Tu Le
Chloe Morgan
Gabriel Santana
Natalia Troncoso
Ching-Chi Tu

Table of Contents

Introduction

Education has always played an invaluable role in societal development. As Eleanor Roosevelt once famously said: “Education is essential to good citizenship and it is important to life because it enables people to contribute to their community and their country”. It gives people the skills and tools they need to learn and navigate the world. These skills include the ability to read, write and communicate, complete tasks and work with others. As time passes, the world as we know it becomes even more competitive, making access to higher education indispensable. The United States is home to the finest universities in the world; unfortunately, many people around the world are unable to access such significant investment due to the burden of expensive tuition and fees. The College Board recently reported that a “moderate” college budget for an in-state public university for the 2016-2017 academic year averaged $24,610 and a private university averaged $49,320. The question is: what goes into these costs? With the tuition and fees of American universities more than doubling over the past two decades, it is crucial we determine the driving forces of this drastic trend.

Problem Statement

A research analysis on potential factors that influence the tuition cost for medium to large sized universities in the United States. The potential factors or variables investigated in this project include the following: total applicants, total admissions, total students enrolled, percent of Freshmen students submitting SAT and ACT scores, SAT and ACT composite 25th and 75th percentile scores, total percent of students admitted, admission yields, part-time enrollment, full-time enrollment, undergraduate enrollment, graduate enrollment, graduation rate within four years, percent of freshmen receiving any financial aid, endowment assets, highest degree offered, geographic region, control of institution, degree of urbanization and historically black college/university. The objective of this project is to identify the most influential factors in predicting the tuition and fees of a specific university based on the aforementioned criteria via statistical tests and analyses. With these tests and analyses, we look to answer the following questions: Which of these twenty-three variables are significant and truly affect the tuition price per university? If so, how and why do they have such influence? By answering these big questions, we look to guide and inform incoming freshmen to pick their future home effectively and economically.

Data Description

The data for this research was taken from the resource section of the Tableau website. Please refer to the References section for the complete list of sources used to develop this project. The original data set taken from the Tableau website contained data points for approximately fifteen hundred (1500) universities in the United States with one hundred forty-five (145) variables all from the year 2013-2014. These data points were manually entered into an excel spreadsheet and then converted into a CSV file to import into R and Minitab. Due to a large volume of observational units and variables, the first step in our data preparation process was to narrow down the number of variables by choosing those that truly affect the tuition and fees of mid to large sized universities. To maintain the accuracy of our regression, the next step was to further reduce the scope of the analysis by reducing the number of observational values. The following steps were taken to reduce the number of observational values:

1. Removed unidentified observational values with missing information in the selected variables.
2. Ordered the observational values by the total number of students enrolled in 2013-2014.
3. Selected the top five hundred universities based on the number of students enrolled in 2013-2014 to narrow down the analysis to just medium to large sized universities.

A summary of the twenty-three variables (including Tuition and fees) investigated in the project is discussed as follows:

I. Dependent Variable:

1. ***Tuition and fees (tuition):*** Cost of tuition and fees in dollars of mid to large universities.

II. Quantitative Variables:

1. ***Total applicants (appTotal):*** Total number of freshmen that applied to a mid-size university. Reasonable values for this value are in the range (0, ∞).
2. ***Total admissions (adminTotal):*** Total number of potential freshmen admissions per university in the United States. Reasonable values for this variable are in the range of (0, ∞).

3. ***Total enrolled (enrollTotal):*** Number of incoming freshmen enrolled in the university. Reasonable values for this variable are in the range of (0, ∞).
4. ***% Freshmen submitting SAT scores (percentSAT):*** Percent of incoming freshmen that used SAT scores when applying to universities. Reasonable values for this variable are in the range of [0.00, 1.00]
5. ***% Freshmen submitting ACT scores (percentACT):*** Percent of incoming freshmen that used ACT scores when applying to universities. Reasonable values for this variable are in the range of [0.00, 1.00].
6. ***SAT Composite 25th percentile score (SAT25):*** Sum of incoming freshmen's Reading and Math scores in the top 25th percentile. Reasonable values for this variable are in the range of [400, 800].
7. ***SAT Composite 75th percentile score (SAT75):*** Sum of the Reading and Math scores in the top 75th percentile. Reasonable values for this variable are in the range of [400, 800].
8. ***ACT Composite 25th percentile score (ACT25):*** Incoming freshmen ACT scores in the top 25th percentile. Reasonable values for this variable are in the range of [0, 36].
9. ***ACT Composite 75th percentile score (ACT75):*** Incoming freshmen ACT scores in the top 25th percentile. Reasonable values for this variable are in the range of [0, 36].
10. ***Total percent admitted (percentAdmit):*** Total percent of freshmen admissions accepted per university. Reasonable values for this variable are in the range of [0.00, 1.00].
11. ***Admissions yield – total (adYield):*** Percent of students who choose to enroll in a college or university after being admitted.
12. ***Full-time enrollment (FTEnroll):*** Number of full-time enrolled students per university in the United States. Reasonable values are in the range of (0, ∞).
13. ***Part-time enrollment (PTEnroll):*** Number of part-time enrolled students per university in the United States. Reasonable values are in the range of (0, ∞).
14. ***Undergraduate enrollment (UNEnroll):*** Total number of enrolled undergraduate students in a specific university. Reasonable values are in the range of of (0, ∞).
15. ***Graduate enrollment (GDEnroll):*** Total number of enrolled graduate students in a specific university. Reasonable values are in the range of (0, ∞).
16. ***Graduation rate - Bachelor degree within 4 years (GDRate):*** This variable was calculated by dividing the number of students graduating by the total number of students

enrolled that year. It is important to specify that this rate represents the rate for students earning a bachelor degree within four years. Reasonable values are in the range of [0.00, 1.00].

17. ***% of Freshmen receiving any financial aid (freshmenAid):*** Percent of freshmen earning any type of financial aid while studying in a mid-size university. Reasonable values are in the range of [0.00, 1.00].
18. ***Endowment assets (en)****:* Amount of endowment assets in dollars donated to universities by ex-alumnus. Reasonable values are in the range of [0, ∞).

III. Qualitative Variables:

1. ***Highest degree offered (degree):*** This categorical variable measures the highest degree offered for mid-size universities. It can be classified as:
    a. Bachelor's Degree (B)
    b. Doctor's degree-professional practice (DP)
    c. Doctor's degree-other (DO)
    d. Doctor's degree-research/scholarship (DR)
    e. Doctor's degree-research/scholarship and professional practice (DRP)
    f. Master's Degree (M)
2. ***Geographic region (geo):*** This variable demonstrates the geographic region in which the university is located. It can be classified as:
    a. Far West (FW)
    b. Great Lakes (GL)
    c. Mid-East (ME)
    d. New England (NE)
    e. Plains (P)
    f. Rocky Mountains (RM)
    g. Southeast (SE)
    h. Southwest (SW)
3. ***Control of institution (CoI):*** This is a binary variable that records whether the university is either a private or public institution. It can be classified as: Public (0) or Private not for profit (1).

4. ***Historically Black College or University (histoBlack):*** This binary variable indicates whether the mid-size university contained only black students or not. It can be classified as: No (0) or Yes (1).
5. ***Degree of urbanization (DoU):*** Categorical variable indicating the degree of urbanization per university in the United States. This variable can be classified as: City (C), Rural (R), Suburb (S), or Town (T).

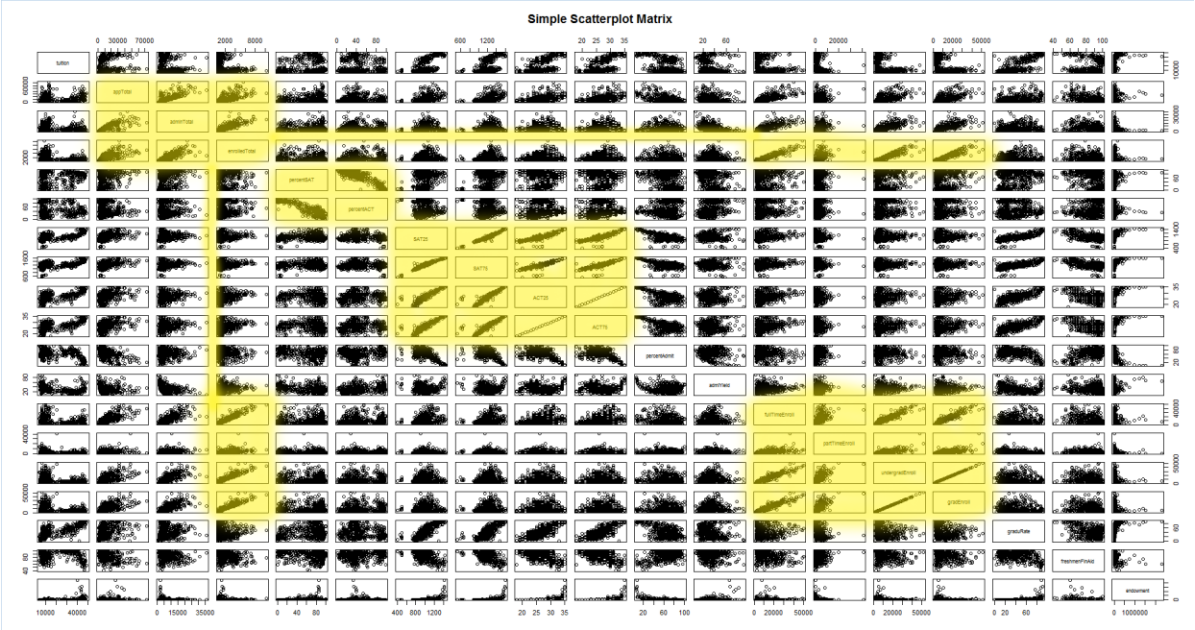
Model Building and Analysis

I. First-Order Model:

After collecting information regarding tuition costs and fees for a vast amount of mid to large sized universities in the United States, our data was then classified and organized in a way where we could analyze the importance of several factors and how these could affect the tuition and fees of universities. Through multiple replications using Minitab Software and R, we wish to be able to accurately predict and forecast the tuition and fees of a university given other relative information and an alpha of 0.05. For the first-order model, we formulated a stepwise regression of all twenty-three variables and found that twelve of them are indeed significant (see Appendix A). We noticed that it did not select a few variables that showed a clear correlation with tuition and fees (e.g. SAT25 and SAT75). Additionally, a few of the variables such as FTEnroll and percentACT have very high Variance Inflation Factor (VIFs) with values of 8.92 and 7.67 respectively. These high values give a sign for multicollinearity; therefore, variable interactions must be considered. In the section below, we will look at these relations and apply methods to reduce or eliminate multicollinearity.

I.I Fixing Multicollinearity

To visually see the multicollinearity among these different variables, we created a Matrix Plot (Figure 1) with all the variables combined. The ones highlighted in yellow show clear signs of multicollinearity or dependency due to the fact that the they demonstrate a linear relationship. If these variables were to be kept unaltered, the model would be inaccurate and we risk failing to get a solution since we want to deal with independent variables.

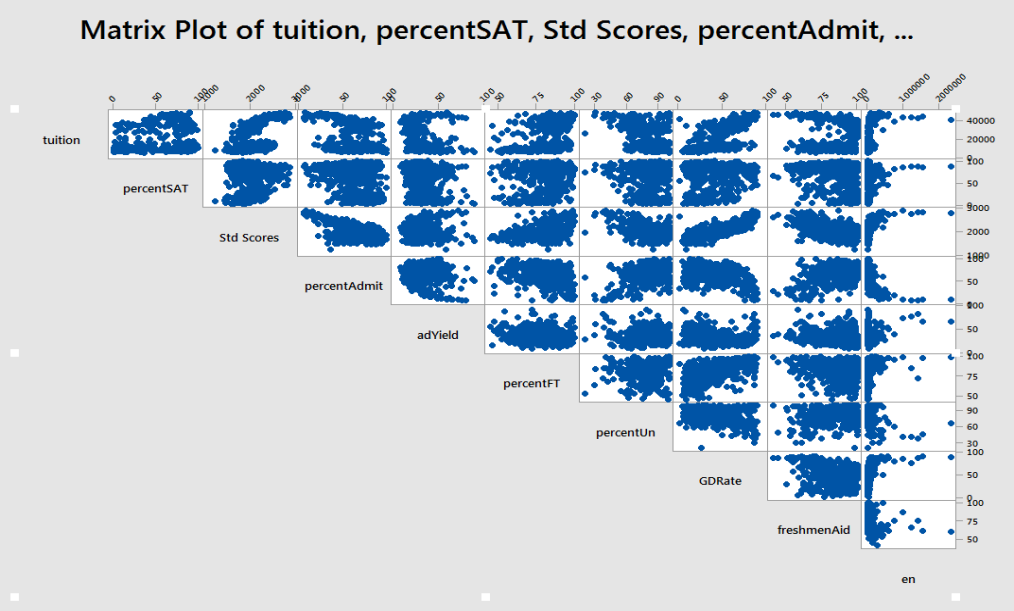


The following variables were identified to have a high VIF: appTotal, adminTotal, enrollTotal, percentACT, SAT25, SAT75, ACT25, ACT75, FTEnrolled, PTEnrolled, UNEnroll, GDEnroll. To eliminate the dependency among these variables we performed the following steps:

1. Deleted the following variables: appTotal, adminTotal, and enrollTotal since they will be covered by already existing percentAdmin and by adYield.
2. Deleted the percentACT column and kept the percentSAT column as these two were highly correlated. The decision process of choosing to keep percentSAT over percentACT was arbitrary and as we develop the model, we will keep this deleted variable in mind.
3. Converted the ACT scores to SAT equivalents (Appendix 3) and summed it with the SAT scores. After converting adding them up, we created a new variable named Standardized Scores (stdScores) which represented this union. Furthermore, since this new variable covers the 25th percentile SAT and ACT scores, we removed these variables. Additionally, we removed the 75th percentile of SAT and ACT scores as they are highly dependent on the 25th percentile scores.
4. Added a new variable called Percent full time (percentFT) which replaces the full-time and part-time variables named FTEnrolled and PTEnrolled respectively. Moreover, we

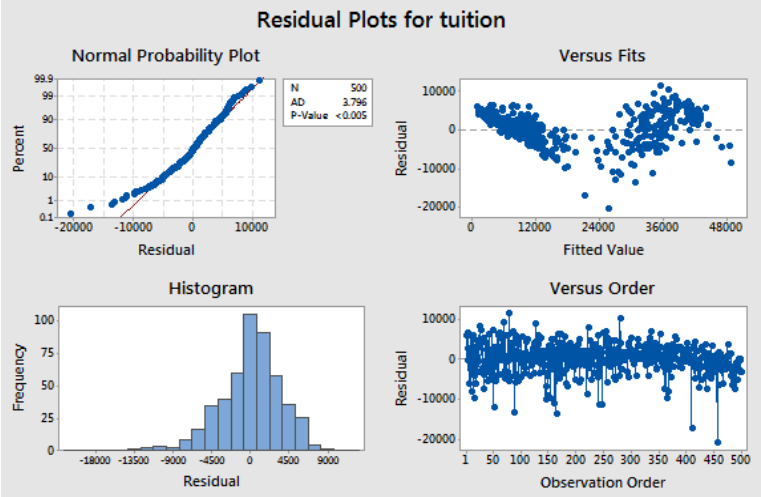
made another variable called Percent Undergrad (percentUN) which takes the place of undergraduate and graduate enrollment named UNEnroll and GDEnroll respectively.

After fixing the dependency among our independent variables, we created a second matrix plot with the new nineteen variables and saw a significant reduction in multicollinearity as seen in Figure 2 below. We then proceeded to perform another stepwise regression using the updated variables which showed a minimal and non-significant reduction in the adjusted R^2. This is a result of the trade-off for lower VIF values (Appendix C).



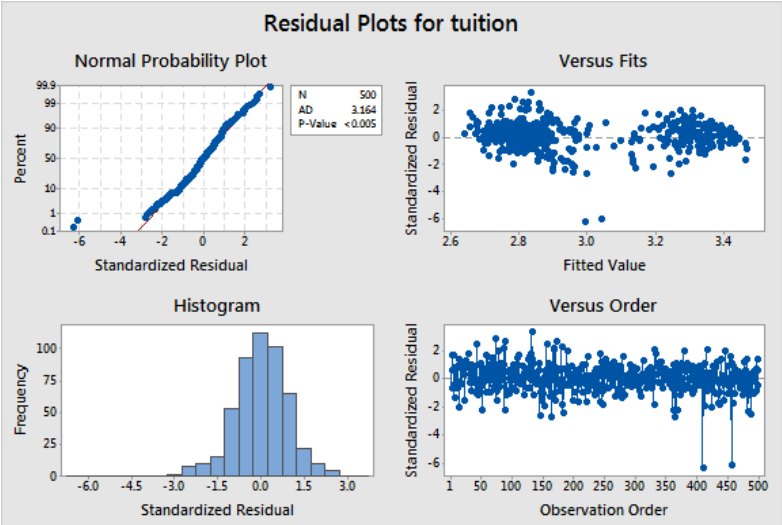
I.II Residual Analysis

Out of the fourteen variables (nine quantitative and five qualitative variables), the stepwise regression identified nine variables to be reasonably significant. These variables are: percentSAT, StdScores, adYield, percentUn, GDRate, CoI, DoI, en, and geo. Now that we have selected the most significant variables and reduced multicollinearity, we proceeded to perform the residual analysis (Figure 3). In the residual analysis, the Normal Probability Plot in the top left quadrant shows a potential violation of normality. This violation was confirmed by the Anderson-Darling results next to the plot which displays a p-value of < 0.005, which is less than any alpha (α).



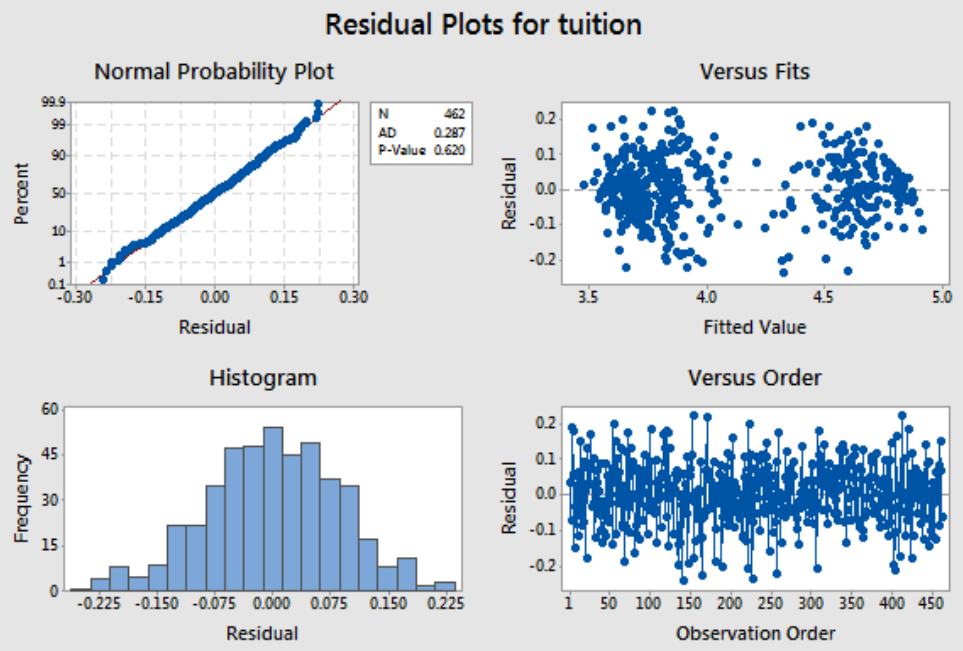
I.II.I *Y* transformation

To fix this normality violation, we carried out the optimal transformation of the tuition and fees. The result that minitab provided was a Box-Cox transformation with lambda (ƛ) = 0.114732. Unfortunately, the transformation was not enough to fix the violation as can be seen in the Residual Analysis below (Figure 4). The p-value remains < 0.005 and this raises suspicion that the lack of normality is being caused by unusual observations. Refer to unusual observations for further details.



I.III Pitfalls

After running an unusual observation analysis, we identified 38 observations that are potentially affecting the normality of our distribution (see Appendix D). After the removal of these unusual observations, we applied a Box-Cox transformation with $\lambda = 0.146789$. The normality of the distribution was restored, as reflected by Anderson-Darling test that identified a p-value of 0.620 which is greater than any alpha (α) value. The normality of the distribution is further reinforced by the histogram in the bottom left quadrant of the Residual Analysis since it resembles an almost perfect bell curve with expected value at zero.



II: Second-Order Model

We know that by applying a Box-Cox transformation and deleting the unusual observations, the normality violation is fixed. However, the residual errors plot doesn't look random. Now, we will go ahead and run the stepwise on the transformed tuition variable with all the nine quantitative and five qualitative significant variables selected after reducing multicollinearity (see Appendix B). We will also include interaction terms of the three most significant variables, which have the highest t-values, with the rest of the variables selected by the stepwise regression run on the fourteen updated variables (see Appendix C). Finally, we will include variables StdScores squared and freshmenAid squared as they seem to have an squared interaction with tuition (see Appendix B).

Figure 3: Shows result for running these interaction and quadratic terms.

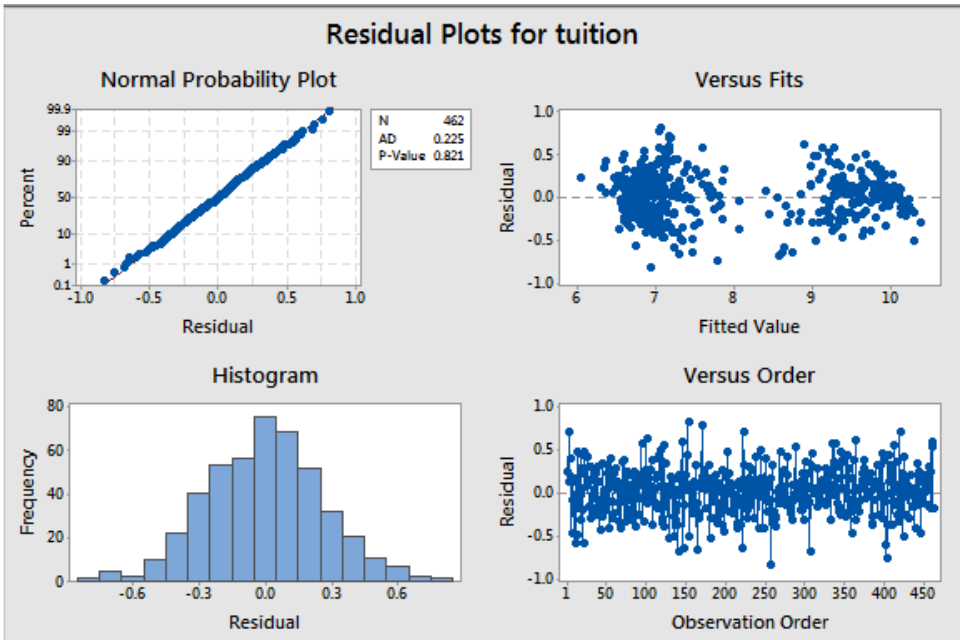
```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
3231.31  94.87%     94.78%      94.43%


Coefficients

Term                   Coef   SE Coef  T-Value  P-Value   VIF
Constant              10843      1425     7.61    0.000
percentUn             -60.6      14.5    -4.17    0.000  1.65
freshmenAid2        -0.1603    0.0989    -1.62    0.106  1.68
percentSAT*en      0.000193  0.000031     6.21    0.000  2.08
stdScores*GDRate    0.10999   0.00734    14.99    0.000  8.64
adYield*percentUN     0.390     0.221     1.76    0.078  2.86
adYield*GDRate       -3.718     0.474    -7.84    0.000  7.39
CoI
  1                   19979       492    40.64    0.000  2.48
histoBlack
  1                   -2053       717    -2.86    0.004  1.08


Regression Equation

CoI  histoBlack
0    0           tuition = 10843 - 60.6 percentUn - 0.1603 freshmenAid2
                           + 0.000193 percentSAT*en + 0.10999 stdScores*GDRate
                           + 0.390 adYield*percentUN - 3.718 adYield*GDRate
```

We now proceed to run a summary of the suggested best variables in Figure 3 above, but we first need to include the remaining variables needed to keep the principle of hierarchy. The prediction equation now includes all these terms plus its lower order versions (see Appendix J).

II.I Residual Analysis

A Residual Analysis was created in order to confirm that error is indeed random (Figure 5). Even though normality assumption still holds, the interaction and quadratic terms added to the model failed to give us a random error plot. Therefore, we decided to reduce the data set as explained in section III below.

Figure 5



III: First Order Model of Reduced Data Set

III.I Restructuring the model data

After testing many different interaction terms and squared variables and no improvements were made on the second model to make the Residual plot looked random, we decided to re-evaluate the data again. The fact that the structure of the residual plots was divided into two regions raise us the suspicion that there is some division in characteristics of our dataset. This division might be from one of our categorical values. After examining the nature of our variables and dataset, we think that the Control of Institution (CoI) which indicates whether the observed college is public or private is the potential factors that cause variation in college tuition.

To verify the hypothesis that CoI variables affect the random pattern of our residual plots, we fit all 18 quantitative variables and 4 qualitative variables from our initial dataset into first order regression model and exclude the CoI variable. The figure 6 below shows that the four types of residual plots of fitted regression model without the CoI variable. We can observe that the bimodal pattern of errors in the residual vs fitted value has disappeared. On the other hand, the figure 7 below shows that as long as we add the CoI variable into our regression model, the bimodal pattern of error starts to appear again. This further enforces our belief that CoI variable is really the reason why the error distribution of our regression model is not random, no matter how many transformation in y variables and addition of interaction terms and squared variables we have.

Figure 6

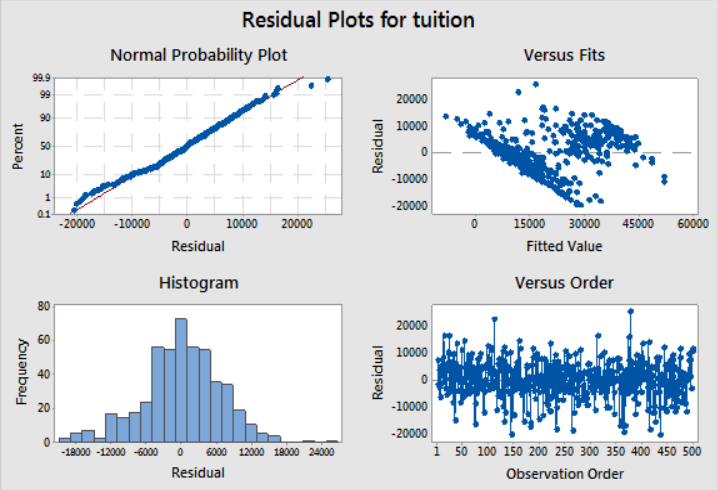
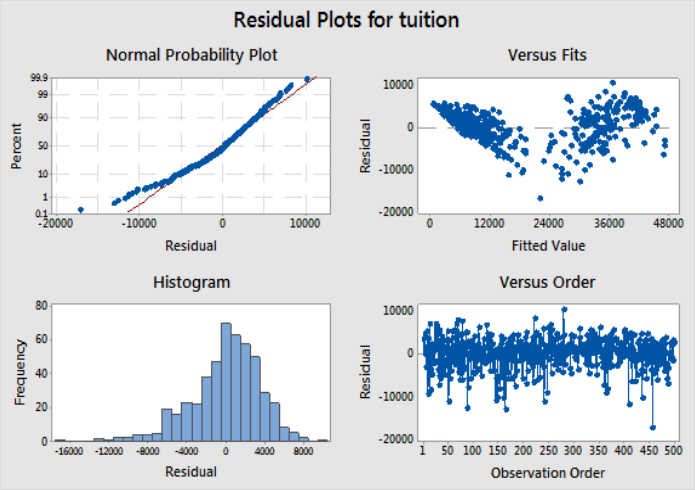


Figure 7



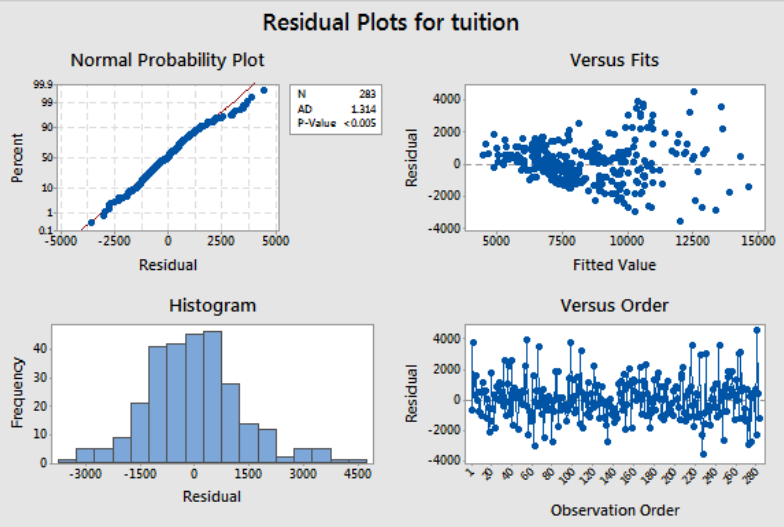
We followed instructor's recommendation and decided to refine the scope of our project to encompass only public institutions, removing private universities from the observations. We then proceeded to run a stepwise regression with the updated variable set including the following quantitative variables: percent of freshmen submitting SAT scores, standardized scores, percent admitted, admission yield, percent of full time students, percent of undergraduate students, graduation rate within four years, percent of freshmen receiving financial aid, and endowment per alumni. The following qualitative variables were also included: geographic region, historical black college or university, highest degree offered, and degree of

urbanization. Note that because observations for private universities were removed, Control of Institution is no longer a variable.

III.II Variable Selection and Screening Techniques

The stepwise regression (refer to appendix E) narrowed the variable scope down to the following quantitative variables: percentSAT, stdScores, percentAdmitted, adYield, percentFT, percentUN, and GDRate. As for the qualitative variables, highest degree offered and historical black college or university were deemed as statistically uninfluential variables and therefore removed. The graphic result of the residual analysis shown in the figure 8 below. From the error plot in the top right quadrant we can identify that the variance increase as the fitted values increase, creating a fanning out pattern. This pattern is the key to determining which type of transformation is to be applied to this regression model.

Figure 8

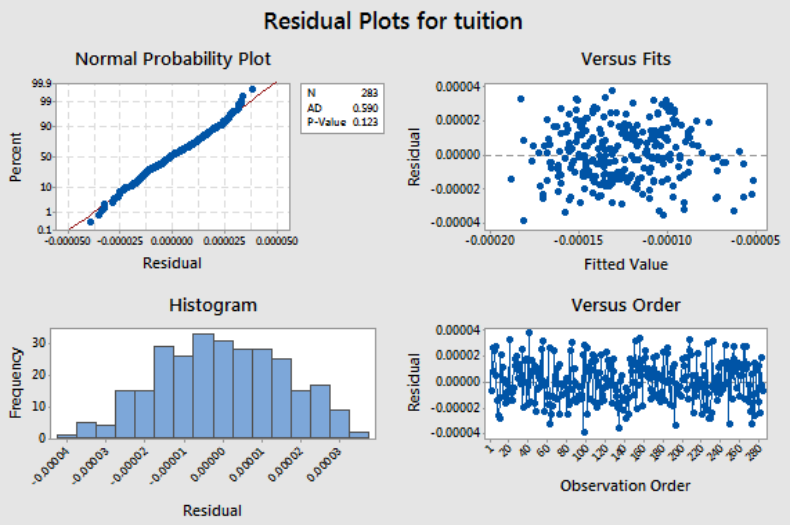


III.III Model Transformation and Normality Testing

After the removal of outlying points (refer to appendix F) and applying an optimal Box-Cox transformation with a λ = -0.764884 the result was a random error distribution as can be seen in the figure 9 displayed below. The normality Probability Plot also displays a satisfying p-value for the Anderson-Darling Normality test despite the tradeoff created by reducing the number of observational units. We can nonetheless identify certain observations that could be considered outliers, especially along the tails of the

normality plot. Nevertheless, the histogram shows that the mean error value is approximately zero, as expected.

Figure 9



To test and confirm the significance of each variable, both quantitative and qualitative, we analyzed the VIF values after running the transformed regression model and saw that these variables were in fact significant due to their relatively low VIFs. Refer to appendix G for details on the specific VIF values.

Conclusion and Recommendation

From our regression analysis, we can conclude that the significant variables that most strongly influence tuition and fees on a public university level are the following: the percent of students submitting SAT scores, the institution's standardized scores (on both 25th and 75th percentile levels), percent of students admitted, admissions yield, percent of full time students, percent of undergraduate students, graduation rate within a four year period, geographic location, and degree of urbanization.

The final regression model is the following:

-tuition^-1 = -0.000150 - 0.000000 percentSAT + 0.000000 Std Scores + 0.000000 percentAdmit - 0.000001 adYield + 0.000000 percentFT - 0.000001 percentUn + 0.000001 GDRate + 0.0 geo_FW + 0.000013 geo_GL + 0.000010 geo_ME + 0.000017 geo_NE - 0.000019 geo_P - 0.000022 geo_RM - 0.000012 geo_SE - 0.000006 geo_SW + 0.0 DoI_C - 0.000000 DoI_S - 0.000008 DoI_T

Note that the aforementioned model reflects a transformation of the dependent tuition variable. In addition, the residual analysis for the model passes the normality and randomness assumptions as showed in the sections above. After splitting the data, there was no need to interaction terms or quadratic terms to fix residual assumptions.

Some practical applications of this study include:

- Financial budgeting for parents and students planning for university expenses.
- Government regulations of prices minimum and maximum charges by public universities.
- Measurement of impact that each variable has on tuition.

Data Validation

To validate the model, let's consider predicting the tuition cost of a public school with the following values for its predicting variables:

percentSAT = 80	GDRate = 70	geo_RM = 0
StdScores = 2200	geo_FW = 0	geo_SE = 1
percentAdmin = 32	geo_GL = 0	geo_SW = 0
adYield = 65	geo_ME = 0	DoI_C = 1
percentFT = 85	geo_NE = 0	DoI_S = 0
percentUN = 70	geo_P = 0	DoI_T= 0

Plugging the numbers:

-tuition^-1= -0.000150 - 0.000000 (80) + 0.000000 (2200) + 0.000000 (32) - 0.000001 (65)+ 0.000000 (85)- 0.000001 (70) + 0.000001 (70)+ 0.0 (0)+ 0.000013 (0)+ 0.000010 (0)+ 0.000017 (0)- 0.000019 (0)- 0.000022 (0)- 0.000012 (1)- 0.000006 (0)+ 0.0 (1)- 0.000000 (0)- 0.000008 (0)

Tuition = 16200

Compare this answer to the data we have Flagler College -St Augustine tuition is 16180.

Real data:

	Flagler College-St Augustine	Predicting	Variation
percentSAT	72	80	10%
Std.Scores	2060	2200	6.30%
percentAdmin	50	32	36%
adYield	26	65	60%

percentFT	96.72%	85%	12.11%
percentUN	100%	70%	30%
GDRate	46	70	34.28%
geo_FW	0	0	
geo_Gl	0	0	
geo_ME	0	0	
geo_NE	0	0	
geo_P	0	0	
geo_RM	0	0	
geo_SE	1	1	
geo_SW	0	0	
DoI_C	0	1	
DoI_S	1	0	
DoI_T	0	0	

One detail to note about our dataset is the mean of our tuition data is 18108.79 and the variation is 14000.84 which is 77% of our mean value. Therefore, the large deviation of our predicted value from actual value of the random observation can be justified by nature of our dataset with big variation.

Appendix

A. The picture below shows the stepwise results applied to all 23 original variables. That is, to the 18 quantitative variables and 5 qualitative variables predicting tuition.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3706.51	93.27%	92.99%	92.45%

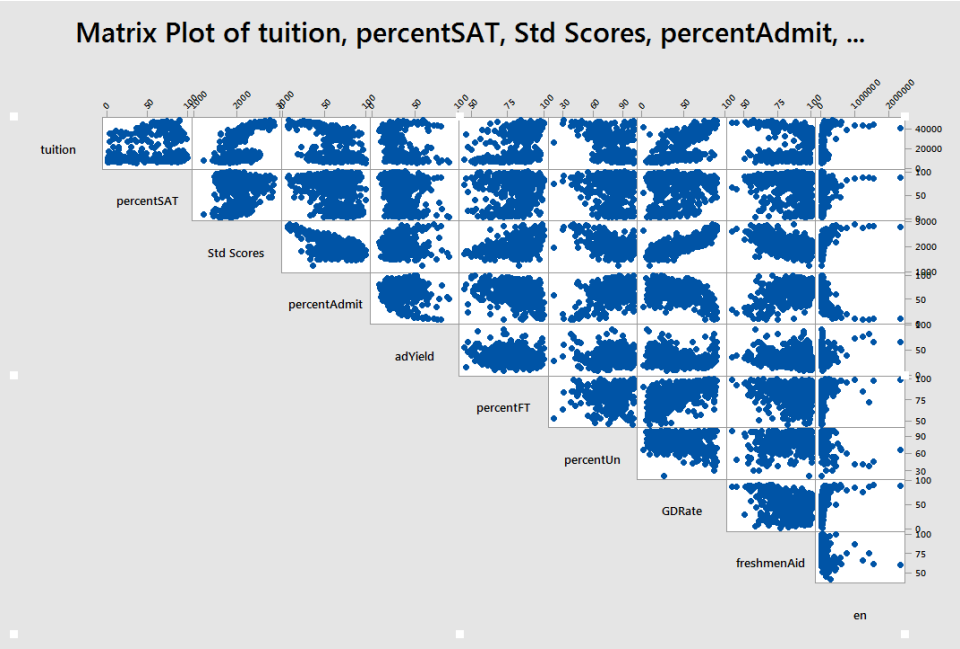
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2870	2322	1.24	0.217	
adminTotal	0.1480	0.0681	2.18	0.030	4.85
percentSAT	-62.8	13.8	-4.56	0.000	7.15
percentACT	-59.2	15.8	-3.74	0.000	7.65
ACT25	828.8	94.1	8.81	0.000	4.45
adYield	-48.4	19.7	-2.46	0.014	2.63
FTEnroll	-0.3314	0.0521	-6.36	0.000	8.92
GDEnroll	0.4312	0.0819	5.26	0.000	3.58
GDRate	88.8	15.9	5.59	0.000	4.87
en	0.00277	0.00127	2.18	0.030	1.73
geo					
GL	-2297	798	-2.88	0.004	2.52
ME	-1039	705	-1.47	0.141	2.96
NE	-1135	840	-1.35	0.177	2.02
P	-4570	1045	-4.37	0.000	2.31
RM	-5752	1068	-5.39	0.000	1.67
SE	-4176	685	-6.09	0.000	3.40
SW	-1779	769	-2.31	0.021	1.76
CoI					
1	17909	542	33.04	0.000	2.47
DoI					
R	-678	1437	-0.47	0.637	1.04
S	-1521	433	-3.51	0.000	1.24
T	-1732	499	-3.47	0.001	1.27

Regression Equation

tuition = 2870 + 0.1480 adminTotal - 62.8 percentSAT - 59.2 percentACT + 828.8 ACT25
- 48.4 adYield - 0.3314 FTEnroll + 0.4312 GDEnroll + 88.8 GDRate + 0.00277 en
+ 0.0 geo_FW - 2297 geo_GL - 1039 geo_ME - 1135 geo_NE - 4570 geo_P - 5752 geo_RM
- 4176 geo_SE - 1779 geo_SW + 0.0 CoI_0 + 17909 CoI_1 + 0.0 DoI_C - 678 DoI_R
- 1521 DoI_S - 1732 DoI_T

B. The picture below shows the Matrix Plot for the 9 new variables after updates to remove collinearity were done.



C. The picture below shows the stepwise results applied to all 14 new variables. That is, to the 9 updated quantitative variables and 5 qualitative variables predicting tuition. Out of the 14 variables, the 9 variables below were the ones

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
3945.99	92.33%	92.06%	91.44%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3267	2826	1.16	0.248	
percentSAT	-27.48	8.50	-3.23	0.001	2.41
Std Scores	6.91	1.20	5.76	0.000	4.01
adYield	-69.9	16.6	-4.20	0.000	1.66
percentUn	-51.0	15.9	-3.22	0.001	1.49
GDRate	89.4	16.6	5.38	0.000	4.71
en	0.00398	0.00134	2.97	0.003	1.71
geo					
GL	-1917	844	-2.27	0.024	2.49
ME	1166	676	1.73	0.085	2.40
NE	1231	817	1.51	0.132	1.68
P	-3264	1100	-2.97	0.003	2.26
RM	-5380	1131	-4.76	0.000	1.65
SE	-3050	712	-4.28	0.000	3.24
SW	-1661	816	-2.03	0.042	1.75
CoI					
1	19605	502	39.03	0.000	1.87
DoI					
R	-263	1530	-0.17	0.864	1.04
S	-1230	457	-2.69	0.007	1.22
T	-1196	520	-2.30	0.022	1.21

Regression Equation

tuition = 3267 - 27.48 percentSAT + 6.91 Std Scores - 69.9 adYield - 51.0 percentUn
+ 89.4 GDRate + 0.00398 en + 0.0 geo_FW - 1917 geo_GL + 1166 geo_ME + 1231 geo_NE
- 3264 geo_P - 5380 geo_RM - 3050 geo_SE - 1661 geo_SW + 0.0 CoI_0 + 19605 CoI_1
+ 0.0 DoI_C - 263 DoI_R - 1230 DoI_S - 1196 DoI_T

D. The figure below shows the 38 unusual observations suggested by minitab.

Fits and Diagnostics for Unusual Observations

Original Response

Obs	tuition	Fit
3	14240	10369
4	37530	29009
21	4558	6461
54	9386	6587
68	7573	10608
72	46870	36560
101	34120	25906
120	31334	23995
129	38941	45878
135	20724	28294
141	14410	21408
147	16170	22171
153	14240	9764
154	7222	10037
159	46752	46027
164	43498	51498
165	7230	10881
171	13728	9473
180	7327	10340
189	8128	11673
222	15150	21187
223	14240	10196
227	22683	32537
257	4404	6802
307	6343	9293
309	8376	6052
402	6410	9220
403	8340	12150
413	12327	8359
420	14096	10347
424	8061	11132

E. The figure below shows the details to the stepwise regression performed after the removal of private institutions from the observational data set.

Regression Analysis: tuition versus percentSAT, Std Scores, percentAdmit, adYield, ...

Method

Categorical predictor coding (1, 0)

Stepwise Selection of Terms

Candidate terms: percentSAT, Std Scores, percentAdmit, adYield, percentFT, percentUn, GDRate, freshmenAid, en, degree, geo, histoBlack, DoI

	----Step 1----		----Step 2----		----Step 3----		-----Step 4----	
	Coef	P	Coef	P	Coef	P	Coef	P
Constant	5536		6002		10972		11632	
GDRate	99.69	0.000	93.58	0.000	86.05	0.000	90.05	0.000
geo			1386	0.000	-1350	0.000	-2294	0.000
percentUn					-55.2	0.000	-52.5	0.000
percentSAT							-11.22	0.003
adYield								
DoI								
percentAdmit								
percentFT								
Std Scores								
S	1766.91		1577.85		1525.20		1504.92	
R-sq	47.81%		59.38%		62.18%		63.31%	
R-sq(adj)	47.63%		58.24%		60.98%		62.01%	
R-sq(pred)	46.94%		56.70%		59.25%		60.24%	
Mallows' Cp	158.84		73.54		51.52		43.83	

	-----Step 5----		-----Step 6----		-----Step 7----		-----Step 8----	
	Coef	P	Coef	P	Coef	P	Coef	P
Constant	12741		12320		11534		10138	
GDRate	85.24	0.000	86.19	0.000	89.22	0.000	80.28	0.000
geo	-1917	0.000	-1879	0.000	-2124	0.000	-1918	0.000
percentUn	-53.2	0.000	-44.9	0.000	-49.3	0.000	-57.6	0.000
percentSAT	-12.42	0.001	-14.96	0.000	-13.75	0.000	-13.12	0.001
adYield	-29.59	0.001	-30.30	0.001	-28.78	0.001	-25.52	0.004
DoI			-798	0.001	-796	0.001	-803	0.001
percentAdmit					15.25	0.013	16.54	0.007
percentFT							24.9	0.030
Std Scores								
S	1479.29		1449.19		1435.92		1426.38	
R-sq	64.68%		66.34%		67.07%		67.63%	
R-sq(adj)	63.29%		64.77%		65.41%		65.87%	
R-sq(pred)	61.54%		62.74%		63.24%		63.64%	
Mallows' Cp	34.12		23.84		19.56		16.81	

	-----Step 9----	
	Coef	P
Constant	8213	
GDRate	70.39	0.000
geo	-1974	0.000
percentUn	-53.8	0.000
percentSAT	-13.24	0.000
adYield	-27.15	0.002
DoI	-783	0.001
percentAdmit	18.13	0.004
percentFT	24.7	0.032
Std Scores	0.997	0.107
S	1422.25	
R-sq	67.93%	
R-sq(adj)	66.07%	
R-sq(pred)	63.74%	
Mallows' Cp	16.21	

α to enter = 0.15, α to remove = 0.15

F. The figure below shows the 10 unusual observations suggested by minitab for the first order model of the reduced observation set.

Fits and Diagnostics for Unusual Observations

Original Response

Obs	trans(y)	Fit
5	67.51	78.26
7	84.23	73.29
16	96.88	80.98
112	93.38	80.78
120	66.36	83.80
155	79.64	92.28
156	91.52	77.93
181	94.22	81.45
246	111.03	92.35
271	80.35	93.87

G. The figure below shows the 10 unusual observations suggested by minitab for the first order model of the reduced observation set.

Model Summary for Transformed Response

S	R-sq	R-sq(adj)	R-sq(pred)
0.0000166	73.93%	72.37%	70.47%

Coefficients for Transformed Response

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-0.000150	0.000021	-7.09	0.000	
percentSAT	-0.000000	0.000000	-3.13	0.002	2.61
Std Scores	0.000000	0.000000	2.83	0.005	2.83
percentAdmit	0.000000	0.000000	6.17	0.000	1.41
adYield	-0.000001	0.000000	-4.96	0.000	1.61
percentFT	0.000000	0.000000	3.01	0.003	2.23
percentUn	-0.000001	0.000000	-5.66	0.000	1.58
GDRate	0.000001	0.000000	7.08	0.000	3.62
geo					
GL	0.000013	0.000005	2.56	0.011	2.61
ME	0.000010	0.000004	2.51	0.013	2.03
NE	0.000017	0.000005	3.23	0.001	1.77
P	-0.000019	0.000006	-3.05	0.003	2.66
RM	-0.000022	0.000006	-3.55	0.000	2.01
SE	-0.000012	0.000004	-2.90	0.004	3.81
SW	-0.000006	0.000005	-1.41	0.160	2.17
DoI					
S	-0.000000	0.000003	-0.10	0.918	1.36
T	-0.000008	0.000003	-3.09	0.002	1.26

J. The figure below shows the equation that has the best variables after adding quadratic and interaction terms.

```
Regression Equation

tuition^0.215615 = 6.849 - 0.001149 percentSAT + 0.000014 Std Scores - 0.00809 adYield
                   - 0.00462 percentUn + 0.0119 freshmenAid + 0.000000 en
                   - 0.000064 freshmenAid2 + 0.000043 percentSAT*GDRate
                   + 0.000000 percentSAT*en + 0.000009 stdScores*GDRate
                   + 0.000003 adYield*percentUN - 0.000103 adYield*GDRate + 0.0 CoI_0
                   + 1.8923 CoI_1 + 0.0 histoBlack_0 - 0.2290 histoBlack_1 - 0.00540 GDRate
```

K. Correlation Matrix to Detect Correlation Term

Correlation: percentSAT, Std Scores, percentAdmit, adYield, percentUn, percentFT, GDRate, ...

	percentSAT	Std Scores	percentAdmit	adYield	percentUn
Std Scores	0.110				
	0.018				
percentAdmit	-0.230	-0.580			
	0.000	0.000			
adYield	-0.332	-0.091	-0.042		
	0.000	0.050	0.364		
percentUn	-0.058	-0.453	0.296	0.054	
	0.210	0.000	0.000	0.249	
percentFT	0.188	0.426	-0.336	-0.249	0.161
	0.000	0.000	0.000	0.000	0.000
GDRate	0.225	0.830	-0.488	-0.301	-0.364
	0.000	0.000	0.000	0.000	0.000
freshmenAid	-0.265	-0.486	0.466	-0.069	0.122
	0.000	0.000	0.000	0.140	0.009
en	0.073	0.642	-0.559	0.112	-0.323
	0.119	0.000	0.000	0.016	0.000
CoI_1	0.115	0.452	-0.285	-0.372	-0.393
	0.014	0.000	0.000	0.000	0.000
DoI_C	-0.037	0.095	-0.076	0.084	-0.229
	0.431	0.041	0.102	0.072	0.000
DoI_S	0.197	0.031	0.003	-0.227	0.046
	0.000	0.512	0.956	0.000	0.324
FW	0.315	-0.056	-0.055	-0.212	0.051
	0.000	0.227	0.241	0.000	0.275

GL		-0.281	0.058	0.064	-0.006	-0.017
	0.000		0.213	0.173	0.898	0.714
ME		0.361	0.139	-0.134	-0.192	-0.053
		0.000	0.003	0.004	0.000	0.259
NE	0.269		0.090	-0.069	-0.133	-0.090
		0.000	0.054	0.140	0.004	0.053
PP		-0.394	0.028	0.145	0.119	-0.004
		0.000	0.550	0.002	0.010	0.940
RM		-0.205	-0.038	0.187	0.022	0.031
		0.000	0.412	0.000	0.643	0.500
SE		-0.237	-0.121		-0.050	0.295	0.073
		0.000	0.009	0.286	0.000	0.118	

	percentFT	GDRate	freshmenAid	en	CoI_1
GDRate	0.534				
	0.000				
freshmenAid	-0.250	-0.370			
	0.000	0.000			
en	0.337	0.554	-0.345		
	0.000	0.000	0.000		
CoI_1	0.236	0.628	0.083	0.396	
	0.000	0.000	0.076	0.000	
DoI_C	-0.061	0.012	-0.084	0.040	0.046
	0.190	0.797	0.071	0.395	0.322
DoI_S	0.037	0.094	-0.022	0.009	0.092
	0.424	0.044	0.633	0.845	0.048
FW	0.159	-0.045	-0.204	-0.094	-0.053
	0.001	0.335	0.000	0.043	0.259

GL	-0.041	0.045	0.041	0.049	0.054
	0.375	0.340	0.381	0.292	0.245
ME	0.172	0.277	-0.044	0.054	0.222
	0.000	0.000	0.344	0.248	0.000
NE	-0.015	0.168	-0.064	0.106	0.102
	0.751	0.000	0.171	0.022	0.029
PP	-0.032	0.017	0.096	-0.010	-0.030
	0.492	0.718	0.039	0.831	0.523
RM	-0.170	-0.129	-0.072	-0.065	-0.125
	0.000	0.005	0.120	0.165	0.007
SE	-0.025	-0.202	0.150	-0.058	-0.132
	0.597	0.000	0.001	0.212	0.004

	DoI_C	DoI_S	FW	GL	ME
DoI_S	-0.661				
	0.000				
FW	0.080	0.001			
	0.087	0.988			
GL	-0.037	0.021	-0.145		
	0.423	0.653	0.002		
ME	-0.136	0.128	-0.183	-0.189	
	0.003	0.006	0.000	0.000	
NE	-0.093	0.182	-0.116	-0.119	-0.150
	0.047	0.000	0.013	0.011	0.001
PP	-0.006	-0.076	-0.097	-0.100	-0.126
	0.899	0.101	0.037	0.032	0.007
RM	0.067	-0.106	-0.073	-0.076	-0.095
	0.152	0.023	0.116	0.105	0.040
SE	0.077	-0.100	-0.230	-0.237	-0.299

	0.100	0.031	0.000	0.000	0.000

	NE	PP	RM
PP	-0.080		
	0.087		
RM	-0.060	-0.051	
	0.197	0.278	
SE	-0.189	-0.158	-0.120
	0.000	0.001	0.010

Based on the output above, we will carry any pairs of variables that has correlations bigger than 0.4. Therefore, some potential interaction terms that we will preceed in the screening process for the second-order model are percentAdmit*stdScore, percentUn*StdScore, percentFT*StdScore, GDRate*StdScore, freshmenAid*StdScore, freshmenAid*percetAdmit, en*StdScore, en*percentAdmit, CoI_1*StdScore, GDRate*percentFT, en*GDRate, CoI_1*GDRate,

References

Data set collected from: https://public.tableau.com/en-us/s/resources
https://www.reference.com/education/education-important-life-2c3b80038e953b9f
http://www.collegedata.com/cs/content/content_payarticle_tmpl.jhtml?articleId=10064