Predicting Graduate School Chance of Admit

Group 2: Piper Jeong Ho Kim, Zoe Wang, Marie Wiegele, Matthew Woods

Abstract

The relationship between various aspects of graduate school applications were examined in order to determine the best model to predict chance of admission. The question investigated how GRE score, TOEFL score, strength of letter of recommendation and personal statement, undergraduate GPA, research experience, and university ranking affect an individual's chance of admission into graduate school. Five hundred undergraduate students applying to graduate school were examined. Although all the variables were related, undergraduate GPA was most strongly related to chance of admission. The models created show the particular importance of performing well during undergraduate schooling in order to maximize an individual's chances of admission into graduate school.

Method

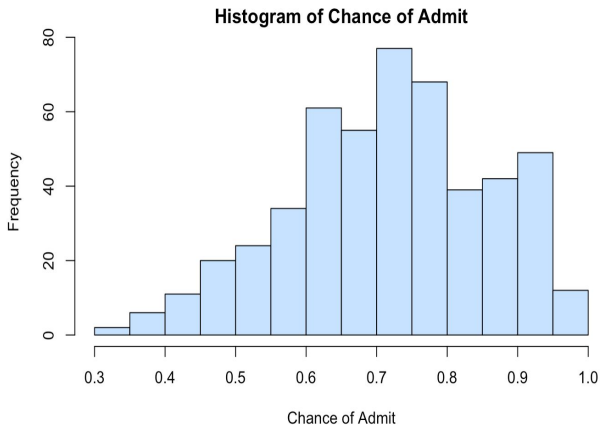
Data

Our data was taken from the given *college* dataset. The data included five hundred observations and eight variables, with each row or observation representing an individual separated by a serial number. No transformations were performed on the data given to us.

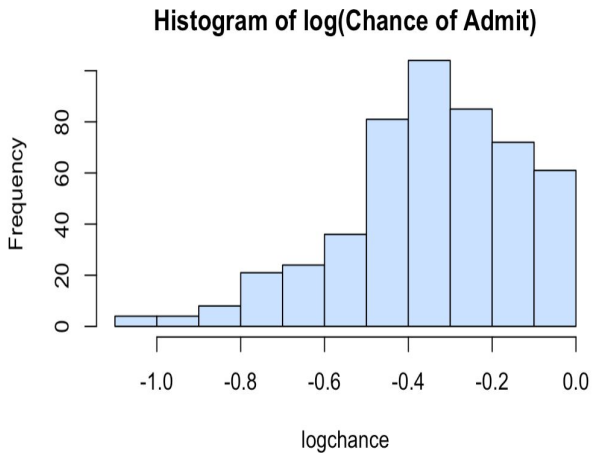
Dependent Variable

The dependent variable chosen was Chance.of.Admit. Chance.of.Admit represents a candidate's chances of gaining admission into graduate school. This is represented numerically as a decimal. The mean chance found was 0.7214, and the range of values in our dataset spanned from 0.34 to 0.97. Upon looking at the histogram provided in Figure 1, it can be seen that

Chance of Admit was slightly skewed left. In an attempt to fix this and to make it the most normal it could be, we applied a log transformation. However, this only made the resulting data more skewed, as can be seen in Figure 2.



(Figure 1: Histogram before transformation)



(Figure 2: Histogram after log transformation)

Independent Variables

GRE Score is a numerical predictor with a maximum value of 340. This represents a student's GRE score, or the combined total of the quantitative and verbal sections of the exam. TOEFL Score is a numerical predictor with a maximum value of 120. University Rating is a categorical variable with five levels, which represents the undergraduate university of a given candidate's rating based on the perspective of a particular Indian student. A top tier university is coded as 5, an above average university as a 4, an average university as a 3, a below average university as a 2, and a weak tier university as a 1. Statement of Purpose, or SOP, is a numerical value out of five which indicates the strength of a student's statement of purpose. In this case, five is the strongest, and zero indicates weakest. Similarly, Letter of Recommendation, or LOR, is a numerical variable of a value ranging from zero to five, which represents the strength of a student's letter of recommendation. Again, five represents the strongest and zero weakest. Undergraduate GPA, or CGPA, is a numerical predictor with a range of zero to ten. Unlike

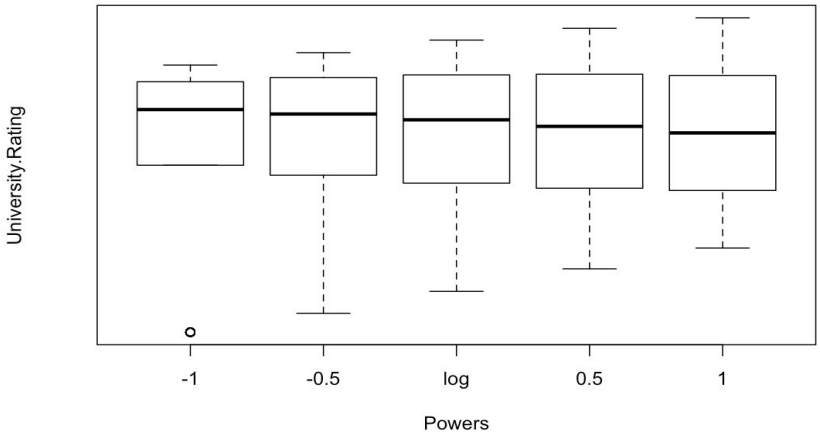
traditional 4.0 scale GPA, this is an average calculated based on the scale shown in Table 1. Lastly, Research Experience is a categorical variable with two levels which represent whether or not a student has prior research experience. A response of "Yes" is coded as 1, while a response of "No" is coded as 0.

Grade	Grade Point Equivalent
A	9-10
B	8-8.9
C	7-7.9
D	6-6.9
F	Below 6

(*Table 1: CGPA conversion chart)*

Transformations

When looking at the boxplots of each predictor and the outcome variable, we noticed that most variables did not need a transformation in order to be approximately normally distributed. The only variable's transformation which positively affected the R^2 and Adjusted R^2 for our standardized model was a log transformation of University Rating.

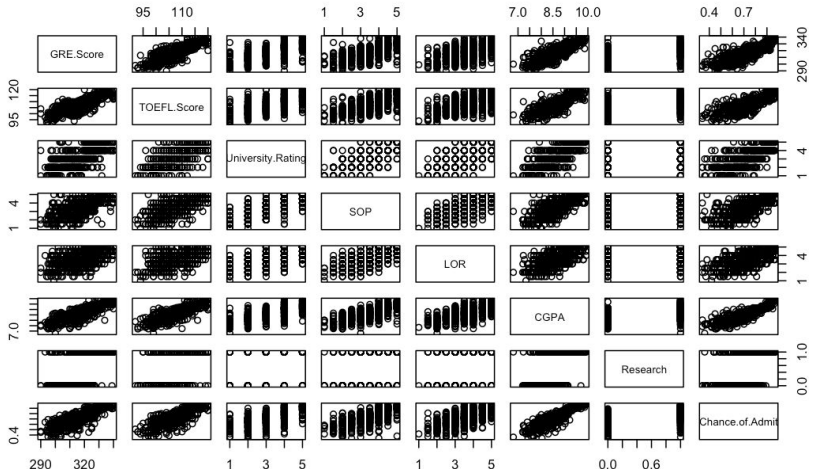


(Figure 3: Symbox boxplots for University.Rating)

Exploratory Data Analysis

Scatterplot Matrix

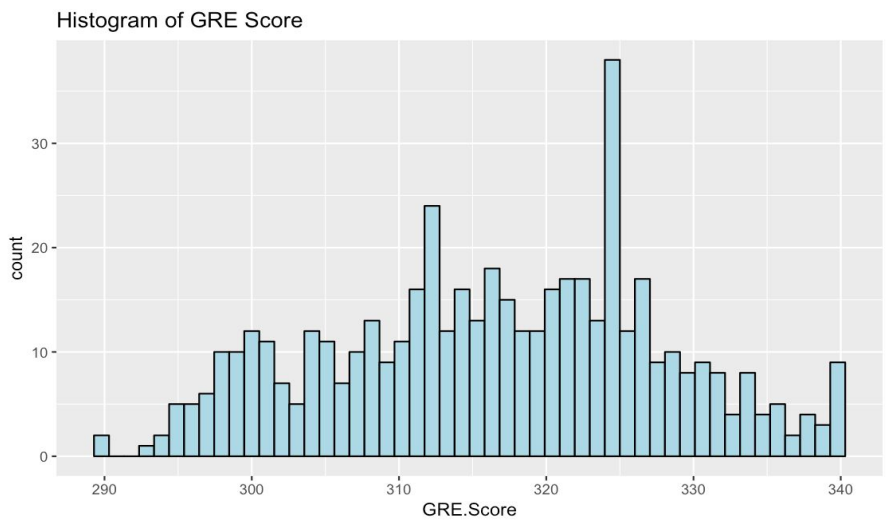
The scatterplot matrix of the variables presented in Figure 3 shows the linearity between variables. The only exception is Research, as it is binary.



(Figure 4: Scatterplot matrix of the variables)

GRE Score

The GRE Score data is approximately normally distributed with no obvious skew, as seen in Figure 4. Based on Table 2, as a student's chance of admission (the dependent variable) increases, the average GRE Score increases as well, indicating a direct relationship between the two variables.



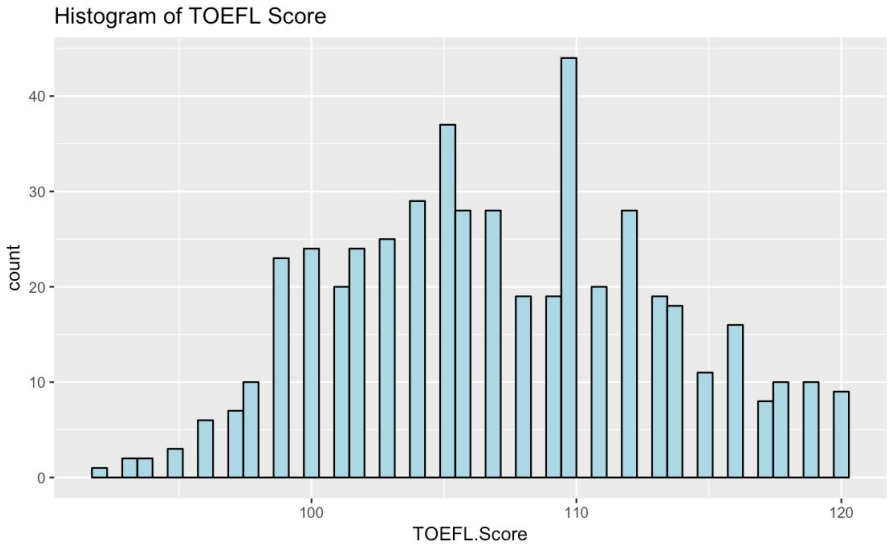
(*Figure 5: Histogram of GRE Score)*

Chance of Admit	34% - 62.9% (1st quantile)	63% - 72.169% (2nd quantile)	72.17% - 81.9% (3rd quantile)	82% - 97% (4th quantile)
Average GRE Score	305.3417	310.9310	319.0305	328.8271

(Table 2)

TOEFL Score

The histogram of TOEFL Score looks approximately normally distributed with no obvious skew, as shown in Figure 5. According to Table 3, TOEFL scores follow a similar pattern as GRE scores in that it has a direct relationship with Chance of Admit. As students' chance of admit increases, their average TOEFL score increases as well.



(Figure 6: Histogram of TOEFL Score)

Chance of Admit	34% - 62.9% (1st quantile)	63% - 72.169% (2nd quantile)	72.17% - 81.9% (3rd quantile)	82% - 97% (4th quantile)
Average TOEFL Score	101.7083	104.2155	107.8702	114.0677

(Table 3)

University Rating

When comparing the average chance of admit, average GRE score, average TOEFL, and average GPA of the students who applied to each university rating, it is obvious that those who applied from top level schools (University Rating of 4 or 5) had higher averages in measures of academic aptitude, which appears to lead to higher chances of admit, as seen in the tables below.

University Rating	1	2	3	4	5
Average Chance of Admit	56.21%	62.61%	70.29%	80.16%	88.81%

(Table 4: University ratings vs. their average chance of admit)

University Rating	1	2	3	4	5
Average GRE Score	304.91	309.13	315.03	323.30	327.89

(Table 5: University ratings vs. their average GRE score)

University Rating	1	2	3	4	5
Average TOEFL Score	100.21	103.44	106.31	110.96	113.44

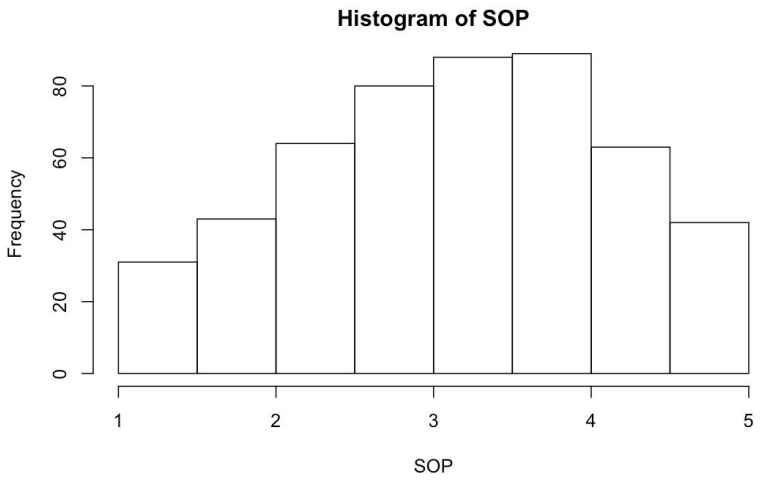
(Table 6: University ratings vs. their average chance of admit)

University Rating	1	2	3	4	5
Average GPA	7.798529	8.177778	8.500123	8.936667	9.278082

(Table 7: University ratings vs. their average chance of admit)

Statement of Purpose

The distribution of SOP strength looks to be approximately normal with a slight right skew, as seen in Figure 6. Table 8 shows that stronger SOPs have higher chances of admission.



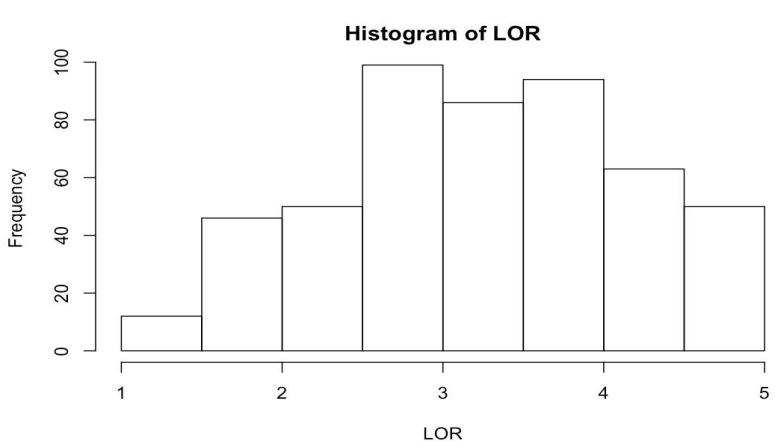
(Figure 7: Histogram of Statement of Purpose)

Chance of Admit	34% - 62.9% (1st quantile)	63% - 72.169% (2nd quantile)	72.17% - 81.9% (3rd quantile)	82% - 97% (4th quantile)
Average SOP	2.48	3.05	3.48	4.36

(Table 8: Average SOP per quantile of Chance of Admit)

Letter of Recommendation

As the strength of a student's letter of rec goes up, their chance of admit does as well - as seen by the average LOR strength increasing as the chance of admit increases in Table 9.



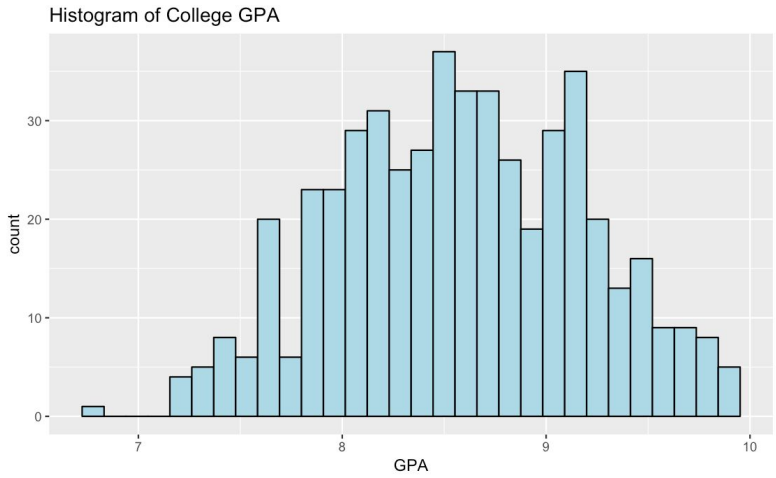
(Figure 8: Histogram of Letter of Recommendation)

Chance of Admit	34% - 62.9% (1st quantile)	63% - 72.169% (2nd quantile)	72.17% - 81.9% (3rd quantile)	82% - 97% (4th quantile)
Average LOR	2.68	3.28	3.58	4.29

(Table 9: Average LOR by quantiles of Chance of Admit)

GPA

The histogram of college GPA looks normal, as seen in Figure 8. Table 10 and Table 11 show that as chance of admit increases, so does average GPA. Additionally, those with higher GPAs have higher average chance of admit.



(Figure 9: Histogram of College GPA)

Chance of Admit	34% - 62.9% (1st quantile)	63% - 72.169% (2nd quantile)	72.17% -81.9% (3rd quantile)	82% - 97% (4th quantile)
Average GPA	7.91	8.34	8.65	9.31

(Table 10: Average GPA per quartile of Chance of Admit)

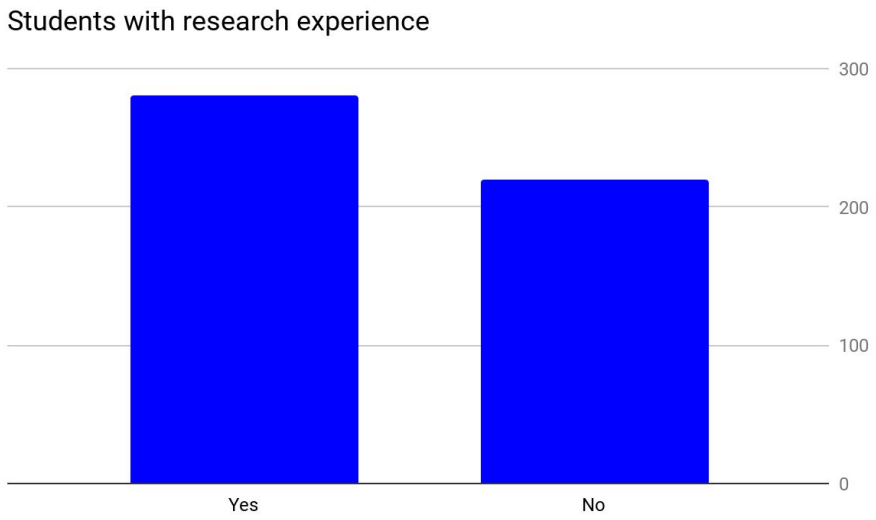
GPA	F: Below 6	D: 6-6.9	C: 7-7.9	B: 8-8.9	A: 9-10

Average Chance of Admit	N/A	36.00%	53.13%	68.66%	87.18%

(Table 11: Average Chance of Admit by Undergraduate GPA)

Research

Students with research experience have, on average, over 15% higher chance of admission than students without, as shown in Table 12.

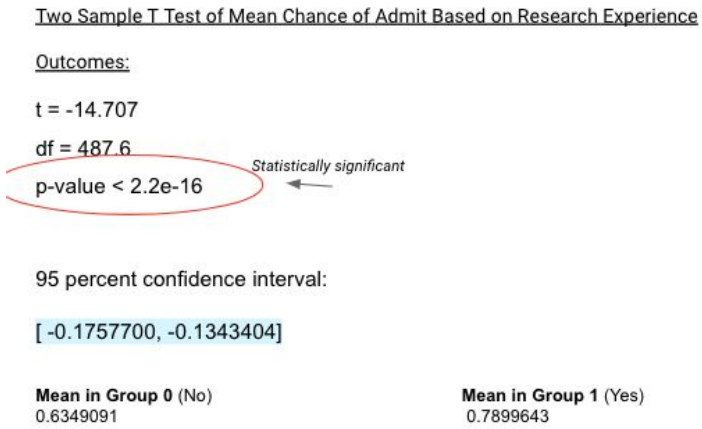


(Figure 10: Frequencies of each level of Research)

Research	Yes	No
Average Chance of Admit	79.00%	63.49%

(Table 11: Average Chance of Admit per level of Research)

In order to ascertain that this difference is significant, we conducted a 2 sample t-test. According to the two sample t test of the means (of chance of admit based on the two research groups yes or no), it was shown that the difference in means among the two groups was statistically significant. Figure 10 shows the results of our test.



(Figure 11: Two Sample T-test)

Correlation Matrix

As shown by the correlation matrix in Figure 11, there are pretty high correlations among most of the variables-- indicating a high amount of multicollinearity. This makes sense considering our variables are measures of academic success, which tend to have high correlations with one another. Our model attempts to account for multicollinearity by standardizing the variables.

	GRE Score	TOEFL Score	University Rating	SOP	LOR	GPA	Research	Chance of Admit
GRE Score	1	0.83	0.64	0.61	0.52	0.83	0.56	0.81
TOEFL Score		1	0.65	0.64	0.54	0.81	0.47	0.79
Univeristy Rating			1	0.73	0.61	0.71	0.43	0.69
SOP				1	0.66	0.71	0.41	0.68
LOR					1	0.64	0.37	0.65
GPA						1	0.5	0.88
Research							1	0.55
Chance of Admit								1

(Figure 12: Correlation matrix of the unstandardized variables)

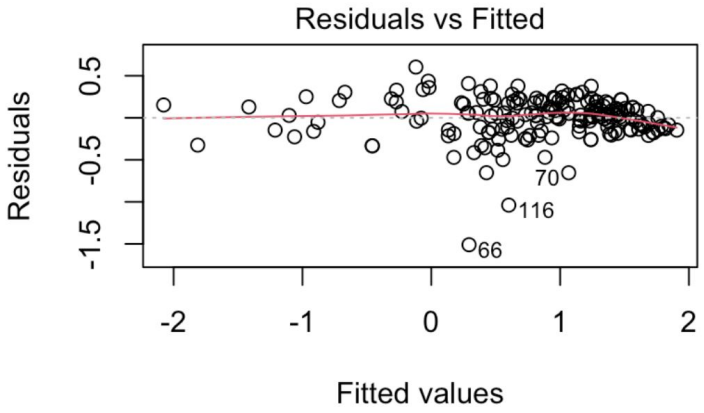
Final Model

$$\text{Z Chance of Admit} = \hat{\beta}_0 + \hat{\beta}_1(\text{Z GRE Score}) + \hat{\beta}_2(\log(\text{University.Rating})) + \hat{\beta}_3(\text{Research}) + \hat{\beta}_4(\text{Z TOEFL.Score}) + \hat{\beta}_5(\text{Z CGPA}) + \hat{\beta}_6(\text{Z LOR}) + \hat{\beta}_7(\text{Z SOP}) + \hat{\beta}_8(\text{GRE.Score: SOP})$$

After review of the exploratory data analysis, we decided on our final model, as shown above. In this model, we predict standardized Chance of Admit with the standardized variables GRE Score, TOEFL Score, GPA, Letter of Recommendation, and Statement of Purpose. The other explanatory variables are the log transform of University Rating, the binary variable Research, and the interaction effect between GRE Score and Statement of Purpose.

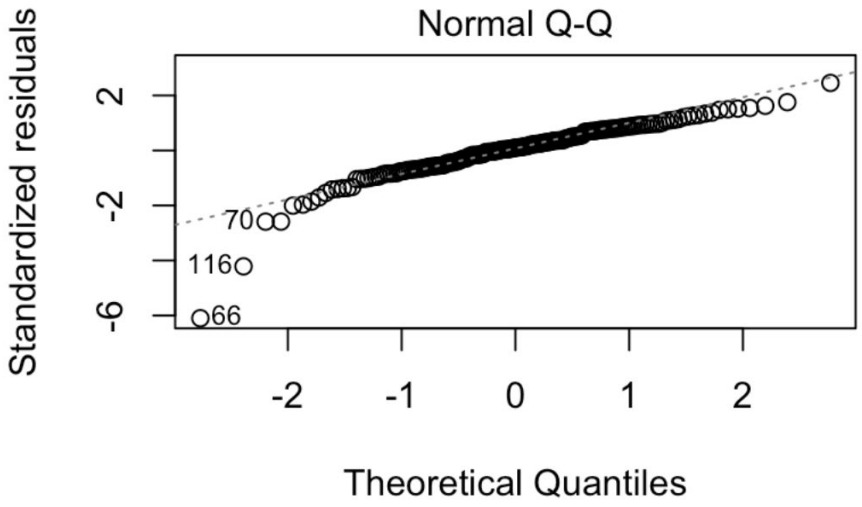
Model Assumptions

Constant Variance



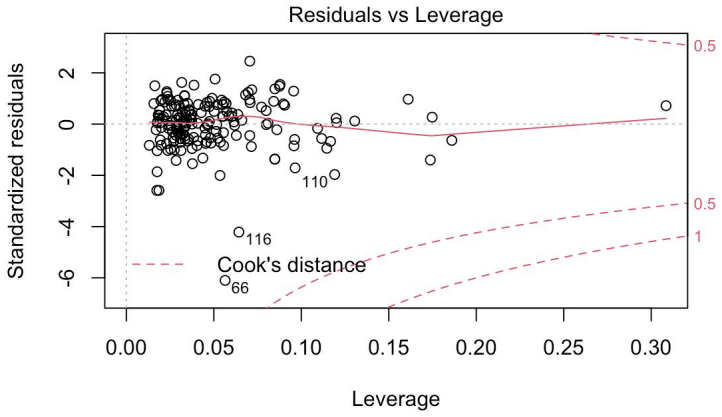
The residuals follow a clear curved pattern which indicates that the error variance of the model is not constant. Furthermore, the p-value of the ncv test for this model was equal to 1.5205e-7. Therefore, we reject the null hypothesis of constant error variance and conclude that the model does not meet the assumption of constant error variance.

Normality of Errors



The normal Q-Q plot very clearly shows that, apart from a few outliers, the standardized residuals plotted against quantiles lie on a straight line. This indicates that the model meets the assumption of normality of errors.

High Leverage Points



The model performs well here. No errors are above Cook's distance. There are only two bad leverage points whose standardized residuals fall outside of the acceptable range of (-2 to 2), which represent a very small portion of our total 500 observations.

Multicollinearity

Variable	VIF
GRE Score	4.376733
log(University.Rating)	1.185065
TOEFL Score	3.233684
Research	1.472924
Statement of Purpose	1.845045
Letter of Recommendation	1.648822
GPA	4.283192
GRE Score:SOP	1.634008

(*Table 12: VIFs of the predictors)*

All of the variables have a VIF less than 5, as seen in Table 12. This indicates that there is only moderate multicollinearity present in the model. Thus, our model performs very well by this metric.

Results

The Adjusted R^2 value for this model was 0.8938. This means that approximately 89% of the observed variation can be explained by the model's inputs (modified for number of predictors). The R^2 before adjustment is 0.8986, and the small difference between the two can be an indication that this model does not overfit. Additionally, the AIC of our model was 29.2061. This low value indicates that overfitting is unlikely. All of our variables had p-values close to zero, indicating that they are all significant predictors. As for the variables, since our variables are standardized, we were able to compare the magnitudes of the coefficients of the variables to see which one had the largest effect, as presented in Table 13. Since GPA had the largest magnitude, we can say that the GPA was the most significant variable. The significance of

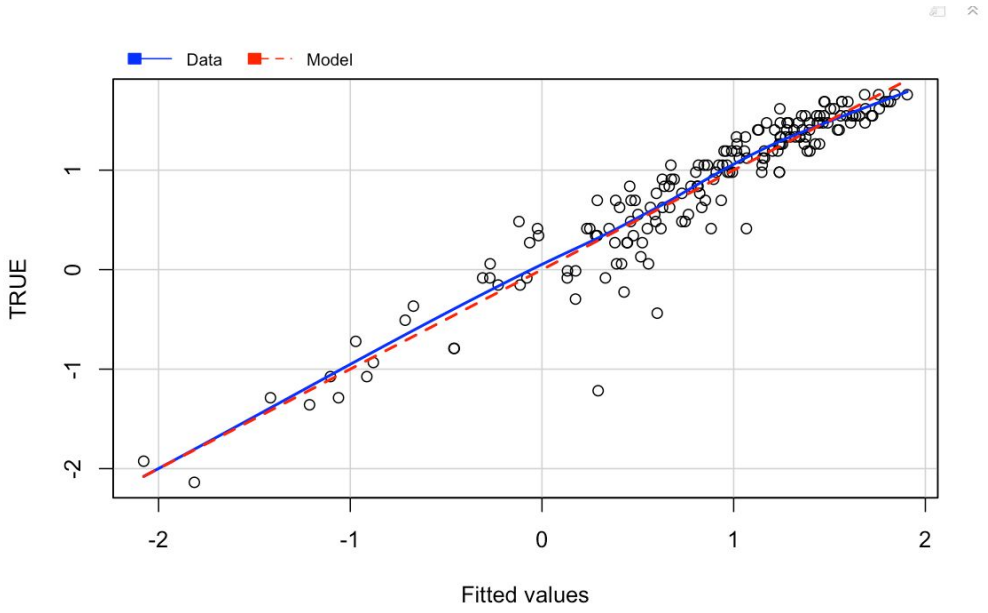
the untransformed coefficients means that for 0.16296 increase in GPA, Chance of Admit increases by 1. However, this does not make sense in context of what Chance of Admit represents, so to scale it back one decimal, an increase of 0.016 in GPA corresponds with an increase of 0.1 in Chance of Admit. This applies for the other untransformed numerical predictors and their respective coefficients. For Research, the coefficient signifies the difference in Chance of Admit between the two groups. We would expect those who did research to have a 0.154 higher chance than those who did not have prior research experience. For the interaction effect, because the coefficient is statistically significant, the SOP scores had a decreasing effect on GRE scores because the partial slope is negative.

	Estimate	Std. Error	t value	Pr (>|t|)
(Intercept)	-0.01770	0.03274	-0.540	0.589581
GRE.Score	0.16296	0.04661	3.497	0.000602
log(University.Rating)	0.16161	0.05596	2.888	0.004383
Research	0.15412	0.03000	5.138	7.59e-07
TOEFL.Score	0.22927	0.04184	5.480	1.52e-07
CGPA	0.35174	0.04890	7.193	1.96e-11
LOR	0.06072	0.03017	2.013	0.045741
GRE.Score: SOP	-0.09426	0.02439	-3.865	0.000158

(*Table 13: Summary table for MLR Model)*

MMP plot

The marginal model plot of our model shows the LOESS line and the model are virtually indistinguishable, indicating that our model is a good fit for E(Y|x)



Cross Validation

Unfortunately the R Code was able to run a correlation test in order to find the strength of the correlation between our y hat values and our y values. Without functioning code, we are unable to comment on the strength of our prediction, however we feel the statistical significance of our coefficients, mmp plot, R^2 and R^2 Adjusted adequately reflects the strength of the model.

```
trainingsample <- collegedata[sample(collegedata$Serial.No., length(collegedata$Serial.No.)/2),]
otherhalf <- collegedata[!(collegedata$Serial.No. %in% trainingsample$Serial.No.),]


attach(trainingsample)
CVmodel1 <- lm(Chance.of.Admit~ GRE.Score + TOEFL.Score + CGPA +Research + log(University.Rating) + LOR
+ SOP + GRE.Score:SOP, data = trainingsample)

summary(CVmodel)


yhat = -1.0658561 + GRE.Score*0.0008599 + TOEFL.Score*0.0048620 + CGPA*0.1021322  + Research
  *0.0226619 +  log(University.Rating)*0.0018402+ LOR*0.0215732 + SOP*-0.0620822+
GRE.Score:SOP*0.0002174

cor(yhat, Chance.of.Admit, use = "complete.obs")
```

```
Error in cor(yhat, Chance.of.Admit, use = "complete.obs") :
  incompatible dimensions
```

R output for attempt at cross validation

Logistic Model

For the Logistic Model, the predicted variable of "Chance of Admit" was coded into a binary variable with a chance of admission below the median (72%) being coded as "0" and a chance admission above the median being coded as "1". The model includes undergraduate GPA (numerical) and Research (categorical). Table 14 presents the model's output.

Coefficients:	**Estimate**	**Std. Error**	**Z Value**	**P Value**
Intercept	-39.0990	3.7163	-10.521	<2e^-16***
Undergraduate GPA	4.5061	.4357	10.343	<2e^-16***

Research1	1.3934	.2841	4.951	7.38e-07 ***

Table 14: R output of Logistic Model

Interpretation

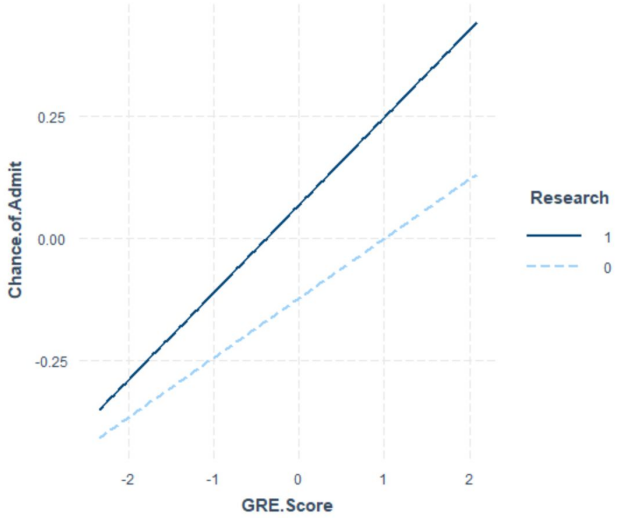
The coefficient of 4.5061 for Undergraduate GPA means that for each one unit increase in Undergraduate GPA, the log odds of a chance of admission above 72% is 4.2433-- which comes out to be 90.56791. The coefficient of 1.3934 for Research1 means that keeping all else constant, for students with research experience, the log(odds ratio) of having a chance of admission above the median is 1.3934-- which means that the odds of having a chance of admission above 72% vs. below 72% with research experience is 4.028524.

Conclusion

We found evidence to support that all of the given aspects of the graduate school application process were important in a candidate's chance of admission. Our investigation shows that of the independent variables, the most important in maximizing chance of admission was undergraduate GPA. Additionally, we saw that chances of admission greatly differ between those who had prior research experience and those who did not as GRE score increased. In context, this means that the better a candidate performs on the GRE exam, the more they benefit from research experience in regards to their chance of admission. This can be seen in Figure 15, which shows the slopes of the two lines representing the levels of research experience diverging gradually as GRE score increases.

A shortcoming of the study was our inability to uphold the assumption of constant variance. A possible improvement could be to try to eliminate skewness in Chance of Admit somehow, although transforming it did not help. Additionally, we could obtain more predictors by looking at other aspects of the application process, such as strength of essays other than the

personal statement. There is also some bias in the University Rating variable as it is not based on statistics but instead on the opinion of one particular student. Thus, it could be improved by being calculated through the use of information such as average salary of graduates or the university's admission rates. Due to the fact that we were unable to troubleshoot the code for running cross validation, we were unable to examine the strength of the model as thoroughly as we would have liked. Finally, since there are very few students falling in the category of 1 for University Rating, we could pool together levels 1 and 2 into one factor in order to make the distribution more normal.



(Figure 17: Interaction between GRE Score and Research)

Appendix

P

	GRE.Score	TOEFL.Score	University.Rating	SOP	LOR	CGPA	Research	Chance.of.Admit
GRE.Score		0	0	0	0	0	0	0
TOEFL.Score	0		0	0	0	0	0	0
University.Rating	0	0		0	0	0	0	0
SOP	0	0	0		0	0	0	0
LOR	0	0	0	0		0	0	0
CGPA	0	0	0	0	0		0	0
Research	0	0	0	0	0	0		0
Chance.of.Admit	0	0	0	0	0	0	0	

(Figure A: R output for p-values of correlation matrix)

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

Residuals:
     Min       1Q   Median       3Q      Max
-1.50984 -0.13645  0.02455  0.17395  0.60356

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           -0.01770    0.03274  -0.540 0.589581
GRE.Score              0.16296    0.04661   3.497 0.000602 ***
log(University.Rating)  0.16161    0.05596   2.888 0.004383 **
Research               0.15412    0.03000   5.138 7.59e-07 ***
TOEFL.Score            0.22927    0.04184   5.480 1.52e-07 ***
CGPA                   0.35174    0.04890   7.193 1.96e-11 ***
LOR                    0.06072    0.03017   2.013 0.045741 *
SOP                    0.23267    0.03665   6.349 1.93e-09 ***
GRE.Score:SOP         -0.09426    0.02439  -3.865 0.000158 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2548 on 169 degrees of freedom
  (322 observations deleted due to missingness)
Multiple R-squared:  0.8986,    Adjusted R-squared:  0.8938
F-statistic: 187.1 on 8 and 169 DF,  p-value: < 2.2e-16
```

(Figure B: R output of summary for MLR model)

```
Call:
glm(formula = COABinary ~ CGPA + Research, family = "binomial")

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.77933  -0.47208   0.07184   0.39097   2.27476

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -39.0990     3.7163 -10.521  < 2e-16 ***
CGPA          4.5061     0.4357  10.343  < 2e-16 ***
Research      1.3934     0.2814   4.951 7.38e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 691.58  on 499  degrees of freedom
Residual deviance: 323.24  on 497  degrees of freedom
AIC: 329.24

Number of Fisher Scoring iterations: 6
```

(Figure C: R output of summary for Logistic Model)