STA302 Methods of Data Analysis I

Final Report

Peiyu Li 1004975627

Introduction

Admission rate for colleges and universities are various across United States. This can be due to many reasons such as cost per academic year, size of faculty members, or number of branch campuses. It is important for both institutions and students to understand which factors can affect admission rate, and those stakeholders would also want to predict admission rate using significant variables. It helps students to better consider which of the institutions to apply, and also helps institutions to balance the size of faulty and number of students.

This project investigates which of the factors shows the most significant relationship with admission rate that can explain the variation better, and find a multiple linear regression model which can be used to predict. The purpose of this model is making various stakeholders understand this relationship, and making good predictions using different data. Therefore, the final model should be simple enough for people to understand, but it should also be complicated enough to make good prediction.

Methods

1. Variable Selection

The dataset was divided equally into a training dataset and a validation dataset. In order to select some variables to be in the model, I first fit all variables in a linear model using training dataset. By checking VIF values, variables with VIF greater than 10 that obviously could cause collinearity should not be in the model. There were still several predictors with VIF larger than 5 which I thought they could also have negative impact and should not be in the model. After this step, I got 17 possible variables.

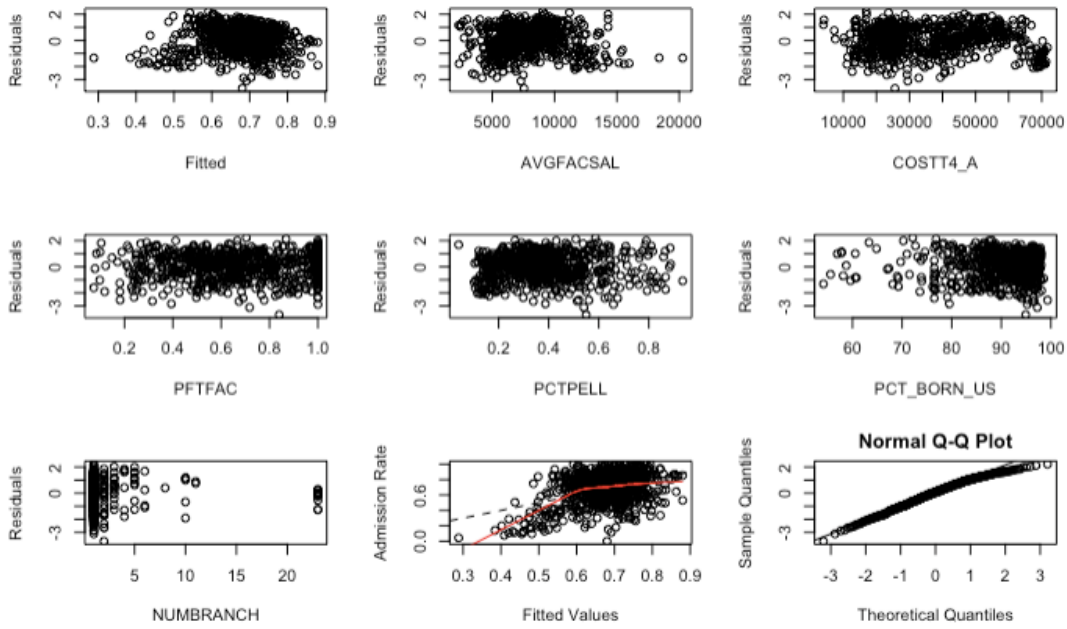
I checked all combinations of these variables and got two candidate models after comparing their adjusted R-squared, AIC and BIC. Then used R to do stepwise selection from those 17 predictors with both AIC and BIC, it gave me the same two model as above. Compared the summary table of these two models, one contained 6 variables while another one contained 11 varibles. Both of them showed similar adjusted R-squared and other criterions values and they are both significant. This could show the longer model might be over-fitting, and left with six variables.

2. Model Validation

To decide if the model I chose is valid, I need to fit the testing dataset into this model. The final model I got is a model transformed both predictors and response. After fitted testing data, predictors that were significant before were still showed significance and the p-value is the same as with training dataset. Also, if use testing data to predict, the mean squared error between the fitted values and the prediction was 0.063 which is small. These evidences confirmed that the final model with six predictors was valid.

3. Model Violations/Diagnostics

After selected six predictors, the residual plots shown below shows some clusters, the Normal QQ plot was not very linear, and the red curve shows that this model was not linear. Then transformations were necessary.



Using powertransform mothod, response and five predictors needed to do transformation. Since the sample size was large, there were 50 leverage points and 77 outliers in this model. Although there were none of them exceed the curoff for Cooks' Distance, but 54 observations are influential to their own fitted values, and over 50 observations would affect each of the predictor coefficients. However, since I did not have enough information about the data, it was hard to tell if some observations are problematic and did not know how to deal with those leverage points.

Results

1. Description of Data

The data contained 1508 observations, half of them are in the traning dataset. The below summary table gave an overview of each variable and divided the variables into numerical and categorical variables. Most of numerical variables measured rates or percentage so their mean and standard deviation were small. I also calculated the percentage of each category for each categorical variable. For variables contained value only 0 and 1, it is obvious that most observations fall into the category 0.

Numerical Variables

	NUMBRANCH	ADM_RATE	COSTT4_A	AVGFACSAL	PFTFAC
Mean	1.63E+00	6.76E-01	3.58E+04	7.90E+03	6.80E-01
SD	2.95E+00	2.02E-01	1.56E+04	2.42E+03	2.40E-01
	PCTPELL	**UG25ABV**	**INC_PCT_LO**	**PAR_ED_PCT_1STGEN**	**FEMALE**
Mean	3.78E-01	1.65E-01	3.57E-01	3.16E-01	5.82E-01
SD	1.63E-01	1.50E-01	1.51E-01	1.02E-01	1.19E-01
	MD_FAMINC	**PCT_WHITE**	**PCT_BLACK**	**PCT_ASIAN**	**PCT_HISPANIC**
Mean	5.10E+04	7.99E+01	1.10E+01	2.75E+00	1.05E+01
SD	2.23E+04	1.25E+01	1.07E+01	2.79E+00	1.78E+01
	PCT_BA	**PCT_GRAD_PROF**	**PCT_BORN_US**	**POVERTY_RATE**	**UNEMP_RATE**
Mean	1.60E+01	9.09E+00	9.04E+01	9.27E+00	3.51E+00
SD	3.51E+00	2.82E+00	7.37E+00	7.33E+00	1.09E+00

	Categorical Variables									
CONTROL	Category	1	2	3						
	Percentage	37.53%	58.89%	3.58%						
REGION	Category	1	2	3	4	5	6	7	8	9
	Percentage	9.68%	19.10%	15.65%	11.41%	24.54%	6.23%	2.65%	7.69%	3.05%
	Category	0	1							
HBCU	Percentage	95.89%	4.11%							
PBI		98.94%	1.06%							
TRIBAL		99.87%	0.13%							
HSI		89.92%	10.08%							
WOMENONLY		99.60%	0.40%							

I also made a boxplot to better visualize numerical variables that measured rates and percentage in Appendix(1).

2. Process of Obtaining Final Model

I checked the VIF for the full model that contained all variables and selected variables with VIF values less than 10. The checked the VIF for this smaller model again, and chose varibles with VIF values less than 5 which gave me 17 predictors. I got two

candidate model with number of predictors 6 and 11 after comparing all combination of those 17 variables and chose the one with highest adjusted R-squared and smalles AIC and BIC values. Then I used R to do stepwise selection using both AIC and BIC, since it gave a combination of both forward and backward selecion methods. The stepwise methods using AIC gave the same 11-variable model, and stepwise using BIC gave the same 6-variable model. The summary of these two models are shown below. Both models had very similar residual standard error, adjusted R-squared and p-value. Since our goal is to get a simple model that could be better explained, the 6-variable model would be better, and the other one seemed over-fitting.

```
Call:
lm(formula = ADM_RATE ~ AVGFACSAL + COSTT4_A + PFTFAC + PCTPELL +
    PCT_BORN_US + NUMBRANCH + HSI + PAR_ED_PCT_1STGEN + PCT_BA +
    UGZSABV + TRIBAL, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64989 -0.12032  0.01505  0.13080  0.41687

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.857e-01  1.712e-01   2.253 0.024529 *
AVGFACSAL        -2.308e-05  4.025e-06  -5.733 1.43e-08 ***
COSTT4_A         -2.784e-06  5.775e-07  -4.820 1.74e-06 ***
PFTFAC           -1.054e-01  3.038e-02  -3.470 0.000551 ***
PCTPELL          -2.381e-01  6.666e-02  -3.571 0.000378 ***
PCT_BORN_US       5.815e-03  1.288e-03   4.514 7.41e-06 ***
NUMBRANCH         7.088e-03  2.311e-03   3.068 0.002237 **
HSI               5.741e-02  2.848e-02   2.016 0.044134 *
PAR_ED_PCT_1STGEN 3.085e-01  1.174e-01   2.629 0.008747 **
PCT_BA            6.797e-03  3.046e-03   2.232 0.025938 *
UGZSABV          -9.501e-02  6.128e-02  -1.551 0.121440
TRIBAL            2.784e-01  1.849e-01   1.505 0.132630
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1841 on 742 degrees of freedom
Multiple R-squared:  0.1827,    Adjusted R-squared:  0.1705
F-statistic: 15.08 on 11 and 742 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = ADM_RATE ~ AVGFACSAL + COSTT4_A + PFTFAC + PCTPELL +
    PCT_BORN_US + NUMBRANCH, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68034 -0.12318  0.01559  0.13950  0.40853

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.525e-01  1.337e-01   5.630 2.55e-08 ***
AVGFACSAL   -2.076e-05  3.846e-06  -5.398 9.06e-08 ***
COSTT4_A    -2.950e-06  5.202e-07  -5.671 2.03e-08 ***
PFTFAC      -1.104e-01  2.902e-02  -3.803 0.000155 ***
PCTPELL     -2.021e-01  6.105e-02  -3.310 0.000978 ***
PCT_BORN_US  3.694e-03  1.099e-03   3.361 0.000815 ***
NUMBRANCH    6.616e-03  2.322e-03   2.850 0.004495 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1854 on 747 degrees of freedom
Multiple R-squared:  0.1656,    Adjusted R-squared:  0.1589
F-statistic:  24.7 on 6 and 747 DF,  p-value: < 2.2e-16
```

To obtain the final model, transformation was needed. All variables and response except PCTFAC were suggested to do power transformations by R. The transformation table is shown below. Finally I got the final model: Admission Rate^2 =1.04 – 2.5e-03*sqrt(AVGFACSAL) – 9.4e-04*sqrt(COSTT4_A) – 0.01*PFTFAC – 0.01*sqrt(PCTPELL) + 1.9e-19* PCT_BORN_US^9 – 0.09*NUMBRACH^-6.

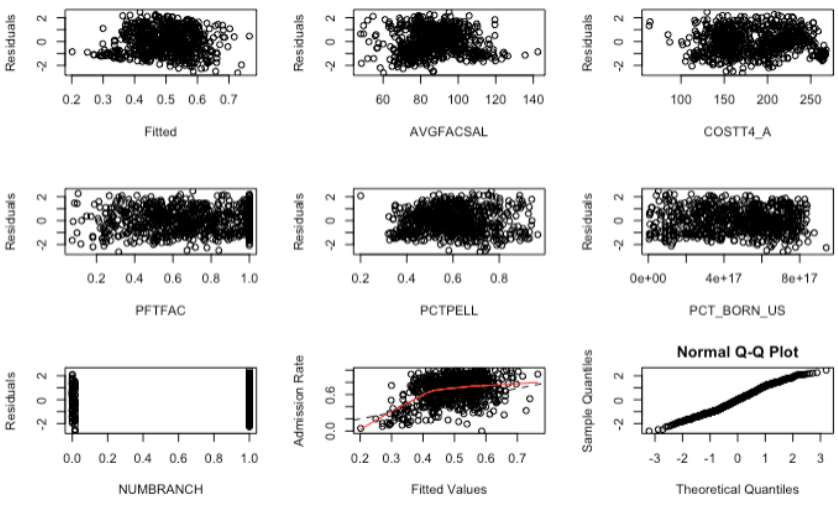
```
'data.frame':   754 obs. of  7 variables:
 $ NUMBRANCH  : int  1 2 1 1 1 1 1 1 1 1 ...
 $ ADM_RATE   : num  1.143 1.292 1.144 1.387 0.749 ...
 $ COSTT4_A   : int  39383 35965 46635 33685 69860 20780 39367 27109 49502 32800 ...
 $ AVGFACSAL  : int  6928 6548 9390 5585 11806 6315 6709 8601 7673 4192 ...
 $ PFTFAC     : num  0.718 0.544 1 0.248 0.945 ...
 $ PCTPELL    : num  0.395 0.461 0.309 0.534 0.117 ...
 $ PCT_BORN_US: num  94.7 95.1 93.7 87.9 86.4 ...
bcPower Transformations to Multinormality
            Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
ADM_RATE       2.2630        2.00       1.8962       2.6297
AVGFACSAL      0.5391        0.50       0.3911       0.6870
COSTT4_A       0.4365        0.50       0.2895       0.5834
PFTFAC         1.0535        1.00       0.8837       1.2233
PCTPELL        0.3888        0.50       0.2727       0.5049
PCT_BORN_US    8.9137        8.91       7.9430       9.8845
NUMBRANCH     -5.7264       -5.73      -6.1648      -5.2879

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)

Likelihood ratio test that no transformations are needed
```

3. Goodness of Final Model

I already showed the residual plots, observed vs. fitted values plot and normal QQ plot above in Methods Section, and it showed non-linearity. After transformation, constant variance was fixed, normality was better, and linearity was better but still not very linear as shown below. Since I need to find a multiple linear model for this data, it was the best I can got. The points were randomly scattered in the residual plots so constant variace was satisfied. Points are also closely scattered around the normal qq line which indicated normality was also satisfied. Linearity was better than before.



The pairwise plot for predictors shown in Appendix(2) also looks linear. I used testing dataset to fit this final model, and it gave me similar results for the significance of each variables and the model, and even smaller residual standard error and slightly larger adjusted R-squared. The prediction error usinf testing dataset was 0.063, and it further proved the final model I got was valid.

After doing model diagnostics, I found 50 leverage points and 77 outliers that had the potential to affect the model, though none of them exceed the Cooks' Distance cutoff. Specifically, 54 observations might affect their own fitted values and ober 50 observations were influential to each of the predictor coefficients. However, I had no further information to tell me if some obeservations are problematic, so I could not remove anyone of them.

Discussion

1. Final Model Interpretation and Importance

The final model I got is ADM_RATE^2 =1.04 – 2.5e-03*sqrt(AVGFACSAL) – 9.4e-04*sqrt(COSTT4_A) – 0.01*PFTFAC – 0.01*sqrt(PCTPELL) + 1.9e-19* PCT_BORN_US^9 – 0.09*NUMBRACH^-6. For interpretation, one of the examples is to say for 1 percent increase in proportion of full-time faculty member(PCTFAC), there should be on average a 0.01 decrease in the sqaured value of admission rate, which sqaure root of average faculty salary(AVGFACSAL), square root of average cost of attendance per academic year(COSTT4_A), square root of percentage of undergraduates receiving Pell grant(PCTPELL), percent of population from students that was born in the US (PCT_BORN_US) to the power of 9, and the value of number of branch campuses(NUMBRANCH) to the power of -6 are still unchanged.

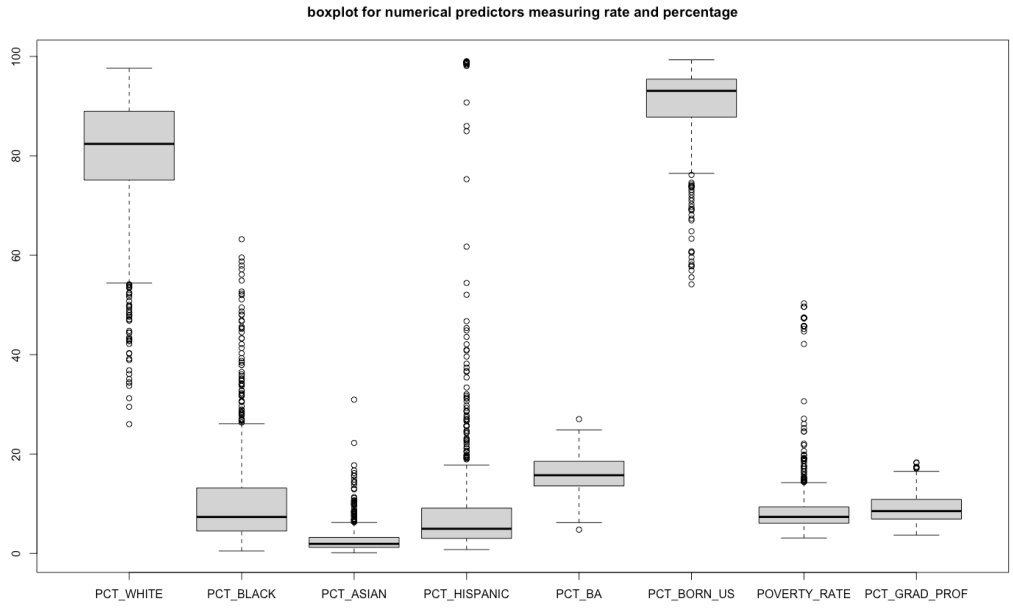
This interpretation might seem complicated but actually is not, since the basic idea is to fix other 5 variables while changing a value for the other one, and will get the average changing in response. The power trasformations are only to satisfy the assumptions and fix non-linearity, so it only affects calculation. The goal for this project is to get a good model easy to understand and complicated to predict. The final model I got satisfiy this purpose. It is simple so that just plug in the predictors value then can get the average estimated values for admission rate, and also can make good prediction. The stakeholders can understand how admission rate is affected.

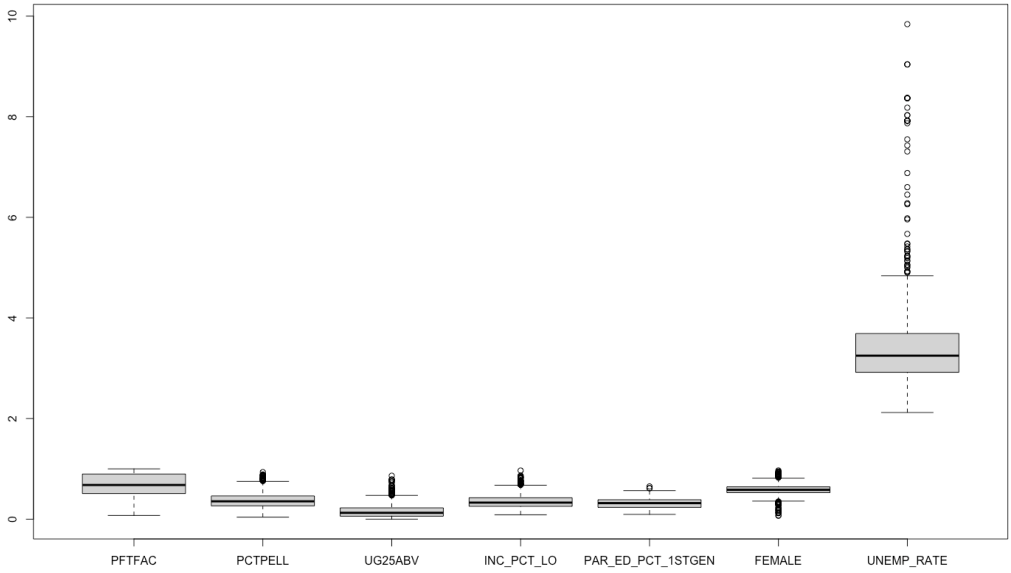
2. Limitations of Analysis

One of the limitations for the model is the data was not linear even after transformation. It might be better if use other models. Since the goal for this project was to get a multiple linear model, linear model did not fit very well and adjusted R-squared was small. The linear model could not explain much variation of the data. This problem could lead to not really good prediction, and the complicated transformations make it hard to interprete. There were also many influential points for the final model that I did not deal with, they had potential to influence the model and the coeffitients.

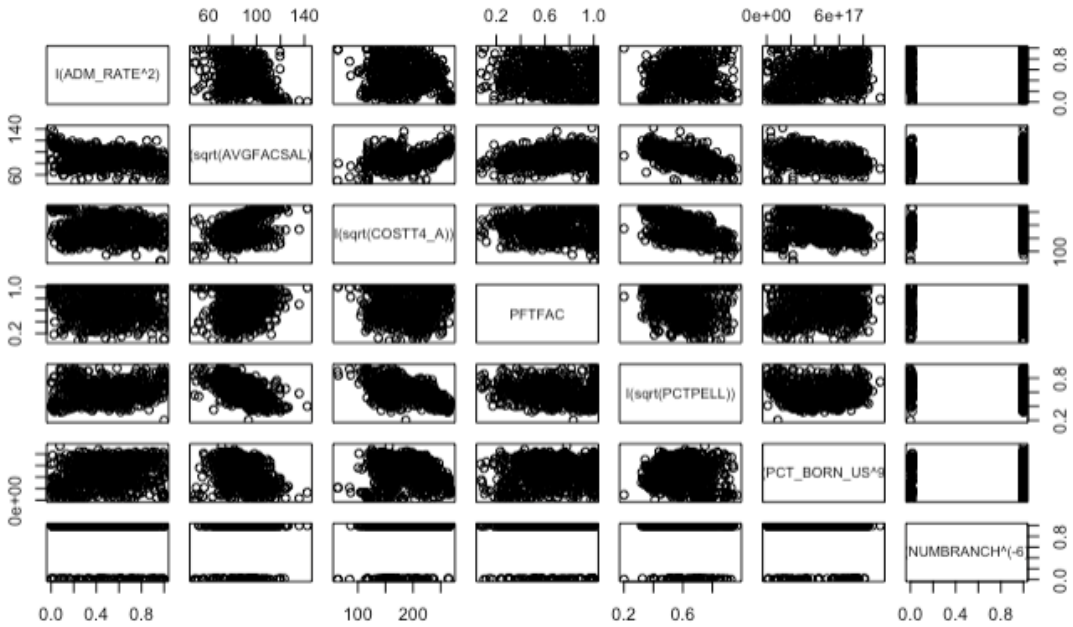
APPENDIX

1.




0
2
4
6
8
10
PFTFAC
PCTPELL
UG25ABV
INC_PCT_LO
PAR_ED_PCT_1STGEN
FEMALE
UNEMP_RATE

2.


I(ADM_RATE^2)
I(sqrt(AVGFACSAL))
I(sqrt(COSTT4_A))
PFTFAC
I(sqrt(PCTPELL))
PCT_BORN_US^9
NUMBRANCH^(-6