Topics in Computer Science – Machine Learning

CSCI 6905 Spring 2018, Group 1

**Multiple Linear Regression with Backward Elimination**

William Clark

Sumati Kulkarni

Babak Maleki Shoja

Venkatesh Reddy Pala

Vishwa Patel

https://lh3.googleusercontent.com/Fbw3or2c1n0LWeh8KnpszbWNmIbahCiYK5qOCOAdU0pL-N50s__3zO8c0OstdkQ43Y17Ze2q1rSSPrTLyj9zMXjBJDUvOv_nI4ONnpeZoglLFbmBMUXYqO_t0cGfGm6D4YSZILC9RvdNyOCYqw

Table of Contents

[**Learning Goals**](https://docs.google.com/document/d/1TjLMUhfJBWyM-008eOZssdetg8k5EJlJroykulQfWbM/edit#heading=h.30j0zll)[**3**](https://docs.google.com/document/d/1TjLMUhfJBWyM-008eOZssdetg8k5EJlJroykulQfWbM/edit#heading=h.30j0zll)

1. Multiple Linear Regression

1.1 Introduction [4](https://docs.google.com/document/d/1TjLMUhfJBWyM-008eOZssdetg8k5EJlJroykulQfWbM/edit#heading=h.3znysh7)

1.2 50 Startup Companies Profit Prediction 5

1.3 What is Backward Elimination? 5

1.4 Mechanism of Backward Elimination 6

1.5 Steps of Backward Elimination 6

2[.](https://docs.google.com/document/d/1TjLMUhfJBWyM-008eOZssdetg8k5EJlJroykulQfWbM/edit#heading=h.tyjcwt) Implementation of MLR 7

2.1 Results 16

2.2 [Conclusion](https://docs.google.com/document/d/1TjLMUhfJBWyM-008eOZssdetg8k5EJlJroykulQfWbM/edit#heading=h.2s8eyo1) [1](https://docs.google.com/document/d/1TjLMUhfJBWyM-008eOZssdetg8k5EJlJroykulQfWbM/edit#heading=h.2s8eyo1)6

[**Individual Contributions**](https://docs.google.com/document/d/1TjLMUhfJBWyM-008eOZssdetg8k5EJlJroykulQfWbM/edit#heading=h.17dp8vu)[1](https://docs.google.com/document/d/1TjLMUhfJBWyM-008eOZssdetg8k5EJlJroykulQfWbM/edit#heading=h.17dp8vu)7

[**Team Summary**](https://docs.google.com/document/d/1TjLMUhfJBWyM-008eOZssdetg8k5EJlJroykulQfWbM/edit#heading=h.26in1rg)18

# Learning Goals

In this assignment, the team developed a model for dataset which contains data of 50 companies in3 states. The descriptive features contains 1) R&D Spend by Company 2) Administration and 3) Marketing. Based on all of these descriptive features we then decide how much profit that company made. So profit is the Target Feature, The purpose of this assignment was to get familiar with applying a prediction model to an existing problem, how we can evaluate the goodness of fit and to determine features that can be excluded from the data.

# Multiple Linear Regression – 50 Startup Companies

https://lh3.googleusercontent.com/Fbw3or2c1n0LWeh8KnpszbWNmIbahCiYK5qOCOAdU0pL-N50s__3zO8c0OstdkQ43Y17Ze2q1rSSPrTLyj9zMXjBJDUvOv_nI4ONnpeZoglLFbmBMUXYqO_t0cGfGm6D4YSZILC9RvdNyOCYqw

### Introduction

Multiple linear regression (MLR) is an extension of simple linear regression.  In the previous chapter we considered a single dependent variable, *y*, and a single independent variable *x*.  MLR is used when there are two or more independent variables where the model using population information is

***yt = β0 + β1x1t + β2x2t + β3x3t + ……. + βkxkt***

It should not be surprising that most interesting phenomena are too complex to be modeled using just a single independent variable.  MLR allows a much more comprehensive model than does simple LR and should provide superior predictions.

**The assumptions of MLR**

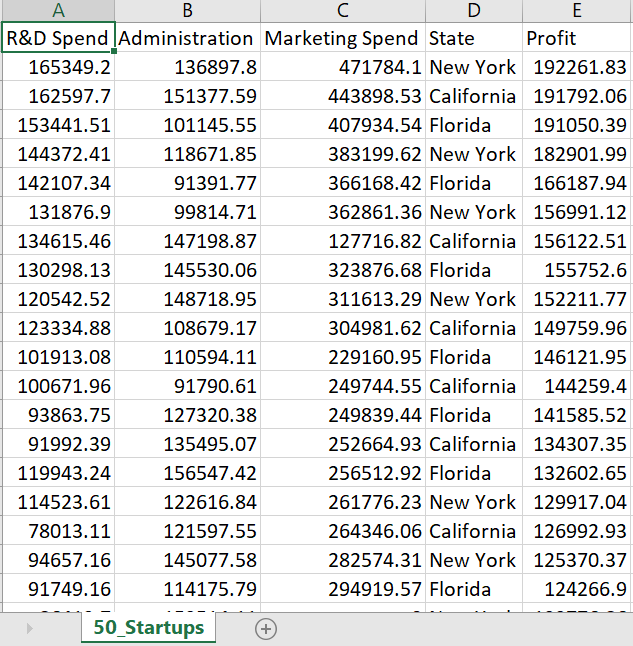
MLR uses some assumptions.  These are necessary for the mathematics to work out properly.  The assumptions used in MLR:

1. The error terms *εi* are normally distributed
2. The error terms are independent of past error terms, that is ***Ε (εi| εi-1 , εi-2,..) = Ε(εi)***
3. The populations all have equal variances, σ1 = σ2 = σ3 = …
4. The independent variables are not correlated with each other.

In the least-squares model, the best-fitting line for the observed data is calculated by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values. The least-squares estimates b0,…,bt are usually computed by statistical software.

### 50 Startup Companies Profit Prediction

The dataset we have used for this project consists of 50 different companies located in 3 different states 1) California 2) New York 3) Florida. The other descriptive features are “R&D budget”, “Administration” and “Marketing” which consists of amount of money which 50 companies spend on these different fields. The target feature is to calculate the amount of Profit these companies made based on all of these descriptive features. Here is how the dataset looks like:



### What is Backward Elimination?

* The idea of Backward Elimination is to remove independent variables that are not statistically significant.
* If your dataset is huge, this could make a great difference, because your model can run with less data.
* Our goal here is to find a group of independent variables that all big impact to the dependent variable.

### Mechanism of Backward Elimination

1. Select a significant level (i.e.: Significant Model = 0.05 ; If the P value is greater than this significant level, then we will remove it)
2. First fit ALL variables to the model.
3. Find the P values for ALL variables.
4. Remove the variable with the largest P value.
5. Fit the model with a variable removed from Step 4.
6. Repeat Step 4 & 5, until all P values are smaller than the significant level defined in Step 1.
7. Model is ready.

### **Steps of Backward Elimination**

**Step-1:** Select a significant level to stay in the model (e.g. Significant Level = 0.05).

**Step-2:** Fit the full model with all possible predictors.

**Step-3:** Consider the predictor with the highest P-value. If P > SL, go to STEP-4 otherwise go to **FINISH.**

**Step-4:** Remove the predictor.

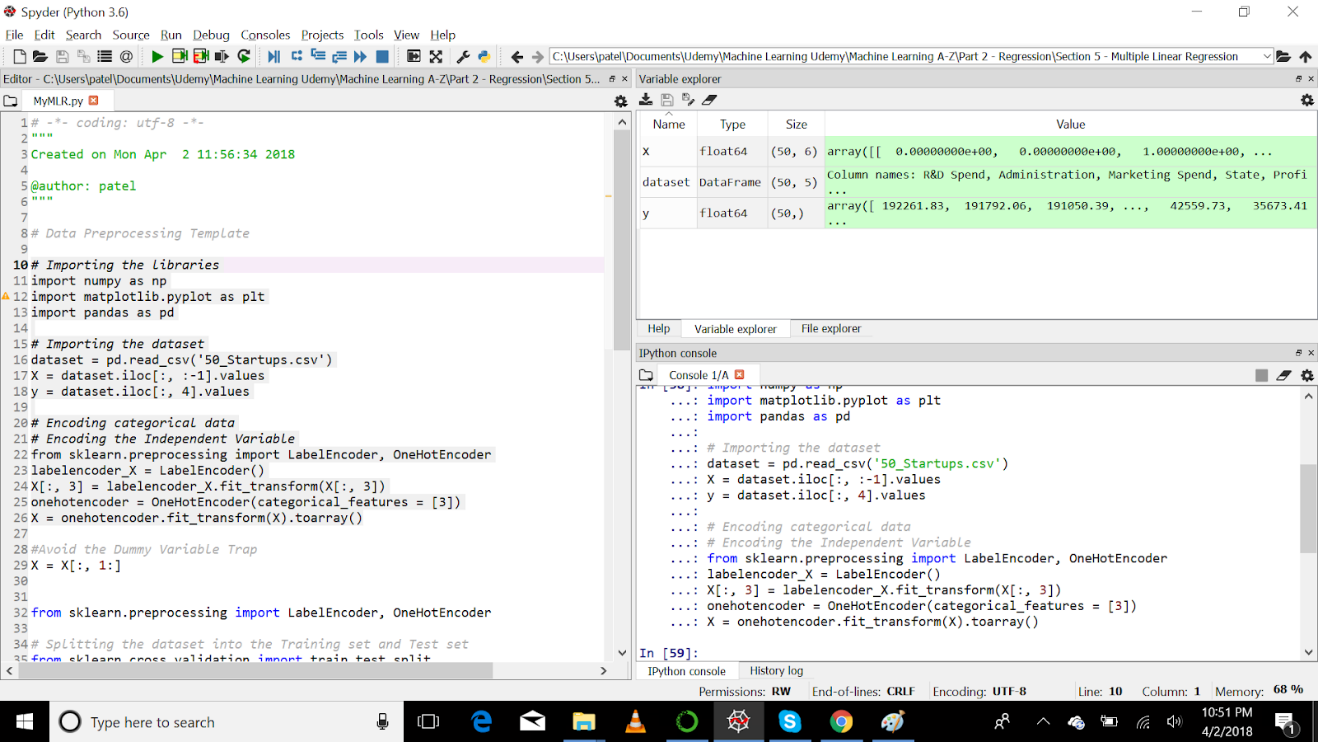
**Step-5:** Fit model without this variable and go back to **Step-3.**

**FINISH:** Your model is ready.

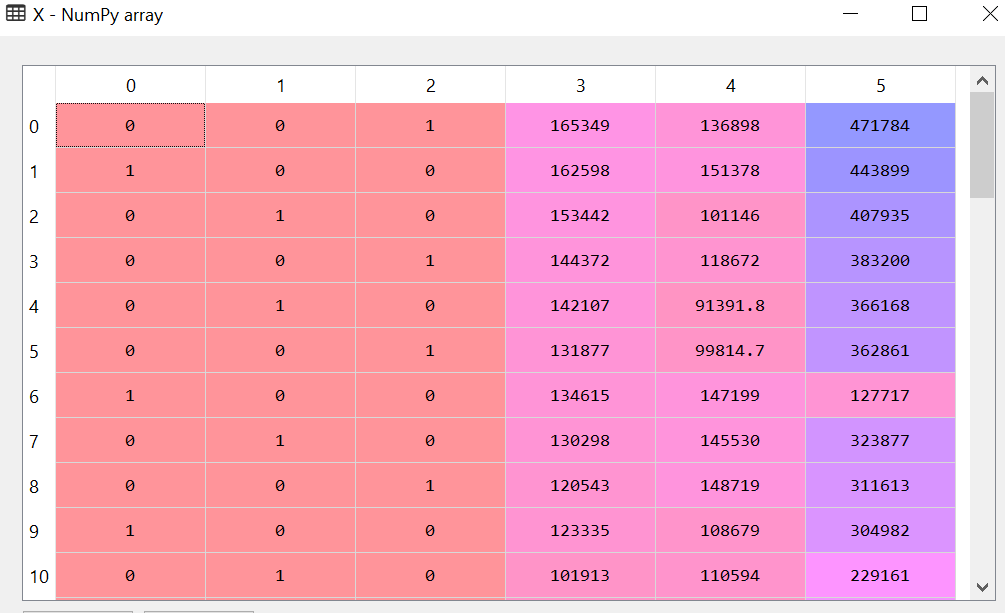
We can follow these steps to perform Backward Elimination for Multiple Linear Regression.

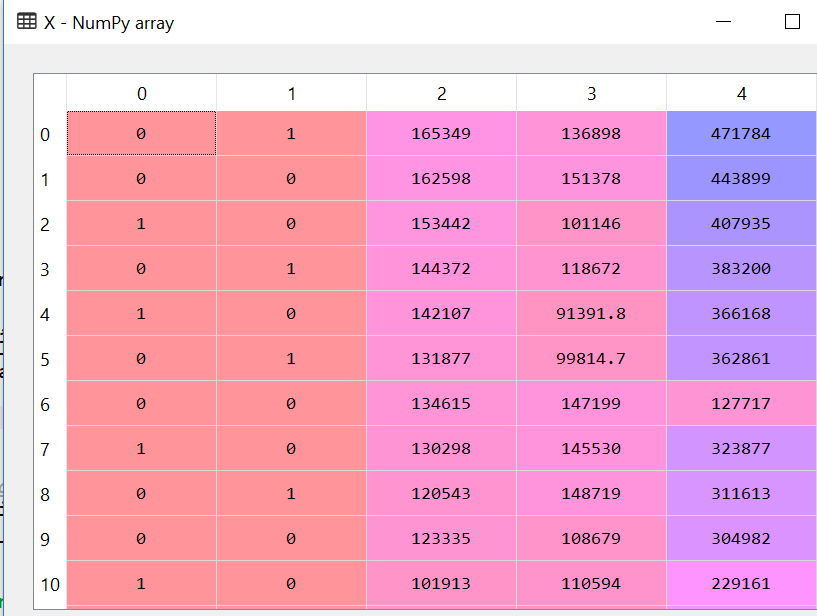
### Implementation:

Below mentioned are some screenshots of the code and the results we got while performing the Multiple Linear Regression with Backward Elimination.

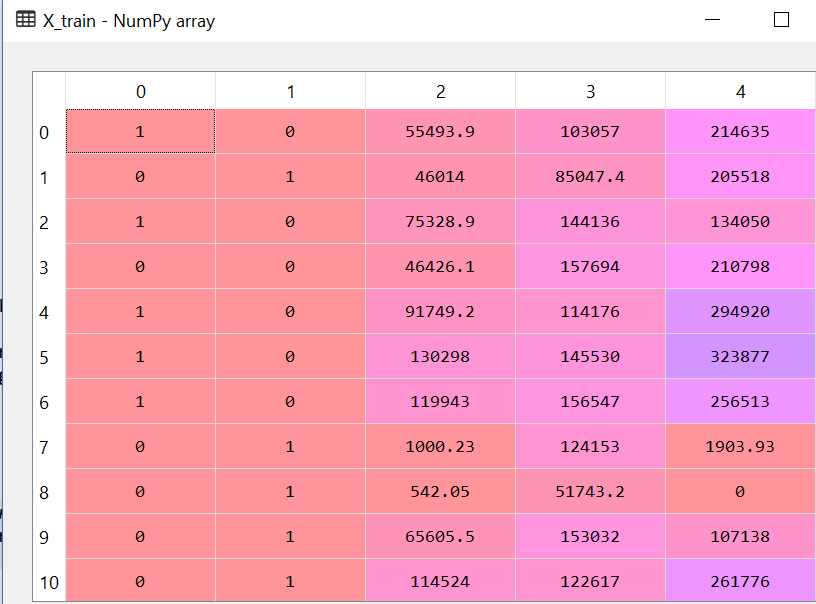


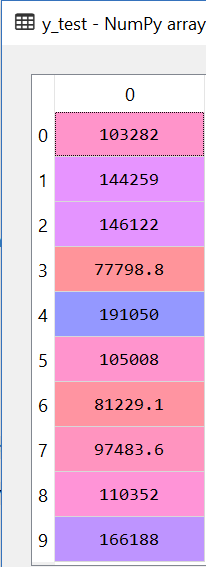
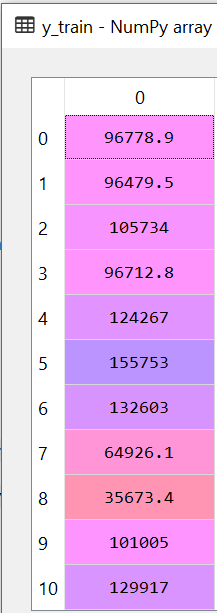
In above screenshot, we have imported dataset from CSV file and stored into two different Matrices which are X(Descriptive Features) and y(Target Features). The variable values can be seen on right side of the photo.

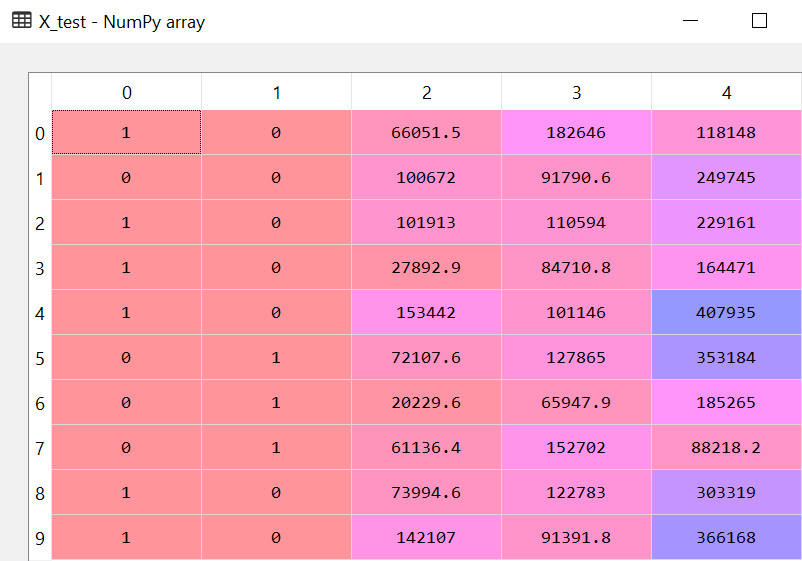




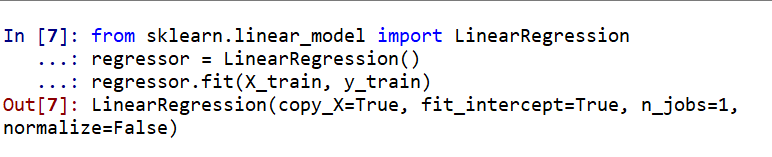
Above mentioned photos, one contains 5 columns and other contains 4 columns. The reason is that in first picture we converted State column into numeric column and then created dummy variable for all 3 states 1) California 2) Florida 3) New York. But, in the second picture, we can see that after creating dummy variable, we have dropped one dummy column because it causes redundancy problem as for example if we have two dummy variables then if you take any one then its meaning will be same as other column for example, if one column has data like {0,1,0,1} and {1,0,1,0} for California state and Florida State respectively. Then when we look at first column {0,1,0,1} we can say that, if there is 0 then it is Florida State and 1 California State. Hence it will be lead to error in calculation not in syntax error. So it is wise to drop one dummy variable column. That is what we have done in our implementation.



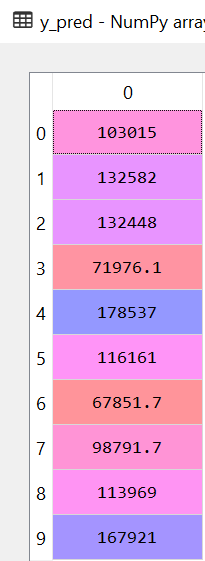




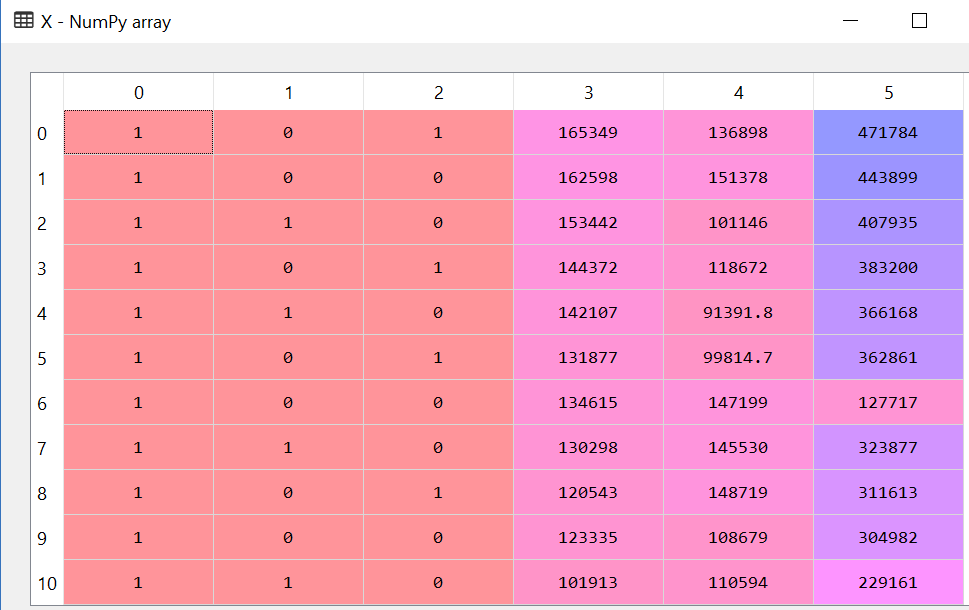
We can see the train and test dataset for our project. We have divided the dataset, giving 20% of data in testing and remaining 80% of data in training the model.

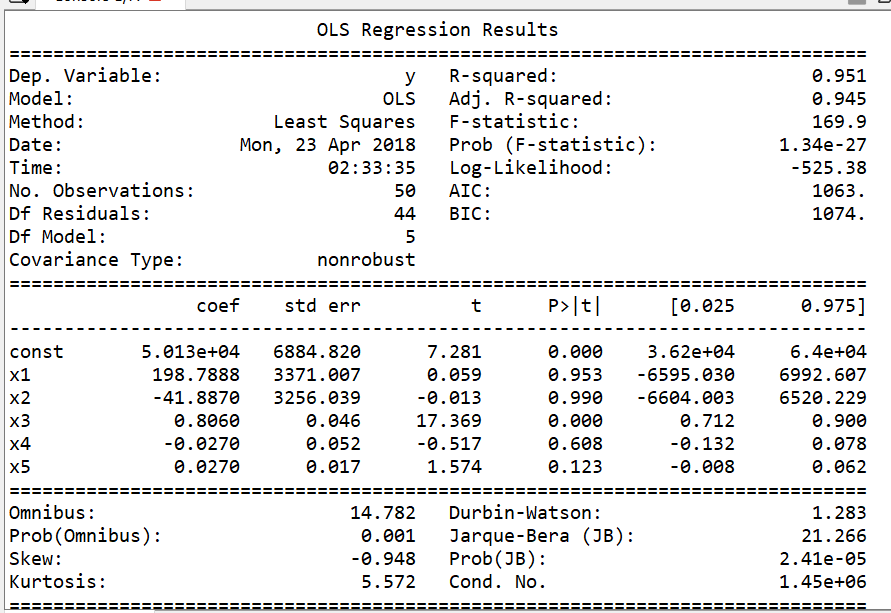


After splitting dataset, we can now train our model using X\_train and y\_train (descriptive features and target features) respectively. We are using here Linear Regression class which can also be used for Linear Regression model but in order to perform Linear Regression we just have to stop at next point where as for Multiple Linear Regression, we have to continue to work on our dataset.



Above mentioned are predicted values for X\_test matrix. This shows the values which are predicted by our model which was trained before by X\_train data. The predicted values are stored in a matrix called “y\_pred”.



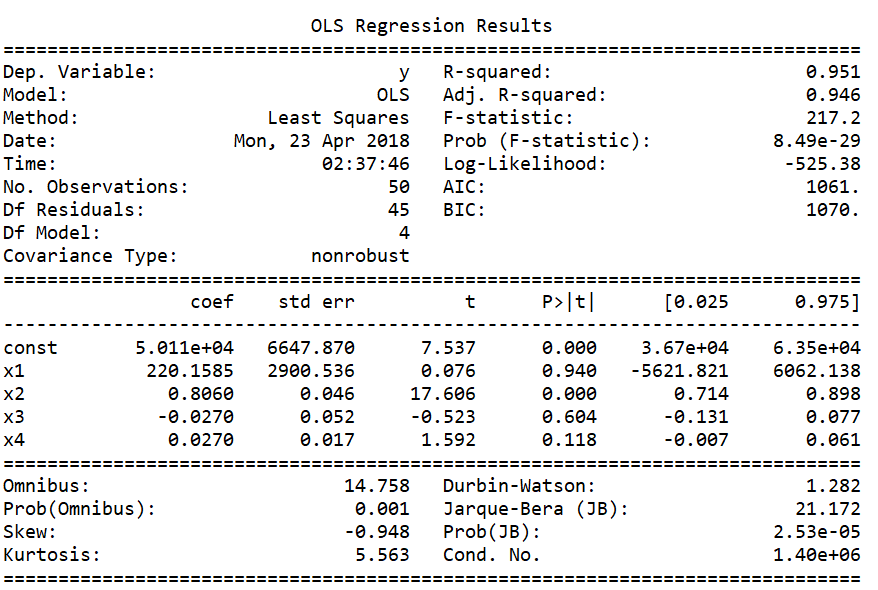


Here, we have used “statsmodels” api in order to first add the whole column of “1”. This is necessary as in the equation of Multiple Linear Regression

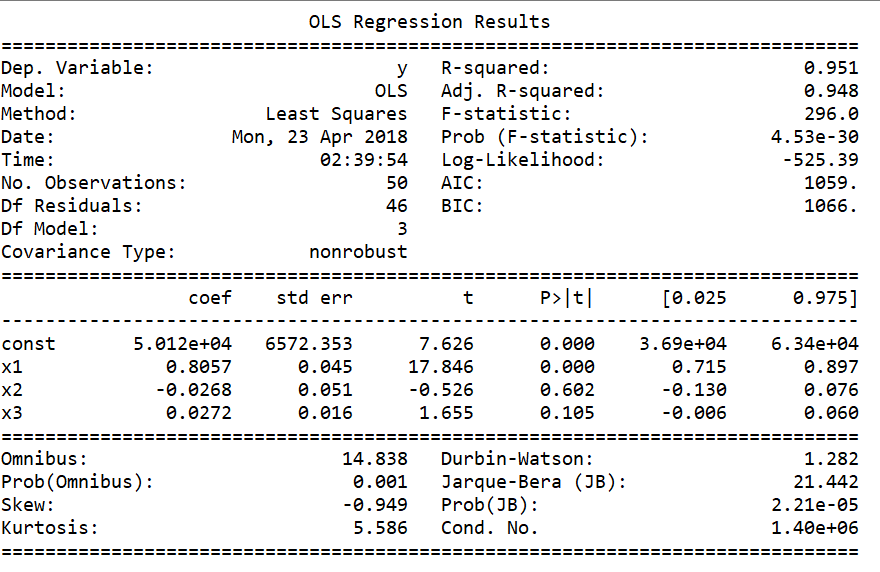
https://lh3.googleusercontent.com/1iBjPQ7LX24Dwp3pyrCxebBwNBjkjmjW0PddJZ3vQHajNh5zd8U2tfXu0NdUWX6Ld3ichd1q6MHpU5o7cY0gPmmI7H3Czpe0sshxemVoPY166pT4M1b6y3y04rZJKpefY3_iWWjsNR5iITE-CQ

B0 is the constant. Hence if a feature is selected whose value is 0 then whole equation becomes 0 by default. Hence, in order to prevent that situation, we have to add a whole column of “1” which will be considered as constant value if the rest of the part of the equation becomes “0” then we will still have the default value.

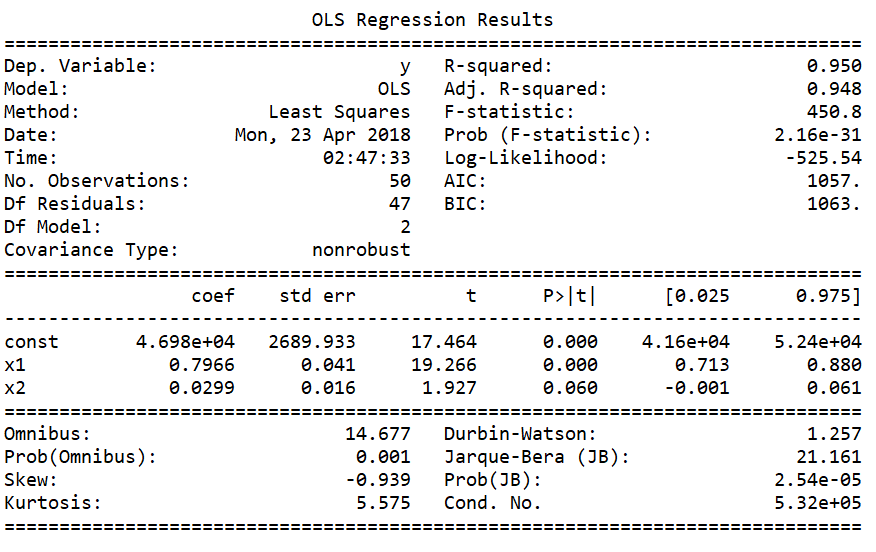
Then we are getting a new array “X\_opt” which consists of all columns existing in X matrix.  Then we take regressor object and call method OLS and give “y” as endog which is the target feature and “X\_opt” as the descriptive feature. Once this is done then call “summary()” method, so that we can know the p-value, t-value squared-R etc. All of these calculations are mentioned on right side of screenshot. The idea here is to calculate which column’s p-value is greater than Significant Value = 0.05 (for our project). If p-value is greater than 0.05 then in next step we have to remove that column from the “X\_opt” and again give that array as input in next step. This way we are removing the columns which have really less impact on predicting target value. Hence, the name “Backward Elimination”.



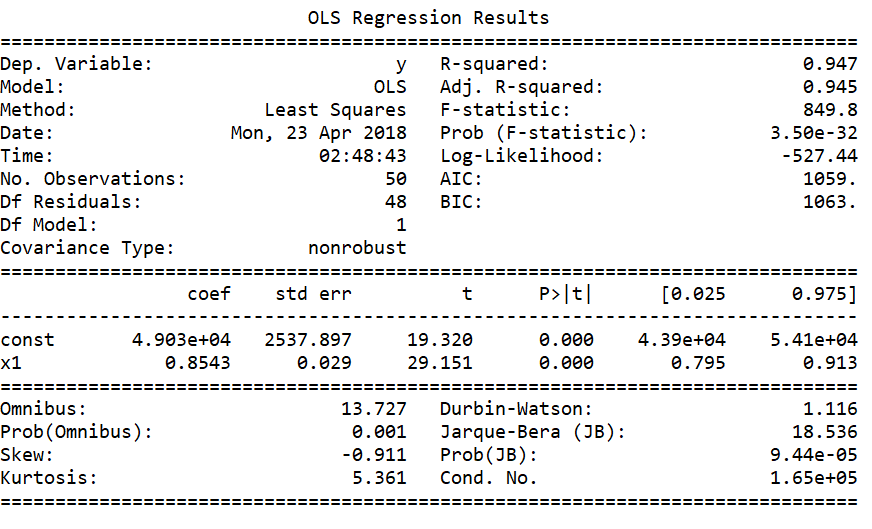
Backward Elimination step is performed here by removing “2nd column”. So we notice difference in above 2 photos, that column having 0.990 as its P value is not present in above mentioned set of results.



Backward Elimination step is performed here by removing “1st column”.



Backward Elimination step is performed here by removing “4th column”.



Backward Elimination step is performed here by removing “5th column”.

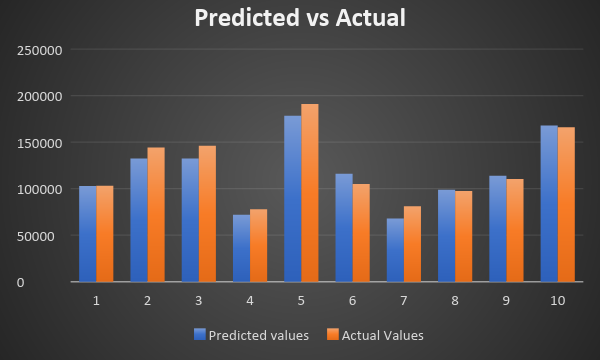
### Results:

Hence, at the end of this step, we can see the p-value of 3rd column is less than significant value (note that the p-value is 0.000 that doesn’t mean the value is 0 it means that the value is far less than 0 which cannot be described in 3 decimal points after 0). Hence, we can say that 3rd column has comparatively more impact on target feature (statistically) than the rest of the columns (descriptive features). As Multiple Linear Regression is the Error base learning, we can also see the error calculated for each set of columns in the above screenshots on right side.

Final step, for calculating the accuracy of the model, we have used R2 technique. After applying, we got R2 = 0.934706847328. We calculate that between two variables, y\_test (which is the actual values for X\_test matrix holding descriptive features) and y\_pred (which holds the value predicted by model for X\_test). From the R2 we can say that our model is predicting values closest to the actual values. Usually, the value of R must be between 0-1.

### Conclusion:

Here both Actual values and Predicted values are compared in chart. From the clustered chart we can see that the values are very close. So, model predicted values for test cases nearly as accurate as actual values. In below graph we can see that blue line is predicted values while orange is Actual values.



### **Individual Contributions**

|  |  |
| --- | --- |
| **William Clark** | Provided invaluable contributions to the completion of the tasks assigned to the group  Reviewed Report |

|  |  |
| --- | --- |
| **Sumati Kulkarni** | Provided invaluable contributions to the completion of the tasks assigned to the group  Drafted report  Reviewed Report |

|  |  |
| --- | --- |
| **Babak Maleki Shoja** | Provided invaluable contributions to the completion of the tasks assigned to the group  Implementation  Drafted Report  Reviewed Report |

|  |  |
| --- | --- |
| **Venkatesh Reedy Pala** | Provided invaluable contributions to the completion of the tasks assigned to the group  Drafted Report  Reviewed Report |

|  |  |
| --- | --- |
| **Vishwa Patel** | Provided invaluable contributions to the completion of the tasks assigned to the group for this project  Implementation  Drafted report  Reviewed Report |

### Team Summary

This assignment helped us get a better understanding of using Python to implement Multiple Linear Regression algorithm with Backward Elimination. As shown in the implementation section we were able to build Multiple Linear Regression algorithm with Backward Elimination in Python. Our team did a great job working and collaborating to successfully complete this project.