STA 302 Method of Data Analysis

Final project

Prediction to Admission Rate

Jiayue Wu (1004768165)

Introduction section

The admission rated of universities and colleges in United States can be drastically different. This can be due to a variety of reasons, for example, difference in facilities, tuition, or even demographics of applicants themselves. This project aims to review admission rates for 1508 colleges and universities in the U.S. to investigate which of the factors/variables in the provided dataset can explain the variation best. The goal of this project is not only building a model complicated enough that can make good predictions but also simple enough for various stakeholders to understand.

Methods Section

1. Variable Selection:

I construct a full model without the variable that absolutely will not have relationship with admission rate at the very beginning. i.e. Institution's unit ID; Institution's name; Postcode; Number of branch campuses.
To select variables more precisely, I use VIF function to check the Multicollinearity of remaining variables. Model we build may have a number of problems if some variables are multicollinear, i.e. wrong sign of coefficients; non-significant predictors with F-test highly significant; standard errors of the regression coefficients, etc. Highly correlated variable will have the same high VIF. I choose 10 to be the cur-off. The number of factor variable do not mean a lot, I would not consider their VIF here. I discarded those whose VIF > 10.

Table of VIF (except for factor variables.)

Variable name	COSTT4_A	AVGFACSAL	PCTPELL	UG25ABV	INC_PCT_LO	PAR_ED_PCT_1ST GEN
VIF	6.740013	2.714045	4.510026	2.591449	15.024636	5.614295

Variable name	FEMALE	MD_FAMINC	PCT_WHITE	PCT_BLACK	PCT_ASIAN	PCT_HISPANIC
VIF	1.266362	9.558917	41.286542	32.229619	9.941444	18.518285

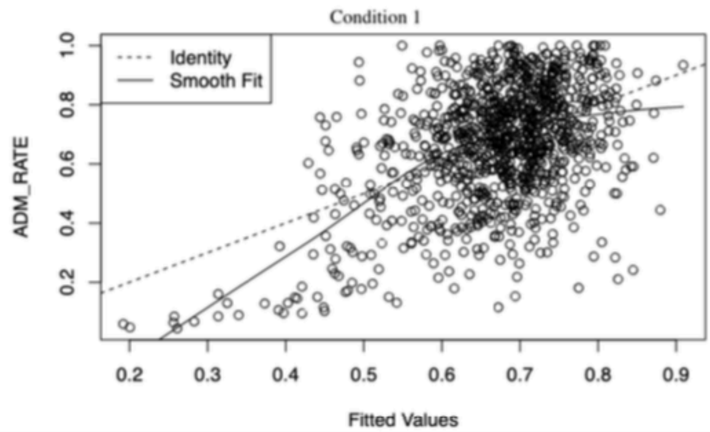
Variable name	PCT_BA	PCT_GRAD_PROF	PCT_BORN_US	POVERTY_RATE	UNEMP_RATE
VIF	9.061429	9.654855	7.031091	23.308562	12.709268

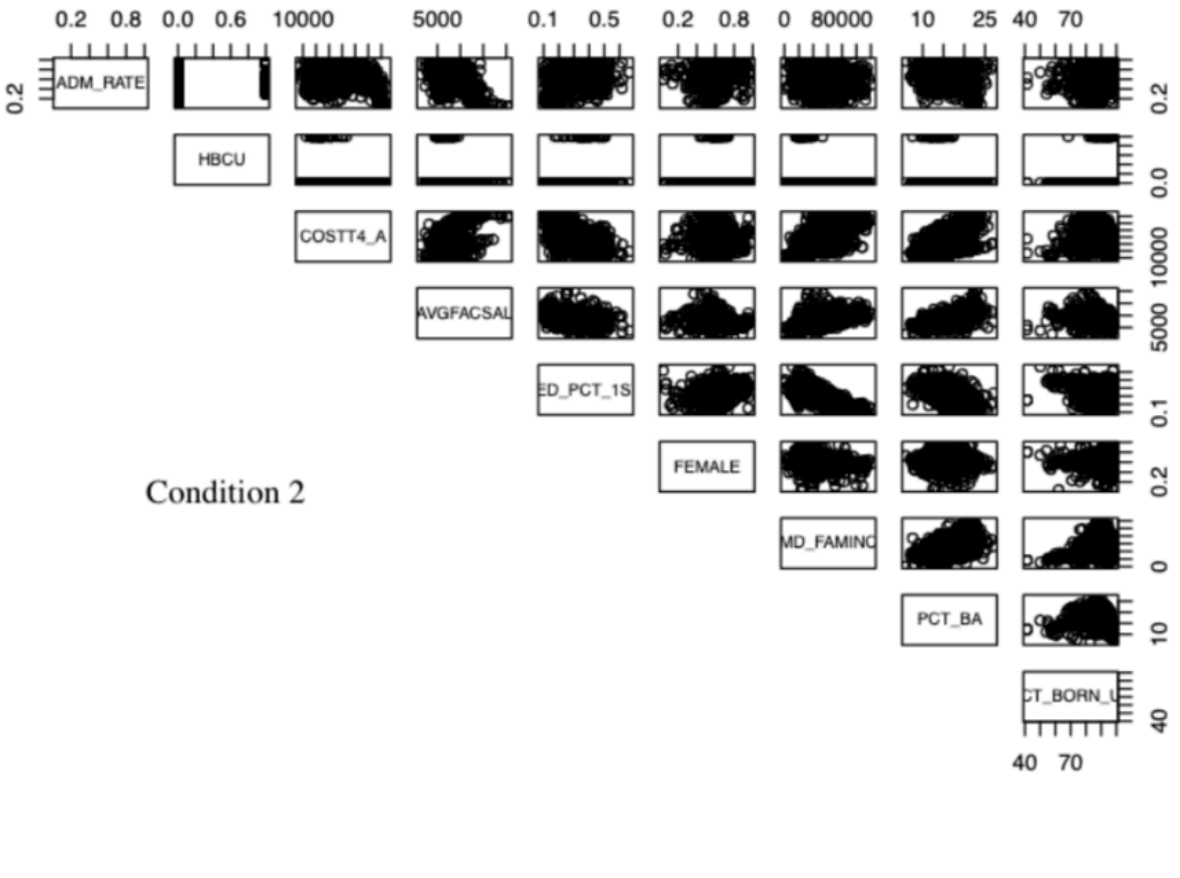
Then I use 3 ways of variable selection to select more significant variables: Forward selection, Backward selection, and Stepwise selection.

It happens that the Backward selection select same variables with Stepwise selection. I will just compare how the forward model and backward model behave.
I improve both of the model and compare their final behaviors. (more details about improving is on obtaining final model part). After comparing leverage points, Adjusted R^2, AIC, BIC, Adjusted AIC, I figured out the final model where I will choose variables from.

2. Model validation
At the very beginning, I randomly dividing the original data into two independent sets: training dataset(75%) and testing dataset(25%). All the model are built based on the training dataset.
The prediction error of our final model is 0.03694553, which is pretty low, it has a great validation.

3. Model violations/Diagonitics.
I check the residual's condition first. If satisfied, I can use the residual plots to tell us how we can fix our incorrect model. I draw a fitted value plot to check condition 1 and use pair function to check condition 2.

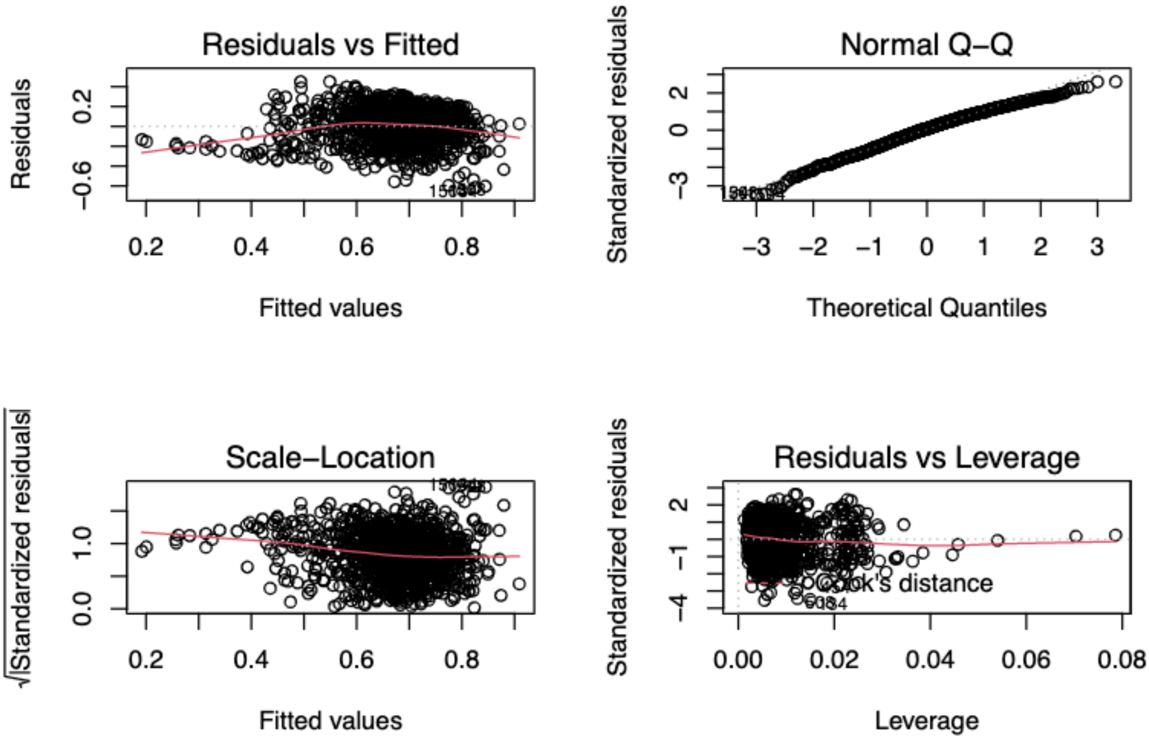




Condition 2

The points are randomly scattered around the function g. Condition 1 holds.
Except for factor variables, all others seems to do not have relation with each other. Condition 2 for holds.
I use **plot function** to test Diagonities:

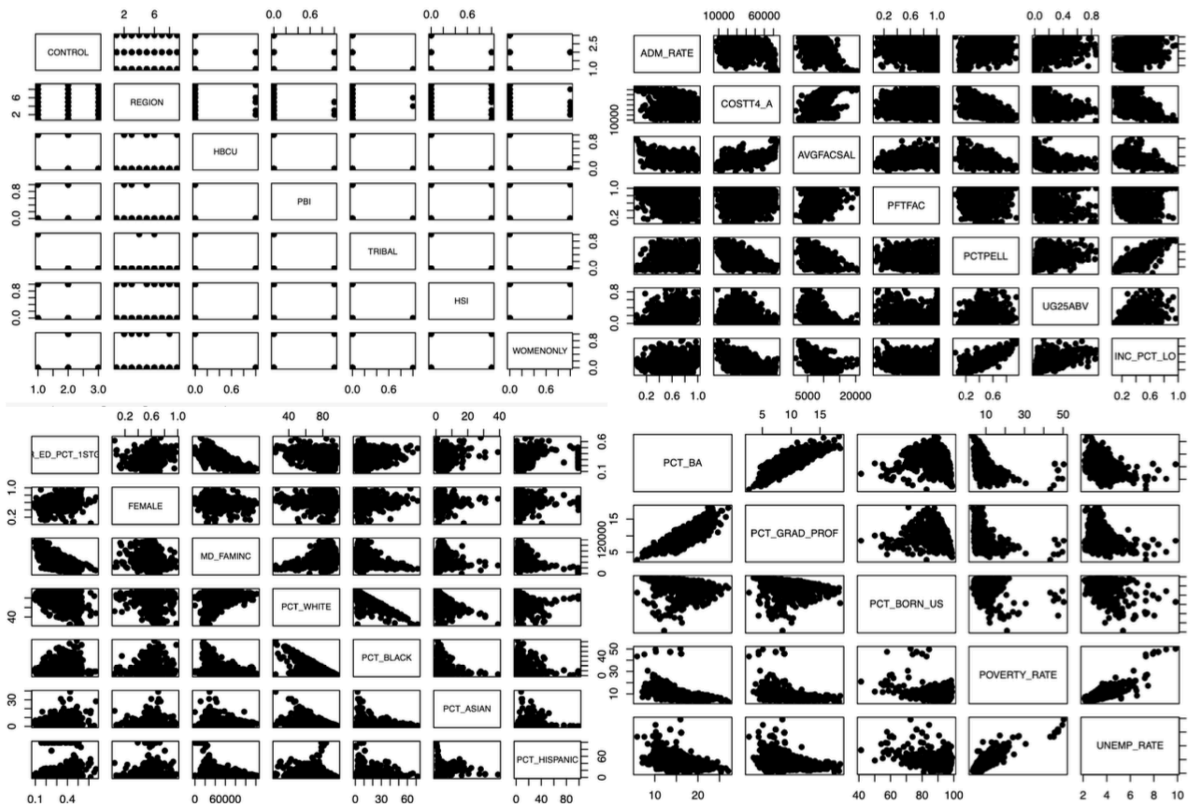
Diagonitics



(1) Residuals vs Fitted
Both residuals plots seem do not have relationship and is constant. I conclude that the linearity assumption holds.
(2) Normal QQ plot
Both of the points of two models on Normal QQplot lift off the line at the ends and wiggle around the line a lot, but it’s not crazy. I conclude that it follows the model assumption of Normality.

Results Section

1. Description of Data:
There are 30 variables in the data set, with 1 response (Admission rate) and other 29 variables.



For the first 4 few in the dataset, ‘UNITID’, ‘INSTNM’, ‘STABBR’, ‘NUMBRANCH’ are characteristic variables, which cannot be plot into scatterplot. For the next 7, i.e. ‘CONTROL’, ‘REGION’, ‘HBCU’, ‘PBI’, ‘TRIBAL’, ‘HSI’, ‘WOMENONLY’ are factor variables, the number of them means some specify indicators instead of numeric meaning. One thing to note that factor variable in the remaining variables should not be directly put in the model but with function as.factor(). ‘ADM_RATE’ is the response. ‘COSTT4_A’ is the average cost; ‘AVGFACSAL’ is the average salary; are all variables that larger than 0 and even 1. For the following 6 variables, i.e. ‘PFTFAC’; ‘PCTPELL’, ‘UG25ABV’, ‘INC_PCT_LO’, ‘PAR_ED_PCT_1STGEN’, ‘FEMALE’, ‘PCT_WHITE’; ‘PCT_BLACK’; ‘PCT_ASIAN’; ‘PCT_HISPANIC’ are all percentage numeric variable, where the last four indicates percentage of area students from. For the following 5 variables are all % variables, and it is clear that the variable ‘PCT_BA’ should have linear relationship with ‘PCT_GRAD_PROF’; ‘POVERTY_RATE’ should have linear relationship with ‘UNEMP_RATE’.

2. Process of Obtaining Final Model

After the VIF selection and out with original forward model and backward model, I improve both of the methods by calculating hii, figuring out outliers and refit each of the model without those bad leverage points. Because outliers’ response value has a disproportionate effect on the estimated regression line.

I also rebuild a reduced model for each of them by deleting the less significant variables in these two models. While ANOVA function suggests original models are better.

Then I do a transformation to both of selection models. So that I have more choices can compare their validation.

I use 3 main ways to compare these models:

Compare leverages (details in appendix):

(1) Compare Cook distance, which shows the influence of each observation on the fitted response values.

(2) Compare DFFITS, which measures how the predicted value at the i^{th} observation changes when the i^{th} observation is deleted.
(3) Compare DFBEATS, which indicates the effect that deleting each observation has on the estimates for the regression coefficients.

While they behave similarly here.

Compare four possible criteria.

(1) Compare Adjusted R^2: it will increase only if the new term improves the model more than would be expected by chance. We prefer Adjusted R^2 larger.
(2) Compare AIC: which balances the goodness of fit of the model with a penalty term reflecting how complex the model is. We prefer AIC smaller.
(3) Compare AIC_C: similar for AIC. We prefer AIC_c smaller.
(4) Compare BIC: similar with AIC. We prefer BIC smaller.

Compare prediction error

Aforementioned, I randomly dividing the original data into two independent sets: training dataset(75%) and testing dataset(25%). I use these two models to test how it behave on the other 25% testing dataset.

Table of data support to compare model

Model name	Forward model	Backward model	Forward transformed model	Backward transformed model
Adjusted R^2	0.2067791	0.2095082	0.1514041	0.131471
AIC	-712.6038	-717.2743	-181.653	-158.1422
AIC_c	-712.4022	-717.0736	-181.4514	-158.0403
BIC	-662.5731	-667.1983	-131.6231	-123.089
Prediction error	0.0362223	0.03694553	0.06305111	0.06119004

Lastly, I conclude that backward selection model behaves best. The transformed model has too high prediction error that I would not consider them anymore. Although backward model has slightly higher prediction error than forward model. But backward model has largest Adjusted R^2 and least AIC, AIC_C, BIC. I believe it would have a better prediction than others when facing population.

3. Goodness of Final Model

The prediction error of final model is pretty small(0.03694553), the final model has been validated correctly. The better behavior of AIC, BIC, AIC_C, Adjusted R^2 of final model than other candidates also suggests better prediction. Aforementioned, this model does not contain multicollinear part, satisfies the violation of residuals and assumption we can test.

Discussion Section

1. Final model interpretation and importance

The final model I figure out is

ADM_RATE = (2.406e-01) – (8.612e-02)*HBCU - (4.046e-06)*COSTT4_A - (2.171e-05)*AVGFACSAL + (4.226e-01)* PAR_ED_PCT_1STGEN + (1.836e-01)*FEMALE + (2.725e-06)*MD_FAMINC + (7.027e-03)*PCT_BA + (2.969e-03)*PCT_BORN_US

Hold all other variables constant,

(1) If all predictors in this model is 0, the admission rate of non-historically black is

(2.406e-01)+ (8.612e-02); while historically black one is 2.406e-01.
(2) 1 unit increase in average annual total cost of attendance will lower admission rate 4.046e-06.
(3) 1 unit increase in Average faculty salary will lower admission rate 2.171e-05.
(4) 1 unit increase in Percentage of first-generation students will higher admission rate 4.226e-01.
(5) 1 unit increase in proportion of female students will higher admission rate 1.836e-01.
(6) 1 unit increase in median family students will higher admission rate 2.725e-06.
(7) 1 unit increase in percentage students with bachelor's degree over age 25 will higher admission rate 7.027e-03.
(8) 1 unit increase in the percentage of students born in US will higher admission rate 2.969e-03.

By comparing with other model, this model has lowest AIC, BIC, AIC_C, and highest Adjusted R^2. The prediction error on testing data also very low. I conclude that this is the 'best' model I can figure out so far using my knowledge from STA302.
However, the admission rate seems do not have large relevant to region/demographic which I guessed they may have from introduction part.

2. Limitations of Analysis

Although this model is the 'best' model I can build so far, the R^2 of this model is not large enough that support it always provide good prediction. Since what I have learned so far is simple linear regression model and what I can build is limited, some of variables may be better to fit in a multiple regression model. The selection of variables may have bias. Moreover, although I have tried to decline the variable size to make the model simpler, there are still 8 variables in the model I build. When use in in the real life, it may be overfitting.

(1450 words)

Appendix

When comparing model, forward model behave similar to backward model on influential points part. More details will be showed here.

DFFITS of forward model

```
##  571  220 1470    4  454   34  638 1449  168  721    1  352  531 1426  376  808
##   36   47   54   65  104  110  130  151  159  183  192  204  217  228  246  248
## 1326 1438 1452  767 1216  508  106 1110 1501  110 1233  812  950  800 1238 1477
##  266  267  276  309  337  338  357  368  383  386  387  389  400  425  435  469
##  692 1469 1485 1439 1458  306  461 1375   62   39 1455 1453 1450   73  453  134
##  492  510  516  522  524  528  537  552  558  564  566  575  580  583  595  598
## 1037  809 1504  438 1489  877   72  145  776  233  109  181  992 1457   93  473
##  628  637  642  686  687  707  710  722  735  770  775  784  833  838  858  862
##  759 1242 1338 1447  189   76  299 1266  249  671  272  866 1492  664 1134
##  863  888  905  923  959  990  991  993 1011 1046 1051 1056 1062 1071 1080
```

DFFITS of backward model

```
##   22  571  220 1470    4  454   34  638 1449  168  721    1  352  531 1426  196
##   18   37   48   55   66  106  112  132  153  161  185  194  206  219  229  238
##  376  808 1326 1438 1452  104  767 1216  508  106 1110 1501  110 1233  812  950
##  248  250  269  270  279  290  312  340  341  360  371  386  389  390  392  403
## 1353   59 1323  692 1439 1458  306  461  245 1375   62 1455 1453   73  453  134
##  445  454  470  496  526  528  532  541  543  556  562  570  579  587  599  602
##  809 1504  438 1489  704  877   72  145  233  521  109  181  898  416  992 1457
##  641  646  690  691  695  711  714  726  774  776  779  788  826  837  839  844
##   93  473  759 1242 1338 1447  201   76 1266  831  249 1351  664 1134
##  864  868  869  894  911  928  993  995  998 1002 1016 1039 1076 1085
```

DFFITS model shows forward model has 1 more influential point.

DFBETAS of forward model

```
## 1167  571 1398 1470  496 1300 1215  454   34  638 1449 1172  696  721  748 1426
##   18   36   48   54   63   81   90  104  110  130  151  161  176  183  189  228
##  376 1295 1359 1438 1452  231 1216  508   84  106  525 1110 1233  950   59  459
##  246  259  264  267  276  311  337  338  356  357  363  368  387  400  450  495
## 1439  897 1375   62 1455 1453 1450   73  134  444  241 1037  738 1504   55  744
##  522  540  552  558  566  575  580  583  598  602  609  628  632  642  646  671
## 1489  749  634  877   72  145  233  109 1445  416  992 1457  758   93  314 1025
##  687  690  706  707  710  722  770  775  794  831  833  838  853  858  877  879
##  801  914 1447  201 1266  960  249 1020   65 1212
##  893  917  923  988  993 1007 1011 1021 1022 1029
```

DFBETAS of backward model

```
##  204  571  839 1470  821  454  784 1449  721 1257 1426  196 1430   74 1326 1438
##    7   37   50   55   78  106  136  153  185  189  229  238  246  256  269  270
## 1452  104 1441  767  508 1501  110 1233  812  950  800 1238 1477 1443 1435 1439
##  279  290  294  312  341  386  389  390  392  403  428  438  473  477  512  526
## 1458 1437 1455 1453 1450   73  762  810  809   55 1489  704   72  763  145  776
##  528  553  570  579  584  587  597  640  641  650  691  695  714  721  726  739
##  109  181 1425  992   93  867   87 1447  768 1280   70  769  132  299   91  831
##  779  788  789  839  864  895  900  928  937  939  945  983  984  996 1000 1002
##  755  866  664    2
## 1033 1061 1076 1080
```

DFBETAS shows forward has 6 more influential points.

In summary, forward model and backward model have similar behavior on influential points.