Is there Equity in our Higher Education System?

Section 6
Joseph Margolis

Introduction

Introduction to Topic

My topic is looking into equity distribution in colleges and universities across the United States. One of the most important qualities of a college or university is how they set a person up for life after they graduate. Does the student get set up for a job upon graduation and if so, how much does that job make the student? Is it enough to make up for the school costs? That idea is the main factor that I am looking at in the schools from the given data set to be able to judge both the location of the schools that set their students up best for life after college, and whether minority serving institutions are successful in bringing equity in this regard to a system that is overwhelmingly by the white majority. This looks into two possible lenses for equity in our colleges and universities, whether or not minority groups are offered the same level as schools as the heavily white schools and if the top schools for post graduation success are evenly distributed throughout the countries.

Research Questions

The first question that I will explore is about which regions of the country provide colleges and universities that leave students with the highest median wage six years after enrollment, while accounting for the predominant degree awarded from that school. The other question looks into how the percentage of minority serving institutions changes as we look at schools with higher median wages for their students six years after enrollment, also while accounting for the predominant degree awarded from that school.

Introduction to Data

The data in this report comes from the U.S. Department of Education and the Institute for Education and Professional Development (IEPD) and it is mainly for prospective college students to help guide decision making on what school is right for them. The data covers 93 different variables, including majors provided and how many take them, location of schools, salaries of alumni, etc., for 3,676 different colleges and universities in the United States and United States territories, making up a sizable portion of the total number of colleges and universities in the U.S. which is around 5,300 schools. The data was collected through an observational study, by obtaining numbers from the schools themselves, tax records from the schools, and financial aid forms completed by the schools. The target population of the data are prospective college students who are trying to decide where they want to attend college, with the cases in this dataset representing the schools with 90 different variables that can be used to separate all the schools from each other, displaying where the schools are, what kind of classes they offer, and how they set up students for after they graduate.

Multiple Linear Regression Modeling

Variable Descriptions

First, my outcome variable for my linear model is the median wage of students from a school six years after enrollment. This variable is used to judge one aspect of the quality of schools, that being how well they prepare students for life after graduation. This is a quantitative outcome variable that is measured in U.S. dollars.

My first explanatory variable I am using is categorical and represents the region of the country each school is from, breaking up the country into the Northeast, North-Central or Midwest, South, and West regions. This is in order to be able to explore the research question of particularly where in the country these higher quality schools are.

The next explanatory variable, which is also categorical, that I am using to create a linear model is the predominant degree earned from that school, the options being a certificate degree, associates degree, and bachelors degree. This is in order to divide up the types of schools so that a higher level institution is not being compared to lower level schools that are not intended to give higher salaries after graduation.

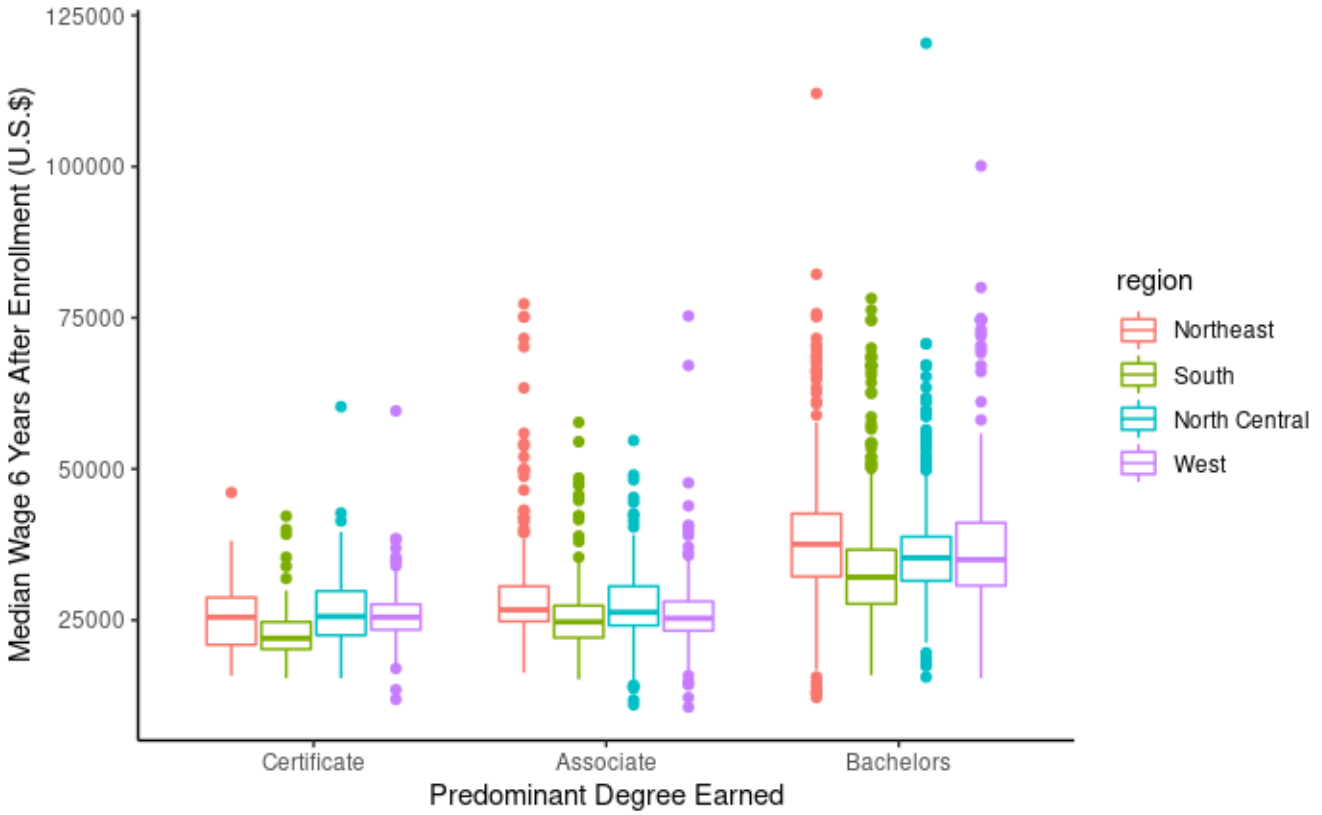
Another explanatory variable being used is the quantitative variable representing the average tuition paid of the school. This is another variable that can hopefully be controlled in order to see where the higher quality schools are compared to their price, since when in conversation about equity in the school system tuition needs to be accounted for those who cannot afford more expensive schooling.

The final explanatory variable I have included in my linear regression model is the size of the school population, another quantitative variable. Once again this is just another variable that if I am looking at the region spread of these high salary earning schools, it is a good variable to control because as schools get larger there is more variance in the wages students earn, and for the most part the trend with the schools has been if there are outliers in wage the outliers are higher than the majority of the population leaving larger schools more likely to have those outliers bringing their median salary up.

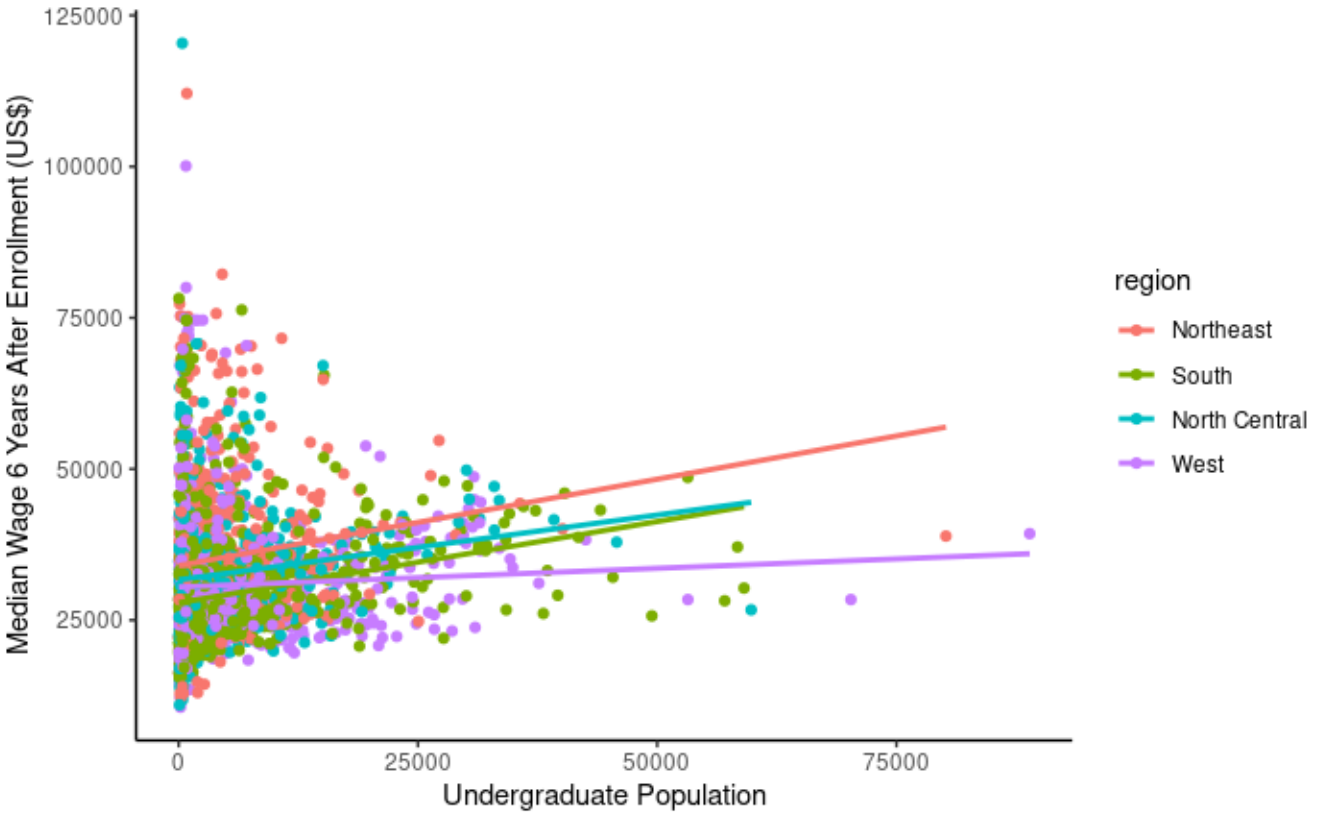
Sample Description

I did not filter my data for outliers because it is unfair to say, in the context of my research question which looks into whether or not people have these high quality schools that prepare for after college well, that schools that are present should not count just because they are much larger or smaller than other schools or just because they are expensive or tend to lead to higher salaries. It takes away too much from the question being considered. The cases were somewhat filtered though because the cases represent 3,676 colleges and universities from the entire United States and United States territories, but through my region variable I focus on only schools in the continental U.S., removing 339 schools in the process. This makes up a population of around 3,337 schools being observed out of nearly 5,300 schools in the country.

Visualization



This colored box-plot displays the median and IQR spreads of the median wages earned six years after enrollment at the schools from each region, grouping them by the predominant degree awarded from that school. Each degree level shows similar patterns when it comes to the medians of the median wages from each region as for all of them the northeast has the highest median wages six years after enrollment, followed by the north central region, then the west, with the south having the lowest median wages 6 years after enrollment. At the certificate and bachelor degree levels though, the west and north central show very similar medians for the median wages, but at the certificate degree level, a larger north central gives the north central a higher 75th percentile median wage and a lower 25th percentile median wage, while the opposite is observed at the bachelor degree level where the west has the greater spread. Another important note shown by this model is that all regions at all degree levels have median wage six years after enrollment outliers above the majority of the data while very few have outliers below.



This scatterplot displays the median wage earned six years after enrollment from schools in each region as the undergraduate populations of the schools increases. As seen by the regression lines in the model, on average each of the regions shows an increase in median wages as the undergrad population increases, with a relatively weak relationship. The steepest increase is seen in the northeast, followed by the south, then the north central region, and lastly the least average increase in median wage per undergraduate student is seen in the West. The outliers in this plot are seen to be at schools with higher populations and, once again, higher median wages six years after enrollment.

Model Selection

When deciding upon a model, I started with four models to explore which variables may have a causal effect on my outcome variable the median wage earned six years after enrollment. The original model just looked at two variables, the region in the U.S. the school was located in and the predominant degree earned from that school. The next model was an interaction model using those same two variables. The third model I used, which I called my “full” model, had four variables added together, the degree and region variables from before plus a variable for the net tuition per student of the school and a variable for the undergrad population. The final model or my “sub” model was the same as the full model only excluding the variable for the undergrad population.

The first test I ran on these variables looked at the adjusted r-squared values of each model, specifically which one was the greatest. The highest r-squared value belonged to the full model at 0.32, with the sub model following with an r-squared value of 0.30. This means that the full model was able to explain 32% of the variance in the median wage earned 6 years after enrollment in this dataset. I also looked for the model displaying the lowest residual standard error, which once again was the full model with a residual standard error of 8,420 over 3,240, meaning that over the 3,240 schools included

in this model, on average the model's predicted median wage six years after enrollment was $8,240 above or below the actually wage seen in the dataset.

Finally I created added variable plots for each of the models to see if any of the models showed variables that were unnecessary or redundant. In all of the plots, besides the interaction model, all the variables showed significant slope and therefore did not show redundancy or need to be removed from the model. The interaction added variable plot showed redundancy from the region variables, but with that plot already being ruled out from the r-squared and residual standard error values I left that alone. Instead what the other plots told me was that all the variables showed relationship to the median wage earned six years after enrollment, ultimately leading me to decide upon the full model that contained all of the variables

Final Model Statement

$$E[Median\ Wage\ Earned\ 6\ Years\ After\ Enrollment \mid Region,\ Degree,\ Net\ Tuition,\ Undergrad\ Population] = \beta_0 + \beta_1(South) + \beta_2(MidWest) + \beta_3(West) + \beta_4(Associate) + \beta_6(Bachelor) + \beta_7(Tuition) + \beta_8(Undergrad$$

Fitted Model

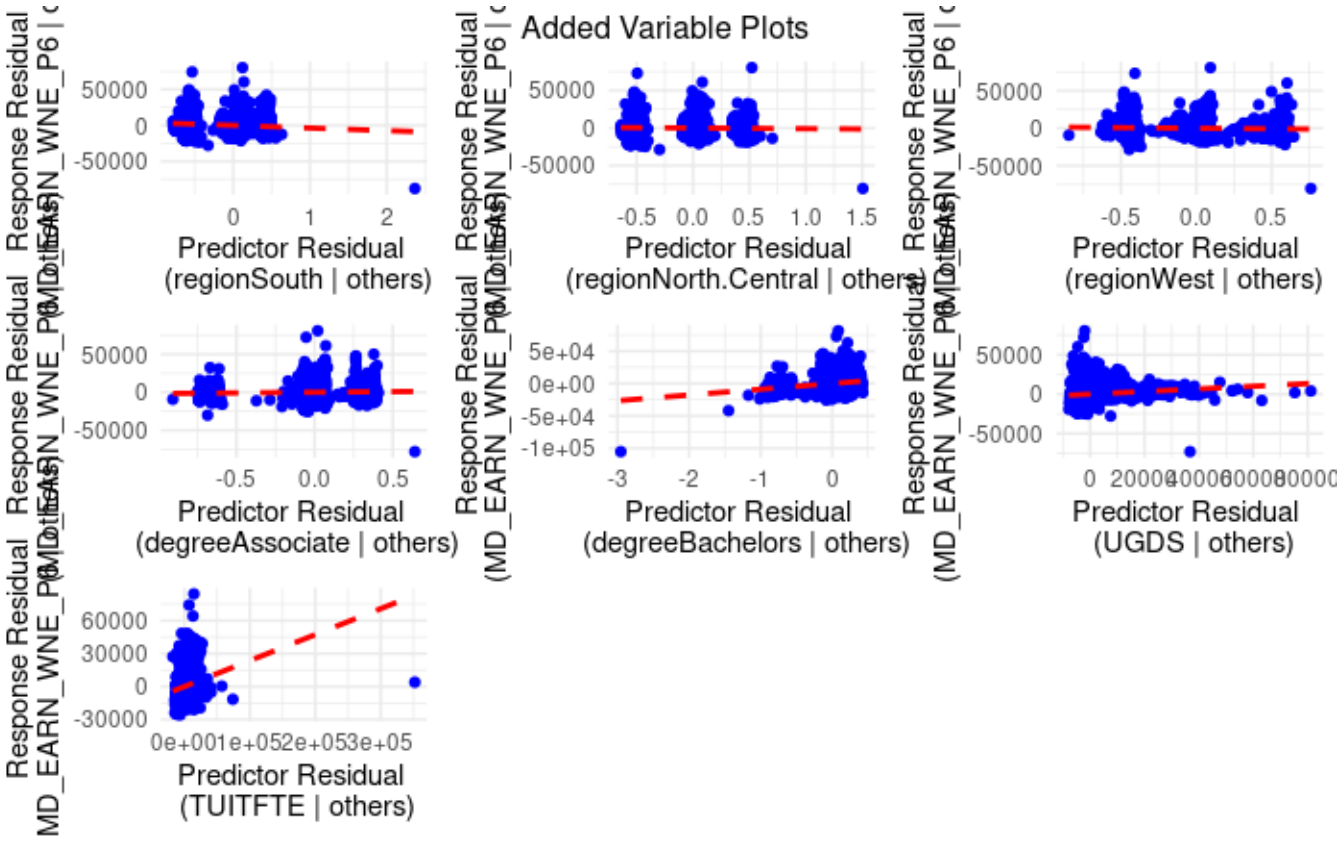
Slope Estimates	Estimates	Standard Errors	95% Confidence Intervals
South	**-3669.02**	**421.55**	**-4495.56 - -2842.49**
Midwest	**-926.77**	**446.31**	**-1801.86 - -51.72**
West	**-1525.57**	**477.03**	**-2460.88 - -590.27**
Associate	**1654.24**	**463.96**	**744.56 - 2563.92**
Bachelor	**8868.59**	**450.69**	**7984.92 - 9752.25**
Net Tuition	**0.17**	**0.021**	**0.125 - 0.209**
Undergrad Population	**0.24**	**0.016**	**0.205 - 0.27**

Interpretations

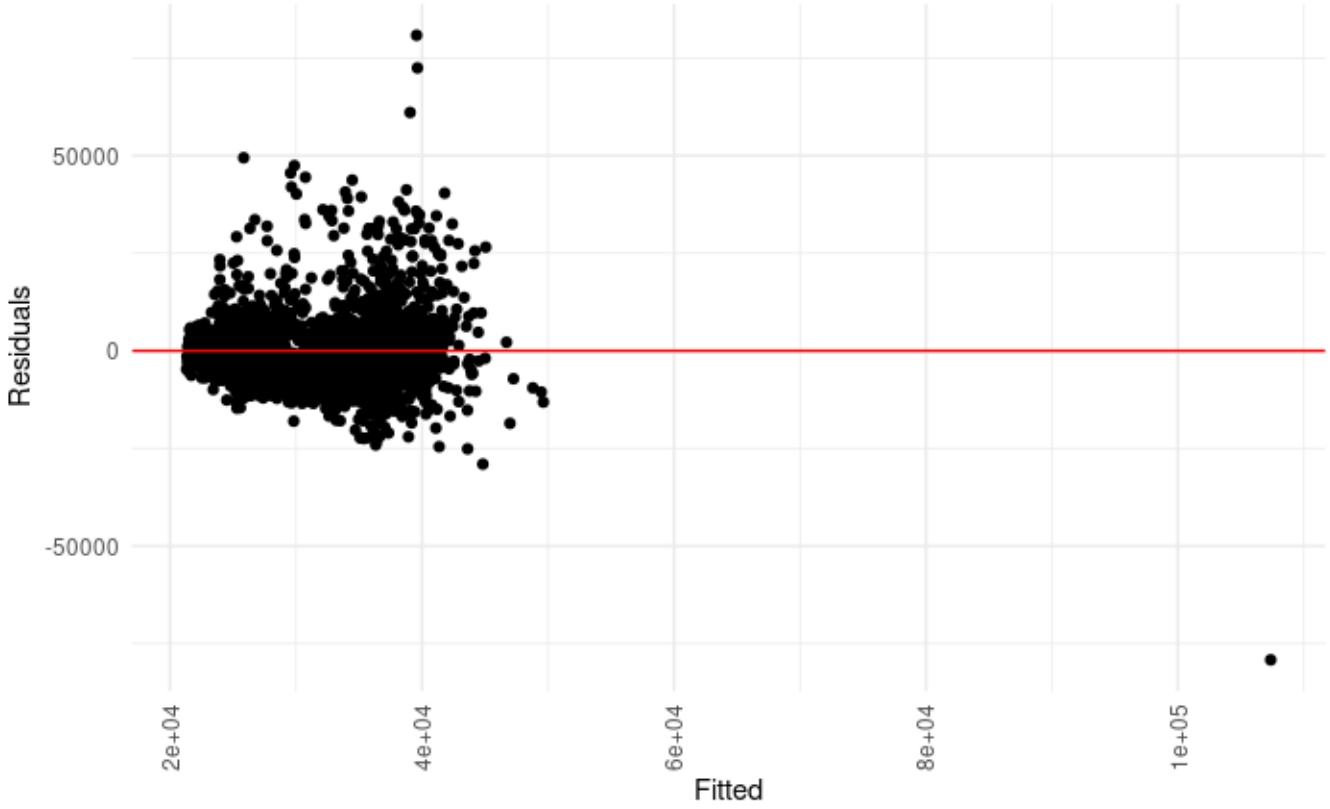
The first thing the estimates tell us about the regions is that the northeast, which is not seen in the table because it is represented by the intercept, has the highest average median wage six years after enrollment while holding the predominant degree earned, the net tuition, and the undergrad population of the schools constant. The next highest median wage is observed in the midwest which shows an estimated average median wage of $926.77 less than the northeast holding all other variables constant. The west follows showing estimated average wages of about $1,525.57 less than the northeast, with the south displaying the lowest average median wages six years after enrollment at about $3,669.02 less than the northeast, once again for the south and west while holding the predominant degree earned, the net tuition, and the undergrad population constant.

Since these are just estimates, we can get a better idea of whether or not a relationship exists by making a 95% confidence interval which would tell us that if we collected many many samples of similar data to represent the full population we would expect that 95% of the samples would produce estimates in the given intervals. The confidence interval for the midwest tells us that we are 95% confident that the true average decrease in median wage six years after enrollment, when holding the other variables constant, when moving from the northeast to each other region is between $52 and $1802 for the midwest, $745 and $2564 for the west, and finally $2842 and $4496 in the south. Along with that all the region variables have p-values lower than the significance threshold 0.05, meaning that if we decided to accept the null hypothesis as true and believe that there is no relationship between region and median wage six years after enrollment there is a less than 5% chance of us getting estimates as extreme or more than what we received in this study, a significantly low enough probability, along with none of the confidence intervals containing 0 is enough to reject the null hypothesis.

Model Evaluation



This added variable plot displays first of all, that all of the variables included in the model are correct to be included. We can tell because all of the variables show significant slopes, where a no slope would be a sign of redundancy and would need to be a variable that should be removed.Each of the variables show relatively strong linear relationships, but there are outliers left in the data, but in the context of this research question I think it is unfair to leave out the outliers as in many cases the outliers are more important since they tend to lean towards higher median wage, meaning that people living in that region have that close access to a high quality school.



I also looked into the residuals of my linear model, first creating a residual plot of the model. As seen in them mode, besides a few outliers and slightly more variance above the residual line than below, the residual plot shows little sign of any pattern and an equal spread as the fitted values increase. When looking into residual values, the model ended up showing an r-squared value of 0.32, telling us that only 32% of the variance in median wages was explained by the model, which is not great but was better than the other models I had looked into. This model also had the lowest residual standard error with a residual standard error of $8,420, meaning that we can assume with 95% confidence that the true median wage from a specific school is around $16,840 above or below the predicted value of our model. This model is the best of the options given, but when making predictions from it the error in the model should be remembered.

Multiple Logistic Regression Modeling

Variable Descriptions

My outcome variable is a binary, categorical variable that represents whether or not the school is a minority serving institution. I am attempting to explain this variable using three explanatory variables, the first of which is the median wage earned six years after enrollment as used before, since this once again explains the quality of the school as I look into whether or not minority serving institutions have similar median wages as the normal admittance colleges and universities.

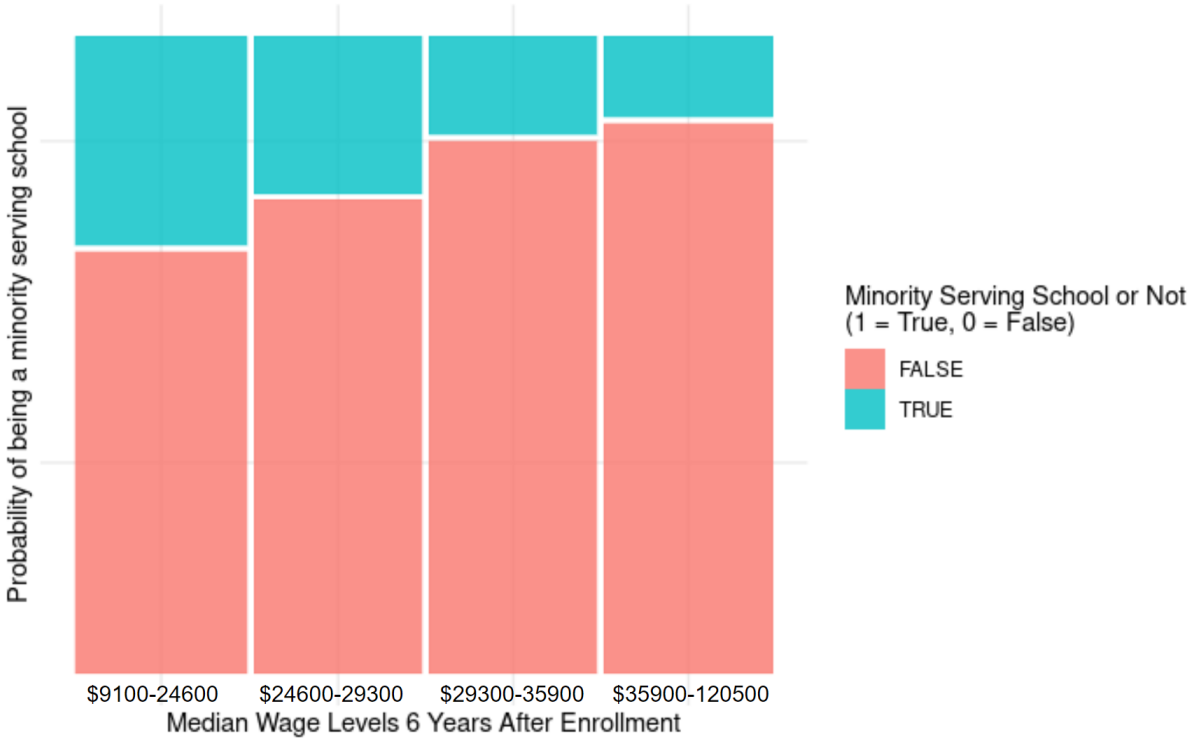
The next variable is the predominant degree earned, as similar to the median wages, this variable can hopefully give insight into whether or not the minority serving institutions are offering the same quality of degrees to their students as the non-minority serving institutions. The final explanatory

variable I used is the undergraduate population as something to explore because it is not fair to say that a 2,000 person minority serving institution is doing as much to bridge the education gap between minorities and white people as a 40,000 student minority serving institution.

Sample Description

Once again outliers were not removed from the dataset because the context of the research question would make that an unfair jump saying that a larger school or a school that leads to higher salaries just does not exist, and in many cases are more important than the other schools. Still cases did get filtered out through not having data in a particular section, being only graduate degree schools, or not being a part of the continental united states lens that I was focusing on. This left a total number of schools being represented at 3,262 schools.

Visualization



This stacked bar plot displays that as a school's median wage six years after enrollment for their students increases, it becomes less and less likely that the school is a minority serving institution. In the lowest quartile group for median wages, about 33% of the schools are minority serving institutions, but when you move up to the 25-50th percentile of median wages only about 25% of those schools are minority serving. The 50-75th percentile group of median wages shows an even greater decline with only around 10% of those schools being minority serving, followed lastly by the highest median wage group with only around 8% of them being minority serving. This displays a consistent decrease in the percentage of minority serving institutions as median wage earned from the school increases.

Model Selection

When selecting a model, I started with three options. My original model explored the outcome variable of whether or not the school is minority serving by looking at the variables for median wage

and degree. My “full" model added onto the original model by including two more variables which were the undergrad population and the region the schools were from and finally my “sub” model nested inside the full model, only removing the regions variable. The first thing I did to compare these model was create a predicted vs. actual boxplot for each of the three models, trying to find the model with the least overlap between the predictions of the actual minority serving institutions and the ones that are not. When doing this I found that my original model had noticeably the most overlap between the two boxplots, while the sub and full models were too similar to tell the difference, leading me towards removing the original model.

Next I calculated false positive, false negative, and the accuracy rates of each of the models. Once again the sub model and the full models had very similar results, with the full model only having false positive and false negative rates that were 1% better than the sub model and an accuracy that was only 1% higher than the accuracy of the sub model. I then ran a test to find the p-values of all of the variables in the full model, and what i found was that the midwest region showed a p-value of slightly over 0.1, well above the significance threshold of 0.05, meaning that there is over a 10% chance of getting test statistics for that variable as extreme as were seen or greater if the null hypothesis that there is no relationship between being in the northeast and the midwest and the odds of a school being a minority serving institution were true. Since this is well above the significance threshold, and also in attempts to keep the model as simple as possible, with results in the full and sub models, I ultimately decided to use the sub model for the rest of the study.

Final Model Statement

$$log(Odds[Minority\ Serving\ Institution|Wage\ 6\ Years\ After\ Enrollment,\ Degree, Undergrad\ Population\ ]) = \beta_0 + \beta_1(MedWage) + \beta_2(Associate) + \beta_3(Bachelor) + \beta_4(UndergradPop)$$

Fitted Model

Odds Ratios	Estimates	95% Confidence Intervals
Median Wage 6 Years After Enrollment	**0.9999039**	**0.9998889 - 0.9999185**
Associate’s Degree	**1.7537**	**1.3473 - 2.2963**
Bachelor's Degree	**2.0319**	**1.5361 - 2.7046**
Undergraduate Population	**1.0000718**	**1.0000595 - 1.0000844**

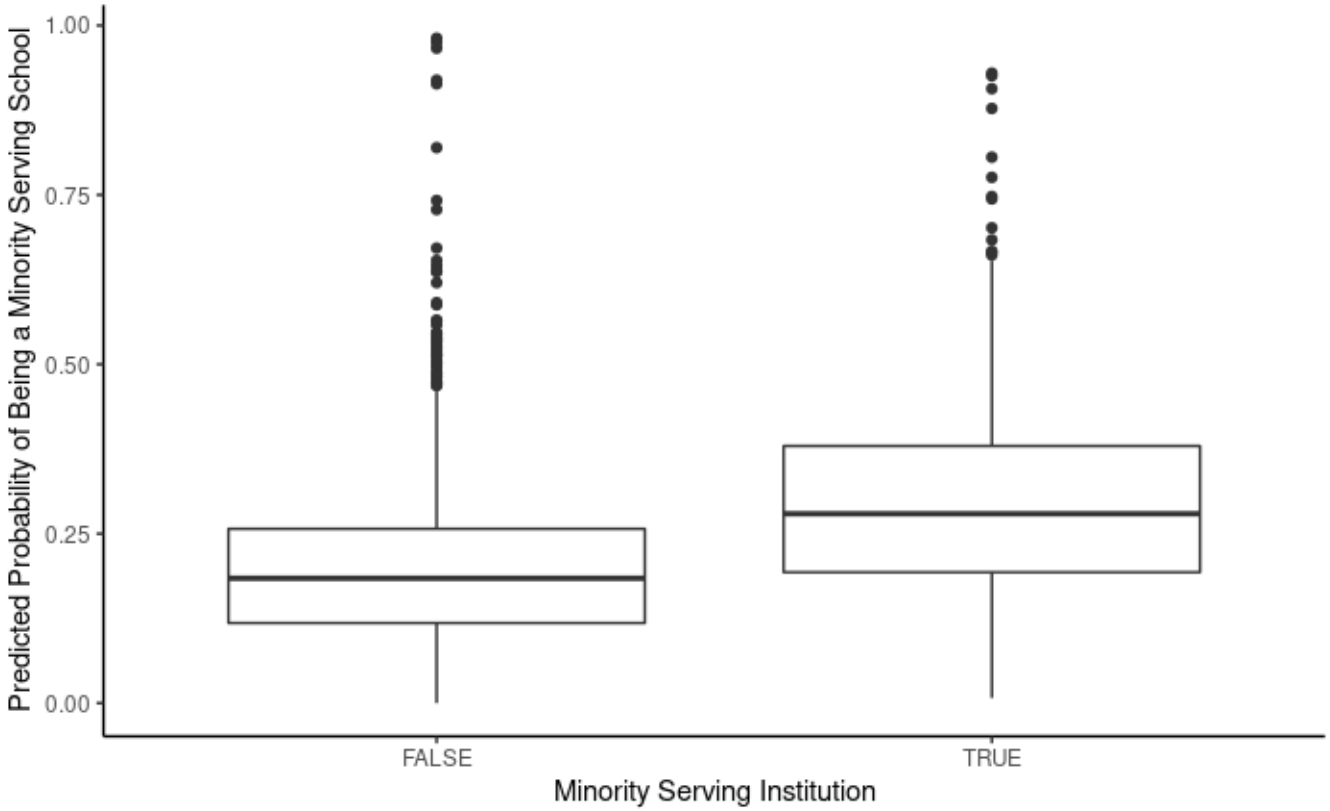
Interpretations

The first coefficient to look at is the coefficient for the median wage six years after enrollment, which is 0.9999039, telling us that for each dollar increase in the median wage, the odds of that school being a minority serving institution is multiplied by a factor of 0.9999039, while holding the predominant degree earned and the undergraduate population constant. This may not seem like a significant value, but the confidence interval still does not include 0 as we are 95% certain that the true multiplicative factor on the odds of being a minority serving institution when increasing the median wage earned after graduation by a student is between 0.9998889 - 0.9999185, once again while holding the predominant

degree earned and the undergraduate population constant. This still might not seem like much but when discussing a change of a single dollar on a scale that stretches to the hundreds of thousands this factor actually proves practically significant.

The next variables of importance are the variables for different degrees. The associate degree coefficient tells us that when going from a school that predominantly awards certificate degrees to a school that predominantly awards associate degrees,we estimate the odds of that school being a minority serving institution increase by a multiplicative factor of 1.8, with us being 95% certain that the true multiplicative factor lies between 1.35 and 2.30 if we hold the median wage and undergrad population constant. We notice an even bigger increase in odds when moving up to a school that predominantly awards bachelors degrees, as we estimate that if median wage and undergrad population are held constant, going from the certificate degree level to a predominantly bachelor degree school the odds of the school being a minority serving institution is multiplied by a factor of 2.03 with us being 95% certain the true multiplicative factor on the odds of being a minority school falls between 1.54 and 2.70.

Model Evaluation



The first thing I looked at when evaluating this model was its predicted probability boxplots shown above, which were mainly in order to choose a threshold level from studying where the center between the median of minority serving institutions and normal admittance schools based on the probability of our model predicting it to be a minority serving institution. The chosen threshold was 0.23, which was used to find first false positive and false negative rates. The false positive rate derived was about 0.33, telling us that about 33% of schools that were not minority serving institutions were falsely predicted to be minority serving. The false negative rate was very similar at about 0.36,

meaning that of the schools that were minority serving institutions, about 36% of them were falsely predicted to be normal admittance schools. This gave an overall accuracy of the model to be about 68% of the schools had accurate predictions of their type from their model. This tells us the model is not perfect at predicting whether or not a school is minority serving and should therefore be looked at with some sort of skepticism when evaluating its outcomes.

Conclusions

General Takeaways

The most important takeaway from my linear model is what regions of the United States have the most nearby access to high quality schools that set students up to have the highest median wage six years after enrollment possible. What the linear model told us is that we expect there to be a relationship between where you are and the amount of money you will make after college, displaying that schools in the northeast have the highest expected average median salaries, followed by the midwest, then the west, and lastly the south has the lowest average median salaries amongst there schools. All of this is looked at through a lens where the predominant degree awarded, the undergrad population, and the net tuition per student were all held constant.

The first important takeaway from the logistic model that explored the probability of a school being a minority serving institution looked into the effect of median wage six years after enrollment on the probability of a school being minority serving, with predominant degree level and undergrad populations constant. The model predicted a negative impact of median wage on the probability of a school being minority serving, meaning that the better a school sets a student up for after college the less likely the school is minority serving. The other spectrum though was seen when median wage was held constant alongside undergrad population and instead we looked at the predominant degree awarded from a school's affect on the school's odds of being a minority serving institution and found that as the degree level increased, the predicted odds of being minority serving increased as well very significantly. This tells us that a greater percentage of the minority serving schools are in the higher education levels than in lower degree levels.

All this tells us, that at least when looking through the lens of median wage six years after enrollment and predominant degree earned from schools to judge their quality, there are certainly signs of inequity. Certain regions, particularly the northeast region, of the country seem to have more access to these higher quality colleges and universities than other parts, and also it seems that the minority serving institutions in America tend to set their students up to make lower wages than normal admitting schools on average.

Limitations

As an observational study we first must understand that no causal conclusions can be made without having any type of controlled experiment done. Along with that, since a lot of the data is coming directly from the colleges and universities, it may have been in schools' best interest to either leave some data out or potentially not participate in the study at all if their data negatively represents the school, resulting in some non-response bias that could inflate some of the averages found in this data set. To add on to that further schools may have also given numbers that are slightly inaccurate that could boost how the school looks, boosting the averages in the data set further. With the overall population of colleges and universities in America being very well represented by this data set, this

being an observational study, and the potential for non-response bias to have occurred, we should view this data with some skepticism.

Neither of the models looked at in this study were perfect predictors of their outcomes. The linear model first of all has limitations seen through only around 35% of the variation in the median wage outcomes were explained by the model, as there are schools on both ends of the quality spectrum in all regions of the country. Therefore when it is said that the South has the lowest estimated average median wage six years after enrollment, it is not to say that top quality schools do not exist in the south, but instead that they are either less frequent or there are more schools of lower quality. The similar idea is seen in my logistic model which only showed around 68% accuracy in its testing. This means that around 32% of predictions that are made by the model are expected to be inaccurate and misleading, even though false positive and false negative rates are similar. Therefore any outcomes seen from these models should be looked at with skepticism.

It should also be noted that when assessing equity in our higher education systems, more should be taken into account that what has been looked at in this study. For example, with my linear model which looks into the quality of a school, median wage six years after enrollment is not the only way to measure a school's quality. That can also be measured by test scores, happiness of the students, cleanliness and in many other variables that were not described in this dataset. Similarly in the logistic model median wage and degree levels are not the only way to assess if minority serving institutions are as beneficial to their students as normally admitting schools.