Analysis of Variance (ANOVA)

Introduction

- Generalization for n samples (arbitrary number).
- ANOVA tests if 'at least one' of samples is significantly different ~~than~~ from the others.
- Some examples:
  - Testing 3 different designs of a website, sample for every design has its own: $\bar{X}, s, n$

web version 1	web version 2	web version 3
$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$
s_1	s_2	s_3
n_1	n_2	n_3

  - Testing 5 different products in olive trees.

- $H_0 \Rightarrow \mu_1 = \mu_2 = \dots = \mu_k$; $H_1 \Rightarrow$ At least one is significantly different, don't know which one in advance.

- Hand-made example:

one way anova (one 'dimension')

Sample 1	Sample 2	Sample 3
3	5	5
2	3	6
1	4	7

$n = 3$

$m = 3$

① Compute global mean

$$\bar{\bar{X}} = \frac{1+2+3+3+4+5+5+6+7}{9} = 4$$

$$d.f = 9 - 1 = 8$$

② Compute SST

$$SST = (3-\overset{\bar{\bar{X}}}{4})^2 + (2-4)^2 + (1-4)^2 + (5-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2 + (7-4)^2 = 30$$

③ Compute sample means

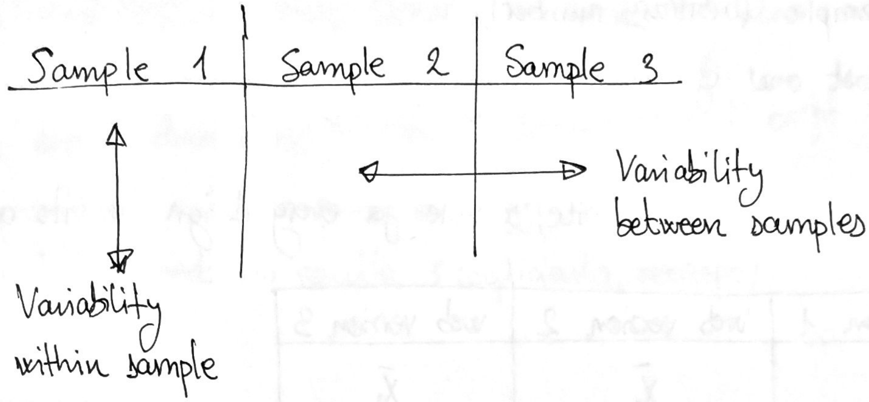
$$\bar{X}_1 = \frac{1+2+3}{3} = 2 \quad \bar{X}_2 = \frac{3+4+5}{3} = 4 \quad \bar{X}_3 = \frac{5+6+7}{3} = 6$$

④ Compute SSW (Sum of Squares Within Samples)

$$SSW = (3-\overset{\bar{X}}{2})^2 + (2-2)^2 + (1-2)^2 + (5-4)^2 + (3-4)^2 + (4-4)^2 + (5-6)^2 + (6-6)^2 + (7-6)^2 = 6$$

$$\begin{matrix} m \cdot n & & m \\ = N & - & k \end{matrix}$$

$$d.f = \underset{\text{diff. samples}}{m} \cdot (\underset{\text{sample mean}}{n-1}) = 3 \cdot (3-1) = 6$$



⑤ Compute SSB (Sum of Squares Between Samples)

$$SSB = \left[(\bar{X}_1 - \bar{\bar{X}})^2 \cdot \underset{\overset{\shortparallel}{3}}{n_1} + (\bar{X}_2 - \bar{\bar{X}})^2 \cdot \underset{\overset{\shortparallel}{3}}{n_2} + (\bar{X}_3 - \bar{\bar{X}})^2 \cdot \underset{\overset{\shortparallel}{3}}{n_3}\right] \cdot \cancel{n} = (2-4)^2 \cdot 3 + (4-4)^2 \cdot 3 + (6-4)^2 \cdot 3 = 24$$

$$d.f = m - 1 = 2$$

⑥ Compute F statistic (F distribution) Fischer - Snedecor

$$MSW = \frac{SSW}{d.f_{ssw}} = \frac{6}{6} = 1$$

$$MSB = \frac{SSB}{d.f_{SSB}} = \frac{24}{2} = 12$$

$$F = \frac{MSB}{MSW}$$

$$F \sim \frac{X_i^2}{X_i^2} = \frac{12}{1} = 12$$

$p = 0.008$ $\alpha = 0.05$

⑦ Perform test $p < \alpha \rightarrow$ Reject null hypothesis!!

Linear Regression

Introduction

- One of the simplest 'Machine learning' algorithms.
- Very good for modeling <u>linear relationships</u> in data. There is no free lunch in Data Science.

It has some advantages:

- Simple and not too prone to overfit. explain overfit a little
- Easy to interpret its results. (coefficients, intercept)
- Analytical solution

And drawbacks:

- Too simple to capture non-linear complex relationships out of the box.
- Analytical solution not suitable for big data problems.

$$\underbrace{y}_{\text{dependent variable}} = \overbrace{\beta_0}^{\text{intercept}} + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_n \cdot X_n$$

independent variables / features: $X_1, X_2, \ldots, X_n$

coefficients: $\beta_1, \beta_2, \ldots, \beta_n$

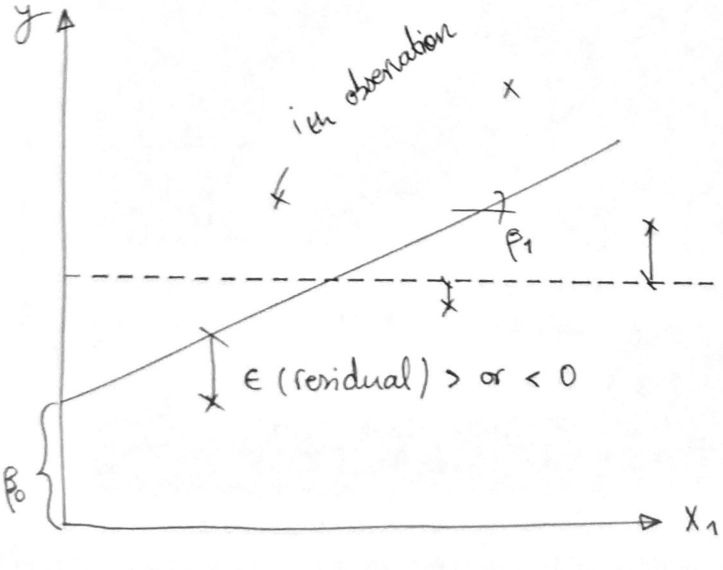
- Some examples
  - \+ salary vs years of experience
  - \+ price of an apartment vs m^2

$$y = \beta \cdot X \rightarrow \beta = (X^T \cdot X)^{-1} \cdot X^T \cdot y$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} 1 & \text{—— } X_1 \text{ ——} \\ 1 & X_{1_s} \cdots X \cdots \\ 1 & \text{—— } X_m \text{ ——} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_n \end{bmatrix}$$

For intercept

Explanation (Single feature version)



$$y_i = \beta_0 + \beta_1 \cdot X_1^i + \epsilon^i$$

① Compute residuals "ϵ^i" for every point.

② Compute Sum of Squared Error:

$$SSE = (\epsilon^1)^2 + (\epsilon^2)^2 + \cdots + (\epsilon^m)^2$$

eq. SSW in ANOVA! | ¿ Problems? Units^2
Grow with m

③ Compute RMSE: How good is my model?

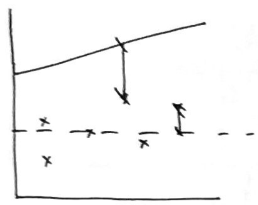
SSE is hard to interpret and growing with samples! Let's normalize:

$$RMSE = \sqrt{SSE / N}$$

④ How to compare models?

$$\text{Compute } SST = \sum_{i}^{m} (X_1^i - \bar{X}_1)^2 \quad \text{and} \quad R^2 = 1 - \frac{SSE}{SST}$$

¿ Can be <0? Yes! If model performs worse than just taking the mean!



SSE >> SST $\quad R^2 < 0$

Features significance

$$t \text{ statistic} = \frac{coef}{\underbrace{std.error}_{\frac{s}{\sqrt{n}}}}$$

↑↑ (want it high or low, extreme)

p (probability of getting such an extreme result by chance) ↓↓

Colinearity can exist, compute Pearson's R and remove correlated features.

[0, ±1]