Paying for Prestige

Rory G. Quinlan[1*]

[1*]Economics Department, University of Pittsburgh, Pittsburgh, PA, USA.

Corresponding author(s). E-mail(s): RoryQuinlan@pitt.edu;

Abstract

This paper isolates and uses key factors to predict tuition for public and private four-year universities across the United States within the same year. Institutional, quality of life, and crime rate metrics will be observed and used to explain the differences between current university tuition rates. The data used for this model consists of 60 observations, each representing a four-year university in the US. These 60 observations are a sample of the top 400 US universities. This study employs seven predictors in a linear model to predict tuition variability across universities and utilizes the leave-one-out cross-validation technique. The model was able to account for over 85% of the variability of the tuition rates, and found that the only significant institutional metric contributing to tuition cost was the university's national ranking.

Keywords: Tuition Prediction, Higher Education, University Ranking, Cross-sectional Data, Generalized Linear Models (GLM)

1 Introduction

Today, student loan debt in the United States totals over 1.7 trillion USD (Hanson, 2023). Tuition hikes are rampant in today's higher education system; since 1980, the cost of attending a four-year college has increased by over 180% (McGurran, 2023). These tuition hikes only further the student debt burden. Clearly, college prices in the US have been rising steadily on average, and student loan debt has followed, but is there a discernible pattern in the tuition system among US colleges *today*?

2 Statement of Purpose

This paper isolates and uses key factors to predict tuition for public and private four-year universities across the United States within the same year. Institutional, quality of life, and crime rate metrics will be observed and used to explain the differences between current university tuition rates. While much research has been conducted to explain the time series data of tuition increases, research to explain the variation between tuition prices while holding time constant is scarce.

Understanding the variations of tuition can add insight to much of the time series research, as these studies are usually only aggregates of tuition averages. As well as establish a market value for a college, given key predictors. With this model, you can determine if a college is over- or under-priced compared to the market value, allowing you to maximize your return on investment and minimize your personal student loan debt

3 Literature Review

Deleeuw (2012) researched the link between unemployment and enrollment in higher education. Her study consisted of a single community college, Monroe County Community College, as two-year institutions are typically more sensitive to tuition increases than their four-year counterparts (Hearn, 1998; Leslie Brinkman, 1987). She compared total credit hour enrollment at MCCC and the unemployment rate from 1980 to 2012. The study found that the unemployment rate is a statistically significant predictor of total credit hours. As the unemployment rate rises, enrollment also rises. Deleeuw posits that individuals seek education when job opportunities are scarce

S. Wahyuddin et al. (2019) analyzed factors affecting tuition in a private university. This study investigates a single Indonesian private university, STMIK Dipanegara Makassar, from 2010 to 2018. S. Wahyuddin et al. found that inflation significantly affected tuition and the number of enrolled students. However, it also found that neither the regional minimum wage nor the number of students enrolled affected tuition.

Lucca et al. (2018) examined a causal association between student credit expansion and the rise in college tuition. They examined the implementation of policies that increased the federal student loan (subsidized and unsubsidized) cap in over 5,000 unique institutions from 2002 to 2012. Their results suggest that while federal student loan cap policies contribute to tuition sensitivity, they do not directly contribute to

the rising cost of tuition. Instead, the results indicate that student loans function as intended- a tool to help students afford the rising tuition.

All these articles attempt to explain institutional changes over time using time series data. Deleeuw (2012) and S. Wahyuddin et al. (2019) only observe a single institution, while Lucca et al. (2018) observe a much larger sample. Deleeuw uses enrollment as the response variable, whereas S. Wahyuddin et al. use it as an explanatory variable in their model. While S. Wahyuddin et al. did not find any significant effect of enrollment on tuition, Deleeuw found that the unemployment rate (a quality of life metric) does affect enrollment (an institutional metric). It is important to note that Deleeuw defines enrollment as the total number of credit hours, whereas S. Wahyuddin et al. define it as the number of unique students enrolled. Definitional differences may account for nuance in the explanatory power of the number of enrolled students to the tuition rate.

Lucca et al. (2018) support the finding of S. Wahyuddin et al. (2019); S. Wahyuddin et al. found that inflation was their best explanatory factor in predicting tuition. Lucca et al. establish that federal student loan cap expansions- a possible confounding variable in S. Wahyuddin et al. 's model- do not have a causal link to tuition increases.

4 Methods

4.1 Data Collection

The data used for this model consists of 60 observations, each representing a four-year university in the US. These 60 observations are a sample of the top 400 US universities. Selecting from the top 400 gives a large enough population to see the variation in institutional metrics while remaining relevant and applicable to the largest number of students by excluding fringe universities. Each university has university-specific (institutional) metrics, such as student-to-faculty ratio, number of undergrads, institutional rank, etc., and university-adjacent (quality of life and crime) metrics; these metrics are from the city that the university is located in, such as crime rate, cost of living, or median income The query yielded 5,829 job postings, each with a unique job ID to ensure non-duplication. Each job record includes structured fields, such as agency, job series, and location, as well as unstructured text, such as duties, qualifications, and job descriptions.

The data used in this paper is an aggregate of three public datasets available on Kaggle: *US College data* published in 2020 by Yash Gupta, scrapped from NCES (National Center for Education and Statistics, *FIPS US quality of life data* published by Zach Vaughan (2016), and scrapped from the FCC FIPS reports, and *United States Crime Rates by City* published by Kabhish Mahesh in 2022, scrapped from City-Data.com. These data sets have reputable sources, were already cleaned to exclude missing observations, and are prepared to be easily exportable.

4.2 Variables

This study uses seven predictors in a linear model to predict variability in tuition across universities defined as follows:

- Institutional control is a categorical variable that states whether a university is public or private. Public is coded as 0, and private is coded as 1
- *Rank*, a number assigned to each college according to the US News and World Report; the lower the number, the "higher rank" or more prestigious the university
- Number of undergraduates, the number of undergraduates currently attending that university (2020)
- *Unemployment*, the percentage of the city the univeristy is located in unemployment
- *Median income*, median income reported by the Economic Policy Institute (2022)
- *Median income*, median income reported by the Economic Policy Institute (2022)
- *Diversity Rank (by race)*, the ranking number of the city by diversity. The more racially diverse the lower (numerically) the ranking number
- *Expend*, the amount of money the school uses per student

4.3 Statistical Modeling and Analysis Plan

To assess the variation of tuition among universities today a combination of Institutional, quality of life, and crime rate metrics were aggregated for 60 universities. The response variable (Y) represents the annual tuition to attend university. Using stepwise selection, the following maximum linear model was generated:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \varepsilon$$

Where the variables are defined as follows:

- x_1: Institutional control (0 = public, 1 = private)
- x_2: National institutional rank
- x_3: Median income of residents in the surrounding city
- x_4: Diversity rank (by race) of the surrounding city
- x_5: Unemployment rate of the surrounding city
- x_6: Number of undergraduate students at the university
- x_7: Expenditure per student for the university

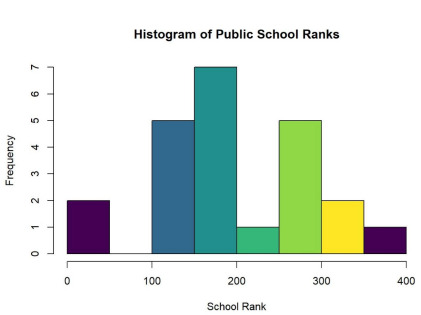
To assess the validity of this maximum model, we used leave-one-out cross-validation (LOOCV). While computationally intensive, this method gives the most accurate result for our smaller sample size.

5 Results

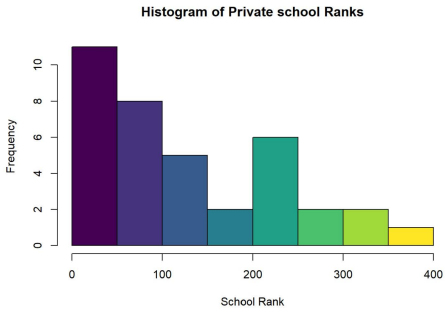
5.1 Descriptive Statistics

The 60 observations use whether the university is public or private as a control variable. 37 schools are private universities and 23 are public. The rank distribution of

private universities is right-skewed, with a tendency for a higher (better) rank. Public university rankings display a more normal distribution. Figures 1(a) and 1(b) are histograms that display private and public university ranking distributions described above.



(a) Descriptive statistics for public universities



(b) Descriptive statistics for private universities

Fig. 1: Histogram of Rank by University type

Table 1: Institutional descriptive statistics for public universities
$n = 23$

Variable	**Mean**	**SD**
Tuition (annual)	26162.91	8922.40
Expenditure	8922.40	2150.88
Number of Undergrads	15493.83	6470.21

Table 2: Institutional descriptive statistics for private universities
$n = 37$

Variable	**Mean**	**Median**	**SD**
Tuition (annual)	47879.76	51360	12793.9
Expenditure	15012.92	10813	10726.2
Number of Undergrads	5227.054	81899.65	3264.94

Table 1 and 2 displays the three institutional predictors of tuition in the model, because whether the university is public or private is our control variable, the left table

contains the three predictors from the 37 private universities and the right side contains four key predictors from public universities. Since the private universities have a right skew, the median value was included in the table as a more robust measure of center.

5.2 Inferential Statistics

Table 3 displays the model summary of each predictor discussed above. Institutional rank, Institutional control, and Median income were all significant at above the .001 significance level. Diversity rank was significant at the .05, and unemployment was significant at the .1 level. Of all predictors, rank is the most significant, by at least two orders of magnitude above all other predictors. Overall the model has a large explanatory power of tuition, with an adjusted R squared value of 0.87. This explanatory power is further validated by the cross validation (LOOCV). After validation the model still performed with a 0.85 R squared value, indicating there aren't any severe overfitting problems, and could be a useful model in the future with more data. Table 4 displays the results of the cross validation.

usepackagetabularx booktabs placeins

Table 3: Regression Results

	Dependent variable: Tuition
Rank	-87.514^{***}
	(10.174)
Institutional Control (public=1)	$-11,292.310^{***}$
	(2,497.191)
Median Income	0.164^{***}
	(0.037)
Diversity Rank	-1.909^{*}
	(1.017)
Unemployment	$173,051.400^{**}$
	(84,461.880)
number Undergrads	-0.274
	(0.166)
Expend	0.099
	(0.107)
Constant	$43,034.530^{***}$
	(6,028.025)
Observations	60
R^2	0.891
Adjusted R^2	0.876
Residual Std. Error	5,493.054 (df = 52)
F Statistic	60.432^{***} (df = 7, 52)

Note: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 4: Cross Validation Results

Intercept	RMSE	Rsquared	MAE
TRUE	5988.835	0.8513063	4797.676

5.3 Model Assumption Validations

The model was evaluated using standard diagnostics including residual and Q-Q plots shown in figure 2 to assess the assumptions of linear regression. Overall, the assumptions are met reasonably well, though some mild concerns are present.

The spread of residuals remains fairly consistent across the fitted values, though there is some mild heteroskedasticity; it appears at higher tuition levels there is a tighter clustering of residuals. In addition, the Q-Q plot shows that most standardized residuals follow the theoretical normal distribution, with the exception of a few outliers (notably universities 18, 19, and 44) that exert disproportionate influence on the model.

These concerns are not severe enough to invalidate the findings, and the knowledge gained from the model remains more valuable than disregarding the results entirely. However, they do suggest that caution should be taken when interpreting the estimates. To address these issues, robust standard errors should be used, providing more reliable inference even in the presence of mild assumption violations.

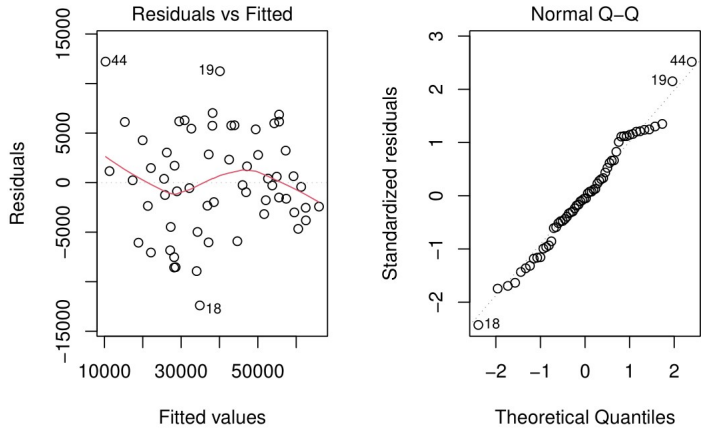


Fig. 2: Residual plot (left) and normal QQ plot (right) of selected model

6 Conclusion

6.1 Interpretation of Results

These results show that we can explain most of the tuition variability while holding time constant. Tuition increases as the school rank gets better (lower in number). This relationship is expected because more desirable (prestigious) schools can leverage this and charge their students a premium. What's impressive is the significance of this relationship. In a simple linear regression, Rank accounts for 65Variables in this model that were insignificant tell us just as much about the higher education market as the ones that were. Besides Rank, no other institution metrics were statistically significant at the 0.1 confidence level (excluding the private vs. public control). Student-to-faculty ratios, graduation rate, or percentage of alumni donating to the university were irrelevant to predicting tuition despite these metrics being widely touted and used in recruitment brochures and websites. Today's college students pay for Prestige rather than a school that invests in them.

6.2 Limitations

While the model performed well after cross-validation, it is a limited sample size to capture the wide variance in college tuition prices today, especially considering the differences between public and private institutions. Additionally, aggregating the data leads to some variability in the collected data over the years. The range of years is small (less than five years), so it is unlikely that the market predictors shifted during this timeframe. However, this range does increase the possibility of error in predicting tuition at a static point in the market. Within the model, there is also slight heteroskedasticity in the regression (displayed in Figure 4), possibly because typically only private institutions' tuition reaches the upper range of tuition. Lastly, the tuition prices used are only "sticker prices", and does not include the prices that students actually pay (after scholarship granted by the school, grant, and other miscellaneous financial aid). Universities have a high ability of implementing first degree price dicrimination. While this leads additional surplus to the universities, it reduces price transparency making academic research reguarding tuition rates by universities nearly impossible to calculate; the decreased price transparency could also decrease decrease perspective students price sensetivity, as there is too much individual variation to make valid comparisons to between students.

6.3 Future Work

Future research can focus on the differences between public and private institution tuition rates while controlling for variables like the significant ones in this model to better understand the true implications of these two types of institutions. Alternatively, the model could be validated with past or future data to see if the market variances for each time frame were motivated by the same factors as today or if the market shifts to other significant predictors. Additional literature is also needed for more accurate tuition rate calculations to increase price transparency and therefore sensetivity in students.

6.4 Data Availability

The analysis scripts and study findings are also publicly available at:
https://github.com/RoryQo/Tuition-Variation-Research

References

[1] DeLeeuw, Jamie. "Unemployment rate and tuition as enrollment predictors." (2012): 1–13.

[2] Gupta, Yash. (2020, Month of dataset creation). *US College Data*. Retrieved 2/21/2024 from https://www.kaggle.com/datasets/yashgpt/us-college-data/dataData

[3] Hanson, M. (2023, August 20). Student Loan Debt Statistics. Education Data Initiative. https://educationdata.org/student-loan-debt-statistics

[4] Hearn, J.C. (1988). Attendance at higher-cost colleges: Ascribed, socioeconomic, and academic influences on student enrollment patterns. *Economics of Education Review*, 7(1), 65–76.

[5] Leslie, L. L., & Brinkman, P. T. (1987). Student price response in higher education. *Journal of Higher Education*, 58, 181–204.

[6] Lucca, D. O., Nadauld, T., & Shen, K. (2018). Credit Supply and the Rise in College Tuition: Evidence from the Expansion in Federal Student Aid Programs. *The Review of Financial Studies*, 32(2). https://doi.org/10.1093/rfs/hhy069

[7] McGurran, B. (2023, May 9). College Tuition Inflation: Compare The Cost Of College Over Time (A. Hahn, Ed.). Forbes Advisor; Forbes. https://www.forbes.com/advisor/student-loans/college-tuition-inflation/

[8] Vaughan, Zach. (2016, January). City/Zip/County/FIPS- Quality of Life (US). Retrieved 2/21/2024 from https://www.kaggle.com/datasets/zachvaughan/city-zip-county-fips-quality-of-life-us

[9] Wahyuddin S., Fauzi Insan Estiko, & Estiko Rijanto. (2019). Analysis of Factors Affecting Tuition Fee in a Private University: A Data Mining Using VAR Model. *IOP Conference Series: Materials Science and Engineering*, 662(2), 022050. https://doi.org/10.1088/1757-899x/662/2/022050

[10] Wickham, H., & Bryan, J. (2023). *readxl: Read Excel Files*. R package version 1.4.3. https://CRAN.R-project.org/package=readxl

[11] Wickham, H., François, R., Henry, L., & Müller, K. (2022). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.10. https://CRAN.R-project.org/package=dplyr