Statistical Analysis of Admission into Statistics Graduate Program

MATH 167R Final Project Report

By: Amelia Le

Introduction and Background

Individuals who are interested in pursuing higher educations are often interested in the decision-making process that admission committee goes through to decides which applicants would be a best fight for their graduate program. Often time, there are several factors that influences the committee decisions to admit an applicant over another. Admission committee looks at different attributes such as: GRE, GPA, recommendation letters, and work experience when reviewing a potential candidate's application[1] and each factor is weighted differently depending on the program's requirements. To better understand how individuals' attributes affect their admission chances, this final project aims to answer the following research questions:

1. Is there a significant difference in the average GPA and test scores of applicants based on their citizenship status? In addition, is there a significant difference in the proportion of foreign applicants who were accepted to one of the top 10 statistics PhD programs in America compared to their American counterpart?
2. What are the most influential factors that affects an applicant's admission chance?
3. Can we build a linear regression model that can successfully predict an individual's admission chance?

Two different datasets on graduate program admissions were used to answer these research questions. The first dataset contains the information on individuals who applied to a statistics graduate program and posted about their experience onto the website GradCafe between the year 2011 and 2019. This dataset contains reported information about the applicants' GPA (4.0 scale), GRE scores (out of 170 or 6), institutions they applied to, the year they applied (2011-2019), their application status (Accepted/Rejected), the degree type (masters or PhD), and their status (American, International, or International with a US degree). The GradCafe statistics csv file[2] can be downloaded from the Gradcafe-Admissions-Data GitHub respiratory that contains a collection of admission data on a multitude of different programs collected from the GradCafe. The second dataset is from the "A Comparison of Regression Models for Prediction of Graduate Admissions"[3] paper published in 2019 in the IEEE journal. The admission dataset can be retrieved from Kaggle under the name " Graduate Admission 2"[4]. The admission dataset contains information on the GPA, GRE score, TOEFL score, Strength of Statement of Purpose (SOP), Strength of Letter of Recommendation (LOR), and undergraduate university ranking (Tier 1 to Tier 5), and the chance of admission to a particular master program (0-1) for 500 anonymous applicants. This dataset was constructed from the perspective of an Indian foreign student who want to predict the admission into a graduate program. The creator of the admission dataset does not tell us which master program these anonymous applicants are applying to nor does it tell us how the chance of admission are calculated. Therefore, the inference on the influence of any predictor variables would be skewed toward the perspective of the individual who created the dataset rather than what is true.

---

[1] Kurt, Daniel. "What Counts When Applying to Grad School: GPA vs. Work Experience." *Investopedia*, Investopedia, www.investopedia.com/articles/personal-finance/020315/applying-grad-school-gpa-vs-work-experience.asp. Accessed 9 Dec. 2023.
[2] Evans, James. "Gradcafe-Admissions-Data." *GitHub*, 21 Aug. 2020, github.com/evansrjames/gradcafe-admissions-data.git. Accessed 4 Dec. 2023.
[3] M. S. Acharya, A. Armaan and A. S. Antony, "A Comparison of Regression Models for Prediction of Graduate Admissions," 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 2019, pp. 1-5, doi: 10.1109/ICCIDS.2019.8862140.
[4] Acharya, Mohan S. "Graduate Admission 2." Kaggle, 28 Dec. 2018, www.kaggle.com/datasets/mohansacharya/graduate-admissions.

Exploratory Data Analysis

A) Grad Café Dataset

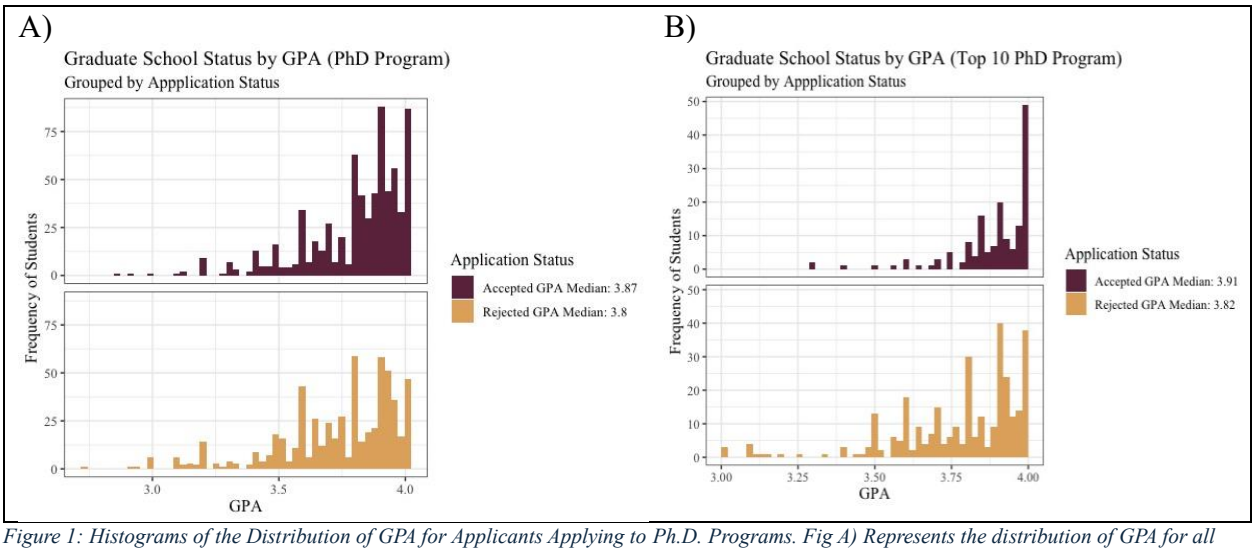

Figure 1: Histograms of the Distribution of GPA for Applicants Applying to Ph.D. Programs. Fig A) Represents the distribution of GPA for all applicants who applied to a Statistics Ph.D. program from 2011 to 2019. Fig B) Represents the distribution of GPA for all applicants who applied to one of the Top 10 Statistics Ph.D. programs.

In Figure 1 above, both histograms show a strong skewness to the left, where most PhD applicants possess a competitive GPA that is closer to the sample median compared to the number of applicants who had an undergraduate GPA below a 3.75. This skewness in the number of applicants based on GPA could mean several things. The first explanation could be that individuals with GPA below 3.75 are less likely to apply to a PhD program due to either feeling under-prepared or not competitive enough to apply. This would explain the smaller number of applicants in that GPA group. The second explanation could be that the sampling pool is highly biased since it is sampled from a pool of applicants who voluntarily submitted their statistics onto GradCafe. Therefore, there is a high probability that the GradCafe dataset reflects individuals that can be classified as competitive applicants. As a result, the distribution of individuals with certain GPA below the sample median are not being accurately represented in the histograms above for the year 2011 to 2019.

Most interestingly, we see extreme outliers in both the histograms for accepted PhD applicants, where individuals with uncharacteristically low GPA were successfully accepted into a PhD program during their application cycles. Table-1 below contains information on the testing scores and PhD programs of two sets of three individuals with the lowest GPA in their respective pool of applicants. When looking at individuals with the lowest GPA accepted to a doctorate program, two of the three individuals have GPAs below the 3.0 cutoff requirement that most graduate programs imposes. Contrastingly, we see that their GRE scores either fall within or above the expected percentile for applicants within their field of graduate study. However, lack the information on other attributes about the applicants that could offset their low GPA point average such as: work experience, letter of recommendation, and other notable achievements[5] that the admission committee took into consideration when admitting the applicants.

[5] Barr, Philippe. "Does GPA Matter for Grad School Applications?" *Get The Insider's Edge On Graduate School Admissions* -, 23 Oct. 2023, admit-lab.com/blog/does-gpa-matter-for-grad-school-applications/.

Table-1: Outliers Applicants Accepted to a Ph.D. Program from 2011 to 2019

All Statistics Ph.D. Programs							Top 10 Statistics Ph.D. Programs							
Institution	**Year**	**GPA**	**GRE Verbal**	**GRE Quant**	**GRE Writing**	**Program Rank**	**Institution**	**Year**	**GPA**	**GRE Verbal**	**GRE Quant**	**GRE Writing**	**Program Rank**	**Program Averages**
University Of Iowa College Of Education	2016	2.9	167	160	5.5	30th[6]	University Of Michigan, Ann Arbor	2018	3.3	163	158	4	7th[7]	None
Rutgers University	2013	2.9	162	158	3.5	24th	Duke University	2019	3.3	157	170	None	7th	Verbal: 162 Quant: 167 GPA: 3.7[8]
Johns Hopkings(JHU)	2012	3.0	800	730	3.5	None	University Of Michigan, Ann Arbor	2018	3.4	162	163	4	7th	None

B) Chance of Admission Dataset

Since the purpose of the admission dataset is to construct a linear regression model to make future prediction on future observations and make inferences on key features that impacts an individual's chance of admission, we are interested in observing whether the data satisfy the linear regression assumptions of linear independency, constant variance, and normality.

In Figure 2 below, we see from the scatterplots that our response variable, chance of admission, have a positive sigmoidal relationship with each of the quantitative predictors in the data. The scatterplots show that the average chance of admission increases proportional to the predictor variables until an inflection point is reached, at which point the average chance of admission began to level off as it approached 100%. Since we want to fit a linear regression model using this data, we will have to transform either the response or predictor variables to address the non-linearity relationship between the two variables. In addition, the histogram shows that on average the chance of admission to this master program is higher for individuals with research experience than those who do not. This shows that research experience is an important factor that admission committee consider when deciding between applicants who have similar academic background.

[6] Best Statistics Graduate Programs - Top Science Schools - US News Rankings, www.usnews.com/best-graduate-schools/top-science-schools/statistics-rankings. Accessed 8 Dec. 2023.
[7] Best Statistics Graduate Programs - Top Science Schools - US News Rankings, www.usnews.com/best-graduate-schools/top-science-schools/statistics-rankings. Accessed 8 Dec. 2023.
[8] The Graduate School. "All Departments: Phd Admissions and Enrollment Statistics." *The Graduate School*, gradschool.duke.edu/about/statistics/all-departments-phd-admissions-and-enrollment-statistics/. Accessed 8 Dec. 2023.

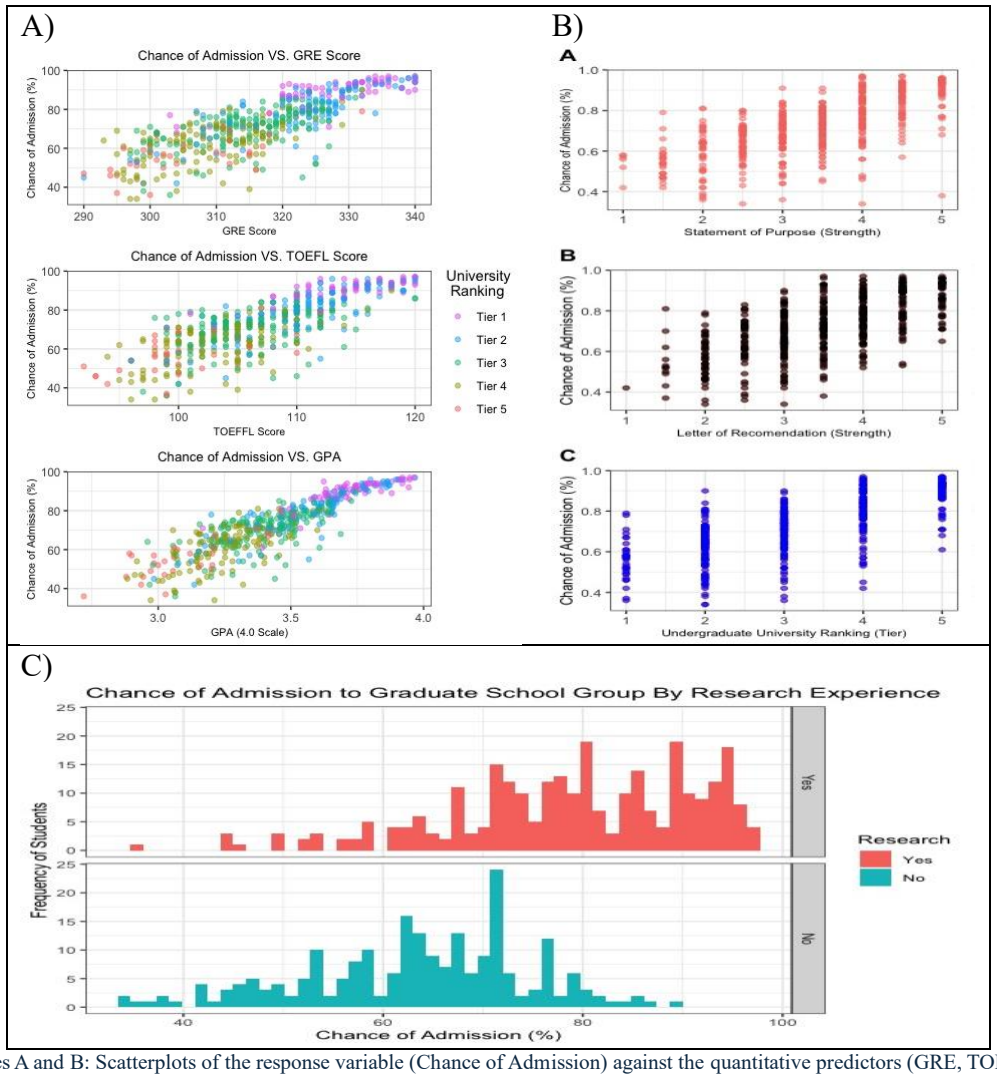

Figure 2: Boxes A and B: Scatterplots of the response variable (Chance of Admission) against the quantitative predictors (GRE, TOEFL, GPA, SOP, LOR, and University Ranking). Box C: Histograms showing the distribution of chance of admission group by research experience.

The Q-Q and residual plots in Figure 3 show that the assumptions of normality and constant variance are being violated. The Q-Q plot shows that there is heavy deviation from the 45° degree line at the lower left corner of the plot, which signify that the distribution of the admission dataset is heavily skewed to one side. The skewness in the distribution make sense in the context of this dataset since we expect fewer individuals with a lower chance of admission to apply to the program. In addition, the residual plot shows a funnel pattern where the residuals approach zero as the fitted value increases. To address the violations of all three assumptions, a squared transformation was done on the response variable to help linearize the relationship between the response and predictor variables and to help stabilize the variance.

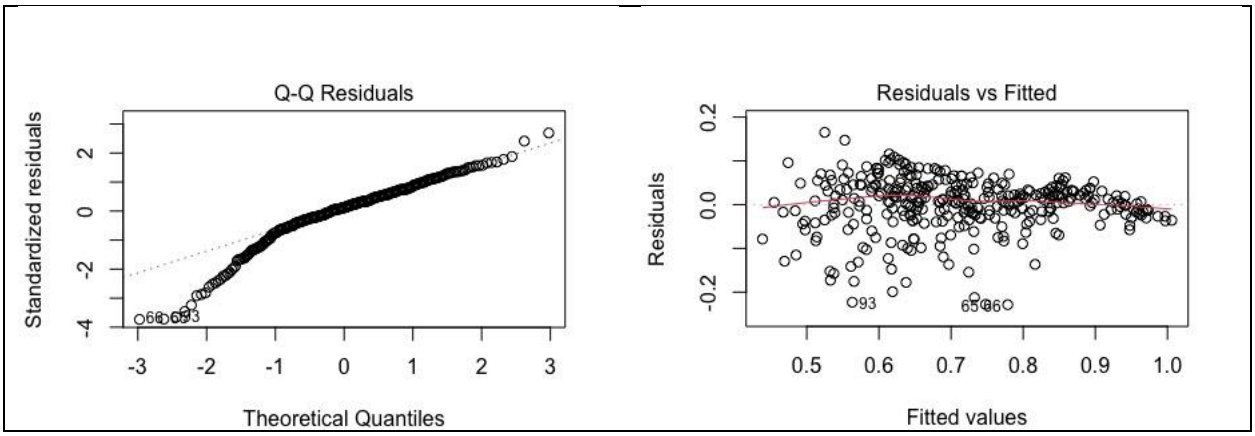

Figure 3: QQ and Residual plots for the full untransformed model.

Results and Discussion

A) Research Question 1

Table -2: Hypothesis testing on the difference in averages of quantitative attributes of admitted PhD applicants between the different citizenship status.

Admission To Statistics PhD Programs For 2011-2019 Cycles				
All Statistics Program				
Hypothesis Testing	**GPA**	**GRE Verbal Score**	**GRE Quant Score**	**GRE Writing Score**
$H_0: \mu_A = \mu_I \ vs. \ H_A: \mu_A \neq \mu_I$	P(|t| ≥ 0.65) = 0.52 $CI^{95\%}$ = [-0.024, 0.048]	P(|t| ≥ 2.7447) = 0.006 $CI^{95\%}$ = [6.12, 36.97]	P(|t| ≥ 1.91) = 0.056 $CI^{95\%}$ = [-0.50, 38.34]	P(|t| ≥ 2.61) = 0.009 $CI^{95\%}$ = [0.14, 1.05]
$H_0: \mu_A = \mu_U \ vs. \ H_A: \mu_A \neq \mu_U$	P(|t| ≥ 1.13) = 0.26 $CI^{95\%}$ = [-0.021, 0.078]	P(|t| ≥ -1.44) = 0.15 $CI^{95\%}$ = [-69.21, 10.85]	P(|t| ≥ -1.9569) = 0.052 $CI^{95\%}$ = [-82.89, 0.48]	P(|t| ≥ 2.82) = 0.005 $CI^{95\%}$ = [0.26, 1.49]
$H_0: \mu_A = (\mu_I + \mu_U) \ vs. \ H_A: \mu_A \neq (\mu_I + \mu_U)$	P(|t| ≥ 1.11) = 0.27 $CI^{95\%}$ = [-0.014, 0.049]	P(|t| ≥ 0.35) = 0.72 $CI^{95\%}$ = [-15.6, 22.41]	P(|t| ≥ -0.23499) = 0.81 $CI^{95\%}$ = [-24.18, 19.01]	P(|t| ≥ 3.74) = 0.0002 $CI^{95\%}$ = [0.33, 1.069]

Table-2 shows that there is no significant difference in the average GPA of applicants applying to a Statistics PhD program regardless of their citizenship status. However, we do see a noticeable difference in the average GRE quant scores between American applicants versus international applicants and international applicants with a US college degree. Where applicants in the two later groups consistently score higher on average on the quant portion than their American counterparts. However, American applicants on average score significantly higher in the GRE writing portion than both international applicants and international applicants with a US degree. We expect American applicants to score significantly higher in the writing portion than their counterparts since English is their primary language compared to the majority international applicants. Therefore, the hypothesis test confirming that American applicants scoring higher on average in the GRE writing portion is not a surprised.

Table-3: Hypothesis testing on the difference in averages of quantitative attributes of admitted master applicants between the different citizenship status.

Admission To Statistics Master Programs For 2011-2019 Cycles				
All Statistics Program				
Hypothesis Testing	**GPA**	**GRE Verbal Score**	**GRE Quant Score**	**GRE Writing Score**
$H_0: \mu_A = \mu_I$ *vs.* $H_A: \mu_A \neq \mu_I$	P(|t| ≥ -1.44) = 0.15 $CI^{95\%}$ = [-0.12, 0.019]	P(|t| ≥ -0.89) = 0.36 $CI^{95\%}$ = [-41.86, 15.65]	P(|t| ≥ -1.7655) = 0.079 $CI^{95\%}$ = [-67.97, 3.81]	P(|t| ≥ 3.37) = 0.0008 $CI^{95\%}$ = [0.57, 2.18]
$H_0: \mu_A = \mu_U$ *vs.* $H_A: \mu_A \neq \mu_U$	P(|t| ≥ -2.7178) = 0.007 $CI^{95\%}$ = [-0.159, -0.025]	P(|t| ≥ -1.80) = 0.072 $CI^{95\%}$ = [-55.81, 2.49]	P(|t| ≥ -2.3204) = 0.021 $CI^{95\%}$ = [-75.84, -6.154]	P(|t| ≥ 0.775) = 0.43 $CI^{95\%}$ = [-0.96, 2.22]
$H_0: \mu_A = (\mu_I + \mu_U)$ *vs.* $H_A: \mu_A \neq (\mu_I + \mu_U)$	P(|t| ≥ -2.3) = 0.022 $CI^{95\%}$ = [-0.138, -0.0108]	P(|t| ≥ -1.807) = 0.071 $CI^{95\%}$ = [-43.16, 1.80]	P(|t| ≥ -2.70) = 0.007 $CI^{95\%}$ = [-64.03, -10.09]	P(|t| ≥ 1.71) = 0.09 $CI^{95\%}$ = [-0.145, 2.06]
$H_0: \mu_I = \mu_U$ *vs.* $H_A: \mu_I \neq \mu_U$	P(|t| ≥ -1.52) = 0.12 $CI^{95\%}$ = [-0.092, 0.011]	P(|t| ≥ -2.51) = 0.013 $CI^{95\%}$ = [-50.08, 22.96]	P(|t| ≥ -0.386) = 0.69 $CI^{95\%}$ = [-54.39, 36.55]	P(|t| ≥ -1.05) = 0.29 $CI^{95\%}$ = [-2.15, 0.65]

Table-3 confirms some of the findings we found in Table-2, where international applicants consistently score significantly higher on average on the GRE quant portion than their American counterparts. Looking at the 95% confident interval for the difference in the average GRE quant scores between American applicants and all international applicants is between 10.09 and 65.03 points.

Table -1: Proportion of Applicants Accepted to a Top 10 Statistics program based on their citizenship status

Proportion of Applicants Accepted to a Top 10 Statistics Ph.D Programs by Citizenship Status			
Hypothesis Testing	**Proportions**	**Test Statistics**	**Confident Interval**
$H_0: p_A = p_I$ *vs.* $H_A: p_A \neq p_I$	$P_A = \frac{82}{82+187} = 0.3048$ $P_I = \frac{48}{48+66} = 0.421$	P(|Z| ≥ 4.319) = 0.03769	[-0.228, -0.0039]
$H_0: p_A = (p_I + p_U)$ *vs.* $H_A: p_A \neq (p_I + p_U)$	$P_A = \frac{82}{82+187} = 0.3048$ $P_{I+U} = \frac{48+22}{48+66+22+58} = 0.3608$	P(|Z| ≥ 1.3586) = 0.2438	[-0.146, 0.0355]

Table-4 shows evidence that the proportion of American applicants accepted to one of the top 10 statistics Ph.D program is equal to the proportion of international applicants accepted based on the total number of applicants applied to those program between the year 2011 and 2019.

B) Research Question 2

$$\widehat{y^2} = 0.0033051x_{GRE} + 0.0043766x_{TOEFL} + 0.0264269x_{LOR} + 0.4189314x_{GPA} + 0.0348930_{Research_{yes}} + \epsilon \qquad \epsilon \sim N(\mu, \sigma^2)$$

Eq 1: Linear Regression formula for the final transformed model

The results from the three forward, backward, and stepwise variables selection methods agreed that the best linear regression models must contains the following five predictors: GRE score, TOEFL score, Strength of Letter of Recommendation, GPA, and research experience. From the formula for the multiple linear regression shown in Eq.1, we see that the three most influential variables in predicting an individual's admissions chance are GPA, having research experience, and letter of recommendation. With GPA being the most influential variable, it can be interpreted from the final model that the chance of

admission increases on average by $\sqrt{0.419}$ or with an additional increase in the individual's grade point average given that other variables stay constant. It seems that for this dataset, GPA has the most influence on the chance of admission for an individual applying to their chosen master program. In addition, having research experience raises a person chance of admission on average by $\sqrt{0.035}$ or approximately 18.7% given that all other predictors stay constant. While an additional increase in the score of a letter of recommendation increases the chance of admission on average by $\sqrt{0.0264}$ or approximately 16.2%. Based on the evidence found in this dataset, GPA, LOR, and research experience are the most important factors in determining an applicant's chance of getting into their chosen master program. These three predictors are important quantifiers that admission officers traditionally use to judge an individual's fit for their graduate program.

C) Research Question 3

Based on the scatterplots of predicted chance of admission plotted against observed chance of admission, we see that there is a strong linear relationship between the two variables. This positive linear relationship shows that our final model was able to accurately predict an individual's chance of admission given their GPA, GRE score, TOEFL score, LOR score, and their research experience. With the testing and training error approximately equal, means that we were able to fit our linear regression model on a training dataset that is representative of the testing dataset. Therefore, our final model was able to perform well at predicting the chance of admission for the testing dataset.

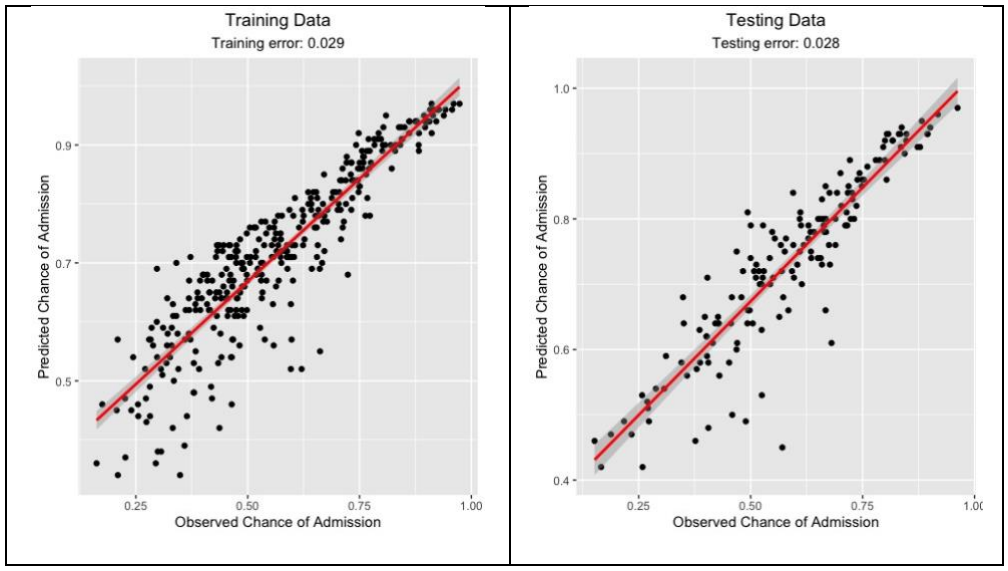

Figure 4: Scatterplots of Predicted against Observed Chance of Admission for the training and testing datasets.

Conclusion

In conclusion, this final report found the following results: international applicants on average have higher score for the GRE quant portion than their American counterparts, regardless of whether they are applying to a PhD or a master program in Statistics. While the American counterparts usually score higher on average in both the verbal and writing portion, most noticeably for applicants applying to a PhD program. My initial hypothesis that the proportion of American applicants to the top 10 PhD programs would be higher than the proportion of international applicants was rejected, and there is evidence showing that the proportions of both types of applicants are not significantly different from each other. I

found that the predictors that have the most influence on the chance of admission for an indiviuals are GPA, GRE scores, TOEFL scores, research experience, and letter of recommendation. With GPA, research experience, and letter of recommendation being the most influential among the 5 predictors.

Appendix A

Table ***-Error! No text of specified style in document.****-2: Average GPA and GRE scores of applicants accepted into a Statistics Master program between 2011-2019 application cycles grouped by all Statistics Master programs and Top 10 Statistics Master programs for each of the citizenship status.*

Admission To Statistics Master Programs For 2011-2019 Cycles											
All Statistics Program						**Top 10 Statistics Program**					
Status	**GPA**	**GRE Verbal Score**	**GRE Quant Score**	**GRE Writing Score**	**Counts**	**Status**	**GPA**	**GRE Verbal Score**	**GRE Quant Score**	**GRE Writing Score**	**Counts**
A	3.690329	180.9259	185.5720	4.794198	243	A	3.739605	218.5263	225.9474	4.447368	76
I	3.742569	194.0275	217.6514	3.417431	109	I	3.777391	179.3043	193.8696	3.489130	46
U	3.782899	207.5870	226.5725	4.165145	138	U	3.844340	192.4528	197.2453	5.452642	53

Table ***-Error! No text of specified style in document.****-3: Average GPA and GRE scores of applicants accepted into a Statistics PhD program between 2011-2019 application cycles grouped by all Statistics PhD programs and Top 10 Statistics PhD programs for each of the citizenship status.*

Admission To Statistics PhD Programs For 2011-2019 Cycles											
All Statistics Program						**Top 10 Statistics Program**					
Status	**GPA**	**GRE Verbal Score**	**GRE Quant Score**	**GRE Writing Score**	**Counts**	**Status**	**GPA**	**GRE Verbal Score**	**GRE Quant Score**	**GRE Writing Score**	**Counts**
A	3.813720	189.2585	199.1087	4.565700	414	A	3.882439	204.3049	212.0366	4.567073	82

I	3.801818	167.7102	180.1875	3.963068	176	I	3.913542	171.5625	181.4792	3.979167	48
U	3.785204	218.4388	240.3163	3.688776	98	U	3.879091	264.9091	280.6818	3.681818	22
I + U	3.795876	185.854	201.6934	3.864964	274	I + U	3.902714	200.9	212.6571	3.885714	70

*Table-**Error! No text of specified style in document.**-4: Numbers of applicants who were rejected or accepted to a graduate program grouped by all Statistics programs or Top 10 Statistics program for each of the citizenship status.*

Status	Accepted				Rejected			
	Master		PhD		Master		PhD	
	All	Top 10	All	Top 10	All	Top 10	All	Top 10
A	243	76	414	82	81	36	296	187
I	109	46	176	48	89	31	160	66
O	7	3	11	5	7	3	5	2
U	138	53	98	22	86	33	139	58
Total:	497	178	699	157	263	103	600	313

Appendix B

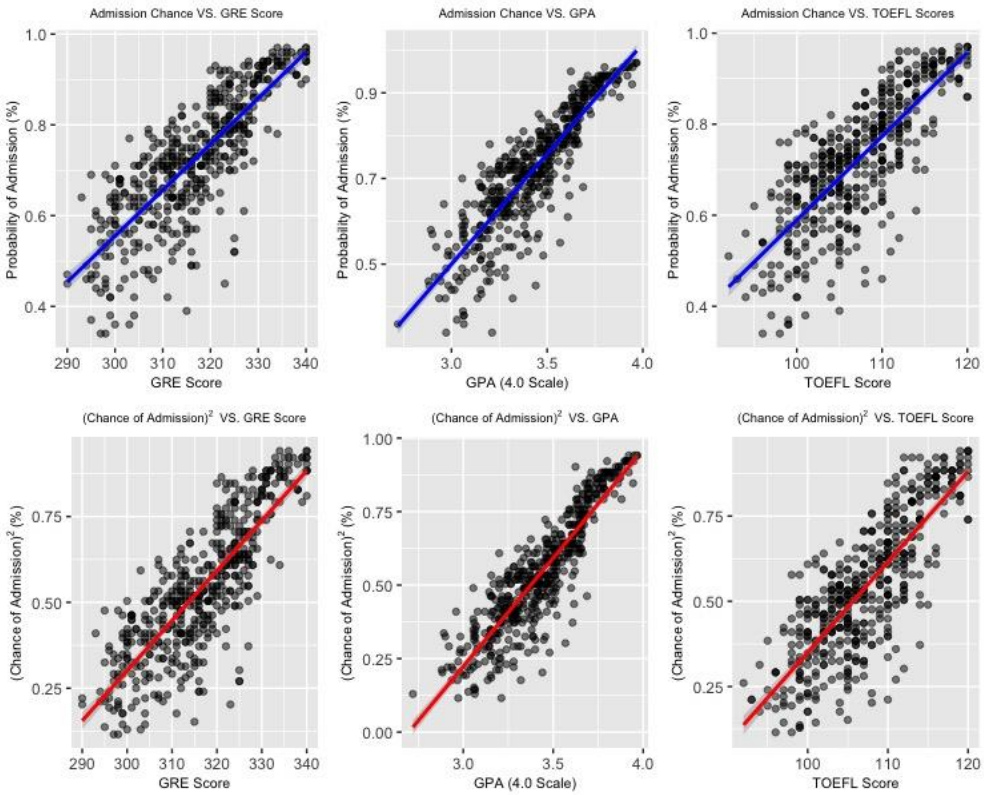

Figure 5: Scatterplots of chance of admission against continuous quantitative predictors for before and after transforming response variable.

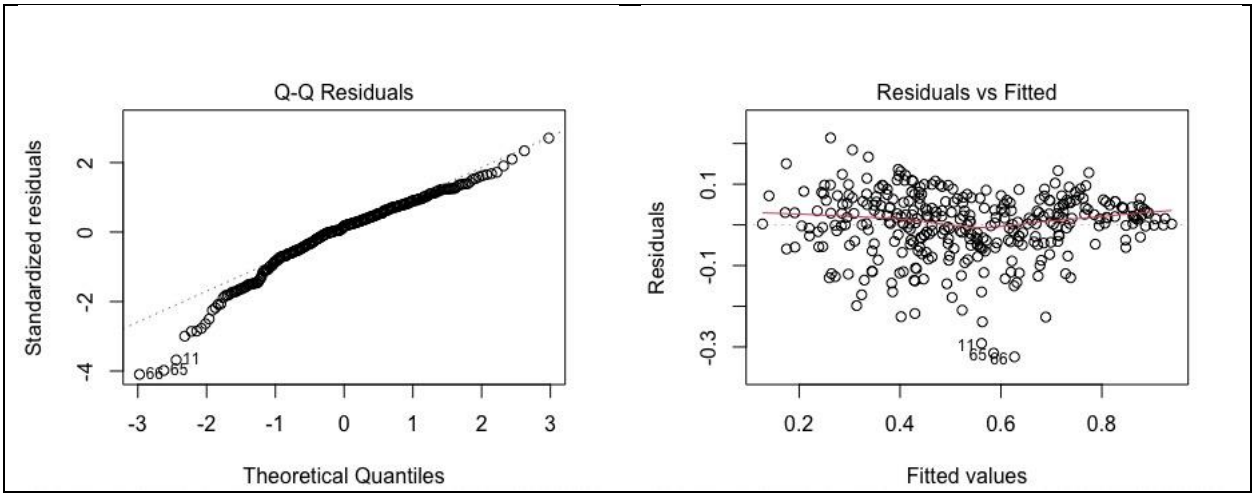

Figure 6: Q-Q and Residual plots for the final fitted linear regression model.

$$\hat{y} = 2.316e^{-3}x_{GRE} + 2.236e^{-3}x_{TOEFL} + 1.860e^{-2}x_{LOR} - 4.561e^{-5}x_{SOP} + 2.856x_{GPA} + 2.058e^{-2}{}_{Research\_Yes} - 7.165e^{-3}x_{Rank:2} + 2.182e^{-4}x_{Rank:3} + 3.494e^{-3}x_{Rank:4} + 2.599e^{-2}x_{Rank:5} + \epsilon \qquad \epsilon \sim N(\mu, \sigma^2)$$

Eq 2: Linear Regression Formula for the full model

*Table**Error! No text of specified style in document.**-5: Comparing estimated parameters of the full and final model*

Statistics	Full Model	Final Model
$\widehat{\beta_{GRE}}$	$2.316e^{-03}$	0.0033051
$\widehat{\beta_{TOEFL}}$	$2.236e^{-03}$	0.0043766
$\widehat{\beta_{LOR}}$	$1.860e^{-02}$	0.0264269
$\widehat{\beta_{SOP}}$	$-4.561e^{-05}$	
$\widehat{\beta_{GPA}}$	2.856	0.4189314
$\widehat{\beta_{Researchyes}}$	$2.058e^{-02}$	0.0348930
$\widehat{\beta_{Uni\_rank_2}}$	$-7.165e^{-03}$	
$\widehat{\beta_{Uni\_rank_3}}$	$2.182e^{-04}$	
$\widehat{\beta_{Uni\_rank_4}}$	$3.494e^{-03}$	
$\widehat{\beta_{Uni\_rank_5}}$	$2.599e^{-02}$	
$se(\widehat{\beta_0})$	$1.311e^{-01}$	
$se(\widehat{\beta_{GRE}})$	$6.164e^{-04}$	0.0007908
$se(\widehat{\beta_{TOEFL}})$	$1.104e^{-03}$	0.0013891
$se(\widehat{\beta_{LOR}})$	$5.209e^{-03}$	0.0062581
$se(\widehat{\beta_{SOP}})$	$5.485e^{-03}$	
$se(\widehat{\beta_{GPA}})$	$2.939e^{-02}$	0.0357687
$se(\beta_{Researcc\ Yes})$	$8.255e^{-03}$	0.0104667
$se(\widehat{\beta_{Uni\_rank_2}})$	$1.450e^{-02}$	
$se(\widehat{\beta_{Uni\_rank_3}})$	$1.519e^{-02}$	
$se(\widehat{\beta_{Uni\_rank_4}})$	$1.804e^{-02}$	
$se(\widehat{\beta_{Uni\_rank_5}})$	$2.045e^{-02}$	
MS_{RES}	0.0039	0.0064
R^2	0.8192	0.8472
Adjusted R^2	0.8137	0.8449