Introducción a IA.

En el caso general: $y = f(\bar{x}, \bar{\beta})$ $\quad \bar{x} = (\underbrace{x_1, x_2, \dots, x_n}_{\text{features}})$ $\quad \hat{\beta} = (x^t x)^{-1} x^t y$

Para el análisis de hoy consideramos el caso lineal simple $\Rightarrow \bar{x} \in \mathbb{R}$,

con esto, tenemos $Y = \beta_0 + \beta_1 X$.

Ejemplos de modelos lineales: $\to$ $y = \beta_0 + \beta_1 X$

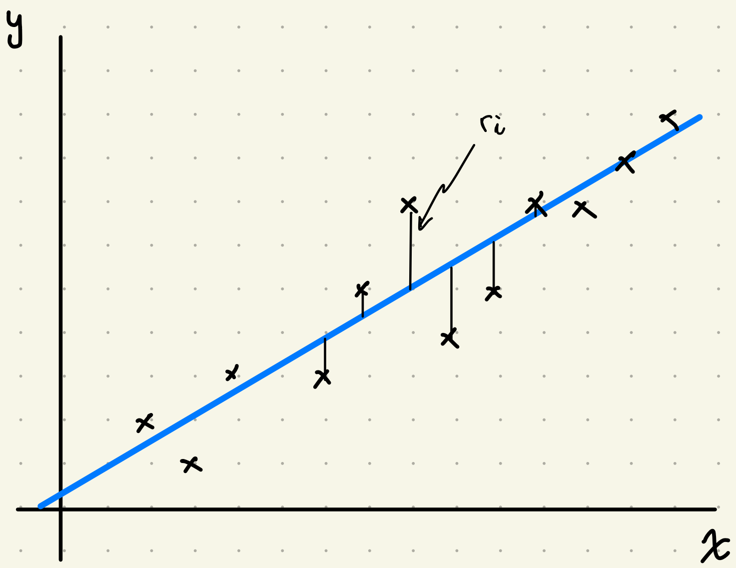
↳ $y = \beta_0 x^{\beta_1} \to \ln(y) = \ln(\beta_0) + \beta_1 \ln(x)$

↳ exp(·)

↳ modelo generalizado

$y = \beta_0 + \beta_1 X + \beta_2 X^2$ $\to$ modelo lineal con los β
no lineal con x.
(linealizable *)

- Analizando la regresión simple.



recordemos que:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$r_i = y_i - \hat{y}_i \leftarrow \text{i-esimo residuo}$$

Suma de cuadrados de residuos $\boxed{RSS = \sum_i r_i^2}$

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_i \underbrace{(y_i - \beta_0 - \beta_1 x_i)^2}_{①}$$

$$\begin{cases} \partial_{\beta_0} \sum_i ①^2 = -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) & \longrightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \partial_{\beta_1} \sum_i ①^2 = -2 \sum (y_i - \beta_0 - \beta_1 x_i)\, x_i & \longrightarrow \hat{\beta}_1 = \dfrac{\frac{1}{N} \sum (x_i - \bar{x})(y_i - \bar{y})^2}{\underbrace{\frac{1}{N} \sum_i (x_i - \bar{x})^2}_{\widehat{var}(x)}} \end{cases}$$

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{\widehat{cov}(X,Y)}{\widehat{var}(X,Y)} = \underbrace{\hat{\rho}_{XY}}_{\text{coef. de correlación de Pearson}} \dfrac{S_Y}{S_X} \end{cases}$$

S_Y, S_X → Estimadores de la varianza

$$cov(X,Y) = \mathbb{E}(X,Y)^2 \quad , \quad cov(X,X) = Var(X)$$

Ahora analizamos los errores:

$$y_i = \hat{y}_i + r_i$$

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + r_i$$

$$\sum_{i=0}^{n} (y_i - \bar{y})^2 = \sum_{i=0}^{n} (\hat{y}_i - \bar{y}) + \sum_{i=0}^{n} r_i^2$$

$$TSS = ESS + RSS$$

TSS → taza de variabilidad total

ESS ↳ suma de explicados al cuadrado

RSS ↳ suma de cuad. de residuos

Con este approach puedo construir test estadísticos para validar el modelo. (lo vemos más adelante).

¿Cómo se que mi modelo es bueno? → Error residual (RSE)

↳ Estadístico R^2

↳ " F

+ $RSE = \sqrt{\dfrac{RSS}{n-2}} = \sqrt{\dfrac{\Sigma (y_i - \hat{y})^2}{n-2}}$ → me interesa que sea bajo.

+ R^2 → mide la bondad del ajuste.

↳ equivalente al coef. de correlación al cuadrado entre x e y.

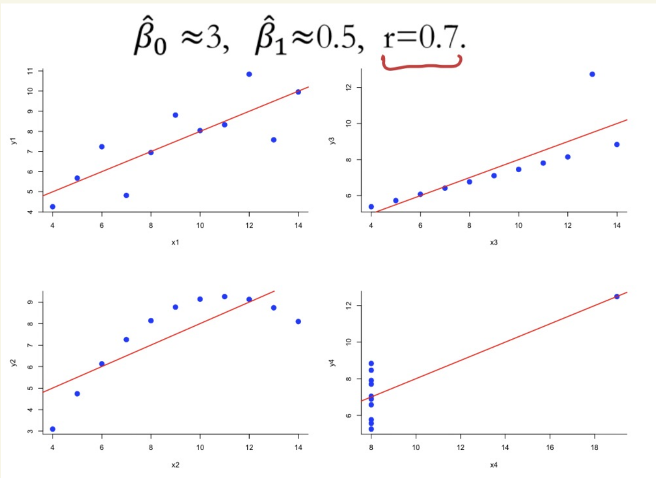
$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = \hat{\rho}_{xy}^{\,2} \quad \text{(esto en regresión simple)}$$

+ R^2 no depende de las unidades de y, solo de las proporciones.

+ $R^2 \in [0,1]$

+ $R^2 \approx 1$ es un "buen" ajuste.

(NO SIEMPRE) ⟶



- Estadístico F:

Armamos la tabla ANOVA del ajuste:

Fuentes de Variacion	SS	gl	cuad medios	
Explicada (ESS)	$\sum_i (\hat{y}_i-\bar{y})^2$	1	$\sum_i (\hat{y}_i-\bar{y})^2$	F
Residual (RSS)	$\sum_i r_i^2$	$n-2$	$s_R^2 = \frac{\sum r_i^2}{n-2}$	
Variab. total (TSS)	$\sum_i (y_i-\bar{y})^2$	$n-1$		

construimos $F = \frac{ESS}{s_R^2}$ $\rightarrow$ si F es muy grande (la var. explicada es muy grande respecto a la no explicada) vamos a rechazar Ho con Ho siendo:

TH: $\underbrace{H_0: \beta_1 = 0}_{\text{Hipótesis nula}} \;\wedge\; \underbrace{H_1: \beta_1 \neq 0}_{\text{hip. Alternativa}}$ $\rightsquigarrow P(F > f_{gl,\alpha})$

$f_{1,n-2;1-\alpha}$

Como interpretamos las hipótesis:

+ Hipótesis nula (H_0): Con esto vamos a suponer que el modelo de regresión lineal afín no explica mejor la varianza que un modelo constante.

+ Hipótesis Alternativa (H_1): El modelo obtenido explica mejor (es estadísticamente significativo) la varianza en la respuesta que un modelo constante.

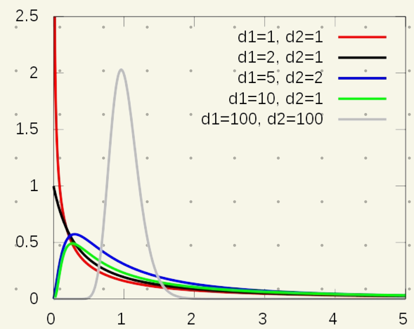
- Como calculamos el F-test?

1. Armamos la tabla ANOVA.

2. Obtenemos el estadístico $F_s = \frac{ESS}{s_R^2}$

3. Calculamos el f_{crit} (p-valor) a partir de los parametros obtenidos:

$$f \sim F_{gl_1, gl_2}$$



4. para un dado α (tipicamente $\alpha = 0.05$):

buscamos f_{crit} tal que:

$$P(F > f_{crit}) = \alpha$$

5. Si $F_s > f_{crit} \Rightarrow$ Reclazo Ho.