Linear regression and Correlation

Realized by:

HANNANI Mohamed
https://mhannani.codes/

Supervised by:

El OIRRAK Ahmed

Agenda

01 Introduction

02 What is correlation ?

03 Type of relationship

04 What is linear regression ?

05 Coefficient of Determination

Introduction

There may be complex and unknown relationships between the variables in your dataset.

It is important to discover and quantify the degree to which variables in your dataset are dependent upon each other.

This knowledge can help you better prepare your data to meet the expectations of machine learning algorithms, such as linear regression, whose performance will degrade with the presence of these interdependencies.

What is correlation ?

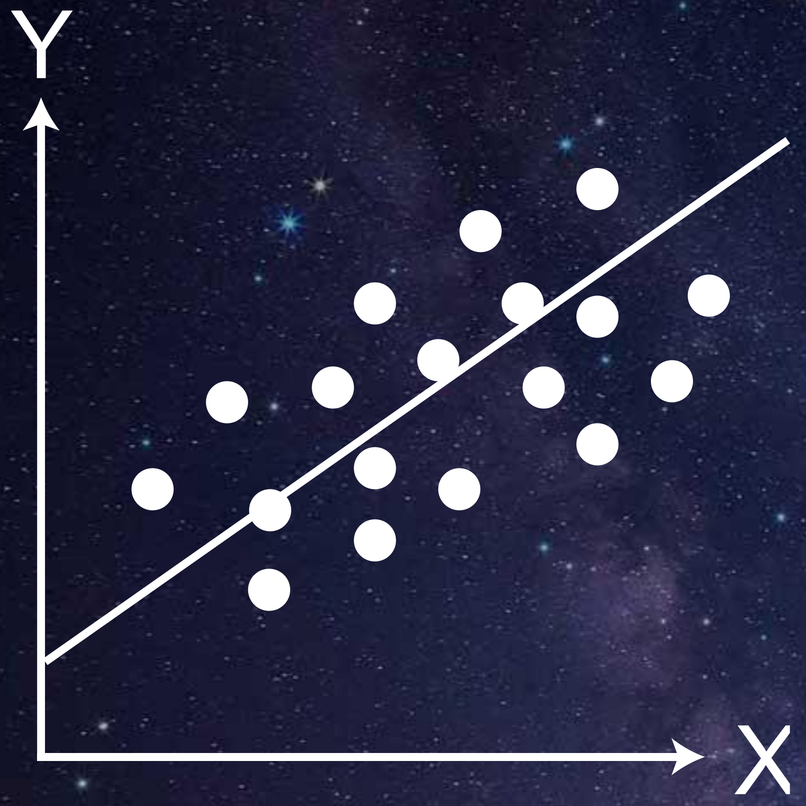
Semantically, Correlation means **Co**-together and **Relation**.

Is a statistical technique which tells us if two variables are related.

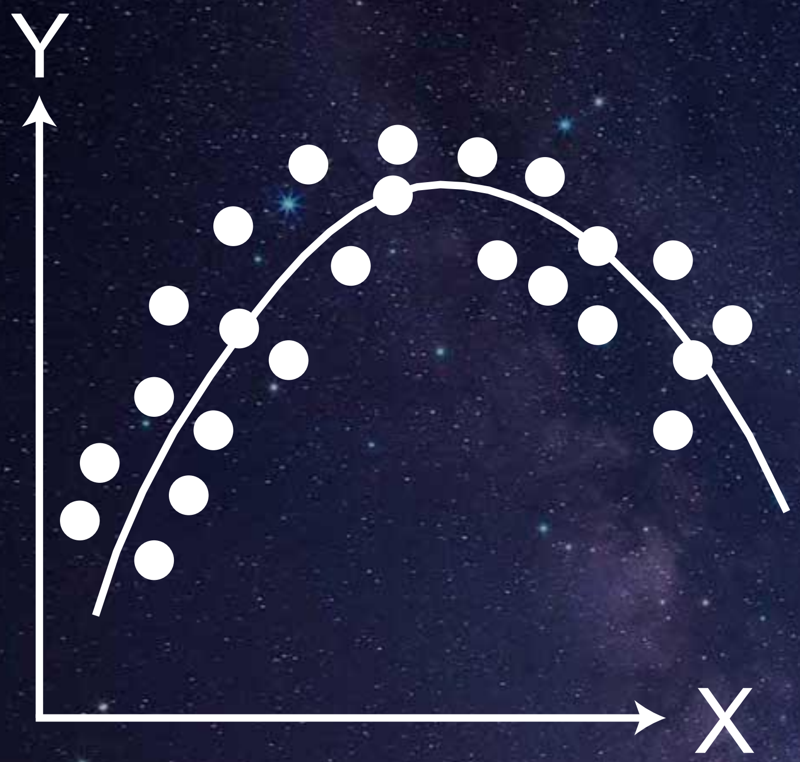
$$X \longrightarrow Y$$

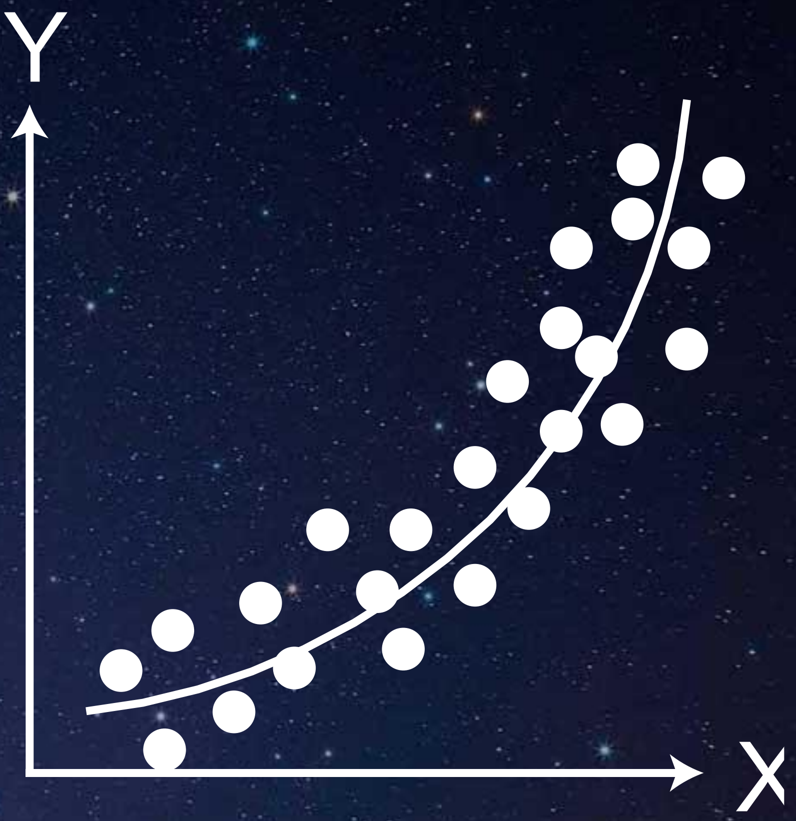
Type of relationship

Linear relationship

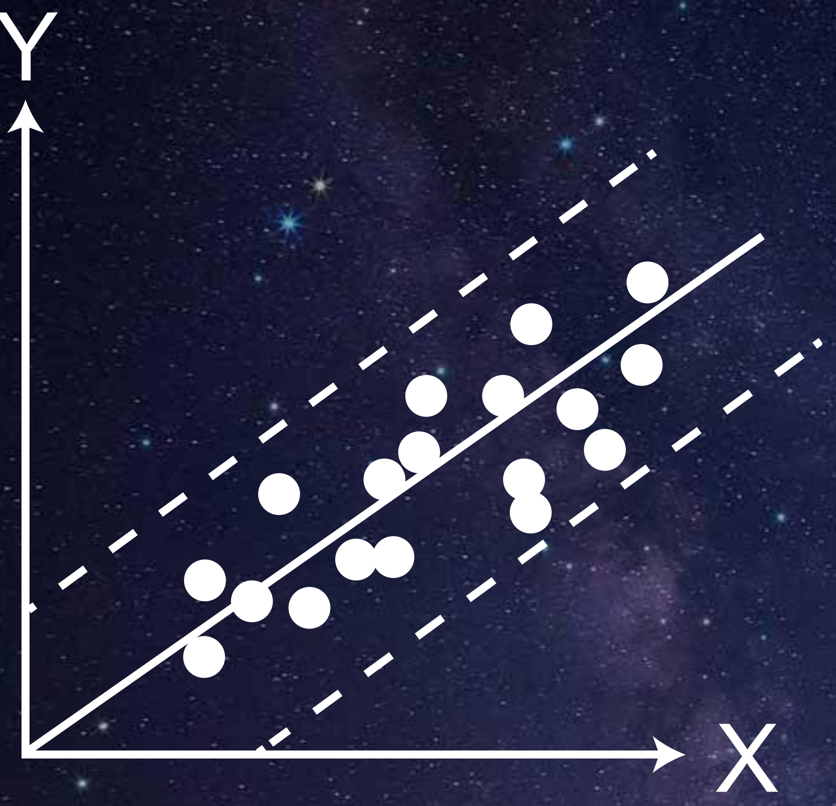


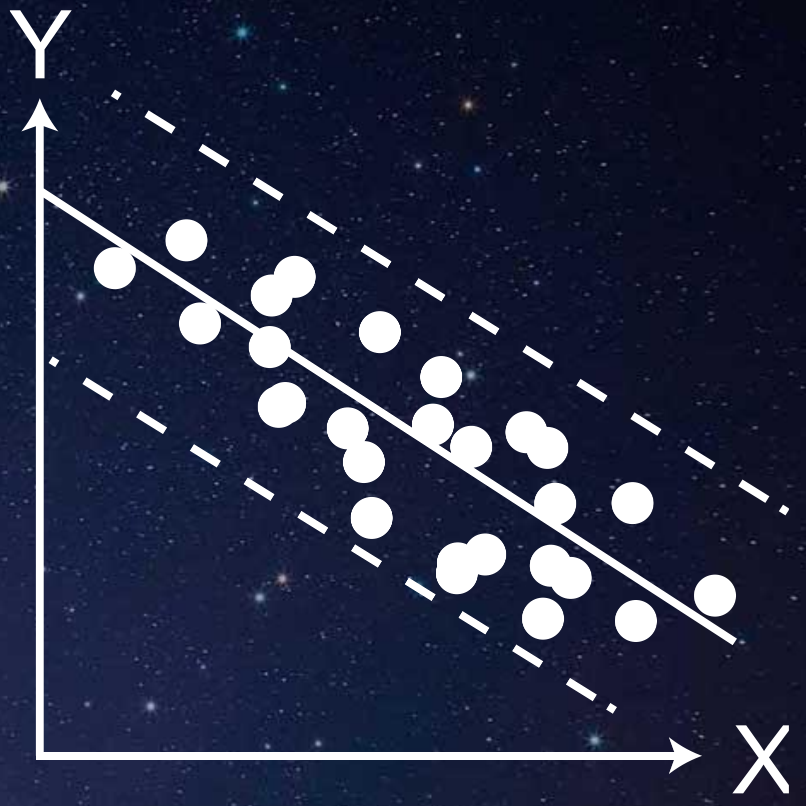
Curvilinear relationship



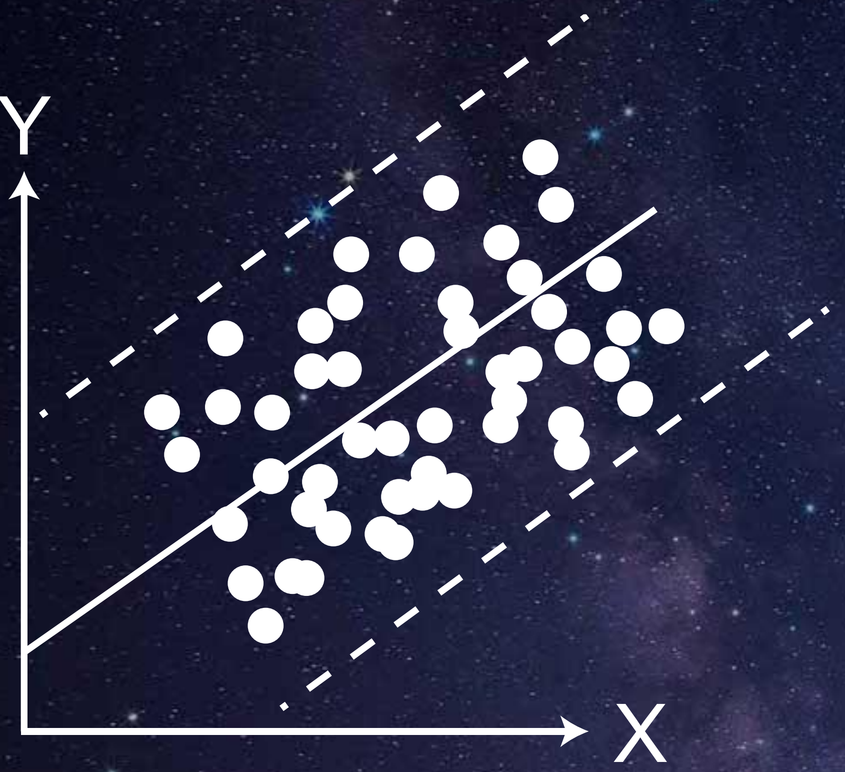


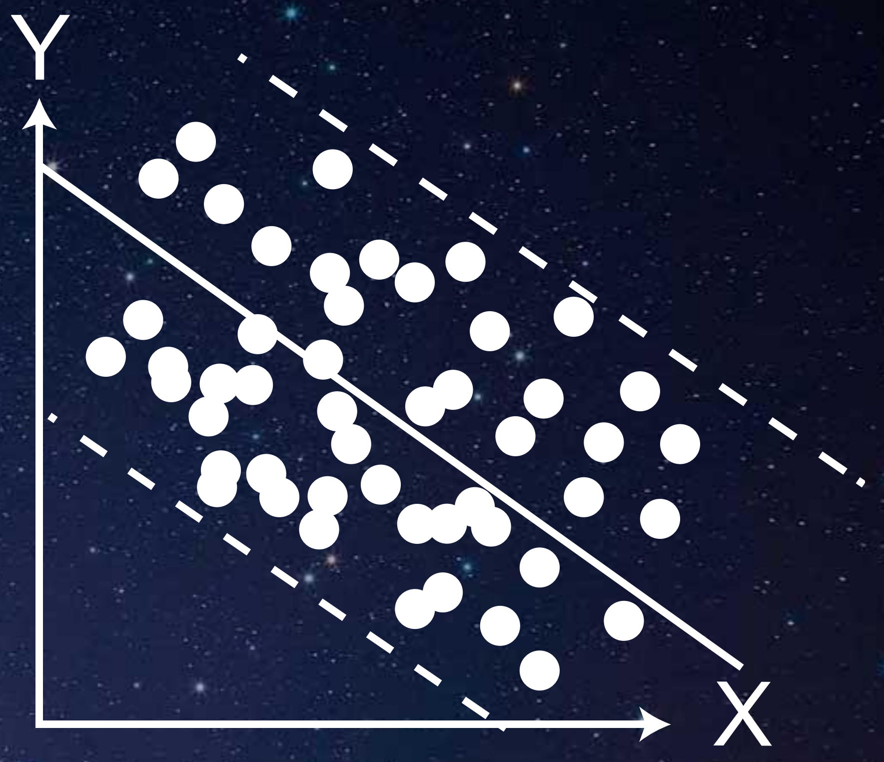
Strong relationship



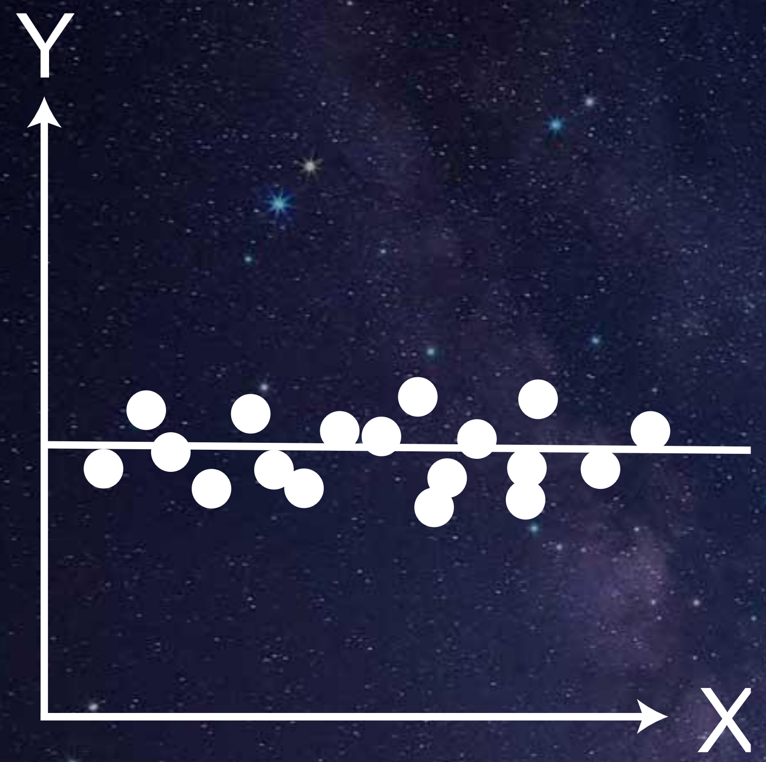


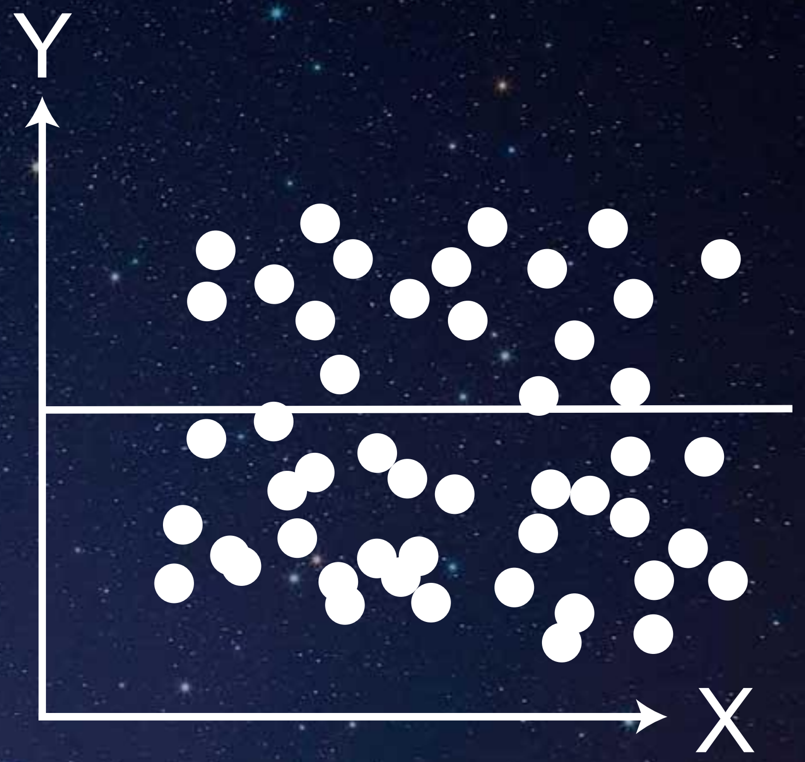
Weak relationship





No relationship





How to measure the correlation degree between two variables ?

PEARSON CORRELATION

Measures the degree of linear association between two interval scaled variables analysis of the relationship between two quantitative outcomes.

$$r = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum(X - \overline{X})^2 \cdot \sum(Y - \overline{Y})^2}}$$

$$\begin{aligned} Where,\ \overline{X} &= mean\ of\ X\ variable \\ \overline{Y} &= mean\ of\ Y\ variable \end{aligned}$$

Assumptions

- ❑ Each observation should have a pair of values.
- ❑ Each variable should be continuous.
- ❑ Each variable should be normally distributed.
- ❑ It should be an absence of outliers.

SPEARMAN CORRELATION

Measures of statistical dependence between two variables.

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(R(x_i) - R(y_i))^2}{n(n^2-1)} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$\begin{aligned} Where,\ R(x_i) &= rank\ of\ x_i \\ R(y_i) &= rank\ of\ y_i \\ n &= number\ of\ pairs \end{aligned}$$

Example

IQ, X	Hours of TV per week, Y
106	7
86	0
100	27
101	50
99	28
103	29
97	20
113	12
112	6
110	17

Example

IQ, X	Hours of TV per week, Y	Rank X	Rank Y	d	d^2
106	7	4	8	-4	16
86	0	10	10	0	0
100	27	7	4	3	9
101	50	6	1	5	25
99	28	8	3	5	25
103	29	5	2	3	9
97	20	9	5	4	16
113	12	1	7	-6	36
112	6	2	9	-7	49
110	17	3	6	-3	9

Example

$$\rho = 1 - \frac{6 * 194}{10(10^2 - 1)}$$

$$\rho = -\frac{29}{165}$$

$$\rho = -0.175757575\ldots$$

Regression analysis

Set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables

Linear regression

Linear approach to modelling the relationship between a scalar response and one or more explanatory.

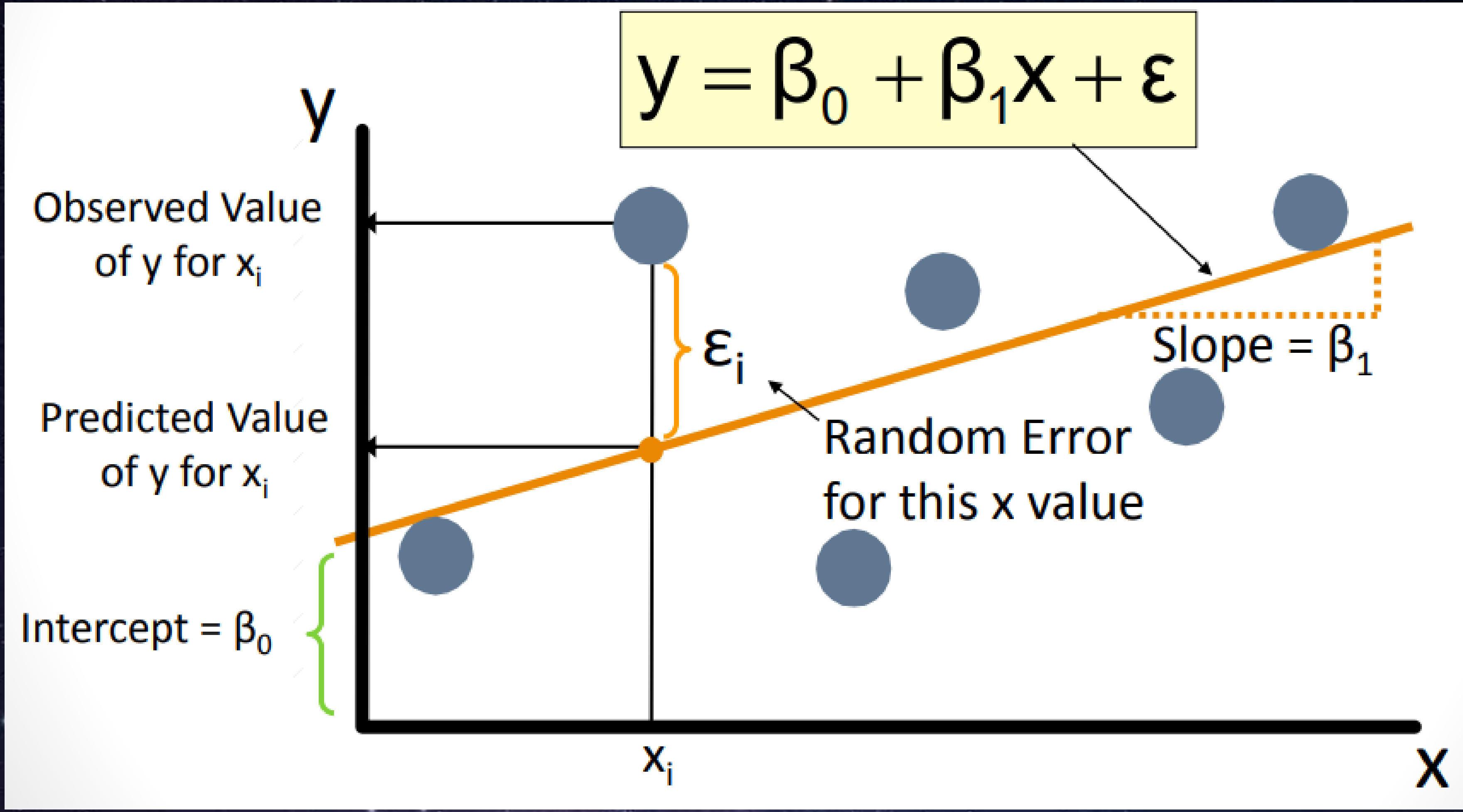
$$Y = \beta_0 + \beta_1 X_i + \epsilon_i$$

Simple linear regression

Linear regression model with a single explanatory variable.

Assumptions

- ❑ The relationship between X and Y is linear
- ❑ Y is distributed normally at each value of X
- ❑ The variance of Y at every value of X is the same



$y = \beta_0 + \beta_1 x + \varepsilon$
y
Observed Value of y for x_i
Predicted Value of y for x_i
Intercept = β_0
ε_i
Slope = β_1
Random Error for this x value
x_i
x

Take that regression line; Estimate y by substituting xi from data; Is it exactly same as yi?

Variation About a Regression Line

The total variation about a regression line is the sum of the squares of the differences between the y-value of each ordered pair and the mean of y.

$$Total\ variation = \sum (y_i - \bar{y})^2$$

Explained variation

The explained variation is the sum of the squares of the differences between each predicted y-value and the meanof y.

$$Explained\ variation = \sum (\hat{y}_i - \bar{y})^2$$

Unexplained variation

The unexplained variation is the sum of the squares of the differences between the y-value of each ordered pair and each corresponding predicted y-value.

$$Unexplained\ variation = \sum (y_i - \hat{y}_i)^2$$

Total variation

The unexplained variation is the sum of the squares of the differences between the y-value of each ordered pair and each corresponding predicted y-value.

$$\boxed{SST} = \boxed{SSE} + \boxed{SSR}$$

SST

Total sum of Squares

Quantifies how much the data points, y_i, vary around their mean, $\bar{y}$.

SSE

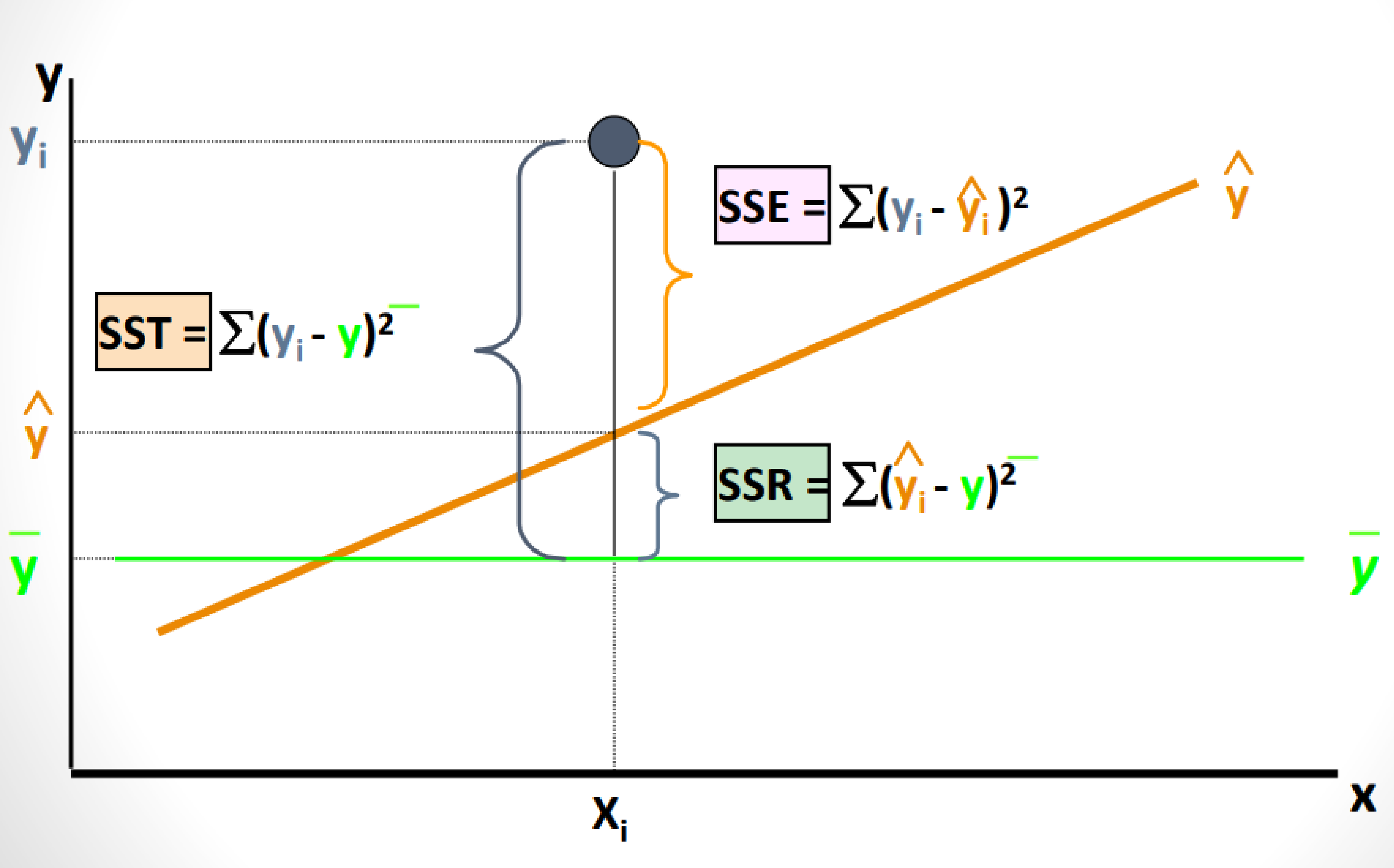
Sum of Squares Error

Quantifies how much the data points, yi, vary around the estimated regression line, y^i.

SSR

Sum of Squares Regression

Quantifies how far the estimated sloped regression line, $\hat{y}_i$, is from the horizontal "no relationship line," the sample mean or $\bar{y}$.



y
y_i
$\hat{y}$
$\bar{y}$
SSE = $\sum(y_i - \hat{y}_i)^2$
SST = $\sum(y_i - \bar{y})^2$
SSR = $\sum(\hat{y}_i - \bar{y})^2$
$\hat{y}$
$\bar{y}$
X_i
x

Coefficient of Determination

The coefficient of determination R^2 is the ratio of the explained variation to the total variation.

The coefficient of determination is also called R-squared.

$$R^2 = \frac{Explained\ variation}{Total\ variation}$$

Resources

- https://www.kaggle.com/kiyoung1027/correlation-pearson-spearman-and-kendall
- https://online.stat.psu.edu/stat462/node/95/
- https://www.colorado.edu/amath/sites/default/files/attached-files/ch12_0.pdf
- https://www.westga.edu/academics/research/vrc/assets/docs/scatterplots_and_correlation_notes.pdf
- http://hpc.ilri.cgiar.org/beca/training/AdvancedBFX2017/Statistics/Correlation_regression_10_6_17.pdf
- https://github.com/rasbt/pattern_classification/blob/master/resources/latex_equations.md
- https://en.wikipedia.org/wiki/Linear_regression#cite_note-Freedman09-1
- https://en.wikipedia.org/wiki/Normal_distribution
- https://machinelearningmastery.com/how-to-use-correlation-to-understand-the-relationship-between-variables/
- https://en.wikipedia.org/wiki/Simple_linear_regression
- https://en.wikipedia.org/wiki/Regression_analysis


THANK YOU