Henry Cooper
Economic Data Analysis
Written Report

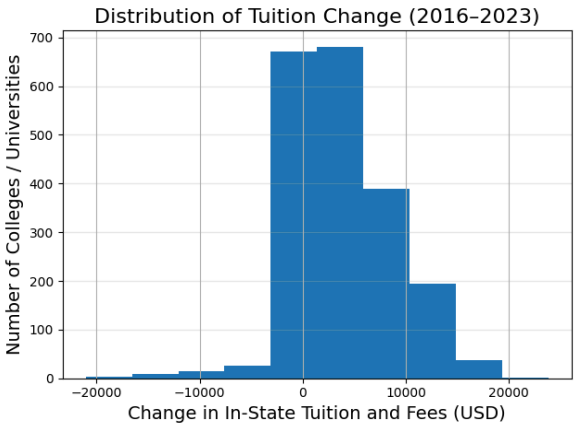
Enrollment & Tuition (With Institution Control)

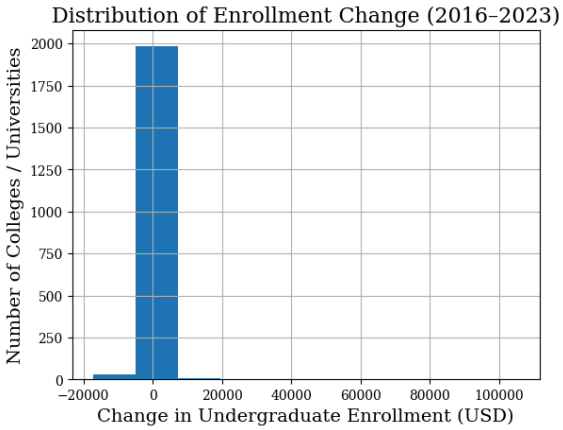
1. Introduction & Question

When looking for an economic research question to ask, something that immediately sparked was questions about education and the role economics play in our education system here in the US. Considering we are doing this project in college, at a public university, all pursuing undergraduate degrees, naturally this became our target demographic. One of the most major economic factors of pursuing an undergraduate degree, and deciding where and how to get an upper level education (public or private), comes down to tuition costs.Our official research question is as stated: How do changes in in-state tuition effect changes in undergraduate enrollment, and does this relationship differ between public and private institutions?

2. Data description & Sources

The data collected for this research project came primarily from the National Center for Education Statistics, and specifically within this IPEDS (Integrated Postsecondary Education Data System) data was taken to look into undergraduate college data across the US. I was able to make a custom data set through the system to compare two school years specifically, 2016 and 2023. By looking at changes between two years we can see the specific relationship between tuition change and the effect this has on enrollment numbers. The custom data set included 3 main variables outside of year, these being control of institution (either public or private non-profit specifically), published in-state tuition and fees for the academic year, and undergraduate enrollment. I was able to make two custom data sets using these variables and then merge the data in python to create two new variables: tuition difference and enrollment difference. We were able to look at about 2000 institutions in total after data cleaning and removing any school with an NaN value for either tuition or enrollment data. Our histograms of these two critical variables are below:





3. Methodology

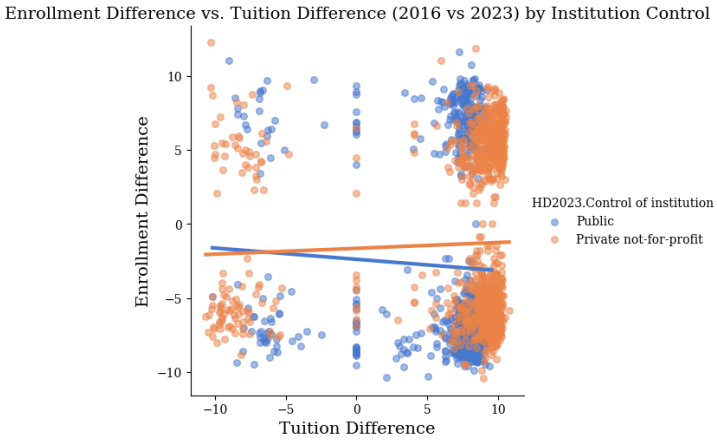
We ran a multivariate linear regression model that included change in tuition as one predictor variable, our categorical variable in institution type, and our interaction term.

$\Delta\ Enroll = \beta_0 + \beta_1(\Delta Tuition) + \beta_2(Public) + \beta_3(\Delta Tuition \times Public) + \epsilon$

This model shows that change in enrollment is our outcome variables and we have three predictor variables with β_1 correlating with our change in tuition slope and our β_2 categorical variable coded as public (public = 1 was coded as public, public = 0 is coded as private) making private schools the reference category, correlating with the change in tuition slope being the β_1 representing private school tuition changes. Our public school slope is seen through the use of the interaction term β_3 which is added to the base slope to get our public slope - this interaction being very important in seeing institutional differences through their slopes. Of course there are still limitations as we cannot account for every variable related to enrollment, and our enrollment is general undergraduate enrollment (I could not find in-state) while tuition is only in-state, meaning our data is not a perfect insight into this relationship.

4. Results & Analysis

In order to make a scatterplot that was not too cluttered and could help with linearity, we had to transform the data to make it more digestible visually and help see trends. Because we had meaningful negative values (when tuition or enrollment went down), we could not run a basic log transformation but rather ran an arcsinh transformation to deal with all our values. Our scatterplot/regression results came out like this:



	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.6488	0.292	-5.654	0.000	-2.220	-1.077
C(Q('HD2023.Control of institution'))[T.Public]	-0.7399	0.540	-1.370	0.171	-1.798	0.319
arcsinh_tuition_difference	0.0396	0.032	1.234	0.217	-0.023	0.102
arcsinh_tuition_difference:C(Q('HD2023.Control of institution'))[T.Public]	-0.1157	0.069	-1.684	0.092	-0.250	0.019

Due to limitations in our data that caused things like heteroskedasticity, which is clear considering our scatterplot certainly has apparent visual limitations, when I ran the OLS regression I used robust standard errors to account for this issue, making our results more accurate. The result coded as C(Q('HD2023.Control'))[T.Public] is our control which at a p-value of 0.171 is not statistically significant, arcsinh_tuition_difference is at a p-value of 0.217 (which is the private school/base tuition difference slope) is also not statistically significant. However the final row shows our interaction term at a p-value of 0.092 which is much closer and below a threshold of 0.1 which is a threshold of less strength but sometimes considered relatively significant.

5. Conclusions

Overall, it is clear to see that based on our p-values all lacking true statistical significance, we cannot reject our null hypothesis that there is no relationship between change in tuition and enrollment, and no related difference between public and private schools. Our tuition difference p-value was well above 0.05, and our tuition-related slope for public schools which uses the base slope (private schools which is a p-value of 0.217) and the interaction term (a better p-value of 0.092), we cannot say that the relationship between tuition and enrollment exists confidently for public schools either. In terms of differences in slope between schools (our interaction), a p-value of 0.092 does suggest marginal evidence that there may be legitimate differences between institutional types in terms of slope, but not enough to be confident in a statistically significant relationship.

6. References

IPEDS data: “Your Primary Source for Information on U.S. Colleges, Universities, and Technical and Vocational Institutions.” *IPEDS*, nces.ed.gov/ipeds/. Accessed 12 Dec. 2025.