ESTADÍSTICA II
Tema 3: Regresión lineal

- Regresión lineal simple: Planteamiento del problema. Recta de regresión. Estimación. Modelo.
- Regresión lineal múltiple: Modelo. Estimadores de mínimos cuadrados. Inferencia sobre los parámetros del modelo. Análisis de la varianza. Coeficientes de determinación. Contrastes de hipótesis lineales. Análisis de influencia.
- Variables regresoras cualitativas: modelo unifactorial

El problema de regresión simple

Observamos dos variables, X e Y, en una muestra de n individuos: $(x_1, y_1), \ldots, (x_n, y_n)$. 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* o *dependiente*.
- 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.

Ejemplo (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

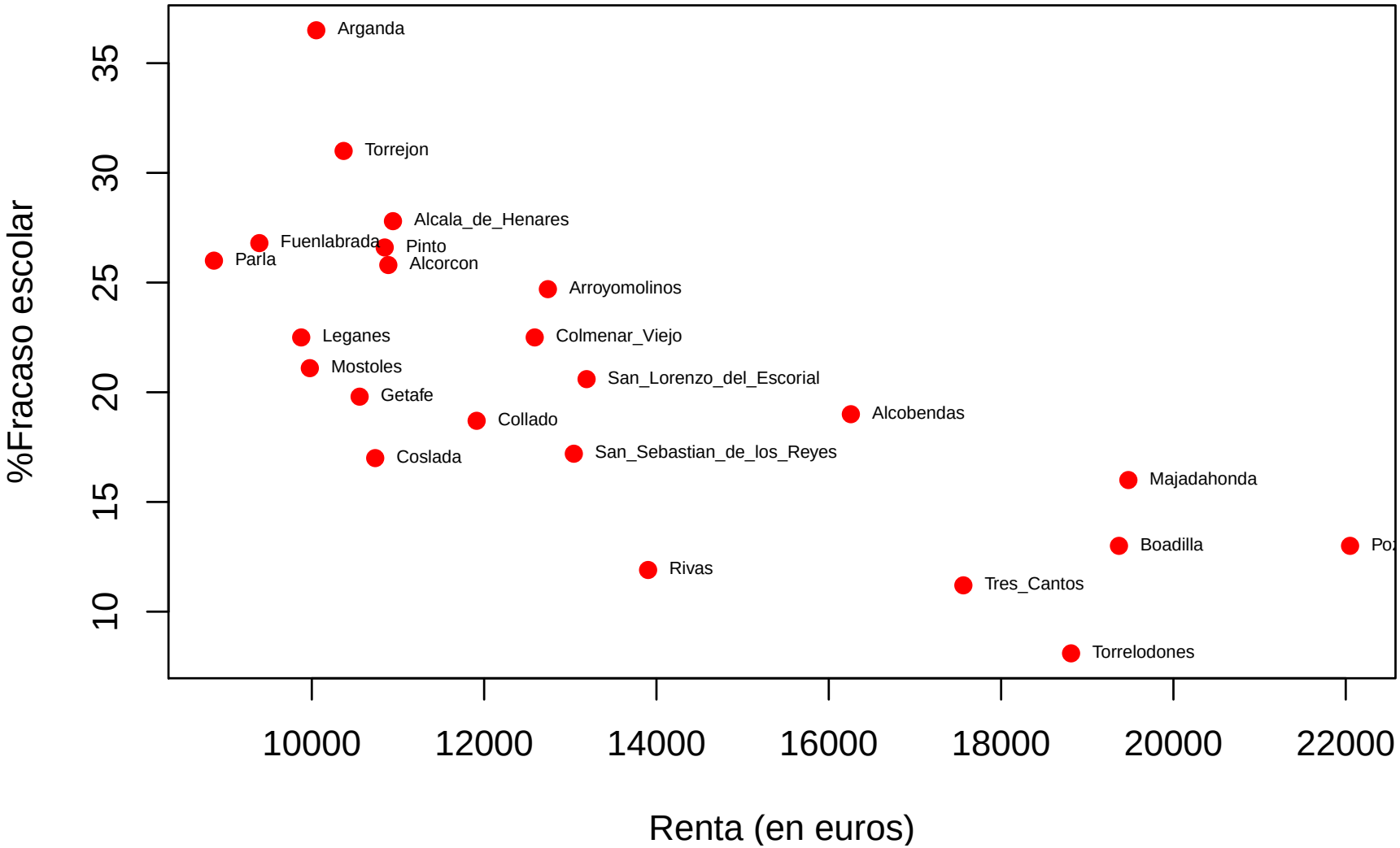
Fracaso escolar en la Comunidad de Madrid

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

Ejemplo (fracaso escolar): Diagrama de dispersión



Recta de regresión (regresión lineal simple)

Matemáticamente, la *relación* más simple entre las dos variables es la de tipo *lineal*:

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

La regresión lineal aparece frecuentemente en contextos como la *calibración* (análisis instrumental, quimiometría):

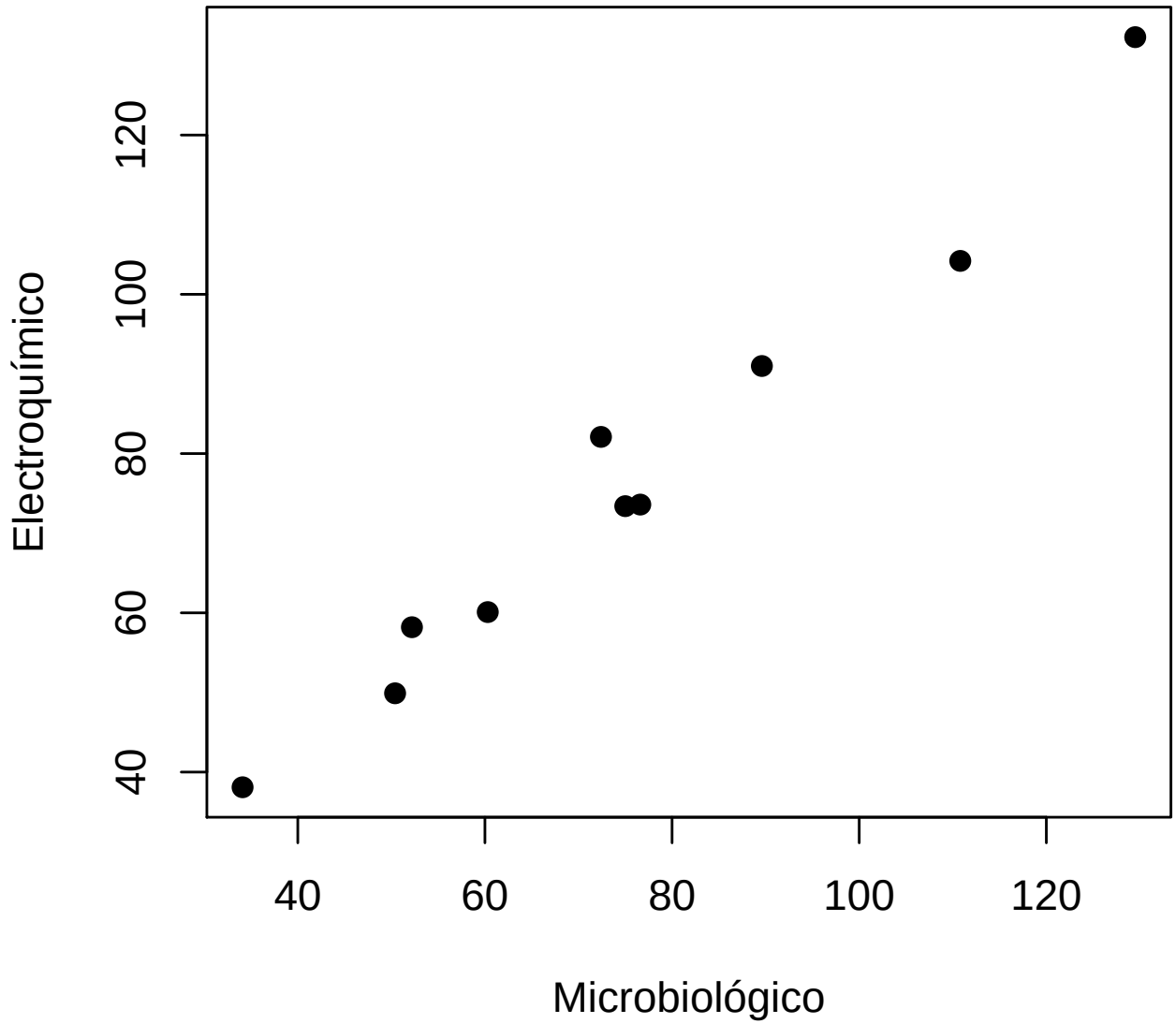
Se toman una serie de materiales de los que se conoce la concentración (X) de un cierto analito. Estos patrones de calibración se miden (Y) en el instrumento analítico bajo las mismas condiciones que posteriormente se utilizarán con los materiales de ensayo.

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

Ejemplo (monensina): Marecek *et al.* (1991) desarrollaron un nuevo método electroquímico para determinar rápidamente la concentración de monensina, un antibiótico poliéter, en las cubas de fermentación donde se produce. El método estándar, un análisis de actividad microbiológica, era complicado y consumía mucho tiempo. Se tomaron muestras en diez cubas de fermentación y se midió la concentración (en ppt) de monensina en cada una de ellas utilizando ambos métodos:

Muestra	Microbiológico	Electroquímico
1	129.5	132.3
2	89.6	91.0
3	76.6	73.6
4	52.2	58.2
5	110.8	104.2
6	50.4	49.9
7	72.4	82.1
8	75.0	73.4
9	34.1	38.1
10	60.3	60.1

Ejemplo (monensina): Diagrama de dispersión



Si estimamos β_0 y β_1 mediante $\hat{\beta}_0$ y $\hat{\beta}_1$, la predicción de la variable respuesta Y para el valor del regresor $X = x_i$ es:

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

Las predicciones $\hat{y}_i$ también se denominan *valores ajustados* o *previstos*.

Unos buenos estimadores deben ser tales que los *errores de predicción* o *residuos*

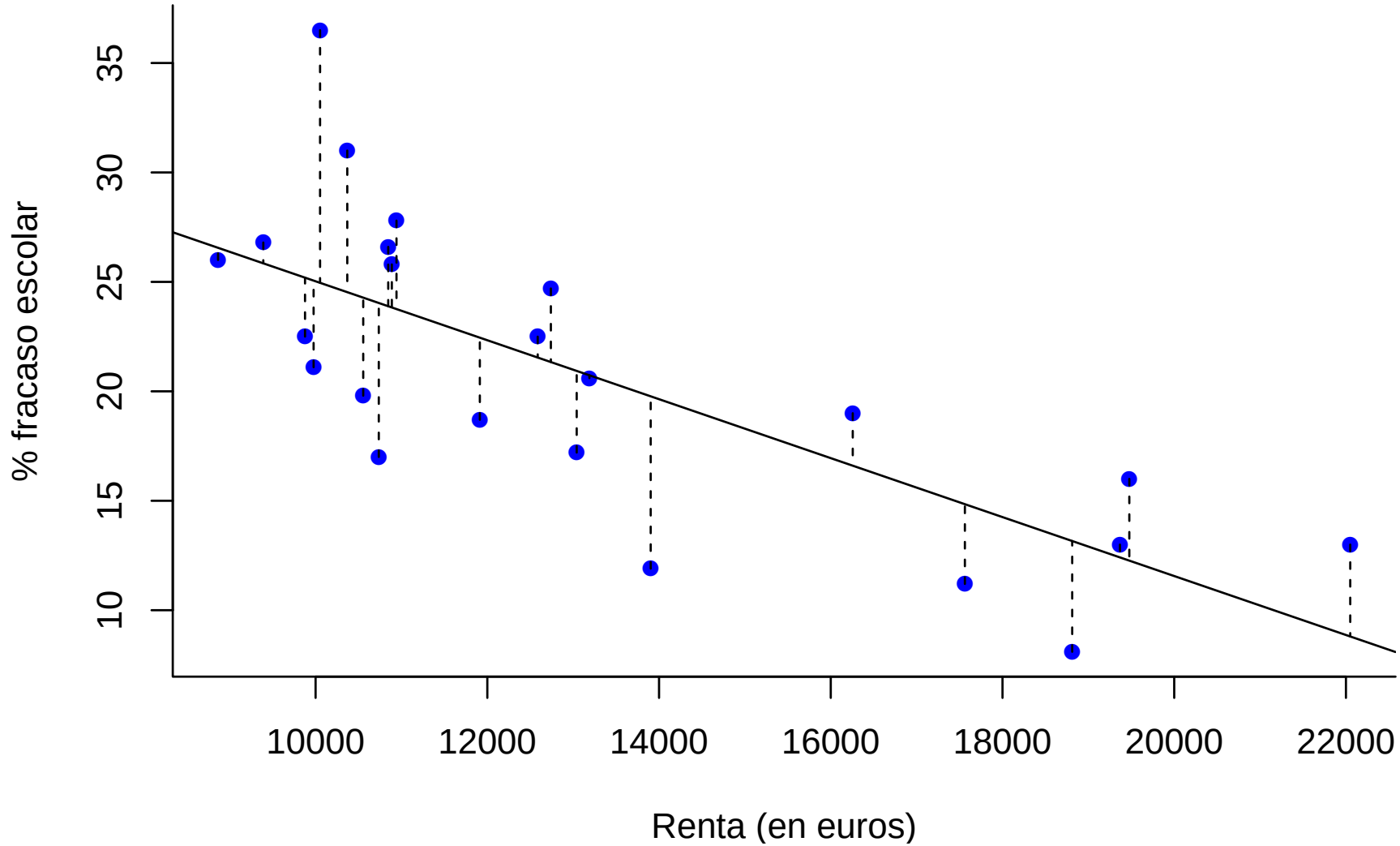
$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

sean pequeños.

En la *recta de regresión de mínimos cuadrados* los valores $\hat{\beta}_0$ y $\hat{\beta}_1$ minimizan *la suma de cuadrados residual*:

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

Ejemplo (fracaso escolar): Diagrama de dispersión y residuos



Estimadores de mínimos cuadrados:

Pendiente

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} = r\sqrt{\frac{S_{yy}}{S_{xx}}}.$$

donde $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$, $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$, $S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2$, y $r = S_{xy}/\sqrt{S_{xx}S_{yy}}$ es el coeficiente de correlación lineal de Pearson.

Término independiente

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

Recta de mínimos cuadrados

$$\hat{y} - \bar{y} = \frac{S_{xy}}{S_{xx}}(x - \bar{x}) \Leftrightarrow \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Mínimos cuadrados como promedio de pendientes:

Reescribimos la expresión del estimador de mínimos cuadrados:

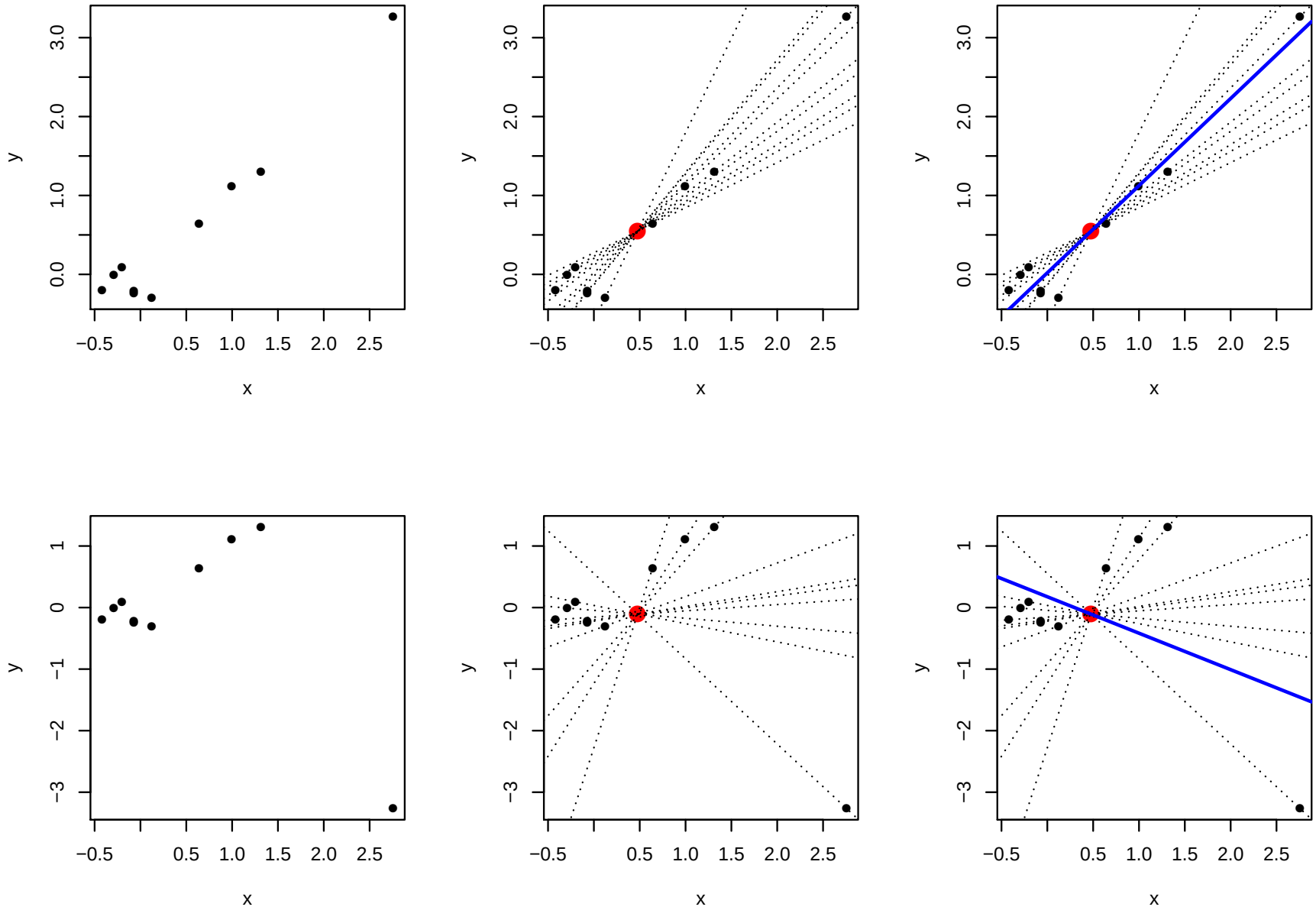
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{S_{xx}} \left(\frac{y_i - \bar{y}}{x_i - \bar{x}} \right) = \sum_{i=1}^{n} w_i \left(\frac{y_i - \bar{y}}{x_i - \bar{x}} \right),$$

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 $\bar{x}$.

Ejemplo (`MinimosCuadrados.R`)**:** 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 (ICs y contrastes) sobre los parámetros, suponemos que se cumple el siguiente modelo:

$$(Y|X = x) = \beta_0 + \beta_1 x + \epsilon,$$

donde:

- La perturbación aleatoria ϵ (cuyos valores observados son los residuos e_i) tiene valor esperado cero: $\mathbb{E}(\epsilon) = 0$.
  $\Leftrightarrow$ *Linealidad:* $\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x$
- La varianza de la perturbación no depende del regresor: $\mathbb{V}(\epsilon) = \sigma^2$.
  $\Leftrightarrow$ *Homocedasticidad:* $\mathbb{V}(Y|X = x) = \sigma^2$.
- La perturbación tiene distribución normal: $\epsilon \sim \mathrm{N}(0, \sigma^2)$.
  $\Leftrightarrow$ *Normalidad:* $(Y|X = x) \sim \mathrm{N}(\beta_0 + \beta_1 x, \sigma^2)$

Además suponemos una hipótesis de *independencia* entre los individuos de la muestra

$$(X_1, Y_1), \ldots, (X_n, Y_n) \quad \text{independientes} \quad i = 1, \ldots, n,$$

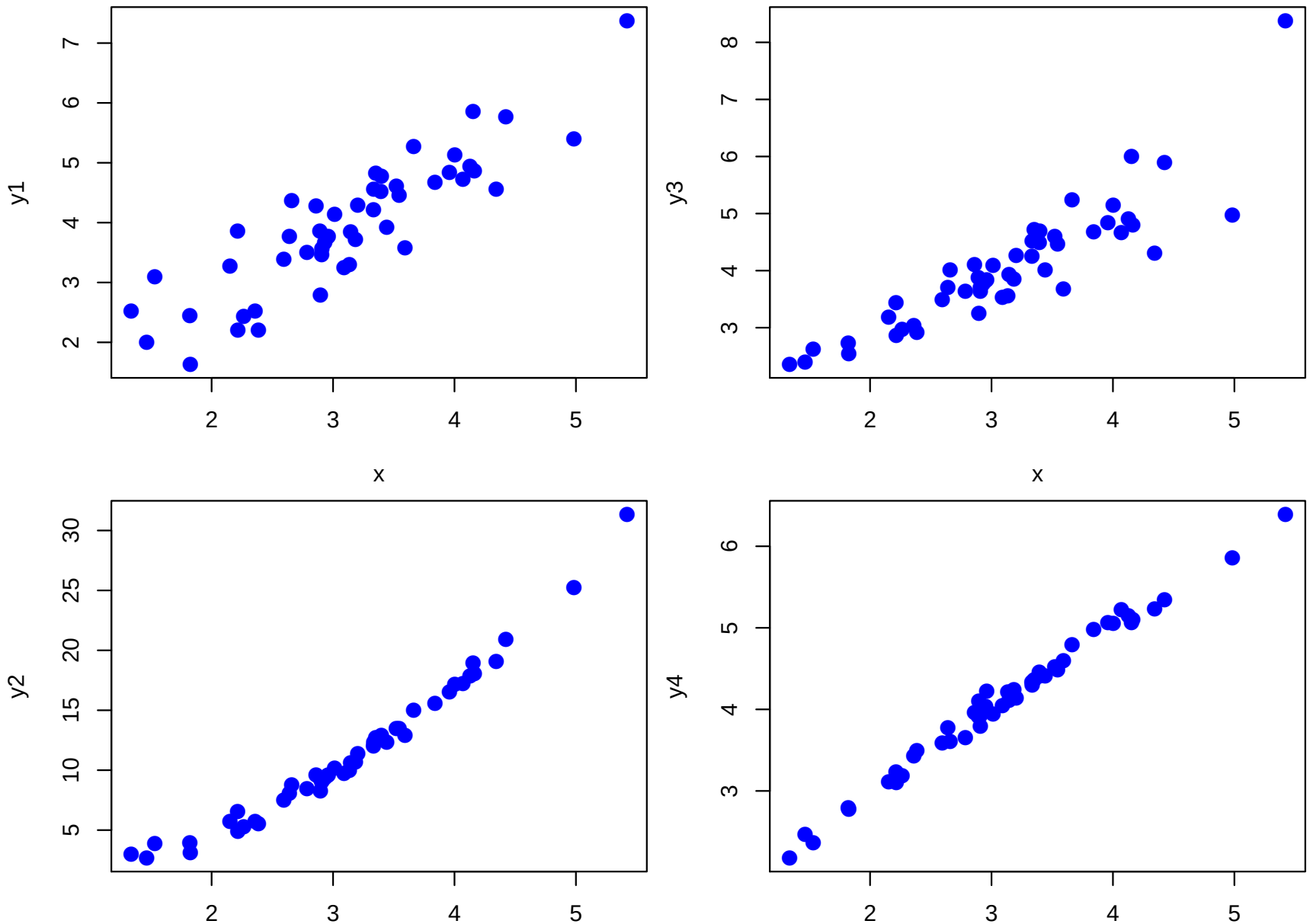
$$\Leftrightarrow X_1, \ldots, X_n, \epsilon_1, \ldots, \epsilon_n \quad \text{independientes} \quad i = 1, \ldots, n.$$

Es decir, suponemos que las perturbaciones del modelo satisfacen

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

En consecuencia, si los datos siguen este modelo, los residuos $e_1, \ldots, e_n$ deberían ser coherentes con estas hipótesis.

Ejemplo (`ModeloRegLinSimple.R`): ¿En cuáles de estas situaciones se satisface el modelo?



Ejemplo (`SimulRegLinSimple.R`): Un ejercicio de simulación

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

Entonces el modelo es

$$(Y|X = x) = x + \epsilon,$$

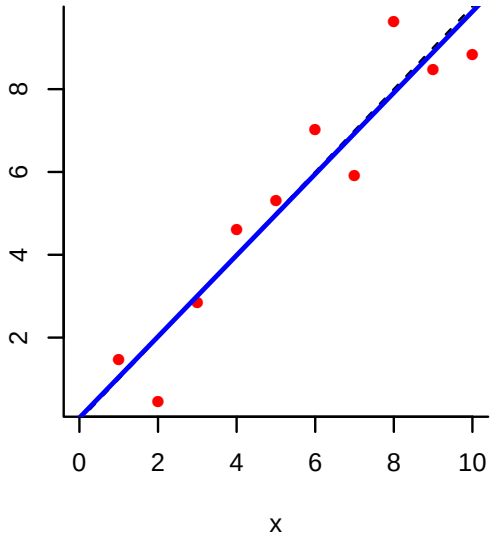
donde la perturbación ϵ tiene distribución normal estándar.

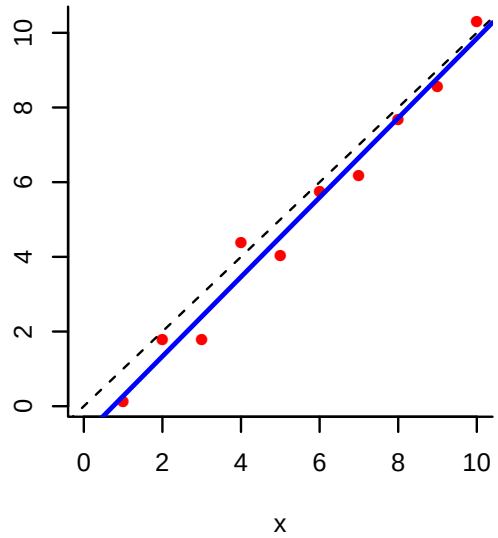
Experimento de simulación:

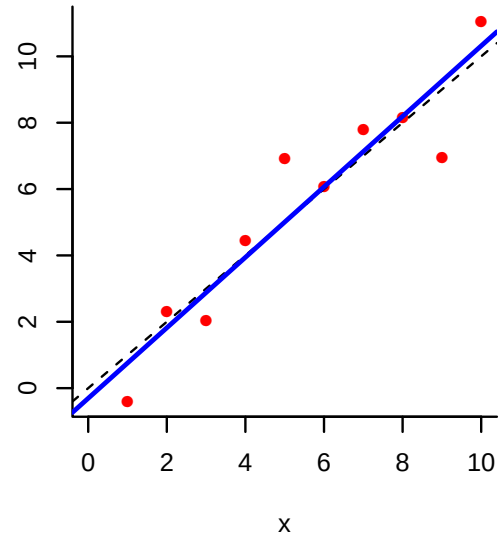
• Fijamos $x_i = 1, 2, \ldots, 10$ ($n = 10$) y generamos las respuestas correspondientes de acuerdo con este modelo.

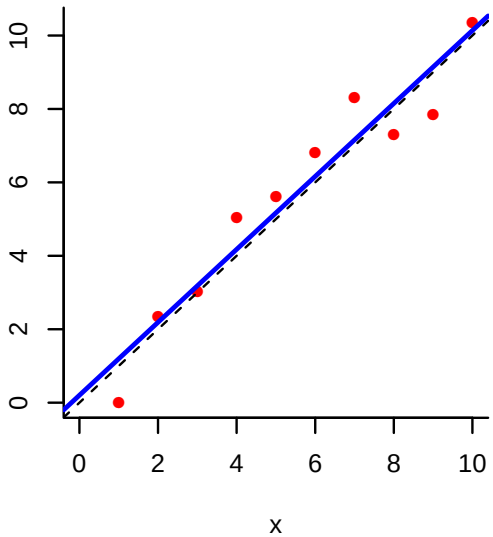
• Posteriormente calculamos la recta de mínimos cuadrados y la representamos junto con la "verdadera recta" $y = x$ (la función de regresión poblacional $\mathbb{E}(Y|X = x) = x$).

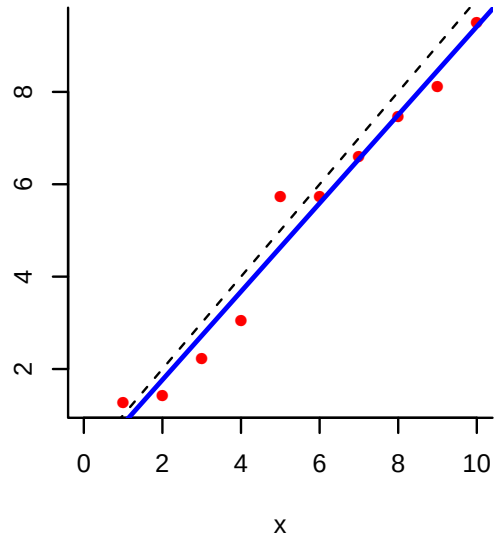
Repetimos 6 veces el experimento:

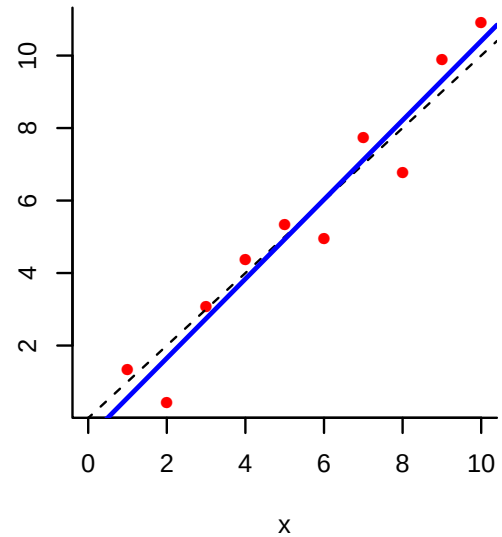




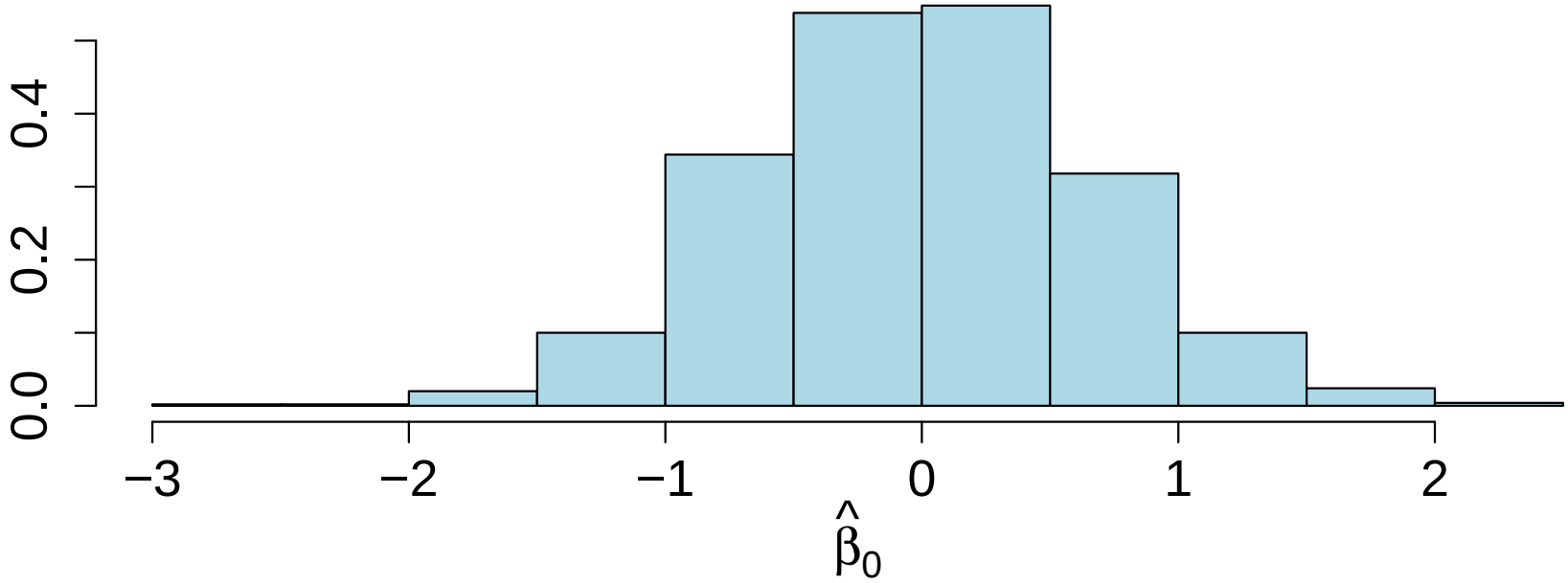


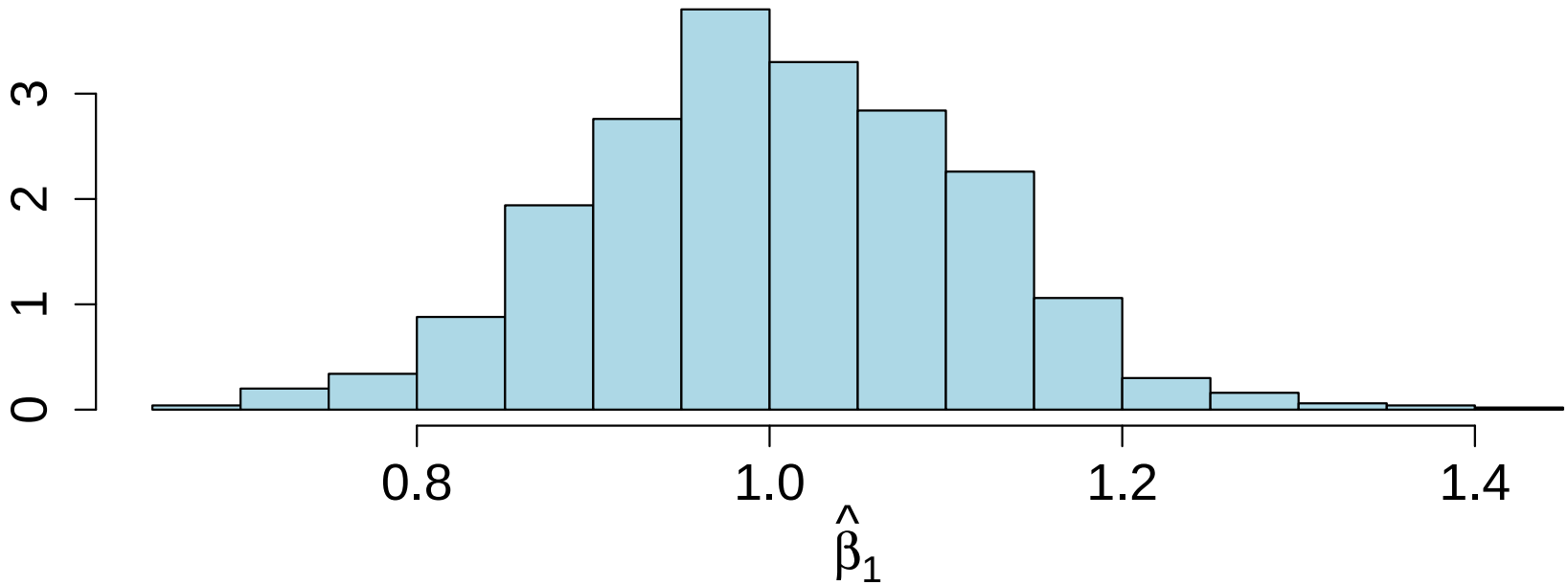






Repetimos 1000 veces el experimento:





Estimación de la varianza

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

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

Se divide por $n-2$ en lugar de n para que el estimador sea insesgado:

$$\mathbb{E}(S_R^2) = \sigma^2$$

es decir, no infraestime sistemáticamente la verdadera varianza.

De hecho, demostraremos que

$$\frac{(n-2)s_R^2}{\sigma^2} \sim \chi^2_{n-2}.$$

Distribución de los estimadores de mínimos cuadrados

Bajo las hipótesis del modelo se cumple:

- $\hat{\beta}_1$ tiene distribución normal de media β_1 y varianza σ^2/S_{xx}.

$$\frac{\hat{\beta}_1 - \beta_1}{s_R\sqrt{\frac{1}{S_{xx}}}} \sim t_{n-2} \Rightarrow \text{IC}_{1-\alpha}(\beta_1) = \left(\hat{\beta}_1 \mp t_{n-2,\alpha/2}\, s_R \sqrt{\frac{1}{S_{xx}}}\right)$$

- $\hat{\beta}_0$ tiene distribución normal de media β_0 y varianza $\sigma^2(1/n + \bar{x}^2/S_{xx})$.

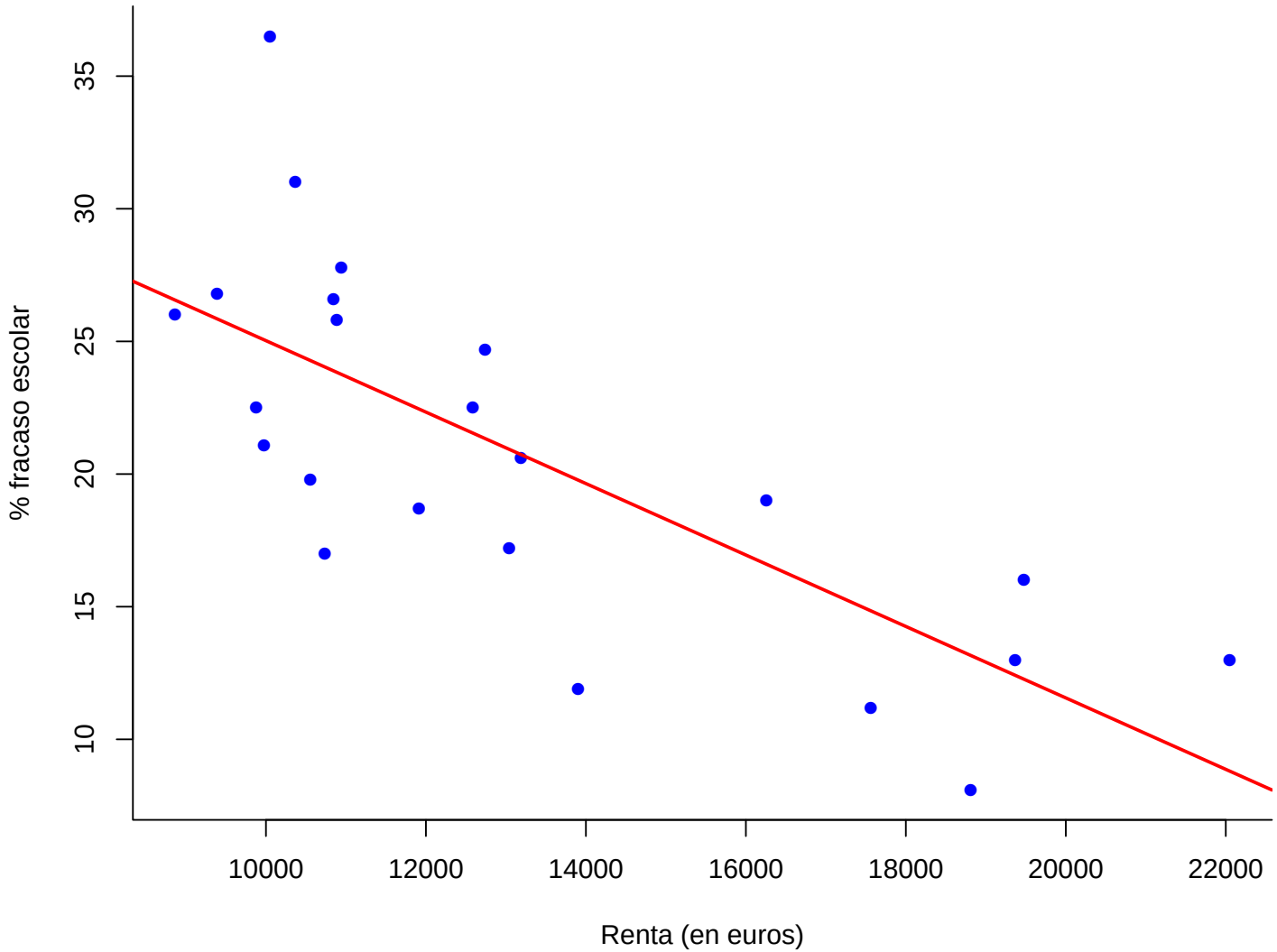
$$\frac{\hat{\beta}_0 - \beta_0}{s_R\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim t_{n-2} \Rightarrow \text{IC}_{1-\alpha}(\beta_0) = \left(\hat{\beta}_0 \mp t_{n-2,\alpha/2}\, s_R \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right)$$

- El vector $(\hat{\beta}_0, \hat{\beta}_1)'$ tiene distribución normal bidimensional y $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x}\sigma^2/S_{xx}$.

Ejemplo (fracaso escolar): Ajuste del modelo con R

```
Datos = read.table("RentaFracaso.txt",header=T)
Renta = Datos$Renta
Fracaso = Datos$Fracaso_escolar
regresion <- lm(Fracaso ~ Renta, data = Datos)
summary(regresion)

Call:
lm(formula = Fracaso ~ Renta, data = Datos)

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) 38.4944272  3.6445192  10.562 7.37e-10 ***
Renta       -0.0013467  0.0002659  -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
```

```
par(mar=c(4,4,1,1))
plot(Renta, Fracaso, pch=16, col="blue", bty="l",
     ylab="% fracaso escolar",xlab="Renta (en miles de euros)")
abline(regresion, col="red", lwd=2)
```



Estimación y predicción

Un modelo de regresión sirve para estimar $E(Y|X = x_0)$ y para predecir futuros valores de Y para un valor x_0 de X.

Los problemas de *estimación* y *predicción* son distintos, aunque las expresiones matemáticas que aparecen en su resolución son casi iguales. En el primero intentamos obtener un estimador de $E(Y|X = x_0) = \beta_0 + \beta_1 x_0$, que es un número fijo aunque desconocido.

En el problema de predicción de $Y_0 = Y|X = x_0$ estamos interesados en conocer, para un valor x_0 fijo de X, el valor correspondiente de Y. $Y|X = x_0$ es una variable aleatoria.

Al final estimaremos $E(Y|X = x_0)$ y predeciremos $Y_0 = Y|X = x_0$ mediante el mismo valor, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_0$, pero *el error de estimación y el de predicción son distintos*.

Estimación de la media condicionada

Queremos estimar $E(Y_0) = E(Y|X = x_0) = \beta_0 + \beta_1 x_0$, el valor promedio de la respuesta cuando $X = x_0$.

Un estimador razonable es

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \bar{y} + \hat{\beta}_1 (x_0 - \bar{x}).$$

Se trata de un estimador centrado: $E(\hat{y}_0) = E(Y|X = x_0)$.

Además

$$\text{IC}_{1-\alpha}(E(Y_0)) = \left(\hat{y}_0 \mp t_{n-2,\alpha/2}\, s_R \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right).$$

Predicción de la respuesta

En el problema de predicción deseamos prever $Y_0 = (Y|X = x_0)$, la respuesta cuando la variable independiente es igual a x_0.

Si conociéramos $E(Y|X = x_0)$ podríamos utilizar esta esperanza como predicción de $Y_0 = (Y|X = x_0)$. Entonces ya tenemos una primera fuente de error debido a la propia variabilidad de $Y|X = x_0$ en torno a su media.

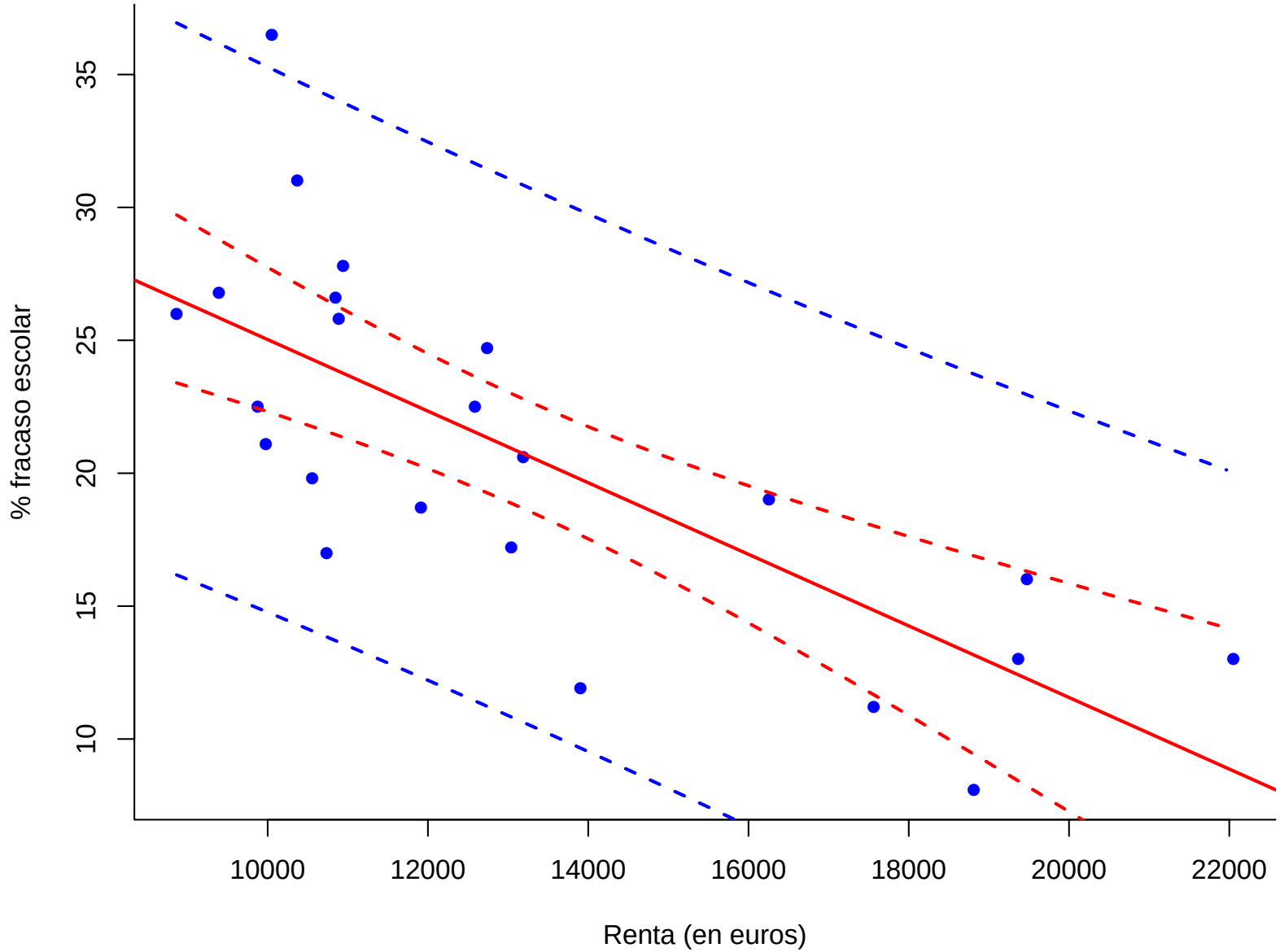
Además, como $E(Y|X = x_0)$ es desconocida, la estimamos mediante $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$: segunda fuente de error en la predicción. Por tanto, finalmente predecimos $Y|X = x_0$ mediante $\hat{y}_0$.

Un intervalo de confianza para la predicción de $Y|X = x_0$ es

$$\mathrm{IC}_{1-\alpha}(Y_0) = \left(\hat{y}_0 \mp t_{n-2,\alpha/2}\, s_R \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right).$$

Ejemplo (fracaso escolar): Bandas de confianza para estimación y predicción

```
valores.x0 <- seq(min(Renta), max(Renta), 100)
datos <- data.frame(Renta = valores.x0)
confianza <- predict(regresion, datos, interval=c("confidence"))
prediccion <- predict(regresion, datos, interval=c("prediction"))

# Representación gráfica
plot(Renta, Fracaso, pch=16, col="blue", bty="l",
     ylab="% fracaso escolar",xlab="Renta (en 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)
```


% fracaso escolar
10
15
20
25
30
35
10000
12000
14000
16000
18000
20000
22000
Renta (en euros)

Regresión lineal múltiple

En cada uno de los n individuos de una muestra observamos una variable respuesta Y y k regresores $\mathbf{X} = (X_1, \ldots, X_k)'$. La muestra es

$$(y_i, \mathbf{x}_i') = (Y_i, x_{i1}, x_{i2}, \ldots, x_{ik}), \quad i = 1, \ldots, n.$$

Modelo de regresión lineal múltiple:

$$(Y|\mathbf{X} = \mathbf{x}) = (Y|X_1 = x_1, \ldots, X_k = x_k) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon,$$

donde

- *Linealidad:* $\mathbb{E}(\epsilon) = 0 \Leftrightarrow \mathbb{E}(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
- *Homocedasticidad:* $\mathbb{V}(\epsilon) = \sigma^2 \Leftrightarrow \mathbb{V}(Y|\mathbf{X} = \mathbf{x}) = \sigma^2$
- *Normalidad:*
  $\epsilon \sim \mathrm{N}(0, \sigma^2) \Leftrightarrow (Y|\mathbf{X} = \mathbf{x}) \sim \mathrm{N}(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \sigma^2)$

Además suponemos

- *Independencia* entre los individuos de la muestra

$$(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n) \quad \text{independientes} \quad i = 1, \ldots, n,$$

$$\Leftrightarrow \mathbf{X}_1, \ldots, \mathbf{X}_n, \epsilon_1, \ldots, \epsilon_n \quad \text{independientes} \quad i = 1, \ldots, n.$$

  Es decir, suponemos que las perturbaciones del modelo satisfacen

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

- Los regresores $X_1, \ldots, X_k$ son linealmente independientes entre sí (no hay *colinealidad*).
- $n \geq k + 2$ (por lo menos hay tantas observaciones como parámetros a estimar en el modelo).

El modelo admite una expresión equivalente en forma matricial:

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}}_{\mathbb{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}_{\boldsymbol{\epsilon}}$$

o

$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

donde $\mathbb{X}$ es la *matriz del diseño* y $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. O, lo que es igual,

$$\mathbf{Y} \sim \mathrm{N}_n(\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$$

Una interpretación geométrica

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

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

El modelo equivale a suponer
$(\mathbf{Y}|\mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_n = \mathbf{x}_n) \sim \mathrm{N}_n(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$, donde $\boldsymbol{\mu} \in \mathcal{V}$.

Estimación de los parámetros del modelo:

Mediante el *criterio de mínimos cuadrados*, los estimadores son los valores $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_k)'$ para los que se minimiza

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

Observación: La predicción $\hat{\mathbf{y}} := \mathbb{X}\hat{\boldsymbol{\beta}}$ es la *proyección ortogonal* de $\mathbf{y}$ sobre $\mathcal{V}$.

Estimadores de mínimos cuadrados:

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

La matriz $\mathbb{X}'\mathbb{X}$ es invertible si se cumplen las hipótesis básicas de ausencia de colinealidad y $n \geq k+2$.

Residuos:

$$\mathbf{e} = (e_1, \ldots, e_n)' = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y},$$

donde $\mathbf{H} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$ (llamada *matriz proyección* o *hat matrix*) es simétrica e idempotente. $\mathbf{H}$ es la matriz de proyección ortogonal sobre $\mathcal{V}$.

Ecuaciones normales:

$$\mathbb{X}'(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{0}_{k+1} \Leftrightarrow \mathbb{X}'\mathbf{e} = \mathbf{0}_{k+1}$$

Por tanto, los residuos tienen $n-(k+1)$ grados de libertad.

Estimación de la varianza: Mediante la *varianza residual*

$$s_R^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2.$$

Propiedades de los estimadores

- El vector $\hat{\boldsymbol{\beta}}$ tiene distribución normal $(k+1)$-dimensional con esperanza $\boldsymbol{\beta}$ y matriz de covarianzas $\sigma^2(\mathbb{X}'\mathbb{X})^{-1}$.
- $\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \mathbf{y}'(\mathbf{I}_n - \mathbf{H})\mathbf{y}$.
- $\dfrac{(n-k-1)s_R^2}{\sigma^2} \sim \chi^2_{n-k-1}$.
- s_R^2 y $\hat{\boldsymbol{\beta}}$ son independientes.

Inferencia sobre los parámetros del modelo

- Distribución: Para todo $j = 0, \ldots, k$,

$$\frac{\hat{\beta}_j - \beta_j}{\text{s.e.}(\hat{\beta}_j)} \sim t_{n-k-1},$$

donde s.e.$(\hat{\beta}_j)$ es el *error típico* (*standard error*) del coeficiente β_j:

$$(\text{s.e.}(\hat{\beta}_j))^2 = s_R^2 \, q_{jj}$$

y q_{jj} es el elemento $j+1$ de la diagonal de $(\mathbb{X}'\mathbb{X})^{-1}$.

- Intervalos de confianza: Para todo $j = 0, 1, \ldots, k$,

$$\text{IC}_{1-\alpha}(\beta_j) = \left(\hat{\beta}_j \mp t_{n-k-1;\alpha/2} \, \text{s.e.}(\hat{\beta}_j)\right).$$

- Contrastes de hipótesis individuales sobre los coeficientes

Satisfecho el modelo de regresión lineal múltiple, 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 de rechazo de H_0 al nivel de significación α es

$$R_j = \{|t(\beta_j)| > t_{n-k-1;\alpha/2}\},$$

siendo $t(\beta_j) = \hat{\beta}_j/\text{s.e.}(\hat{\beta}_j)$ el 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.

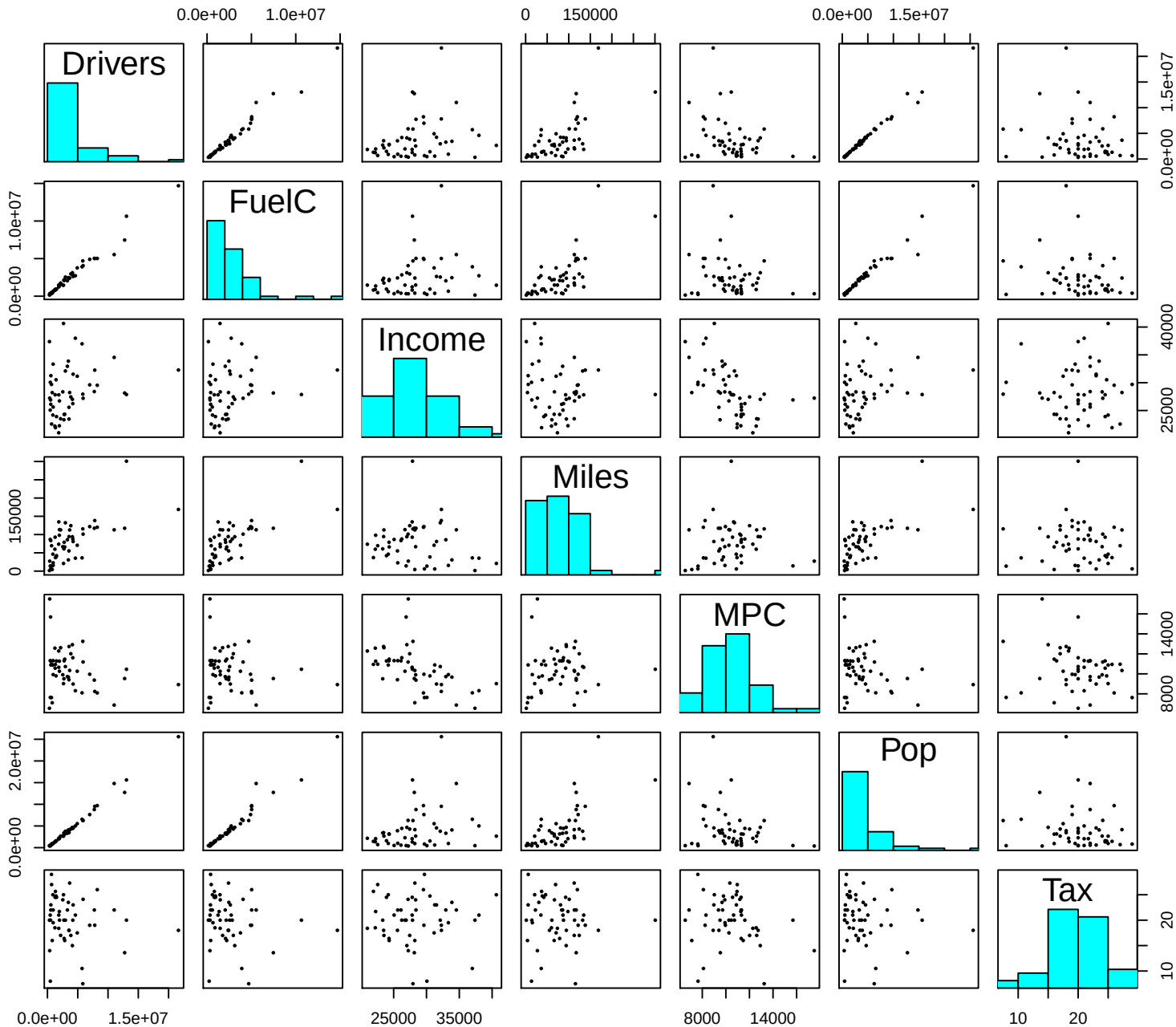
```
load("combustible.Rdata")
head(fuel2001) # Para ver por pantalla las primeras líneas de datos
```

	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

La primera columna son las abreviaturas postales (2 dígitos) de los estados de EEUU.

`Drivers`	Nº de permisos de conducir en el estado
`FuelC`	Gasolina vendida para conducción (en miles de galones)
`Income`	Renta per capita (año 2000)
`Miles`	Millas de autovías del Federal-aid highway program en el estado
`MPC`	Millas conducidas per capita (estimación)
`Pop`	Población con edad mínima de 16 años
`Tax`	Tasa impositiva estatal sobre la gasolina (en centavos por galón)

```
# Función para poner histogramas en la diagonal
# del diagrama de dispersión múltiple:

panel.hist <- function(x, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(usr[1:2], 0, 1.5) )
  h <- hist(x, plot = FALSE)
  breaks <- h$breaks; nB <- length(breaks)
  y <- h$counts; y <- y/max(y)
  rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
}

# Diagrama de dispersión múltiple con histogramas en la diagonal

pairs(fuel2001,pch=16,cex=0.5,oma=c(2,2,2,2),
      diag.panel = panel.hist,bg = "light blue")
```


Drivers
FuelC
Income
Miles
MPC
Pop
Tax
0.0e+00
1.0e+07
1.5e+07
2.0e+07
0
150000
25000
35000
40000
8000
14000
10
20

Ejemplo (combustible EE.UU.): 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)
Call:
lm(formula = FuelC ~ Drivers+Income+Miles+MPC+Tax, data=fuel2001)

Residuals:
     Min       1Q   Median       3Q      Max
-1517423  -161111    19930   173532  1085471

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.844e+05  8.102e+05  -0.598 0.552903
Drivers      6.144e-01  2.229e-02  27.560  < 2e-16 ***
Income       7.526e+00  1.611e+01   0.467 0.642587
Miles        5.813e+00  1.587e+00   3.664 0.000652 ***
MPC          4.643e+01  3.488e+01   1.331 0.189820
Tax         -2.114e+04  1.298e+04  -1.629 0.110298
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 394100 on 45 degrees of freedom
Multiple R-squared:  0.9808,    Adjusted R-squared:  0.9787
F-statistic: 459.5 on 5 and 45 DF,  p-value: < 2.2e-16
```

Ejemplo (combustible EE.UU.):

Ejercicio: Lleva a cabo los contrastes de la forma $H_0 : \beta_j = 0$ para todos los coeficientes del modelo ($\alpha = 0.05$).

Desviación típica residual s_R:

```
sR2 = sum((reg$residuals)^2)/reg$df.residual
sqrt(sR2)
[1] 394072.4
```

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 contraste de la regresión. Análisis de la varianza

Bajo el modelo de regresión lineal múltiple, 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)

La *suma de cuadrados total* (*total sum of squares*)

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \mathbf{y}'(\mathbf{I}_n - \mathbf{M})\mathbf{y},$$

donde $\mathbf{M} = \mathbf{1}_n\mathbf{1}_n'/n$ y $\mathbf{I}_n - \mathbf{M}$ es la *matriz de centrado*, mide la variabilidad total en la respuesta.

La *suma de cuadrados del modelo de regresión* (*model sum of squares*)

$$\text{MSS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \mathbf{y}'(\mathbf{H} - \mathbf{M})\mathbf{y}$$

mide la parte de la variabilidad explicada por el modelo.

La *suma de cuadrados de los errores o residual* (*residual sum of squares*)

$$\text{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \mathbf{y}'(\mathbf{I}_n - \mathbf{H})\mathbf{y}$$

mide la parte de la variabilidad no explicada por el modelo.

Hacemos un análisis de la varianza, es decir, examinamos qué proporción de la TSS es explicada por el modelo de regresión utilizando la *descomposición de la variabilidad*:

$$\begin{aligned} \mathbf{y}'(\mathbf{I}_n - \mathbf{M})\mathbf{y} &= \mathbf{y}'(\mathbf{H} - \mathbf{M})\mathbf{y} + \mathbf{y}'(\mathbf{I}_n - \mathbf{H})\mathbf{y} \\ \|\mathbf{y} - \mathbf{M}\mathbf{y}\|^2 &= \|\mathbf{H}\mathbf{y} - \mathbf{M}\mathbf{y}\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \\ \text{TSS} &= \text{MSS} + \text{RSS} \end{aligned}$$

Comparamos la variabilidad explicada por el modelo con la no explicada mediante el estadístico del contraste:

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

Tabla ANOVA para el contraste de la regresión:

	SS	df	MS	F
Modelo	MSS	k	$\frac{\text{MSS}}{k}$	F
Residual	RSS	$n-k-1$	s_R^2	
Total	TSS	$n-1$		

Bajo $H_0 : \beta_1 = \ldots = \beta_k = 0$, el estadístico F sigue una distribución $F_{k,n-k-1}$.

La región de rechazo de $H_0 : \beta_1 = \ldots = \beta_k = 0$ al nivel de significación α es

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

Ejemplo (combustible EE.UU.):

```
# Regresión sin regresores, sólo con término independiente
model0 = lm(FuelC ~ 1,data=fuel2001)
anova(model0,reg)
```

```
Analysis of Variance Table

Model 1: FuelC ~ 1
Model 2: FuelC ~ Drivers + Income + Miles + MPC + Tax
  Res.Df        RSS Df Sum of Sq      F    Pr(>F)
1     50 3.6379e+14
2     45 6.9882e+12  5 3.568e+14 459.52 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\text{TSS} = 3.6379\text{e}+14 \qquad \text{RSS} = 6.9882\text{e}+12$$

$$n-1 = 50 \qquad n-k-1 = 45$$

$$F = 459.52 \qquad \text{p-valor} < 2.2\text{e}-16$$

El coeficiente de determinación

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

$$R^2 = \frac{\text{MSS}}{\text{TSS}}.$$

Podemos interpretar R^2 como un coeficiente de correlación múltiple entre Y y las k variables regresoras.

Propiedades:

- $0 \leq R^2 \leq 1$.
- Cuando $R^2 = 1$, $\hat{y}_i = y_i$ para todo i.
- Cuando $R^2 = 0$, $\hat{y}_i = \bar{y}$ para todo $i = 1, \ldots, n$.
- En regresión simple $R^2 = r^2$.
- Se cumple que $F = \dfrac{R^2}{1 - R^2} \dfrac{n - k - 1}{k}$.

El coeficiente de determinación ajustado

Un inconveniente del coeficiente de determinación para comparar distintos modelos de regresión entre sí es que 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*

$$\bar{R}^2 = 1 - \frac{\text{RSS}/(n-k-1)}{\text{TSS}/(n-1)} = 1 - \frac{s_R^2}{s_y^2}.$$

$\bar{R}^2$ sólo aumenta al introducir un nuevo regresor en el modelo si la varianza residual disminuye.

Ejemplo (combustible EE.UU.):

```
reg = lm(FuelC ~ Drivers+Income+Miles+MPC+Tax,data=
    fuel2001)
Multiple R-squared:  0.9808, Adjusted R-squared:  0.9787

reg1 = lm(FuelC ~ Drivers,data=fuel2001)
Multiple R-squared:  0.9704, Adjusted R-squared:  0.9698

reg2 = lm(FuelC ~ MPC+Tax,data=fuel2001)
Multiple R-squared:  0.1002, Adjusted R-squared:  0.06276
```

Contraste de hipótesis lineales

Queremos contrastar $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{0}$, siendo $\mathbf{A}$ una matriz $p \times (k+1)$ con $\text{rg}(\mathbf{A}) = p < k+1$.

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

$$H_0 : \beta_1 = \beta_2; \beta_0 = 0 \Leftrightarrow \mathbf{A}\boldsymbol{\beta} = \mathbf{0},$$

donde

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

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:

RSS_0, la variabilidad no explicada (residual) bajo el modelo reducido, M0;

RSS, la variabilidad no explicada (residual) bajo el modelo completo, M.

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

Bajo $H_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{0}$, se cumple

$$F = \frac{(\text{RSS}_0 - \text{RSS})/p}{\text{RSS}/(n-k-1)} \sim F_{p,n-k-1}.$$

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

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

En particular, podemos comparar dos modelos anidados.

Ejemplo (combustible EE.UU.):

```
reg = lm(FuelC ~ Drivers+Income+Miles+MPC+Tax,data=fuel2001)
reg1 = lm(FuelC ~ Drivers,data=fuel2001)
anova(reg1,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
```

$$F = \frac{\frac{RSS_0 - RSS}{p}}{\frac{RSS}{n-k-1}} = \frac{\frac{1.0775 \cdot 10^{13} - 6.9882 \cdot 10^{12}}{4}}{\frac{6.9882 \cdot 10^{12}}{45}} = 6.0962$$

Interpretación de los contrastes

Contraste global (F)	**Contrastes individuales (t)**	**Conclusión**
Modelo explicativo	Todas las X_i explicativas	Nos quedamos con todas las X_i
Modelo explicativo	Algunas X_i explicativas	Nos quedamos con las X_i explicativas
Modelo explicativo	Ninguna X_i explicativa	Colinealidad
Modelo no explicativo	Todas las X_i explicativas	Colinealidad
Modelo no explicativo	Algunas X_i explicativas	Colinealidad
Modelo no explicativo	Ninguna X_i explicativa	Modelo no adecuado para describir la relación entre Y y $X_1, \ldots, X_K$.

Análisis de influencia

En algunos conjuntos de datos, los estadísticos calculados pueden cambiar mucho si se elimina un individuo de la muestra. Entonces decimos que se trata de un *dato influyente*.

El *análisis de influencia* estudia cambios en los resultados de la regresión cuando se perturban ligeramente los datos. La perturbación de datos más habitual es eliminar de la muestra los datos uno a uno. A continuación se estudia la influencia de ese individuo comparando el análisis a partir de la muestra completa con aquél resultante de eliminar el dato.

Utilizamos la notación con el subíndice (i) para indicar "con el i-ésimo dato eliminado":

$$\hat{\boldsymbol{\beta}}_{(i)} = (\mathbb{X}'_{(i)}\mathbb{X}_{(i)})^{-1}\mathbb{X}'_{(i)}\mathbf{Y}_{(i)}.$$

Ejemplo (cuencas fluviales): Para estudiar la relación entre calidad del agua y uso del terreno, Haith (1976) observó las siguientes variables en 20 cuencas fluviales del estado de Nueva York (EEUU), obteniendo los datos del fichero `NewYorkRivers.txt`:

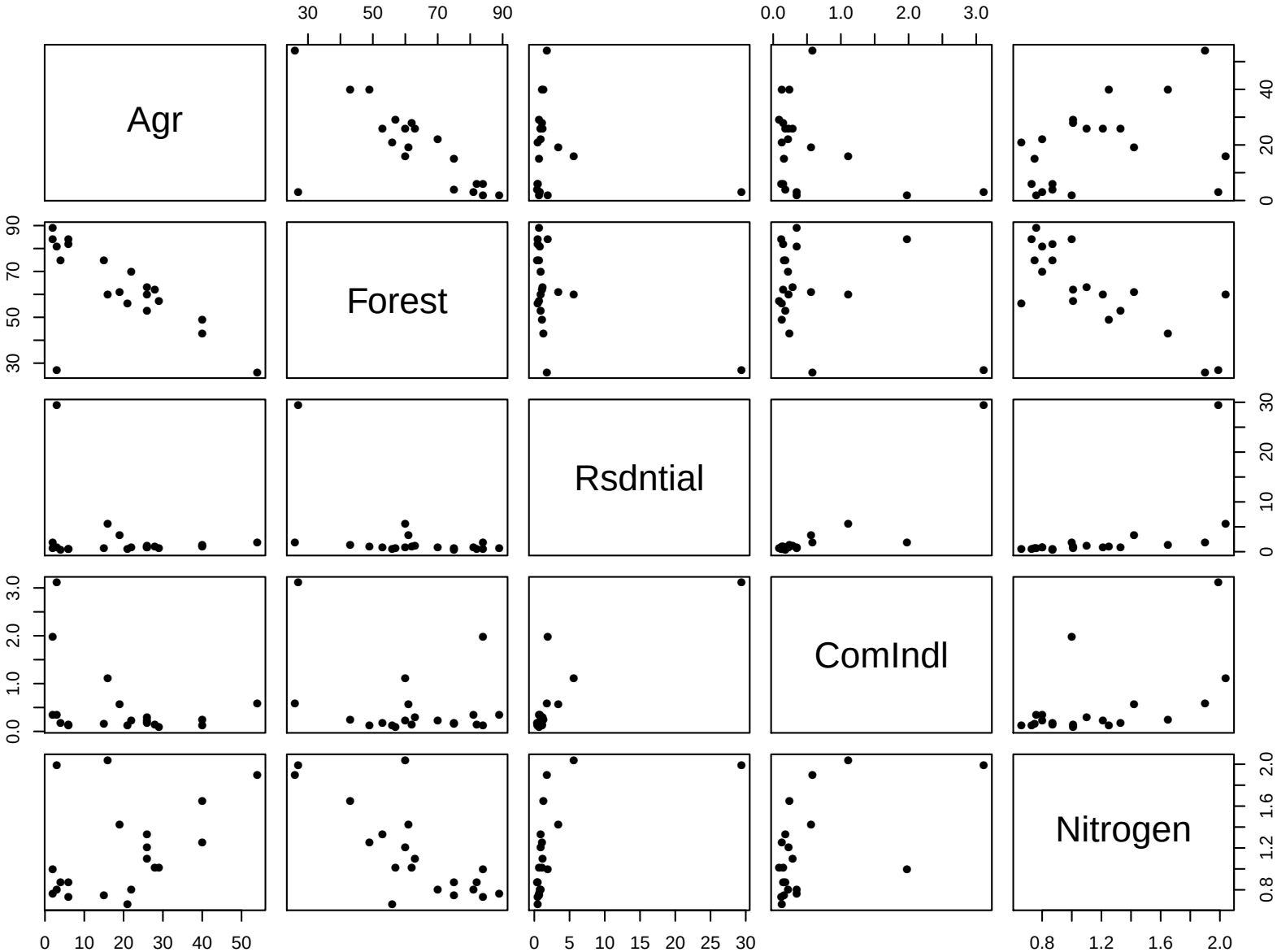
Y	$=$	Concentración de nitrógeno (en mg/l), promedio de medidas realizadas a intervalos regulares durante primavera, verano y otoño (`Nitrogen` en `NewYorkRivers.txt`)
X_1	$=$	Porcentaje de superficie terrestre para uso agrícola (`Agr` en `NewYorkRivers.txt`)
X_2	$=$	Porcentaje de superficie terrestre cubierta por bosque (`Forest` en `NewYorkRivers.txt`)
X_3	$=$	Porcentaje de superficie terrestre para uso residencial (`Rsdntial` en `NewYorkRivers.txt`)
X_4	$=$	Porcentaje de superficie terrestre para uso comercial o industrial (`ComIndl` en `NewYorkRivers.txt`)

```
Datos = read.table("NewYorkRivers.txt",header=T,sep="\t")
```

Ejemplo (cuencas fluviales):

Río	Y	X_1	X_2	X_3	X_4
Olean	1.10	26	63	1.2	0.29
Cassadaga	1.01	29	57	0.7	0.09
Oatka	1.90	54	26	1.8	0.58
Neversink	1.00	2	84	1.9	1.98
Hackensack	1.99	3	27	29.4	3.11
Wappinger	1.42	19	61	3.4	0.56
Fishkill	2.04	16	60	5.6	1.11
Honeoye	1.65	40	43	1.3	0.24
Susquehanna	1.01	28	62	1.1	0.15
Chenango	1.21	26	60	0.9	0.23
Tioughnioga	1.33	26	53	0.9	0.18
West Canada	0.75	15	75	0.7	0.16
East Canada	0.73	6	84	0.5	0.12
Saranac	0.80	3	81	0.8	0.35
Ausable	0.76	2	89	0.7	0.35
Black	0.87	6	82	0.5	0.15
Schoharie	0.80	22	70	0.9	0.22
Raquette	0.87	4	75	0.4	0.18
Oswegatchie	0.66	21	56	0.5	0.13
Cohocton	1.25	40	49	1.1	0.13

Ejemplo (cuencas fluviales): `pairs(Datos[,-1],pch=16)`



Ejemplo (cuencas fluviales): Contrastes individuales

```
reg = lm(Nitrogen~Agr+Forest+Rsdntial+ComIndl,data=Datos)
summary(reg)

Call:
lm(formula = Nitrogen ~ Agr + Forest + Rsdntial + ComIndl, data = Datos)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49404 -0.13180  0.01951  0.08287  0.70480

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.722214   1.234082   1.396   0.1832
Agr          0.005809   0.015034   0.386   0.7046
Forest      -0.012968   0.013931  -0.931   0.3667
Rsdntial    -0.007227   0.033830  -0.214   0.8337
ComIndl      0.305028   0.163817   1.862   0.0823 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2649 on 15 degrees of freedom
Multiple R-squared:  0.7094,    Adjusted R-squared:  0.6319
F-statistic: 9.154 on 4 and 15 DF,  p-value: 0.0005963
```

Ejemplo (cuencas fluviales): Contraste global

```
reg0 = lm(Nitrogen~1,data=Datos)
anova(reg0,reg)
Analysis of Variance Table

Model 1: Nitrogen ~ 1
Model 2: Nitrogen ~ Agr + Forest + Rsdntial + ComIndl
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     19 3.6226
2     15 1.0527  4    2.5699 9.1542 0.0005963 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Las regresiones simples muestran ausencia de relación entre la respuesta y las variables (esto se observa un poco también en el gráfico).
- La regresión múltiple, sin embargo, indica que conjuntamente las variables son significativas.
- ¿Por qué se produce la paradoja?

Como $\mathbf{e} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$, se cumple que $\mathbb{V}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$. Por tanto, $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$, con $\mathbf{H} = (h_{ij})$.

El *potencial* (*leverage*) de un dato muestral $\mathbf{x}_i$ es h_{ii}, que además está estrechamente relacionado con la distancia de Mahalanobis:

$$h_{ii} = \frac{1}{n} + \frac{1}{n-1} d_M^2(\mathbf{x}_i, \bar{\mathbf{x}}) = \frac{1}{n} + (\mathbf{x}_i - \bar{\mathbf{x}})'(\tilde{\mathbb{X}}'\tilde{\mathbb{X}})^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}),$$

siendo $\tilde{\mathbb{X}} = (x_{ij})_{i=1,\ldots,n, j=1,\ldots,k}$ la matriz del diseño sin la columna de 1's.

El potencial es una medida de lo distante que está el regresor $\mathbf{X} = \mathbf{x}_i$ de los valores restantes del regresor $\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_n$.

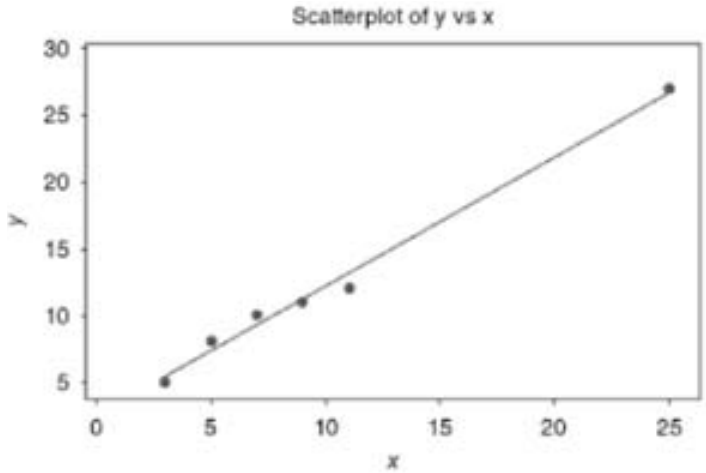


Figure 4.1 Example of a pure leverage point.

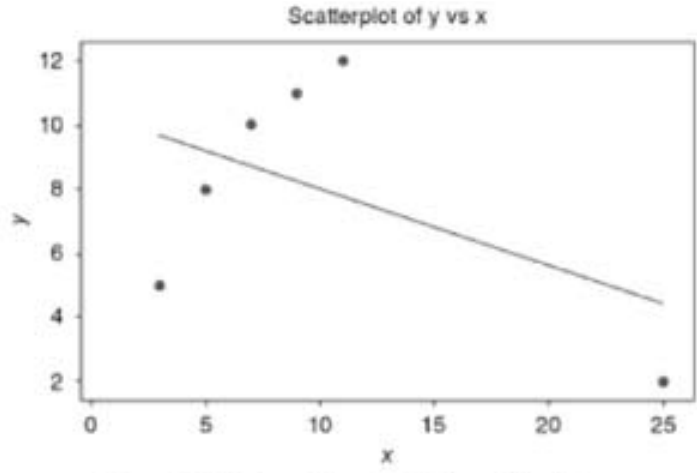
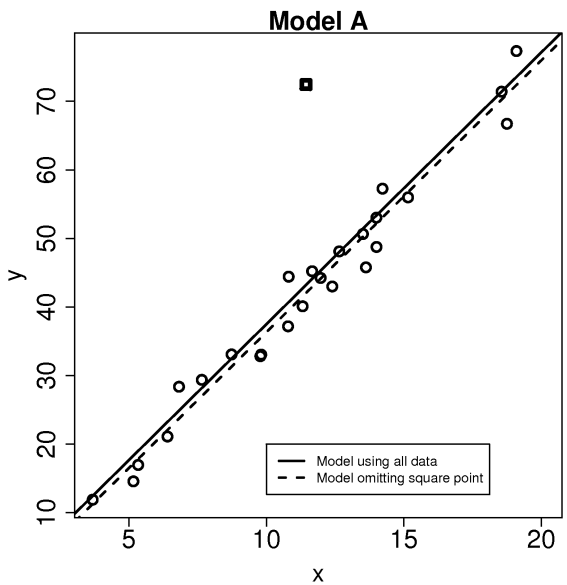
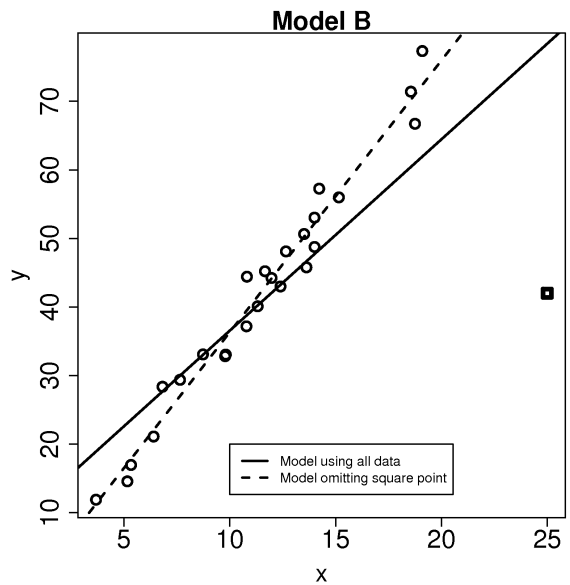
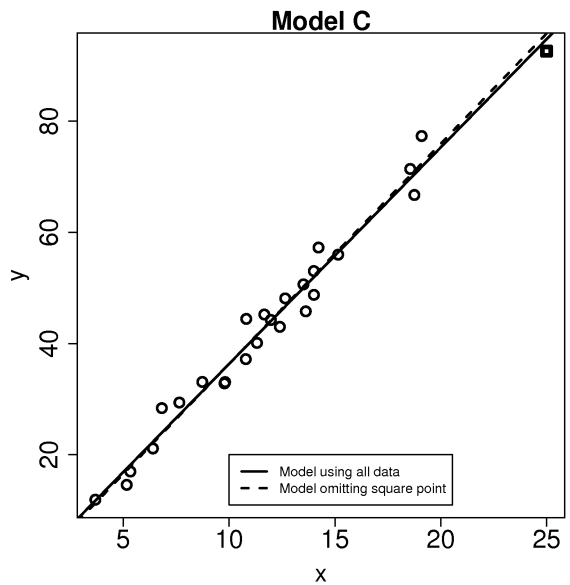


Figure 4.2 Example of an influential point.







La *distancia de Cook* (Cook 1977) mide cómo cambia el vector de estimadores $\hat{\boldsymbol{\beta}}$ cuando se elimina cada observación, es decir, compara $\hat{\boldsymbol{\beta}}$ con cada $\hat{\boldsymbol{\beta}}_{(i)}$, $i = 1, \ldots, n$.

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

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

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'\mathbb{X}'\mathbb{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{(k+1)s_R^2} = \frac{\|\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}}\|^2}{(k+1)s_R^2},$$

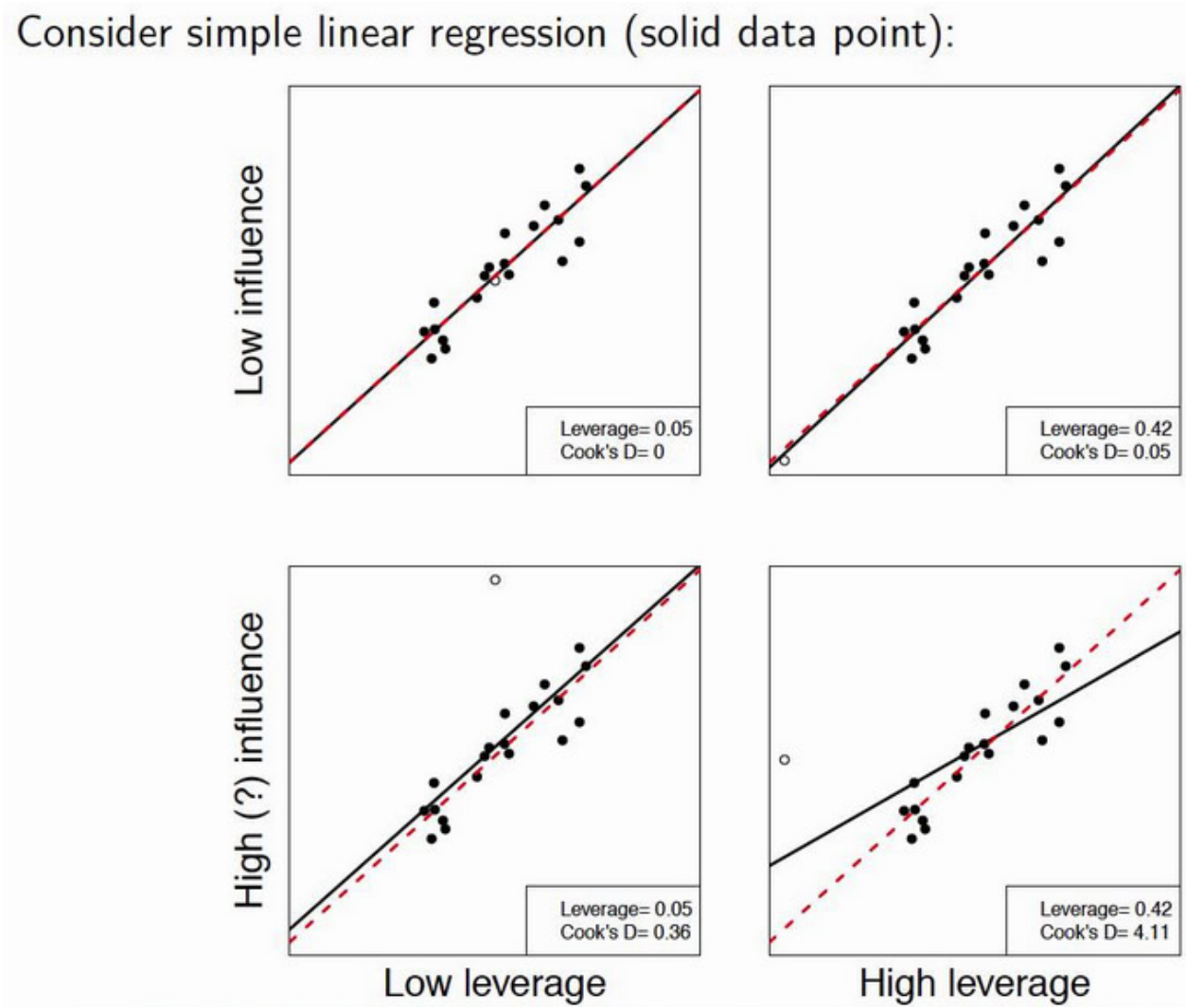
siendo $\hat{\mathbf{y}}_{(i)} = \mathbb{X}\boldsymbol{\beta}_{(i)}$.

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.

La distancia de Cook está relacionada con el potencial y los residuos:

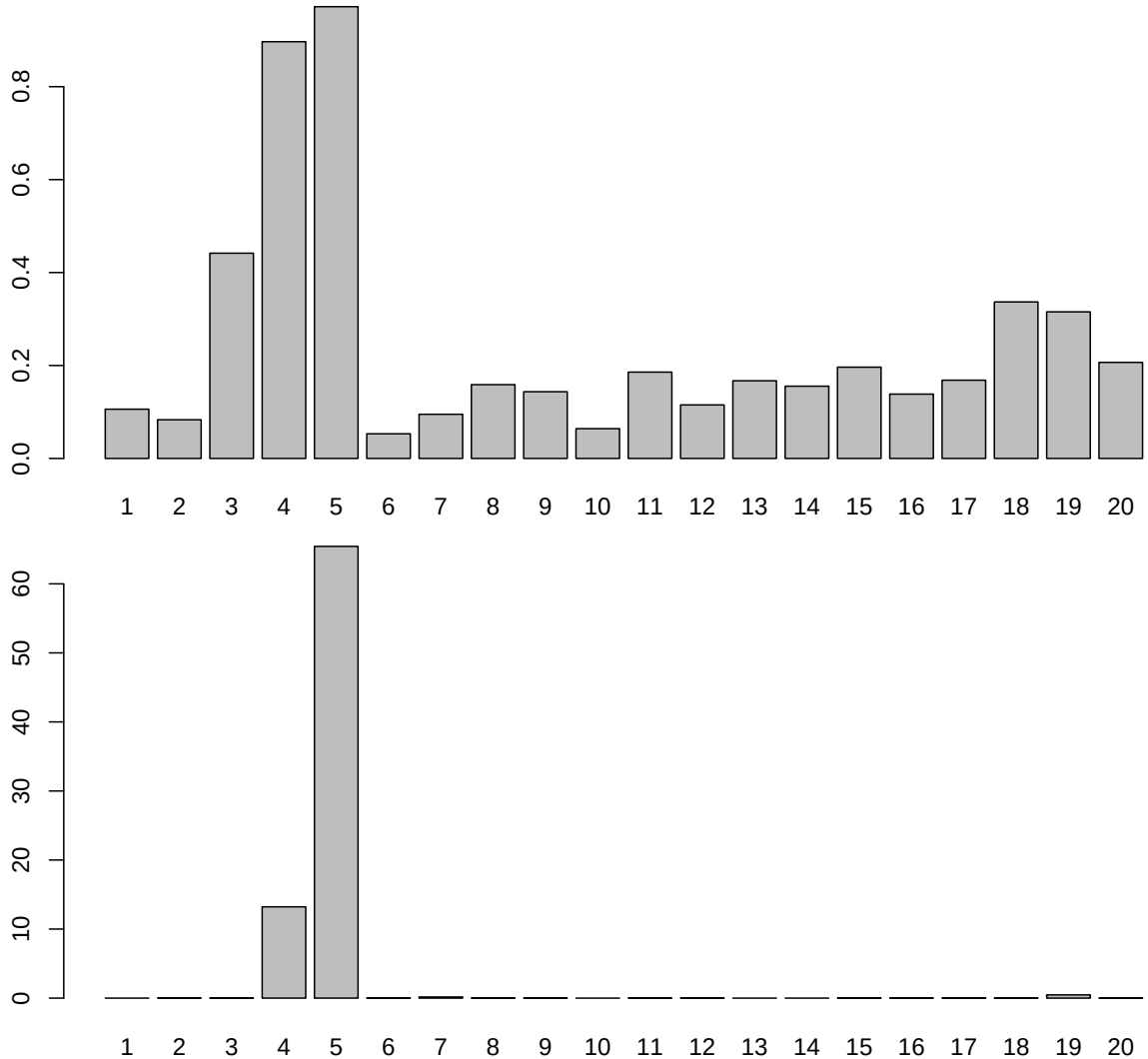
$$D_i = \frac{1}{k+1} r_i^2 \frac{h_{ii}}{1-h_{ii}},$$

donde $r_i = e_i/(s_R\sqrt{1-h_{ii}})$ son los residuos estandarizados.



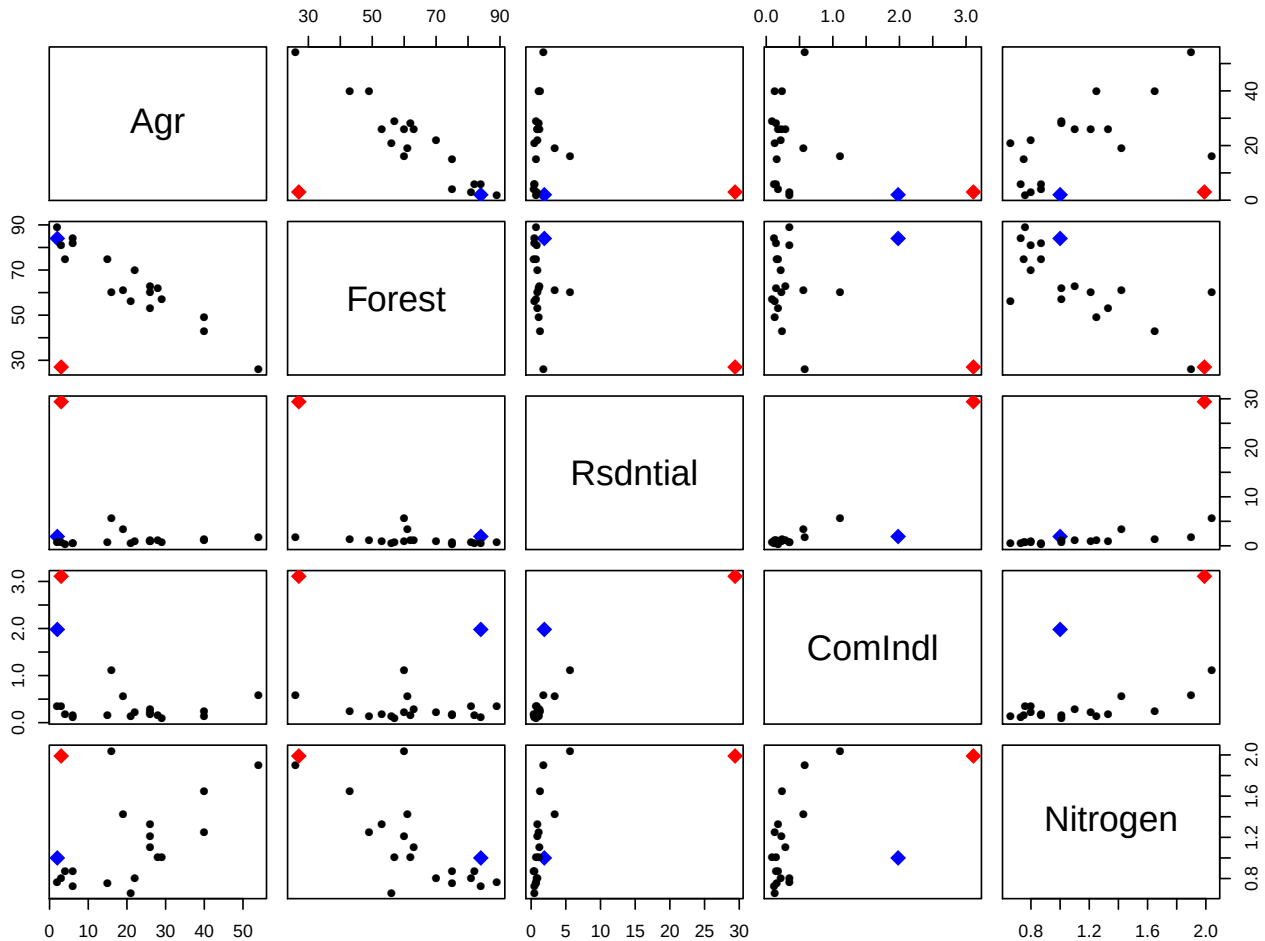
Ejemplo (cuencas fluviales): Potencial h_{ii} e influencia D_i

```
potencial <- hatvalues(reg)
cook <- cooks.distance(reg)
barplot(potencial, col="blue", xlab="Potencial")
barplot(cook, col="blue", xlab="Distancia de Cook")
```



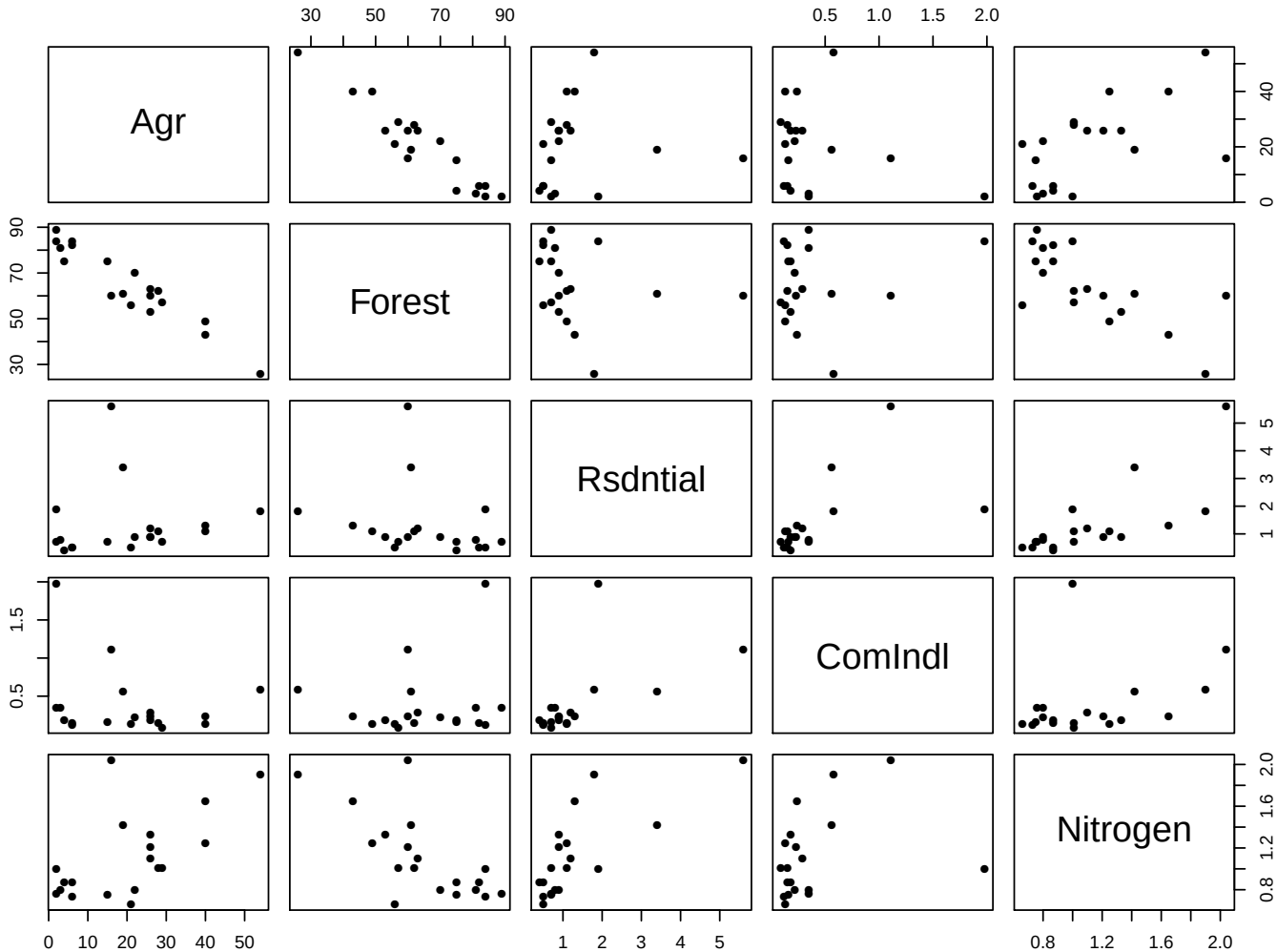
Ejemplo (cuencas fluviales):

```
pairs(Datos[,-1],oma=c(2,2,2,2),
      col=c(rep("black",3),"blue","red",rep("black",15)),
      pch=c(rep(16,3),18,18,rep(16,15)),
      cex=c(rep(1,3),2,2,rep(1,15)))
```



Ejemplo (cuencas fluviales): Resultados sin la observación 5

```
pairs(Datos[-5,-1],pch=16)
```



Ejemplo (cuencas fluviales):

```
regSin5 = lm(Nitrogen~Agr+Forest+Rsdntial+ComIndl,data=Datos[-5,])
summary(regSin5)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.626014   0.781091   2.082  0.05620 .
Agr          0.002352   0.009539   0.247  0.80881
Forest      -0.012760   0.008815  -1.448  0.16976
Rsdntial     0.181161   0.044390   4.081  0.00112 **
ComIndl      0.075618   0.113957   0.664  0.51775
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1676 on 14 degrees of freedom
Multiple R-squared:  0.864,     Adjusted R-squared:  0.8252
F-statistic: 22.24 on 4 and 14 DF,  p-value: 6.055e-06

cor(Datos[-5,c(2,3)])

               Agr     Forest
Agr     1.0000000 -0.9490166
Forest -0.9490166  1.0000000
```

Ejemplo (cuencas fluviales):

```
regSin5SinAgr = lm(Nitrogen~Forest+Rsdntial+ComIndl,data=Datos[-5,])
summary(regSin5SinAgr)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.812896   0.183000   9.907 5.65e-08 ***
Forest      -0.014831   0.002596  -5.712 4.11e-05 ***
Rsdntial     0.177784   0.040881   4.349 0.000573 ***
ComIndl      0.073646   0.110060   0.669 0.513571
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1623 on 15 degrees of freedom
Multiple R-squared:  0.8634,    Adjusted R-squared:  0.8361
F-statistic: 31.61 on 3 and 15 DF,  p-value: 9.948e-07
```

Variable regresora cualitativa: modelo unifactorial

Ejemplo (fertilizantes): 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:

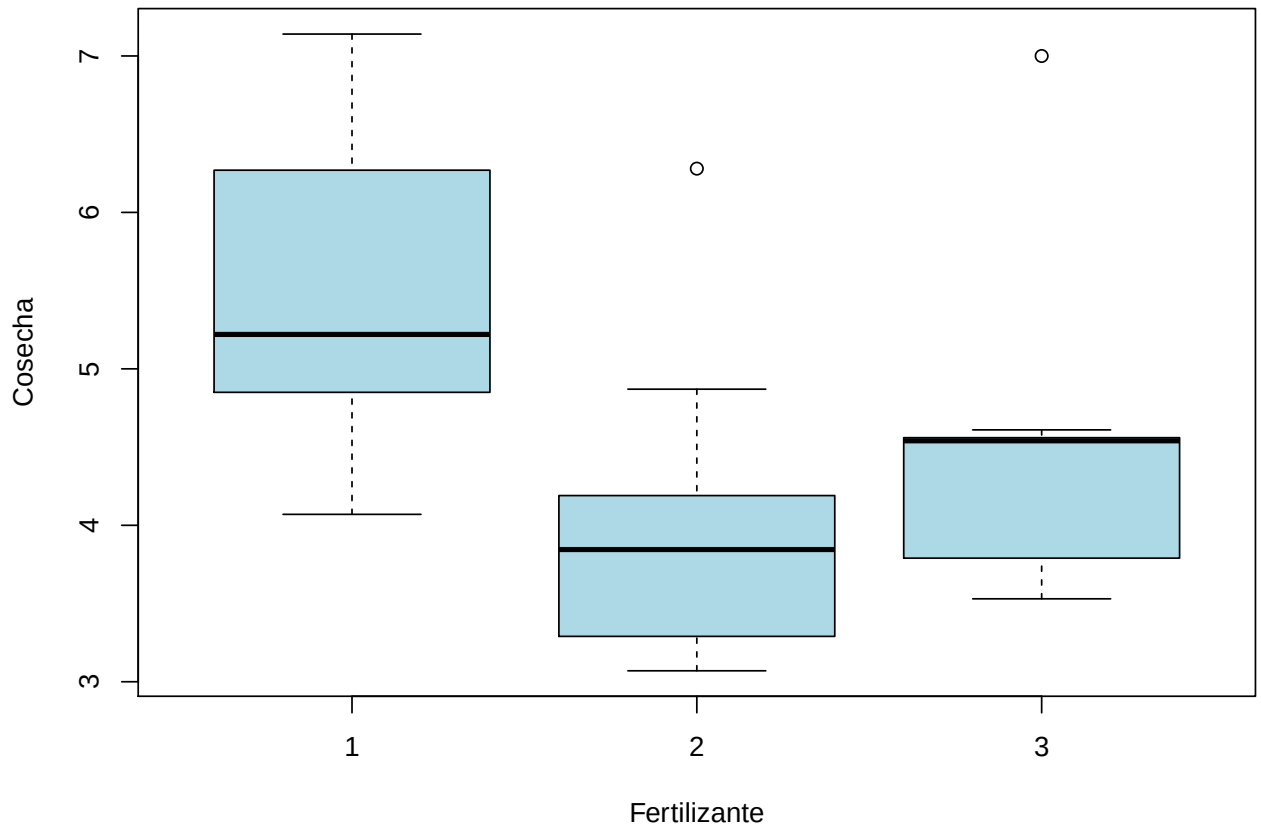
Fert. 1	6.27	5.36	6.39	4.85	5.99	7.14	5.08	4.07	4.35	4.95
Fert. 2	3.07	3.29	4.04	4.19	3.41	3.75	4.87	3.94	6.28	3.15
Fert. 3	4.04	3.79	4.56	4.55	4.55	4.53	3.53	3.71	7.00	4.61

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).

Ejemplo (fertilizante):

```
ej.cosecha = read.table("cosecha.txt", header=TRUE)
boxplot(ej.cosecha$cosecha~ej.cosecha$fertilizante,
        col="lightblue",xlab="Fertilizante",
        ylab="Cosecha")
```



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}	$\bar{y}_{1\bullet}$	s_1
2	y_{21}	y_{22}	$\cdots$	y_{2n_2}	$\bar{y}_{2\bullet}$	s_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
k	y_{k1}	y_{k2}	$\cdots$	y_{kn_k}	$\bar{y}_{k\bullet}$	s_k

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

Muestra	n_i	$\bar{y}_{i\bullet}$	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,

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

- β_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).

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,

$$Y_{ij} \sim \mathrm{N}(\beta_i, \sigma^2).$$

Observemos que el modelo unifactorial (1) se puede expresar como un modelo de regresión $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, donde

- $\mathbf{Y} = (Y_{11}, Y_{12}, \ldots, Y_{kn_k})'$
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$
- $\boldsymbol{\epsilon} = (\epsilon_{11}, \epsilon_{12}, \ldots, \epsilon_{kn_k})'$

¿Cuál es la matriz de diseño $\mathbb{X}$?

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

La respuesta prevista con el modelo es $\hat{y}_{ij} = \hat{\beta}_i = y_{i\bullet} = \dfrac{1}{n_i}\displaystyle\sum_{j=1}^{n_i} y_{ij}$.

El residuo es $e_{ij} = y_{ij} - \hat{y}_{ij}$.

Una estimación insesgada de σ^2 es la varianza residual

$s_R^2 = \dfrac{1}{n-k}\text{RSS}$, siendo $\text{RSS} = \displaystyle\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2$.

Los g.l. de los residuos son $n-k$ porque $\sum_{j=1}^{n_i} e_{ij} = 0$, para $i = 1, \ldots, k$.

Formulación equivalente del modelo (1): $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, donde $\sum_{i=1}^{k} \alpha_i = 0$.

μ es el valor esperado de la respuesta Y sin tener en cuenta el efecto del factor.

α_i es el efecto *incremental* del factor i sobre el nivel medio de la respuesta, es decir, la parte de la respuesta esperada (con respecto al nivel global μ) que es debida a que el nivel del factor es i.

La estimación de los parámetros por mínimos cuadrados es:

$$\hat{\mu} = y_{\bullet\bullet} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \qquad \hat{\alpha}_i = y_{i\bullet} - y_{\bullet\bullet}$$

Contraste de igualdad de medias $H_0 : \beta_1 = \cdots = \beta_k$

Modelo reducido:

$$\text{RSS}_0 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2 = \text{TSS} \qquad \text{con} \qquad y_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}$$

Modelo completo:

$$\text{RSS} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2 = \sum_{i=1}^{k}(n_i - 1)s_i^2$$

Incremento de variabilidad:

$$\begin{aligned}\text{RSS}_0 - \text{RSS} &= \text{TSS} - \text{RSS} = \text{MSS} \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2.\end{aligned}$$

Tabla ANOVA del modelo unifactorial

Fuente de variación	Suma de cuadrados	gl	Cuadrados medios	Estadístico
Explicada	MSS	$k-1$	$\frac{\text{MSS}}{k-1}$	$F = \frac{\text{MSS/(k-1)}}{\text{RSS}/(n-k)}$
Residual	RSS	$n-k$	$s_R^2 = \frac{\text{RSS}}{n-k}$	

La región de rechazo de $H_0 : \beta_1 = \cdots = \beta_k$ es $R = \{F > F_{k-1,n-k;\alpha}\}$.

Ejemplo (fertilizante): Tabla ANOVA y contraste

```
# La variable fertilizante debe ser un factor
fertilizante = factor(ej.cosecha$fertilizante)
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
```

Ejemplo (fertilizante): Tabla ANOVA y contraste haciendo una regresión con un regresor tipo factor

```
mod0 = lm(cosecha~1,data=ej.cosecha)
modC = lm(cosecha~fertilizante-1,data=ej.cosecha)
# -1 para quitar término independiente
anova(mod0,modC)

Analysis of Variance Table

Model 1: cosecha ~ 1
Model 2: cosecha ~ factor(fertilizante) - 1
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     29 36.445
2     27 25.622  2    10.823 5.7024 0.008594 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hemos visto que el modelo de análisis de la varianza con un factor equivale a un modelo de regresión sin término independiente β_0 y con tantos regresores binarios (0-1, indicando ausencia-presencia del nivel del factor) como niveles del factor. A estos regresores se les llama *variables ficticias* o *dummy variables*.

Cuando el modelo de regresión utiliza tanto regresores cuantitativos como cualitativos (factores), entonces el modelo y su estimación se complican de manera evidente por la codificación del factor en el modelo mediante variables ficticias y por la posible interacción entre factor y regresores.

En la práctica actual es frecuente utilizar como método de predicción los árboles de regresión y versiones *ensemble* de los mismos (*bagging*, *boosting*, *random forests*,...). Estas técnicas están explicadas, por ejemplo, en Izenman (2008).

Referencias

Cook, R.D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics*, 19, 15–18.

Izenman, A.J. (2008). *Modern Multivariate Statistical Techniques. Regression, Classification, and Manifold Learning*. Springer.

Marecek, V., Janchenova, H., Brezina, M. (1991). *Anal. Chim. Acta*, 244, 15–19.

Peña, D. (2002). *Regresión y diseño de experimentos*. Alianza.
- Cap. 2: "El análisis de la varianza"
- Cap. 5: "El modelo de regresión simple"
- Cap. 6: "Diagnosis y predicción en el modelo de regresión lineal simple"
- Cap. 7: "El modelo general de regresión"

Rencher, A.C., Schaalje, G.B. (2008). *Linear Models in Statistics*. Wiley.
- Cap. 6: "Simple Linear Regression"
- Cap. 7: "Multiple Regression: Estimation"

Weisberg, S. (2005). *Applied Linear Regression*. Wiley.
- Cap. 9: "Outliers and Influence"