Regression

Penyusun Modul: Chairul Aulia

Editor: Citra Chairunnisa

Regression

"Regression analysis is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another — the effect of a price increase upon demand, for example, or the effect of changes in the money supply upon the inflation rate." *Source: Sykes (1993)*.

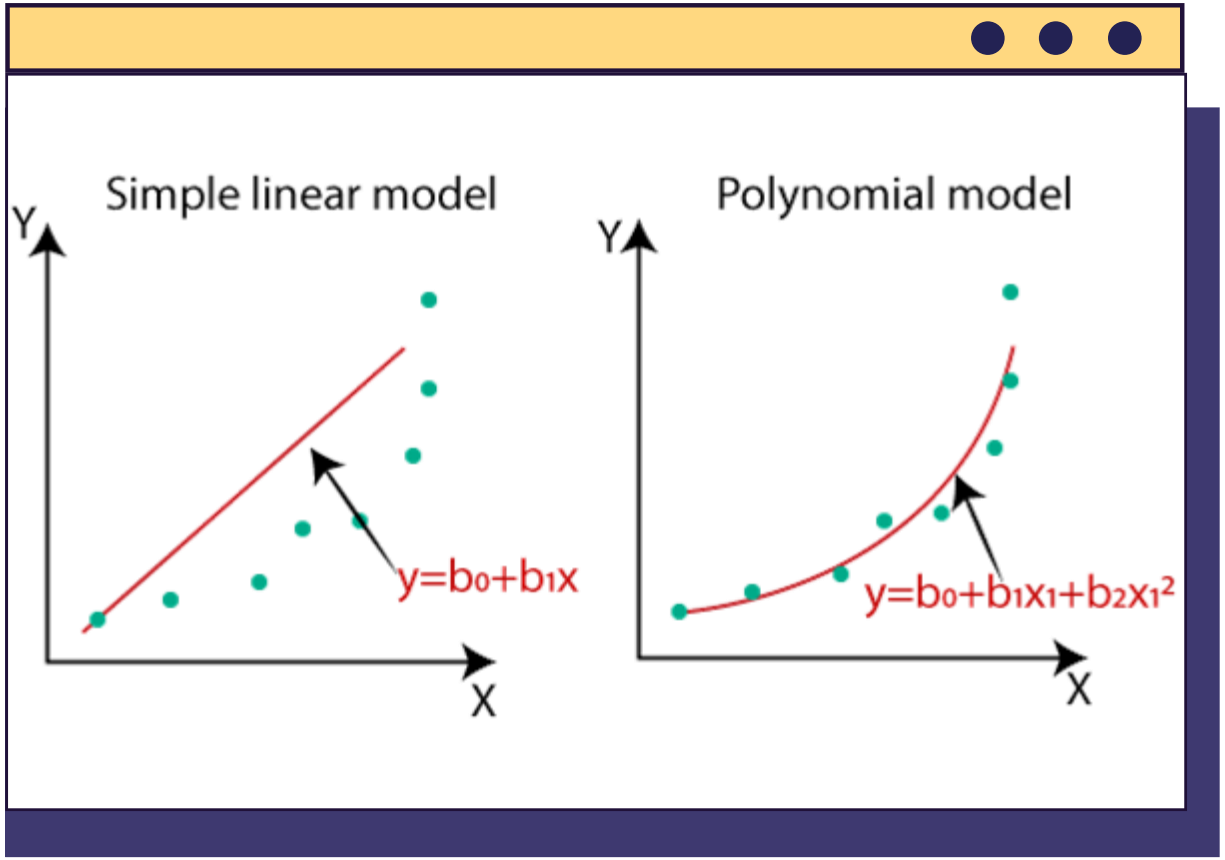
REGRESSION

Regression in machine learning consists of mathematical methods that allow data scientists to **predict** a **continuous outcome (y)** based on the value of one or more **predictor variables (x)**.

Keyword: Continuous Data

TYPES OF REGRESSION

1. Linear Regression
2. Polynomial Regression



Linear Regression

Linear regression finds the linear relationship between the dependent variable and one or more independent variables using a best-fit straight line.

The distance between each point and the line is minimized to achieve the best fit line.

You might have heard the term “Logistic Regression”, but it is actually one of the *classification* algorithms.

Least Square Method

How to obtain the best fit line?
To help us get a better understanding let's simulate the process of finding best fit line intuitively from https://phet.colorado.edu/sims/html/least-squares-regression/latest/least-squares-regression_en.html

Error/loss is calculated by subtracting the actual value from the predicted one. Since the result from subtracting might be negative, we square the difference to make it a positive value.

In regression we want to find the weights where the error is minimized

Best Fit Line

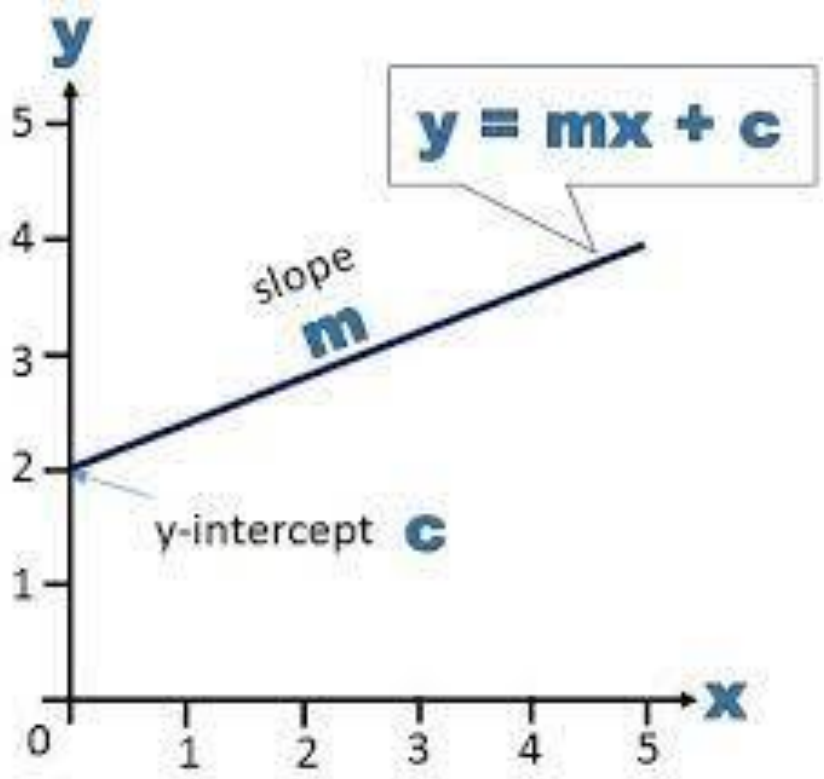
$$y = mx + c$$

y = dependent variable

m= the slope of the line

x = independent variable

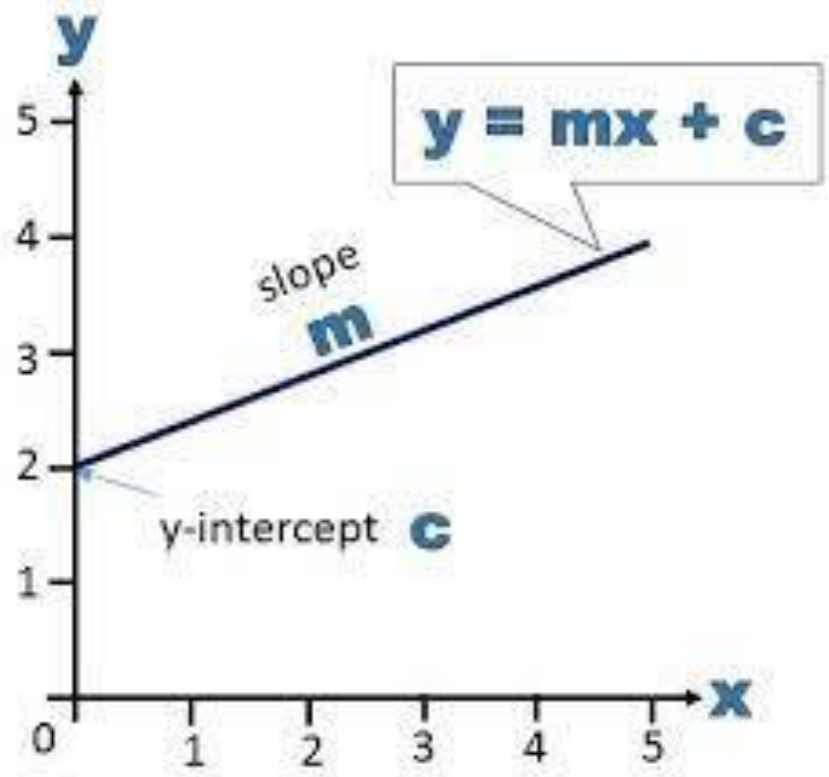
c = y-intercept



Calculating the weights

$$m = \frac{\sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{n} (x_i - \overline{X})^2}$$

$$c = \bar{Y} - m\bar{X}$$



Example 1

Let’s see implementation of linear regression without Scikit Learn in **linear_regression_implementation** file on your JupyterLab and try it out

We’ll be using **headbrain** dataset for this

Example 2

Let’s make another linear regression model!
Now we will predict salary based on years of working experience!

We’ll be using **SalaryData** dataset for simple linear regression

Example 3

How about multiple linear regression? Let's predict startup profit based on R&D spend and marketing spend!

We'll be using **50_Startups** for this example

Your Turn

Create linear regression models to predict startup profit based on

1. Only 1 feature (R&D spend, Administration or Marketing spend)
2. All 3 features (R&D spend, Administration and Marketing spend)