Study the Students' Choice of the U.S. Higher Education

Yao Zhang

Introduction

The decision of whether going to attend college and which higher education institution students should enroll in is the most complicated and important decision for any American family. Many researches show the strong evidence about the relationships between receiving the higher education and better performances lifetime such as more successful career and higher incomes. The Department of Education released the most updated college scorecard this September to increase the transparency of college performance and guide the Americans to make more rational choices of enrolling in any U.S. higher education institutions.

I would like to explore this dataset by studying the following topic: the effect of institution features (e.g. two/four year college, private/public schools etc.) and students' demographics on students' earnings after completion. Additionally, due to the high dimensions and resourceful information this dataset provides, I would also like to explore the potential relationships between all available variables. However, I decided not to study another proposed topic "the completion rate discrepancy across the students groups with different demographics" after screening the dataset more carefully. Because the student's demographic and earnings information were collected at students level instead of institution level, while the completion rate was collected at institution level, which doesn't make sense use the data from lower hierarchical to estimate the higher hierarchical data.

Methodologies

I would like to study the after school earnings and the dataset contain two measurements of earnings: "md_earn_wne_p6" and "md_earn_wne_p6", which are representing for median earnings of students working 6 and 8 years after entry, so they could be as dependent variable in my study.

Firstly, I conducted an exploratory data analysis by checking the descriptive statistics of all variables and conducing the principal component analysis and clustering analysis. Secondly, I used variables "md_earn_wne_p6" as the dependent variable to build the regression model for it. I assume that the result for 6-year-after-school earnings will be similar to 8-year-after-school earnings, so "md_earn_wne_p6" could be used as a measurement for the after school earnings. When building the regression models, my goal is to using cross validation to find the lowest MSE. I used forward/backward stepwise and LASSO to select my independent variables, modify the linear model by using polynomial regression, splines regression, adding interaction terms or transforming the dependent variable, and check the normality and independence assumptions. Thirdly, I would like to use the graphic models to examine the covariate relationships among all variables. Finally, I also want to use other methods such as linear mixed-effect model, hierarchical generalized linear model, classification decision tree and random forest tree to further explore my dataset.

The R packages I will use include "VIM", "mice", "corrplot", "ggplot2", "reshape", "stats", "cluster", "MASS", "glmnet", "boot", "car", "quantreg", "nlme",

"hglm", "tree", and "randomForest". Each of them will be explained more detailed in the following sections of each step. All R codes could be found in the Appendix A.

Data Preparation

Data Profiling in Excel

This original dataset is extracted from CollegeScoreCard (CSC) dataset year 2005, which is a huge merged file that composed of data from different sources. The original data contains 1729 features and 6845 observations. After preliminary checking the description of 1729 features, I picked up 66 features that I am interested in with Excel. Among them, I combined "C150_4" and "C150_L4" to one column as "C150" that indicates the completion rate for first-time, full-time students. The dataset is saved as " CSC2005.csv". The description of each variable could be found in the Appendix B.

Data Cleaning in R

The dataset "CSC2005.csv" is loaded in R as the data frame "CSC2005". Due to the privacy issues, many columns have data that are "PrivacyCompressed" in the original dataset. In order to let the R handle the missing variables correctly, I replaced strings such as "NULL" and "PrivacyCompressed" with "NA". Then I checked the class for each column and convert "factor" to "numeric" for columns containing analyzable information. After looking at the percentage of missing value for each variable, I decided to delete features "SAT_AVG_ALL" and "ADM_RATE_ALL" since they have over 50% of data are missing. Additionally, I removed the "C150_4" and "C150_4L" as the "C150" contains the information of these two variables.

As it was mentioned in the previous section, I would like to build regression model on variable "md_earn_wne_p6". However, based on the result of missing value checking, unfortunately, there are about 20% of "md_earn_wne_p6" data are missing. In order to make sure the analysis valid, I created a new dataset "6yearearning2005.csv" by removing the observations with missing values of dependent variable "md_earn_wne_p6" and imputing missing values of other independent variables, and another new dataset "8yearearning2005.csv" was created by same approach.

Exploratory Data Analysis

Descriptive Statistics

The dataset "6yearearning2005" has 5427 observations and 60 variables, except for "INSTNM", "CITY" and "STABBR", all variables are either numeric or integer. As the the boxplot of variables and descriptive statistic summary of dataset shows (Figure 1), all variables have different scales, so I scaled the data to make sure they have same means. Additionally, some variables such as "PCIP_**"have a lot of large outliers (Figure 2).

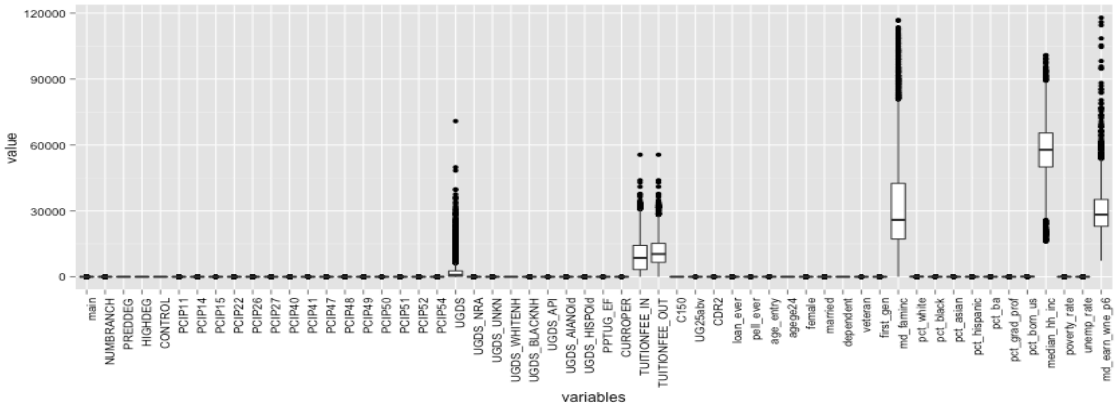
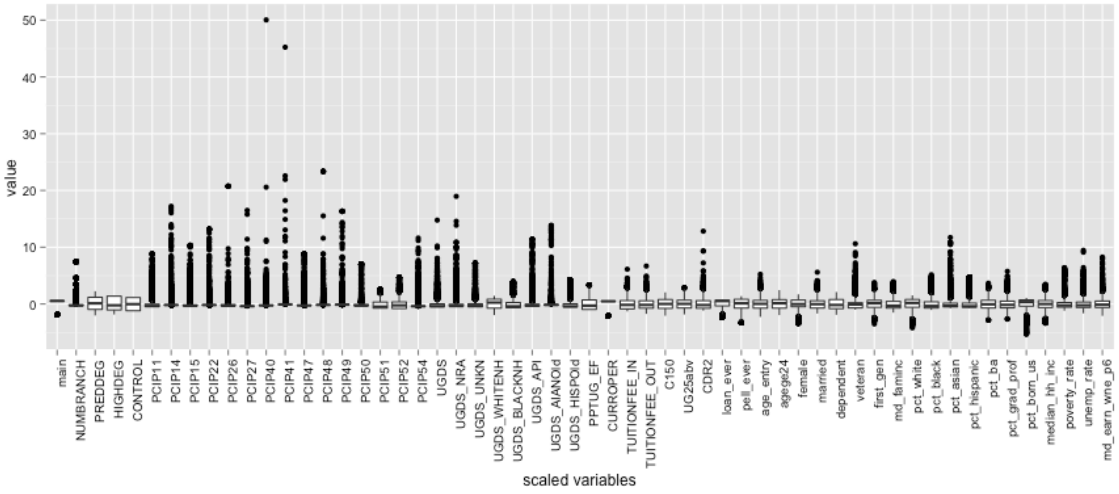


Figure 1



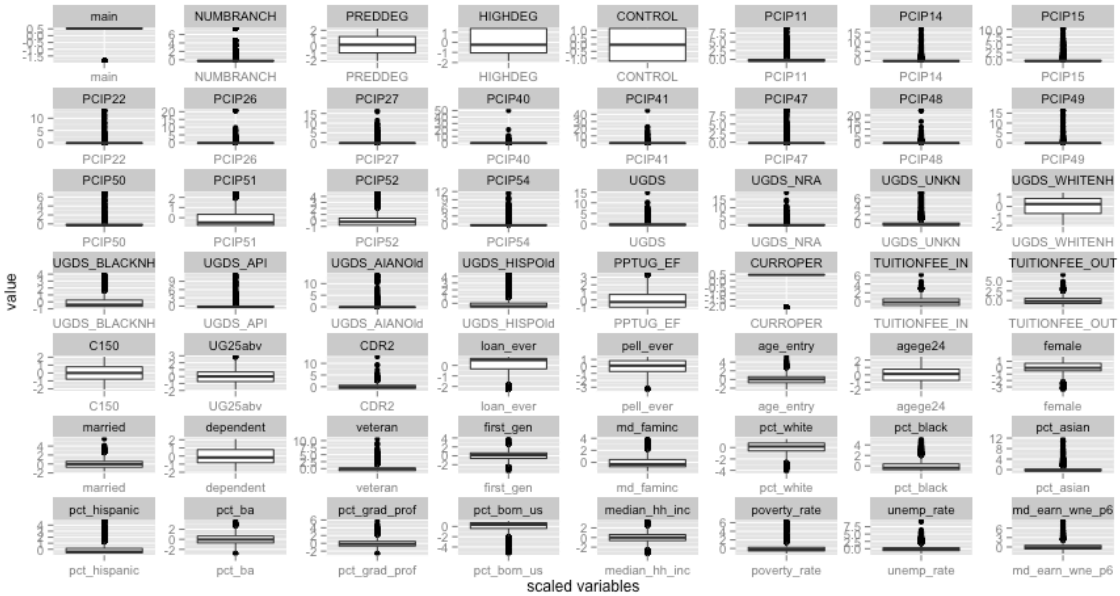


Figure 2

By visualizing the distribution of all variables, we could see that all “PCIP_**” variables, “UGDS_**” except for “UGDS_WHITENH”, “pct_*” except for “pct_ba” and “pct_born_us” variables all have very right skewed distribution. Variables such as “C150”, “UGS25abv”, “first_gen” and “pct_ba” are almost nearly normal distributed (Figure 3).

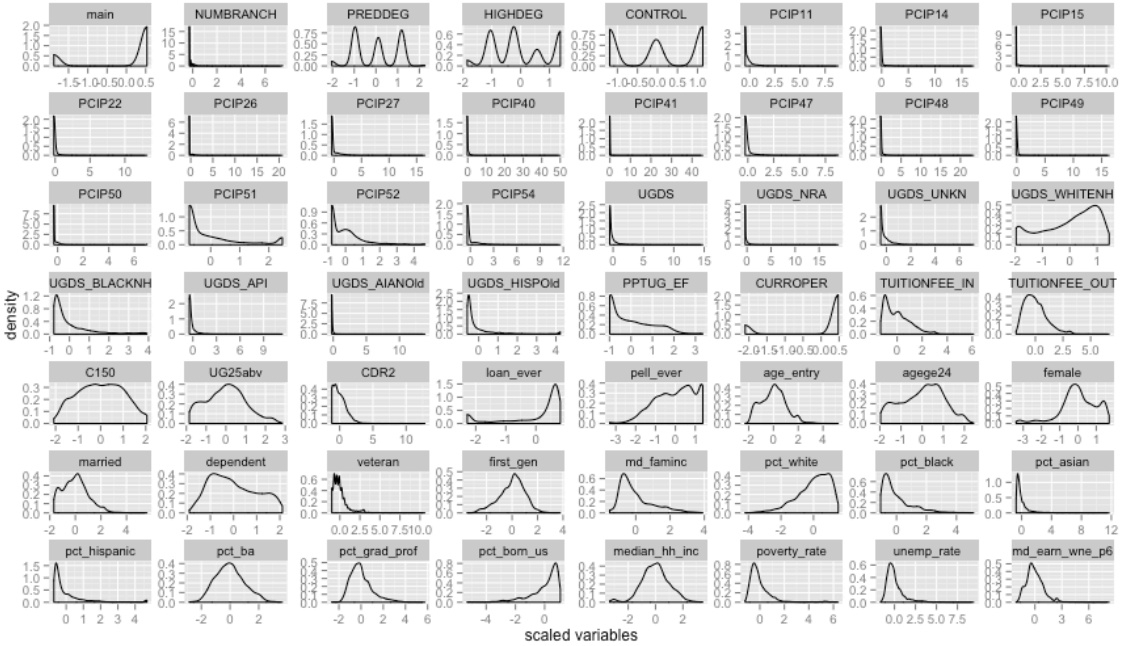


Figure 3

The corrplot provides an initial idea about the correlation between these variables. The graph tells that “NUMBRANCH” and “main”, “PREDDEG” and “HIGHDEG”, “UGDS_WHITENH” and “pct_white”, “UGDS_BLACKNH” and “pct_black”, “UGDS_API” and “pct_asian”, “UGDS_HISPOld” and “pct_hispanic”, “TUITION_IN” and “TUITION_OUT”, “loan_ever” and “pell_ever”, “UG25abv” and “dependent”, “age_entry” and “agege25”, “first_gen” and “md_faminc”, “pct_ba” and “pct_grad_prof”, “poverty_rate” and “unemp_rate” are highly correlated (Figure 4).

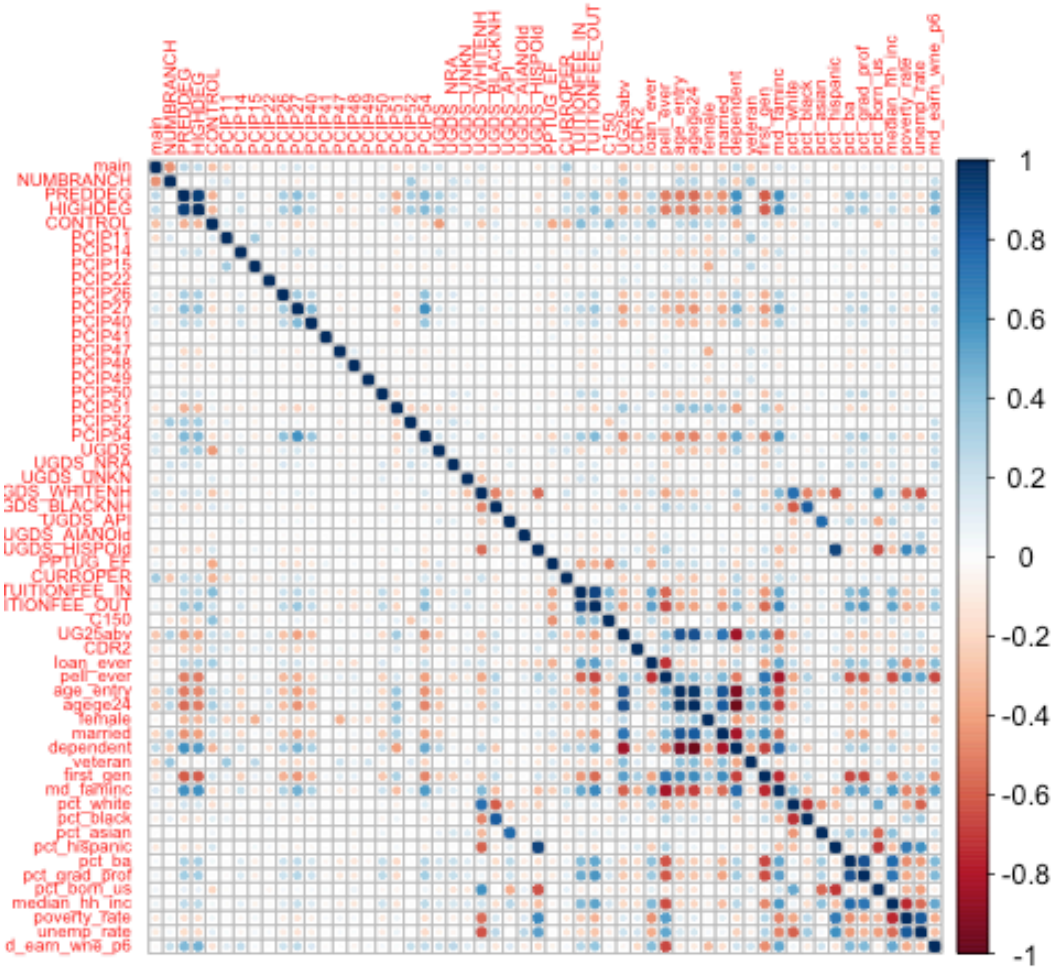


Figure 4

PCA and Clustering

I used principal component analysis to summarize all the variables, using as few as possible components. The scree plot showed us below the first component explains 21% variation and the second component only explains about 7% of the variations (Figure 5).

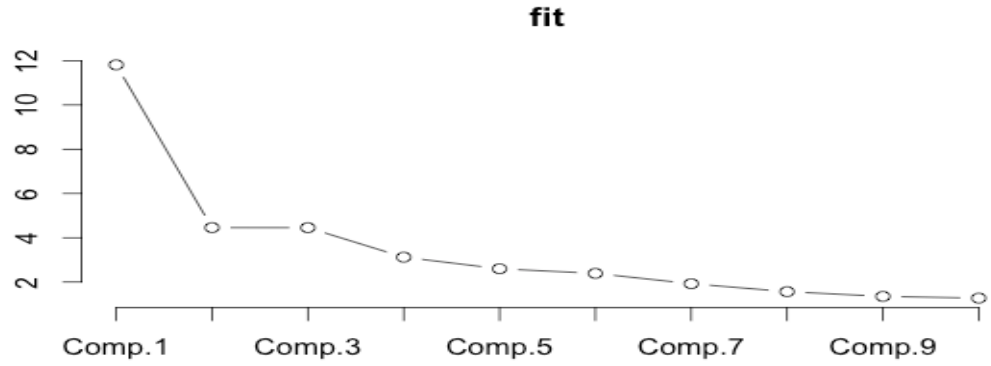


Figure 5

According to the loadings results, variables are loading averagely (between 0.1 and 0.2) on the first component, “PREDDEG”, “HIGHDEG”, “PCIP26”, “PCIP27”, “PCIP40”, “PCIP51”, “PCIP54”, “UGDS_WHITENH”, “TUITION_IN”, “TUITION_OUT”, “UG25abv”, “CDR2”, “loan_ever”, “pell_ever”, “age_entry”, “agege24”, “married”, “dependent”, “first_gen”, “md_faminc”, “pct_white”, “pct_ba”, “pct_grad_prof”, “pct_born_us”, “median_hh_inc”, “poverty_rate”, “unemp_rate”, “md_earn_wne_p6” have relatively high correlation with the first principle component. I also used “biplot” to lie all original variables up on the first two components (Figure 6, left). Because of too many variables and observations, the plot is a little bit messy, but we still could see variables such as “md_faminc” and “dependent” load very strongly on the first component, which as the loading scores proved, too. We also could find that the institution 11508303 displayed as an outlier on the second component, and the scores of this institution on the second component is -8.22440438, which explains the plot.

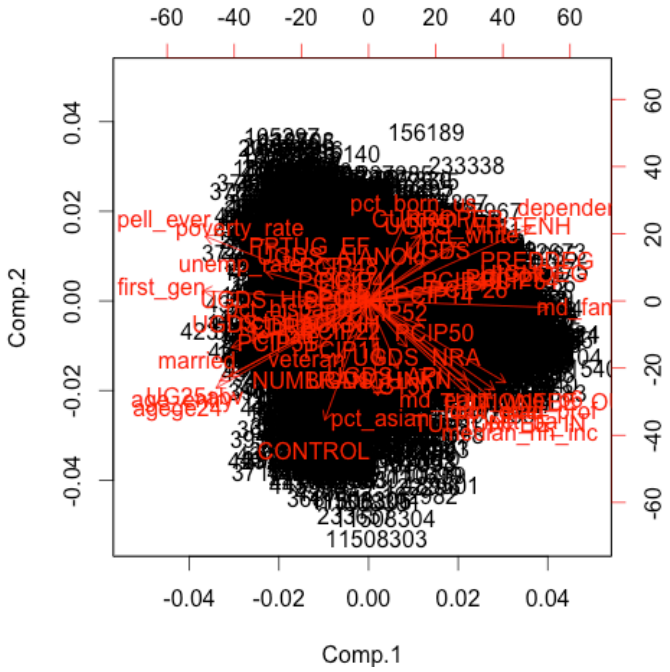


Figure 6

For clustering, I picked up the first ten institutions to see the pattern. In the complete linkage and single linkage cluster, institutions 100663 and 100830 are grouped together, which indicates these two schools are similar on the aspects variables provided. The institution 100690 is separated from other 9 institutions (Figure 7).

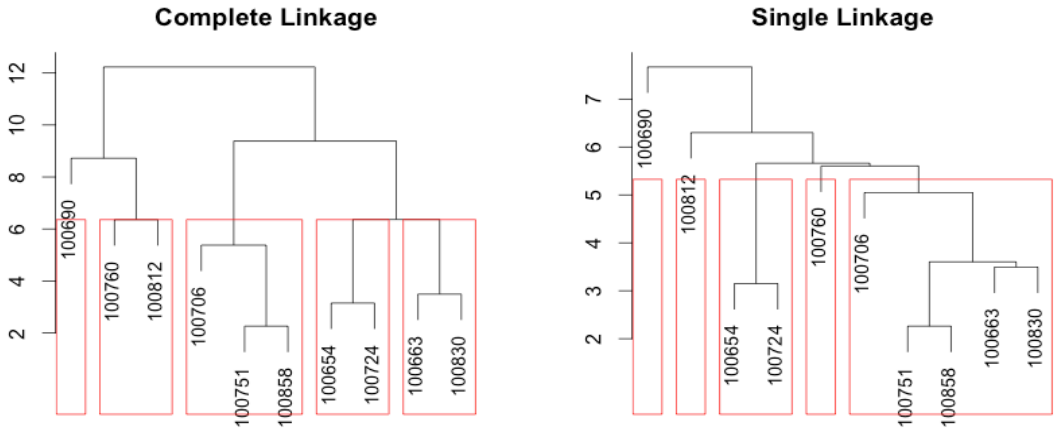


Figure 7

In the k-k clustering analysis, I used three clusters; each of them includes 2773, 1067, and 1587 observations respectively. According to the plot below, only 29.05% of the point variability is explained by the two components (Figure 8). This is not a good clustering, since there are too many overlaps. For example, I checked the "CONTROL" among three clusters, one of the "main" value have 782, 286, and 130 in three clusters.

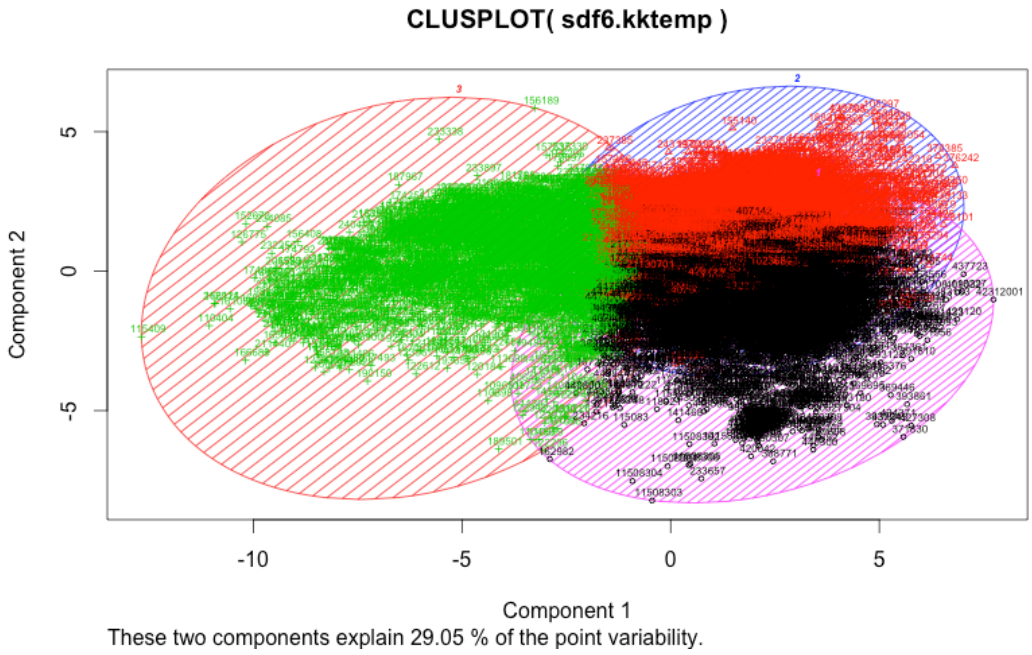


Figure 8

Linear Regression

Since I'm interested in predicting "md_earn_wne_p6", I plotted bivariate graphs including "md_earn_wne_p6". According to the graph, there is no obvious transformation needed for the model, but it seems that "UGDS_NRA" might have non-linearity relationship with "md_earn_wne_p6" (Figure 9).

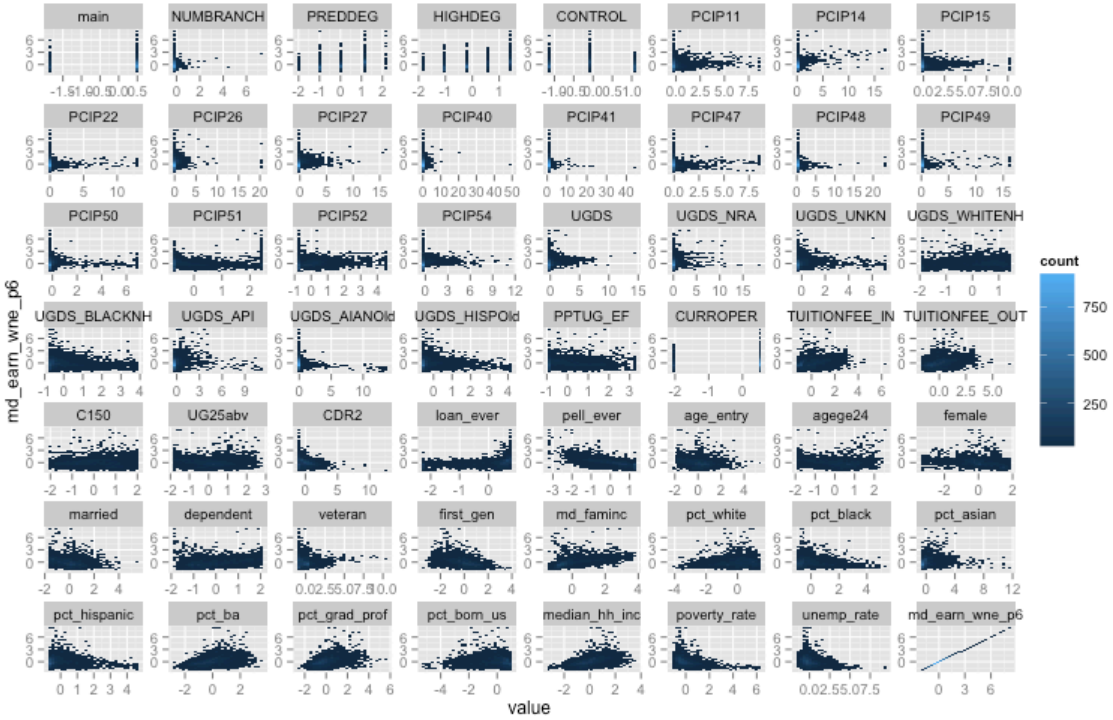


Figure 9

Variable Selection

I used forward/backward stepwise and LASSO to select variables, and also used the k-fold (k=5) cross validation to compare which result can obtain the lowest MSE. The model obtained by the forward stepwise selection includes 43 variables; backward stepwise selection gave a model of 45 variables; LASSO model used all independent variables to explain the dependent variable. However, the cross validation MSE of three models didn't vary a lot from each at all. Therefore, I decided to use those 43 variables from the forward stepwise selection for model modification later. The temporary linear model is below:

md_earn_wne_p6 ~ pell_ever + veteran + pct_born_us + CONTROL + PCIP51 + NUMBRANCH + PCIP14 + pct_white + PCIP52 + PCIP11 + dependent + female + HIGHDEG + median_hh_inc + PCIP47 + PCIP50 + UGDS_NRA + md_faminc + age_entry + agege24 + PCIP22 + pct_ba + UGDS_API + PREDDEG + UGDS_AIANOld + PCIP49 + PCIP15 + PCIP27 + UGDS + UG25abv + CDR2 + UGDS_UNKN + UGDS_BLACKNH + married + main + C150 + PCIP26 + loan_ever + PCIP40 + first_gen + poverty_rate + PCIP48 + pct_hispanic

And the cross validation MSE of this model is about 0.2588.

Model Modification

I plotted the residuals with respect to each covariate for checking if any nonlinear transformations are necessary. According to the graph, most of the variables have linear relationships with response. There might be some non-linearity with respect to UGDS_NRA and PCIP26 variables. Thus, I tried to use polynomial regression to see whether I can obtain the even lower MSE. The lowest MSE around 0.2546 is achieved at the 2nd degree of polynomial for PCIP26, but 1st degree of UGDS_NRA, which means that only the PCIP26 needs to be transformed. Comparing with the linear model, the polynomial regression improved the MSE performance just a little bit (from 0.2588 to 0.2571), which does not appear to be significant enough. Then I used the splines for variable PCIP26 to find if other transformation could bring lower cross validation MSE, but MSE I obtains is 0.2605, which is even higher than the one I got from the polynomial regression model. As it was mentioned a earlier in the exploratory analysis section, variables "NUMBRANCH" and "main", "loan_ever" and "pell_ever", "UG25abv" and "dependent", "age_entry" and "agege25", "first_gen" and "md_faminc" are highly correlated, so I added the interaction terms successively. Eventually, the model with interaction terms of "NUMBRANCH:main" and "first_gen:md_faminc" achieved lower MSE of 0.2501. I saved the model as "interlm":

md_earn_wne_p6 ~ pell_ever + veteran + pct_born_us + CONTROL +
PCIP51 + NUMBRANCH + PCIP14 + pct_white + PCIP52 + PCIP11 +
dependent + female + HIGHDEG + median_hh_inc + PCIP47 + PCIP50 +
UGDS_NRA + md_faminc + age_entry + agege24 + PCIP22 + pct_ba +
UGDS_API + PREDDEG + UGDS_AIANOld + PCIP49 + PCIP15 + PCIP27 +
UGDS + UG25abv + CDR2 + UGDS_UNKN + UGDS_BLACKNH + married +
main + C150 + PCIP26 + loan_ever + PCIP40 + first_gen + poverty_rate +
PCIP48 + pct_hispanic - PCIP26 + poly(PCIP26,2) + NUMBRANCH:main +
first_gen:md_faminc

I plotted the residual Q-Q plot for the model "interlm", as it can be seen from the qqplot, some of the neighborhoods' errors are off by 6-14 sigma. After removing 9 institutions that have a significant influence on the coefficients with cook's distance, the residual Q-Q plot still has a very heavy tail (Figure 10.1).

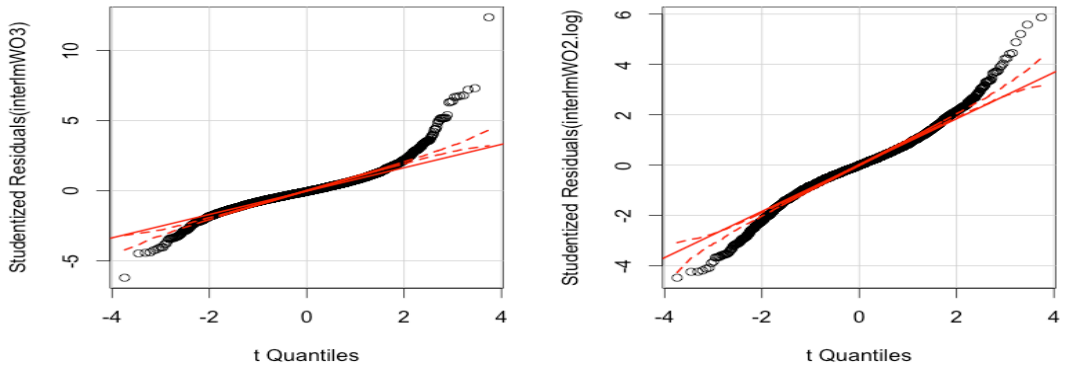


Figure 10.1 Figure 10.2

Therefore, I decided to transform the response "md_wne_p6" as log(md_wne_p6) to explore a more convincible model. The cross validation MSE of the model that uses log(md_wen_p6) achieves even lower as 0.2498, and the residual Q-Q plot looks much better than the one before, but still has heavy tails even after removing 6 outliers (Figure 10.2). Additionally, the p-value of K-S normality test (sample size is larger than 5000) is 0.007681 (< 0.05), so the residual of model doesn't have a very good normal distribution.

Therefore, I decided to use the quantile regression based on the above model. Firstly, I tried the 0.25 quantile and 0.75 quantile, and use ANOVA to see whether the quantile regression is necessary. The test result shows that a quantile regression is needed since the p-value is very small. Then a quantile regression from 0.05 to 0.95 quantile was conducted and the coefficients was plotted (Figure 11). We can see that coefficients of variables such "pell_ever", "pct_born_us", "CONTROL", and "C150" vary a lot from quantile 0.05 to 0.95. I used the K-S to test the normality assumption again and the p-value is 0.0375, which is significant at 99% confidence level. Therefore, the quantile regression works better.

Figure 11

Other Methods

Undirected Graphs by "huge" package

Undirected graphs describe the conditional independence among many variables. Each node of the graph represents a single variable and no edge between two variables implies that they are conditional independent given all other variables. Firstly, I applied the nonparanormal transformation to the data since the dataset fails the normality assumption as it was discussed above. Then the graph is estimated using the default methods "mb" with lambda of 0.05, 0.4, and 0.95.

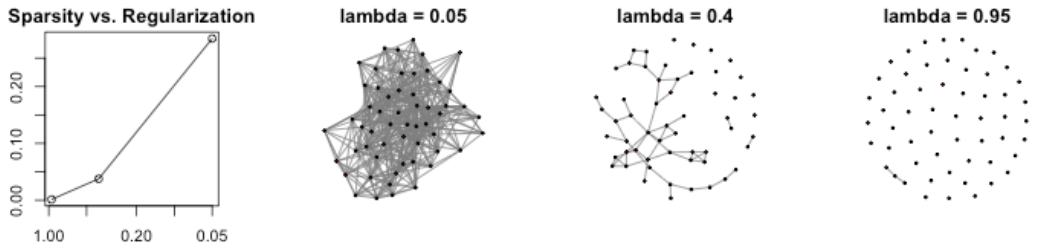


Figure 12

The graph becomes more sparsely as the penalty parameter lambda is increased. As it shows in the plot (Figure 12), as the lambda increases, the number of edges between nodes decreases, which indicates that fewer variables are conditional independent given to all other variables as the penalty parameters increase.

Linear Mixed-Effects Model (LMM) by "nlme" package

Because the linear model and the generalize linear model assume there are only fixed effects in the model, but actually I'm not sure about the reality for this dataset, I tried the LMM and regarded the variables "main" and "NUMBRANCH" as the random effects. I downloaded the "lmm" package and used "lme" function for the previous linear model. There are two levels for "main" variable and 20 levels for "NUMBRANCH" variable. In the output of the model, the description of random effect is basically the measure of variance at different levels of "main" and "NUMBRANCH" expressed as standard deviations. According to the result, there was quite a bit variance between institutions are/not main campus (StdDev= 0.1088082) and number of branches the institution has (StdDev= 0.2518706), and there was also quite a bit residual variance (StdDev= 0.4528748) between other variables that were used in the model. After printing out the coefficients, I found that every level of "main" and "NUMBRACH" have different estimated intercepts, while the linear model without random effects always have the constant intercepts.

Hierarchical Generalized Linear Model (HGLM) by "hglm" package

"hglm" is used to fit hierarchical generalized linear models. It can be used for linear mixed models and generalized linear models with random effects for a variety of links and a variety of distributions for both the outcomes and the random effects. In this case, since the residuals are not normal distributed, the HGLM would be a more valid

approach. “log(md_earn_wne_6)” was set as the response Y, “main” and “NUMBRANCH” were treated as random effects Z1 and Z2 that are highly correlated, other variables except for “main” and “NUMBRACH” were viewed as fixed effects. The summary of the model includes “the summary of the fixed effects estimates”, which indicated the significant variables and the coefficients of them. “The summary of the random effects estimates” provides the information of random effects including coefficients and standard errors. The model has two random effects; there are two dispersion parameters 0.009724 and 0.108083, so we had larger variance component for the “NUMBRANCH” correlation than the “main”.

Classification Decision Tree by “tree” package

Since I would like to predict the earnings after school – variable “md_earn_wne_6” based on other variables, I changed that variable to categorical variable (“High”). Additionally, I used the non-scaled dataset, since it is easier to interpret the node of the tree. I split the earnings value to Yes/No based on whether the value is greater than 50000 or not, and attached the High variable to the original dataset. In order to evaluate how good the tree, I used random sampling to split the dataset to training data and testing data, and fitted the tree model using training data, obtaining the below tree model (Figure 13). In order to check how the model is doing, I used testing data to predict with the tree model and calculated the misclassification error. The error here is just around 1.91%, which is very low. This tree model is a pretty good classification tree to tell which attending which institution could bring, and what types of students tend to have higher earnings after school.

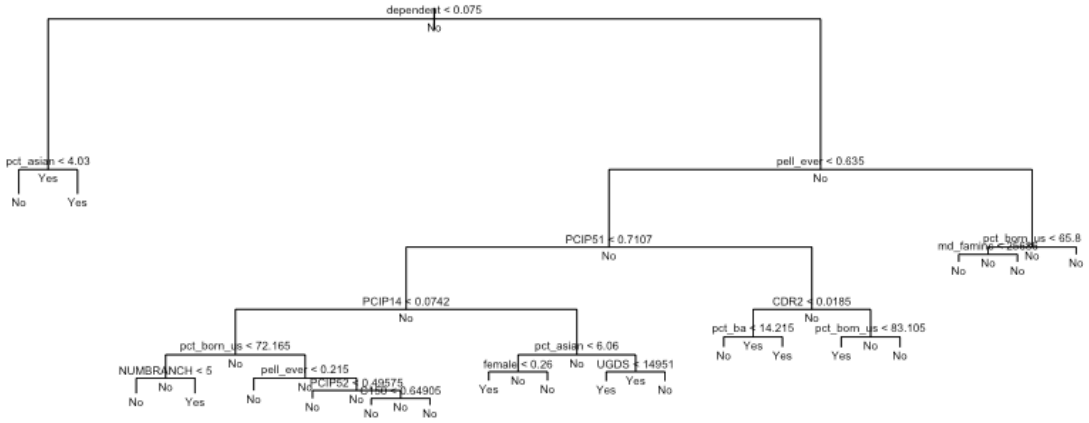


Figure 13

Random Forest by “randomForest” package

This package implements Breiman's random forest algorithm for classification and regression. I used the non-scaled data for this method, too. 500 trees were grown and 4 variables are randomly sampled at each split. One of the model outputs indicated that variables “PREDDEG”, “HIGHDEG”, “PCIP51”, “UGDS”, “UDGS_NRA”, “UGDS_API”, “TUITION_OUT”, “UG25abv”, “CDR2”, “pell_ever”, “loan_ever”, “age_entry”, “agege24”, “female”, “dependent”, “veteran”, “first_gen”, “md_faminc”,

"pct_asian", "pct_ba", "pct_grad_prof", "median_hh_inc", and "poverty_rate" are significantly important, which has similar patter with the linear regression variable selection result.

Limitation and Potential Further Analysis

Firstly, too many missing values in the dataset decrease the data quality. The imputed missing values may cause the bias. Secondly, many variables are not just simply normal distributed, so the assumption for linear model failed. Thirdly, the earnings data were collected from students who participated financial aid programs; in other words, they are not really randomly selected from all students of each institution. Therefore, the earnings data may not represent the overall after school earnings for each institution. Thus, the above models could only be used as a reference for school selection, not an absolute principle.

Further analysis could be conducted in more complicated ways. For example, we could separate schools from different states and conduct the analysis spatially, or divide schools in different layers based on some criteria then conduct regression methods at different layers. Additionally, because there is another variable "md_earn_wne_8" in the original dataset, which could be viewed as another response for future analysis. A comparison between models with similar but not identical earnings response could be a good justification for model selection and decisions on school choosing.

Data Source:
https://collegescorecard.ed.gov/data/

References:

Better Information for Better College Choice & Institutional Performance retrieved from https://collegescorecard.ed.gov/assets/BetterInformationForBetterCollegeChoiceAndInstitutionalPerformance.pdf

Chapter 17 Undirected Graphic Models *The Elements of Statistical Learning.* Second Edition

https://cran.r-project.org/web/packages/huge/huge.pdf

Breiman L., Friedman J. H., Olshen R. A., and Stone, C. J. (1984) *Classification and Regression Trees.* Wadsworth.

Manual On Setting Up, Using, And Understanding Random Forests V3.1
https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf.

Lars Ronnegard, Xia Shen and Moudud Alam (2010). **hglm: A Package for Fitting Hierarchical Generalized Linear Models**. *The R Journal*, **2**(2), 20-28.

Package 'nlme'
https://cran.r-project.org/web/packages/nlme/nlme.pdf