early makers
em lyon business school

Machine Learning

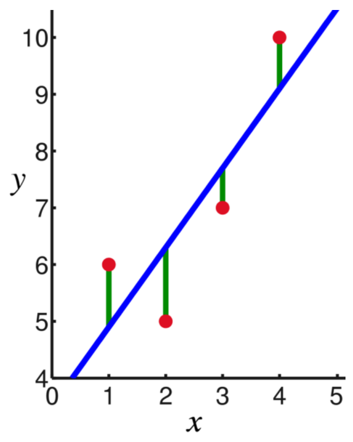
2. Linear regression

Nicolas Gartner

Part 1: Simple linear regression

Regression problems (recall from last course)

Linear regression (or curve fitting)



In linear regression, the observations (**red**) are assumed to be the result of random deviations (**green**) from an underlying relationship (**blue**) between a dependent variable (*y*) and an independent variable (*x*).

Examples of methods:

- **Least square algorithms:** a method where the sum of the squares of the residuals made in the results of every single equation is minimized.
- **Bayesian linear regression:** an approach to linear regression in which the statistical analysis is undertaken within the context of Bayesian inference

$$Posterior\ distribution = \frac{prior\ distribution\ \times likelihood}{model\ evidence}$$

Ex. : maximum likelihood or maximum a posteriori estimation

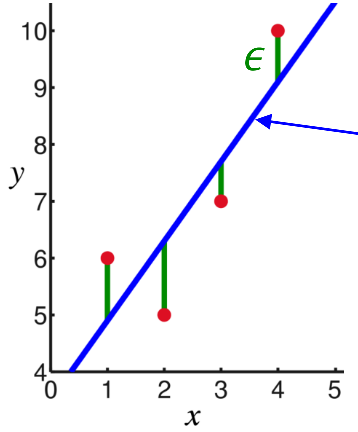
This can also be made with multiple variables (and input x would become a vector)

$prior\ distribution$: initial set of parameters (things that you want to learn)
likelihood: similarity of the considered sample, from which you want to compute something, to the prior samples (to the prior distribution) considered, from which your algorithm has learned.
$model\ evidence$: represents how well it seems that the model is correct
posterior distribution: the new set of parameters

Linear regression principia

Fixed equation from that relates X (input) to y (output) with the acceptance of ϵ (an irreducible error)

$$y = f(X) + \epsilon$$



The function that relates x to y could be:

- $f(X) = \alpha X + \beta$
- $f(X) = \alpha X + \theta X^2 + \beta$
- $f(X) = \alpha e^{-i\omega X} + \beta$

For the rest of the presentation, we will only focus on that simple case

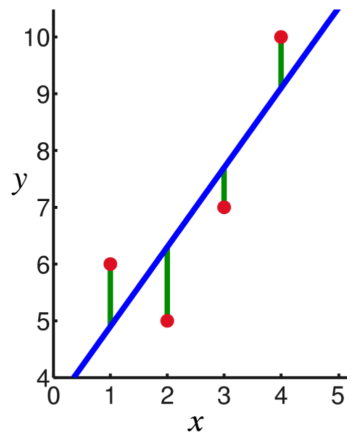
Linear regression objective is to estimate the best matching $f(X)$ which is called estimate, written $\hat{y} = \widehat{f(X)}$

So the objective is to minimize the residual error $e = y - \hat{y}$ for every sample X

Don't confuse e with ϵ:

$$\underline{e} = (\alpha X + \beta) + \epsilon - (\hat{\alpha} X + \hat{\beta}) = (\alpha - \hat{\alpha})X + (\beta - \hat{\beta}) + \underline{\epsilon}$$

Linear regression least square resolution



- Most used principle is to minimize the sum of squared residuals:

Number of observations

$$minimize \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Solution of that minimization is:

$$\boxed{\hat{\beta} = \bar{y} - \hat{\alpha}\bar{X} \qquad \text{and} \qquad \hat{\alpha} = \frac{\sum_{i=1}^{n} X_i y_i - n\bar{X}\bar{y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}}$$

With $\bar{X}$ the mean value of samples X and $\bar{y}$ the mean value of observations.

This gives you the estimate function: $\widehat{f(X)} = \hat{\alpha}X + \hat{\beta} = \hat{y}$ which allows to make predictions

GDP per capita

Amount of gold medals in Olympic games

GDP: gross domestic product

A basic example of linear regression using least square method

X (input)	-2	-1	0	1	2	3	4
y (output)	-5	-3	-1	1	3	5	7

$$\widehat{f(X)} = \hat{\alpha}X + \hat{\beta}$$

- First let's compute $\hat{\alpha}$:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} X_i y_i - n\bar{X}\bar{y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} = \frac{63-7}{35-7} = \frac{56}{28} = 2$$

$$\sum_{i=1}^{n} X_i y_i = (-2 \times -5) + (-1 \times -3) + \cdots + (4 \times 7) = 10 + 3 + 0 + 1 + 6 + 15 + 28 = 63$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{7}(-2 + (-1) + \cdots + 3 + 4) = 1$$

$$\bar{y} = 1$$

$$\sum_{i=1}^{n} X_i^2 = (-2 \times -2) + (-1 \times -1) + \cdots + (4 \times 4) = 4 + 1 + 0 + 1 + 4 + 9 + 16 = 35$$

A basic example of linear regression using least square method

X (input)	-2	-1	0	1	2	3	4
y (output)	-5	-3	-1	1	3	5	7

$$\widehat{f(X)} = \hat{\alpha}X + \hat{\beta}$$

- Then compute $\hat{\beta}$:

$$\hat{\beta} = \bar{y} - \hat{\alpha}\bar{X} = 1 - 2 = -1$$

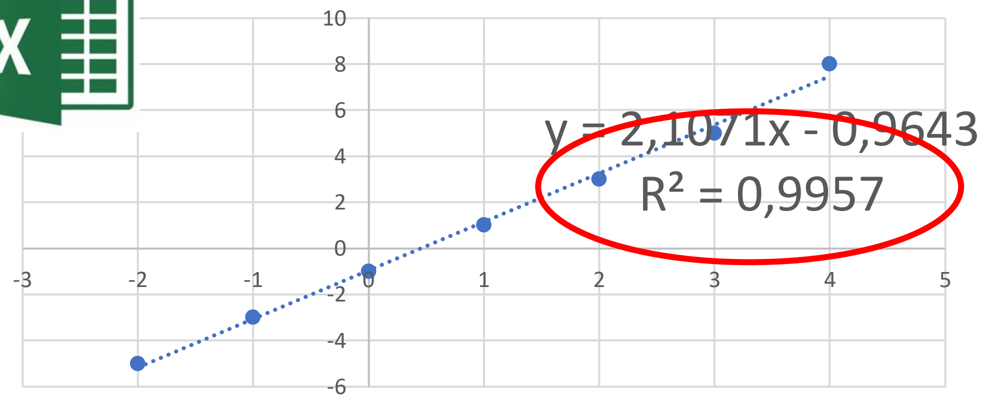
$\bar{X} = 1$, $\bar{y} = 1$ and $\hat{\alpha} = 2$

$$\boxed{\widehat{f(X)} = 2X - 1}$$

Tools to analyze your results

X (input)	-2	-1	0	1	2	3	4
y (output)	-5	-3	-1	1	3	5	7

$$\boxed{\widehat{f(X)} = 2X - 1}$$



What is R^2 ?

It indicates you a correlation score.

Best possible score is 1.0 which means perfect correlation.

Worst score is the lowest value (which can even be negative)

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Real values Predicted values

Tools to analyze your results

X (input)	-2	-1	0	1	2	3	4
y (output)	-5	-3	-1	1	3	5	7

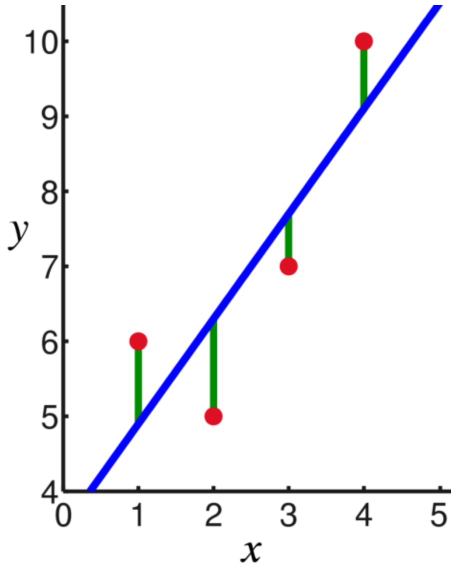
$$\boxed{\widehat{f(X)} = 2X - 1}$$

The mean squared error

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

- Gives a value for the error
- Get the square root to have an order of magnitude of the potential error

Assumptions for applying linear regression



- There is a linear relationship between the dependent variable (y) and the independent variable (x).
- The observations (y_i) are selected independently and randomly from the population.
- For ordinary least squares or Bayesian regression :
  - Residuals should be normally distributed with a mean of 0 and variance σ
  - If not the case some techniques allow the weighting of input data

Videos

- For next class, please watch: https://youtu.be/9saL47Nuguw

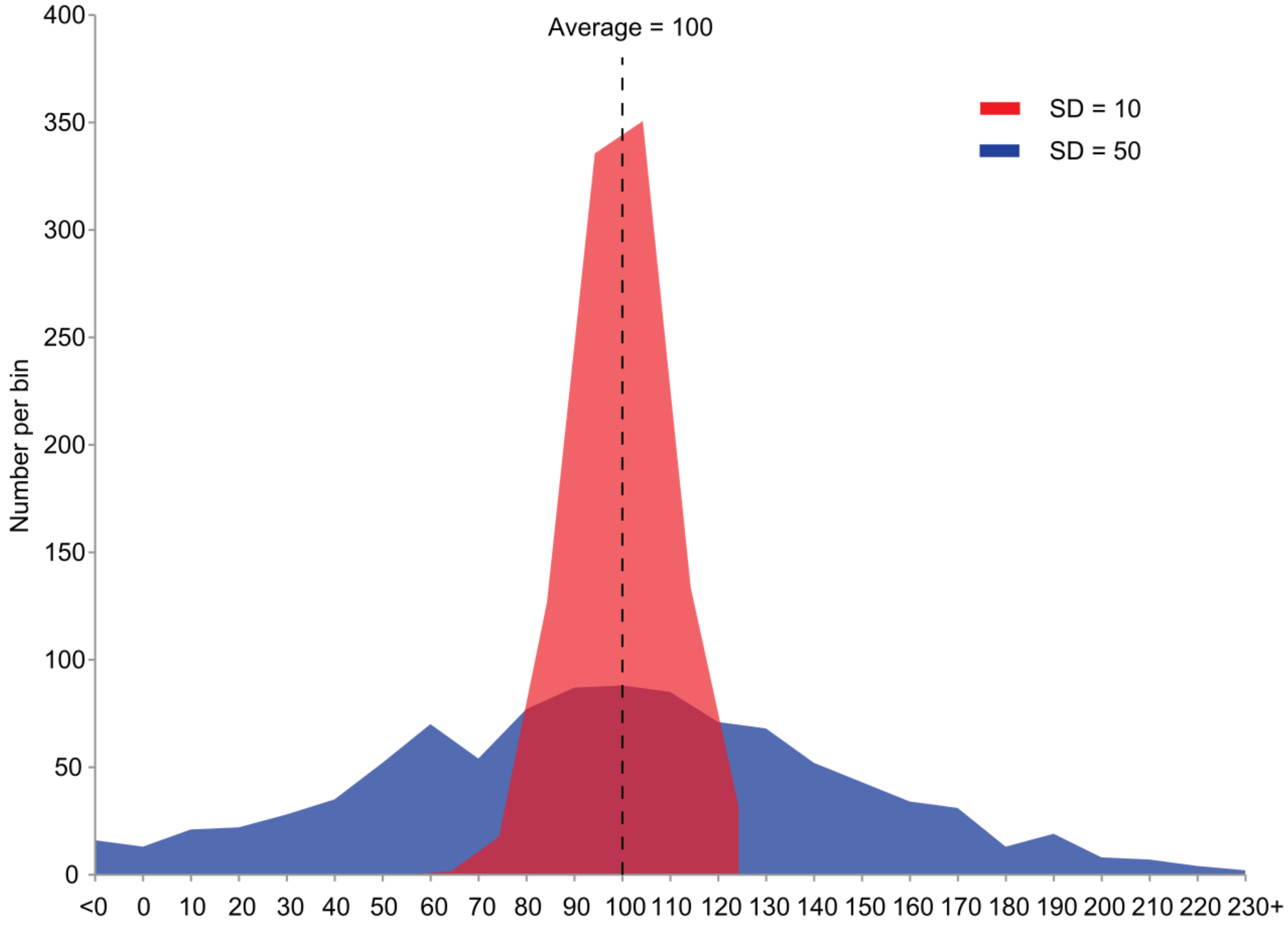
The Problem With Linear Regression | Data Analysis

- Additionally, and **optionally**, you might have a look to: https://youtu.be/eq7KF7JTinU

- Only if you have time, maybe save it for later
- 9h course
- Very similar to what we will see in this class
- Can help you understand better

Annex

Variance



- A measure of the dispersion of the data
- It is the square of the standard deviation (SD)