STA212 Project 2

Frank Liu & David Qu

May 7, 2020

Abstract

The goal of this project is to analyze and build a model for predictions of whether a student can be accepted into graduate school by using six predictors. The dataset we use is from Kaggle websites called "Admission" that include 400 information of students. We conduct EDA process, and we explore potential predictors, check the linearity, outliers, and multicollinearity. We decide to not transform any predictors and keep all predictors in the rough model. We applied both hypothesis tests and Nested Likelihood Ratio Test to refine our rough model. Then we conduct BSS, and get the same refined model as the Nested Likelihood Ratio Test. Other indicators like AIC and residual deviance confirmed our model. All conditions of logistic regression are met. Our final model is:

$log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -3.989979 + 0.002264 \times GRE + 0.804038 \times GPA - 0.675443 \times Rank2 - 1.340204 \times Rank3 - 1.551464 \times Rank4$

The result matches our intuition, which means higher GRE score, higher GPA, and more prestigious undergraduate college all contribute to a higher log odds of getting into graduate school. Limitations of this model is the sample size (400) isn't large enough for the model to be a credible reference. But we certainly can gain some insights from it for future students who plan to apply for graduate school. Predictors that we removed (SES, gender, race) do play a role in the admission process, but they are not significant enough to be included.

Data

The data that we choose for this project is about the information of students who applied to graduate school, including whether or not they were accepted (response variable), their college GPA, GRE score, gender, races, social economic status(SES), and the prestige of their undergraduate institution (explanatory variables). We want to use all given predictors to build a logistic regression model for whether a students will be accepted into a graduate school. The data comes from Kaggle websites by searching the keywords" college admission", and it is called Admission, uploaded by Thokala Eswar Chand on Oct, 2019. The APA citation of this dataset is:

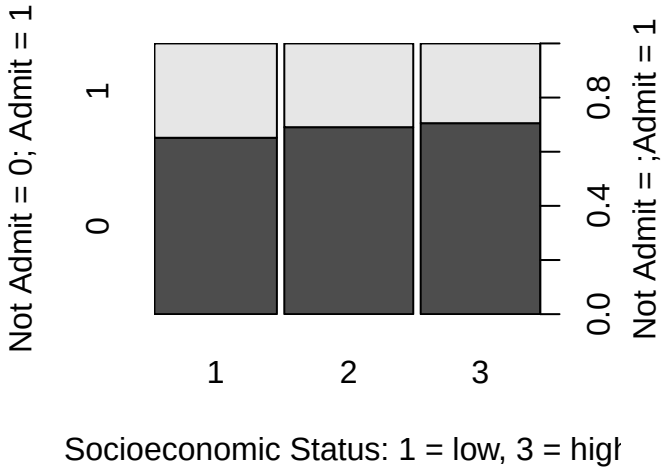
Admission. (n.d.). Retrieved May 2, 2020, from https://kaggle.com/eswarchandt/admission

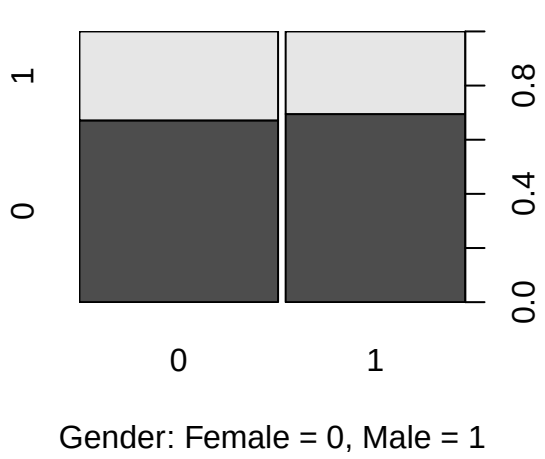
Each row in the dataset represents an application submitted by a student. The response variables is categorical: Admit / Not admit. Other predictors include: Old Version of GRE Score: (numeric: 200 - 800); College GPA: (numeric: 0.00 - 4.00); Socioeconomic Status: (categorical:l 1 - low, 2 - medium, 3 - high); Gender: (categorical: 0-female 1-male); Race: (categorical 1- Hispanic, 2- Asian, 3 - African-American); and Rank: (categorical 1 through 4. Student's original college institutions with a rank of 1 have the highest prestige, while those with a rank of 4 have the lowest). There are 400 rows in our data set and after checking, there's no missing data.

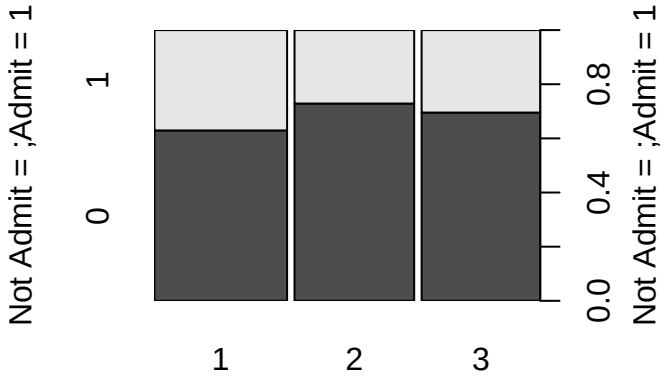
Exploratory Data Anlysis

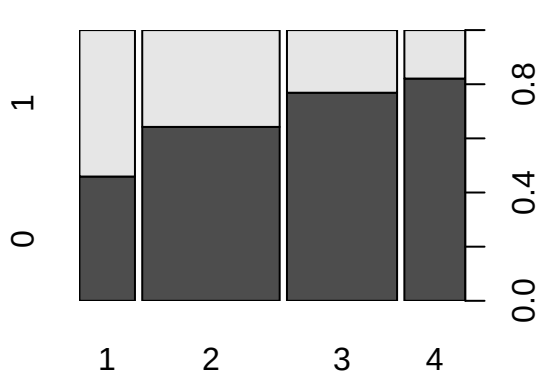
Exploring Potential Predictors

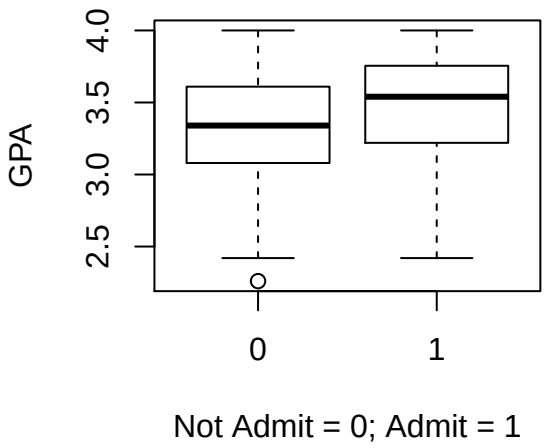
In this section, all potential predictors will plot with the response variables to see if they should be included in the rough model. Specifically, since our response variables is binary, side-by-side box plot will be made with numeric variables, such as GRE score and GPA, and mosaic plot will be made with categorical variables, such as Socioeconomic Status, Gender, Race, and Rank.

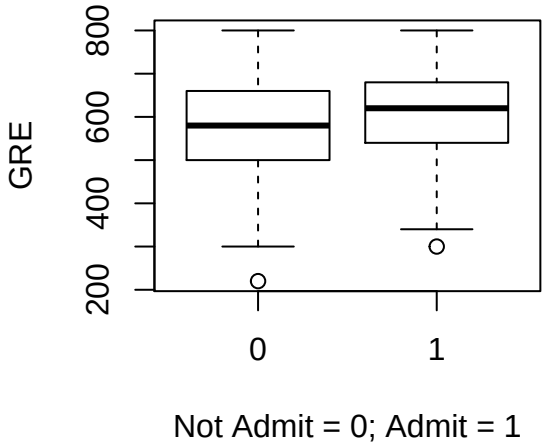

Not Admit = 0; Admit = 1
0
1
0.0
0.4
0.8
1
2
3
Socioeconomic Status: 1 = low, 3 = high


Not Admit = 0; Admit = 1
0
1
0.0
0.4
0.8
0
1
Gender: Female = 0, Male = 1


Not Admit = 0; Admit = 1
0
1
0.0
0.4
0.8
1
2
3
Race: 1 = Hispanic, 2 = Asian, 3 = African Am


Not Admit = 0; Admit = 1
0
1
0.0
0.4
0.8
1
2
3
4
Rank: 1 = high, 4 = low

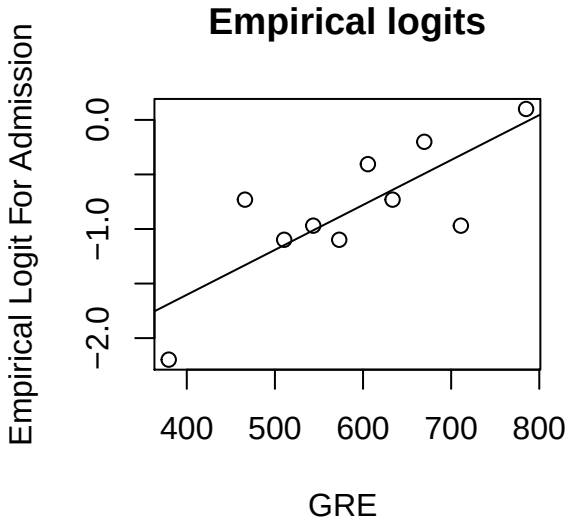


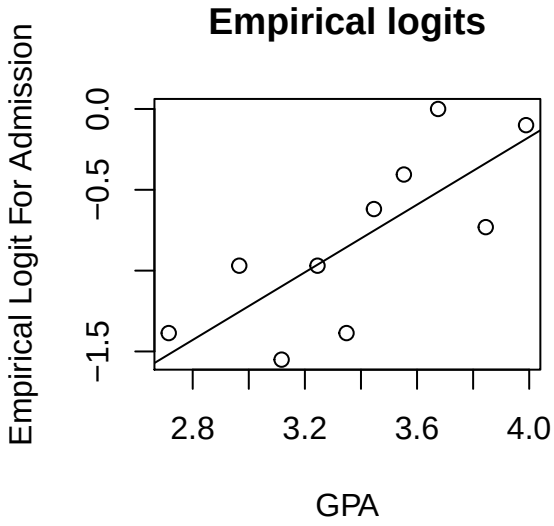


By looking at the plots, we decide to save all the predictors for our rough model, because they may all influence the probability of being admitted. According to the side-by-side box plot, although GRE and GPA seems to be less effective since there's overlap in boxes, we are still unsure if they should be excluded from our model. It is always safe to keep it at this moment. According the mosaic plot, different levels of rank and race have an apparent difference in admitted or not. For SES and gender, even the difference is small, we cannot guarantee that they are definitely not a predictor. Hence, at this point, we decide to keep all 6 predictors in our rough model.

Linearity and Multicollinearity

I will make the empirical logit plot for all potential numeric predictors with the response variables to see if the variables meet the linear(shape) condition.





We think the condition of linearity is satisfied because all the data points or bins are randomly distributed around the best fit line, and there is no obvious outlier. We conclude that we do not need to apply any transformation on our numeric variables because both GRE and GPA seem to meet the conditions of logistic regression. For categorical variables, we will not change their levels as well, because all the levels have a similar amount of data points, and they all indicate a difference in the proportion of being admitted.

Just in case, we applied log transformation to GRE and GPA and compare the deviance of the transformed model with original model to make sure transformation is not necessary. For our rough model that include all six predictors, the Residual deviance is 453.93. The residual deviance with log transformed GRE and log transformed GPA are 453.80 and 454.13. We see a tiny drop in model with log GPA, so either form of predictors will be fine. We decide to keep these two numeric variables as original form to make the model tidy.

```
## [1] 453.9313
```

```
## [1] 453.8011
```

```
## [1] 454.125
```

Since we have only two numeric variables, we will check the correlation between these two numeric variables(GPA and GRE score).

```
## [1] 0.3842659
```

And the correlation is 0.3843, which is less than 0.6. Hence, we don't need to worry about the effect of multicollinearity.

Check for outliers

```
## 316
## 316
```

```
## [1] 453.9313
```

```
## [1] 450.0343
```

We fit a logistic regression model for our two numerical variables. And we calculate the standard residuals for each of the model. We found that there is a general outlier, but we don't have any extreme outlier. By removing this data point, our deviance of the rough model drops from 453.93 to 450.03 After looking at this data point, we decide not to remove it. The application of this student is admitted. Although he has a low GRE and GPA, other factors may have an impact on his application, such as socioeconomic status, race, and gender. That's very much like what Wake Forest did, who review students holistically and focus more from their achievements outside of classroom rather than focus too much on standardized score or GPA. Therefore, we will not change our dataset.

Therefore, our rough model would be:

$log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \beta_0 + \beta_1 SES2 + \beta_2 SES3 + \beta_3 GenderMale + \beta_4 RaceAsian + \beta_5 RaceAfrican + \beta_6 Rank2 + \beta_7 Rank3 + \beta_8 Rank4 + \beta_9 GPA + \beta_{10} GRE$

Model Selection

For model selection, we first test the significance of predictors in our rough model. We based on summary of our model and select predictors that have a p-value larger than 0.05. We consider to remove the predictors that are not significant, and we derive our first model (smaller model). Then we compare the smaller model with the rough model by conducting a Nested Likelihood Ratio Test.

The predictors that have a p-value larger than 0.05 are SES2, SES3, Gender_Male, Race2, and Race3. We remove the predictors SES, gender, and race to build a smaller model.

Nested Likelihood Ratio Test:

```
##                            Estimate Std. Error z value  Pr(>|z|)
## (Intercept)             -3.5491260  1.1638778 -3.0494 0.0022930
## gre                      0.0022880  0.0011048  2.0709 0.0383709
## gpa                      0.8146862  0.3359995  2.4247 0.0153225
## as.factor(ses)2         -0.1350207  0.2768419 -0.4877 0.6257497
## as.factor(ses)3         -0.2580326  0.2849643 -0.9055 0.3652052
## as.factor(Gender_Male)1 -0.1916110  0.2304171 -0.8316 0.4056443
## as.factor(Race)2        -0.4856333  0.2825868 -1.7185 0.0857004
## as.factor(Race)3        -0.3130361  0.2752386 -1.1373 0.2554018
## as.factor(rank)2        -0.7115677  0.3216470 -2.2123 0.0269485
## as.factor(rank)3        -1.3607226  0.3503117 -3.8843 0.0001026
## as.factor(rank)4        -1.5804684  0.4229561 -3.7367 0.0001864
##
## n = 400 p = 11
## Deviance = 453.93133 Null Deviance = 499.97652 (Difference = 46.04519)
```

```
##                     Estimate Std. Error z value  Pr(>|z|)
## (Intercept)      -3.9899791  1.1399509 -3.5001 0.0004650
## gre               0.0022644  0.0010940  2.0699 0.0384651
## gpa               0.8040375  0.3318193  2.4231 0.0153879
## as.factor(rank)2 -0.6754429  0.3164897 -2.1342 0.0328288
## as.factor(rank)3 -1.3402039  0.3453064 -3.8812 0.0001039
## as.factor(rank)4 -1.5514637  0.4178316 -3.7131 0.0002047
##
## n = 400 p = 6
## Deviance = 458.51749 Null Deviance = 499.97652 (Difference = 41.45903)
```

```
## Analysis of Deviance Table
##
## Model 1: as.factor(admit) ~ gre + gpa + as.factor(rank)
## Model 2: as.factor(admit) ~ gre + gpa + as.factor(ses) + as.factor(Gender_Male) +
##     as.factor(Race) + as.factor(rank)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       394     458.52
## 2       389     453.93  5   4.5862   0.4684
```

The hypotheses are: H_0 : stick with refined model. H_a : move to rough model.

The p-value for the test is 0.4684, which is obviously higher than the cutoff. Therefore, we fail to reject the null hypothesis. We do not have strong evidence to move to the rough model(larger model), and we stick with the refined model (smaller model).

We also perform the method of best subset selection(BSS) to choose our final model according to some metric. Since we are constructing a multiple logistic model, we are looking for the model with the smallest AIC.

```
## Loading required package: leaps
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
##
## Call:  glm(formula = y ~ ., family = family, data = Xi, weights = weights)
##
## Coefficients:
##        (Intercept)                 gre                 gpa  `as.factor(rank)2`
```

```
##          -3.989979            0.002264            0.804038           -0.675443
## `as.factor(rank)3`  `as.factor(rank)4`
##          -1.340204           -1.551464
##
## Degrees of Freedom: 399 Total (i.e. Null);  394 Residual
## Null Deviance:       500
## Residual Deviance: 458.5     AIC: 470.5
```

```
## [1] 475.9313
```

By using the code of BSS, the best model is the same as the model we derived previously, which include rank, GRE, and GPA. It has an AIC of 467.4. Our rough model has an AIC of 475.93, which means our final model indeed has a lower AIC and it is more preferable.

Final Model and Conditions for Inference

```
##                     Estimate Std. Error z value  Pr(>|z|)
## (Intercept)      -3.9899791  1.1399509 -3.5001 0.0004650
## gre               0.0022644  0.0010940  2.0699 0.0384651
## gpa               0.8040375  0.3318193  2.4231 0.0153879
## as.factor(rank)2 -0.6754429  0.3164897 -2.1342 0.0328288
## as.factor(rank)3 -1.3402039  0.3453064 -3.8812 0.0001039
## as.factor(rank)4 -1.5514637  0.4178316 -3.7131 0.0002047
##
## n = 400 p = 6
## Deviance = 458.51749 Null Deviance = 499.97652 (Difference = 41.45903)
```

Our final model is:

$log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -3.989979 + 0.002264 \times GRE + 0.804038 \times GPA - 0.675443 \times Rank2 - 1.340204 \times Rank3 - 1.551464 \times Rank4$

All the conditions for logistic regression is achieved by our model. 1) We have a binary response variable. 2) As we have checked in EDA, the shape of our model is approximately linear. For 3) Randomness and 4) Independence, since we obtained our dataset from Kaggle, we have no clue how the data is collected. There's no clear description of the data collection process, but from the objective and background of these data sources, it's safe to assume that this data is from the education department in the US. Because the institution (educational department) is credible, we assume that the dataset is collected properly and the assumption of independence and randomness is achieved. That means the sample of students are collected at random, and the admission of one student into graduate school will not influence the chance of another student getting into graduate school. Therefore, all conditions for logistic regression model are satisfied.

Analysis and Conclusion

After deciding our final model, we find that our response variable, the log odds of being admitted to a graduate school, is related to GRE score, GPA, and the rank of the applicant's undergraduate school. Notes that a higher log odds also indicate a higher probability. In particular, the lower prestige of the rank of the school (as rank variable increase from 1 to 4), the more negative is the slope of the predictor. For GRE and GPA, they have a positive slope, which means the probability of being admitted is higher if the applicant has a higher score in GRE and GPA. The result of model matches our intuition. GRE score, GPA, and rank of school all indicate a higher probability of being admitted. Also notice that the predictors like gender, socioeconomic status, and race that in our rough model are removed in our final model. This tells that the application system is relatively fair, that all the predictors that are unrelated to academics are insignificant

in our model. Even they are part of the consideration of whether or not students being admitted, they are not significant enough to be included at least in our model. Our model is very reliable because we meet all the conditions of logistic regression. Also, the final model decided by BSS is the same as the model if we look at the significance of the predictors, which means that it is clearly the best option. The only problem that we concern is the source of the dataset, because it only has 400 data points and we don't know if the sample is large enough to describe the situation of all applicant.

Our model could be an appropriate reference for formulating strategies for students who wants to apply graduate school in the future.