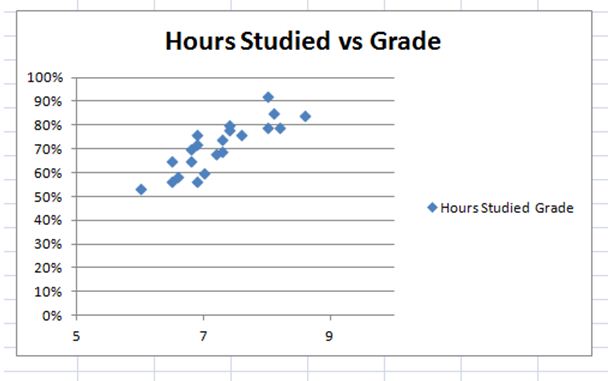
Linear Regression

Linear regression attempts to model the relationship between two variables by using a linear equation to the observed data. Linear regression has a wide range of real life applications.

For example;



There is a positive correlation between hours of study and grade. If change in variable x causes a change in variable y, x is called explanatory variable and y dependent variable.

To model data we find the best fit line.



Looks like a dark line is the best fit based visual observation. From here, we can find the slope and intercept of the dark line.Let’s assume that y=mx+b where m is slope and b y is the intercept and let’s say we found that m=9 and b = 40. Our model will be y = 9x+40. We can now use a new input and estimate output.

What will be the grade of a student who worked 3 hours?

Y = 9\*3+40=67

Of course, we made a lot of strong assumptions but this is the idea of linear regression.

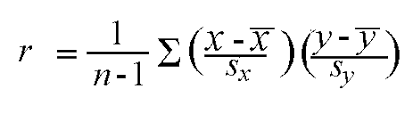
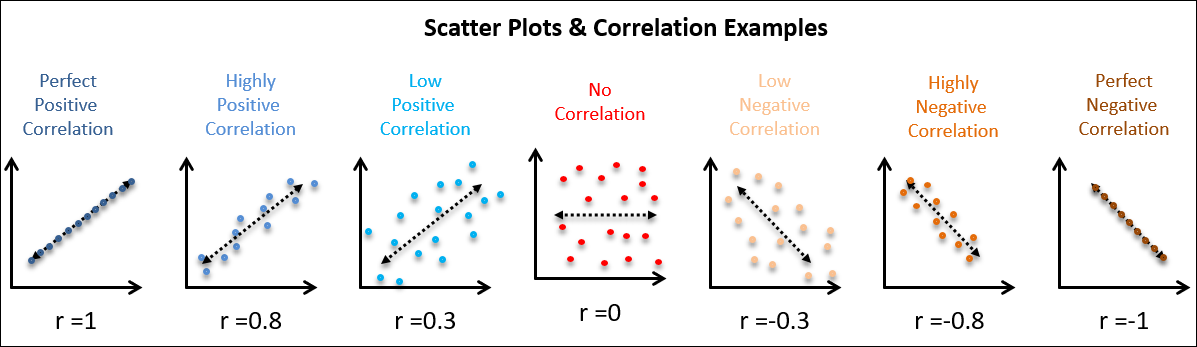
In real life it is impossible to visualize data properly. For example, if there is a large data the graph may look like a dot.

To be able to find the best model and minimize the mistakes we need to use advanced techniques.

Correlation

The strength of the associations between variables.

The value of r is between1 and -1 inclusive.



How to use r in simple linear regression to develop model

y = β0+ β1 x

y is the dependent variable

X is the explanatory variable

β0 is the y intercept

β1 is the slope

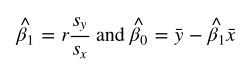
Regression Equations

Since we wouldn’t know the model we need to find the equation for the least-square regression line.



β0 hat is the estimated y intercept

β1 hat is the estimated slope



We can use the above formula and find the least square regression equation.

Let’s assume that we have found r=0.92 and standard deviation of x,y are 2,20 and mean of x,y 6, 89 respectively.

From here β1 is 0.92\*2 / 5 and β0 is 89 - β1 \* 6

β1 = 0.368 β0 = 86.792

y\_hat = 0.368 \* x + 86.792

We can now find any outputs for given inputs. Nice.

(There should be constraint on the above equation to be viable in real life. For example, we can’t plug in negative numbers or students can’t receive 900 for working extra hours)

How well does the line of the equation fit the data?

One of the ways to assess fitness is coefficient of variation or coefficient of determination (R = r^2).

We can basically find this value by taking the square root of correlation coefficient.

R tells what percent of domains explained by the model.

Another way of calculation of R;

First find Res SS, the sum of square of the residuals components

Second find Reg SS, the sum of square of the regression components

R = reg ss/ (reg ss+ res ss)

Multiple Linear Regression





Similar logic as linear regression. However, we have multiple explanatory variables and still try to minimize the above quantity.

Python Sklearn Implementation

Import libraries

Load data

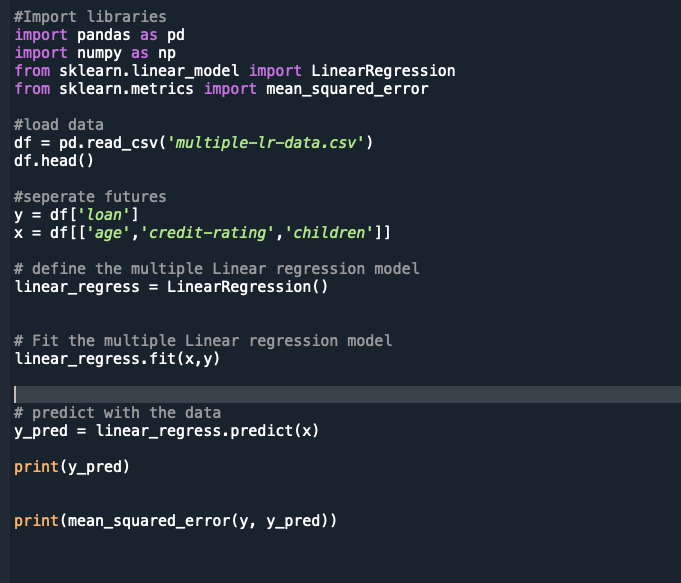
Separate futures and target

(Split data into train and test for some cases to find accuracy, Also we may need to normalize input for some ML algorithm)

Define the model

Fit the Model

Predict



Loss function by using gradient decent

<https://towardsdatascience.com/linear-regression-using-gradient-descent-97a6c8700931>

Linear Regression with Spark from Scratch