2018 Spring STAT448 Advanced Data Analysis

Final Report: Analysis of University Information

Submitted to:

Prof. Yeonjoo Park
Department of Statistic
University of Illinois at Urbana-Champaign

Report Prepared By:

Group 8
Jinran Yang
(jinrany2)
Department of Industrial and Enterprise Systems Engineering
University of Illinois at Urbana-Champaign

May 3, 2018

1. INTRODUCTION

1.1 Description of Data:

This dataset is taken from the ASA Statistical Graphics Section's 1995 Data Analysis Exposition contains information on over 1300 American colleges and universities. The U.S. News data contains information on tuition, room & board costs, SAT or ACT scores, application/acceptance rates, graduation rate, student/faculty ratio, spending per student, and a number of other variables for 1300+ schools.

Citation: Index of /datasets/colleges

Number of variables: 35

1. FICE (Federal ID Number)
2. College Name
3. State (Postal Code)
4. Public/Private indicator (public = 1, private = 2)
5. Average Math SAT score
6. Average Verbal SAT score
7. Average Combined SAT score
8. Average ACT score
9. First quartile - Math SAT
10. Third quartile - Math SAT
11. First quartile - Verbal SAT
12. Third quartile - Verbal SAT
13. First quartile - ACT
14. Third quartile - ACT
15. Number of applications received
16. Number of applicants accepted
17. Number of new students enrolled
18. Pct. new students from top 10% of H.S. class
19. Pct. new students from top 25% of H.S. class
20. Number of full time undergraduates
21. Number of part time undergraduates
22. In-state tuition
23. Out-of-state tuition
24. Room and board costs
25. Room costs
26. Board costs
27. Additional fees
28. Estimated book costs
29. Estimated personal spending
30. Pct. of faculty with Ph.D.'s
31. Pct. of faculty with terminal degree
32. Student/faculty ratio
33. Pct.alumni who donate
34. Instructional expenditure per student
35. Graduation rate

1.2 Description of the analysis:

This report mainly focuses on seeing whether there are associations between tuition and any other variables, such as university types, acceptance rates, student/faculty ratio, percentage of new students from top 10% of high school class, room and board costs, etc.

Since the difference between the in-state tuition and out-state tuition is significant for public university, I use the arithmetic average of them as the general tuition to do the analysis.

Besides, I also create two variables in the analysis -- accept rate and enrollment rate which are calculated by number of application accepted divided by number of application received and number of new students enrolled divided by number of application accepted respectively.

Then, I removed all the missing value. After that, we have 267 universities' information.

1.3 Basic Descriptive Statistic

In this data set, we only have one categorical variable -- Public/Private indicator. So, I run *pro univariate* sort by university type to get the basic descriptive statistic of the overall tuition.

Table 1 basic descriptive statistic of the tuition of Public University

Moments			
N	58	**Sum Weights**	58
Mean	4684.51724	**Sum Observations**	271702
Std Deviation	1852.20569	**Variance**	3430665.9
Skewness	3.00630221	**Kurtosis**	13.2376864
Uncorrected SS	1468340660	**Corrected SS**	195547956
Coeff Variation	39.5388808	**Std Error Mean**	243.206522

Table 2 basic descriptive statistic of the tuition of Private University

Moments			
N	209	**Sum Weights**	209
Mean	11866.823	**Sum Observations**	2480166
Std Deviation	3582.51114	**Variance**	12834386.1
Skewness	0.46085346	**Kurtosis**	-0.387009
Uncorrected SS	3.21012E10	**Corrected SS**	2669552304
Coeff Variation	30.189303	**Std Error Mean**	247.807477

According to the table above, we can know that there are 58 public universities and 209 private universities' information in this data set. The mean of tuition of public university is 4685 which is much lower than that of private university which is 11867.

Besides, the standard deviation of tuition of private university is much higher than that of public university, which might imply that some private university might have extremely high tuition compared to other private university.

What is more, based on the result of the tests for normality, we can also find that both of them are not normally distributed. As a result, I choose the nonparametric method (proc npar1way) to do a location test for the tuition of public university and private university.

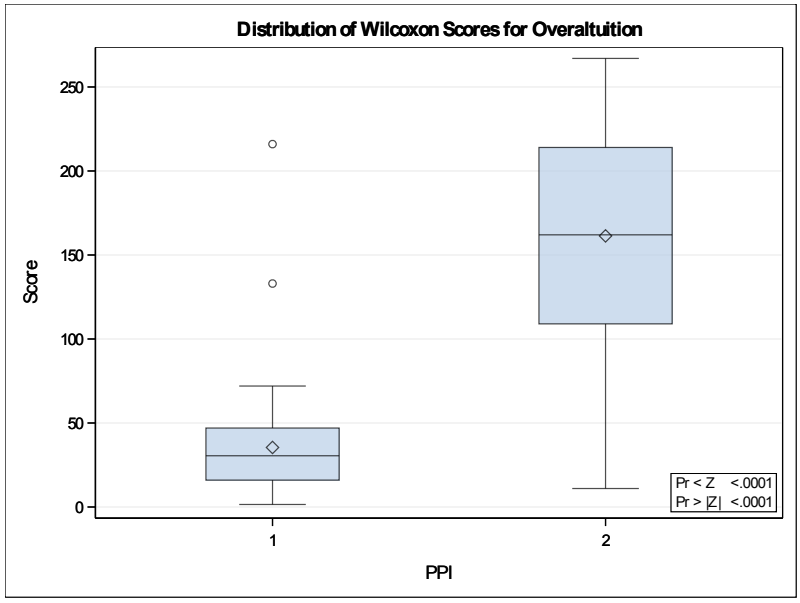

Fig. 1. Location of tuition of Public University (PPI=1) and Private University (PPI=2)

As we can see from the graph above, generally the tuition of private university is much higher than that of the public university. There only exist two extreme points which means there are two public universities' tuition are higher than others.

2. METHODS

2.1 ANOVA

ANOVA can tell us which variables are significant to the tuition. The first thing we should do is to choose an appropriate way to the ANOVA. According to the basic descriptive statistics of the data set, we know that the number of public university is different from the number of private university so that we are going to deal with the unbalanced data set. As a result, I should choose *proc glm* to run the ANOVA.

First, I used *proc glmselect* to do a stepwise model selection to find the variables significant to the tuition. Then, I check the diagnostics panel. According to the graph of Cook's distance, there seems exist some influential points. So, I set 4/N (N is the number of observations which is 267) as a cut off to remove the influential points.

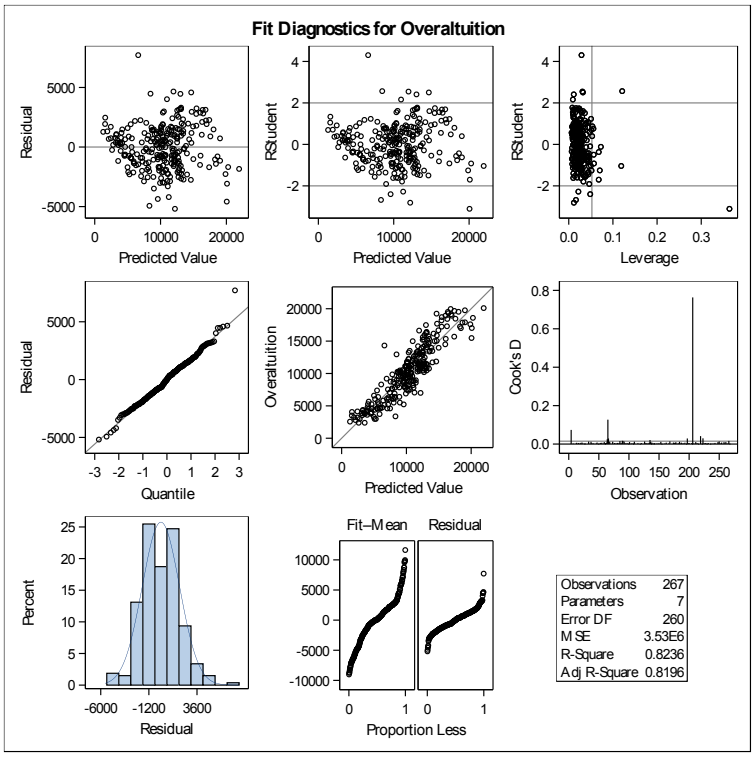

Fig. 2. Diagnostics panel of the first model

After three times removal of influential points, I get diagnostics panel as follow. As we can see, the histogram now is more symmetric and bell-shaped than before. In the RStudent graph, majority of residuals are randomly distributed between [-2,2]. And the new model can explain 88.75% variation of tuition which is quite a lot.

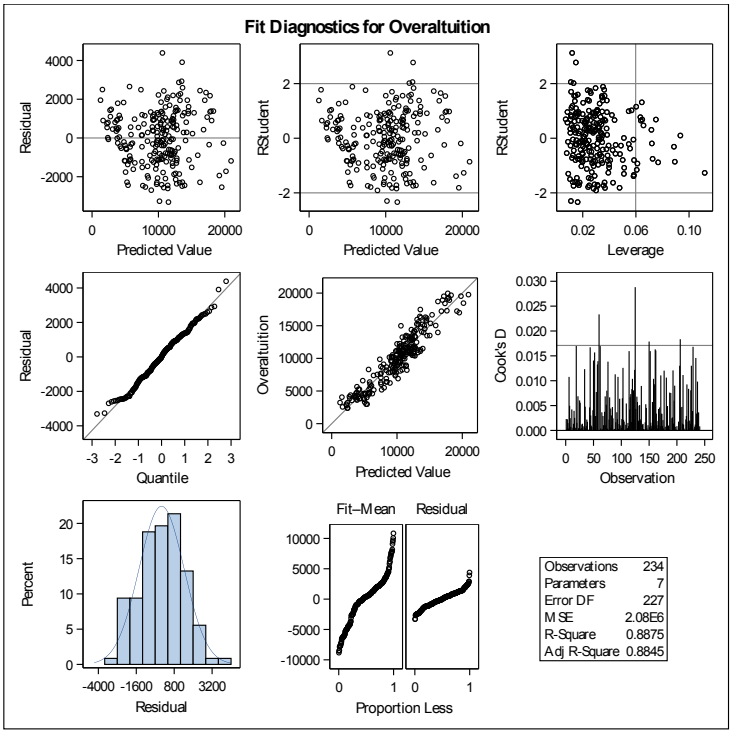

Fig. 3. Diagnostics panel of the model after removing all outliers

Table 3 Result of the final model of ANOVA

R-Square	Coeff Var	Root MSE	Overaltuition Mean
0.887501	14.26371	1441.030	10102.77

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PPI	1	673437436.1	673437436.1	324.30	<.0001
ACSS	1	62439112.8	62439112.8	30.07	<.0001
RBCost	1	11625551.2	11625551.2	5.60	0.0188
PctDonate	1	9806718.6	9806718.6	4.72	0.0308
InstExpend	1	288382301.5	288382301.5	138.87	<.0001
enrolrate	1	47856305.7	47856305.7	23.05	<.0001

We can tell from the Table 3, the type of university, students' average combined SAT score, Room and Board costs, Pct.alumni who donate, Instructional expenditure per student and enrollment rate are significant to the tuition.

2.2 Linear Regression

We know which variables are significant to tuition by ANOVA, but we still do not know how they contribute the tuition and what relationships between them and tuition. Thus, I do a linear regression and compare the result with the result of ANOVA.

Similar to ANOVA, I do a stepwise model selection and I also check the VIF in order to ensure there is no multicollinearity problem. And then, I use 4/N (N is the number of observations which is 267) as a cut off to remove the outliers. After three times removal, I get a model which can explain 89.63% variation of tuition.

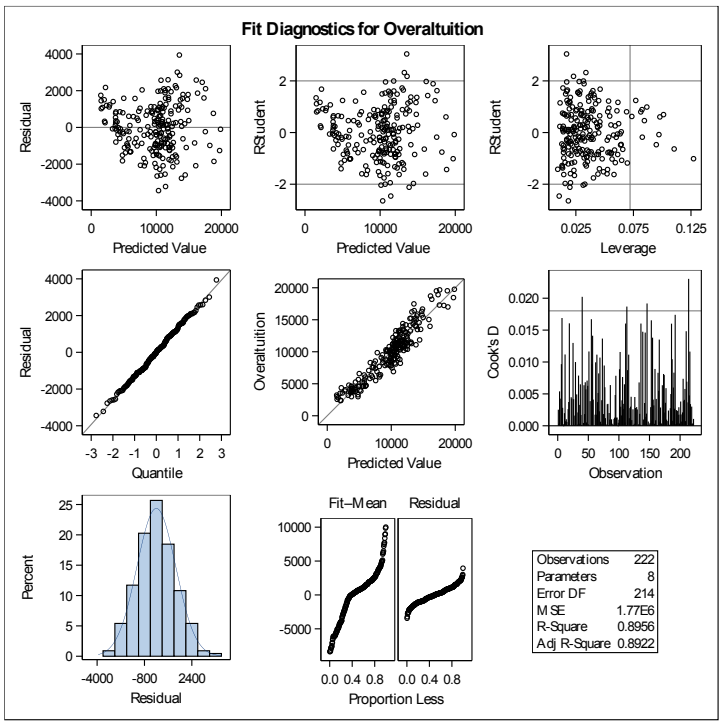

Fig. 4 Diagnostics panel of final model of Linear Regression

As we can see, the histogram is symmetric and bell-shaped. Majority of residuals are randomly distributed between [-2,2] in RStudent graph. The Quantile plot is almost a straight line. As a result, the residuals seem to be normally distributed and we can trust the result we get from the linear regression.

Table 4 Result of the final model of Linear Regression

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	-6558.41435	1280.64830	-5.12	**<.0001**	0
InstExpend	1	0.52619	0.03979	13.22	**<.0001**	1.97543
PPI	1	**4410.63964**	275.30509	16.02	**<.0001**	1.72559
ACSS	1	**5.35548**	1.23194	4.35	**<.0001**	1.98266
PersonalEst	1	-0.61471	0.17578	-3.50	**0.0006**	1.32244
enrolrate	1	**-3782.94464**	825.04094	-4.59	**<.0001**	1.21949
GradRate	1	**22.27386**	6.74395	3.30	**0.0011**	1.54261
AddFee	1	-1.53053	0.43162	-3.55	**0.0005**	1.15721

After removing all the outliers, some variables which are original significant become insignificant. Thus, I remove them and again refit the model and then check whether there exist influential points. Eventually, it turns out no outliers and then this is the final model.

Since all VIF value are less than 10, there is no multicollinearity issue in this model. It can be found that average combined SAT score and graduation rate have strong positive relationship with tuition which might imply that students of university with high tuition tend to study harder and get better academic performance. And the tuition of private university (PPI=2) generally higher than that of public university (PPI =1) by 4277. On the other hand, the enrollment rate has a really strong negative relationship with tuition which imply that students who have already be accepted by the university might choose not to enroll that university because of high tuition.

2.3 Generalized Linear Model

Since tuition is a positive continuous variable, I do a generalized linear model which allow response variable follow gamma distribution and I use log as the link function. I, at the beginning, do a model selection to find the variables which are significant to tuition. And then after removing all the outliers, I check the residual plots. We find that majority of residuals are randomly distributed between [-2,2]. As a result, this model should be fine.

Comparing to the result of linear regression, we find that the result of the generalized linear model is nearly the same as the result of linear regression. The only difference is that the linear regression considers the personal estimated cost is also significant to the

tuition. However, there are too many factors influence the personal estimated cost such as living style, university or living location, individual consumption concept, family society status and etc. which make it hard to explain it. Thus, generally we cannot draw a clear conclusion about the relationship with personal estimated cost and tuition. We can only assume that it might have close relationship with the location of the university.

Table 5 Result of generalized linear regression

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	6.9062	0.1213	6.6685	7.1440	3241.87	<.0001
PPI	1	0.7633	0.0265	0.7113	0.8152	830.45	<.0001
ACSS	1	0.0006	0.0001	0.0003	0.0008	22.63	<.0001
AddFee	1	-0.0001	0.0000	-0.0002	-0.0000	6.44	0.0112
InstExpend	1	0.0000	0.0000	0.0000	0.0000	127.31	<.0001
GradRate	1	0.0025	0.0007	0.0011	0.0038	13.23	0.0003
enrolrate	1	-0.5002	0.0804	-0.6577	-0.3427	38.73	<.0001
Scale	0	55.7815	0.0000	55.7815	55.7815		

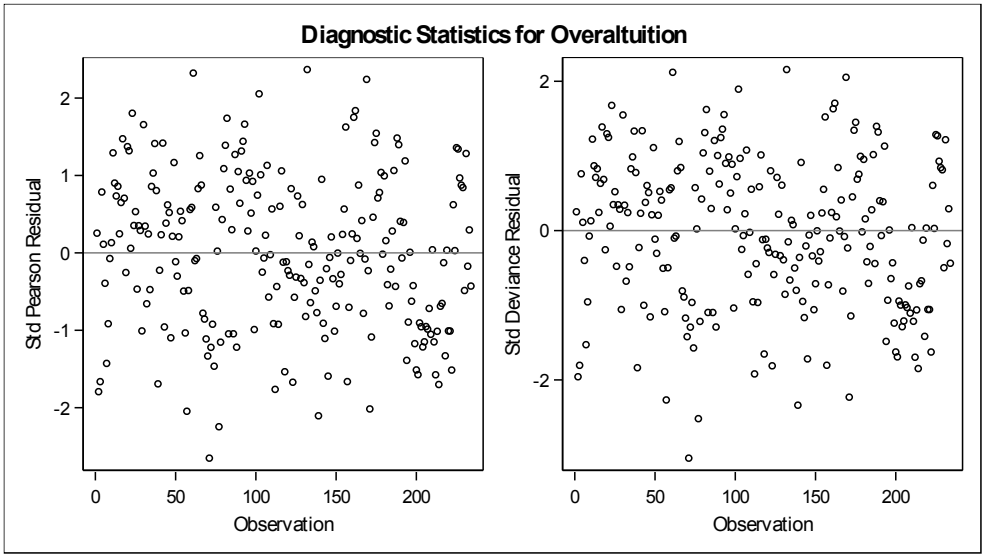

Fig. 5 Residuals plot of generalized model selection

2.4 Principal Component Analysis

Principal component analysis can help us to find the hidden information about tuition in this data set. In order to eliminate the influence of the difference variance among variables, I use correlation matrix to do PCA.

Based on the Scree Plot, I choose knee point which is 3 to draw a score plot. And we can also see that three principal components can explain over 60% variance of the tuition. Since the eigenvector is too long, I put it in appendices.

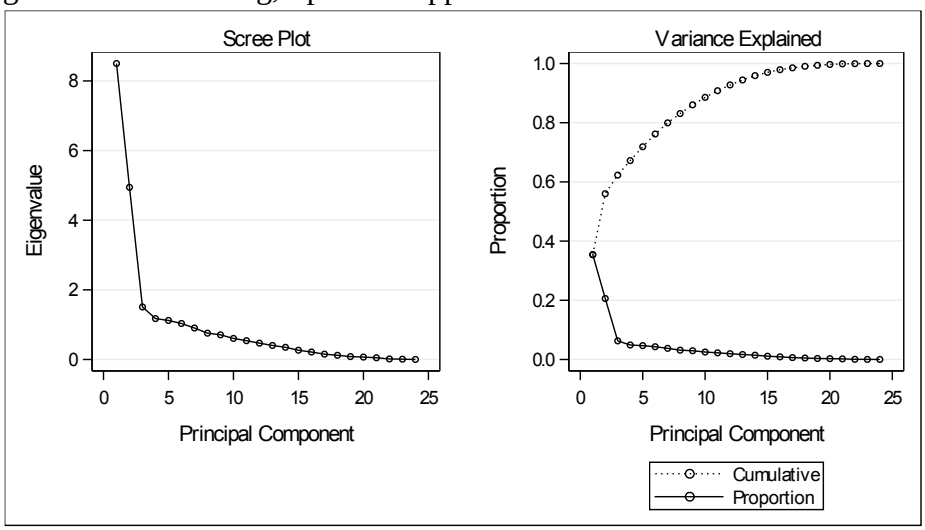

Fig. 6 Scree Plot and Variance Explained

According to the values of eigenvectors, I interpret the principal components as follow:

PC1: + Student academic performance + Tuition

PC2: + Size of university

- Tuition

PC3: + (number of new students enrolled) / (number of accepted)

- Room and Board costs

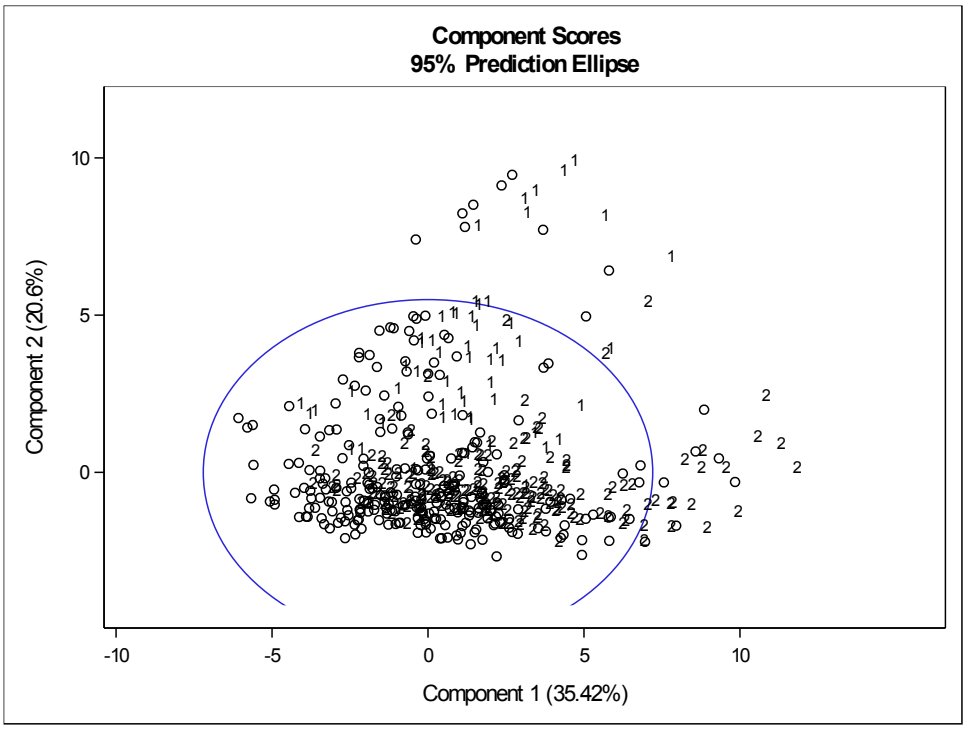

Fig. 7 Score Plot of PC 1 and PC 2

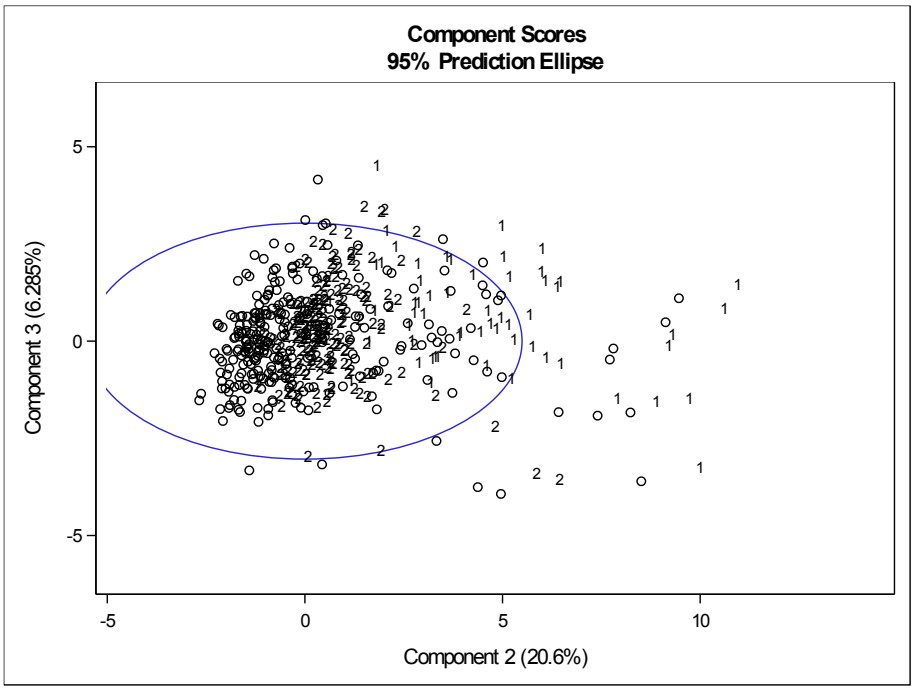

Fig. 8 Score Plot of PC 2 and PC 3

For PC 1, as we can see, some private university have really large PC1 which implies that some private universities students' academic performance is outstanding and the tuitions are really high.

For PC 2, some public university have really large PC2 which implies that contrast between the university size and the tuition is significant. It might imply that the public universities have lower tuition and larger population of students. Besides, there also exist a few private universities that have relative larger PC 2. It might imply that the university tuition is high however, the population of student is small.

For PC 3, It seems no difference between the private university and public university.

3. CONCLUSION

At the beginning, by descriptive analysis, we know the tuition of private university is much higher than that of public university. Then, we use ANOVA to find which variables are significant to tuition. We find that the type of university (public of private), students' average combined SAT score, Room and Board costs, Pct.alumni who donate, Instructional expenditure per student and enrollment rate are significant to the tuition. But we do not know how these variables are correlated to the tuition. As a result, I use linear regression to find which variables are significant to tuition and how they correlated to the tuition and also compare the result of linear regression with the result of ANOVA.

Based on the result of linear regression, we know that average combined SAT score, graduation rate has a strong positive relationship with tuition which might imply that students of university with higher tuition tends to study harder and have a better academic performance. And the tuition of private university (PPI=2) generally higher than that of public university (PPI =1) by 4277. On the other hand, the enrollment rate has a really strong negative relationship with tuition which might imply that students who have already be accepted by the university will choose not to attend because of high tuition. As we can see, the result of ANOVA generally agrees with the result of linear regression.

Then, I do a generalized linear regression which assumes the response variable is gamma distributed. I also do a model selection, remove all influential points and check the residuals plots. Then, after comparing the result of generalized linear regression and the linear regression, it turns out that they are very much close to each other. Generally, we get nearly the same result from these two different methods.

At last, I do a principal component analysis to find the hidden information about the tuition of the data set. The result of PCA reveals that the tuition and students' academic performance are outstanding in private university. On the other hand, the contrast of tuition and the size of university is significant in public university. It can be interpreted as the tuition of public university is lower, however, the public university accept much more students than private university. There are also a few private university which have larger PC 2. It can be interpreted as these university have really high tuition but a small population of students.

In conclusion, the tuition has close relationship with the type of university, student academic performance, the location (living expenses), the size of university.

4. APPENDICES

4.1 Variables labels

When coding in SAS, I use abbreviation to represent variables. My variables and it corresponding label are as follow:

Name – College Name

PPI – Public Private Indicator

AMSS – Average Math SAT score

AVSS – Average Verbal SAT score

ACSS – Average Combined SAT score

AAS – Average ACT Score

FQMS – First Quartile Math SAT

TQMS – Third Quartile Math SAT

FQVS – First Quartile Verbal SAT

TQVS – Third Quartile Verbal SAT

FQA – First Quartile ACT

TQA – Third Quartile ACT

Num_app_rec – Number of applications received

Num_app_acc – Number of applications accepted

Enroll – Number of new students enrolled

TopHS10 – Pct. New students from top 10% of H.S. class

TopHS25 – Pct. New students from top 25% of H.S. class

Under – Number of full time undergraduates

PartUnder – Number of part time undergraduates

InTuition – In State Tuition

OutTuition – Out of State Tuition

RBCost – Room and Board costs

AddFee – Additional Fee

BookEst – Estimated Book Costs

PersonalEst – Estimated Personal Spending

PctPhD – Pct. Of faculty with Ph.D.'s

PctTerminal – Pct. of faculty with terminal degree

SFRatio – Student/faculty ratio

PctDonate – Pct.alumni who donate

InstExpend – Instructional expenditure per student

GradRate – Graduation rate

4.2 Result of generalized linear regression

Model Information		
Data Set	WORK.DIAGNOSTICS1	Predicted Values and Diagnostic Statistics
Distribution	Gamma	
Link Function	Log	
Dependent Variable	Overaltuition	

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	227	4.0695	0.0179
Scaled Deviance	227	227.0000	1.0000
Pearson Chi-Square	227	3.9696	0.0175

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Scaled Pearson X2	227	221.4313	0.9755
Log Likelihood		-1998.8513	
Full Log Likelihood		-1998.8513	
AIC (smaller is better)		4011.7027	
AICC (smaller is better)		4012.1983	
BIC (smaller is better)		4035.8899	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	6.9062	0.1213	6.6685	7.1440	3241.87	<.0001
PPI	1	0.7633	0.0265	0.7113	0.8152	830.45	<.0001
ACSS	1	0.0006	0.0001	0.0003	0.0008	22.63	<.0001
AddFee	1	-0.0001	0.0000	-0.0002	-0.0000	6.44	0.0112
InstExpend	1	0.0000	0.0000	0.0000	0.0000	127.31	<.0001
GradRate	1	0.0025	0.0007	0.0011	0.0038	13.23	0.0003
enrolrate	1	-0.5002	0.0804	-0.6577	-0.3427	38.73	<.0001
Scale	0	55.7815	0.0000	55.7815	55.7815		

4.2 PCA- eigenvalue and three eigenvectors

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	8.50035898	3.55583060	0.3542	0.3542
2	4.94452838	3.43613653	0.2060	0.5602
3	1.50839186	0.33349178	0.0628	**0.6231**
4	1.17490007	0.05333352	0.0490	0.6720
5	1.12156655	0.08748047	0.0467	0.7187
6	1.03408608	0.13009652	0.0431	0.7618
7	0.90398956	0.14799720	0.0377	0.7995
8	0.75599236	0.04778526	0.0315	0.8310
9	0.70820710	0.10292800	0.0295	0.8605
10	0.60527910	0.06566779	0.0252	0.8857
11	0.53961131	0.07097776	0.0225	0.9082
12	0.46863355	0.06544046	0.0195	0.9277
13	0.40319309	0.05306594	0.0168	0.9445
14	0.35012715	0.08390456	0.0146	0.9591
15	0.26622259	0.05235662	0.0111	0.9702
16	0.21386597	0.06073571	0.0089	0.9791

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
17	0.15313026	0.03304178	0.0064	0.9855
18	0.12008848	0.03686918	0.0050	0.9905
19	0.08321930	0.01392247	0.0035	0.9940
20	0.06929683	0.02015605	0.0029	0.9969
21	0.04914078	0.03301444	0.0020	0.9989
22	0.01612634	0.00726633	0.0007	0.9996
23	0.00886001	0.00767572	0.0004	1.0000
24	0.00118429		0.0000	1.0000

Eigenvectors			
	Prin1	**Prin2**	**Prin3**
AMSS	0.314640	0.035842	0.186079
AVSS	0.317298	-.034773	0.172822
ACSS	0.322208	0.002879	0.184216
AAS	0.309119	-.010480	0.141211
Num_app_rec	0.138186	0.367335	-.222365
Num_app_acc	0.096029	0.387901	-.259783
Enroll	0.071523	0.414734	-.114836
TopHS10	0.299387	-.010214	0.186801
TopHS25	0.294001	0.017733	0.194326
Under	0.055281	0.425378	-.080349
PartUnder	-.018823	0.284938	-.045738
RBCost	0.168910	-.085535	**-.415313**
AddFee	0.056004	0.122102	0.030906
BookEst	0.041326	0.084597	0.209191
PersonalEst	-.072244	0.222329	0.218207
PctPhD	0.234503	0.100891	-.021562
PctTerminal	0.217879	0.093889	-.071845
SFRatio	-.145045	0.222206	0.150765
PctDonate	0.191529	**-.186217**	-.013694
InstExpend	0.250468	-.081484	-.166531
GradRate	0.199917	-.094989	-.074489
acceptrate	-.165840	-.080222	-.100274
rollrate	-.095376	0.110455	0.531406
Overaltuition	0.223519	**-.247527**	-.226311