Modelo Lineal

Estadística (M)

Modelización y Predicción

Algunas 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).

Modelización y Predicción

Algunas 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 técnicas 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.

Modelización y Predicción

Algunas 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 técnicas 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.

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.

Aprendizaje Estadístico

Algunos ejemplos de aprendizaje estadístico:

Aprendizaje Estadístico

Algunos ejemplos de aprendizaje estadístico:

- Predecir si un paciente hospitalizado tendrá un segundo infarto de miocardio o no teniendo en cuenta mediciones clínicas, dietas y variables demográficas.
- Predecir los precios que tendrán en 6 meses las acciones de ciertas compañías a partir de mediciones del rendimiento de las compañías y datos macroeconómicos.
- 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.

Vamos a considerar el último ejemplo:

En 97 pacientes que van a ser tratados con una prostatectomía radical se miden las siguientes variables:

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

El objetivo es poder predecir el logaritmo del volumen del tumor ($y =$**lcavol**).

Diagramas de Dispersión

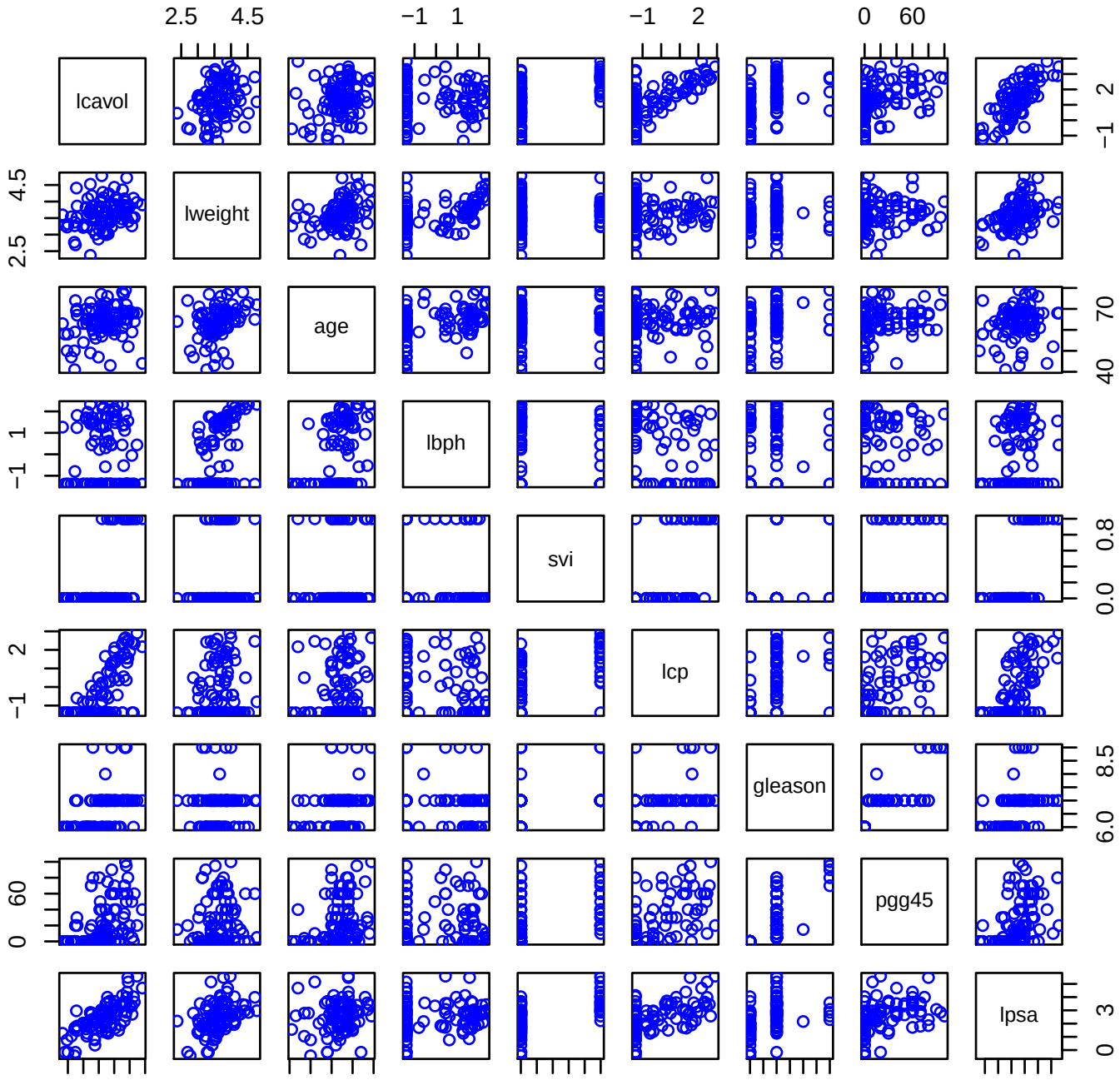
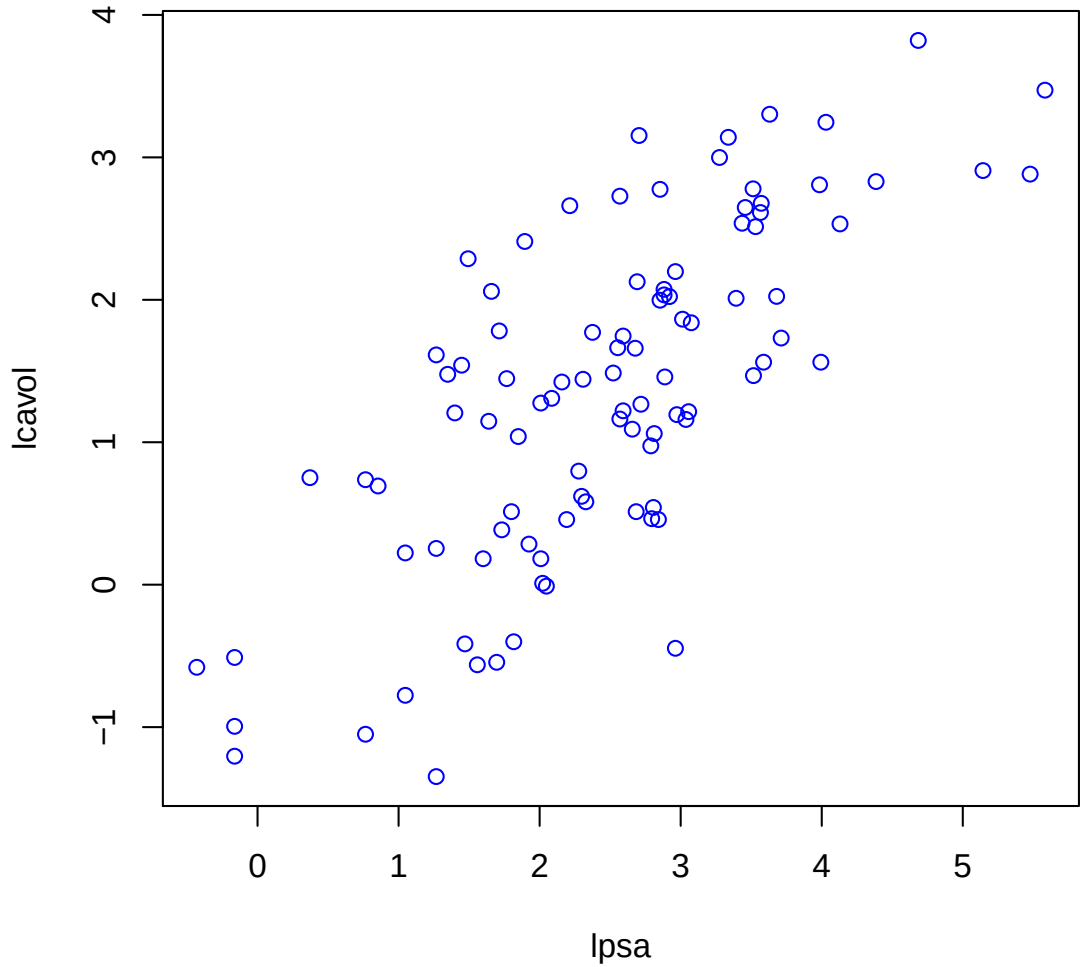


Diagrama de Dispersión



Modelos de regresión

Buscamos un modelo que exprese a la variable de respuesta en términos de las otras variables presentes (covariables).

Cuando hablamos de un modelo nos referimos a una expresión matemática que sea válida aproximadamente y que describa en algún sentido el comportamiento de la variable de interés en función de las demás variables predictoras.

En general, identificamos con la letra y a la variable dependiente. El modelo pretende describir cómo el comportamiento de $E(y)$ varía bajo condiciones cambiantes del vector de p covariables $\mathbf{x}$.

En un modelo de regresión se postularía:

$$y = f(x_1, x_2, x_3, \ldots, x_p) + \epsilon$$

o en general

$$y = f(\mathbf{x}) + \epsilon$$

Modelos de regresión

$$y = f(\mathbf{x}) + \epsilon$$

Las posibles funciones de regresión f pertenecen a una clase $\mathcal{F}$ tan grande que es frecuente que se simplifique el problema suponiendo cierta forma o ciertas propiedades de la función de regresión f.

Una forma de simplificar el problema suponiendo que la familia $\mathcal{F}$ puede expresarse en función de un número finito de constantes desconocidas, a estimar, llamadas parámetros, que controlan el comportamiento del modelo. En este sentido diremos que el modelo de regresión es paramétrico.

Modelos de regresión

Para simplificar, pensemos que tenemos dos covariables: x_1 y x_2.

Algunos ejemplos de modelos paramétricos y no paramétricos cuando hay dos covariables son

Modelos no paramétricos

(i) $y = f(x_1, x_2) + \varepsilon$ donde $f(x_1, x_2)$ es una función continua.

(ii) $y = f(x_1, x_2) + \varepsilon$ donde $f(x_1, x_2)$ es una función continua y derivable.

(iii) $y = f_1(x_1) + f_2(x_2) + \varepsilon$, f_i funciones continuas.

(iv) $y = f(x_1, x_2) + \varepsilon$ donde $f(x_1, x_2)$ es monótona creciente en x_1 y x_2.

Modelos de regresión

Modelos paramétricos

(i) $y = \alpha + \beta x_1 + \gamma x_2 + \varepsilon$

(ii) $y = \alpha \log(x_1) + \beta \log(x_2) + \gamma\, x_1^3 + \delta\ \mathsf{sen}(x_2) + \varepsilon$

(iii) $y = \alpha x_1^{\beta} x_2^{\delta} + \varepsilon$

(iv) $y = \alpha e^{\beta x_1} + \gamma e^{\delta x_2} + \varepsilon$

Modelo Lineal

Uno de los modelos más sencillos es el **modelo lineal**, en el que los parámetros intervienen como simples coeficientes de las variables independientes o de funciones de éstas.

Es el caso de:

(i) $y = \alpha + \beta x_1 + \gamma x_2 + \varepsilon$

(ii) $y = \alpha \log(x_1) + \beta \log(x_2) + \gamma\, x_1^3 + \delta\ \mathsf{sen}(x_2) + \varepsilon$

Modelo Lineal

En estos dos ejemplos $f(x)$ es **lineal** en los **parámetros**. No es el caso, por ejemplo, de $f(x) = \alpha e^{-\beta x}$, conocido como creciemiento exponencial, ya que no es lineal como función de los parámetros α o β.

Algunos ejemplos sencillos de modelos lineales dependientes de una sola variable son:

$$\begin{aligned} f(x) &= \alpha + \beta x \\ f(x) &= \alpha + \beta x + \gamma x^2 \\ f(x) &= \alpha + \beta \log(x) \end{aligned}$$

Modelo Lineal

En general, en un modelo lineal tendremos que la $i-$ésima respuesta y_i está asociada a un vector de covariables $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ de la siguiente forma

$$y_i = \theta_1 x_{i1} + \cdots + \theta_p x_{ip} + \epsilon_i \, .$$

es decir

$$f(\mathbf{x}_i) = \theta_1 x_{i1} + \cdots + \theta_p x_{ip} = \boldsymbol{\theta}' \mathbf{x}_i$$

siendo $\boldsymbol{\theta}' = (\theta_1, \ldots, \theta_p)$.

Modelo Lineal

En general, en un modelo lineal tendremos que la $i-$ésima respuesta y_i está asociada a un vector de covariables $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ de la siguiente forma

$$y_i = \theta_1 x_{i1} + \cdots + \theta_p x_{ip} + \epsilon_i \, .$$

es decir

$$f(\mathbf{x}_i) = \theta_1 x_{i1} + \cdots + \theta_p x_{ip} = \boldsymbol{\theta}' \mathbf{x}_i$$

siendo $\boldsymbol{\theta}' = (\theta_1, \ldots, \theta_p)$.

Eventualmente

- las covariables podrían ser funciones de otras variables, tal como ocurre en el caso ii) y en nuestro ejemplo.
- $x_{i1} = 1$ para todo i, en este caso decimos que el modelo tiene intercept u ordenada al origen.

Enfoque matricial

respuesta $y \longleftrightarrow p$ variables explicativas x_j

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

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

$$\begin{aligned} y_i &= \theta_1 x_{i1} + \cdots + \theta_p x_{ip} + \epsilon_i \quad i = 1, \ldots, n \\ E(\varepsilon_i) &= 0 \\ V(\varepsilon_i) &= \sigma^2 \\ cov(\varepsilon_i, \varepsilon_j) &= 0 \quad i \neq j \end{aligned}$$

donde, $\theta_1, \ldots, \theta_p$ son p parámetros desconocidos a estimar.

En forma matricial plantearíamos

$$y_i = \theta_1 x_{i1} + \cdots + \theta_p x_{ip} + \epsilon_i \quad i = 1, \ldots, n$$

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

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ . \\ . \\ \theta_p \end{pmatrix} \qquad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{pmatrix}$$

$$\Downarrow$$

$$\underset{n \times 1}{\mathbf{Y}} = \underset{n \times p}{\mathbf{X}} \underset{p \times 1}{\boldsymbol{\theta}} + \underset{n \times 1}{\boldsymbol{\epsilon}}$$

La matriz $\mathbf{X} \in \Re^{n \times p}$ recibe el nombre de matriz de regresión o de diseño.

En general, se elige de tal forma que tenga rango máximo, es decir $\text{rg}(\mathbf{X}) = p$, sin embargo esto no siempre es posible, como en el caso de algunos diseños tratados en análisis de la varianza (ANOVA).

Trataremos el caso de rango completo.

La teoría que veremos no necesita que la primera columna sea de 1's, es decir que el modelo tenga intercept, por lo tanto estudiaremos el caso general.

Algunos ejemplos: Modelo de regresión simple

En el caso más sencillo de regresión simple tendríamos

$$y_i = \theta_1 + \theta_2 x_i + \epsilon_i \quad 1 \leq i \leq n$$

$$p = 2 \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix} \qquad \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$$

Modelo de regresión a través del origen

En el caso más sencillo de regresión simple tendríamos

$$y_i = \theta_1 + \theta_2 x_i + \epsilon_i \quad 1 \leq i \leq n$$

$$p = 2 \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix} \qquad \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$$

Modelo de posición

$$y_i = \mu + \epsilon_i \quad 1 \leq i \leq n$$

$$p = 1 \qquad \mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ 1 \end{pmatrix} \qquad \theta = \mu$$

Modelo de 2 muestras

$$y_{i1} = \mu_1 + \epsilon_{i1} \quad 1 \leq i \leq n_1$$

$$y_{i2} = \mu_2 + \epsilon_{i2} \quad 1 \leq i \leq n_2$$

$$p = 2 \qquad \mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ . & . \\ . & . \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ . & . \\ . & . \\ 0 & 1 \end{pmatrix} \qquad \boldsymbol{\theta} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

Modelo de regresión cuadrática

$$y_i = \theta_1 + \theta_2 x_i + \theta_3 x_i^2 + \epsilon_i \quad 1 \leq i \leq n$$

$$p = 3 \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ . & . & . \\ . & . & . \\ 1 & x_n & x_n^2 \end{pmatrix} \qquad \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix}$$

Propiedades de vectores y matrices aleatorias

Dada una matriz $\mathbf{V}$ $(r \times s)$ de variables aleatorias conjuntamente distribuidas $\{V_{ij}\}$ con esperanza finita, definimos la matriz o vector de esperanzas como:

$$\{E(\mathbf{V})\}_{ij} = E(V_{ij})$$

En el caso del modelo Ω, esto nos permite decir que el vector de errores es tal que

$$E(\boldsymbol{\epsilon}) = \mathbf{0}$$

y que

$$E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = E\begin{pmatrix} \epsilon_1\epsilon_1 & \epsilon_1\epsilon_2 & \dots & \epsilon_1\epsilon_n \\ \epsilon_2\epsilon_1 & \epsilon_2\epsilon_2 & \dots & \epsilon_2\epsilon_n \\ \dots & & \dots & \\ \dots & & \dots & \\ \epsilon_n\epsilon_1 & \epsilon_n\epsilon_2 & \dots & \epsilon_n\epsilon_n \end{pmatrix} = \sigma^2\mathbf{I}$$

Lema 1: Sean $\mathbf{A} \in \Re^{q \times r}$, $\mathbf{B} \in \Re^{s \times t}$ y $\mathbf{C} \in \Re^{q \times t}$ matrices constantes y $\mathbf{V}$ una matriz aleatoria de dimensión $r \times s$, entonces:

$$E(\mathbf{AVB} + \mathbf{C}) = \mathbf{A}E(\mathbf{V})\mathbf{B} + \mathbf{C}.$$

Matriz de Covarianza

Sea $\mathbf{v} = (v_1, \ldots, v_n)'$ un vector aleatorio de variables con $E(v_i) = \mu_i$ y varianza finita. Definimos la matriz de covarianza de $\mathbf{v}$ como:

$$\{\boldsymbol{\Sigma}_{\mathbf{v}}\}_{ij} = Cov(\mathbf{v}_i, \mathbf{v}_j) = E[(v_i - \mu_i)(v_j - \mu_j)]$$

Podemos escribirla como:

$$\boldsymbol{\Sigma}_{\mathbf{v}} = E[(\mathbf{v} - \boldsymbol{\mu})(\mathbf{v} - \boldsymbol{\mu})']$$

donde $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$.

En este sentido, como $E(\boldsymbol{\epsilon}) = \mathbf{0}$, entonces hemos visto que

$$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2\mathbf{I}$$

Usaremos el siguiente resultado basado en las propiedades de linealidad de la esperanza:

Lema 2: Sean $\mathbf{A} \in \Re^{m \times n}$, una matriz constante, $\mathbf{d}$ un vector de constantes y $\mathbf{v}$ un vector aleatorio n–dimensional con matriz de covarianza $\mathbf{\Sigma_v}$. Si $\mathbf{w} = \mathbf{A}\mathbf{v} + \mathbf{d}$, entonces:

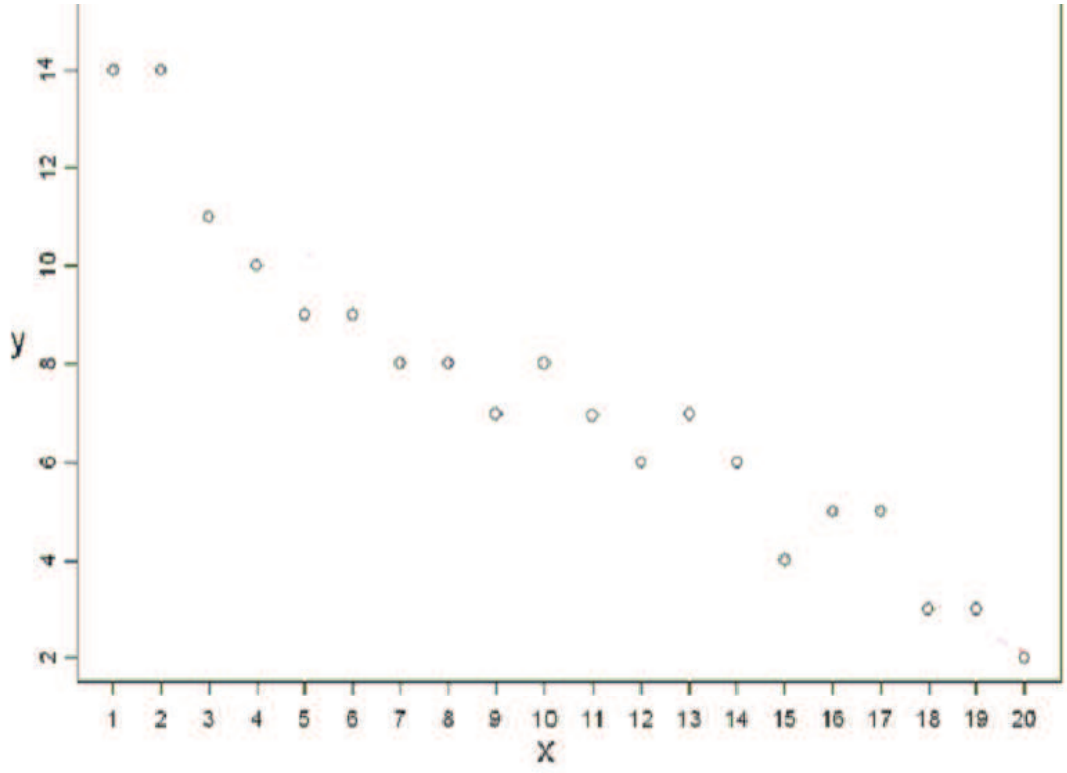
$$\mathbf{\Sigma_w} = \mathbf{A}\mathbf{\Sigma_v}\mathbf{A}' .$$

El modelo que presentamos más arriba puede escribirse como:

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

o equivalentemente

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


y
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
X

¿Cómo estimamos los parámetros?

Mínimos Cuadrados

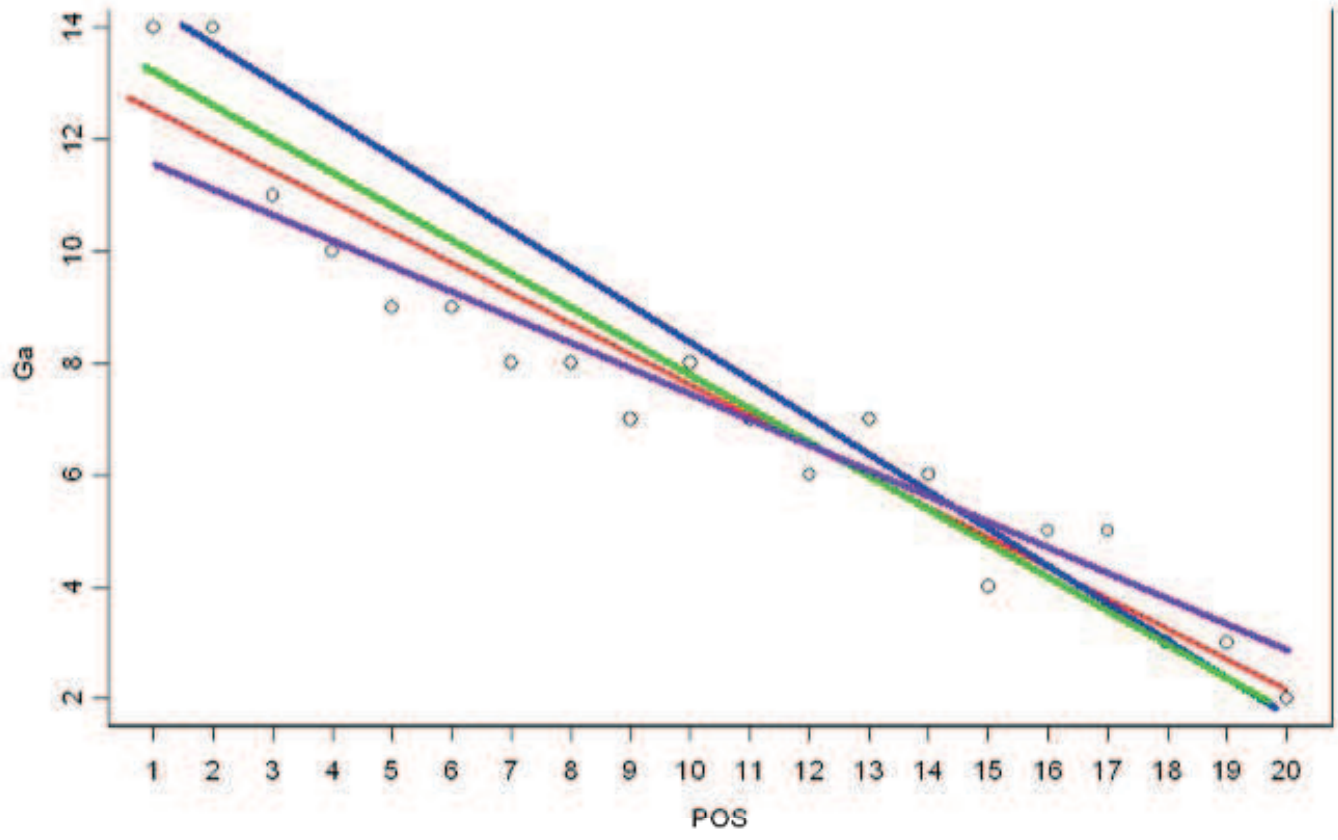
Si los puntos en un gráfico parecen seguir una recta como en el gráfico, el problema es elegir la recta que mejor ajusta los puntos.

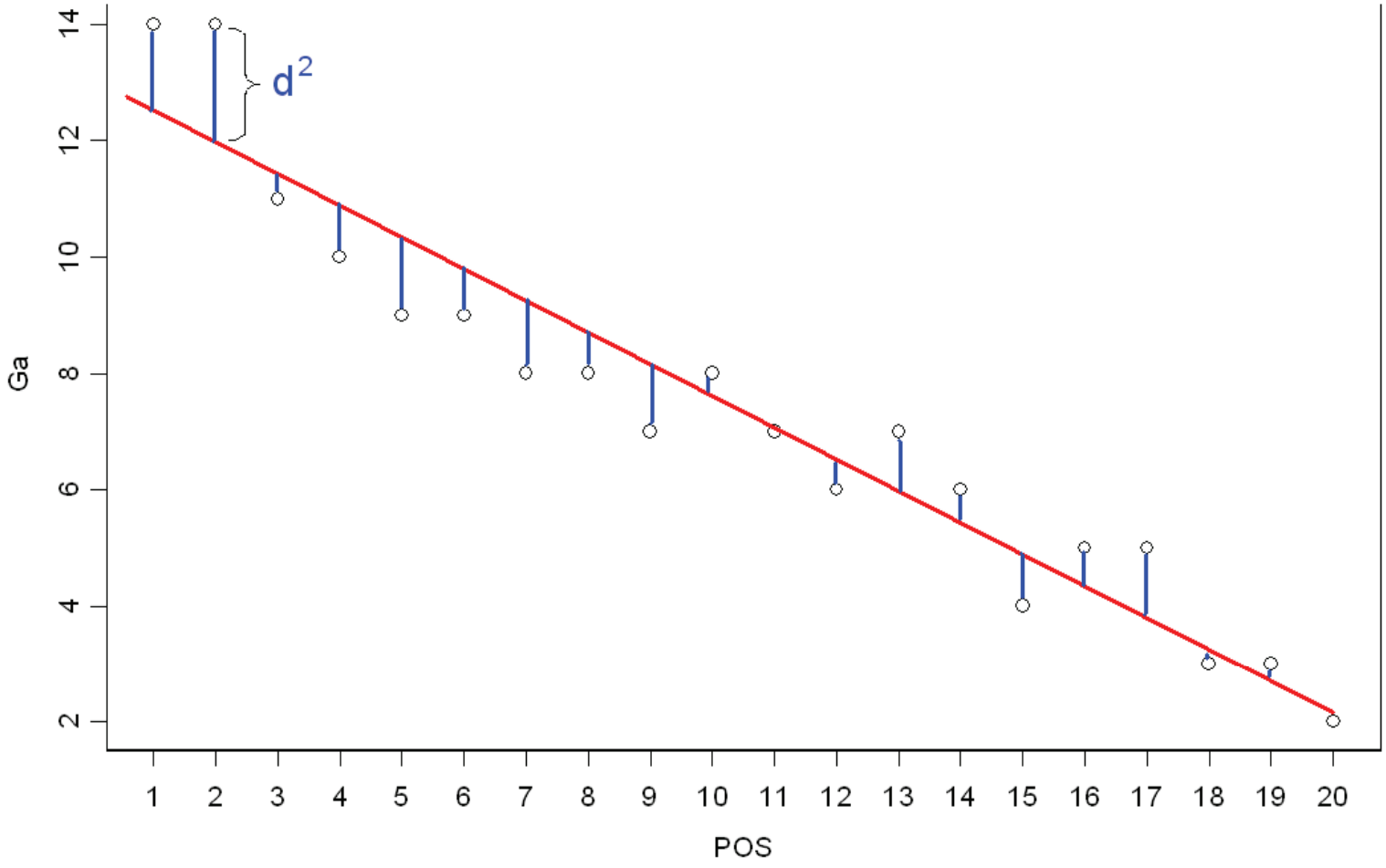
Tendremos en cuenta:

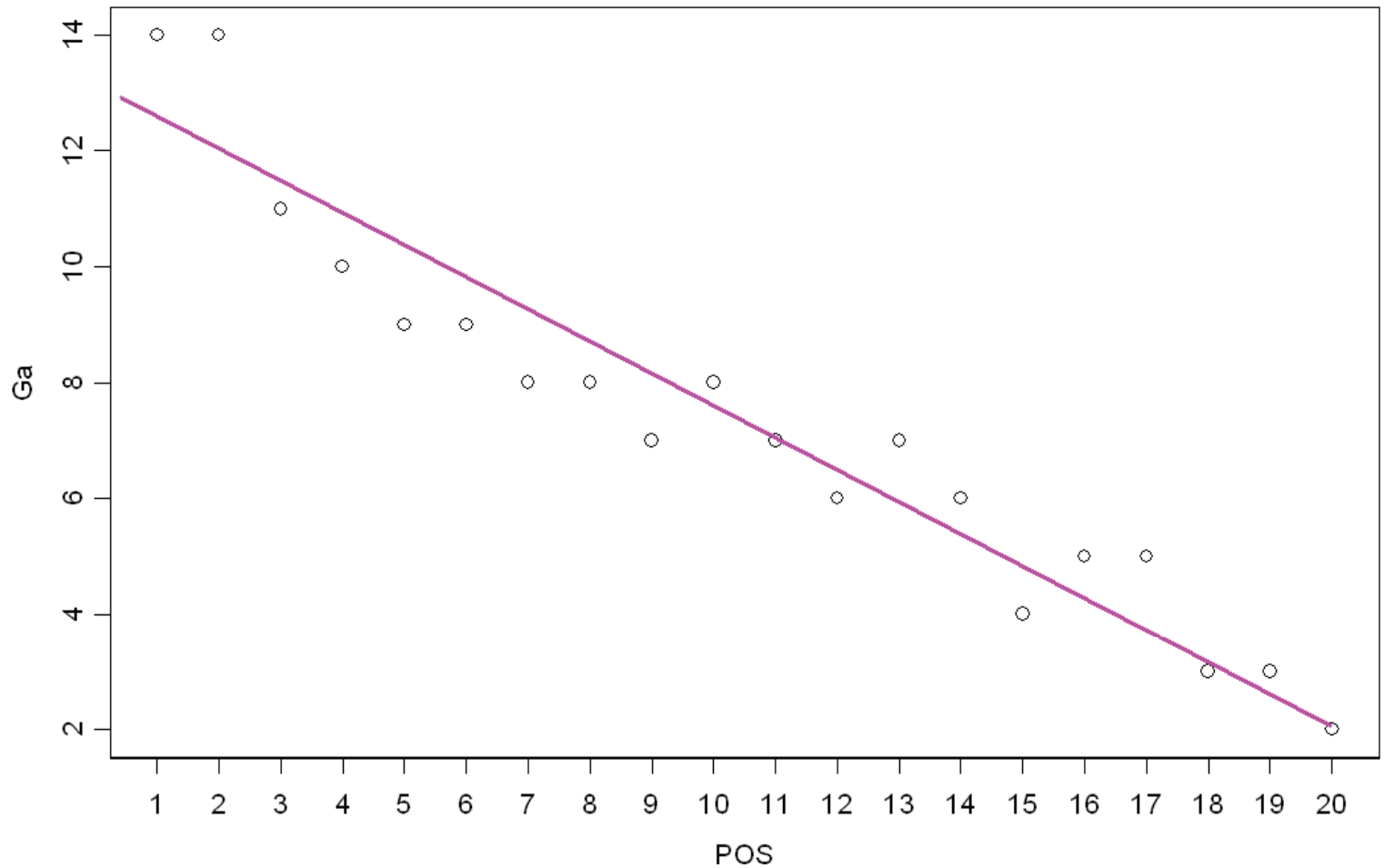
a) tomar una distancia promedio de la recta a todos los puntos

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


Ga
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
POS


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
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

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

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

y los errores son tales que $E(\varepsilon_i) = 0$, $V(\varepsilon_i) = \sigma^2$, $Cov(\varepsilon_i, \varepsilon_j) = 0$, es decir son no correlacionados y la función f es conocida salvo por los parámetros $\theta_1 \dots \theta_p$.

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

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

$$\sum_{i=1}^{n} (y_i - f(\mathbf{x}_i, b_1 \dots, b_p))^2$$

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

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

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 $\mathbf{b} \in \Re^p$, el vector de residuos es

$$\mathbf{Y} - \mathbf{X}\mathbf{b}\,.$$

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

Llamemos

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

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).

Veremos que 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:

$$\frac{\partial \mathcal{S}(\mathbf{b})}{\partial b_k} = -2 \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij} b_j) x_{ik} = 0$$

Ecuaciones Normales

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

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

$$\frac{\partial \mathcal{S}(\mathbf{b})}{\partial b_k} = -2 \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} x_{ij} b_j) x_{ik} = 0$$

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

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

reordenando la suma

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

Ecuaciones Normales

Como

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

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

$$n\widehat{\theta}_1 + \widehat{\theta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \widehat{\theta}_p \sum_{i=1}^{n} x_{ip} \quad = \quad \sum_{i=1}^{n} y_i$$

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

Ecuaciones Normales

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

El estimador de mínimos cuadrados cumple estas p ecuaciones, entonces

$$\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\theta}} = \mathbf{X}'\mathbf{Y}\,,$$

que se conocen como ecuaciones normales.

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

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

Ejemplo

En el caso de regresión simple tendríamos

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

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

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

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

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

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

Ejemplo

Entonces

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

y por otro lado

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

Interpretación Geométrica

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

Nuestro modelo plantea

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

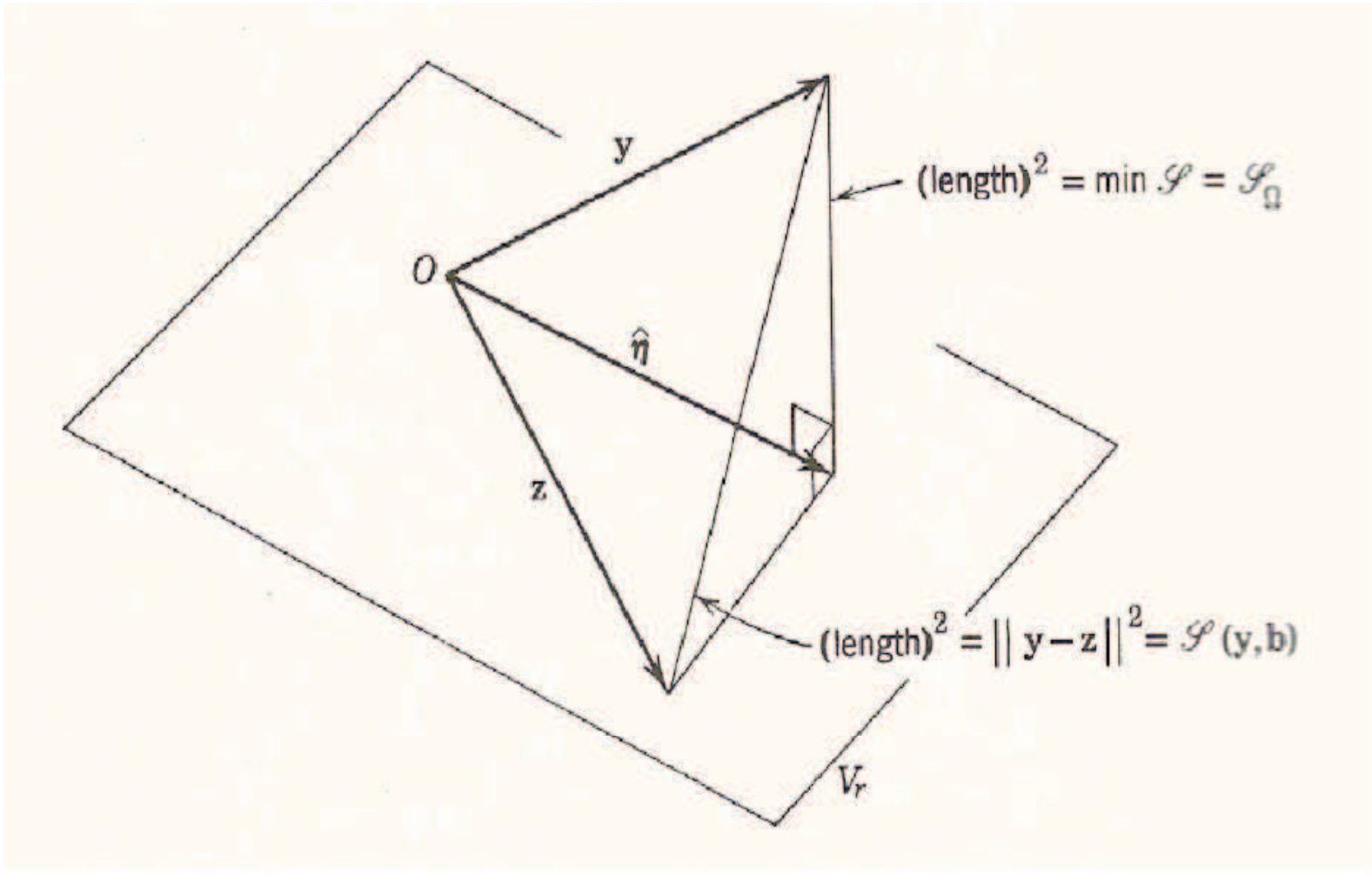
Luego, si

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

llamando a $\mathbf{x}^i$ la i–ésima columna de $\mathbf{X}$, Tenemos

$$\boldsymbol{\eta} = \theta_1\mathbf{x}^1 + \theta_2\mathbf{x}^2 + \cdots + \theta_p\mathbf{x}^p$$

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


y
O
$\hat{\eta}$
z
V_r
(length)^2 = min $\mathscr{S}$ = $\mathscr{S}_\Omega$
(length)^2 = $||\,\mathbf{y}-\mathbf{z}\,||^2$ = $\mathscr{S}(\mathbf{y},\mathbf{b})$

Entonces

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

El mínimo se alcanza en la proyección ortogonal de Y sobre V_p: $\widehat{\mathbf{Y}}$ o $\widehat{\boldsymbol{\eta}}$, 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:

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

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

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

En este caso existe la inversa de $\mathbf{X}'\mathbf{X}$, pues $\text{rg}(\mathbf{X}'\mathbf{X}) = \text{rg}(\mathbf{X}) = p$.

De las ecuaciones normales queda:

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

entonces

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

En consecuencia el vector de residuos es:

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

donde $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \in \Re^{n \times n}$.

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 $E(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$.
- $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\theta}}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

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

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

Bajo el modelo Ω, tenemos que

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

es un estimador insesgado de σ^2.

Lema Auxiliar: Sea $\mathbf{x}$ un vector aleatorio n–dimensional y sea $\mathbf{A} \in \Re^{n \times n}$ una matriz simétrica. Si $E(\mathbf{x}) = \boldsymbol{\mu}$ y su matriz de covarianza es $\boldsymbol{\Sigma}_{\mathbf{x}}$ entonces

$$E(\mathbf{x}'\mathbf{A}\mathbf{x}) = tr(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$$