Regression Analysis Spring-2019 Final Project

Factors Affecting Graduate Admissions

By,

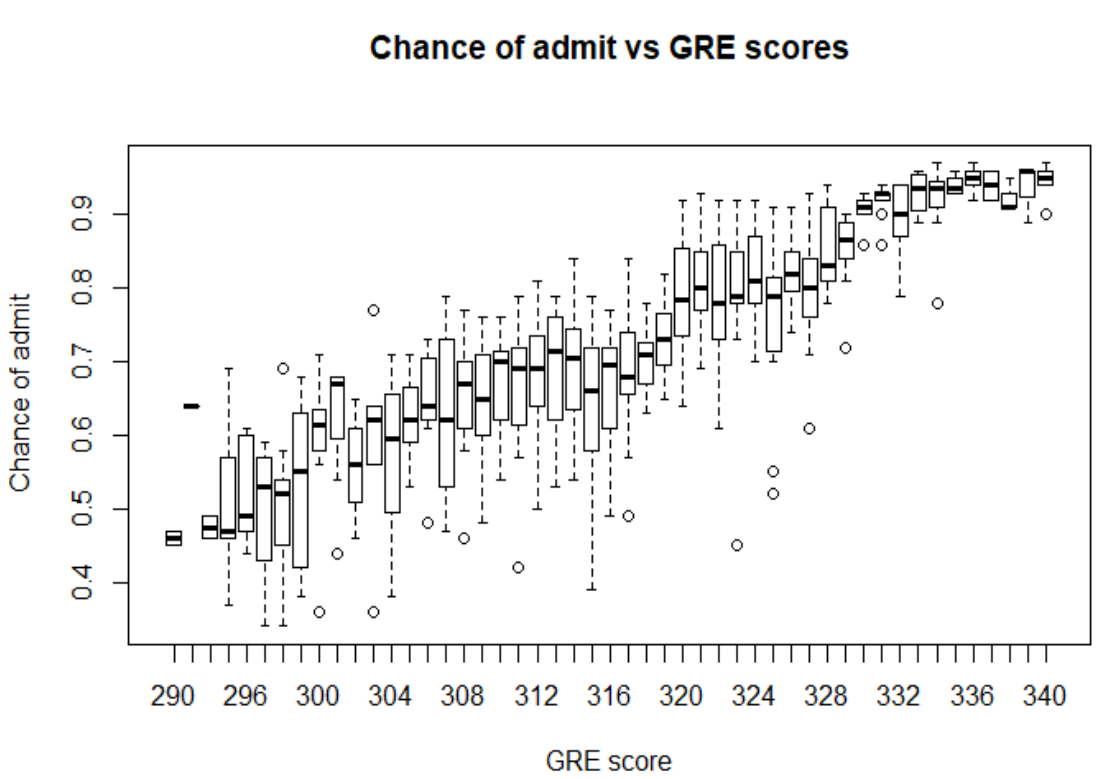
Krishnamurthy Prajwal Chadaga

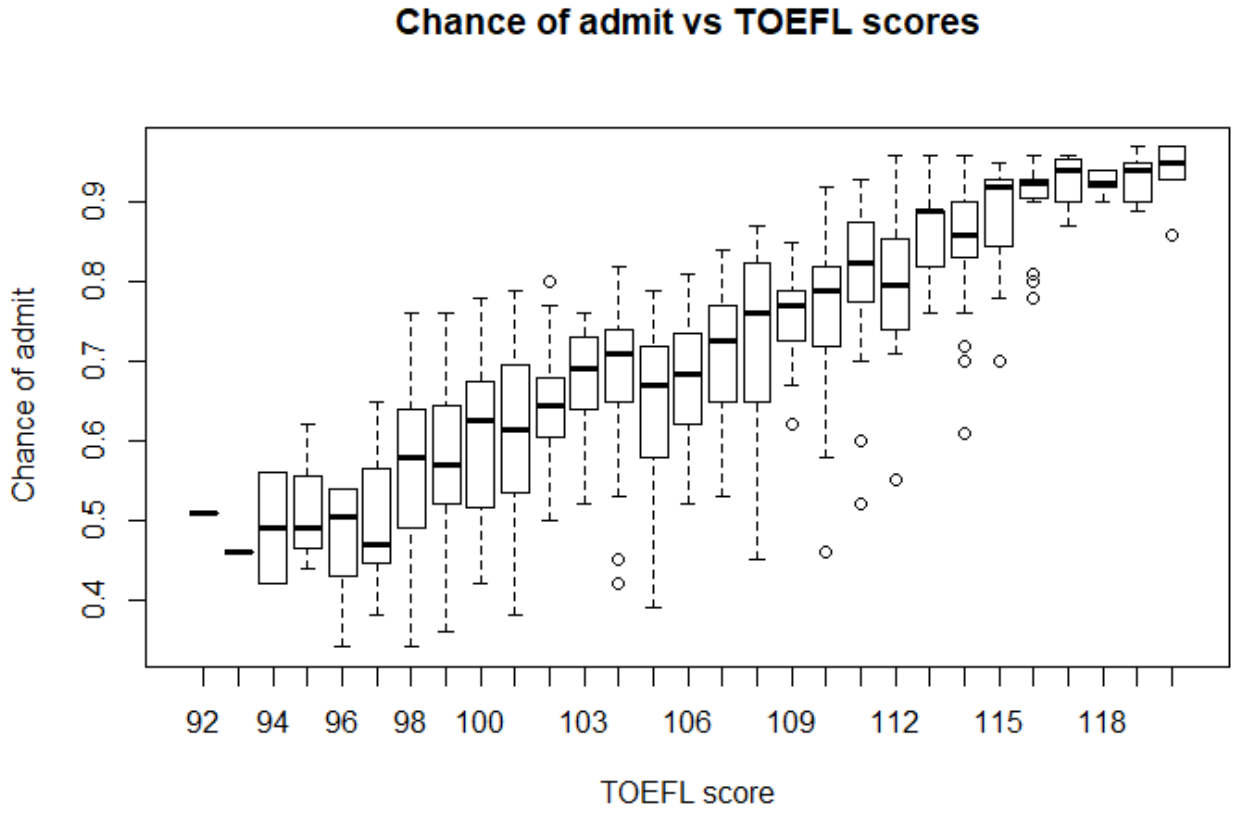
Purpose

- To analyze the factors affecting the admission of a student into a college
- To build a model containing all these factors and the analyze significance of the various factors
- To build a final model after removing the insignificant factors and analyzing the new model
- The factors that I've considered are GRE Score (out of 340), TOEFL Score (out of 120), University rating (1-5), Statement of Purpose (SOP) rating (1-5), Letter of Recommendation (LOR) rating (1-5), Cumulative GPA (out of 10) and Research (1 if any research done or 0 otherwise)
- The dataset I have used is derived from Kaggle and contains 500 entries. The dataset is inspired by the UCLA graduate dataset

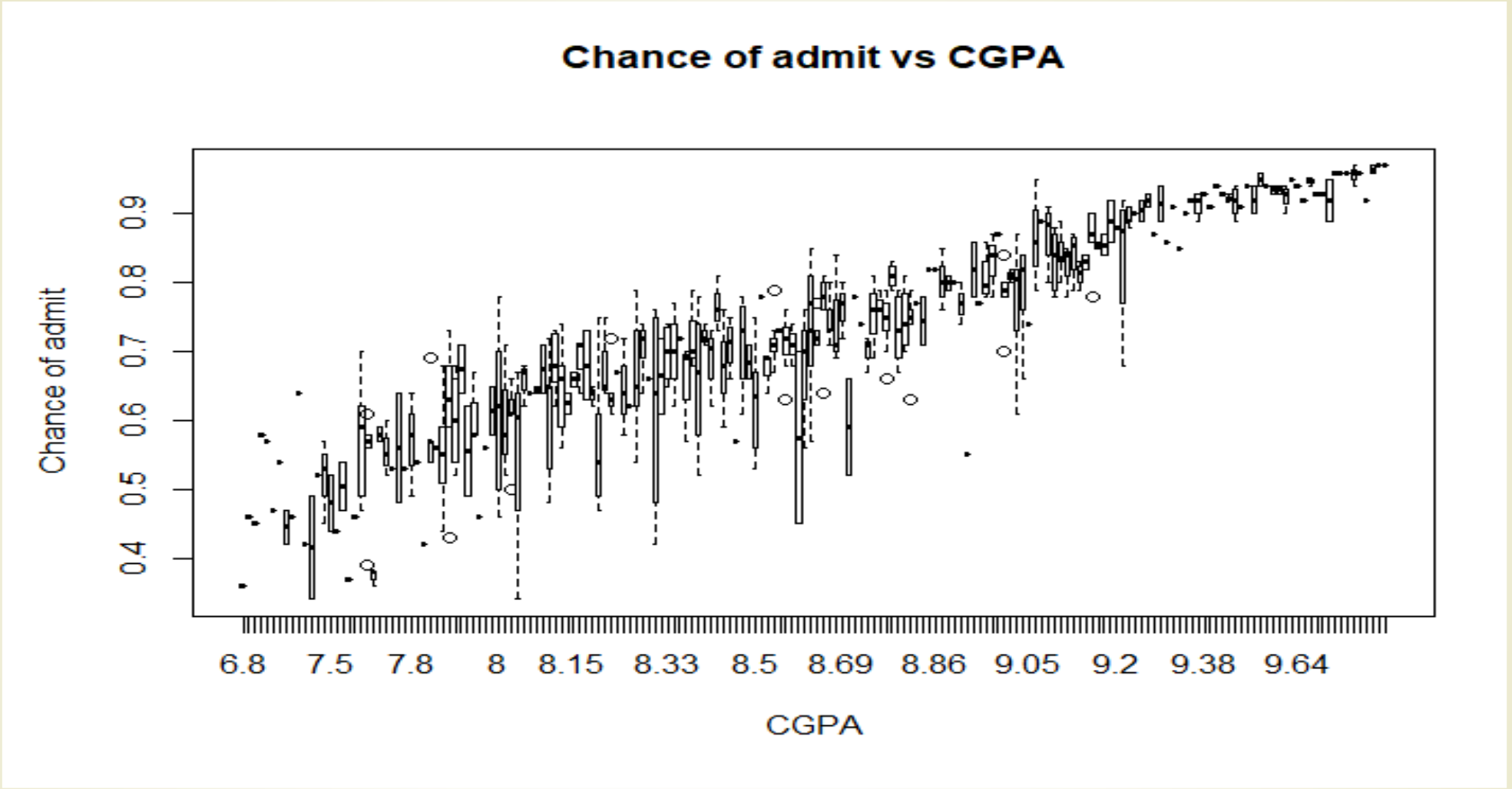
Important Statistics

- Multiple students have perfect GRE (340/340), TOEFL (120/120), SOP (5/5), LOR(5/5) and Research (1/1) scores
- The highest Chance of admit is 0.97 (97%) and incidentally this student also has the highest CGPA of 9.92 and perfect GRE and TOEFL scores and Research value of 1
- Students who have the highest SOP, LOR, University rating of 5 and have done research (1), incidentally don't have a perfect GRE/TOEFL score or the highest CGPA. This indicates that all these factors are completely independent of each other
- The students with the lowest Chance of Admit (0.34/34%) don't have the lowest value of any of the above mentioned factors indicating that it's not a straight-forward linear relationship between the Chance of Admit and the independent factors

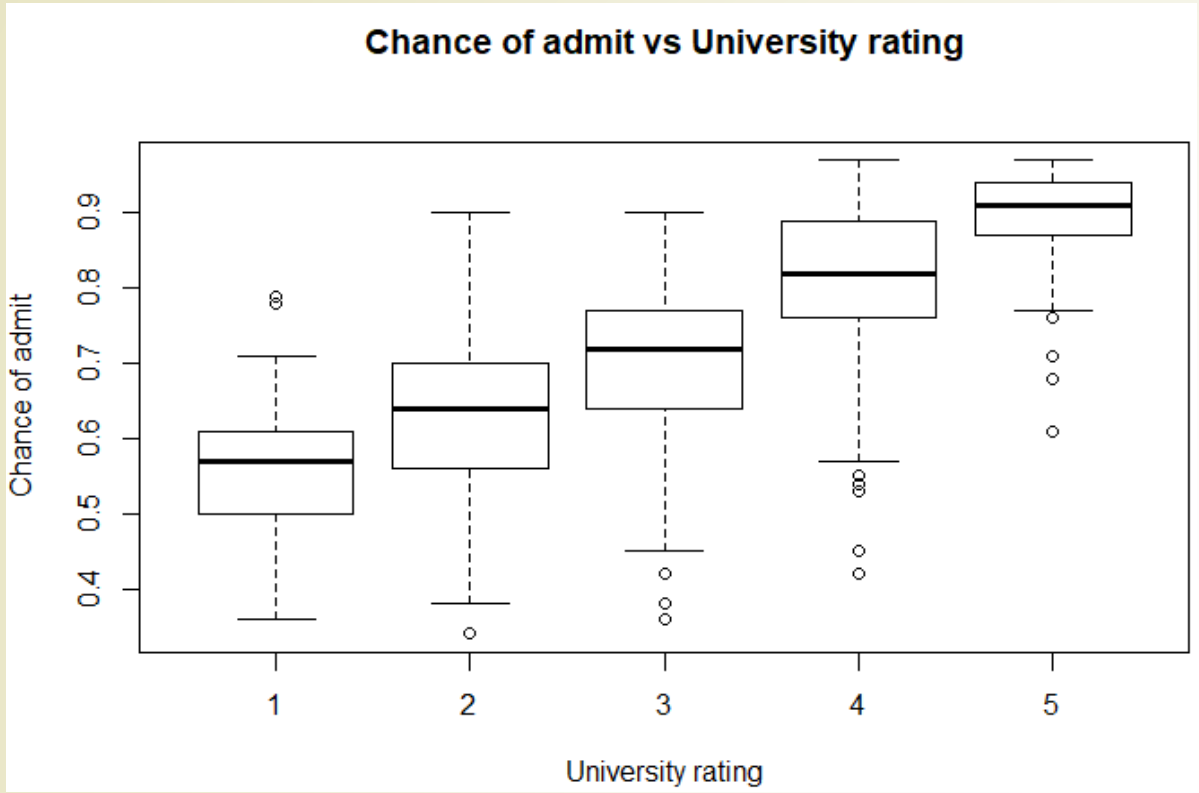


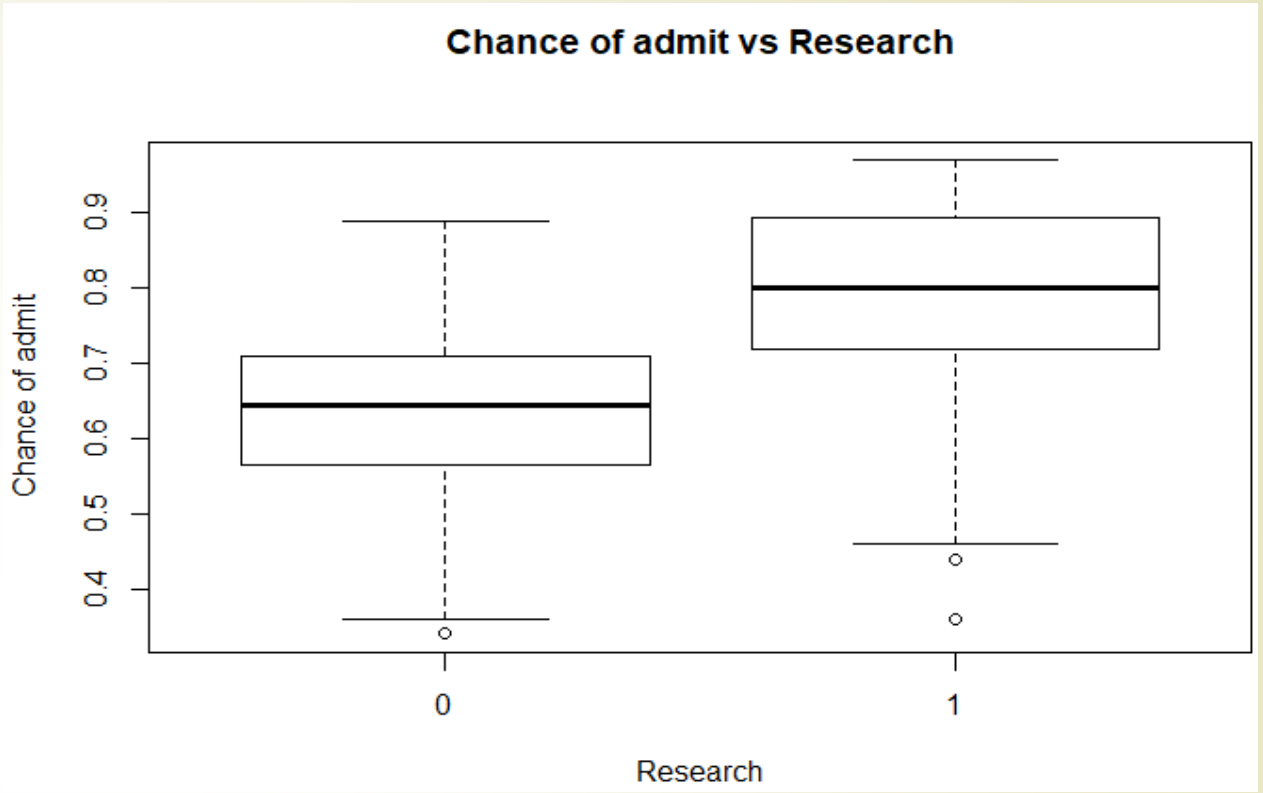


- The above box plots indicate that the Chance of admit doesn't have a linear relationship with both the GRE and the TOEFL scores which implies that, with an increase in these scores the chance of admit doesn't really
- There are also a lot of outliers present in both these plots

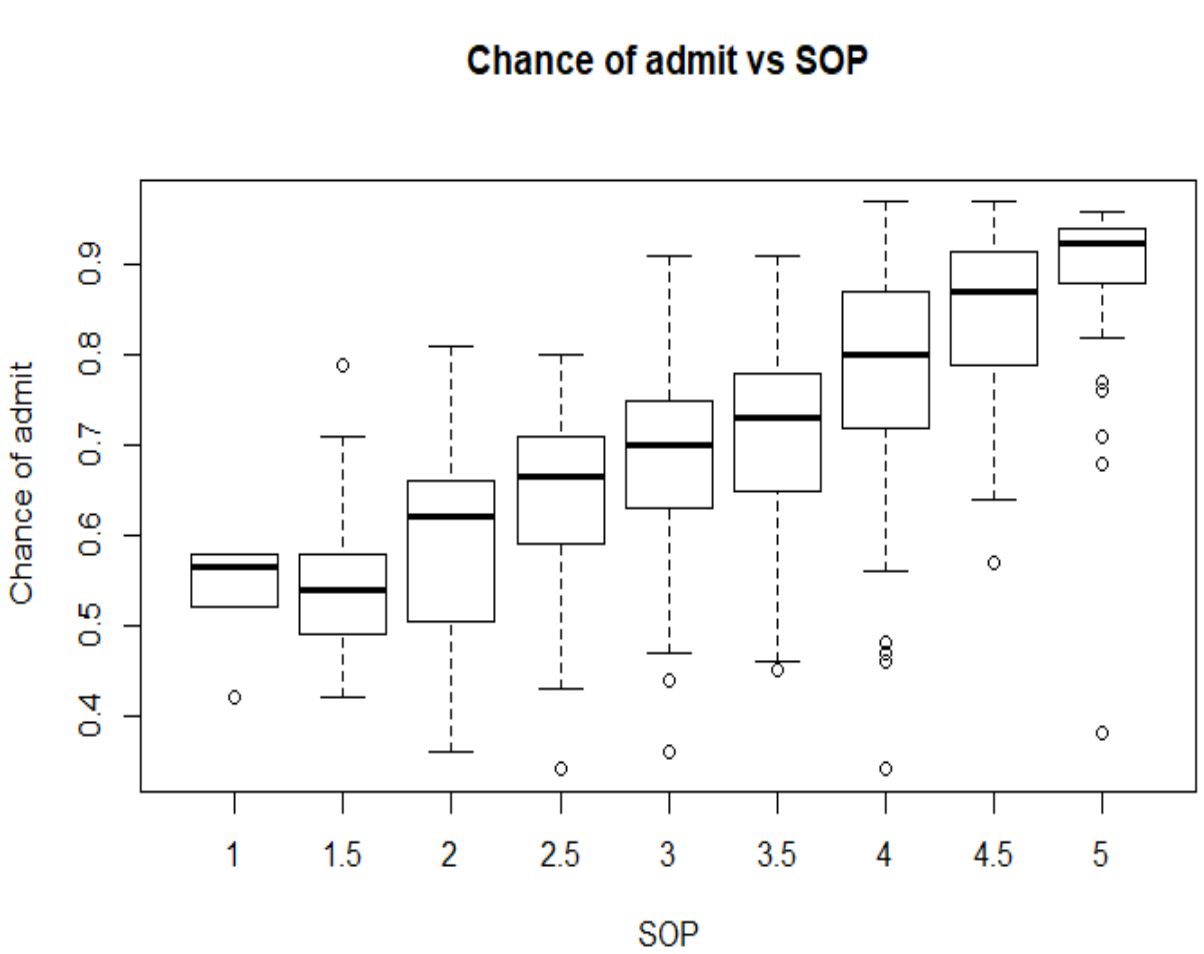


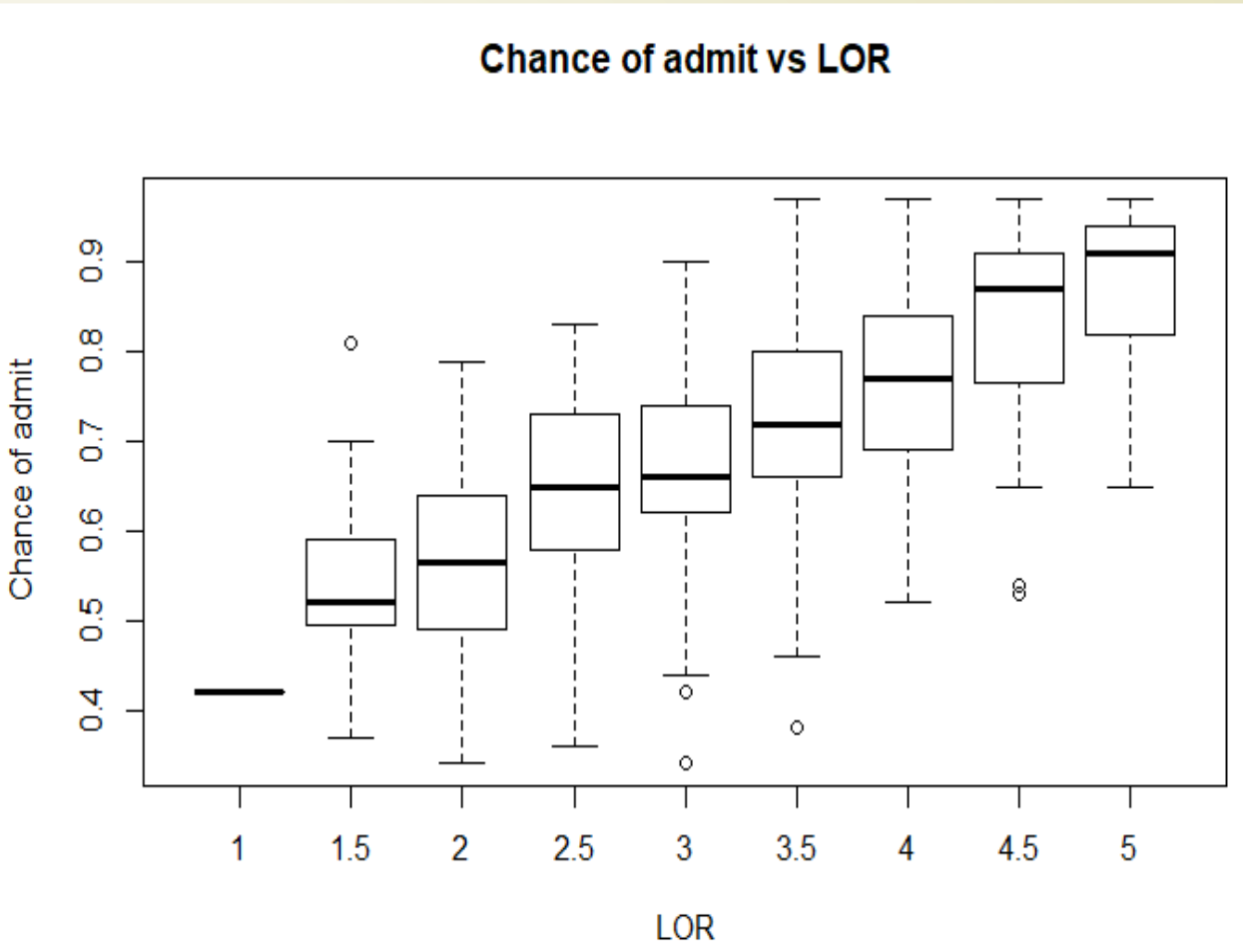
- As with GRE and TOEFL scores, the above box plot also indicates that the Chance of admit doesn't have a linear relationship with CGPA, which implies that an increase in the CGPA score doesn't imply that the chance of admit also increases
- This plot contains the most outliers as the CGPA value ranges from 6.8 to 9.92 with increases of 0.01





- Here you can notice a linear relationship between University ratings, Research and Chance of Admit, which implies that Chance of Admit increases with the increase in value of both these factors
- These plots have very few outliers when compared to GRE, TOEFL and CGPA scores
- The median value of Chance of Admit for Research=1 is ~0.8 (80%) and for Research=0 is ~0.6(60%)
- The median values of Chance of Admit for University ratings of 1,2,3,4 and 5 are ~0.57,0.65,0.72,0.8 and 0.92 respectively





- There is a linear relationship between Chance of Admit and SOP, LOR
- There are very few outliers for both of these factors
- SOP score of 5 has the most outliers
- LOR scores of 3 and 4.5 both have 2 outliers each

MLR Model 1

```
Call:
lm(formula = Chance_of_Admit ~ GRE_Score + TOEFL_Score + University_Rating +
    SOP + LOR + CGPA + Research, data = admit)

Residuals:
      Min        1Q    Median        3Q       Max
-0.266657 -0.023327  0.009191  0.033714  0.156818

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.2757251  0.1042962 -12.232  < 2e-16 ***
GRE_Score          0.0018585  0.0005023   3.700 0.000240 ***
TOEFL_Score        0.0027780  0.0008724   3.184 0.001544 **
University_Rating  0.0059414  0.0038019   1.563 0.118753
SOP                0.0015861  0.0045627   0.348 0.728263
LOR                0.0168587  0.0041379   4.074 5.38e-05 ***
CGPA               0.1183851  0.0097051  12.198  < 2e-16 ***
Research           0.0243075  0.0066057   3.680 0.000259 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

s: 0.05999 on 492 degrees of freedom
Multiple R-squared: 0.8219,
Adjusted R-squared: 0.8194
F-statistic: 324.4 on 7 and 492 DF,  p-value: < 2.2e-16
```

- The original model consists of all the 7 variables
- The standard error value (0.0599/5.99%) and the F-stat values (324.4) indicate that this is a good fit
- The R2 value of the model is 0.822. This value is high enough to indicate that this a good model
- The equation for this model is:

Chance_of_Admit = -1.276 + (0.00186 * GRE_Score + (0.00278 * TOEFL_Score) + (0.006 * University_Rating) + (0.0015 * SOP) + (0.017 * LOR) + (0.118 * CGPA) + (0.0243 * Research)

- I have used an α value of 0.05 for my analysis
- P-value suggests that the variables University_Rating and SOP are insignificant

Confidence Intervals

- The confidence intervals of Chance_of_Admit for a student with the best profile i.e., GRE_Score=340, TOEFL_Score=120, University_Rating=4, SOP=4.5, LOR=4, CGPA=9.91 and Research=1 is:

```
      fit       lwr      upr
0.9853652 0.9700299 1.000701
```

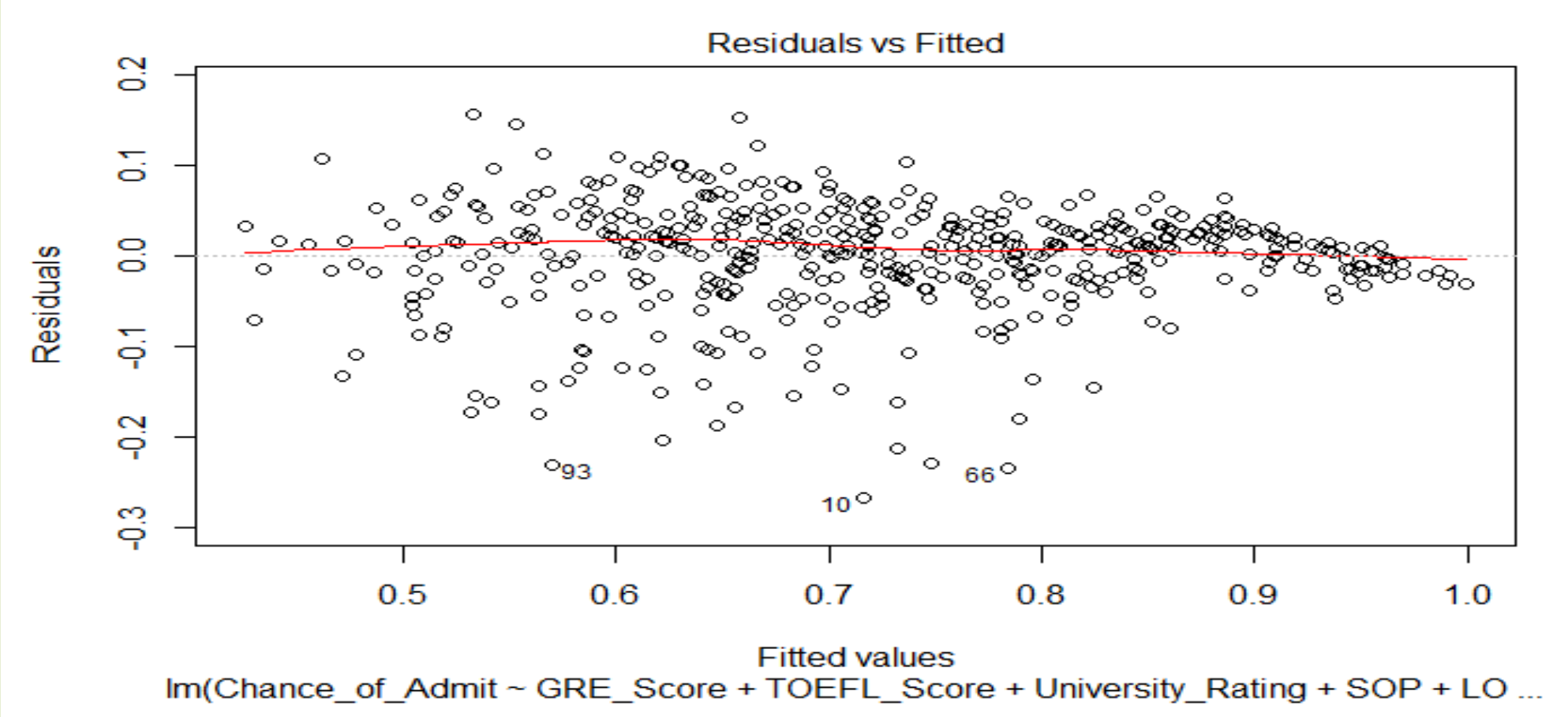
- The confidence intervals of Chance_of_Admit for a student with an average profile i.e., GRE_Score=310, TOEFL_Score=105, University_Rating=3, SOP=3, LOR=3, CGPA=7.5 and Research=0 is:

```
      fit       lwr      upr
0.5531457 0.5356834 0.570608
```

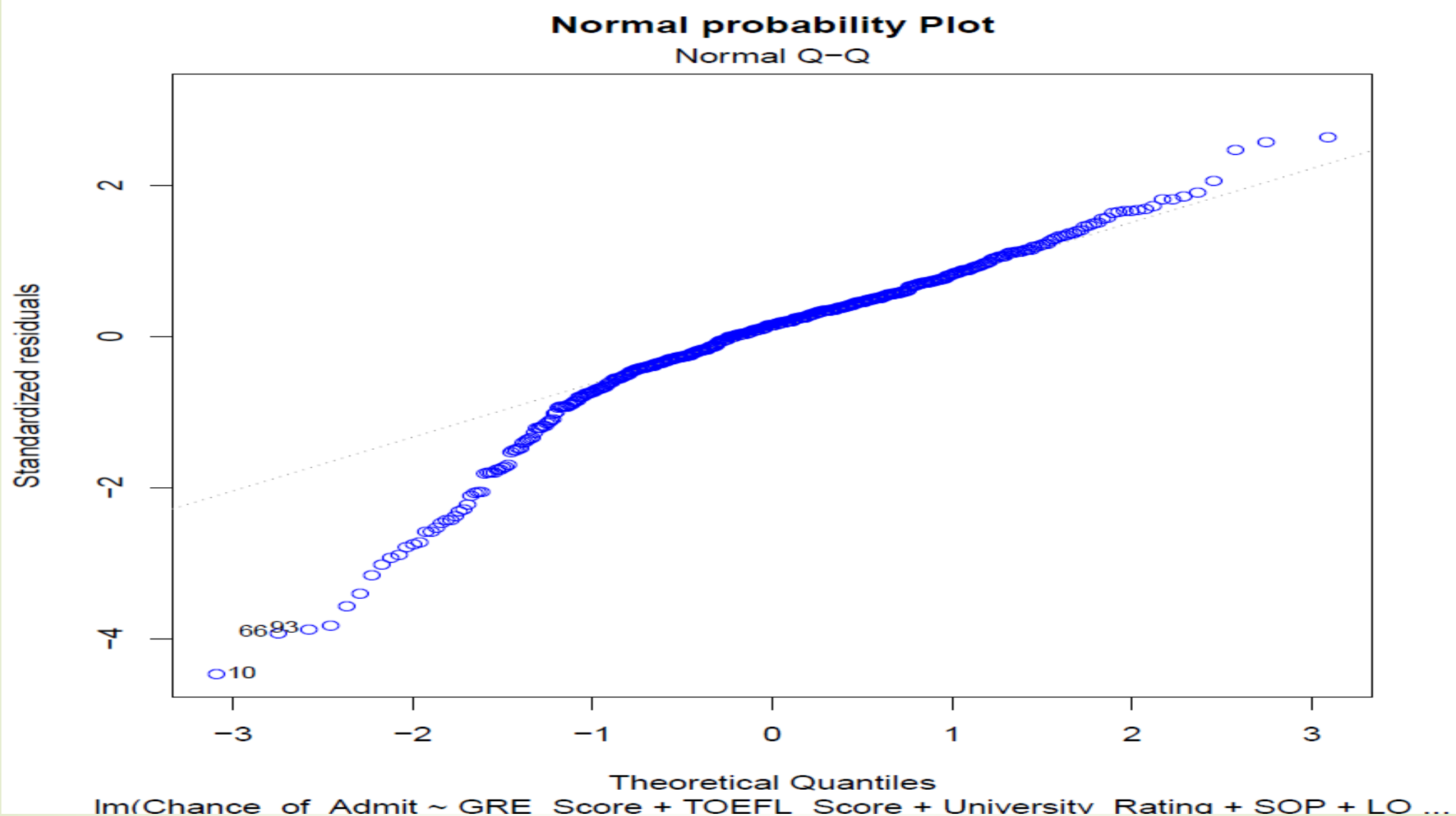
- The confidence intervals of Chance_of_Admit for a student with the worst profile i.e., GRE_Score=290, TOEFL_Score=92, University_Rating=1, SOP=1, LOR=1, CGPA=6.8 and Research=0 is:

```
      fit       lwr       upr
0.3482199 0.3303188 0.3661209
```

Model Analysis

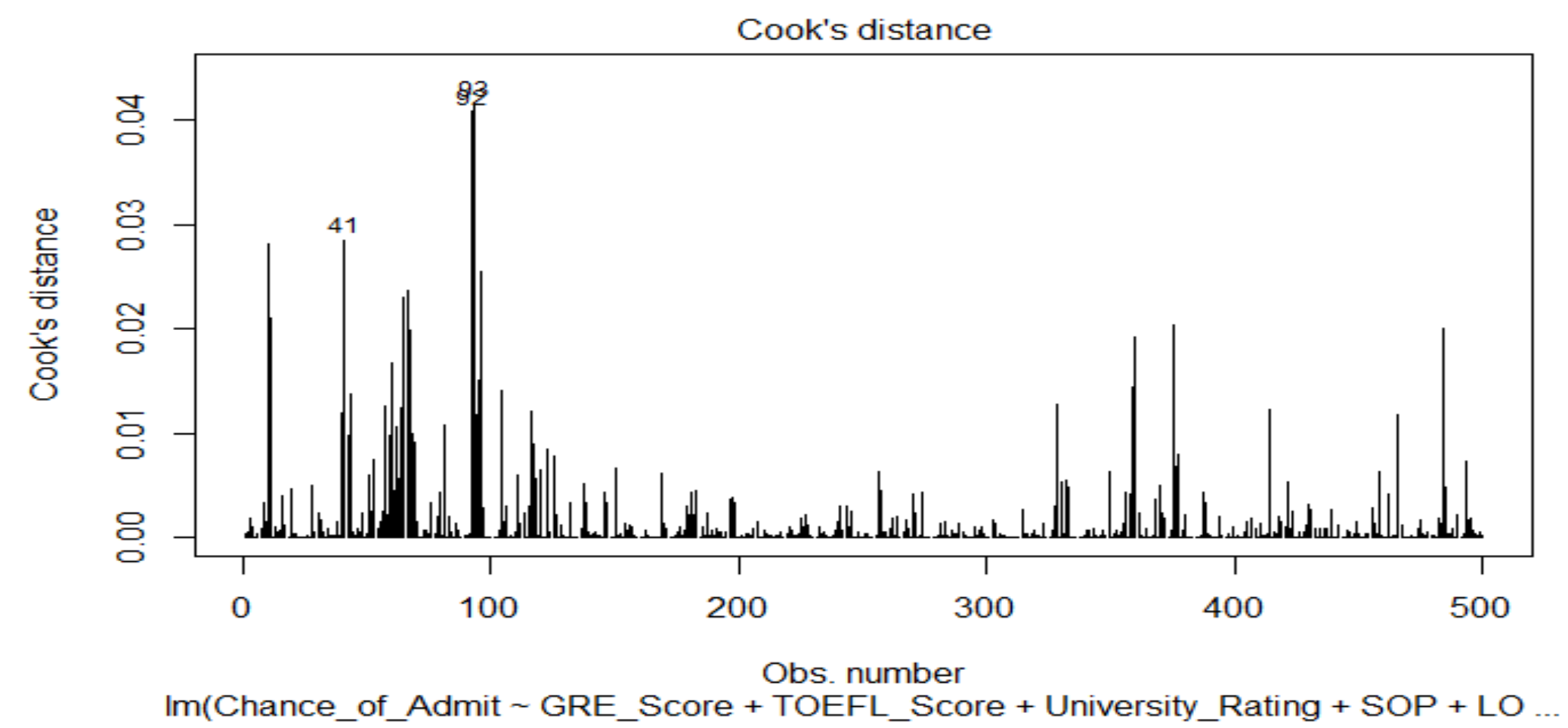


By visual inspection, 3 points (10, 66 and 93) seem to have high values, but as it is not standardized it does not confirm that those are outliers.



- The plot indicates that the data is partially normal, with huge deviations at both ends
- Normal probability plot indicates heavy-tailed distribution as points above the line are in lower percentile and points below the line are in higher percentile which implies that non-linearity might exist

Measure of Influence



- Plotting of Cook's D values indicates that 3 points(41,92 and 93) are in question. Among them points 92 and 93 seem to be outliers
- DFFIT values: From the DFFIT table, I noticed that there are a lot of points that have a DFFIT value more than the cutoff value of 2*(p/n)^0.5 = 0.237. The 3 points mentioned above are also present among these points. Hence, we can deduce that the points 92 and 93 are outliers

Residual Analysis

- R-Student: Alpha=0.05, therefore, t(alpha/2,n-p-1) = t(0.025,492) = ~1.96. This is the cutoff value. Hence, comparing with 1.96, points there are several points (10, 11, 41-43, 414, 387, 375, 376, 377, 360, 359, 150, 116, 104, 92-96,81,60-62 and 64-68) that can be considered as outliers. Cook's D also suggests that the point 41,92 and 93 are outliers.

- Standardized residuals: points 10,11,41,60,65-67 and 93 have a value >3. Hence, these points are outliers.

- Studentized residuals: The points mentioned above are the leverage points, and these are the outliers in the graph.

Model building

- After evaluating all possible models:

(Intercept)	GRE_Score	TOEFL_Score	University_Rating	SOP	LOR	CGPA	Research	SSE	RSQ	adjR2	Cp	BIC
1	1	1	1	0	1	1	1	1.770810	0.8218570	0.8196889	6.120850	-819.0821
1	1	1	1	1	1	1	1	1.770375	0.8219007	0.8193668	8.000000	-812.9903
1	1	1	0	0	1	1	1	1.782708	0.8206601	0.8188449	7.427398	-821.9484
1	1	1	0	1	1	1	1	1.779163	0.8210167	0.8188384	8.442201	-816.7291

- The model obtained by excluding the factor SOP (modl3_temp) and the model obtained by excluding both factors SOP and University_Rating (modl2_temp), both have similar values of adjR2, Mallow's Cp statistic, Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC)
- However, the PRESS stat value of modl2_temp(1.826) is slightly higher than the PRESS stat value of modl3_temp(1.821)
- Hence, I have selected modl2_temp (the model built excluding both SOP and University_Rating) using the above method

- After running the Forward selection, Backward deletion and Stepwise regression model building methods I arrived at the same final model in each case, consisting of the factors GRE_Score, TOEFL_Score, LOR, CGPA and Research
- Factors University_Rating and SOP were removed from the original model as these factors were insignificant

```
Call:
lm(formula = Chance_of_Admit ~ GRE_Score + TOEFL_Score + LOR +
    CGPA + Research, data = admit)

Residuals:
      Min         1Q     Median         3Q        Max
-0.265965 -0.023835  0.008003  0.035543  0.158379

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.3357018  0.0990753 -13.482  < 2e-16 ***
GRE_Score    0.0018892  0.0005024   3.760 0.000190 ***
TOEFL_Score  0.0030174  0.0008619   3.501 0.000506 ***
LOR          0.0193203  0.0037939   5.092 5.04e-07 ***
CGPA         0.1229798  0.0093018  13.221  < 2e-16 ***
Research     0.0251649  0.0065988   3.814 0.000154 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

s: 0.06007 on 494 degrees of freedom
Multiple R-squared: 0.8207,
Adjusted R-squared: 0.8188
F-statistic: 452.1 on 5 and 494 DF,  p-value: < 2.2e-16
```

- The new model has a standard error of 0.06/6% and an R2 value of 0.82. Both these values are similar to the initial model which indicates that this new model is as good as the previous one at predicting values, if not better
- The F-stat value has increased to 452.1
- P-value of all the variables is considerably less than 0.05. This indicates that all the variables in the final model are significant. GRE_Score and CGPA have the least p-value of ~ 2e-16 each. This indicates that these 2 factors are the most influential in determining the Chance_of_Admit
- The equation of the final model is:

Chance_admit = -1.336 + (0.0019 * GRE_Score) + (0.003 * TOEFL_Score) + (0.019 * LOR) + (0.123 * CGPA) + (0.025 * Research)

- CGPA and Chance_of_Admit have the highest correlation of 0.882 followed by GRE_Score and Chance_of_Admit with a correlation of 0.81.

```
> cor(admit$GRE_Score,admit$Chance_of_Admit)
[1] 0.8103506
> cor(admit$TOEFL_Score,admit$Chance_of_Admit)
[1] 0.7922276
> cor(admit$LOR,admit$Chance_of_Admit)
[1] 0.6453645
> cor(admit$CGPA,admit$Chance_of_Admit)
[1] 0.8824126
> cor(admit$Research,admit$Chance_of_Admit)
[1] 0.545871
```

Conclusion

- Chance of admission of a student into a graduate school depends on the student's GRE and TOEFL scores, LOR rating, CGPA and whether the student has done a Research or not
- GRE Score and CGPA are the most important factors among the one's mentioned above
- The Student's SOP and University Rating are not significant factors in determining the chance of a student getting admitted into a graduate school
- Final model is:

Chance_admit = -1.336 + (0.0019 * GRE_Score) + (0.003 * TOEFL_Score) + (0.019 * LOR) + (0.123 * CGPA) + (0.025 * Research)