method. This approach involves trimming off more extreme values, which helps minimize the sum of squared errors by focusing on a subset of the data that is less affected by outliers. By reducing the influence of outliers even further, the LTS method can potentially provide a more accurate and reliable model fit.

3 Graduate Admission Prediction

The aim of this report is to provide a comprehensive analysis of the Graduate Admissions dataset to predict the likelihood of graduate admissions based on various academic and personal attributes.

3.1 About Dataset

The dataset is designed to predict the likelihood of graduate admissions for students applying to Master's programs. The context is specific to an Indian perspective, which may reflect regional educational practices and standards.
The dataset includes several key parameters that are important in the admissions process. These parameters are:

- **GRE Scores**: Scores out of 340, representing the Graduate Record Examination scores.
- **TOEFL Scores**: Scores out of 120, representing the Test of English as a Foreign Language scores.
- **University Rating**: A rating out of 5, indicating the quality of the university.
- **Statement of Purpose (SOP)**: Strength of the SOP on a scale of 1 to 5.
- **Letter of Recommendation (LOR)**: Strength of the LOR on a scale of 1 to 5.
- **Undergraduate GPA (CGPA)**: GPA out of 10.
- **Research Experience**: Binary variable indicating whether the student has research experience (0 or 1).
- **Chance of Admit**: The probability of admission, ranging from 0 to 1.

3.1.1 Citation

> 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.

3.2 Prepare Data

3.2.1 Import Libraries and Set Up Data

To begin the analysis, necessary libraries are imported, and the dataset is loaded into the R environment in **Code 3.2.1**.

The result shows that the dataset contains 400 observations and 9 attributes, as displayed in the output below.

Output

```
Serial.No.  GRE.Score  TOEFL.Score  University.Rating  SOP  LOR  CGPA
1           337        118          4                   4.5  4.5  9.65
2           324        107          4                   4.0  4.5  8.87
3           316        104          3                   3.0  3.5  8.00
4           322        110          3                   3.5  2.5  8.67
5           314        103          2                   2.0  3.0  8.21
6           330        115          5                   4.5  3.0  9.34

            Research  Chance.of.Admit
1                  1             0.92
2                  1             0.76
3                  1             0.72
4                  1             0.80
5                  0             0.65
6                  1             0.90
[1] 400 9
```

3.2.2 Check Missing Values

The dataset is inspected for missing values to ensure data integrity. If missing values are detected, appropriate methods such as imputation or removal will be applied. From the output of **3.2.2**, we observe that it does not exist any missing values.

Output

```
Serial.No.0 GRE.Score0 TOEFL.Score0 University.Rating0 SOP0 LOR0 CGPA0
   Research0 Chance.of.Admit0
```

3.2.3 Detect and Handle Outliers

The Research variable indicates whether a student has research experience or not, which is a binary variable with values 0 or 1. So we convert this variable to a factor in **3.2.3** before identifying any potential outliers .

We generate box plots in Figure 1 for each of the other variables GRE.Score, TOEFL.Score, University.Rating, SOP, LOR, and CGPA in **Code 3.2.4**.

None of the boxplots show extreme outliers that are far from the whiskers, indicating that the data is generally well-behaved. There are a few outliers, specifically in the LOR and CGPA variables. In this case, since the outlier is minimal, we could remove the outliers to improve the model's predictive performance **Code 3.2.5**.

3.3 Descriptive Statistics

In this section, we summarize the key statistical measures for the variables under consideration, including the mean, median, standard deviation, and range.

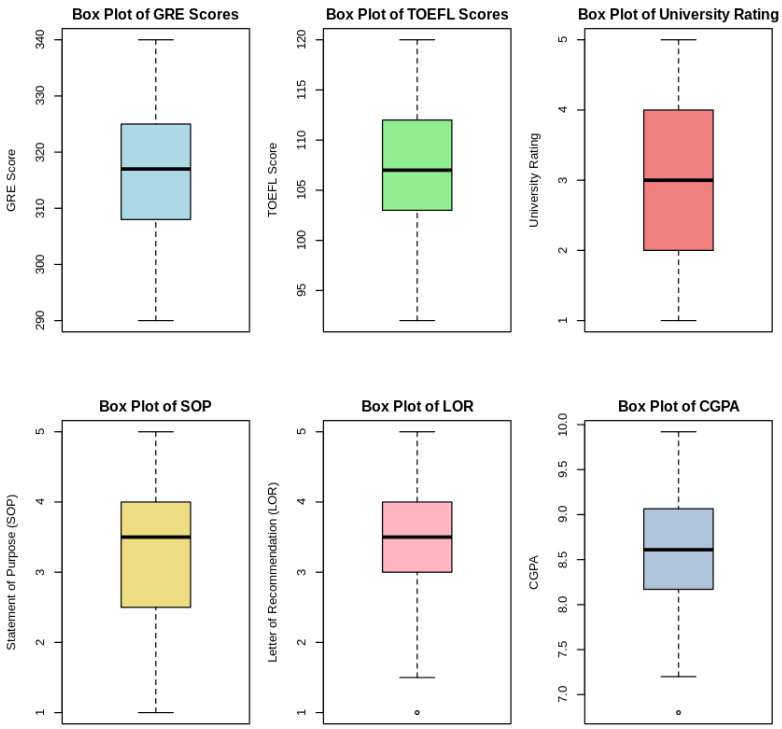


Figure 1: Boxplot of Variables

3.3.1 Summary Statistics

Table 1 presents the summary statistics for the quantitative variables in the dataset, including GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, and Chance of Admit.

Table 1: Summary Statistics for Quantitative Variables

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
GRE Score	290	309	317	317	325	340
TOEFL Score	92	103	107	107.5	112	120
University Rating	1	2	3	3.10	4	5
SOP	1	2.5	3.5	3.41	4	5
LOR	1.5	3	3.5	3.47	4	5
CGPA	7.2	8.20	8.63	8.61	9.07	9.92
Chance of Admit	0.36	0.64	0.73	0.73	0.83	0.97

3.3.2 Histograms and Density Plots

To visualize the distribution of the quantitative variables, we generate histograms overlaid with density plots for each variable in **3.2.6**. Figure 2 shows the histograms for GRE Score, TOEFL Score, CGPA, University Rating, and LOR.

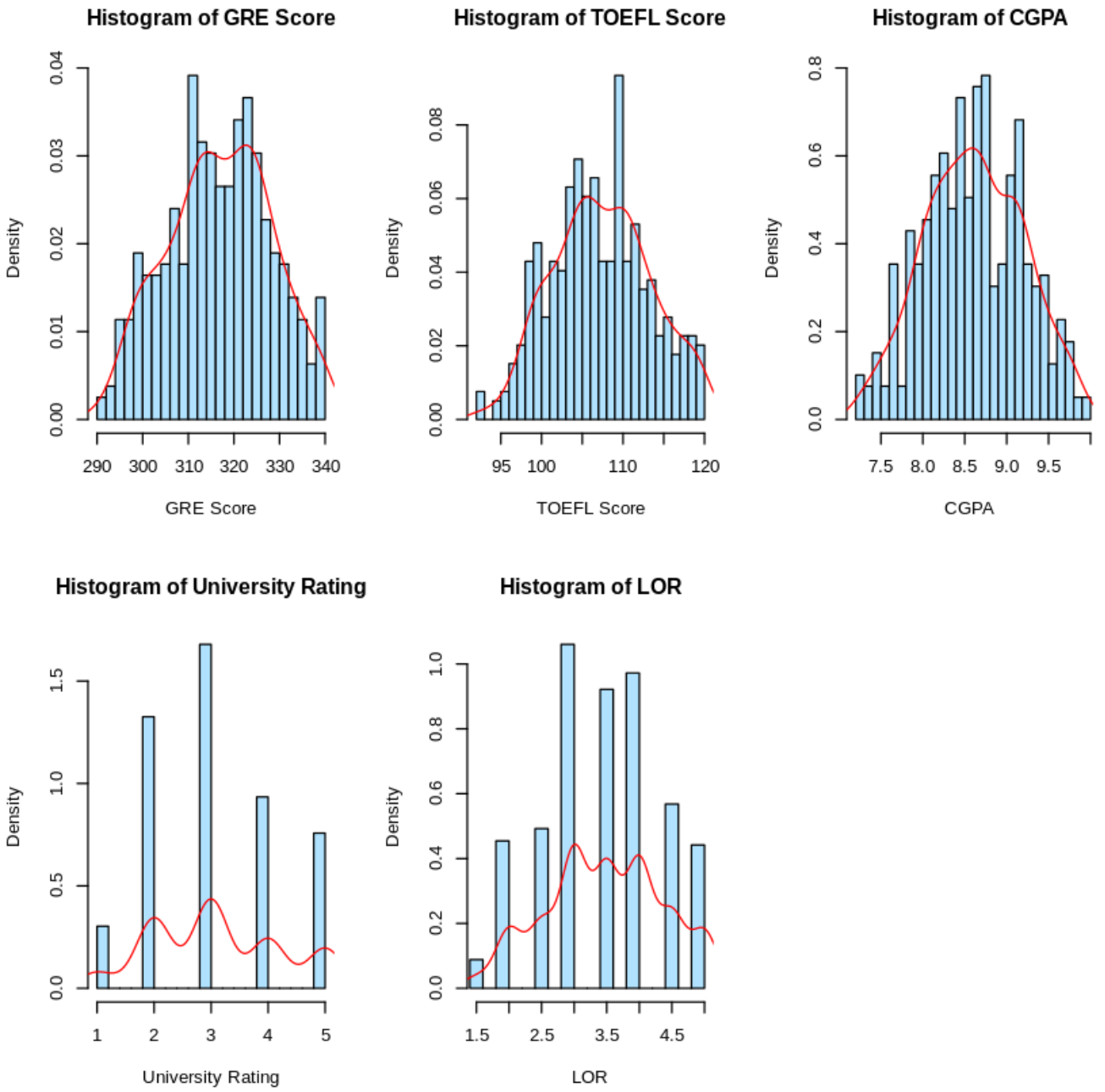


Figure 2: Histograms with Density Plots of Quantitative Variables

For most variables, the distribution is approximately normal, with slight deviations observed in the TOEFL Score and CGPA.

3.3.3 Distribution of the Qualitative Variable

We generated a pie chart to show the distribution of the ‘Research‘ variable in **3.2.7**. The chart 3 is divided into two segments representing the two categories of the ‘Research‘ variable:

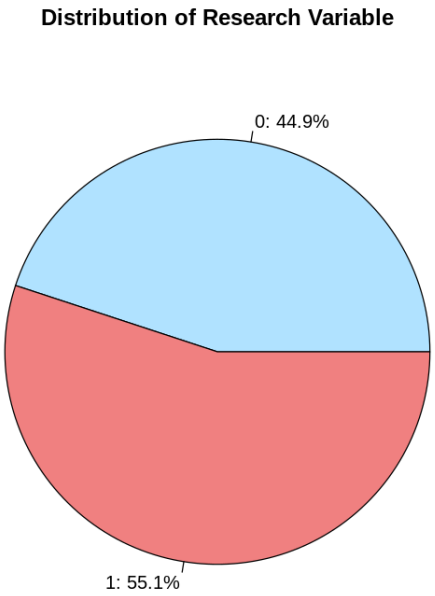


Figure 3: Distribution of Research Variable

- **0**: This segment represents the percentage of instances where 'Research = 0' (i.e., the candidate did not have research experience). This accounts for 44.9% of the data.
- **1**: This segment represents the percentage of instances where 'Research = 1' (i.e., the candidate had research experience). This accounts for 55.1% of the data.

The chart indicates that there is a relatively balanced distribution between the two categories, with slightly more candidates having research experience (55.1%) than those without (44.9%). This balance is important for analyzing the impact of research experience on the dependent variable 'Chance of Admit' because it ensures that both categories are well-represented.

3.4 Linear Regression Model

3.4.1 Data Splitting

To evaluate the performance of our linear regression model, the dataset was split into training and testing sets in **3.2.8**. This split helps in assessing how well the model generalizes to unseen data.

First, we set a seed to ensure the reproducibility of the results. Next, we randomly selected 80% of the data to be used as the training set, which will be used to train the linear regression model. The remaining 20% of the data was reserved as the testing set, which will be used to evaluate the model's performance.

3.4.2 Linear Regression Full Model

A linear regression model was fitted to the training data to predict the **Chance of Admit** using the following predictors: **GRE Score**, **TOEFL Score**, **University Rating**, **Statement of Purpose (SOP)**, **Letter of Recommendation (LOR)**, **CGPA**, and **Research Experience**. The model is specified in **3.2.9**

The output of the linear regression model is summarized below:

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

Residuals:
      Min         1Q     Median         3Q        Max
-0.241613 -0.023589  0.008477  0.034834  0.150241

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.3426740  0.1323573 -10.144  < 2e-16 ***
GRE.Score          0.0020801  0.0006429   3.235  0.00135 **
TOEFL.Score        0.0027798  0.0011608   2.395  0.01722 *
University.Rating  0.0072008  0.0053757   1.340  0.18138
SOP               -0.0005393  0.0062405  -0.086  0.93119
LOR                0.0197461  0.0061008   3.237  0.00134 **
CGPA               0.1178112  0.0130777   9.009  < 2e-16 ***
Research           0.0164261  0.0086233   1.905  0.05772 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06163 on 312 degrees of freedom
Multiple R-squared:  0.812,     Adjusted R-squared:  0.8078
F-statistic: 192.6 on 7 and 312 DF,  p-value: < 2.2e-16
```

- The model has a high **Multiple R-squared** value of 0.812, which suggests that approximately 81.2% of the variability in the **Chance of Admit** is explained by the predictors in the model.
- The **Adjusted R-squared** value of 0.8078 indicates that the model remains strong even after adjusting for the number of predictors.
- The **F-statistic** of 192.6 and the associated p-value of less than 2.2e-16 indicate that the overall model is statistically significant.
- Significant predictors in the model include **GRE Score** (p-value = 0.00135), **TOEFL Score** (p-value = 0.01722), **LOR** (p-value = 0.00134), and **CGPA** (p-value $<$ 2e-16). These variables have a meaningful impact on the **Chance of Admit**.
- **University Rating**, **SOP**, and **Research Experience** do not have statistically significant coefficients, suggesting that they may not have a strong linear relationship with the **Chance of Admit** in this model.

These results highlight the importance of academic performance indicators such as **GRE Score**, **TOEFL Score**, and **CGPA** in predicting the likelihood of admission to graduate programs.

3.4.3 Model Reduction and Multicollinearity Analysis

After fitting the full linear regression model, we evaluated the presence of multicollinearity among the predictors using the Variance Inflation Factor (VIF). The VIF values for the full model are as follows:

```
GRE.Score          4.52
TOEFL.Score        3.98
University.Rating  3.02
SOP                3.09
LOR                2.42
CGPA               4.95
Research           1.51
```

Generally, a VIF value greater than 5 suggests a problematic level of multicollinearity. The predictors **GRE.Score** and **CGPA** exhibited VIF values close to 5, indicating potential multicollinearity.

To address this issue, we performed stepwise model reduction, removing predictors with higher VIF values while monitoring the impact on the model's performance.

3.4.4 Reduced Model 1: Excluding CGPA

In the first reduced model, **CGPA** was removed due to its high VIF value. The resulting model is as follows:

```
model_red1 <- lm(Chance.of.Admit ~ GRE.Score + TOEFL.Score +
                 University.Rating + SOP + LOR + Research, data = trainData)
```

The summary of **Reduced Model 1** shows:

```
Residual standard error: 0.06907
Multiple R-squared: 0.7632
Adjusted R-squared: 0.7586
```

The VIF values improved slightly, particularly for **GRE.Score** and **CGPA**, though some multicollinearity persisted.

3.4.5 Reduced Model 2: Excluding GRE.Score

In **Reduced Model 2**, we further excluded **GRE.Score**, another variable with a high VIF value:

```
model_red2 <- lm(Chance.of.Admit ~ TOEFL.Score + University.Rating +
                 SOP + LOR + Research, data = trainData)
```

The summary of **Reduced Model 2** shows:

```
Residual standard error: 0.07355
Multiple R-squared: 0.7305
Adjusted R-squared: 0.7263
```

VIF values were further reduced, and the model remained statistically significant, with a small reduction in the adjusted R-squared value.

3.4.6 Reduced Model 3: Excluding SOP

Continuing the reduction process, **SOP** was excluded in **Reduced Model 3**, leading to the following model:

```
model_red3 <- lm(Chance.of.Admit ~ TOEFL.Score + University.Rating +
                 LOR + Research, data = trainData)
```

The summary of **Reduced Model 3** shows:

```
Residual standard error: 0.07357
Multiple R-squared: 0.7296
Adjusted R-squared: 0.7261
```

The VIF values were further lowered, and the model's adjusted R-squared remained stable.

3.4.7 Final Reduced Model: Excluding University Rating

Finally, in **Reduced Model 4**, we excluded **University Rating**, resulting in the following simplified model:

```
model_red4 <- lm(Chance.of.Admit ~ TOEFL.Score + LOR + Research,
                 data = trainData)
```

The summary of **Reduced Model 4** shows:

```
Residual standard error: 0.07544
Multiple R-squared: 0.7147
Adjusted R-squared: 0.712
```

The VIF values in this model were all below 2, indicating minimal multicollinearity:

```
TOEFL.Score  1.63
LOR          1.49
Research     1.32
```

The final reduced model includes **TOEFL.Score**, **LOR**, and **Research** as predictors. Despite the stepwise removal of variables, this model still explains approximately 71.2% of the variance in the **Chance of Admit**. This model was chosen for its simplicity, improved VIF values, and statistically significant predictors.

While the VIF values in `model_red4` are all below 2, indicating minimal multicollinearity, several diagnostics were conducted to assess the model's assumptions:

```
 lag Autocorrelation D-W Statistic p-value
    1     0.004344417      1.990565   0.916
  Alternative hypothesis: rho != 0

         studentized Breusch-Pagan test

data:  model_red4
BP = 18.124, df = 3, p-value = 0.0004147

         Shapiro-Wilk normality test

data:  model_red4$residuals
W = 0.9503, p-value = 6.304e-09
```

Durbin-Watson Test for Autocorrelation

- **Null Hypothesis (H_0)**: There is no autocorrelation in the residuals ($\rho = 0$).
- **Alternative Hypothesis (H_1)**: There is autocorrelation in the residuals ($\rho \neq 0$).

The Durbin-Watson statistic is 1.9906 with a p-value of 0.916. Since the p-value is greater than the significance level (typically 0.05), we fail to reject the null hypothesis. This suggests that there is no significant autocorrelation in the residuals.

Breusch-Pagan Test for Heteroscedasticity

- **Null Hypothesis (H_0)**: The residuals have constant variance (homoscedasticity).
- **Alternative Hypothesis (H_1)**: The residuals do not have constant variance (heteroscedasticity).

The Breusch-Pagan test statistic is 18.124 with a p-value of 0.0004147. Since the p-value is less than the significance level (0.05), we reject the null hypothesis. This indicates the presence of heteroscedasticity in the residuals, meaning that the variance of the errors is not constant.

Shapiro-Wilk Test for Normality

- **Null Hypothesis (H_0)**: The residuals are normally distributed.
- **Alternative Hypothesis (H_1)**: The residuals are not normally distributed.

The Shapiro-Wilk W statistic is 0.9503 with a p-value of 6.304e-09. Since the p-value is much smaller than 0.05, we reject the null hypothesis, indicating that the residuals are not normally distributed.

The results show that while the residuals do not show autocorrelation, they show significant evidence of heteroscedasticity and non-normality. These issues suggest that the model might need further refinement or transformation to better meet the assumptions of linear regression.

3.5 Model Transformation

Following the diagnostics of the reduced model `model_red4`, which indicated issues with heteroscedasticity and non-normality of residuals, a Box-Cox transformation was applied to the dependent variable `Chance.of.Admit` in **3.2.10**. The optimal lambda value, determined using the `boxcox` function, was found to be approximately 2.27 as observed in Figure 4.

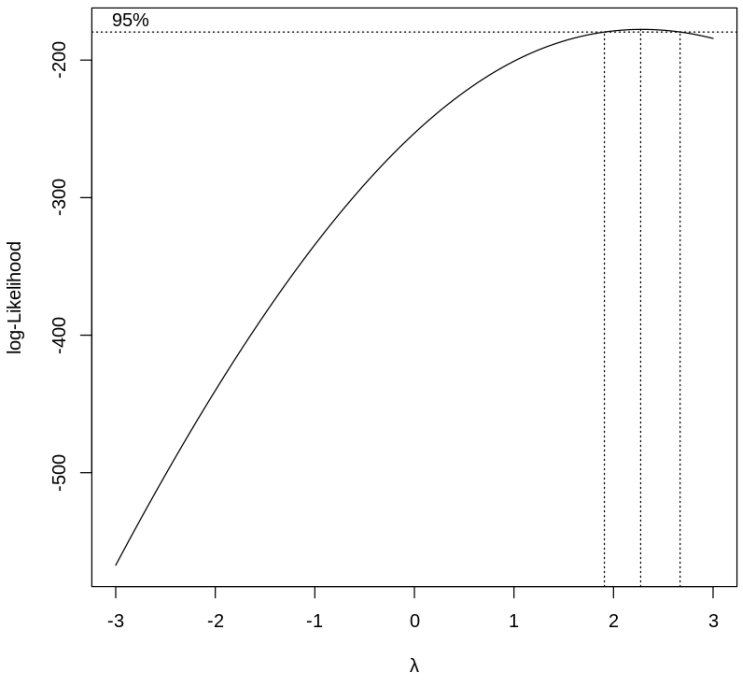


Figure 4: Box-Cox Transformation Plot

A new model, `model_best`, was fitted using `(Chance.of.Admit)`$^{2.27}$:

```
Call:
lm(formula = (Chance.of.Admit)^(best_lam_new) ~ TOEFL.Score +
    LOR + Research, data = trainData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38636 -0.05790  0.00899  0.07728  0.29266

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.928291   0.121865 -15.823  < 2e-16 ***
TOEFL.Score  0.020087   0.001255  16.002  < 2e-16 ***
LOR          0.070508   0.008102   8.703  < 2e-16 ***
Research     0.064760   0.013641   4.747 3.13e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
17
18 Residual standard error: 0.1043 on 316 degrees of freedom
19 Multiple R-squared:  0.7505,    Adjusted R-squared:  0.7481
20 F-statistic: 316.9 on 3 and 316 DF,  p-value: < 2.2e-16
```

After applying the Box-Cox transformation, the model's Multiple R-squared improves to 0.7505, and the Adjusted R-squared increases to 0.7481. However, the residual standard error increases to 0.1043. The predictors remain highly significant after the transformation.

3.6 Diagnostic Tests for Final Model

```
1  TOEFL.Score 1.62623284234583 LOR 1.49247307185952 Research
      1.324096427304
2  lag Autocorrelation D-W Statistic p-value
3     1      -0.0184738      2.034658   0.708
4   Alternative hypothesis: rho != 0
5
6           studentized Breusch-Pagan test
7
8  data:  model_best
9  BP = 5.9545, df = 3, p-value = 0.1138
10
11          Shapiro-Wilk normality test
12
13 data:  model_best$residuals
14 W = 0.9808, p-value = 0.000278
```

3.6.1 Multicollinearity

```
VIF Values:
TOEFL.Score: 1.626
LOR: 1.492
Research: 1.324
```

The VIF values for the predictors in `model_best` remain below 2, indicating minimal multicollinearity.

3.6.2 Autocorrelation

The Durbin-Watson test was conducted to check for autocorrelation in the residuals:

```
lag Autocorrelation D-W Statistic p-value
   1      -0.0184738      2.034658   0.768
```

The Durbin-Watson statistic is 2.0347 with a p-value of 0.768, suggesting that there is no significant autocorrelation in the residuals.

3.6.3 Homoscedasticity

The Breusch-Pagan test was performed to test for heteroscedasticity:

```
BP = 5.9545, df = 3, p-value = 0.1138
```

With a p-value of 0.1138, the test does not provide sufficient evidence to reject the null hypothesis of homoscedasticity, implying that the residuals are homoscedastic.

3.6.4 Normality of Residuals

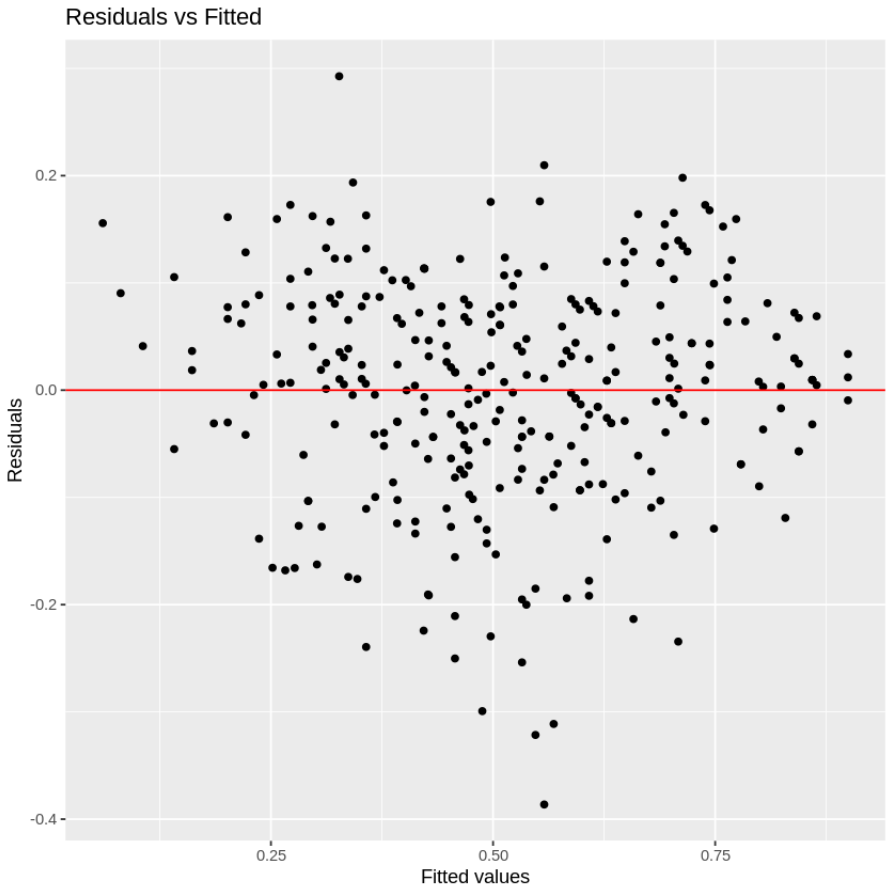
The Shapiro-Wilk test was used to check the normality of the residuals:

```
W = 0.9808, p-value = 0.000278
```

The p-value is 0.000278, indicating that the residuals are not normally distributed. However, given the large sample size, this deviation from normality may not significantly impact the validity of the model's inferences.

3.7 Visualization on Plot

We plot the figure using **3.2.11** to evaluate the assumptions of the linear regression model, particularly focusing on the assumptions of linearity, homoscedasticity (constant variance), and the independence of errors.

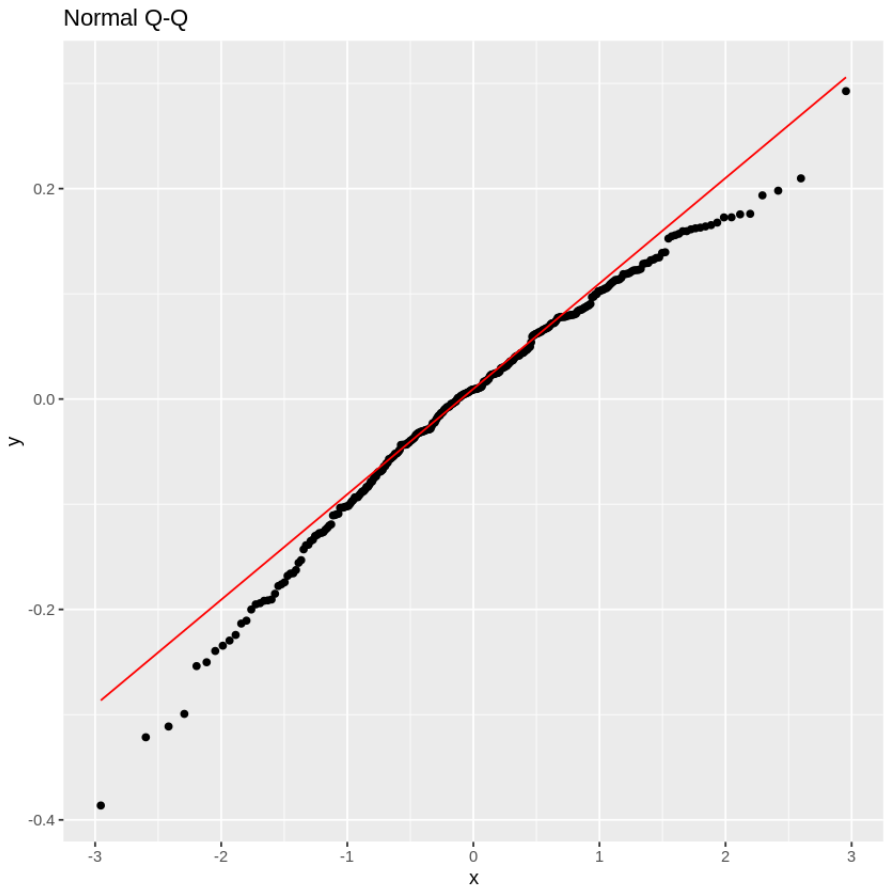


From the plot of residuals versus Fitted values, we can find some information:

- Firstly, we can see the points seem to significantly vary away from the red line due to lambda being 2.27. This suggests that our model needs another transformation to make the points fit closer to the line.

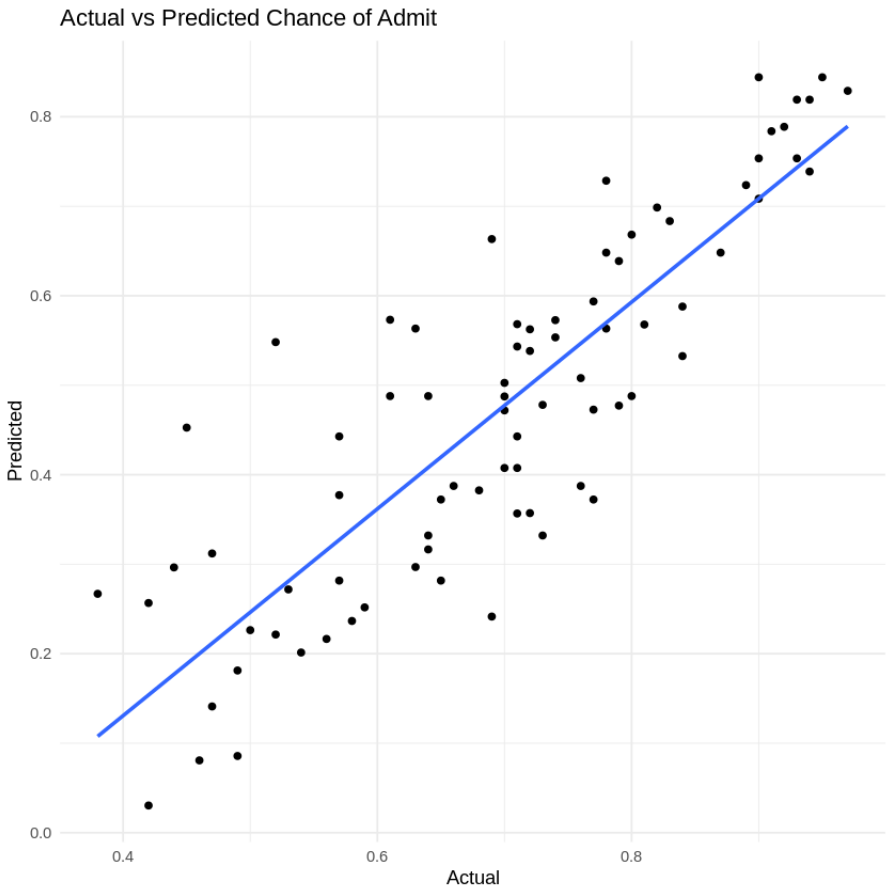
- Secondly, If we draw a line from top to bottom, left to right where the points group together. We can see at around 0.3, that line varies from 0.2 to 0, and at around 0.5, that line shifts down, from 0.1 to -0.14. This means that the residuals do change throughout entire values but not too much, so we can assume that heteroscedasticity is a minor issue, which is also derived from the Breusch-Pagan test.

- Thirdly, we can see some cluster points (the group of points that overlap each other) at around 0.5 of fitted values, which indicates an interaction between variables. However, visually, the cluster is not too many to form a clear pattern so we can conclude that interaction is not considerable. Besides, the cluster seems to be tightly close to the red line and does not deviate significantly away from overall points.



From the Q-Q residuals plot, we can see that the normality of the data is not quite good. From the half bottom to the half upper, the points seem to slightly deviate from the red line. This means that our data is not normal and considered using other transformations. However, we have already used BoxCox transformation, so we will need another model that less relies on normality, mean, and variances.

3.8 Model Evaluation on Test Data



The performance of the final transformed model `model_best` was evaluated on the test dataset in **3.2.12** and **3.2.13**. The following metrics were computed:

- **Mean Absolute Error (MAE):** 0.0631
- **Mean Squared Error (MSE):** 0.0074
- **Root Mean Squared Error (RMSE):** 0.0861
- **R-squared (R^2):** 0.6754
- **Mean value of test data:** 0.6981

MAE gives us 0.0631 means on average, our predictions off 0.0631%. Also, RMSE means that our predictions off 0.0861%. The R-squared is 0.6754 which means that about 67.54% of data can be explained by our model. If we take that error compared with the mean value for percentages, we will have:

- $\frac{0.0631}{0.6981} = 9.04\%$ means predicted values is 9.04% different from actual mean values, which is relatively high. However, as we use 10% error for the standard so we assume this error is acceptable.

- $\frac{0.0861}{0.6981} = 12.33\%$ means predicted values are 12.33% different from actual mean values, which is relatively high. This means that our model even has a decent R-squared, is still not good

enough for prediction. So we consider using other models which are not affected by means, or variances.

While the R^2 value on the test data is slightly lower compared to the training data, this is expected due to the model's generalization to unseen data. Overall, the transformed model `model_best` demonstrates strong predictive performance, making it a suitable choice for predicting the `Chance.of.Admit`.

3.9 Conclusion

The regression model was fitted using the Box-Cox transformation with the best lambda (λ) on the dependent variable, *Chance of Admit*. The model incorporates *TOEFL Score*, *Letter of Recommendation (LOR)*, and *Research Experience* as independent variables. The results show that all three predictors are statistically significant with p-values less than 0.001, indicating strong evidence against the null hypothesis for each predictor.

- **TOEFL Score**: The coefficient for *TOEFL Score* is positive, suggesting that higher TOEFL scores are associated with an increased *Chance of Admit*. This finding is intuitive, as higher language proficiency should increase the likelihood of admission.
- **LOR**: The *LOR* coefficient is also positive and statistically significant, indicating that stronger recommendations correlate with higher admission chances. This reflects the importance of endorsements from academic or professional references in the admissions process.
- **Research Experience**: Similarly, the *Research Experience* coefficient is positive and significant, suggesting that candidates with research experience are more likely to be admitted. This aligns with the expectations that research involvement is a strong indicator of a candidate's capability for graduate studies.

However, the residual analysis indicates some areas for caution:

- The residuals appear to be heteroscedastic, suggesting that the variance of errors is not constant across all levels of fitted values.
- The Shapiro-Wilk test for normality of residuals has a p-value below 0.05, indicating that the residuals are not perfectly normally distributed, although the violation is minor.
- The Durbin-Watson statistic suggests no significant autocorrelation in the residuals, supporting the validity of the model.

3.9.1 Suggestions

Although the model performs well, there is room for improvement:

- **Model Complexity**: Consider exploring non-linear models or interaction terms to capture more complex relationships in the data.
- **Feature Engineering**: Additional features, such as the quality of undergraduate institutions or specific course grades, could be incorporated to enhance the model's predictive power.

- **LTS Method**: The Least Trimmed Squares (LTS) method could be employed to enhance the robustness of the model. By trimming off extreme values that contribute to the sum of squared errors, LTS minimizes the influence of outliers and could lead to a more reliable estimation of the regression coefficients.
- **Data Collection**: Gathering more data, particularly from different regions, could improve the generalizability of the model.

Further research could also focus on understanding why certain predictors, such as `CGPA` or `SOP`, did not perform as expected in the model. Additionally, the non-normality of residuals could be addressed by applying more advanced statistical techniques or data transformations.

4 Appendix

Code 1.2.1

```
library(ggplot2)
library(dplyr)
library(moments)
library(corrplot)
library(car)
library(lmtest)
library(faraway)
library(MASS)
ori_data<-read.csv("auto_mpg.csv",sep = ";")
head(ori_data)
dim(ori_data)
```

Code 1.2.2

```
str(ori_data)
```

Code 1.2.3

```
ori_data$horsepower <- as.numeric(ori_data$horsepower)
ori_data  <- ori_data [, !(names(ori_data ) %in% c("car_name"))]
dim(ori_data)
```

Code 1.2.4

```
apply(is.na(ori_data),2,which)
```