Predicting Admission Chance to Random University: A Logistic Regression Project

By: Steven Rivera

Introduction:

A search engine inquiry for school name and 'what are my chances' will bring will result in hundreds to thousands of results. From undergraduate to specific graduate programs to law, medicine, nursing, etc., there is a collective anxiety for all applicants that wish to start on their career path to individuals trying to change careers through education.

Certainly, there will always be anecdotes and comparisons of raw data like GPA and GRE, work and volunteer experience, however this project will attempt to create a multiple regression model to predict admission to Random University using a limited, prepared and cleaned dataset.

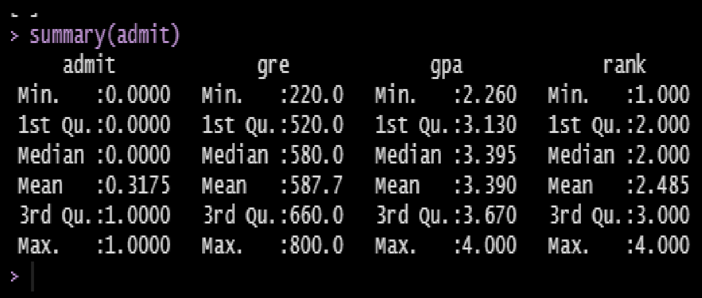
The Data:

The dataset is called ADMIT and was acquired through course DSC 423: Regression Analysis taught by Dr. Jonathan Gemmell a professor of computer science and data science at DePaul University. The dataset is in the .csv format. It contains 400 rows representing students and four (4) columns. The dependent variable (Y) 'admit' is a binomial variable where 0 = not admitted and 1 = admitted. The independent variables are GRE scores with a minimum score of 220 and a maximum score of 800, GPA on a standard 4.0 scale and finally rank of the student's high school on a scale of 1 to 4 (1 being the highest rank and 4 being the lowest) which I take to be levels much like divisions of college basketball teams.

Please note that while this dataset is through the previously mentioned course, this simple project to showcase multiple linear regression is personal and was not in any way part of coursework for either submission or grading purposes and is my attempt to practice logistic regression on a controlled and minimal dataset. As such, all the provided data is valid and was received in working order with the only change making the high school ranking variable a factor for the four (4) levels (three (3) dummy variables).

Data Analysis:

To start with analyzing and trying to understand the data, I pulled up a summary on the entire data set as well as histograms of the numerical data GRE and GPA. While interesting, this gives us minimal insight into the probability of admittance. As we see here in the summary, of the 400 students that applied, 127 were admitted which gives applicants a 31.75% chance of getting into Random University. While not nearly as low as the Ivy League schools (somewhere around 6%), Random University is more selective than the average college in the United States at 68%.[1]



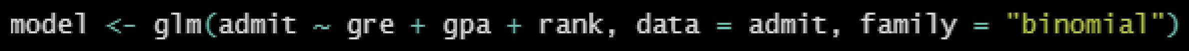
```
> summary(admit)
     admit             gre             gpa             rank      
 Min.   :0.0000   Min.   :220.0   Min.   :2.260   Min.   :1.000  
 1st Qu.:0.0000   1st Qu.:520.0   1st Qu.:3.130   1st Qu.:2.000  
 Median :0.0000   Median :580.0   Median :3.395   Median :2.000  
 Mean   :0.3175   Mean   :587.7   Mean   :3.390   Mean   :2.485  
 3rd Qu.:1.0000   3rd Qu.:660.0   3rd Qu.:3.670   3rd Qu.:3.000  
 Max.   :1.0000   Max.   :800.0   Max.   :4.000   Max.   :4.000  
> 
```

Additionally, both GRE and GPA histograms are both positive and skewed to the right, which makes sense as college applicants at more selective institutions need to be more competitive to increase their chances.

[1] https://www.usnews.com/education/best-colleges/the-short-list-college/articles/colleges-with-the-lowest-acceptance-rates

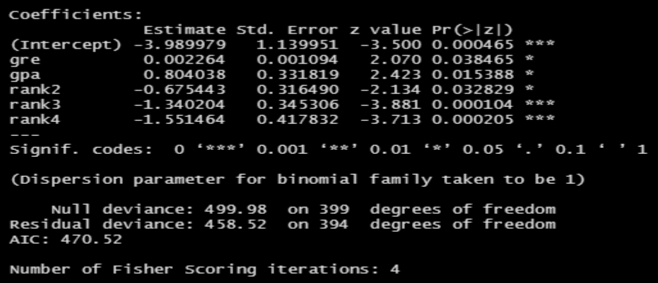
This is confirmed looking at the summary of the data with means that suggest a competitive application cycle. I should note here that rank is harder to pin down in terms of initial analysis.

The Model:
When I began this project, I knew it would be an attempt to predict probabilities with a binary/binomial dependent variable. As such, I started the model building process by crafting a general linear model in the binomial family. Since there are so few independent variables, I included all of them initially to see how well it performed.



```
model <- glm(admit ~ gre + gpa + rank, data = admit, family = "binomial")
```

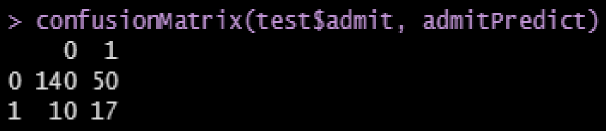
After building and running the model in R-Studio using the R programming language, we are provided with our summary output. All the resulting P-Values look good and pass the test using a 5% alpha. AIC can be ignored here because it is relative to model2. GRE being 0.002264 we can say that for every positive unit change in GRE, the log odds of admittance increase by 0.002264 OR .2%. Likewise, looking at GPA with an output of 0.804038 we can say with this model that for every unit change in GPA, the log odds of admittance changes by 80.4%



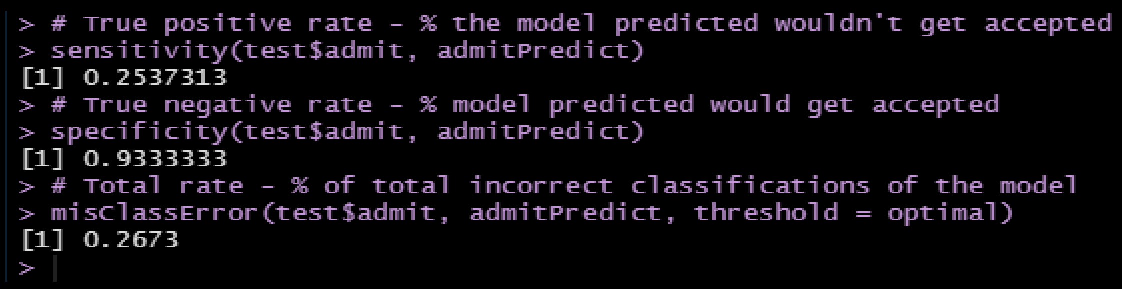
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -3.989979   1.139951  -3.500 0.000465 ***
gre          0.002264   0.001094   2.070 0.038465 *  
gpa          0.804038   0.331819   2.423 0.015388 *  
rank2       -0.675443   0.316490  -2.134 0.032829 *  
rank3       -1.340204   0.345306  -3.881 0.000104 ***
rank4       -1.551464   0.417832  -3.713 0.000205 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 458.52  on 394  degrees of freedom
AIC: 470.52

Number of Fisher Scoring iterations: 4
```

Since GRE in the first model has a P-Value of 0.038 it can be accepted in a standard 5% alpha, however I created a second model without this value just to determine the effects on removing. While there is a positive shift in the P-Values, and a change from 470 to 472.88 AIC value I determine here that it is not a significant enough change in output to keep GRE out of the model.

To further evaluate the model, I created a confusion matrix as a test of the accuracy, validity and robustness of the model. To do this, I created a training and testing set and utilized a 50/50 probability test. This is a way to build a probability test model with the optimal cutoff to maximize accuracy. Here we can see the true positives at 140 which says our model predicted the student was admitted and it was true. The false positives are 50 which means our model predicted these 50 students would be admitted, however they were not admitted. Even though we do have nearly three times higher prediction accuracy, I believe we can do better in the future. This output of the model is further enhanced by testing for sensitivity, specificity and the total rate of misclassification errors as shown below with a positive accuracy of the model at 73.27%.



```
> confusionMatrix(test$admit, admitPredict)
    0  1
0 140 50
1  10 17
```



```
> # True positive rate - % the model predicted wouldn't get accepted
> sensitivity(test$admit, admitPredict)
[1] 0.2537313
> # True negative rate - % model predicted would get accepted
> specificity(test$admit, admitPredict)
[1] 0.9333333
> # Total rate - % of total incorrect classifications of the model
> misClassError(test$admit, admitPredict, threshold = optimal)
[1] 0.2673
> 
```

Conclusion:
Given the output of the model, we can say that GRE, GPA and school ranking are statistically significant factors on admission to Random University, however the training and testing data for the model, while valid, should be improved in some way as to be more helpful in predicting admission.

One of the main causes, I suspect, is the limited amount of data in this dataset. Having only 400 individual applicants is not nearly enough to make the solid, verifiable model and argument for the weight of the independent variables. Further, testing against only a single university does not give us the most useful or practical scope of the overall issue plaguing message boards and social media during application season. I would posit that we might be better served looking at all universities vs applicant in X University's rank tier based on US News' ratings and ranking assessment performed annually.

Additionally, we may want to look at the motivations of the student that would have applied to Random University and if it is regionality, cost, student interests via their interests (i.e. intended college major) vs specialty of the university. This would give us additional insight into why students are attracted to apply and provide a better understanding by Random University on how and where monies are allocated in the future (sports vs sciences vs advertising).