Introducción a la Inteligencia Artificial

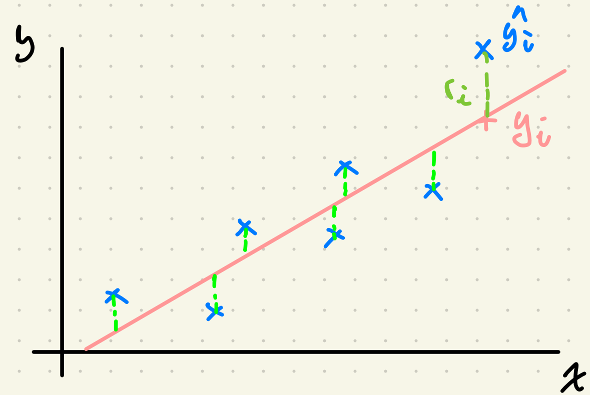
Clase ~~4~~ 3

.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 $\propto \sum_i r_i^2$

$r_i \sim \mathcal{N}(0, \sigma_r)$

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

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

$$\bar{r} = \underbrace{(\mathbb{I} - H)}_{\text{simétrica, idempotente, } tr = n-p}\bar{y}$$

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

⊕ 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:

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

$\hat{C}_{xy}$

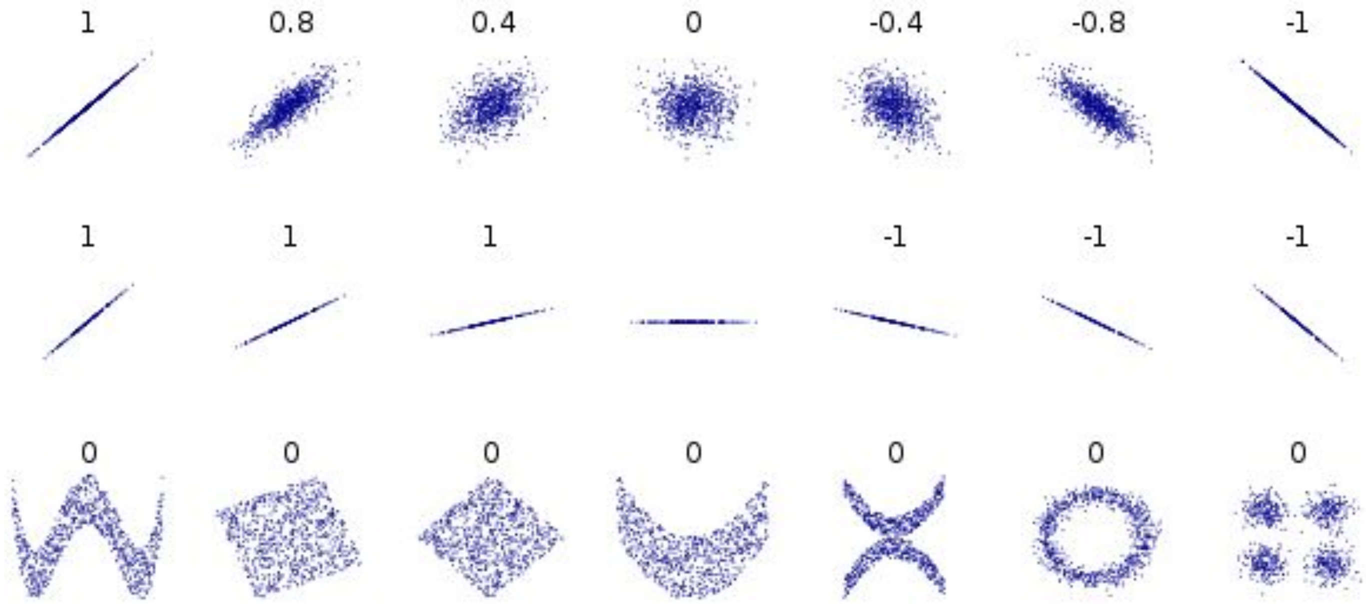
$\hat{\sigma}_x^2$

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

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.

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

① TSS: tasa de variabilidad total

② ESS: tasa de variabilidad explicada

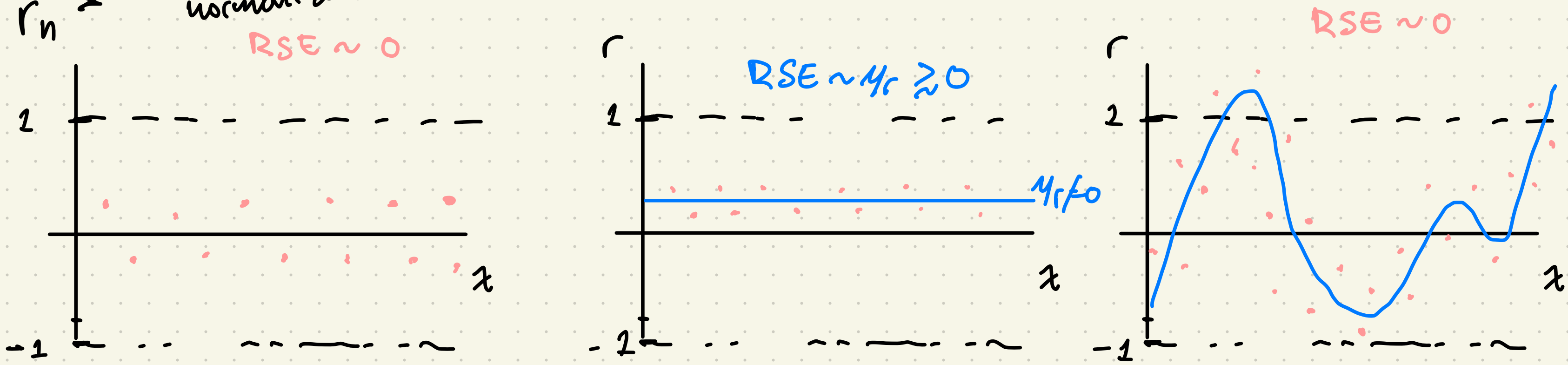
③ RSS: suma de los residuos al cuadrado

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

r_n ← residuo normalizado



• 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:

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

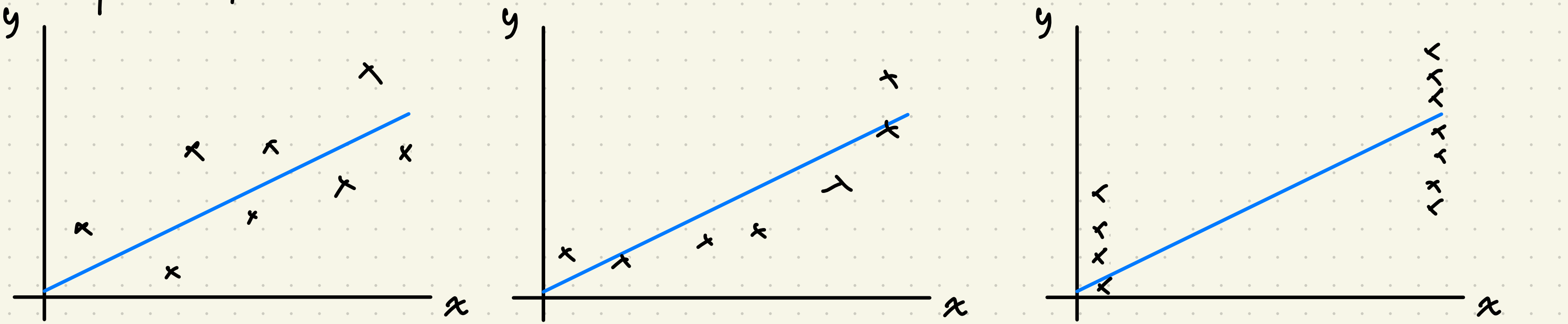
1 es valido únicamente bajo el régimen de reg. simple

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

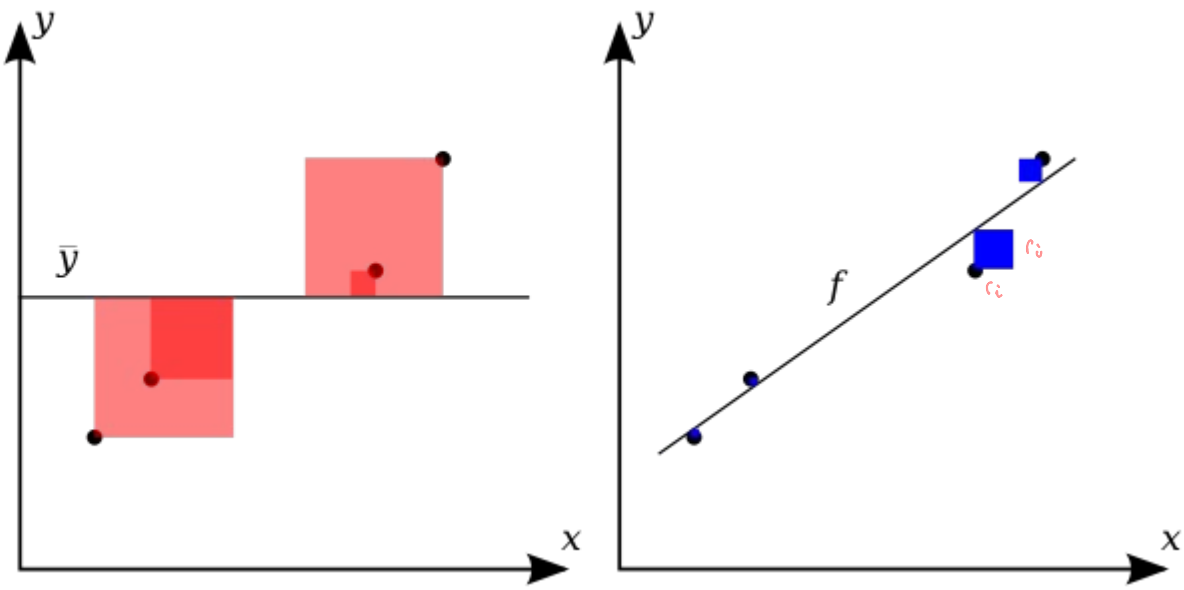
  si hay una reg lineal

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

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



Regresión Lineal - R2



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

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

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

¿Similar a σ2?

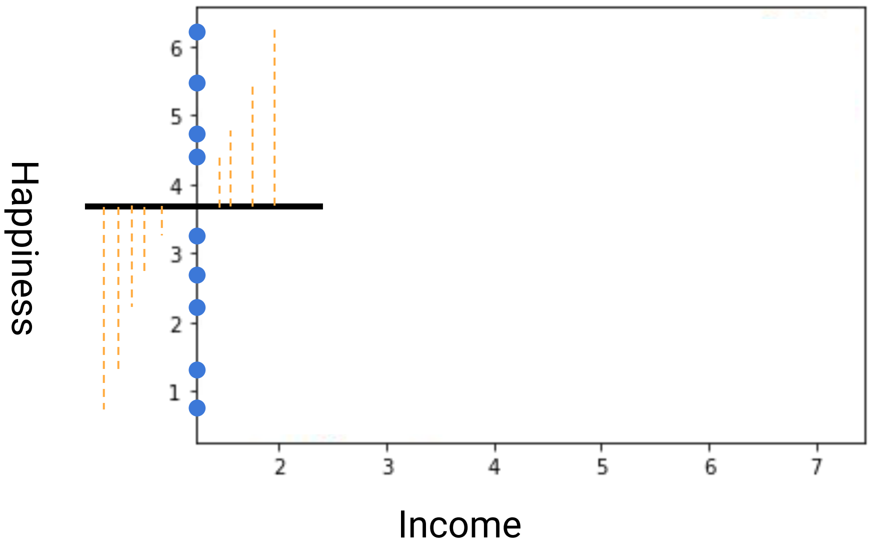
Regresión Lineal - R2

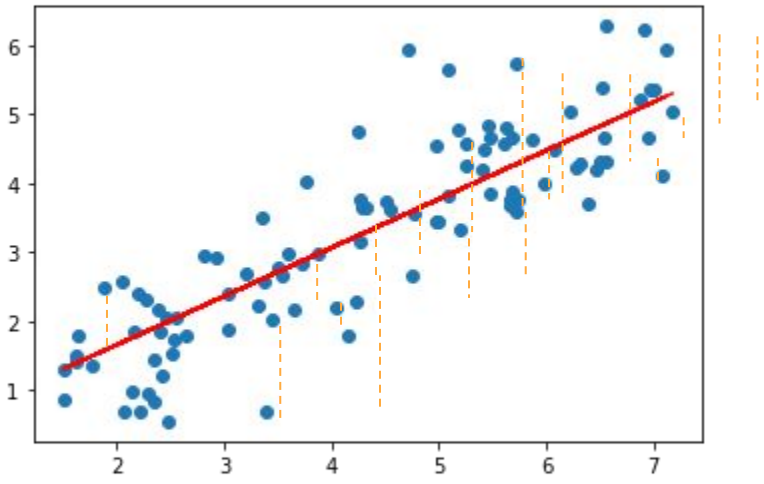
$$
\begin{aligned}
R^2 &= 1 - \frac{SS_{res}}{SS_{tot}} \\
&= 1 - \left(\frac{SS_{res}}{SS_{tot}} * \frac{n}{n}\right) \\
&= \boxed{1 - \boxed{\frac{\sigma_{res}}{\sigma_{tot}}}}
\end{aligned}
$$

Proporción de varianza no explicada

Proporción de varianza explicada

Regresión Lineal - R2





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

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

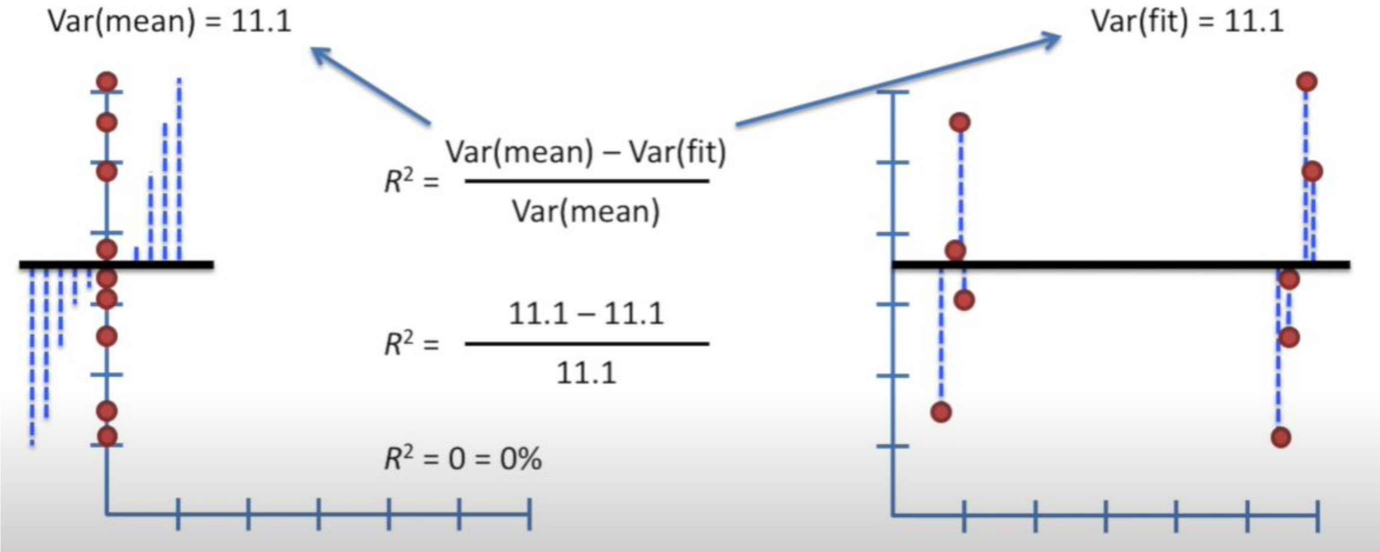
$$SS(fit) = (happiness - lr\_fit)^2$$

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

.UBAfiuba FACULTAD DE INGENIERÍA

Regresión Lineal - R2

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



Fuente: StatQuest with Josh Starmer

Regresión Lineal - R2

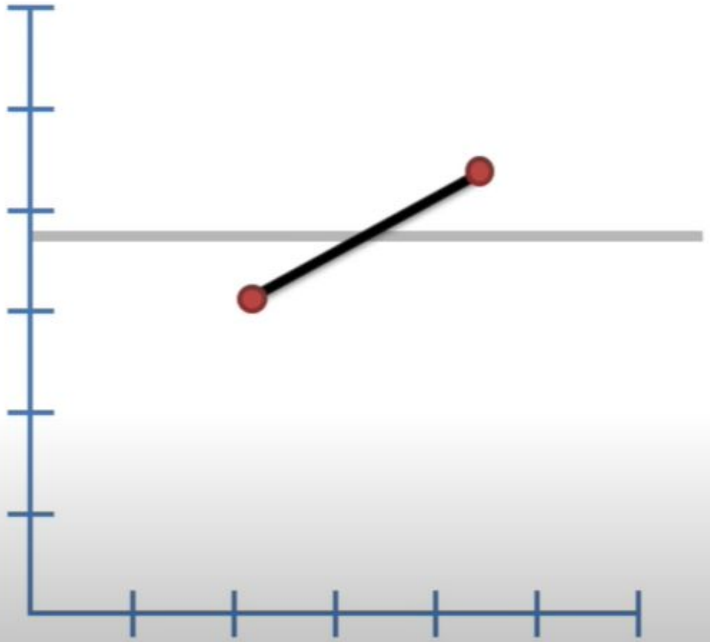
$$F = \frac{\textit{Varación en happiness explicada por income}}{\textit{Variación en happiness no explicada por income}}$$

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

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

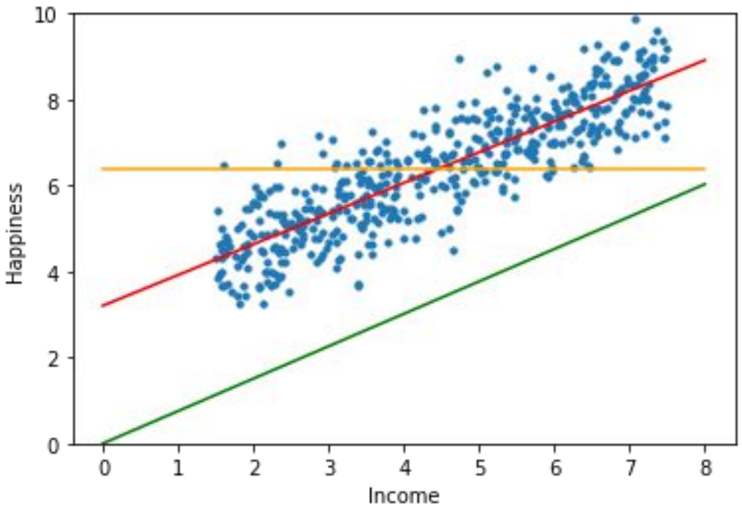
Regresión múltiple **con ordenada** →Correlación entre observación y predicción

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

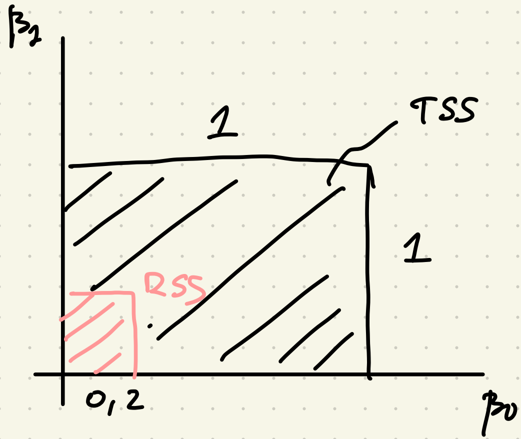
R2 inflation

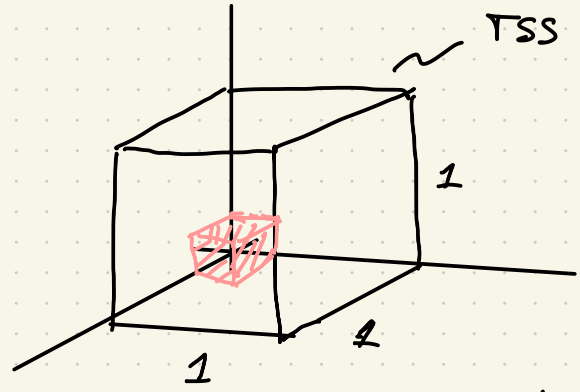
(caso regresión multivariada)

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

caso regresión lineal simple

regresión lineal multivariada





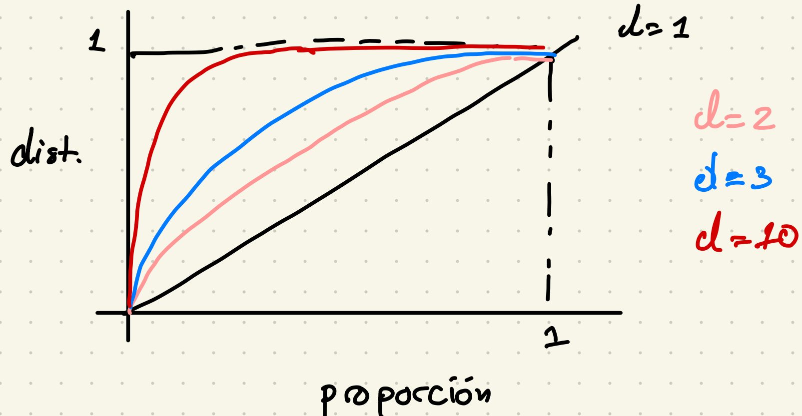
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

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

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

$y = \beta_0 + \beta_1 x$

$\boxed{\text{R2 Inflation}}$



Estadístico F:

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.

con esto planteamos nuestro test de hipótesis:

$$TH \begin{cases} H_0 : \beta_j = 0 \quad \forall j = 0 \\ H_1 : \beta_j \neq 0 \quad j \in [1, p] \end{cases}$$

Aquí buscamos rechazar H_0 con una significancia de α

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

Repaso

¿Qué es un test de hipótesis?

Un test de hipótesis es un procedimiento estadístico que permite tomar una decisión acerca de una afirmación o supuesto (hipótesis) con base en los datos recolectados. Existen dos tipos principales de hipótesis:

1. Hipótesis nula (H0): Propone que no hay una relación significativa o un efecto notable entre dos fenómenos. Es la hipótesis que se asume verdadera hasta que se demuestre lo contrario.
2. Hipótesis alternativa (H1): Sostiene que hay una relación significativa o un efecto notable entre dos fenómenos. Esta hipótesis se acepta si los datos recolectados proporcionan suficiente evidencia en contra de la hipótesis nula.

El proceso de prueba en sí consiste en recoger datos, analizarlos y luego decidir si la evidencia respalda la hipótesis nula o la alternativa. Si la evidencia va en contra de la hipótesis nula (basado en un nivel de significancia preestablecido, usualmente 0.05), "rechazamos" la hipótesis nula y aceptamos la alternativa. Por otro lado, si la evidencia no es suficientemente fuerte, "aceptamos" la hipótesis nula, que no significa necesariamente que la hipótesis nula sea verdadera, sino que no tenemos suficientes pruebas para rechazarla. Finalmente, es interesante notar que estas conclusiones están sujetas a errores. Por ejemplo, podríamos rechazar una hipótesis nula verdadera (Error tipo I) o aceptar una hipótesis nula falsa (Error tipo II).

¿Cómo calculamos el F test? $\rightarrow$ Scipy.stats.f or scipy.stats.f_oneway

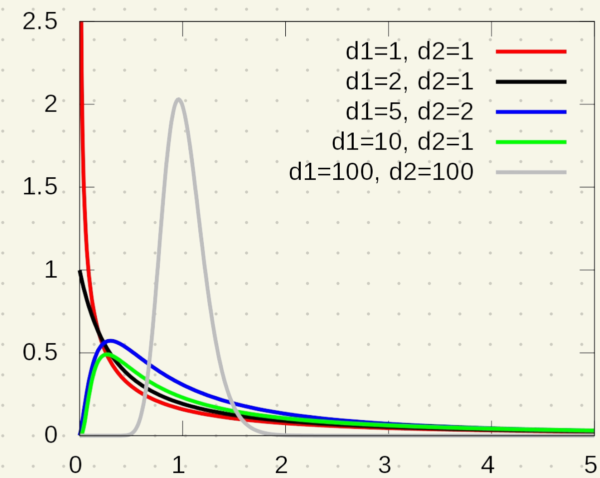
1. armamos la tabla ANOVA

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

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

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

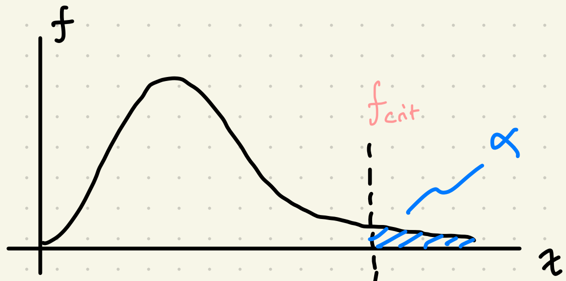
gl: grados de libertad



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

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

4



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).

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

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

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

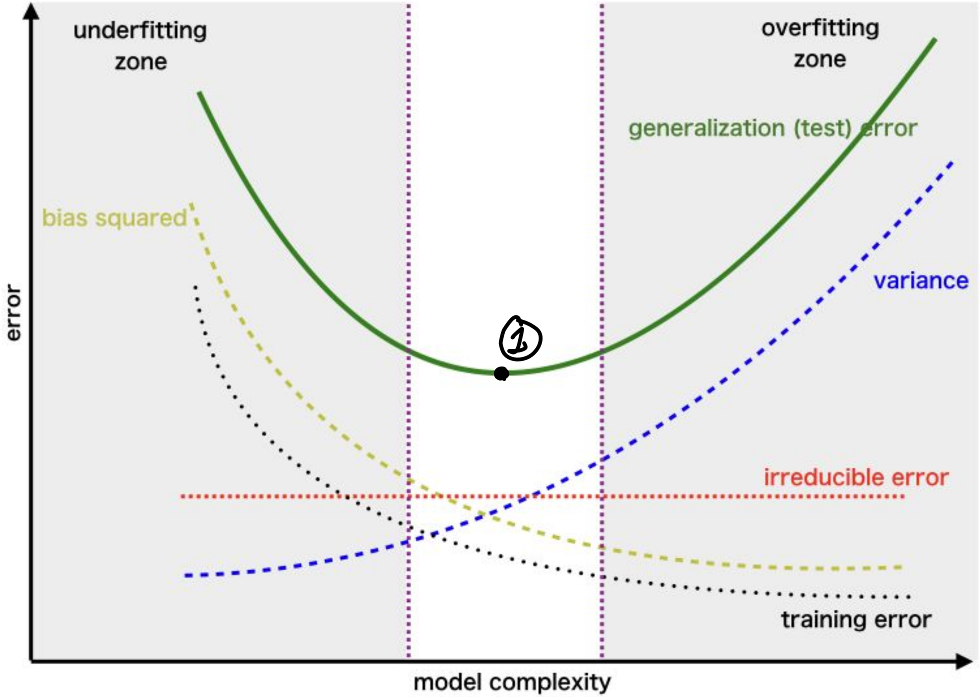
1. error de estimación.
2. sesgo.

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

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

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

cota cramer-rao de varianza de un estimador

Nota adicional:

Existen múltiples test de interés en la regresión lineal, en los vamos a tener relacionados a los param. del modelo (los β's), estos buscan responder:

+ ¿cual es la validez del modelo? + ¿Cuales son variables importantes?

En gral planteamos:

- $H_0: \bar{\bar{R}}\bar{\beta} = \bar{r}$ vs. $H_1: \bar{\bar{R}}\bar{\beta} \neq \bar{r}$ donde $\bar{\bar{R}} \in \mathbb{R}^{J\times p} \wedge \bar{r} \in \mathbb{R}^{J\times 1}$

Algunos casos de interés:

$H_0: \beta_1 = \beta_2 = \dots = 0 \quad (\beta_j = 0 \ \forall j \in [1, p-1])$ test de bondad de la regresión

Aquí:

$$\bar{\bar{R}} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \in \mathbb{R}^{p-1, p} \qquad \bar{r} = [0, \dots, 0]$$

Bibliografía

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