

DATA ANALYTICS PROJECT

Building Regression Model

S Suhaasini

II Year, BE Computer Science, St Joseph Institute of Technology, Chennai

May 22, 2022

Contents

[1. ABSTRACT 2](#_Toc104137807)

[2. Introduction 2](#_Toc104137808)

[3. Existing Method 3](#_Toc104137809)

[4. Proposed method with Architecture 5](#_Toc104137810)

[5. Methodology 6](#_Toc104137811)

[6. Implementation 6](#_Toc104137812)

[7. Conclusion 10](#_Toc104137813)

**DATA ANALYTICS PROJECT**

BUILDING A REGRESSION MODEL

# ABSTRACT

Data Analytics is scientific process of transforming data into insights for making better decisions. There are four stage of analytics. Descriptive analytics describe in detail about an event that has occurred in the past; diagnostic analytics aid an analyst to dig deeper into an issue so that they can arrive at the source of a problem; predictive analytics tries to predict a trend from many different but co-dependent variables; prescriptive analytics which recommends an optimal solution

In building prediction model regression analysis is one of the powerful statistical tools. Regression Analysis is used to ascertain the probable form of the relationship between variables with the ultimate objective to predict or estimate the value of one variable corresponding to a given value of another variable. There are several regression techniques used to find the relation between the independent variable(s) and the dependent variable, *viz* Linear; Logistic; Polynomial; Stepwise, Ridge, Lasso, Elasticnet.

In the given data set 50\_Startups data there are 3 independent variables administrative spend, marketing spend and R&D spend and one dependent variable profit. It’s multivariate distribution and to start with we use Multiple Linear Regression (MLR) model.

We use Python – An anaconda with Spyder as the IDE. We import the data set and start with forward selection method. We split the data as test and training on 80:20 basis and build a multiple linear regression model (MLR). We observe the relationship in the test and predicted values and find that there is a relation.

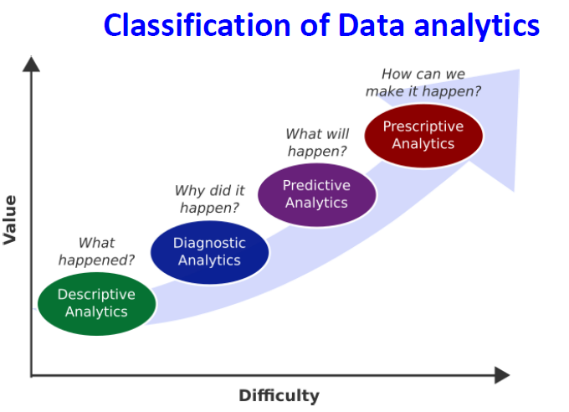
The main goal is to find an optimal team of independent variables, so that each variable in the independent team has a great impact on the dependent variable profit. That is each independent variable of the team is a powerful predictor that is highly statistically significant and has as effect on the dependent variable profit, which may be positive for an increase in 1 unit of profit or negative for a decrease in 1 unit of profit. For this, we will incorporate backward elimination. Based in the P value with Significant level (5%) we eliminate the less significant variables one by one.

Finally we observe that the independent variable R&D spend has the strongest impact on dependent variable Profit and hence we retain that alone in the linear regression model

# Introduction

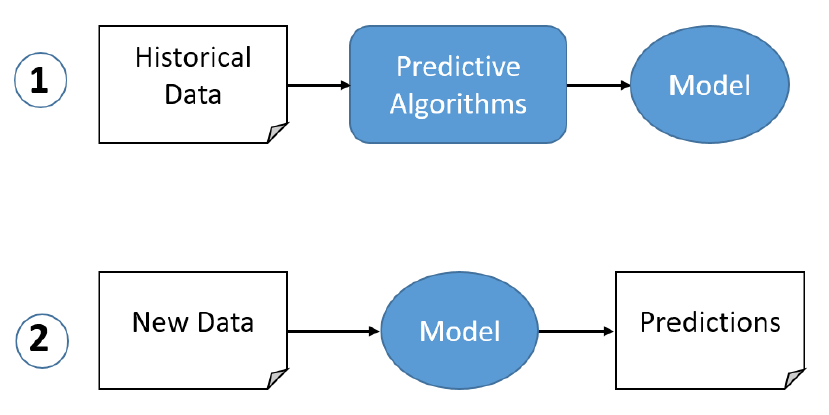
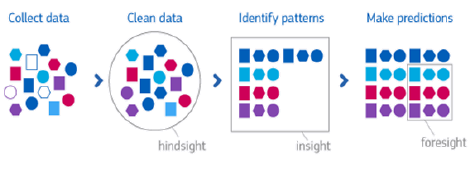
Data Analytics is defined as “the scientific process of transforming data into insights for making better decisions.” Data Analytics, is the use of data, information technology, statistical analysis, quantitative methods, and mathematical or computer-based models to help managers gain improved insight about their business operations and make better, fact-based decisions – James Evans

Data analytics is the process of examining, transforming, and arranging raw data in a specific way to generate useful information from it. It is allows for the evaluation of data through analytical and logical reasoning to lead to some sort of outcome or conclusion. It is a multi-faceted process that involves a number of steps, approaches, and diverse techniques

Based on the phase of workflow and the kind of analysis required, there are four major types of data analytics.

* ***Descriptive analytics:***
* describe in detail about an event that has occurred in the past
* ***Diagnostic analytics***
* aid an analyst to dig deeper into an issue so that they can arrive at the source of a problem
* ***Predictive analytics***
* Many different but co-dependent variables are analysed to predict a trend
* ***Prescriptive analytics***
* tells what decision to make to optimize the outcome

For the purpose of predicting statistical tools and techniques are used to build a predictive model



# Existing Method

In building prediction model regression analysis is one of the powerful statistical tools. Regression Analysis is used to ascertain the probable form of the relationship between variables with the ultimate objective to predict or estimate the value of one variable corresponding to a given value of another variable.

*Regression* is a statistical measure used in finance, investing and other disciplines that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables). The regression equation enables to predict value of variable ‘y’ for a given value of ‘x’ and vice versa and give us a better summary of the relationship between the two variables.

For e.g. if we have regression line between tax and income; for a given income we can find out the tax amount.

#### Regression Lines – Properties

* When there is perfect correlation i.e. when r = ± 1, the two regression lines coincide (become one line).
* The farther the two regression lines from each other, the lesser is the degree of correlation.
* The two regression lines always meet at point (‾X,‾Y)

#### Types of regression

* Linear Regression:
  + Linear relation between the independent variable and the dependent variable.
* Logistic Regression
  + Logistic regression is used to find the probability of event=Success and event=Failure. We use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1
* Polynomial Regression
  + A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation: y=a+b\*x**^2**
* Stepwise Regression
  + This form of regression is used when we deal with multiple independent variables. In this technique, the selection of independent variables is done with the help of an automatic process, which involves *no* human intervention.
  + The aim of this modeling technique is to maximize the prediction power with minimum number of predictor variables. It is one of the method to handle higher dimensionality of data set
* Ridge Regression:
  + Ridge Regression is a technique used when the data suffers from multicollinearity ( independent variables are highly correlated).
* Lasso Regression:
  + Similar to Ridge Regression, Lasso (Least Absolute Shrinkage and Selection Operator) also penalizes the absolute size of the regression coefficients. In addition, it is capable of reducing the variability and improving the accuracy of linear regression models.
* Elasticnet Regression:
  + ElasticNet is hybrid of Lasso and Ridge Regression techniques. It is trained with L1 and L2 prior as regularizer. Elastic-net is useful when there are multiple features which are correlated. Lasso is likely to pick one of these at random, while elastic-net is likely to pick both

#### Several methods to build a model:

* Backward-Elimination
* Forward Selection
* Bi-directional Elimination
* Score Comparison

#### Backward Elimination:

Step1: The first step in the starting is to select a significance level (SL).

Step2: Now fit the whole model with all the possible predictors

Step3: Look for the highest P-value. If P>SL, then move to step 4, otherwise the model is ready.

Step4: Eliminate the Predictor.

Step5: Fit the model excluding that variable.

#### Forward Selection:

Step1: To enter a model, select a significance level (e.g. SL = 0.05).

Step2: Now fit all the simple regression models, and select the one with the lowest P-value.

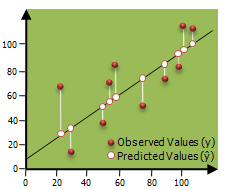
Step3: Preserve this variable, and fit all the promising models with one predictor added to one’s that we are already having.

Step4: Consider the model which is having the lowest p-value, and if P<SL then, go to step 3, otherwise the model is ready.

# Proposed method with Architecture

In this project (data set 50 startups) has three independent variables administrative spend, marketing spend and R&D spend and one dependent variable profit. It’s multivariate distribution and to start with we use Multiple Linear Regression (MLR) model.

#### Linear Regression

Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line).

It is represented by an equation Y=a+b\*X + e, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

**Multiple Linear Regression**

Multiple linear regression is an enhancement of simple linear regression. Two or more features is used in prediction.

Any dataset with n no. of observations, p independent variables and y as the response-dependent variable the regression line for p features can be mathematically written as;

F(xi)= m0+m1xi1 +m2xi2+………+mpxip

Where, F(Xi) is the response of the predicted value and m0,m1,m2,..mp is the regression coefficients.

Generally, it adds an error to the data which is called as the residual error and modifies the equation as follows;

F(xi)= m0+m1xi1 +m2xi2+………+mpxip+ei

yi=F(xi)+ei

#### Forward Selection:

Step1: To enter a model, select a significance level (e.g. SL = 0.05).

Step2: Now fit all the simple regression models, and select the one with the lowest P-value.

Step3: Preserve this variable, and fit all the promising models with one predictor added to one’s that we are already having.

Step4: Consider the model which is having the lowest p-value, and if P<SL then, go to step 3, otherwise the model is ready.

**Assumptions in this Regression model:**

**Linearity**:  The relation between the dependent and independent variables is assumed to be linear.

**Homoscedasticity**: The feature variable maintain constant variance.

**Multivariate Normality**: It is assumed in Multiple Regression that the residuals are normally distributed.

**Multicollinearity**: In this model, it is assumed that there is an either little or no multicollinearity in the data.

# Methodology

#### A stepwise implementation of the Multiple Linear Regression (MLR) model:

Step1: Data Pre-processing:

1. Import Libraries.
2. Import Dataset.
3. Encode Categorical Data and avoid a Dummy Variable Trap.
4. Split Dataset into a Training set and Test set.

Step2: Fit Multi Linear Regression to the Training set.

Step3: Predict the Results of the Test set.

# Implementation

#### Step 1: Data Pre-processing:

We use Anaconda – Spyder (Python3.9) for this data analysis. We start Syder and create a new file CODE.py. We will first **import the libraries**. Then we **import the dataset** - 50\_Startups.csv.

#importing the libraries

import numpy as np

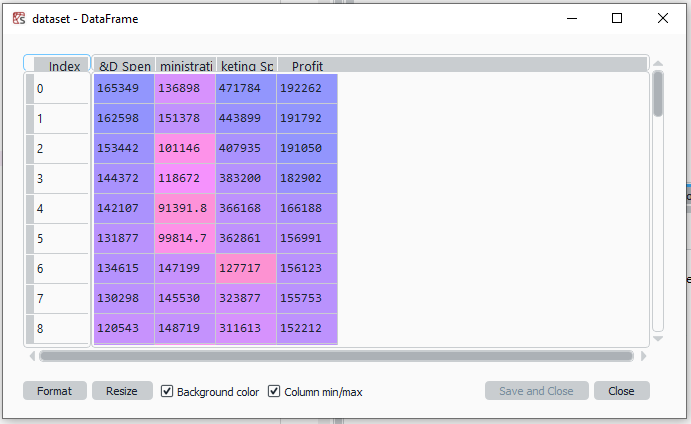
import pandas as pd

import matplotlib.pyplot as plt

#importing the dataset using the Pandas library

dataset= pd.read\_csv('50\_Startups.csv')

The dataset of data of 50 companies. It has the information for a particular financial year regarding money spent on R&D, Administration, Marketing, and profit earned.



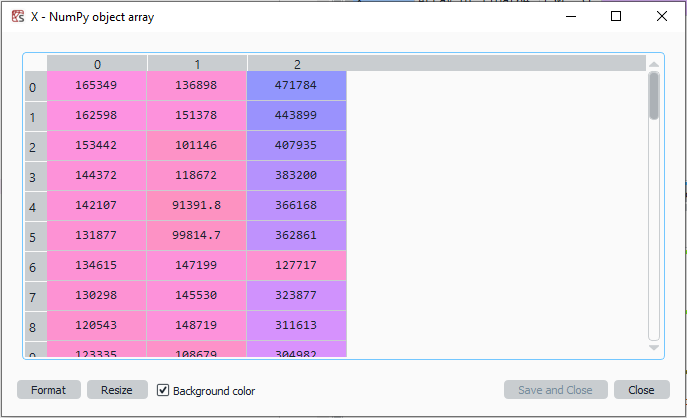
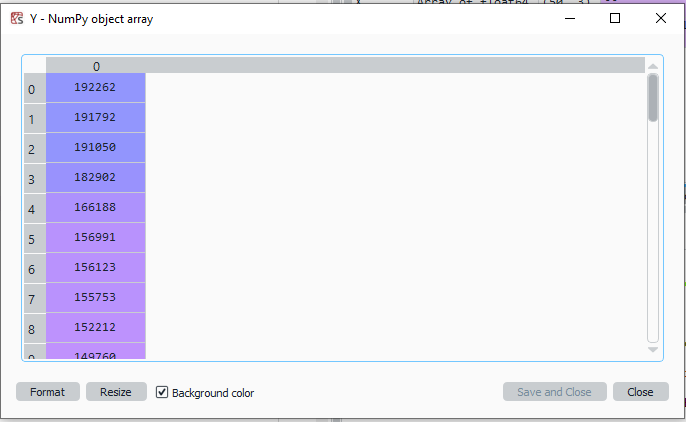
Now we extract the feature matrix. Here, X is the matrix of the independent variables, and Y is the matrix of the dependent variable.

#extracting matrix features

X= dataset.iloc [:,:-1]. values

Y= dataset.iloc [: , 3].values

Since variable X as the matrix contains same data types we can view in the Variable Explorer.



Normally, if there are categorical variables it will cause a problem in machine learning model equations. To encode it we use LabelEncoder to encode that particular column However, since **there are no categorical variables** we do not have to perform the step of encoding Categorical Data

Now we will **split** the dataset into the training set and the test set. The test set will contain the ten observations, and the training set will have 40 observations.

# Splitting the dataset into the Training set and Test set

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, Y, test\_size = 0.2, random\_state = 0)

#### Step2: Fit Multi Linear Regression to the Training set.

Now, will fit the multiple linear regression to the training set. We will use the fit method to fit the regressor object in the training set.

# Fitting Multiple Linear Regression to the Training set

from sklearn.linear\_model import LinearRegression

regressor = LinearRegression()

regressor.fit(X\_train, y\_train)

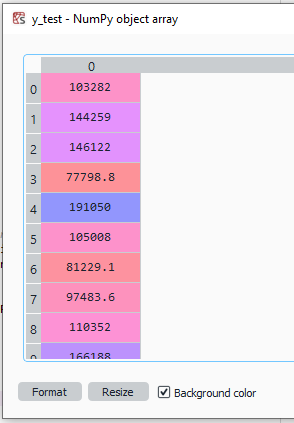
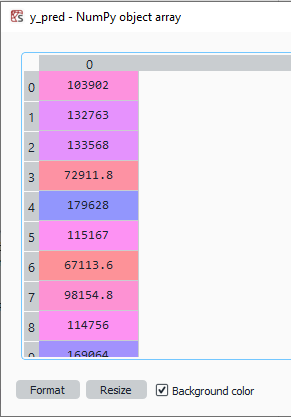
#### Step3: Predict the Results of the Test set.

Now the next step of this model is the prediction of test set results.

# Predicting the Test set results

y\_pred = regressor.predict(X\_test)

On execution of the above code, we will observe the following predicted value of the test set. We can also compare the real values with the predicted values by clicking at Y\_test in variables explorer’s pane.



From the above images, it’s seen that there is a multilinear dependency between the independent variables as well as the dependent variables.

When we built this model, we actually used all the independent variables. But if among these independent variables some of them are highly statistically significant, means they have a great impact on the dependent variable of profit and some that are not statistically significant at all, which means even if we remove these variables from the model, we will still get a fair prediction.

#### Backward elimination

The main goal is to find an optimal team of independent variables, so that each variable in the independent team has a great impact on the dependent variable profit. That is each independent variable of the team is a powerful predictor that is highly statistically significant and has as effect on the dependent variable profit, which may be positive for an increase in 1 unit of profit or negative for a decrease in 1 unit of profit. For this, we will incorporate backward elimination.

We first import the statsmodel library.

import statsmodels.api as sm

We will now add a column of 1 to the matrix X or matrix of independent variables, which will correspond to x0=1 associated with constant b0. For this, we will use the append function to add the column of 1 to the matrix.

X=np.append(arr = np.ones((50,1)).astype(int), values=X, axis=1)

As we want to add the column of 1 at the beginning to the matrix, so instead of adding the column to the matrix, we are adding the matrix to the column. Because when we were adding the column to the matrix, it was getting added at the end. So, we choose to add the matrix to the column of 1. Here we have used axis=0, as we want to add a line.

We are actually proceed with Backward Elimination:

We will first create a new matrix of features **X\_opt** which will be the optimal features of the matrix, i.e., a matrix containing an optimal team of independent variables that have a high impact on the dependent variable.

X\_opt=X[:, [0,1,2,3]]  
regressor\_OLS=sm.OLS(endog = Y, exog=X\_opt).fit()  
regressor\_OLS.summary()

We then created a new regressor object for the new class of statsmodel library. We are fitting ordinary least square algorithm to X\_opt and Y. And to get the significant value for the independent variable summary function is used.

**Output:**

OLS Regression Results

==============================================================================

Dep. Variable: y R-squared: 0.951

Model: OLS Adj. R-squared: 0.948

Method: Least Squares F-statistic: 296.0

Date: Sun, 22 May 2022 Prob (F-statistic): 4.53e-30

Time: 10:30:10 Log-Likelihood: -525.39

No. Observations: 50 AIC: 1059.

Df Residuals: 46 BIC: 1066.

Df Model: 3

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

const 5.012e+04 6572.353 7.626 0.000 3.69e+04 6.34e+04

x1 0.8057 0.045 17.846 0.000 0.715 0.897

x2 -0.0268 0.051 -0.526 0.602 -0.130 0.076

x3 0.0272 0.016 1.655 0.105 -0.006 0.060

==============================================================================

Omnibus: 14.838 Durbin-Watson: 1.282

Prob(Omnibus): 0.001 Jarque-Bera (JB): 21.442

Skew: -0.949 Prob(JB): 2.21e-05

Kurtosis: 5.586 Cond. No. 1.40e+06

==============================================================================

From the output given above, it can be seen that we got P-value for each independent variable. We will now compare the P-value with the significant level (SL=0.05), to decide if we need to remove it from out model or not.

Now the highest P-value is **60%**, so we will remove **x2**which is the **Admin spend** at index=2 and again check for the highest P-value.

X\_opt=X[:, [0,1,3]]

regressor\_OLS=sm.OLS(endog = Y, exog=X\_opt).fit()

regressor\_OLS.summary()

The model is finally getting ready, may be index 1 & 3 which is the R&D spend and Marketing spend composed of independent variables makes the best team to predict the profit.

OLS Regression Results

==============================================================================

Dep. Variable: y R-squared: 0.950

Model: OLS Adj. R-squared: 0.948

Method: Least Squares F-statistic: 450.8

Date: Sun, 22 May 2022 Prob (F-statistic): 2.16e-31

Time: 11:10:33 Log-Likelihood: -525.54

No. Observations: 50 AIC: 1057.

Df Residuals: 47 BIC: 1063.

Df Model: 2

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

const 4.698e+04 2689.933 17.464 0.000 4.16e+04 5.24e+04

x1 0.7966 0.041 19.266 0.000 0.713 0.880

x2 0.0299 0.016 1.927 0.060 -0.001 0.061

==============================================================================

Omnibus: 14.677 Durbin-Watson: 1.257

Prob(Omnibus): 0.001 Jarque-Bera (JB): 21.161

Skew: -0.939 Prob(JB): 2.54e-05

Kurtosis: 5.575 Cond. No. 5.32e+05

==============================================================================

The highest P-value is here 0.000 and 0.060 and which is x1 where below SL of 5% is the lowest.

Thus we observe that the independent variable R&D spend has the strongest impact on dependent variable Profit and hence we retain that alone in the linear regression model.

# Conclusion

Therefore, it can be seen that the R&D spend variable to make a very powerful predictor of profit. So, finally, the optimal team of the independent variable that can predict profit with the highest statistical significance the strongest impact is composed of only one independent variable i.e., R&D spend.