Predicting the Actual Cost of University Attendance with U.S. News and World Report's University Statistics

SAM SPEAKE, Marquette University

Seeing as it is nearly impossible to know what the real cost of attending a postsecondary institution is without investing time and money into the application process, I set out to attempt to create a linear regression to predict the cost using the *U.S. News and World Report*'s available college census data. From this endeavour I found that, of many predictors, the only statistically significant ones were the school's tuition and the percentage of the school's students who are receiving aid. This inability to create accurate predictions is likely due to three factors: the lack of individual data when trying to create individualized predictions, the countless confounding factors, and the inadequacy of the *U.S. News and World Report*'s college ranking system.

ACM Reference Format:
Sam Speake. 2019. Predicting the Actual Cost of University Attendance with U.S. News and World Report's University Statistics. *J. ACM* 37, 4, Article 111 (August 2019), 7 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

It is no secret that the cost of going to college has skyrocketed in the United States and continues to climb, yet despite the vast financial information and calculators available, it can be extremely difficult to pinpoint how much attending a specific university will actually cost. Seeing as most universities wont tell applicants what kind of aid they will receive until long after they have applied and, hopefully, been accepted, not to mention the steep application fees, I set out to see if there is a way to predict the true cost of attending a specific university before applying. For consistency's sake, I will define the cost of attending a specific university as the out of pocket expenses after scholarships and aid have been taken into account, including both tuition and housing.

Attempting to calculate a rough estimate of the cost of college before applying, without individual data, introduces an immediate roadblock, the removal of individuality. Hence, any two students at the same school are likely facing different financial challenges due to differences in high school scores and familial financial standing permitting for different levels of scholarship and aid to be awarded. Because of this, it becomes quite difficult to create an accurate model of cost based solely on statistics about the universities themselves. Nonetheless, I set out on my attempt to lighten some of the burden from the college application process.

2 LITERARY REVIEW

2.1 Cost of College Around the World

Around the world, attending a quality university is typically exceedingly affordable. *Insider* looked at the average cost of college in twenty-eight first world countries around the globe, and found that nearly a third had no tuition costs, while nearly all the remaining colleges had tuition costs equalling less than half of that in the United States [7]. With this in mind, it would seem logical to

Author's address: Sam Speake, sam.speake@marquette.edu, Marquette University.

0004-5411/2019/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

think that strong efforts are being made to lower the cost of school in the United States in order to remain competitive with these countries in the future. In reality, as *CNBC* points out, quite the contrary is happening. From the time the article was published, the total student loan debt had increased by 150 percent in the previous decade, totaling over $1.4 trillion [2].

2.2 United States College Affordability

While the aforementioned figure of 1.4 trillion dollars in American student loans is a staggering figure in and of itself, it does not do much for realizing the actual impact the rising cost of college has on an individual. There tends to be a common view of paying for college where a student works full time over the summer in between semesters, and is able to successfully pay their way through school graduating with little to no debt. With the average cost of college in 2017, the *US Congress Joint Economic Committee* found that a student at a public 4-year institution would have to make $38.63 an hour working full time over the summer to afford their school, while a student a private 4-year institution would have to make $87.25 an hour [6]. The committee elaborated that this often prohibits students from being as economically active as a successful economy requires, as even students who graduate with a valuable degree often find themselves unable to make big life steps, such as buying a home, in favour of paying towards looming student loans, all the while students who do not graduate may find themselves unable to achieve financial freedom for the majority of their lives [6].

While attempts have been made to bridge the socioeconomic gap this paywall creates through need-based aid, the reality is that college remains exceedingly unaffordable for students with limited resources [1]. Additionally, as Baum points out, the typical metrics used to measure affordability are wildly inadequate, as a degree of little value made quite cheap is still unaffordable due to a likely nonexistent return on interest, whereas a valuable degree made far less expensive, yet still more expensive than the prior exposition, becomes affordable due to the large return on interest [1]. This is supported by the finding that individuals with an accredited postsecondary degree will make, on average, over a million dollars more than someone with identical circumstances (bar the degree) throughout their lifetimes [6].

2.3 The Role of the U.S. News and World Report

Every year, *U.S. News* takes a census of over three hundred United States universities, collecting data points about the institutions such as class size, tuition, faculty to student ratio, etc. Then, they compile and compare this data to create and release many sets and subsets of university rankings. While it can be nice to have this list for many reasons, whether you are a hopeful student, a school counselor, or a university administrator, recent analysis has suggested that these ranking actually cause far more harm than good.

As *U.S. News* claims to base these rankings purely on academic quality, it would be expected that slight differences in ranking display insignificant differences in academic quality [3]. Contrary to that logical intuition, The Journal of Economic Behavior and Organization found that arbitrary thresholds tend to have a significant impact on the amount of applications the university receives [4]. Of these thresholds, the most significant is the rank 50 cut off, where, historically, there is a "2-6% discontinuous drop in applications" [4]. Because of this, the *U.S. News and World Report* college rankings tend to play a larger part than you would think in the typical happenings of many universities.

With the manufactured importance of these rankings being so high, many institutions diligently try to game the system. This may be done through increasing spending on factors which likely hold a lot of weight, thus increasing tuition, driving up applications rates in order to appear more

selective, making score submissions for tests such as the ACT and SAT optional, or a plethora of other tactics [5].

3 DATA CLEANING

To attempt to tackle this question I used the University Statistics data set from *Kaggle*, it is a relatively small data set at 347 KB, containing 19 data points about each of 311 universities, taken from the *U.S. News and World Report*'s college census. The data set is available at https://www.kaggle.com/theriley106/university-statistics.

3.1 Initial Cleaning

The data set is stored in a JSON file, and when initially loaded into RStudio required quite a bit of cleaning. To start, I made the data frame a matrix in order to transpose it due to the rows being attributes and the columns being instances. Once that was fixed, I removed any attributes which were not nominal, ordinal, count, or continuous data, apart from a few used for university identification. Keeping school names was useful so that I could employ my empirical knowledge as reference in order to make sure that my regressions and visualizations made sense. Additionally, I removed any null columns and replaced null data with N/A values. Finally, due to the transpose the row names were gibberish, so I replaced them with indices since the instances were identifiable through the university names.

4 ANALYSIS AND RESULTS

4.1 Initial Visualizations

To start out I wanted to see if there was any relationship between the tuition of a school and the cost after aid. When I mocked up the initial visualization, the data seemed completely unrelated, but *Fig. 1* shows that, when colored by institutional control, public schools tend to increase in cost after aid as tuition increases. What a riveting discovery.

Following this visualization I was interested to see whether or not location had an impact on cost. As the data set included the zipcode of each university, I sought to merge the data with the *IRS*'s fiscal census data. Quickly, I realized that the census data was somehow quite incomplete, missing data pertaining to the zipcodes of the majority of the schools surveyed by the *U.S. News and World Report*. In lieu of this, I merged the data with state longitude and latitude data in order to create *Fig. 2*, hoping still to find a pattern related to location.

Having seen that the states which typically had a higher average tuition loosely represented a map of average cost of livings, I turned back towards my initial visualization, *Fig. 1*, to dive deeper.

4.2 Regression

Having seen that tuition and cost after aid seemed to be somewhat correlated, I decided to create a linear regression using all the continuous attributes and the categories private and public in order to see what variables actually had any significance on cost after aid. The summary statistics of the coefficients are visible in *Table 1*. Of all the possible regressors, only tuition and percent receiving aid ended up being statistically significant for cost after aid at the assumed alpha value of 0.05, I have marked these in the table with ***. Finally, I created the final linear regression, visible in equation (1), using the two significant variables.

$$CostAfterAid = 0.27 * Tuition - 0.022 * PercentReceivingAid + 32410 \quad (1)$$

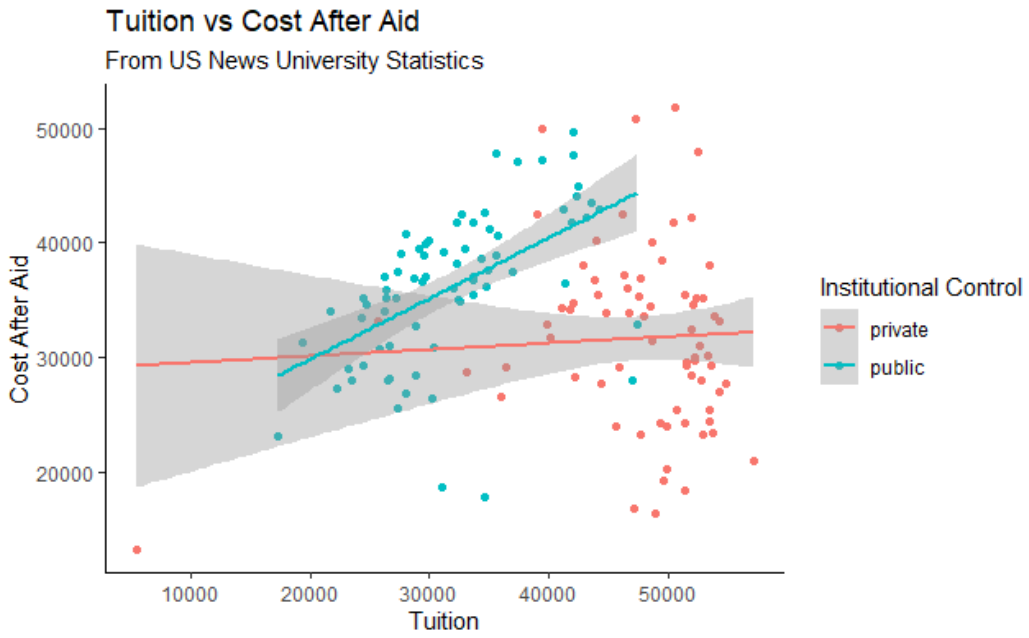

Fig. 1. Tuition vs Cost After Aid by Institutional Control.

Table 1. Initial Regressors for Cost After Aid

Variable	Estimate	P-Value
Average ACT	3.08e+04	0.188
Average SAT	8.98e+01	0.907
Enrollment	-1.51e+01	0.604
Tuition	5.96e-01	2.08e-08***
Acceptance Rate	3.25e+01	0.54
Percent Receiving Aid	-2.54e+02	3.16e-06***
Rank	6.07e+01	0.111
Average High School GPA	-7.26e+02	0.869
Public Control	5.23e+02	0.824

4.3 Final Visualizations

Now knowing which attributes were significant for cost after aid, it became easy to visualize the inherent relationships. *Fig. 3* is quite similar to *Fig. 1.* in that it displays the cost after aid by tuition, colored by institutional control. To paint a more accurate picture of the data, though, I employed small multiples, dividing the data set using three thresholds of the percent of students who receive aid.

Fig 3. seems to show that more expensive private schools tend to give better aid, especially when more than fifty percent of students are receiving aid, and as public schools tend to give more aid, they tend to give better aid up until the few public schools above fifty percent are considered. With this in mind I created my final visualization which shows that there is a relatively strong correlation

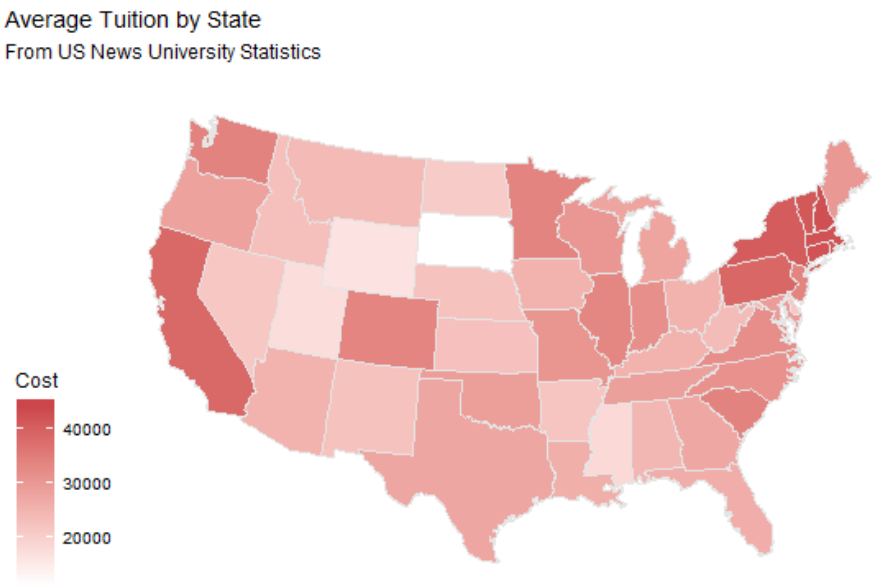

Fig. 2. Average Tuition by State.

between the percent of students that receive aid at a school, and the average percent of tuition that is covered by aid at that school, visible in *Fig. 4*.

5 CONCLUSIONS

As visible in the previous section, tuition and percent receiving aid are both strongly significant predictors for cost after aid. For tuition, this is a relatively common sense finding, seeing that as tuition increases, it makes sense that cost would increase. For percent receiving aid, there are two likely explanations for this correlation. The first is that if a university is giving aid to a higher percent of its students, it is quite likely that the university is also giving higher quality aid, thus quality of aid and percent receiving aid would likely be correlated. The second possible explanation is that this is more so the effect of a conflation of the variables. To show this I will use an example, if a university gave all of its students 20% off their tuition, the cost after aid would drop by 20%, whereas if a school gave 20% of their students 50% off their tuition, the cost after aid would only drop by 10%. Hence, it is difficult to be sure of what exactly cost after aid is representing.

Despite the strength of the aforementioned correlations, from the data present I do not believe that an accurate predictor of an individual's cost to attend a university can be created without employing individual user data, just taking statistics about the university into account. I do not believe this to be a complete impossibility, though, more so a result of a plethora of confounding factors. Factors such as the vast lack of data from schools outside the top 150 ranks, seeing as unranked schools are probably less likely to submit data as they have far less to gain from the rankings, or such as the many data points within a university that would skew their averages away from what the average student could expect, such as the vast amount of huge sports scholarships given by division one schools.

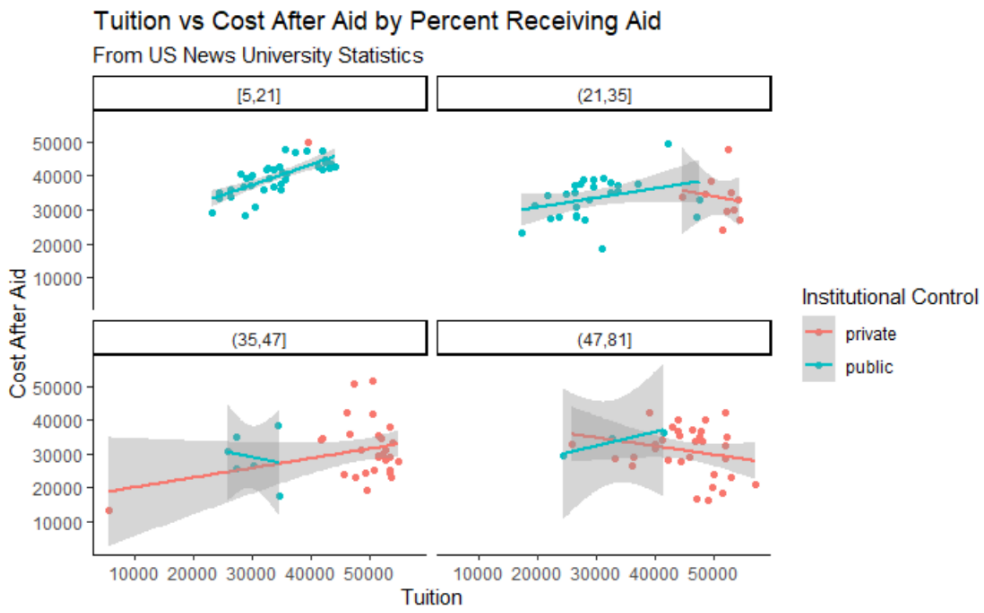

Fig. 3. Tuition vs Cost After Aid by Institutional Control, Separated by Percent Receiving Aid.

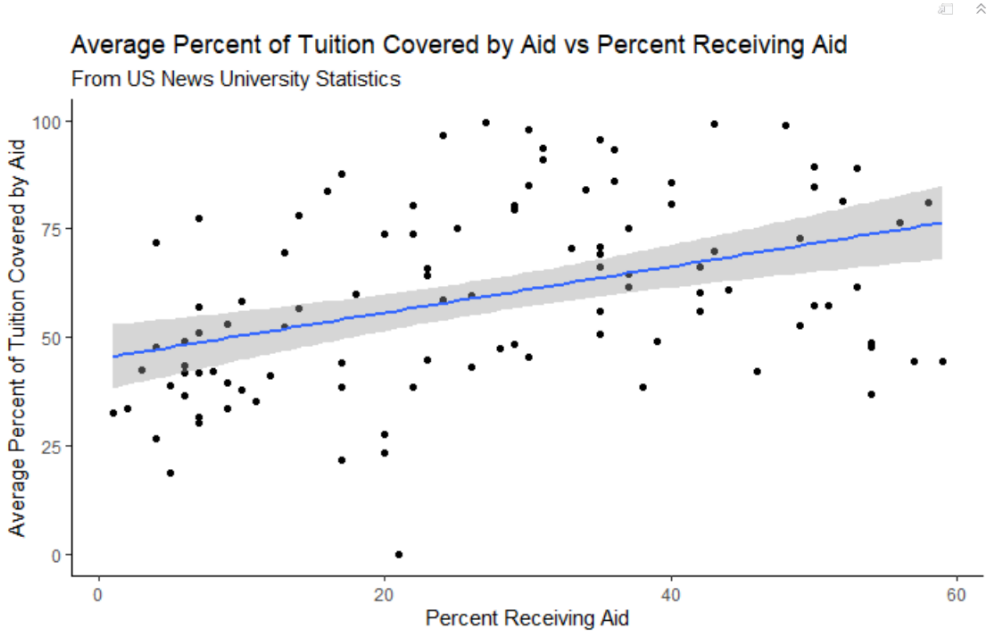

Fig. 4. Average Percent of Tuition Covered by Aid vs Percent Receiving Aid.

Furthermore, I think that the final issue lies in the data itself due to the inadequacies of the *U.S. News* college census outlined in Section 2.3. Empirical knowledge can be used to expand on the

examples laid out in that section, in order to draw possible correlations. For example, universities such as Marquette have decided to go ACT/SAT optional under the pretense that it helps students who have had less opportunities get admitted. While this is probably a true factor, the big implication of the policy is that the school's average ACT and SAT scores will rise, thus likely increasing the schools rank, which can be used as an excuse to raise tuition seeing as the academic quality has increased in the public eye, despite remaining likely stagnant. The same is true of a school doing promotional work in order to drive up the amount of applications it receives. While nothing about the school has actually changed, the acceptance rate will appear more selective and, once again, the school will likely rank better. Due to the weight that the *U.S. News and World Report*'s college rankings carry, outlined in section 2.3, it would appear that these academic slights of hand are happening throughout countless facets of American postsecondary institutions.

Hence, due to the lack of individualized data, the countless confounding factors, and the inadequacy of the *U.S. News and World Report*'s college ranking system, it would seem that an objective college cost prediction model is unlikely to be able to be created without major changes within the industry of postsecondary academia.

REFERENCES

[1] Sandy Baum. 2017. *How Should We Think About College Affordability?* Retrieved December 11, 2019 from https://www.higheredtoday.org/2017/05/16/think-college-affordability/

[2] Jessica Dickler. 2017. *Student loan balances jump nearly 150 percent in a decade.* Retrieved December 11, 2019 from https://www.cnbc.com/2017/08/29/student-loan-balances-jump-nearly-150-percent-in-a-decade.html

[3] Scott Jaschik. 2017. *Where 'U.S. News' Rankings Have Influence on Potential Applicants.* Retrieved December 11, 2019 from https://www.insidehighered.com/admissions/article/2017/08/28/study-documents-where-us-news-rankings-have-influence

[4] Andrew G. Meyer, Andrew R. Hanson, and Daniel C. Hickman. 2017. Perceptions of institutional quality: Evidence of limited attention to higher education rankings. *Journal of Economic Behavior & Organization* 142 (Oct. 2017), 241–258. https://doi.org/10.1016

[5] John Tierney. 2013. *Your Annual Reminder to Ignore the U.S. News and World Report College Rankings.* Retrieved December 11, 2019 from https://www.theatlantic.com/education/archive/2013/09/your-annual-reminder-to-ignore-the-em-us-news-world-report-em-college-rankings/279103/

[6] Ranking Member U.S. Senator Martin Heinrich. 2017. *The College Affordability Crisis in America.* Retrieved December 11, 2019 from https://www.jec.senate.gov/public/_cache/files/5270bffa-c68e-44f0-ac08-693485083747/the-college-affordability-crisis-in-america.pdf

[7] Katie Warren. 2018. *Here's what college costs in 28 countries around the world.* Retrieved December 11, 2019 from https://www.insider.com/cost-of-college-countries-around-the-world-2018-6