GRADUATION ADMISSION PREDICTION BY MULTIPLE LINEAR REGRESSION

Doan Viet Anh | BDS21DON043 | Statistical Analysis

Contents

I. Understanding data

1. Introduction

“How can I get into graduate school?” This must be an intriguing question for students who want to pursue their postgraduate degrees (Master’s and doctorate programs) in top institutions across the world. Certainly, there are various factors, which need to be considered when applying such as GRE scores, TOEFL scores, research experience and so on. Therefore, with the aim to help students in shortlisting universities, this report focuses on predicting the probability of getting admission based on their profiles. The method for this problem is multiple linear regression and the language is R.

The dataset for this problem contains the following attributes.

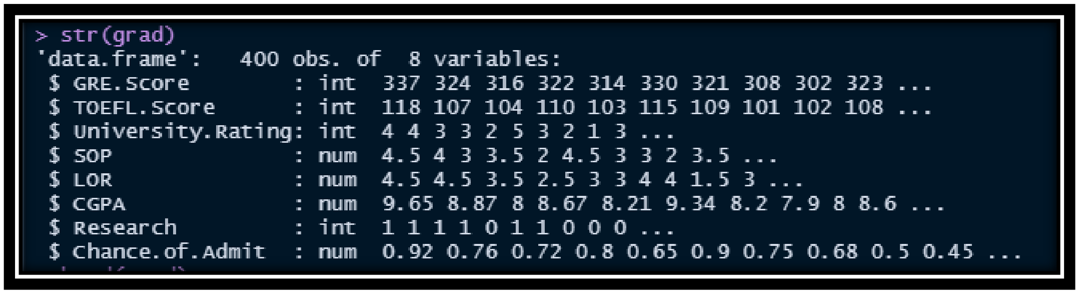
1. GRE Scores (out of 340)
2. TOEFL Scores (out of 120)
3. University Rating (out of 5)
4. Statement of Purpose and Letter of Recommendation Strength (out of 5)
5. Undergraduate GPA (out of 10)
6. Research Experience (either 0 or 1)
7. **Chance of Admit** (ranging from 0 to 1)

Chance of Admit is the response variable and other attributes are independent variables.

2. Overview of data

There are 400 observations and 8 variables. Most of the variables are in integer or numeric as such will not have to be concerned with factor variables. Also, from the null value test in R, there is no missing value in the dataset.

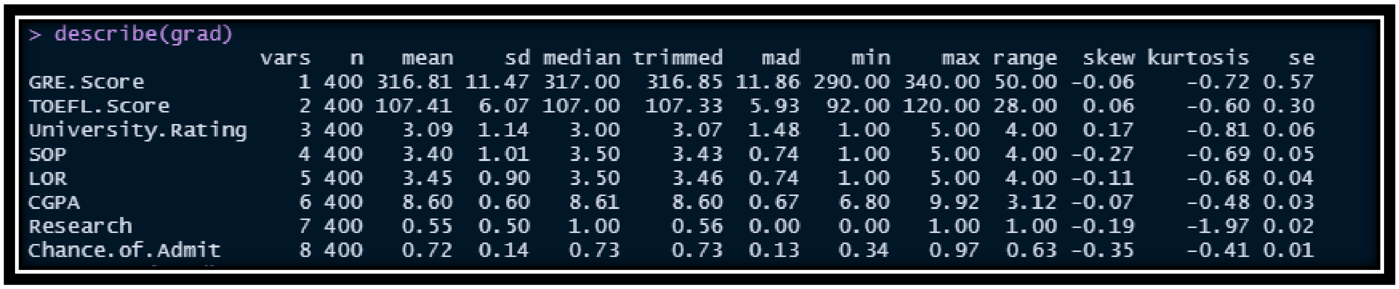
Note: the data frame of the data is called ‘grad’.



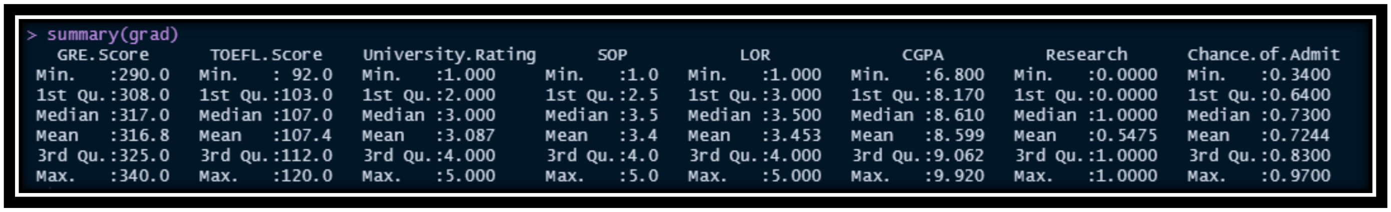
```
> str(grad)
'data.frame':   400 obs. of  8 variables:
 $ GRE.Score        : int  337 324 316 322 314 330 321 308 302 323 ...
 $ TOEFL.Score      : int  118 107 104 110 103 115 109 101 102 108 ...
 $ University.Rating: int  4 4 3 3 2 5 3 2 1 3 ...
 $ SOP              : num  4.5 4 3 3.5 2 4.5 3 3 2 3.5 ...
 $ LOR              : num  4.5 4.5 3.5 2.5 3 3 4 4 1.5 3 ...
 $ CGPA             : num  9.65 8.87 8 8.67 8.21 9.34 8.2 7.9 8 8.6 ...
 $ Research         : int  1 1 1 1 0 1 1 0 0 0 ...
 $ Chance.of.Admit  : num  0.92 0.76 0.72 0.8 0.65 0.9 0.75 0.68 0.5 0.45 ...
```

3. Statistic for data

Let us find the descriptive statistic for the data



```
> describe(grad)
                  vars   n   mean    sd median trimmed   mad    min    max range  skew kurtosis   se
GRE.Score            1 400 316.81 11.47 317.00  316.85 11.86 290.00 340.00 50.00 -0.06    -0.72 0.57
TOEFL.Score          2 400 107.41  6.07 107.00  107.33  5.93  92.00 120.00 28.00  0.06    -0.60 0.30
University.Rating    3 400   3.09  1.14   3.00    3.07  1.48   1.00   5.00  4.00  0.17    -0.81 0.06
SOP                  4 400   3.40  1.01   3.50    3.43  0.74   1.00   5.00  4.00 -0.27    -0.69 0.05
LOR                  5 400   3.45  0.90   3.50    3.46  0.74   1.00   5.00  4.00 -0.11    -0.68 0.04
CGPA                 6 400   8.60  0.60   8.61    8.60  0.67   6.80   9.92  3.12 -0.07    -0.48 0.03
Research             7 400   0.55  0.50   1.00    0.56  0.00   0.00   1.00  1.00 -0.19    -1.97 0.02
Chance.of.Admit      8 400   0.72  0.14   0.73    0.73  0.13   0.34   0.97  0.63 -0.35    -0.41 0.01
```



```
> summary(grad)
   GRE.Score      TOEFL.Score    University.Rating      SOP            LOR             CGPA          Research        Chance.of.Admit
 Min.   :290.0   Min.   : 92.0   Min.   :1.000     Min.   :1.0   Min.   :1.000   Min.   :6.800   Min.   :0.0000   Min.   :0.3400
 1st Qu.:308.0   1st Qu.:103.0   1st Qu.:2.000     1st Qu.:2.5   1st Qu.:3.000   1st Qu.:8.170   1st Qu.:0.0000   1st Qu.:0.6400
 Median :317.0   Median :107.0   Median :3.000     Median :3.5   Median :3.500   Median :8.610   Median :1.0000   Median :0.7300
 Mean   :316.8   Mean   :107.4   Mean   :3.087     Mean   :3.4   Mean   :3.453   Mean   :8.599   Mean   :0.5475   Mean   :0.7244
 3rd Qu.:325.0   3rd Qu.:112.0   3rd Qu.:4.000     3rd Qu.:4.0   3rd Qu.:4.000   3rd Qu.:9.062   3rd Qu.:1.0000   3rd Qu.:0.8300
 Max.   :340.0   Max.   :120.0   Max.   :5.000     Max.   :5.0   Max.   :5.000   Max.   :9.920   Max.   :1.0000   Max.   :0.9700
```

From the above table, we can see that:

- Regarding the response variable (Chance.of.Admit):

  1. The skewness is negative (= -0.35)

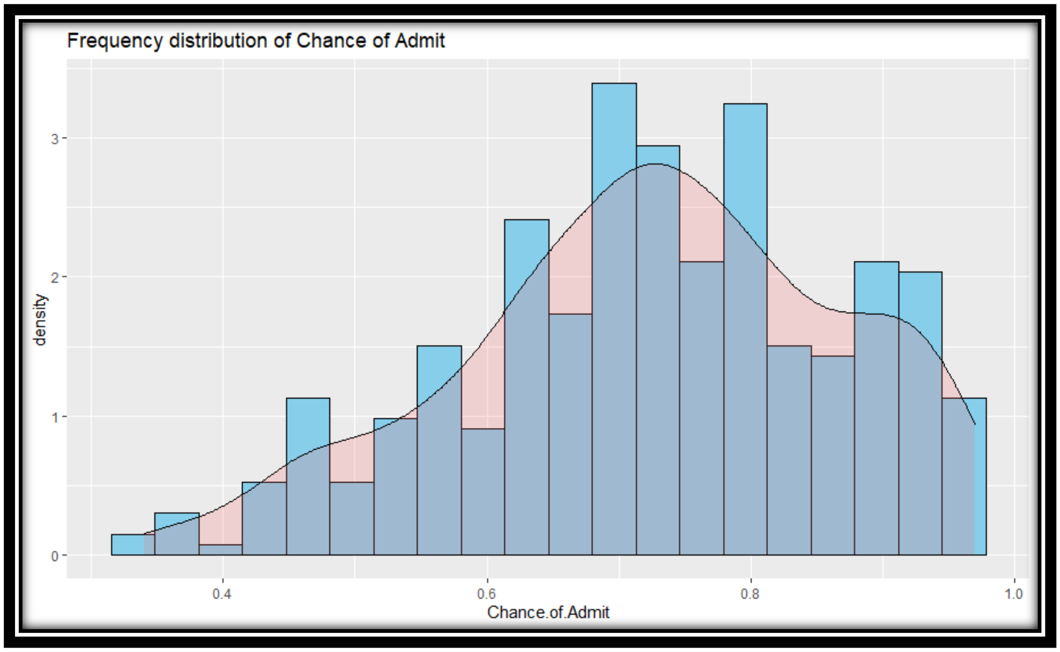
  2. Mean = 0.7244 < Median = 0.7300

- These indicate that the distribution of response variable is negatively skewed. We will approach this issue later in this report.

4. Visualizing response variable’s distribution

Frequency distribution of response variable (y)

Again, the graph below clearly indicates the negative skewness of variable y’s distribution.



II. Regression Analysis

A multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \varepsilon$$

♦ For Hypothesis testing and the setting of confidence limits, we also assume that ε is normally distributed.

♦ 4 Assumptions of linear regression

1. Linear relationship between Xs, and y
2. Independence: The residuals are independence
3. Homoscedasticity: The residuals have constant variance at every level of x
4. Normality: The residuals follow normal distribution

1. Correlation

a. Definition

Correlation coefficient between two random variables X and Y, usually denoted by r_{XY} is a numerical measure of linear relationship between them and is defined as:

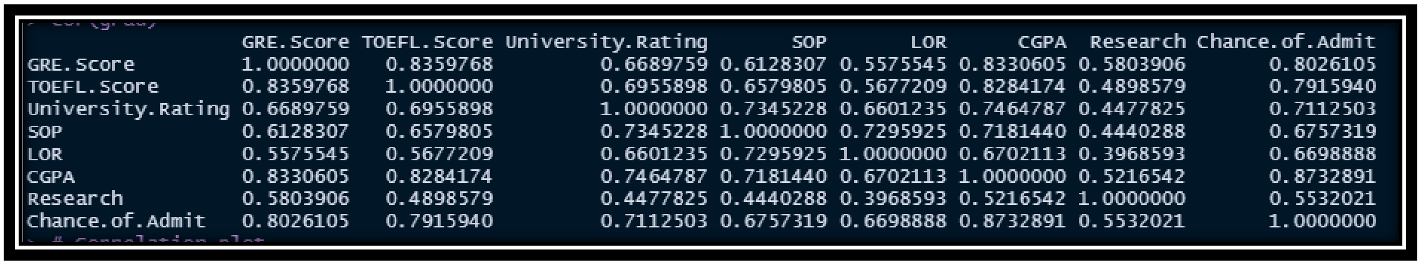
$$r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- r_{XY} provided a measure of linear relationship between X and Y.

- It is a measure of degree of relationship

b. Correlation Matrix

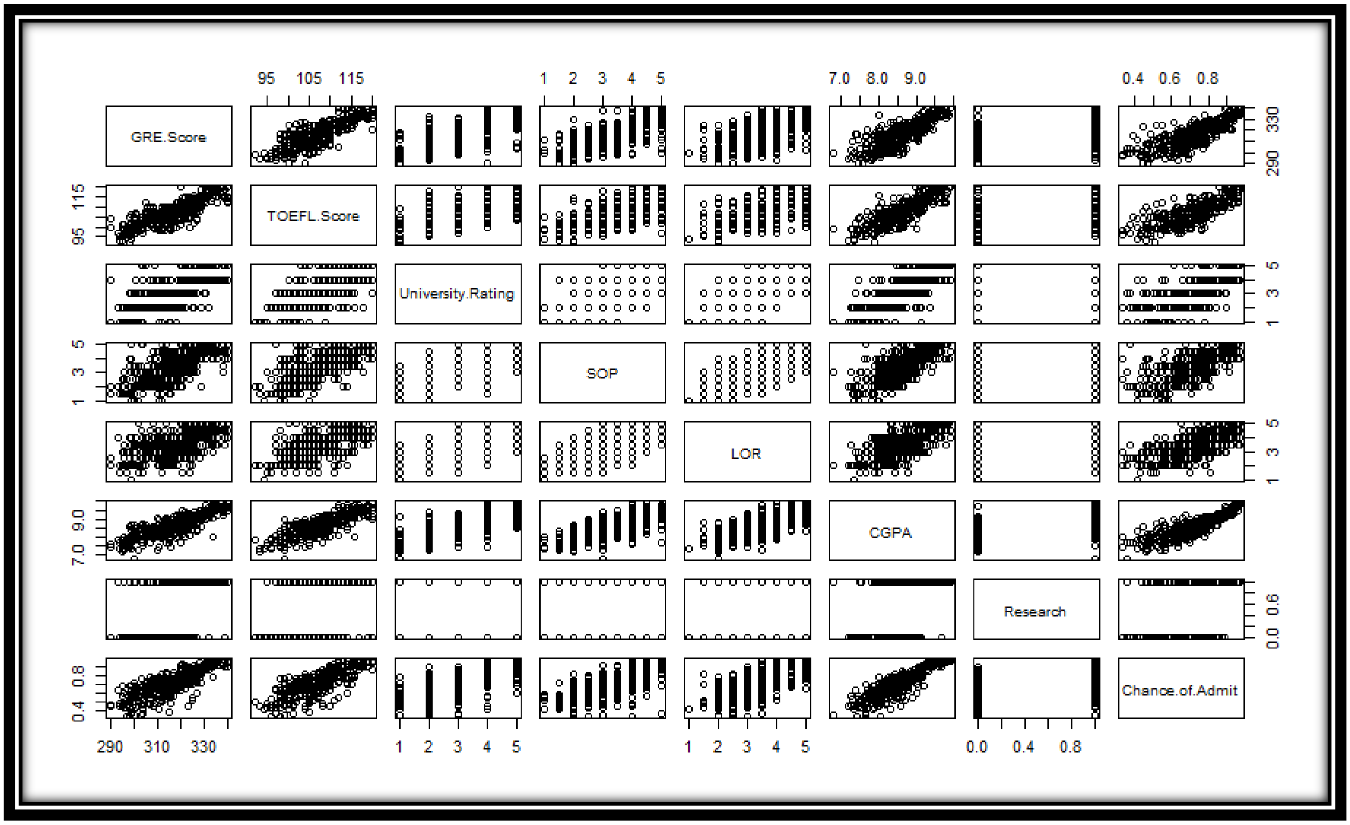
Correlation matrix is a table which simply displays correlations among variables. Let us look at the Chance.of.Admit column, we can see that most of the variables are in high correlation (near 1) with the response variable. This indicates that linear regression model is suitable for this problem.



```
                  GRE.Score TOEFL.Score University.Rating       SOP       LOR      CGPA  Research Chance.of.Admit
GRE.Score         1.0000000   0.8359768         0.6689759 0.6128307 0.5575545 0.8330605 0.5803906       0.8026105
TOEFL.Score       0.8359768   1.0000000         0.6955898 0.6579805 0.5677209 0.8284174 0.4898579       0.7915940
University.Rating 0.6689759   0.6955898         1.0000000 0.7345228 0.6601235 0.7464787 0.4477825       0.7112503
SOP               0.6128307   0.6579805         0.7345228 1.0000000 0.7295925 0.7181440 0.4440288       0.6757319
LOR               0.5575545   0.5677209         0.6601235 0.7295925 1.0000000 0.6702113 0.3968593       0.6698888
CGPA              0.8330605   0.8284174         0.7464787 0.7181440 0.6702113 1.0000000 0.5216542       0.8732891
Research          0.5803906   0.4898579         0.4477825 0.4440288 0.3968593 0.5216542 1.0000000       0.5532021
Chance.of.Admit   0.8026105   0.7915940         0.7112503 0.6757319 0.6698888 0.8732891 0.5532021       1.0000000
```

c. Correlation Plot

Correlation plot helps us to visualize correlation between variables.



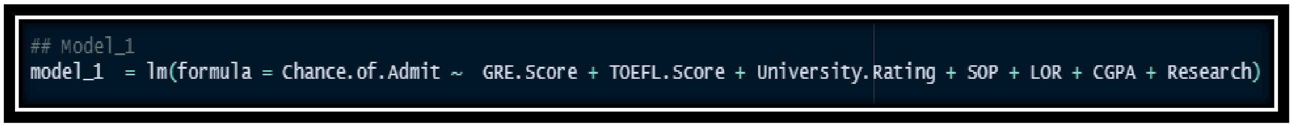
d. Summary

The last column in both correlation matrix and correlation plot shows that there are high correlations between response variable and each independent variable. This indicates that the multiple linear regression model should be a great fit for this problem. However, there could be some multicollinearity between independent variables. We will tackle this issue later in the report.

2. Model 1 – A simple approach

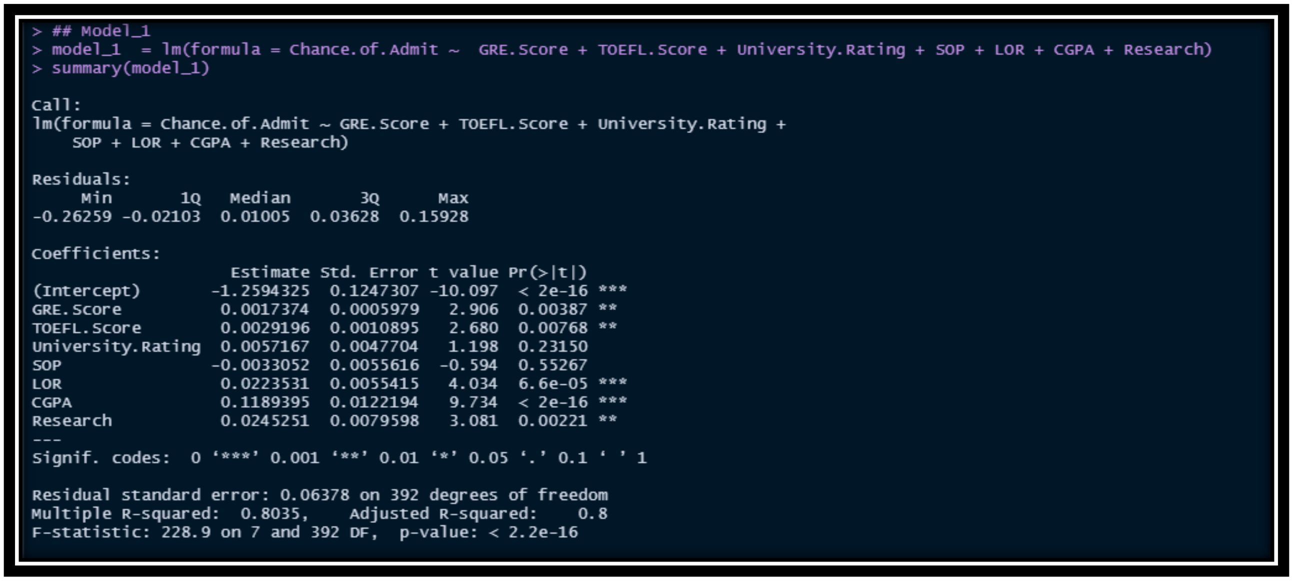
a. Model evaluation

In the first model, we take all the independent variables into consideration as the code below:



```
## Model_1
model_1  = lm(formula = Chance.of.Admit ~  GRE.Score + TOEFL.Score + University.Rating + SOP + LOR + CGPA + Research)
```

This is the evaluation of the model 1:



```
> ## Model_1
> model_1  = lm(formula = Chance.of.Admit ~  GRE.Score + TOEFL.Score + University.Rating + SOP + LOR + CGPA + Research)
> summary(model_1)

Call:
lm(formula = Chance.of.Admit ~ GRE.Score + TOEFL.Score + University.Rating +
    SOP + LOR + CGPA + Research)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.26259 -0.02103  0.01005  0.03628  0.15928 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -1.2594325  0.1247307 -10.097  < 2e-16 ***
GRE.Score          0.0017374  0.0005979   2.906  0.00387 ** 
TOEFL.Score        0.0029196  0.0010895   2.680  0.00768 ** 
University.Rating  0.0057167  0.0047704   1.198  0.23150    
SOP               -0.0033052  0.0055616  -0.594  0.55267    
LOR                0.0223531  0.0055415   4.034  6.6e-05 ***
CGPA               0.1189395  0.0122194   9.734  < 2e-16 ***
Research           0.0245251  0.0079598   3.081  0.00221 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06378 on 392 degrees of freedom
Multiple R-squared:  0.8035,    Adjusted R-squared:    0.8 
F-statistic: 228.9 on 7 and 392 DF,  p-value: < 2.2e-16
```

Criteria: p-value < 0.05

According to the output, let us look at the last column, the independent variables University.Rating and SOP shows p-value > 0.05, which indicates that they have less or no impact on the output. In other words, we should take these 2 attributes out of the model for further analysis.

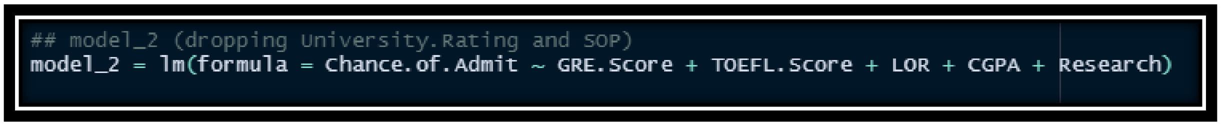
Criteria: r-squared > 0.6

The multiple R-squared of model 1 is 0.8035. This means 80.35% of the variation in the dependent variable can be explained by independent variables.

3. Model 2 – Removing insignificant attributes.

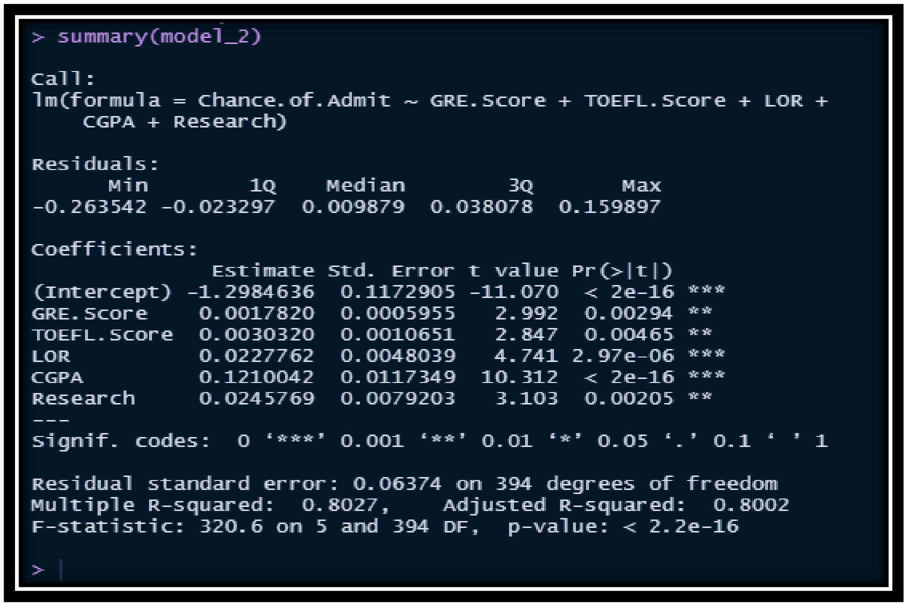
a. Model evaluation

We continue the analysis with model 2 with unimportant attributes are removed. This is followed by the code below.



```
## model_2 (dropping University.Rating and SOP)
model_2 = lm(formula = Chance.of.Admit ~ GRE.Score + TOEFL.Score + LOR + CGPA + Research)
```

This is the evaluation of model 2:



```
> summary(model_2)

Call:
lm(formula = Chance.of.Admit ~ GRE.Score + TOEFL.Score + LOR + 
    CGPA + Research)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.263542 -0.023297  0.009879  0.038078  0.159897 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.2984636  0.1172905 -11.070  < 2e-16 ***
GRE.Score    0.0017820  0.0005955   2.992  0.00294 ** 
TOEFL.Score  0.0030320  0.0010651   2.847  0.00465 ** 
LOR          0.0227762  0.0048039   4.741 2.97e-06 ***
CGPA         0.1210042  0.0117349  10.312  < 2e-16 ***
Research     0.0245769  0.0079203   3.103  0.00205 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06374 on 394 degrees of freedom
Multiple R-squared:  0.8027,	Adjusted R-squared:  0.8002 
F-statistic: 320.6 on 5 and 394 DF,  p-value: < 2.2e-16

> 
```

Criteria: p-value < 0.05

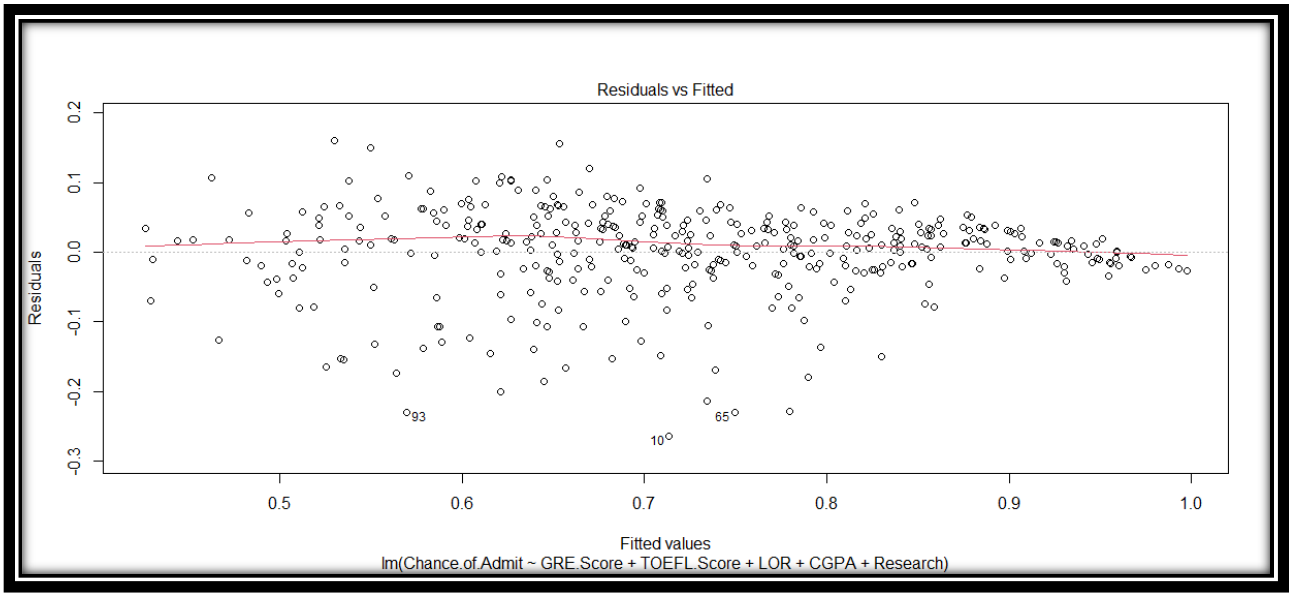
All of the independent variables have p-value <0.05

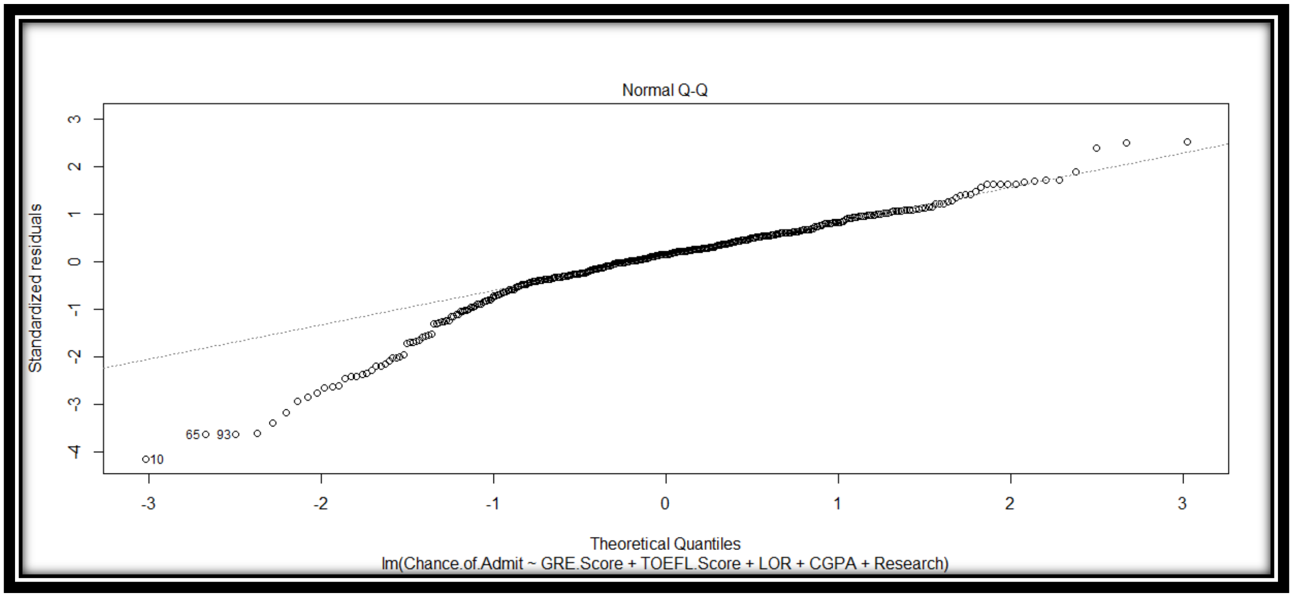
Criteria: R-squared > 0.6

The R-squared of model 2 is roughly same compared to that of model 1.

b. Residual Analysis

After fitting the model, we check residual vs fitted plots and Normal QQ plot to see if linear regression assumptions are satisfied.





Inference: From both residual vs fitted plot and Normal QQ plot, we see the assumptions that variance of residual is constant and the data is linearly distributed are correct. However, there are rooms for improvement:

1. The red line can be closer to the dot line standardized residual y = 0
2. There are some outliers in the Normal Q-Q plot

We will address this problem in the final model.

4. Model 3 - Removing multicollinearity using VIF

a. Variance Inflation Factor

Measures the correlation between each independent variable with other independent variables

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2} = \frac{1}{\mathrm{Tolerance}}$$

Where R_i is the coefficient for regressing x_i on other x’s

Criteria: VIF > 5 indicates multicollinearity.

b. Stepwise Regression

Method: Removing highly correlated variable – Stepwise Regression

Approach

- A null model is developed without any predictor variable x. In null model, the predicted value will be the overall mean of y.
- Then predictor variables x’s are added to the model sequentially.
- After adding each new variable, the method also removes any variable that no longer provide an improvement in the model fit.

$$\boxed{AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)}$$

Where:

n: number of observations

sigmoid_hat: estimate of error or residual variance

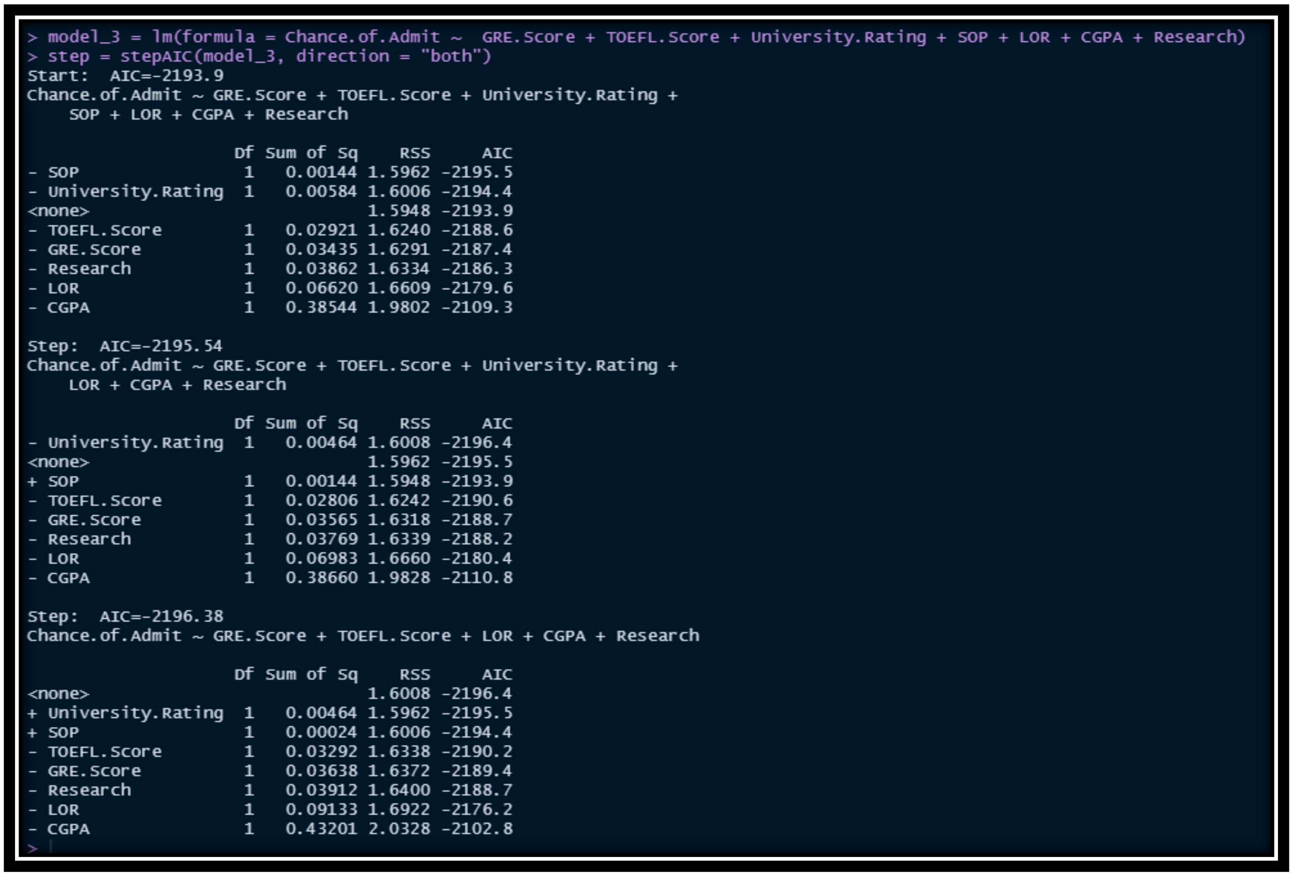
d: number of x variables included in the model

RSS: Residual sum of squares

c. Model evaluation

- **Stepwise Regression**:

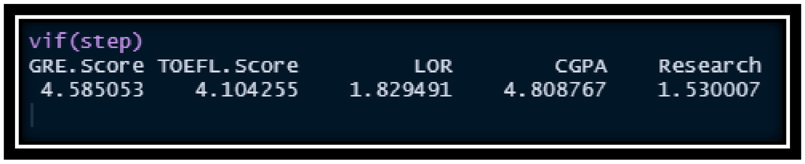
+ model 3 is similar to model 1 because Step AIC considers all the input parameters



```
> model_3 = lm(formula = Chance.of.Admit ~  GRE.Score + TOEFL.Score + University.Rating + SOP + LOR + CGPA + Research)
> step = stepAIC(model_3, direction = "both")
Start:  AIC=-2193.9
Chance.of.Admit ~ GRE.Score + TOEFL.Score + University.Rating + 
    SOP + LOR + CGPA + Research

                    Df Sum of Sq    RSS     AIC
- SOP                1   0.00144 1.5962 -2195.5
- University.Rating  1   0.00584 1.6006 -2194.4
<none>                           1.5948 -2193.9
- TOEFL.Score        1   0.02921 1.6240 -2188.6
- GRE.Score          1   0.03435 1.6291 -2187.4
- Research           1   0.03862 1.6334 -2186.3
- LOR                1   0.06620 1.6609 -2179.6
- CGPA               1   0.38544 1.9802 -2109.3

Step:  AIC=-2195.54
Chance.of.Admit ~ GRE.Score + TOEFL.Score + University.Rating + 
    LOR + CGPA + Research

                    Df Sum of Sq    RSS     AIC
- University.Rating  1   0.00464 1.6008 -2196.4
<none>                           1.5962 -2195.5
+ SOP                1   0.00144 1.5948 -2193.9
- TOEFL.Score        1   0.02806 1.6242 -2190.6
- GRE.Score          1   0.03565 1.6318 -2188.7
- Research           1   0.03769 1.6339 -2188.2
- LOR                1   0.06983 1.6660 -2180.4
- CGPA               1   0.38660 1.9828 -2110.8

Step:  AIC=-2196.38
Chance.of.Admit ~ GRE.Score + TOEFL.Score + LOR + CGPA + Research

                    Df Sum of Sq    RSS     AIC
<none>                           1.6008 -2196.4
+ University.Rating  1   0.00464 1.5962 -2195.5
+ SOP                1   0.00024 1.6006 -2194.4
- TOEFL.Score        1   0.03292 1.6338 -2190.2
- GRE.Score          1   0.03638 1.6372 -2189.4
- Research           1   0.03912 1.6400 -2188.7
- LOR                1   0.09133 1.6922 -2176.2
- CGPA               1   0.43201 2.0328 -2102.8
> 
```

+ We can see that the optimized model in step AIC is corollary to model 2.

- **VIF**:



```
vif(step)
GRE.Score TOEFL.Score         LOR        CGPA    Research 
 4.585053    4.104255    1.829491    4.808767    1.530007 
```

Criteria: VIF < 5:

VIF values of all attributes are satisfied after step AIC.

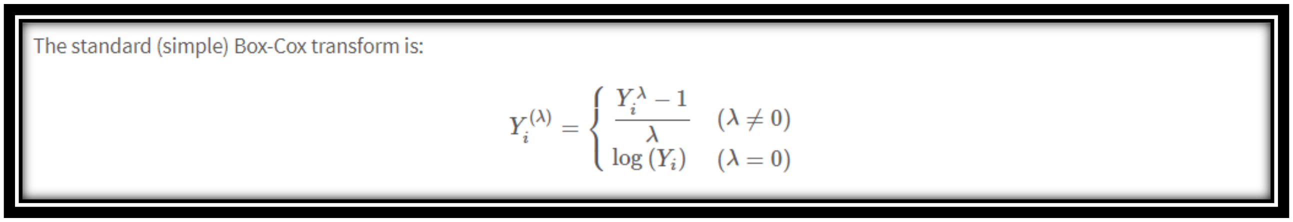
5. Final model – Improving model using Box Cox Transformation and Stepwise Regression

a. Box Cox Transformation

Introduction

This procedure finds the appropriate Box-Cox power transformation for a dataset containing a pair of variables that are to be analyzed by linear regression. The aims is to modify the distributional shape of the response variable so that the residuals are more normally distributed. This is done so that tests and confidence limits that require normality can more appropriately be used.

Formula:



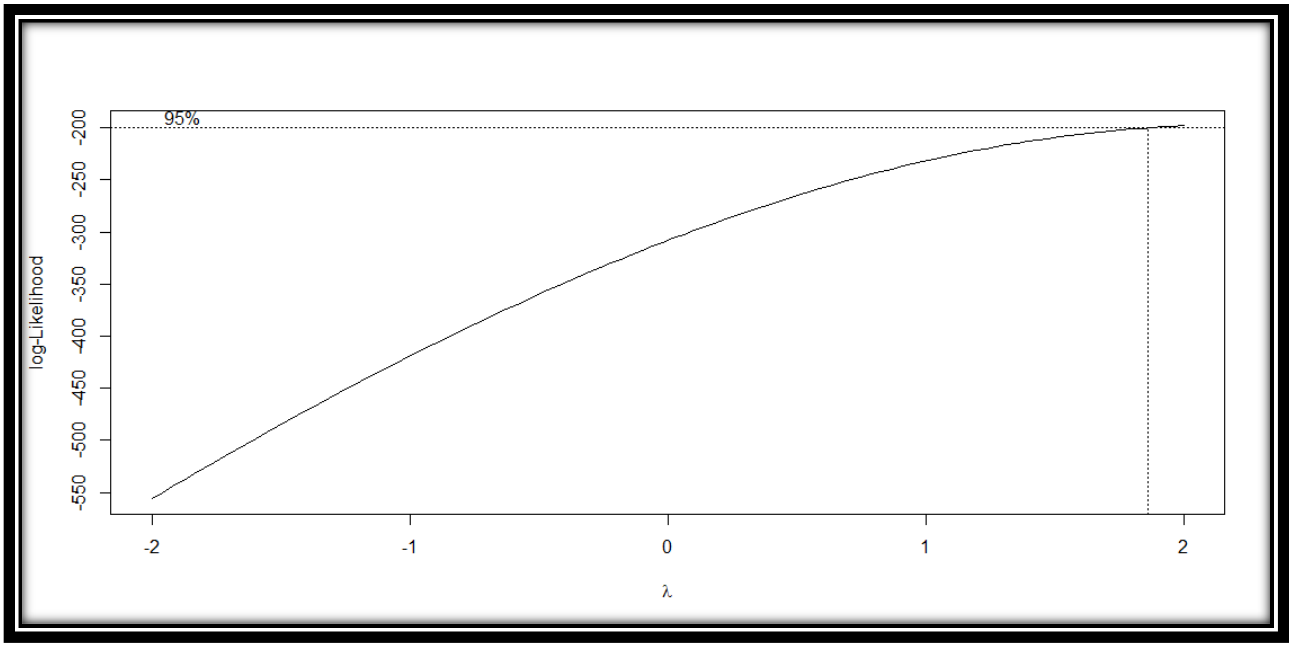
The standard (simple) Box-Cox transform is:

$$Y_i^{(\lambda)} = \begin{cases} \dfrac{Y_i^{\lambda} - 1}{\lambda} & (\lambda \neq 0) \\ \log\left(Y_i\right) & (\lambda = 0) \end{cases}$$

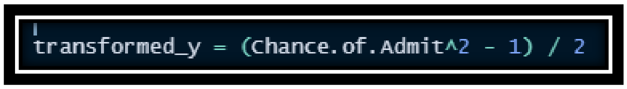
Likelihood Estimation:

The likelihood for a given λ is inversely proportional to the square root of the mean square error of the residuals from the linear regression. The likelihood function is maximized when this value is minimized.

Choosing lambda:



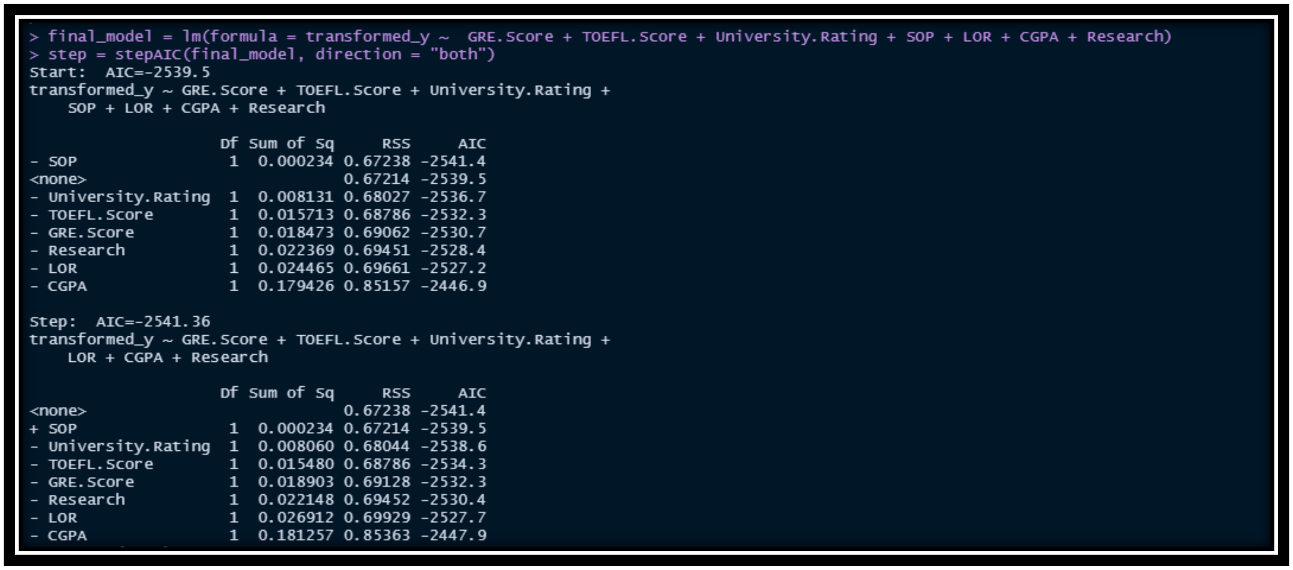
Look at the graph, we can see that the optimized lambda is 2. The following code transforms y.



```
transformed_y = (Chance.of.Admit^2 - 1) / 2
```

b. Model Evaluation

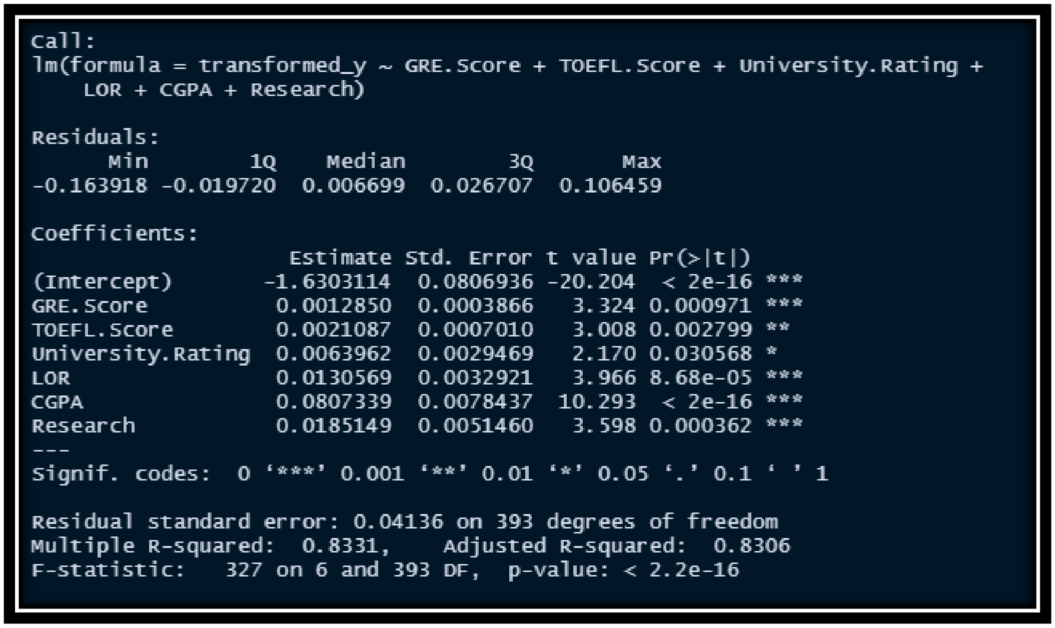
Step AIC:



```
> final_model = lm(formula = transformed_y ~  GRE.Score + TOEFL.Score + University.Rating + SOP + LOR + CGPA + Research)
> step = stepAIC(final_model, direction = "both")
Start:  AIC=-2539.5
transformed_y ~ GRE.Score + TOEFL.Score + University.Rating + 
    SOP + LOR + CGPA + Research

                    Df Sum of Sq     RSS     AIC
- SOP                1  0.000234 0.67238 -2541.4
<none>                           0.67214 -2539.5
- University.Rating  1  0.008131 0.68027 -2536.7
- TOEFL.Score        1  0.015713 0.68786 -2532.3
- GRE.Score          1  0.018473 0.69062 -2530.7
- Research           1  0.022369 0.69451 -2528.4
- LOR                1  0.024465 0.69661 -2527.2
- CGPA               1  0.179426 0.85157 -2446.9

Step:  AIC=-2541.36
transformed_y ~ GRE.Score + TOEFL.Score + University.Rating + 
    LOR + CGPA + Research

                    Df Sum of Sq     RSS     AIC
<none>                           0.67238 -2541.4
+ SOP                1  0.000234 0.67214 -2539.5
- University.Rating  1  0.008060 0.68044 -2538.6
- TOEFL.Score        1  0.015480 0.68786 -2534.3
- GRE.Score          1  0.018903 0.69128 -2532.3
- Research           1  0.022148 0.69452 -2530.4
- LOR                1  0.026912 0.69929 -2527.7
- CGPA               1  0.181257 0.85363 -2447.9
```

After transforming the response variable, only SOP attribute is removed.

Summary:



```
Call:
lm(formula = transformed_y ~ GRE.Score + TOEFL.Score + University.Rating + 
    LOR + CGPA + Research)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.163918 -0.019720  0.006699  0.026707  0.106459 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -1.6303114  0.0806936 -20.204  < 2e-16 ***
GRE.Score          0.0012850  0.0003866   3.324 0.000971 ***
TOEFL.Score        0.0021087  0.0007010   3.008 0.002799 ** 
University.Rating  0.0063962  0.0029469   2.170 0.030568 *  
LOR                0.0130569  0.0032921   3.966 8.68e-05 ***
CGPA               0.0807339  0.0078437  10.293  < 2e-16 ***
Research           0.0185149  0.0051460   3.598 0.000362 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04136 on 393 degrees of freedom
Multiple R-squared:  0.8331,	Adjusted R-squared:  0.8306 
F-statistic:   327 on 6 and 393 DF,  p-value: < 2.2e-16
```

Criteria: p-value < 0.05

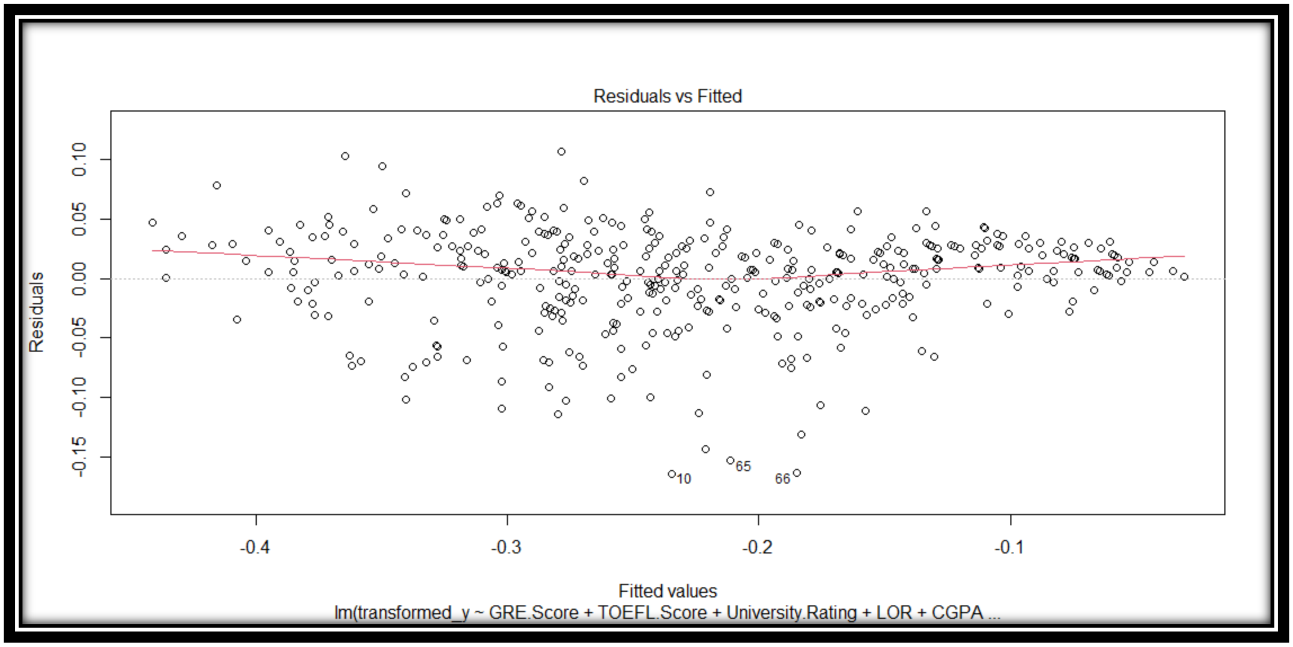
All attributes has satisfied p-value

Criteria: R-squared > 0.6

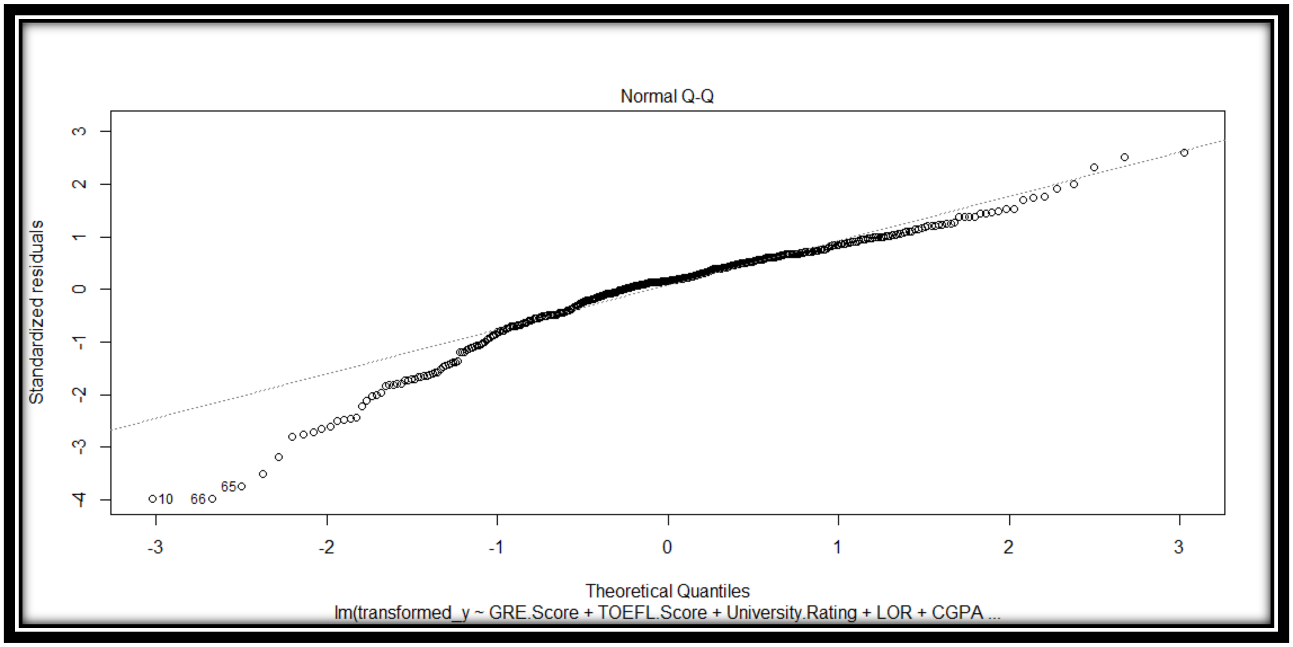
R-squared has increased 3.04% compared to model 2. This improvement is significant.

c. Residual Analysis

Residual vs Fitted plot:

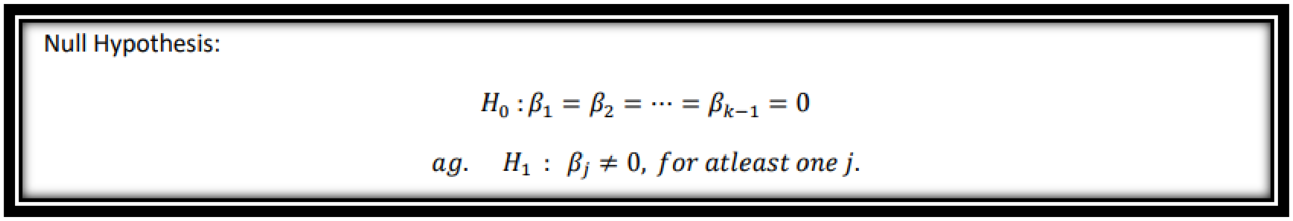


Normal QQ plot:



Inference: From both residual vs fitted plot and Normal QQ plot, we see the assumptions that variance of residual is constant, and the data is linearly distributed are correct. However, we still cannot tackle the problem of the red line and the outliers.

d. Hypothesis Testing on final model

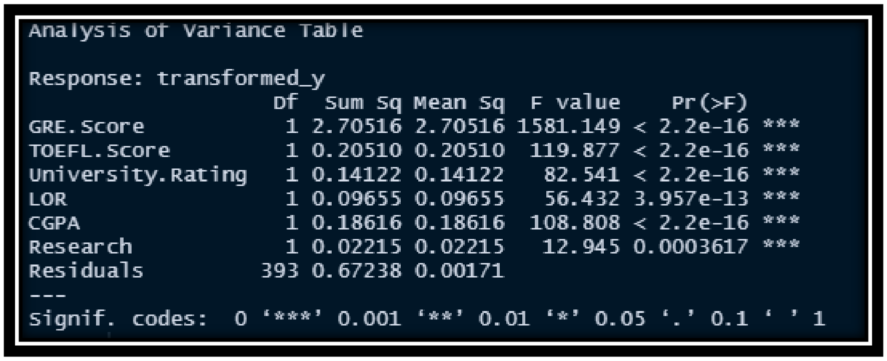


Null Hypothesis:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0$$

$$ag. \quad H_1 : \beta_j \neq 0, \; for\, atleast\, one\, j.$$

ANOVA output:



```
Analysis of Variance Table

Response: transformed_y
                   Df  Sum Sq Mean Sq  F value    Pr(>F)    
GRE.Score           1 2.70516 2.70516 1581.149 < 2.2e-16 ***
TOEFL.Score         1 0.20510 0.20510  119.877 < 2.2e-16 ***
University.Rating   1 0.14122 0.14122   82.541 < 2.2e-16 ***
LOR                 1 0.09655 0.09655   56.432 3.957e-13 ***
CGPA                1 0.18616 0.18616  108.808 < 2.2e-16 ***
Research            1 0.02215 0.02215   12.945 0.0003617 ***
Residuals         393 0.67238 0.00171                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Criteria: p-value < 0.05

Since all the p values are significant, we reject the null hypothesis. The final model can be used for predictions

III. Conclusion

a. Summary

We have improved model through several methods and testing to have the final accuracy of 83.31%. Therefore, the final model should be helpful in predicting the chance of admission. Besides, there are other methods for improving model we can also apply such as outlier treatment, fixing nonlinearity, interactions checking.

b. Reference

- Mohan S Acharya, Asfia Armaan, Aneeta S Antony: A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019
  https://www.kaggle.com/datasets/mohansacharya/graduate-admissions

- Box-cox transformation algorithm
  http://www.css.cornell.edu/faculty/dgr2/_static/files/R_html/Transformations.html#1_motivation

c. Appendix

R-Script (Submitted along with this report)

GitHub link:

- https://github.com/DVANH0302/Graduate-Submission-2-