Multiple Linear Regression on Admission Rate

Hanzi Jiang

Introduction

College and university admission rates differ from school to school. Understanding what factors attribute to this variation is important, particularly when stakeholders wish to predict a future admission rate given a trend or a change in the admission practices of schools. For example, school boards may want to know how the admission rate will change if the poverty rate of their school districts decreases, so that they can prepare for changing educational needs. A model is built in this project, primarily used to describe and understand what affects admission rates. It can also be used to predict admission rates if some predictors change. This model is simple enough to understand, while complicated enough to make good predictions. It also satisfies all necessary properties. The dataset contains information for admission rate and 29 other variables on 1508 colleges and universities in the US.

Method

Since the dataset was large, it was randomly split into a training and a validation dataset (754/754). This provided a way of evaluating the performance of the final models. The model was built from the training dataset. To find out if linear regression analysis was appropriate, and to verify independence, scatterplots of ADM_RATE versus each numerical predictor were made.

Model assumptions, multicollinearity (by VIF) and extreme observations were checked for the initial full model. Sets of four diagnostic plots were generated for each numerical predictor. The overall F-test indicates an overall significant linear relationship. Individual T-tests showed some of the predictors were significant, so it was reasonable to do a predictor selection.

The all subsets method was not suitable due to a large number of variables. Forward, backward and stepwise selections for AIC and BIC were preformed instead. A combination of these selection methods was needed because each one did not explore all models. Different methods were likely to yield different models. Models with high adjusted R-squared and low AIC, BIC, corrected-AIC were preferred. All of the criteria were important because they measured different aspects of model fit with different penalties for complexity. BIC and corrected-AIC were valued more, since they penalized complexity harder. Simpler models that could be understood by stakeholders were preferred. Predictors were attempted to be added to the chosen best models, yet the increase in adjusted R-squared did not compensate complexity and the increase in BIC. Predictors in the models were then attempted to be removed, one at a time, until simplicity no longer compensated the drop in R-squared or rise in AIC, BIC or corrected-AIC. Partial F test was performed to see if predictors excluded in the final model could be dropped from the initial full model simultaneously.

Model validity was checked for each chosen model. Sets of four diagnostic plots were generated for each model. Residuals vs Fitted plot was used to determine linearity. Equally spread residuals around the horizontal line without patterns indicated linearity being met. Normality was satisfied when the residuals in the Normal Q-Q plot followed a straight line. Constant spread of the residuals in the Scale-Location plot indicated constant variance was met. Residuals vs Leverage plot gave information on influential and extreme observations. All violations could be attempted to be corrected by performing common transformations or Box-Cox transformations. Luckily, no transformation was needed. Multicollinearity could lead to inflated variances of the regression coefficients, making the regression surface unstable. It was proven to be absent by VIF. Leverage points and outliers were checked. Points with the most influence by dfbetas were removed and added back to see the difference of regression coefficients. Models with no observation that greatly impact the regression line were preferred.

Candidate models were tested on the validation dataset to see if they perform just as well as in the training dataset. The selection methods specified above gave biased estimated coefficients and larger test statistics, so a validation step was particularity necessary. The influence of each data points on each regression coefficients was calculated. It was found no specific observation had a great impact on the regression lines of both the training and the validation models. No point would dramatically affect the model fit if removed. The regression coefficients of the validation model were compared with the ones of the training model. Minor differences were found. Adjusted R-squared, AIC, BIC, and corrected-AIC were computed for the validation models. A good model should yield similar outputs because the two datasets were independent. The model that equally best fitted both the two datasets and with no observation(or less) that could highly impact the regression coefficients was chosen.

Result

Figure 1 shows a summary of the dataset with a description for each variable. First 11 except NUMBRANCH are categorical, the others are numerical. From the scatterplots of response versus each predictor (Appendix Figure 1), it is clear a linear trend is present and independence is satisfied. REGION was found to be a summary of STABBR. Since it was not clear which one was more useful, both were kept. The initial full model included all predictors except UNITID and INSTNM because they were unique names and IDs of institutions but not reasonable predictors. The overall F test was significant (p-value < 2.2e-16).

Models suggested by forward, backward, stepwise sections for AIC and BIC and their adjusted R-squareds, AICs, BICs, corrected-AICs are shown in Figure 2. AIC forward and stepwise models have far more predictors. Their slightly higher adjusted R-squared (< 2%) and lower AIC and corrected-AIC does not compensate the higher BIC and complexity. These two models were abandoned. AIC backward model has 2 more predictors than BIC backward, roughly the same adjusted R-squared, AIC and corrected-AIC, and a slightly higher BIC. Therefore the AIC backward model is abandoned. Even though the BIC forward model was slightly worse in every aspect, it had one less predictor than the BIC backward model. Both BIC forward and backward models are kept. BIC backward and stepwise models contained the same predictors.

Seven more models were made, each with one predictor removed from the BIC backward model (Appendix Figure 2). The model with NUMBRANCH removed (call it as model 1) was chosen because it had similar results as the original BIC backward model. All other models either resulted in a much lower adjusted R-squared, much higher AIC, BIC, or corrected-AIC. The BIC forward model was abandoned because even though it had the same number of predictor as model 1, it had worse AIC, BIC, and corrected-AIC.

The same procedure was repeated. Six more models were made by removing one predictor from model 1 (Appendix Figure 3). Removal of HSI or PFTFAC reduced a small amount of adjusted R-squared, but the increase in AIC, BIC, and corrected-AIC is moderate, so all three models were kept. Call the model with HSI removed model 2, with PFTFAC removed model 3. Any further removal of the predictor made the models fit much worse.

The influential points by dfbetas for each of the three models are shown in Figure 3 in red. Model 1 and 3 have observations 157, 186, 439, 553 as both leverage points and outliers. Model 2 has 186, 578 as both leverage points and outliers. Each model contains roughly the same number of extreme observations. All three models had similar diagnostic plots (Figure 4). The Residuals vs Fitted plots are centred around the horizontal line y = 0 and with a roughly equal spread. The left goes up a bit due to a few outliers, not an indication of non-linearity or unequal variances. The Scale-Location plots have equal spread around a horizontal line with no distinct pattern. A few outliers cause the left tail to curve up a bit, but it is not a serious problem. Points in the Normal Q-Q plot lie on the dashed line. The change in estimated coefficients would be trivial if the outliers in the Residual-Leverage plot were removed. All three models satisfy model assumptions, with no influential point that altered the regression coefficients greatly. However, partial F tests preferred the initial full model over the final models (p-value = 0.0008752, 0.0001258, 0.0001318), meaning some useful predictors were missing in the models. Better model could be obtained by exploring more models.

The models were tested on the validation dataset. Both the training and validation models have a similar number of leverage points, outliers, and influential points. The significance of each coefficient did not change much, and regression coefficients of each model with the most influential point removed did not alter (Figure 5). No observation had a disproportionate effect on the regression coefficients for both the training model and the validation model. The regression coefficients of model 1 fluctuated by ± 0.038 when tested on the validation dataset. The ones of model 2 fluctuated by 0.576. The ones of model 3 fluctuated by 0.066. The adjusted R-squared, AIC, BIC, and corrected-AIC are slightly higher in the validation dataset. Model 2 was chosen as the final model because it was simpler, it performed equally well as the other models, and its regression coefficients differed by an acceptable amount. Most regression coefficients in the training model 2 and all in the validation model 2 were significant. It was overall significant (p-value < 2.2e-16). The minor difference in the outputs might be due to how the two datasets were sampled in R. The final model is:

$$\hat{Y} = 1.09 - 0.1157 \cdot X_1 - 0.06353 \cdot X_2 - 2.916 \times 10^{-5} \cdot X_3 - 0.09773 \cdot X_4 - 2.82 \times 10^{-3} \cdot X_5 - 2.054 \times 10^{-3} \cdot X_6$$

Discussion

$\hat{Y} = 1.09 - 0.1157 \cdot X_1 - 0.06353 \cdot X_2 - 2.916 \times 10^{-5} \cdot X_3 - 0.09773 \cdot X_4 - 2.82 \times 10^{-3} \cdot X_5 - 2.054 \times 10^{-3} \cdot X_6$

$\hat{Y}$ is estimated admission rate;
$X_1 = 1$ for private non-profit institution, 0 otherwise;
$X_2 = 1$ for private for-profit institution, 0 otherwise;
X_3 is average faculty salary;
X_4 is proportion of full-time faculty members;
X_5 is percentage of Black students;
X_6 is percentage of Asian students.

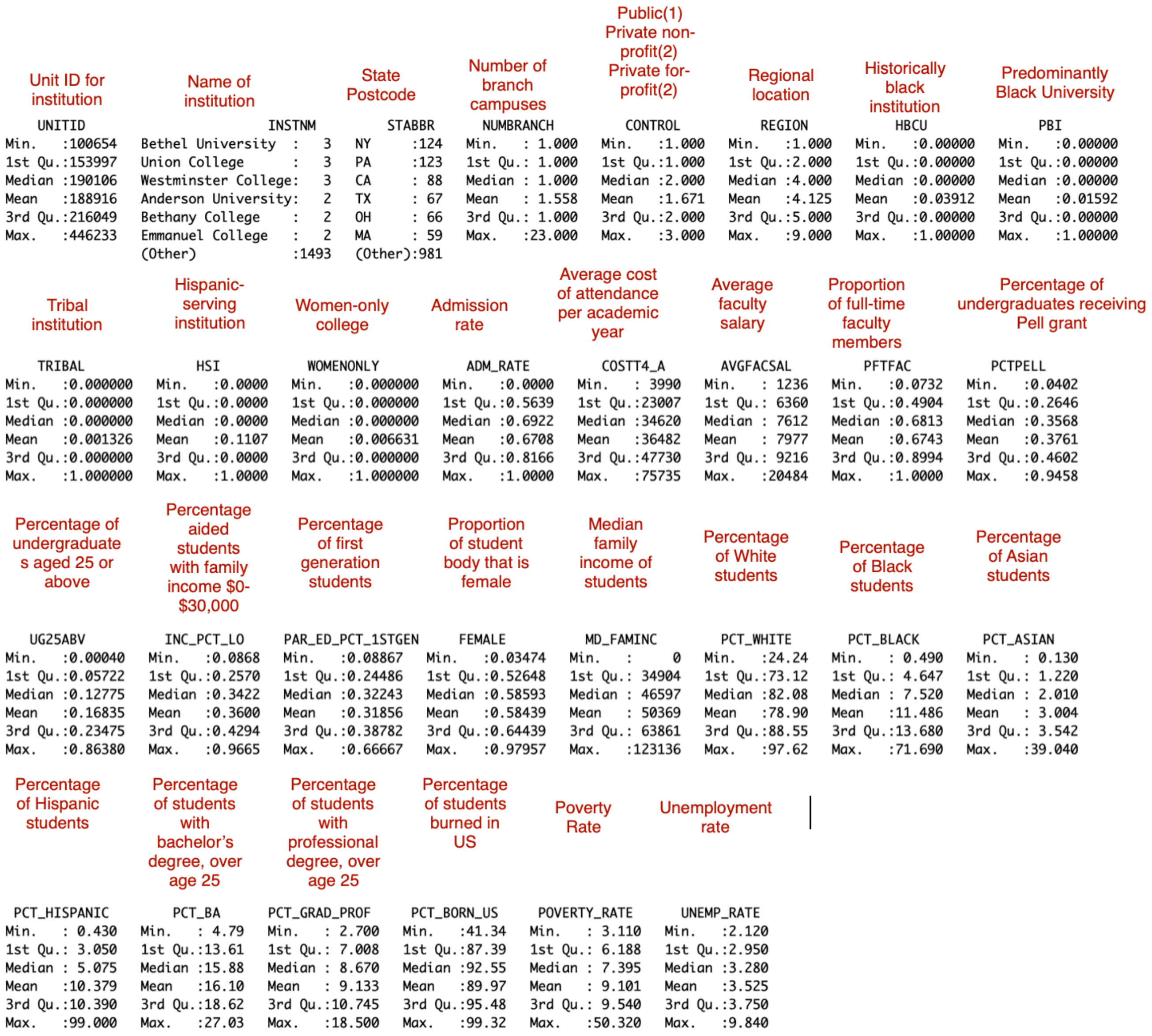
- When $X_1 \dots X_6$ are zeros, the admission rate is expected to be 1.09 on average.
- When $X_3 \dots X_6$ are fixed, the admission rate is expected to on average decrease by 0.1157 if the institution is private non-profit instead of public, or increase by 0.1157 if public instead of private non-profit, or decrease by 0.06353 if private for-profit instead of public, or increase by 0.06353 if public instead of private for-profit, or decrease by 0.05271 if private for-profit instead of private non-profit, or increase by 0.05271 if private non-profit instead of private for-profit.
- When $X_1, X_2, X_4 \dots X_6$ are fixed, the admission rate is expected to decrease by 2.916×10^{-5} on average with each unit increase in average faculty salary.
- When $X_1 \dots X_3, X_5, X_6$ are fixed, the admission rate is expected to decrease by 9.773×10^{-4} on average with each percent increase in proportion of full-time faculty members.
- When $X_1 \dots X_4, X_6$ are fixed, the admission rate is expected to decrease by 2.82×10^{-5} on average with each percent increase in the proportion of Black students.
- When $X_1 \dots X_5$ are fixed, the admission rate is expected to decrease by 2.054×10^{-5} on average with each percent increase in the proportion of Asian students.

The model is simple, as it only requires five pieces of information. It is highly interpretable, and is able to explain a substantial amount (19.42%) of variability in the responses. It gives good predictions. It is also overall significant (p-value < 2.2e-16). The regression model is important as it allows shareholders to understand what influence admission rate the most. They can predict future admission rates given a change in any of the predictors.

The model gives valid prediction only when values of $X_1 \dots X_6$ yield an estimated admission rate in the rage [0,1]. The partial F test did not support dropping the predictors not in the final model from the initial full model simultaneously. This means some useful predictors are not included in the final model. Better models would be found if more models were explored. The R-squared is not very high. It is unable to capture some variations in the responses, and prediction intervals will be large. The predictor selection method leads to biases in parameters. This results in larger test statistics, making null hypothesis to be rejected when it should not. The selection method did not explore all possible combinations of predictors. Forward, backward, and stepwise selection each returns only one model. It is possible that the returned models are high in only AIC or BIC, but not adjusted R-squared or corrected-AIC. Models selected based solely on AIC tend to be overfitted. The order of parameter entry/deletion also has a great impact on the chosen model. Differences in regression coefficients is seen in the validation process, so the model might not perform that well on other datasets. This might due to how datasets are sampled by R or the biased selection method. Therefore, the final model may not be the best.

Figures and Tables

Figure 1. Summary and description of each variable.


Unit ID for institution
Name of institution
State Postcode
Number of branch campuses
Public(1) Private non-profit(2) Private for-profit(2)
Regional location
Historically black institution
Predominantly Black University
Tribal institution
Hispanic-serving institution
Women-only college
Admission rate
Average cost of attendance per academic year
Average faculty salary
Proportion of full-time faculty members
Percentage of undergraduates receiving Pell grant
Percentage of undergraduates aged 25 or above
Percentage aided students with family income $0-$30,000
Percentage of first generation students
Proportion of student body that is female
Median family income of students
Percentage of White students
Percentage of Black students
Percentage of Asian students
Percentage of Hispanic students
Percentage of students with bachelor's degree, over age 25
Percentage of students with professional degree, over age 25
Percentage of students burned in US
Poverty Rate
Unemployment rate

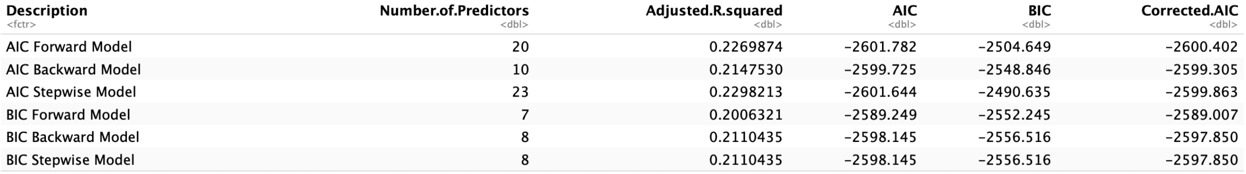
```
     UNITID                          INSTNM         STABBR      NUMBRANCH          CONTROL          REGION            HBCU               PBI
 Min.   :100654   Bethel University  :   3   NY     :124   Min.   : 1.000   Min.   :1.000   Min.   :1.000   Min.   :0.00000   Min.   :0.00000
 1st Qu.:153997   Union College      :   3   PA     :123   1st Qu.: 1.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:0.00000   1st Qu.:0.00000
 Median :190106   Westminster College:   3   CA     : 88   Median : 1.000   Median :2.000   Median :4.000   Median :0.00000   Median :0.00000
 Mean   :188916   Anderson University:   2   TX     : 67   Mean   : 1.558   Mean   :1.671   Mean   :4.125   Mean   :0.03912   Mean   :0.01592
 3rd Qu.:216049   Bethany College    :   2   OH     : 66   3rd Qu.: 1.000   3rd Qu.:2.000   3rd Qu.:5.000   3rd Qu.:0.00000   3rd Qu.:0.00000
 Max.   :446233   Emmanuel College   :   2   MA     : 59   Max.   :23.000   Max.   :3.000   Max.   :9.000   Max.   :1.00000   Max.   :1.00000
                  (Other)            :1493   (Other):981
```

```
     TRIBAL              HSI            WOMENONLY           ADM_RATE         COSTT4_A        AVGFACSAL         PFTFAC          PCTPELL
 Min.   :0.000000   Min.   :0.0000   Min.   :0.000000   Min.   :0.0000   Min.   : 3990   Min.   : 1236   Min.   :0.0732   Min.   :0.0402
 1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:0.000000   1st Qu.:0.5639   1st Qu.:23007   1st Qu.: 6360   1st Qu.:0.4904   1st Qu.:0.2646
 Median :0.000000   Median :0.0000   Median :0.000000   Median :0.6922   Median :34620   Median : 7612   Median :0.6813   Median :0.3568
 Mean   :0.001326   Mean   :0.1107   Mean   :0.006631   Mean   :0.6708   Mean   :36482   Mean   : 7977   Mean   :0.6743   Mean   :0.3761
 3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:0.000000   3rd Qu.:0.8166   3rd Qu.:47730   3rd Qu.: 9216   3rd Qu.:0.8994   3rd Qu.:0.4602
 Max.   :1.000000   Max.   :1.0000   Max.   :1.000000   Max.   :1.0000   Max.   :75735   Max.   :20484   Max.   :1.0000   Max.   :0.9458
```

```
    UG25ABV           INC_PCT_LO     PAR_ED_PCT_1STGEN      FEMALE           MD_FAMINC        PCT_WHITE        PCT_BLACK        PCT_ASIAN
 Min.   :0.00040   Min.   :0.0868   Min.   :0.08867   Min.   :0.03474   Min.   :     0   Min.   :24.24   Min.   : 0.490   Min.   : 0.130
 1st Qu.:0.05722   1st Qu.:0.2570   1st Qu.:0.24486   1st Qu.:0.52648   1st Qu.: 34904   1st Qu.:73.12   1st Qu.: 4.647   1st Qu.: 1.220
 Median :0.12775   Median :0.3422   Median :0.32243   Median :0.58593   Median : 46597   Median :82.08   Median : 7.520   Median : 2.010
 Mean   :0.16835   Mean   :0.3600   Mean   :0.31856   Mean   :0.58439   Mean   : 50369   Mean   :78.90   Mean   :11.486   Mean   : 3.004
 3rd Qu.:0.23475   3rd Qu.:0.4294   3rd Qu.:0.38782   3rd Qu.:0.64439   3rd Qu.: 63861   3rd Qu.:88.55   3rd Qu.:13.680   3rd Qu.: 3.542
 Max.   :0.86380   Max.   :0.9665   Max.   :0.66667   Max.   :0.97957   Max.   :123136   Max.   :97.62   Max.   :71.690   Max.   :39.040
```

```
  PCT_HISPANIC        PCT_BA       PCT_GRAD_PROF     PCT_BORN_US     POVERTY_RATE      UNEMP_RATE
 Min.   : 0.430   Min.   : 4.79   Min.   : 2.700   Min.   :41.34   Min.   : 3.110   Min.   :2.120
 1st Qu.: 3.050   1st Qu.:13.61   1st Qu.: 7.008   1st Qu.:87.39   1st Qu.: 6.188   1st Qu.:2.950
 Median : 5.075   Median :15.88   Median : 8.670   Median :92.55   Median : 7.395   Median :3.280
 Mean   :10.379   Mean   :16.10   Mean   : 9.133   Mean   :89.97   Mean   : 9.101   Mean   :3.525
 3rd Qu.:10.390   3rd Qu.:18.62   3rd Qu.:10.745   3rd Qu.:95.48   3rd Qu.: 9.540   3rd Qu.:3.750
 Max.   :99.000   Max.   :27.03   Max.   :18.500   Max.   :99.32   Max.   :50.320   Max.   :9.840
```

Figure 2. Performance of models returned by forward, backward and stepwise of AIC and BIC



Description	Number.of.Predictors	Adjusted.R.squared	AIC	BIC	Corrected.AIC
<fctr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
AIC Forward Model	20	0.2269874	−2601.782	−2504.649	−2600.402
AIC Backward Model	10	0.2147530	−2599.725	−2548.846	−2599.305
AIC Stepwise Model	23	0.2298213	−2601.644	−2490.635	−2599.863
BIC Forward Model	7	0.2006321	−2589.249	−2552.245	−2589.007
BIC Backward Model	8	0.2110435	−2598.145	−2556.516	−2597.850
BIC Stepwise Model	8	0.2110435	−2598.145	−2556.516	−2597.850

Figure 3. Influential points in model 1, mode 2, and model 3

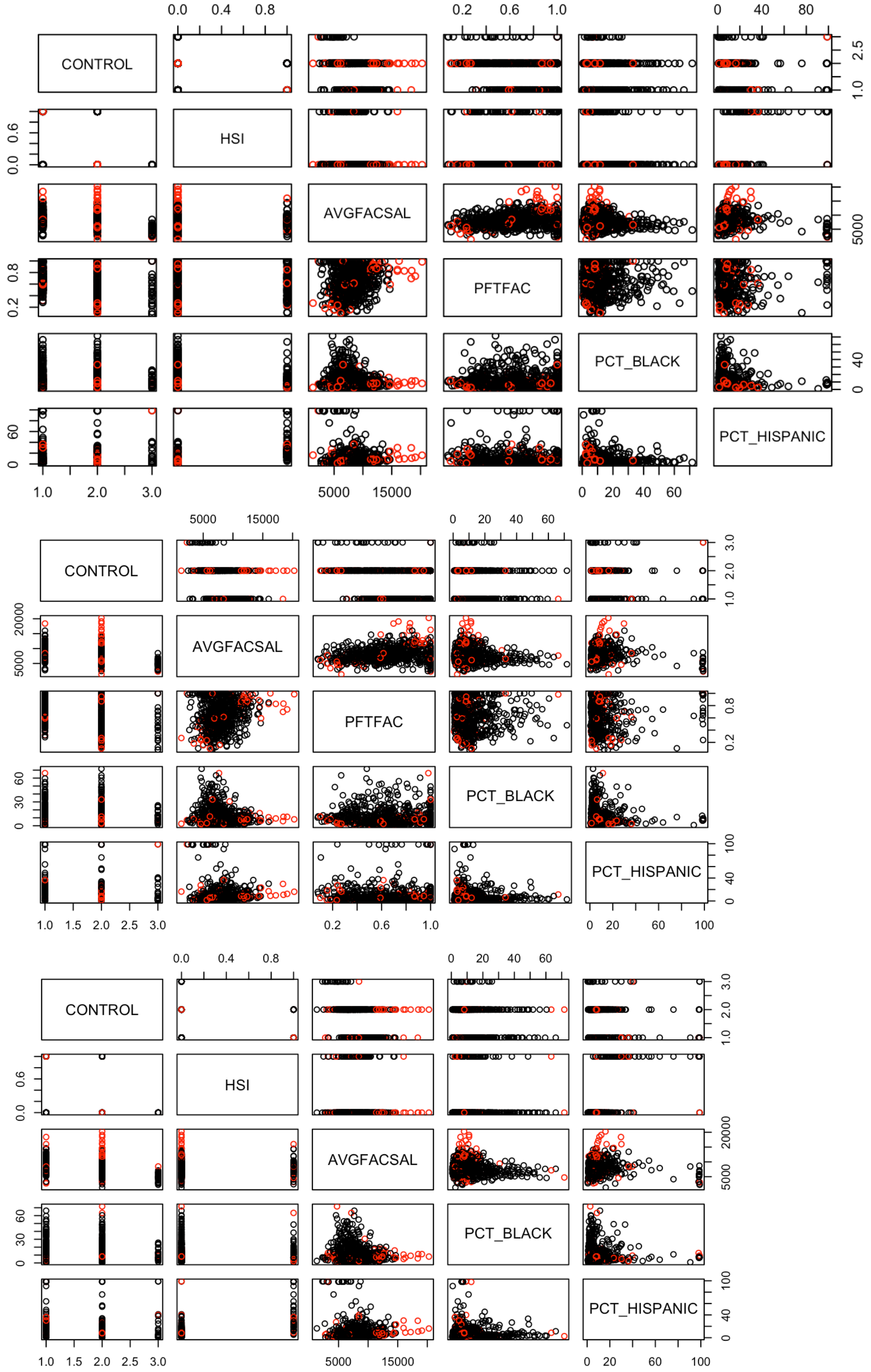


Figure 4. Diagnostics plots for model 1, model 2, and model 3 (bottom)

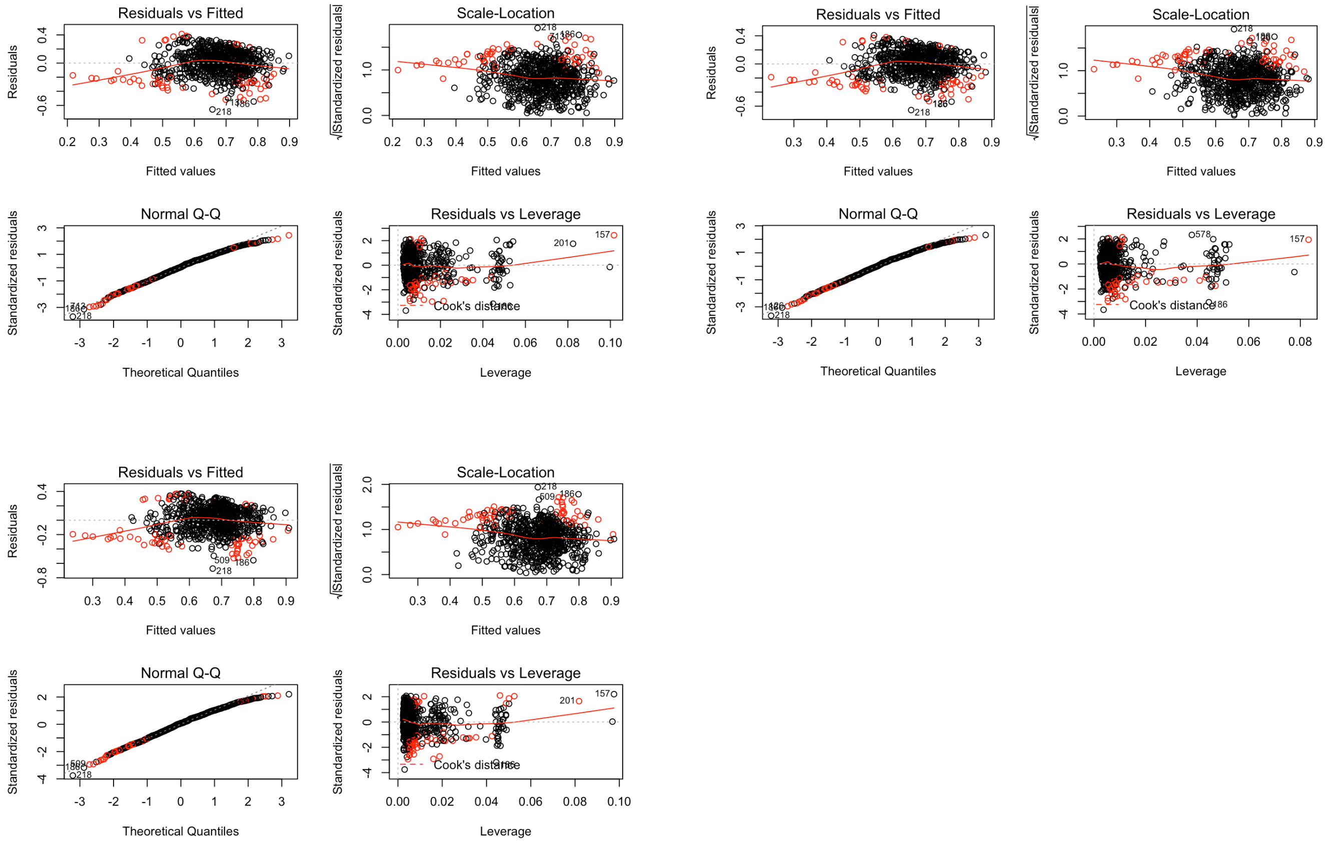


Figure 5. Regression Coefficients for model 1, model, and model 3, models with the most influential point removed, the validation models, validation models with the most influential point removed.

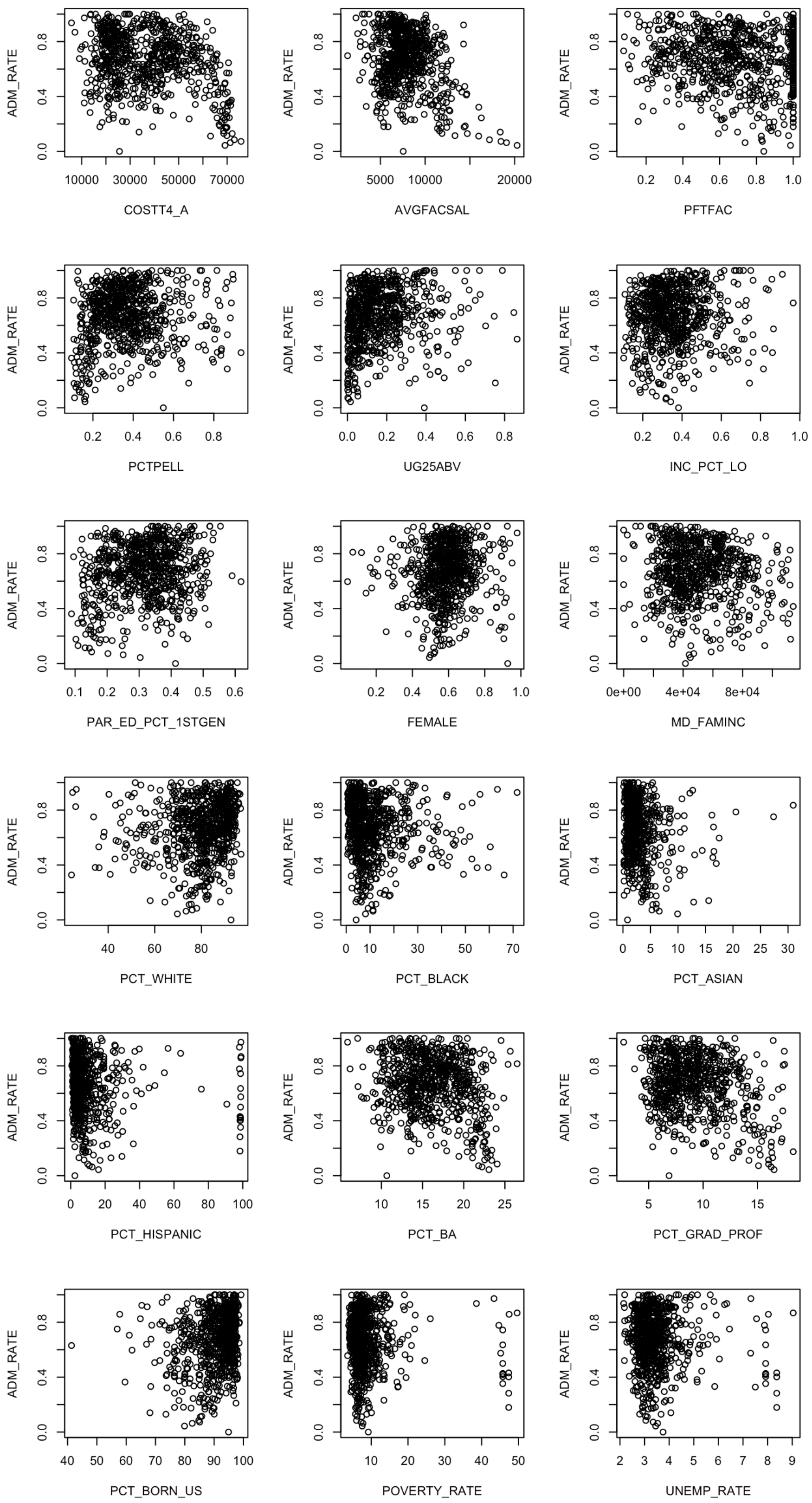
Predictor/Coeffcient	Model 1 Training	Model 1 Training Observation #2634 Removed	Model 1 Validation	Model 1 Validation Observation #3222 Removed
(Intercept)	1.086e+00 ***	1.086e+00 ***	1.112e+00 ***	1.112e+00 ***
CONTROL = 2	-1.136e-01 ***	-1.136e-01 ***	-1.033e-01 ***	-1.033e-01 ***
CONTROL = 3	-4.050e-02	-4.050e-02	-1.058e-01 **	-1.058e-01 **
HSI = 1	9.123e-02 **	9.123e-02 **	5.364e-02 .	5.364e-02 .
AVGFACSAL	-2.928e-05 ***	-2.928e-05 ***	-3.547e-05 ***	-3.547e-05 ***
PFTFAC	-8.86e-02 **	-8.86e-02 **	-5.065e-02 .	-5.065e-02 .
PCT_BLACK	-2.93e-03 ***	-2.93e-03 ***	-3.111e-03 ***	-3.111e-03 ***
PCT_HISPANIC	-3.126e-03 ***	-3.126e-03 ***	-2.166e-03 ***	-2.166e-03 ***

Predictor/Coeffcient	Model 2 Training	Model 2 Training Observation #72 Removed	Model 2 Validation	Model 2 Validation Observation #2554 Removed
(Intercept)	1.090e+00 ***	1.090e+00 ***	1.117e+00 ***	1.117e+00 ***
CONTROL = 2	-1.157e-01 ***	-1.157e-01 ***	-1.084e-01 ***	-1.084e-01 ***
CONTROL = 3	-6.353e-02	-6.353e-02	-1.211e-01 **	-1.211e-01 **
AVGFACSAL	-2.916e-05 ***	-2.916e-05 ***	-3.524e-05 ***	-3.524e-05 ***
PFTFAC	-9.773e-02 ***	-9.773e-02 ***	-5.843e-02 *	-5.843e-02 *
PCT_BLACK	-2.820e-03 ***	-2.820e-03 ***	-3.124e-03 ***	-3.124e-03 ***
PCT_HISPANIC	-2.054e-03 ***	-2.054e-03 ***	-1.402e-03 ***	-1.402e-03 ***

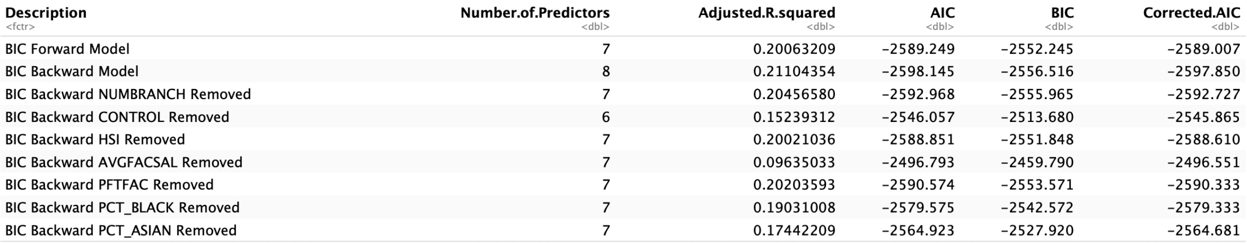
Predictor/Coeffcient	Model 3 Training	Model 3 Training Observation #2770 Removed	Model 3 Validation	Model 3 Validation Observation #3254 Removed
(Intercept)	1.025e+00 ***	1.025e+00 ***	1.076e+00 ***	1.076e+00 ***
CONTROL = 2	-1.087e-01 ***	-1.087e-01 ***	-9.931e-02 ***	-9.931e-02 ***
CONTROL = 3	-2.044e-02	-2.044e-02	-8.651e-02 *	-8.651e-02 *
AVGFACSAL	1.004e-01 ***	1.004e-01 ***	6.234e-02 *	6.234e-02 *
PFTFAC	-2.973e-05 ***	-2.973e-05 ***	-3.574e-05 ***	-3.574e-05 ***
PCT_BLACK	-2.875e-03 ***	-2.875e-03 ***	-3.062e-03 ***	-3.062e-03 ***
PCT_HISPANIC	-3.231e-03 ***	-3.231e-03 ***	-2.262e-03 ***	-2.262e-03 ***

Appendix

Appendix Figure 1. Scatterplots of admission rate against each numerical candidate predictor

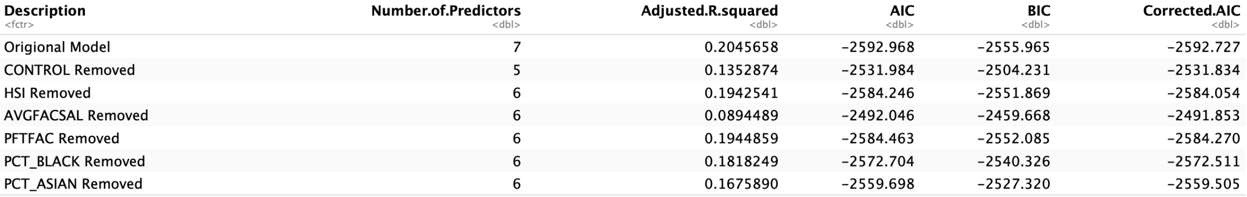


Appendix Figure 2. Performance of models BIC forward, BIC backward, and BIC backward with one additional predictor removed



Description	Number.of.Predictors	Adjusted.R.squared	AIC	BIC	Corrected.AIC
<fctr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
BIC Forward Model	7	0.20063209	−2589.249	−2552.245	−2589.007
BIC Backward Model	8	0.21104354	−2598.145	−2556.516	−2597.850
BIC Backward NUMBRANCH Removed	7	0.20456580	−2592.968	−2555.965	−2592.727
BIC Backward CONTROL Removed	6	0.15239312	−2546.057	−2513.680	−2545.865
BIC Backward HSI Removed	7	0.20021036	−2588.851	−2551.848	−2588.610
BIC Backward AVGFACSAL Removed	7	0.09635033	−2496.793	−2459.790	−2496.551
BIC Backward PFTFAC Removed	7	0.20203593	−2590.574	−2553.571	−2590.333
BIC Backward PCT_BLACK Removed	7	0.19031008	−2579.575	−2542.572	−2579.333
BIC Backward PCT_ASIAN Removed	7	0.17442209	−2564.923	−2527.920	−2564.681

Appendix Figure 3. Performance of models BIC backward with NUMBRANCH removed, and 6 other models each with one additional predictor removed



Description	Number.of.Predictors	Adjusted.R.squared	AIC	BIC	Corrected.AIC
<fctr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Origional Model	7	0.2045658	−2592.968	−2555.965	−2592.727
CONTROL Removed	5	0.1352874	−2531.984	−2504.231	−2531.834
HSI Removed	6	0.1942541	−2584.246	−2551.869	−2584.054
AVGFACSAL Removed	6	0.0894489	−2492.046	−2459.668	−2491.853
PFTFAC Removed	6	0.1944859	−2584.463	−2552.085	−2584.270
PCT_BLACK Removed	6	0.1818249	−2572.704	−2540.326	−2572.511
PCT_ASIAN Removed	6	0.1675890	−2559.698	−2527.320	−2559.505