Predicting admission rates of colleges and universities in the United States

Yifan Xu

Jun 25, 2020

1. Introduction

1.1 Introduction of the study

The admission rate of colleges and universities in the USA has dramatically fluctuated, and various factors, such as school reputation, financial support, and applicant profiles, are arguably regarded as factors contributing to such fluctuation.

Divided into 3 categories: school identifiers, school characteristics, and applicant characteristics, 29 factors were collected and studied in this report to identify the factors which strongly affected the admission rate. The main goal was to fit a model based on these identified factors to know how the admission rate could be affected by these factors.

The importance of this study was twofold: 1) colleges and universities in the USA could use the model to make predictions on admission rates, and 2) reasonable actions could be taken to improve admission activities in the future according to the prediction of the model.

2. Methods

2.1 Variable Selection

There were 30 variables (1 response and 29 predictor variables) in total. For 29 predictor variables, we first removed several text variables which just described the characteristic of institutions and had no predictive effect on admission rate.

Then, we used multiple linear regression (MLR) analysis since we wanted to find the relationship between admission rate and multiple predictors. We fitted a model by MLR with all remaining predictors to see how each one contributed to the model.

Next we checked the multicollinearity of each variable by their Variance Inflation Factor (VIF) values and deleted variables with strong multicollinearity (whose VIF was greater than 10). Otherwise, they would increase the variance of estimated coefficients, which may make the interval of prediction larger and make the prediction less meaningful. Therefore, we could ensure that there was no strong correlation among the remaining predictors.

Afterwards, in order to choose one best predictor subset among the remaining predictors, we used a penalized-likelihood criteria: Bayesian Information Criterion (BIC). Since there were too many predictors, BIC would be less overfitting compared with Akaike's Information Criterion (AIC). The smaller the BIC, the better the model. We used two selection methods: Forward Selection and Backward Selection. The aim of these two methods was to find the resulting model with the lowest BIC.

If we got two different models by these two methods, we would choose the one with higher adjusted R^2 as the final model. Since the larger adjusted R^2, the better the model fitted observed values.

2.2 Model Validation

To ensure the final model provide reasonable predictions, it is necessary to divide the data set before starting to build the model, since we wanted to make sure the model obtained was not only appropriate for this particular dataset, but still accurate when predicting unknown samples.

Therefore, we randomly used 60% of the data as a training set and 40% as a test set. The training set was used in model building process to obtain the final model, while the test set was regarded as a new independent dataset.

After getting the model, we used the test dataset to fit a new model with the same predictors and expected similar results from two datasets. To be specific, we checked whether estimated coefficients and standard errors of these coefficients were similar and whether each predictor was still significant (has a strong linear relationship with the response) between two datasets.

The model would be valid if there were quite small differences in model fits between training and test datasets.

2.3 Model Violations/Diagnostics

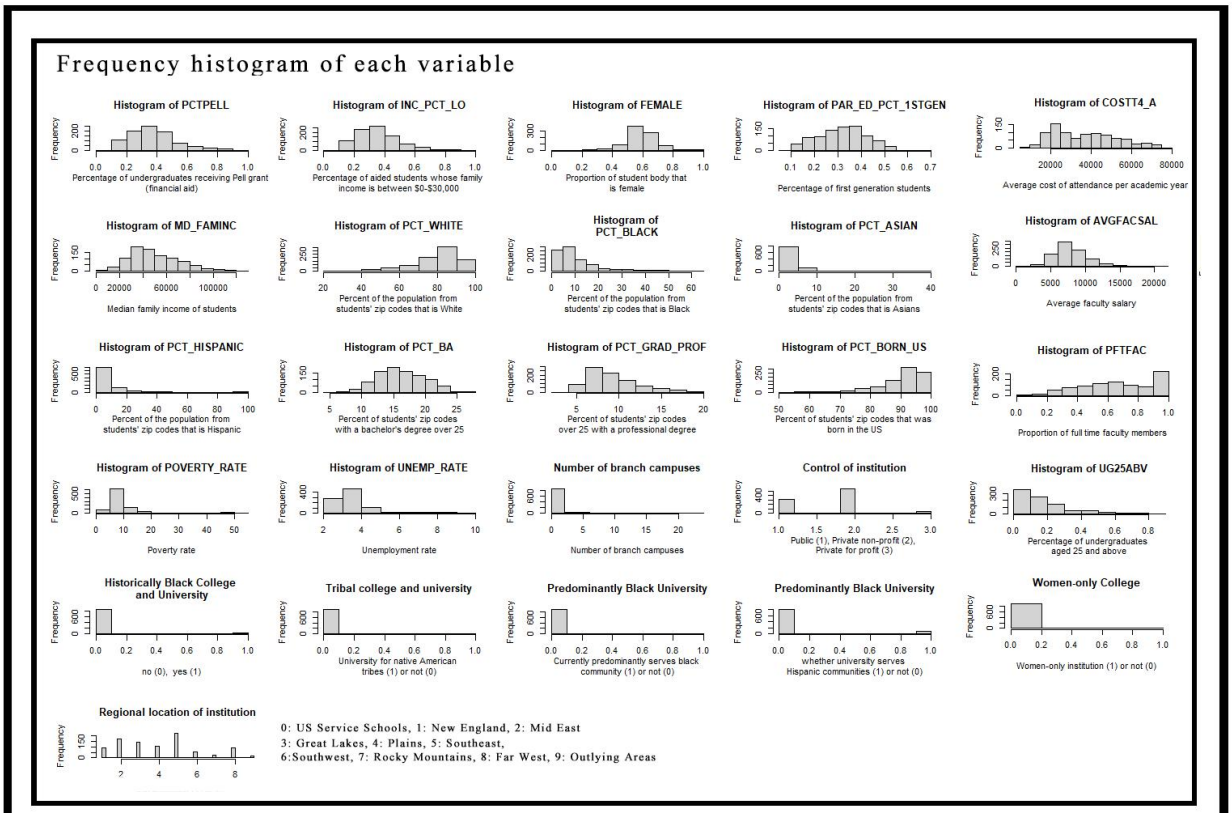
After fitting any model, we need to check 4 assumptions. Only when all assumptions are satisfied can we believe what model tells us. If there are no obvious systematic patterns or clusters in the residual (difference between the observed value and the predicted value) plot, we can conclude that common variance, linearity and independence of errors are satisfied. If Normal QQ plot indicates a one-to-one relationship between quantiles in the residuals and the standard normal, then normality of errors holds. Also, if assumptions are not violated, then there is no need to transform the data.

Next, we found out the observations which could have large influence on our model and determined whether they actually had inappropriate influence. Thus, we identified influential observations by DFFITS values. Large absolute DFFITS value (bigger than 2*square root of [p+1 divided by n], where p is the number of predictors and n is the number of observations) suggests that it has much influence on its own prediction. By comparing the original model and the model without some influential points, we could know how these points affected our model.

3. Results

3.1 Description of Data

The data was collected from 1508 colleges and universities in the USA. There were 29 predictors, which were based on different identifiers and characteristics of the school and different applicant characteristics. 3 of them were just text variables, 8 were categorical variables, and the rest were numerical variables.



Summary table of each numerical variable

	vars	n	mean	sd	median	trimmed	mad	min	max
ADM_RATE	1	904	0.67	0.20	0.69	0.68	0.19	0.05	1.00
COSTT4_A	2	904	36751.51	15474.45	35477.50	35748.02	17942.43	5810.00	75735.00
AVGFACSAL	3	904	7883.68	2426.99	7481.00	7682.32	2030.42	1236.00	20484.00
PFTFAC	4	904	0.67	0.25	0.67	0.68	0.31	0.07	1.00
PCTPELL	5	904	0.38	0.16	0.36	0.36	0.15	0.04	0.94
UG25ABV	6	904	0.17	0.16	0.13	0.14	0.12	0.00	0.86
INC_PCT_LO	7	904	0.36	0.14	0.35	0.35	0.12	0.10	0.97
PAR_ED_PCT_1STGEN	8	904	0.32	0.10	0.32	0.32	0.11	0.09	0.65
FEMALE	9	904	0.59	0.12	0.59	0.59	0.09	0.03	0.98
MD_FAMINC	10	904	50178.55	21839.32	46175.00	48609.50	20686.72	0.00	121852.50
PCT_WHITE	11	904	79.06	12.74	82.14	80.80	10.82	24.59	97.62
PCT_BLACK	12	904	11.43	10.89	7.43	9.31	5.84	0.57	63.22
PCT_ASIAN	13	904	2.98	3.15	2.01	2.38	1.36	0.14	39.04
PCT_HISPANIC	14	904	9.76	14.96	5.05	6.56	3.79	0.63	99.00
PCT_BA	15	904	16.17	3.58	15.93	16.14	3.66	4.79	27.03
PCT_GRAD_PROF	16	904	9.15	2.88	8.64	8.88	2.68	3.83	18.31
PCT_BORN_US	17	904	90.18	7.43	92.65	91.39	5.29	54.15	99.32
POVERTY_RATE	18	904	8.91	6.15	7.36	7.82	2.27	3.11	50.32

Figure 1: Information of each variable

From figure 1 above, we could see the summary for each variable and its median, mean as well as the frequency histogram. Also, plot 7 in appendix showed the scatterplot between each predictor and admission rate.

We noticed that these institutions were almost not Historically Black College and University, Predominantly Black University, Tribal college and university, Hispanic-serving institution or Women-only College. Also, most of them were private non-profit. Besides, the poverty rate, which was used to describe income levels of student's home neighbourhood, was mostly concentrated around 8%. In addition, the average faculty salary of institutions was mostly between 5000 and 12500.

3.2 Process of Obtaining Final Model

Besides five text variables (respectively the unit ID of each institution, the numeric identifier, name, control and location of each institution), we used all the remaining variables to fit a full model with the training data.

According to table 2, for variables in the orange box, their VIF values were too large, so we deleted them. Then we used the remaining predictors to get a reduced model and multicollinearity was alleviated.

	GVIF	Df
as.factor(NUMBRANCH)	1.606456	10
as.factor(HBCU)	2.279508	1
as.factor(PBI)	1.247122	1
as.factor(TRIBAL)	1.474856	1
as.factor(HSI)	2.344181	1
as.factor(WOMENONLY)	1.079917	1
PFTFAC	1.254438	1
AVGFACSAL	2.399604	1
PCTPELL	4.124481	1
UG25ABV	2.377269	1
FEMALE	1.248325	1
COSTT4_A	2.128053	1
PAR_ED_PCT_1STGEN	5.378853	1
MD_FAMINC	9.205276	1
PCT_ASIAN	6.597078	1
PCT_BA	6.732705	1
PCT_GRAD_PROF	6.390189	1
PCT_BORN_US	6.443774	1
UNEMP_RATE	7.415303	1
POVERTY_RATE	17.452240	1
PCT_HISPANIC	14.081486	1
INC_PCT_LO	13.846155	1
PCT_WHITE	23.132765	1
PCT_BLACK	18.362901	1

VIF value for each variable in full model

	GVIF	Df
as.factor(NUMBRANCH)	1.399689	10
as.factor(HBCU)	1.315008	1
as.factor(PBI)	1.053233	1
as.factor(TRIBAL)	1.025600	1
as.factor(HSI)	1.865662	1
as.factor(WOMENONLY)	1.073758	1
PFTFAC	1.245186	1
AVGFACSAL	2.429114	1
PCTPELL	3.981770	1
UG25ABV	2.274053	1
FEMALE	1.214947	1
COSTT4_A	2.175585	1
PAR_ED_PCT_1STGEN	5.068834	1
MD_FAMINC	4.709817	1
PCT_ASIAN	2.213883	1
PCT_BA	5.728496	1
PCT_GRAD_PROF	5.843037	1
PCT_BORN_US	3.651402	1
UNEMP_RATE	3.245927	1

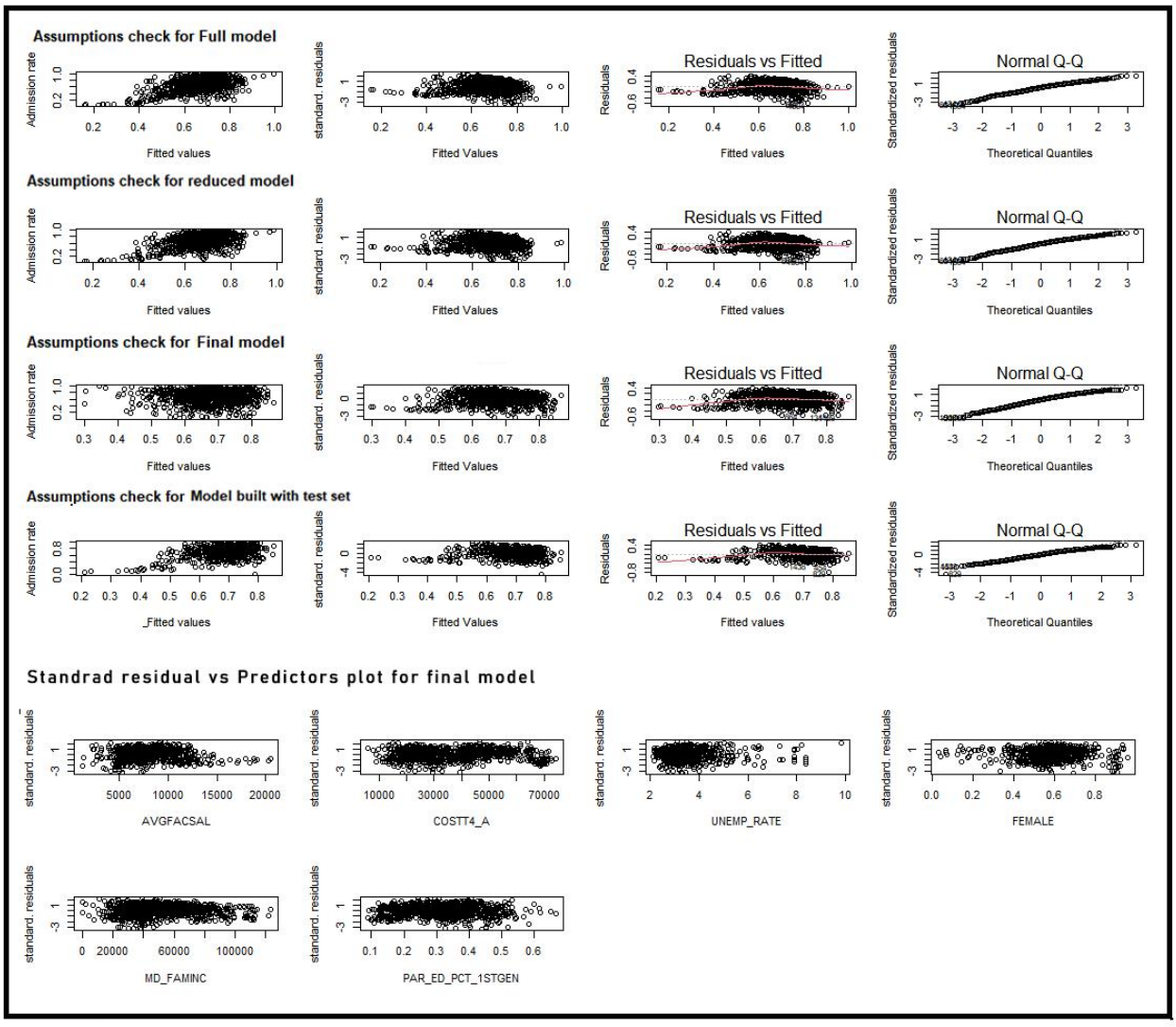
New VIF values for remaining predictors

Table 2: VIF values of predictors

Then we conducted model selection to obtain models with the lowest BIC value. For the above reduced model, the best subsets of predictors selected by forward and backward selection were both the same. Therefore, we used these predictors to obtain the final model (see table 6 in appendix).

3.3 Goodness of Final Model

The graph below illustrated that assumptions for all fitted models were satisfied. There were no obvious patterns or clusters in the residual plots, and from each Normal QQ plot, a one-to-one relationship was visible.



Plot 3: Assumptions check for all fitted models

Next, we conducted model validation. Compared to the model built with the training set, the model built with the test set just lost the significance of one variable: UNEMP_RATE. We noticed that the estimated coefficients of each predictor and residual standard errors were similar. Also, R^2 was just 3.2% lower (see table 4). Overall, the differences between 2 models were quite small, meaning that the final model had the ability to provide reasonable predictions.

	Model	R.square	Significant_predictors	Residual_SD
1	With training dataset	21.60	6	0.1765
2	With test set	17.59	5	0.1829

Coefficients:	Estimate
(Intercept)	7.971e-01
AVGFACSAL	-2.140e-05
COSTT4_A	-4.292e-06
UNEMP_RATE	-2.934e-02
MD_FAMINC	2.564e-06
PAR_ED_PCT_1STGEN	2.833e-01
FEMALE	1.337e-01

Estimated coefficients with training set	

Coefficients:	Estimate
(Intercept)	7.441e-01
AVGFACSAL	-2.605e-05
COSTT4_A	-3.254e-06
UNEMP_RATE	-1.076e-02
MD_FAMINC	2.274e-06
PAR_ED_PCT_1STGEN	2.702e-01
FEMALE	1.701e-01

With test set	

Table 4: Information comparison between 2 model fitts

For model diagnosis, the DFFITS values of 58 observations were greater than 0.176, indicating that they were influential observations.

Among these 58 observations, we found 3 with sufficiently large DFFITS values. Based on their information, there were not enough contextual reasons to remove them. Comparing the model with and without these 3 observations (see table 5), we figured out that these 2 models were similar, thus those influential observations did not change the model too much.

Information indicated by DFFITS value:

Cutoff: 0.1760
Number of influential observations: 58
Number of influential observations bigger than 2*cutoff: 3

Three strong influential observations:

	X	UNITID	INSTNM	STABBR	NUMBRANCH	CONTROL	REGION	HBCU	PBI
63	208	110486	California State University-Bakersfield	CA	1	1	8	0	0
134	508	126359	Bel-Rea Institute of Animal Technology	CO	1	3	7	0	0
1427	3784	241410	Pontifical Catholic University of Puerto Rico-Ponce	PR	3	2	9	0	0

	TRIBAL	HSI	WOMENONLY	ADM_RATE	COSTT4_A	AVGFACSAL	PFTFAC	PCTPELL	UG25ABV	INC_PCT_LO	PAR_ED_PCT_1STGEN
63	0	1	0	0.2302	16714	9585	0.5124	0.6272	0.2022	0.5266055	0.5817223
134	0	0	0	0.2415	23043	5023	0.6250	0.4971	0.4206	0.4844720	0.3586207
1427	0	1	0	0.9171	15161	4165	0.7018	0.8063	0.1019	0.7680233	0.3084023

	FEMALE	MD_FAMINC	PCT_WHITE	PCT_BLACK	PCT_ASIAN	PCT_HISPANIC	PCT_BA	PCT_GRAD_PROF	PCT_BORN_US
63	0.6426606	27270.5	60.16	7.33	4.15	38.32	10.01	4.86	80.24
134	0.8788820	30229.5	83.02	4.83	2.80	13.33	20.38	10.24	90.24
1427	0.5843023	10990.5	83.86	5.24	0.14	98.95	15.95	8.48	72.69

	POVERTY_RATE	UNEMP_RATE
63	16.39	6.60
134	6.18	2.94
1427	50.32	9.84

	Model	R.square	Significant predictors	Residual SD
1	With 3 observations	0.2160	6	0.1765
2	Without 3 observations	0.2243	6	0.1748

Coefficients:

	Estimate
(Intercept)	7.971e-01
AVGFACSAL	-2.140e-05
COSTT4_A	-4.292e-06
UNEMP_RATE	-2.934e-02
MD_FAMINC	2.564e-06
PAR_ED_PCT_1STGEN	2.833e-01
FEMALE	1.337e-01

With 3 observations	

Coefficients:

	Estimate
(Intercept)	7.909e-01
AVGFACSAL	-2.104e-05
COSTT4_A	-4.346e-06
UNEMP_RATE	-3.176e-02
MD_FAMINC	2.583e-06
PAR_ED_PCT_1STGEN	3.030e-01
FEMALE	1.462e-01

Without 3 observations	

Table 5: Information of 3 strong influential observations

4. Discussion

4.1 Final Model Interpretation and Importance

Our final model is:

ADM_RATE_estimated = 7.97e-01 - 2.14e-05***AVGFACSAL** + 1.337e-01***FEMALE** + 2.56e-06***MD_FAMINC** -2.934e-02***UNEMP_RATE** -4.292e-06***COSTT4_A** + 2.833e-01 ***PAR_ED_PCT_1STGEN** (Plot 8 shows pairwise relationships among all variables.)

From the model, the following factors: average faculty salary, the proportion of female student body and first-generation students, unemployment rate, average cost of attendance per academic year, as well as median family income of students, have strong predictive effect on the admission rate.

For each variable, the coefficient nearby is the average change in the admission rate for every additional one unit increase, when all other variables remain unchanged. By adding these information of an institution to the equation，**ADM_RATE_estimated** predicts the admission rate of that institution.

In practice, if a university or college in the US wants to increase its admission rate, it can take some measures to decrease the average cost of attendance per academic year or the average faculty salary in the future.

4.2 Limitations of Analysis

For the model built, there are two lingering problems. The first is that the adjusted R^2 is not high enough, meaning that the model only explained 21.07% of variation in responses, and thus the prediction may not as accurate as expected.

Second, when the same model was built with the test dataset, the significance of one variable was lost. Therefore, the final model was not completely independent with the training dataset, slightly affecting the accuracy of the estimated coefficients and thus prediction of our model.

References

Katherine Daignault. STA302/1001: Methods of Data Analysis 1. Week 3 Part B Materials. Slides P13.

Katherine Daignault. STA302/1001: Methods of Data Analysis 1. Week 6 Part B Materials. Slides P32.

Appendix

The analysis was done using the software R Studio.

Below are plots and tables which referred in the main sections above.

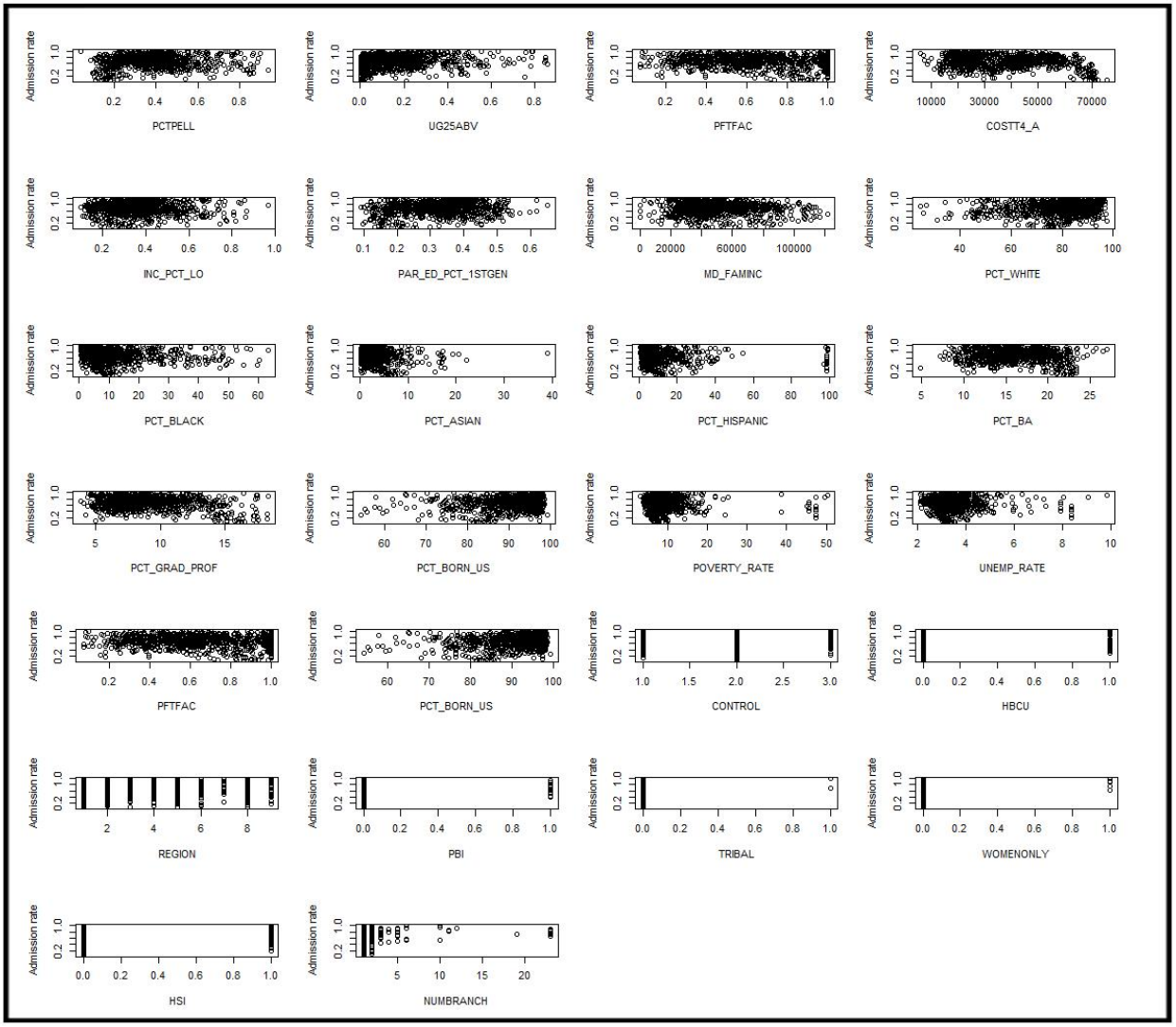
Model built by Backward Selection with the lowwest BIC:

```
Call:
lm(formula = ADM_RATE ~ AVGFACSAL + FEMALE + COSTT4_A + PAR_ED_PCT_1STGEN +
    MD_FAMINC + UNEMP_RATE, data = sample1_train)

Coefficients:
      (Intercept)          AVGFACSAL             FEMALE           COSTT4_A
        7.971e-01         -2.140e-05          1.337e-01         -4.292e-06
PAR_ED_PCT_1STGEN          MD_FAMINC         UNEMP_RATE
        2.833e-01          2.564e-06         -2.934e-02
```

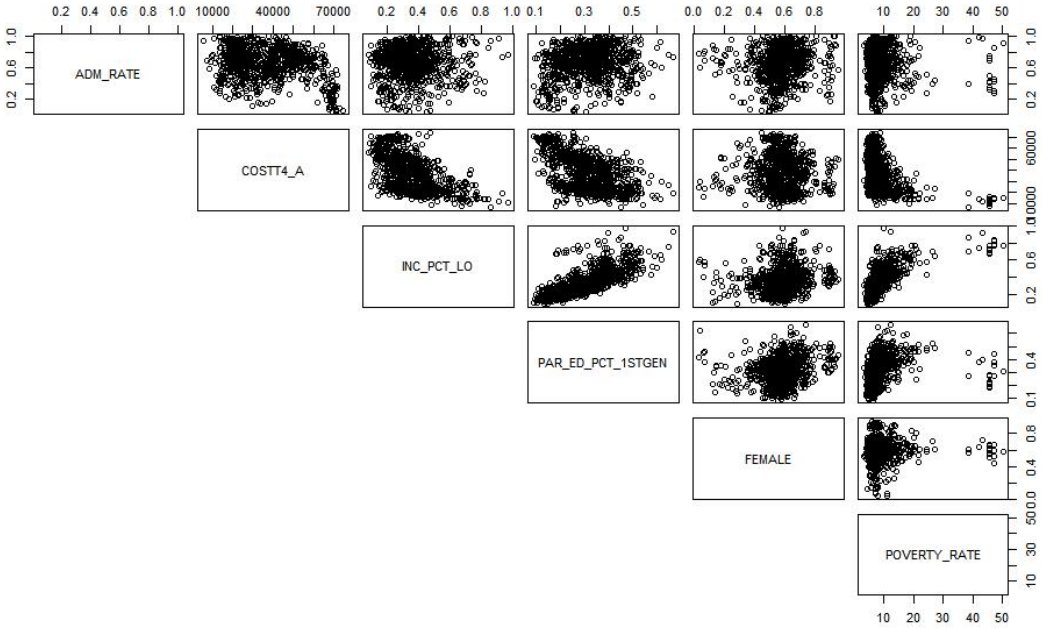
Model built by Forward Selection with the lowwest BIC:

```
Call:
lm(formula = ADM_RATE ~ AVGFACSAL + COSTT4_A + UNEMP_RATE + MD_FAMINC +
    PAR_ED_PCT_1STGEN + FEMALE, data = sample1_train)

Coefficients:
      (Intercept)          AVGFACSAL           COSTT4_A         UNEMP_RATE
        7.971e-01         -2.140e-05         -4.292e-06         -2.934e-02
        MD_FAMINC  PAR_ED_PCT_1STGEN             FEMALE
        2.564e-06          2.833e-01          1.337e-01
```

Table 6: Results of 2 selection methods



Plot 7: Scatterplot between each predictor and admission rate



Plot 8: Pairwise relationships among all variables