Modelos Lineales Generalizados: aplicaciones en R

Modelos Lineales Generalizados: un enfoque aplicado

Ana M. Bianco Jemina García
(anambianco@gmail.com) (jeminagarcia@gmail.com)

1. Modelo Lineal

Modelización y Predicción

Algunos de los métodos estadísticos más extendidos se ocupan de la modelización de datos y de la predicción.

Muchas de estas técnicas estadísticas se encuadran en lo que hoy se conoce como aprendizaje estadístico (AE).

El AE abarca una vasta cantidad de procedimientos que ayudan a comprender los datos cuando se analizan varias variables al mismo tiempo, ya sea postulando modelos o encontrando relaciones entre las variables o estructuras que ayudan a su comprensión.

Modelizació y Predicción

Los métodos de AE pueden reunirse en dos grandes grupos:

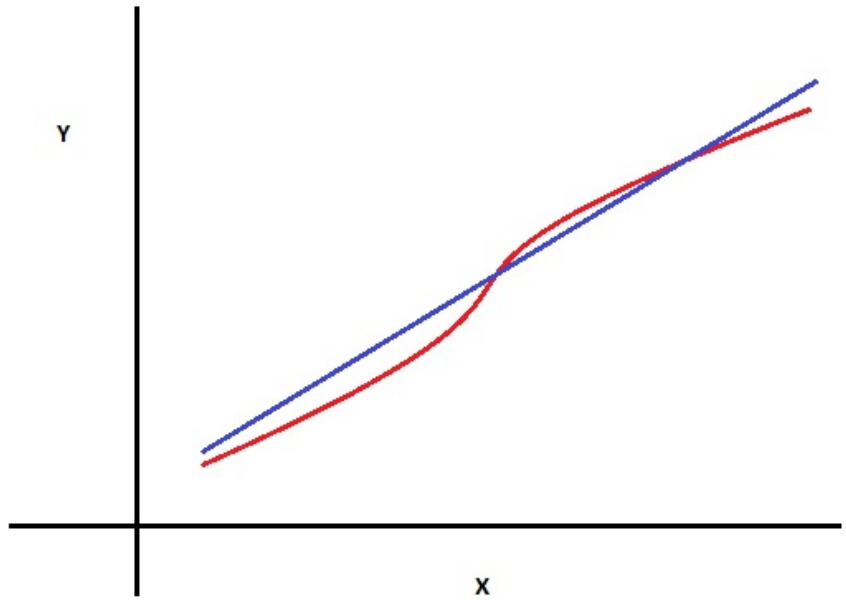
- Aprendizaje Supervisado: Aquí una de las variables es identificada como una respuesta

- Aprendizaje No Supervisado: todas las variables cumplen un rol análogo.

El modelo lineal puede pensarse como la aproximación simple al aprendizaje supervisado.

Un modelo muy difundido de ajuste y predicción, que puede verse como es un caso particular de GLM.

Modelo Lineal: aproximación



En general, la verdadera relación es más compleja!!

Aprendizaje Estadístico

Algunos ejemplos:

- Predecir si un paciente hospitalizado tendrá un segundo infarto de miocardio o no teniendo en cuenta mediciones clínicas, dietas y variables demográicas.
- Predecir los precios que tendrán en 6 meses las acciones de ciertas compañías a partir de mediciones del rendimiento de las compañías y datos económicos.
- Estimar la cantidad de glucosa en sangre que tendrá un individuo diabético a partir del espectro de adsorción infra-rojo de la sangre.
- Identificar los factores de riesgo de cáncer de próstata, usando mediciones clínicas y variables demográficas.

Ganancia de un transistor

En un estudio se miden $y =$ ganancia de un transistor en un circuito integrado entre el emisor y el receptor (hFE) junto con dos variables que pueden ser controladas. Se realizaron 14 medidiones.

- $x_1 =$ tiempo de conducción en minutos
- $x_2 =$ iones $\times 10^{14}$

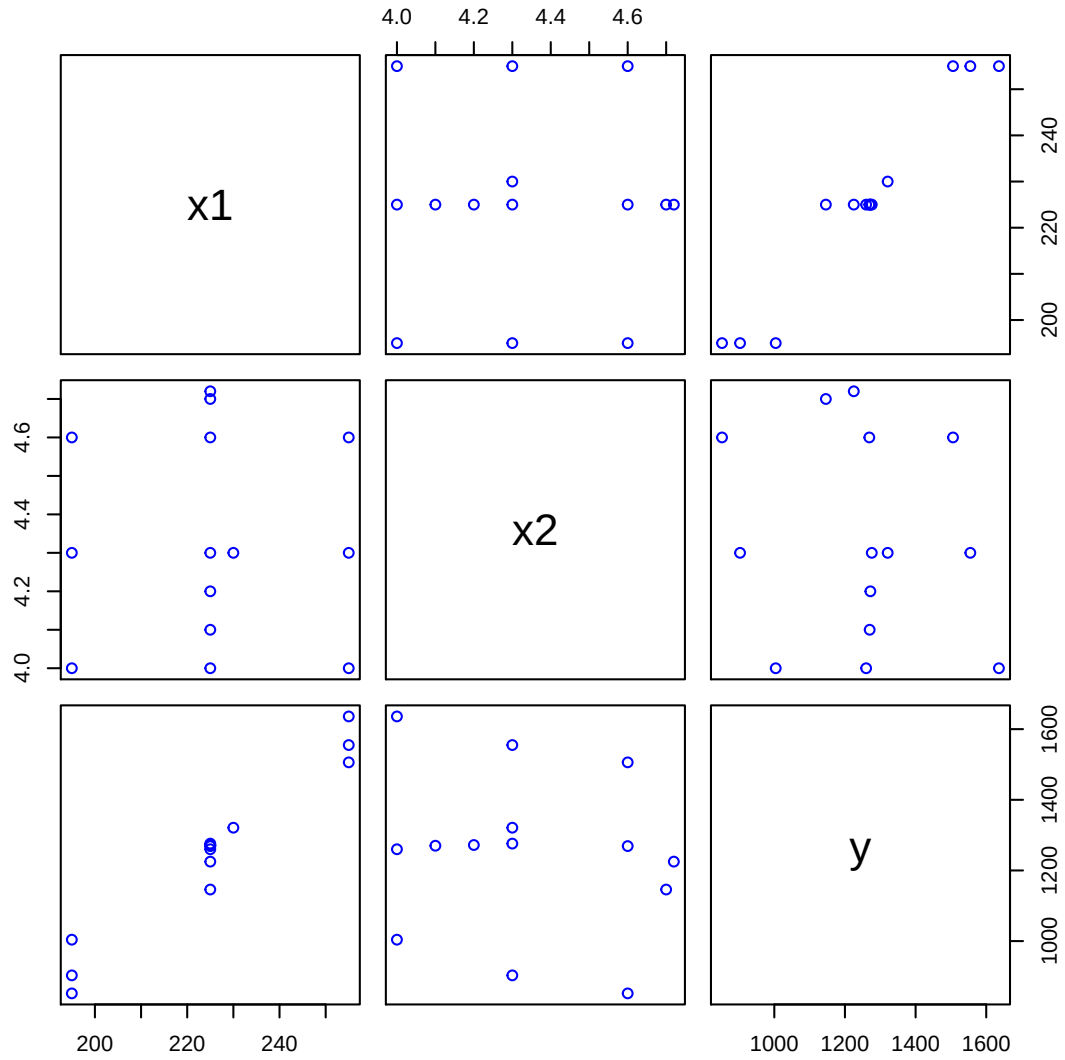
Ganancia de un transistor

```
> iones<- read.table("iones.txt",header=T)
> iones<- iones[,2:4]
> attach(iones)
> iones
```

```
    x1   x2    y
1  195 4.00 1004
2  255 4.00 1636
3  195 4.60  852
4  255 4.60 1506
5  225 4.20 1272
6  225 4.10 1270
7  225 4.60 1269
8  195 4.30  903
9  255 4.30 1555
10 225 4.00 1260
11 225 4.70 1146
12 225 4.30 1276
13 225 4.72 1225
14 230 4.30 1321
```

Ganancia de un transistor

```
> pairs(iones,col="blue")
```



Datos de Prostata

En 97 pacientes que van a tener una prostatectomía radical se miden las siguientes variables:

- $x_1 =$ **lcavol**: log del volumen del tumor
- $x_2 =$ **lweight**: log del peso de la próstata
- $x_3 =$ **age**: edad
- $x_4 =$ **lbph**: log de la cantidad de hiperplasia prostática benigna
- $x_5 =$ **svi**: invasión seminal (si o no)
- $x_6 =$ **lcp**: logaritmo de la penetración capsular
- $x_7 =$ **gleason**: score de Gleason
- $x_8 =$ **pgg46**: porcentaje de scores de Gleason 4 or 5.
- $x_9 =$ **lpsa**: log del PSA (PSA: Antígeno prostático)

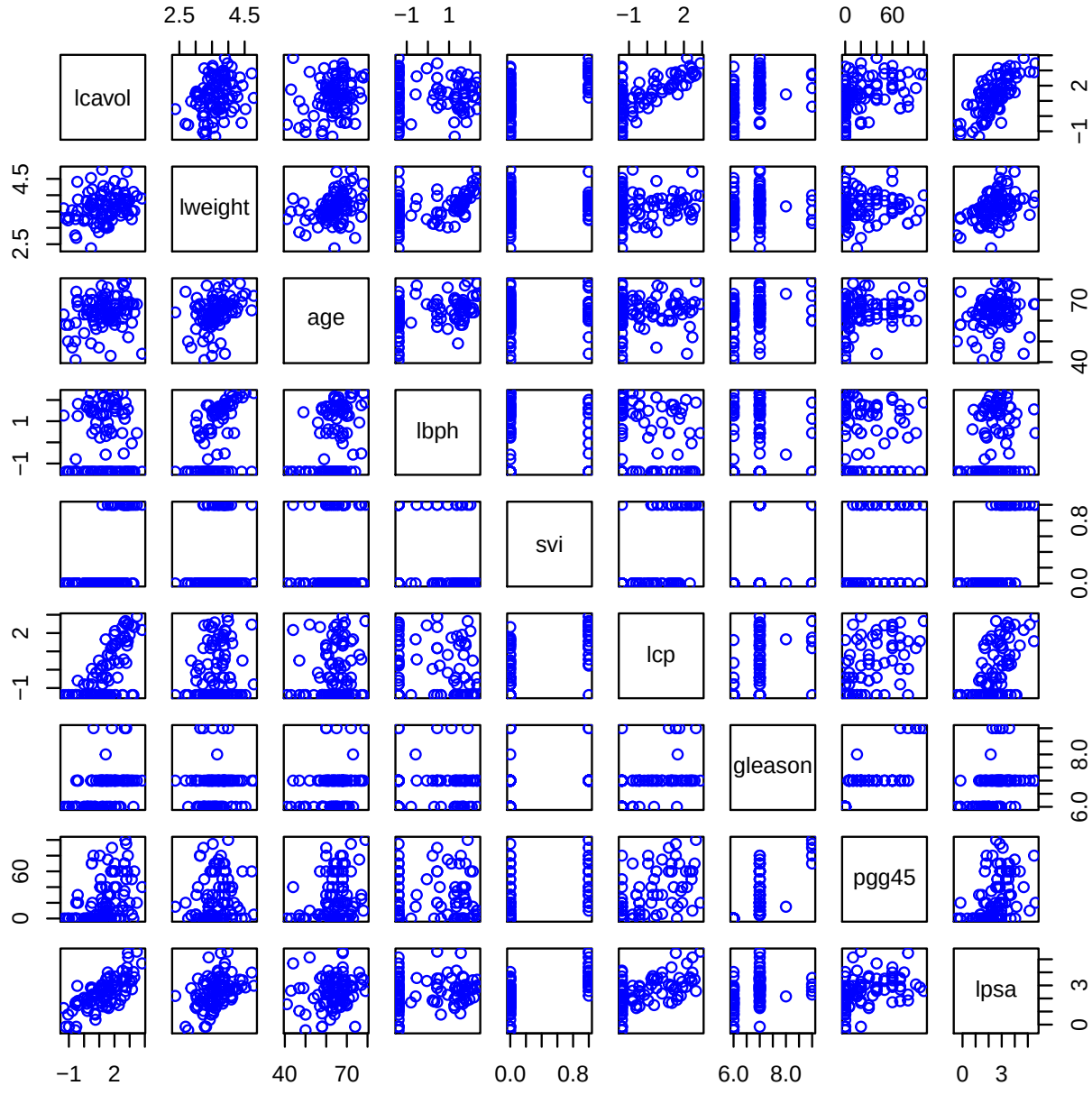
El objetivo es predecir el log del PSA.

Datos de Prostata: Diagramas de Dispersión

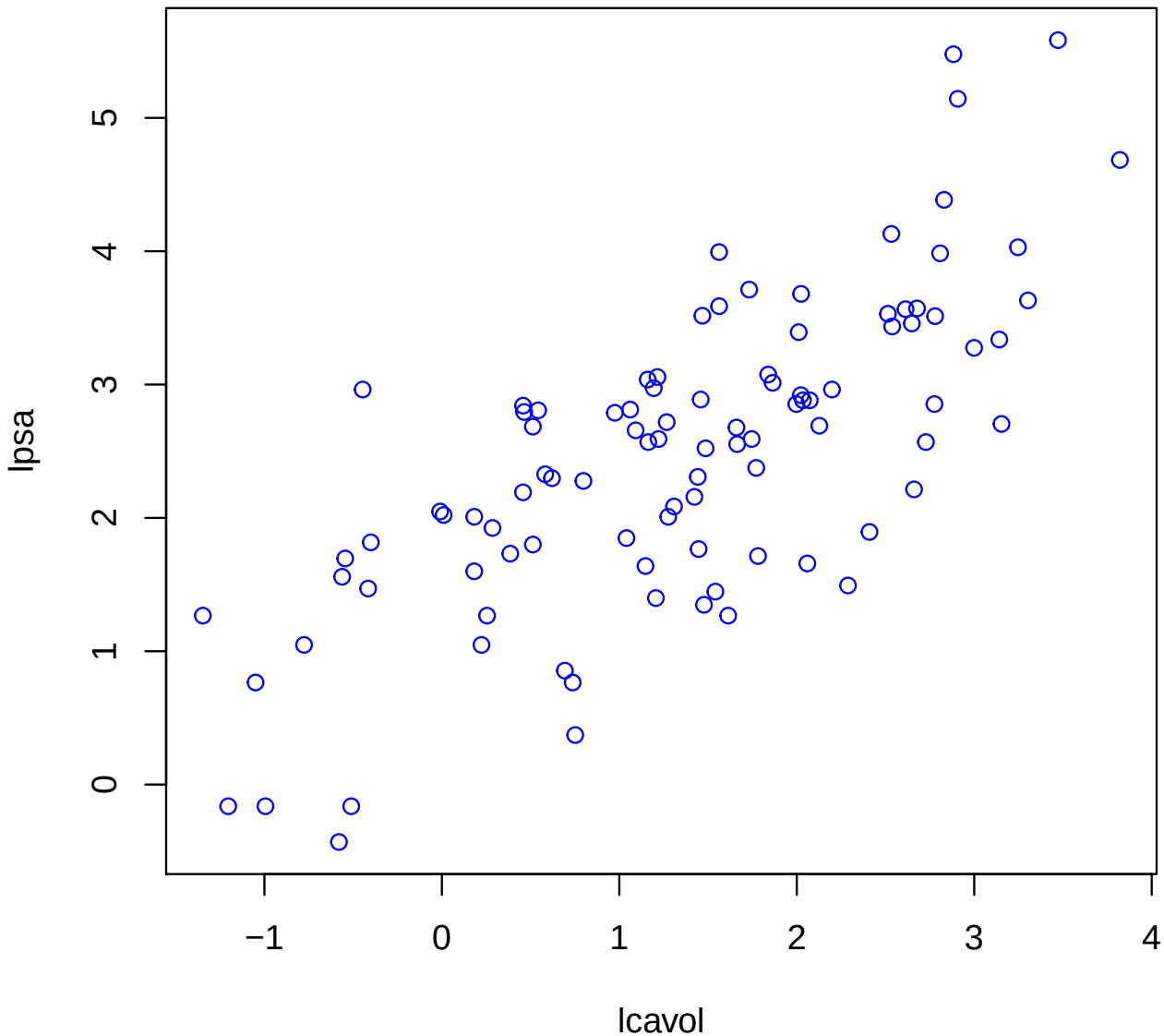
```
> setwd("C:\\Users\\Ana\\Dropbox\\Ana\\GLM\\2019\\Doctex")
> datos.ori<- read.table("prostate.txt",header=T)
> datos<- datos.ori[,1:9]
> attach(datos)
> print(datos[1:6,],digits=3)
```

```
  lcavol lweight age  lbph svi   lcp gleason pgg45   lpsa
1 -0.580    2.77  50 -1.39   0 -1.39       6     0 -0.431
2 -0.994    3.32  58 -1.39   0 -1.39       6     0 -0.163
3 -0.511    2.69  74 -1.39   0 -1.39       7    20 -0.163
4 -1.204    3.28  58 -1.39   0 -1.39       6     0 -0.163
5  0.751    3.43  62 -1.39   0 -1.39       6     0  0.372
6 -1.050    3.23  50 -1.39   0 -1.39       6     0  0.765
```

```
> pairs(datos,col="blue")
```



```
> plot(lcavol,lpsa,col="blue")
```



Modelo Lineal Simple

En regresión lineal interesa establecer la relación entre una variable dependiente y y otras p variables: $x_1, \ldots, x_p$ explicativas.

El caso más sencillo es cuando tenemos una sola variable esplicativa:

$$y = \beta_0 + \beta_1 x + \epsilon$$

donde

- β_0: oredenada al origen o intercept
- β_1: pendiente
- ϵ: error

Modelo Lineal Simple

En regresión lineal interesa establecer la relación entre una variable dependiente y y otras p variables: $x_1, \ldots, x_p$ explicativas.

El caso más sencillo es cuando tenemos una sola variable esplicativa:

$$y = \beta_0 + \beta_1 x + \epsilon$$

donde

- β_0: oredenada al origen o intercept
- β_1: pendiente
- ϵ: error

El modelo es lineal en los parámetros β_0 y β_1. De manera que

$$y = \beta_0 + \beta_1 x^2 + \epsilon$$

también es un modelo lineal.

Modelo Lineal Simple

β_0, β_1 son 2 parámetros desconocidos a estimar.

Para ello suponemos que disponemos de una muestra $(y_1, x_1), \ldots, (y_n, x_n)$ independientes tales que

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

donde

- $E(\epsilon) = 0$
- $V(\epsilon) = \sigma^2$
- ϵ_i son no correlacionados: $Cov(\epsilon_i, \epsilon_j) = 0$ si $i \neq j$

¿Cómo estimamos los parámetros?

Mínimos Cuadrados

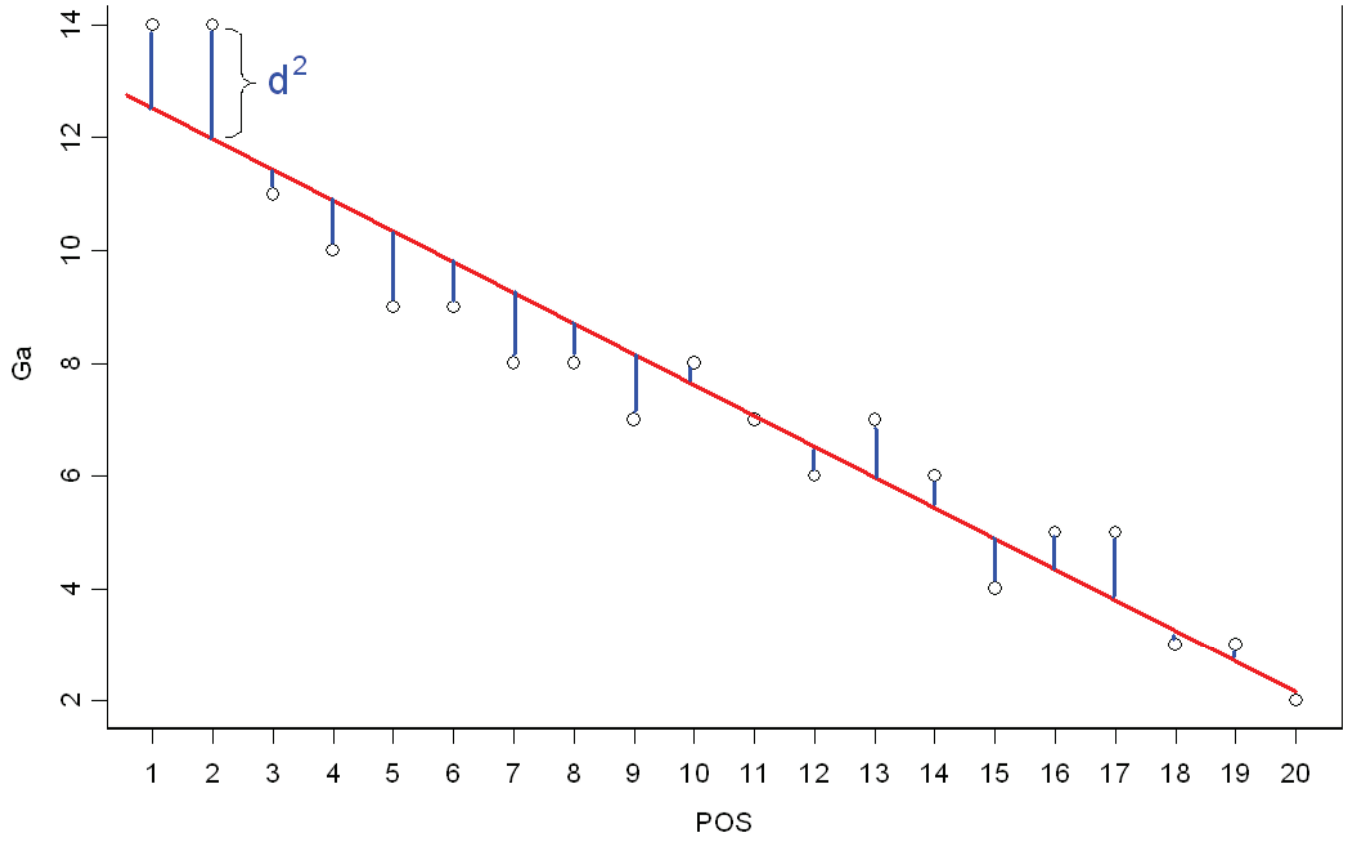
Si los puntos en un gráfico parecen seguir una recta, el problema es elegir la recta que mejor ajusta los puntos.

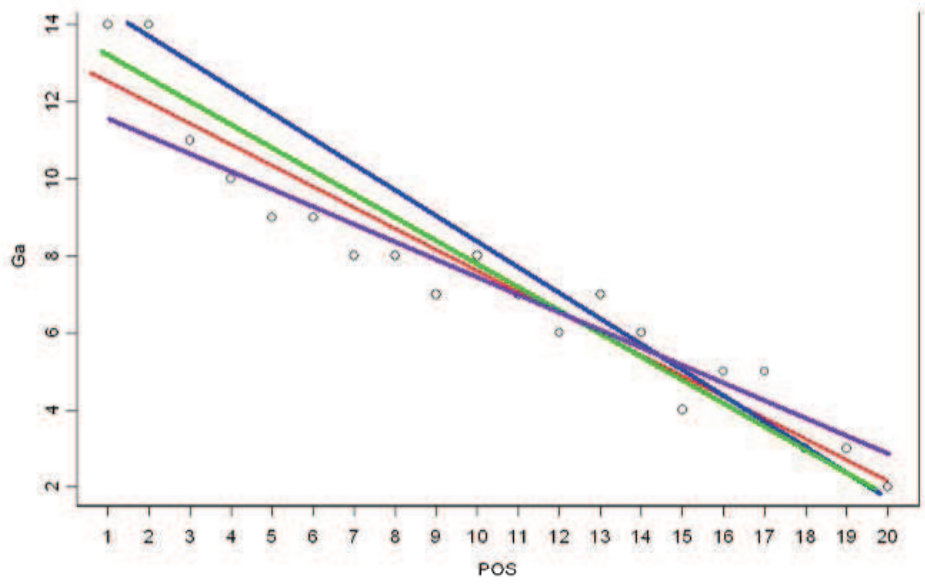
Tendremos en cuenta:

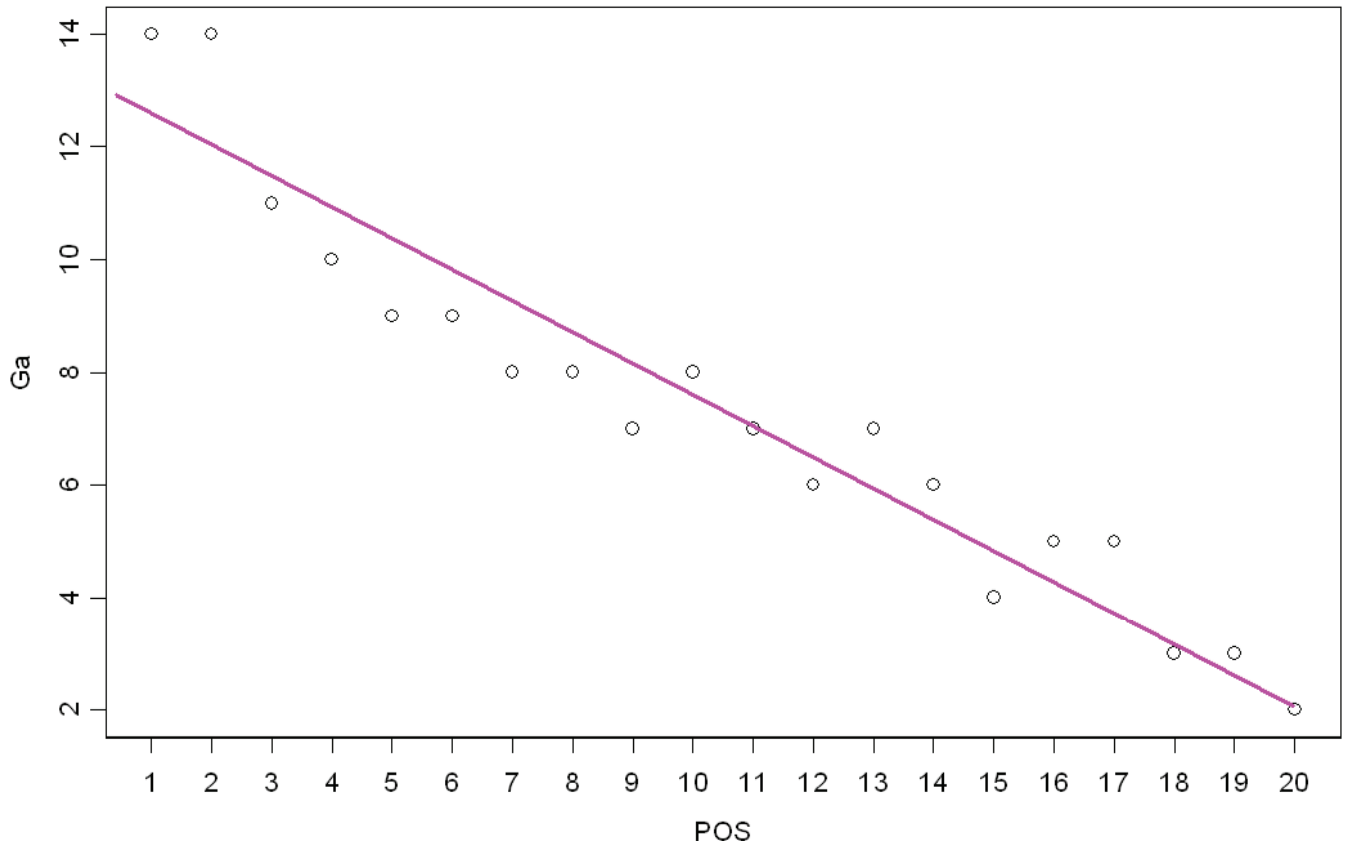
a) tomar una distancia promedio de la recta a todos los puntos

b) mover la recta hasta que esta distancia promedio sea la menor posible.

Si tenemos(y_i, x_i), $1 \leq i \leq n$, y, en forma genérica, queremos predecir la respuesta y a partir de la covariable x usando una recta, podríamos definir el error cometido en cada punto como la distancia vertical del punto a la recta.


d2
Ga
POS
2
4
6
8
10
12
14
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20


Ga
POS
2
4
6
8
10
12
14
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20


Ga
14
12
10
8
6
4
2
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
POS

Estimadores de mínimos cuadrados

$$\min_{b_0,b_1} \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

Si definimos

$$\mathcal{S}(b_0, b_1) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

para hallar el mínimo debemos derivar el igualar a 0:

$$\begin{aligned} \frac{\partial \mathcal{S}}{\partial b_0}(b_0, b_1) &= -2 \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i)) = 0 \\ \frac{\partial \mathcal{S}}{\partial b_1}(b_0, b_1) &= -2 \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i)) x_i = 0 \end{aligned}$$

Estimadores de mínimos cuadrados

$$\min_{b_0,b_1} \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

Si definimos

$$\mathcal{S}(b_0, b_1) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

Desarrolando quedan las **ecuaciones normales**

$$\begin{aligned} n\widehat{\beta}_0 + \left(\sum_{i=1}^{n} x_i\right) \widehat{\beta}_1 &= \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right) \widehat{\beta}_0 + \left(\sum_{i=1}^{n} x_i^2\right) \widehat{\beta}_1 &= \sum_{i=1}^{n} y_i x_i \end{aligned}$$

Estimadores de mínimos cuadrados

Entonces

$$\begin{aligned}
\widehat{\beta}_0 &= \overline{y} - \overline{x}\widehat{\beta}_1 \\
\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}
\end{aligned}$$

Notacion Matricial

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad 1 \leq i \leq n$$

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ \dots & \quad \dots \\ y_n &= \beta_0 + \beta_1 x_n + \epsilon_n \end{aligned}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{pmatrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

Notacion Matricial

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad 1 \leq i \leq n$$

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{pmatrix}$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

Ecuaciones Normales

Las 2 ecuaciones normales pueden escribirse como

$$\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$$

Cuando $\mathbf{X}'\mathbf{X}$ es no singular, la solución es única y resulta

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Para tener en cuenta...

En el caso de regresión simple tendríamos

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ x_1 & x_2 & x_3 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}$$

El sistema sería

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

Para tener en cuenta...

La inversa resulta

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n\sum_{i=1}^{n}(x_i - \overline{x})^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix}$$

y además

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

y por lo tanto

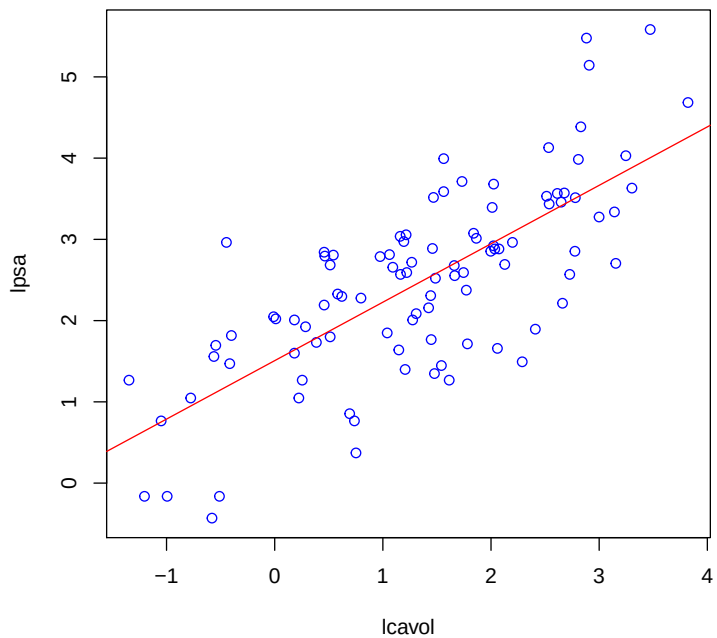
$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = \frac{1}{n\sum_{i=1}^{n}(x_i - \overline{x})^2} \begin{pmatrix} (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i y_i) \\ n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i) \end{pmatrix}$$

Datos de Prostata: lcavol vs. lpsa

```
> salida<- lm(lpsa~lcavol)


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.50858    0.19419  -2.619   0.0103 *
lpsa         0.74992    0.07109  10.548   <2e-16 ***
---
NA

Residual standard error: 0.8041 on 95 degrees of freedom
Multiple R-squared:  0.5394, Adjusted R-squared:  0.5346
F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16
```

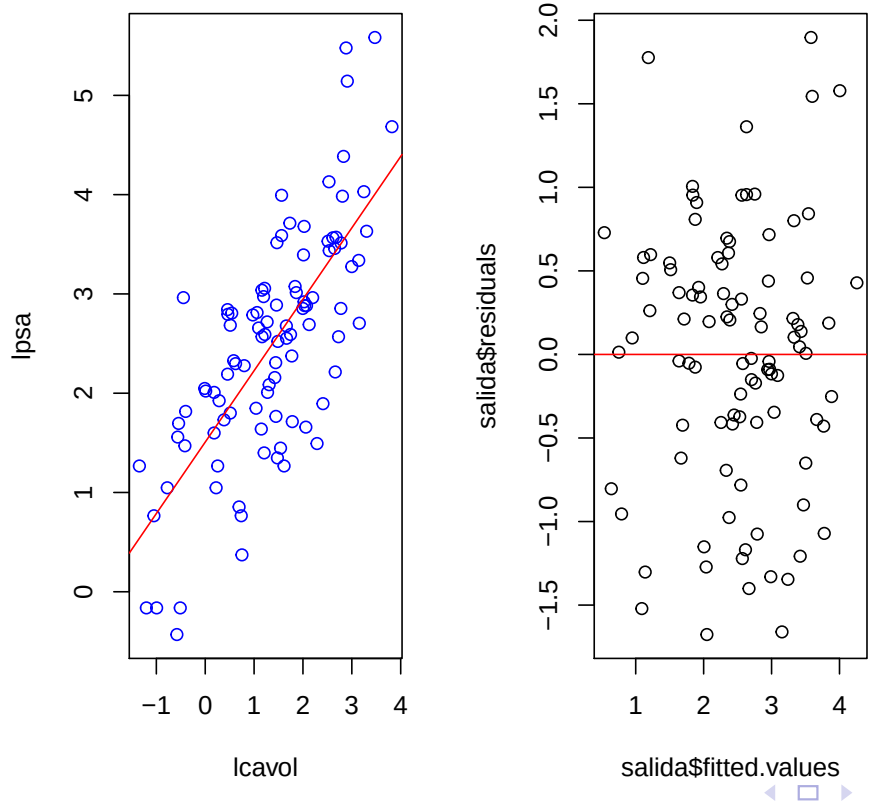
Datos de Prostata: lpsa vs. lcavol

```
> salida$coefficients

(Intercept)      lcavol
  1.5072975   0.7193204

>   plot(lcavol,lpsa,col="blue")
>   abline(lm(lpsa~lcavol),col="red")
```



Datos de Prostata: lcavol vs. lpsa

```
> par(mfrow=c(1,2))
> plot(lcavol,lpsa,col="blue")
> abline(lm(lpsa~lcavol),col="red")
> plot(salida$fitted.values,salida$residuals)
> abline(h=0,col="red")
```



Predichos y Residuos

- i-ésimo valor predicho $\quad \widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$
- i-ésimo residuo $\quad r_i = y_i - \widehat{y}_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)$

Explicando la Variabilidad

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 \\ \sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n} r_i^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 \end{aligned}$$

Sumas de cuadrados

Explicando la Variabilidad

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2.$$

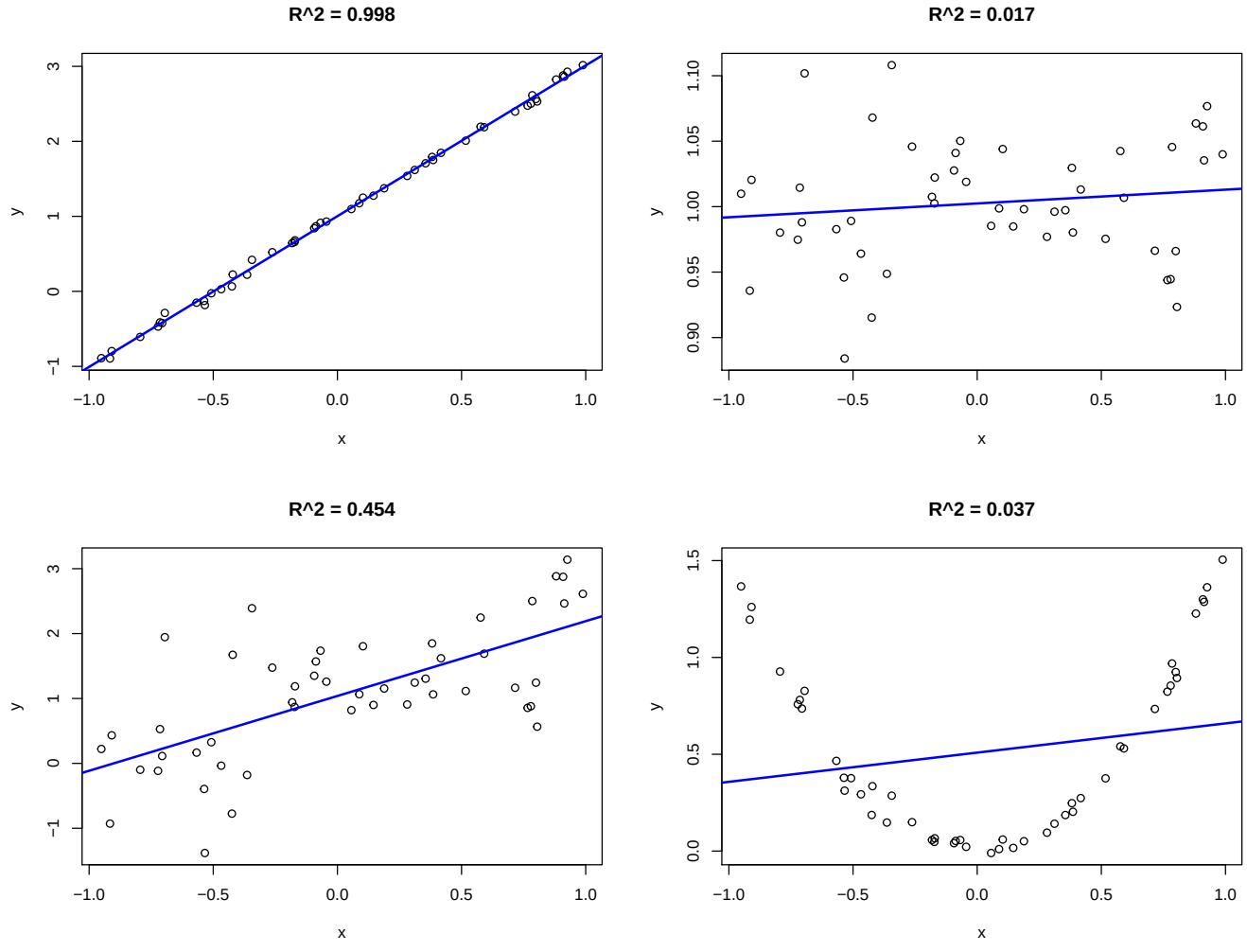
Puede escribir como SCT = SCE + SCR donde

- $SCT = \sum_{i=1}^{n}(y_i - \overline{y})^2$ variabilidad total: variabilidad total de la respuesta **Y**
- $SCE = \sum_{i=1}^{n} r_i^2$ variabilidad residual: variabilidad de **Y** no explicada por el modelo de regresión.
- $SCR = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$ variabilidad de la regresión: variabilidad de **Y** explicada por el modelo de regresión

Coeficiente R^2

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

- $0 \leq R^2 \leq 1$
- Es la proporción de variación explicada por la regresión
- $R^2 \approx 1 \Rightarrow$ el modelo explica muy bien
- $R^2 \approx 0 \Rightarrow$ el modelo provee un ajuste pobre


R^2 = 0.998
R^2 = 0.017
R^2 = 0.454
R^2 = 0.037
x
y

Caso de más covariables

La notación matricial es muy adecuada para tratar el caso más general.

respuesta $y \longleftrightarrow p$ variables explicativas x_j

Supondremos $x_j, 1 \leq j \leq p$ son determinísticas.

Muestra $(x_{i1}, \ldots, x_{ip}, y_i)$, $1 \leq i \leq n$ que cumplen el modelo

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \quad i = 1, \ldots, n \\ E(\epsilon_i) &= 0 \\ V(\epsilon_i) &= \sigma^2 \\ cov(\varepsilon_i, \varepsilon_j) &= 0 \quad i \neq j \end{aligned}$$

donde, $\beta_0, \beta_1, \ldots, \beta_p$ son $p+1$ parámetros desconocidos a estimar

Caso general

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p} + \epsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \cdots + \beta_p x_{2p} + \epsilon_2 \\
\cdots &\quad \cdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_p x_{np} + \epsilon_n
\end{aligned}
$$

El estimador de mínimos cuadrados de $\beta_0, \beta_1 \ldots \beta_p$ minimiza

$$
\mathcal{S}(b_0, b1, \ldots, b_p) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}))^2
$$

Caso general

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p} + \epsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \cdots + \beta_p x_{2p} + \epsilon_2 \\
\cdots & \quad \cdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_p x_{np} + \epsilon_n
\end{aligned}
$$

$$
\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ & \cdots & & \cdots & \\ & \cdots & & \cdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{pmatrix}
$$

$$
\Downarrow
$$

$$
\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon
$$

Caso general

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \quad i = 1, \ldots, n$$

En el caso general tenemos

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ & \ldots & & \ldots & \\ & \ldots & & \ldots & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_p \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{pmatrix}$$

$$\Downarrow$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

Caso general

El estimador de mínimos cuadrados de $\beta_0, \beta_1 \ldots \beta_p$ minimiza

$$\mathcal{S}(b_0, b1, \ldots, b_p) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}))^2$$

Derivando e igualando a 0 obtenemos las **ecuaciones normales**.

Los estimadores de mínimos cuadrados $\widehat{\beta}_0, \ldots, \widehat{\beta}_p$ cumplen:

$$\begin{aligned} \frac{\partial \mathcal{S}(\mathbf{b})}{\partial b_0} &= -2 \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \mathbf{b}) = 0 \\ \frac{\partial \mathcal{S}(\mathbf{b})}{\partial b_k} &= -2 \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \mathbf{b}) x_{ik} = 0 \quad k = 1, \ldots, p \end{aligned}$$

siendo $\mathbf{x}_i' = (1, x_{i1}, \ldots, x_{ip})$ a i-ésima fila de la matriz de diseño $\mathbf{X}$

Ecuaciones Normales

$\widehat{\beta}_0, \ldots, \widehat{\beta}_p$, los estimadores de mínimos cuadrados cumplen:

$$
\begin{aligned}
n\widehat{\beta}_0 + \widehat{\beta}_1 \left(\sum_{i=1}^{n} x_{i1} \right) + \cdots + \widehat{\beta}_p \left(\sum_{i=1}^{n} x_{ip} \right) &= \sum_{i=1}^{n} y_i \\
\widehat{\beta}_0 \left(\sum_{i=1}^{n} x_{ik} \right) + \widehat{\beta}_1 \left(\sum_{i=1}^{n} x_{i1} x_{ik} \right) + \cdots + \widehat{\beta}_p \left(\sum_{i=1}^{n} x_{ip} x_{ik} \right) &= \sum_{i=1}^{n} y_i x_{ik} \;\; k = 1, \ldots, p
\end{aligned}
$$

Ecuaciones Normales

Estas $p+1$ ecuaciones pueden escribirse como

$$\mathbf{X}'\mathbf{X}\widehat{\beta} = \mathbf{X}'\mathbf{Y}$$

que se conocen como ecuaciones normales

Cuando $\mathbf{X}'\mathbf{X}$ es no singular, la solución es única y resulta

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

Propiedades

Bajo los supuestos que hemos mencionado:

- $\widehat{\boldsymbol{\beta}}$ es insesgado: $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ (sesgo)
- $V(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ (varianza)
- $\widehat{\boldsymbol{\beta}}$ es una combinación lineal de $\mathbf{Y}$ (linealidad)
- $\widehat{\boldsymbol{\beta}}$ es el estimador lineal insesgado de menor varianza (optimalidad)

Predichos y Residuos

- i-ésimo valor predicho $\quad \widehat{y}_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_0 x_{ip}$

- i-ésimo residuo $\quad r_i = y_i - \widehat{y}_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_0 x_{ip})$

Explicando la Variabilidad: igual que antes...

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 \quad = \quad \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$$

Inferencia

Para hacer inferencia sobre los coeficientes del modelo (tests de hipótesis e intervalos de confianza) es necesario hacer un supuesto adicional:

Normalidad de los errores: $\epsilon_i \sim N(0, \sigma^2)$

$\longrightarrow$ **media 0, homoscedásticos e independientes**

Tradicionalmente se hacen tests de hipótesis para determinar la significación de cada uno de los coeficientes (y los IC asociados) y también se realiza un test para comprobar la significación de toda la regresión.

Inferencia individual

Para hacer inferencia, tests de hipótesis e intervalos de confianza, sobre los coeficientes del modelo es necesario hacer un supuesto adicional:

$$\text{Normalidad de los errores:} \epsilon_i \sim N(0, \sigma^2)$$

$\longrightarrow$ **media 0, homoscedásticos e independientes**.

Cuando $\mathbf{X}$ tiene rango completo puede probarse que:

$$\widehat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

Regresión lineal simple queda:

$$\begin{aligned} Var(\widehat{\beta}_0) &= \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right) \\ Var(\widehat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \end{aligned}$$

$\Rightarrow \sigma$ necesitaremos estimarla

Estimación de σ^2

Un estimador insesgado de σ^2 es

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n-(p+1)} = \frac{\sum_{i=1}^{n} r_i^2}{n-p-1}$$

Bajo el supuesto $\epsilon_i \sim N(0, \sigma^2)$

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{var}(\widehat{\beta}_j)}} \sim t_{n-p-1} \quad j = 0, 1, \ldots, p$$

Por lo tanto: un intervalo de nivel exacto $1-\alpha$

$$\widehat{\beta}_j \pm t_{n-p-1,\alpha/2}\sqrt{\widehat{var}(\widehat{\beta}_j)}$$

Regresión Lineal Simple

Así, en el caso más simple obtendríamos los siguientes Intervalos de Confianza de nivel $1-\alpha$

$$\widehat{\beta}_0 \quad \pm \quad t_{n-2,\alpha/2}\,\widehat{\sigma}\,\sqrt{\left(\frac{1}{n}+\frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)}$$

$$\widehat{\beta}_1 \quad \pm \quad t_{n-2,\alpha/2}\,\widehat{\sigma}\,\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$

Tests de Hipótesis

Bajo el supuesto $\epsilon_i \sim N(0, \sigma^2)$

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{var}(\widehat{\beta}_j)}} \sim t_{n-p-1} \quad j = 0, 1, \ldots, p$$

Para testear: $H_0 : \beta_j = 0 \quad vs. \quad H_1 : \beta_j \neq 0$ basta comparar con el percentil deseado de una t_{n-p-1} a

$$\frac{\widehat{\beta}_j}{\sqrt{\widehat{var}(\widehat{\beta}_j)}}$$

Rechazamos H_0 a nivel α si

$$\left| \frac{\widehat{\beta}_j}{\sqrt{\widehat{var}(\widehat{\beta}_j)}} \right| > t_{n-p-1,\alpha/2}$$

Analicemos una salida

```
> salida.iones<-lm(y~x1+x2)
> summary(salida.iones)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-44.584 -24.565  -3.266  22.330  63.201

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -520.0767   192.1071  -2.707  0.02039 *
x1            10.7812     0.4743  22.730 1.35e-10 ***
x2          -152.1489    36.6754  -4.149  0.00162 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.93 on 11 degrees of freedom
Multiple R-squared:  0.9798, Adjusted R-squared:  0.9762
F-statistic: 267.2 on 2 and 11 DF,  p-value: 4.742e-10
```

Test de Significación de toda la regresión

Este test chequea la hipótesis

$$H_0 : \beta_1 = \beta_2 = \quad = \beta_p = 0 \quad vs. \quad H_1 : \exists \text{ al menos un } \beta_j \neq 0$$

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$$

Bajo normalidad, tenemos que

$$F = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2/p}{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2/(n-p-1)} \sim \mathcal{F}_{p,n-p-1},$$

Test de Significación de toda la regresión

Este test chequea la hipótesis

$$H_0 : \beta_1 = \beta_2 = \quad = \beta_p = 0 \quad vs. \quad H_1 : \exists \text{ al menos un } \beta_j \neq 0$$

En el ejemplo, tenemos:

```
F-statistic 267.2 on 2 and 11 DF, p-value 4.742 e-10
```

Por lo que el estadístico F concluye que el modelo de regresión es altamente significativo.

R^2 ajustado

Para completar la lectura de la salida, tenemos que introducir al R^2 ajustado.

El coeficiente R^2 tiene la particularidad de que a medida que se agregan nuevas variables, va aumentando.

Esta no es una característica deseable cuando se desean comparar modelos con distinto número de variables o complejidad. Para ello se suele usar una versión corregida que penaliza por la cantidad de parámetros del modelo:

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

R^2_{adj} tiene en cuenta la complejidad del modelo