# Multiple Linear Regression (Module -7)

**Problem Statement: -**

An Analytics Company has been tasked by a crucial job of finding out what factors does affect a startup company and will it be profitable to do so or not. For this they have collected some historical data and would like to apply supervised predictive learning algorithm such as Multilinear regression on it and provide brief insights about their data. Predict Profit, given different attributes for various startup companies.

Solution:

There are five basic steps when you’re implementing Multiple linear regression:

1. Import the packages and classes you need.
2. Provide data to work with and eventually do appropriate transformations.
3. Create a regression model and fit it with existing data.
4. Check for collinearity between independent variables and between independent and dependent variables .
5. Check for overfitting issue
6. Apply the model for predictions.

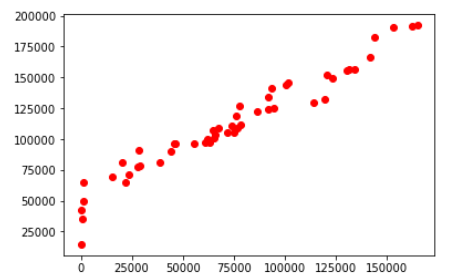
The first step is to import the package numpy and

import numpy as np

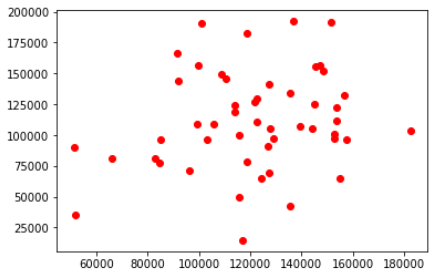
Now, you have all the functionalities you need to implement linear regression.

relavancy check of dependent variable with independent variables by scatter plot

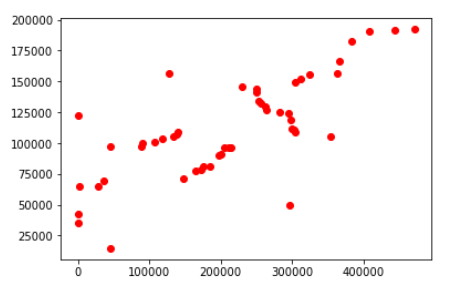
#scatterplot R&D Spend(independent) vs profit



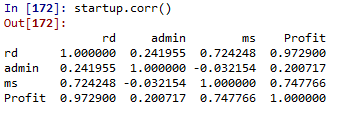
#scatterplot administration spend vs profit



#Scatter plot of marketing spend vs profit

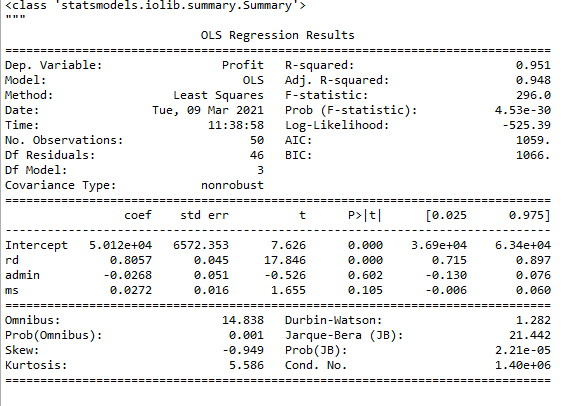


correlation between variables is obtained by correlation matrix



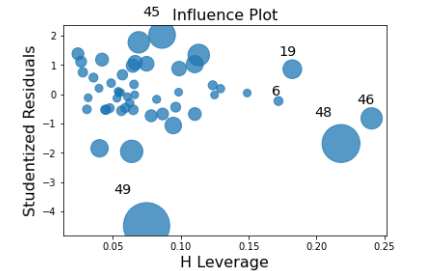
The class statsmodel.formula.api  will be used to perform linear regression and make predictions accordingly.





from statsmodels.api import sm

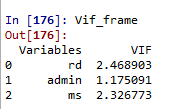
influence\_plot is plotted for model



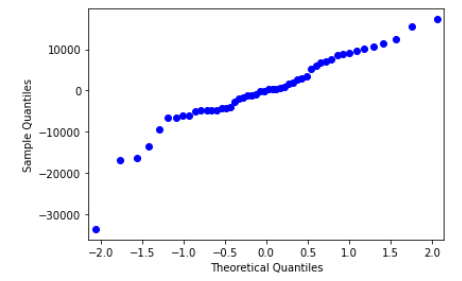
Check for Colinearity to decide to remove a variable using VIF

Assumption: VIF > 10 = colinearity

calculating VIF's values of independent variables



QQ plot is plotted for residuals



Data is splitted in to test and train with 20% test data



RMSE value for test data is obtained by

