Elizabeth Morgan

Ryan Phelps

Econ3339

9 November 2020

College Tuition Factors; A Regression Analysis

The price of college is a largely scrutinized factor of American economics. The debate surrounding the factors that cause the largely varying prices in education has caused many to simply accept it as is. There are, of course, theories: on campus amenities, college size, religious affiliation, location, school popularity, the list goes on. However, there is no readily available definable proof for price. This paper will not fill that requirement, either, and it is unknown how long until there is one that successfully is.

For this project, I used information from two government databases, the first being the Integrated Postsecondary Education Data System, or IPEDS, and the second being the Equity in Athletics Data Analysis, or EADA. I knew from the beginning I wanted my dependent variable to be price per student, and I gathered 32 independent variables for 265 colleges across the United States for comparisons. I limited colleges to only include public or private not-for-profit colleges, and only colleges with a student population of 5,000 or more students, and limited data to the 2018-2019 school year or 2018 fiscal year, depending on applicability. Technically, this totals 479 colleges, but I had to cut back to 265 due to the lack of IPEDS data for certain categories, such as average ACT scores or average student aid. I am confident in my sample size, having over 250 samples pushes my model far past the necessary size for student’s t and a normal distribution assumption. Of course, a larger sample is always better, but I fully believe my sample, as well as the spread across the United States it represents, is beyond adequate for this model’s purposes.

An abundance of my gathered unusable variables were sports-centric, this such as participation count, total and separated by sex, and various assistant coach salaries broken down into different subsections. These largely limited my use of the EADA database, leaving my model to heavily rely on IPEDS instead. I also attempted to find correlation by comparing data pieces from the databases before running regression, such as finding the average salary paid to a head coach per student, but this did not improve my model either. I attempted upward of 35 models using various assortments of independent variables and comparison variables.

As shown in figure one, my final model contains 12 independent variables, and not shown is my one dependent variable, total price. The following paragraph will contain the IPEDS definition of each variable, possibly edited for clarity. The dependent variable, total price for in-state students living, is the cost of attendance for full-time, first-time degree seeking in-state undergraduates living on campus, including in-state tuition and fees, books and supplies, on campus room and board, and other campus expenses. The first independent variable, graduation rate total cohort is the graduation rate of first-time, full-time degree seeking students who completed their degree within 150% of the school catalog’s normal estimated time. Student-to-faculty ratio is the total FTE students not in graduate or professional programs divided by total instructional staff not teaching in graduate or professional programs. FTE is calculated by the full amount of full-time students plus 1/3 of part-time students. ACT difference is a personal calculation variable I made, which contains the difference between the 75th percentile ACT score and the 25th percentile ACT score for a college. The 25th and 75th percentile ACT scores are measures of the average test scores accepted into a college, specifically measuring the percentage of applicants who were accepted with a score equal to or less than the score, for 25% and 75% of accepted applicants respectively. Full-time fall 2017 cohort includes the full-time first-time degree seeking students enrolled at the institution in the fall of the prior year, which is used as a measurement of retention between academic years. Grand total (all students undergraduate total) is the grand total of men and women enrolled, as undergraduates, for credit at a university

The following paragraph will contain the EDEA definitions for variables taken from their database instead, possibly edited for clarity. Grnd\_total\_revenue and grnd\_total\_expense is the total revenue or expense for both men’s and women’s sports teams for a fiscal year at a university. Hdcoach\_salary\_women and hdcoach\_salary\_men is the average pay of coaches for all women’s or men’s sports at a university, respectively. Tot\_expense\_all\_notalloc is the total revenues and expenses for both men and women’s teams not allocated by gender or sport.

As seen in figure one, my adjusted R2 for my best model is 0.328. This is the highest adjusted R2 I was able to produce from any of my test regressions, which lead to this being my chosen best model. Adjust R2, in this case 0.328, interprets to mean that 32.8% of the variation of tuition cost can be explained by my dependent variables, and the model as a whole. Using a .1 confidence interval to determine statistical significance, my model has six significant variables, meaning, that the various dependent variables are useful in explaining the variation in the dependent variable, in this case, cost.

The statistically significant dependent variables are graduation rate total cohort, with a p-value of 0.00, student-to-faculty ratio, with a p-value of 0.010, ACT difference, with a p-value of 0.012, grnd\_total\_revenue, with a p-value of 0.031, hdcoach\_salary\_women, with a p-value of 0.072, and grnd\_total\_expense , with a p-value of 0.097. Since all of the p-values for these variables are less than the chosen significance level, .1, they are all determined to be statistically significant, and contribute to the deviation in cost.

Some of the associated coefficients seem expended beyond usability, such as the graduation rate total cohort stating that for every 118.15% increase in students graduating on time, the price goes up a single dollar. Other significant variables say the following changes lead to a $1 raise in price for a year: a decrease of 207 in student to faculty, an increase of ACT difference of 562 points, an increase in average women’s head coach salary of 1.2 cents. The student to faculty ratio change was unexpected and is, quite frankly, astronomical, with the absolute worst, or highest, student to faculty ratio in the entirety of the gathered data is only one student to thirty academic faculty members. As well, the ACT difference score is impossible, considering the highest possible ACT score is a 36.

My assembled model has four variables considered irrelevant, as well as two non-computed variables. The two non-computed are the ACT composite 25th percentile and 75th percentile, due to an error on my part of including them both in a regression along with the ACT difference variable, creating a perfect multicollinearity between them, and effectively cancelling out the composite scores for the difference variable. For irrelevant variables, it contains hdcoach\_salaray\_men, with a p-value of 0.239, Full-time fall 2017 cohort, with a p-value of .457, Grand total (All students Undergraduate total), with a p-value of .692, and tot\_expense\_all\_notalloc, with a p-value of .702.

I think that women’s coach salary might affect price while men’s does not might be that colleges tend to scale women’s sports with their size more, while men’s sports are more popularized, leading colleges to allocate more funding towards them despite the college’s actual student-based income. I believe this may also be the case with tot\_expense\_all\_notalloc, of the exuberant funding, especially in the south, of things like college football, disregarding the actual college’s income.

Finally, variables I believe may have improved the model if I had the time or means to acquire them. An estimated expense of living for the town the college is located in, a state’s student aid eligibility requirements and loan APRs, and one that would be, quite possibly, impossible to acquire but I think might be extremely helpful, would be the notoriety of a college within its state, and possible bordering states. How popular is the college, even on the other side of the state, what is the public opinion on the school, what are common associations made with the school, etc.

Figure one 