Analysing the Factors that Influence the Graduate School Admission Decision

Jingwen Deng 1004713081

Dec.22, 2020

This study can be found on: https://github.com/jingwennnn/Analysing-the-Factors-that-Influence-the-Graduate-School-Admission-Decision

Abstract

The U.S. Graduate school is the dream goal that many undergraduates pursued. However, the number of the position is limited in each Graduate school program. No one knows the exact standard of the Graduate school admission office. On the other hand, the process of application is costly in both time and money. In this paper, multiple linear regression models, logistic regression models, and classification trees are fitted on *Chance.of.Admit* using the *Admission* dataset (Acharya, 2018) to analyze the critical factors that affect the application decision. After fitting models, a propensity score matching with a logistic regression approach is introduced to identify the causal inference between the research experience and the level of the probability of being admitted.

Keywords

Graduate School, Graduate School Admission, GPA, TOEFL Score, Research, Linear Regression, Logistic Regression, Classification Tree, Propensity Score Matching, Casual Inference

1. Introduction

During each year, there are a significant number of undergraduate students chasing a graduate school goal to obtain a deeper understanding of their academic or career path in the U.S. Universities receive plenty of applications from all around the world while the admission space is limited. The admission office of a particular university usually considers identical variables to choose applicants. Some elite universities have admission preferences for minority students, athletes, and legacies. (Espenshade, T.) From undergraduates' point of view, the process of filling in the admission is very costly in terms of time and money. Undergraduates need to prepare for extra examinations, ask for recommendation letters, and at the same time keeping the GPA relatively high to fulfill the application requirements. This paper addresses and evaluates the key factors that influence one's application decision the most.

The UCLA Graduate Dataset inspires the dataset (Acharya, 2018), called "Admission," used in this paper. It includes a variable called *Chance.of.Admit* ranging from 0-1 as a score that the dataset creator gives to based on other variables collected from Indian undergraduates who applied for U.S. graduate schools. In this paper, by fitting a multiple linear regression model and logistic regression models on the response variable *Chance.of.Admit*, the essential factors that affect the graduate school admission decisions can be determined. Also, training and testing datasets are used to examine logistic regression models' accuracy and goodness. After fitting the logistic model, a classification tree is plotted to visualize the critical factors and the decision better.

In this study, observational data (*Admission*) is used since it is more feasible than experimental data in the real-life since experimental design data costs plenty of time, money, and other resources to obtain. Therefore, making causal inference using observational data becomes meaningful from a practical perspective. The casual links between the key factors that affect admission are discussed through propensity score matching (Alexander, 2020) in this study after analyzing the fitted models. Propensity score matching is a statistical technique used for cleaning data and eventually making the original data look similar to an experimental design data by matching the treated and controlled observations on the estimated probability of being treated (propensity score). (Alexander, 2020)

The detail about the data used, the fitted models, and the propensity score in this study are discussed in the Methodology section (Section 2). Results of the data summaries fitted models, fitted classification tree, and propensity score analysis are stated in the Result section (Section 3). Finally, the summary of findings with a conclusion, weaknesses, and future work is discussed in the Discussion section. (Section 4)

2. Methodology

2.1 Data

The dataset used in this study is created by Mohan S Acharya. and can be found from kaggle.com. (Acharya, 2018) This data, called it *Admission*, has 500 observations with nine variables. Meanwhile, data of 500 people with seven features of $GRE.Score$, $TOEFL.Score$, $University.Rating$, SOP, LOR, $CGPA$, and $Research$ are collected from Indian undergraduates by Mohan S Acharya (Acharya, 2018). The specific meanings of these seven features are listed below.

1. $GRE.Score$: The GRE Score is a numerical value out of 340. GRE General Test contains three parts Verbal Reasoning, Quantitative Reasoning, and Analytical Writing. (ETS GRE, 2020)
2. $TOEFL.Score$: The TOEFL Score is a numerical value out of 120. TOEFL is the world's premier English-language test for university study, work, and immigration. (ETS TOEFL, 2020)
3. $University.Rating$: An integer score from 0 to 5 where 0 is the lowest and 5 is the highest. The rating is determined by the preference order of a student for universities.
4. SOP: The strength of Statement of Purpose (out of 5)
5. LOR: The strength of Letter of Recommendation (out of 5)
6. $CGPA$: The grade point average in the older format. (out of 10)
7. $Research$: Research Experience (0 = no research experience; 1 = have research experience)

The $Serial.No.$ is a specific number assigned to each Indian undergraduate, just like a unique order number. Since $Serial.No.$ is not an influential feature, it is removed from the dataset. The analysis focus on variable $Chance.of.Admit$ in the *Admission* dataset. $Chance.of.Admit$ is a variable, ranging from 0 to 1 that the dataset creator gives to each observation based on the performance of other variables collected from Indian undergraduates who applied for U.S. graduate schools. $Chance.of.Admit$ is the critical feature throughout this study.

One additional categorical variable is called $Admit.Prob$ is created based on whether the value of $Chance.of.Admit$ is greater than the average of $Chance.of.Admit$. If the value of $Chance.of.Admit$ is greater than the average of $Chance.of.Admit$, $Admit.Prob$ is categorized as $High$ representing the admitted probability is high. If not, $Admit.Prob$ is classified as Low, indicating the accepted possibility is low. This variable is prepared to analyze logistic regression, decision tree, and propensity score matching in the following section.

Now we have nine variables in *Admission* with seven features of individuals, $Chance.of.Admit$ as the numerical target and $Admit.Prob$ as the categorical target. These nine variables are chosen to continue the study since each of the variables is unique and significant. The following Table 1 is the *Admission* data visualization, which shows the first six observations of the dataset.

Table 1: Data Visualization

GRE.Score	TOEFL.Score	University.Rating	SOP	LOR	CGPA	Research	Chance.of.Admit	Admit.Prob
337	118	4	4.5	4.5	9.65	1	0.92	high
324	107	4	4.0	4.5	8.87	1	0.76	high
316	104	3	3.0	3.5	8.00	1	0.72	low
322	110	3	3.5	2.5	8.67	1	0.80	high
314	103	2	2.0	3.0	8.21	0	0.65	low
330	115	5	4.5	3.0	9.34	1	0.90	high

2.2 Model

Since we are interested in finding the factors that influence the Graduate school admission result, Multiple Linear Regression (MLR) models and logistic regression models are generated using the R language in R Markdown. (R, 2019). Functions of $lm()$ and $glm()$ are used to create the fitted MLR models and logistic regression models respectively. A decision tree is built to predict the level of admission probability using function $rpart()$ and $as.party()$.

2.2.1 Multiple Linear Regression (MLR)

In the beginning, a full MLR model with all seven predictors is generated to select the significant variables for the MLR model. Based on the summary table of the full MLR model in Appendix #1, $GRE.Score$, $TOEFL.Score$, LOR, $CGPA$, and $Research$ are chosen as predictors for the MLR model.

- MLR Model Equation:

  $$y_{Chance.of.Admit} = \beta_0 + \beta_1 X_{GRE.Score} + \beta_2 X_{TOEFL.Score} + \beta_3 X_{LOR} + \beta_4 X_{CGPA} + \beta_5 X_{Research_1} + \epsilon_i$$

  $Research$ is a categorical variable in the data, so a dummy variable $X_{Research_1}$ is used in the model equation. $X_{Research_1} = 1$ only when the person has research experience, otherwise, $X_{Research_1} = 0$. The response variable is $y_{Chance.of.Admit}$ in this MLR model, and all other X are predictors of Y. β_0 is the intercept representing when all predictors take a value of zero, what the chance of admission would be. β_1, β_2, β_3, and β_4 indicates the average change in chance of admission when $GRE.Score$, $TOEFL.Score$, LOR, and $CGPA$ increase by one unit, respectively. β_5 is the difference in average change of chance of admission between those with research experience and without research experience. When a person has research experience, the average change in the chance of admission is β_5.

2.2.2 Logistic Regression Models

A new binary variable $Admit$ is introduced with the same meaning with variable $Admit$ to fit a logistic regression model.Prob.$If the admission probability is high$(Admit.Prob = high\$), then $Admit = 1$. If the admission probability is low ($Admit.Prob = low$), then $Admit = 0$. $Admit$ is the response variable here. The following are the basic descriptions of four logistic regression models that are fitted in this report.

- $GLM1$: a fitted full logistic regression model with all seven predictors ($GRE.Score$, $TOEFL.Score$, $University.Rating$, SOP, LOR, $CGPA$, and $Research$) on the full dataset.
- $GLM2$: a fitted full logistic regression model with all seven predictors ($GRE.Score$, $TOEFL.Score$, $University.Rating$, SOP, LOR, $CGPA$, and $Research$) on the training dataset.
- $GLM3$: a fitted reduced logistic regression model with three predictors ($GRE.Score$, $CGPA$, and $Research$) on the full dataset.
- $GLM4$: a fitted reduced logistic regression model with three predictors ($GRE.Score$, $CGPA$, and $Research$) on the training dataset.

First, $GLM1$ is fitted. The dataset is then randomly split into a training and testing dataset, where training contains 80% of the observations and testing includes the left 20% observations. Then, $GLM2$ is fitted on the training dataset. The fitted $GLM2$ model is used to make predictions of *Admit* in the testing data afterward. The predicted *Admit* is classified back to *high* if the numerical probability of predicted *Admit* is more significant than 0.5, and *low* on the other hand. Tables of confusion matrix can be generated with the number of correct classifications on the diagonal and the wrong classifications off the diagonal. Besides, the accuracy of $GLM2$ can be determined by comparing the predicted level of admission probability and the actual level of admission probability in the training data and evaluates by the testing data.

Similarly, $GLM3$ is fitted using the original data, and $GLM4$ is fitted using the training data. We can find the accuracy of $GLM4$ using a similar approach as $GLM2$. Using $GLM1$ and $GLM3$ to predict the admission probability level in the original data, comparing the results by confusion matrix, the accuracy of $GLM1$ and $GLM3$ can be found. Finally, the accuracy of these four logistic models is obtained separately.

- Full Logistic Regression Model Equation: ($GLM1$ and $GLM2$)

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_{GRE.Score} + \beta_2 X_{TOEFL.Score} + \beta_3 X_{University.Rating} + \beta_4 X_{SOP} + \beta_5 X_{LOR} + \beta_6 X_{CGPA} + \beta_7 X_{Research_1}$$

  $log(\frac{p}{1-p})$ is the odds of the level of admission probability, which is also called "log odds". p is the probability of the level of admission probability. β_0 is the constant intercept. β_1, β_2, β_3, β_4, β_5 and β_6 represents the average change in log odds when there is one unit increase in the GRE score, TOEFL score, ratings of University, the strength of the proposal, the strength of recommendation letters, and the CGPA, respectively. β_7 is the coefficient of dummy variable *Research* indicating the change in log odds if an undergraduate had research experience.

- Reduced Logistic Regression Model Equation: ($GLM3$ and $GLM4$)

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_{GRE.Score} + \beta_2 X_{CGPA} + \beta_3 X_{Research_1}$$

  Similarly, $log(\frac{p}{1-p})$ is the odds of the level of admission probability. p is the probability of the level of admission probability. β_0 is the constant intercept. β_1 and β_2 demonstrate the average change in log odds when there is one unit increase in the GRE score and the CGPA, respectively. β_3 is the coefficient of dummy variable *Research* representing the change in log odds if an undergraduate had research experience.

2.2.3 Classification Tree

Classification Tree is a supervised learning statistical method that is useful when the response y is categorical. In this study, a classification tree is built to predict individuals' level of admission probability (whether it is *high* or *low*) using all seven predictors.

2.2.4 Propensity Score Matching

Propensity scores help with the underlying issue of causal inference by isolating the confounding variables' adjustment and analyzing the treatment impact. This statistical technique makes the observational data similar to experimental data by matching observations with similar propensity scores and dropping those un-matched observations.

3. Results

3.1 Data

Table 2: Summary Statistics of Level of Admission Probability

Admission Probability Level	Number of Observations	GRE Score	TOEFL Score	CGPA	Research Proportion
high	248	324.484	111.210	9.021	0.831
low	252	308.587	103.238	8.139	0.294

- Table 2 shows the summary statistics between two groups, high admission probability, and low admission probability. The number of observations in each group is shown in the second column. The average GRE score, average TOEFL Score, and average CGPA of each group are indicated in columns three to five, respectively. The last column tells the proportion of the number of people that have research in each group.

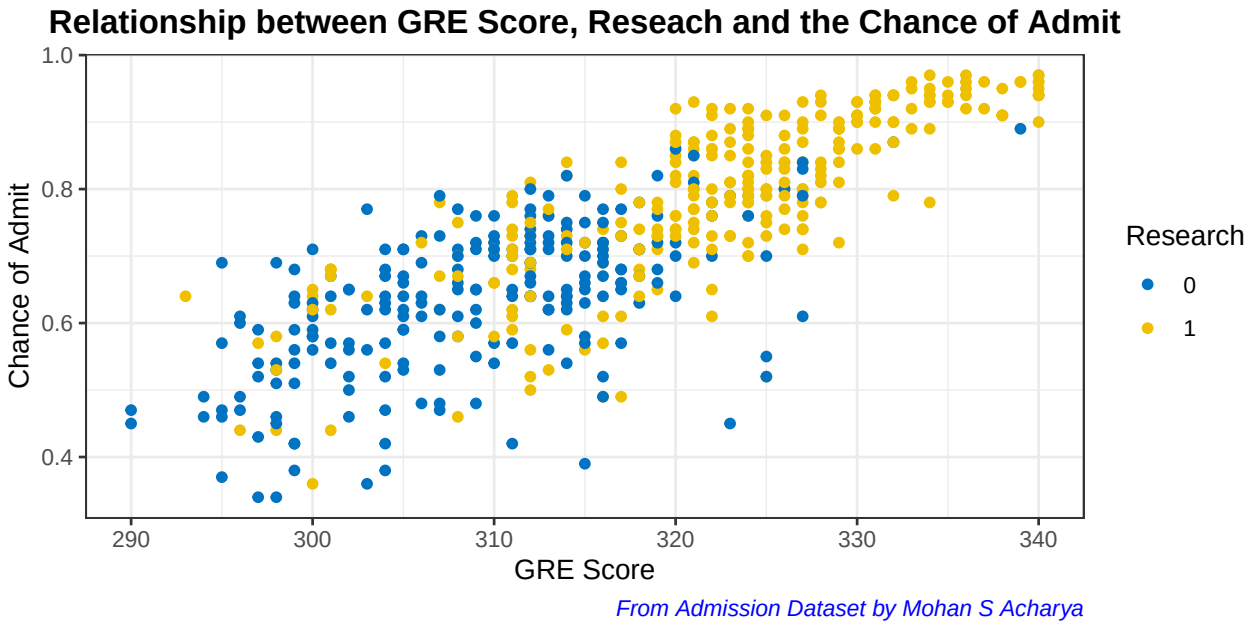


Figure 1: Relationship between GRE Score, Reseach and the Chance of Admit

- In Figure 1, the x-axis represents the GRE Score, while the y-axis represents the Chance of Admit. Datapoints with numerical values are indicated in the above scatterplot. The feature of research experience is identified through different colors in Figure 1. Yellow data points stand for has some research experiences, and blue data points mean no research experience.

3.2 Model

3.2.1 Multiple Linear Regression (MLR)

Table 3: Summary Table of MLR Model

term	estimate	std.error	statistic	p.value
(Intercept)	-1.3357	0.0991	-13.4817	0e+00
GRE.Score	0.0019	0.0005	3.7604	2e-04
TOEFL.Score	0.0030	0.0009	3.5009	5e-04
LOR	0.0193	0.0038	5.0924	0e+00
CGPA	0.1230	0.0093	13.2210	0e+00
as.factor(Research)1	0.0252	0.0066	3.8135	2e-04

Table 3 is a summary table for the fitted MLR model with five predictors, including the estimate, standard error, statistics, and p-values for each of the five predictors. Based on Table 3, the fitted MLR model equation can be obtained:

$$\hat{y}_{Chance.of.Admit} = -1.336 + 0.002X_{GRE.Score} + 0.003X_{TOEFL.Score} + 0.019X_{LOR} + 0.123X_{CGPA} + 0.025X_{Research_1}$$

3.2.2 Full Logistic Regression Model on Original Data ($GLM1$)

Table 4: Summary Table of Full Logistic Regression Model on Original Data (GLM1)

term	estimate	std.error	statistic	p.value
(Intercept)	-60.0449	7.7208	-7.7771	0.0000
GRE.Score	0.0813	0.0294	2.7652	0.0057
TOEFL.Score	0.0396	0.0484	0.8183	0.4132
University.Rating	0.3199	0.2225	1.4378	0.1505
SOP	0.4922	0.2715	1.8130	0.0698
LOR	0.2540	0.2384	1.0656	0.2866
CGPA	3.0281	0.6149	4.9245	0.0000
as.factor(Research)1	1.2065	0.3422	3.5257	0.0004

Table 4 is a summary table for the fitted logistic regression model ($GLM1$) with seven predictors. The estimate, standard error, statistics, and p-values for each of the seven predictors are shown in the table. Based on Table 4, the fitted logistic regression model equation can be obtained:

$$\begin{aligned} log(\frac{\hat{p}}{1-\hat{p}}) = &-60.045 + 0.081X_{GRE.Score} + 0.040X_{TOEFL.Score} + 0.320X_{University.Rating} \\ &+ 0.492X_{SOP} + 0.254X_{LOR} + 3.028X_{CGPA} + 1.206X_{Research_1} \end{aligned}$$

3.2.3 Full Logistic Regression Model on Training Data ($GLM2$)

Table 5: Summary Table of Full Logistic Regression Model on Training Data (GLM2)

term	estimate	std.error	statistic	p.value
(Intercept)	-60.8380	8.5921	-7.0807	0.0000
GRE.Score	0.0902	0.0338	2.6666	0.0077
TOEFL.Score	0.0276	0.0550	0.5013	0.6162
University.Rating	0.3634	0.2543	1.4288	0.1531
SOP	0.3985	0.3027	1.3164	0.1880
LOR	0.2803	0.2652	1.0569	0.2906
CGPA	2.9613	0.6982	4.2415	0.0000
as.factor(Research)1	1.1431	0.3810	3.0003	0.0027

Table 5 is a summary table for the fitted logistic regression model ($GLM2$) on the training dataset with seven predictors. The estimate, standard error, statistics, and p-values for each of the seven predictors are shown in the table. Based on Table 5, the fitted logistic regression model equation can be obtained:

$$log(\frac{\hat{p}}{1-\hat{p}}) = -60.838 + 0.090X_{GRE.Score} + 0.028X_{TOEFL.Score} + 0.363X_{University.Rating} + 0.398X_{SOP} + 0.280X_{LOR} + 2.961X_{CGPA} + 1.143X_{Research_1}$$

3.2.4 Reduced Logistic Regression Model on Original Data ($GLM3$)

Table 6: Summary Table of Reduced Logistic Regression Model on Original Data (GLM3)

term	estimate	std.error	statistic	p.value
(Intercept)	-63.9141	7.4216	-8.6119	0e+00
GRE.Score	0.0938	0.0255	3.6750	2e-04
CGPA	3.9182	0.5454	7.1847	0e+00
as.factor(Research)1	1.1972	0.3224	3.7130	2e-04

Table 6 is a summary table for the fitted reduced logistic regression model ($GLM3$) on the original dataset with three predictors. The estimate, standard error, statistics, and p-values for each of the three predictors are indicated in the table. Based on Table 6, the fitted logistic regression model equation can be obtained:

$$log(\frac{\hat{p}}{1-\hat{p}}) = -63.914 + 0.0938X_{GRE.Score} + 3.918X_{CGPA} + 1.197X_{Research_1}$$

3.2.5 Reduced Logistic Regression Model on Training Data ($GLM4$)

Table 7: Summary Table of Reduced Logistic Regression Model on Training Data (GLM4)

term	estimate	std.error	statistic	p.value
(Intercept)	-63.9029	8.2693	-7.7277	0.0000
GRE.Score	0.0945	0.0290	3.2581	0.0011
CGPA	3.9000	0.6222	6.2680	0.0000
as.factor(Research)1	1.1045	0.3611	3.0585	0.0022

Table 7 is a summary table for the fitted reduced logistic regression model ($GLM4$) on the training dataset with three predictors. The estimate, standard error, statistics, and p-values for each of the three predictors are indicated in the table. Based on Table 7, the fitted logistic regression model equation can be obtained:

$$log(\frac{\hat{p}}{1-\hat{p}}) = -63.903 + 0.0945X_{GRE.Score} + 3.900X_{CGPA} + 1.105X_{Research_1}$$

Table 8: Accuracy of Logistic Regression Models

GLM1	GLM2 on Training	GLM2 on Testing	GLM3	GLM4 on Training	GLM2 on Testing
0.88	0.51	0.9	0.88	0.9	0.86

Table 8 shows the accuracy of each model predicting the corresponding dataset. The accuracy in the first column and the fourth column is the accuracy when using $GLM1$ and $GLM3$ to predict the level of admission probability in the original dataset. The second column, "GLM2 on Training" indicates the accuracy when using $GLM2$ to predict the level of admission probability in the training dataset. The third column, "GLM2 on Testing" indicates the accuracy when using $GLM2$ to predict the level of admission probability in the testing dataset. The last two columns have similar meanings to the second and third columns.

3.2.6 Classification Tree

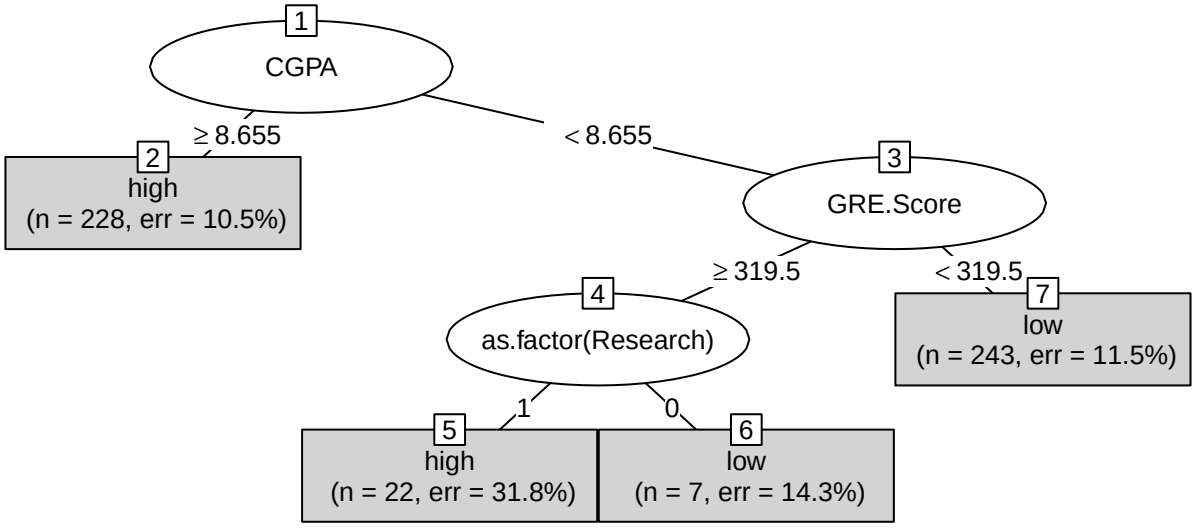


Figure 2: Classification Tree of Admission Probability Level

Figure 2 is the classification tree to predict the level of admission probability, including all of the seven factors in the dataset. The classification tree has seven nodes, and four of these nods are terminal nodes. Terminal nodes are indicated by rectangular boxes and nonterminal nodes by ovals.

3.2.7 Propensity Score Matching with Logistic Regression Model

	(1)
(Intercept)	-63.112 ***
	(7.506)
GRE.Score	0.092 ***
	(0.026)
CGPA	3.877 ***
	(0.547)
as.factor(Research)1	1.189 ***
	(0.321)
N	440
logLik	-138.612
AIC	285.224

*** p < 0.001; ** p < 0.01; * p < 0.05.

The propensity score is the probability of being treated, which in this study, which is the propensity score is the probability of given a research experience. The above is the table of the propensity score regression showing that all predictors are significant.

4. Discussion

4.1 Data Summary

- From Table 2, the number of observations in both groups is similar, which indicates *Admission* is non-biased data. In other words, the *Admission* dataset provides different perspectives, and the model generated by this data can be relatively convincible. Besides, all of the average GRE score, average TOEFL Score, and the average CGPA with high admission probability are significantly higher than the other group with low admission probability. Most of the people with high admission probability had research experience before (83.1%). However, only 29.4% of people in the group with low admission probability achieved some research experiences. There is a relatively large difference (53.7%) in the proportion of research experience between the two groups.
- From Figure 1, a significant upward trend exists indicating the GRE score factor and the chance of admission are positively correlated. When a undergraduates' GRE score increases, their chance of admission tends to increase simultaneously. The yellow data points are clustering in the top right corner while the blue data points are in the middle and bottom left corner, and if we look at this plot vertically, yellow data points are above blue data points in most cases. This phenomenon demonstrates that people who have research experiences will likely have a higher chance of admission than those who do not have research experience.

4.2 Model Summary

4.2.1 MLR

- Based on Table 3 and the fitted MLR model equation, the fitted MLR model's intercept is -1.336, which means the value of chance of admission is -1.336 when all other predictors take in a value of zero. When GRE score, TOEFL score, the strength of the letter of recommendation, and CGPA increase by one unit, the chance of admission will increase by 0.002, 0.003, 0.019, and 0.123, respectively. When a person has research experience, the admission chance value increases by 0.025 while the other factors remain unchanged. Moreover, the adjusted R^2 of this fitted MLR model is 0.8188 (see Appendix #2). This fitted model has explained an 81.88% variation of the chance of admission, indicating the goodness of fit of this model. Based on this fitted MLR model, all of these five predictors account for the change in admission chance.
- The model assumptions are checked by plotting the model. (see Appendix #3) The Residuals vs. Fitted plot is used to check the linear relationship assumptions. Here, we have a horizontal line without distinct patterns, which indicates a linear relationship. The Normal Q-Q plot is used to examine whether the residuals are normally distributed. Almost all the residuals points follow the straight dashed line in the Normal Q-Q plot, so we assume the residuals are normally distributed. The homoscedasticity can be checked by the Scale-Location plot, which is the residuals' homogeneity of variance. (Kassambara, 2018) The corresponding plot in Appendix #3 is a good indication of homoscedasticity.(Kassambara, 2018) Finally, from the Residuals vs. Leverage plot, a few influential cases may influence the regression result, but here I assume all model assumptions are satisfied.
- Additionally, the Variance Inflation Factors (VIF) scores of the fitted MLR model predictors are checked. The VIF scores identify the correlation between independent variables and their strength. VIF greater than 5 represents high multicollinearity where the coefficients are poorly estimated, and the p-values are questionable. (Daignault, 2020) Since all scores are lower than 5, all of these five predictors are not correlated, so the model result is meaningful.

4.2.2 Logistic Regression Models

- Since $GLM1$ and $GLM2$ contains seven predictors, the potential problem of overfitting exists. Overfitting data means that the original fitness is too well to use on predictions other than the current dataset cases. In other words, the model is not suitable for further data. Remove some predictors that can reduce the influence of overfitting. Therefore, only those three significant variables are chosen to generate a reduced logistic regression model. (The p-values in Tables 4 and 5 indicate that only the GRE score, CGPA, and research experience are significant.)
- For model validation propose, training and testing datasets are randomly generated from the original dataset. They are resulting in a training dataset with 400 observations and a testing dataset with 100 observations. Using the fitted logistic model, which is generated based on the training dataset to predict the testing data observations, allows us to generate the confusion matrix (See Appendix #6) and measure the fitted model's accuracy. In Table 8, we can see the accuracy of each logistic regression model. When fitting a full logistic regression on the training data, the accuracy decreases, which indicates the model may have overfitted the data. The accuracy of the same model used to predict the original data, training data, and testing data should not significantly differ. If there is a large difference, then the corresponding fitted model may be overfitting and not valid for most cases. Since the accuracy when using $GLM4$ to predict the level of admission probability in the training dataset is 0.9; the accuracy when using $GLM4$ to predict the level of admission probability in the testing dataset is 0.86; the accuracy that using the same model to predict the original dataset is 0.88, the reduced logistic model ($GLM3$) are chosen.

- From the model equations listed in the result section, $\hat{p}$ can be calculated for the probability of admission level. Based on $GLM3$, for every one unit increases in the GRE score, the log odds increase by 0.0938. For every one unit increases in the CGPA, the log odds increase by 3.918. Moreover, if there is research experience, the log odds increase by 1.197. GRE score, CGPA, and research experience are considered the most significant factors for admission probability level. For the model assumptions check in this case, although there is some small violation in assumptions, we assumed this model meets all of the model assumptions.

4.2.3 Classification Tree

- Figure 2 shows that the most critical factors that affect the classification of the level are the CGPA, GRE score, and research experience. The classification tree predicts that if one's CGPA is greater or equal to 8.655, it is classified as a high probability of acceptance. If the CGPA is lower than 8.655, we should look at the GRE score and the research experience. When the GRE score is less than 319.5, this person is predicted to have a low probability of acceptance. On the other hand, when a person's GRE score $\geq$ 319.5 and he has research experience, this person is predicted to have a high probability of being admitted. If he does not have any research experience, he is classified as having a low acceptance chance. n is the number of observations that are classified under each terminal, and the corresponding err is the proportion in the terminal that is misclassified by this decision tree.

4.2.4 Propensity Score Matching with Logistic Regression Model

- Since an observational study is used in this study, the non-randomization of subjects to interventions may cause uncertainties and may be the source of the effects rather than the intervention or procedure alone. Thus, a propensity score matching technique is introduced in this study to adjust the data by balancing the covariates between the treatment groups. (Lanza, 2013) After these steps, there will be clear proof that the difference in results is due to the difference in outcomes. The covariates observed are balanced at each propensity score value; this means that the covariates' distributions are the same for students in the treated and control groups of the same propensity score. (Lanza, 2013) If two observations have a similar propensity score with different research experience indicator value in this technique, then these two observations are matched. All of the un-matched observations are removed in the matching dataset. By doing this, the data becomes less biased and more similar to an experimental study, which produces a more accurate and convincible result.

4.3 Conclusion

- To conclude, the most significant factors that affect the U.S. Graduate school admission decision are GRE score, CGPA, and research experience. A logistic regression model with these three predictors is useful in predicting the chance of admission since it has a relatively high accuracy of 88%. The classification tree gives the same result. This logistic regression model is validated by splitting the dataset into training and testing data, fitting the models using the training dataset to predict the testing. The propensity score result indicates that people with research experience are more likely to have a high admission chance. The causal inference exists between the research experience and the level of the probability of admission.
- From the AIC criterion, the multiple linear regression model has the smallest AIC value (Appendix #5), which determines it has the best fit for the data among all models in this study. AIC evaluates the model's goodness of fit of the data. The five predictors, GRE score, TOEFL score, the strength of recommendation letter, CGPA, and research experience, have explained about 82% of the variation in the level of admission chance. Therefore, the TOEFL score and the recommendation letter's strength can be treated as minor factors that affect the U.S. Graduate school admission result.

4.4 Weaknesses

- The size of the data is not very large (500 observations), so there is a chance that our dataset is not very representative of the group of undergraduates. Especially during the propensity score matching step, only 220 pairs (400 observations left) in the matching dataset may cause some small deviation in the model results. Some further cleaning processes, such as recategorizing the response variables based on a specific self-chosen-scale, may affect the outcome. Also, the variable chance of admission is created by Mohan S Acharya using his perspective. Thus, there is the possibility of the existence of bias while assigning the chance of admitting score. Moreover, this is observational data with features collected from Indian undergraduates who applied for U.S. graduate schools, which may not represent the undergraduates throughout the world.
- There are some violations in the model assumptions, although it is assumed to meet all assumptions. For the residual plots of logistic regression, patterns exist, so there are some assumption violations. This may be due to the wrong selected variable and the dependence between variables that did not notice. These small violations may decrease the precision of our models and causes uncertainties. Some influential cases should also be carefully considered and determined whether to include in the analysis or not. Moreover, the models are chosen based on my intuition, so it may not be the best model, leading to prediction's inaccuracy.

4.5 Next Steps

- Improve the current models by applying some transformation (such as taking the log, square, square root, etc.) on the selected variables to satisfy the model assumptions. Fitting another model using different techniques such as Bayesian Models on this data and check if it produces a different result. Try to recategorized the response variable using a different scale and see what the outcome is.
- Since the data size is not very large, we can try to find some other related dataset online and combine into a more extensive data set if possible. If there is no similar data set online, we can consider building an online survey and posting it on the Internet to collect data related to this topic. Consider designing surveys in a way that attracts people, such as awarding small prizes.

References

- Caetano, Samantha-Jo "STA304H1: Surveys, Sampling and Observational Data Weekly Slides." Department of Statistical Sciences University of Toronto. December 2020.
- Alexander, Rohan "STA304H1: Surveys, Sampling and Observational Data Weekly Slides." Department of Statistical Sciences University of Toronto. December 2020.
- Daignault, Katherine. "STA302/1001: Methods of Data Analysis 1 Weekly Slides." Department of Statistical Sciences University of Toronto. June 2020.
- Acharya, M. (2018, December 28). Graduate Admission 2. Retrieved December 22, 2020, from https://www.kaggle.com/mohansacharya/graduate-admissions
- 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
- Your Machine Learning and Data Science Community. (n.d.). Retrieved December 22, 2020, from https://www.kaggle.com/
- Alexander, R. (2020, November 05). Difference in differences. Retrieved December 22, 2020, from https://www.tellingstorieswithdata.com/06-03-matching_and_differences.html
- Espenshade, T., Chung, C., & Walling, J. (2004, December 21). Admission Preferences for Minority Students, Athletes, and Legacies at Elite Universities*. Retrieved December 22, 2020, from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0038-4941.2004.00284.x
- ETS GRE (2020). About the GRE® General Test. (n.d.). Retrieved December 22, 2020, from https://www.ets.org/gre/revised_general/about
- ETS TOEFL (2020). The TOEFL® Family of Assessments. (n.d.). Retrieved December 22, 2020, from https://www.ets.org/toefl
- Kassambara, Visitor, & Mann, T. (2018, March 11). Linear Regression Assumptions and Diagnostics in R: Essentials. Retrieved December 22, 2020, from http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/
- Lanza, S., Moore, J., & Butera, N. (2013, December). Drawing causal inferences using propensity scores: A practical guide for community psychologists. Retrieved December 22, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4098642/
- Cite R : R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Cite "tidyverse": Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Cite "ggplot2" : H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Cite "dplyr" : Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr
- Cite "lme4": Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.
- Cite "knitr": Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.27.
- Cite "kableExtra": Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra

- Cite “devtools”: Hadley Wickham, Jim Hester and Winston Chang (2020). devtools: Tools to Make Developing R Packages Easier. R package version 2.3.2. https://CRAN.R-project.org/package=devtools
- Cite “jtools”: Long JA (2020). *jtools: Analysis and Presentation of Social Scientific Data*. R package version 2.1.0, <URL: https://cran.r-project.org/package=jtools>.
- Cite “pander”: Gergely Daróczi and Roman Tsegelskyi (2018). pander: An R ‘Pandoc’ Writer. R package version 0.6.3. https://CRAN.R-project.org/package=pander
- Cite “car”: John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/
- Cite “rpart”: Terry Therneau and Beth Atkinson (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. https://CRAN.R-project.org/package=rpart
- Cite “partykit”: Torsten Hothorn, Achim Zeileis (2015). partykit: A Modular Toolkit for Recursive Partytioning in R. Journal of Machine Learning Research, 16, 3905-3909. URL http://jmlr.org/papers/v16/hothorn15a.html
- Cite “broom”: David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.0. https://CRAN.R-project.org/package=broom

Appendix

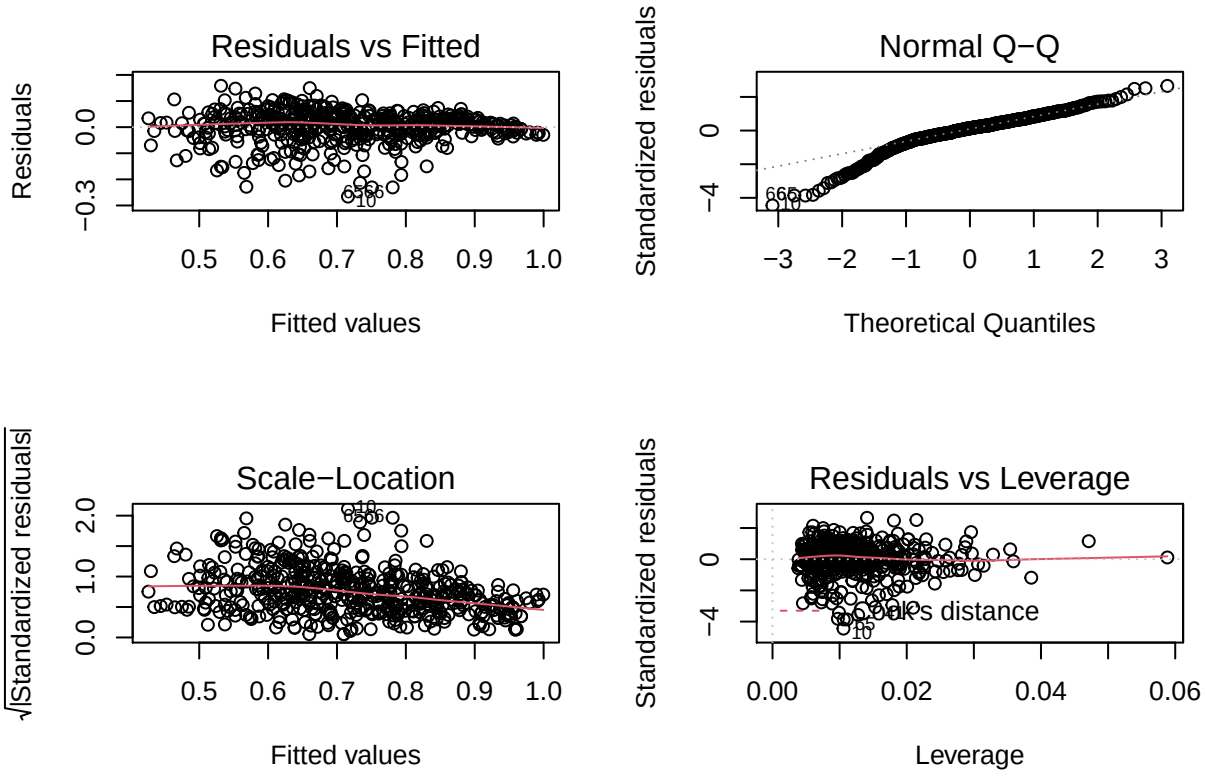
1. The Full Multiple Linear Regression Model

Table 9: Summary Table of Full MLR Model

term	estimate	std.error	statistic	p.value
(Intercept)	-1.2757	0.1043	-12.2317	0.0000
GRE.Score	0.0019	0.0005	3.6998	0.0002
TOEFL.Score	0.0028	0.0009	3.1842	0.0015
University.Rating	0.0059	0.0038	1.5628	0.1188
SOP	0.0016	0.0046	0.3476	0.7283
LOR	0.0169	0.0041	4.0743	0.0001
CGPA	0.1184	0.0097	12.1982	0.0000
as.factor(Research)1	0.0243	0.0066	3.6798	0.0003

2. The Adjusted R^2 of MLR Model is 0.8188449.

3. MLR Model Assumption Check



4. Data Visualization after Propensity Score Matching

Table 10: Data Visualization after Propensity Score Matching

GRE	TOEFL	U Rating	SOP	LOR	CGPA	Research	Admit Chance	Admit.Prob	.fitted
290	104	4	2.0	2.5	7.46	0	0.45	low	0.0356
290	100	1	1.5	2.0	7.56	0	0.47	low	0.0376
294	93	1	1.5	2.0	7.36	0	0.46	low	0.0520
295	93	1	2.0	2.0	7.20	0	0.46	low	0.0530
295	96	2	1.5	2.0	7.34	0	0.47	low	0.0573
293	97	2	2.0	4.0	7.80	1	0.64	low	0.0595

Where "U Rating" = $University.Rating$ and "Admit Chance" = $Chance.of Admit$

5. AIC of the MLR Model is -1385.299.

6. Confusion Matrices

Table 11: Confusion Matrix GLM1

	high	low
high	215	26
low	33	226

Table 12: Confusion Matrix GLM2 on Training

	high	low
high	173	24
low	28	175

Table 13: Confusion Matrix GLM2 on Testing

	high	low
high	41	4
low	6	49

Table 14: Confusion Matrix GLM3

	high	low
high	215	29
low	33	223

Table 15: Confusion Matrix GLM4 on Training

	high	low
high	176	25

	high	low
low	25	174

Table 16: Confusion Matrix GLM4 on Testing

	high	low
high	38	5
low	9	48

The number of correct classifications is on the diagonal, and the wrong classifications are off the diagonal.