Tema 3: Regresión lineal

José R. Berrendero

Universidad Autónoma de Madrid

Estructura de este tema

- El modelo de regresión lineal simple
- Regresión lineal múltiple
- Estimadores de mínimos cuadrados
- Inferencia sobre los parámetros del modelo
- Análisis de la varianza
- Contrastes de hipótesis lineales
- Diseño de experimentos: modelo unifactorial

El problema de regresión simple

Observamos dos variables, X e Y, el objetivo es analizar la relación existente entre ambas, de forma que podamos predecir o aproximar el valor de la variable Y a partir del valor de la variable X.

- La variable Y se llama **variable respuesta**
- La variable X se llama **variable regresora o explicativa**

En un problema de regresión (a diferencia de cuando calculamos el coeficiente de correlación) el papel de las dos variables no es simétrico.

Fracaso escolar y nivel de renta en la CAM

EL PAÍS, martes 18 de octubre de 2005

El fracaso escolar es más alto en las zonas con menor renta

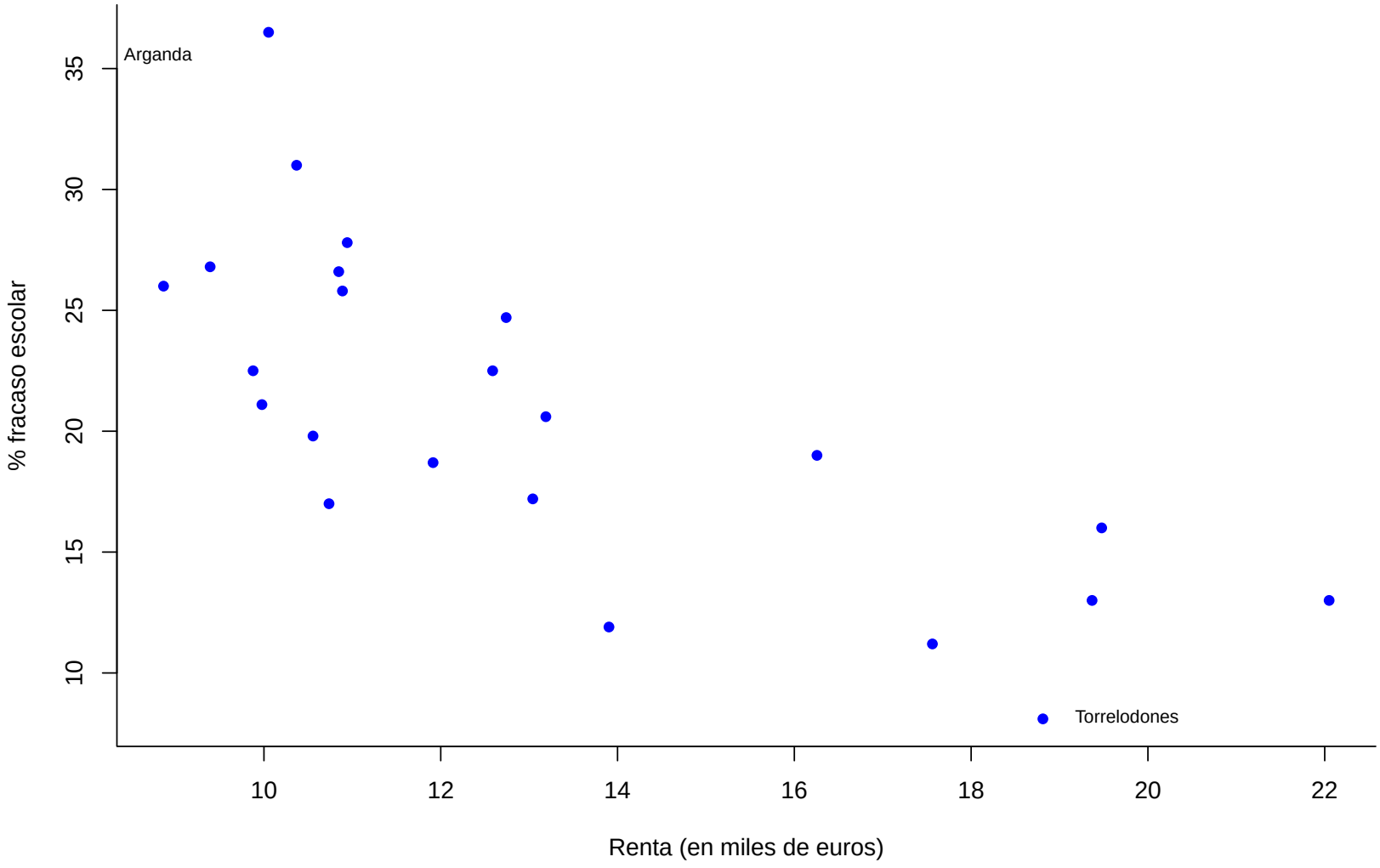
Fracaso escolar en la Comunidad de Madrid

Renta per cápita bruta media en 2003: 13.095 euros

	CURSO 2003/2004 Renta (euros)	Fracaso escolar (%)
Parla	8.864	26,0
Fuenlabrada	9.391	26,8
Leganés	9.877	22,5
Móstoles	9.977	21,1
Arganda	10.052	36,5
Torrejón	10.369	31,0
Getafe	10.555	19,8
Coslada	10.736	17,0
Pinto	10.846	26,6
Alcorcón	10.888	25,8
Alcalá de Henarés	10.942	27,8
Collado	11.913	18,7
Colmenar Viejo	12.587	22,5
Arroyomolinos	12.740	24,7
S. Sebastián de los Reyes	13.041	17,2
S. Lorenzo del Escorial	13.189	20,6
Rivas	13.903	11,9
Alcobendas	16.256	19,0
Tres Cantos	17.562	11,2
Torrelodones	18.812	8,1
Boadilla	19.368	13,0
Majadahonda	19.477	16,0
Pozuelo	22.050	13,0

EL PAÍS

Diagrama de dispersión



Recta de regresión

Frecuentemente, la relación entre las dos variables es lineal:

$$Y_i \approx \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n.$$

Problema estadístico: estimar los parámetros β_0 y β_1 a partir de los datos (x_i, Y_i), $i = 1, \ldots, n$.

Si estimamos β_0 y β_1 mediante $\hat{\beta}_0$ y $\hat{\beta}_1$, la predicción de la variable respuesta Y_i en función de la regresora x_i es:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Unos buenos estimadores deben ser tales que los errores de predicción

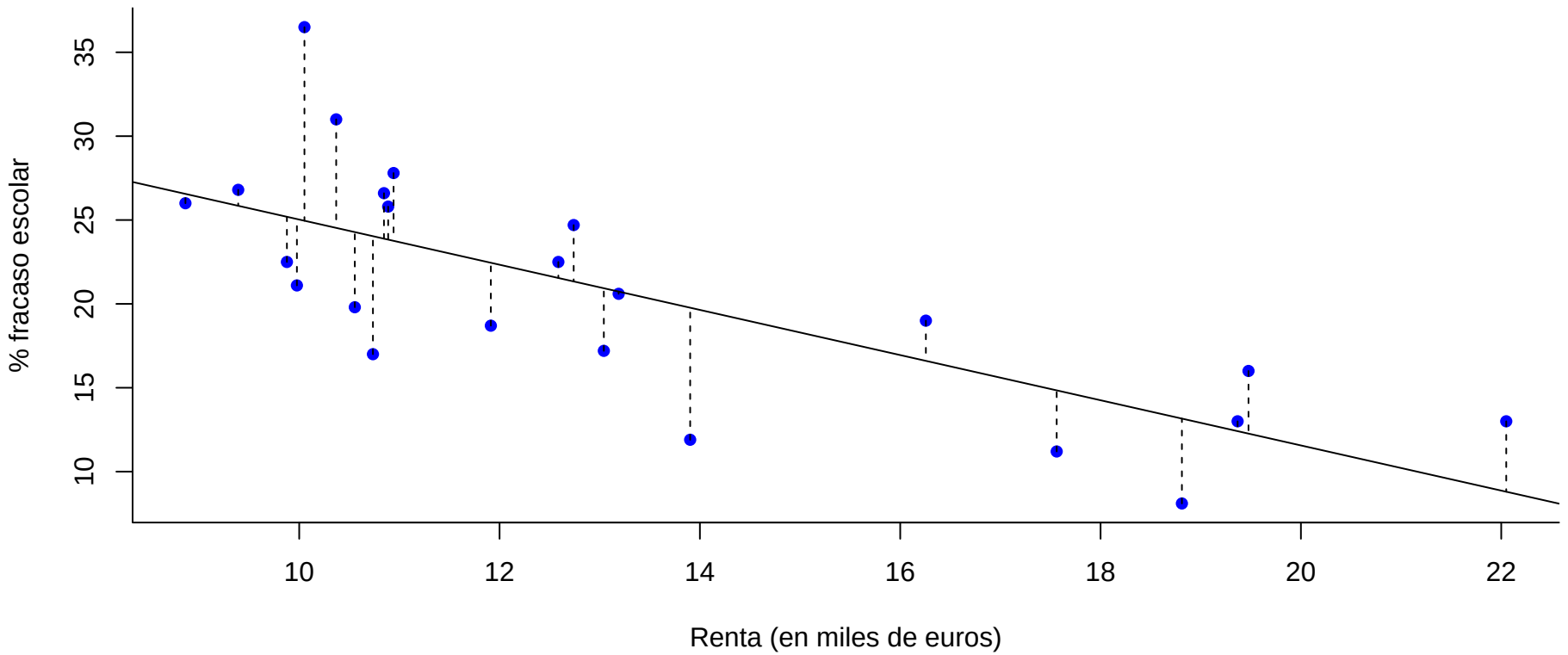
$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

sean pequeños.

La recta de mínimos cuadrados

La recta de regresión de mínimos cuadrados viene dada por los valores $\hat{\beta}_0$ y $\hat{\beta}_1$ para los que se minimiza:

$$\sum_{i=1}^{n}[Y_i - (\beta_0 + \beta_1 x_i)]^2.$$



Estimadores de mínimos cuadrados

Pendiente

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} = r\frac{S_y}{S_x}.$$

donde $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})$ y $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Término independiente

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

Recta de mínimos cuadrados

$$y - \bar{y} = \frac{S_{xy}}{S_{xx}}(x - \bar{x})$$

A los errores $e_i = Y_i - \hat{Y}_i$ se les llama **residuos**

A las predicciones $\hat{Y}_i$ se les llama **valores ajustados**

Mínimos cuadrados como promedio de pendientes

Vamos a reescribir la expresión del estimador de mínimos cuadrados:

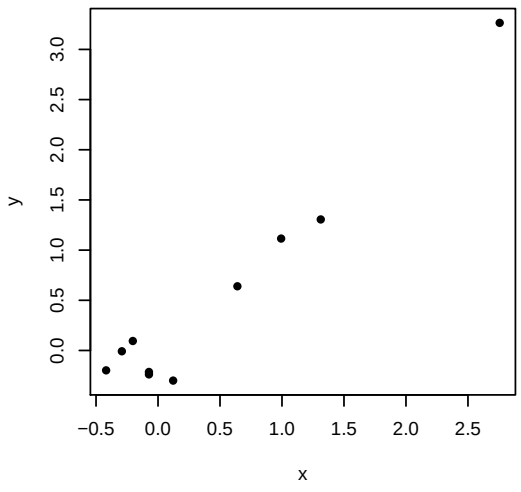
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{S_{xx}} \left(\frac{Y_i - \bar{Y}}{x_i - \bar{x}} \right) = \sum_{i=1}^{n} w_i \left(\frac{Y_i - \bar{Y}}{x_i - \bar{x}} \right),$$

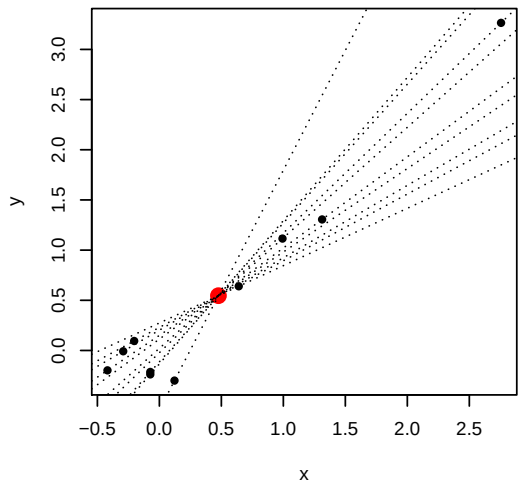
donde $w_i = (x_i - \bar{x})^2 / S_{xx}$.

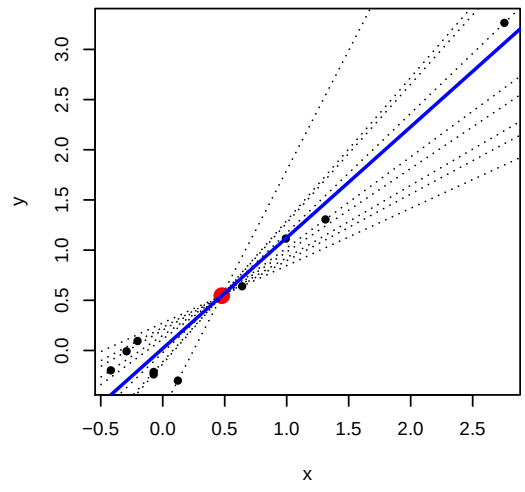
El estimador de la pendiente es una media ponderada de las pendientes de las rectas que unen cada punto (x_i, Y_i) con el vector de medias $(\bar{x}, \bar{Y})$.

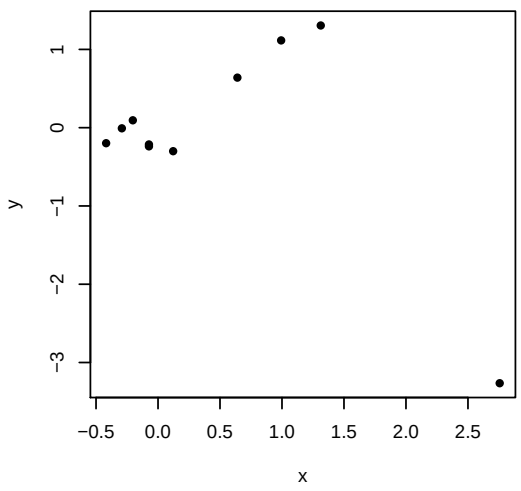
La ponderación w_i que recibe cada punto (x_i, Y_i) es mayor cuanto mayor es la distancia entre x_i y $\bar{x}$.

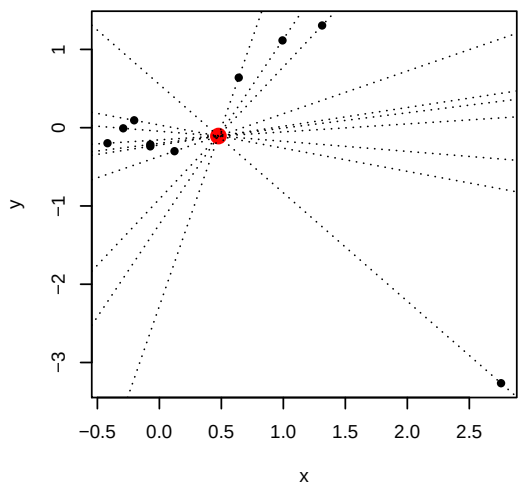
Mínimos cuadrados como promedio de pendientes

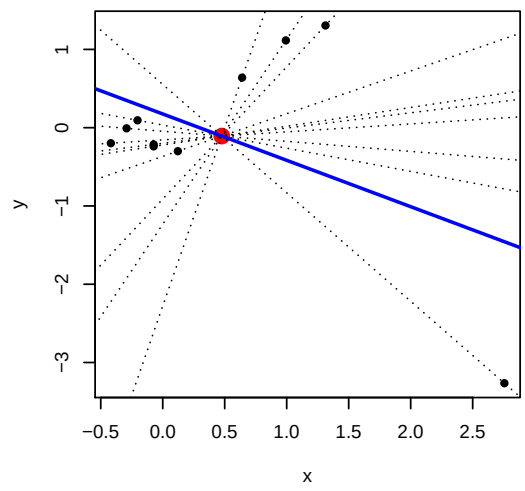












Mínimos cuadrados como promedios de respuestas

Otra expresión alternativa del estimador de la pendiente:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{S_{xx}} = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S_{xx}}\right) Y_i \equiv \sum_{i=1}^{n} \alpha_i Y_i.$$

El estimador de la pendiente es una combinación lineal de las respuestas Y_i.

El valor absoluto de los coeficientes, $|\alpha_i|$, también aumenta con la distancia entre x_i y $\bar{x}$.

Calcula:

- $\sum_{i=1}^{n} \alpha_i$
- $\sum_{i=1}^{n} \alpha_i x_i$
- $\sum_{i=1}^{n} \alpha_i^2$

El modelo de regresión lineal simple

Para poder hacer inferencia (IC y contrastes) sobre los parámetros, suponemos que se verifica el siguiente modelo:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

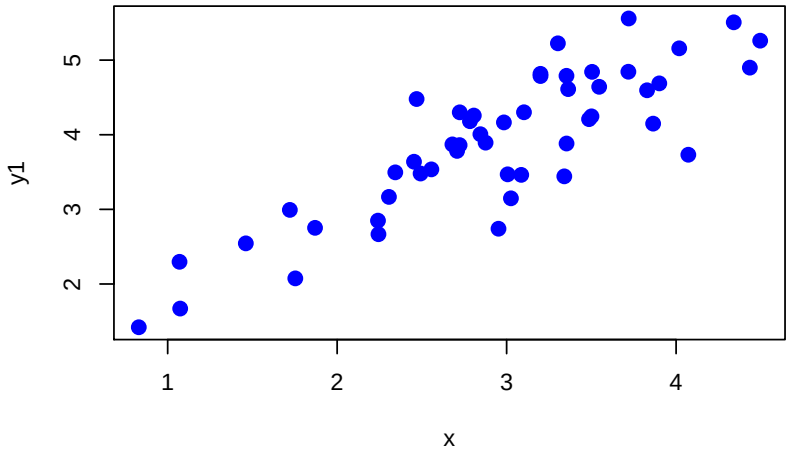
donde:

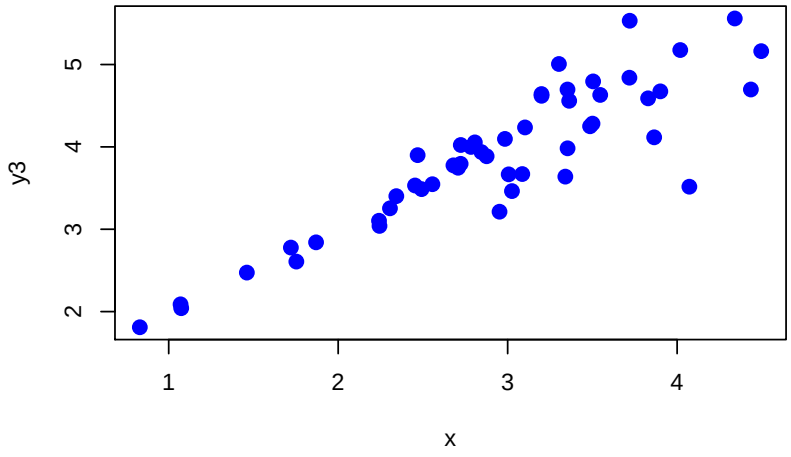
- El valor esperado de los errores ϵ_i es cero.
- Todos los errores tienen la misma varianza σ^2.
- Los errores tienen distribución normal.

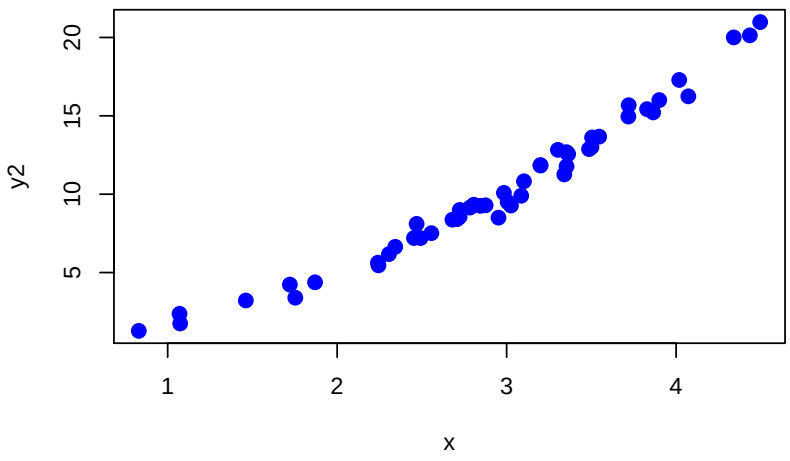
En resumen:

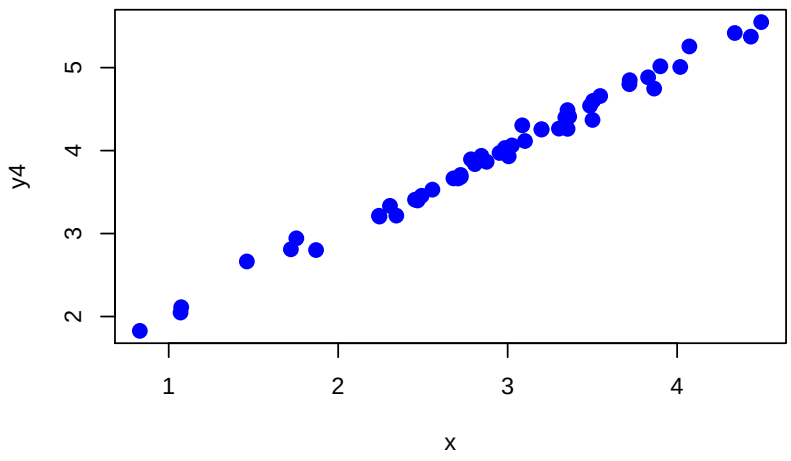
$$\epsilon_1, \ldots, \epsilon_n \quad \equiv \mathrm{N}(0, \sigma^2) \ \text{ independientes}$$

¿En cuáles de estas situaciones se verifica el modelo?









Simulación

- Supongamos que $\sigma = 1$, $\beta_0 = 0$ y $\beta_1 = 1$

- Entonces el modelo es

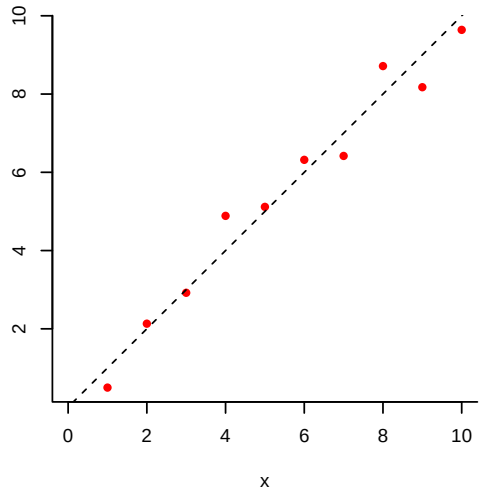
$$Y_i = x_i + \epsilon_i,$$

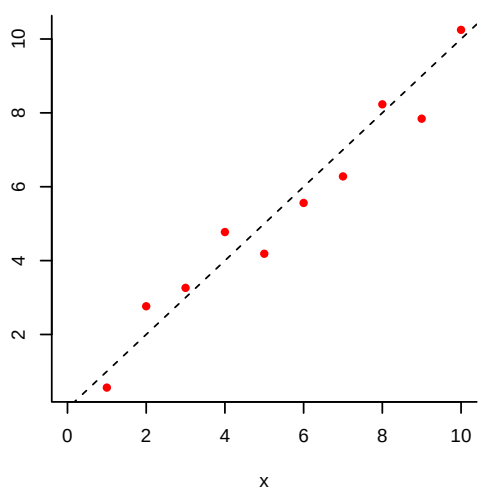
  donde los errores ϵ_i tienen distribución normal estándar y son independientes

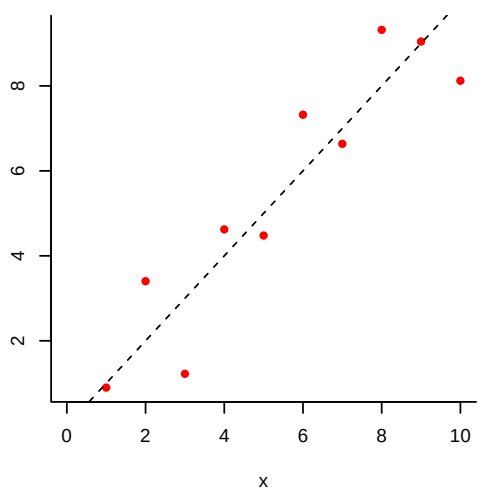
- Fijamos $x_i = 1, 2, \ldots, 10$ ($n = 10$) y generamos las respuestas correspondientes de acuerdo con este modelo

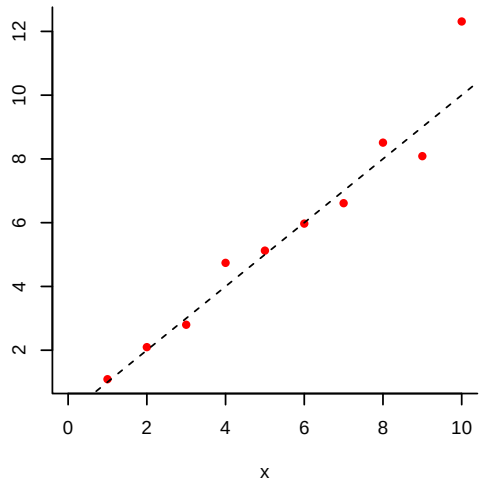
- Posteriormente calculamos la recta de mínimos cuadrados y la representamos junto con la *verdadera recta* $y = x$

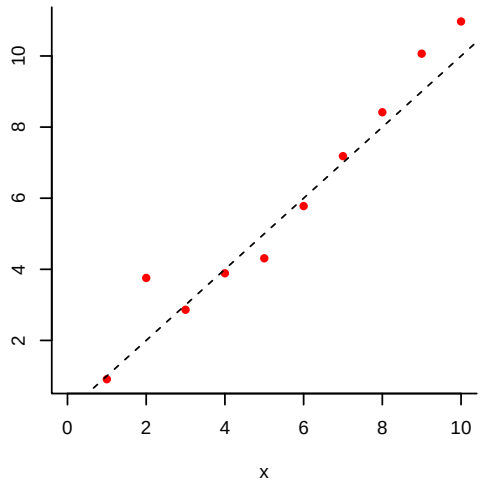
Repetimos seis veces el experimento

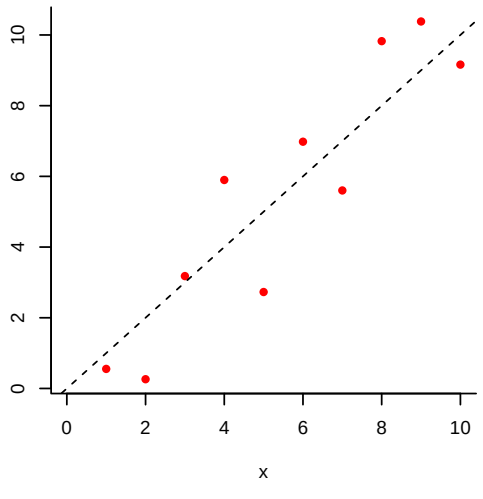




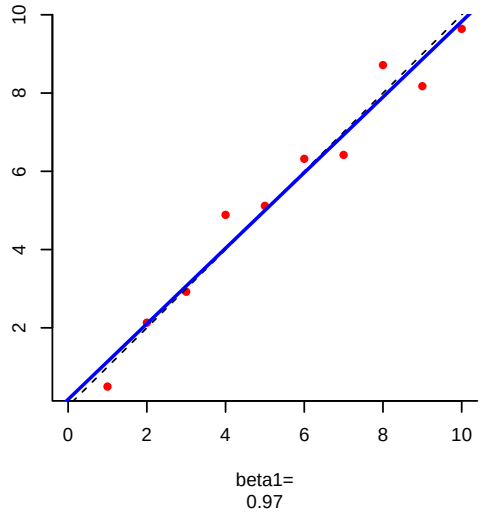


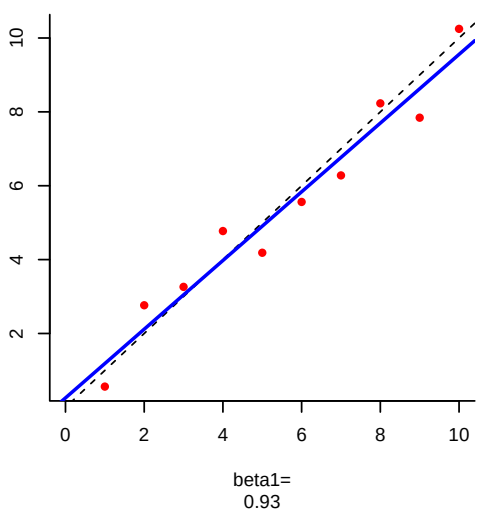


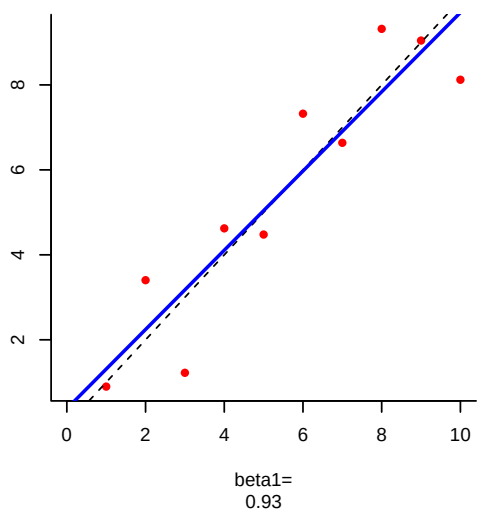


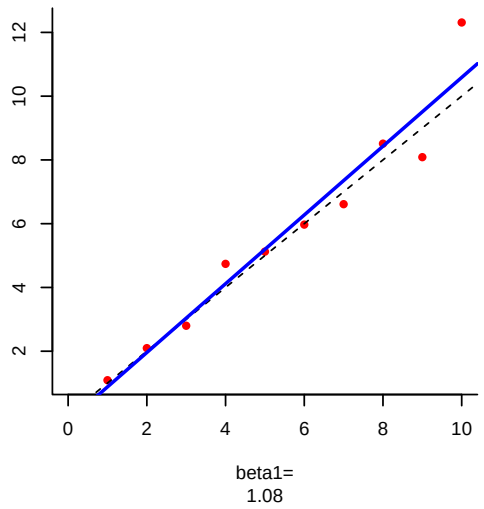


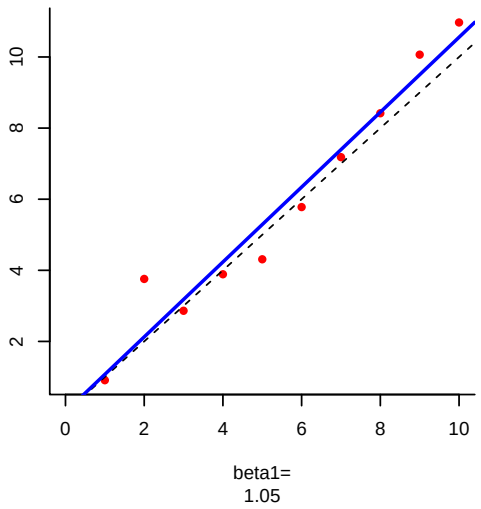
Repetimos seis veces el experimento

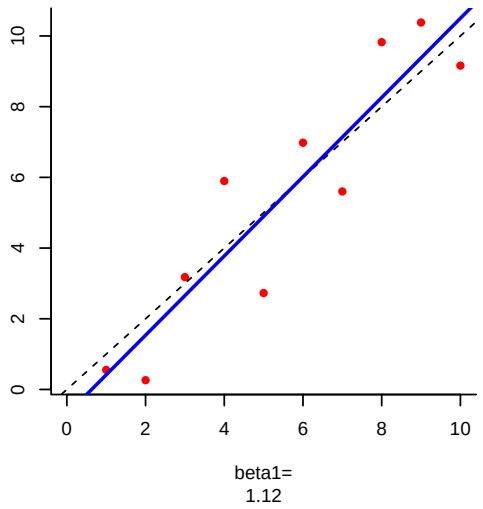




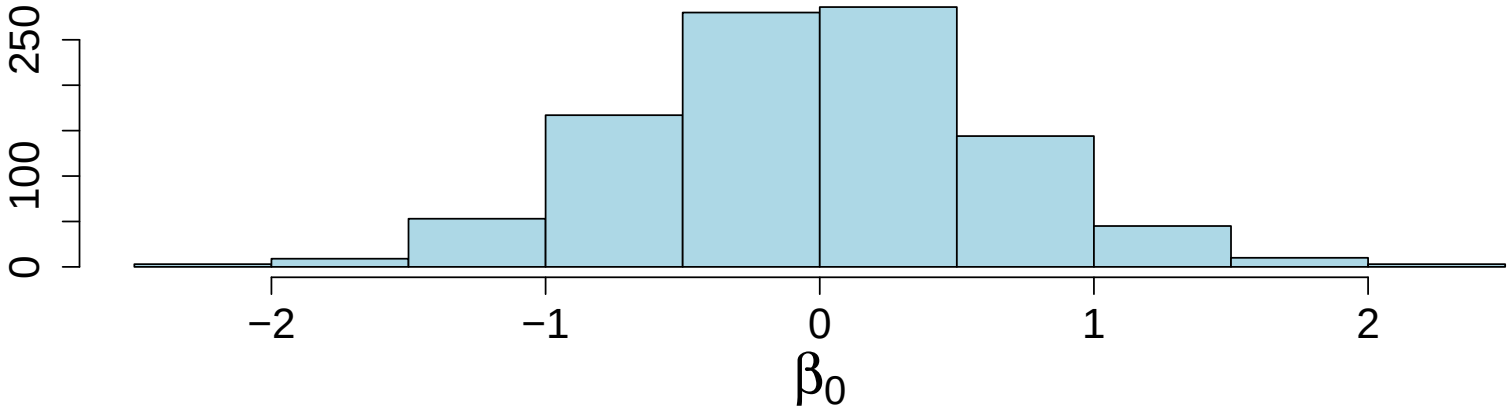


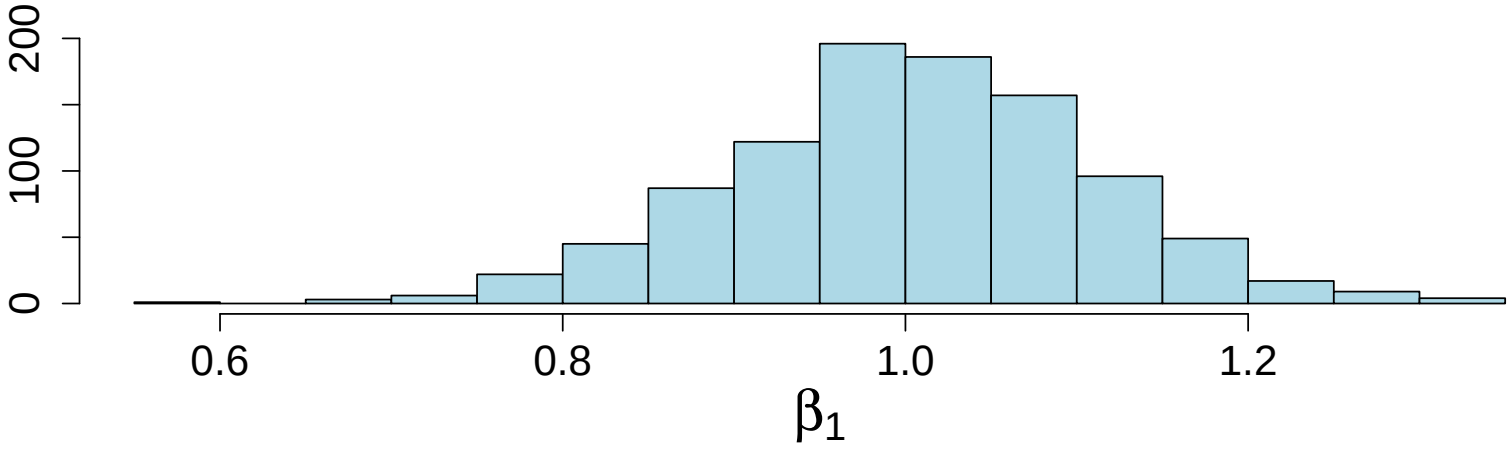






Repetimos mil veces el experimento





Estimación de la varianza

- La varianza de los errores, σ^2, se estima mediante la **varianza residual**:

$$S_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2}{n-2}$$

- Se divide por $n-2$ en lugar de n para que el estimador sea insesgado (es decir, no infraestime sistemáticamente la verdadera varianza).

- De hecho, demostraremos que $(n-2)S_R^2/\sigma^2 \equiv \chi^2_{n-2}$.

Distribución de los estimadores de mínimos cuadrados

Bajo las hipótesis del modelo se verifica:

- $\hat{\beta}_1$ tiene distribución normal de media β_1 y varianza σ^2/S_{xx}.
- $\hat{\beta}_0$ tiene distribución normal de media β_0 y varianza $\sigma^2(1/n + \bar{x}^2/S_{xx})$.
- El vector $(\hat{\beta}_0, \hat{\beta}_1)'$ tiene distribución normal bidimensional y $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x}\sigma^2/S_{xx}$.

Renta y fracaso escolar: ajuste del modelo con **R**

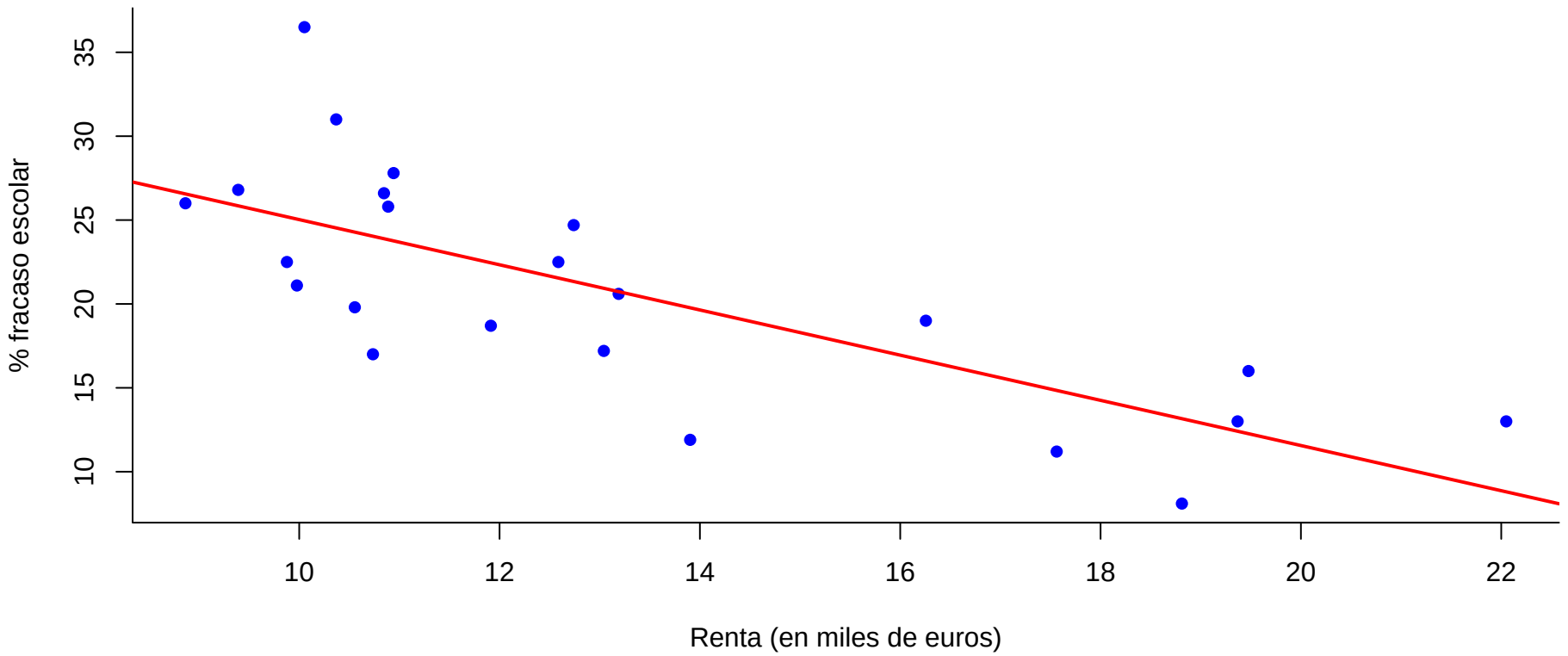
```
regresion <- lm(Fracaso ~ Renta, data = fracasoCAM)
summary(regresion)
```

Renta y fracaso escolar: ajuste del modelo con **R**

```
##
## Call:
## lm(formula = Fracaso ~ Renta, data = fracasoCAM)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8717 -3.7421  0.5878  3.0368 11.5423
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.4944     3.6445  10.562 7.37e-10 ***
## Renta        -1.3467     0.2659  -5.065 5.14e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.757 on 21 degrees of freedom
## Multiple R-squared:  0.5499, Adjusted R-squared:  0.5285
## F-statistic: 25.66 on 1 and 21 DF,  p-value: 5.138e-05
```

Representación gráfica

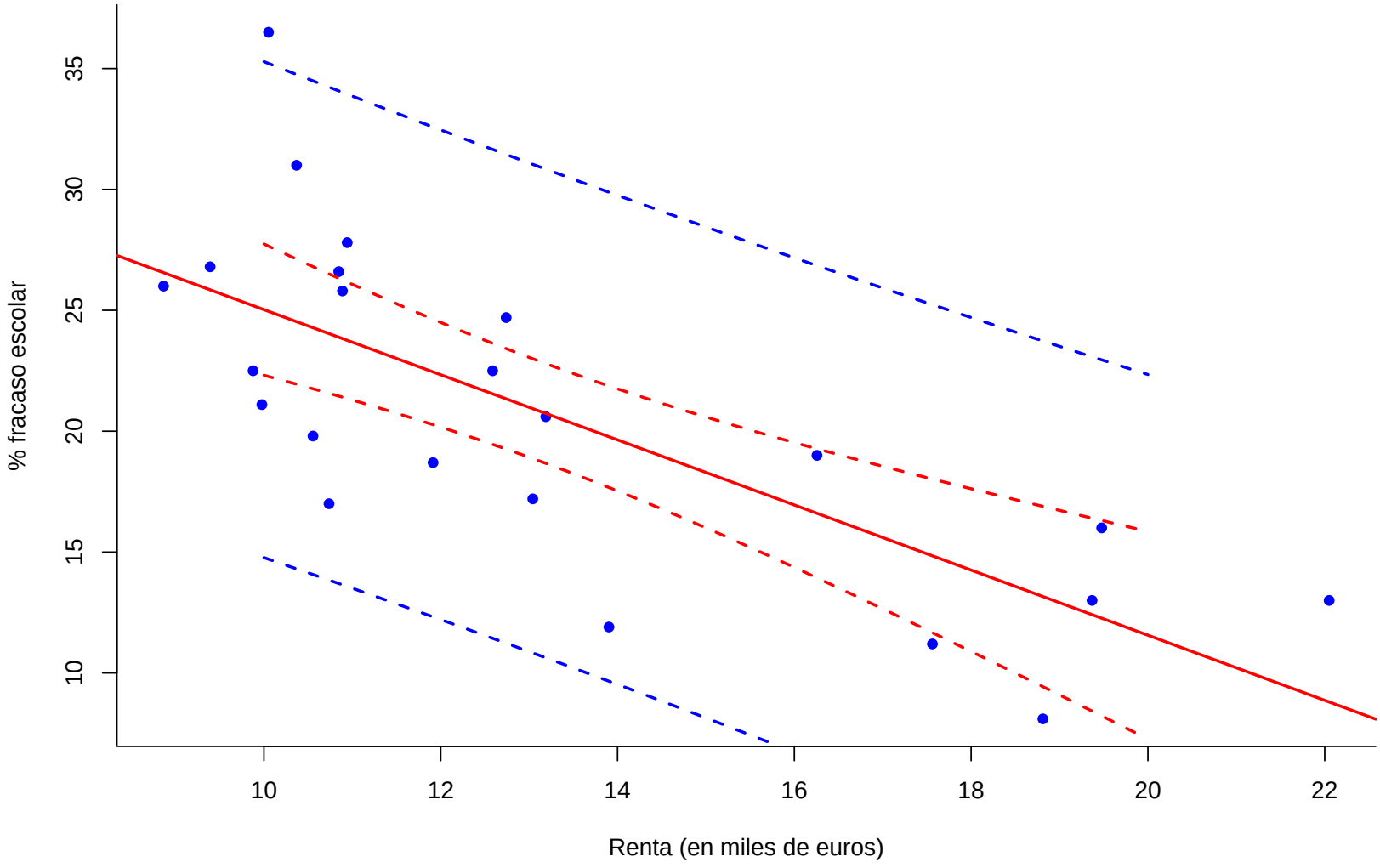
```
plot(fracasoCAM$Renta, fracasoCAM$Fracaso,
     pch=16,col='blue', bty='l',
     ylab='% fracaso escolar',xlab='Renta (en miles de euros)')
abline(regresion, col='red', lwd=2)
```



Intervalos de confianza y de predicción en x_0

```
valores.x0 <- seq(10, 20, 0.1)
datos <- data.frame(Renta = valores.x0)
confianza <- predict(regresion, datos, interval=c('confidence'))
prediccion <- predict(regresion, datos, interval=c('prediction'))

# Representación gráfica
plot(fracasoCAM$Renta, fracasoCAM$Fracaso,
     pch=16, col='blue', bty='l',
     ylab='% fracaso escolar',xlab='Renta (en miles de euros)')
abline(regresion, col='red', lwd=2)
lines(valores.x0, confianza[,2], lty=2, col='red', lwd=2)
lines(valores.x0, confianza[,3], lty=2, col='red', lwd=2)

lines(valores.x0, prediccion[,2], lty=2, col='blue', lwd=2)
lines(valores.x0, prediccion[,3], lty=2, col='blue', lwd=2)
```

Intervalos de confianza y de predicción en x_0



El modelo de regresión lineal múltiple

Tenemos una muestra de n observaciones de las variables Y y $X_1, \ldots, X_k$. Para la observación i, tenemos el vector $(Y_i, x_{i1}, x_{i2}, \ldots, x_{ik})$.

Modelo de regresión lineal múltiple

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \ldots, n,$$

donde las variables de error ϵ_i verifican

Hipótesis

- ϵ_i tiene media cero, para todo i.
- $\text{Var}(\epsilon_i) = \sigma^2$, para todo i (homocedasticidad).
- Son variables independientes.
- Tienen distribución normal.
- $n \geq k + 2$ (hay más observaciones que parámetros).
- Las variables X_i son linealmente independientes entre sí (no hay *colinealidad*).

El modelo de regresión lineal múltiple

Las hipótesis se pueden reexpresar así: las observaciones de la muestra son independientes entre sí con

$$Y_i \equiv \mathrm{N}(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}, \sigma^2),\ \text{ß} = 1, \ldots, n.$$

Forma matricial

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & & & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

De forma más compacta, el modelo equivale a

$$Y = X\beta + \epsilon, \quad \epsilon \equiv \mathrm{N}_n(0, \sigma^2 \mathbb{I}_n) \Leftrightarrow Y \equiv \mathrm{N}_n(X\beta, \sigma^2 \mathbb{I}_n)$$

Una interpretación geométrica

Sea $\mathcal{V} \subset \mathbb{R}^n$ el subespacio vectorial generado por las columnas de la matriz de diseño X ($\dim(V) = k + 1$).

$$\mu \in \mathcal{V} \Leftrightarrow \text{Existe } \beta \in \mathbb{R}^{k+1} \text{ tal que } \mu = X\beta$$

El modelo equivale a suponer $Y \equiv \mathrm{N}_n(\mu, \sigma^2 \mathbb{I}_n)$, donde $\mu \in \mathcal{V}$.

¿Cómo estimarías β a partir de Y y X?

Estimación de los parámetros del modelo

Criterio de mínimos cuadrados

Los estimadores son los valores $\hat{\beta}_0, \ldots, \hat{\beta}_k$ para los que se minimiza

$$\|Y - X\beta\|^2 = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})]^2.$$

Observación: $\hat{Y} \equiv X\hat{\beta}$ es la **proyección ortogonal** de Y sobre $\mathcal{V}$.

Ecuaciones normales

Si $e = Y - \hat{Y}$ es el vector de residuos,

$$X'(Y - \hat{Y}) = 0 \Leftrightarrow X'e = 0$$

Estimadores de mínimos cuadrados

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

Propiedades

- $\hat{\beta}$ es el estimador de máxima verosimilitud (EMV) de β ya que la función de verosimilitud es:

$$L(\beta, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\|Y - X\beta\|^2\right\}.$$

- El EMV de σ^2 es

$$\hat{\sigma}^2 = \frac{\|y - X\hat{\beta}\|^2}{n} = \frac{\|e\|^2}{n} = \frac{\sum_{i=1}^n e_i^2}{n}$$

- El vector $\hat{\beta}$ tiene distribución normal $(k+1)$-dimensional con vector de medias β y matriz de covarianzas $\sigma^2(X'X)^{-1}$.

Estimación de la varianza

Varianza residual

Es la suma de los residuos al cuadrado, corregida por los gl apropiados

$$S_R^2 = \frac{\|Y - \hat{Y}\|^2}{n-k-1} = \frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2.$$

Propiedades

- $\|Y - \hat{Y}\|^2 = Y'(I-H)Y$, donde $H = X(X'X)^{-1}X'$ (la llamada *hat matrix*) es simétrica e idempotente (es la matriz de proyección ortogonal sobre $\mathcal{V}$).
- $(n-k-1)S_R^2/\sigma^2 \equiv \chi^2_{n-k-1}$.
- S_R^2 y $\hat{\beta}$ son independientes.

Inferencia sobre los parámetros del modelo

Distribución

Todos los estimadores $\hat{\beta}_j$ verifican:

$$\frac{\hat{\beta}_j - \beta_j}{\text{error típico de } \hat{\beta}_j} \equiv t_{n-k-1}.$$

Intervalos de confianza

Para cualquier $j = 0, 1, \ldots, k$,

$$\text{IC}_{1-\alpha}(\beta_j) = \left(\hat{\beta}_j \mp t_{n-k-1;\alpha/2} \times \text{error típico de } \hat{\beta}_j\right).$$

Inferencia sobre los parámetros del modelo

Contraste de hipótesis

Queremos determinar qué variables X_j son significativas para explicar Y.

$$H_0 : \beta_j = 0 \qquad (X_j \text{ no influye sobre } Y)$$
$$H_1 : \beta_j \neq 0 \qquad (X_j \text{ influye sobre } Y)$$

La región crítica de cada H_0 al nivel de significación α es

$$R = \left\{ \frac{|\hat{\beta}_j|}{\text{error típico de } \hat{\beta}_j} > t_{n-k-1;\alpha/2} \right\}.$$

El cociente $\hat{\beta}_j/(\text{error típico de } \hat{\beta}_j)$ se llama estadístico t asociado a β_j.

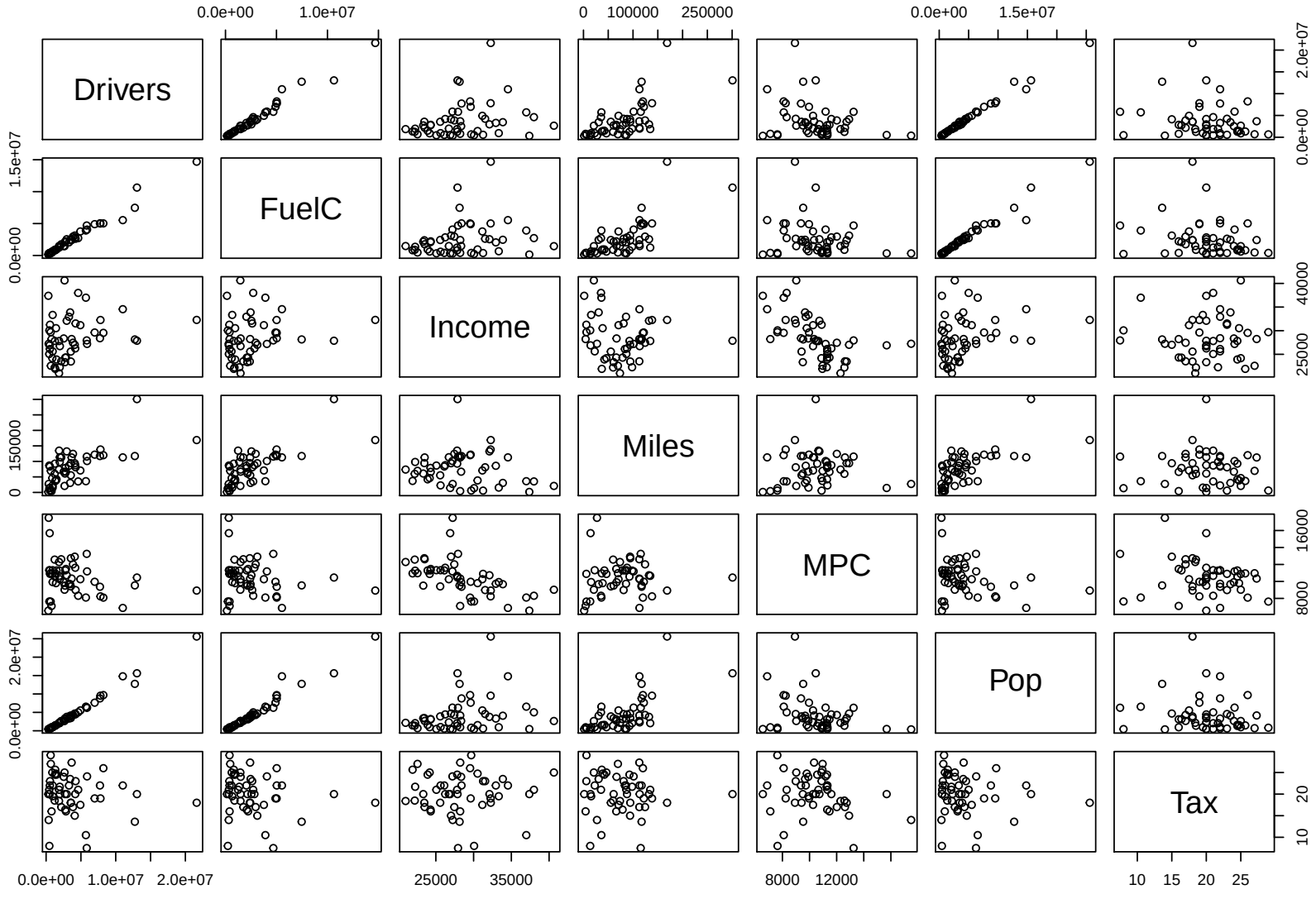
Ejemplo: consumo de combustible en EE.UU.

Los datos `fuel2001` (en el fichero `combustible.RData`) corresponden al consumo de combustible (y otras variables relacionadas) en EE.UU.

```
load('combustible.Rdata')
head(fuel2001)
```

```
##      Drivers    FuelC Income  Miles      MPC      Pop  Tax
## AL   3559897  2382507  23471  94440 12737.00  3451586 18.0
## AK    472211   235400  30064  13628  7639.16   457728  8.0
## AZ   3550367  2428430  25578  55245  9411.55  3907526 18.0
## AR   1961883  1358174  22257  98132 11268.40  2072622 21.7
## CA  21623793 14691753  32275 168771  8923.89 25599275 18.0
## CO   3287922  2048664  32949  85854  9722.73  3322455 22.0
```

Ejemplo: matriz de diagramas de dispersión



Ejemplo: ajuste del modelo

(No se incluye `Pop` por ser prácticamente proporcional a `Drivers`)

```
reg <- lm(FuelC ~ Drivers+Income+Miles+MPC+Tax,
          data=fuel2001)
summary(reg)
vcov(reg)[1:4,1:4]
```

El comando `vcov` produce la matriz de covarianzas (estimadas) de los estimadores de los coeficientes $\hat{\beta}_j$, es decir, $S_R^2(X'X)^{-1}$.

La desviación típica residual fue $S_R = 394100$ con 45 gl.

Ejemplo: ajuste del modelo

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-484395.3400	810159.0498	-0.60	0.5529
Drivers	0.6144	0.0223	27.56	0.0000
Income	7.5260	16.1075	0.47	0.6426
Miles	5.8135	1.5867	3.66	0.0007
MPC	46.4323	34.8794	1.33	0.1898
Tax	-21143.0574	12979.2804	-1.63	0.1103

	(Intercept)	Drivers	Income	Miles
(Intercept)	656357685963.74	-120.86	-10533450.28	-188275.57
Drivers	-120.86	0.00	-0.11	-0.03
Income	-10533450.28	-0.11	259.45	7.98
Miles	-188275.57	-0.03	7.98	2.52

Ejercicios

- Lleva a cabo los contrastes de la forma $H_0 : \beta_j = 0$ para todos los coeficientes del modelo ($\alpha = 0.05$).
- ¿Cuál es la matriz de covarianzas estimada del vector $(\hat{\beta}_2, \hat{\beta}_3)'$?
- Contrasta $H_0 : \beta_2 = \beta_3$
- Calcula una elipse de confianza al 95% para el vector $(\hat{\beta}_2, \hat{\beta}_3)'$.

Ejercicios

Contrastes individuales

Basta comprobar si los p-valores son mayores o menores que α. A nivel 0.05 son significativas: `Drivers` y `Miles`

Matriz de covarianzas estimada de $(\hat{\beta}_2, \hat{\beta}_3)'$

$$\begin{pmatrix} 259.45 & 7.98 \\ 7.98 & 2.52 \end{pmatrix}$$

Por lo tanto, el error típico de $\hat{\beta}_2 - \hat{\beta}_3$ es:

$$\text{ET}(\hat{\beta}_2 - \hat{\beta}_3) = \sqrt{259.45 + 2.52 - 2 \times 7.98} \approx 15.6847$$

Contraste de $H_0 : \beta_2 = \beta_3$

$$t = \frac{|\hat{\beta}_2 - \hat{\beta}_3|}{\text{ET}(\hat{\beta}_2 - \hat{\beta}_3)} \approx \frac{1.7125}{15.6847}$$

que es menor que $t_{45;0.025} = 2.014$. Por lo tanto se acepta H_0.

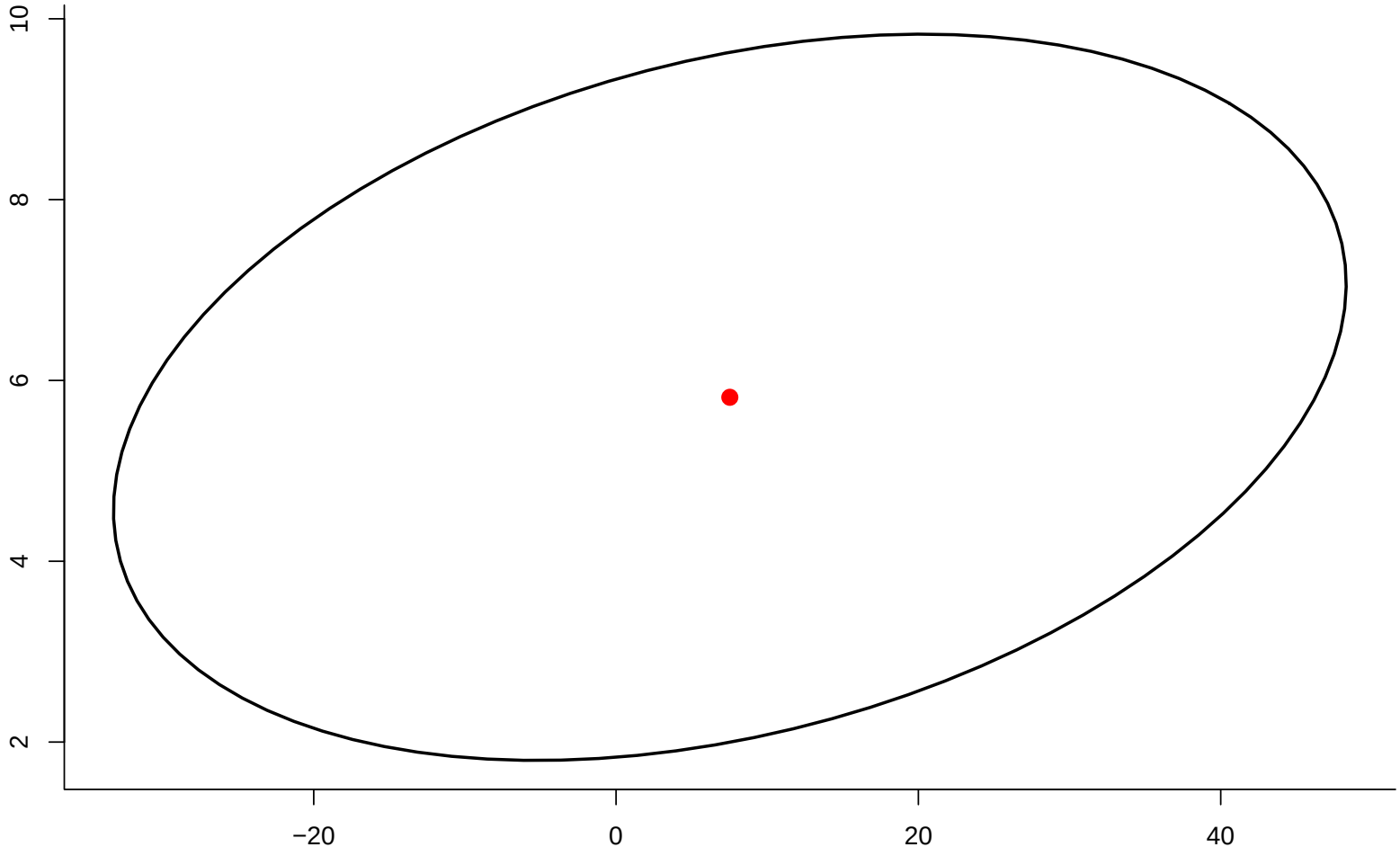
Ejercicios

Elipse de confianza al 95 % para el vector $(\hat{\beta}_2, \hat{\beta}_3)'$

$$(7.52 - \beta_2, 5.81 - \beta_3) \begin{pmatrix} 259.45 & 7.98 \\ 7.98 & 2.52 \end{pmatrix}^{-1} (7.52 - \beta_2, 5.81 - \beta_3)' \leq 2F_{2,45;0.05}.$$

$$(7.52 - \beta_2, 5.81 - \beta_3) \begin{pmatrix} 0.004 & -0.014 \\ -0.014 & 0.44 \end{pmatrix} (7.52 - \beta_2, 5.81 - \beta_3)' \leq 6.408.$$

Ejercicios



Predicción

```
nuevo.dato <- data.frame(2718209.0, 27871.0, 78914.0, 10458.4, 20.0)
names(nuevo.dato) <- names(fuel2001)[-c(2,6)]
nuevo.dato
```

```
##   Drivers Income Miles     MPC Tax
## 1 2718209  27871 78914 10458.4  20
```

```
predict(reg, nuevo.dato, interval='confidence')
```

```
##       fit     lwr     upr
## 1 1916963 1795684 2038242
```

```
predict(reg, nuevo.dato, interval='prediction')
```

```
##       fit     lwr     upr
## 1 1916963 1114048 2719878
```

El análisis de la varianza

Suma de cuadrados total

$\text{SCT} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = Y'(I - M)Y$ mide la variabilidad total en la respuesta.

Suma de cuadrados de la regresión

$\text{SCR} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = Y'(H - M)Y$ mide la parte de la variabilidad **explicada** por el modelo.

Suma de cuadrados de los errores o residual

$\text{SCE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = Y'(I - H)Y$ mide la parte de la variabilidad **no explicada** por el modelo.

Descomposición de la variabilidad

$$\begin{aligned} Y'(I - M)Y &= Y'(H - M)Y + Y'(I - H)Y \\ \|Y - MY\|^2 &= \|HY - MY\|^2 + \|Y - \hat{Y}\|^2 \\ \text{SCT} &= \text{SCR} + \text{SCE} \end{aligned}$$

La tabla de análisis de la varianza

```
anova(reg)
```

```
## Analysis of Variance Table
##
## Response: FuelC
##            Df     Sum Sq    Mean Sq   F value    Pr(>F)
## Drivers     1 3.5301e+14 3.5301e+14 2273.2167 < 2.2e-16 ***
## Income      1 6.7563e+11 6.7563e+11    4.3507 0.0426945 *
## Miles       1 2.1698e+12 2.1698e+12   13.9723 0.0005216 ***
## MPC         1 5.2927e+11 5.2927e+11    3.4082 0.0714577 .
## Tax         1 4.1208e+11 4.1208e+11    2.6536 0.1102978
## Residuals 45 6.9882e+12 1.5529e+11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El contraste de la regresión

Queremos contrastar

$H_0 : \beta_1 = \ldots = \beta_k = 0$ (el modelo no es explicativo: **ninguna de las variables explicativas influye en la respuesta**)

$H_1 : \beta_j \neq 0$ para algún $j = 1, \ldots, k$ (el modelo es explicativo: **al menos una de las variables X_j influye en la respuesta**)

Comparamos la variabilidad explicada con la no explicada mediante el estadístico F:

$$F = \frac{\text{SCR}/k}{\text{SCE}/(n-k-1)}.$$

Bajo H_0, el estadístico F sigue una distribución $F_{k,n-k-1}$.

La región de rechazo de H_0 al nivel de significaci'on α es

$$R = \{F > F_{k,n-k-1;\alpha}\}$$

El coeficiente de determinación

Es una medida de la bondad del ajuste en el modelo de regresón:

$$R^2 = \frac{\text{SCR}}{\text{SCT}}.$$

Propiedades

- $0 \leq R^2 \leq 1$.
- Cuando $R^2 = 1$ existe una relación exacta entre los valores ajustados y la variable respuesta.
- Cuando $R^2 = 0$, $\hat{Y}_i = \bar{Y}$ para todo $i = 1, \ldots, n$.
- Podemos interpretar R^2 o como un coeficiente de correlación múltiple entre Y y las k variables regresoras.
- En regresión simple $R^2 = r^2$.
- Se verifica $F = \dfrac{R^2}{1 - R^2} \dfrac{n - k - 1}{k}$.

El coeficiente de determinación ajustado

El coeficiente de determinación para comparar distintos modelos de regresión entre sí tiene el siguiente inconveniente:

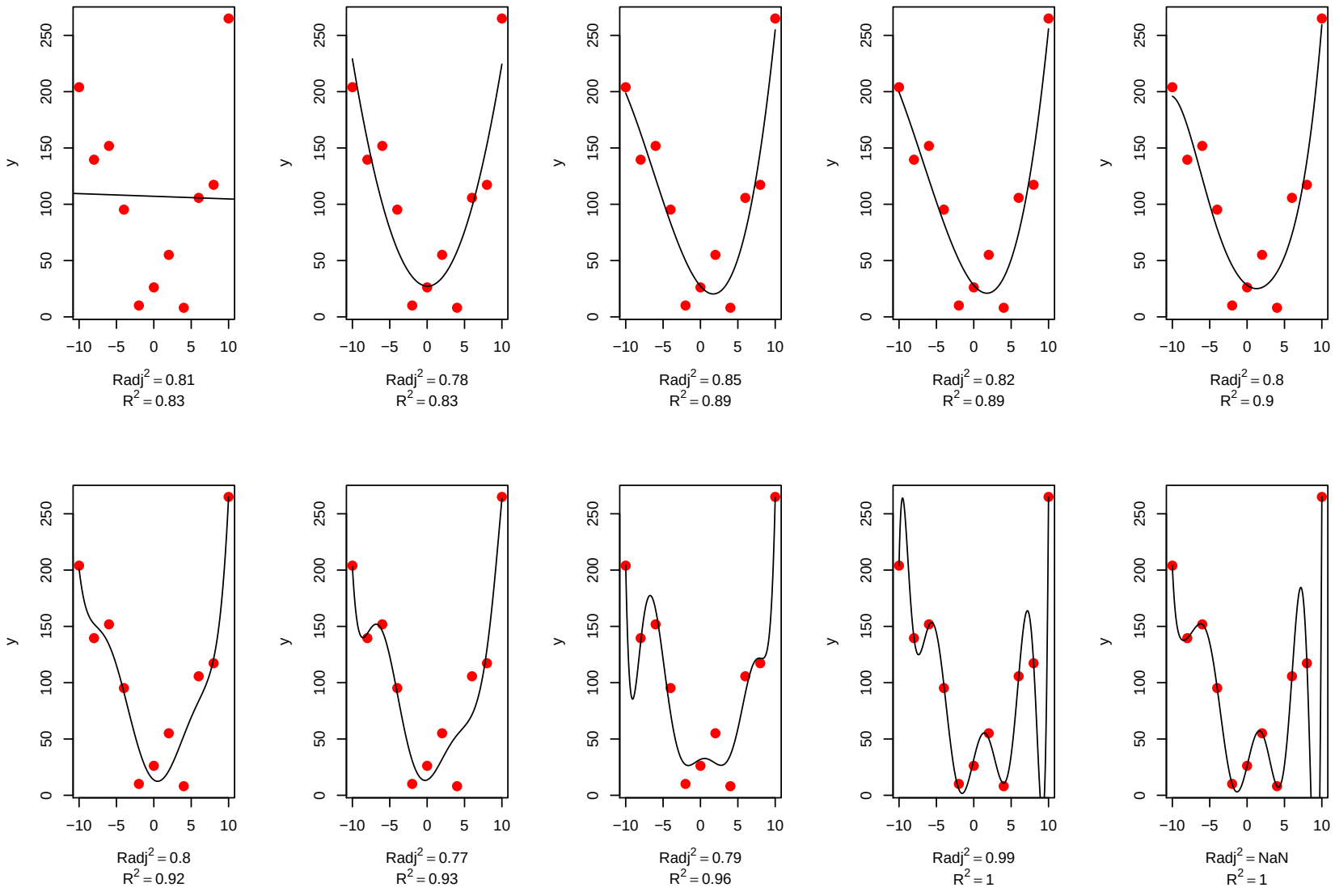
Siempre que se añade una nueva variable regresora al modelo, R^2 aumenta, aunque el efecto de la variable regresora sobre la respuesta no sea significativo.

Por ello se define el **coeficiente de determinación ajustado** o corregido por grados de libertad

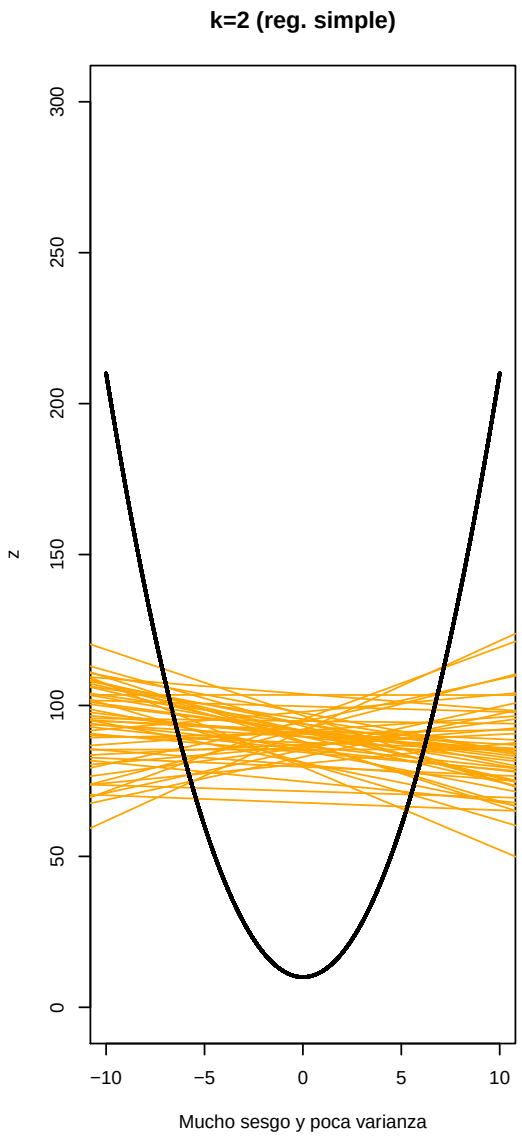
$$\bar{R}^2 = 1 - \frac{\text{SCE}/(n-k-1)}{\text{SCT}/(n-1)} = 1 - \frac{S_R^2}{\text{SCT}/(n-1)} = 1 - \frac{S_R^2}{S_y^2}$$

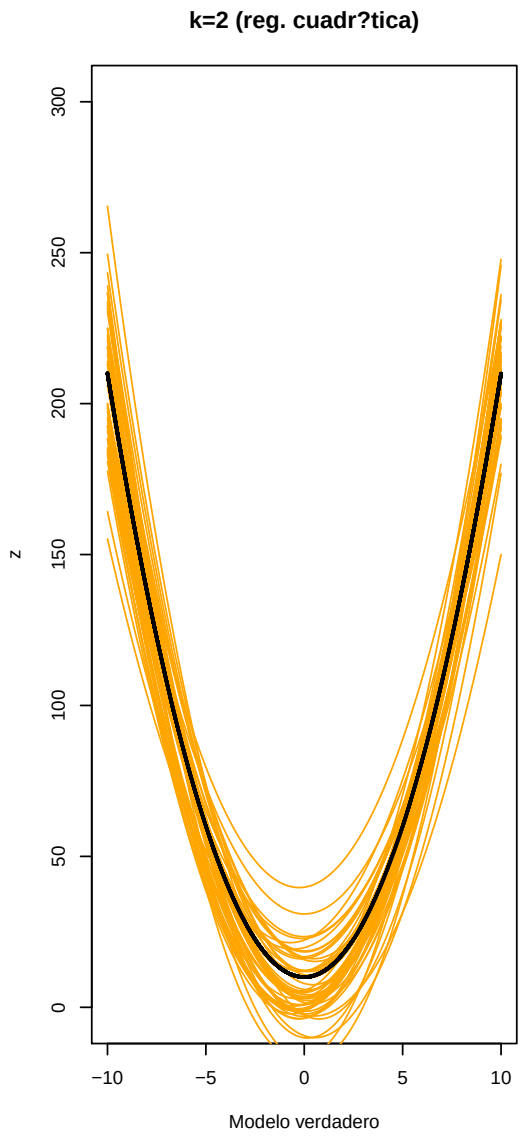
$\bar{R}^2$ solo disminuye al introducir una nueva variable en el modelo si la varianza residual disminuye.

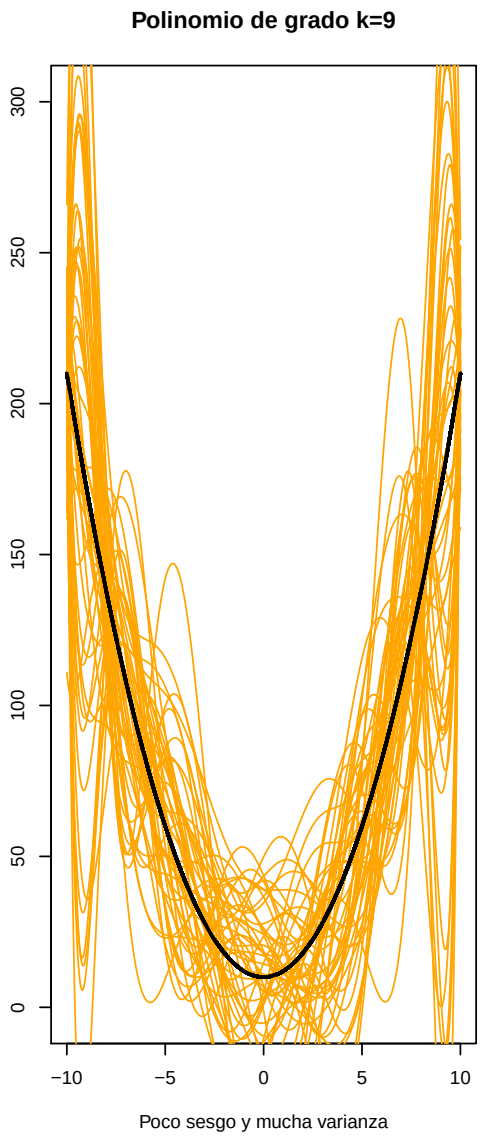
Regresión polinómica y sobreajuste



Curvas estimadas a partir de 50 muestras de 10 puntos







Contraste de hipótesis lineales

Queremos contrastar $H_0 : A\beta = 0$, donde A es una matriz $p \times (k+1)$ con rango$(A) = p < k+1$.

Por ejemplo, en el modelo $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$, podríamos estar interesados en contrastar

$$H_0 : \beta_1 = \beta_2; \beta_0 = 0 \Leftrightarrow A\beta = 0,$$

donde

$$A = \begin{pmatrix} 0 & 1 & -1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

Si H_0 fuese cierta, habría que ajustar el modelo más simple $Y_i = \beta_1 \tilde{x}_{i1} + \beta_3 x_{i3} + \epsilon_i$, con $\tilde{x}_{i1} = x_{i1} + x_{i2}$.

Llamaremos **modelo reducido** (M0) al modelo lineal que resulta de imponer las restricciones de H_0.

Principio de incremento de la variabilidad relativa

La idea básica de este principio es considerar:

- SCE_0, la variabilidad no explicada (residual) bajo el modelo reducido.
- SCE, la variabilidad no explicada (residual) bajo el modelo completo.

Siempre se cumple $\text{SCE}_0 > \text{SCE}$.

- Se rechaza H_0 cuando se pierde mucho al considerar M0 en lugar de M, es decir, cuando el cociente

$$\frac{\text{SCE}_0 - \text{SCE}}{\text{SCE}}$$

sea suficientemente grande.

Contraste de hipótesis lineales

Bajo $H_0 : A\beta = 0$, se verifica

$$\frac{(\mathrm{SCE}_0 - \mathrm{SCE})/p}{\mathrm{SCE}/(n-k-1)} \equiv F_{p,n-k-1}$$

Por lo tanto la región crítica del contraste para un nivel α es

$$R = \left\{ \frac{(\mathrm{SCE}_0 - \mathrm{SCE})/p}{\mathrm{SCE}/(n-k-1)} > F_{p,n-k-1;\alpha} \right\}.$$

Tabla ANOVA en **R**

Supongamos un modelo con, por ejemplo, $k = 3$ variables regresoras: x_1, x_2 y x_3.

	Df	Sum Sq	Mean Sq	F value
x1	1	$\texttt{SC}_1$	$\texttt{SC}_1$	$\texttt{SC}_1/\texttt{MCE}$
x2	1	$\texttt{SC}_{12}$	$\texttt{SC}_{12}$	$\texttt{SC}_{12}/\texttt{MCE}$
x3	1	$\texttt{SC}_{123}$	$\texttt{SC}_{123}$	$\texttt{SC}_{123}/\texttt{MCE}$
Residuals	$n-k-1$	SCE	$\texttt{MCE} = \texttt{SCE}/(n-k-1)$	

- $\texttt{SC}_1 = \texttt{SCE}_0 - \texttt{SCE}$,
  donde (M_0) es $Y = \beta_0 + \epsilon$ y (M) es $Y = \beta_0 + \beta_1 x_1 + \epsilon$.
- $\texttt{SC}_{12} = \texttt{SCE}_0 - \texttt{SCE}$,
  donde (M_0) es $Y = \beta_0 + \beta_1 x_1 + \epsilon$ y (M) es $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.
- $\texttt{SC}_{123} = \texttt{SCE}_0 - \texttt{SCE}$,
  donde (M_0) es $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ y (M) es
  $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$.

Tabla ANOVA en **R**

```
# Modelo completo
reg <- lm(FuelC ~ Drivers+Income+Miles+MPC+Tax,
          data=fuel2001)
# Modelo reducido
reg0 <- lm(FuelC ~ Drivers, data=fuel2001)
anova(reg0)
```

```
## Analysis of Variance Table
##
## Response: FuelC
##            Df     Sum Sq    Mean Sq F value    Pr(>F)
## Drivers     1 3.5301e+14 3.5301e+14  1605.4 < 2.2e-16 ***
## Residuals 49 1.0775e+13 2.1990e+11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparación de dos modelos anidados

```
anova(reg0, reg)
```

```
## Analysis of Variance Table
##
## Model 1: FuelC ~ Drivers
## Model 2: FuelC ~ Drivers + Income + Miles + MPC + Tax
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1     49 1.0775e+13
## 2     45 6.9882e+12  4 3.7868e+12 6.0962 0.0005231 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F = \frac{\frac{\text{SCE}_0 - \text{SCE}}{p}}{\frac{\text{SCE}}{n-k-1}} = \frac{\frac{1.0775 \cdot 10^{13} - 6.9882 \cdot 10^{12}}{4}}{\frac{6.9882 \cdot 10^{12}}{45}} = 6.0962$$

Análisis de influencia: ejemplo

Experimento para estudiar la cantidad de cierto medicamento en el hígado de una rata, tras recibir una dosis oral. La dosis recibida fue de 40 mg por kg de peso corporal. Tras cierto tiempo se sacrifican las ratas y se mide la cantidad de medicamento en su hígado.

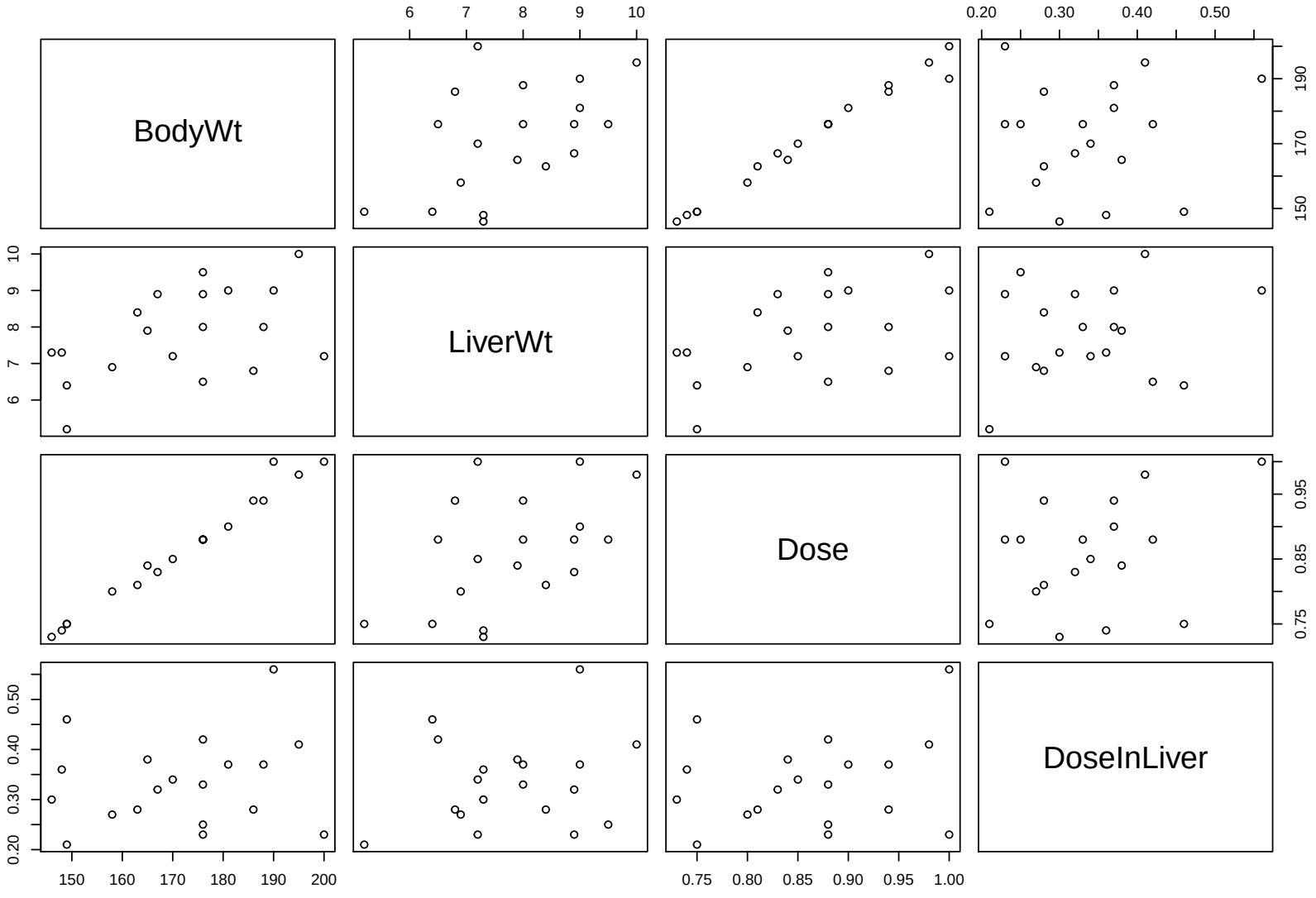
Hay 19 observaciones, tres variables regresoras y una variable respuesta:

- `BodyWt`: Peso de la rata en g
- `LiverWt`: Peso del hígado en g
- `Dose`: Dosis relativa recibida por la rata (fracción de la máxima dosis)
- `DrugInLiver`: Proporción de la dosis en el hígado.

Modelo:

$$\texttt{DoseInLiver} = \beta_0 + \beta_1 \texttt{BodyWt} + \beta_2 \texttt{LiverWt} + \beta_3 \texttt{Dose} + \epsilon$$

Diagramas de dispersión



Regresiones simples

```
##
## Call:
## lm(formula = DoseInLiver ~ BodyWt, data = ratas)
##
## Residuals:
##       Min         1Q     Median         3Q        Max
## -0.128342 -0.060647 -0.008889  0.046916  0.209763
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.1962346  0.2215825   0.886    0.388
## BodyWt      0.0008105  0.0012862   0.630    0.537
##
## Residual standard error: 0.08999 on 17 degrees of freedom
## Multiple R-squared:  0.02283,    Adjusted R-squared:  -0.03465
## F-statistic: 0.3971 on 1 and 17 DF,  p-value: 0.537
```

Regresiones simples

```
##
## Call:
## lm(formula = DoseInLiver ~ LiverWt, data = ratas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12129 -0.05790 -0.00805  0.03739  0.20724
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.22037    0.13573   1.624    0.123
## LiverWt      0.01471    0.01718   0.856    0.404
##
## Residual standard error: 0.08913 on 17 degrees of freedom
## Multiple R-squared:  0.04134,    Adjusted R-squared:  -0.01505
## F-statistic: 0.7331 on 1 and 17 DF,  p-value: 0.4038
```

Regresiones simples

```
##
## Call:
## lm(formula = DoseInLiver ~ Dose, data = ratas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.13761 -0.06211 -0.00427  0.04850  0.19239
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.1330     0.2109   0.631    0.537
## Dose          0.2346     0.2435   0.963    0.349
##
## Residual standard error: 0.08864 on 17 degrees of freedom
## Multiple R-squared:  0.05178,    Adjusted R-squared:  -0.004002
## F-statistic: 0.9283 on 1 and 17 DF,  p-value: 0.3488
```

Regresión múltiple

```
##
## Call:
## lm(formula = DoseInLiver ~ BodyWt + Dose, data = ratas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12333 -0.07416  0.01238  0.04884  0.12668
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.285517   0.191267   1.493   0.1550
## BodyWt      -0.020444   0.007838  -2.608   0.0190 *
## Dose         4.125330   1.506472   2.738   0.0146 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07654 on 16 degrees of freedom
## Multiple R-squared:  0.3347, Adjusted R-squared:  0.2515
## F-statistic: 4.024 on 2 and 16 DF,  p-value: 0.0384
```

Observaciones

- Las tres regresiones simples muestran ausencia de relación entre la respuesta y las variables (también el gráfico).
- La regresión múltiple, sin embargo, indica que conjuntamente `BodyWt` y `Dose` son significativas. Lo mismo ocurre si añadimos `LiverWt` del modelo.
- Pero `BodyWt` y `Dose` miden esencialmente lo mismo.
- ¿Por qué se produce la paradoja?

Potencial de un punto

El potencial (*leverage*) de un punto, h_i, es el correspondiente elemento de la diagonal de la *matriz sombrero* H.

Los potenciales determinan la varianza de los residuos:

$$\text{Var}(e_i) = \sigma^2(1 - h_i).$$

El potencial está estrechamente relacionado con la distancia de Mahalanobis:

$$h_i = \frac{1}{n} + \frac{1}{n-1} d_M^2(x_i, \bar{x}).$$

Aquí, x_i y $\bar{x}$ no incluyen la coordenada (igual a uno) correspondiente a β_0.

La distancia de Cook

La distancia de Cook mide cómo cambia el vector de estimadores $\hat{\beta}$ cuando se elimina cada observación.

Para ello, se utiliza la distancia de Mahalanobis (estandarizada) entre $\hat{\beta}$ y $\hat{\beta}(i)$.

Si recordamos que la matriz de covarianzas de $\hat{\beta}$ se puede estimar con $S_R^2(X'X)^{-1}$, tenemos que la distancia de Cook es:

$$D_i = \frac{[\hat{\beta} - \hat{\beta}(i)]' X'X [\hat{\beta} - \hat{\beta}(i)]}{(k+1)S_R^2}$$

La distancia de Cook

- Para calibrar los valores obtenidos se compara con las tablas de la distribución $F_{k+1,n-k-1}$. En general observaciones tales que $D_i \geq 1$ pueden ser relevantes.

- Se puede escribir en términos de los valores ajustados:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j(i) - \hat{Y}_j)^2}{(k+1)S_R^2}$$

- También está relacionado con el potencial y los residuos:

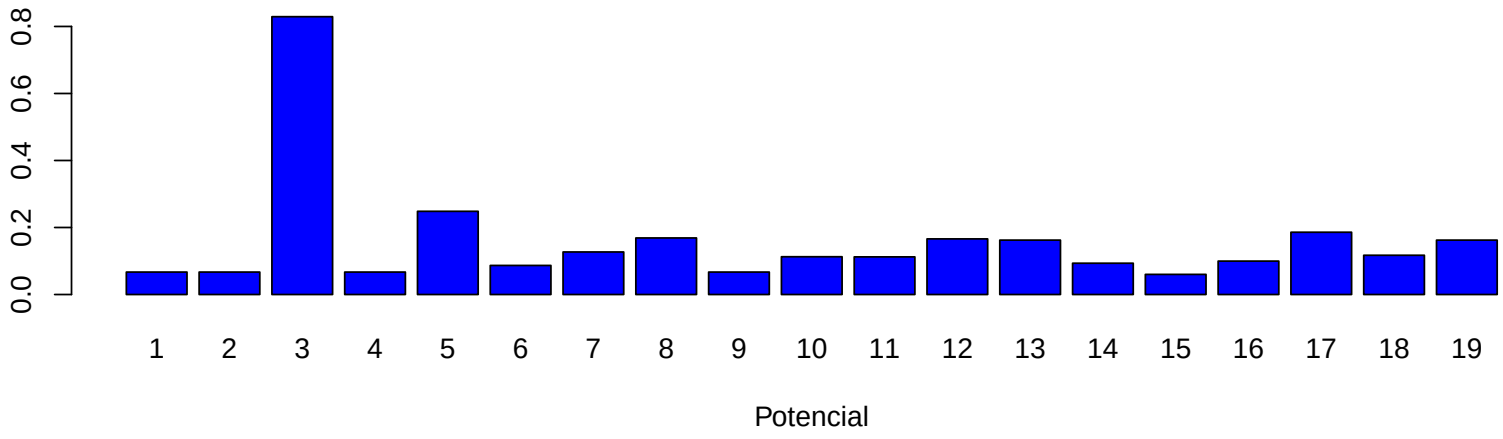
$$D_i = \frac{1}{k+1} r_i^2 \frac{h_i}{1-h_i},$$

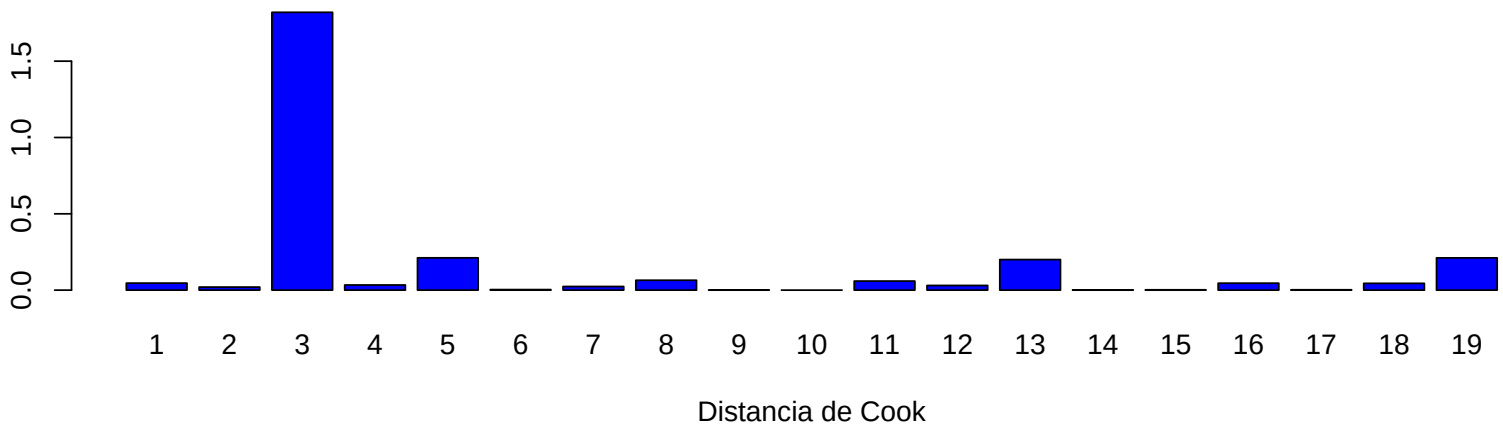
donde $r_i = e_i/(S_R\sqrt{1-h_i})$ son los residuos estandarizados.

Análisis de influencia en el ejemplo

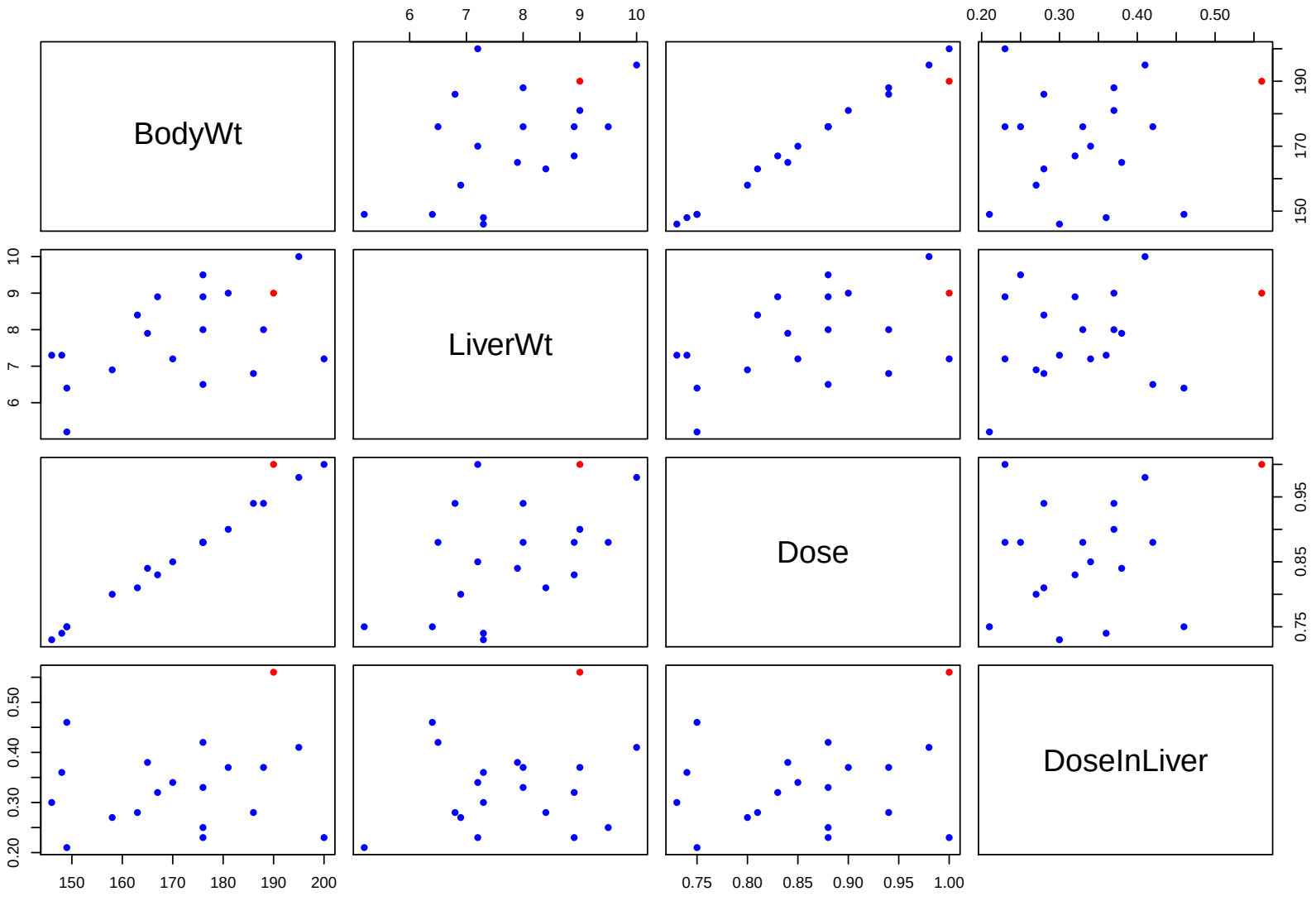
```
potencial <- hatvalues(reg.multiple)
cook <- cooks.distance(reg.multiple)

layout(1:2)
barplot(potencial, col='blue', xlab='Potencial')
barplot(cook, col='blue', xlab='Distancia de Cook')
```

Análisis de influencia en el ejemplo





Análisis de influencia en el ejemplo



Resultados sin la tercera observación

```
##
## Call:
## lm(formula = DoseInLiver[-3] ~ BodyWt[-3] + Dose[-3], data = rata
##
## Residuals:
##       Min         1Q     Median         3Q        Max
## -0.116828 -0.057680  0.006176  0.048894  0.133172
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.33199    0.19540   1.699    0.110
## BodyWt[-3]  -0.00444    0.01693  -0.262    0.797
## Dose[-3]     0.87516    3.40021   0.257    0.800
##
## Residual standard error: 0.07622 on 15 degrees of freedom
## Multiple R-squared:  0.004829,   Adjusted R-squared:  -0.1279
## F-statistic: 0.03639 on 2 and 15 DF,  p-value: 0.9643
```

Variable regresora cualitativa: modelo unifactorial

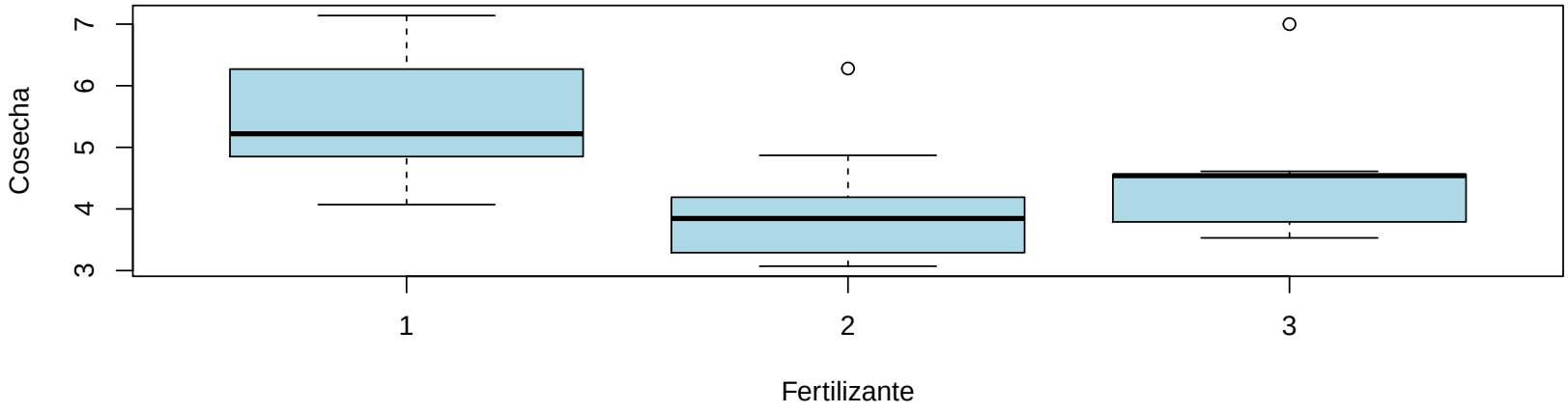
En un estudio para comparar la eficacia de tres fertilizantes se utiliza cada uno de ellos en 10 parcelas (asignando aleatoriamente cada parcela a uno de los tres fertilizantes) y posteriormente se registra el peso en toneladas de la cosecha resultante en cada parcela. Los datos son:

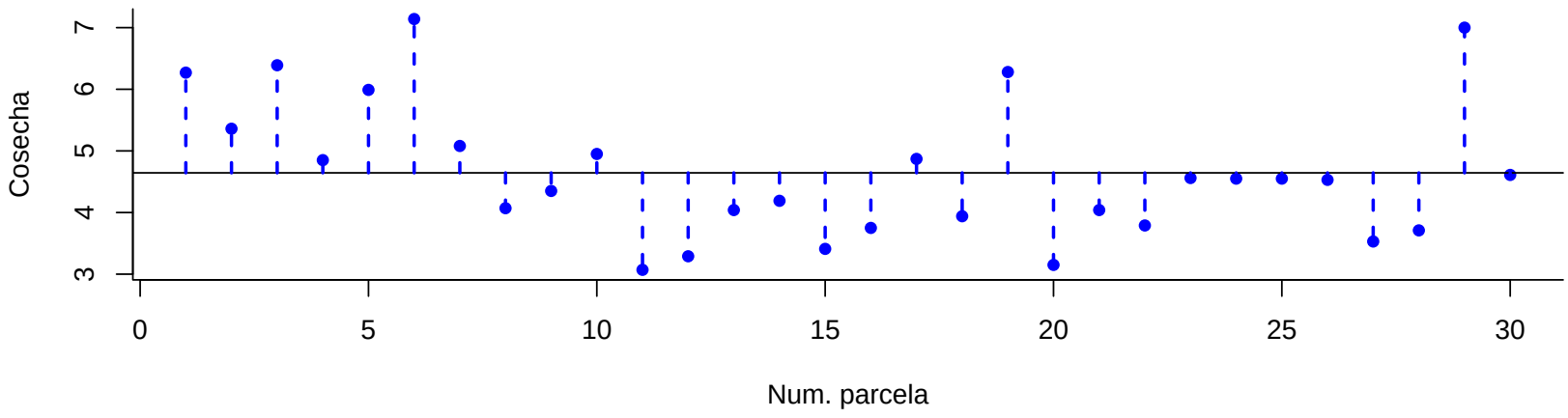
Fert. 1	6.27	5.36	6.39	4.85	5.99	7.14	5.08	4.07	4.35	4.95
Fert. 2	3.07	3.29	4.04	4.19	3.41	3.75	4.87	3.94	6.28	3.15
Fert. 3	4.04	3.79	4.56	4.55	4.55	4.53	3.53	3.71	7.00	4.61

Una variable explicativa cualitativa se llama **factor**. Los valores que toma se llaman **niveles**. En este modelo los niveles son los distintos **tratamientos** que aplicamos a las **unidades experimentales**.

En el ejemplo tenemos un factor (el tipo de fertilizante) que se presenta en tres niveles o tratamientos, que se aplican a las unidades experimentales (las parcelas).

Descripción de los datos





Notación

Disponemos de respuestas correspondientes a k niveles del factor, n_i es el tamaño muestral del grupo i y $n = n_1 + \cdots + n_k$ es el número total de respuestas.

Muestra	Respuestas				Medias	Desv. típicas
1	Y_{11}	Y_{12}	$\cdots$	Y_{1n_1}	$\bar{Y}_{1.}$	S_1
2	Y_{21}	Y_{22}	$\cdots$	Y_{2n_2}	$\bar{Y}_{2.}$	S_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
k	Y_{k1}	Y_{k2}	$\cdots$	Y_{kn_k}	$\bar{Y}_{k.}$	S_k

En el ejemplo: $k = 3$, $n_i = 10$, $n = 30$.

Muestra	n_i	$\bar{Y}_{i.}$	S_i
1	10	5.445	0.976
2	10	3.999	0.972
3	10	4.487	0.975

Formulación del modelo unifactorial

Si Y_{ij} representa la respuesta j para el nivel i,

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad i = 1, \ldots, k, \quad j = 1, \ldots, n_i.$$

- μ_i es el nivel medio de la respuesta para el nivel i del factor.
- ϵ_{ij} es la variable de error que recoge el resto de variables que influyen en la respuesta. Estas variables son independientes y tienen distribución normal con media 0 y desviación típica σ.
- **Homocedasticidad**: La desviación típica es la misma para todos los niveles del factor.

Otra forma equivalente de escribir lo mismo:

Para $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, las variables Y_{ij} son independientes y, además,

$$Y_{ij} \equiv \mathrm{N}(\mu_i;\, \sigma^2)$$

Formulación del modelo unifactorial

El modelo unifactorial se puede expresar en la forma $Y = X\beta + \epsilon$, donde

- $Y = (Y_{1,1}, Y_{1,2}, \ldots, Y_{k,n_k})'$
- $\beta = (\mu_1, \ldots, \mu_k)'$
- $\epsilon = (\epsilon_{1,1}, \epsilon_{1,2}, \ldots, \epsilon_{k,n_k})'$

¿Cuál es la matriz de diseño X?

¿Cuánto vale $X'X$? ¿Cuánto vale $\hat{\beta}$?

Contraste de igualdad de medias

Objetivo: Contrastar $H_0 : \mu_1 = \cdots = \mu_k$

Modelo reducido:

$$\text{SCE}_0 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 = \text{SCT}.$$

Modelo completo:

$$\text{SCE} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 = \sum_{i=1}^{k} (n_i - 1) S_i^2.$$

Incremento de variabilidad:

$$\text{SCE}_0 - \text{SCE} = \text{SCT} - \text{SCE} = \text{SCR} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2 = \sum_{i=1}^{k} n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2.$$

Contraste de igualdad de medias

En esta situación, ¿cuáles son los gl de SCR y SCE?

Para contrastar $H_0 : \mu_1 = \cdots = \mu_k$ se compara $\text{SCR}/(k-1)$ con $\text{SCE}/(n-k)$ mediante el cociente:

$$F = \frac{\text{SCR}/(k-1)}{\text{SCE}/(n-k)}.$$

Se rechaza H_0 en la región crítica:

$$R = \{F > F_{k-1,n-k;\alpha}\}$$

Tabla ANOVA del modelo unifactorial

Fuente de variación	Suma de cuadrados	gl	cuadrados medios	estadístico
Explicada	$\sum_{i=1}^{k} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$	$k-1$	$\frac{\sum_{i=1}^{k} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2}{k-1}$	F
Residual	$\sum_{i=1}^{k}(n_i - 1)S_i^2$	$n-k$	$\frac{\sum_{i=1}^{k}(n_i-1)S_i^2}{n-k}$	

Con los datos del ejemplo:

Fuente de variación	Suma de cuadrados	gl	cuadrados medios	estadístico
Explicada				
Residual				

Solución con **R**

```r
# La variable fertilizante debe ser un factor
resultado = aov(cosecha ~ fertilizante)
summary(resultado)
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## fertilizante  2  10.82   5.411   5.702 0.00859 **
## Residuals    27  25.62   0.949
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```