A Multilinear Regression Model for College Admission Rate Prediction and Understanding in the United States

Wang Zhenyu

June 2020

1 Introduction

Admission rate is an important measurement for universities and colleges in the United States. Understanding factors affecting admission rates helps the government to implement more targeted education policies towards different types of post-secondary institutions. Also, predicting admission rates by other factors helps the school to modify its admission policies more reasonably.

The goal of this study is to develop a statistical model based on the sample provided by the **COLLEGEBOARD**. This model should allow stakeholders without statistical background interpret, and let them understand factors that influence admission rates. Also, it can be used to make a reliable prediction in practice. In this study, I only use the techniques of the multilinear regression model and its relevant knowledge.

2 Methods

2.1 Initial Variable Selection

Plot the data in scatterplots to understand relationships between each predictor and the reponse. At this stage, I select a wide range of potential predictors to build the model, they include:

1. Numeric and dummy predictor variables in the sample
2. Interaction terms and Main Effect Terms by combination of original numeric and dummy variables.

I do not consider meaningless index variables. For simplicty, I transform categorical variables with k categories ($k > 2$) into $k-1$ separated binary dummy variables. I decide to add new terms by contextual reasons, features of the data (skewness), **error and trial method**. In most situations, I cannot find all important terms at one time. If I want to add new terms in later steps, then I have to come back this step, this is due to conditional nature of the regression, adding new terms changes the whole relationship.

2.2 Build Training and Testing Datasets

After selecting variables, I split the sample in half randomly into two datasets. One is for model training (2.3 - 2.6), the other is for model validation (2.7). From now on, I use the training dataset.

2.3 Model Diagnostic

First, I build a multilinear regression model based on the dataset.

2.3.1 2 Basic Conditions

1. To verify conditional mean response is a single function (g) of a linear combination of the predictors, I make a plot of the response against the fitted values. If the points are randomly scattered around the identity function, then the linearity condition is satisfied. Otherwise, if I have a reasonable idea of function g, then I modify non-linearity by applying g^{-1} to the response.

2. To verify conditional mean of each predictor is a linear function with another predictor, I make a scatterplot matrix of all the predictors, and if all pairwise relationships have a linear pattern (or at least has no evidence of a non-linear relationship), then this condition is satisfied.

If these basic conditions fail, I come back to 2.1 to re-select variables.

2.3.2 Multilinear Regression Model Assumptions

1. To check the linearity between conditional mean of the response and perdictors, I look at standardized residual plots of predictors and response, any systematic pattern in the plots, such as a curve, indicates violation of linearity. To handle this, I apply the **box-cox method** and **transformation** on variables to fix non-linearity, or **add new terms**.

2. To check the constant variance, I still use standardized residual plots. Clusters of residuals have obvious separation from the rest in the plots indicate non-constant variance of errors. To handle this, I apply variance **(stabilizing) transformation** on the response variable.

3. To verify normality, I look at the Normal Probability Plot. If the relationship between quantiles in the residuals and standard normal is not approximately one-to-one, then normality fails. To handle this, I apply the **box-cox method** and **transformation** on variables.

2.3.3 Multicollinearity

Multicollinearity of a predictor is measured by *Variance Inflation Factor* (VIF). If any variable has a VIF greater than 5, collinearity becomes problem, then I drop

redundant variables with high VIF to handle multicollinearity. I pay more attention to variables that are contextually related. Since dropping variables change nature of the regression, I go back to the beginning of 2.3 to redo all model diagnostics.

2.3.4 Outliers and Leverage Points

To find influential points, I use Cook's distance, DFFITS, methods. Due to large number of predictors, it is unrealistic to use DFBETA, since it only measures influence a data stage on a single predictor. I do not remove influential points at this point. If the existence of these points lead to a very bad result and I have a good contextual reason to remove them, I consider remove them in 2.6 - 2.7, and the removal should be discussed.

2.4 Testing Overall Sigificance

After all violations are fixed in 2.3, I use F test to determine determine whether there is a linear association between the mean response and the predictors in the model overall. If there is no significance relationship, then I go back to 2.1 and re-select other variables.

2.5 Automated Stepwise Selection

If the model passes the test, I use the stepwise Selection Method approach is inappropriate since this avoids problem caused by conditional nature of the regression. Due to the large number of predictors, it is unrealistic to use the All Possible Subsets. With this approach, I choose both AIC and BIC criteria to obtain two possible 'best' models, which gives a more comprehensive result. At this stage, go back to 2.3 to make sure all model assumptions of two new models are still satisfied.

2.6 Partial F test (Contextual Variable Selection)

Automated methods may remove variables that are important by context so I use partial F tests to determine whether those should still be included, similarly, if there is contextual reason to remove any variable at this stage, I use Partial F test to test if I can drop those predictors from the model simultaneously. Again, go back to 2.3 to make sure all model assumptions of two new models are still satisfied.

2.7 Model Comparison

There will be two candidates selected by AIC and BIC criteria. I choose one of them based on:

1. The model that has minimal differences including estimated coefficients between training dataset and testing dataset is better.
2. The model with higher adj R^2, lower AIC, BIC is better.
3. The model does not lose too much significance is better.

2.8 Model Validation

To validate the final model, I build a new model based on the testing dataset. I need to do model diagnostics (2.3) for testing models as well. If a model loses significance for many variables or has other problematic observations, then I conclude this model is invalid. To handle this, I use the similar **model diagnostic** in 2.3, and check if two data sets have too many differences, it is be helpful that testing and training datasets have similar characteristics. If it is problematic observation is not caused by datasets, then I go back to 2.1 to re-select the variables.

3 Results

3.1 Data Description

There are 1508 observations in the sample. Each observation records 30 variable of a university or college. After splitting categorical variables, there are 40 variables. Their statistical summaries are in **Table 1**, and the distributions of numeric variables are in **Figure 1**. Here are some important features,

1. NUMBRANCH variable has 3^{rd} quantile 1, indicates most schools have 1 branch in the sample.
2. UG25ABV, PCT_BLACK, PCT_ASIAN, PCT_HISPANIC, POVERTY_RATE, UNEMP_RATE variables are highly left skewed.
3. PCT_WHITE, PCT_BORN_US are heavily right skewed. (Understanding skewness is important for variable transformation).
4. COSTT4_A variable has a high variability.
5. All dummy variables have median value of 0, except CONTROL2.

6. Dummy variable CONTROL2 leads to a significant different trends of other variables. From **Figure 2**, it is obvious to see two clusters caused by it.

3.2 Process of Obtaining Final Model

3.2.1 Initial Variable Selection

Because of finding 3.1.6, interaction terms 'COSTT4_A:CONTROL_2' and 'PCT_WHITE:CONTROL_2' as well as original variables are added to initial full model.

3.2.2 Model Assumption Verification

In general, all model assumptions are verified, a few variables need power transformations to improve their performance. Transformation are recorded in **Table 2 Transformation** column.

3.2.3 Multicolinearity

Dropping a few variables with high VIF makes every rest variable's VIF approximately lower than 5. Dropped Variables are recorded in **Reason for Dropping Column**.

3.2.4 Overall Significance

At significance level of 0.05, the F test indicates there is a strong evidence to reject that there is no linear association between the predictors and the mean response overall. There is at least one regression coefficient for predictors being non-zero. ($p < 0.05$, F = 8.2, df = 34, 719)

3.2.5 Automated Stepwise Variable Selection

Result of Automated Stepwise Variable Selection is recorded in **Table 3 Variable Selection row**.

3.2.6 Contextual Variable Selection

Automated variable selection (BIC criteria) drops interaction the term PCT_WHITE:CONTROL_2 , but it has contextual importance to be in the model. At significance level of 0.5, Partial F test indicates there is strong evidence that we cannot drop PCT_WHITE:CONTROL_2 from the model selected by BIC criteria ($p < 0.05$, F = 5.4332, df = 1).

3.3 Goodness of Final Model

3.3.1 Model Comparison

See detailed numeric result in **Table3**. Model selected by BIC criteria is chosen to be the final model for the following reasons:

1. AIC selected model needs 8 more predictors than BIC ones.
2. Although AIC selected model beats BIC ones adj R^2 in terms of adj R^2, the difference is not significant.
3. The performance of AIC selected model is worse in the validation. Five predictors lose their significance in the testing dataset.

3.3.2 Final Expression

$$Y \sim \sqrt{NUMBRANCH} + \text{AVGFACSAL} + \sqrt{PAR\_ED\_PCT\_1STGEN} + \text{FEMALE} + PCT\_WHITE^4 + \text{MD\_FAMINC} + PCT\_WHITE^4 \times \text{CONTROL\_2}$$

Check **Table 3 Coefficient Significance Row** for coefficients.

3.3.3 Model Validation

First, all coefficients of models built on the training dataset and the testing dataset have the same trend and similar quantities pairwisely. Second, no predictor loses its significance in the model built on testing dataset. Thridly, although there are decrease in adj R^2 and increases in AIC and BIC, this difference is controllable, for instance, adj R^2 is still greater than 0.2. At significance level of 0.05, F test indicates there is a strong evidence to reject that there is no linear association between the predictors and the mean response overall. ($p < 0.05$, F = 30.28, df =7, 746). Model assumptions are also verified (see **APPENDIX Figure 3**), there is no serious problematic observations. Therefore, the final model is validated.

4 Discussion

4.1 Interpretation

I will discuss the model result in several aspects. First, **financial factor** is decisive on admission rate. For 10000 dollars increase in average faculty salary, we expect to see on average a 30% decrease in the admission rate when other predictors are constant. On the other hand, for 10000 dollars increase median family income of students, we expect to see on average a 2.4% increase in the admission rate when other predictors are constant. The influence of faculty salary

is counterintuitive, since we always expect schools to increase number of admitted students to balance the cost of faculty salaries.

Second, **student recourse factors** also play an important role. For 1% increase in poportion of female students, we expect to see on average 0.17% increase in the admission rate when other predictors are constant. Also, for 1% increase in white students (to the power of 4) proportion, we expect to see on average 0.3% increase in the admission rate when other predictors are constant. Lastly, for 1% increase in first generation students proportion, we expect to see on average 0.57% increase in admission rate when other predictors are constant. About half predictors are about students recourse.

The quality and **quantity of the school** are essential to the admission rate. for 1% increase in first generation students proportion, for 1 more (square root of) branch of the school, we expect to see 4% increase admission rate, when other predictors are at a constant value. Having branch schools increases the students quantity, but may decrease education quality due to decentralization of the education recourse. Schools should keep a balance between quantity and quality of education recourse to achieve an appropriate admission rate.

Lastly, admission rate of a school highly depends on the **controller** of that school. There is an 0.18% average decrease in admission rate for non-profit private controlled schools compared to other schools for a 1% proportion increase in white students. From this, the education department should take action to boost the public education system to offset the inequality in admission brought by private schools.

4.2 Limitation

4.2.1 The Existence of Influential Points

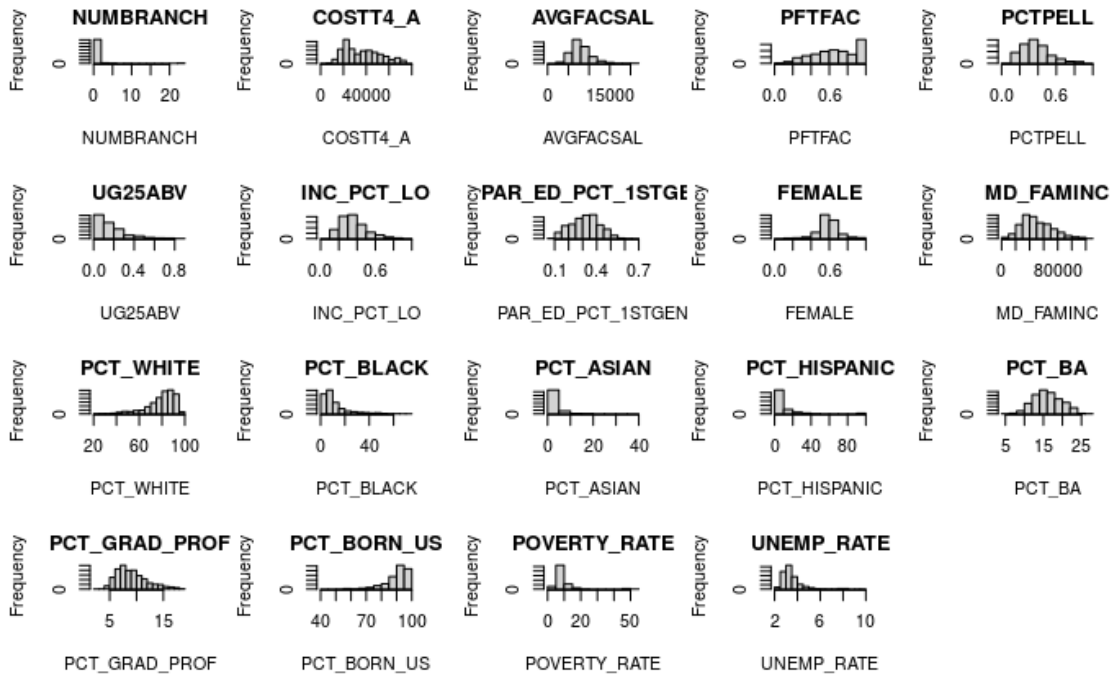
There exist some influential points (Method 2.3.4), for instance, Saint Elizabeth College of Nursing with zero admission rate. Some of them are influential to the performance of the model and eliminate similarities two datasets. But I do not have a justifiable and contextual reason to remove them based on the sample itself. In the future, it is worthwhile to implement more comprehensive research to extreme data points (schools).

4.2.2 A Trade-Off between Interpretability and Functionally

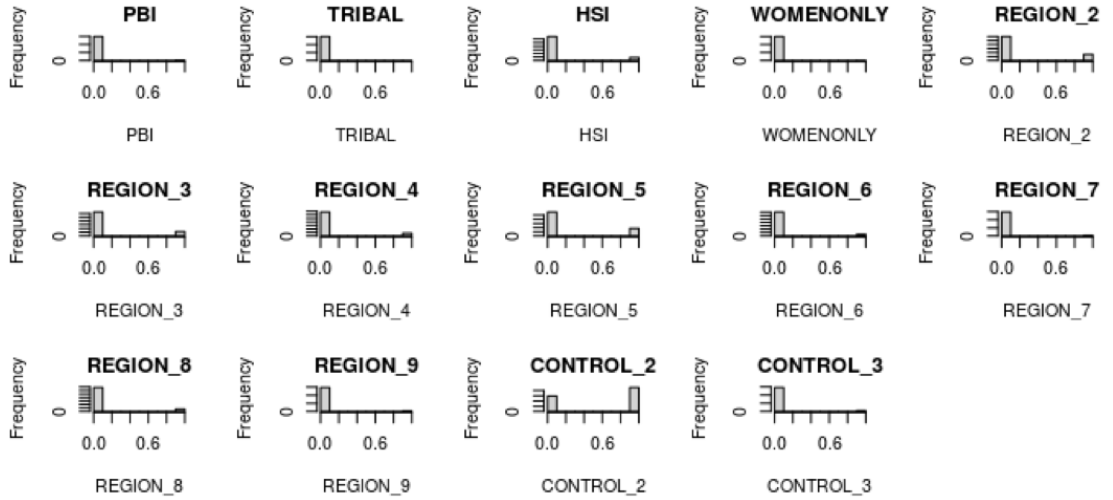
The final model is easy to understand. However, there is a trade-off between between having a interpretable model and one that is functionally more appropriate. Consider a model with orignal predictors in the final model plus an extra **log(AVGFACSAL)** term, this model's predicted values fit the actual response bet-

ter (See Appendix Figure 4). The fitted line is closer to identity line. Though we can not interpret this model properly due to the conditional nature of regression, but this model gives more precise prediction.

4.3 Tables and Figures

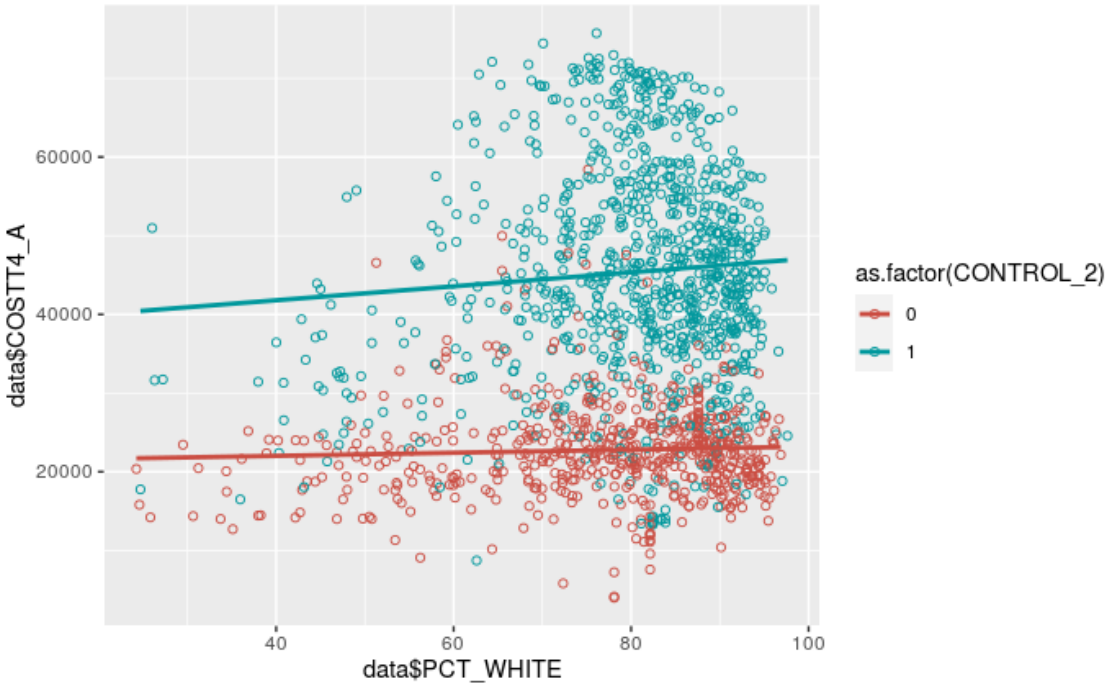


(a) Numeric Variables

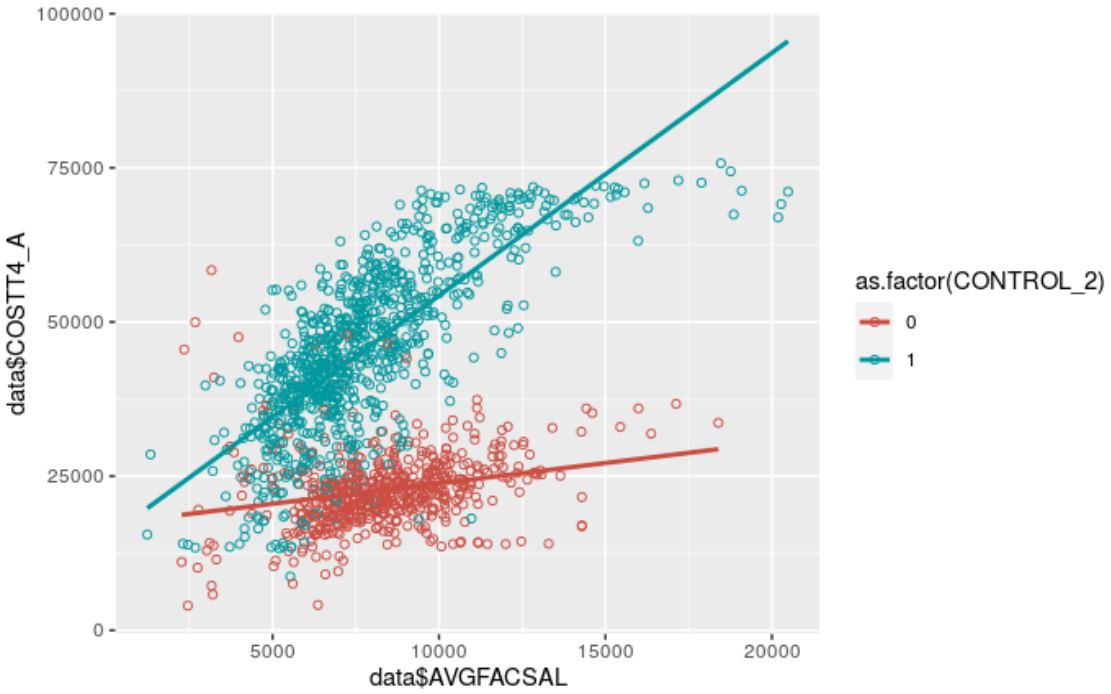


(b) Dummy Variables

Figure 1: Distributions (Histograms) of Variables



(a) PCT_WHITE vs. COSTT4_A



(b) AVGFACSAL vs. COSTT4_A

Figure 2: Grouped Scatter Plots Between Variables
(Grouped by Factor CONTROL2)

Table 1: Statistical Summary of Variables

Variable Name	Variance	Mean	Median
NUMBRANCH	7.690764	1.557692	1
COSTT4_A	248314309	36482.09	34620.5
AVGFACSAL	6046888	7976.757	7612
PFTFAC	0.060652	0.6743416	0.3568
PCTPELL	0.02556201	0.376095	0.376095
UG25ABV	0.02358009	0.1683536	0.1683536
INC_PCT_LO	0.02131717	0.3600454	0.12775
PAR_ED_PCT_1STGEN	0.01036554	0.3185573	0.3422238
FEMALE	0.01471482	0.5843872	0.5859253
MD_FAMINC	480626911	50368.61	46597.25
PCT_WHITE	171.5796	78.8967	82.075
PCT_BLACK	119.8285	11.48639	7.52
PCT_ASIAN	10.87038	3.003667	2.01
PCT_HISPANIC	262.4847	10.37873	5.075
PCT_BA	12.67531	16.10111	15.885
PCT_GRAD_PROF	8.117469	9.13303	8.67
PCT_BORN_US	62.74977	89.96862	92.55
POVERTY_RATE	42.3012	9.101041	7.395
UNEMP_RATE	1.032374	3.524542	3.28
HBCU	0.03761887	0.03912467	0
PBI	0.01567222	0.01591512	0
TRIBAL	0.00132538	0.00132626	0
HSI	0.09854411	0.1107427	0
WOMENONLY	0.006591697	0.0066313	0
REGION2	0.1626363	0.204244	0
REGION3	0.1320944	0.1564987	0
REGION4	0.09175395	0.102122	0
REGION5	0.1832368	0.2413793	0
REGION6	0.06253179	0.06697613	0
REGION7	0.02458025	0.02519894	0
REGION8	0.08535411	0.09416446	0
REGION9	0.02331824	0.02387268	0
CONTROL2	0.2386177	0.6074271	1
CONTROL3	0.03083752	0.03183024	0

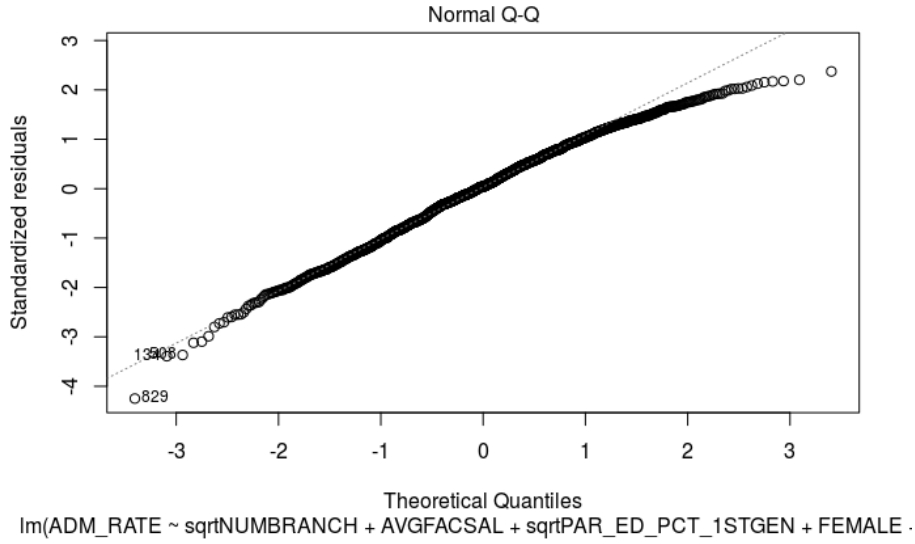
Table 2: Variable Selection

Variable Name	Transformation	In the Final Model	Reason for Dropping
UNITID	\	NO	Meaningless
INSTNM	\	NO	Meaningless
STABBR	\	NO	Meaningless
NUMBRANCH	Square Root	YES	\
CONTROL	2 Dummy Vari∼	YES	\
REGION	8 Dummy Vari∼	NO	Automated Selection, Model Comparison
HBCU	1 Dummy Vari∼	NO	Automated Selection
PBI	1 Dummy Vari∼	NO	Automated Selection
TRIBAL	1 Dummy Vari∼	NO	Automated Selection
HSI	1 Dummy Vari∼	NO	Model Comparison
WOMENONLY	1 Dummy Vari∼	NO	Automated Selection
COSTT4_A	\	NO	Model Comparison
AVGFACSAL	\	YES	\
PFTFAC	\	NO	Model Comparison
PCTPELL	\	NO	Automated Selection
UG25ABV		NO	Model Comparison
INC_PCT_LO	\	NO	Collinearity
PAR_ED_PCT_1STGEN	Square Root	YES	\
FEMALE	\	YES	\
MD_FAMINC	\	YES	
PCT_WHITE	Powered(4)	YES	\
PCT_BLACK	\	NO	Collinearity
PCT_ASIAN	\	NO	Collinearity
PCT_HISPANIC	\	NO	Collinearity
PCT_BA	\	NO	Automated Selection
PCT_GRAD_PROF	\	NO	Collinearity
PCT_BORN_US	\	NO	Model Comparison
POVERTY_RATE	\	NO	Automated Selection
UNEMP_RATE	\	NO	Model Comparison

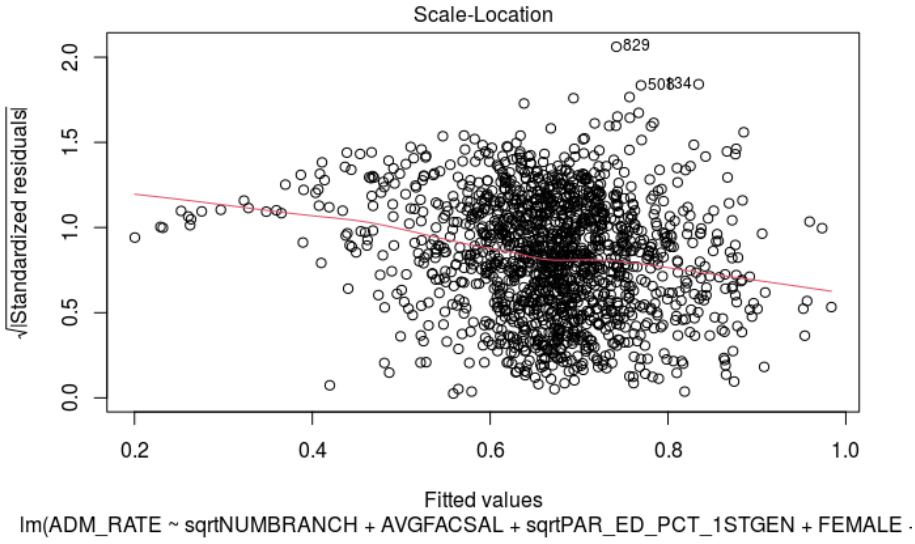
Table 3: Model Comparison

Model	AIC (Train)	AIC (Test)	BIC (Train)	BIC (Test)
Variable Selection	Intercept NUMBRANCH COSTT4_A AVGFACSAL PFTFAC PA_ED_PCT_1.. FEMALE PCT_WHITE PCT_BORN_US UNEMP_RATE POVERTY_RATE MD_FAMINC HBCU TRIBAL HSI CONTROL_2 PCT_WHITE: CON.._2	Same As AIC (Train)	Intercept NUMBRANCH AVGFACSAL PA_ED_PCT_1.. FEMALE PCT_WHITE MD_FAMINC PCT_WHITE: CON.._2	Same As BIC (Train)
Coefficient, Significance	5.6e-01 4.2e-02 -1.8e-06 -2.7e-05 -5.3e-02 3.0e-01 1.6e-01 2.3e-09 2.6e-03 4.2e-02 -8.4e-03 1.8e-06 -6.4e-02(NS) 3.6e-01 5.5e-02 1.0e-02(NS) -1.3e-09	5.6e-01 2.1e-02 -1.2e-06(NS) -2.6e-05 -5.4e-02 2.1e-01(NS) 1.3e-01 1.8e-09 9.9e-04(NS) 7.6e-03(NS) -5.7e-03 1.5e-06 -1.4e-02(NS) 3.5e-02(NS) 5.7e-02 -2.1e-02 -1.5e-09	7.9e-01 4.0e-02 -3.0e-05 5.7e-01 1.7e-01 3.1e-09 2.4e-06 -1.8e-09	3.6e-01 2.3e-02 -2.6e-05 4.6e-01 1.4e-01 2.4e-09 2.2e-06 -2.2e-09
adj R^2	0.259	0.2248	0.2396	0.2139
AIC	-2653.589	-2651.197	-2589.834	-2577.561
BIC	-2511.203	-2574.957	-2535.932	-2609.569
Note	NS stands for not significant			

5 APPENDIX

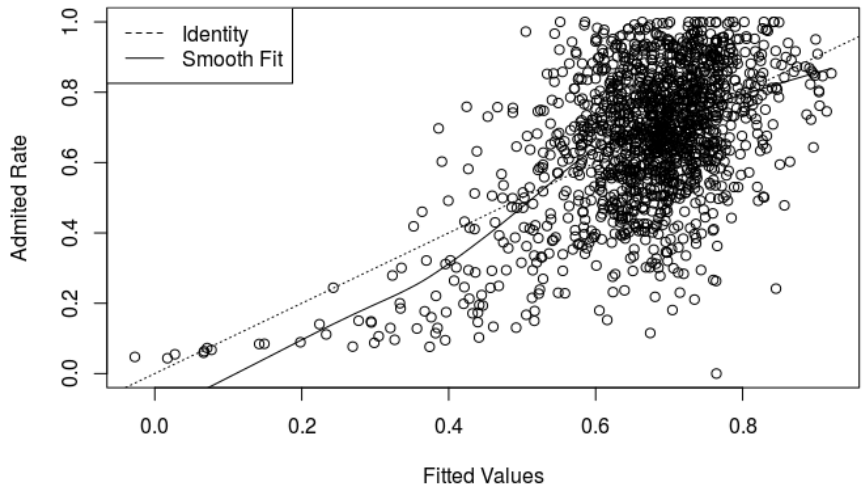


(a) Normal QQ Plot

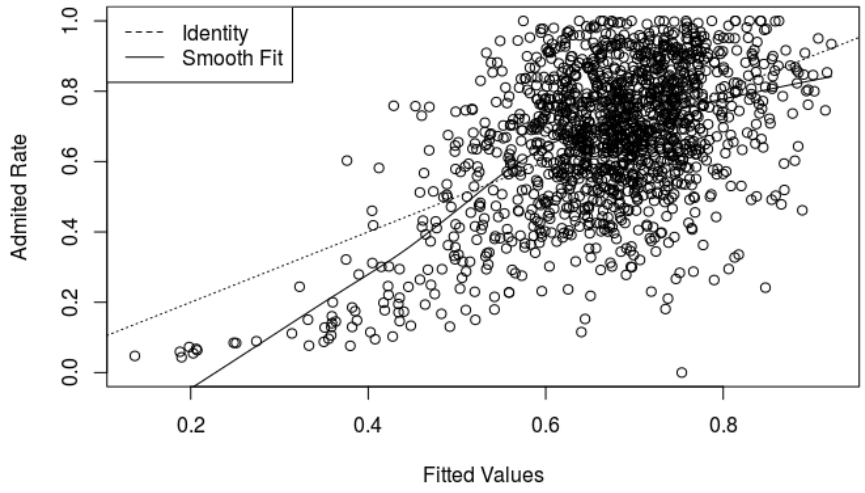


(b) Standardized Residual Plot

Figure 3: Final Model Assumption Verification



(a) With Extra log(AVGFACSAL) Term



(b) Without eEtra log(AVGFACSAL) Term

Figure 4: Fitted Value vs. Response scatterplot