IMMUNE
CODING INSTITUTE

Supervised Machine Learning

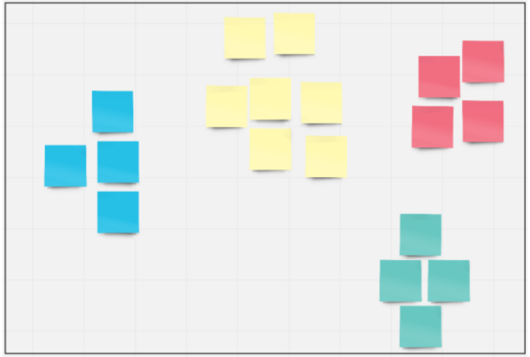
Module 3

What did we learn from module 3.1?

What did we learn from module 3.1?

4 min

- Each on your own
- Populate with ideas, concepts, examples…



6 min

- All together
- Cluster similar ideas, enrich board, prepare to tell your story

IMMUNE CODING INSTITUTE

What did we learn from module 3.1?

- You will be working in **teams**:
  - **TEAM 1 (Mónica):**
    - Daniel Rey
    - Laura Martín
    - Samuel Carballo
    - Mauricio Asperti
    - Marcelo Araujo
    - Isabel Hita
  - **TEAM 2 (Juan):**
    - Marcos García
    - Ignacio Cifuentes
    - María Dolores Carmena
    - Fernando Rodríguez
    - Ayose Sosa Guerra
  - **TEAM 3 (Miguel):**
    - Vittoria Reale
    - Rubén Farias
    - José Pascual
    - Ángel Moya
    - Kay Kozaronek
    - Miguel García

facilitator -> timing, everybody speaks, go,go,go!

presenter -> summarizes results

Module 3 Summary

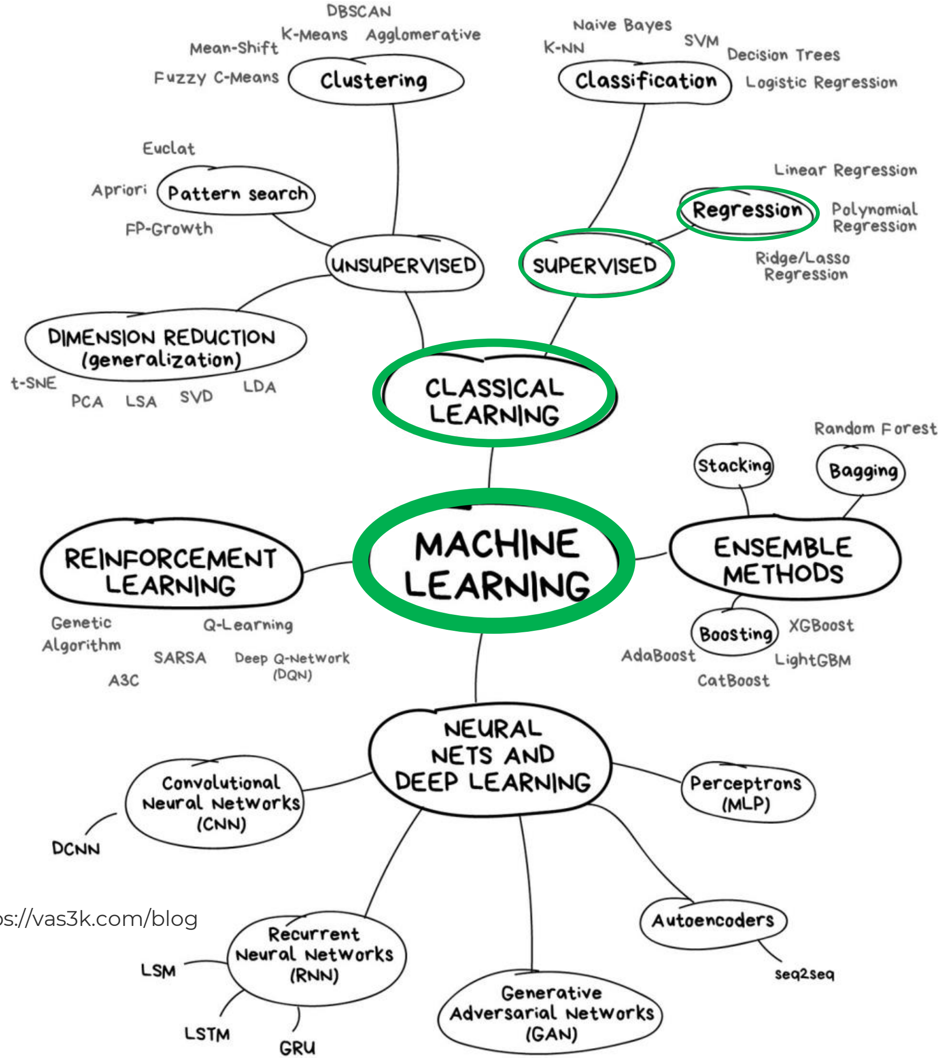
SESSION	TITLE	TEACHER
1	ML Foundations	Juan
2	**Regression Introduction and Practice**	**Juan**
3	Classification Introduction and Practice	Carlos
4	Feature Engineering and Selection for ML	Carlos
5	Advanced Supervised Models 1	Carlos
6	Advanced Supervised Models 2	Carlos
7	Hands-on Practice	Carlos

IMMUNE
CODING INSTITUTE

Supervised ML

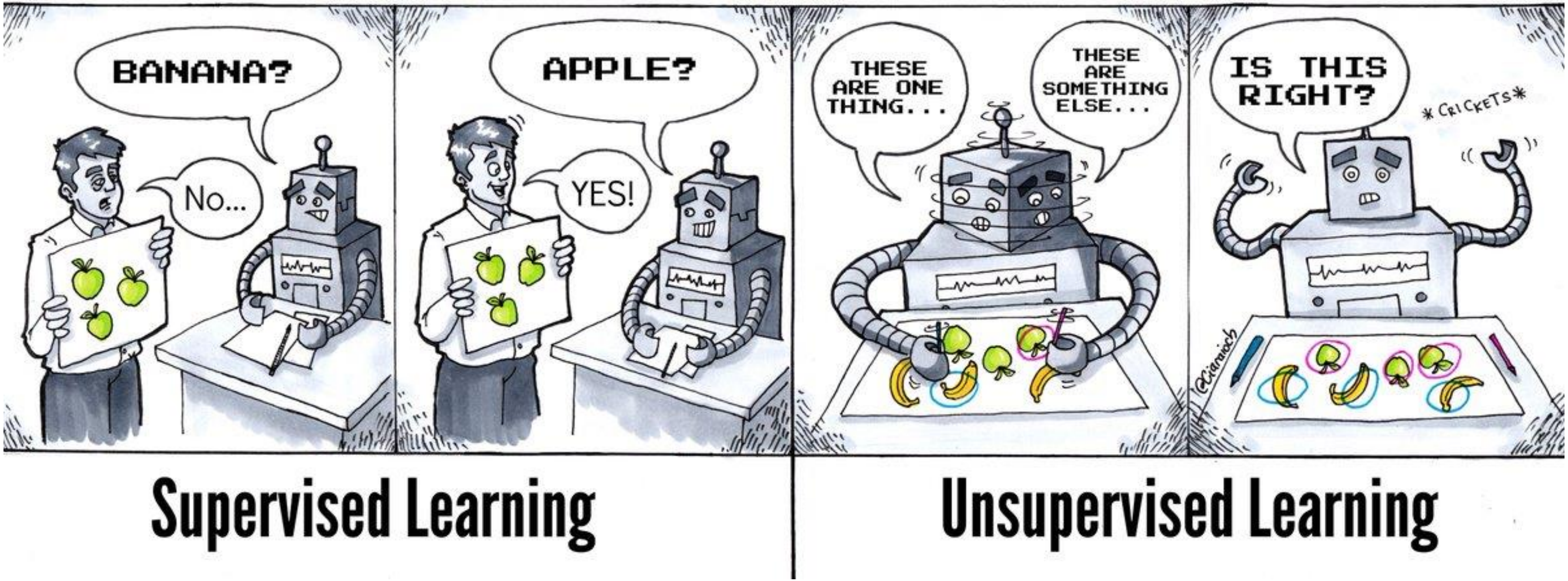
Module 3.2

Types of machine learning 🧐



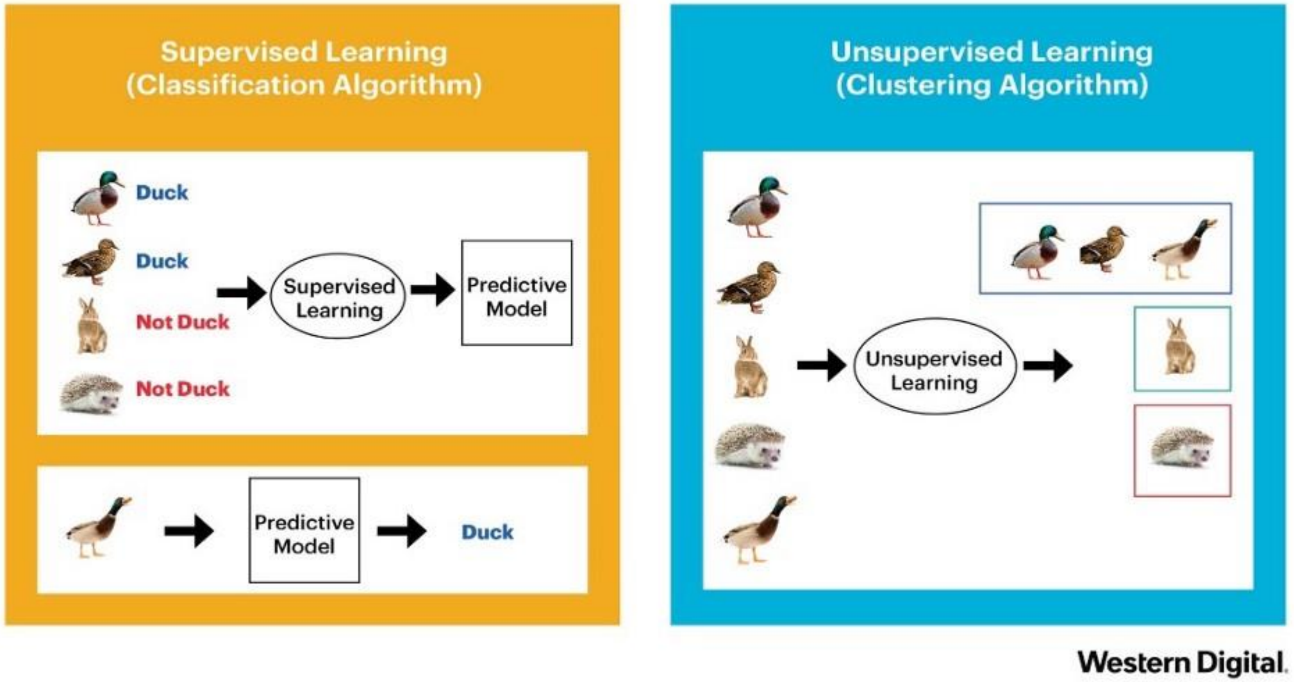
https://vas3k.com/blog

Supervised vs. Unsupervised learning

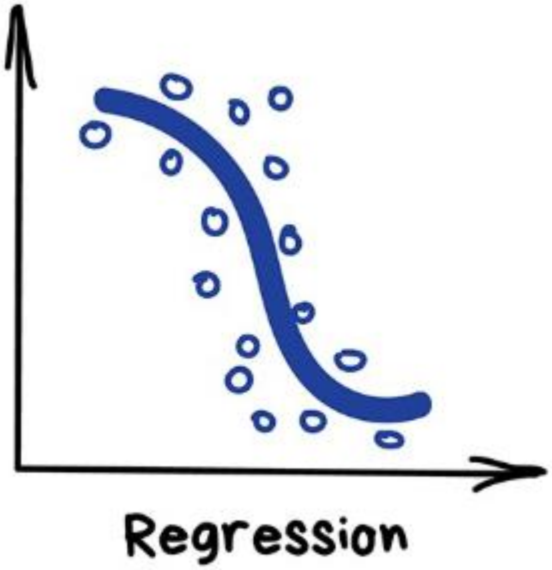


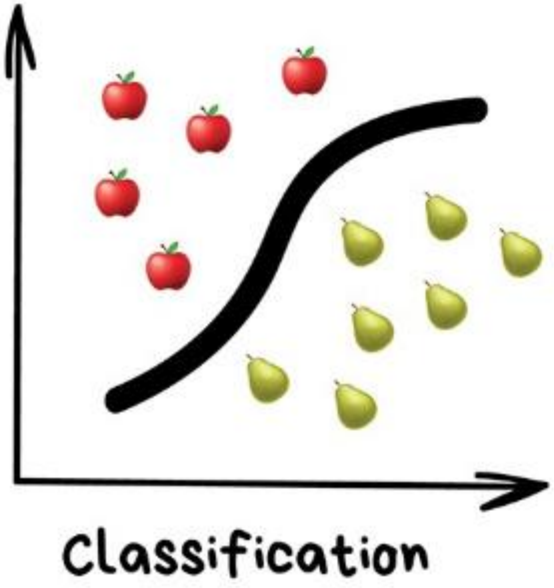
https://pbs.twimg.com/media/DsCTvc3XQAE7Njb.jpg

Supervised vs. Unsupervised learning



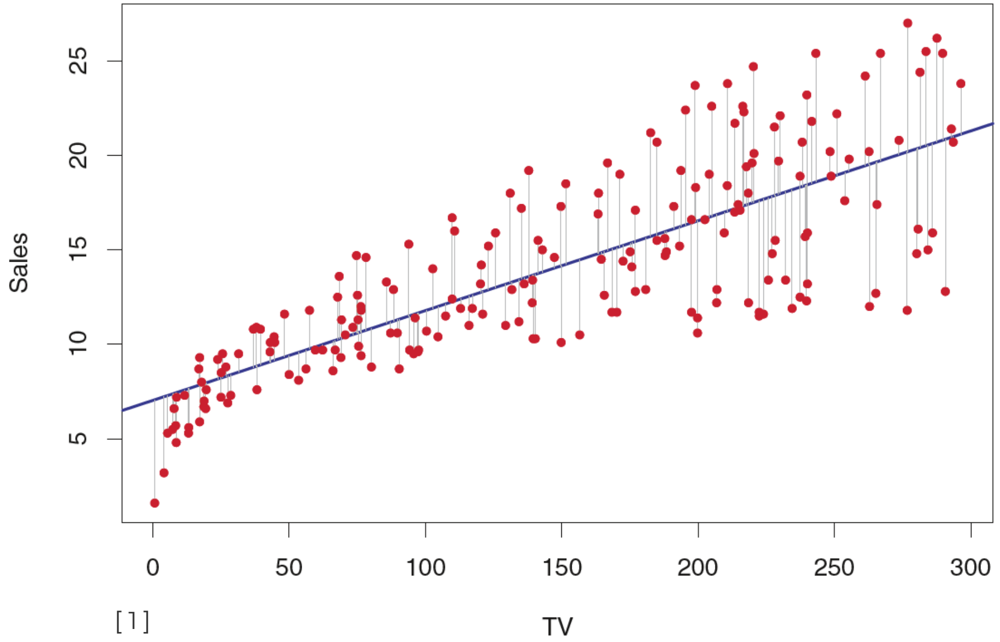
Regression vs. Classification





https://vas3k.com/blog

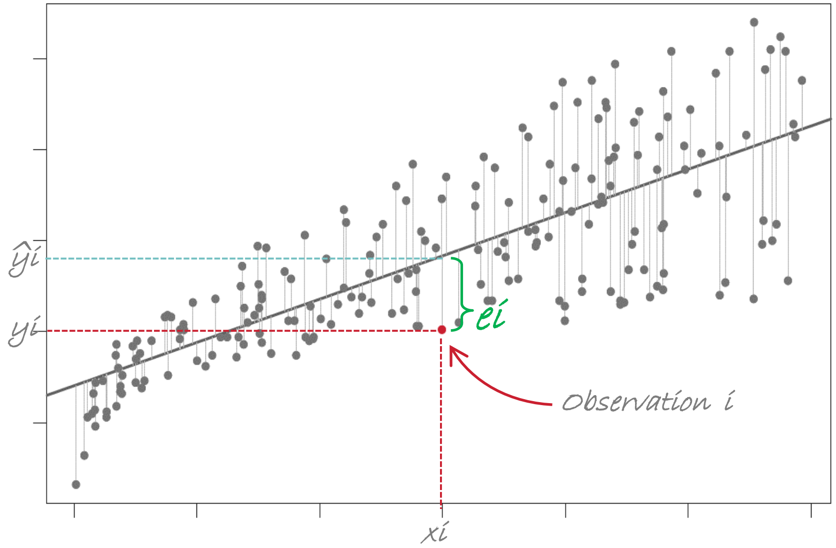
Simple Linear Regression



[1]

- Quantitative predictions
- Single input variable
- $Y \approx \beta_0 + \beta_1 X$
- β_0, β_1:
  - Constant and unknown
  - Coefficients/parameters
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- $\hat{\beta}_0, \hat{\beta}_1$:
  - Calculated from training data
  - Reduce closeness

Estimate Coefficients



[1]

LEAST SQUARES

$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

ith residual **>** $e_i = y_i - \hat{y}_i$

Residual Sum of Squares:

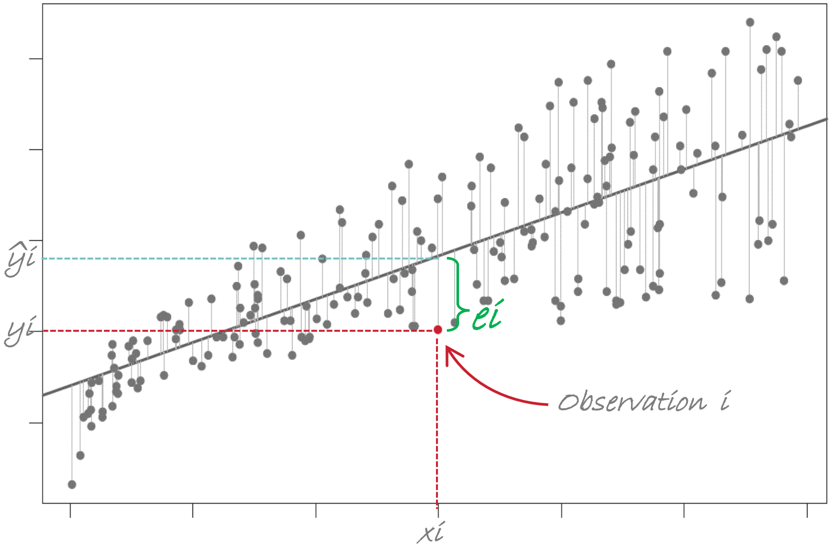
$$\begin{aligned} \mathrm{RSS} &= e_1^2 + e_2^2 + \cdots + e_n^2 \\ &= \left(y_1 - \hat{\beta}_0 + \hat{\beta}_1 x_1\right)^2 + \cdots + \left(y_n - \hat{\beta}_0 + \hat{\beta}_1 x_n\right)^2 \end{aligned}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

IMMUNE CODING INSTITUTE

Estimate Coefficients



[1]

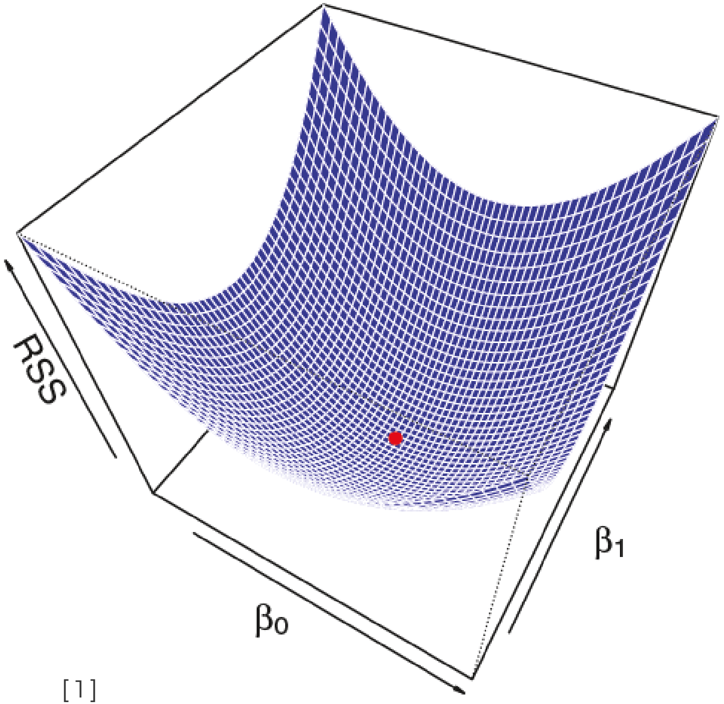
LEAST SQUARES MATRIX APPROACH

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$$

$$\mathrm{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = \left(\mathrm{X}^{\top}\mathrm{X}\right)^{-1}\mathrm{X}^{\top}\mathbf{y}$$

Estimate Coefficients



[1]

- This diagram shows how different values for each regression coefficient determine RSS value.
- We can see how there is a single solution for the global minimum of the loss function

SciKit-Learn

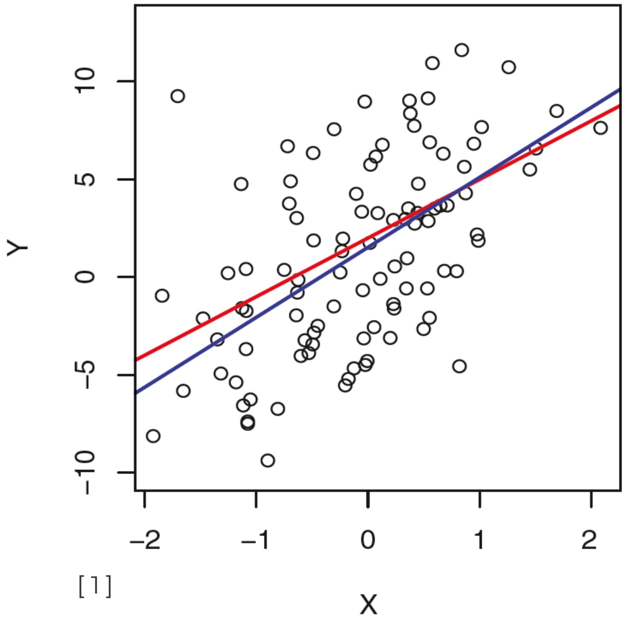
- Free machine learning library for Python
- Started in 2007, first public release in 2010
- Community driven project, however institutional and private grants help to assure its sustainability
- Used for data modeling, not loading, manipulating, summarizing...
- Focuse on usability, medium scale projects
- Who uses SciKit-Learn?

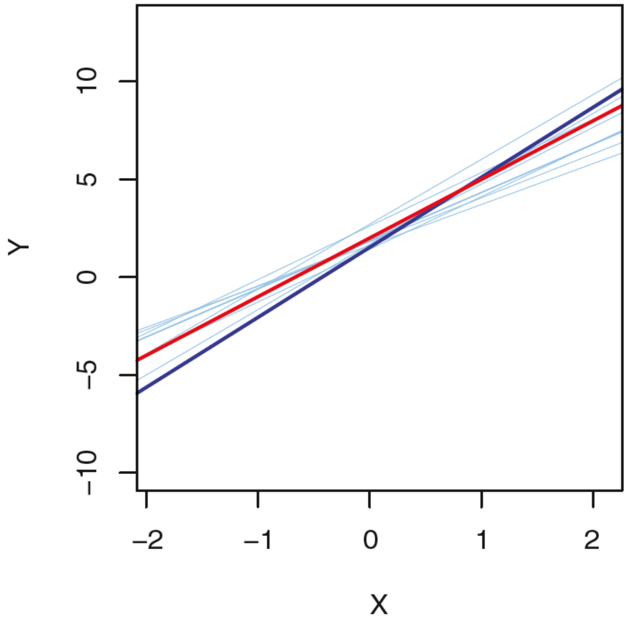
https://scikit-learn.org/

IMMUNE
CODING INSTITUTE

Accuracy Assesment

$$Y = 2 + 3X + \epsilon$$





[1]

LEFT: Population regression line vs. one ramdom sample least squares line

RIGHT: Population regression line vs. 10 random samples least squares lines.

Population regression line

Least squares line

IMMUNE CODING INSTITUTE

Example: Mean Accuracy Assesment 🧐

- Sample mean is equal to:

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- To find the standard error of the sample mean, we find its variance first:

$$Var(\hat{\mu}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(y_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

- To find the **standard error** of the sample mean, we find its variance first:

$$SE(\hat{\mu}) = \sqrt{Var(\hat{\mu})} = \frac{\sigma}{\sqrt{n}}$$

Coefficient Accuracy Assesment

Same approach for least squares coefficients:

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right]$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

...where: $\sigma^2 = \mathrm{Var}(\epsilon)$

We can estimate σ^2 from data:

$$\sigma = RSE = \sqrt{\frac{RSS}{n-2}}$$

(residual standard error)

Coefficient Accuracy Assesment

- If $\beta_1 = 0 \Rightarrow Y = \beta_0 + \epsilon$ and therefore there is no relationship between Y and X

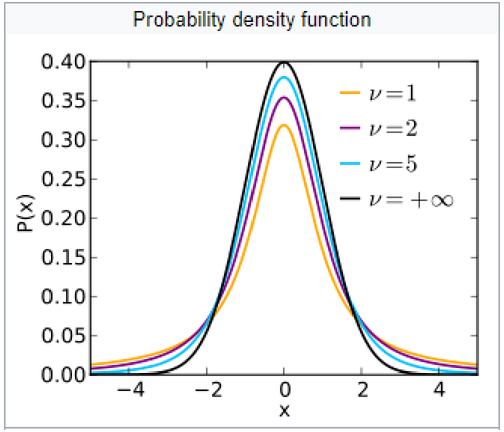
- Test **null hypothesis** of:
  H_0: There is no relationship between X and Y -> β_1=0
  H_a: There is some relationship between X and Y -> $\beta_1 \neq 0$

- t-statistic: $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$

- If β_1=0, t will have a **t-distribution** with n−2 **degrees of freedom**

- Probability of observing any number equal to |t| or larger in absolute value -> **p-value**

Student's t

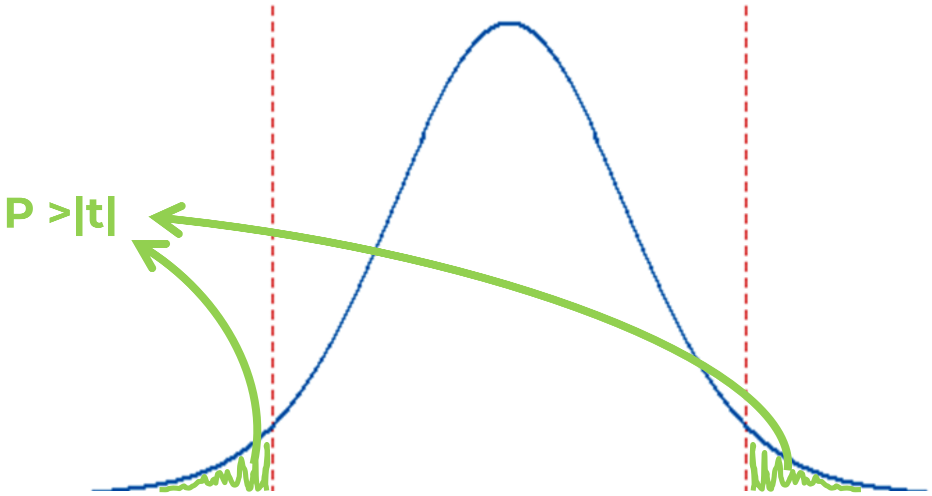


https://en.wikipedia.org/wiki/Student%27s_t-distribution

Coefficient Accuracy Assesment

The lower the probability (p-value), the higher the evidence against the null hypothesis

Typical p-value cutoffs: 5%-1%



Student distribution

IMMUNE
CODING INSTITUTE

Model Accuracy Assesment

RSE (residual standard error)

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- Estimate of the standard deviation of ϵ
- Average amount the response will deviate from true regression line
- Measured in response units
- Lack of fit of the model

R2

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

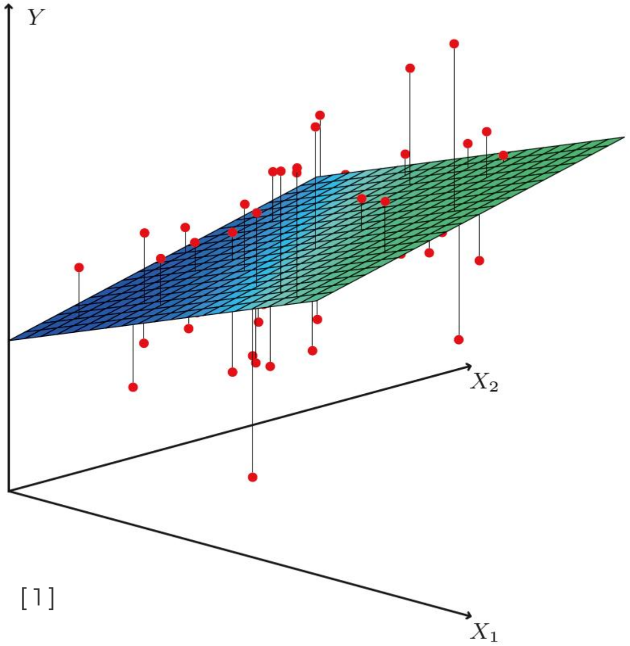
$$\mathrm{TSS} = \sum(y_i - \bar{y})^2$$

- **TSS >** (total sum of squares) total variance in the response Y
- **RSS >** amount of variability that is left unexplained after performing the regression
- **R2 >** proportion of variability in Y that can be explained using X

Multiple Linear Regression

- Linear regression extensión
- Multiple input variables:
- $Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$
- β_j **>** average effect on Y of a one unit increase in X_j
- Least squares coefficient estimation:

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$$



[1]

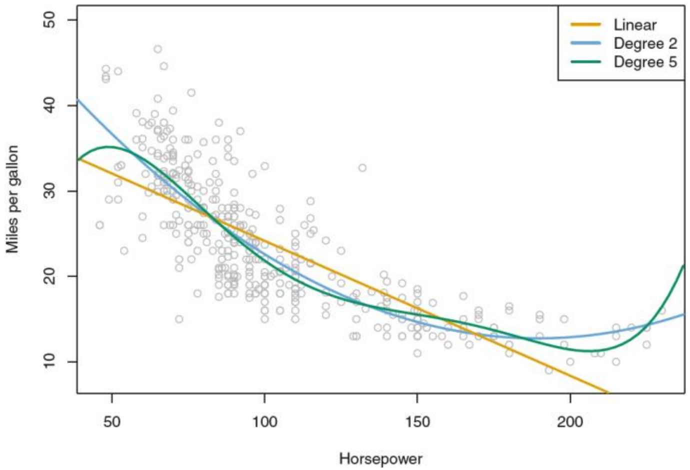
IMMUNE CODING INSTITUTE

Regression Extensions

- Extensions of the linear model:
  - Removing the additive assumption:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

  - Removing the linear assumption -> **polynomial regression**



$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i$$

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

References

[1] G. James, D. Witten, T. Hastie, R. Tibshirani. An Introduction to Statistical Learning with Applications in R. Springer, 2017.

[2] T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical: Data Mining, Inference and Prediction. Springer, 2009.

IMMUNE

CODING INSTITUTE