Modeling Project: Choosing a post-secondary institution

Lucinda Sparks

Stats for Business and Economics 211

Rosemary Walker

April 21, 2010

Table of Contents

I. Introduction

The United States has one of the most comprehensive school systems. It has thousands of elementary, middle, and high schools. It also has a vast network of community colleges, vocational schools, and Division I, II, and III colleges. With so many options for post-secondary education, how do schools get rather large numbers of students enrolled at their school? What might factor into the students' decision? Could it be tuition, financial aid, graduation rate, or something else? Many variables could influence or have an effect on the enrollment for schools. If administrators want to boost their school's enrollment and bring in more money, they must know which variables affect enrollment the most.

II. Model

The population regression equation for this model is as follows:

$$y= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 +$$

The dependent variable is the number of students enrolled in the school. The independent variables are tuition, financial aid, graduation rate, the type of school, and ACT score.

It is expected that tuition will have a negative sign. As the cost of a school continues to increase, it is expected that fewer students can afford school and therefore there will be a smaller enrollment. As for financial aid, the sign should be positive, because as students get more money for school it should be more feasible for them to attend thus increasing enrollment. The graduation rate should have a positive sign, too, since students should be more inclined to attend a successful institution. The type of school with regards to choosing public or private institutions could be harder to predict. Public school is recorded as 1 and private as 0. It is expected that type would have a positive coefficient because private schools tend to be more expensive to attend. Lastly, it is predicted that ACT will have a negative sign. This prediction is based on the assumption that as the average ACT scores increase there will be fewer students to achieve that score. Kansas dictates that colleges require a minimum ACT score of only 21 to be admitted.

For this model to accurately estimate the coefficients, it must fulfill five basic assumptions which are:

a. Errors have a mean of zero

b. Errors are normally distributed

c. No heteroskedasticity

d. No autocorrelation

e. No multicollinearity

III. Data

The data was found in a website located at www.campuscorner.com. It was a guide for students that included data on colleges within the United States. This one website contained data on all of the variables used in this project.

The dependent variable is enrollment. Five independent variables were included as they were expected to have a strong effect on enrollment. The first variable is the average tuition per student, measured in dollars. The second variable is the graduation rate which is a percentage that is recorded in decimal form. The third variable is the average amount of financial aid and is also measured in dollars. The type of school is the 4^{th} variable and is categorical data recorded as either 1 for public or 0 for private. The last variable is the average ACT score of the students within the institution and is measured in points.

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Enrollment	50	49062	1315	50377	15772.56	12573.401
Tuition	50	28613	2721	31334	13644.40	10202.426
Graduation rate	50	.6900	.2900	.9800	.652600	.1732382
Financial aid	50	20378	1324	21702	7933.38	6487.479
ACT score	50	16	16	32	24.72	3.351
Valid N (listwise)	50					

In this particular sample, the average graduation rate is 65.3%, and the typical university admits students with an average ACT score of 24.72. These colleges also give the students an average of $7,933.38 in financial aid to help with the tuition bill that averages $13,644.40 a year. The average enrollment of the schools within this sample is 15,772.56 students.

IV. Results

OLS Estimate of College Enrollment
Dependent Variable = Number of students enrolled
(p-values in parentheses)

* = Significant at 10%
** = Significant at 5%
*** = Significant at 1%

Explanatory Variable	I Standardized Coefficient Estimate	II Standardized Coefficient Estimate
Tuition	.228 (.695)	-.841** (.024)
Financial Aid	-1.557** (.026)	
Graduation Rate	26313.021** (.029)	26154.047** (038)
Type of School	14666.534** (.016)	9667.441* (.097)
ACT Score	1914.350*** (.008)	1532.095** (.034)
Constant	*	*
R-square	.674	.634
F	18.155 (.000)	19.478 (.000)
Sample Size	50	50

The first model is the best one to use when predicting enrollment. Some variables appear to be insignificant in the first model. When they are removed from the model and tested through a subset test, they are found to be significant. That being said, tuition is not significant in the first model at a 5% or 10% level. All of the other variables are significant at a 5% level.

The results of the independent variables were extremely interesting. Tuition, which was believed to be the most important determining factor, was not significant in the first model. However, in the second model, when financial aid was dropped due to multicollinearity, tuition became significant at the 5% level. Besides the significance of certain variables, the signs of the independent variables were peculiar. Tuition's sign was positive which would indicate that as school cost more it would have higher enrollment. Financial aid was negative which would indicate that the more help available to students to pay for college would result in a lower enrollment. Lastly, the ACT was positive. This indicates that the higher the scores achieved the more students that were attending. While to me this does not make sense, I cannot say that it is wrong. I really do not know which should be right. These odd signs of the independent variables could indicate multicollinearity.

When checked for multicollinearity, this project's variables were heavily correlated with each other. The majority of the variables were correlated with each other. There were a couple of pair wise correlations that barely passed the test of having less than .50 correlations. The functional form test showed that the model was miss-specified in the first part of the test. In addition, the disturbances in the data are non normal. Due to this, the confidence intervals and hypothesis tests cannot be performed in a normal way.

V. Conclusion

Overall, the variables in this model which were tuition, financial aid, graduation rate, type of school, and average ACT score seemed to affect enrollment in school. The final equation for the model included all of the original variables. While tuition appeared to be insignificant, this could be due to high multicollinearity with financial aid. It was still a significant factor in the model even though the test "proved" tuition to be insignificant. The model was miss-specified, had multicollinearity, and the errors were not normal. On the other hand, there was neither heteroskedasticity nor autocorrelation. Due to these violations of assumptions the hypothesis tests and confidence intervals will be inaccurate if computed normally. Also, the variances of the OLS estimators are too large. The estimators will be sensitive to changes in sample sizes and elimination of independent variables. The violation of assumptions has caused the estimator to no longer be unbiased. Thus, the model is not BLUE. It can still be used, but it is no longer the best linear unbiased estimator. In order for it to be BLUE, the necessary corrections must be made to fix the functional form, normality, and multicollinearity. Once this has been done, the model is ready to predict an educational institution's enrollment.

VI. Reference

Work Cited

Top Colleges & Education Guides for Students. Web. 24 Jan. 2010. <http://www.campuscorner.com/>.

VII. Appendices

A. Computer Printouts

Coefficients[a]

Model	Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1 (Constant)	-47393.958	12597.689		-3.762	.000
Tuition per student	.228	.576	.185	.395	.695
Graduation rate	26313.021	11675.582	.363	2.254	.029
Financial Aid	-1.557	.674	-.803	-2.312	.026
Type of School	14666.534	5853.981	.587	2.505	.016
Average ACT Score	1914.350	690.196	.510	2.774	.008

ANOVA[b]

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	5.217E9	5	1.043E9	18.155	.000[a]
Residual	2.529E9	44	5.748E7		
Total	7.746E9	49			

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.821[a]	.674	.636	7581.275	1.655

Correlations

		Number of students enrolled	Average In-State Tuition per student	Graduation rate	Financial Aid	Type of School	Average ACT Score
Pearson Correlation	Number of students enrolled	1.000	-.495	.092	-.418	.646	.032
	Tuition per student	-.495	1.000	.656	.952	-.885	.721
	Graduation rate	.092	.656	1.000	.713	-.417	.834
	Financial Aid	-.418	.952	.713	1.000	-.768	.788
	Type of School	.646	-.885	-.417	-.768	1.000	-.477
	Average ACT Score	.032	.721	.834	.788	-.477	1.000
Sig. (1-tailed)	Number of students enrolled	.	.000	.264	.001	.000	.413
	Tuition per student	.000	.	.000	.000	.000	.000
	Graduation rate	.264	.000	.	.000	.001	.000
	Financial Aid	.001	.000	.000	.	.000	.000
	Type of School	.000	.000	.001	.000	.	.000

Regression for Model 2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.796[a]	.634	.601	7938.819

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.910E9	4	1.228E9	19.478	.000[a]
	Residual	2.836E9	45	6.302E7		
	Total	7.746E9	49			

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-32909.525	11444.649		-2.876	.006
	Average In-State Tuition per student	-.841	.360	-.683	-2.338	.024
	Graduation rate	26154.047	12226.008	.360	2.139	.038
	Type of School	9667.441	5696.520	.387	1.697	.097
	Average ACT Score	1532.095	701.701	.408	2.183	.034

Regression- yhat_2

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.523E9	6	9.205E8	17.804	.000[a]
	Residual	2.223E9	43	5.170E7		
	Total	7.746E9	49			

Regression- yhat_2 and yhat_3

ANOVA[b]

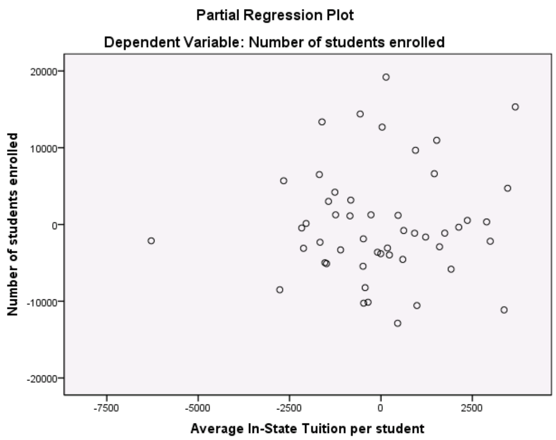
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.534E9	7	7.905E8	15.005	.000[a]
	Residual	2.213E9	42	5.268E7		
	Total	7.746E9	49			

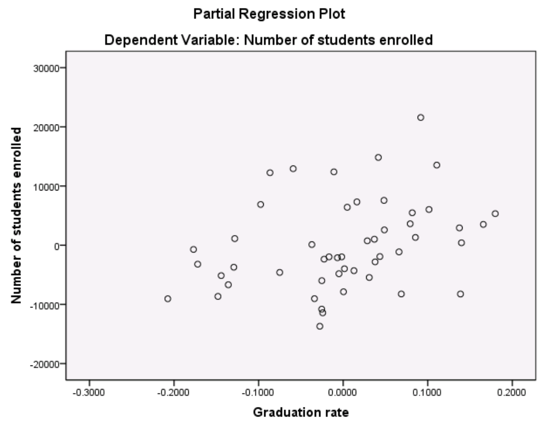
	N	Sum
e2	50	2.53E9
ranktype_e2	50	7.47E10
Valid N (listwise)	50	

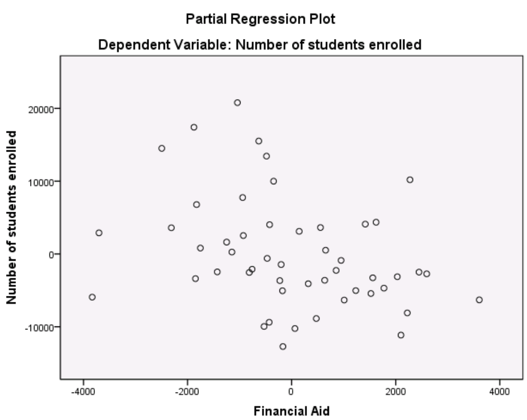
Tests of Normality

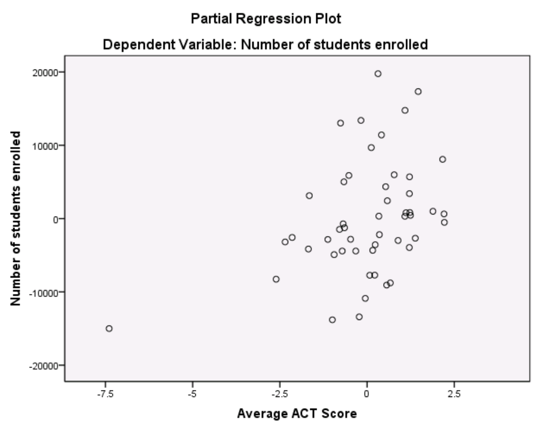
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	.135	50	.023	.954	50	.050

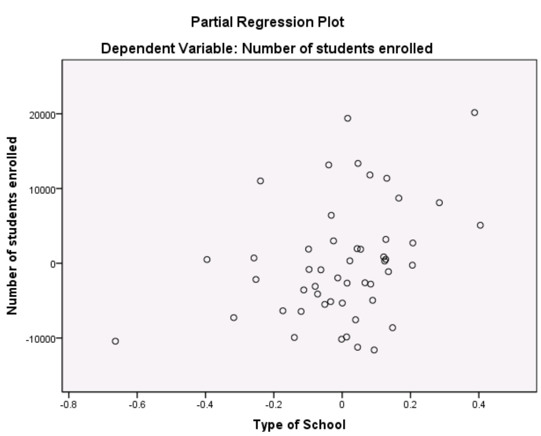
a. Lilliefors Significance Correction

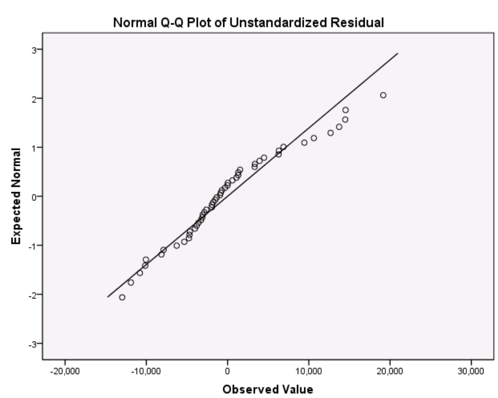

Partial Regression Plot
Dependent Variable: Number of students enrolled
Number of students enrolled
Average In-State Tuition per student


Partial Regression Plot
Dependent Variable: Number of students enrolled
Number of students enrolled
Graduation rate


Partial Regression Plot
Dependent Variable: Number of students enrolled
Number of students enrolled
Financial Aid


Partial Regression Plot
Dependent Variable: Number of students enrolled
Number of students enrolled
Average ACT Score


Partial Regression Plot
Dependent Variable: Number of students enrolled
Number of students enrolled
Type of School


Normal Q-Q Plot of Unstandardized Residual
Expected Normal
Observed Value

B. Answers to Parts 1 and 2

Modeling Project: Part One

Symbol	**Variable**
x_1	Tuition
x_2	Graduation Rate
x_3	Financial Aid
x_4	Type of School
x_5	ACT score

1. $\hat{y}$= -47393.958 +.228x_1 + 26313.021x_2 – 1.557x_3 + 14666.534x_4 + 1914.350x_5
   (12597.689) (.576) (11675.582) (.674) (5853.981) (690.196)
2. R^2= .674; 67.4% of the variation in the number of students enrolled can be explained by Tuition, Graduation Rate, Financial Aid, Type of School, and ACT score.
3. H_0: $\beta_1=\beta_2=\beta_3=\beta_4=\beta_5=0$
   H_a: At least one $\beta_i \neq 0$, i= 1,2,3,4,5
   Reject H_0 if p-value <.05
   P-value = .000
   Reject the null
   At a 5% significance level at least one of the following explains enrollment: Tuition, Graduation Rate, Financial Aid, Type of School, and ACT score.
4. H_0: $\beta_1=0$
   H_a: $\beta_1 \neq 0$
   Reject H_0 if p-value < .05
   P-value= .695
   Do not reject null
   At a 5% significance level, Tuition is not a significant cause of the number of students enrolled.
5.

Variable	**P-value**	**Significant at 5%? 10%?**
Tuition	.695	N/N
Graduation Rate	.029	Y/Y
Financial Aid	.026	Y/Y
Type of School	.016	Y/Y
ACT score	.008	Y/Y

6. $\hat{y}$= -32909.525 -.841x_1 + 26154.047x_2 + 9667.441x_4 + 1532.095x_5
   (11444.649) (.360) (12226.008) (5696.520) (701.701)
   NOTE: Financial Aid was dropped even though it was significant, because it had high multicollinearity with tuition (Correlation = .952). It is believed that tuition is a more important factor in determining enrollment.
7. H_0: $\beta_4=0$
   H_a: $\beta_4 \neq 0$
   Reject H_0 if $F_c > F_{1,44,.05}$
   $F_{1,44,.05}$=4.08
   F_c= 5.34
   5.34 > 4.08
   Reject the null

At a 5% significant level, it appears that financial aid does affect enrollment. Even though financial aid is almost perfectly correlated with tuition, it should be included in the equation since it is significant. The full regression model should be used when analyzing the data.

8. ^{2}=.601; 60.1% of the variation in the number of students enrolled can be explained by Tuition, Type of School, Graduation Rate, and ACT score, adjusted for degrees of freedom.
9. H_0: $\beta_1=\beta_2=\beta_4=\beta_5=0$
   H_a: At least one $\beta_i \neq 0$, i= 1,2,4,5
   Reject H_0 if p-value <.05
   P-value = .000
   Reject the null

At a 5% significance level at least one of the following explains enrollment: Tuition, Graduation Rate, Type of School, and ACT score.

10.

Variable	**P-value**	**Significant at 5%? 10%?**
Tuition	.024	Y/Y
Graduation Rate	.038	Y/Y
Type of School	.097	N/Y
ACT score	.034	Y/Y

11. B_1: All else held equal, a 1 dollar increase in tuition will result in an estimated .841 student decrease in enrollment.
   B_2: All else held equal, a 1% increase in the graduation rate will result in an estimated 26154.047 student increase in enrollment
   B_4: All else held equal, on average Public Schools will have 14666.53 higher enrollment than private schools
   B_5: All else held equal, a 1 point increase in average ACT scores will result in an estimated 701.701 student increase in enrollment.

12.

Variable	**95 % Confidence Interval**
Tuition	(-1.566, -.116)
Graduation Rate	(1529.602, 50779.492)
Type of School	(-1805.940, 21140.821)
ACT score	(118.797, 2945.393)

13. In repeated sampling, 95% of the Confidence Intervals computed in this manner will contain the true population slope coefficient. With 95% confidence, the population parameter for tuition, β_1, lies between -1.566 and -1.16.

Modeling Project: Part Two

1. In order for the OLS estimators to be BLUE, they must have:
   a. Error have a mean of 0
   b. Errors are normally distributed
   c. No autocorrelation
   d. No heteroskedasticity
   e. No multicollinearity

2. The model has multicollinearity if
   a. Any pair wise correlation > .50

b. There is a large F and small t's
c. Any VIF is > 10 or
d. Average VIF is significantly greater than 1

My modeling project has all four signs of multicollinearity for example:
In the table below, the correlations that are greater than .50 are highlighted. There is multicollinearity between many of the variables.
The F-statistic is large compared to the t-statistics.
The VIF for In-State Tuition is 29.458 and for Financial Aid it is 16.280
The average VIF = 12.238

3. H_0: $\rho = 0$
H_a: $\rho > 0$
Reject H_0 if $d < 1.34$, accept H_0 if $d > 1.77$, inconclusive if $1.34 \leq d \leq 1.77$
$d = 1.655$

	Number of students enrolled	Average In-State Tuition per student	Graduation rate	Financial Aid	Type of School	Average ACT Score
Number of students enrolled	1	-.495**	.092	-.418**	.646**	.032
Average In-State Tuition per student	-.495**	1	.656**	.952**	-.885**	.721**
Graduation rate	.092	.656**	1	.713**	-.417**	.834**
Financial Aid	-.418**	.952**	.713**	1	-.768**	.788**
Type of School	.646**	-.885**	-.417**	-.768**	1	-.477**
Average ACT Score	.032	.721**	.834**	.788**	-.477**	1

Inconclusive
At a 5% significance level, there is not sufficient evidence to conclude anything about autocorrelation. It is inconclusive.

4. H_0: No heteroskedasticity
H_a: Heteroskedasticity
Reject H_0 if $Q > 1.645$
h= ———— = 29.5257

Q= —— $^{1/2}$ (29.53- — = 1.4
Do not reject. There does not appear to be heteroskedasticity in the type of school.

5. H_0: ${}_1 = 0$
H_a: ${}_1 \neq 0$
Reject H_0 if $F_c > F_{1,43,.05}$
F=4.08
F_c= ———————————— = 5.919
5.919 > 4.08
Reject the null hypothesis. It appears that the model is misspecified at the 5% level. There is no need to perform the second half of the test due to the null has already been rejected.

6. H_0: Disturbances are normal
H_a: Disturbances are not normal

Reject H_0 if p-value < .05
P-value= .023
Reject H_o
At a 5% significance level, the errors do not appear to be normal.

7. Some of the assumptions have been violated. For example, there is multicollinearity among the variables which as caused tuition to be insignificant and a few variables appear to have the wrong sign. This also means that the variances of the OLS estimators are too large. It could cause any of the variables to appear to be insignificant; however in this case it only affected one variable. The other significance of multicollinearity is that the estimators might be very sensitive to changes in sample size and elimination of "insignificant" variables. The model has also been misspecified. It does not have the correct functional form. While this is not an assumption, it still poses a problem. Due to these two violations, multicollinearity and misspecification, the estimator is no longer unbiased.

The errors in the data are not normal. Even though this assumption has been violated the estimator can still be blue. Since the disturbances are not normal, the confidence intervals and tests on the $\beta_{i,}$ for linear relationships, must be performed in alternative ways.

As for the assumption of autocorrelation, it neither passed nor failed. The test for autocorrelation was inconclusive. Therefore, nothing can be concluded about the presence of autocorrelation.

Overall, the model is not BLUE, has violated multicollinearity and normality, and is misspecified.