¿Cómo estimamos los parámetros?

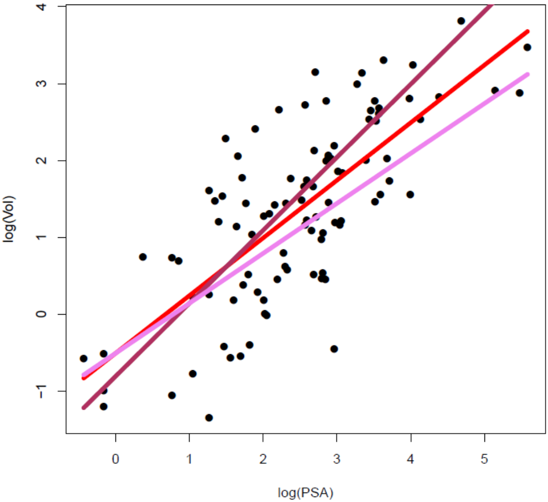
Mínimos Cuadrados

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

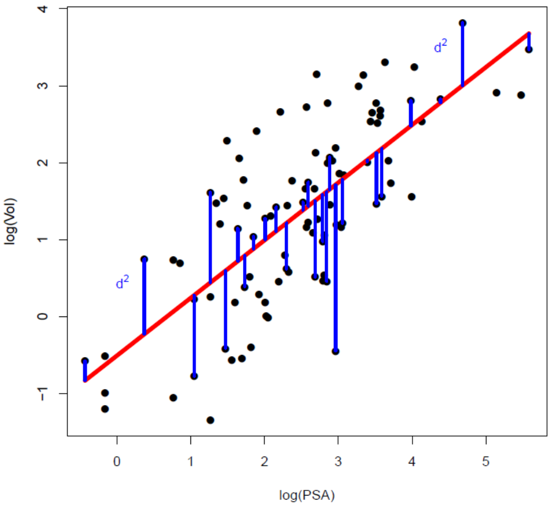
Tendremos en cuenta:

a) tomar una distancia promedio de la recta a todos los puntos
b) elegir la recta con esta distancia promedio lo más pequeña posible.

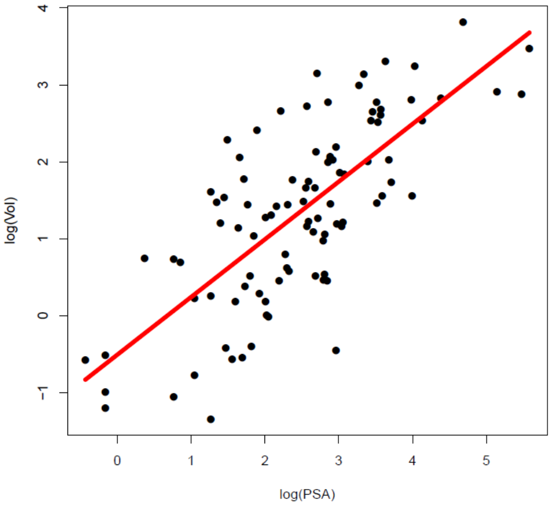
Si tenemos$(\mathbf{x}_i, y_i)$, $1 \leq i \leq n$, y queremos predecir y a partir de x usando una recta, podríamos definir el error cometido en cada punto como la distancia vertical del punto a la recta.



log(Vol)
log(PSA)
4
3
2
1
0
−1
0
1
2
3
4
5



d²
d²
log(Vol)
log(PSA)
-1
0
1
2
3
4
0
1
2
3
4
5


log(Vol)
log(PSA)
-1
0
1
2
3
4
0
1
2
3
4
5

Método de Mínimos Cuadrados

Supongamos que tenemos un modelo que depende de p parámetros. Sean $(\mathbf{x}_i, y_i)$ tales que

$$y_i = m(\mathbf{x}_i, \theta_1 \ldots \theta_p) + \varepsilon_i$$

y los errores son tales que $\mathbb{E}(\varepsilon_i) = 0$, $\mathbb{V}(\varepsilon_i) = \sigma^2$, $\mathbb{C}ov(\varepsilon_i, \varepsilon_j) = 0$, es decir son no correlacionados y la función m es conocida salvo por los parámetros $\theta_1 \ldots \theta_p$.

Estimamos $\theta_1 \ldots \theta_p$ minimizando la suma de cuadrados residual . El estimador de mínimos cuadrados $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_p)$ cumple

$$\widehat{\boldsymbol{\theta}} = \underset{\mathbf{b} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \left[y_i - m(\mathbf{x}_i, b_1 \ldots, b_p) \right]^2$$

Método de Mínimos Cuadrados

Estimamos $\theta_1 \ldots \theta_p$ minimizando la suma de cuadrados residual

$\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_p)$ es el estimador de mínimos cuadrados si minimiza en $\mathbf{b} = (b_1, \ldots, b_p)^{\mathsf{t}}$

$$\sum_{i=1}^{n} \left(y_i - m(\mathbf{x}_i, b_1 \ldots, b_p)\right)^2$$

Por ejemplo, en el caso de la regresión simple en el que $m(x, \theta_1, \theta_2) = \theta_1 + \theta_2\, x$, minimizaremos:

$$\frac{1}{n} \sum_{i=1}^{n} \left[y_i - (b_1 + b_2 x_i)\right]^2 .$$

Esta suma se llama la suma de cuadrados residual para la recta y la recta resultante recta de cuadrados mínimos.

Caso Lineal

Dado el vector $\mathbf{b} \in \mathbb{R}^p$, el vector de residuos es

$$\mathbf{r}(\mathbf{b}) = \mathbf{Y} - \mathbf{X}\mathbf{b}\,.$$

El estimador de mínimos cuadrados de $\theta_1 \ldots \theta_p$ minimiza

$$\sum_{i=1}^{n} (y_i - b_1 x_{i1} - \cdots - b_p x_{ip})^2 = \|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2\,,$$

donde $\|\mathbf{u}\|^2 = \mathbf{u}^{\mathsf{t}}\mathbf{u} = \sum_{i=1}^{n} u_i^2$.

Llamemos

$$\mathcal{S}(\mathbf{b}) = \|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2 = (\mathbf{Y} - \mathbf{X}\mathbf{b})^{\mathsf{t}}(\mathbf{Y} - \mathbf{X}\mathbf{b})$$

Un conjunto de funciones de $\mathbf{Y}$, $\widehat{\theta}_1 = \widehat{\theta}_1(\mathbf{Y})$, $\widehat{\theta}_2 = \widehat{\theta}_2(\mathbf{Y}), \ldots$ $\widehat{\theta}_p = \widehat{\theta}_p(\mathbf{Y})$ que minimiza $\mathcal{S}(\mathbf{b})$ es el estimador de mínimos cuadrados de $\boldsymbol{\theta}$ (LS).

El LS siempre existe, aunque no siempre es único.

Ecuaciones Normales

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

Los estimadores de mínimos cuadrados $\widehat{\theta}_1, \ldots, \widehat{\theta}_p$ cumplen:

$$\frac{\partial \mathcal{S}(\mathbf{b})}{\partial b_k}\Big|_{\mathbf{b}=\widehat{\boldsymbol{\theta}}} = -2 \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} b_j \right) x_{ik}\Big|_{\mathbf{b}=\widehat{\boldsymbol{\theta}}} = 0$$

Ecuaciones Normales

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

Los estimadores de mínimos cuadrados $\widehat{\theta}_1, \ldots, \widehat{\theta}_p$ cumplen:

$$\frac{\partial \mathcal{S}(\mathbf{b})}{\partial b_k}\Big|_{\mathbf{b}=\widehat{\boldsymbol{\theta}}} = -2\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij} b_j\right) x_{ik}\Big|_{\mathbf{b}=\widehat{\boldsymbol{\theta}}} = 0$$

Por lo tanto, para $1 \leq k \leq p$

$$\sum_{i=1}^{n} y_i x_{ik} = \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij} x_{ik} \widehat{\theta}_j$$

Ecuaciones Normales

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

Los estimadores de mínimos cuadrados $\widehat{\theta}_1, \ldots, \widehat{\theta}_p$ cumplen:

$$\frac{\partial \mathcal{S}(\mathbf{b})}{\partial b_k}\Big|_{\mathbf{b}=\widehat{\boldsymbol{\theta}}} = -2 \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} b_j \right) x_{ik}\Big|_{\mathbf{b}=\widehat{\boldsymbol{\theta}}} = 0$$

Por lo tanto, para $1 \leq k \leq p$

$$\sum_{i=1}^{n} y_i x_{ik} = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij} x_{ik} \widehat{\theta}_j$$

reordenando la suma

$$\sum_{i=1}^{n} y_i x_{ik} = \sum_{j=1}^{p} \widehat{\theta}_j \sum_{i=1}^{n} x_{ij} x_{ik}$$

Ecuaciones Normales

Como

$$\sum_{i=1}^{n} y_i x_{ik} \quad = \quad \sum_{j=1}^{p} \widehat{\theta}_j \sum_{i=1}^{n} x_{ij} x_{ik} \quad 1 \leq k \leq p$$

cuando el modelo tiene intercept, y recordemos que $x_{i1} = 1$ para todo i, luego

$$\begin{aligned} n\widehat{\theta}_1 + \widehat{\theta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \widehat{\theta}_p \sum_{i=1}^{n} x_{ip} &= \sum_{i=1}^{n} y_i \\ \widehat{\theta}_1 \sum_{i=1}^{n} x_{ik} + \widehat{\theta}_2 \sum_{i=1}^{n} x_{i2} x_{ik} + \cdots + \widehat{\theta}_p \sum_{i=1}^{n} x_{ip} x_{ik} &= \sum_{i=1}^{n} y_i x_{ik} \quad k = 2, \ldots, p \end{aligned}$$

Ecuaciones Normales

$$\sum_{j=1}^{p} \widehat{\theta}_j \sum_{i=1}^{n} x_{ij} x_{ik} \quad = \quad \sum_{i=1}^{n} y_i x_{ik} \quad 1 \leq k \leq p$$

El estimador de mínimos cuadrados cumple estas p ecuaciones, entonces

$$\mathbf{X}^{\mathsf{t}} \mathbf{X} \widehat{\boldsymbol{\theta}} = \mathbf{X}^{\mathsf{t}} \mathbf{Y} \,,$$

que se conocen como ecuaciones normales.

Si $\mathbf{X}^{\mathsf{t}} \mathbf{X}$ es no singular, la solución es única y resulta

$$\boxed{\widehat{\boldsymbol{\theta}} = \left(\mathbf{X}^{\mathsf{t}} \mathbf{X} \right)^{-1} \mathbf{X}^{\mathsf{t}} \mathbf{Y} \,.}$$

Ecuaciones Normales

$$\sum_{j=1}^{p} \widehat{\theta}_j \sum_{i=1}^{n} x_{ij} x_{ik} \quad = \quad \sum_{i=1}^{n} y_i x_{ik} \quad 1 \le k \le p$$

El estimador de mínimos cuadrados cumple estas p ecuaciones, entonces

$$\mathbf{X}^{\mathsf{t}}\mathbf{X}\widehat{\boldsymbol{\theta}} = \mathbf{X}^{\mathsf{t}}\mathbf{Y}\,,$$

que se conocen como ecuaciones normales.

Si $\mathbf{X}^{\mathsf{t}}\mathbf{X}$ es no singular, la solución es única y resulta

$$\boxed{\widehat{\boldsymbol{\theta}} = \left(\mathbf{X}^{\mathsf{t}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{t}}\mathbf{Y}\,.}$$

Al vecor de valores predichos lo notamos:

$$\boxed{\widehat{\mathbf{Y}} = \mathbf{X}^{\mathsf{t}}\widehat{\boldsymbol{\theta}}\,.}$$

Ejemplo - Ejercicio: Comprobar

En el caso de regresión simple tendríamos

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

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

El sistema sería

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

Ejemplo

La inversa resulta

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

y además

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

y por lo tanto

$$\widehat{\boldsymbol{\theta}} = \begin{pmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \end{pmatrix} = \frac{1}{n\displaystyle\sum_{i=1}^n (x_i - \overline{x})^2} \begin{pmatrix} (\displaystyle\sum_{i=1}^n y_i)(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)(\sum_{i=1}^n x_i y_i) \\ n\displaystyle\sum_{i=1}^n x_i y_i - (\sum_{i=1}^n y_i)(\sum_{i=1}^n x_i) \end{pmatrix}$$

Ejemplo - Ejercicio: Comprobar

Entonces

$$\boxed{\widehat{\theta}_1 = \overline{y} - \overline{x}\widehat{\theta}_2}$$

y por otro lado

$$\boxed{\widehat{\theta}_2 = \frac{\displaystyle\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\displaystyle\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2}}$$

Interpretación Geométrica

Asumamos que $\mathbf{X}$ tiene rango completo.

Nuestro modelo plantea

$$\Omega: \quad \begin{aligned} \mathbb{E}(\mathbf{Y}) &= \mathbf{X}\boldsymbol{\theta} \\ \boldsymbol{\Sigma}_{\mathbf{Y}} &= \sigma^2\mathbf{I} \end{aligned}$$

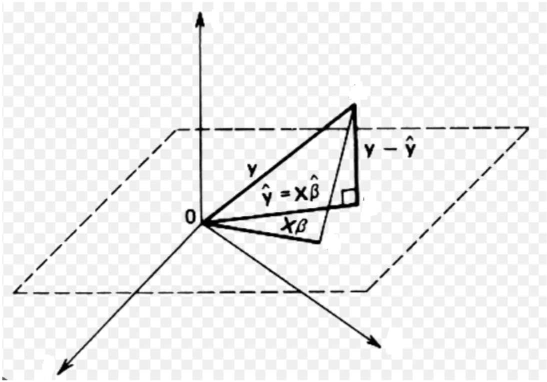
Luego, si llamamos

$$\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\theta}$$

llamando a $\mathbf{x}^{(i)}$ la i–ésima columna de $\mathbf{X}$. Nuestro modelo asume que

$$E(\mathbf{Y}) = \theta_1\mathbf{x}^{(1)} + \theta_2\mathbf{x}^{(2)} + \cdots + \theta_p\mathbf{x}^{(p)}$$

Es decir que $E(\mathbf{Y}) \in \mathcal{V} =$ subespacio generado por las p columnas de $\mathbf{X}$: $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(p)}$ y r es rg$(\mathbf{X})$.



y
$\hat{y} = X\hat{\beta}$
$y - \hat{y}$
0
$X\beta$

Entonces

$$\min_b \mathcal{S}(\mathbf{b}) = \min_b \|\mathbf{Y} - \mathbf{X}\mathbf{b}\|^2 = \min_{\mathbf{z}\in V} \|\mathbf{Y} - \mathbf{z}\|^2$$

El mínimo se alcanza en la proyección ortogonal de Y sobre V: $\widehat{\mathbf{Y}}$, a quien podemos escribir como $b_1\mathbf{x}^{(1)} + b_2\mathbf{x}^{(2)} + \cdots + b_p\mathbf{x}^{(p)}$, siempre existe y es única, aunque los b_i pueden no serlo.

En términos de las ecuaciones normales tenemos que:

$$\begin{aligned} \mathbf{X}^{\mathsf{t}}\mathbf{X}\widehat{\boldsymbol{\theta}} &= \mathbf{X}^{\mathsf{t}}\mathbf{Y} \\ \mathbf{X}^{\mathsf{t}}\widehat{\mathbf{Y}} &= \mathbf{X}^{\mathsf{t}}\mathbf{Y} \end{aligned}$$

Caso en que $rg(\mathbf{X}) = p$

En este caso existe la inversa de $\mathbf{X}^{\mathsf{t}}\mathbf{X}$, pues $\mathrm{rg}(\mathbf{X}^{\mathsf{t}}\mathbf{X}) = \mathrm{rg}(\mathbf{X}) = p$.

Como de las ecuaciones normales:

$$
\begin{aligned}
\mathbf{X}^{\mathsf{t}}\mathbf{X}\widehat{\boldsymbol{\theta}} &= \mathbf{X}^{\mathsf{t}}\mathbf{Y} \\
\\
\widehat{\boldsymbol{\theta}} &= (\mathbf{X}^{\mathsf{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{t}}\mathbf{Y}
\end{aligned}
$$

entonces

$$
\mathbf{X}\widehat{\boldsymbol{\theta}} = \mathbf{X}(\mathbf{X}^{\mathsf{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{t}}\mathbf{Y} = \mathbf{P}\mathbf{Y} = \widehat{\mathbf{Y}}
$$

En consecuencia, el vector de residuos es:

$$
\begin{aligned}
\mathbf{r}(\widehat{\boldsymbol{\theta}}) &= \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\theta}} \\
&= \mathbf{Y} - \mathbf{X}(\mathbf{X}^{\mathsf{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{t}}\mathbf{Y} \\
&= \mathbf{Y} - \mathbf{P}\mathbf{Y} = (\mathbf{I} - \mathbf{P})\mathbf{Y}
\end{aligned}
$$

donde $\mathbf{P} = \mathbf{X}(\mathbf{X}^{\mathsf{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{t}} \in \mathbb{R}^{n \times n}$.

Propiedades del Estimador de Mínimos Cuadrados

Usando la notación matricial escribimos el modelo como

$$\begin{aligned} \Omega: \quad \mathbf{Y} &= \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \\ E(\boldsymbol{\epsilon}) &= 0 \\ \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} &= \sigma^2 \mathbf{I} \end{aligned}$$

Lema: Si se cumple el modelo Ω, tenemos que

- $\widehat{\boldsymbol{\theta}}$ es un estimador insesgado de $\boldsymbol{\theta}$, es decir $E(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$.
- $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\theta}}} = \sigma^2(\mathbf{X}^{\mathsf{t}}\mathbf{X})^{-1}$

Estimación de σ^2

Las varianzas de los estimadores dependen del diseño y σ^2, que es desconocida.

Dado que $\sigma^2 = E(\epsilon^2)$, parece natural estimarla mediante el promedio de los cuadrados de los residuos. El vector de residuos es

$$\begin{aligned} \mathbf{r} &= \mathbf{Y} - \widehat{\mathbf{Y}} \\ &= \mathbf{Y} - \mathbf{P}\mathbf{Y}\,, \end{aligned}$$

Veremos que bajo el modelo Ω

$$\boxed{s^2 = \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{n-p} = \frac{\|\mathbf{Y} - \mathbf{P}\mathbf{Y}\|^2}{n-p} = \frac{\mathbf{Y}^{\mathsf{t}}(\mathbf{I} - \mathbf{P})\mathbf{Y}}{n-p}}$$

es un estimador insesgado de σ^2.