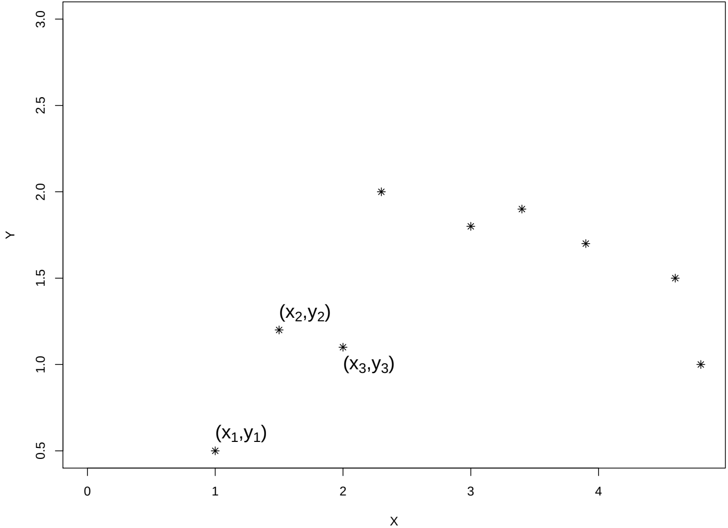
## Los datos

- Nos limitamos de momento a una variable respuesta $Y$ y una variable explicativa $X$.
- Presentación de los datos:

| $X$ | $Y$ |
|---|---|
| $x_1$ | $y_1$ |
| $x_2$ | $y_2$ |
| $\vdots$ | $\vdots$ |
| $x_n$ | $y_n$ |

- Empezamos por una nube de puntos

# Nube de puntos



# Ajuste

## Decidimos ajustar una curva de una determinada forma funcional

- Por ejemplo, una recta: $Y = aX + b$.
- Por ejemplo, una parábola: $Y = a_0 + a_1 X + a_2 X^2$.
- En general, especificamos una familia paramétrica:

$$x \mapsto f(\theta, x) \quad \theta = (\theta_1, \ldots, \theta_k),$$

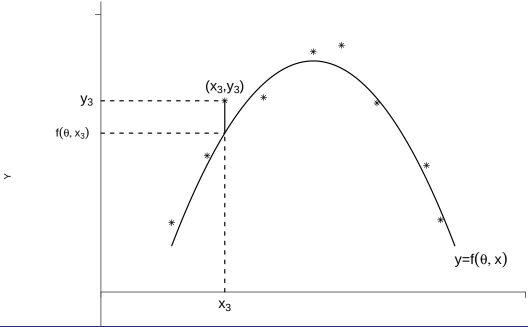
$\theta$ es el vector de parámetros.

## Nuestro objetivo

Buscamos la función de la familia que "mejor" se ajusta a la nube $\Leftrightarrow$ Debemos encontrar el valor concreto de $\theta$ que corresponde a esa función óptima

# Nuestro concepto de “mejor”: el criterio de mínimos cuadrados

Buscamos minimizar la suma de los cuadrados de las distancias verticales entre la curva y los datos de la nube de puntos.



# Ajuste por mínimos cuadrados

La suma de cuadrados:

$$SC(\theta) = (y_1 y_1 - f(\theta\theta, x_1 x_1))^2 + \cdots + (y_n y_n - f(\theta\theta, x_n x_n))^2.$$

Buscamos el valor $\hat{\theta}$ de $\theta$ que minimiza $\theta \mapsto SC(\theta)$.

## Nota

En $SC(\theta)$, $x_1, \ldots, x_n$ e $y_1, \ldots, y_n x_1, \ldots, x_n$ e $y_1, \ldots, y_n$ son valores fijados concretos. Sólo $\theta\theta$ es variable.

## Para la minimización

- Si la forma de $f$ es simple, (lineal respecto a $\theta_1, \ldots \theta_k$), tenemos fórmulas explícitas para calcular $\hat{\theta}$. (Regresión lineal)
- Si la forma de $f$ es más complicada, se debe recurrir a minimización numérica (Regresión no lineal)

## Algunos términos

- La curva de ecuación $y = f(\hat{\theta}, x)$: curva ajustada.
- Los valores $\hat{y}_1 = f(\hat{\theta}, x_1), \ldots, y_n = f(\hat{\theta}, x_n)$: valores ajustados.
- Las distancias verticales entre los puntos observados y la curva ajustada: los residuos $e_1, \ldots, e_n$. Tenemos

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n.$$

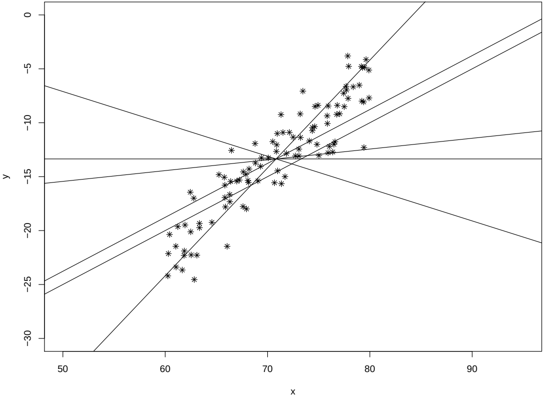
- La suma de cuadrados $SC(\hat{\theta}) = \sum_{i=1}^n e_i^2$ se llama suma de cuadrados residuales.
- La varianza de los residuos: varianza residual

$$s_e^2 = \frac{1}{n-1} \sum_{i=1}^n (e_i - \bar{e})^2.$$

# Guión

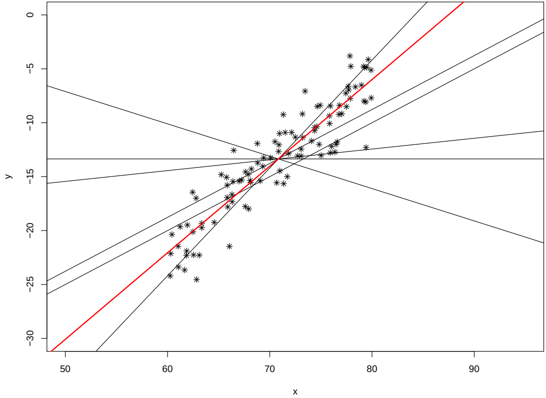
# Recta $y = ax + b$.

Entre todas las posibles rectas:



# Recta $y = ax + b$.

Buscamos la mejor:



# Ecuación de la recta ajustada

## La suma de cuadrados:

$$\theta = (a, b), \quad SC(\theta) = SC(a, b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2.$$

## La minimización

- Los candidatos para el mínimo:

$$\frac{\partial}{\partial a} SC(a, b) = 0, \qquad \frac{\partial}{\partial b} SC(a, b) = 0.$$

- Obtenemos

$$\hat{a} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} \qquad \hat{b} = \bar{y} - \hat{a}\bar{x}.$$

# Ecuación de la recta ajustada

$$\boxed{\hat{a} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} \qquad \hat{b} = \bar{y} - \hat{a}\bar{x}.}$$

Con nuestros datos:

| $X$ | $Y$ |
|---|---|
| $x_1$ | $y_1$ |
| $x_2$ | $y_2$ |
| $\vdots$ | $\vdots$ |
| $x_n$ | $y_n$ |

Necesitaremos calcular:

La media de $X$, $(\bar{x})$. La media de $y$, $(\bar{y})$. La media del produ

La media de $X^2$, $(\overline{x^2})$. La media de $Y^2$, $(\overline{y^2})$.

$$\boxed{\hat{a} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - (\bar{x})^2} \qquad \hat{b} = \bar{y} - \hat{a}\bar{x}.}$$

Introducimos:

$$s_{xy} = \frac{n}{n-1}(\overline{xy} - \bar{x}\bar{y}),$$

que llamamos la covarianza de $X$ e $Y$.
Por lo tanto:

$$\boxed{\hat{a} = \frac{s_{xy}}{var(x)} \qquad \hat{b} = \bar{y} - \hat{a}\bar{x}.}$$

Otra forma de escribir la ecuación:

$$\boxed{y - \bar{y} = \frac{s_{xy}}{var(x)}\left(x - \bar{x}\right).}$$

# Tipos de asociación lineal

La covarianza puede ser positiva o negativa, pero del mismo signo que la pendiente $\hat{a}$.

- Covarianza positiva = asociación positiva: cuando crece una variable crece la otra.
- Covarianza negativa = asociación negativa: cuando crece una variable decrece la otra.

# Bondad del ajuste

## Propiedades de los residuos

- La media de los residuos es nula.
- La varianza residual es

$$s_e^2 = s_y^2 \left(1 - \frac{(s_{xy})^2}{s_x^2 s_y^2}\right).$$

$$s_e^2 = s_y^2 \left(1 - \frac{(s_{xy})^2}{s_x^2 s_y^2}\right).$$

## Correlación lineal

Introducimos

- $r = \frac{s_{xy}}{s_x s_y}$: coeficiente de correlación (de Pearson) de $X$ e $Y$.

# Bondad del ajuste

$$s_e^2 = s_y^2 \left(1 - \frac{(s_{xy})^2}{s_x^2 s_y^2}\right)$$

$$s_e^2 = s_y^2 \left(1 - \frac{(s_{xy})^2}{s_x^2 s_y^2}\right) \Rightarrow \boxed{s_e^2 = s_y^2 \left(1 - R^2\right).}$$

## Propiedades de $r$ y $R^2$

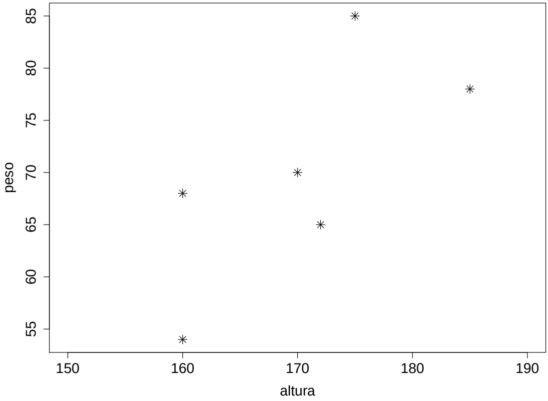
- $0 \leq R^2 \leq 1$.
- Cuanto más próximo a 1 esté $R^2$, mejor ajuste.
- $R^2 = 1 \Rightarrow s_e^2 = 0 \Rightarrow$ ajuste perfecto.
- $-1 \leq r \leq 1$.
- $r$ es del mismo signo que la pendiente.
- Cuanto más próximo a $\pm 1$ esté $r$, mejor ajuste.

# Ejemplo

Queremos estudiar la relación entre el peso y la altura en un grupo de individuos. Los datos son

| Peso(kg) | 54 | 70 | 65 | 78 | 68 | 85 | $Y$ |
|---|---|---|---|---|---|---|---|
| Altura(cm) | 160 | 170 | 172 | 185 | 160 | 175 | $X$ |

# Ejemplo, nube de puntos



# Cálculos

$$\bar{x} = \frac{160+170+\ldots+175}{6} = 170.33, \qquad \bar{y} = \frac{54+70+\ldots+85}{6} = 70,$$

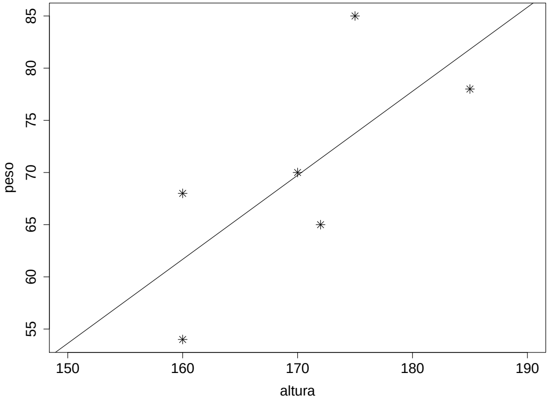
$$\overline{x^2} = \frac{160^2+170^2+\ldots+175^2}{6} = 29089, \qquad \overline{y^2} = \frac{54^2+70^2+\ldots+85^2}{6} = 4995.7,$$

$$\overline{xy} = \frac{160\times 54+170\times 70+\ldots+175\times 85}{6} = 11984.2$$

$$\begin{aligned}
s_x^2 &= \frac{n}{n-1}(\overline{x^2} - (\bar{x})^2) = \frac{6}{5}[29089 - (170.33)^2] \simeq 90.7,\\
s_y^2 &= \frac{n}{n-1}(\overline{y^2} - (\bar{y})^2) = \frac{6}{5}[4995.7 - (70)^2] \simeq 144.8,\\
s_{xy} &= \frac{n}{n-1}(\overline{xy} - (\bar{x})(\bar{y})) = \frac{6}{5}[11984.2 - 170.33 \times 70] \simeq 73.
\end{aligned}$$

$$y - 70 = \frac{73}{90.7}(x - 170.33), \qquad \boxed{y = 0.80x - 67.1}.$$

# Ejemplo, nube de puntos con recta ajustada



# Ejemplo, bondad del ajuste

$$r = \frac{s_{xy}}{s_x\, s_y} = \frac{73}{\sqrt{90.7}\sqrt{114.8}} \simeq 0.715,$$

lo que implica que $R^2 \simeq 0.51$, mal ajuste.
Se suele considerar buen ajuste a partir de $R^2 \geq 0.8$ aprox.

# Predicción

**Disponemos de un modelo ajustado, lo usamos para predecir**

$x_0$ valor no observado de $X$, nuestra predicción del valor de la respuesta $Y$ será:

$$y_{x_0} = \hat{a}\, x_0 + \hat{b}.$$

Ejemplo: ¿A qué peso correspondería una altura de 180cm?
$peso = 0.8 altura - 67.1 \Rightarrow peso \simeq 0.80 \times 180 - 67.1 \simeq 76.9 kg$

**¡CUIDADO!**

Es peligroso extrapolar nuestro modelo lejos del rango observado de valores de $X$.
Ejemplo: ¿a qué peso correspondería la altura de un niño de 80 cm?

# Transformaciones que permiten linealizar

## Contexto

- Hay fórmulas explícitas para el ajuste de modelos lineales. Para modelos no lineales, en general se usan algoritmos numéricos.
- Algunos modelos específicos no lineales se pueden ajustar con los métodos lineales.

## Veremos:

- Modelo exponencial $y = be^{ax}$
- Modelo potencial $y = bx^a$.

# Modelo exponencial

## Ecuación del modelo exponencial

$$y = be^{ax}, \qquad b > 0$$

- Si $a > 0$, crecimiento exponencial.
- Si $a < 0$, decrecimiento exponencial.

# Ajuste del modelo exponencial

## Modelo original y modelo transformado

Modelo teórico original

$$y = be^{ax}$$

$$\xrightarrow{\text{aplico ln}}$$

Modelo transformado

$$\ln(y) = \ln(b) +$$

$$y' = b' + a'x'$$

Llamemos las variables transformadas $Y' = \ln(Y)$, y $X' = X$, tenemos: $Y' = a'X' + b'$.

# Procedimiento

Modelo transformado: $Y' = \ln(Y)$, y $X' = X$, tenemos:

$$Y' = a'X' + b'.$$

Añadimos una columna a nuestros datos:

| $X$ | $Y$ | $Y' = \ln(Y)$ |
|---|---|---|
| $x_1$ | $y_1$ | $\ln(y_1)$ |
| $x_2$ | $y_2$ | $\ln(y_2)$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $x_n$ | $y_n$ | $\ln(y_n)$ |

# Procedimiento (II)

- Ajustamos ahora una recta de $Y'$ sobre $X$.
- Hacemos la transformación inversa del modelo ajustado para obtener el ajuste original.

## Ejemplo

Queremos ajustar un modelo exponencial a los siguientes datos

| $X$ | $Y$ | $Y' = \ln(Y)$ |
|---|---|---|
| 2.3 | 2.92 | 1.07 |
| 5 | 3.69 | 1.31 |
| 7.1 | 6.19 | 1.82 |
| 8 | 6.36 | 1.85 |

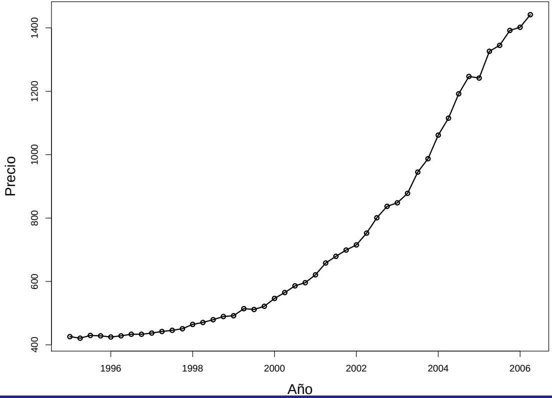
## Algunas transformaciones útiles

- Obtenemos $y' = 0.148x' + 0.682$
- es decir que $\ln(y) = 0.148x + 0.682$,
- lo que implica que

$$y = e^{0.148x}e^{0.682} = 1.18e^{0.148x}.$$

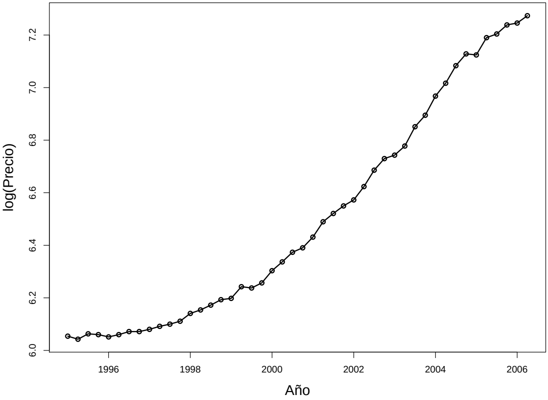
# Modelo exponencial: ejemplo

Evolución del precio del metro cuadrado de la vivienda en la Región de Murcia entre 1995 y 2006 ( datos cuatrimestrales, fuente: Ministerio de la Vivienda )



# Modelo exponencial: ejemplo

¿Qué deberíamos observar al representar el logaritmo de $Y$ en función de $X$?



# Modelo potencial

## Ecuación del modelo potencial

$$y = bx^a \qquad x > 0$$

La forma de la curva depende del valor de $a$.

# Ajuste del modelo potencial exponencial

## Modelo original y modelo transformado

Modelo teórico original

$$y = bx^a$$

$\xrightarrow{\text{aplico ln}}$

Modelo transformado

$$\ln(y) = \ln($$

$$y' = b' + a'$$

Llamemos las variables transformadas $Y' = \ln(Y)$, y $X' = \ln(X)$, tenemos: $Y' = a'X' + b'$.

# Procedimiento

Modelo transformado: $Y' = \ln(Y)$, y $X' = \ln(X)$, tenemos:

$$Y' = a'X' + b'.$$

Añadimos dos columnas a nuestros datos:

| $X$ | $Y$ | $X' = \ln(X)$ | $Y' = \ln(Y)$ |
|---|---|---|---|
| $x_1$ | $y_1$ | $\ln(x_1)$ | $\ln(y_1)$ |
| $x_2$ | $y_2$ | $\ln(x_2)$ | $\ln(y_2)$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_n$ | $y_n$ | $\ln(x_n)$ | $\ln(y_n)$ |

# Procedimiento

- Ajustamos ahora una recta de $Y'$ sobre $X'$.
- Hacemos la transformación inversa del modelo ajustado para obtener el ajuste original.

## Ejemplo

Queremos ajustar un modelo potencial a los siguientes datos

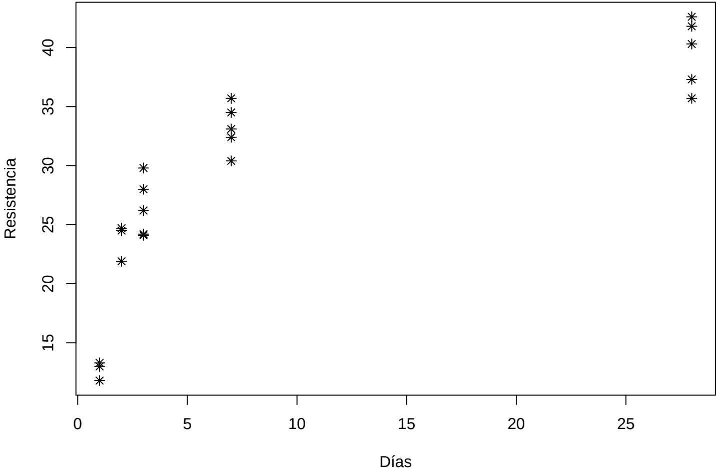
| $X$ | $Y$ | $X' = \ln(X)$ | $Y' = \ln(Y)$ |
|---|---|---|---|
| 3 | 10.3 | 1.1 | 2.3 |
| 7.34 | 13.5 | 2 | 2.6 |
| 20.1 | 18.2 | 3 | 2.9 |
| 54.6 | 24.5 | 4 | 3.2 |

- Obtenemos $y' = 0.298x' + 2.006$,
- es decir que $\ln(y) = 0.298\ln(x) + 2.006$,
- lo que implica que

$$y = e^{0.298\ln(x)}e^{2.006} = 7.433x^{0.298}.$$

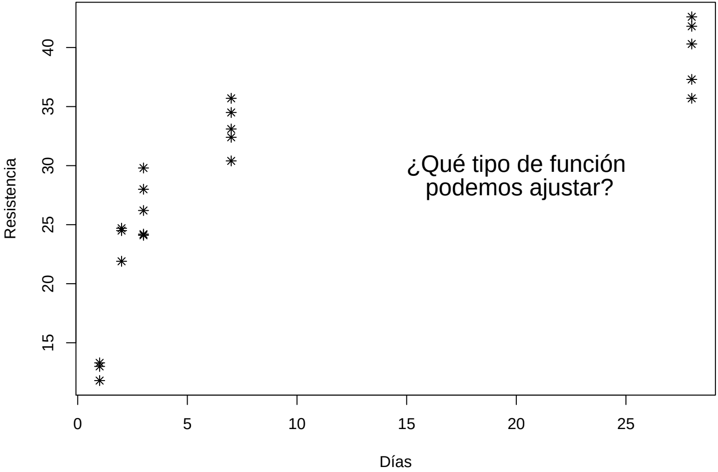
# Un último ejemplo

Volvemos a la resistencia del cemento en función del tiempo de fraguado:



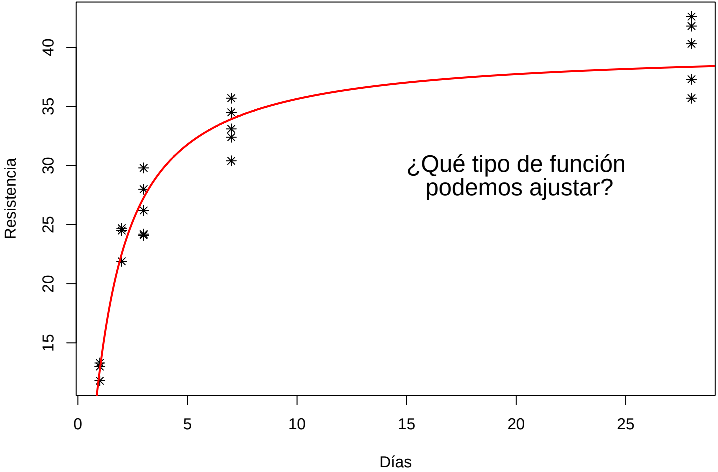
# Un último ejemplo

Volvemos a la resistencia del cemento en función del tiempo de fraguado:



# Un último ejemplo

Volvemos a la resistencia del cemento en función del tiempo de fraguado:



$$R = C\exp^{-k/t}$$

¿Interpretación de $C$?

$$\ln(R) = 3.688 - 1.146/t$$

$$\Rightarrow \boxed{R = 39.96\exp^{-1.146/t}.}$$