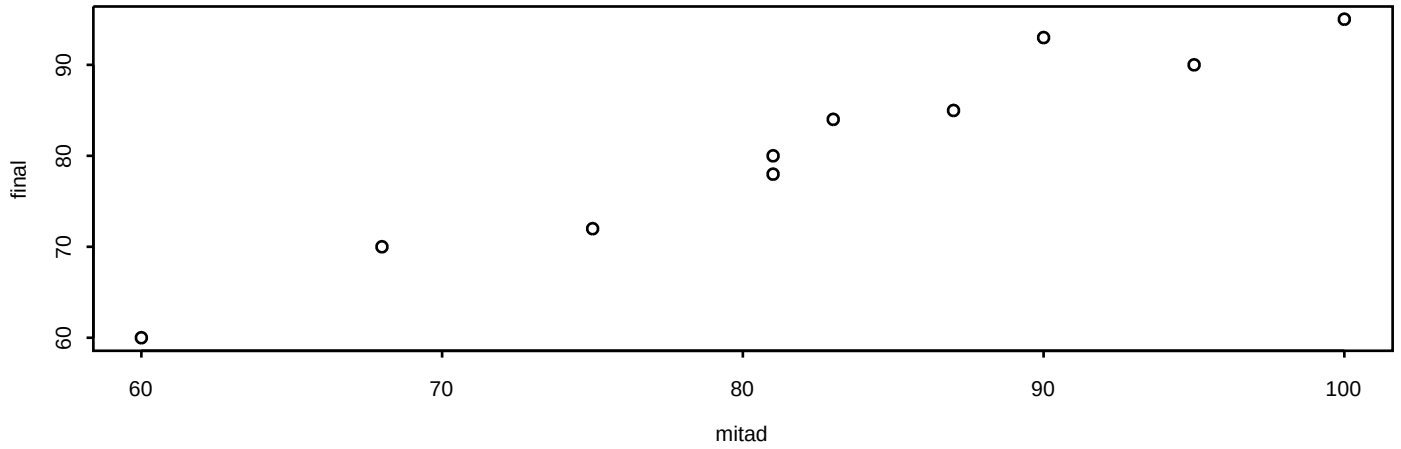
Modelo Lineal

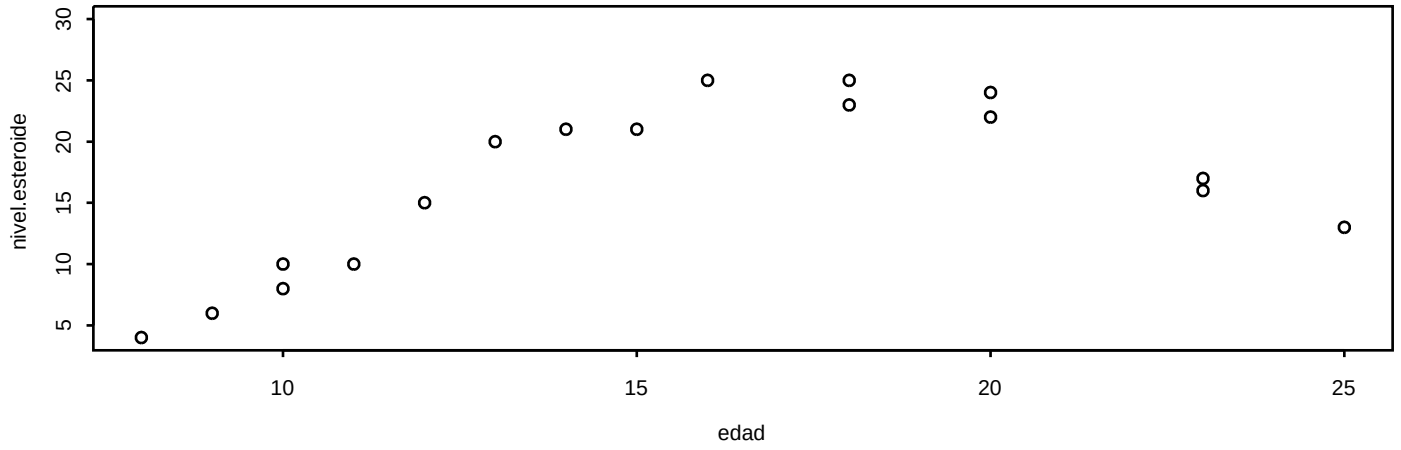
En regresión lineal interesa establecer la relación entre una variable dependiente Y y otras p variables: $X_1, \ldots, X_p$. Esta metodología es ampliamente usada en problemas de economía, de la industria y de ciencias en general. Por ejemplo:

- en mujeres de 8 a 25 años se desea relacionar la edad y la cantidad de esteroides presentes en plasma.
- dadas las evaluaciones de mitad y de fin de año de alumnos que participan en un estudio de rendimiento, se quiere relacionar la performance de los alumnos en los dos exámenes. El objetivo es poder predecir en situaciones similares cómo le irá a un alumno en la evaluación final a partir de lo que se observa en la evaluación de mitad de curso.

- un ingeniero está interesado en la relación entre la cantidad de óxido que se forma en un metal calcinado en un horno y la temperatura de horneado y el tiempo expuesto a dichas temperaturas.

En los dos primeros ejemplos podríamos tener gráficos como los siguientes:


final
60
70
80
90
mitad
60
70
80
90
100


nivel.esteroide
5
10
15
20
25
30
edad
10
15
20
25

En los dos primeros ejemplos consideramos sólo dos variables, mientras que en el tercero hay 3 variables involucradas.

En general tendremos:

- y: *variable dependiente.*
- x: *variables independientes (predictoras, regresoras o covariables).*

Buscaremos un modelo que exprese a la variable dependiente en términos de las variables independientes.

Cuando hablamos de un modelo nos referimos a una expresión matemática que describa en algún sentido el comportamiento de la variable de interés en función de las demás variables, es decir, las covariables.

En general, identificaremos con la letra Y (y) a la variable dependiente. El modelo pretende describir cómo el comportamiento de $E(Y)$ varía bajo condiciones cambiantes de las otras variables.

En nuestro caso, supondremos, al menos en un principio, que $V(Y)$ no es afectada por estas condiciones cambiantes, es decir toma un valor constante σ.

Bajo el supuesto de que *otras variables* aportan información sobre la variable Y, éstas variables son incorporadas al modelo como variables independientes.

Identificaremos con $\mathbf{X} = (X_1, \ldots, X_p)'$ ($\mathbf{x} = (x_1, \ldots, x_p)'$) a las variables independientes. Estas podrían ser variables aleatorias o constantes conocidas. En general, trabajaremos bajo este último caso y más adelante lo extenderemos al caso de variables aleatorias.

Una forma general de plantear el modelo es expresando a la media de la distribución de Y como una $g(\mathbf{x})$. En el caso de covariables alestorias como

$$E(Y|\mathbf{X} = \mathbf{x}) = g(\mathbf{x}) \quad \text{para } \mathbf{x} \in D\,,$$

o en el caso de covariables fijas como

$$Y = g(X_1, \ldots, X_p) + \varepsilon\,,$$

según el caso, donde la función g en general no será conocida y $E(\varepsilon) = 0$.

Los modelos de este tipo se llaman modelos de regresión. Las posibles funciones de regresión g pertenecen a una clase $\mathcal{G}$ tan grande que es frecuente que se simplifique el problema suponiendo cierta forma o ciertas propiedades de la función de regresión g.

Una forma de simplificar el problema suponiendo que la familia $\mathcal{G}$ 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.

Se dirá que el modelo de regresión es no parámétrico si la familia $\mathcal{G}$ no puede expresarse en un número finito de parámetros.

Algunos ejemplos de modelos paramétricos y no paramétricos cuando hay dos variables independientes X_1 y X_2.

Modelos paramétricos

(i) $Y = \theta_1 X_1 + \theta_2 X_2 + \theta_3 + \varepsilon$

(ii) $Y = \theta_1 e^{\theta_2 X_1} + \theta_3 e^{\theta_4 X_2} + \varepsilon$

(iii) $Y = \theta_1 X_1^{\theta_2} X_2^{\theta_3} + \varepsilon$

(iv) $Y = \theta_1 \log X_1 + \theta_2 \log X_2 + \theta_3 X_1^3 + \theta_4 \text{ sen } X_2 + \varepsilon$

Modelos no paramétricos

(i) $Y = g(X_1, X_2) + \varepsilon$ donde $g(X_1, X_2)$ es una función continua.

(ii) $Y = g(X_1, X_2) + \varepsilon$ donde $g(X_1, X_2)$ es una función continua y derivable.

(iii) $Y = g(X_1, X_2) + \varepsilon$ donde $g(X_1, X_2)$ es monótona creciente en X_1 y X_2.

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) $Y = \theta_1 X_1 + \theta_2 X_2 + \theta_3 + \varepsilon$

(iv) $Y = \theta_1 \log X_1 + \theta_2 \log X_2 + \theta_3 X_1^3 + \theta_4 \text{ sen } X_2 + \varepsilon$

En todos estos ejemplos $g(x)$ es **lineal** en los **parámetros**. No es el caso, por ejemplo, de $g(x) = \beta_o e^{-\beta_1 x}$, conocido como creciemiento exponencial, ya que no es lineal como función de los parámetros β_o o β_1.

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

$$\begin{aligned} g(x) &= \beta_o + \beta_1 x \\ g(x) &= \beta_o + \beta_1 x + \beta_2 x^2 \\ g(x) &= \beta_o + \beta_1 \log x \end{aligned}$$

En las situaciones más complejas Y depende de un conjunto de p variables $(x_1, \ldots, x_p)$, por lo tanto tendremos

$$g(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{ip-1} \, .$$

Eventualmente, las $x_i's$ podrían ser funciones de otras variables, tales como $W_1 = \log X_1$, $W_2 = \log X_2$, $W_3 = X_1^3$, etc., tal como ocurre en el caso iv).

También podríamos introducir variables explicativas que sean categóricas como las dummies que sólo toman los valores 0 y 1 y que sirven, como ya veremos, para indicar las distintas categorías de una variable categórica. Este caso es de especial interés pues permite tratar en el marco del modelo lineal el problema de comparar la media de más de dos poblaciones, que se conoce como **Análisis de la Varianza**.

Una vez establecido el modelo, nos interesará:

- Estimar los parámetros desconocidos: β_j y σ
- Testear hipótesis del tipo

$$H_o : \beta_j = 0 \qquad \text{o} \qquad H_o : c'\beta = \delta$$

- Intervalos de confianza para los parámetros o combinaciones lineales de los mismos.
- Predicción
- Chequeo de supuestos
- Identificación de datos atípicos.
- Medidas de ajuste
- Criterios para la selección de modelos.

Enfoque matricial

respuesta $y \longleftrightarrow p-1$ variables explicativas x_j

Por ahora, supondremos $x_j, 1 \leq j \leq p-1$ determinísticas.

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

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

donde, $\beta_0, \beta_1, \ldots, \beta_{p-1}$ son p parámetros desconocidos a estimar.

Este modelo tiene intercept u ordenada al origen, eventualmente podríamos saber que es 0, en cuyo caso plantearíamos

$$y_i = \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{ip-1} + \epsilon_i \quad i = 1, \ldots, n$$

En el caso general tenemos

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

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_{p-1} \end{pmatrix} \qquad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{pmatrix}$$

$$\Downarrow$$

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

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).

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.

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 delo modelo Ω, esto nos permite decir que el vector de errores es tal que

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

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: 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:

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

Podemos escribirla como:

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

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

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

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

Usaremos frecuentemente el siguiente

Lema: 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 $\Sigma_{\mathbf{v}}$. Si $\mathbf{w} = \mathbf{A}\mathbf{v} + \mathbf{d}$, entonces:

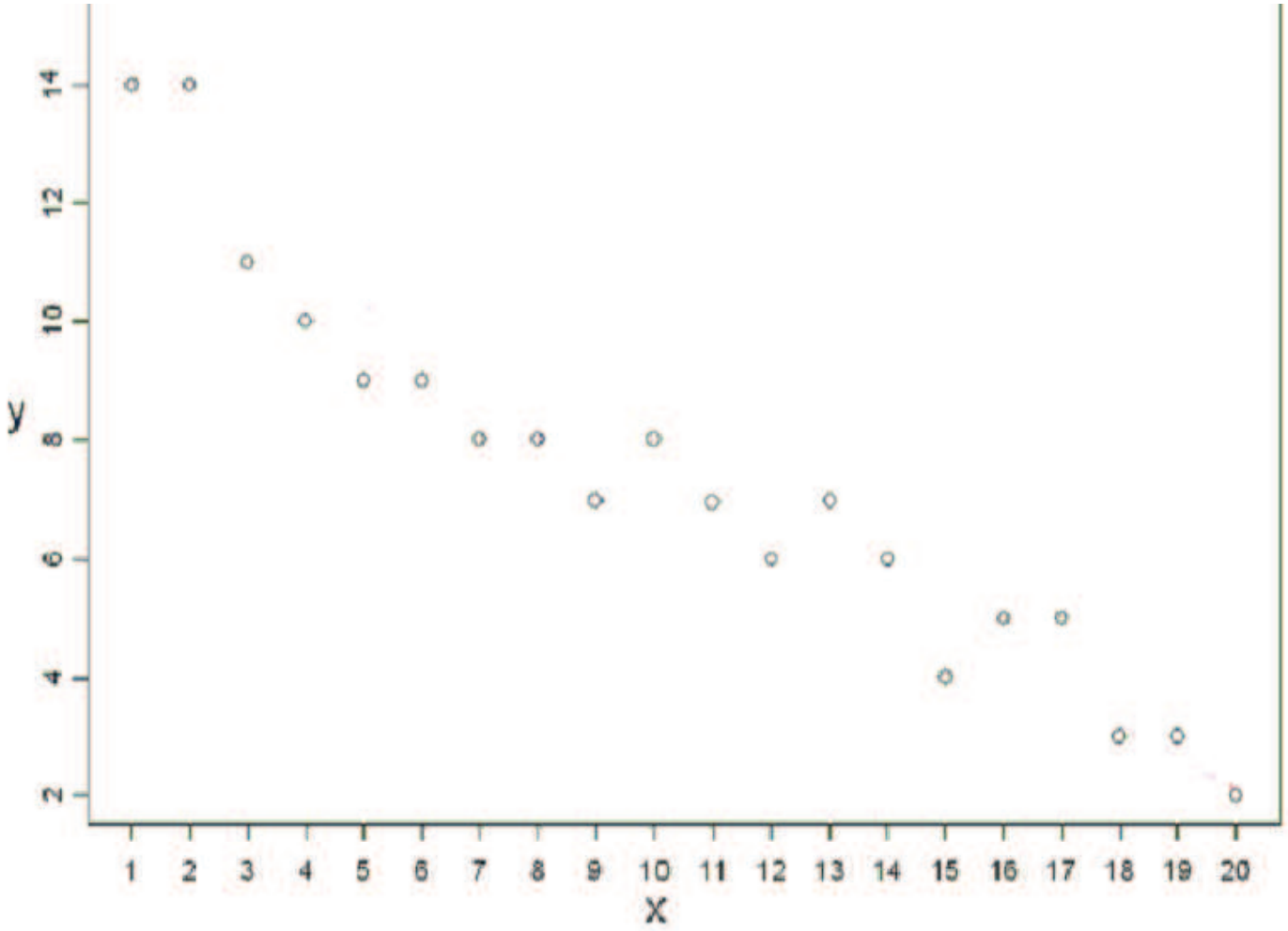
$$\Sigma_{\mathbf{w}} = \mathbf{A}\Sigma_{\mathbf{v}}\mathbf{A}' .$$

El modelo que presentamos más arriba puede escribirse como:

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

o equivalentemente

$$\Omega : E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} \quad \Sigma_{\mathbf{Y}} = \sigma^2\mathbf{I}$$


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

¿Cómo estimamos los parámetros?

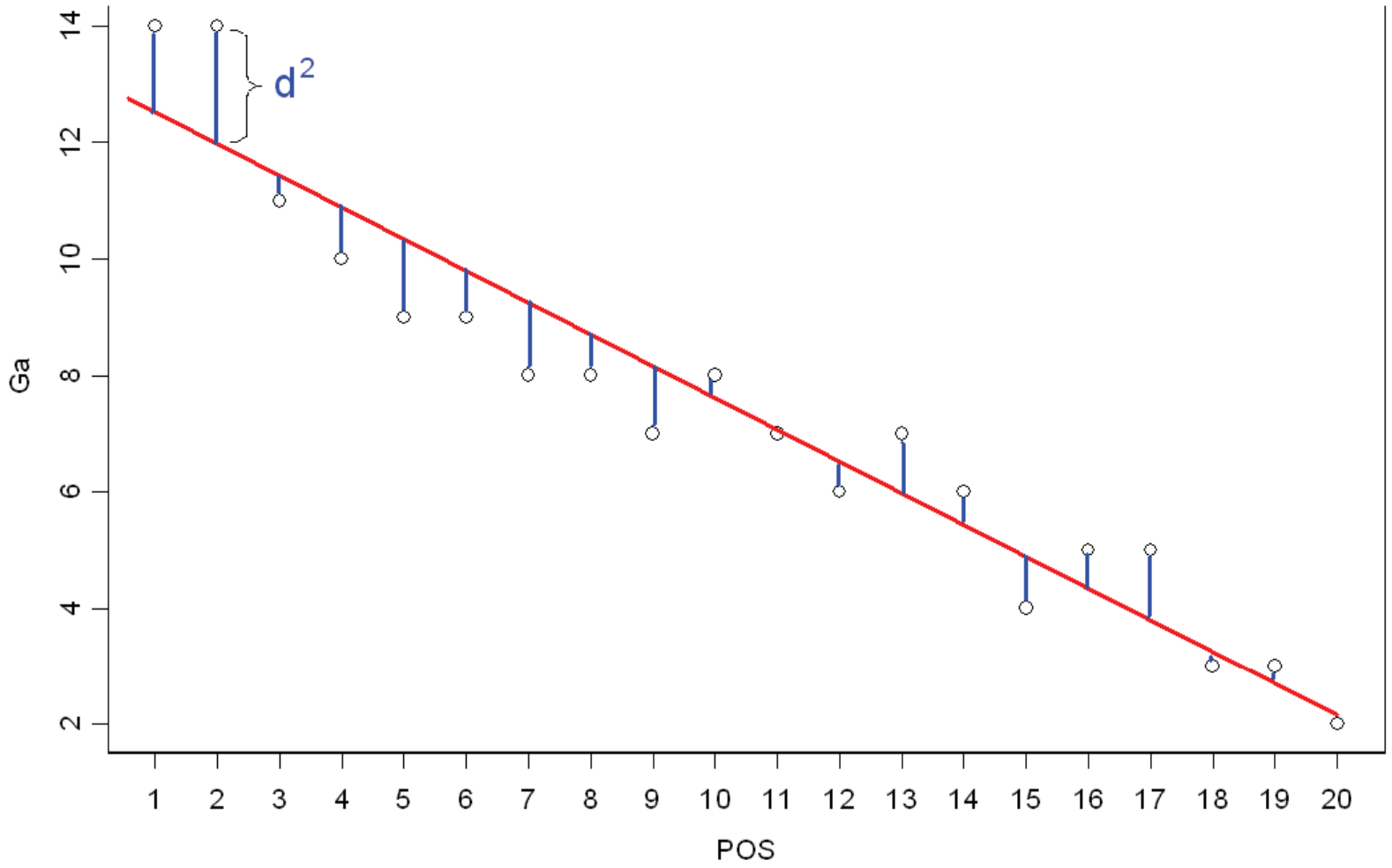
Mínimos Cuadrados

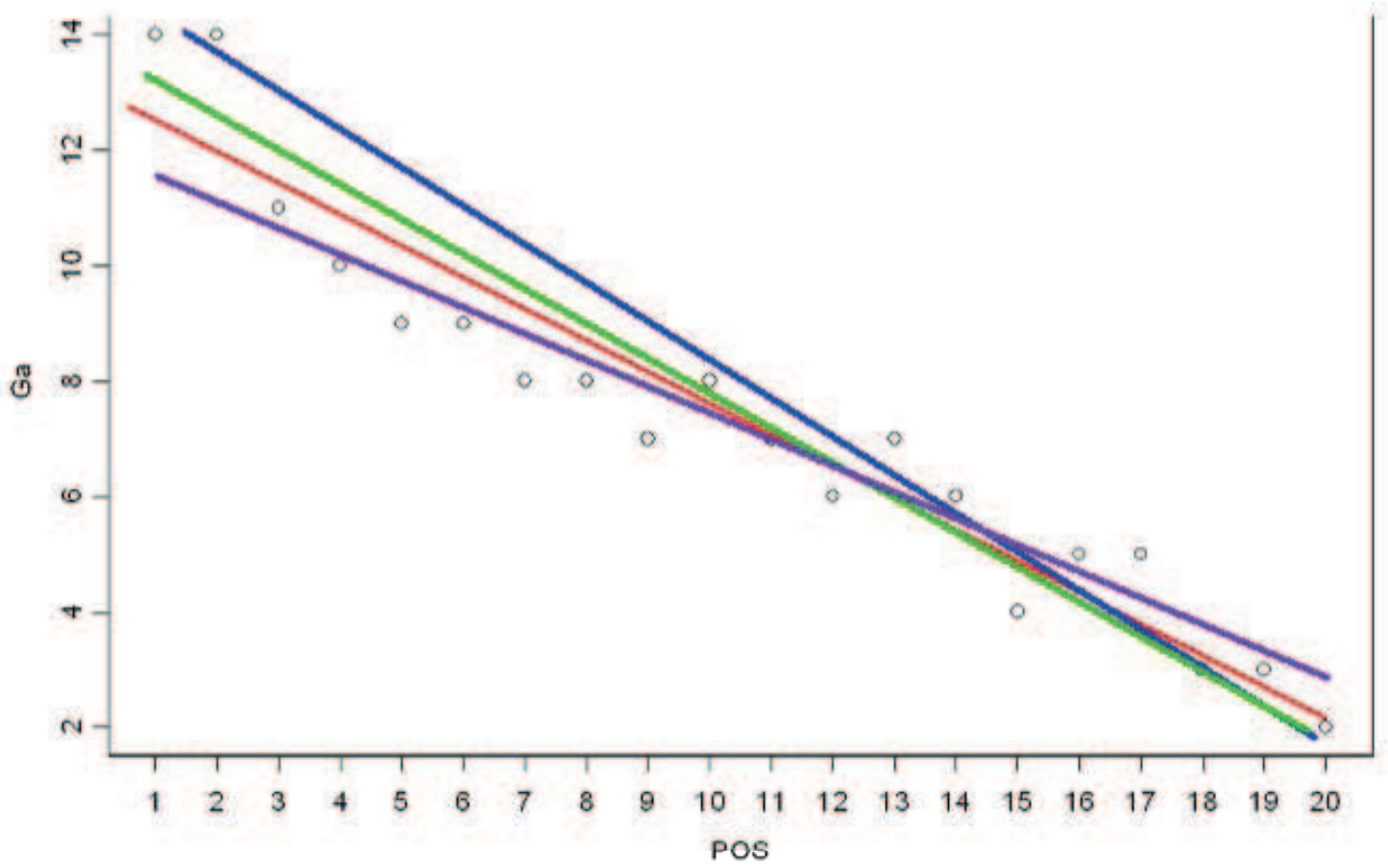
Si los puntos en un gráfico parecen seguir una recta, el problema es elegir la recta que mejor ajusta los puntos.

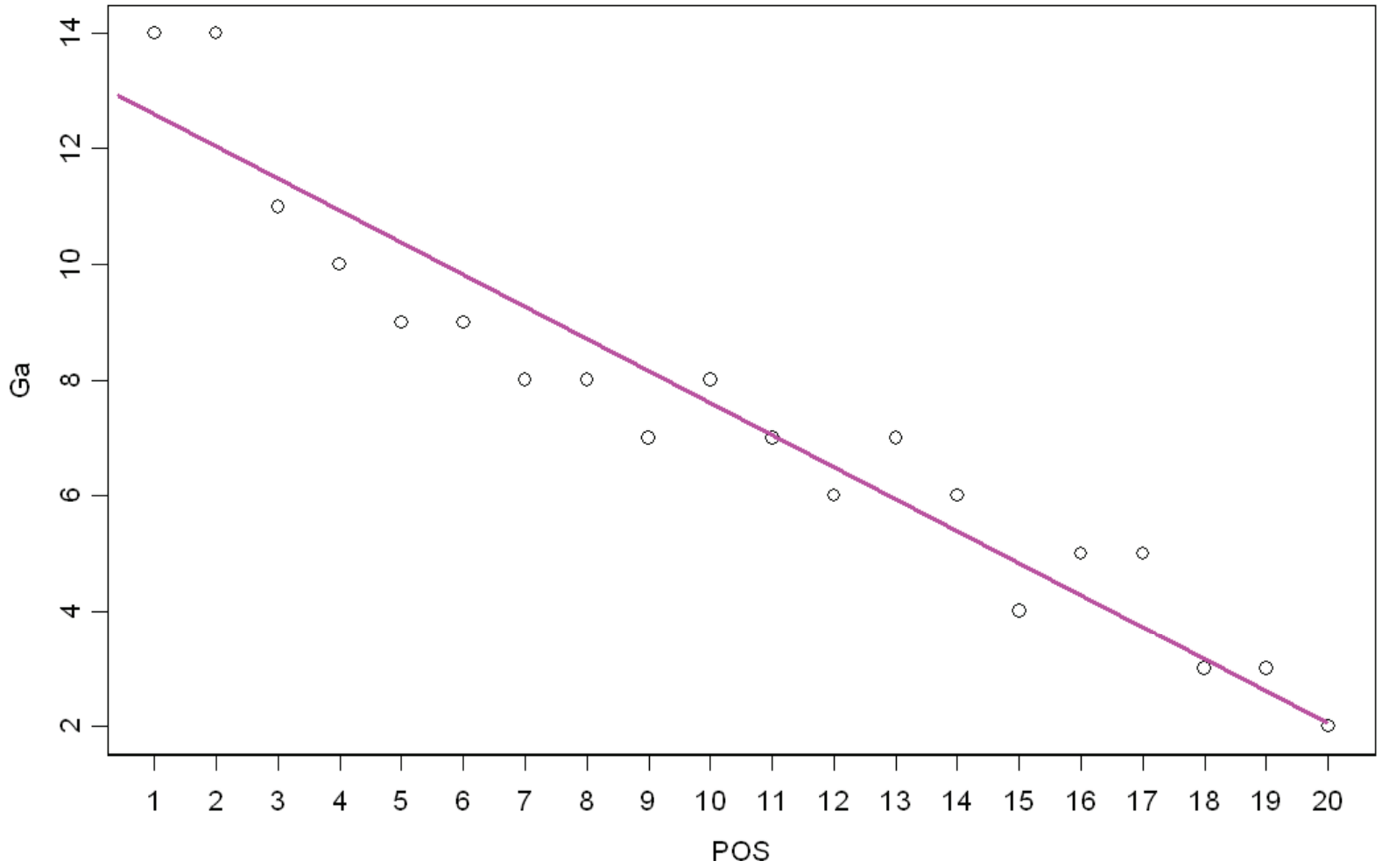
a) tomar una distancia promedio de la recta a todos los puntos

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.


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
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


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

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

$$y_i = g(\mathbf{x}_i, \beta_1 \ldots \beta_p) + \varepsilon_i$$

$E(\varepsilon_i) = 0$, $V(\varepsilon_i) = \sigma^2$, ε_i son independientes y la función g es conocida salvo por los parámetros $\beta_1 \ldots \beta_p$.

Estimamos $\beta_1 \ldots \beta_p$ minimizando la suma de cuadrados residual, o sea $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ es el estimador de mínimos cuadrados si minimiza

$$\sum_{i=1}^{n} (y_i - g(\mathbf{x}_i, \beta_1 \ldots \beta_p))^2$$

En el caso de la regresión simple en el que $g(x, \beta_1, \beta_2) = \beta_1 + \beta_2 x$, minimizaremos:

$$\frac{1}{n} \sum_{i=1}^{n} [y_i - (\beta_1 + \beta_2 x_i)]^2 .$$

Esta medida promedio se llama la suma de cuadrados residual del error para la recta. Fue inicialmente propuesta por Gauss. La recta de regresión así definida produce la menor suma de cuadrados residual para el error de predecir y a partir

de x y por esta razón se la suele llamar recta de mínimos cuadrados.

Consideremos para cada vector $\mathbf{b} \in \Re^p$ el vector de residuos

$$\mathbf{Y} - \mathbf{Xb}\,.$$

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

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

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

Llamemos

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

Definición: un conjunto de funciones de $\mathbf{Y}$, $\widehat{\beta}_1 = \widehat{\beta}_1(\mathbf{Y})$, $\widehat{\beta}_2 = \widehat{\beta}_2(\mathbf{Y}), \ldots$ $\widehat{\beta}_p = \widehat{\beta}_p(\mathbf{Y})$ que minimice $\mathcal{S}(\mathbf{b})$ es el estimador de mínimos cuadrados de $\boldsymbol{\beta}$ (LS).

Veremos que el LS siempre existe, pero no siempre es único.

Derivando e igualando a 0 obtenemos las **ecuaciones normales** . Los estimadores de mínimos cuadrados $\widehat{\beta}_1, \ldots, \widehat{\beta}_p$ 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 $1 \leq k \leq p$

$$\begin{aligned} \sum_{i=1}^{n} Y_i x_{ik} &= \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij} x_{ik} b_j \\ \sum_{i=1}^{n} Y_i x_{ik} &= \sum_{j=1}^{p} b_j \sum_{i=1}^{n} x_{ij} x_{ik} \end{aligned}$$

Si el modelo tiene intercept, y lo escribimos como antes en términos de $\beta_0, \ldots, \beta_{p-1}$, los estimadores $\widehat{\beta}_i$ cumplen

$$\begin{aligned} n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \ldots + \widehat{\beta}_{p-1} \sum_{i=1}^{n} x_{ip-1} &= \sum_{i=1}^{n} y_i \\ n\widehat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i1} x_{ik} + \ldots + \widehat{\beta}_{p-1} \sum_{i=1}^{n} x_{ip-1} x_{ik} &= \sum_{i=1}^{n} y_i x_{ik} \qquad k = 1, \ldots, p-1 \end{aligned}$$

Estas p ecuaciones pueden escribirse como

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

que se conocen como ecuaciones normales.

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

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

Ejemplo: En el caso de regresión simple tendríamos

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

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

El sistema sería

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

La inversa resulta

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

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

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

entonces

$$\boxed{b_0 = \overline{y} - \overline{x} b_1}$$

y por otro lado

$$\boxed{b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}$$

Interpretación Geométrica

Nuestro modelo plantea

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

Luego, si

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

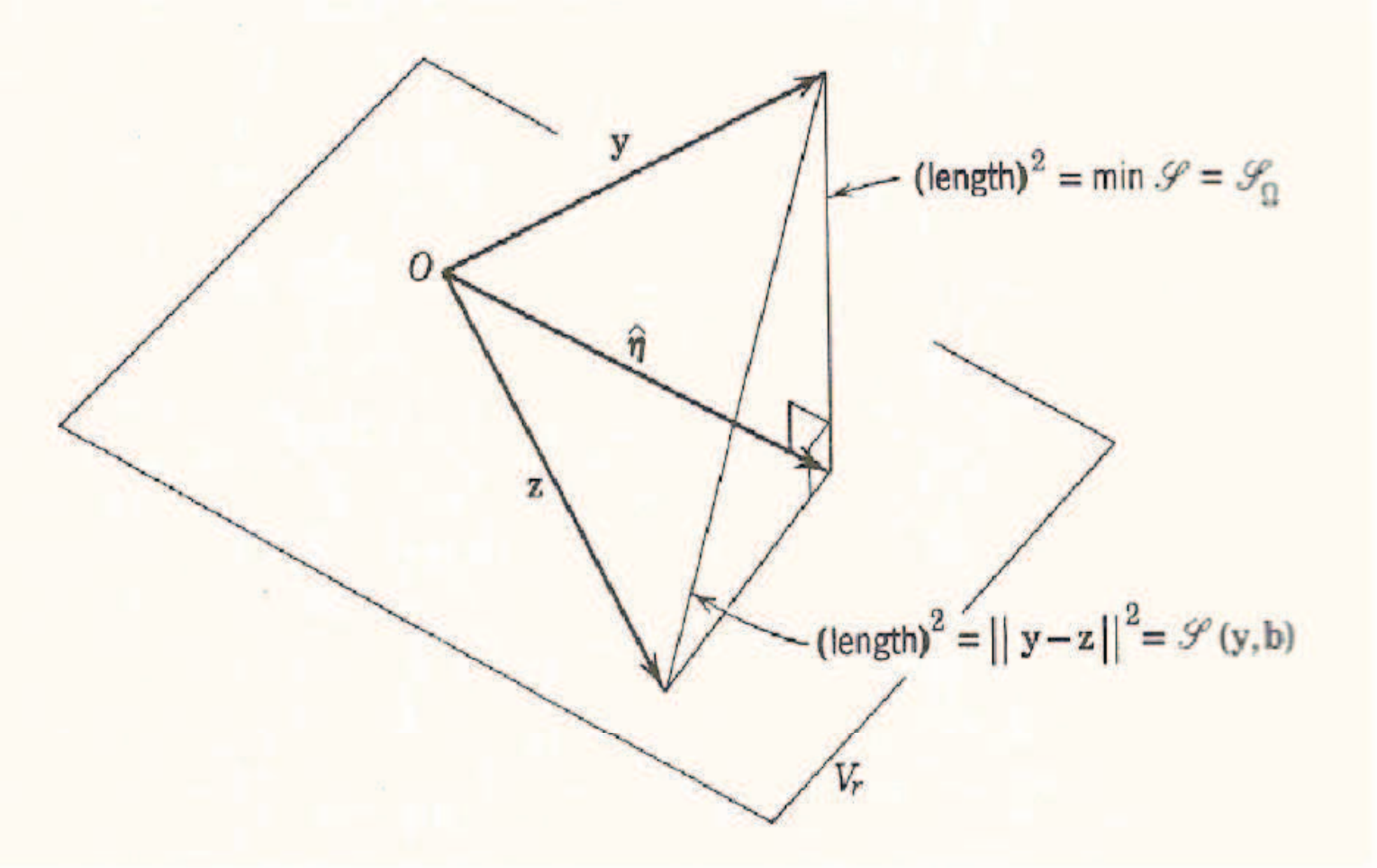
si $\mathbf{x}^i$ es la i–ésima columna de $\mathbf{X}$ entonces

$$\boldsymbol{\eta} = \beta_1\mathbf{x}^1 + \beta_2\mathbf{x}^2 + \ldots + \beta_p\mathbf{x}^p$$

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

Entonces

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


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

y sabemos que se alcanza en $\widehat{\boldsymbol{\eta}} = b_1\mathbf{x}^1 + b_2\mathbf{x}^2 + \ldots + b_p\mathbf{x}^p$ la proyección ortogonal de Y sobre V_r, que sabemos que siempre existe y es única, aunque los b_i pueden no serlo.

En términos de las ecuaciones normales tenemos que:

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$$

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

Dados $\{b_1, \ldots, b_p\}$ funciones de $\mathbf{Y}$ serán un conjunto de estimadores de mínimos cuadrados (EMC) si y sólo si $\mathbf{X}'\mathbf{b} = \widehat{\boldsymbol{\eta}}$, es decir satisfacen las ecuaciones normales.

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:

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

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

entonces

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

En consecuencia el vector de residuos es:

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

donde $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \in \Re^{n\times n}$ es la matriz de **proyección sobre el espacio generado por las columnas de X**. Suele llamarse a esta matriz de proyección **P** o **H** (hat matrix).

Propiedades de P

Matriz simétrica e idempotente, es decir: $\mathbf{P} = \mathbf{P}' = \mathbf{P}^2$. $\mathbf{I} - \mathbf{P}$ también es simétrica es idempotente, es decir también es una matriz de proyección y proyecta sobre el ortogonal de V_r.

Lema:

i) **P** y $\mathbf{I} - \mathbf{P}$ son simétricas e idempotentes

ii) $\text{rg}(\mathbf{P}) = \text{tr}(\mathbf{P}) = p$ y $\text{rg}(\mathbf{I} - \mathbf{P}) = \text{tr}(\mathbf{I} - \mathbf{P}) = n - p$

iii) $(\mathbf{I} - \mathbf{P})\mathbf{X} = \mathbf{0}$

Suma de Cuadrados

Tenemos que

$$\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \;=\; \|\mathbf{Y} - \mathbf{PY}\|^2$$

Notemos que obtenemos el Teorema de Pitágoras. En efecto,

$$\begin{aligned}
\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 = \|\mathbf{Y} - \mathbf{PY}\|^2 &= \|(\mathbf{I} - \mathbf{P})\mathbf{Y}\|^2 \\
&= \mathbf{Y}'(\mathbf{I} - \mathbf{P})'(\mathbf{I} - \mathbf{P})\mathbf{Y} \\
&= \mathbf{Y}'(\mathbf{I} - \mathbf{P})\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{PY} \\
&= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{P}'\mathbf{PY} \\
&= \|\mathbf{Y}\|^2 - \|\mathbf{PY}\|^2 \\
&= \|\mathbf{Y}\|^2 - \|\widehat{\mathbf{Y}}\|^2 = \|\mathbf{Y}\|^2 - \|\widehat{\boldsymbol{\eta}}\|^2
\end{aligned}$$

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

Propiedades del Estimador de Mínimos Cuadrados

Usando la notación matricial podemos escribir el modelo como

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

Lema: Si se cumple el modelo Ω, tenemos que

- $\widehat{\boldsymbol{\beta}}$ es un estimador insesgado de $\boldsymbol{\beta}$, es decir $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
- $\Sigma_{\widehat{\boldsymbol{\beta}}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

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

Propiedades

Bajo el modelo Ω

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \boldsymbol{\epsilon} \\ E(\boldsymbol{\epsilon}) &= 0 \\ \Sigma_{\boldsymbol{\epsilon}} &= \sigma^2\mathbf{I} \end{aligned}$$

tenemos que

- $E(\widehat{\mathbf{Y}}) = \mathbf{X}\beta$
- $\Sigma_{\widehat{\mathbf{Y}}} = \sigma^2\mathbf{P}$
- $E(\mathbf{r}) = \mathbf{0}$
- $\Sigma_{\mathbf{r}} = \sigma^2(\mathbf{I} - \mathbf{P})$

Si llamamos p_{ij} a los elementos de $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ tenemos que

$$p_{ij} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j$$

donde $\mathbf{x}_i$ representa la i–ésima fila de $\mathbf{X}$.

Luego:

$$\begin{aligned} Var(\widehat{y}_i) &= \sigma^2 p_{ii} \\ Var(r_i) &= \sigma^2(1 - p_{ii}) \\ Cov(r_i, r_j) &= -\sigma^2 p_{ij}\,, \end{aligned}$$

por lo tanto

$$Corr(r_i, r_j) = -\frac{p_{ij}}{\sqrt{1 - p_{ii}}\,\sqrt{1 - p_{jj}}}$$

Proposición: Dados $1 \leq i, j \leq n$ tenemos que

i) $0 \leq p_{ii} \leq 1$

ii) $-\frac{1}{2} \leq p_{ij} \leq \frac{1}{2}$ si $i \neq j$

Como ya vimos $Var(\widehat{y}_i) = \sigma^2 p_{ii}$, una consecuencia inmediata es que

$$Var(\widehat{y}_i) \leq Var(y_i) = \sigma^2 \, .$$

Una propiedad interesante es que **P** es invariante por transformaciones lineales no singulares de la forma $\mathbf{X} \to \mathbf{XA}$, donde $\mathbf{A} \in \Re^{p \times p}$ y $\text{rg}(\mathbf{A}) = p$. Este tipo de transformaciones es útil, por ejemplo, si queremos realizar un cambio de unidades en las covariables.

Respecto a las propiedades de invariancia, podemos ver que si

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

para $\mathbf{A} \in \Re^{p \times p}$ no singular, $\lambda \in \Re$ y $\boldsymbol{\gamma} \in \Re^p$, entonces

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}(\mathbf{XA}, \mathbf{Y}) &= \mathbf{A}^{-1}\widehat{\boldsymbol{\beta}} && \text{Invariancia por transformaciones afines} \\
\widehat{\boldsymbol{\beta}}(\mathbf{X}, \lambda\mathbf{Y}) &= \lambda\widehat{\boldsymbol{\beta}} && \text{Invariancia por cambios de escala} \\
\widehat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y} + \mathbf{X}\boldsymbol{\gamma}) &= \widehat{\boldsymbol{\beta}} + \boldsymbol{\gamma} && \text{Invariancia por cambios de regresión}
\end{aligned}$$

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

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

Bajo el modelo Ω, tenemos que

$$\boxed{s^2 = \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2}{n-p} = \frac{\|\mathbf{Y} - \mathbf{PY}\|^2}{n-p}}$$

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 $\Sigma_{\mathbf{x}}$ entonces

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

Respecto del diseño

• Covariables aleatorias

Si las covariables son aleatorias suponemos que tenemos los vectores $(\mathbf{x}_i, y_i)$ i.i.d. que satisfacen el modelo

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$$

donde los ϵ_i son i.i.d., con $E(\epsilon_i) = 0$ y $Var(\epsilon_i) = \sigma^2$ e independientes de $\mathbf{x}_i \sim F$.

El análogo de suponer que $\mathbf{X}$ tiene rango completo es asumir que la distribución de $\mathbf{x}$ no está concentrada en ningún hiperplano, es decir

$$P(\mathbf{a}'\mathbf{x} = 0) < 1 \quad \forall \mathbf{a} \neq \mathbf{0}$$

Esta condición se cumple, por ejemplo, si $\mathbf{x}$ tiene densidad.

En este caso, $\widehat{\boldsymbol{\beta}}$ está bien definido y las fórmulas que vimos para esperanza y varianza de $\widehat{\boldsymbol{\beta}}$ son válidas condicionalmente:

$$E(\widehat{\boldsymbol{\beta}}|\mathbf{X} = \mathbf{x}) = \boldsymbol{\beta} \qquad \Sigma_{\widehat{\boldsymbol{\beta}}|\mathbf{X}=\mathbf{x}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Se puede ver que si $\mathbf{V_x} = E(\mathbf{x}\mathbf{x}')$ existe, entonces para n grande la distribución aproximada de $\widehat{\boldsymbol{\beta}}$ será

$$N_p\left(\boldsymbol{\beta}, \frac{\sigma^2 \mathbf{V}_\mathbf{x}^{-1}}{n}\right)$$

Cuando el modelo tiene intercept, podemos escribirlo como:

$$y_i = \beta_0 + \mathbf{x}_i'\boldsymbol{\beta}_1 + \epsilon_i$$

donde β_0 es la intercept y $\boldsymbol{\beta}_1$ es el vector de pendientes. En este caso resulta

$$\sigma^2 \mathbf{V}_\mathbf{x}^{-1} = \sigma^2 \begin{pmatrix} 1 + \boldsymbol{\mu}_\mathbf{x}' \Sigma_\mathbf{x}^{-1} \boldsymbol{\mu}_\mathbf{x} & -\boldsymbol{\mu}_\mathbf{x}' \Sigma_\mathbf{x}^{-1} \\ -\Sigma_\mathbf{x}^{-1} \boldsymbol{\mu}_\mathbf{x} & \Sigma_\mathbf{x}^{-1} \end{pmatrix}$$

con $\boldsymbol{\mu}_\mathbf{x} = E(\mathbf{x})$ y $\Sigma_\mathbf{x}$ matriz de covarianza de $\mathbf{x}$.

• Estructura Ortogonal en la matriz de Diseño

Supongamos que podemos dividir a la matriz $\mathbf{X}$ en k conjuntos de columnas ortogonales: $\mathbf{X}_1, \ldots, \mathbf{X}_k$, de manera que

$$\mathbf{X} = [\mathbf{X}_1 \ldots \mathbf{X}_k]$$

La correspondiente división en los parámetros daría

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)'$$

Luego podemos escribir:

$$E(\mathbf{Y}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \ldots + \mathbf{X}_k\boldsymbol{\beta}_k$$

Como las columnas de $\mathbf{X}_i$ son ortogonales a las de $\mathbf{X}_j$ si $i \neq j$, tenemos que $\mathbf{X}_i'\mathbf{X}_j = 0$, luego

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & 0 & \dots & 0 \\ 0 & \mathbf{X}_2'\mathbf{X}_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \mathbf{X}_k'\mathbf{X}_k \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_1'\mathbf{Y} \\ \mathbf{X}_2'\mathbf{Y} \\ \dots \\ \mathbf{X}_k'\mathbf{Y} \end{pmatrix}$$

entonces

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y} \\ (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{Y} \\ \dots \\ (\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\mathbf{Y} \end{pmatrix} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \\ \dots \\ \widehat{\boldsymbol{\beta}}_k \end{pmatrix}$$

en consecuencia el estimador de $\boldsymbol{\beta}_i$ no cambiará si alguno de los otros $\boldsymbol{\beta}_j$ se iguala a 0, es decir si se remueve del modelo.

¿Cómo resulta la suma de cuadrados?

$$\mathbf{Y}'\mathbf{Y} - \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} = \mathbf{Y}'\mathbf{Y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} = \mathbf{Y}'\mathbf{Y} - \sum_{j=1}^{k} \widehat{\boldsymbol{\beta}}_j'\mathbf{X}_j'\mathbf{Y}$$

Por lo tanto si en el modelo ponemos algún $\boldsymbol{\beta}_i = 0$, el único cambio en la suma de cuadrados es que el término de $\widehat{\boldsymbol{\beta}}_i'\mathbf{X}_i'\mathbf{Y}$ no aparece:

$$\mathbf{Y}'\mathbf{Y} - \sum_{\substack{j=1 \\ j \neq i}}^{k} \widehat{\boldsymbol{\beta}}_j'\mathbf{X}_j'\mathbf{Y}$$

En el caso más sencillo, cada $\mathbf{X}_i$ consta de una única columna y resulta:

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

y la suma de cuadrados queda

$$\mathbf{Y}'\mathbf{Y} - \sum_{j=1}^{k} \widehat{\boldsymbol{\beta}}_j'\mathbf{X}_j'\mathbf{Y} = \mathbf{Y}'\mathbf{Y} - \sum_{j=1}^{k} \widehat{\boldsymbol{\beta}}_j^2\mathbf{X}_j'\mathbf{X}_j$$