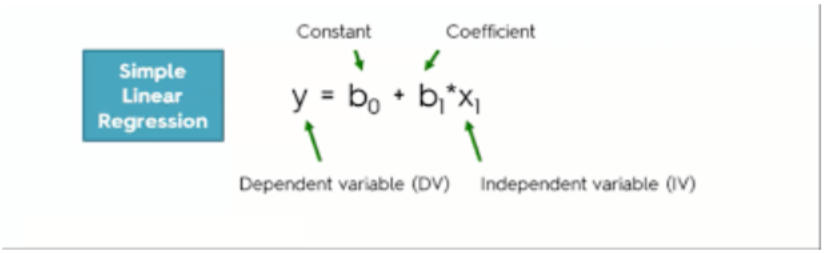
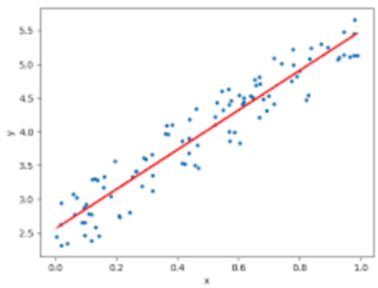
Representation of Linear Regression



Linear Regression equation

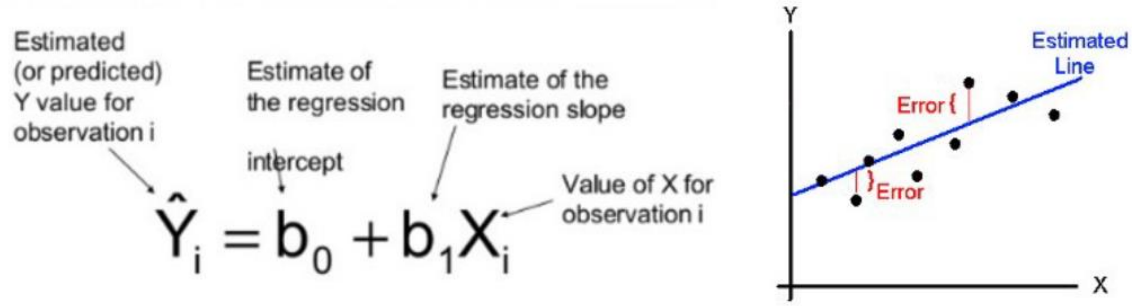
It's a method to predict a target variable by fitting the *best linear relationship* between the dependent and independent variable.



Linear Regression

In the above diagram we can clearly see the linear relationship between dependent variable y and independent variable x and that linear realationship is denoted by the best fit line.

What is best fit line?



A line is considered to be best fit line when the square of the summation of all the errors is minimum.So we can say that the distance between the best fit point and the actual point should be minimum.

The cost function of a linear regression is root mean squared error or mean squared error. They are both the same; just we square it so that we don't get negative values.

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

https://www.topcoder.com/thrive/articles/introduction-to-linear-regression#:~:text=The%20Cost%20Function%20of%20Linear%20Regression%3A&text=The%20cost%20function%20is%20the,error%20or%20mean%20squared%20error.

Use this link to get more deep understanding

Multiple Linear Regression: it's simple as its name, to elucidate the connection between the target variable and two or more explanatory variables. Multiple linear regression is used to do any kind of predictive analysis as there is more than one explanatory variable.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{\boxed{p}} X_{\boxed{p}} + \varepsilon$$

number of predictors

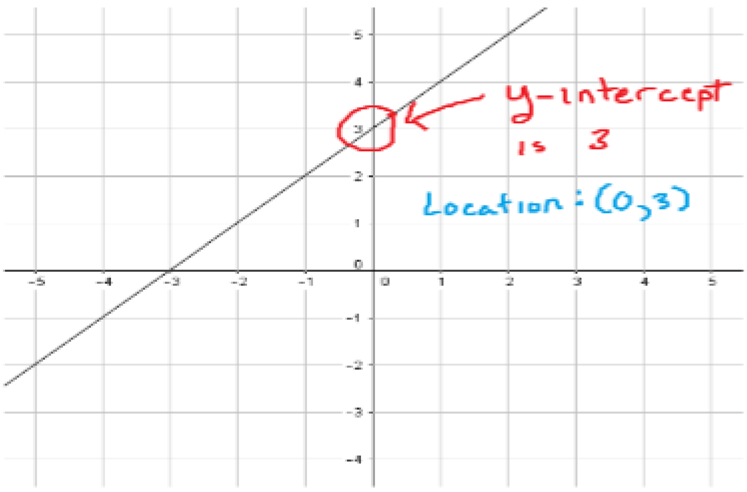
Understanding slope and intercept in regression:

Slope: Slope is what tells you how much your target variable will change as the independent variable increases or decreases.
The formula of the slope is y=mx+b

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

Intercept: The y-intercept is wherever the regression curve y=mx+b crosses the y axis (where x=0), and is denoted by b.
The formula to calculate intercept is b= y -mx

When slope and intercept are going to be placed into the formula y=mx+b, then you may get the description of the best-fit line.



How to check if the selected model is right or not.

It is very difficult to find which model is perfect for a given dataset, but here are some of the methods from which we can find the accuracy of the model.

1.R square and adjusted R square-: In regression to check the accuracy of the model we use R square and adjusted R square and for classification we use Confusion matrix.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

SSres-> sum of residual or error, SStot->sum of average total

If R square value is closer to 1 then it is a good model and if its less than 0 then the best fit line is worse than average.

As we keep on adding independent features R square value increases that means our model has a good accuracy, but there may be a case where the independent feature is not correlated with dependent feature. In that case we use adjusted R square.

$$\text{Adjusted R Squared Formula} = 1 - \left[\frac{(1 - R^2) \times (n - 1)}{(n - k - 1)} \right]$$

k-> independent feature, n-> whole sample size

Here the attributes which are not correlated to the dependent feature does not contribute to the accuracy and adjusted R square is always less than or equal to R square value.

https://www.analyticsvidhya.com/blog/2020/07/difference-between-r-squared-and-adjusted-r-squared/

learn from this blog for r vs adj r

code implementation

https://anaconda.org/sanchitiitr/linear-regression-model/notebook

https://medium.com/analytics-vidhya/machine-learning-models-linear-regression-58855efb2355