Analysis of startup companies’ profitability using multiple regression

Illinois Institute of Technology

Submission Date: 11/30/2022

MATH 564/484 Applied Statistics/Regression

Chao Yang cyang72@hawk.iit.edu

Albert Mandizha amandizha@hawk.iit.edu

Contribution: 50%-50%

**Abstract**

In this project, we applied multiple regression into the research on startup companies’ performance based on their profitability. Multiple regression is a type of supervised learning. We measured the performance of a startup company by predicting its profit, based on Research and Development Spending, Marketing spending, State and Administrative Costs. We clean and encode categorical data in the data processing stage, fit the data on the multiple linear regression model and perform lasso in the data analysis stage to perform feature selection. After training the model we deployed our model on test data and gave us a 98% accuracy, hence we concluded that multiple regression is a good fit on predicting the profit of startup companies. It could be established that companies that spend more on Marketing and Research and Development experienced a high return on profitability level.

Terminology: multiple regression, supervised learning, lasso

**Contents**

[1.0 Introduction 3](#_Toc120747027)

[1.2 Objectives of the Project 3](#_Toc120747028)

[1.2.1 Research Questions 3](#_Toc120747029)

[1.2.2 Hypothesis 3](#_Toc120747030)

[2.0 Data Collection and Preparation 3](#_Toc120747031)

[2.1 Data Source 3](#_Toc120747032)

[2.2 Data Preparation 3](#_Toc120747033)

[3.0 Methodology 4](#_Toc120747034)

[3.1 Introduction 4](#_Toc120747035)

[3.2. Step 1 Analyzing the correlation and direction of the Data 4](#_Toc120747036)

[3.3. Step 2 Model Estimation and Fitting 5](#_Toc120747037)

[3.3.1 Multiple linear regression formula 5](#_Toc120747038)

[3.2.2 Model Fitting 5](#_Toc120747039)

[3.2.3 Checking for Homoscedasticity 6](#_Toc120747040)

[3.2.4 Feature selection 6](#_Toc120747041)

[3.3 Step 3 Model Evaluation 6](#_Toc120747042)

[4.0 Analysis and Results 7](#_Toc120747043)

[4.1 Importing and Plotting the Dataset 7](#_Toc120747044)

[4.2 Data Histograms 7](#_Toc120747045)

[4.3 Density Plot 9](#_Toc120747046)

[4.4 Checking & Treating Outliers 10](#_Toc120747047)

[4.5 Correlation Analysis 11](#_Toc120747048)

[4.6 Model Fitting 12](#_Toc120747049)

[4.7 Hypothesis test on coefficients 12](#_Toc120747050)

[4.8 Feature Selection 13](#_Toc120747051)

[5.0 Model Accuracy 15](#_Toc120747052)

[6.0 Conclusion 15](#_Toc120747053)

[References 16](#_Toc120747054)

[Appendix 16](#_Toc120747055)

# 1.0 Introduction

As a subcategory of machine learning and artificial intelligence, supervised learning, also known as supervised machine learning, is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately.

The startup company is a newly formed business with particular momentum behind it based on perceived demand for its product or service. We would like to find a metric of measuring startup companies. There are a few metrics to track growth, customer acquisition cost, retention rate and customer lifetime revenue. In our idea, business should show its value in profit, so we decided to measure the startup company by its profit. From the dataset we collected, we treat profit as an independent variable, other variables will be counted as dependent variables. In supervised learning, multiple linear regression is compatible with our dataset so that we apply it in our analysis. Multiple regression, a type of supervised learning, focuses on using the independent variable whose values are known to predict the value of the single dependent variable.

In the data processing stage, we parse and encode categorical data, plot the variables and check the missing data, it shows our dataset is clean and neat. In the data analysis stage, we splitted the data for train and test sets, in order to enhance the prediction accuracy and interpretability for the fitted multiple regression, we perform lasso regression because it can perform both variable selection and regularization. In the model accuracy stage, we compared the accuracy of test sets and train sets.

## 1.2 Objectives of the Project

* To establish the relationship between Profit and its independent variables.
* To build a prediction regression model, that predicts the profitability of a new startup.

### 1.2.1 Research Questions

1. Is there a relationship between the dependent variable and independent variables?
2. Can the linear regression model predict the profitability of a new start-up?

### 1.2.2 Hypothesis

Is there a relationship between the dependent variable and independent variables?

* H0 There is no relationship between the independent and dependent variable.
* H1 There exists a relationship between at least one of the independent variable and the dependent variable

# 2.0 Data Collection and Preparation

## 2.1 Data Source

The data used in the project has been obtained from [www.kaggle.com/datasets/amineoumous/50-startups-data](http://www.kaggle.com/datasets/amineoumous/50-startups-data). (Amine Oumous 2020).

## 2.2 Data Preparation

The dataset that's we see here contains data about 50 startups. It has 5 columns: “R&D Spend”, “Administration”, “Marketing Spend”, “State”, “Profit”.

The first 3 columns indicate how much each startup spends on Research and Development, how much they spend on Marketing, and how much they spend on administration cost, the state column indicates which state the startup is based in, and the last column states the profit made by the startup (Amine Oumous 2020).

# 3.0 Methodology

## 3.1 Introduction

The paper used Multiple Linear regression to build a prediction model of profit which can be used by a startup to determine its expected profitability given its Research and development spending, Marketing Spending, State and Administration Costs. Inorder to use Multiple linear regression the following assumptions were made:

1. **Homogeneity of variance** (homoscedasticity): the size of the error in our prediction doesn’t change significantly across the values of the independent variable.
2. **Independence of observations**: the observations in the dataset were collected using statistically valid [sampling methods](https://www.scribbr.com/methodology/sampling-methods/), and there are no hidden relationships among variables.
3. **Correlation** was established between (Amine Oumous 2020)features: In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model. If two independent variables are too highly correlated (r2 > ~0.6), then only one of them should be used in the regression model.
4. **Normality:** The data follows a distribution. Histograms of each dependent variable will be done to test whether each variable follows a normal distribution.
5. **Linearity:** the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor.

## 3.2. Step 1 Analyzing the correlation and direction of the Data

First, a scatter plot was used to analyze the data and check for directionality and correlation of data**.** In our project we will use the scatter plot to determine the direction of data. The data to be determined is R.D Spending, State, Administration Costs, and Marketing Spending.

A correlation Matrix will be used to evaluate the relationship between independent variables. If the correlation value is greater than 0.6 between two dependent variables one of the variables will be dropped since they will be highly correlated.

## 3.3. Step 2 Model Estimation and Fitting

The second step of regression analysis is to fit the regression line. Mathematically least square estimation is used to minimize the unexplained residual (Uyanik 2013).The multiple regression formula is explained as below:

### 3.3.1 Multiple linear regression formula

According to ( (Kutner, et al. 2005) the formula for a multiple linear regression is:

y = {\beta_0} + {\beta_1{X_1}} + … + {{\beta_n{X_n}} + {\epsilon}

* y = the predicted value of the dependent variable
* B_0 = the y-intercept (value of y when all other parameters are set to 0)
* B_1X_1 = the regression coefficient (B_1) of the first independent variable (X_1) (a.k.a. the effect that increasing the value of the independent variable has on the predicted y value)
* … = do the same for however many independent variables you are testing
* B_nX_n = the regression coefficient of the last independent variable
* \epsilon = model error (also known as how much variation there is in our estimate of y)

To find the best-fit line for each independent variable, multiple linear regression calculates three things:

* The regression coefficients that lead to the smallest overall model error.
* The *t* statistic of the overall model.
* The associated [*p* value](https://www.scribbr.com/statistics/p-value/) (how likely it is that the *t* statistic would have occurred by chance if the [null hypothesis](https://www.scribbr.com/statistics/null-and-alternative-hypotheses/#definition) of no relationship between the independent and dependent variables was true).

It then calculates the *t* statistic and *p* value for each regression coefficient in the model.

### 3.2.2 Model Fitting

The data is split into test and training data sets. The model will be fitted using 80% of the data that is the training data from our dataset. After fitting the model on the training data and obtaining the regression model the model will then be deployed on the test data (20%) to assess model performance.

### 3.2.3 Checking for Homoscedasticity

We will check that our model is actually a good fit for the data, and that we don’t have large variation in the model error. The Residual Vs Fitted Plot, Q-Q plot, Scale location and residual vs leverage will help us check our model for good fit.

### 3.2.4 Feature selection

Lasso Regression is a type of linear regression that uses shrinkage. The project will use lasso regression to perform feature selection with the coefficient of irrelevant features being reduced to 0. Lasso performs regularization which is an important concept to avoid overfitting, in scenarios where the trained and test data are much varying.

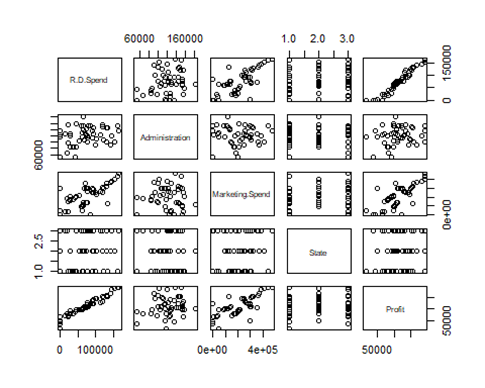
## 3.3 Step 3 Model Evaluation

The last step for the linear regression analysis is the test of significance and model accuracy. Linear regression uses two tests to test whether the found model and the estimated coefficients can be found in the general population the sample was drawn from. Firstly, the F-test will be used to test the overall model.

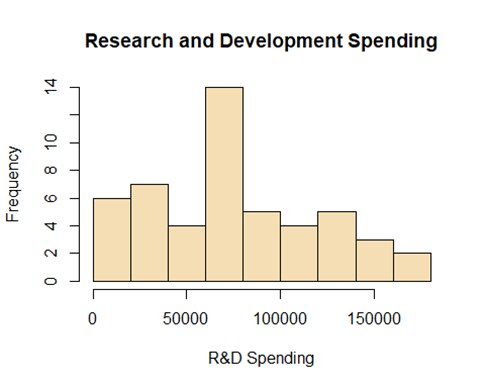
The null hypothesis is that the independent variables, Marketing Spend, R.D Spending, State and Administrative Costs have no influence on the dependent variable. In other words, the F-tests of the linear regression tests whether the R²=0. Secondly, multiple t-tests analyze the significance of each coefficient and the intercept. The t-test has the null hypothesis that the coefficient/intercept is zero.

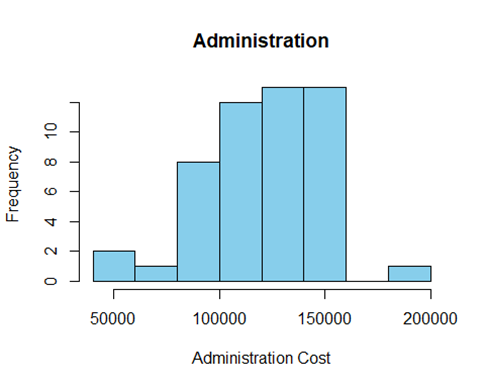
# 4.0 Analysis and Results

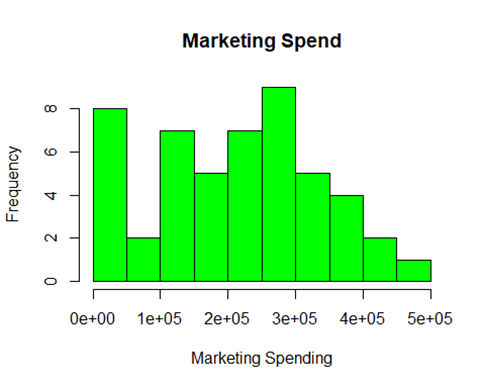
## 4.1 Importing and Plotting the Dataset

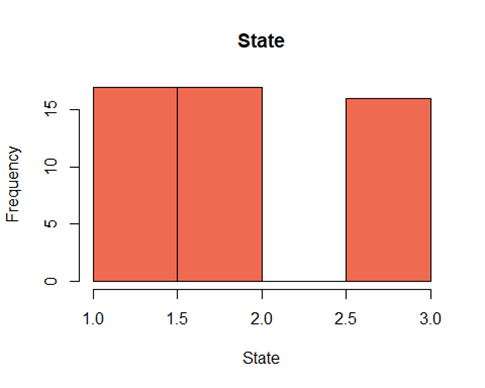
****

## 4.2 Data Histograms

****

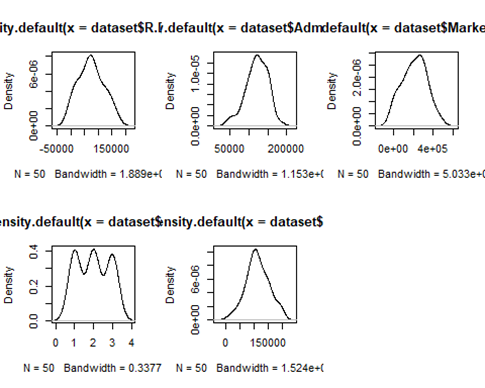
****

****

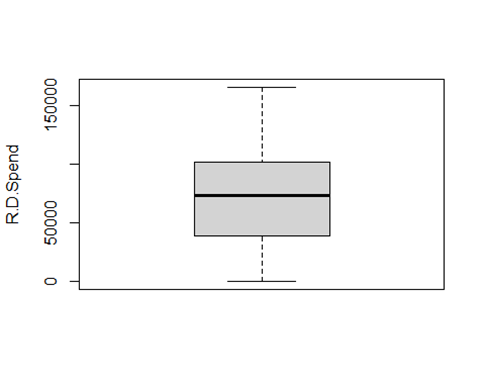
****

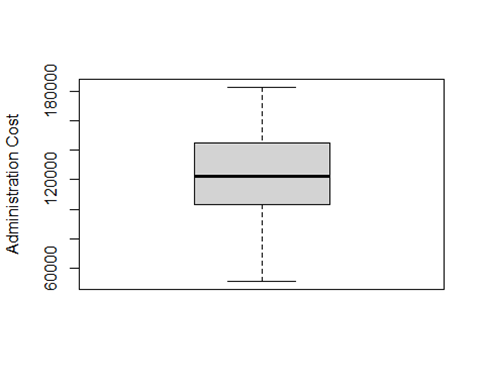
## 4.3 Density Plot

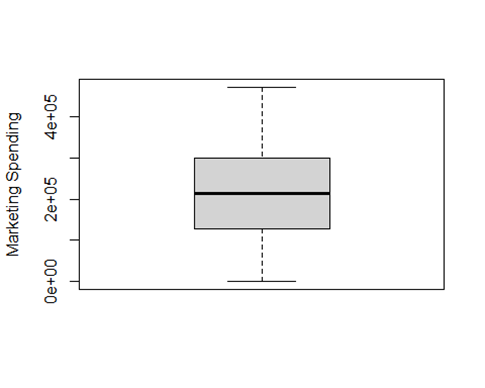
Density curves allow us to quickly see whether or not a graph is left skewed, right skewed, or has no skew. We see there are no features that have no skew.



## 4.4 Checking & Treating Outliers

****

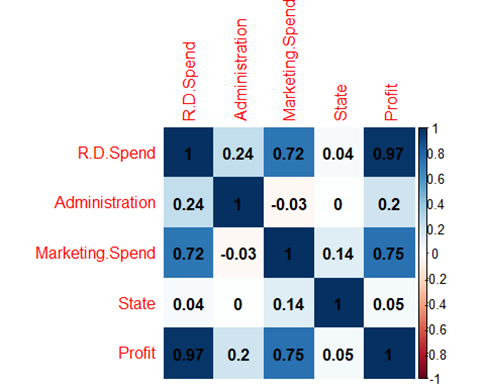
****

****

No Outliers can be noted from the above data provided. This mean that all of our features do not have any outliers.

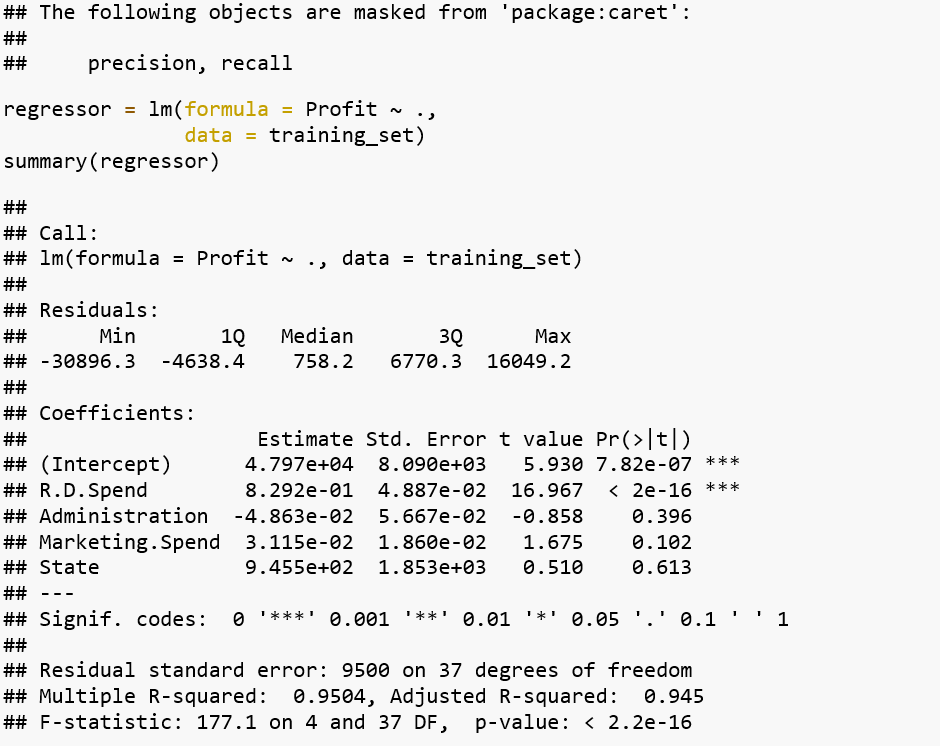
## 4.5 Correlation Analysis

In the data processing stage, correlation plots show that R&D spend has a high correlation with profit, which means for the startup company who spends more on research and development, they will gain more advantage for their product in the market competition, that’s benefit to profit as a feedback. In the perspective of correlation, we found that marketing spend has high correlation with profit too, which corresponds to the current market situation, a product with a high exposure rate, has high probability to be sold and hence profit can go up.



## 4.6 Model Fitting

To further study the relationship between profit and other dependent variables, we fit the data into multiple regression in the analysis stage. From the fitted linear regression, R&D spend has a significant value in its t-value which means the R&D spend has a great impact on the profit. Marketing spend has an impact on the profit too, however, it’s not as great as R&D spend.



## 4.7 Hypothesis test on coefficients

**4.7.1 Hypothesis**

* H0: Bi = 0, i =0,1,2,3,4
* Ha: Bi != 0

**4.7.2 Decision Criterion**

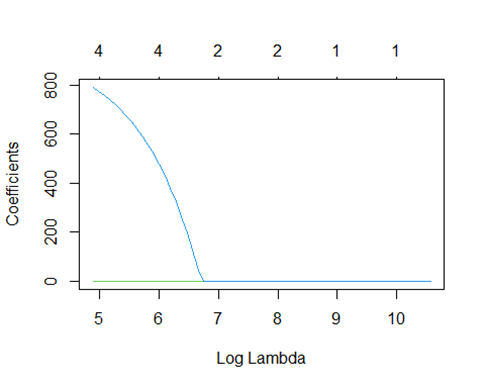
* Reject Ha fs F0 > F (0.05,4,37)
* F critical = qf(p=.05, df1=4, df=37, lower.tail=FALSE)
* F\_statistic = 177.1
* F critical < F statistic

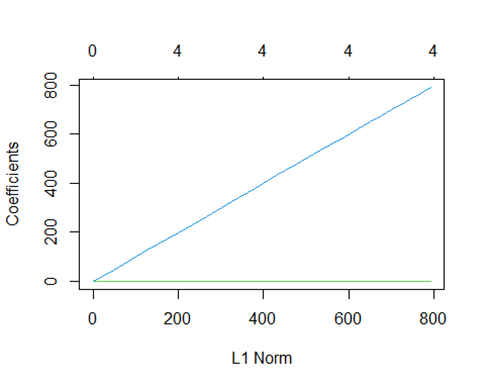
**4.7.3 Conclusion**

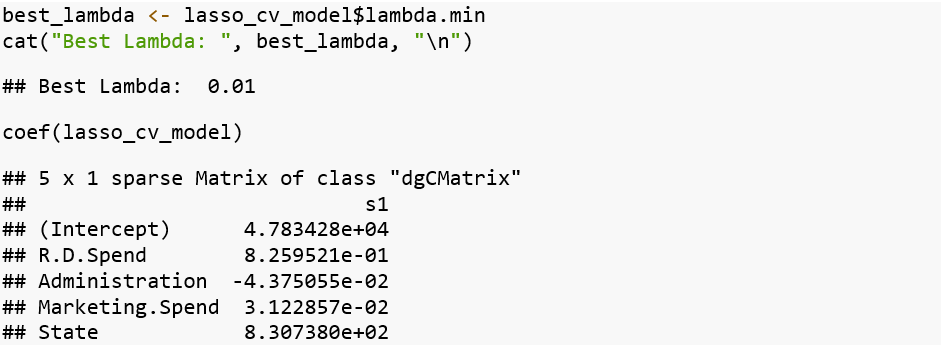
Hence we reject H0, which means there are relationships between independent variables and dependent variables. From the calculations of the F ratio it can be established that there is a relationship between Profit and its independent variables.

## 4.8 Feature Selection

Lasso regression was conducted and all features were selected as none of the features coefficient was reduced to Zero. This shows the significance of all the features in the model. Lasso model was performed on both the training and test data. The lasso used the shrinkage approach. From Lasso it can be concluded that all the features that are R and D Spending, Administration Costs, State and Marketing Spending are important features in explaining the variance in Profitability of start-up firms.

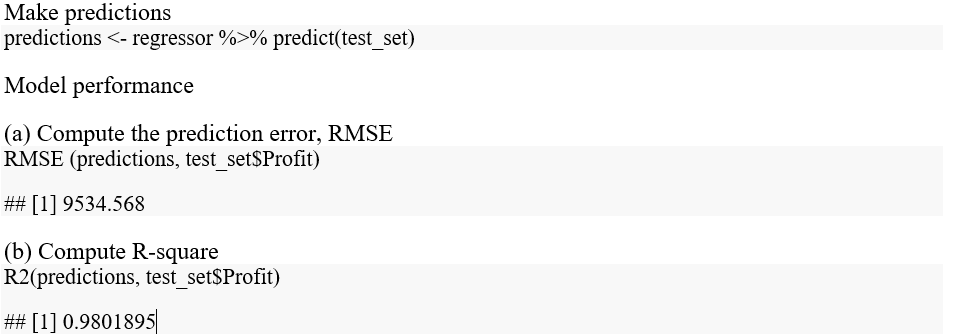






The resulting model has coefficients for all predictor variables. All the coefficients for other predictors got smaller with the exception of the Intercept. This was expected as lasso performs feature selection through shrinking irrelevant coefficients to 0

# 5.0 Model Accuracy



From the output above, the R2 is 0.98, meaning that the observed and the predicted outcome values are highly correlated, which is very good.

The prediction error RMSE is 9534.568, representing an error rate of 9534.568/mean(test\_set$Profit) = 8.3%, which is good.

# 6.0 Conclusion

**6.1 Current Model**

* R&D spend and Advertising are important factors for improving a startup companies’ profitability. Especially increasing of R&D spend can help the company achieve better performance in profit
* Lasso regression analysis is helpful when we need to perform both variable selection and regularization, with shrinkage of coefficients to 0, all the coefficients for other predictors get smaller as expected.
* According to our R squared value, 98% of the variation in the output variable is explained by the input variables which prove that our data model makes a good prediction

**6.1. Future work**

In terms of future work the project may look into increasing the number of features and also consider other independent variables that may help explain the uncaptured variance by the model.

# References

Amine Oumous . 2020. *Kaggle 50 Startups Data.* September 28. Accessed November 22, 2022. https://www.kaggle.com/datasets/amineoumous/50-startups-data.

Fang, Tingting, and Risto Lahdelma. 2016. "Evaluation of a multiple linear regression model and SARIMA model in forecasting heat demand for district heating system." *Applied Energy* 544-552.

Kutner, Michael H., Christopher J. Nachtsheim, John Neter, and William Li. 2005. *Applied Linear Statistical Models.* New York: Library of Congress Cataloging-in-Publication Data.

Uyanik, Gulden Kaya. 2013. "A Study on multiple linear regression analysis." *Procedia-Social and Behavioural Sciences* 234-240.

# Appendix

Git Hub Link to project files.

<https://github.com/cyangIIT/applied_stat_project>