

Regresión lineal

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n,$$

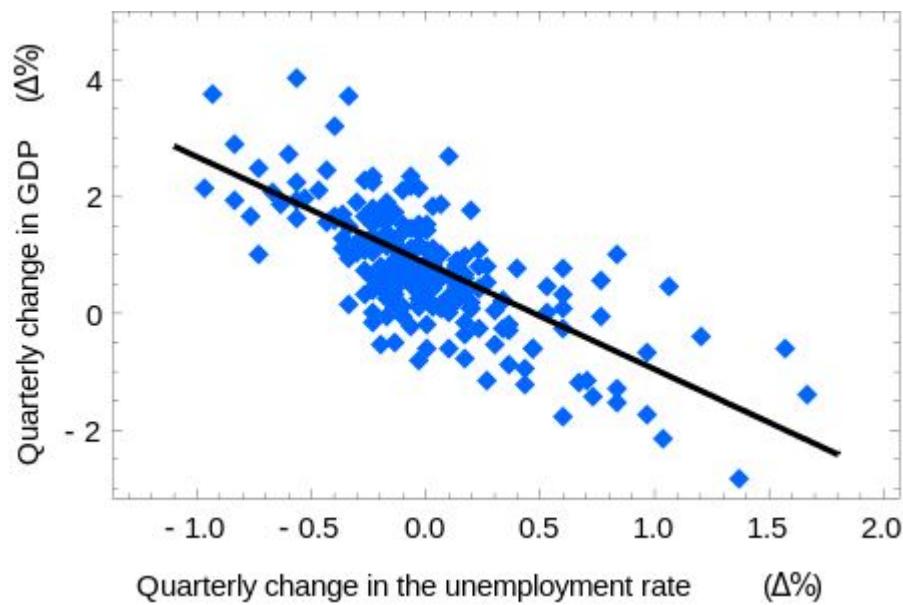
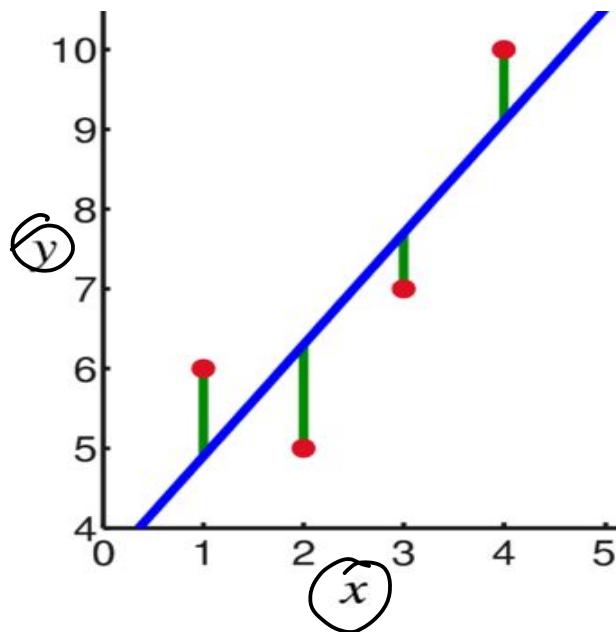
$$\hat{y} = \sum_{j=0}^r \beta_j x_j$$



Regresión Lineal

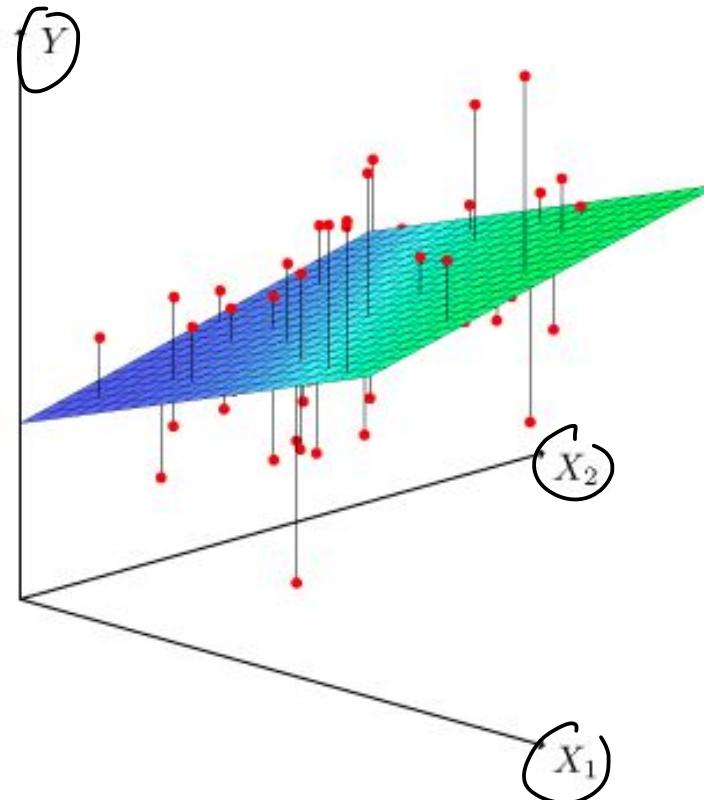
$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n,$$

En ésta clase vamos a ver el framework teórico detrás de la gran mayoría de los modelos de Machine Learning: aprendizaje estadístico. Para ello, vamos a utilizar como modelo base la regresión lineal.



Regresión Lineal - Teoría

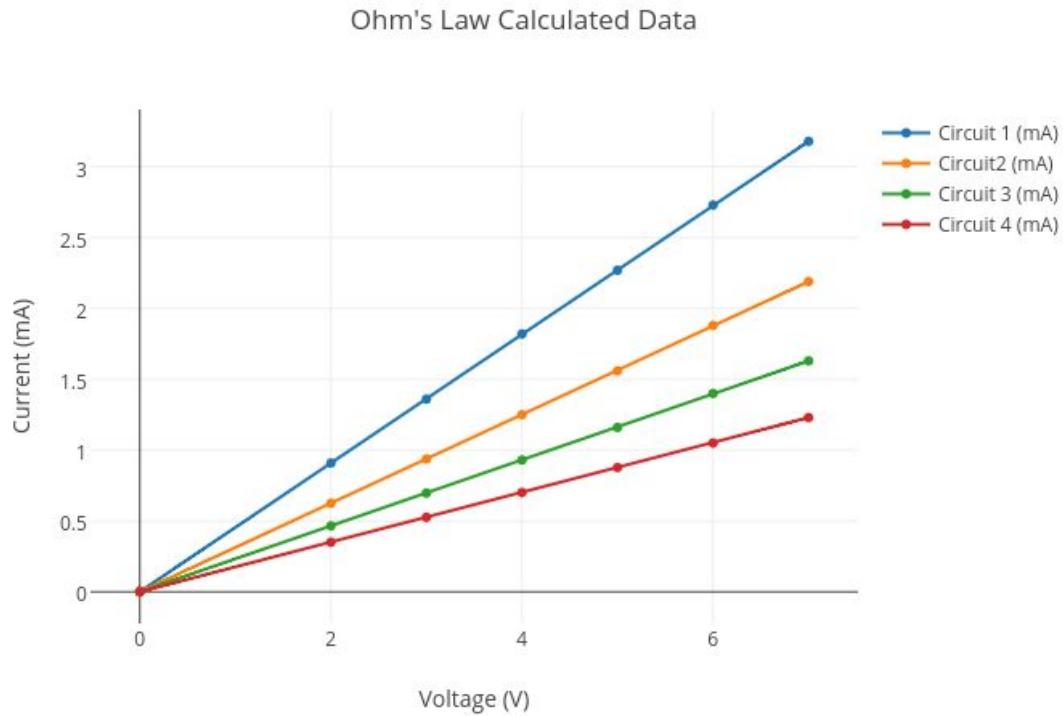
Regresión Lineal $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n,$



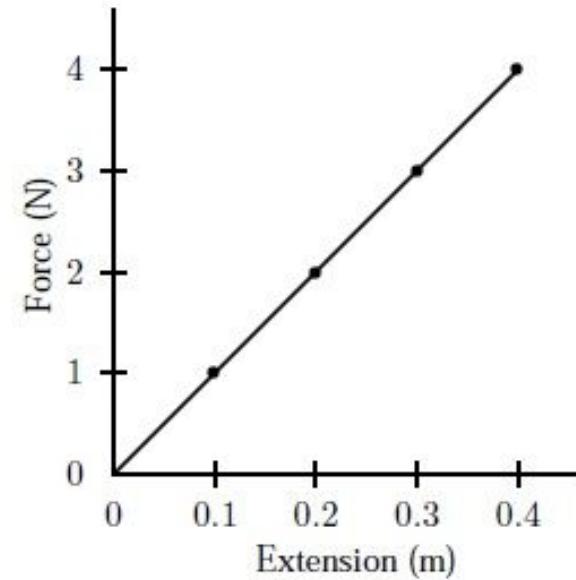
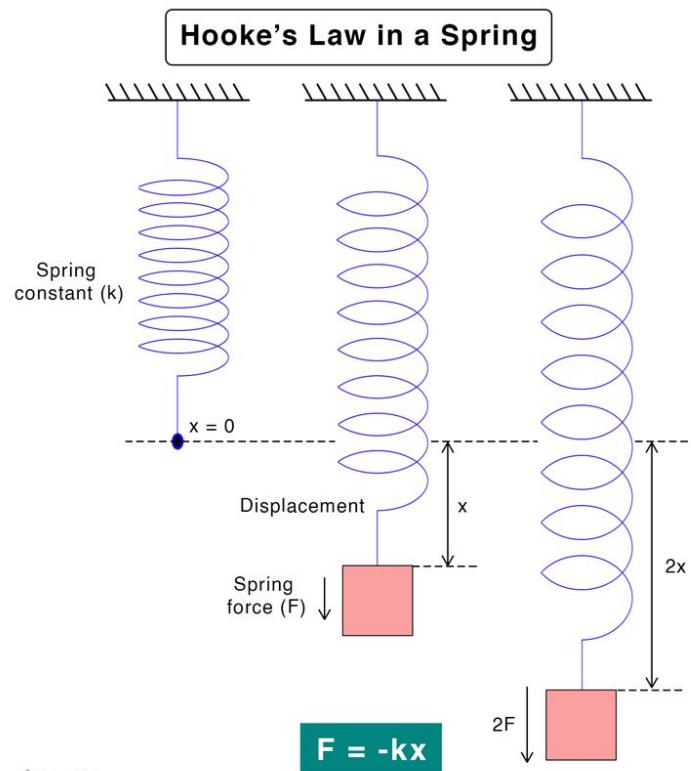
Ley de Ohm

$$I = V/R$$

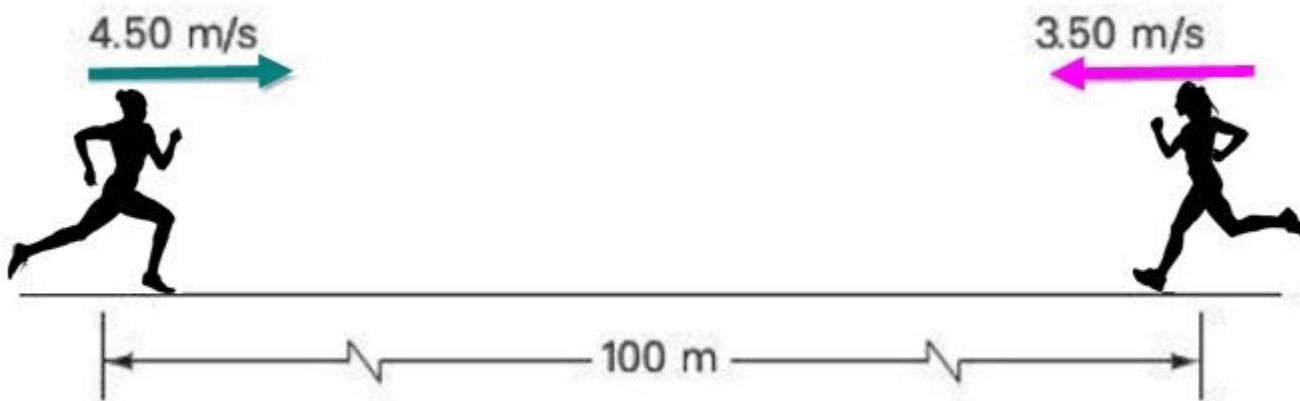
R constante



Ley de Hooke



Movimiento rectilíneo uniforme



$$x(t) = x(t_0) + V * t$$

Población de parásitos



Ejemplo: En un estudio sobre la población de un parásito se hizo un recuento de parásitos en 15 localizaciones con diversas condiciones ambientales.

Los datos obtenidos son los siguientes:

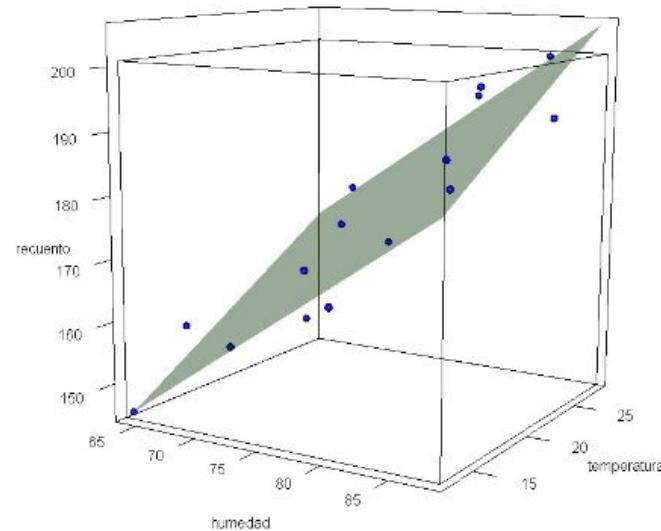
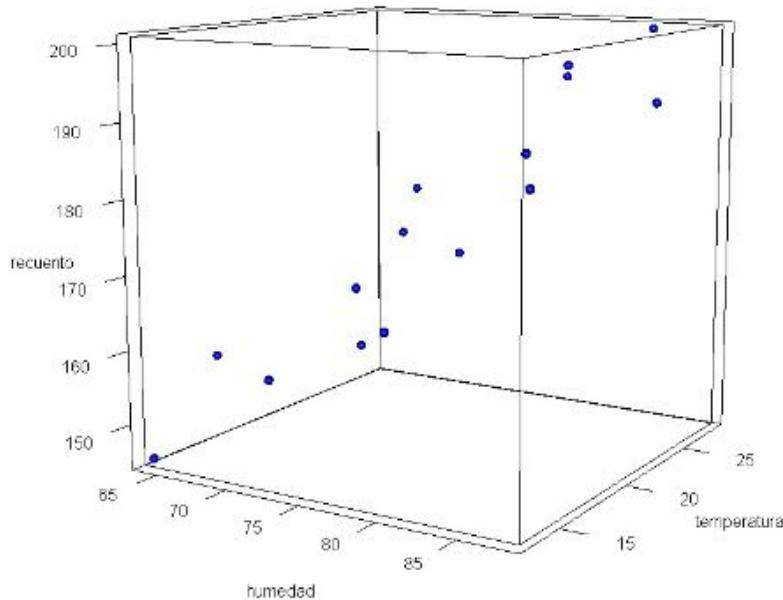
Temperatura	15	16	24	13	21	16	22	18	20	16	28	27	13	22	23
Humedad	70	65	71	64	84	86	72	84	71	75	84	79	80	76	88
Recuento	156	157	177	145	197	184	172	187	157	169	200	193	167	170	192

Fuente:

Población de parásitos

$$\text{Recuento} = \beta_0 + \beta_1 \text{Temperatura} + \beta_2 \text{Humedad} + \epsilon$$

$$\text{Recuento} = 25.7115 + 1.5818 \text{Temperatura} + 1.5424 \text{Humedad}$$



Jamboard



consideremos un conjunto de datos (observaciones) x_1, \dots, x_n con $x_i \in \mathbb{R}^d$ son las mediciones del sistema. Además vamos a considerar $y_i \in \mathbb{R}$ el conj de respuestas del sistema. llamamos a \bar{x} **variables regresoras** e y **variable de resp / dependiente**.

En gral en ml buscamos encontrar una relación entre y & \bar{x} :

$$y = f(\bar{x}; \theta) + \varepsilon \leftarrow \text{comp. aleatorio}$$

en fn de aproximación

En regresión buscamos inferir $\hat{y} = \hat{f}(x)$, la precisión de la estimación podemos separarla en dos componentes **reducible** (depende de los datos) e **irreducible**

Como buscamos mejorar el error reducible, tenemos que optimizarlo. Suponemos \bar{x} fijo y f conocida, voy a calcular una forma de error, en particular el **Error cuadrático medio**:

$$\mathbb{E}(\gamma - \hat{\gamma})^2 = \mathbb{E}(f(x) + \varepsilon - \hat{f}(x))^2$$

(*) mediante supuestos.

$$= \underbrace{\mathbb{E}(f(x) - \hat{f}(x))^2}_{\text{error reducible}} + \underbrace{\mathbb{E}(\varepsilon)^2}_{\text{Error irreducible}}$$

$\varepsilon \sim N(0, \sigma)$ indep.
 ε no depende de X .

la f más sencilla que podemos pensar es una comb. lineal de los param. (es simple, es barata, es explicable, \sim precisa)

$$\hat{f}(\bar{x}, \bar{p}) = \beta_0 + \sum_{i=1}^D \beta_i x_i$$

Supuestos del modelo lineal:

0. Existe una relación lineal entre X e Y .
1. Los regresores son independientes $\rightarrow P(x_1, \dots, x_n) = P(x_1) \dots P(x_n)$
2. Ausencia de colinealidad $\rightarrow \beta(i,j), i \neq j / \beta_i x_i + \beta_j x_j = x_k \forall k$
3. El proceso de generación de datos es homocelástico \rightarrow (2)

① \rightarrow los E_i iid γ no dependen de los datos ($E_i \sim N(0, \sigma^2) \forall i$)

con estos supuestos limitamos la familia de f 's que modelen el sist. γ con esto podemos tomar 3 métodos:

② MSE (Mean Square error) \rightarrow Enfoque Empírico

③ ML (Maximum Likelihood) \rightarrow Enf. probabilístico

④ MAP (Maximum a posteriori) \rightarrow " bayesiano.

① MSE

partimos de un dataset $D = \{(x_i, y_i) \mid i \in [1, \dots, K] \quad x_i \in \mathbb{R}^{m \times 1}\}$ y construimos el error:

$$E(y - \hat{y})^2 = E(\beta) = \sum_{i=1}^K (y_i - \hat{f}(\beta_i))^2 = \sum_{i=1}^K \left(y_i - \beta_0 - \sum_{j=1}^m x_{ij} \cdot \beta_j \right)^2 \quad (1)$$

a $x_j = [x_1, \dots, x_m]$ le voy a agregar un 1 como prefijo para representar a β_0 $\Rightarrow x_j' = [1, x_1, \dots, x_m]$ con esto:

reescribimos ①:

$$\begin{aligned}\mathcal{E}(\beta) &= \sum_{i=1}^k \left(y_i - \sum_{j=0}^m x_{ij} \cdot \beta_j \right)^2 \\ &= (\bar{y} - \bar{x}\bar{\beta})^t (\bar{y} - \bar{x}\bar{\beta})\end{aligned}$$

$$\bar{y} = [y_0, \dots, y_k]$$

$$\bar{\beta} = [\beta_0, \beta_1, \dots, \beta_m]$$

$$\bar{x} = \begin{bmatrix} 1 & x_{11} & \dots & x_{m1} \\ 1 & x_{21} & \dots & x_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{k1} & \dots & x_{mk} \end{bmatrix}$$

Vamos a minimizar ② $\rightarrow \partial_{\bar{\beta}} \mathcal{E}(\beta) = 0$

$$\begin{aligned}\partial_{\beta} \mathcal{E} &= \partial_{\beta} [(\bar{y} - \bar{x}\bar{\beta})^t (\bar{y} - \bar{x}\bar{\beta})] = -2\bar{x}^t(\bar{y} - \bar{x}\bar{\beta}) = 0 \\ &= \bar{x}^t(\bar{y} - \bar{x}\bar{\beta}) = \bar{x}^t\bar{y} - \boxed{\bar{x}^t\bar{x}} \cdot \bar{\beta} \quad \xrightarrow{\text{X}^t \cdot \text{X matriz de diseño}}\end{aligned}$$

$$\hat{\beta} = (\bar{x}^t\bar{x})^{-1} \cdot \bar{x}^t\bar{y}$$

`np.inv(X.T @ X).dot(X.T @ y)`

$$\hat{y} = X\hat{\beta} = \underbrace{X(X^t X)^{-1} X^t}_H y \quad \hookrightarrow \hat{y} = H y$$

la parte más difícil (y costosa) de esto es calcular $(X^t X)^{-1}$. Sobre todo si $k \gg m$ (y viceversa) Vemos que no existe la inversa \Rightarrow ③

② → para estos casos utilizamos *pseudovolos inversos*

(2) Maximum Likelihood (método de máxima verosimilitud)

bajo las cond. de la reg. lineal estamos diciendo que existe una distrib. de y condicionada para cada x , $P(y/x=x, \bar{\beta}, \sigma^2) \sim N$

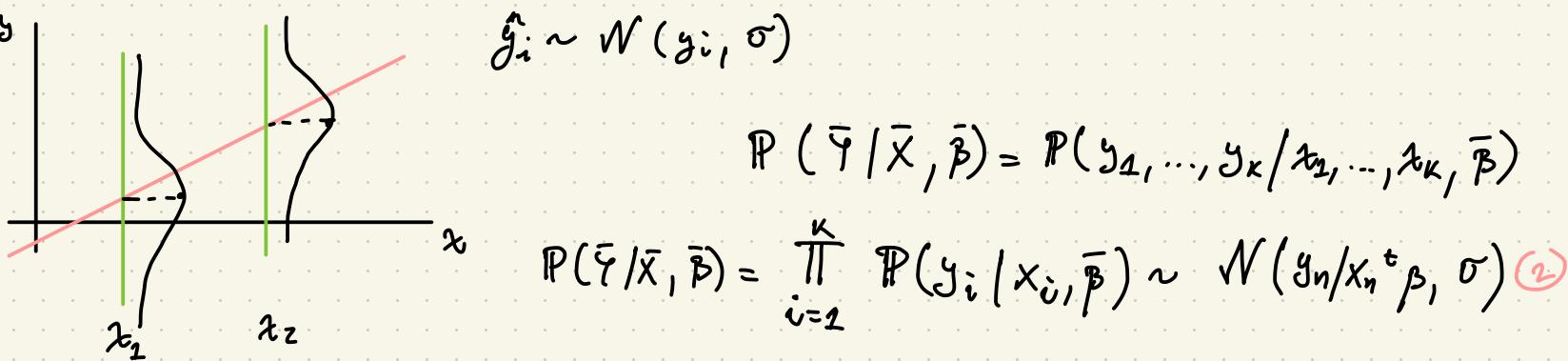
Dado los pares $(x_1, y_1), \dots, (x_K, y_K)$ podemos escribir lo siguiente:

$$\prod_{i=1}^K P(y_i/x_i, \beta, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(y_i - \beta_0 - \sum_j \beta_j x_{ij})^2}{2\sigma^2}} \quad (1)$$

Esta función la conocemos como fn. de verosimilitud $L(\bar{\beta}, \sigma)$ de los parámetros y los datos. La forma funcional proviene de propagar la distrib. que conocemos $\epsilon_i \sim N(0, \sigma^2)$.

Con esto buscamos $\max_{\beta} L$:

$$\exists \hat{\beta} / \max_{\beta} L \rightarrow \text{partimos de } y_i = f(x_i, \beta) + \epsilon$$



$$\hat{y}_i \sim N(y_i, \sigma)$$

$$P(\bar{Y} | \bar{X}, \bar{\beta}) = P(y_1, \dots, y_k | x_1, \dots, x_k, \bar{\beta})$$

$$P(\bar{Y} | \bar{X}, \bar{\beta}) = \prod_{i=1}^k P(y_i | x_i, \bar{\beta}) \sim N(y_i | x_i^T \bar{\beta}, \sigma^2)$$

con esto $\hat{\beta}_{ML} = \arg \max (\textcircled{2}) :$

$$\hat{\ell} = \arg \max_{\bar{\beta}} \prod_{i=1}^k P(y_i | x_i, \bar{\beta}, \sigma^2)$$

Si intentamos maximizar $\hat{\ell}$ el problema se torna muy complicado.
Por esto utilizaremos la versión logarítmica (log-likelihood) :

Log likelihood

$$\ell(\bar{\beta}) = \sum_{i=1}^k \frac{1}{2\sigma^2} (y_i - x_i^T \bar{\beta})^2 = \frac{1}{2\sigma^2} \underbrace{(\bar{y} - \bar{x}\bar{\beta})^T (\bar{y} - \bar{x}\bar{\beta})}_{\|\bar{y} - \bar{x}\bar{\beta}\|^2}$$

$$\ell(\bar{\beta}) = \frac{1}{2\sigma^2} \|\bar{y} - \bar{x}\bar{\beta}\|^2 \quad \textcircled{3}$$

optimizamos ③ :

$$\partial_{\beta} l = 0 \rightarrow \partial_{\beta} \left(\frac{1}{2\sigma} (\bar{y} - \bar{x}\bar{\beta})^t (\bar{y} - \bar{x}\bar{\beta}) \right) = 0$$

$$\partial_{\bar{\beta}} (y^t y - 2y^t x \beta + \beta^t x^t x \beta) = 0$$

$$0 - 2y^t x + 2\beta^t x x^t = 0$$

$$-y^t x + \beta^t x^t x = 0 \rightarrow \hat{\beta}_{ML} = (x^t x)^{-1} x^t y$$

Lo mas interesante de esto es que ML y Empírico obtienen el mismo resultado, y ademas es de solución cerrada.

MAP (Maximum a posteriori) Enfoque Bayesiano

En los métodos que vimos anteriormente no ponemos suposiciones sobre los parámetros θ . El método MAP propone asumir la distribución 'a priori' $p(\theta)$. Esto, restringe los valores que pueden tomar. Vamos a considerar $p(\theta) \sim \mathcal{N}(0, 1)$, esto va a limitar el valor de $\theta \in [-2, 2]$ con alta probabilidad (esto es $\pm 2\sigma_\theta$). Teniendo el dataset (X, Y) , en vez de maximizar la fn. de verosimilitud, vamos a buscar los parámetros θ que maximizan la distribución a posteriori $p(\theta | X, Y)$. Si aplicamos el teorema de Bayes:

Teorema de Bayes:

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

$$P(\theta | X, Y) = \frac{P(Y | X, \theta) P(\theta)}{P(Y | X)}$$

M1

En la ec. M1 vamos a buscar

θ_{MAP} que maximize la distib. a posteriori.

Vamos a utilizar un truco similar al log usado en ML.

$$\log(P(\theta | X, Y)) = \log(P(Y | X, \theta)) + \log(P(\theta)) + \text{cte.} \quad \text{M2}$$

no depende de θ

Para encontrar θ_{MAP} , planteamos:

$$\theta_{MAP} \in \operatorname{argmin} \{-\log P(Y | X, \theta) - \log P(\theta)\}$$

Para esto vamos a considerar:

$$-\partial_{\theta} \log p(\theta | x, y) = -\partial_{\theta} \log p(y | x, \theta) - \partial_{\theta} \log p(\theta)$$

Sabiendo que $p(\theta) \sim \mathcal{N}(\phi, b^2 \mathbb{I})$, $\phi = [0, \dots, 0] \in \mathbb{R}^D$; $b^2 \mathbb{I} = \begin{bmatrix} b & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & b \end{bmatrix}$ podemos obtener:

$$-\partial_{\theta} \log p(\theta | x, y) = \partial_{\theta} \left(\frac{1}{2\sigma^2} (y - \Phi \theta)^T (y - \Phi \theta) + \frac{1}{2b^2} \theta^T \theta + \text{cte} \right) \quad (\text{M3})$$

donde Φ es la matriz de features $[\mathbb{1}^T, \bar{x}] = \begin{bmatrix} 1 & x_1 & \dots & x_n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_D & \dots & x_n \end{bmatrix}$

A partir de (M3):

$$-\partial_{\theta} \log P(\theta | x, y) = \frac{1}{\sigma^2} (\theta^T \Phi^T \Phi - y^T \Phi) + \frac{1}{b^2} \theta^T$$

tomando $-\partial_{\theta} \log P(\theta | x, y) = 0$

$$\frac{1}{\sigma^2} (\theta^T \Phi^T \Phi - y^T \Phi) + \frac{1}{b^2} \theta^T = 0 \implies \theta^T \left(\frac{1}{2\sigma^2} \Phi^T \Phi + \frac{1}{b^2} \mathbb{I} \right) - \frac{1}{\sigma^2} y^T \Phi = 0$$

Continuando:

$$\Theta^t \left(\Phi^t \Phi + \frac{\sigma^2}{b^2} \mathbb{I} \right) = y^t \Phi \Rightarrow \Theta^t = y^t \Phi \left(\Phi^t \Phi + \frac{\sigma^2}{b^2} \mathbb{I} \right)^{-1}$$

Con esto obtenemos el estimador MAP

$$\Theta_{MAP} = \left(\Phi^t \Phi + \frac{\sigma^2}{b^2} \mathbb{I} \right)^{-1} \Phi^t y$$

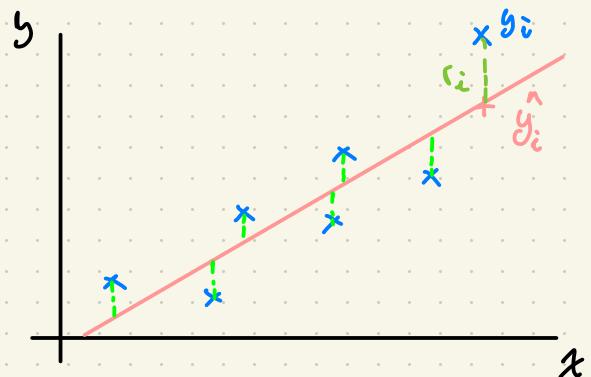
Si vemos el resultado obtenido es muy similar al obtenido previamente salvo por el término $\sigma^2/b^2 \mathbb{I}$. Este término nos asegura que el término a invertir sea simétrico y definido estricto positivo. Esto asegura la existencia de la inversa $\Rightarrow \Theta_{MAP}$ tiene solución única.

Finalmente, Θ_{MAP} tiene un efecto regularizador sobre los parámetros que luego aprovecharemos.

Índice

1. Notas clase anterior
2. Análisis de la regresión lineal (R^2)
3. Descomposición Bias-Variance
4. Práctica

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



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

con esto podemos definir los residuos

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

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

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

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

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

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

simétrica,

idempotente,

$$tr = n-p$$

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

⊕

⊕ Acá \bar{y} y \bar{x} representan los promedios.

$$\textcircled{1} \rightarrow \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

$$\sum (y_i x_i - \bar{\beta}_0 x_i - \bar{\beta}_1 \bar{x}_i^2) = 0$$

$$\sum (y_i x_i - \bar{y} x_i - \bar{\beta}_1 \bar{x}_i - \bar{\beta}_1 \bar{x}_i^2) = 0$$

Desarrollando (ver apunte) llegamos a:

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

C_{xy}

$\hat{\sigma}_x^2$

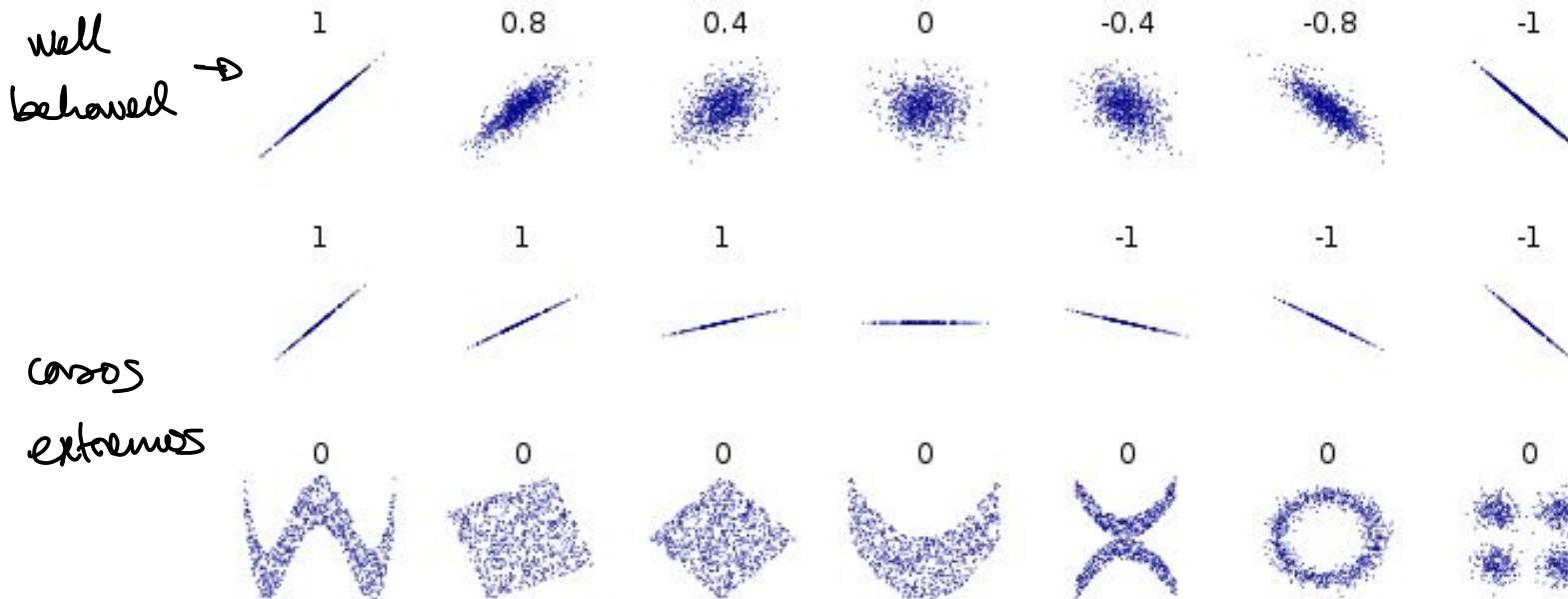
$$\left\{ \begin{array}{l} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \frac{\text{covar}(x, y)}{\text{Var}(x)} = \rho_{xy} \end{array} \right.$$

coef. de correlación
lineal de Pearson

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

Coeficiente de correlación de Pearson

(Lineal)

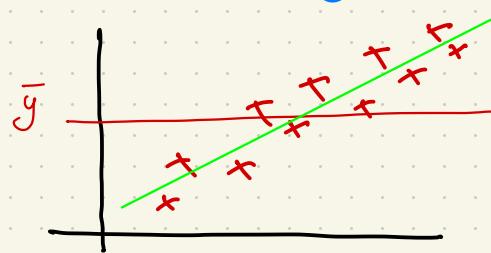


Analizamos los errores de la regresión:

$$y_i = \hat{y}_i + r_i \quad \bar{y} : \text{media mvestral (pomeclio)}$$

/ el término
creado
de confor-
por indep.

- ① TSS : varza de Variabilidad total
 - ② ESS : varza de Variabilidad explicada
 - ③ RSS : Suma de los residuos al cuadrado

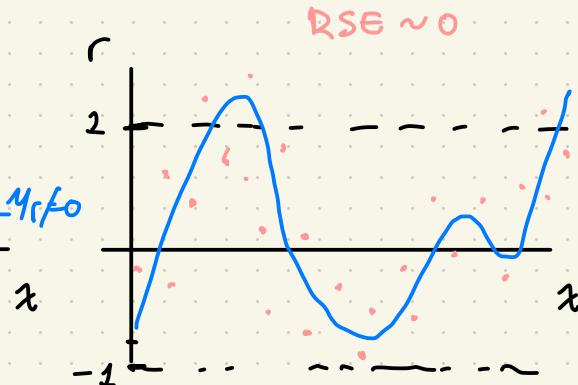
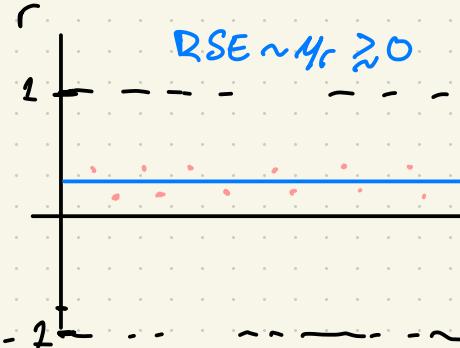
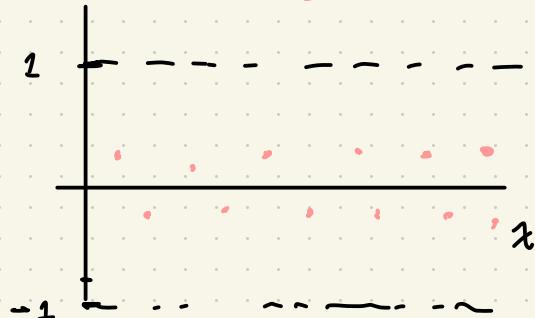


Estos valores me permiten construir métricas de bondad de ajuste para diagnosticar mi modelo.

métricos posibles → error residual (RSE)
R² (coef. de pearson)
Est. F (Análisis "Avanzado")

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

r_n \leftarrow residuo
normalizado
 $RSE \approx 0$



- Si RSE es bajo entonces nuestro ajuste es bueno
puede que sea
- RSE no es sensible a la distribución y/o tendencia funcional de los residuos.

Coeficiente de determinación - “R cuadrado”

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

RSS
TSS

SS = Sum of squares

res = residuos

tot = total

+ coef de pearson R^2 :

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \stackrel{\textcircled{1}}{=} (\beta_{xy})^2$$

① es válido únicamente bajo el régimen de reg. simple

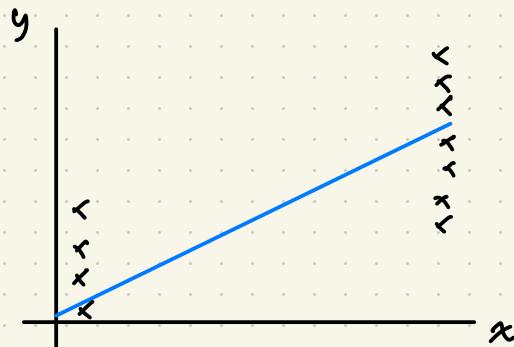
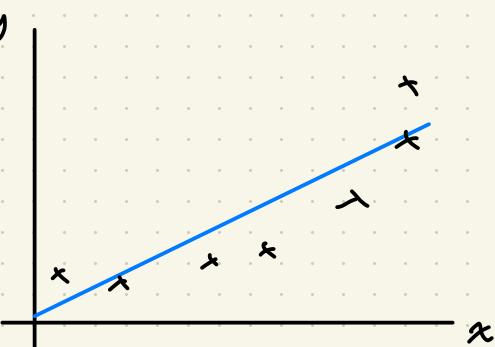
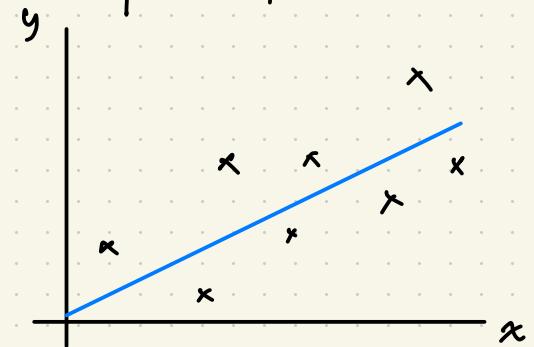
. R^2 no depende de las escalas, solo de penle de los proporciones.

. $R^2 \in [0, 1]$ $\Rightarrow R^2 \approx 1$ ms el modelo es "bueno"

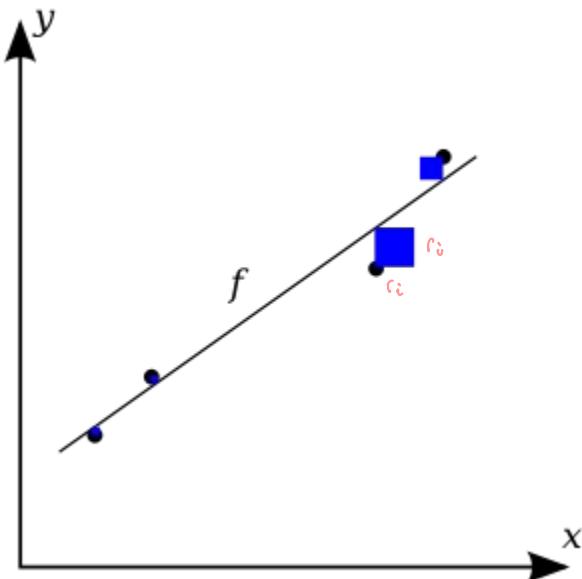
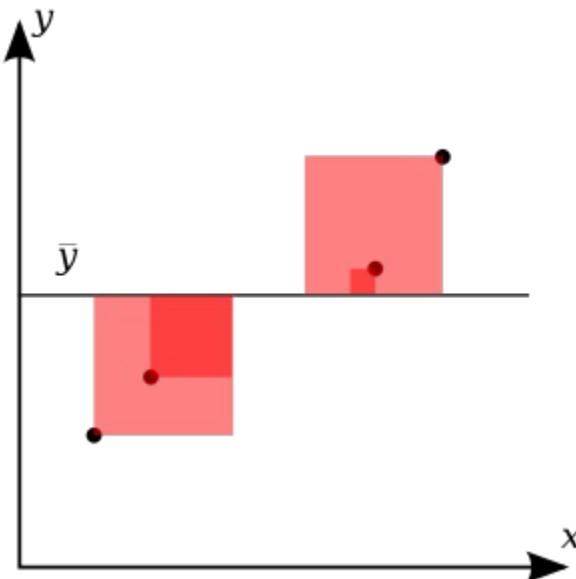
si hay una
reg lineal

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

$$\beta_0 \approx 3, \beta_1 \approx 0.5, R^2 \approx 0.7$$



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

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

¿Similar a σ^2 ?

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

Regresión Lineal - R2

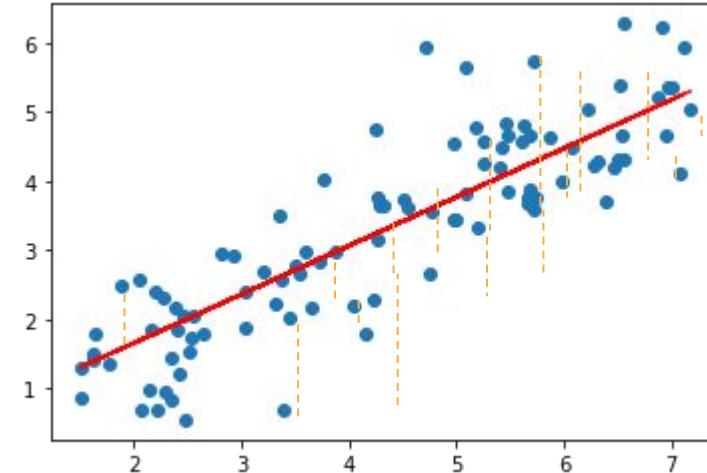
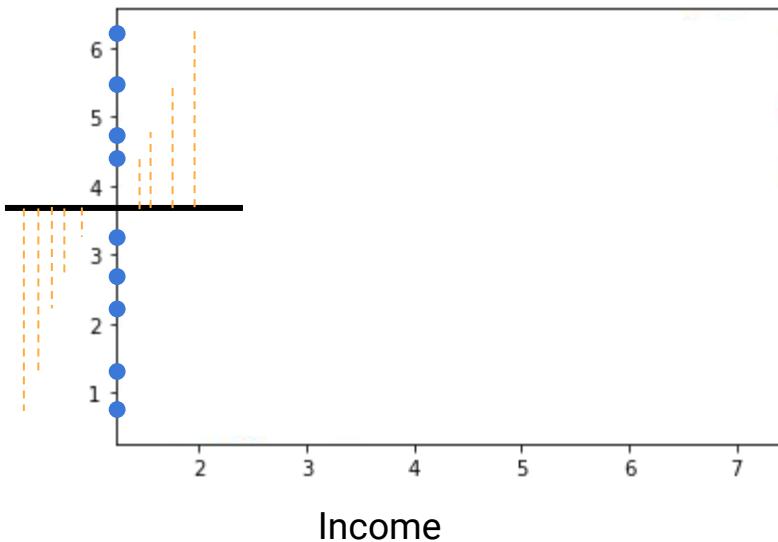
$$\begin{aligned} R^2 &= 1 - \frac{SS_{res}}{SS_{tot}} \\ &= 1 - \left(\frac{SS_{res}}{SS_{tot}} * \frac{n}{n} \right) \\ &= 1 - \boxed{\frac{\sigma_{res}}{\sigma_{tot}}} \end{aligned}$$

Proporción de varianza no explicada

Proporción de varianza explicada

Regresión Lineal - R2

Happiness



$$SS(\text{media}) = (happiness - \text{media})^2$$

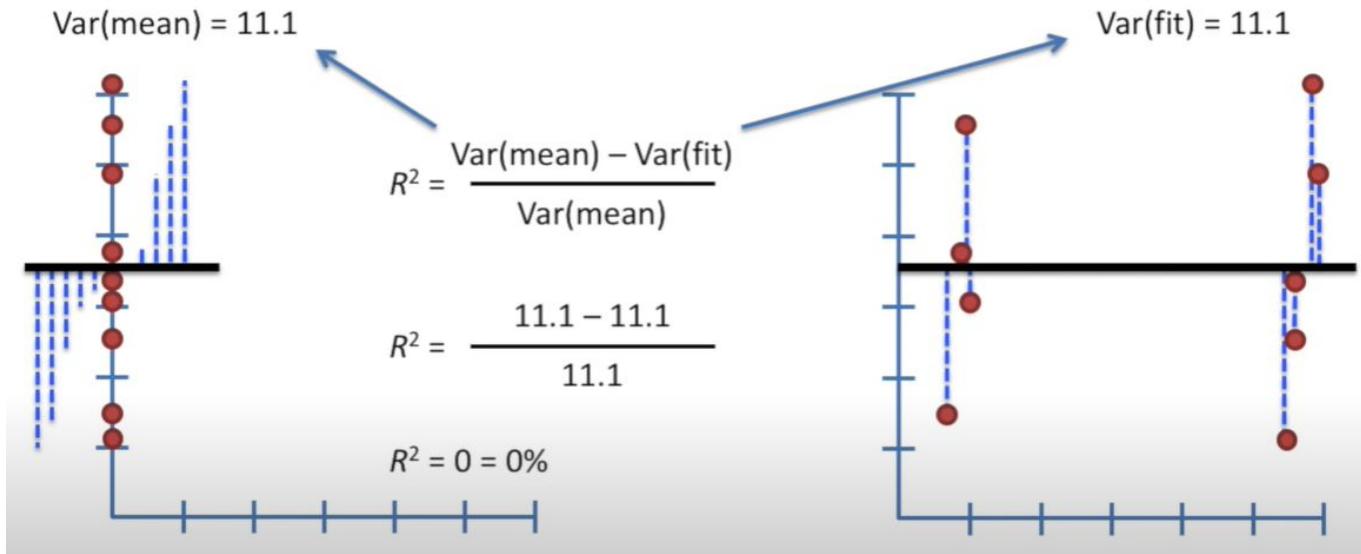
$$\text{Variación}(\text{media}) = \frac{(happiness - \text{media})^2}{n}$$

$$SS(\text{fit}) = (happiness - lr_fit)^2$$

$$\text{Variación}(\text{fit}) = \frac{(happiness - lr_fit)^2}{n}$$

Regresión Lineal - R²

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



Fuente: StatQuest with Josh Starmer

Regresión Lineal - R²

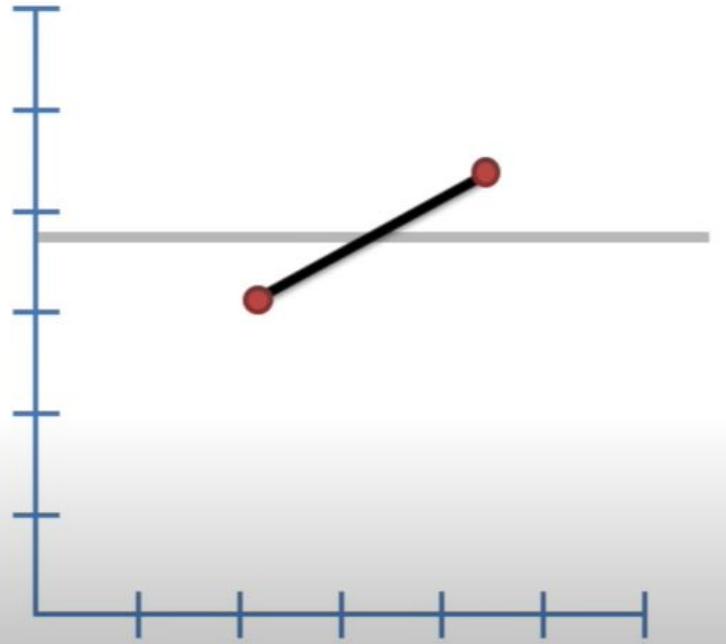
$$F = \frac{\text{Varación en happiness explicada por income}}{\text{Variación en happiness no explicada por income}}$$

$$\text{SS(mean)} = 10$$

$$\text{SS(fit)} = 0$$

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

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



Fuente: StatQuest with Josh Starmer

R2 y el coeficiente de correlación de Pearson

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

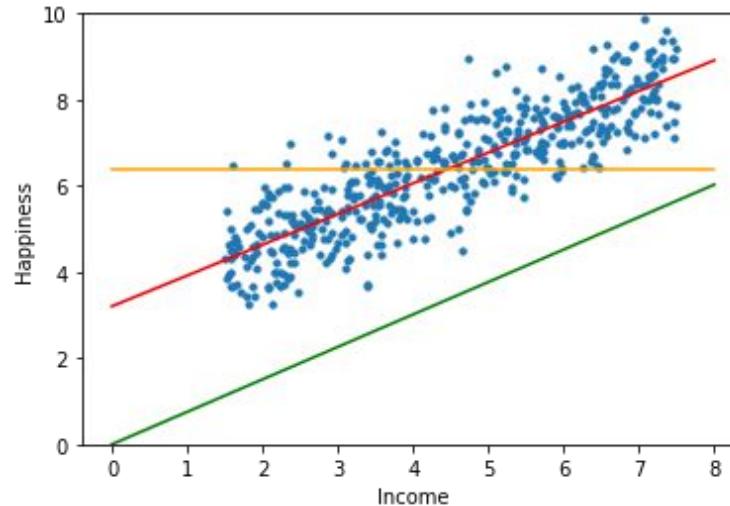
Regresión múltiple **con ordenada** → Correlación entre observación y predicción

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

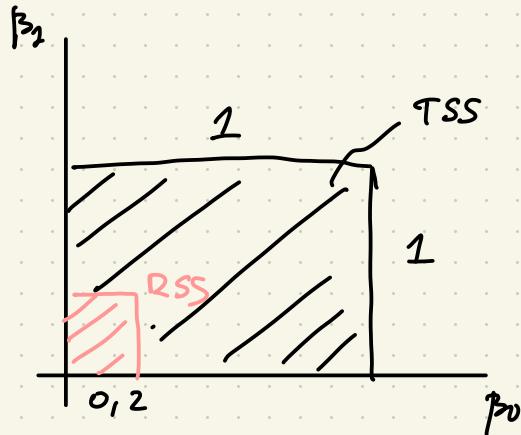
R2 inflation

(caso regresión multivariada)

↳ Esta relacionado con el efecto de la maldición de la dimensionalidad

↑ cantidad de predictores → ↑R2 → F-test para comparación válida entre modelos.

caso regresión lineal simple

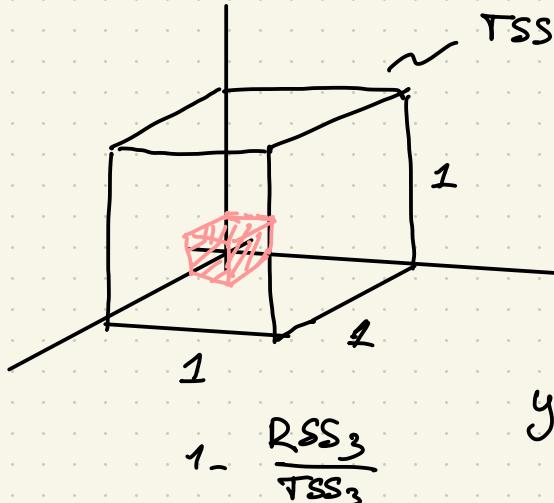


$$R^2 = 1 - \frac{RSS_2}{TSS_2}$$

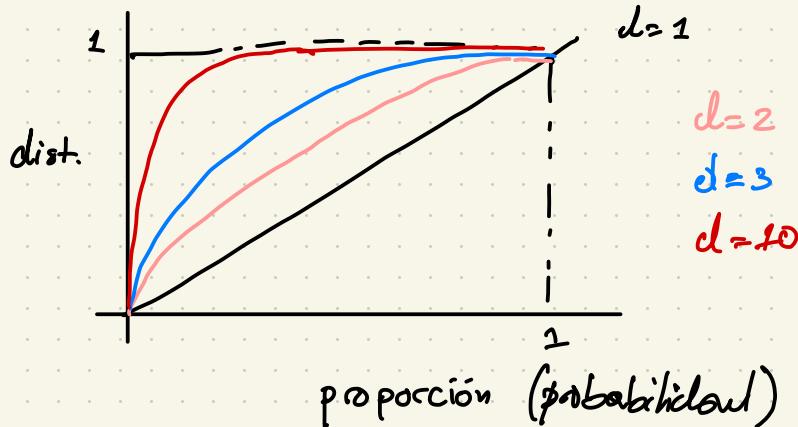
$$y = \beta_0 + \beta_1 x$$

R₂ Inflation

regresión lineal multivariada



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



proporción (probabilidad)

Estadístico F:

Vamos a suponer que $X \in \mathbb{R}^{n \times p}$ donde n es la cantidad de 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:

- $H_0: \beta_j / \beta_0 = 0$; $H_A: \beta_j / \beta_0 \neq 0$
- $H_0: \beta_i = \beta_j$; $H_A: \beta_i \neq \beta_j$ - etc.

con esto planteamos nuestro test de hipótesis:

$$T_H \begin{cases} H_0 : \beta_j = 0 \quad \forall j \neq 0 & \text{Aquí buscamos rechazar } H_0 \\ H_1 : \beta_j \neq 0 \quad j \in \{1, p\} & \text{con una significancia de } \alpha \end{cases}$$

$$T_H \sim F_{\alpha, \beta} \text{ (distrib. F)} \text{ nos buscamos } P(F \geq f_{crit})$$

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 (H_0): 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 (H_1): 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? → Scipy.stats.f or scipy.stats.f_oneway

1. armamos la tabla ANOVA

2. obtenemos el estadístico $F = \frac{ESS}{S^2_r}$

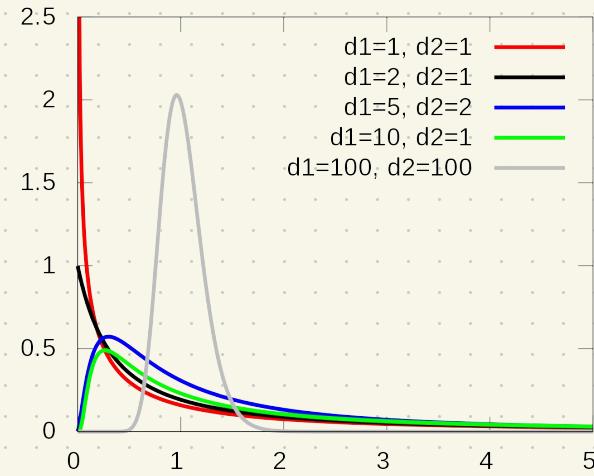
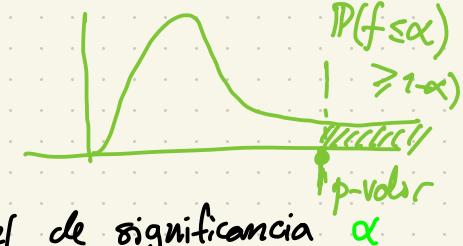
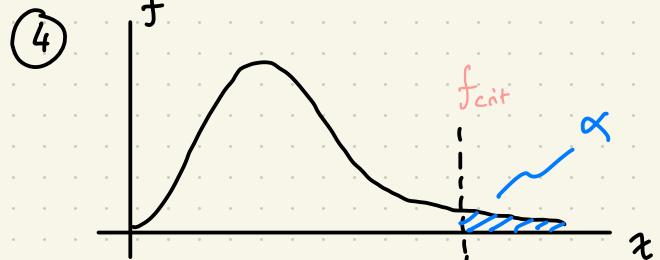
3. Encontramos f_{crit} (p-valor), definimos el nivel de significancia α

$$f \sim F_{gl_{ESS}, gl_{RSS}}$$

gl: grados de libertad

4. buscamos $Pr(\tilde{F} \geq f_{crit}) = \alpha$

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



Ahora queremos validar para un dato j si el coef. es bueno:

$$\left\{ \begin{array}{l} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{array} \right. , \text{ para un } j \text{ fijo}$$

Esto es comparar: $\left\{ \begin{array}{l} y = \beta_0 + \sum_{j=1}^{k-1} \beta_j x_j + \sum_{l=k+1}^p \beta_l x_l \\ y = \beta_0 + \beta_1 x_1 + \dots + \beta_j x_j + \dots \end{array} \right.$

$$RSS' = \sum_{i=1}^N (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \dots - \hat{\beta}_{j-1} x_{j-1} - \hat{\beta}_{j+1} x_{j+1} - \dots - \hat{\beta}_p x_p)$$

RSS con $\beta_j = 0$

$$F = \frac{(RSS - RSS')/1}{RSS/(n-p)} ; \text{ bajo } H_0 \sim f_{n-p} = (t_{n-p})^2$$

Vamos a calcular p-valor para un dcho α :

$$F \geq t_{\text{crit}} \text{ (p-valor)}$$

Con esto yo tengo la significancia de mi β_j , en un Summary clásico se ve lo siguiente

```
adv.lm=lm(sales~TV+radio+newspaper)
```

```
summary(adv.lm)
```

Call:

```
lm(formula = sales ~ TV + radio + newspaper)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.8277	-0.8908	0.2418	1.1893	2.8292

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.938889	0.311908	9.422	<2e-16 ***
TV	0.045765	0.001395	32.809	<2e-16 ***
radio	0.188530	0.008611	21.893	<2e-16 ***
newspaper	-0.001037	0.005871	-0.177	0.86

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared: 0.8972, Adjusted R-squared: 0.8956
F-statistic: 570.3 on 3 and 196 DF, p-value: < 2.2e-16

Erros estándar estimados
Para los parámetros β .

Estadísticos correspondientes a
los test individuales para cada
 β . Estos tests tienen distribución
 t_{196}

P-valores correspondientes a
los test individuales.

Este modelo plantea:

$$\begin{aligned} \text{sales} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{radio} \\ + \beta_3 \text{newspaper} \end{aligned}$$

Otras medidas a tener en cuenta

Mean Absolute Error

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

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2. \rightarrow \text{Mean Square Error}$$

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

↳ Root Mean Square Deviation

Bias-Variance Tradeoff

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

1. error de estimación.

2. sesgo.

$$\hat{\beta} - \beta = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \gamma - \beta$$

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

$$= (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \varepsilon$$

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

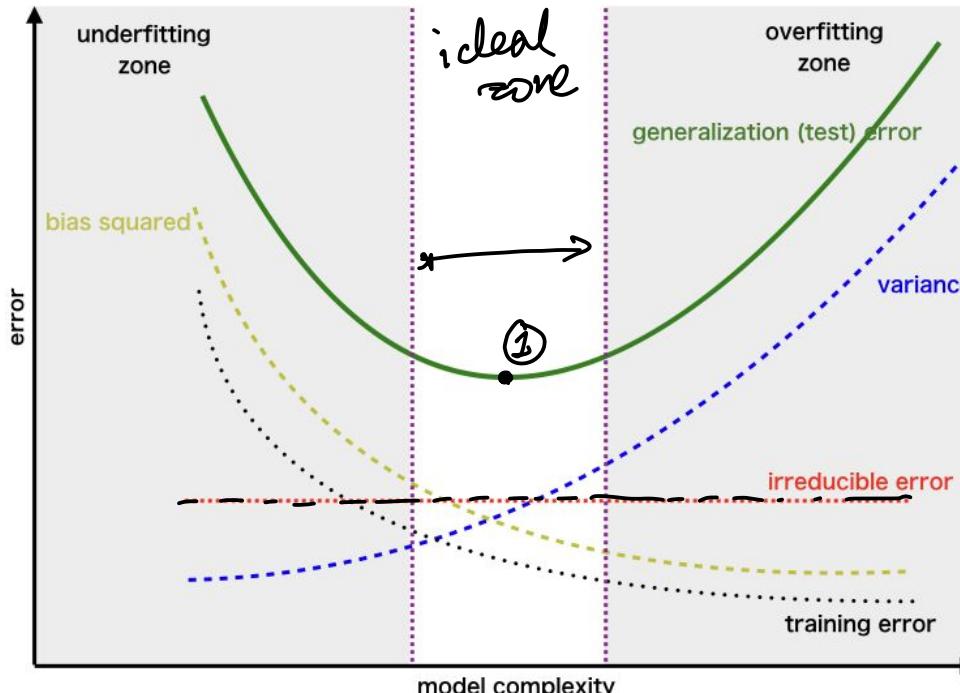
$$MSE = \underbrace{Bias(\hat{f})^2 + Var(\hat{f})}_{\text{Error de sesgo}} + \sigma_\epsilon^2$$

↑
Error de sesgo

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

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

Bias-Variance Tradeoff



② este punto se obtiene por optimización (early-stop) y la implementación depende del problema y del modelo.

cota cramer-raf
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:

+ ¿cuál es la validez del modelo?

+ ¿Cuáles son variables importantes?

En general planteamos:

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

Algunos casos de interés:

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

test de bondad de la regresión

Aquí:

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

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