TEMA 3 — MODELO DE REGRESIÓN LINEAL

X	Y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

var. explicativa (regresión) — respuesta var. dep (papeles no simétricos)

X_1	X_2	$\cdots$	X_k	Y
x_{11}	x_{21}	$\cdots$	x_{k1}	y_1
x_{12}	x_{22}	$\cdots$	x_{k2}	y_2
$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_{1n}	x_{2n}	$\cdots$	x_{kn}	y_n

explicativas — respuesta

<u>Recordatorio</u> (Recta de regresión) : $x_1y_1, \dots, x_ny_n$

Calcular β_0 y β_1 para las que se minimizan los errores verticales entre la nube de puntos y la recta $y = \beta_0 + \beta_1 x$

$$ECM(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

$$\longrightarrow \begin{cases} \hat{\beta}_1 = \dfrac{cov_{XY}}{V_X} \\ \hat{\beta}_0 = \bar{y} - \dfrac{cov_{XY}}{V_X}\cdot\bar{x} \end{cases}$$

Recta regresión: $y - \bar{y} = \dfrac{cov_{xy}}{V_x}(x - \bar{x}) \longleftrightarrow y = \hat{\beta}_0 + \hat{\beta}_1 x$

Errores/residuos: $e_i = y_i - \underbrace{(\hat{\beta}_0 + \hat{\beta}_1 x_i)}_{= \hat{y}_i} \qquad i = 1, \dots, n$

$$ECM(\hat{\beta}_0, \hat{\beta}_1) = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

1. Modelo de regresión lineal simple

$$\begin{array}{cc} X & Y \\ x_1 & y_1 \\ | & | \\ x_n & y_n \end{array} \qquad Y|X=x := \beta_0 + \beta_1 x + \varepsilon$$

$\mathbb{E}(\varepsilon) = 0$

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

Habitual: $\varepsilon \sim N(0,\sigma^2)$

Recuerdo: (X,Y) normal bidim. $\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{yx} & \sigma_y^2 \end{pmatrix}$

$Y|X=a \sim N(\tilde{\mu}, \tilde{\sigma}^2)$

$$\begin{cases} \tilde{\mu} = \mu_Y + \sigma_{xy} \dfrac{1}{\sigma_x^2}(a-\mu_x) & \longleftarrow \text{lineal en } a \\ \tilde{\sigma}^2 = \sigma_Y^2 - \dfrac{\sigma_{xy}}{\sigma_x^2} & \longleftarrow \text{no depende de } a \end{cases}$$

Modelo:

$$\begin{array}{cc} x_1 & y_1 \\ | & | \\ x_n & y_n \end{array}$$

$\underbrace{x_1, \dots, x_n}_{\text{fijos}}$

variables aleatorias

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, \dots, n$

Condiciones:

$\longrightarrow \mathbb{E}(\varepsilon_i) = 0 \quad \forall i$

$\longrightarrow V(\varepsilon_i) = \sigma^2 \quad \forall i$

$\longrightarrow \operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j$

Muy habitual: $\varepsilon_i \sim N(0,\sigma^2)$ indep.

Versión matricial

$$\underbrace{\begin{pmatrix} y_1 \\ | \\ y_n \end{pmatrix}}_{\mathbb{Y}} = \underbrace{\begin{pmatrix} 1 & x_1 \\ | & | \\ 1 & x_n \end{pmatrix}}_{\mathbb{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}_{\vec{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ | \\ \varepsilon_n \end{pmatrix}}_{\vec{\varepsilon}}$$

$\mathbb{Y} = \mathbb{X}\vec{\beta} + \vec{\varepsilon}$

$\mathbb{E}(\vec{\varepsilon}) = \vec{0} \qquad \operatorname{cov}(\vec{\varepsilon}) = \sigma^2 I_n$

$\Rightarrow \mathbb{E}(\mathbb{Y}) = \mathbb{X}\cdot\vec{\beta}$

$\operatorname{cov}(\mathbb{Y}) = \sigma^2 I_n$

Habitual: $\vec{\varepsilon} = N(\vec{0}, \sigma^2 I_n) \Longrightarrow \mathbb{Y} \sim N(\mathbb{X}\cdot\vec{\beta}, \sigma^2 I_n)$

Parámetros: $\beta_0, \beta_1, \sigma^2$

1.1. - ESTIMACIÓN para $\vec{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ (Mínimos cuadrados)

Buscamos $\vec{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ para el que $f(\vec{\beta}) = (\vec{y} - \vec{x}\vec{\beta})^T(\vec{y} - \vec{x}\vec{\beta}) =$

$$= \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

$$f(\vec{\beta}) = y^T y - y^T x\beta - \beta^T x^T y + \beta^T x^T x \beta =$$

$$= y^T y - 2\beta^T x^T y + \beta^T x^T x \beta$$

$$\frac{\partial f}{\partial \beta} = -2x^T y + 2x^T x \beta = \vec{0} \iff x^T x \beta = x^T y \iff$$

$$\iff \boxed{\hat{\beta} = (x^T x)^{-1} x^T . y}$$

Obs: $x^T x = \begin{pmatrix} 1 & \text{—} & 1 \\ x_1 & \text{—} & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ | & | \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & n\overline{x^2} \end{pmatrix} = n \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{pmatrix}$

$$\det(x^T x) \longleftrightarrow V_x \neq 0$$

$$(x^T x)^{-1} = \frac{1}{n V_x} \begin{pmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

$$\begin{cases} \hat{\beta}_1 = \dfrac{1}{n V_x} \displaystyle\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \dfrac{\mathrm{cov}_{xy}}{V_x} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

1.2. - VALORES AJUSTADOS, RESIDUOS y SUMAS DE CUADRADOS

Ya tenemos $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$ $\qquad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad i = 1, \dots, n$

$$\hat{y} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} \longrightarrow \hat{y} = x\hat{\beta} = \underbrace{x(x^T x)^{-1} x^T}_{= H \text{ matriz hat}} y = Hy$$

H es → simétrica $H^T = H$
→ idempotente $H^2 = H$

Residuos: $\begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \longrightarrow e = y - \hat{y} = \underbrace{(I_n - H)}_{\text{simétrica, idempotente}} y$ $\qquad e_i = y_i - \hat{y}_i \quad i = 1, \dots, n$

Obs: 1) $X^T \cdot e = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, porque $X^T (I_n - H) y =$

$$= \underbrace{\left(X^T - X^T X (X^T X)^{-1} X^T \right)}_{=0} y = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

También sabemos $\begin{pmatrix} 1 & \text{—} & 1 \\ x_1 & \text{—} & x_n \end{pmatrix} \cdot e = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow$

$$\Rightarrow \begin{cases} 1^T \cdot e = 0 \Longrightarrow \sum_{i=1}^n e_i = 0 \\ x^T e = 0 \Longrightarrow \sum_{i=1}^n x_i e_i = 0 \end{cases}$$

2) $\hat{y}^T \cdot e = 0$, porque $[H \cdot (I_n - H)] y = H - H^2 = H - H = 0$

¿Cuán grandes son los residuos?

$$e^T \cdot e = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)^2 = y^T (I_n - H) y$$

Observar que $\sum e_i^2$ dividido por $\begin{cases} n \\ n-2 \end{cases}$, será estimación de σ^2

Observación sobre sumas de cuadrados:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n \left((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \right)^2 + 2 \underbrace{\sum_{i=1}^n e_i (\hat{y}_i - \bar{y})}_{=0}$$

(salvo n, es V_y variación total de las y_i)

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\substack{\text{suma de cuadrados total} \\ (\text{TSS} - \text{SCT})}} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\substack{\text{suma de cuadrados explicada por modelo} \\ (\text{MSS} - \text{SCM})}} + \underbrace{\sum_{i=1}^n \underbrace{(y_i - \hat{y}_i)^2}_{e_i^2}}_{\substack{\text{suma de cuadrados residual} \\ (\text{RSS} - \text{SCR})}}$$

Se define entonces el coeficiente de determinación:

$$R^2 = \frac{SCM}{SCT} = 1 - \frac{SCR}{SCT}$$

<u>Obs</u>: para el caso bidimensional $R^2 = \rho_{xy}^2$

1.3 – Propiedades de los estimadores

no $\begin{matrix} x_1 & y_1 \\ | & | \\ x_n & y_n \end{matrix}$ sino $\begin{matrix} x_1 & Y_1 \\ | & | \\ x_n & Y_n \end{matrix}$ $= \mathbb{Y}$

$$\mathbb{Y} = X.\beta + \varepsilon = \begin{pmatrix} \varepsilon_1 \\ | \\ \varepsilon_n \end{pmatrix}$$

$$\begin{cases} \mathbb{E}(\varepsilon) = 0 \\ \mathrm{cov}(\varepsilon) = \sigma^2 I_n \end{cases}$$

quizás en algún momento $\varepsilon \sim N(0, \sigma^2 I$

$$\begin{cases} \mathbb{E}(\mathbb{Y}) = X.\beta \\ \mathrm{cov}(\mathbb{Y}) = \sigma^2 I_n \end{cases}$$

Tenemos <u>estimadores</u>:

$$\hat{\beta} = (X^T X)^{-1} X^T \mathbb{Y}$$

$$\begin{cases} \hat{\beta}_1 = \dfrac{1}{nV_x} \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) \\ \hat{\beta}_0 = \bar{Y} - \bar{x}\hat{\beta}_1 \end{cases}$$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

Miramos algunas propiedades, empezando por la media:

$$\mathbb{E}(\hat{\beta}_1) = \frac{1}{nV_x} \sum_{i=1}^{n} (x_i - \bar{x}) \underbrace{\mathbb{E}(Y_i)}_{\beta_0 + \beta_1 x_i} = \frac{1}{nV_x} \Big[\beta_0 \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})}_{=0} + \beta_1 \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})}_{= nV_x} \Big]$$

$= \beta_1$ ($\hat{\beta}_1$ es un estimador insesgado de β_1)

$$\mathbb{E}(\hat{\beta}) = (X^T X)^{-1} X^T \underbrace{\mathbb{E}(\mathbb{Y})}_{X.\beta} = \beta$$

($\hat{\beta}$ estimador insesgado de β)

Hemos hecho los cálculos en una línea usando la notación matricial.

$$\mathrm{Cov}(\hat{\beta}) = (X^T X)^{-1} X^T \underbrace{\mathrm{cov}(\mathbb{Y})}_{\sigma^2 I_n} \left[(X^T X)^{-1} X^T \right]^T = \sigma^2 (X^T X)^{-1} X^T . X (X^T X)^{-1} =$$

$$= \sigma^2 (X^T X)^{-1} = \sigma^2 \frac{1}{nV_x} \begin{pmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

caso bidim.

— RECUERDO:

$$\begin{matrix} x_1 & y_1 \\ | & | \\ x_n & y_n \end{matrix} \qquad \mathbb{Y} = X\beta + \varepsilon \qquad \mathbb{Y} = \begin{pmatrix} Y_1 \\ | \\ Y_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$\mathbb{E}(\varepsilon) = \vec{0}$$

$$\operatorname{cov}(\varepsilon) = \sigma^2 I_n \qquad X = \begin{pmatrix} 1 & x_1 \\ | & | \\ 1 & x_n \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ | \\ \varepsilon_n \end{pmatrix}$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \qquad \hat{\beta} = (X^T X)^{-1} X^T \mathbb{Y} \longrightarrow \begin{cases} \mathbb{E}(\hat{\beta}) = \beta \\ \operatorname{cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \end{cases}$$

$\mathbb{E}(\hat{\beta}_0) = \beta_0$, $\mathbb{E}(\hat{\beta}_1) = \beta_1$ ✓ insesgados

$$V(\hat{\beta}_0) = \frac{\sigma^2}{nV_x}\overline{x^2} = \frac{\sigma^2}{nV_x}\left[\frac{1}{n} + \frac{\overline{x}^2}{nV_x}\right]$$

$$V(\hat{\beta}_1) = \frac{\sigma^2}{nV_x}$$

$$\operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\overline{x}}{nV_x}\sigma^2$$

¿Estimador para σ^2?

$$e^T e = \sum_{i=1}^{n} e_i = \mathbb{Y}^T (I_n - H)\mathbb{Y} \qquad H = X(X^TX)^{-1}X^T$$

$$\operatorname{rang}(H) = \operatorname{tr}(H) = \operatorname{tr}(X(X^TX)^{-1}X^T) = \operatorname{tr}(X^TX(X^TX)^{-1}) =$$

$$= \operatorname{tr}(I_2) = 2$$

$$\operatorname{rang}(I_n - H) = \operatorname{tr}(I_n - H) = n-2$$

¿Media de $\mathbb{Y}^T(I_n - H)\mathbb{Y}$?

general: A simétrica, $n\times n$

$$\begin{pmatrix} X_1 \\ | \\ X_n \end{pmatrix} = \mathbb{X} \text{ vector aleat.} \begin{cases} \mathbb{E}(\mathbb{X}) = \mu = \begin{pmatrix} \mu_1 \\ | \\ \mu_n \end{pmatrix} \\ \operatorname{cov}(\mathbb{X}) = \Sigma \end{cases}$$

forma cuadrática $\mathbb{X}^T A \mathbb{X} \longrightarrow$ variable aleatoria

$$\boxed{E(X^T A X) = tr(A\Sigma) + \mu^T A \mu}$$

ALGO QUE TUVIMOS QUE HACER EN EL TEMA 1

Comentario: A simétrica, $X \begin{cases} M \\ \Sigma \end{cases}$

$$\boxed{V(X^T A X) = 2\,tr(A\Sigma A\Sigma) + 4\mu^T A \Sigma A \mu}$$

En nuestro caso $Y \begin{cases} \vec{0} \\ \sigma^2 I_n \end{cases}$

$$V(Y^T(I_n - H)Y) = 2\sigma^4 \underbrace{tr(I_n - H)}_{=n-2}$$

$$V(S_R^2) = \frac{2\sigma^4}{n-2}$$

$$Z = \frac{(n-2)S_R^2}{\sigma^2} \begin{cases} E(Z) = n-2 \\ V(Z) = 2(n-2) \end{cases}$$

¿Distribución de $\hat{\beta}_0, \hat{\beta}_1, S_R^2$? A partir de ahora añadimos esta hipótesis:

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \sim N(\vec{0}, \sigma^2 I_n)$$

$$Y \sim N(X\beta, \sigma^2 I_n)$$

Resulta: $\hat{\beta} = (X^T X)^{-1} X^T Y \sim N_2(\beta, \sigma^2 (X^T X)^{-1})$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{nV_x}\right)$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{nV_x}\overline{x^2}\right)$$

Corolario 2, tema 1

$$\frac{(n-1)S_R^2}{\sigma^2} = \frac{1}{\sigma^2} Y^T (I_n - H) Y \sim \chi^2_{n-2}$$

TEOREMA: En regresión lineal simple + normalidad:

- $\hat{\beta} \sim N\left(\beta, \sigma^2 (X^T X)^{-1}\right)$
- $\dfrac{(n-2) S_R^2}{\sigma^2} \sim \chi^2_{n-2}$
- $\hat{\beta}$ y S_R^2 son independientes

Obs:
$$\frac{\hat{\beta}_1 - \beta_1}{S_R \sqrt{1/nV_x}} = \frac{\overbrace{\dfrac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{nV_x}}}^{\sim N(0,1)}}{\underbrace{\sqrt{\dfrac{S_R^2 (n-2)}{\sigma^2}}}_{\sqrt{\chi^2_{n-2}}}} \sqrt{n-2} \sim t_{n-2}$$

COROLARIO:
$$\frac{\hat{\beta}_1 - \beta_1}{S_R \sqrt{1/nV_x}} \sim t_{n-2}, \quad \frac{\hat{\beta}_0 - \beta_0}{S_R \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{nV_x}}} \sim t_{n-2}$$

Podemos ahora conseguir intervalos de confianza para $\beta_0, \beta_1 \Longrightarrow$
$\Longrightarrow$ percentil de t-Student.

Observación (Un último detalle, sobre sumas de cuadrados)

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i^2 + \bar{Y}^2 - 2Y_i\bar{Y}) = \sum_{i=1}^{n} Y_i^2 + n\bar{Y}^2 - 2n\bar{Y}^2 =$$
$$= \sum_{i=1}^{n} Y_i^2 - n\bar{Y}_i^2 = \left(1 - \frac{1}{n}\right)\sum_{i=1}^{n} Y_i^2 - \frac{1}{n}\sum_{i \neq j} Y_i Y_j$$

Adicionalmente: $J_n = \text{ones}(n) = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \longrightarrow \underbrace{\frac{1}{n} J_n}_{\text{rango } 1}$ $\nearrow$ simétrica, $\searrow$ idempotente

$$\Longrightarrow TSS = Y^T \underbrace{\left(I_n - \tfrac{1}{n}J_n\right)}_{\text{simétrica, idempotente, rango } n-1} Y$$

$= RSS + MSS$

$\Rightarrow RSS = Y^T (I_n - H) Y \longrightarrow$ rango $n-2$

$\Rightarrow MSS = Y^T \left(H - \tfrac{1}{n}J_n\right) Y \longrightarrow$ rango 1

Como $(I_n - H)(H - \frac{1}{n}J_n) = H - \frac{1}{n}J_n - H + \frac{1}{n}\underbrace{HJ_n}_{=J_n} = 0$

$$H = X(X^TX)^{-1}X^T$$

$$HX = X(X^TX)^{-1}X^TX = X$$

$$X = \begin{pmatrix} 1 & x_1 \\ | & | \\ 1 & x_n \end{pmatrix} \qquad H\begin{pmatrix}1\\|\\1\end{pmatrix} = \begin{pmatrix}1\\|\\1\end{pmatrix} \longrightarrow HJ_n = J_n$$

$\Rightarrow$ Por el teorema 3, RSS y MSS son independientes.

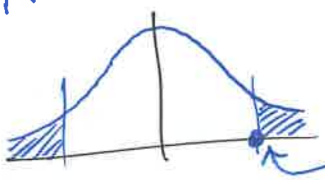
Con todos los preliminares expuestos, empezamos la Estadística

$$\underbrace{\beta_0, \beta_1, \sigma^2}_{\text{desconocidos}} \xrightarrow{\text{generan}} \begin{matrix} x_1 & y_1 \\ | & \\ x_n & y_n \end{matrix} \longrightarrow \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \qquad S_R^2 = \frac{1}{n-2}\sum e_i^2$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad \hat{y} = \begin{pmatrix} \hat{y}_1 \\ | \\ \hat{y}_n \end{pmatrix} \longrightarrow e = \begin{pmatrix} e_1 \\ | \\ e_n \end{pmatrix}$$

1. Intervalos de confianza

Usamos el corolario que relaciona $\hat{\beta}_1, \hat{\beta}_0$ con t-Student.



$t_{\{n-2;\alpha/2\}}$

$$\mathbb{P}\left(\left| \frac{\hat{\beta}_1 - \beta_1}{S_R\sqrt{1/nV_x}} \right| > t_{\{n-2;\alpha/2\}} \right) = \alpha$$

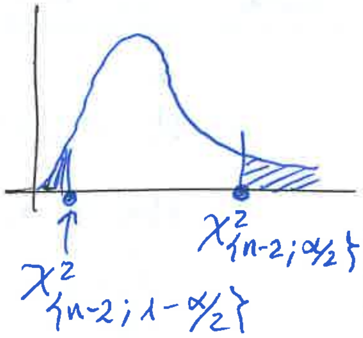
$$1-\alpha = \mathbb{P}\left(\hat{\beta}_1 - t_{\{n-2;\alpha/2\}} S_R\sqrt{1/nV_x} \leq \beta_1 \leq \hat{\beta}_1 + t_{\{n-2;\alpha/2\}} S_R \sqrt{1/nV_x} \right)$$

$$\Rightarrow \boxed{IC_{1-\alpha}(\beta_1) = \hat{\beta}_1 \pm t_{\{n-2;\alpha/2\}} S_R \sqrt{1/nV_x}}$$

Por otro lado, desarrollando análogamente:

$$\Rightarrow \boxed{IC_{1-\alpha}(\beta_2) = \hat{\beta}_2 \pm t_{\{n-2;\alpha/2\}} S_R \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{nV_x}}}$$

Ahora usamos que $\frac{(n-2)S_R^2}{\sigma^2} \sim \chi^2_{n-2}$



$$\mathbb{P}\left(\chi^2_{\{n-2;1-\alpha/2\}} \leq \frac{(n-2)S_R^2}{\sigma^2} \leq \chi^2_{\{n-2;\alpha/2\}}\right) = 1-\alpha$$

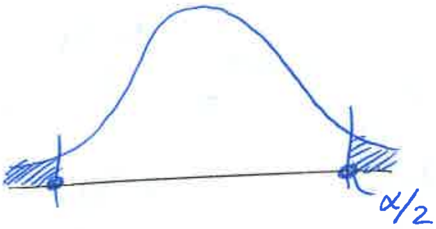
$$\mathbb{P}\left(\frac{(n-2)S_R^2}{\chi^2_{\{n-2;\alpha/2\}}} \leq \sigma^2 \leq \frac{(n-2)S_R^2}{\chi^2_{\{n-2;1-\alpha/2\}}}\right) = 1-\alpha$$

$$\Rightarrow \boxed{IC_{1-\alpha}(\sigma^2) = \left(\frac{(n-2)S_R^2}{\chi^2_{\{n-2;\alpha/2\}}}, \frac{(n-2)S_R^2}{\chi^2_{\{n-2;1-\alpha/2\}}}\right)}$$

② Contraste de hipótesis

$H_0: \beta_1 = 0$

Bajo H_0, $\frac{\hat{\beta}_1}{S_R\sqrt{1/nV_x}} \sim t_{n-2}$



Región de rechazo (con nivel de significación α):

$$\boxed{R = \left\{\left|\frac{\hat{\beta}_1}{S_R\sqrt{1/nV_x}}\right| > t_{\{n-2;\alpha/2\}}\right\}}$$

Alternativa:

$$\left.\begin{array}{l} RSS = Y^T(I_n - H)Y \\ MSS = Y^T(H - \frac{1}{n}I_n)Y \end{array}\right\} \text{— independientes}$$

Si $\beta_1 = 0$: $\quad X\beta = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}\begin{pmatrix} \beta_0 \\ 0 \end{pmatrix} = \beta_0\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$

$$Y \sim N\left(\beta_0\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \sigma^2 I_n\right) = \beta_0\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \underbrace{X}_{\sim N(\vec{0},\sigma^2 I_n)}$$

$$Y^T(I_n - H)Y = \beta_0^2(1\ldots1)(I_n - H)\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \sigma^2 I_n X^T(I_n - H)X \sim \chi^2_{n-2}$$

$$\Rightarrow \left.\begin{array}{l} RSS \sim \chi^2_{n-2} \\ MSS \sim \chi^2_1 \end{array}\right\} \text{indep.} \Rightarrow \frac{MSS/1}{RSS/(n-2)} \sim F_{\{1;n-2\}}$$

F de Fisher

Región rechazo: $\boxed{R = \left\{\frac{MSS}{RSS/(n-2)} > F_{\{1;n-2;\alpha\}}\right\}}$

Queremos estimar $\mathbb{E}(Y|X=x_0) = \beta_0 + \beta_1 x_0$

Lo hacemos con $Z = \hat{\beta}_0 + \hat{\beta}_1 x_0$

$\longrightarrow \mathbb{E}(Z) = \beta_0 + \beta_1 x_0$

$\longrightarrow V(Z) = V(\hat{\beta}_0) + x_0^2 V(\hat{\beta}_1) + 2x_0 \operatorname{Cov}(\hat{\beta}_0, \hat{\beta}_1)$

$$= \sigma^2 \left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{nV_x} \right]$$

Así que
$$\frac{Z - (\beta_0 + \beta_1 x_0)}{S_R \sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{nV_x}}} \sim t_{n-2}$$

Por tanto, el intervalo de confianza para la media condicionada:

$$\boxed{IC_{1-\alpha}(\beta_0 + \beta_1 x_0) = (\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\{n-2;\alpha/2\}} S_R \sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{nV_x}}}$$

Queremos ahora estimar $Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$ $\sim N(0,\sigma^2)$ indep. ← nuevo

Lo hacemos con $Z = \hat{\beta}_0 + \hat{\beta}_1 x_0$

$\longrightarrow \mathbb{E}(Y_0 - Z) = 0$

$\longrightarrow V(Y_0 - Z) = \underbrace{V(Y_0)}_{=\sigma^2} + V(Z) - 2\underbrace{\operatorname{Cov}(Y_0, Z)}_{=0}$

$$= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{nV_x} \right]$$

$$\Longrightarrow \frac{Y_0 - Z}{S_R \sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{nV_x}}} \sim t_{n-2}$$

Por tanto, el intervalo de confianza para la predicción del valor:

$$\boxed{IC_{1-\alpha}(Y_0) = (\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{\{n-2;\alpha/2\}} S_R \sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{nV_x}}}$$

REGRESIÓN LINEAL MÚLTIPLE ($\vec{x}_j \in \mathbb{R}^d$)

Datos $\longrightarrow$ $\underbrace{X_1, \dots, X_k}_{\text{var. regresoras}}, \underbrace{Y}_{\text{var. respuesta}}$ $\quad k=1 \Rightarrow$ RL simple

$$n \left\{ \begin{array}{c} X_1 \cdots X_k \quad Y \\ \underbrace{\begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}}_{k} \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{1} \end{array} \right.$$

Consideramos $\boxed{n \geq k+2}$

Suponemos columnas lin. independientes

Modelo genérico: $Y | X_1 = x_1 \cdots X_k = x_k \equiv \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \qquad \mathbb{Y} = \mathbb{X} \cdot \beta + \varepsilon$$

(notación matricial)

Hipótesis sobre ε: $\begin{cases} \mathbb{E}(\varepsilon) = \vec{0} \\ \operatorname{cov}(\varepsilon) = \sigma^2 I_n \end{cases}$ Casi siempre $\varepsilon \sim N(\vec{0}, \sigma^2 I_n)$

1.) ESTIMADORES PARA PARÁMETROS

Con el mismo argumento (mínimos cuadrados) que en el caso simple, se obtienen los estimadores:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} \text{ de } \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} \text{ dados por } \boxed{\hat{\beta} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{Y}}$$

Obs: $\mathbb{X}$ es $n \times (k+1)$: $\underbrace{\boxed{X^T}}_{(k+1)\times n} \underbrace{\boxed{}}_{n\times(k+1)} = \underbrace{\boxed{}}_{(k+1)(k+1)}$ simétrica y def. pos.

Esto es consecuencia de un resultado general:
Sea B matriz $n \times p$, rango p. $A = B^T B$ es $p \times p$ y simétrica.
Vamos a ver que A es def. pos. Consideramos $\vec{y}^T A \vec{y} \overset{?}{>} 0$
$y^T A y = y^T B^T B y = (By)^T (By) = \|By\|^2 \geq 0$ y es cero solo si
$By = \vec{0}$ (como B rango máx. solo cuando $\vec{y} = \vec{0}$).

Obs: Si suponemos normalidad, los estimadores por máx. vero son los mismos.

TEOREMA (Gauss-Markov): $\mathbb{Y} = \mathbb{X}\cdot\beta + \varepsilon$, $\mathbb{X}$ rango máximo

$$\begin{cases} \mathbb{E}(\varepsilon) = 0 \\ cov(\varepsilon) = \sigma^2 I_n \end{cases}$$

Cada estimador $\hat{\beta}_j$ (de mínimos cuadrados) $j = 0 \dots k$ tiene mínima varianza de entre todos los estimadores lineales e insesgados de β_j.

demostración: El estimador de mínimos cuadrados es $\hat{\beta} = (X^T X)^{-1} X^T \mathbb{Y}$. Comparamos con cualquier otro:

$\tilde{\beta} = A.\mathbb{Y}$ Para que sea insesgado $\mathbb{E}(\tilde{\beta}) = A.\mathbb{E}(\mathbb{Y}) =$

$= \underbrace{AX\cdot\beta = \beta}$ para todo β.

$\Longrightarrow AX = I_{k+1}$

las varianzas de cada $\tilde{\beta}_j$ están en la diagonal

Calculamos $cov(\tilde{\beta}) = cov(A\mathbb{Y}) = A \overbrace{cov(\mathbb{Y})}^{\sigma^2 I_n} A^T = \sigma^2 A A^T$.

¿Qué A minimiza los elementos de esa diagonal?

$$AA^T = \left(A - (X^TX)^{-1}X^T + (X^TX)^{-1}X^T\right)\cdot\left(A - (X^TX)^{-1}X^T + (X^TX)^{-1}X^T\right)^T =$$

$$= \left(A - (X^TX)^{-1}X^T\right)\left(A - (X^TX)^{-1}X^T\right)^T +$$

$$+ (X^TX)^{-1}X^TX(X^TX)^{-1} + \underbrace{\left(A - (X^TX)^{-1}X^T\right)X(X^TX)^{-1}}_{=0} +$$

$$+ \underbrace{\boxed{}}_{=0} =$$

$$= \underbrace{\left[A - (X^TX)^{-1}X^T\right]\left[A - (X^TX)^{-1}X^T\right]^T}_{\text{matriz semidef. pos.}} + (X^TX)^{-1}$$

matriz semidef. pos. $\longrightarrow$ los elementos de la diagonal son ≥ 0 $\Rightarrow$

$\Rightarrow$ se hacen 0 si $A = (X^TX)^{-1}X^T$ ∎

2. PRONÓSTICO Y RESIDUOS

$$\boxed{\hat{Y} = X\cdot\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}_{\substack{\shortparallel \\ H}} \cdot Y = HY}$$

H matriz hat $n\times n$
- $HX = X$
- simétrica
- idempotente
- rango es $k+1$

Los residuos son: $\boxed{\begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = \vec{e}} = Y - \hat{Y} = \underbrace{(I_n - H)}_{\text{rango } n-k-1} Y =$

$= (I_n - H)(X\cdot\beta + \varepsilon) =$

$= X\cdot\beta + \varepsilon - HX\beta - H\varepsilon =$

$= \boxed{(I_n - H)\varepsilon}$

Propiedades

- Perpendicularidad:

  $X^T\cdot e = 0_{k+1}$

  $\hat{Y}^T\cdot e = 0$

- $\mathbb{E}(e) = \vec{0}$

$\mathrm{cov}(e) = \mathrm{cov}((I_n - H)\varepsilon) = (I_n - H)\overbrace{\mathrm{cov}(\varepsilon)}^{\sigma^2 I_n}(I_n - H)^T =$

$= \sigma^2(I_n - H)$ → no es diagonal!

$$SCR = e^Te = \sum_{i=1}^{n} e_i^2 = Y^T(I_n - H)Y$$

El estimador para σ^2 sería $S_R^2 = \frac{1}{n-k-1} SCR =$

$$= \frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-k-1} Y^T(I_n - H)Y$$

3. SUMAS DE CUADRADOS

$$SCT = TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \mathbb{Y}^T \Big(I_n - \frac{1}{n} \overbrace{J_n}^{\text{matriz de unos}}\Big) \mathbb{Y} \quad (\text{RANGO } n-1)$$

$$SCM = MSS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \mathbb{Y}^T \Big(H - \frac{1}{n} J_n\Big) \mathbb{Y} \quad (\text{RANGO } k)$$

$$SCR = RSS = \sum_{i=1}^{n} (Y_i - \hat{Y})^2 = \mathbb{Y}^T (I_n - H) \mathbb{Y} \quad (\text{RANGO } n-k-1)$$

las 3 matrices son sim. e idempot

$$\boxed{TSS = MSS + RSS}$$

$$R^2 = \frac{MSS}{TSS} = 1 - \frac{RSS}{TSS} \quad \text{entre } 0 \text{ y } 1$$

Nota: si se añaden variables regresoras (sin influencia sobre Y), aumenta R^2.

Por esta razón hay quien usa: $\overline{R^2} = 1 - \dfrac{RSS/_{n-k-1}}{TSS/_{n-1}}$

Necesitamos distribución de los estimadores.

MODELO NORMAL: Suponemos que $\varepsilon \sim N_n(0, \sigma^2 I_n)$.

Por tanto, $\mathbb{Y} \sim N(X \cdot \beta, \sigma^2 I_n)$.

Como $\hat{\beta} = (X^T X)^{-1} X^T \mathbb{Y}$, $\quad \hat{\beta} \sim N_{k+1}\big(\beta, \sigma^2 (X^T X)^{-1}\big)$

Como $SCR = \mathbb{Y}^T \underbrace{(I_n - H)}_{\text{rango } n-k-1} \mathbb{Y}$, $\quad H = X (X^T X)^{-1} X^T$

y además

$$(X\beta)^T (I_n - H)(X \cdot \beta) = \beta^T X (I_n - H) X \cdot \beta =$$

$$= \beta^T X^T X \beta - \beta^T \cancel{X^T X} \cancel{(X^T X)^{-1}} X^T X \beta = 0$$

Concluimos que

$$\frac{SCR}{\sigma^2} = \frac{(n-k-1) S_R^2}{\sigma^2} \sim \chi^2_{n-k-1}$$

Además, como $(X^TX)^{-1}X(I_n - H) = 0$

resulta que $\hat{\beta}$ y SCR (o bien S_R^2) son <u>independientes</u>.

Llamemos $\underset{\substack{\uparrow \\ (K+1)\times(K+1)}}{(X^TX)^{-1}} = \begin{pmatrix} q_{00} & & \xi \\ & q_{11} & \\ \xi & & \ddots \; q_{KK} \end{pmatrix}$ q_{ii} es el elemento en la posición $i+1$ de esta matriz.

Cada $\hat{\beta}_j \sim N(\beta_j, \sigma^2 q_{jj})$ $\quad j = 0, \dots, K$

Y por tanto, $\dfrac{\hat{\beta}_j - \beta_j}{S_R\sqrt{q_{jj}}} \sim t_{n-K-1}$ $\quad j = 0, \dots, K$.

Inmediatamente, dada una muestra $\left.\begin{matrix} 1 & \cdots \\ | & \\ 1 & \cdots \end{matrix}\right| \begin{matrix} y_1 \\ \vdots \\ y_n \end{matrix}$

A) $\boxed{IC_{1-\alpha}(\beta_j) = \hat{\beta}_j \pm t_{\{n-K-1;\, \alpha/2\}} S_R\sqrt{q_{jj}} \qquad j = 0, \dots, K}$

B) $H_0 : \beta_j = 0 \qquad j = 1, \dots, K$

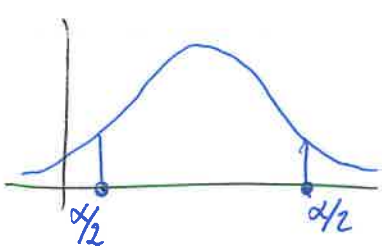
Región de rechazo para nivel de significación α

$$\boxed{R_j(\alpha) = \left\{ \left| \frac{\hat{\beta}_j}{S_R\sqrt{q_{jj}}} \right| > t_{\{n-K-1;\, \alpha/2\}} \right\}}$$

<u>Nota 1</u>: Bajo normalidad, $\hat{\beta}$ es estimador max. VERO de β

Gauss-Markov dice que $\hat{\beta}$ es de mínima varianza de entre estimadores lineales e <u>insesgados</u> de β.
(bajo normalidad)

c) Intervalo de confianza para σ^2:



$$\boxed{IC_{1-\alpha}(\sigma^2) = \left(\frac{(n-K-1)S_R^2}{\chi^2_{\{n-K-1;\, \alpha/2\}}}, \frac{(n-K-1)S_R^2}{\chi^2_{\{n-K-1;\, 1-\alpha/2\}}} \right)}$$

CONTRASTE GLOBAL DE LA REGRESIÓN

$$\boxed{H_0: \beta_1 = \cdots = \beta_K = 0}$$

Bajo H_0, $\mathbb{Y} \sim N(X\cdot\beta, \sigma^2 I_n) \sim N\left(\beta_0 \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}, \sigma^2 I_n\right)$

Se analizan: por un lado $\dfrac{RSS}{\sigma^2} = \dfrac{1}{\sigma^2} \mathbb{Y}^T (I_n - H)\mathbb{Y} \sim \chi^2_{n-K-1}$ tema 1 tma. 1

por otro lado $\dfrac{MSS}{\sigma^2} = \dfrac{1}{\sigma^2} \mathbb{Y}^T \big(\underbrace{H - \tfrac{1}{n} J_n}_{\text{rango } K}\big)\mathbb{Y} \sim \chi^2_K$

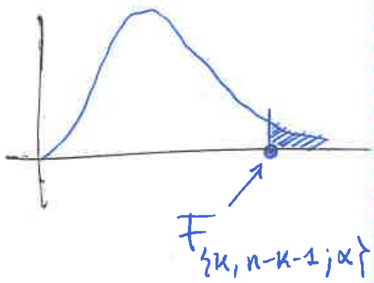
En general, no es χ^2_K pero bajo H_0, sí:

$$\beta_0^2 (1 \cdots 1) \underbrace{\left(H - \tfrac{1}{n} J_n\right)}_{} \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} = 0$$ corolario 2, tema 1

recuerdo $\longrightarrow H \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} = \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$

$\hookrightarrow \frac{1}{n} J_n \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} = \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix}$

Conclusión, bajo H_0: $$\boxed{\dfrac{MSS/K}{RSS/(n-K-1)} \sim F_{K, n-K-1}}$$



Región de rechazo para H_0 con nivel de sign. α:

$$\boxed{R = \left\{ \dfrac{MSS/K}{RSS/(n-K-1)} > F_{\{K, n-K-1; \alpha\}} \right\}}$$

Habitual: tabla ANOVA

	Grados	Sumas cuadrados		
MODELO (regresión)	K	MSS	MSS/K	$\dfrac{MSS/K}{RSS/n-K-1}$
RESIDUOS	$n-K-1$	RSS	$RSS/n-K-1$	p-valor
TOTAL	$n-1$	TSS		

Predicción

X_1	$\cdots$	X_K	Y
X_{11}	$\cdots$	X_{K1}	y_1
$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_{n1}	$\cdots$	X_{Kn}	y_n

Nuevo vector $x_0^T = (x_{01}, \dots, x_{0K})$

¿y_0?

$\Rightarrow \tilde{x}_0^T = (1, x_{01}, \dots, x_{0K})$, ¿y_0?

a) Queremos estimar $E(Y_0 \mid X_1 = x_{01}, \dots, X_K = x_{0K}) = \tilde{x}_0^T \beta =$

$= \beta_0 + \beta_1 x_{01} + \cdots + \beta_K x_{0K}$

$Z = \tilde{x}_0^T \cdot \hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_K x_{0K} = \tilde{x}_0^T (X^T X)^{-1} X^T Y$

$E(Z) = \tilde{x}_0^T (X^T X)^{-1} X^T X \beta = \tilde{x}_0^T \beta$

$V(Z) = \tilde{x}_0^T (X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1} \tilde{x}_0 = \sigma^2 \tilde{x}_0^T (X^T X)^{-1} \tilde{x}_0$

Bajo normalidad $\dfrac{\tilde{x}_0^T \cdot \hat{\beta} - \tilde{x}_0^T \beta}{S_R \sqrt{\tilde{x}_0^T (X^T X)^{-1} \tilde{x}_0}} \sim t_{n-K-1}$

Así que

$$\boxed{IC_{1-\alpha}(\tilde{x}_0^T \cdot \beta) = \tilde{x}_0^T \hat{\beta} \pm t_{\{n-K-1; \alpha/2\}} S_R \sqrt{\tilde{x}_0^T (X^T X)^{-1} \tilde{x}_0}}$$

b) Predecir $Y_0 = \tilde{x}_0^T \beta + \varepsilon_0 \sim N(0, \sigma^2)$ indep. de las $\varepsilon_1, \dots, \varepsilon_n$

con $Z = \tilde{x}_0^T \cdot \hat{\beta}$

$E(Y_0 - Z) = 0$

$V(Y_0 - Z) = \underbrace{V(Y_0)}_{= \sigma^2} + V(Z) = \sigma^2 (1 + \tilde{x}_0^T (X^T X)^{-1} \tilde{x}_0)$

Así que,

$\dfrac{Y_0 - \tilde{x}_0^T \hat{\beta}}{S_R \sqrt{1 + \tilde{x}_0^T (X^T X)^{-1} \tilde{x}_0}} \sim t_{n-K-1}$

$$\boxed{IC_{1-\alpha}(Y_0) = \tilde{x}_0^T \cdot \hat{\beta} \pm t_{\{n-K-1; \alpha/2\}} S_R \sqrt{1 + \tilde{x}_0^T (X^T X)^{-1} \tilde{x}_0}}$$

① INTERPRETACIÓN DE LOS CONTRASTES

Contraste global (F)	Contrastes individuales (t)	
Modelo explicativo	Todas las X_i explicativas	→ ✓
	Algunas de las X_i explicat.	→ te quedas con las explicativas
	Ninguna explicativa	→ ups! colinealidad
Modelo no explicativo	Todas	→ ups! colineal.
	Alguna	→ ups! "
	Ninguna	→ basura

② INTERVALOS DE CONFIANZA SIMULTÁNEOS

$$\mathbb{Y} = X \cdot \beta + \varepsilon \qquad \beta \text{ parámetros}$$

$$\hat{\beta} = (X^T X)^{-1} X^T \mathbb{Y} \quad \text{(estimador)}$$

$X^T X$ es simétrica, def. pos.

Tomamos raíz cuadrada:

$$X^T X = \underset{\text{ortogonal}}{P} \Lambda P^T = \underbrace{P \Lambda^{1/2} P^T}_{(X^T X)^{1/2}} \underbrace{P \Lambda^{1/2} P^T}$$

Consideramos el vector aleatorio $V = (X^T X)^{1/2} (\hat{\beta} - \beta) \Rightarrow$

$$\Rightarrow \begin{cases} \mathbb{E}(V) = \mathbb{0} \\ \operatorname{cov}(V) = (X^T X)^{1/2} \underbrace{\operatorname{cov}(\hat{\beta})}_{\sigma^2 (X^T X)^{-1}} (X^T X)^{1/2} = \sigma^2 I \end{cases} \quad ; \text{ y } V \text{ es normal.}$$

$$\Rightarrow \frac{V^T V}{\sigma^2} = \frac{1}{\sigma^2} (\hat{\beta} - \beta)^T (X^T X) (\hat{\beta} - \beta) \sim \chi^2_{K+1}$$

independientes

por otro lado

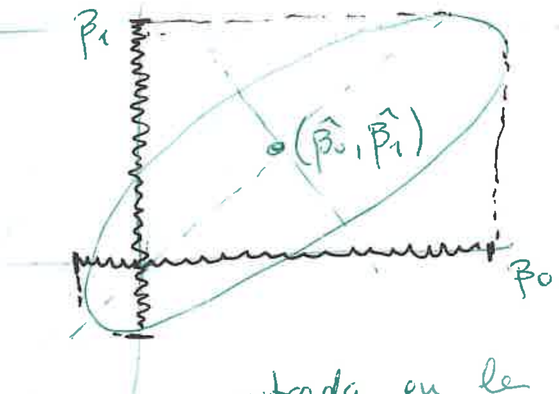
$$\frac{n-K-1}{\sigma^2} S_R^2 \sim \chi^2_{n-K-1}$$

Así que:

$$\frac{V^T V / (K+1)}{S_R^2} \sim F_{K+1,\, n-K-1}$$

Entonces, con probabilidad $1-\alpha$:

$$\underbrace{(\hat{\beta}-\beta)^T \frac{(X^TX)}{S_R^2} (\hat{\beta}-\beta)}_{\text{elipsoide}} \leq (k+1) F_{\{k+1,\, n-k-1;\, \alpha\}}$$



intervalos de confianza proyectados

Proyectando, se obtiene

$$\boxed{IC_{1-\alpha}(\beta_j) = \hat{\beta}_j \pm S_R \sqrt{q_{jj}} \sqrt{(k+1) F_{\{k+1,\, n-k-1;\, \alpha\}}}}$$

q_{jj}: entrada en la diagonal de $(X^TX)^{-1}$

③ VALIDACIÓN DEL MODELO

Lista de hipótesis que hemos hecho:

- → Linealidad de los parámetros
- → $n \geq k+2$
- → no colinealidad de las variables regresoras
- → ε_i media 0
- → ε_i varianza σ^2
- → ε_i incorreladas
- → ε_i normales ($\to$ indep.)

Tenemos los residuos e_i $i=1,\dots,n$

$$\hat{\beta} = (X^TX)^{-1}X^T Y$$

$$\hat{Y} = X\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}_{=H} Y$$

$$e = (I_n - H)Y = (I_n - H)\varepsilon$$

$$\mathbb{E}(e) = 0$$

$$\operatorname{cov}(e) = \sigma^2 (I_n - H)$$

$H = (h_{ij})$

$\sum_{j=1}^{n} h_{ij} = 1 \qquad \sum_{i=1}^{n} h_{ii} = k+1$

$\frac{1}{n} \leq h_{ii} \leq 1 \qquad -\frac{1}{2} \leq h_{ij} \leq \frac{1}{2}, \; i \neq j$

Si todos h_{ij} son "pequeños", los residuos son casi "independientes"

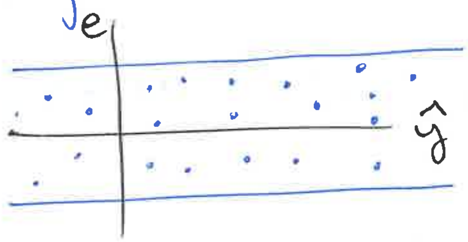
$$V(e_i) = \sigma^2(1-h_{ii})$$

Residuos estandarizados ("a la t")

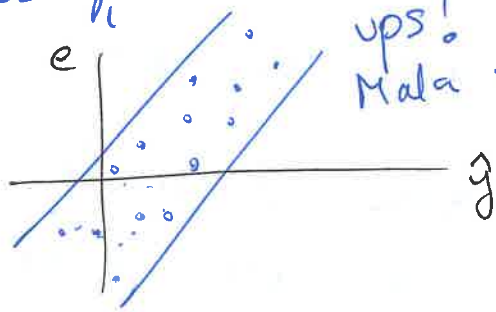
$$e_i^* = \frac{e_i}{S_R\sqrt{1-h_{ii}}} \sim \text{parecido a } N(0,1)$$

Residuos e son perpendiculares a $\hat{\mathbb{Y}}$.

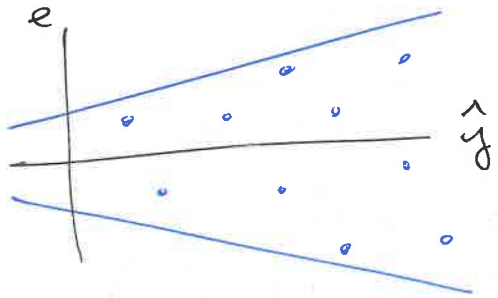
Dibujamos los e_i frente a los $\hat{y}_i$



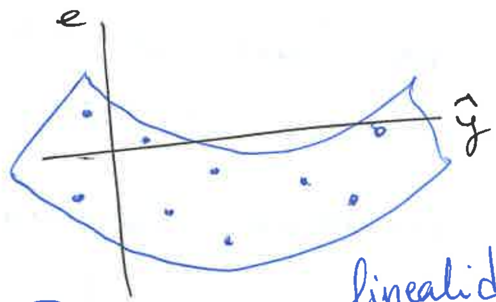
perfecto!



ups! Mala pinta

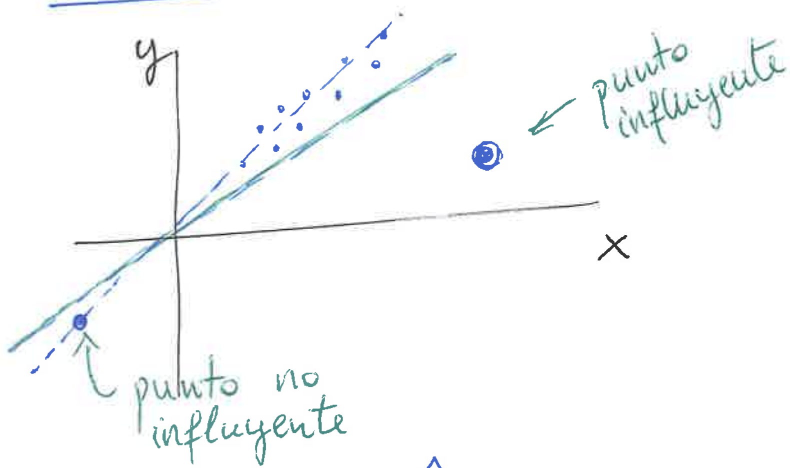


puede que σ^2 no sea constante



Pinta a no linealidad

(4) PUNTOS "INFLUYENTES"



Una observación es <u>influyente</u> si tiene mucho impacto en $\hat{\beta}$ y $\hat{y} = X\hat{\beta}$

$$\hat{\mathbb{Y}} = H\mathbb{Y}$$

$$\hat{y}_i = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j$$

Si h_{ii} es "grande" (próximo a 1), entonces $\hat{y}_i$ viene casi determinado por y_i.

A h_{ii} se le llama el "leverage" (apalancamiento) de la observación i.

Todos los h_{ii} suman $k+1 \Rightarrow$ "en media" son $\frac{k+1}{n}$

Hay quien dice que observación influyente es cuando $h_{ii} > 2\frac{k+1}{n}$.

Formalizamos el concepto de "influencia" midiendo el efecto que tiene quitar esa observación:

Con todos los datos $\longrightarrow \hat{\beta}, \hat{y}$

Quitando el dato "i" $\longrightarrow \hat{\beta}_{(i)}, \hat{y}_{(i)}$

Comparamos (distancia de Cook)

$$D_i = \frac{1}{(k+1)S_R^2}\left(\hat{\beta} - \hat{\beta}_{(i)}\right)^T (X^T X)\left(\hat{\beta} - \hat{\beta}_{(i)}\right) =$$

$$= \frac{1}{k+1}\left(\frac{e_i}{S_R\sqrt{1-h_{ii}}}\right)^2 \frac{h_{ii}}{1-h_{ii}}$$

TEMA 4 CLASIFICACIÓN

Contexto: tenemos unos individuos que pertenecen a ciertas poblaciones $\longrightarrow \Pi_0, \Pi_1, \Pi_2, \dots$

De cada individuo tenemos unas medidas de ciertas magnitudes: $X_1, X_2, \dots, X_K$

Datos:

	X_1	X_2	$\cdots$	X_K	Población
1	•	•	$\cdots$	•	•
2	•	•	$\cdots$	•	•
$\vdots$		$\vdots$			$\vdots$
n	•	•	$\cdots$	•	•

Desarrollamos un procedimiento que clasifique individuos en las poblaciones

- $\to$ el sistema se calibra con los datos disponibles
- $\to$ se usa para clasificar datos nuevos

Sistemas

- $\to$ reglas lineales
- $\to$ regresión logística
- $\to$ árboles de clasificación
- $\to$ redes neuronales
- $\vdots$

Ejemplos	Poblaciones	Variables
Marketing	{compradores, no compradores}	- ingresos - educación - tipo familia - compras anteriores
Medicina	{desarrolla tumor, no}	{edad, hábitos, datos analíticos}
Seguros	{fraude, no fraude}	{ingresos, edad, #tarjetas de crédito, #partes anteriores}
Finanzas	{Quiebra, No quiebra}	{precio acciones, nivel endeudamiento, resultados anteriores}

Formalizamos lo anterior:

$\mathbb{X} = (X_1, \dots, X_K) \longrightarrow$ vector de observaciones

Dos poblaciones $\longrightarrow \Pi_0$ y Π_1

$\begin{cases} \mathbb{X} \text{ se distribuye en } \Pi_0 \text{ con } f_0(x_1 \dots x_K) \\ \mathbb{X} \text{ se distribuye en } \Pi_1 \text{ con } f_1(x_1 \dots x_K) \end{cases}$

Esperamos que f_0 y f_1 sean "distintos"

(si iguales, da igual todo)

Normalmente, de f_0 y f_1 se sabe (o se postula)

cierta información:

$\rightarrow$ parcial $\begin{cases} \mu_1 \\ \mu_0 \end{cases}$ $\begin{cases} \Sigma_0 \\ \Sigma_1 \end{cases}$

$\rightarrow$ (semi) completa

f_0 y f_1 normales con $\begin{cases} \mu_0, \Sigma_0 \\ \mu_1, \Sigma_1 \end{cases}$ alguna desconocida

$\rightarrow$ completa

f_0, f_1 normales con $\left.\begin{cases} \mu_0, \Sigma_0 \\ \mu_1, \Sigma_1 \end{cases}\right\}$ todo conocido

Objetivo: desarrollar un sistema/procedimiento que clasifique "mal" con baja probabilidad

(quizás incluir un "coste" de mala clasificación)

1. La idea de Fisher (1938)

Transformar los datos K-dimensionales en 1-dim, proyectando sobre una dirección adecuada

$\mathbb{X} = (X_1, \dots, X_K)$ $\begin{cases} \rightarrow f_0 \text{ en } \Pi_0 \\ \rightarrow f_1 \text{ en } \Pi_1 \end{cases}$

Suponemos conocidos $\begin{cases} \mu_0, \Sigma_0 \\ \mu_1, \Sigma_1 \end{cases}$ $\Sigma_0 = \Sigma_1 = \Sigma$

Proyectamos datos sobre dirección $a \in \mathbb{R}^K$

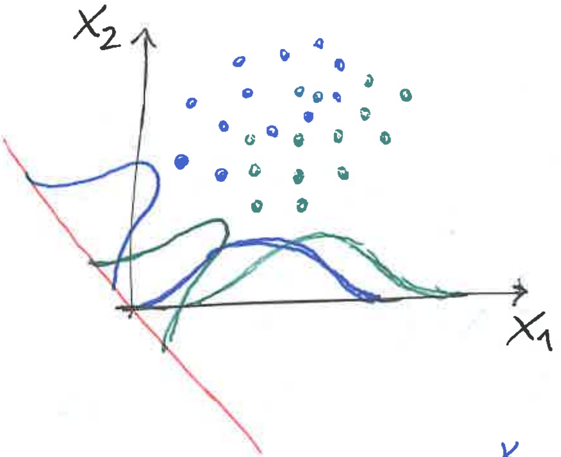
$$Z = a^T X \longrightarrow \text{media} \quad \mathbb{E}(Z) = \begin{cases} a^T\mu_0 & \text{en } \Pi_0 \\ a^T\mu_1 & \text{en } \Pi_1 \end{cases}$$

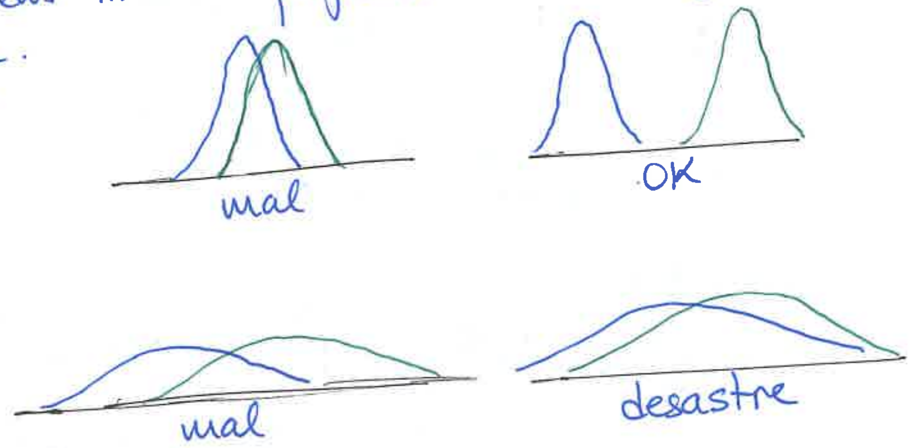
$$\longrightarrow \text{varianza} \quad \mathbb{V}(Z) = a^T \Sigma a$$

Queremos encontrar $a \in \mathbb{R}^K$:

→ distancia entre las medias proyectadas sea <u>grande</u>

→ varianza <u>pequeña</u>.





Buscamos $a \in \mathbb{R}^K$ tal que:

$$f(a) = \frac{|a^T(\mu_0 - \mu_1)|^2}{a^T \Sigma a}$$ sea "máxima" (maximizar $f(a)$)

<u>Obs</u>: si $\lambda \neq 0$, $f(\lambda a) = f(a) \longrightarrow$ buscamos direcciones, y no tanto vectores

<u>Recordatorio</u>: (Desigualdad de Cauchy-Schwarz)

$$\left(\sum x_i y_i\right)^2 \le \left(\sum x_i^2\right)\left(\sum y_i^2\right)$$

$x, y \in \mathbb{R}^K$ no nulos

$(x^T y)^2 \le (x^T x)(y^T y)$ con igualdad $\Leftrightarrow x = cy$ para cierto $c \in \mathbb{R}$.

Generalización: $x, y \in \mathbb{R}^K$, A matriz $K \times K$ sim. def. pos.,

entonces $(x^T y)^2 \le (x^T A x)(y^T A^{-1} y)$ con igualdad si y solo si $\begin{cases} x = c A^{-1} y \\ y = c A x \end{cases}$ para cierto $c \in \mathbb{R}$

demo

Escribimos $A = P\Lambda P^T$, $A^{-1} = P\Lambda^{-1}P^T$

Raíces cuadradas:

$$A = P\Lambda^{1/2}\Lambda^{1/2}P^T = (P\Lambda^{1/2})(P\Lambda^{1/2})^T$$

pero también

$$A = \underbrace{P\Lambda^{1/2}P^T}_{=C} P\Lambda^{1/2}P^T = C.C \qquad (C.C^{-1} = Id)$$

$$A^{-1} = P\Lambda^{1/2}P^T P\Lambda^{-1/2}P^T = C^{-1}C^{-1}$$

Con esto:

$$(x^Ty)^2 = (x^T I_k y)^2 = (x^T C C^{-1} y)^2 = \left[(Cx)^T(C^{-1}y)\right]^2 \leq$$

$$\overset{C\text{-}S}{\leq} ((Cx)^T Cx).((C^{-1}y)^T(C^{-1}y)) = (x^T \underbrace{CC}_{=A} x).(y^T \underbrace{C^{-1}C^{-1}}_{=A^{-1}} y) \qquad \blacksquare$$

TEOREMA: El máximo de $f(a) = \dfrac{|a^T(\mu_0-\mu_1)|^2}{a^T\Sigma a}$ se alcanza en múltiplos de $\boxed{a_m = \Sigma^{-1}(\mu_0-\mu_1)}$

demostración

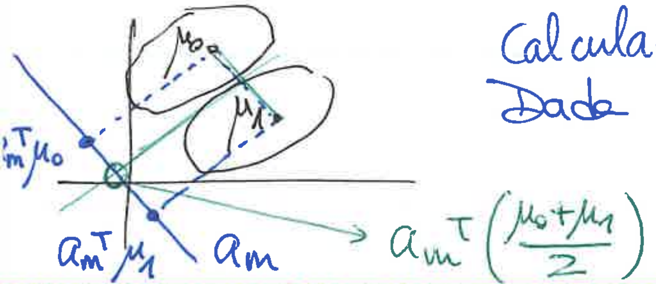
C-S general, con Σ

$$\frac{|a^T(\mu_0-\mu_1)|^2}{a^T\Sigma a} \leq \frac{\cancel{(a^T\Sigma a)}((\mu_0-\mu_1)\Sigma^{-1}(\mu_0-\mu_1))}{\cancel{a^T\Sigma a}}$$

Si tomamos $a_m = c\,\Sigma^{-1}(\mu_0-\mu_1)$:

$$f(a_m) = (\mu_0-\mu_1)\Sigma^{-1}(\mu_0-\mu_1)$$ y ya estaría. $\blacksquare$

Esto nos da una REGLA DE CLASIFICACIÓN:



Calculamos $a_m = \Sigma^{-1}(\mu_0-\mu_1)$

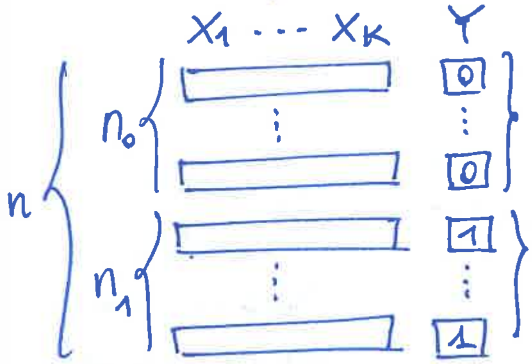
Dada observación $x = (x_1 - x_k)^T$,

si $a_m^T x \geq a^T\left(\frac{\mu_0+\mu_1}{2}\right) \Rightarrow x$ en Π_1

si $a_m^T x <$ " $\Rightarrow x$ en Π_0

Esto último requiere conocer $\longrightarrow \mu_0, \mu_1, \Sigma_0 = \Sigma_1 = \Sigma$

En la práctica:



$\longrightarrow \bar{x}_0, S_0$

$\longrightarrow \bar{x}_1, S_1$

Estimamos:

$\hat{\mu}_0 = \bar{x}_0$

$\hat{\mu}_1 = \bar{x}_1$

$$\hat{\Sigma} = \underbrace{\frac{(n_0-1)S_0 + (n_1-1)S_1}{n_0+n_1-2}}_{S_{pooled}}$$

La regla de clasificación sería:

1. Calculamos $\hat{a}_m = S_{pooled}^{-1}(\bar{x}_0 - \bar{x}_1)$

2. Dado $x = (x_1 \text{ — } x_K)^T$,

si $\hat{a}_m^T x > \hat{a}_m^T \left(\frac{\bar{x}_0 + \bar{x}_1}{2}\right) \longrightarrow x$ en Π_1

si $\hat{a}_m^T x < \hat{a}_m^T \left(\frac{\bar{x}_0 + \bar{x}_1}{2}\right) \longrightarrow x$ en Π_0

2 ¿Qué tal funciona?

Tasa de error aparente: TEA = porcentaje de puntos mal clasificados

(quizás convenga distinguir mal clasificados entre los de Π_0 y los de Π_1)

Trampa: usar los mismos datos para
- $\nearrow$ crear la regla
- $\searrow$ validar su funcionamiento

Alternativas:
- $\longrightarrow$ dividir los datos
  - $\nearrow$ training set
  - $\searrow$ test set
- $\longrightarrow$ validación cruzada

Regresión Logística

$X_1, \dots, X_K \leftarrow$ variables regresoras

$Y \leftarrow$ respuesta (0,1 en nuestro caso)

Estructura de datos

	X_1	—	X_K	Y
dato 1	x_{11}	$\cdots$	x_{1K}	y_1
$\vdots$	$\vdots$		$\vdots$	$\vdots$
dato n	x_{n1}	$\cdots$	x_{nK}	y_n

Modelo $\mathbb{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ y_i serán Bernoullis indeps. cada uno con su param.

$y_i \sim Ber(p_i)$

$p_i = \mathbb{P}(Y_i = 1 \mid x_i)$

$Y \sim Ber(p(x)) \qquad x = (x_1, \dots, x_K)$

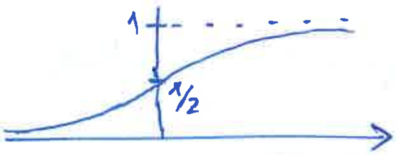
$$¿\ \underbrace{p(x)}_{\text{valores en } [0,1]} = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K}_{\text{valores en } \mathbb{R}}\ ?$$

no funciona

Buscamos función $\mathbb{R} \rightarrow [0,1]$

$$h(x) = \frac{1}{1+e^{-x}}$$

función logística (logit)

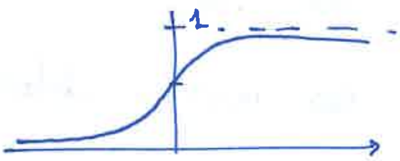


$h(0) = 1/2$

$h(-x) = 1 - h(x)$

$h'(x) = h(x)(1-h(x))$

Alternativa, $\Phi(x)$ (probit)



Modelo LOGIT

$$p(x) = \mathbb{P}(Y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_K x_K)}} = h(\beta_0+\beta_1 x_1+\cdots+\beta_K x_K)$$

1. Interpretación de los parámetros $\beta_0, \beta_1, \ldots, \beta_K$

Razón de probabilidades (odds): $O(x) = \dfrac{p(x)}{1-p(x)} = \dfrac{\frac{1}{1+e^{-(\cdots)}}}{1-\frac{1}{1+e^{-(\cdots)}}} =$

$$= e^{\beta_0+\beta_1 x_1+\cdots+\beta_K x_K} \implies \ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0+\beta_1 x_1+\cdots+\beta_K x_K$$

Variaciones: $x=(x_1,\ldots,x_K) \longrightarrow x+\Delta_j = (x_1,\ldots,x_j+1,\ldots,x_K)$

¿Cuánto cambian los parámetros?

$$\frac{O(x+\Delta_j)}{O(x)} = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_K x_K}\cdot e^{\beta_j}}{e^{\beta_0+\beta_1 x_1+\cdots+\beta_K x_K}} = e^{\beta_j}$$

2. Estimación de parámetros

Datos:

$$\begin{array}{ccc} x_1 & \boxed{x_{11} \cdots x_{1K}} & \boxed{y_1} \\ \vdots & \vdots & \vdots \\ x_n & \boxed{x_{n1} \cdots x_{nK}} & \boxed{y_n} \end{array}$$

Queremos estimar $\beta = \begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_K\end{pmatrix}$

Usamos máx. verosimilitud

$$\text{VERO}(\beta_0,\ldots,\beta_K) = \prod_{i=1}^{n} p(x_i)^{y_i}(1-p(x_i))^{1-y_i}$$

Variamos $\beta_0,\ldots,\beta_K$ para maximizar $\text{VERO}(\cdot)$ explícitamente es una expresión muy complicada $\longrightarrow$ solver (o un mejor maximizador)

(*) Un poco de notación:

$$\beta = (\beta_0, \dots, \beta_K)^T$$

$$x_i = (x_{i1}, \dots, x_{iK})^T \qquad i = 1, \dots, n$$

$$\tilde{x}_i = (1, x_{i1}, \dots, x_{iK})^T \qquad i = 1, \dots, n \qquad \text{(añadimos columna de 1's)}$$

$$\beta^T \cdot \tilde{x}_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_K x_{iK}$$

$$p(x_i) = \frac{1}{1 + e^{-\beta^T \tilde{x}_i}} = h(\beta^T \tilde{x}_i)$$

$$\frac{p(x_i)}{1 - p(x_i)} = e^{\beta^T \tilde{x}_i}$$

$$\begin{cases} h'(x) = h(x)\,(1 - h(x)) \\ \dfrac{\partial}{\partial \beta_j} (\beta^T \tilde{x}_i) = x_{ij} \end{cases}$$

$$\frac{\partial}{\partial \beta_j} p(x_i) = h(\beta^T \tilde{x}_i)\left(1 - h(\beta^T \tilde{x}_i)\right) x_{ij} \qquad \begin{matrix} i = 1, \dots, n \\ j = 1, \dots, K \end{matrix}$$

$$\text{VERO}(\beta) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

$$\log\text{VERO}(\beta) = \sum_{i=1}^{n} \left(y_i \ln(p(x_i)) + (1-y_i) \ln(1 - p(x_i)) \right) =$$

$$= \sum_{i=1}^{n} y_i \ln\left(\frac{p(x_i)}{1 - p(x_i)} \right) + \sum_{i=1}^{n} \ln(1 - p(x_i)) =$$

$$= \sum_{i=1}^{n} y_i \cdot \beta^T \tilde{x}_i + \sum_{i=1}^{n} \ln(1 - p(x_i))$$

$$\frac{\partial}{\partial \beta_j} \log\text{VERO}(\beta) = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} \frac{p(x_i)(1 - p(x_i))}{1 - p(x_i)} x_{ij} =$$

$$= \sum_{i=1}^{n} x_{ij} (y_i - p(x_i))$$

$$\nabla \log \text{VERO}(\beta) = \begin{pmatrix} \frac{\partial}{\partial \beta_0} LV \\ \vdots \\ \frac{\partial}{\partial \beta_k} LV \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & \cdots & & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ & \vdots & & \\ x_{1k} & x_{2k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} y_1 - p(x_1) \\ \vdots \\ y_n - p(x_n) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}}$$

$k+1$ ecuaciones no lineales

$k+1$ incógnitas $\beta_0, \beta_1, \dots, \beta_k$

$\downarrow$ numérico

$\hat\beta_0, \hat\beta_1, \dots, \hat\beta_k$ solución

3. Regla de clasificación

observación $\longrightarrow a = (a_1, \dots, a_k)$

$\tilde a = (1, a_1, \dots, a_k)$

Calculamos $\hat p(a) = h(\hat\beta^T \cdot \tilde a) = \dfrac{1}{1 + e^{-(\hat\beta_0 + \hat\beta_1 a_1 + \cdots)}}$

Si $\hat p(a) > \frac{1}{2}$ $\Downarrow$ clase 1

Si $\hat p(a) < \frac{1}{2}$ $\Downarrow$ clase 2

$$\hat p(a) > \frac{1}{2} \iff \frac{1}{1 + e^{-\hat\beta^T \cdot \tilde a}} > \frac{1}{2} \iff \boxed{\hat\beta^T \cdot \tilde a > 0}$$

obs: ¿Distribución de los estimadores?

Si n grande, como son estimadores de máx-vero,

$$\hat\beta \simeq N(\beta, \text{cov}(\hat\beta)) \longrightarrow (X^T \cdot \hat W \cdot X)^{-1} \longrightarrow \begin{pmatrix} \hat p(x_1)(1-\hat p(x_1)) & & 0 \\ & \ddots & \\ 0 & & \hat p(x_n)(1-\hat p(x_n)) \end{pmatrix}$$

Esto daría un intervalo de confianza (aprox.)

$$IC_{1-\alpha}(\beta_j) = \hat\beta_j \pm z_{\alpha/2} \cdot SE(\hat\beta_j)$$

$\longrightarrow$ standard error

raíz cuadrada del elemento $j+1$ de la diag. de $(X^T \hat W X)^{-1}$

O bien, contrastar $H_0 : \beta_j = 0$

Usamos que $\underbrace{\dfrac{\hat{\beta}_j}{SE(\hat{\beta}_j)}}_{\text{Wald}} \approx N(0,1)$.

REGLAS "ÓPTIMAS" DE CLASIFICACIÓN

Planteamiento: dos poblaciones $\begin{cases} \Pi_0 \\ \Pi_1 \end{cases}$

Queremos clasificar objetos (en Π_0 o en Π_1) en función de unos cuantos atributos/observaciones/medidas $\mathbb{X} = (X_1 - X_K)^T$

Este vector $\mathbb{X}$ se distribuye como $\begin{cases} \rightarrow f_0(\vec{x}) \text{ en } \Pi_0 \\ \rightarrow f_1(\vec{x}) \text{ en } \Pi_1 \end{cases}$ $\vec{x} = (x_1, \dots, x_K)^T$

(razonablemente "diferentes")

Además, tenemos probabilidades "a priori" $p_0, p_1 : p_0 + p_1 = 1$

Una regla de clasificación g: $g : \Omega \subset \mathbb{R}^K \longrightarrow \{0, 1\}$

$$\vec{x} \longmapsto g(\vec{x}) = \begin{cases} 0 \\ 1 \end{cases}$$

Una tal regla divide Ω en: $\Omega = R_0^{(g)} \cup R_1^{(g)}$

Consideramos: $\mathbb{P}(\mathbb{X} \in R_0^{(g)} \mid \Pi_1) = \int_{R_0^{(g)}} f_1(\vec{x})\, d\vec{x}$

$$\mathbb{P}(\mathbb{X} \in R_1^{(g)} \mid \Pi_0) = \int_{R_1^{(g)}} f_0(\vec{x})\, d\vec{x}$$

$$\mathbb{P}\left(\begin{matrix}\text{mala clasificación} \\ \text{con regla } g\end{matrix}\right) = p_1 \int_{R_0^{(g)}} f_1(\vec{x})\, d\vec{x} + p_0 \int_{R_1^{(g)}} f_0(\vec{x})\, d\vec{x}$$

Objetivo: hallar g que minimice esta probabilidad

<u>Observación</u>: $1 = \int_\Omega f_1(\vec{x})d\vec{x} = \int_{R_0(g)} f_1(\vec{x})d\vec{x} + \int_{R_1(g)} f_0(\vec{x})\,d\vec{x}$

$$\Rightarrow \mathbb{P}\begin{pmatrix}\text{mala clasif.}\\ \text{regla } g\end{pmatrix} = P_1\left(1 - \int_{R_1(g)} f_1(\vec{x})d\vec{x}\right) + P_0\int_{R_1(g)} f_0(\vec{x})\,d\vec{x} =$$

$$= P_1 + \int_{R_1(g)} \left(P_0 f_0(\vec{x}) - P_1 f_1(\vec{x})\right)d\vec{x}$$

Esta probabilidad es <u>mínima</u> cuando

$$R_1 = \{x \in \Omega : P_0 f_0(\vec{x}) - P_1 f_1(\vec{x}) \leq 0\} =$$

$$= \left\{x \in \Omega : \frac{f_1(\vec{x})}{f_0(\vec{x})} \geq \frac{P_0}{P_1}\right\}$$

R_0 es la contraria (con $\leq$ aquí)

<u>Nota</u>: Argumento bayesiano

Observamos un $\vec{x} = (x_1 \cdots x_n)$

clasificamos $\vec{x}$ en función de cual sea la mayor probabilidad a posteriori

$$\mathbb{P}(\Pi_1|\vec{x}) = \underbrace{\mathbb{P}(\vec{x}|\Pi_1)}_{f_1(\vec{x})} \cdot \frac{P_1}{p(\vec{x})} = \frac{P_1 f_1(\vec{x})}{p(\vec{x})}$$

$$\mathbb{P}(\Pi_0|\vec{x}) = \mathbb{P}(\vec{x}|\Pi_0) \cdot \frac{P_0}{p(\vec{x})} = \frac{P_0 f_0(\vec{x})}{p(\vec{x})}$$

comparamos estos dos valores

Nota 2 : Incorporamos costes de mala clasificación

	π_0	π_1
π_0	0	$C_{1\|0}$
π_1	$C_{0\|1}$	0

Interesa minimizar coste medio de mala clasificación =

$$= P_1 C_{0|1} \int_{R_0^{(g)}} f_1(\vec{x})\,d\vec{x} + P_0 C_{1|0} \int_{R_1^{(g)}} f_0(\vec{x})\,d\vec{x}$$

Sea g óptima:

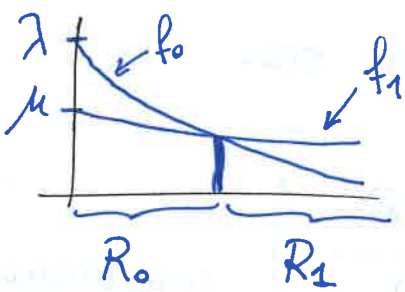
$$R_1^{(g)} = \{ \vec{x} \in \Omega \mid P_0 C_{1|0} f_0(\vec{x}) - P_1 C_{0|1} f_1(\vec{x}) \leq 0 \} =$$

$$= \left\{ \vec{x} \in \Omega \;\middle|\; \frac{f_1(\vec{x})}{f_0(\vec{x})} \geq \frac{C_{1|0}}{C_{0|1}} \cdot \frac{P_0}{P_1} \right\}$$

Ejemplo: $K=1$, $\Omega = \mathbb{R}^+$, X variable

$$\begin{cases} f_0(x) = \lambda e^{-\lambda x}, & x > 0 \\ f_1(x) = \mu e^{\mu x}, & x > 0 \end{cases} \qquad \lambda > \mu \qquad \mathbb{E}(X) = \frac{1}{\lambda}$$

$P_0 = P_1 = 1/2$



$$f_0(x) = f_1(x)$$

$$\lambda e^{-\lambda x} = \mu e^{-\mu x} \Longrightarrow x = \frac{1}{\lambda - \mu} \ln\left(\frac{\lambda}{\mu}\right)$$

Si $P_0 = \frac{1}{10}$, $P_1 = \frac{9}{10}$

$$\Longrightarrow \frac{f_1(x)}{f_0(x)} \geq \frac{1/10}{9/10} = \frac{1}{9} \qquad \ldots$$

El caso de dos poblaciones normales

$$x=(x_1,\dots,x_K) \qquad f_0(x) = N_K(\mu_0,\Sigma_0)\ ,\ P_0 \quad \left(\text{quizás coste } C_{0|1}\right)$$

$$f_1(x) = N_K(\mu_1,\Sigma_1)\ ,\ P_1 \quad \left(\text{"} \ C_{1|0} \ \text{"}\right)$$

$$f_i(x) = \frac{1}{(2\pi)^{K/2}\,|\Sigma_i|^{1/2}}\cdot \exp\left(\frac{-1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right) \qquad i=0,1$$

$$R_1(x) = \left\{x\in\mathbb{R}^K : \frac{f_1(x)}{f_0(x)} \geq \frac{P_0}{P_1}\right\}$$

$$\frac{P_0}{P_1} \leq \frac{|\Sigma_0|^{1/2}}{|\Sigma_1|^{1/2}}\exp\left\{\frac{-1}{2}\left((x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right)\right\}$$

$$\ln\left(\frac{P_0}{P_1}\right) \leq \frac{-1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) - \frac{1}{2}\Big[x^T(\Sigma_1^{-1}-\Sigma_0^{-1})x - 2(\mu_1^T\Sigma_1^{T}-\mu_0^T\Sigma_0^{-1})x$$
$$+\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0\Big]$$

$$\boxed{\ln\left(\frac{P_0}{P_1}\right) \leq \frac{-1}{2}\overbrace{x^T(\Sigma_1^{-1}-\Sigma_0^{-1})x}^{\text{forma cuadrática}} + \overbrace{(\mu_1^T\Sigma_1^{-1}-\mu_0^T\Sigma_0^{-1})x}^{\text{lineal}} \underbrace{-\frac{1}{2}\left[\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right)+\mu_1^T\Sigma_1^{-1}\mu_1-\mu_0^T\Sigma_0^{-1}\mu_0\right]}_{\text{constante}}}$$

Caso particular $\Sigma_1 = \Sigma_0 = \Sigma$

$$\ln\left(\frac{P_0}{P_1}\right) \leq (\mu_1-\mu_0)^T\Sigma^{-1}x + \frac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1+\mu_0)$$

Si $P_0 = P_1 \longrightarrow$ <u>Fisher</u>