STA304final paper

2022/4/25

1 Abstract

It is common for applicants to review a school's basic information and characteristics before they decide to apply for the school. Applicants need to compete with others. Hence, they are willing to know the opportunity of getting an offer. In this paper,we are going to examine the admission rate of a large amount of universities/colleges in U.S and try to predict the admission rate for a given university.

Keywords:admission rate,universities,applicants

2 Introduction

Admission rate is the rate of being accepted. It is calculated by dividing the number of accepted students by total number of applicants. Admission rate varies from university to university and this may be the result of various reasons, including types of university, facilities and equipment of different campus, tuition fee and some other preferences in applicants. We need to decide factors that affects admission rate first, then describe and understand the relationship between admission rate and these factors. The aim of this paper is to fit a linear regression model which describes the relationship between admission rate and predictors most precisely in our population of 1,508 observations of universities and colleges in US. The goal of this paper is to help various stakeholders to understand the relationship by using the model. Also, we should be able to make predictions of admission rate if we are given a new university in future.

3 Data

This data set includes 1,508 observations of universities and colleges in United States. It is derived from a larger collection of measures on schools in the United States (https://collegescorecard.ed.gov/data/)[1].This is the official website of U.S department of education.The original data set contains all cumulative information from 1996 to 2020,and our initial selection of this data set is to take a subset with annually report of 2018 and 30 primary variables.

The data set has 30 variables, with one response variable (ADM_RATE) and 29 possible predictor variables. The variables are divided into 3 categories: school identifiers, school characteristics and applicant distribution(mainly on race).

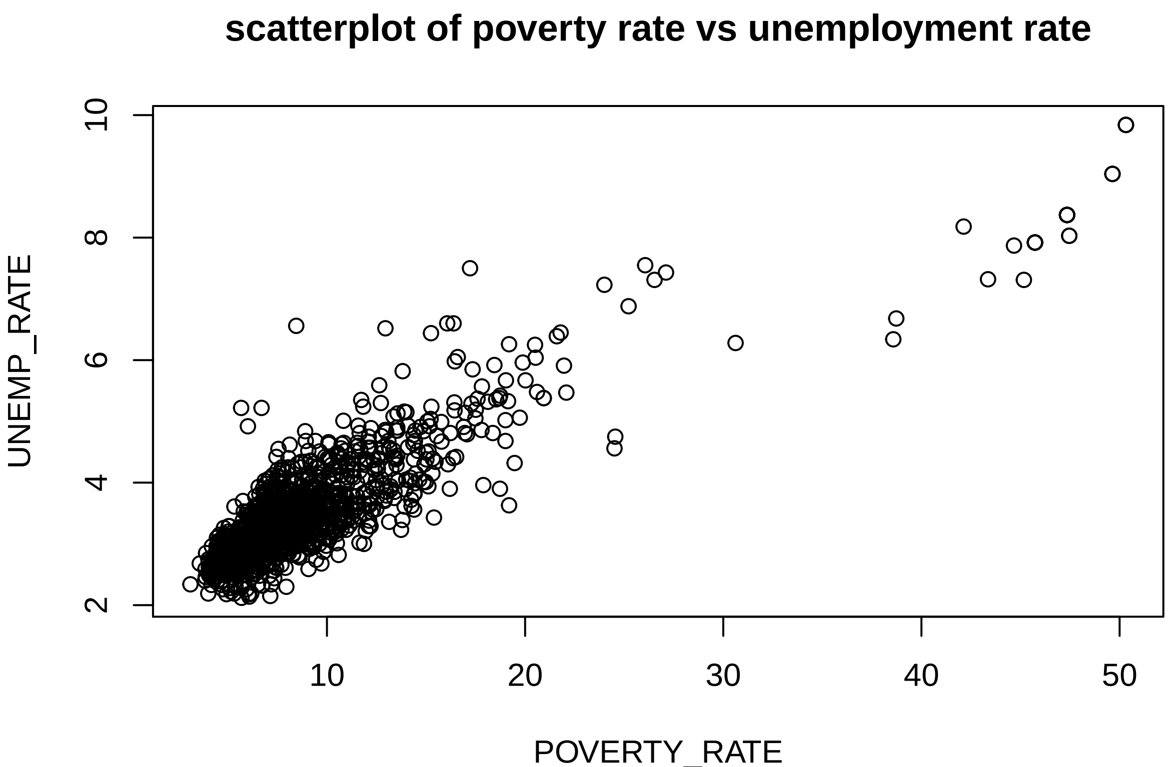
Since there are 29 predictor variables and 1 response variable,we need to examine the predictor variables first. The UNITID,INSTNM are name and coding(identifiers) of schools,hence each university will have a unique UNITID and INSTUM.So these are not predictor variables.The STABBR(the state postcode) is a factor variable with approximately 15 values,this will decrease the accuracy and elegance of the final model.So we exclude this factor.

Now we make quantitative analysis.Specifically, researchers find that the unemployment rate, median wages, and wage inequality in the lower half of the wage distribution all are significant determinants of poverty rates.[2] We test whether the result is similar in our case.

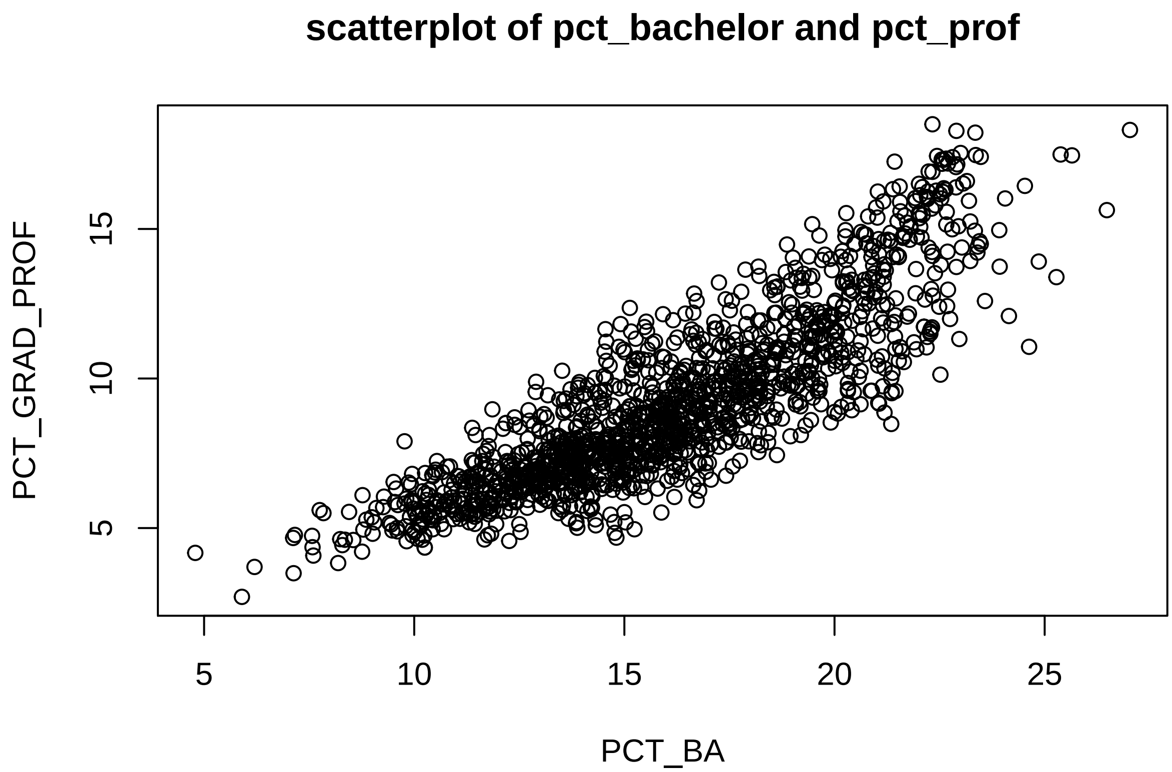
```
##   X UNITID                            INSTNM STABBR NUMBRANCH CONTROL REGION
## 1 1 100654            Alabama A & M University     AL         1       1      5
## 2 2 100663 University of Alabama at Birmingham     AL         1       1      5
## 3 4 100706 University of Alabama in Huntsville     AL         1       1      5
```

```
## 4 5 100724           Alabama State University     AL          1       1      5
## 5 6 100751          The University of Alabama     AL          1       1      5
## 6 9 100830     Auburn University at Montgomery     AL          1       1      5
##   HBCU PBI TRIBAL HSI WOMENONLY ADM_RATE COSTT4_A AVGFACSAL PFTFAC PCTPELL
## 1    1   0      0   0         0   0.8986    22489      7101 0.7411  0.7067
## 2    0   0      0   0         0   0.9211    24347     10717 0.7766  0.3632
## 3    0   0      0   0         0   0.8087    23441      9442 0.6544  0.2698
## 4    1   0      0   0         0   0.9774    21476      7754 0.5826  0.7448
## 5    0   0      0   0         0   0.5906    29424     10225 0.7454  0.1802
## 6    0   1      0   0         0   0.9281    18291      7678 0.9655  0.4584
##   UG25ABV INC_PCT_LO PAR_ED_PCT_1STGEN    FEMALE MD_FAMINC PCT_WHITE PCT_BLACK
## 1  0.0758  0.6020088         0.3658281 0.5640301   23553.0     46.84     47.98
## 2  0.2296  0.4276132         0.3412237 0.6390907   34489.0     69.02     27.76
## 3  0.1842  0.3746337         0.3101322 0.4763499   44787.0     76.38     18.98
## 4  0.0848  0.6146166         0.3434343 0.6134185   22080.5     42.69     52.32
## 5  0.0725  0.2615467         0.2257127 0.6152524   66733.5     75.35     21.06
## 6  0.2244  0.4892262         0.3818961 0.6929481   29671.5     59.97     37.21
##   PCT_ASIAN PCT_HISPANIC PCT_BA PCT_GRAD_PROF PCT_BORN_US POVERTY_RATE
## 1      1.48         3.79  13.00          6.86       94.74        14.88
## 2      1.10         2.03  15.93          8.55       96.50        10.91
## 3      1.42         2.55  17.67          8.91       95.27         9.37
## 4      1.41         4.09  11.81          6.76       94.53        16.96
## 5      1.20         2.41  16.48          9.21       96.08        10.05
## 6      0.91         1.62  14.72          9.09       96.85        13.00
##   UNEMP_RATE
## 1       4.84
## 2       3.45
```

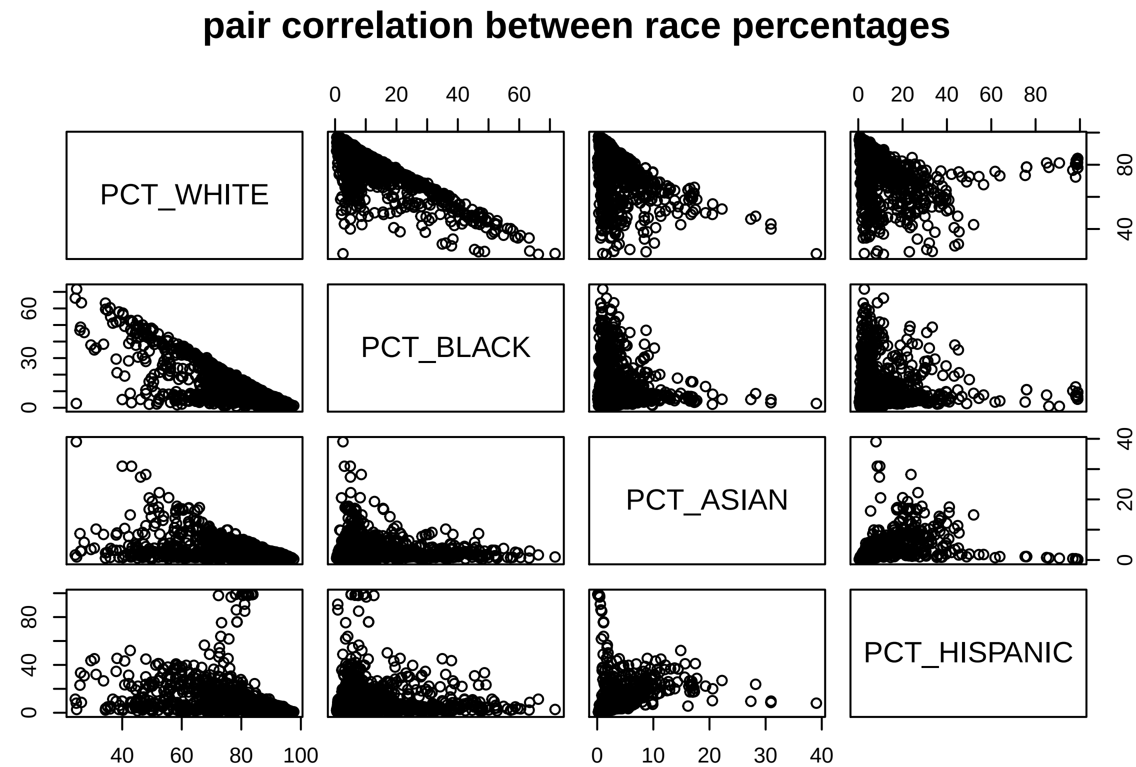
```
## 3       3.64
## 4       4.81
## 5       3.26
## 6       3.79
```



Based on our consensus,we also know that the education level is correlated.Basically,higher bachelor's degree rate will lead to a higher professional rate.We test this assumption.



We also examines the race relation.



The negative correlation seems to be apparent in PCT_WHITE,PCT_BLACK and PCT_WHITE,PCT_ASIAN.

```
##                  PCT_WHITE   PCT_BLACK   PCT_ASIAN PCT_HISPANIC      PCT_BA
## PCT_WHITE       1.00000000 -0.79321428 -0.46243554  -0.23556225  0.06196087
## PCT_BLACK      -0.79321428  1.00000000 -0.04991684  -0.04900010 -0.22940300
## PCT_ASIAN      -0.46243554 -0.04991684  1.00000000   0.15214425  0.39851124
## PCT_HISPANIC   -0.23556225 -0.04900010  0.15214425   1.00000000 -0.06306915
## PCT_BA          0.06196087 -0.22940300  0.39851124  -0.06306915  1.00000000
## PCT_GRAD_PROF   0.01690181 -0.14549584  0.37270975  -0.05398928  0.86069235
## PCT_BORN_US     0.51866148 -0.08369940 -0.63800081  -0.62524196 -0.24568241
## POVERTY_RATE   -0.29479982  0.23942855 -0.11085878   0.82200752 -0.38389814
## UNEMP_RATE     -0.47839161  0.33485076  0.04083329   0.75109275 -0.36708155
```

```
##              PCT_GRAD_PROF PCT_BORN_US POVERTY_RATE  UNEMP_RATE
## PCT_WHITE       0.01690181   0.5186615   -0.2947998 -0.47839161
## PCT_BLACK      -0.14549584  -0.0836994    0.2394286  0.33485076
## PCT_ASIAN       0.37270975  -0.6380008   -0.1108588  0.04083329
## PCT_HISPANIC   -0.05398928  -0.6252420    0.8220075  0.75109275
## PCT_BA          0.86069235  -0.2456824   -0.3838981 -0.36708155
## PCT_GRAD_PROF   1.00000000  -0.3254802   -0.3065118 -0.26339747
## PCT_BORN_US    -0.32548016   1.0000000   -0.3288603 -0.44950144
## POVERTY_RATE   -0.30651181  -0.3288603    1.0000000  0.89583487
## UNEMP_RATE     -0.26339747  -0.4495014    0.8958349  1.00000000
```

So we exclude UNEMP_RATE,PCT_BLACK,PCT_GRAD_PROF first.

Model

We need to split the data set into training set and testing set(75-25 ratio) and exclude the 7 variables as we mentioned.

```
## [1] 24
```

Based on the multiple regression model,we start with full model and use stepwise method to find the appropriate model.

```
##
## Call:
## lm(formula = ADM_RATE ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73398 -0.12397  0.01258  0.13276  0.40221
```

```
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.008e-01  1.919e-01   3.651 0.000273 ***
## NUMBRANCH          5.954e-03  1.864e-03   3.194 0.001444 **
## CONTROL           -5.490e-02  1.832e-02  -2.996 0.002792 **
## REGION             2.104e-03  3.539e-03   0.594 0.552397
## HBCU              -4.937e-02  4.248e-02  -1.162 0.245401
## PBI               -2.417e-02  4.601e-02  -0.525 0.599520
## TRIBAL             1.758e-01  1.283e-01   1.370 0.170896
## HSI                3.695e-02  2.548e-02   1.450 0.147253
## WOMENONLY          1.857e-02  6.532e-02   0.284 0.776272
## COSTT4_A          -1.925e-06  7.279e-07  -2.644 0.008304 **
## AVGFACSAL         -2.807e-05  3.583e-06  -7.834  1.1e-14 ***
## PFTFAC            -6.999e-02  2.427e-02  -2.884 0.003997 **
## PCTPELL            1.722e-02  7.091e-02   0.243 0.808167
## UG25ABV           -2.379e-02  5.440e-02  -0.437 0.661959
## INC_PCT_LO         1.208e-02  1.393e-01   0.087 0.930909
## PAR_ED_PCT_1STGEN  3.211e-01  1.183e-01   2.714 0.006744 **
## FEMALE             1.168e-01  4.674e-02   2.500 0.012558 *
## MD_FAMINC          2.054e-06  7.355e-07   2.793 0.005309 **
## PCT_WHITE          1.119e-03  8.122e-04   1.377 0.168765
## PCT_ASIAN          9.771e-05  3.179e-03   0.031 0.975487
## PCT_HISPANIC      -8.697e-04  1.071e-03  -0.812 0.416858
## PCT_BA             1.413e-03  2.745e-03   0.515 0.606826
## PCT_BORN_US        2.031e-04  1.703e-03   0.119 0.905054
## POVERTY_RATE      -2.218e-03  2.565e-03  -0.865 0.387363
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1764 on 1107 degrees of freedom
## Multiple R-squared:  0.2297, Adjusted R-squared:  0.2137
## F-statistic: 14.35 on 23 and 1107 DF,  p-value: < 2.2e-16
```

More than half of the predictor variables are not significant.Though the overall model is significant,we have to limit the number of predictor variables.

Based on the selection, we have four final models.

Model 1: lm(formula = ADM_RATE ~ AVGFACSAL + CONTROL + POVERTY_RATE + HBCU + COSTT4_A + NUMBRANCH + FEMALE + PFTFAC + MD_FAMINC + HSI + PCT_BORN_US + REGION, data = train_set0) with 12 predictor variables

Model 2: lm(formula = ADM_RATE ~ NUMBRANCH + CONTROL + REGION + TRIBAL + HSI + COSTT4_A + AVGFACSAL + PFTFAC + PAR_ED_PCT_1STGEN + FEMALE + MD_FAMINC + PCT_WHITE + PCT_HISPANIC + PCT_BA, data = train_set0) with 14 predictor variables

Model 3: lm(formula = ADM_RATE ~ AVGFACSAL + CONTROL + POVERTY_RATE + HBCU + COSTT4_A + NUMBRANCH + FEMALE, data = train_set0) with 7 predictor variables

Model 4: lm(formula = ADM_RATE ~ NUMBRANCH + HSI + COSTT4_A + AVGFACSAL + PFTFAC + FEMALE + MD_FAMINC + PCT_WHITE + PCT_HISPANIC, data = train_set0) with 9 predictor variables

Now we compare these models based on the selection criteria:AIC/BIC/AICc

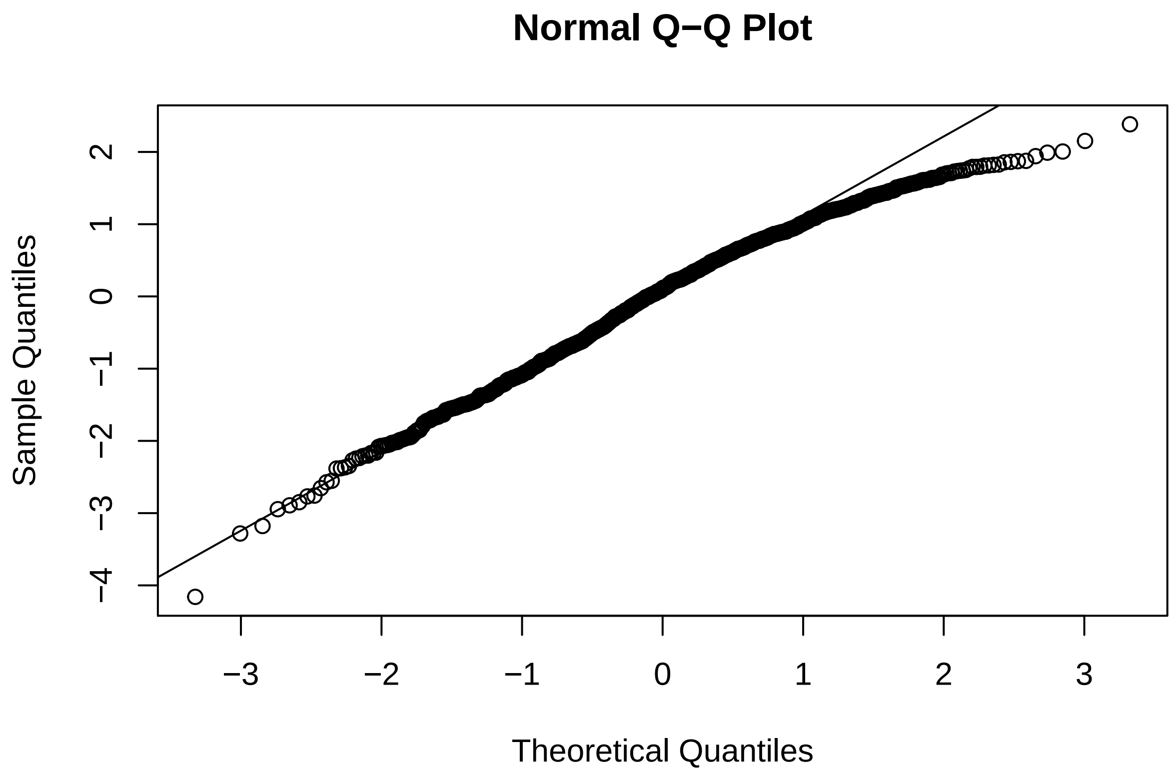
```
##          SSres   Rsq_adj       AIC     AIC_c       BIC  p
## [1,] 34.87981 0.2118837 -3910.692 -3910.316 -3836.260 12
```

```
## [2,] 34.54599 0.2180276 -3917.568 -3917.081 -3833.074 14
## [3,] 35.55605 0.2001810 -3898.974 -3898.814 -3849.696  7
## [4,] 35.28882 0.2047759 -3903.506 -3903.271 -3844.167  9
```

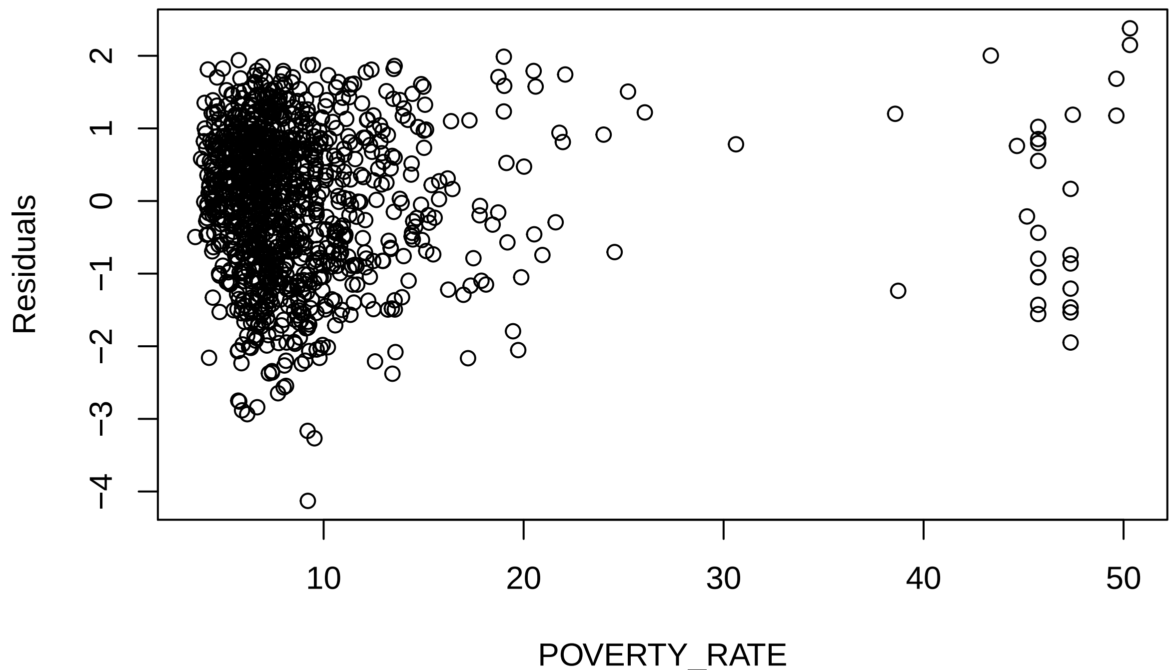
The R^2 value is quite similar in all four cases as well as all selection criteria. We then compare the MAE and RMSE.

```
## Installing package into '/opt/r'
## (as 'lib' is unspecified)

##      MAE  RMSE  model
## 1 0.1402 0.175 model1

##      MAE   RMSE  model
## 1 0.1391 0.1749 model2

##      MAE   RMSE  model
## 1 0.1424 0.1766 model3

##      MAE   RMSE  model
## 1 0.1391 0.1744 model4
```

In this case,we decide to use the model with least predictor variables.We choose model 3.



The normality of the assumption is satisfied.



There's a pattern of residuals in the residual plot. The residuals might be correlated with each other.

Note that there are outliers in the model. We need to exclude the outliers.

```
##
## Call:
## lm(formula = ADM_RATE ~ AVGFACSAL + CONTROL + POVERTY_RATE +
##     HBCU + COSTT4_A + NUMBRANCH + FEMALE, data = train_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.75129 -0.11835  0.01768  0.13139  0.36908
##
```

```
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.698e-01  5.102e-02  19.008  < 2e-16 ***
## AVGFACSAL    -2.371e-05  3.533e-06  -6.713 3.20e-11 ***
## CONTROL      -4.917e-02  2.167e-02  -2.269 0.023470 *
## POVERTY_RATE -7.730e-03  1.973e-03  -3.919 9.51e-05 ***
## HBCU                 NA         NA      NA       NA
## COSTT4_A     -1.940e-06  7.137e-07  -2.718 0.006676 **
## NUMBRANCH     1.140e-02  5.485e-03   2.079 0.037913 *
## FEMALE        1.693e-01  4.807e-02   3.522 0.000448 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1748 on 997 degrees of freedom
## Multiple R-squared:  0.1526, Adjusted R-squared:  0.1475
## F-statistic: 29.92 on 6 and 997 DF,  p-value: < 2.2e-16
```

Excluding the outliers doesn;t necessarily improve the quality of the original fitting model.Since there are approximately 10% of the data are identified as outliers,we decided to include these outliers in the final model.

```
##      MAE   RMSE  model
## 1 0.1456 0.1773 model3
```

```
##      MAE   RMSE  model
## 1 0.1424 0.1766 model3
```

Our final step insures that the RMSE for the training set and testing set are approximately the same. Thus,there's no overfitting or underfitting of the data.

4 Results

Our final model is:

```
##
## Call:
## lm(formula = ADM_RATE ~ AVGFACSAL + CONTROL + POVERTY_RATE +
##     HBCU + COSTT4_A + NUMBRANCH + FEMALE, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73087 -0.12411  0.01996  0.13646  0.41511
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.034e+00  4.276e-02  24.186  < 2e-16 ***
## AVGFACSAL    -2.935e-05  2.971e-06  -9.879  < 2e-16 ***
## CONTROL      -5.661e-02  1.581e-02  -3.580 0.000358 ***
## POVERTY_RATE -5.856e-03  8.796e-04  -6.658 4.35e-11 ***
## HBCU         -1.043e-01  3.033e-02  -3.439 0.000604 ***
## COSTT4_A     -1.743e-06  5.734e-07  -3.039 0.002432 **
## NUMBRANCH     6.393e-03  1.839e-03   3.477 0.000526 ***
## FEMALE        1.264e-01  4.285e-02   2.951 0.003235 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1779 on 1123 degrees of freedom
## Multiple R-squared:  0.2051, Adjusted R-squared:  0.2002
```

```
## F-statistic:  41.4 on 7 and 1123 DF,  p-value: < 2.2e-16
```

All of them are significantly different from zero.

5 Discussion

It is obvious that this predictor variable has 7 predictor variables. No transformation on the predictor variable is implemented,so we may need to test the normality of the data by using histogram for further research.

Based on the introduction we made in first section, it is obvious that the admission rate is related to school identifiers (Number of branch, HSI),school characteristics (COSTT4_A,AVGFACSAL,PFTFAC,FEMALE) and student population characteristics (MD_FAMINC,PCT_WHITE and PCT_HISPANIC). So all these 3 categories are useful in predicting the admission rate for American universities/colleges. The most important variables are AVGFACSAL,FEMALE,NUMBRANCH and MD_FAMINC. The factors related to the student population characteristics are hard to change, but we can view the school characteristics as a guide when we select the universities/colleges.

6 Reference

[1]Data Home: College Scorecard. Data Home | College Scorecard. (n.d.). Retrieved April 27, 2022, from https://collegescorecard.ed.gov/data/

[2]Why poverty persists. NBER. (n.d.). Retrieved April 27, 2022, from https://www.nber.org/digest/jun06/why-poverty-persists

[3]RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.