Extracción y análisis de datos

- MANIPULACIÓN DATOS - PANDAS
  VISUALIZACIÓN
  - dataframes
  - drop
  - loc
  - seaborn

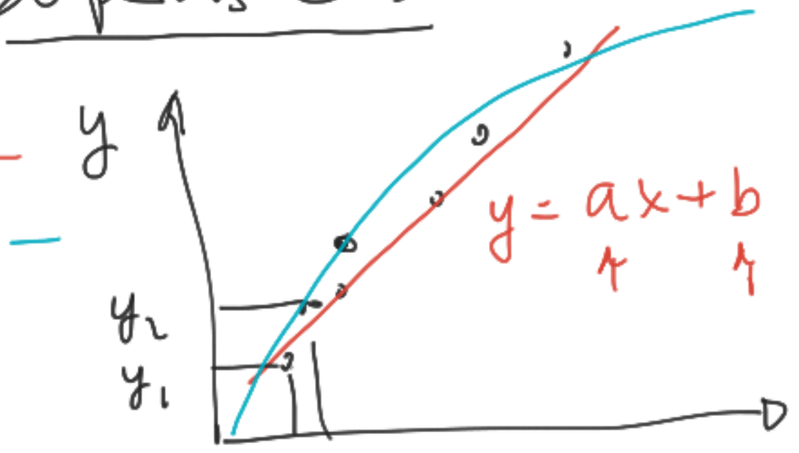
  missing data drop

- AGRUPAMIENTO - CLUSTERING
  - K-means
  - jerárquico
  - GMM
  - GMM + BIC → How many clusters?

- REDUCCIÓN DIMENSIONALIDAD —
  - PCA
  - Selección atributos
  - Proyección datos multi-dimensionales

TÉCNICAS DE REGRESIÓN — supervisadas

- datos variados | lineal — | no lineal —
- selección modelos
  (BIC - permanencia)
- Árboles de decisión



$\{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\}$

muestra

↓

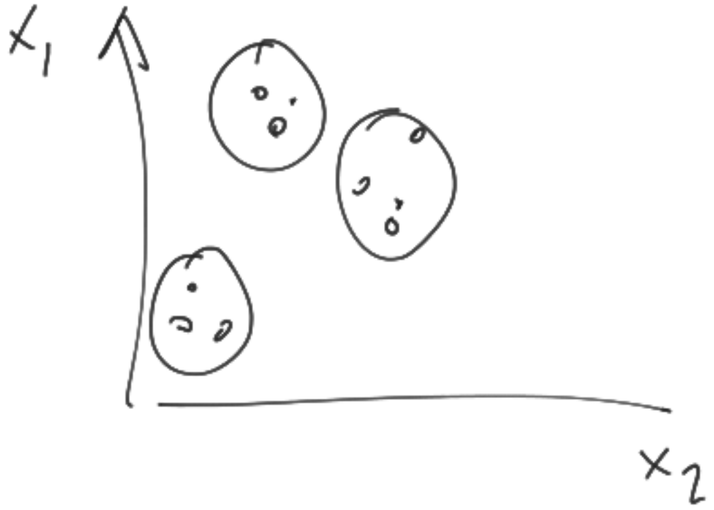
estimación estadística parámetros modelo

a, b

no - supervisadas

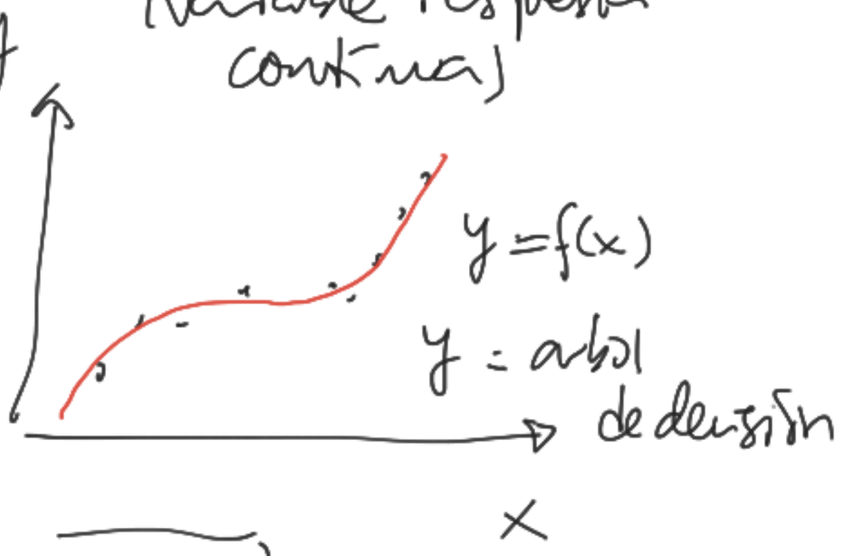
↓

CLUSTERING



supervisadas

Regresión
(variable respuesta continua)



$x \rightarrow$ [árbol de decisión] $\rightarrow \hat{y}$

Clasificación

$A_{N \times M}$

ω - etiqueta clase

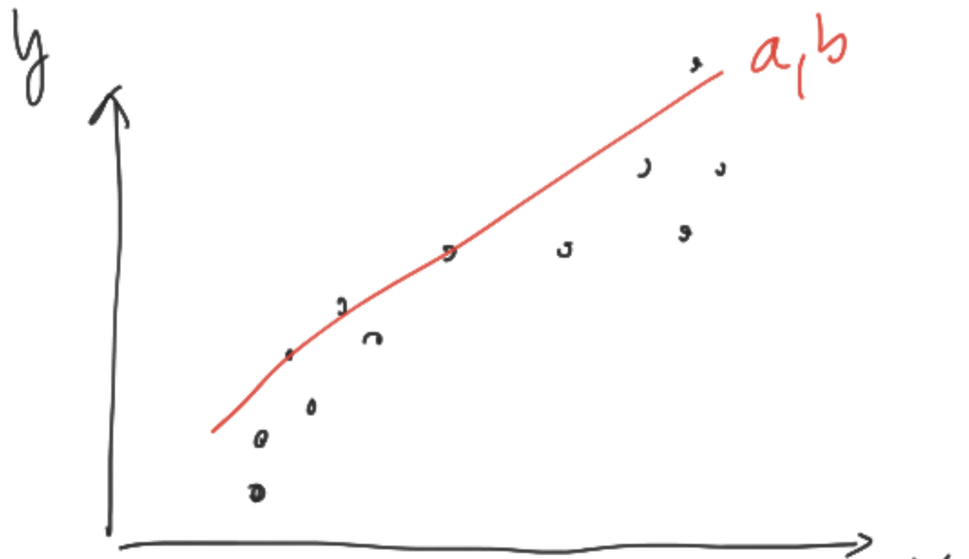
$\vec{x} \rightarrow [C] \rightarrow \hat{\omega}$

DATOS UNIVARIADOS

- estimar
- entrenar
- ajustar

y	x
1d	Nd

fit



DATOS MULTIVARIADOS

atributos

y	$x_1, \dots x_M$
~	
N	$A_{N \times M}$
respuestas	

datos $x_1, x_2, x_3 \dots x_N \longrightarrow \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$

$\tilde{x}_1 = x_1 - \hat{\mu}$

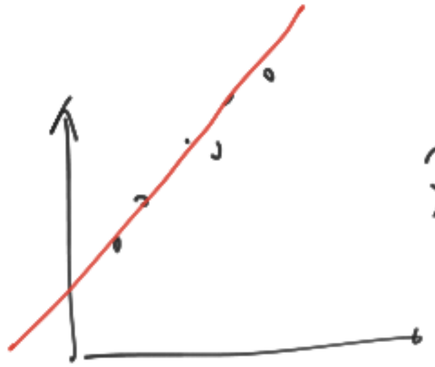
$\tilde{x}_2 = x_2 - \hat{\mu}$ $\longrightarrow$ $\{\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \dots, \tilde{x}_N\}$

$\vdots$

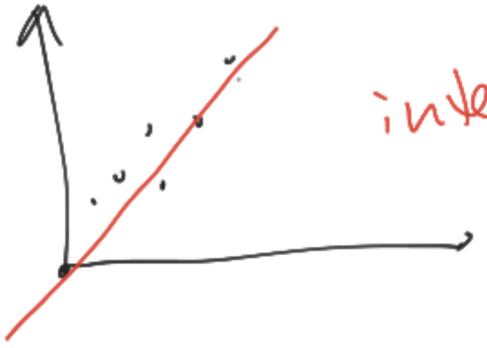
$\tilde{x}_N = x_N - \hat{\mu}$

datos centrados (a)

(media nula)

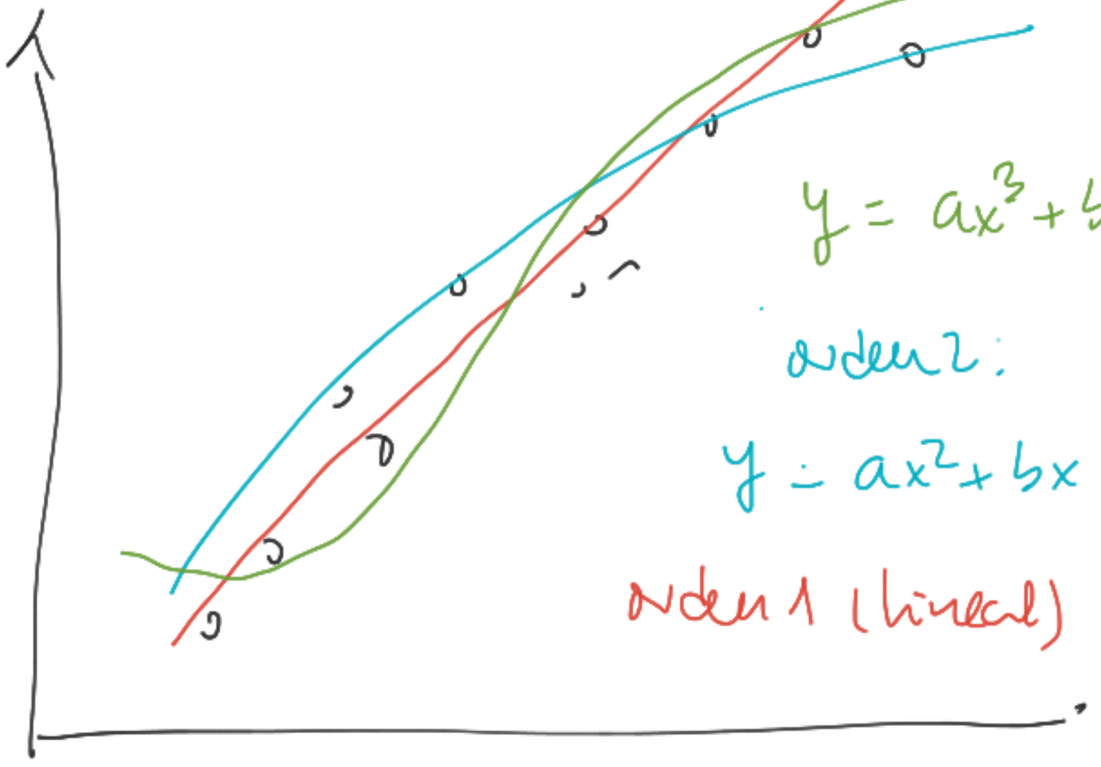


(a, b)



intercepción = 0 (b = 0)

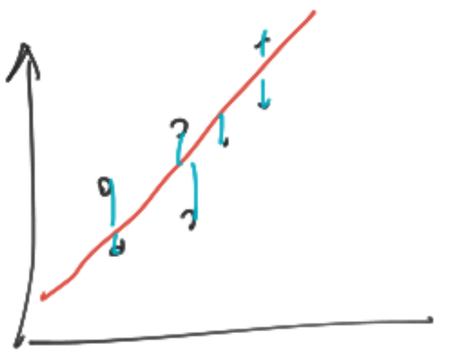
SELECCIÓN MODELOS REGRESIÓN → ÍNDICES PARSIMONIA



ERROR + COMPLEJIDAD

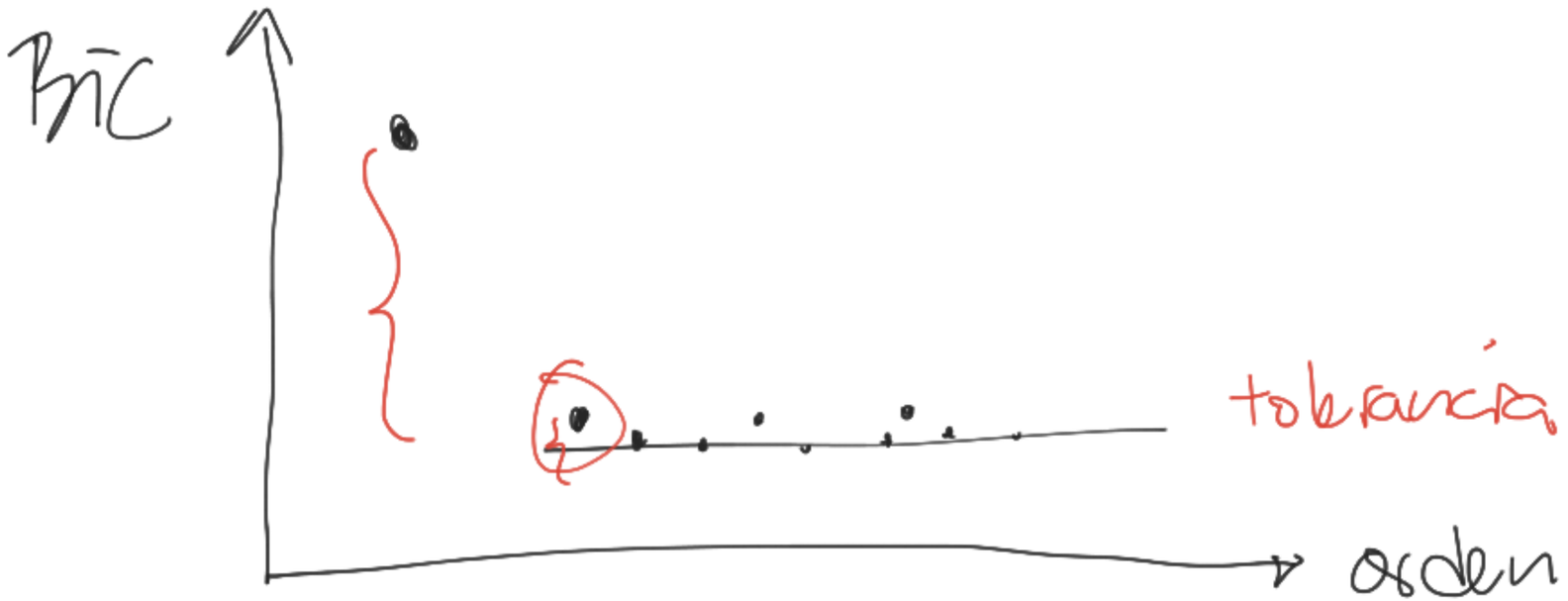
(#parámetros)

(a, b) → 2



$\sum \text{residuos}^2$

- AIC
- BIC


BIC
tolerancia
orden

$$\{x_1, x_2, \dots x_n\} \longrightarrow \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

error estándard estimación media $\hat{\mu}$

$$s.e.m = \frac{s}{\sqrt{n}}$$

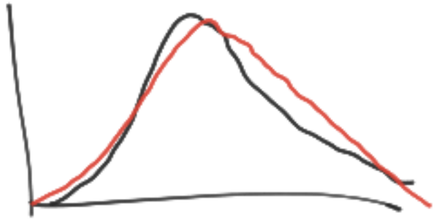
$\alpha = 0.05$ (95%)

Intérvalo Confianza

$$\left[\hat{\mu} - t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}} \; ; \; \hat{\mu} + t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}} \right]$$

tablas distribución t - Student

(~Gaussiana con tamaño n finito)



$y = ax + b$

$\{(x_1, y_1), (x_2, y_2) \dots, (x_n, y_n)\}$

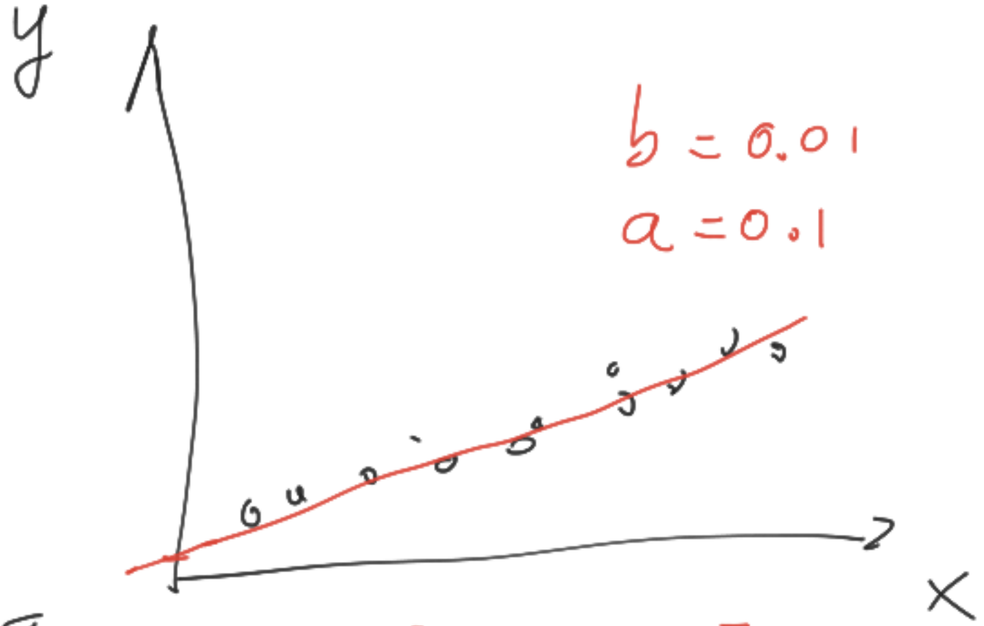
$\hat{a}$ → intervalo contanza 95%

$\hat{b}$ → $[-0.02, 0.02]$

$\times$

$[-0.008, 0.13]$ (!!)

no se puede descater que no haya crecimiento de y con x ←



$[0.05, 0.15]$

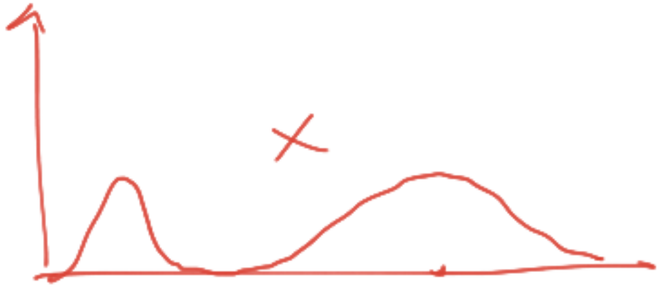
$a > 0$ es estadísticamente significativo

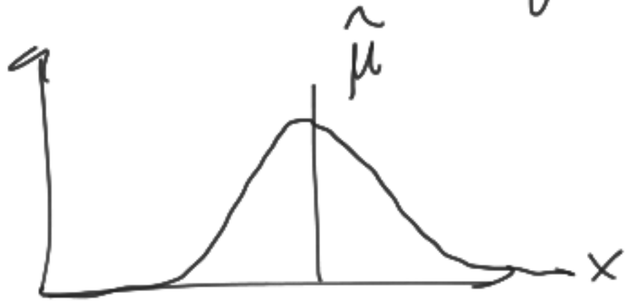
↓

y crece con x

→ heteroscedástico x

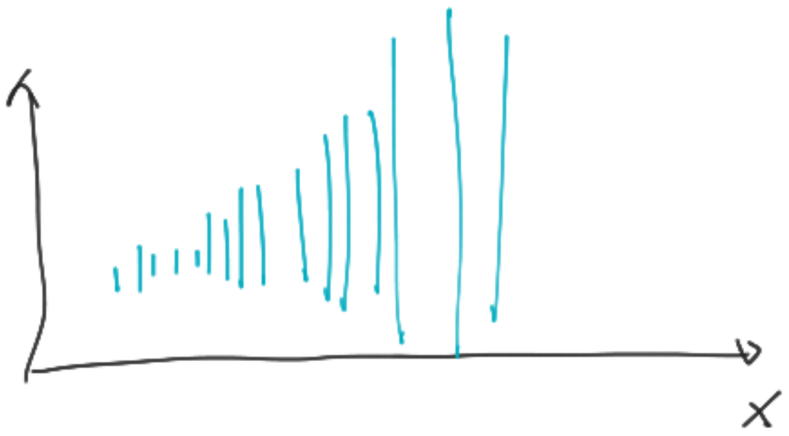
distribución normal - gaussiana

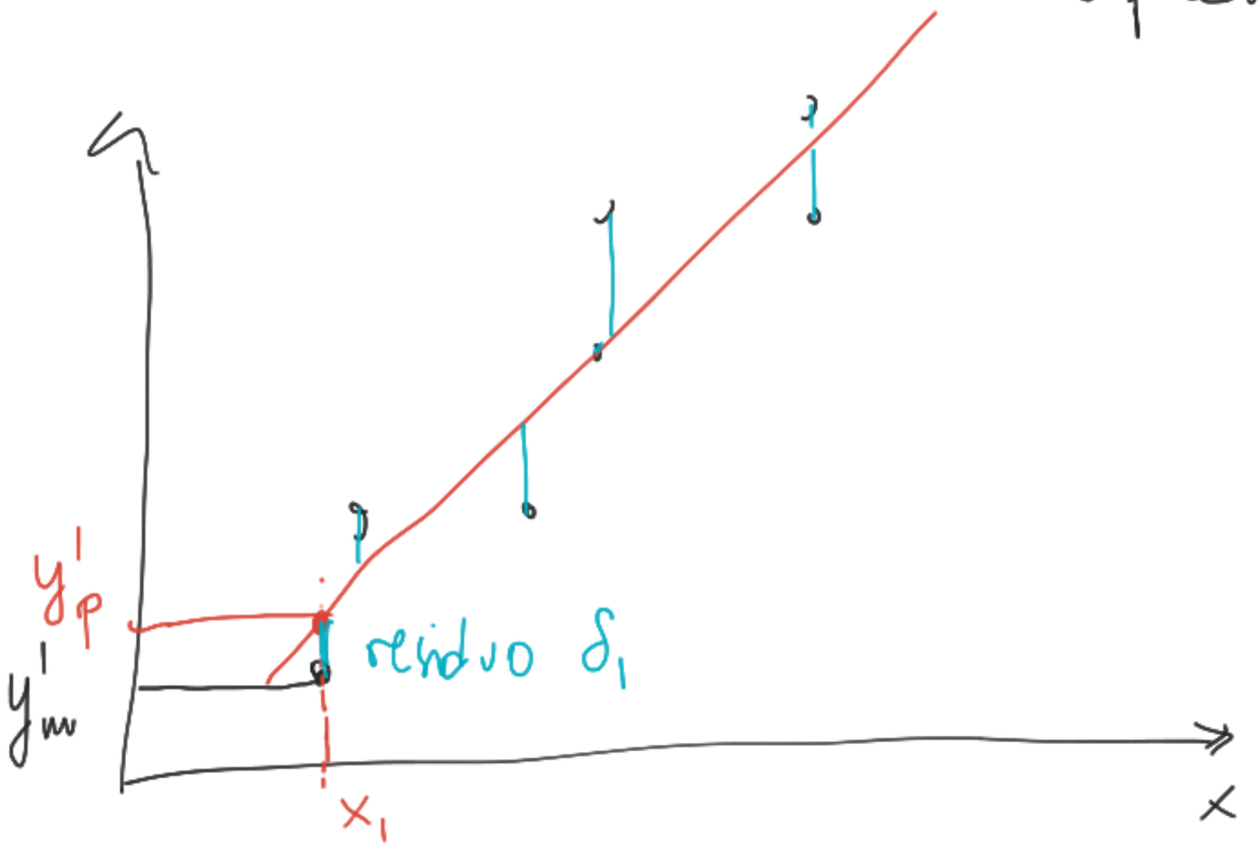




Distribución Residuos

$y_{predicción} - y_{muestra} = \delta$



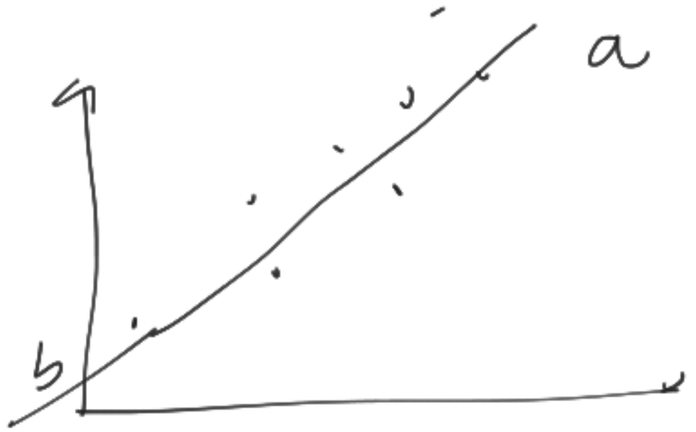


y'_p = model.predict (x_1)

add_column

$$y = (x\ 1) \cdot \underbrace{\begin{pmatrix} a \\ b \end{pmatrix}}_{\beta} + \varepsilon = ax + b + \varepsilon$$

$1 \times 2 \quad 2 \times 1$



Residuos distribución normal

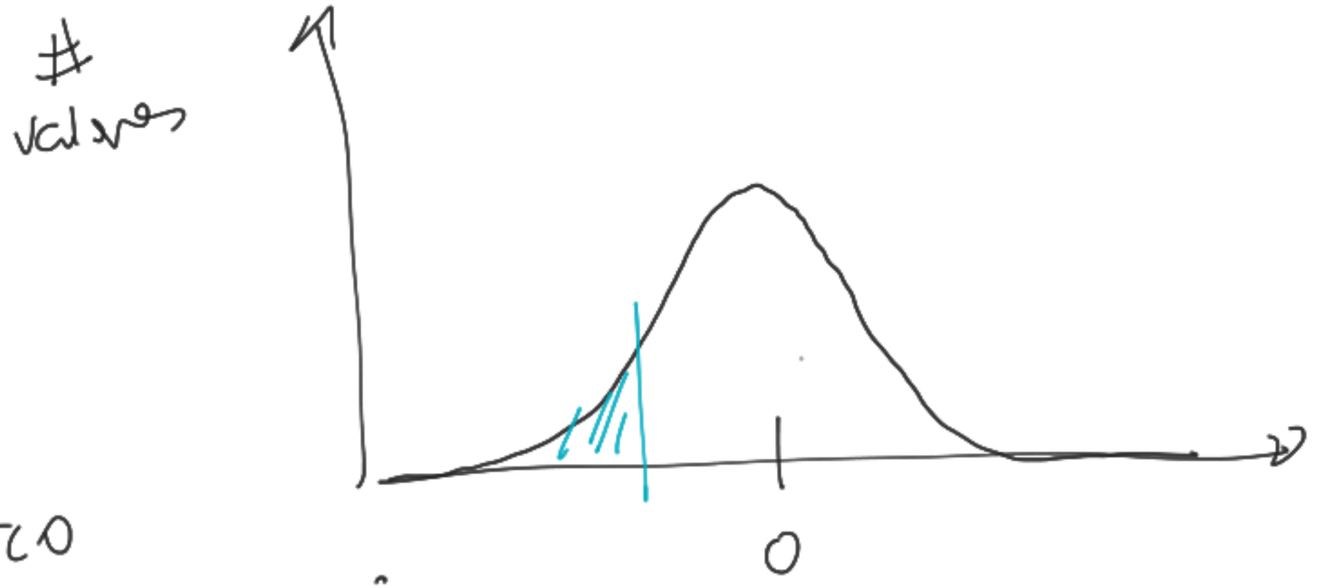
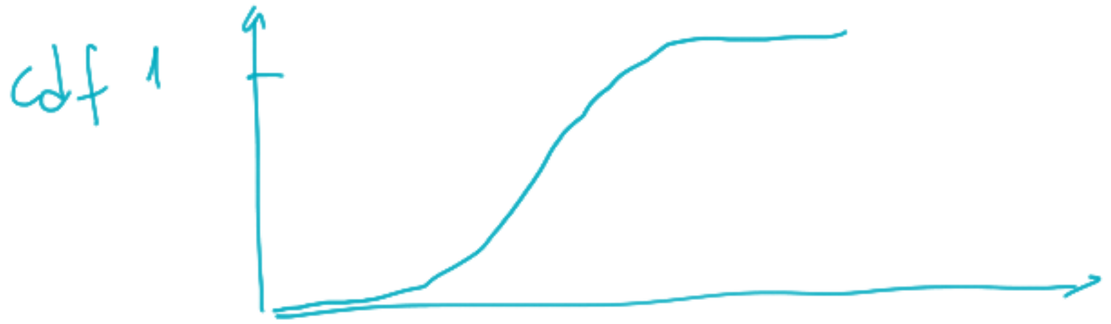
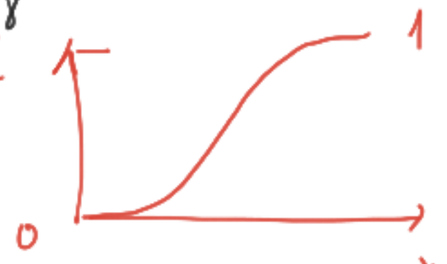


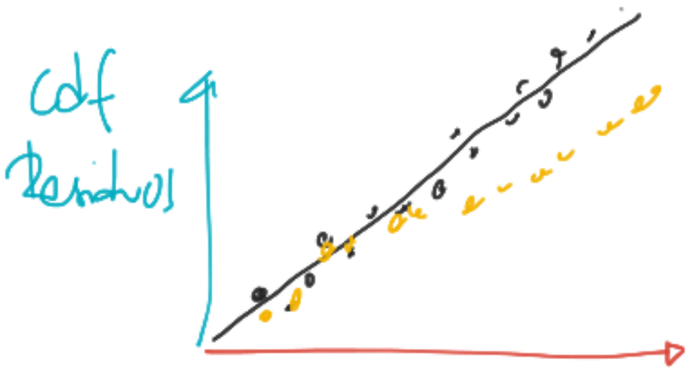
gráfico cuantil - cuantil (q-q plot)





residuos

$$\delta_i = y_i - \hat{y}_i$$



• datos normales

• datos no normales

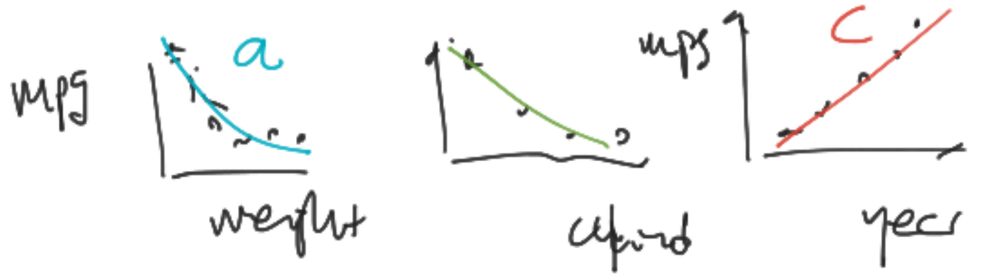
univariada $y = f(x)$

multivariada $y = f(x_1, x_2, x_3 \dots)$

ex: mpg cars →

mpg = f(weight, cylinders, year)
model

pairplot



mpg = (a)·weight + (b)·cylinders + (c)·year + d

a<0 b<0 c>0

Regresión con árboles de decisión (Regression trees)

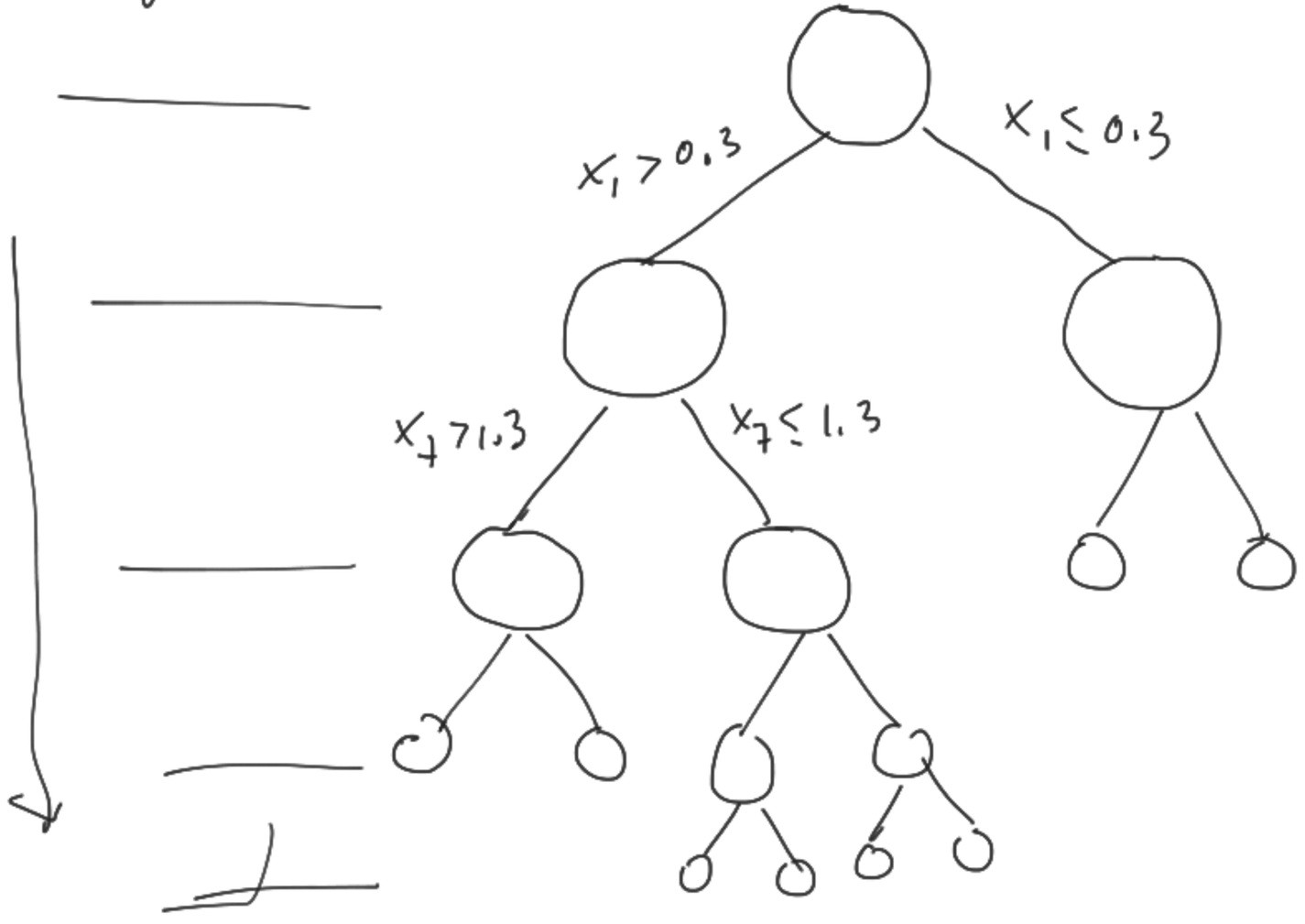
datos x_1 x_2 ... x_n y

obs 1

⋮

obs n

tree depth



Decision tree regressor

mpg ~ weight

Ejercicio:

- load multivariate data
- clean data → NaNs, variables tipo string... (M) → (d)
- Dimensionality Reduction (PCA)
  (90-95% varianza explicada)
- Proyectar datos espacio PCA reducido
- Clustering (k-means, hierarchical, GMM)
- Visualizar resultado