# Introducción a la Inteligencia Artificial

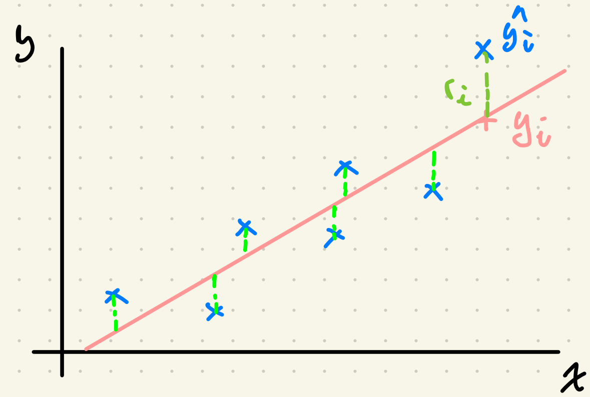
Clase ~~4~~ 3

.UBA fiuba

FACULTAD DE INGENIERÍA

# Índice

Analizamos el caso simple: (1 regresor, 1 var. dep)




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

con esto podemos definir los residuos

$$r_i = y_i - \hat{y}_i \quad \forall i \in [1, \dots, N]$$

bondad del ajuste $\propto \sum_i r_i^2$

$r_i \sim N(0, \sigma_r)$

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

$\bar{r} = \bar{y} - \hat{y} = \bar{y} - \bar{\bar{X}}\hat{\beta} = \bar{y} - \bar{\bar{H}}\bar{y}$

$\bar{r} = \underbrace{(\mathbb{I} - H)}_{}\bar{y}$

simétrica,
idempotente,
$tr = n - p$

$$\begin{cases} \partial_{\beta_0} f = -2 \sum (y_i - \beta_0 - \beta_1 x_i) = 0 \rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \partial_{\beta_1} f = -2 \sum (y_i - \beta_0 - \beta_1 x_i) \cdot x_i = 0 \rightarrow (1) \end{cases}$$

(+)

(+) Acá $\bar{Y}$ y $\bar{X}$ representan los promedios.

(1) $\rightarrow$ $$\sum (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

$$\sum (y_i x_i - \beta_0 x_i - \beta_1 x_i^2) = 0$$

$$\sum (y_i x_i - \bar{y} x_i - \beta_1 x_i - \beta_1 x_i^2) = 0$$

Desarrollando (ver apunte) llegamos a:

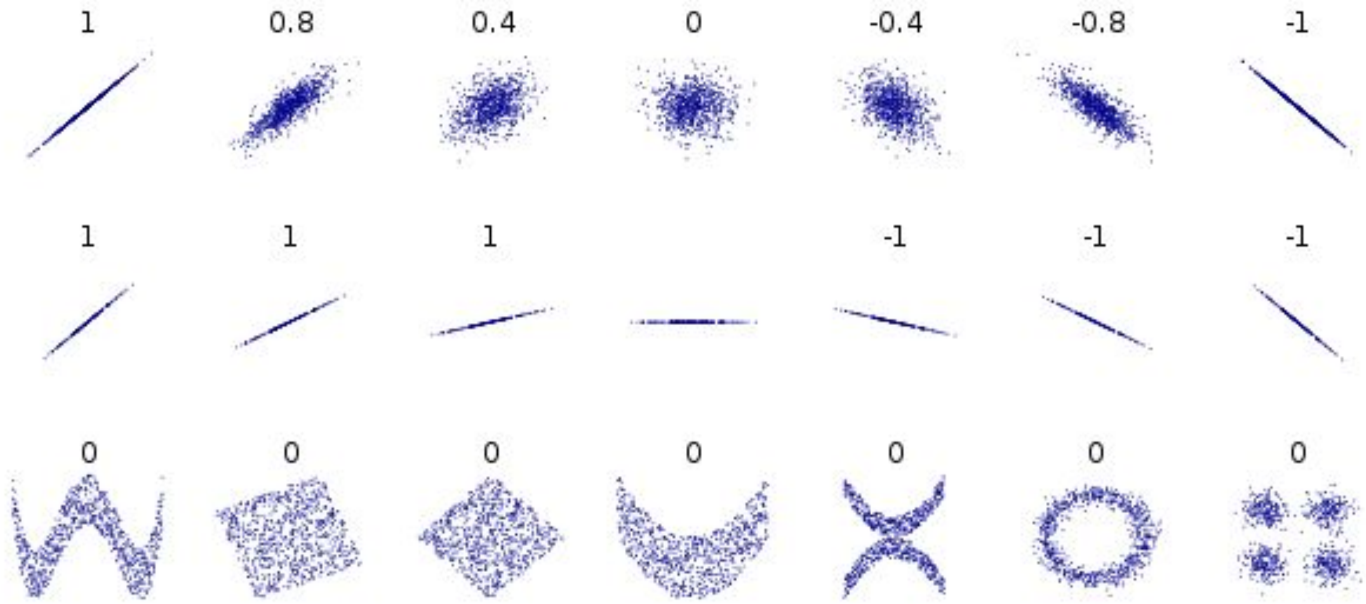
$$\hat{\beta}_1 = \frac{N}{N} \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N}}{\frac{\sum (x_i - \bar{x})^2}{N}}$$

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{\text{covar}(x,y)}{\text{Var}(x)} = \hat{\rho}_{xy} \end{cases}$$

coef. de correlación lineal de Pearson

$$\rho_{xy} \in [-1, 1]$$

## Coeficiente de correlación de Pearson (Lineal)



Analizamos los errores de la regresión:

$y_i = \hat{y}_i + r_i$ $\qquad$ $\bar{y}$ : media muestral (promedio)

el término cruzado desaparece por indep.

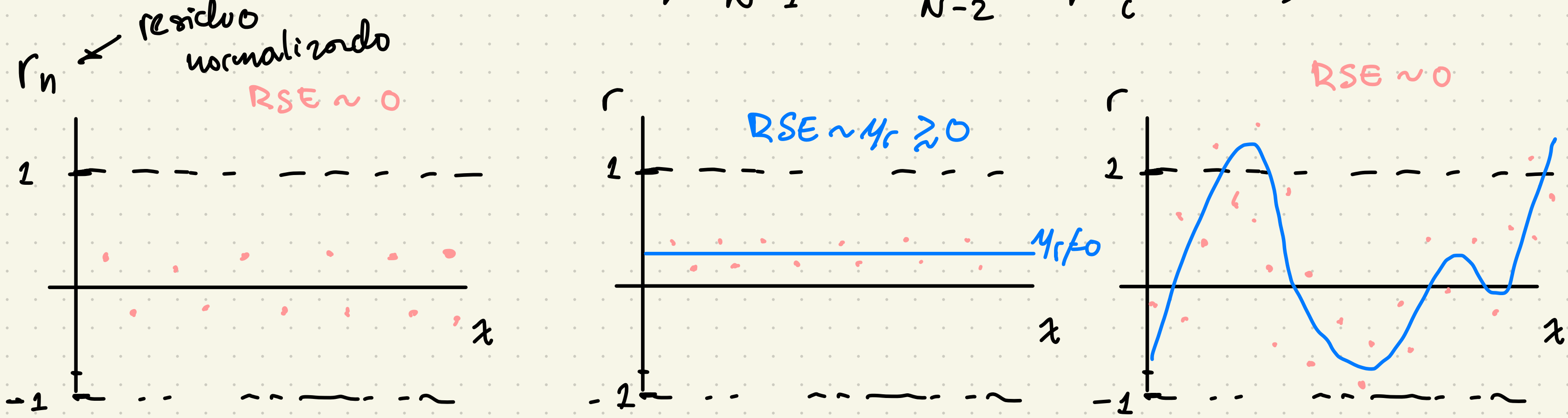
$$y_i - \bar{y} = \hat{y}_i - \bar{y} + r_i \rightsquigarrow \underbrace{\sum_{i=0}^{N} (y_i - \bar{y})^2}_{1} = \underbrace{\sum_{i=0}^{N} (\hat{y}_i - \bar{y})^2}_{2} + \underbrace{\sum_{i=0}^{N} r_i^2}_{3}$$

1. TSS : tasa de variabilidad total
2. ESS : tasa de variabilidad explicada
3. RSS : suma de los residuos al cuadrado

Estos valores me permiten construir métricas de bondad de ajuste para diagnosticar mi modelo.

métricas posibles →
- error residual (RSE)
- $R^2$ (coef. de pearson)
- Est. F (Análisis "Avanzado")

• Error residual: $RSE = \sqrt{\frac{RSS}{N-1}} = \frac{1}{N-2}\sqrt{\sum_i (y_i - \hat{y}_i)^2}$




• Si RSE es bajo ~~entonces~~ puede que nuestro ajuste ~~es~~ sea bueno

• RSE no es sensible a la distribución y/o tendencia funcional de los residuos.

## Coeficiente de determinación - “R cuadrado”

RSS

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

TSS

SS = Sum of squares
res = residuos
tot = total

+ coef de pearson $R^2$:

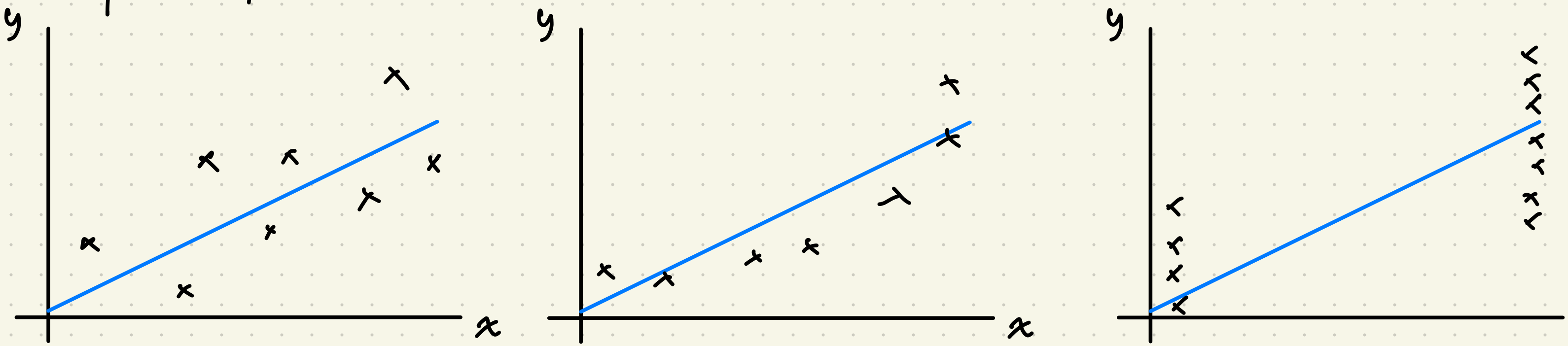
$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \overset{①}{=} (\rho_{xy})^2$$

① es valido únicamente bajo el régimen de reg. simple

- $R^2$ no depende de las escalas, solo depende de las proporciones.
- $R^2 \in [0,1]$ ⇝ $R^2 \sim 1$ ⇝ el modelo es "bueno"

  $R^2 \lesssim 0$ ⇝ el modelo es peor que haber aproximado con la media.

$\beta_0 \sim 3$, $\beta_1 \sim 0{,}5$, $R^2 = 0{,}7$



Estadístico F:

Vamos a armar una tablita (Anova) de nuestro ajuste

| Fuente de variación | suma de los cuadrados | grados de libertad | cuadrados medios |
|---|---|---|---|
| Explicada | $\sum(\hat{y}_i-\bar{y})^2$ | 1 | $S_E = \sum(\hat{y}_i-\bar{y})^2$ |
| residual | $\sum r_i^2$ | $N-2$ | $S_R = \frac{1}{N-2}\sum r_i^2$ |
| total | $\sum(y-\bar{y})^2$ | $N-1$ | |

una vez armada la tabla $\longrightarrow$ calculo mi estadístico $F = \frac{ESS}{S_R^2}$

si F es grande $\longrightarrow$ la variabilidad explicada es muy grande respecto a la var residual.

con esto planteamos nuestro test de hipótesis:

$$TH \begin{cases} H_0 : & \beta_1 = 0 \\ H_1 : & \beta_1 \neq 0 \end{cases}$$

Aquí buscamos rechazar $H_0$ con una significancia de $\alpha$

$TH \sim \mathcal{F}_{\alpha, \beta}$ ( distrib. F ) $\leadsto$ buscamos $\mathbb{P}(F \geq f_{crit})$

¿Cómo calculamos el F test? → Scipy.stats.f or scipy.stats.f_oneway

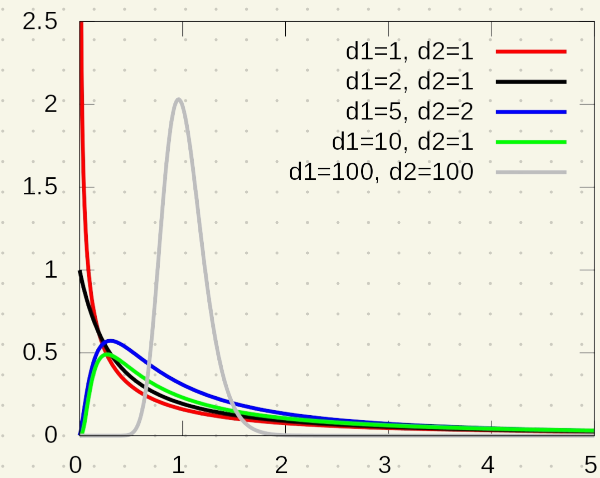
1. armamos la tabla ANOVA

2. obtenemos el estadístico $F = {ESS}/{S_r^2}$

3. Encontramos $f_{crit}$ (p-valor), definimos el nivel de significancia $\alpha$

$$f \sim \mathcal{F}_{gl_{ESS},\, gl_{RSS}}$$

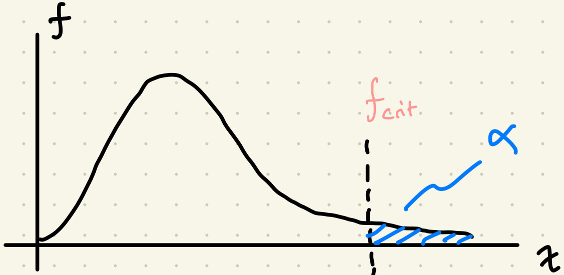
gl: grados de libertad




4. buscamos $\mathbb{P}(\mathcal{F} \geq f_{crit}) = \alpha$

5. si $F \geq f_{crit}$ recharazamos $H_0$

(4)



Adicional: Como determinamos el estadístico F en regresión multivariada:

Vamos a suponer que $X \in \mathbb{R}^{n \times p}$ donde $n$ es la cantidad datos y $p$ la cantidad de regresores (columnas).

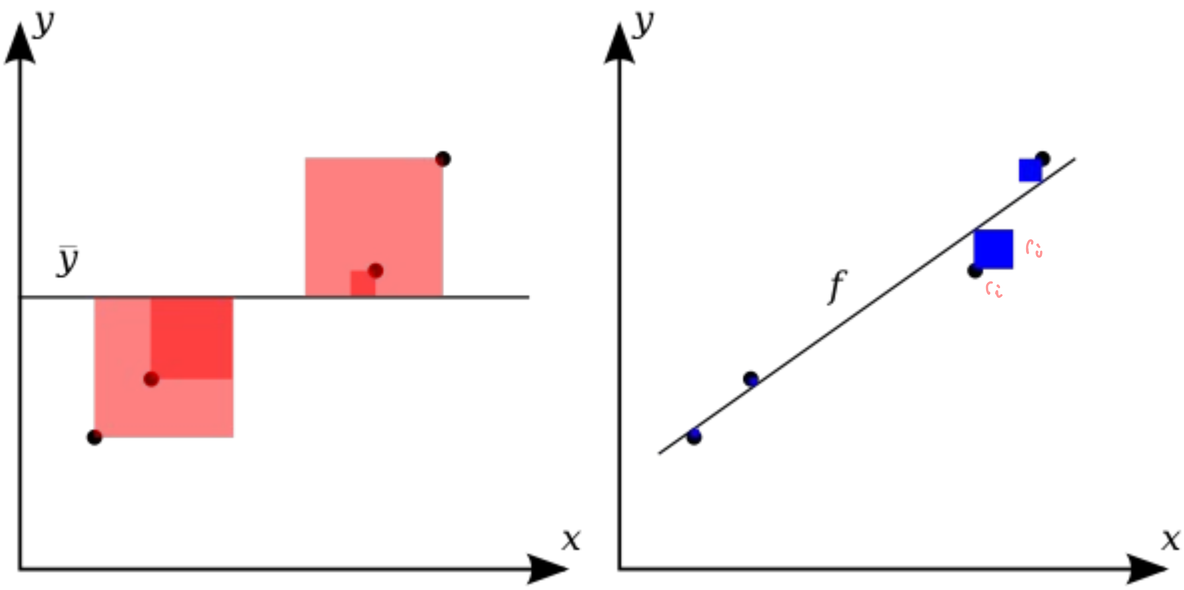
Armamos la tabla ANOVA de la siguiente forma.

| Fuente de variación | suma de los cuadrados | grados de libertad | cuadrados medios |
|---|---|---|---|
| Explicada | $\sum (\hat{y}_i - \bar{y})^2$ | $p-1$ | $S_E = \frac{1}{p-1} \sum_i (\hat{y}_i - \bar{y})^2$ |
| residual | $\sum r_i^2$ | $n-p$ | $S_R = \frac{1}{n-p} \sum_i r_i^2$ |
| total | $\sum (y - \bar{y})^2$ | $n-1$ | |

bajo estas condiciones podemos analizar múltiples test de hipótesis como:

- $H_0: \exists j / \beta_j = 0$; $H_A: \nexists j / \beta_j = 0$ – $H_0: \beta_i = \beta_j$; $H_A: \beta_i \neq \beta_j$ – etc.

## Regresión Lineal - R2




$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

ESS

$$SS_{reg} = \sum_i (f_i - \bar{y})^2$$

RSS

$$SS_{res} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

TSS

$$SS_{tot} = \sum_i (y_i - \bar{y})^2$$

$$= SS_{res} + SS_{reg}$$

¿Similar a σ2?

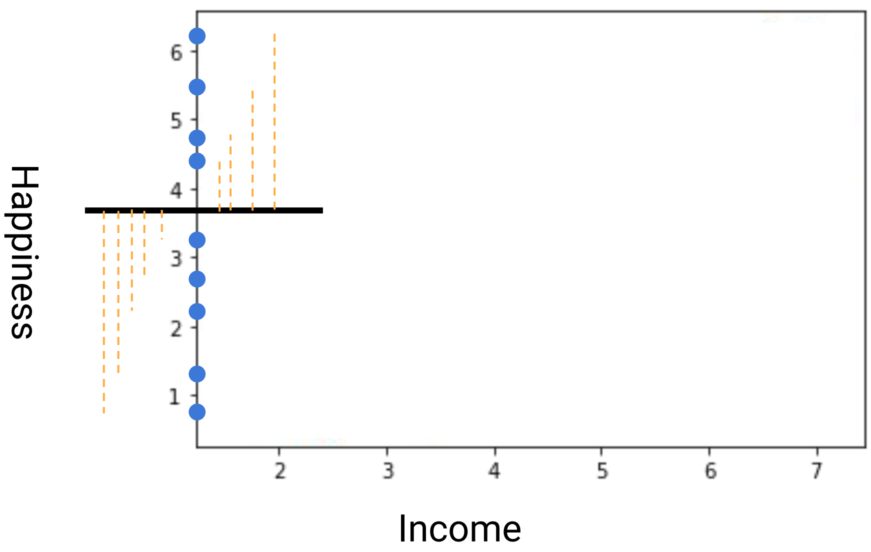
# Regresión Lineal - R2

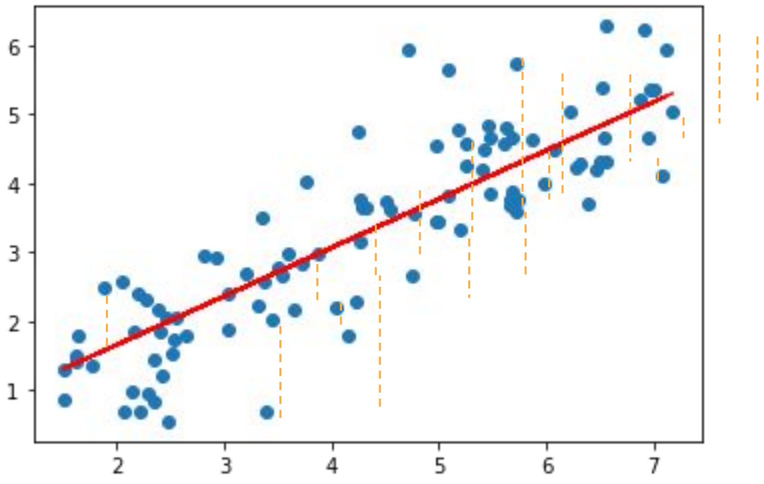
$$
\begin{aligned}
R^2 &= 1 - \frac{SS_{res}}{SS_{tot}} \\
&= 1 - \left(\frac{SS_{res}}{SS_{tot}} * \frac{n}{n}\right) \\
&= \boxed{1 - \boxed{\frac{\sigma_{res}}{\sigma_{tot}}}}
\end{aligned}
$$

Proporción de varianza no explicada

Proporción de varianza explicada

# Regresión Lineal - R2




$$SS(media) = (happiness - media)^2$$

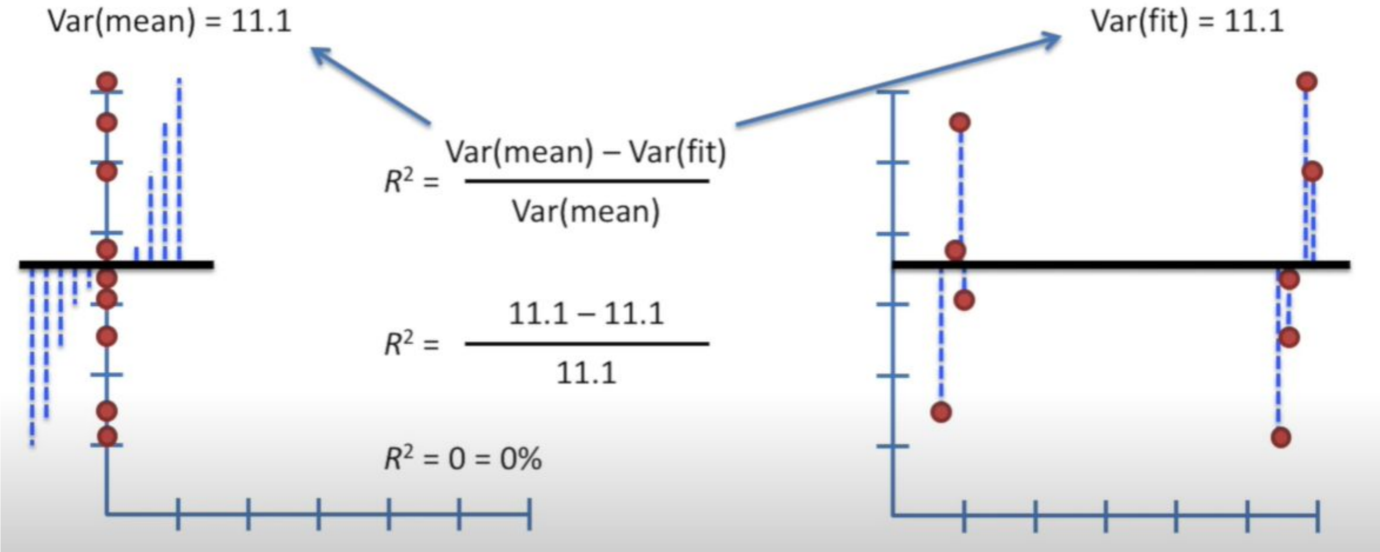
$$Variación(media) = \frac{(happiness - media)^2}{n}$$

$$SS(fit) = (happiness - lr\_fit)^2$$

$$Variación(fit) = \frac{(happiness - lr\_fit)^2}{n}$$

# Regresión Lineal - R2

$$R^2 = \frac{Variación(media) - Variación(fit)}{Variación(media)}$$




Fuente: StatQuest with Josh Starmer

## Regresión Lineal - R2

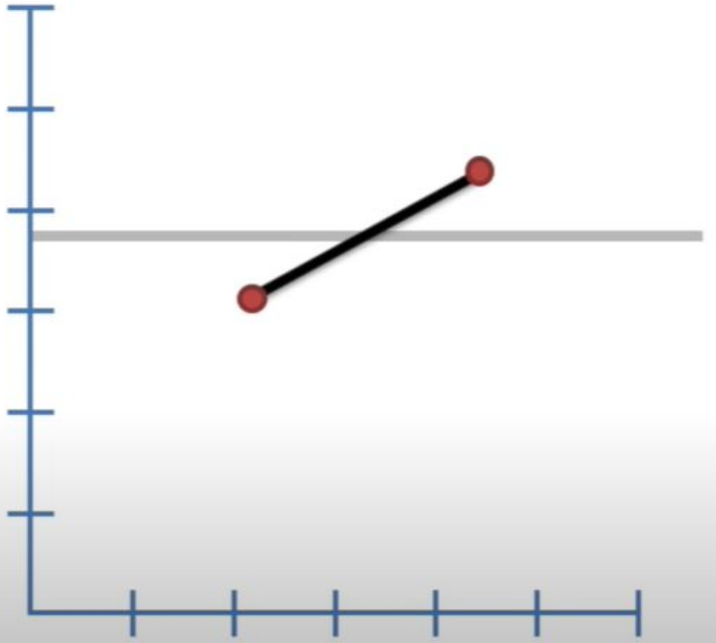
$$F = \frac{Varación\ en\ happiness\ explicada\ por\ income}{Variación\ en\ happiness\ no\ explicada\ por\ income}$$

**SS(mean)** = 10

**SS(fit)** = 0

$$R^2 = \frac{SS(mean) - SS(fit)}{SS(mean)}$$

$$= \frac{100 - 0}{100} = 100\%$$



Fuente: StatQuest with Josh Starmer

## R2 y el coeficiente de correlación de Pearson

$$\rho = \frac{cov(x,y)}{\sigma_x \sigma_y} \in [-1, 1]$$

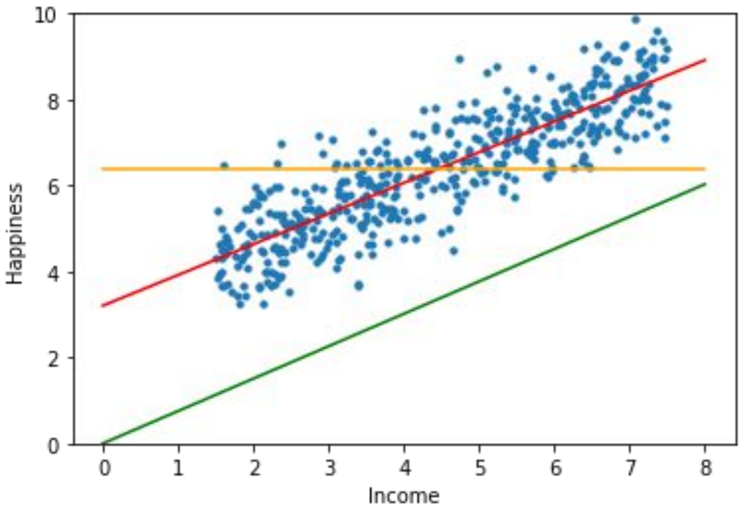
Regresión múltiple **con ordenada** →Correlación entre observación y predicción

Regresión **con ordenada** →Correlación entre variable dependiente e independiente

$$\rho^2 = R^2 \in [0, 1]$$

## ¿R2 negativo?

$$R^2 = 1 - \frac{\sigma_{res}}{\sigma_{tot}} < 0 \leftrightarrow \sigma_{res} > \sigma_{tot}$$

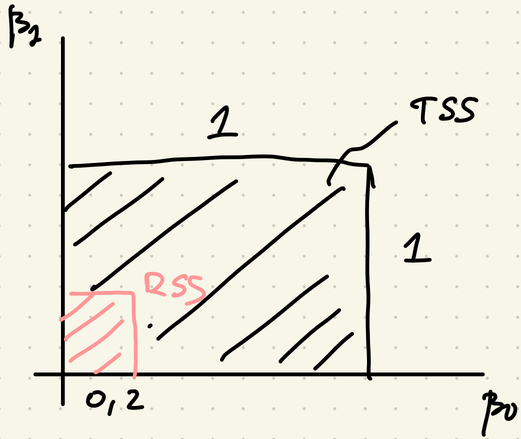



Predecir con el promedio es mejor que el modelo

## R2 inflation

$\uparrow$ cantidad de predictores $\longrightarrow \uparrow$ R2 $\longrightarrow$ F-test para comparación válida entre modelos.

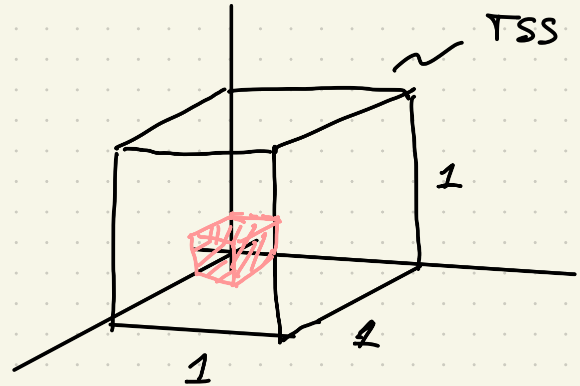
caso regresión lineal simple




$$R^2 = 1 - \frac{RSS_2}{TSS_2}$$

$$y = \beta_0 + \beta_1 x$$

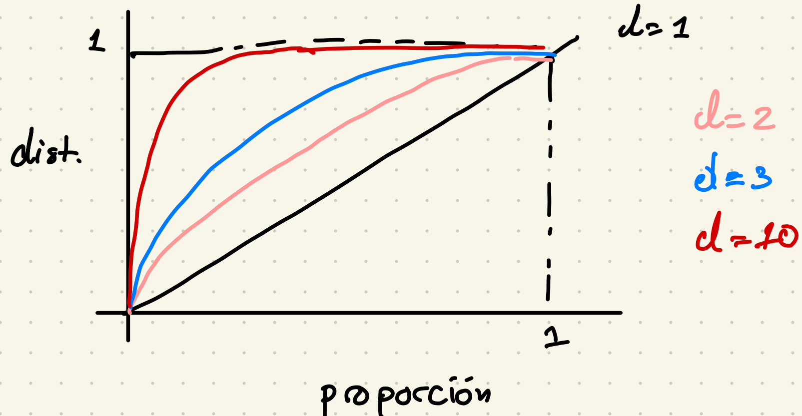
regresión lineal multivariada




$$1 - \frac{RSS_3}{TSS_3}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

R2 Inflation



## Otras medidas a tener en cuenta

Mean Absolute Error

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}.$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

→ Mean Square Error

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}.$$

↳ Root Mean Square Deviation

## Bias-Variance Tradeoff

Cuando utilizamos el **error cuadrático medio** en un modelo de ML, podemos descomponer el mísmo en términos de bias (sesgo) y variance (varianza).

$$MSE = \underbrace{Bias(\hat{f})^2} + Var(\hat{f}) + \sigma_\epsilon^2$$

$$Bias = E[\hat{f} - f]$$

$$Var(\hat{f}) = E[(E[\hat{f}] - \hat{f})^2]$$

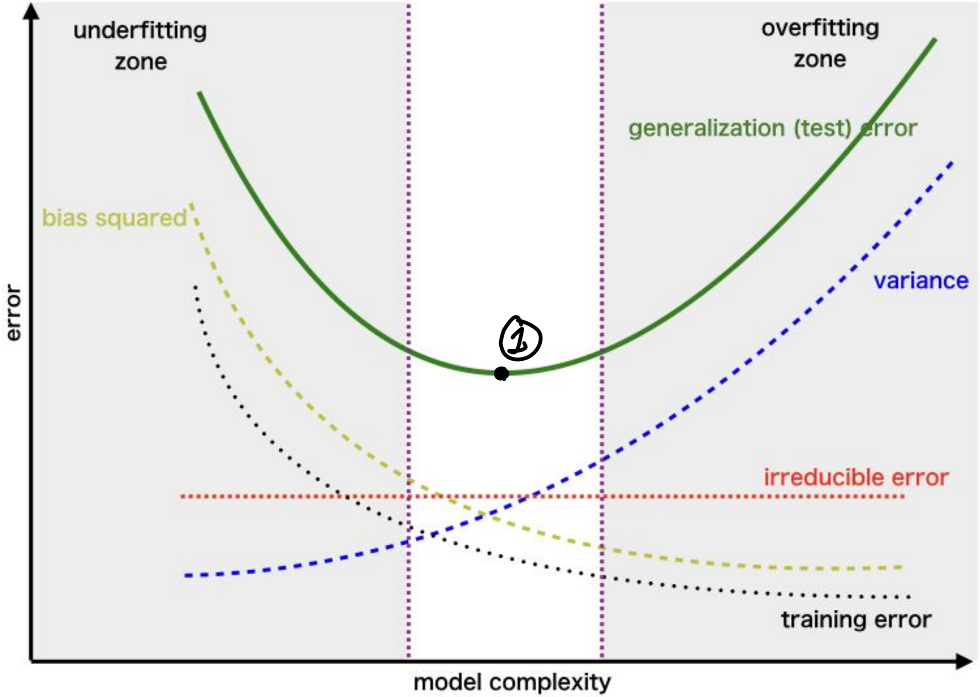
1. error de estimación.
2. sesgo.

$$\hat{\beta} - \beta = (X^t X)^{-1} X^t y - \beta$$
$$= (X^t X)^{-1} X^t (X\beta + \varepsilon) - \beta$$
$$= (X^t X)^{-1} X^t \varepsilon$$

$$E(\hat{\beta} - \beta) = (X^t X)^{-1} X^t \underbrace{E(\varepsilon)}_{=0} = 0$$

## Bias-Variance Tradeoff

(1) este punto se obtiene por optimización (early-stop) y la implementación depende del problema y del modelo.




↳ en gral # iteraciones
# epochs

# Bibliografía


- The Elements of Statistical Learning | Trevor Hastie | Springer
- An Introduction to Statistical Learning | Gareth James | Springer
- Deep Learning | Ian Goodfellow | https://www.deeplearningbook.org/
- Stanford | CS229T/STATS231: Statistical Learning Theory | http://web.stanford.edu/class/cs229t/
- Mathematics for Machine Learning | Deisenroth, Faisal, Ong
- Artificial Intelligence, A Modern Approach | Stuart J. Russell, Peter Norvig