STAT 551 Logistic Regression Assignment

Yuchi Hu

February 10, 2019

Problem 1

It is not useful to model this probability directly as a linear combination of the predictors because some estimated probabilities end up either negative or greater than 1. To handle this problem, we should model this probability using a function that gives outputs between 0 and 1 for all values of the predictor(s). In logistic regression, we use the logistic function, which satisfies this criteria.

Problem 2

Part 2.1

First, we look at the numerical summary of the data.

Table 1: Numerical Summary of MyData

admission	GRE	GPA	Rank	GPAcategorized
Min. :0.0000	Min. :300.0	Min. :2.260	Min. :1.000	High :141
1st Qu.:0.0000	1st Qu.:520.0	1st Qu.:3.130	1st Qu.:2.000	Low : 57
Median :0.0000	Median :590.0	Median :3.385	Median :2.000	Medium:152
Mean :0.3029	Mean :587.8	Mean :3.393	Mean :2.486	NA
3rd Qu.:1.0000	3rd Qu.:660.0	3rd Qu.:3.670	3rd Qu.:3.000	NA
Max. :1.0000	Max. :800.0	Max. :4.000	Max. :4.000	NA

We can see from **Table 1** that the median of *admission* is 0 (did not get admitted to grad school), the median of *GRE* is 590, the median of *GPA* is 3.385, and the median of *Rank* (prestige status of undergrad school) is 2. In terms of *GPAcategorized*, there are many more students with “Medium” or “High” GPA than those with “Low” GPA.

Next, we look at some histograms and bar charts of the variables.

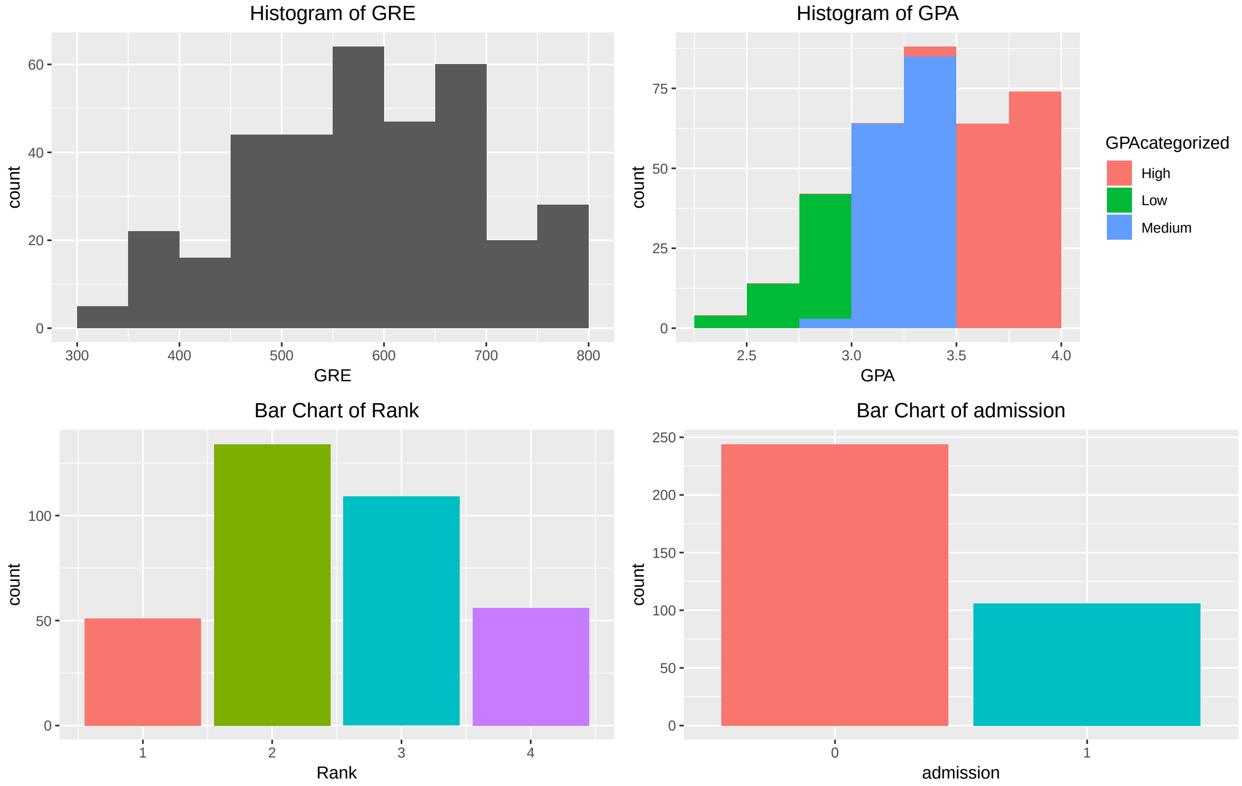

Figure 1: Histograms and bar charts

From the top left plot of **Figure 1**, we can see that *GRE* is approximately normally distributed with most values falling between 450 and 700. The top right plot shows the histogram of *GPA* colored by *GPAcategorized*. We can see that the distribution of *GPA* is left-skewed with more values falling above 3.0 than below it. The bottom left plot shows that the values of *Rank* are mostly 2 and 3. The bottom right plot shows that the values of *admission* are predominantly 0 (did not get admitted to grad school).

To get an idea of the relationship between each predictor and *admission*, we construct some boxplots and mosaic plots.

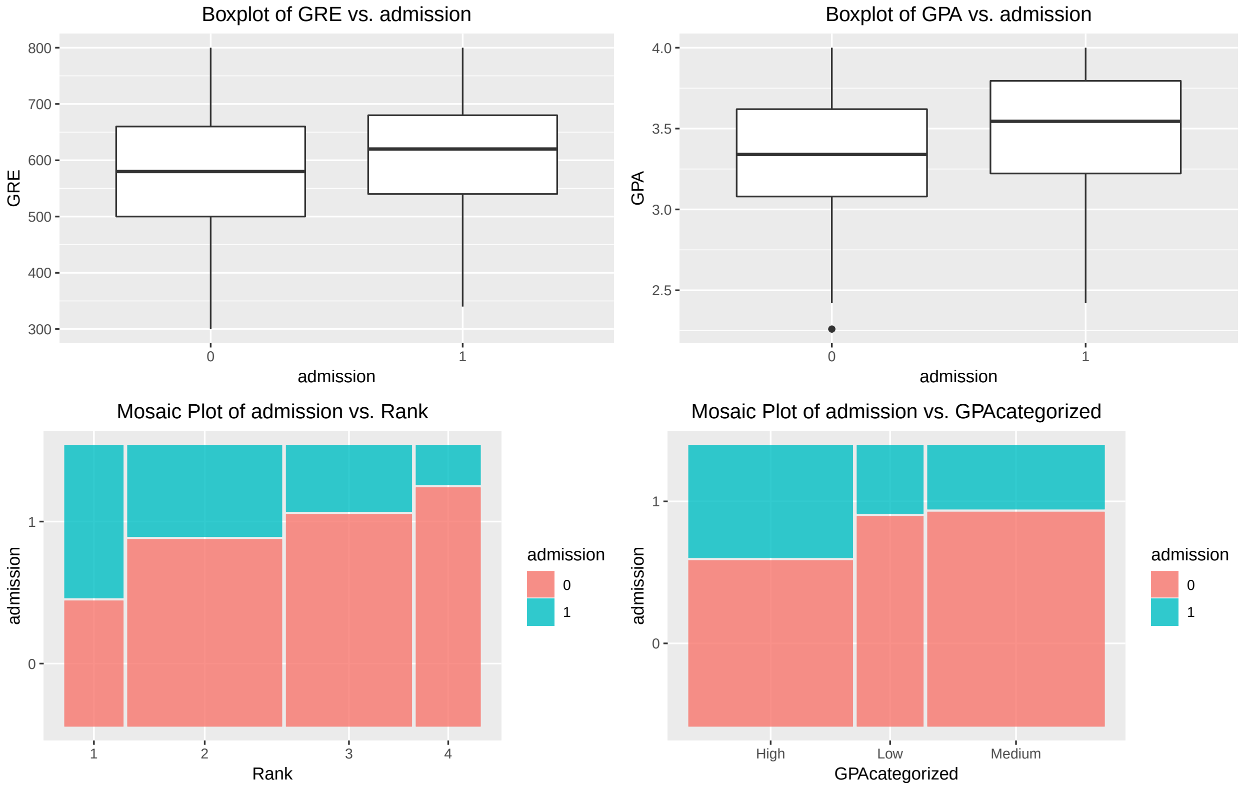

Figure 2: Boxplots and mosaic plots

From the top left and top right plots of **Figure 2**, we can see that both higher values of *GRE* and *GPA* tend to correspond to *admission* = 1 (admitted to grad school). The lower left plot shows that as the value of *Rank* increases (prestige status of the undergrad school decreases), the relative proportion of students who did not get admitted increases. In other words, students from more prestigious undergrad schools are more likely to be admitted to grad school. The lower right plot shows that students are more likely to be admitted with a "High" GPA than with a "Low" or "Medium" GPA although the admission rate for students with "Low" and "Medium" GPA's are about equal.

Based on these plots, we should include all of the variables in a predictive model. **Naturally, we should only use one of *GPA* and *GPAcategorized*.**

Problem 3

Part 3.1

We fit a logistic regression model (logisticModel1) with *admission* as the response and *GRE* and *GPAcategorized* as the predictors. The summary of logisticModel1 is below:

```
##
## Call:
## glm(formula = admission ~ GRE + GPAcategorized, family = binomial,
##     data = MyData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2368  -0.8554  -0.6959   1.2547   2.0463
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.409566   0.752161  -3.204  0.00136 **
## GRE                   0.003185   0.001147   2.778  0.00547 **
## GPAcategorizedLow    -0.447423   0.369577  -1.211  0.22604
## GPAcategorizedMedium -0.635577   0.267484  -2.376  0.01750 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 429.29  on 349  degrees of freedom
## Residual deviance: 409.88  on 346  degrees of freedom
## AIC: 417.88
##
## Number of Fisher Scoring iterations: 4
```

From the summary, we can see that the parameters for *GRE* and *GPAcategorizedMedium* are significant at $\alpha = 0.05$. The parameter for *GPAcategorizedLow* is not significant at $\alpha = 0.05$. The AIC of LogisticModel1 is 417.88.

Two parameters were estimated for the *GPAcategorized* variable in LogisticModel1. The parameter for *GPAcategorizedLow* is -0.447, so the odds of *admission* for a student with "Low" GPA is about exp(-0.447) = 63.9% of the odds of *admission* for a student with "High" GPA. The parameter for *GPAcategorizedMedium* is -0.636, so the odds of *admission* for a student with "Medium" GPA is about exp(-0.636) = 53% of the odds of *admission* for a student with "High" GPA.

The "High" level of *GPAcategorized* is taken as reference.

The parameter for *GRE* is 0.0032, and exp(0.0032) = 100.32%. So, it estimates an increase of 0.32% in the odds of *admission* for each unit increase in *GRE.*

Part 3.2

We change the reference level of *GPAcategorized* to “Low”. Then, we fit the same logistic regression model (logisticModel2). The summary of logisticModel2 is below:

```
##
## Call:
## glm(formula = admission ~ GRE + GPAcategorized, family = binomial,
##     data = MyData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2368  -0.8554  -0.6959   1.2547   2.0463
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.856988   0.707198  -4.040 5.35e-05 ***
## GRE                   0.003185   0.001147   2.778  0.00547 **
## GPAcategorizedHigh    0.447423   0.369577   1.211  0.22604
## GPAcategorizedMedium -0.188154   0.369228  -0.510  0.61034
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 429.29  on 349  degrees of freedom
## Residual deviance: 409.88  on 346  degrees of freedom
## AIC: 417.88
##
## Number of Fisher Scoring iterations: 4
```

We can see that the estimates of the parameters (excluding *GRE*) changed. This is because the reference level of *GPAcategorized* is now “Low” instead of “High”.

The parameter for *GPAcategorizedHigh* is 0.447, so the odds of *admission* for a student with “High” GPA is about exp(0.447) = 156.4% of the odds of *admission* for a student with “Low” GPA. The parameter for *GPAcategorizedMedium* is -0.188, so the odds of *admission* for a student with “Medium” GPA is about exp(-0.188) = 82.8% of the odds of *admission* for a student with “Low” GPA.

Part 3.3

From logisticModel1, for a student with a GRE of 650 and “High” GPA, the probability of being ADMITTED is

$$\hat{p}(admission = 1) = \frac{exp(-2.409566 + 0.003185 * GRE)}{1 + exp(-2.409566 + 0.003185 * GRE)} = \frac{exp(-2.409566 + 0.003185 * 650)}{1 + exp(-2.409566 + 0.003185 * 650)} = 0.416$$

Thus, the probability of being rejected is 1 - 0.416 = **0.584**.

From logisticModel2, for a student with a GRE of 650 and “High” GPA, the probability of being ADMITTED is

$$\hat{p}(admission = 1) = \frac{exp(-2.856988 + 0.447423 + 0.003185 * GRE)}{1 + exp(-2.856988 + 0.447423 + 0.003185 * GRE)} = \frac{exp(-2.409565 + 0.003185 * 650)}{1 + exp(-2.409565 + 0.003185 * 650)} = 0.416$$

Thus, the probability of being rejected is 1 - 0.416 = **0.584**.

We get the same results using predict().

```r
1 - predict(logisticModel1, newdata=data.frame(GRE=650, GPAcategorized='High'), type='response')
```

```
##         1
## 0.5839985
```

```r
1 - predict(logisticModel2, newdata=data.frame(GRE=650, GPAcategorized='High'), type='response')
```

```
##         1
## 0.5839985
```