EDUCATIONAL FACTORS AND THEIR EFFECTS ON COLLEGE TUITION IN PRIVATE INSTITUTIONS ACROSS THE UNITED STATES

Matthew Decker

Department of Economics and Finance
School of Business and Economics
State University of New York at Plattsburgh
101 Broad St, Plattsburgh, NY 12901

May 16, 2022

Abstract

To date, educational and economic factors have caused significant variation of tuition prices of private universities for the 2019 and 2020 fiscal educational year. This paper offers a cross-sectional model observing the causation of increasing college costs across the United States with underlying support from the human capital theory of education. The analysis at hand focuses on educational and institutional variables and their effects on the associated tuition costs for only private institutions. A series of STATA econometric tests were completed in order to determine a model, which further tests were then run for deeper analysis.

Keywords: Cross-sectional, Private Universities, Increasing College Costs, Human Capital Theory, Institutional Factors, Tuition Costs

1. Introduction

Across the United States, educational costs are increasing drastically, and for most, for unknown reasons. This paper will deeply examine the factors that affect tuition costs in the, at most 20, highest-populated private college institutions per state in the United States for the 2019/2020 fiscal school year. The broad microeconomic reasoning behind my model is increasing tuition costs in most educational institutions. If the increasing price of education can be further understood, maybe its path can be altered. Private colleges were used to give more variation in the data due to multiple states having programs such as SUNY, which would result in many of the exact same data points. A similar model was tested previously allowing us to analyze these two model's results, which will be discussed further within the data section. Previous literature on this general area of study seem to underestimate the impact educational and institutional variables have on the price variation of tuition fees across the United States. A large portion of the literature is assessing the problem of accessibility of higher education, which relates to my topic since increasing educational costs is one of the main factors of this accessibility problem. Another popular dependent variable discussed within many pieces of literature review is financial aid. These literature pieces give an in-depth analysis on financial aid, but not it's direct effects with tuition costs. Thus, there is reasonable room for my research topic within this field of study.

2. Literature Review

The first piece of literature review, titled 'Factors Affecting Instructional Costs at Major Research Universities' (Brinkman, 1981) is one of the first pieces of published literatures regarding educational costs and what drives these costs up and down. Both the literature and my research question relate to the fact that the primary goal of the tests performed were to get a better

understanding of variables and their effects on educational costs. Or, as stated in the literature, "Previous studies have uncovered many of the factors that account for costs in higher education. The primary task here is to specify the relative importance of these and additional factors" (Brinkman, 1981). As mentioned, this is very similar to my research due to the independent variable of tuitionfees and Brinkman's instructional costs. Within the literature, it is stated that "the input variables in the study accounted for 71 percent of the variation in costs, with the faculty-student ratio being by far the most influential variable" (Brinkman, 1981). Although it is not the most important variable in my study, it is definitely valuable, as the p-value was 0, showing significance even at the 1% level. A minor gap in the literature is the sample of data Iused. I used all private universities for the entire United States, whereas in the literature was 'consisting of twenty-nine public and twenty-one private institutions' (Brinkman, 1981). I avoided public educational institutions due to the lack of variation in data when considering state-level programs such as SUNY, in New York state. All of the SUNY schools are going to have the same, or very similar tuition costs. Seeing that this was my dependent variable, I decided to look deeper into private institutions. Further into the literature, Brinkman describes his methodology for the tests run on his model and states that the 'The relationships among the variables were analyzed using multiple regression, an approach frequently used in similar circumstances' (Brinkman, 1981). The final significant gap in the literature that I could find, was the dependent variables used for the regressions run. My dependent variable was tuitionfees in order to directly annotate the independent variables effects on the cost of tuition at educational institutions. Within the literature, the model's '…dependent variable in the analysis that follows is the reported expenditures by institution' (Brinkman, 1981), therefore I can see that Brinkman and I had different views of what 'costs' were in our individual studies.

Within the second piece of literature, a more recent study titled 'Institutional Factors Influencing Students' College Choice Decision in Malaysia: A Conceptual Framework' (Ming, 2010) the author attempts to explain institutional factors affecting student's choice of which educational institution they will attend in Malaysia. The first, and most obvious gap in the literature are the geographical locations of the studies done. The model presented in the literature is obviously done for higher educational institutions located in Malaysia, whereas my research is for the United States. Although a gap, this isn't necessarily a bad thing, as this allows us to see if similar variables used are significant in both the United States and Malaysia. A study researched within the literature 'found that financial assistance offered by university as one of the four very important attributes expected from a particular higher education institution of choice' (Ming, 2010). Right away Ican see that the geographical difference is not significant when determining that financial aid is important in determining student's choice of educational institution. Although our dependent variables are moderately different, the largest driving factor, according to my personal research, are tuition costs of the educational institution, therefore this directly has an effect on the choices made by the students. Furthermore, my main independent variable of my model is the financial aid given by the educational institution. As my underlying research suggested, as did this piece of literature, financial aid is significant in both the price of higher education as well as the choices of which institution students choose to study at. The final significant similarity between Ming's study, and my study, is that 'There is a significant positive relationship between cost and college choice decision' (Ming, 2010). Once again, I assumed costs to be tuition prices of the educational institution and used this as the dependent variable in my model. Ming also believed that costs were one of 'the most important elements' (Ming, 2010), in the test and used it as an independent variable in the model.

Within the third piece of literature, titled 'Tuition, Financial Aid, and Access to Public Higher Education: A review of Literature' the author's goal is to analyze the "relationship between rising tuition and access to public higher education in the United States. It reviews research on the relationship between tuition and enrollment in higher education…" (Heller, 1996). Very similarly to my model, Heller is comparing the effects enrollment on tuition costs in the United States, although I have eleven different variables to more deeply analyze the direct relationship with the tuition costs rather than the accessibility of college. As seen in the previous description, the main gap in the literature once again is the overall main study or the dependent variable. As stated, my dependent variable is tuition costs, in the form of tuitionfees, and Heller's dependent variable is accessibility of educational institutions. Another similarity between my research question and the literature in the type of analysis used. Both in the literature and my research the type of analysis used 'was a state cross-sectional analysis' (Heller, 1996). These studies analyzed variable's effects on higher educational costs and accessibility for each individual state for a fixed period of time, if I were to extend this model over multiple years, I would consider the use of time series analysis.

Within the fourth piece of literature, titled 'Investing in Schooling in Chile: The Role of Information About Financial Aid for Higher Education' the author investigates the impacts of providing Chilean children with further information regarding financial aid to analyze its effects on schooling outcomes. Although this piece of literature is the furthest from my specific research question, I can still analyze similarities in the variables used in the overall model. More specifically, on page five of the literature, within the first table offered, it can be seen that the researcher used the retention rate of the schools as an independent variable in their study. Similarly, I used retention rate in order to further identify reasons tuition costs increase, and thus ended up being one of the most significant and most highly correlated variables in my model. A large gap

in the literature, and much like the Malaysia literature, is the geographic significance and the sample of the model. Due to the literature's dependent variable of accessibility, the authors 'randomly assigned over 6,000 eighth graders in 226 poor urban Chilean schools and some of their parents to receive standardized information (in the form of a short DVD program) about financial aid opportunities for higher education' (Dinkelman, 2014). The difference here is that in my model, I am testing for changes in tuition costs, whereas the authors are testing for accessibility. Therefore, it makes sense that they would narrow their research down to specific poor urban schools when analyzing accessibility. Although my research and the literature done are not very similar for this specific piece of literature, it allows there to be room for my research and can give other researchers the opportunity to use my model as a base for their completely different model, while still having similarities in the two.

Within the fifth piece of literature, titled 'Does Financial Aid Affect Institutional Aid? An Analysis of the Role of State Policy on Postsecondary Institutional Pricing Strategies' the author's goal is to determine whether or not institutions alter their pricing in order to receive a larger aid package from the state. The main similarity between this literature and my research question presented is the dependent variable once again. According to the literature, "Institutions have the ability to alter their net-price through two mechanisms, list tuition and fees and institutional aid awards" (Curs, 2010). The similarity is in the variables being used, Curs and his coauthors are using both tuition fees and financial aid awards as dependent variables within his study, whereas tuition fees is my dependent variable, and financial aid given was my main independent variable. The significant gap in the literature, once again, is the main reasoning for the study being done. My model is used to analyze the independent variable's effects on tuition costs, while the literature is being used to determine the greed of higher educational institutions in the form of alteration of

costs to further benefit from aid packages. In conclusion, the literature determined that educational institutions do not alter costs based on benefits of state financial aid packages, but 'Alternatively, individual state and Federal financial aid awards are often calculated based upon estimates of tuition, fees, and other cost of attendance measures of an institution" (Curs, 2010). This goes to show that Curs and the other authors of this literary review, believe there are other institutional 'costs of measures' that can be significant when analyzing the reasoning for variation in educational institutions' costs.

Within the sixth piece of literature, titled 'Income and Financial Aid Effects on Persistence and Degree Attainment in Public Colleges' the author deeply examines the attainability and 'effectiveness of different types of financial aid in promoting student persistence and timely bachelor's degree attainment' (Dowd, 2004). As seen in this excerpt from the literature, there are similarities in the model of the literature and my overall research topic. Student persistence is another way of saying retention rate, or the rate at which student come back to the educational institution and is a very important variable within my model. On top of this similarity in this variable for student persistence, Dowd also uses the different types of financial aid to determine the impact they have on the student retention, which is similar to my financial aid variable although in my model financial aid is the main independent variable for interpreting changing educational costs. Furthermore, within the literature, it is stated that "The effect of tuition pricing and financial aid on persistence has received increasing attention with the development of theories that assign an important role to finances in determining students' college participation decisions" (Dowd, 2004). Dowd continues to go on to say that retention, financial aid, and tuition costs are all gaining traction as the leading cost drivers of educational institutions. The main gap within the literature is once again the overall research topic being studied. As I have previously stated, I am analyzing

the effects specific variables have on the costs of tuition at educational institutions, while Dowd is doing a general analysis of degree attainment and student persistence, and the factors that affect this topic. In conclusion, although the models are different, I can determine that the literature accurately tested their model due to the highly significant correlation coefficient between my variables retentionrate and gradrate, as well as the highly significant correlation coefficient between my variables ln_finaid and retentionrate.

Within the seventh piece of literature, titled 'Econometric Studies of Higher Education' the author does a complete analysis of the econometrics of higher educational institutions, with support from the human capital theory. An immediate similarity between the literature and my research topic, is the underlying economic human capital theory, which states that the larger percentage of educated workers there are in the workforce, the more efficient the economy becomes. In a broader sense, if the trend of increasing tuition costs can be altered, the economy can become more productive. Further into the literature, it is stated that 'studies have also focused on whether the return to higher education depends upon the type of institution that an individual attends' (Ehrenberg, 2004). This is both a direct similarity, and difference in the topics at hand. First, the similarity is the use of the variable for educational institutions and whether or not their type affects the dependent variable of the study. On the other hand, Ehrenberg is analyzing the return on higher education whereas I am analyzing the effects these variables have on the costs of tuition at the educational institution. Furthermore, Ehrenberg discusses the importance of acceptance rate of the educational institution and that previous research 'found that attendance at the most selective private institutions confers extra economic advantages to students' (Ehrenberg, 2004). The level at which universities accept their students, and the overall quantity of students,

typically has a strong correlation with the costs of educational institutions or the rate of returns an educational institution brings with the completion of a degree.

Within the eighth, and final piece of literature review, titled 'Factors Affecting Net Tuition Revenue at Private Colleges. Working Paper Series' analyzes the 'Factors affecting-changes in net-tuition-revenue in private colleges' (Cohen, 1982). This literature review is the most accurate when referring to my research topic of factors affecting the cost of educational institutions tuition costs for private college in the United States. Similarly, to my model, the model in the literature review is a combination of two cross-sectional analyses over two different time periods, whereas my model for one instructional year. The greatest similarity between my research topic and the literature, is the overall topic being studied. As stated in the literature, "As the costs facing higher education continue to escalate, there is a greater need to understand and to project the revenues that institutions will have available for meeting instructional costs", as well as "there are many factors that affect tuition revenue and all of them should be considered together before instituting a new policy' (Cohen, 1982). This is the exact reasoning behind my test, as variables responsible for increasing tuition costs need to be further understood. Whether it be educational, economic, institutional, or a mixture of all three of these variables that are responsible for the changes, the tests must be done if I want to impact this field positively. Later in the literature, it is stated that "College financial aid policies are based on the assumption that student financial assistance can offset the adverse enrollment effect of raising tuition levels" (Cohen, 1982). This also goes to show that financial aid policies are extremely significant when analyzing tuition costs, hence why it is my model's main independent variable, and supports the model strongly as such. Cohen believes as well, that financial aid can be a core driving factor when analyzing educational institutions associated costs. The only gap within the literature that I noticed was the date of the study.

Obviously, this is barely a gap, but within the past ten to twenty years the importance of a higher education has increased drastically, and thus could affect lots of the data collected for the variables in the models shown. Larger economic participation directly affects the need for a higher education, and thus is why me and Cohen have developed these research papers.

In conclusion, there are many similar literature pieces, all of which have a different overall purpose to their study. Two of the above studies were for educational accessibility in foreign countries. One of the largest accessibility issues, according to my research, is the cost of higher education. Though these model's tests aim to find the affects variables have on degree attainment or overall accessibility, they use similar or even some of the same variables that were in my model, all of which can be analyzed and compared to determine similarities and differences between the literature pieces. Of the eight literature pieces compared and explained, two of them were much more accurate to my research topic, rather than those that are analyzing the degree attainment of educational institutions or the overall accessibility of higher education. Of these two, there were many similarities to my topic and model, from the variables, to the dependent variable of the model, to the general framework of the models tested, and one even down to the underlying economic theory of the study that's driving our research question. The significant gaps in the literature were mainly present in the studies that primarily focused on degree attainment or general accessibility, rather than the educational institution's costs.

3. Empirical Model and Estimation

The empirical model estimated for this research question is a right-hand semi-log model. The actual equation of this model, after attempting a variation of a linear, a double log, and a semi-log model is;

$$TuitionFees_i = \beta_0 + \beta_1 lnFinAid_i + \beta_2 lnFamilyIncome_i + \beta_3 GradRate_i + \beta_4 lnEndowment_i + \beta_5 StudentFacultyRatio_i + \beta_6 AcceptanceRate_i + \beta_7 RetentionRate_i + \beta_8 FinType_i - \beta_9 AreaType_i + \beta_{10} SchoolType_i + \beta_{11} lnEnrollment_i + \beta_{12} StateID_i + \beta_{13} GDPpc_i + \varepsilon_i$$

To determine the right-hand semi-log was the best model, a series of regression analyses were performed. The model which produced the highest R-squared and Adjusted R-squared, as well as had the most significant variables, was then chosen as the model best fitted for this socio-economic issue. A series of variables were used to determine the impact they have individually and as a whole on this model. The variables and their descriptions are shown in the table below. Additionally, a robustness check was needed to correct for heteroskedasticity present in my model.

Variable	Description
TuitionFees	The average cost of attending the educational institution per year, for 2019/2020 fiscal year
FinAid	The amount of financial aid the educational institution awards their students
FamilyIncome	The average family income of the County in which the educational institution is located
GradRate	The average graduation rate of the educational institution, in percent format calculated by dividing the country's total GDP by its population.
Enrollment	A variable that shows the total number of students attending the educational institution in the chosen year.
Endowment	The amount of endowment the educational institution receives
StudentFacultyRatio	The number of students divided by the number of faculty of the educational institution
AcceptanceRate	The average rate at which students are accepted into the educational institution
RetentionRate	The rate at which students return after their first year at the educational institution
GDPpc	A variable that shows the GDP per capita of each state for 2019
FinType	A dummy variable for the financial type of the educational institution, being independent (not for profit) (0) or proprietary (for-profit) (1)

AreaType	A dummy variable for whether the educational institution is in a rural (0) or urban (1) area
SchoolType	A dummy variable to determine whether the educational institution is specialized (1) or comprehensive (0)

4. Data

Shown below is the Descriptive Statistics table of the model. The descriptive statistics summarizes or describes the characteristics of a data set.

Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
tuitionfees	754	29286.719	13648.05	1776	59100
gradrate	750	.557	.208	.02	1
acceptancerate	740	.698	.234	.043	1
ln enrollment	748	7.54	1.406	3.045	11.64
retentionrate	745	.795	1.629	0	1
ln endowment	612	18.037	1.958	9.687	24.434
ln familyincome	747	11.049	.264	10.062	12.576
ln finaid	762	16.672	1.824	8.263	21.016
studentfacultyratio	747	12.798	5.39	3	77
schooltype	758	.148	.355	0	1
areatype	767	.737	.441	0	1
fintype	767	.116	.32	0	1
stateid	767	24.014	14.184	1	50
gdppc	767	54389.325	9919.113	35015	75258

For the tuition fees variable, the mean is 29286.719, meaning the average costs of all private college institutions is $29,286.72, the standard deviation is 13648.05, meaning the average

of all tuition costs are within $13,648.05 of the mean, the minimum value is 1776 and the maximum value is 59100, thus there are schools with yearly costs ranging from $1,776 to $59,100.

The variable for the institution's graduation rates has a mean of 0.557 meaning on average, 55.7% of all students enrolled graduate from the private institution. The standard deviation of gradrate has a value of .208, thus all of the recorded graduation rate data is within 20.8% of the mean. Graduation rate also had minimum and maximum values of .02 and 1, meaning that the smallest percentage of graduating students from a university was 2% while the largest was 100%.

A variable for the acceptance rate of the educational institutions has a mean of .698, meaning that the average percent at which a student is accepted into a private institution in the United States is 69.8%. The standard deviation of the acceptance rate is .234, therefore all acceptance rates are within 23.4% of the mean. The minimum value of acceptance rate is 0.043 while the maximum value is 1.0, thus the most difficult school to get accepted into had a rate of 4.3% while the easiest educational institutions to get into had a rate of 100%.

For the logged variable of the educational institution's total enrollment for the chosen years, the mean value was 7.54. This shows that the average educational institution had approximately 4470.79, and after the log is added I get the value as shown above. The standard deviation of the enrollment variable is stated as 1.406. Thus, all of the enrollment values are within 1.4% of the mean. The logged minimum number of students enrolled in a private university is 3.045, while the logged maximum amount of student enrolled is 11.64.

The retention rate variable has a mean value of 0.795, showing that 79.5% of the students who attended the institution returned the following year. The standard deviation for retention rate is 1.629, thus all recorded data for retention rates are within 1.63% of the mean of the data. The minimum retention rate was measured at 0.00 while the maximum was measured at 1.00. This

shows that the minimum return rate for a private educational institution's students was 0% while the maximum return rate was 100%.

The logged variable for the educational institution's endowment for those years had a mean of 18.037. This shows the educational institutions received an average of $537,000,000 but due to the large ranges of this variable's data, a log was used and is shown in the summary table above as 18.037. The standard deviation was 2.2758 showing that all logged endowment values are within 2.28% of the logged mean. The minimum and maximum values for the endowment variable were 12.59811 and 23.12116, this range is much more manageable than the model without the logged variables.

For the logged variable of FamilyIncome, the mean is 11.049. Prior to the log transformation, the average family income had a mean of $65,210, similarly with every logged variable, the range of values were too large, thus taking the log giving us 11.049. The standard deviation of the familyincome variable is 0.264, meaning that all family's incomes are within 0.26% of the mean. This variable also has a minimum value of 10.63181, and a maximum value of 11.68227, which is a much more manageable range of values.

For the logged variable of FinAid, the mean is 16.98725 and prior to log transformation, approximately $49,000,000 in financial aid is given by the institution on average. The standard deviation is 1.824, showing that all logged finaid data points are within 1.824% of the logged mean values. This variable also has a minimum value of 8.263, and a maximum value of 21.016, which is a significantly smaller range than prior to the log transformation.

The variable for student to faculty ratio has a mean value of 12.798, meaning that on average, for every individual faculty member there are 12.8(13) students. The standard deviation for the student to faculty ratio is 5.39, thus all values in the dataset are within 5.39% of the

variable's mean. The minimum and maximum values for the student to faculty ratio are 3 and 77, showing that the minimum students per faculty 3 while the maximum number of students per faculty is 77.

The dummy variable for determining whether the school was a specialized or comprehensive university had a mean of .148, thus 14.8% of the schools are comprehensive while the remaining percent must be specialized schools. Due to its float data type, the minimum and maximum values are 0 and 1, and the regression tests for whichever of the two choices is valued at 1.

The dummy variable for the type of area the educational institution is located in, whether it be an urban or a rural area, has a mean of .737. This shows that approximately 74% of the public institutions researched, were located within urban areas, while the remaining were located in rural areas. This dummy variable also has a standard deviation of .441, thus all of the data is within 44.1% of the mean which makes sense to its 50/50 float nature. This dummy variable has a minimum and maximum value of 0 and 1 once again, due to its float data type.

The dummy variable for the financial type of the educational institution has a mean of 0.116, thus 11.6% of private colleges are for-profit. Therefore, there are significantly more not-for-profit universities across the United States. The standard deviation of this variable is 0.32 meaning that all of the data is within 32% of the mean. Not surprisingly, this number should be lower seeing that the financial types of these educational institutions are so heavily skewed toward not-for-profit.

For my final, and newest addition to the model, a more impactful state-level variable was added in the form of GDPpc. The mean of the gdppc variable is 54389.235, thus the average gross domestic product per capita of the United States is $54,389.24. The standard deviation of this state-

level variable is 9919.113. Thus, all GDPpc values within the data are within $9,913.11 of the mean value. The state with the highest GDPpc value was Massachusetts, with a GDPpc value of 75258, which represents the maximum value within the summary. The state with the lowest GDPpc is Missouri, with a value of 35015, which is also representative of the minimum value within the summary.

Pairwise Correlations

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
(1) tuitionfees	1.000													
(2) gradrate	0.725	1.000												
(3) acceptancerate	-0.489	-0.407	1.000											
(4) ln_enrollment	0.501	0.344	-0.305	1.000										
(5) retentionrate	0.001***	-0.012**	-0.106	-0.010***	1.000									
(6) ln_endowment	0.683	0.671	-0.417	0.645	0.608	1.000								
(7) ln_familyincome	0.157	0.124	0.083*	0.103	-0.074*	0.171	1.000							
(8) ln_finaid	0.680	0.455	-0.387	0.873	-0.010***	0.732	0.018**	1.000						
(9) studentfaculty~o	-0.256	-0.213	0.243	0.154	-0.015**	-0.22	-0.026**	0.054*	1.000					
(10) schooltype	-0.153	-0.019**	0.115	-0.271	-0.012**	-0.10*	0.142	-0.336	-0.033**	1.000				
(11) areatype	0.119	0.112	0.035**	0.237	-0.047**	0.23*	0.382	0.148	0.069*	0.037**	1.000			
(12) fintype	-0.350	-0.223	0.344	-0.286	-0.038**	-0.090*	0.138	-0.404	0.275	0.200	0.096*	1.000		
(13) stateid	0.054*	0.060*	0.088*	-0.101	-0.052*	-0.11*	-0.076*	-0.030**	-0.073*	-0.007***	-0.071*	-0.095*	1.000	
(14) gdppc	0.313	0.249	-0.030**	0.265	-0.022**	0.32*	0.534	0.198	-0.101	0.076*	0.218	-0.009***	-0.044**	1.000

**** p<0.01, ** p<0.05, * p<0.1*

Within this section, I will also include a descriptive analysis of the correlation matrix, which displays all of the pairwise correlation coefficients between all of the variables in the dataset.

In reference to the correlation matrix shown above, the closer the correlation coefficients are to 1, the higher the correlation between the two variables. Any value labeled with an asterisk shows that it is significant at the 10% level, two asterisks show significance at the 5% level, and any value with three asterisks show the strongest significance at or below the 1% significance level. The level of significance tested for can also be altered within STATA to only show the correlations at the specified levels rather than all three together.

When analyzing the correlation matrix further, there are many correlation coefficients with highly significant values as well as highly correlated values. More specifically, there are 19 correlation coefficients that are significant at the 10% level, there are 15 correlation coefficients significant at the 5% level, and there are 5 correlation coefficients highly significant at the 1% level. The correlation coefficient between the variable retentionrate and the variable tuitionfees is highly significant at the 1% level with a value of 0.001. This shows a significant relationship between the two variables, which makes sense, the number of students returning directly has an effect on tuition costs. If less students return, the educational institution must make up for this deficit by increasing prices. Another highly correlated coefficient between two variables is shown between enrollment and retentionrate, which once again is logical. The number of students returning would clearly cause variation in the number of students enrolled within the educational institution. The final highly correlated coefficient between two variables is shown between retentionrate and ln_finaid, with a value of -0.010. The amount of financial aid given by educational institutions is dependent on their enrollment, making this pair very significant in the model. Although they have a negative correlation this isn't necessarily a problem, this just shows

that as retentionrate increases, ln_finaid decreases, this stays true conversely as well. Or in real terms, the more students that return, the less financial aid available for additional students.

On top of these highly significant correlation coefficients, there are also many highly correlated variables. A correlation of -1.0 indicates a perfect negative correlation, and a correlation of 1.0 indicates a perfect positive correlation. The closer to 1 the coefficient's value is, the stronger the relationship between the variables. Further analysis of the correlation matrix shown above uncovers many moderately strong correlation coefficients, which I would consider having a value that would be greater than 0.50.

A strong correlation can be seen between the variables ln_finaid and ln_enrollment, this being the strongest relationship of the entire correlation matrix with a value of 0.873. These variables are very highly correlated due to the nature that educational institutions financial aid given is highly dependent upon the number of students the institution enrolls. Another highly correlated pair of variables are ln_finaid and tuitionfees. Their correlation coefficient was valued at 0.68. These two variables are my main independent variable and my dependent variable, so I expected to see a high correlation between these variables. Applied to a real situation the higher the tuition fees of the educational institution, the more financial aid given across the educational institution, making these variables highly correlated.

Data collected for this model was done through the following informational websites; datausa.io to find most of the educational data, nces.ed.gov for the remainder of the educational variables, and statista.com for the GDP per capita variable data.

As briefly mentioned in the introduction, my previous semester I analyzed a broader version of this paper's research question, being institutional/educational and economic effects on tuition costs in private educational institutions across New York state for the same time period.

Extending this model to the entire United States allows me to interpret and compare the tests of the two models and determine if there is an overall better model. Both models focused on the right-hand semi-log empirical model, as it yielded the best results out of the 4 models tested for.

When first comparing the summary statistics, the significant change in the model is the number of observations per variable. The test for New York state yielded a maximum of 154 observations, whereas the current model has an upwards limit of 767 observations. Apart from the number of observations in the model, the current model has more accurate data when comparing standard deviations. The current models' variables had standard deviations that yielded data to the closer to the mean, showing a higher correlation between the variables and the model.

In reference to the correlation matrix for the current model, there are generally more highly correlated coefficients between variables than there were in the previous model. The largest coefficient between variables in the old model was 0.6323, where the largest coefficient in the current model was 0.873. When referring to the significance of correlation coefficients, the previous model was tested at the 5% level and yielded a larger quantity of significant coefficients between variables. The current model was tested at the 1%, 5%, and 10% levels and has a smaller number of significant coefficients, but those that are significant at the 1% level, are much stronger than those that are significant at the 5% level. The larger number of observations in the current model give more validation to the data and gives overall better results, showing stronger correlations and more significance between variables.

When analyzing the two regression analyses on both models, the current model has a much stronger R-squared value at 0.79 versus the previous model's 0.66. Using the coefficients, I can determine whether or not a variables' impact on the dependent variable is accurate or not. The previous model had a few coefficients that were not accurately represented, for example

retentionrate in the previous model increased tuitionfees by approximately $22,000, which definitely doesn't seem like an accurate representation. The increase in the number of observations in the current model gives the retentionrate variable a much more accurate value of $6,114. This increase in tuition costs is more reasonable and, as explained above, retentionrate is very significant in the model. Looking more closely at the variables, this semester I added more state-level data in the form of the GDP per capita per state in the United States. This type of data was unnecessary when testing for only New York, as all of the data would be the exact same. I also dropped the ZeroFinAid variable. Within New York state, there are multiple institutions that do not offer financial aid to their students, thus the variable was included. Since the current model was the highest populated private educational institutions, there were no universities that did not offer financial aid to their students. When interpreting the two model's p-values, the main independent variable, finaid, was highly significant in both tests, but was slightly more significant in the current model. Also, the endowment variable was previously highly significant, but with the addition of observations, was transformed in the current model, into an insignificant variable. This shows that endowment educational institutions receive is far more significant in private universities in New York state, rather than across the United States. Apart from these minor differences in the regressions, the current model is by far the more reliable model for testing.

Regarding other tests done, neither model had significant levels of multicollinearity as tested with the vif command in Stata. The previous model also had no traces of heteroskedasticity, while the current model does (will be further mentioned in the next section). Finally, an ovtest for omitted variables was done. In result, both models have no omitted variables, but both have non-linearity present.

5. Empirical Results

Right hand semi-log regression

tuitionfees	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
gradrate	19513.865	2265.101	8.62	0	15065.144	23962.585	***
acceptancerate	-4688.687	1282.656	-3.66	0	-7207.858	-2169.516	***
ln_enrollment	-4956.167	441.291	-11.23	0	-5822.874	-4089.46	***
retentionrate	6114.546	2998.742	2.04	.042	224.933	12004.158	**
ln_endowment	278.095	226.956	1.23	.221	-167.653	723.843	
ln_familyincome	4159.804	1163.283	3.58	0	1875.084	6444.523	***
ln_finaid	7800.456	445.393	17.51	0	6925.692	8675.22	***
studentfacultyratio	-651.727	89.087	-7.32	0	-826.697	-476.758	***
schooltype	1349.518	866.669	1.56	.12	-352.644	3051.68	
areatype	-899.877	606.483	-1.48	.138	-2091.026	291.273	
fintype	-2369.626	3434.537	-0.69	.491	-9115.151	4375.898	
stateid	19.546	17.697	1.10	.27	-15.212	54.303	
gdppc	.109	.031	3.51	0	.048	.17	***
Constant	-124513.71	13036.946	-9.55	0	-150118.63	-98908.795	***

Mean dependent var	32882.035	SD dependent var	12604.787
R-squared	0.790	Number of obs	599
F-test	169.635	Prob > F	0.000
Akaike crit. (AIC)	12102.399	Bayesian crit. (BIC)	12163.932

*** p<.01, ** p<.05, * p<.1*

As shown in the regression done above, on the right-hand semi-log empirical formula, I can analyze the data further than the correlation matrix and overall summary statistics of the model.

Analyzing the regression table above, the number of observations within this regression is 599. The Prob > r, or the P-value, is statistically significant at the 1%, 5%, and 10% level with a value of 0.0000 or less than 1%. This means that the overall model is significant, and the null hypothesis can be rejected.

The R-Squared and adjusted R-Squared values were overall superior to the previously tested model with values of 0.6642 and 0.6342. The R-Squared is measured as the proportion of variance within the TuitionFees, or dependent variable, which is predicted by the independent variables and is the overall strength of the association of the model. The adjusted R-Squared is

measured as predictors added to the model explaining some of the variance in the dependent variable or more specifically with the equation $1 - ((1 - Rsq) * ((N - 1) / (N - k - 1))$. The larger the number of observations and the smaller the number of predictors, the larger the difference between the R-squared and the adjusted R-squared values.

The F-value is the mean square model divided by the mean square residual yielding 169.635. Root MSE is measured at 5835.50, being the standard deviation of the error term. SS is the sum of squares associated with the total, model, and residual of the dataset. The MS is defined as the mean squares, or the sum of squares divided by their degrees of freedom for the model.

For the second half of the regression, I can use the coefficient values to interpret the variable's effect on the dependent variable as described below (All of the below situations are possible, if all other independent variables remain constant).

As the graduation rate increases by one unit, the tuition price increases by $19,513.87. The higher percentage of graduates there are within a university, the better the education given by the school, thus increasing the cost of tuition. Although this increase is a bit extreme, the higher the graduation rate of the school, the more dependable the educational institution is. Thus, making these higher tuition fees an opportunity costs for a better chance at graduating with a higher education.

As acceptance rates increase by one unit, the tuition costs associated decrease by $4,688.69 due to the association that the higher the acceptance rate, the lesser quality the education received is when compared to other institutions, making it more affordable. Educational institutions with a large acceptance rate will typically have lower costs also due to the larger number of students that will be accepted to study there, in hand, lowering costs due to higher quantity.

Analyzing the coefficient for enrollment, as enrollment increases by one percent, the associated tuition costs decrease by approximately $4,956.17/100. This is accurate, as the number of students is directly correlated with the costs associated with attending the educational institution. The higher quantity of students, the less the institution has to charge due to the number of students paying. Conversely, if the number of students is lower, the costs have to make up for the deficit of students with increased costs.

As retention rate increases by one unit, the associated tuition fees increase by $6,114.55. This is accurate due to the fact that returning students would directly have an effect on tuition prices. The less students that return the following semester, the more money the educational institution is losing, in which they respond by increasing costs.

As the endowment of the school increases by one percent, the price of tuition increases by $278.10/100. This is an accurate representation due to endowment being the amount of money a university has available to spend on educational resources, thus the more they have to spend, the higher quality of learning received, leading to an increase in tuition costs. Students are willing to pay higher tuition costs if it means they get a higher quality education, with higher quality materials.

As family income increases by one percent, the tuition price increases $4,159.80/100. Respectively, the more money a family makes, the more expensive the tuition will be due to financial awards based on family income. Students fall under a certain bracket dependent on their household incomes, all of which have different tuition rates.

As financial aid is increased by one percent the price of tuition increases by $7,800.46/100. This makes sense, seeing that the more financial aid the university has to offer usually implies a higher cost of tuition. A majority of financial aid is given to pay for the bulk of college expenses,

being tuition costs. The higher the initial tuition costs for the institution, the more financial aid students will require to successfully and comfortably live throughout their higher education.

As the number of students increases compared to the faculty, or the student to faculty ratio, increases by one unit, the associated tuition decreases by $651.73. The more students a campus has compared to the amount of faculty shows a decrease due to the fact that there are less faculty that need to be paid by the institution and more students paying to study there. Conversely, if there are more faculty and less students, the tuition will increase due to the increased amount of salary wages being paid. Furthermore, the decrease in the amount of student paying to study at the educational institution will result in an increase in tuition costs to recover the deficit accumulated form the decrease in enrollment.

As the variable GDP per capita, or GDPpc, increases by one unit, the associated tuitionfees increase by 10.9%. This makes sense due the association of economic success in the state, with the success of the residents within the state. If the economy in a particular state is thriving, it is likely that there are more higher paying jobs available and as mentioned above, the family income of the student is highly significant when determining tuition costs associated with the student. According to the definition of GDP per capita, GDP per capita considers the reflection of such economic health into an individual citizens perspective. Thus, if these individual citizens have more money, educational institutions can get away with increasing tuition prices.

Next, I can annotate the coefficients of the dummy variables in the model. Dummy variables have a float data type, allowing them to be ‘this or that’ or ‘true and false’, all of which take the form of 0 or 1.

The dummy variable for the type of school the institution is, whether it be a specialized institution or a comprehensive institution, shows that if the institution specializes in a certain field

of study, the tuition costs is $1,349.52 higher than institutions that do not speciliaze. Many of these specialized institutions focus on a particular field of study rather than offering a range of educational focuses. Most of these institutions are typically high cost and have high return on investment. Specializations vary from studying law to become a lawyer, to studying medicine to become a doctor, both of which are very important fields of study. These specialized institutions typically have less students and more staff who of which have higher credentials, automatically reducing the amount of money the institution can make, due to larger wage payments. This, in hand, directly causes an increase in the cost of tuition.

The dummy variable for the type of area the institution is located in, shows that if the educational institution is located within an urban area, the tuition fees associated are $899.88 lower than those for institutions in rural areas. Meaning that educational institutions in rural areas are more likely to have a higher tuition cost than that of rural areas. Typically, urban areas have a larger population, meaning more people able to easily attend the educational institution in these areas. And as stated above, a higher enrollment of the educational institution will drive costs down due to the quantity of students paying.

The dummy variable for the financial type of the educational institution shows that if the school is proprietarily and not independently funded, then tuition price will be by $2,369.63 lower than the tuition for independently funded schools. This shows that, if the institution is for-profit, then the overall tuition fees will increase by the stated amount. This is due to the fact that for-profit institutions typically get no financial help apart from donors, as they are a business trying to make money. Whereas not for profit institutions get financial support from the federal and local governments. The less support the institutions are receiving, the higher the costs will be driven to compensate for this lack of financial support.

The standard error column shows the standard errors associated with the coefficients previously talked about. This is calculated by testing whether the parameter is significantly different from 0 by dividing the parameter estimate by the standard error in order to receive the t-values, which are significant if the value is less than +/- the absolute value of 2 for the pre-selected alpha of 5%. Therefore, looking at the regression above, the variables ln_endowment, areatype, and stateid are significant at the 5% level. Considering the t-value for retentionrate is 2.04, I can also conclude that this is significant, as it is still extremely close to 2. Based on t-values, the rest of the variables aren't quite close to 2, meaning they're insignificant in this test, but not unimportant by any means.

Moving on to the p-values shown in the regression demonstrated above. The p-value shows the overall significance of the variable within this regression. At the chosen 1% alpha level, the variables gradrate, acceptancerate, ln_enrollment, ln_familyincome, ln_finaid, studentfacultyratio, and GDPpc are all statistically significant, meaning that they are all smaller than the 1% significance level. Thus, you can reject the null hypothesis. Apart from variables significant at the 1% level, the variable retentionrate is significant at the 5% level. Otherwise the remaining insignificant variables wouldn't even pass at a 10% significance level due to their values being greater than 0.10. Apart from individually annotating them, the overall model is significant at the 1%, 5%, and 10% levels with a value of 0.0000.

Variance inflation factor

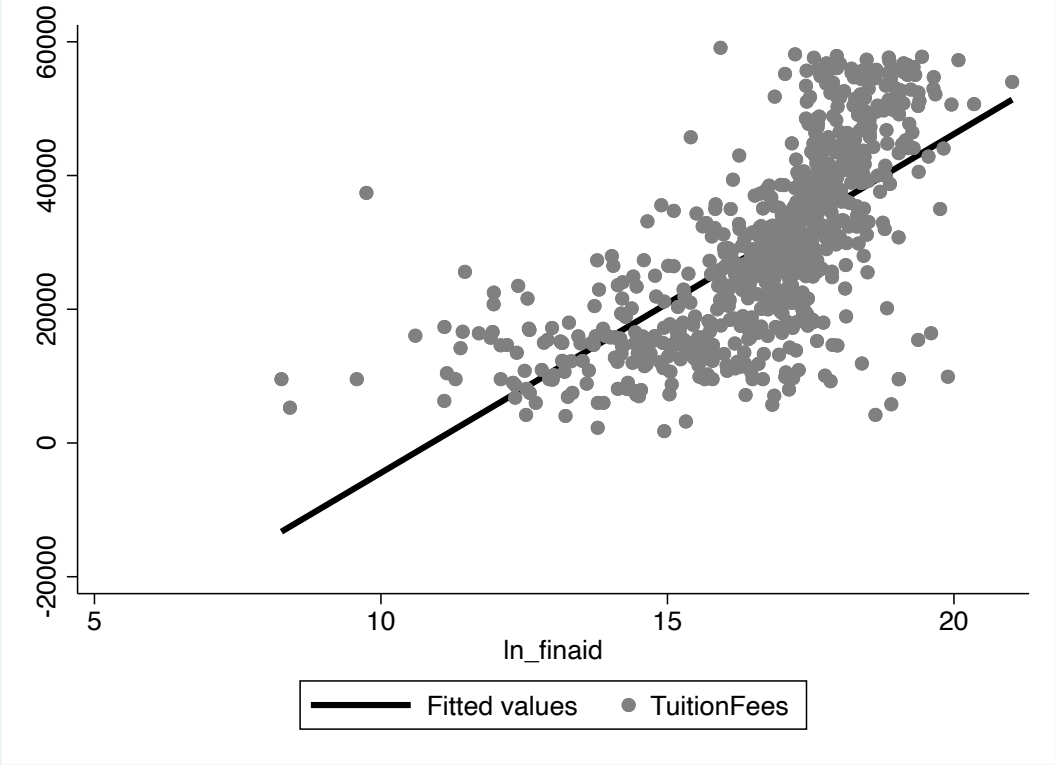
	VIF	1/VIF
ln finaid	5.621	.178
ln enrollment	4.216	.237
ln endowment	3.512	.285
gradrate	3.106	.322
retentionrate	2.469	.405
ln familyincome	1.64	.61
gdppc	1.637	.611
studentfacultyratio	1.541	.649

acceptancerate	1.405	.712
areatype	1.296	.771
schooltype	1.156	.865
stateid	1.072	.933
fintype	1.034	.967
Mean VIF	2.285	.

As shown above, a test was run for finding multicollinearity within the dataset. The estat vif command calculates the variance inflation factors for the independent variables in the model. The VIF is the ratio of variance in a model with multiple independent variables, compared to a model with only one independent variable. Having multicollinearity makes interpreting the coefficients more difficult than it typically would be due to the reduction of the power of the model. As seen within the test, there was a mean VIF of 2.285, which shows that there is very little to no multicollinearity within the model. A general rule of thumb is that if the mean VIF is greater than 10, there is high multicollinearity. Not only is there little to no multicollinearity in the entire model, individually the VIF values for the variables were all also significantly less than 10. Overall having multicollinearity of 2.285 is nothing that needs to be fixed within the model. If the VIF values were over 10, that would show that the independent variables aren't truly independent, and tests would need to be done to resolve this issue.

Breusch,Pagan/Cook,Weisberg test for heteroskedasticity

Assumption: Normal error terms
Variable: Fitted values of tuitionfees
H0: Constant variance
chi2(1) = 17.97
Prob > chi2 = 0.0000



Heteroskedasticity is caused by a systematic change in the variance of residuals over a range of measured values, that goes unnoticed by the system and can alter p-values. From the above Breusch-Pagan and Cook Weisberg test and the associated graph, it is evident that there is a minor heteroskedasticity problem. Visually on the graph, between the fitted values 0 and 4000 a couple of outliers can be seen that would potentially cause heteroskedasticity within the model, as well as looking at the test, the overall P-value of the test was significant with a value of 0.0000, which is much lower than the 5% chosen significance level. Typically, if the P-value of the test falls under the chosen alpha then there is evidence of heteroskedasticity and as shown, I was able to reject the null hypothesis giving us evidence that there are minor traces of heteroskedasticity in this model.

Linear regression
Robustness check

tuitionfees	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
gradrate	19513.865	3239.264	6.02	0	13151.862	25875.867	***
acceptancerate	-4688.687	1379.117	-3.40	.001	-7397.31	-1980.064	***
ln_enrollment	-4956.167	530.714	-9.34	0	-5998.505	-3913.83	***
ln_endowment	278.095	255.383	1.09	.277	-223.484	779.674	
ln_familyincome	4159.804	1203.697	3.46	.001	1795.71	6523.897	***
ln_finaid	7800.456	601.406	12.97	0	6619.278	8981.633	***
retentionrate	6114.546	4351.317	1.41	.16	-2431.56	14660.652	
studentfacultyratio	-651.727	131.41	-4.96	0	-909.82	-393.634	***
schooltype	1349.518	903.725	1.49	.136	-425.423	3124.459	
areatype	-899.877	617.637	-1.46	.146	-2112.933	313.18	
fintype	-2369.626	4924.239	-0.48	.631	-12040.967	7301.714	
stateid	19.546	18.747	1.04	.298	-17.275	56.366	
gdppc	.109	.033	3.28	.001	.044	.174	***
Constant	-124513.71	13725.97	-9.07	0	-151471.89	-97555.533	***

Mean dependent var	32882.035	SD dependent var	12604.787
R-squared	0.790	Number of obs	599
F-test	153.226	Prob > F	0.000
Akaike crit. (AIC)	12102.399	Bayesian crit. (BIC)	12163.932

*** *p<.01,* ** *p<.05,* * *p<.1*

As mentioned previously, the model has minor traces of heteroskedasticity. To fix these minor traces of heteroskedasticity a robustness check regression is performed. A robust regression is an iterative procedure that seeks to identify outliers and minimize their impact on the coefficient estimates. The amount of weighting assigned to each observation in robust regression is controlled by a special curve called an influence function. Overall this test leads to slightly different results due to the fact that it excludes the outliers of the data in the test. After the robust regression was done, it can be seen that more of the individual variables p-values' became insignificant in the model, while others barely changed at all.

Ramsey RESET test for omitted variables
Omitted: Powers of fitted values of tuitionfees
H0: Model has no omitted variables
F (3, 582) = 27.01
Prob > F = 0.0000

After the minor heteroskedasticity problem was solved, an OV test, or omitted variable test was completed in order to find omitted variables within the model. This Ramsey RESET test can be used to look for two types of misspecification. It can look for either omitted variables, or functional form misspecification, depending on the options you choose when you run the test. I chose to test for omitted variables and the result, as shown above, is that there are no omitted variables within the model, but instead there is non-linearity present as shown by the significant P-value. Non-linearity is shown when there is not a straight-line or direct relationship between the independent variable and a dependent variable. While attempting to fix this issue, I transformed all of my variables into logged versions attempting to force the relationships, and still received an answer less than the alpha tested for. Thus, the right-hand semi-log model yielded the best results even with the non-linearity present.

Linear Regression
Linear model

tuitionfees	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
gradrate	33324.068	2651.087	12.57	0	28117.278	38530.857	***
acceptancerate	-6285.836	1675.581	-3.75	0	-9576.711	-2994.961	***
enrollment	-.129	.044	-2.91	.004	-.217	-.042	***
retentionrate	14790.226	3725.752	3.97	0	7472.772	22107.68	***
endowment	0	0	-1.47	.143	0	0	
familyincome	.041	.019	2.15	.032	.003	.078	**
finaid	0	0	4.84	0	0	0	***
studentfacultyratio	-427.071	105.131	-4.06	0	-633.549	-220.592	***
schooltype	-2784.924	1055.901	-2.64	.009	-4858.736	-711.112	***
areatype	195.192	736.992	0.26	.791	-1252.275	1642.659	
fintype	2299.433	4460.208	0.52	.606	-6460.507	11059.373	
stateid	34.195	22.589	1.51	.131	-10.17	78.56	
gdppc	.146	.038	3.80	0	.071	.222	***
Constant	-832.718	3283.887	-0.25	.8	-7282.339	5616.903	

Mean dependent var	32841.748	SD dependent var	12632.862
R-squared	0.665	Number of obs	600
F-test	89.360	Prob > F	0.000
Akaike crit. (AIC)	12406.958	Bayesian crit. (BIC)	12468.515

*** $p<.01$, ** $p<.05$, * $p<.1$

For the very first model, I attempted to use a linear model. The equation of this model would be as shown:

$$TuitionFees = \beta_0 + \beta_1 FinAid_i + \beta_2 FamilyIncome_i + \beta_3 GradRate_i + \beta_4 Endowment_i - \beta_5 StudentFacultyRatio_i - \beta_6 AcceptanceRate_i + \beta_7 RetentionRate_i + \beta_8 FinType_i - \beta_9 AreaType_i + \beta_{10} SchoolType_i - \beta_{11} Enrollment_i + \beta_{12} StateID_i + \beta_{13} GDPpc + \varepsilon_i$$

Overall, these results were not as good as the right-hand semi-log model actually chosen. One of the most important aspects to look at is the adjusted R-squared value as well as the model's significance. As seen in the regression for the linear model above, the adjusted R-squared value is much smaller at 0.665, while the adjusted R-squared for the model used was 0.79. Both of these regressions yielded the same overall significance and show that the model itself is significant, but there are other aspects which show the strength of the regression, such as the significance of variables. Although the schooltype variable was significant in this linear model and insignificant in the chosen model, this single variable isn't more important than the difference in the R-squared between the two regressions.

Linear regression
Left hand semi-log model

ln_tuitionfees	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
gradrate	1.311	.117	11.18	0	1.081	1.542	***
acceptancerate	-.169	.074	-2.28	.023	-.315	-.024	**
enrollment	0	0	-3.47	.001	0	0	***
retentionrate	.301	.165	1.83	.068	-.023	.625	*
endowment	0	0	-1.79	.074	0	0	*
familyincome	0	0	1.78	.075	0	0	*
finaid	0	0	3.43	.001	0	0	***
studentfacultyratio	-.013	.005	-2.88	.004	-.023	-.004	***
schooltype	-.138	.047	-2.95	.003	-.23	-.046	***
areatype	.024	.033	0.74	.458	-.04	.088	
fintype	.201	.197	1.02	.31	-.187	.588	
stateid	.001	.001	1.50	.135	0	.003	
gdppc	0	0	2.83	.005	0	0	***
Constant	9.168	.145	63.08	0	8.883	9.454	***

Mean dependent var	10.306	SD dependent var	0.475
R-squared	0.535	Number of obs	600
F-test	51.836	Prob > F	0.000
Akaike crit. (AIC)	372.324	Bayesian crit. (BIC)	425.087

*** $p<.01$, ** $p<.05$, * $p<.1$

The second empirical formula tested was the left-hand semi-log and the results are given above. The formula for this model is given as;

$$lnTuitionFees = \beta_0 + \beta_1 FinAid_i + \beta_2 FamilyIncome_i + \beta_3 GradRate_i + \beta_4 Endowment_i - \beta_5 StudentFacultyRatio_i - \beta_6 AcceptanceRate_i + \beta_7 RetentionRate_i + \beta_8 FinType_i - \beta_9 AreaType_i + \beta_{10} SchoolType_i - \beta_{11} Enrollment_i + \beta_{12} StateID_i + \beta_{13} GDPpc + \varepsilon_i$$

This model is overall significant, much like the current chosen empirical model, but with an adjusted R-squared value of 0.535, which is significantly lower than the right-hand semi-log formula's value of 0.79. This model was the overall weakest out of the four tested, as only about 54% of the variance of the dependent variable is explained by the models' independent variables. Furthermore, there are less highly significant p-values than the chosen model. Many of the variables are significant at the 10% level whereas in the right-hand semi-log these variables are significant at the 1% level, making the left-hand semi-log a worse overall model.

Linear Regression
Double log model

ln_tuitionfees	Coef.	St.Err.	t-value	p-value	[95% Conf	Interval]	Sig
gradrate	.701	.1	7.03	0	.505	.896	***
acceptancerate	-.085	.056	-1.51	.131	-.196	.026	
ln_enrollment	-.228	.019	-11.75	0	-.266	-.19	***
retentionrate	-.092	.132	-0.69	.488	-.35	.167	
ln_endowment	-.001	.01	-0.11	.914	-.021	.019	
ln_familyincome	.166	.051	3.25	.001	.066	.267	***
ln_finaid	.355	.02	18.14	0	.317	.394	***
studentfacultyratio	-.025	.004	-6.42	0	-.033	-.017	***
schooltype	.045	.038	1.18	.239	-.03	.12	
areatype	-.027	.027	-1.02	.31	-.079	.025	
fintype	-.084	.151	-0.56	.577	-.381	.212	
stateid	.001	.001	1.34	.181	0	.003	
gdppc	0	0	1.91	.057	0	0	*
Constant	4.031	.573	7.03	0	2.905	5.156	***

Mean dependent var	10.308	SD dependent var	0.472
R-squared	0.711	Number of obs	599
F-test	110.931	Prob > F	0.000
Akaike crit. (AIC)	84.033	Bayesian crit. (BIC)	145.566

*** p<.01, ** p<.05, * p<.1

The third and final empirical model tested was the double log model, with the results shown above. The equation for this model can be given as;

$$lnTuitionFees = \beta_0 + \beta_1 lnFinAid_i + \beta_2 lnFamilyIncome_i + \beta_3 GradRate_i + \beta_4 lnEndowment_i - \beta_5 StudentFacultyRatio_i - \beta_6 AcceptanceRate_i + \beta_7 RetentionRate_i + \beta_8 FinType_i - \beta_9 AreaType_i + \beta_{10} SchoolType_i - \beta_{11} lnEnrollment_i + \beta_{12} StateID_i + \beta_{13} GDPpc + \varepsilon_i$$

This was the next best model, given the R-squared is measured at 0.711 compared to our model's R-squared of 0.79. Overall, this would be the next best model to use given the R-squared for the double log model is close to the right-hand semi-log model and would most likely give similar results. Although the R-squared was decently high, and overall close to the current model's R-squared, the number of variables that are insignificant is too large to do further testing on this model. More specifically, throughout this paper we've seen how important the variable retentionrate is, and within the double log model, it is highly insignificant.

Given the large ranges of the collected dataset for specific variables, log transformed variables will change the case from a unit change to a percent change. Therefore, testing all of the possible models with logs and without logs were required to gain an accurate understanding of which gives the best statistical data. Comparing the 4 models shown above allowed for the outputs to be compared. This then permits me to pick the strongest model based on the tests run, which as I know, is the right-hand semi-log model.

If I were to have a major issue with a step-in calculating data, a different model would be used instead. Additionally, omitted variable tests were done on all 4 of the models. These tests resulted in similar outcomes to the chosen model, being that all four models had the same non-linearity problem, making my case stronger for choosing the right-hand semi-log model. These RESET tests were done in order to calculate omitted variables for all of the models to see which

alternative options I would have had if I had encountered an unfixable error in tests for the chosen model.

6. Discussion and Concluding remarks

Variable	Hypoth. Sign of Coefficient	Variable	Actual Sign of Coefficient
TuitionFees		TuitionFees	
FinAid	+	FinAid	+
FamilyIncome	+	FamilyIncome	+
GradRate	+	GradRate	+
Endowment	+	Endowment	+
StudentFacultyRatio	-	StudentFacultyRatio	-
AcceptanceRate	-	AcceptanceRate	-
RetentionRate	+	RetentionRate	+
Enrollment	-	Enrollment	-
StateID	?	StateID	+
FinType	?	FinType	+
AreaType	?	AreaType	-
SchoolType	?	SchoolType	+
GDPpc	+	GDPpc	+

Overall, the results I got did support my theory as well as the literature review. There are many institutional, educational, and economic factors that can cause variation in tuition costs for private colleges in the United States for the 2019/2020 fiscal school year. Many of the literary reviews used very similar variables to support their models. Therefore, I would say that due to the lack of literary review regarding increasing tuition costs specifically, my model is accurately represented by the tests and the underlying research done thus far. If I were to make any specific

changes after doing literature review, I would've added a variable that somehow represented the accessibility of the specific educational institution as well as potentially a variable that shows an average return on the education received in monetary value, since these are very popular topics amongst the educational field of study. All of the variables in my model are capable of causing price variation in tuition costs, while some are more accurately described by their values than others. The main unexpected result calculated in this model was for the omitted variable test. Zero of the four models pass the omitted variable test for non-linearity, thus making using the right-hand semi-log as the main model much easier. I was also surprised to see not a single dummy variable was significant in this model, much like the previous semester's tests on New York private institutions when underlying research stated that these were possible effects on the costs of tuition prices. The coefficients expected signs barely differ from the actual signs as seen above. As I determined, all of the hypothesized signs were accurately represented the way I believed they should have been. Last semester, in my previous model, it had been assumed that the more financial aid an institution has to give out, the higher the initial tuition costs would be, seeing that financial aid is meant to cover a large portion of the tuition costs. Within this model, financial aid is accurately represented by its coefficient within the regression. The dummy variables were given question marks respectively due to the fact that these variables could be either positive or negative but turned out to be both positive and negative. Finally, policy solutions similar to those in England can be put into place. These policies typically put a cap on the amount of tuition the educational institution can ask for from their students, lowering costs when compared to the United States. In conclusion, institutional, educational, and economic factors definitely do have a large effect on the cost of tuition for private colleges across the United States.

References

Brinkman, Paul T. "Factors affecting instructional costs at major research universities." *The Journal of Higher Education* 52.3 (1981): 265-279.

Ming, Joseph Sia Kee. "Institutional factors influencing students' college choice decision in Malaysia: A conceptual framework." *International Journal of Business and Social Science* 1.3 (2010).

Heller, Donald E. "Tuition, Financial Aid, and Access to Public Higher Education: A Review of the Literature." (1996).

Dinkelman, Taryn, and Claudia Martínez A. "Investing in schooling in Chile: The role of information about financial aid for higher education." *Review of Economics and Statistics* 96.2 (2014): 244-257.

Dinkelman, Taryn, and Claudia Martínez A. "Investing in schooling in Chile: The role of information about financial aid for higher education." *Review of Economics and Statistics* 96.2 (2014): 244-257.

Curs, Bradley R., and Luciana Dar. "Does state financial aid affect institutional aid? An analysis of the role of state policy on postsecondary institutional pricing strategies." *An Analysis of the Role of State Policy on Postsecondary Institutional Pricing Strategies (July 16, 2010)* (2010).

Dowd, Alicia C. "Income and financial aid effects on persistence and degree attainment in public colleges." *education policy analysis archives* 12.21 (2004): n21.

Ehrenberg, Ronald G. "Econometric studies of higher education." *Journal of econometrics* 121.1-2 (2004): 19-37.

Cohen, Bethaviva. "Factors Affecting Net Tuition Revenue at Private Colleges. Working Paper Series." (1982).