**Final project**

Investigate factors for the different admission rate of each university and college

Yiyun Sun

**Introduction:**

The admission rates of colleges and universities in the United States is different is always an important topic, and it may be influenced by several different factors. The aim of this project is to figure out which of the variables has huge impact on the admission rate from 1508 observations of each university and college. The model constructed in this project needs to describe how the factors affect admission rate to everyone and predict new acceptance rates as well. Thus, this model was explained clearly and simply in my research.

**Method:**

1. Variables

There are 30 variables in the original dataset. For the predictor variables, there are 9 factor variables, 18 numeric variables, and 2 other variables. The variables collected here are divided into three categories: school identifiers, school features, and applications features. I cannot deny that there exists collinearity between these variables.

1.1 Correlation

To decide which variable is useful for the purpose, I calculated the correlation among each numeric variable and ADM\_RATE (response variable) firstly. Then 6 numeric variables having higher correlation with ADM\_RATE were selected. Next I deleted the variable which has no impact on ADM\_RATE obviously, (Unit ID for institution, Name of institution). Then I deleted several variables which almost no change among 1508 observations. Thus, the research in this project is mainly based on remain 10 variables, where 1 response variable and 9 predictor variables. I constructed a new data frame for these 10 variables, which was named “fp”.

2. Model validation

I divided the original data into 2 set – training set and validation set firstly. Then I checked the model whether assumptions of it were satisfied. Then I checked the multicollinearity in the predictor variables. Next, the overall F test were constructed to see if each variable was significant on the response variable. Then I used stepwise selection which is a method combines both forward and backwards methods, to construct a suitable regression model on the training dataset. BIC was used in this stepwise method to find the best model because BIC is more precise than AIC. From this process, I got a temporary model. Next, I set up a transformation on this mode. Finally, I validated my new transformed model in the test dataset.

3. Diagnostics

To diagnose this model, I firstly check if there exist leverage points, outliers and influential points in this model. Then the assumption needs to be checked. The Normal QQ plot constructed to verify whether normality of errors was satisfied. If there is a one-to-one relationship in the QQ plot, the normality holds. Next, scatterplots between each variable and residual plot for each variable were constructed to check if condition 1 and condition 2 for residual plots holds. Only if both conditions hold, residual plot can be used in multiple linear regression model. Then, the residual plots were constructed to check other assumptions. If there is no clear and discernible pattern in the plots, assumptions hold.

**Results:**

1. Data characteristic

The data used in this project is “fp” – a new data frame from the original data. This new data contains 10 selected variables and 1508 observations for each variable. HBCU, PBI, HSI are categorical variables, others are numeric variables. The top left part of the Table 1 demonstrates the summary of the data.

The top right part of the Table 1 presents the correlation between each numeric variable in the original data and ADM\_RATE variable. It is clearly to see the correlation are all not strong, so I only chose several variables with higher correlation. The variables can be found in the top left part of Table 1.

The lower left part is the scatterplot between each variable, it shows that the linear relationship between variables is really not strong.

2. Stepwise selection

The Table 2 presents the process of the stepwise selection. Since the model with smaller value of BIC is better, the last model with lowest BIC value, -5103.96, is the best. The model in this step I got is ADM\_RATE = -2.162e-05 AVGFACSAL-3.199e-06 COSTT4\_A + 3.366e-03 PCT\_BORN\_US - 1.108e-01 HBCU + 5.451e-03 PCT\_BA + 5.737e-01. This is a 5 predictor variables multiple linear regression model.

3. Analysis of the model

The top part of the Table 3 is the overall F-test of the model I got before. The test is significant and each predictor is also significant, since p-value are all small enough. But the lower part of the Table 3 demonstrates the residual plot of each variable. The residual plot of PCT\_BORN\_US variable shows a trend so the condition 1 of this model may be violated, which means the residual plot cannot used to check assumption. Thus, this model is not very suitable.

4. Transformation

Because of the former model has some defect, I did a transformation on the response variable using Box-Cox method firstly. The top left part of Table 4 presents in the Box-Cox method is approximately 1.5. The lower left part of the Table 4 is the residual plots and Normal QQ plot of this Y- transformed model.

The top right part of Table 4 demonstrates a power transformation on both response variable and predictor variables. The lower right part of Table 4 is the residual plots and Normal QQ plot of this X&Y- transformed model.

Compared with the plots, it is easy to see the residual plot of PCT\_BORN\_US and PCT\_BA variables perform worse than before, which means X&Y- transformed model is not a wise move. But look at the plot of Y- transformed model, it performed better than the original model although not perfect. Overall, Y- transformed model is the best choice.

The model is ADM\_RATE^(1.5) = -2.208e-05 AVGFACSAL -3.693e-06 COSTT4\_A + 3.996e-03 PCT\_BORN\_US -1.294e-01 HBCU + 6.088e-03 PCT\_BA + 4.640e-01, demonstrated in the top left part of the Table 5.

5. Diagnostics

The lower left of the Table 4 presents the plots of the model. There is almost no exact pattern in residual plot and there exists a one-to-one relationship in the Normal QQ plot. This means the assumptions are satisfied in this model. The top right part of the Table 5 presents the leverage points of this model. The lower left part of the Table 5 presents the outlier points of this model and lower right part of the Table 5 presents the influential points. Thus, the model ADM\_RATE^ (1.5) = -2.208e-05 AVGFACSAL -3.693e-06 COSTT4\_A + 3.996e-03 PCT\_BORN\_US -1.294e-01 HBCU + 6.088e-03 PCT\_BA + 4.640e-01 is a proper model.

**Discussion:**

1. Final model

The final model is ADM\_RATE^ (1.5) = -2.208e-05 AVGFACSAL -3.693e-06 COSTT4\_A + 3.996e-03 PCT\_BORN\_US -1.294e-01 HBCU + 6.088e-03 PCT\_BA + 4.640e-01. This means that for every one-unit increase in the average faculty salary, admission rate decreases by about 0.69e-05 when other independent variables are fixed at a constant value. The final model satisfies the assumption and is useful for prediction to some extent.

2. Limitation

Although this model is the best model of the data, there exists some random errors. For example, the residual plot is not that perfect which may cause non-constant variance. Another problem is that there are one categorical predictor variables in this model, which is difficult to interpret. This may cause misunderstanding in the real practice. Last problem is that the coefficients of this model are too small, which lead to little change in response variable. It may reduce practicality of this model.

**Tables and figures:**

|  |  |
| --- | --- |
| ## HBCU PBI HSI ADM\_RATE ## Min. :0.00000 Min. :0.00000 Min. :0.0000 Min. :0.0000 ## 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:0.5639 ## Median :0.00000 Median :0.00000 Median :0.0000 Median :0.6922 ## Mean :0.03912 Mean :0.01592 Mean :0.1107 Mean :0.6708 ## 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:0.0000 3rd Qu.:0.8166 ## Max. :1.00000 Max. :1.00000 Max. :1.0000 Max. :1.0000 ## COSTT4\_A AVGFACSAL PAR\_ED\_PCT\_1STGEN PCT\_BA ## Min. : 3990 Min. : 1236 Min. :0.08867 Min. : 4.79 ## 1st Qu.:23007 1st Qu.: 6360 1st Qu.:0.24486 1st Qu.:13.61 ## Median :34620 Median : 7612 Median :0.32243 Median :15.88 ## Mean :36482 Mean : 7977 Mean :0.31856 Mean :16.10 ## 3rd Qu.:47730 3rd Qu.: 9216 3rd Qu.:0.38782 3rd Qu.:18.62 ## Max. :75735 Max. :20484 Max. :0.66667 Max. :27.03 ## PCT\_GRAD\_PROF PCT\_BORN\_US ## Min. : 2.700 Min. :41.34 ## 1st Qu.: 7.008 1st Qu.:87.39 ## Median : 8.670 Median :92.55 ## Mean : 9.133 Mean :89.97 ## 3rd Qu.:10.745 3rd Qu.:95.48 ## Max. :18.500 Max. :99.32 | ADM\_RATE ## ADM\_RATE 1.00000000 ## COSTT4\_A -0.28735526 ## AVGFACSAL -0.33136253 ## PFTFAC -0.09307594 ## PCTPELL 0.10647585 ## UG25ABV 0.16313361 ## INC\_PCT\_LO 0.03770508 ## PAR\_ED\_PCT\_1STGEN 0.18471323 ## FEMALE 0.09575300 ## MD\_FAMINC -0.08160502 ## PCT\_WHITE 0.15595637 ## PCT\_BLACK -0.08766346 ## PCT\_ASIAN -0.17157648 ## PCT\_HISPANIC -0.09782331 ## PCT\_BA -0.20538689 ## PCT\_GRAD\_PROF -0.28086544 ## PCT\_BORN\_US 0.21457972 ## POVERTY\_RATE -0.03838170 ## UNEMP\_RATE -0.03952999 |
|  |  |

**Table 1: Summary of the new data “fp” and correlation between each variable and response variable.**

|  |
| --- |
| ## Start: AIC=-4850.38## ADM\_RATE ~ 1## ## Df Sum of Sq RSS AIC## + AVGFACSAL 1 6.6073 53.568 -5018.5## + COSTT4\_A 1 4.9688 55.206 -4973.0## + PCT\_GRAD\_PROF 1 4.7469 55.428 -4967.0## + PCT\_BORN\_US 1 2.7707 57.404 -4914.1## + PCT\_BA 1 2.5384 57.637 -4908.1## + PAR\_ED\_PCT\_1STGEN 1 2.0531 58.122 -4895.4## <none> 60.175 -4850.4## + HBCU 1 0.1697 60.005 -4847.3## + PBI 1 0.0014 60.174 -4843.1## + HSI 1 0.0000 60.175 -4843.1## ## Step: AIC=-5018.45## ADM\_RATE ~ AVGFACSAL## ## Df Sum of Sq RSS AIC## + COSTT4\_A 1 2.0060 51.562 -5068.7## + PCT\_BORN\_US 1 0.8125 52.755 -5034.2## + PCT\_GRAD\_PROF 1 0.5833 52.984 -5027.6## + HBCU 1 0.4810 53.087 -5024.7## + PAR\_ED\_PCT\_1STGEN 1 0.2797 53.288 -5019.0## <none> 53.568 -5018.5## + PCT\_BA 1 0.0428 53.525 -5012.3## + PBI 1 0.0127 53.555 -5011.5## + HSI 1 0.0041 53.564 -5011.3## - AVGFACSAL 1 6.6073 60.175 -4850.4## ## Step: AIC=-5068.69## ADM\_RATE ~ AVGFACSAL + COSTT4\_A## ## Df Sum of Sq RSS AIC## + PCT\_BORN\_US 1 0.8577 50.704 -5086.7## + HBCU 1 0.7722 50.789 -5084.1## + PCT\_BA 1 0.2677 51.294 -5069.2## <none> 51.562 -5068.7## + HSI 1 0.0748 51.487 -5063.6## + PBI 1 0.0266 51.535 -5062.2## + PAR\_ED\_PCT\_1STGEN 1 0.0171 51.545 -5061.9## + PCT\_GRAD\_PROF 1 0.0001 51.562 -5061.4## - COSTT4\_A 1 2.0060 53.568 -5018.5## - AVGFACSAL 1 3.6444 55.206 -4973.0## ## Step: AIC=-5086.67## ADM\_RATE ~ AVGFACSAL + COSTT4\_A + PCT\_BORN\_US## ## Df Sum of Sq RSS AIC## + HBCU 1 0.76103 49.943 -5102.2## + PCT\_BA 1 0.39051 50.313 -5091.0## <none> 50.704 -5086.7## + HSI 1 0.09643 50.608 -5082.2## + PCT\_GRAD\_PROF 1 0.04792 50.656 -5080.8## + PAR\_ED\_PCT\_1STGEN 1 0.03449 50.669 -5080.4## + PBI 1 0.01058 50.693 -5079.7## - PCT\_BORN\_US 1 0.85772 51.562 -5068.7## - COSTT4\_A 1 2.05118 52.755 -5034.2## - AVGFACSAL 1 2.37310 53.077 -5025.0## ## Step: AIC=-5102.16## ADM\_RATE ~ AVGFACSAL + COSTT4\_A + PCT\_BORN\_US + HBCU## ## Df Sum of Sq RSS AIC## + PCT\_BA 1 0.30130 49.642 -5104.0## <none> 49.943 -5102.2## + HSI 1 0.03990 49.903 -5096.0## + PAR\_ED\_PCT\_1STGEN 1 0.03844 49.905 -5096.0## + PCT\_GRAD\_PROF 1 0.03745 49.906 -5096.0## + PBI 1 0.01841 49.925 -5095.4## - HBCU 1 0.76103 50.704 -5086.7## - PCT\_BORN\_US 1 0.84651 50.789 -5084.1## - COSTT4\_A 1 2.34242 52.285 -5040.4## - AVGFACSAL 1 2.52137 52.464 -5035.2## ## Step: AIC=-5103.96## ADM\_RATE ~ AVGFACSAL + COSTT4\_A + PCT\_BORN\_US + HBCU + PCT\_BA## ## Df Sum of Sq RSS AIC## <none> 49.642 -5104.0## + PAR\_ED\_PCT\_1STGEN 1 0.21004 49.432 -5103.0## - PCT\_BA 1 0.30130 49.943 -5102.2## + PCT\_GRAD\_PROF 1 0.08705 49.555 -5099.3## + HSI 1 0.06188 49.580 -5098.5## + PBI 1 0.01575 49.626 -5097.1## - HBCU 1 0.67182 50.313 -5091.0## - PCT\_BORN\_US 1 0.95290 50.595 -5082.6## - COSTT4\_A 1 2.59435 52.236 -5034.5## - AVGFACSAL 1 2.79656 52.438 -5028.6  ## ## Call:## lm(formula = ADM\_RATE ~ AVGFACSAL + COSTT4\_A + PCT\_BORN\_US + ## HBCU + PCT\_BA, data = fp)## ## Coefficients:## (Intercept) AVGFACSAL COSTT4\_A PCT\_BORN\_US HBCU PCT\_BA ## 5.737e-01 -2.162e-05 -3.199e-06 3.366e-03 -1.108e-01 5.451e-03 |

**Table 2: The process of the stepwise selection using BIC**

|  |
| --- |
| ## ## Call:## lm(formula = ADM\_RATE ~ AVGFACSAL + COSTT4\_A + PCT\_BORN\_US + ## HBCU + PCT\_BA, data = fp)## ## Residuals:## Min 1Q Median 3Q Max ## -0.70621 -0.11969 0.02156 0.13945 0.38600 ## ## Coefficients:## Estimate Std. Error t value Pr(>|t|) ## (Intercept) 5.737e-01 6.699e-02 8.563 < 2e-16 \*\*\*## AVGFACSAL -2.162e-05 2.350e-06 -9.199 < 2e-16 \*\*\*## COSTT4\_A -3.199e-06 3.610e-07 -8.860 < 2e-16 \*\*\*## PCT\_BORN\_US 3.366e-03 6.270e-04 5.370 9.14e-08 \*\*\*## HBCU -1.108e-01 2.458e-02 -4.509 7.03e-06 \*\*\*## PCT\_BA 5.451e-03 1.805e-03 3.019 0.00258 \*\* ## ---## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1## ## Residual standard error: 0.1818 on 1502 degrees of freedom## Multiple R-squared: 0.175, Adjusted R-squared: 0.1723 ## F-statistic: 63.74 on 5 and 1502 DF, p-value: < 2.2e-16 |
|  |

**Table 3: The process of the stepwise selection using BIC**

|  |  |  |
| --- | --- | --- |
|  | | ## bcPower Transformations to Multinormality  ## Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd ## Y1 1.5064 1.51 1.3762 1.6366 ## Y2 0.4555 0.50 0.3532 0.5578 ## Y3 0.4490 0.50 0.3407 0.5574 ## Y4 -5.5383 -5.54 -5.8178 -5.2588 ## Y5 0.9517 1.00 0.7854 1.1180 ## Y6 9.8023 9.80 9.1175 10.4871 ##  ## Likelihood ratio test that transformation parameters are  equal to 0 ## (all log transformations) ## LRT df pval ## LR test, lambda = (0 0 0 0 0 0) 9370.422 6 < 2.22e-16 ##  ## Likelihood ratio test that no transformations are needed ## LRT df pval ## LR test, lambda = (1 1 1 1 1 1) 16768.18 6 < 2.22e-16 |
|  |  | |

**Table 4: Comparison between two transformations**

|  |  |
| --- | --- |
| Call:  lm(formula = I(ADM\_RATE^(1.5)) ~ AVGFACSAL + COSTT4\_A + PCT\_BORN\_US +  HBCU + PCT\_BA, data = fp)  Residuals:  Min 1Q Median 3Q Max  -0.64181 -0.15941 0.00786 0.16627 0.48749  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 4.640e-01 8.008e-02 5.794 8.35e-09 \*\*\*  AVGFACSAL -2.208e-05 2.809e-06 -7.860 7.30e-15 \*\*\*  COSTT4\_A -3.693e-06 4.316e-07 -8.556 < 2e-16 \*\*\*  PCT\_BORN\_US 3.996e-03 7.495e-04 5.332 1.12e-07 \*\*\*  HBCU -1.294e-01 2.939e-02 -4.404 1.14e-05 \*\*\*  PCT\_BA 6.088e-03 2.158e-03 2.821 0.00485 \*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.2173 on 1502 degrees of freedom  Multiple R-squared: 0.152, Adjusted R-squared: 0.1492  F-statistic: 53.85 on 5 and 1502 DF, p-value: < 2.2e-16 |  |
|  |  |

**Table 5: Diagnostics of the final model**