1-1-2016

Factors of Educational Attainment

Connor Hansen
George Fox University, chansen13@georgefox.edu

Sarah Ng
George Fox University, sng13@georgefox.edu

Factors of Educational Attainment

Connor Hansen, Sarah Ng

George Fox University

Abstract

The following research has been conducted in order to establish both the economic and statistical significance of certain factors that affect the percentage of those who hold a bachelor's degree by age 25. We look at what demographics are/are not important when predicting who will obtain a bachelor's degree by age 25, and list some options we think would increase the percentage of those who hold a bachelor's degree, based on our research. We stipulate that increasing the percent of the population holding a degree will better prepare people for success and a higher standard of living in our rapidly changing economy. To test our hypothesis, we have estimated an equation and run regressions on our data from which we can draw insight about where we ought to focus our efforts in education policy.

Keyword: Educational Attainment, Education Data, State Education Data

JEL Code: I200

Introduction:

The purpose of this research is to examine the factors that may determine one's probability of attaining a bachelor's degree by the age of 25. We aim to understand which socio-economic variables play a significant role in the educational achievement of an individual. Our research questions is as follows: What influences the level of educational attainment, specifically the percentage of the population that holds a bachelor's degree, using state-level data and cross sectional analysis?

It can be argued that having a bachelor's degree in the U.S. is important to increasing one's lifetime earnings, and thus raising one's standard of living. Standard of living issues are one of the greatest challenges facing economists and policy-makers today in the face of an increasing gap between the poor and rich, and a frustrated middle-class. Another key reason to be concerned about the percentage of the population who hold bachelor's degrees is the changing nature of our economy; we are in a period of transition where manual labor, by-and-large, is becoming obsolete as many industries are increasingly substituting capital in for labor. In time of economic stagnation, growing human capital and improving societies capacity for innovation through education are ways to combat said stagnation. A bachelor's degree is a way to increase one's skills and marketability in securing a job in the new economic landscape. Overall, our research is concerned with who should receive the most attention from educational policies.

Many of our variables cannot be changed easily. By looking at some factors which may have a significant effect on whether or not an individual obtains a bachelor's degree, we can gain insight into how policy can be utilized to most effectively to address the problem. For example, if we discovered that racial demographics play a significant role in determining the likelihood of earning a bachelor's degree, we might recommend that educational spending focus additional

funds to get those disadvantaged racial demographics more involved in educational attainment through scholarships and outreach programs.

Presently, we will provide an overview of the data we used for our research -- listing and describing the regressors and dependent variable we chose to include, and justifications for our data sets. Also included will be a table of our data's key statistics. Part three will address our methodology in estimating our regression accompanied by an explanation. Next, we will discuss those results and any economic/statistical significance in our findings, and what they mean to those who formulate policy aimed to increase either the percentage of the population who have a bachelor's degree or aim to increase the lifetime earnings/standard of living of individuals in the new economy. After discussing and interpreting the data, we will make some concluding thoughts on our research.

We found that our data was best explained through a non-linear OLS model that incorporates the natural log of median household income. Across the four regressions that will be presented in this paper, median household income, population density per mile, and unemployment were consistently statistically significant. In regards to policy formation, we believe that the economically significant impact that median household income plays on bachelor's degree attainment can be used to influence, or at least inform, state-level policy makers who recognize the importance that education plays, especially in the new economy that we currently exist in.

Data Overview:

For this research, we are examining a number of economic indicators which we suggest might have an effect on the percentage of the U.S. population over the age of 25 who hold a bachelor's degree. We propose to divide our sample to look at all fifty states in the U.S.

(excluding Washington D.C. for data continuity).The regressors which we intend to test on the dependent variable described earlier are the racial demographics (white, black, Asian, and other), the gender demographics (male or female), the median household income, the population density per square mile, the high school graduation rate, and the unemployment rate.

Data about racial demographics was collected from the Census Bureau from the data set entitled "Annual Estimates of the Resident Population by Sex, Race Alone or in Combination, and Hispanic Origin: April 1, 2010 to July 1, 2015". All data comes from the year 2014.

For our analysis, we have calculated the percentage of each key demographics' presence (white, black, and Asian), and have grouped the remainder of the categories into that of "other". "Other" contains Pacific Islander, Native American, etc. Hispanic has been left out since it is not technically its own category in the data sets we used; it is contained (mostly) within the "white" category.

In regards to gender demographics, we once again calculated the percentage of each state's population who fall under either female or male. This data was acquired from the Census Bureau's data set titled "Annual Estimates of the Resident Population by Single Year of Age and Sex: April 1, 2010 to July 1, 2015". All data comes from the year 2014. It is important to note, however, that due to the low variation in the male-to-female ratio between states, dropping this variable may lead to more accurate results.

Data about the household median income was again collected from the Census Bureau's data set titled "Median Household Income by State", specifically from the year 2014. The data is presented in dollars earned, and adjusted in 2015 real dollars.

Data about the population density per square mile in 2014 was (albeit shamefully, but the other searches yielded no results for free) gathered from Wikipedia's article "List of U.S. States

By Population Density".

We gathered state level data for the percentage of the population that had attained a bachelor's degree for the year 2014 from the U.S. Census Bureau's "Educational Attainment: 2010-2014 American Community Survey 5-Year Estimates". This data is collected yearly through a survey of about 3.5 million households selected from every county in the nation. It is important to note that for the one-year estimates surveys apply to areas with over 65,000 people.

State unemployment rates come from the Bureau of Labor Statistics. The unemployment rate refers to the percent of the population in the labor force that is not employed and is actively seeking employment. The specific set of data within the BLS was the" Local Area Unemployment Statistics" gathered annually for states. The LAUS uses the underlying ideas found in the Current Population Survey and the estimates themselves are created using data from the CPS, the Current Employment Statistics survey, state unemployment insurance systems, and the Census Bureau's American Community Survey.

The data available for measuring high school education is the "Public High School 4-year Adjusted Cohort Graduation Rate, by Selected Student Characteristics and State: 2010-11 through 2013-14." The ACGR refers to the graduation rate of students that attain a regular diploma in the standard four years. It is adjusted for transfer students, emigrates, and deaths. We did not use the data from this set that was sorted by student characteristics. This data was published by the National Center for Education Statistics and gathered by the U.S. Department of Education. Figure 1 is a summary of some key statistics from the research conducted for this study.

Fig. 1: Key Statistics for Data Collected Across the 50 States

	Mean	Median	Minimum	Maximum	St. Dev.
Percentage of Population: Black	0.12	0.09	0.01	0.38	0.10
Percentage of the Population: Asian	0.05	0.03	0.01	0.56	0.08
Percentage of the Population: Female	0.51	0.51	0.47	0.52	0.01
Median Household Income	54962	54613	35521	76165	9109
Population Density Per Square Mile	200.35	105.90	1.30	1210.10	265.51
Unemployment Rate	5.75	6.00	2.70	7.90	1.26
High School Graduation Rate	82.54	84.50	69.00	91.00	5.66
Percentage of Population over 25 that Holds a Bachelor's Degree	18.02	18.05	11.60	23.70	2.73

To further the reader's understanding of what our data looks like, we have supplied a table in *figure 2* to summarize the data from a few notable states. They are ordered in relation to their respective percentage of the population holding a bachelor's degree. West Virginia had the minimum of 11.6% of the population holding a bachelor's degree, Iowa had closest to the mean (18.02%), Wisconsin had closest to the median (18.05%), and Colorado had the maximum at 23.7%. It is interesting to note that at this point, just by examining these four, it appears there is a strong correlation between a state's median income, unemployment rate and percentage of the population holding a bachelor's degree.

Figure 2: Selected States

	West Virginia	Iowa	Wisconsin	Colorado
Percentage of the Population: White	0.95	0.94	0.90	0.92
Percentage of Population: Black	0.05	0.05	0.08	0.06
Percentage of the Population: Asian	0.01	0.03	0.03	0.04
Percentage of the Population: Female	0.51	0.50	0.50	0.50
Median Household Income	39,552.00	57,810.00	58,080.00	60,0940.00
Population Density Per Square Mile	76.70	55.90	106.60	52.60
Unemployment Rate	6.60	4.20	5.40	5.00
High School Graduation Rate	85.00	91.00	89.00	77.00
Percentage of Population over 25 that Holds a Bachelor's Degree	11.60	18.00	18.10	23.70

Methodology:

For this study, we are using cross sectional data. Specifically, we will be looking at the data for our different variables for the year 2014 across all fifty states of the United States. Our estimated regressions can be used to see what has significant, both statistical and economic, impact on education attainment as well as “construct” hypothetical states to see what our

regression will predict about the level of educational attainment. Additionally, we can use our regression to predict what would happen in educational attainment in a particular state given a change in these significant variables.

Before running any of the regressions, we hypothesize that our results will indicate that the percentage of the population that is black and unemployment will have an inverse effect on the regression, i.e. when they increase, the percentage of those holding a bachelor's degree will decrease. Unfortunately, African-American communities tend to have high crime rates, underdeveloped school systems, and less economic resources in general. Additionally, unemployment is associated with an unhealthy economy as a whole in which the focus would be more on getting a job and making a livable wage than investing in one's human capital. We tentatively believe that the following regressors will be positively correlated with our dependent variable: percentage of the population that is Asian, because the Asian culture tends to stress educational achievement; the percentage of the population that is female, because higher education institutions tend to have a higher population of females than males; the median household income, because a higher income corresponds with a higher ability to pay; the population density per mile, because colleges and universities are generally more prevalent and therefore easier to access in highly populated areas; and, the high school graduation rate because in order to earn a bachelor's degree you must first attain a high school diploma or general education diploma.

Our data was collected in such a way that we are unable to run regressions using probit, logit, dummy variables, or interaction terms; therefore, our main tool for modeling will be the ordinary least squares estimator (OLS) for multiple regression.

We will use apply certain techniques that we have learned over the course of the semester

to describe accurate models which can determine the influence of the previously discussed regressors on the level or bachelor's degree attainment by state. Because we are using OLS, we take on the assumptions that the error term has a condition mean of zero given the independent variables, the independent variables are independently and identically distributed draws from their joint distribution, large outliers are unlikely, and there is no perfect multicollinearity (Stock & Watson, p. 201)

The equation we will be using to run multiple regressions on will be:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + u$$

The percentage of a state's population that has attained a bachelor's degree is a function of race, gender, income, population density, the unemployment rate, and high school graduation rate. Let race_black = X_1, race_Asian = X_2, female = X_3, median_HH_income= X_4,pop_density_mile= X_5, unemployment = X_6, and HS_graduation = X_7. 'u' refers to the error term which accounts for the difference between the real Y value and our regression function. This regression excludes the racial qualifier "white" and "male" have been left out of the regression in order to avoid perfect multicollinearity on race and gender. This is the linear equation we start with; however, in our results and interpretations we will discuss deviations and additions that will include leaving out some variables and applying nonlinear variables such a logarithms of selected variables. We believe that median household income, population density per square mile, unemployment, and high school graduation all have the potential to be more accurate as nonlinear variables. In other words, we believe that these variables will have a decreasing marginal impact on the dependent variable.

Results and Interpretations:

For our first test, we ran a regression (1) which included the variables which we discovered were statistically significant from running other regressions. These variables are: the percentage of the population which is female, the median household income, the unemployment rate, and the high-school graduation rate. We used a simple OLS regression with a cross-sectional interpretation. What we found after testing this regression is that all the variables in this regression were comfortably within the range of statistical significance. Upon examination of what the regression says, it was surprising to see that as the percent of the Asian population and/or the high-school graduation rate increases, the probability of an individual getting a bachelor's degree decreases. This doesn't quite make sense, given that the average Asian's IQ is 6% higher than whites -- East Asians with an IQ of 106, and whites with an average IQ of 100 (Jensen & Rushton, 2005). On the same note, it makes sense that an increase in the African-American population decreases the average percent of people who achieve a bachelor's degree, given that African-Americans score lower than both whites and Asians. However, this can't totally explain educational attainment, since IQ scores do not always indicate how an individual may perform in college; there are so many more factors out there that contribute to an individual's education. Among them are median household income, population density, and unemployment. The results for these variables are all intuitive -- as median household income within a state rises, we expect there would be more surplus money to spend on a child's higher education. As unemployment rises, money becomes more tight, and fewer people can afford to send their child to college. As population density increases, educational attainment increases, which makes sense, since it implies that with more people come more resources with which to raise children and prepare them for college. In addition, more densely populated areas tend to

have more skilled labor than farm labor available; it provides a greater opportunity for people to get into higher-paying jobs, thus prompting more people to get a college degree.

Our second regression is made up of all the variables aside from "female" due to the warning from the GRETL software that indicated our data matrix was close to singularity. This warning was resolved when we excluded "female." Conceptually, the differences from state to state regarding the percentage of population that was female was very small, the split was basically 50-50 across the states, and therefore was unlikely to have statistical or economical significance in comparison to the other independent variables. Additionally, since we are looking at the number of total bachelor's degree holders (regardless of gender), we decided this exclusion was appropriate. Our R^2 was at a reassuring .72, which is not significantly different from our other regressions. Just like the first regression, median household income, population density, and unemployment are all statistically significant in this regression.

The third and fourth regressions we ran included the same variables as the second one, namely all the ones we mention in the report aside from female and percentage of the population that is white; however, the difference is that two variables were replaced with their natural logarithm: median household income and population density per capita. Additionally, this meant that the dependent variable also had to be in natural log form. Using the natural log means that this is now a nonlinear function. Conceptually, it makes sense that the effect of median household income and population density per capita would have diminishing marginal returns. For example, an increase in income by one dollar would have a larger effect on the whether or not a bachelor's degree is achieved for a person with a low wage than a high wage. Beyond a certain point, everyone earning said income or higher can afford to go to a university. Similarly, it can be hypothesized that states with more urban areas will have better access to colleges;

however, at some point increasing the density by one person is no longer going to make a noticeable difference. It is important to note that while nonlinear modeling does a better job conceptually at explaining our data, there were difficulties in keeping the regression from giving off warnings about multicollinearity when we introduced non-linear terms into the equation.

The third regression reflects the nonlinear nature of median household income; therefore, the dependent variable of percentage people over the age 25 with a bachelor's degree per state is also in logarithmic form. In this regression, the population density per mile, the unemployment rate, and the natural log of median household income are all significant at the 95% confidence level with the latter being significant at the 99% confidence level. The R-squared, and the adjusted R-squared, are the highest for this regression out of all four at 0.746 and 0.710 respectively. Additionally, this regression has the smallest standard error. This indicates that our hypothesis of the data being better explained with a nonlinear method holds true.

In addition to being statistically significant, in regression 3 the natural log of median household income is economically significant. Because this is a log-log regression, we interpret this data as a 1%change in X correlates with a beta percent change in the dependent variable. In other words, beta1 is the elasticity of Y with respect for X1. For example, if you took the different median incomes from the states with the highest and lowest values, the increase in the state level of bachelor's degree attainment would be predicted at 37%. Clearly, this holds enough economic significance to grab the attention of policymakers who recognize the importance of higher education for their local economies. One way to use this research practically would be to focus on not only providing additional assistance to lower-income households in regards to increasing educational attainment, but also to fix, or at least further, the process of equalizing,

the root causes of low-income. We believe that the significance of median household income signals a greater underlying issue that is not addressed in this report.

The fourth regression switched the natural log of median household income for the natural log of population density per square mile. While the r-squared and adjusted r-squared values are not as high as the 3rd regression, they remain higher than the first 2 linear models. On the same note, the standard error for this regression is lower than when the natural log of median household income was used but smaller than for linear models. Despite this, this model has more variables holding significance than the previous three regressions with percentage of population that is Asian and high school graduation rate now being significant at the 90% level. In this regression, only population density is economically significant. Here, taking the percent change from the min and the max leads to a 31.6% increase in Y. A possible extrapolation from these results would be that states with large rural populations should pay attention to the ease of access their residents have to education and the incentives for making the effort to get somewhere with bachelor's degree programs.

To give a fair criticism about our research, there were a few flaws that may be the reason for some of the surprises in our regressions. We had to deal with and skirt past some problems with possible multicollinearity (there may be some lingering in the data, of which we are unaware). Also, by choosing data from 2014, we are subject to distortions of data from that year -- perhaps it was a bad year for one state, and a year of plenty for another. Another problem could be omitted variable bias -- there are likely other variables out there that could better explain educational attainment. For example, some measure of opportunity or state programs to help fund college education could have been useful. Perhaps a measure of cost of living is missing from our data. Finally, our study may suffer from simultaneous causality. For example,

bachelor's degree attainment may be affecting median income and the unemployment rate. In the end, it is best to take these results with a grain of salt, as they are imperfect, but by no means should these results be dismissed altogether. Figure three contains the results from the 4 most significant regressions that we ran in Gretl.

Fig. 3: OLS regressions where the dependent variable is the percent of the population over 25 that holds a bachelor's degree

Regression	(1)	(2)	(3)	(4)
race_black		−0.540	0.061	-0.121
		(2.95)		(0.185)
race_asian		−2.096	-0.122	-0.233*
		(2.497)	(0.118)	(0.138)
median_HH_income	0.0002***	0.0002		.00001***
	(0.00005)	(0.00004)		(.000002)
ln(median_HH_income)			0.689***	
			(0.126)	
pop_density_mile	0.001**	0.002	0.0001**	
	(0.001)	(0.001)	(0.00001)	
ln_pop_density_mile				0.034***
				(0.011)
unemployment	-0.506***	−0.653	-0.035**	-0.041***
	(0.178)	(0.265)	(0.015)	(0.011)
HS_graduation		−0.063	-0.004	-0.006*
		(0.051)	(0.003)	(0.003)
const	9.506***	15.658	-4.153***	2.796***
	(2.38)	(5.527)	(1.388)	(0.294)
S.E. of regression	1.531	1.547	0.085	0.088
R^2	0.704	0.718	0.746	0.728
F-Statistic	36.71	27.723	32.031	24.738
Adjusted R^2	0.685	0.678	0.71	0.69
p-value (F)	0.000	0.000	0.000	0.000

Conclusion:

After examining the data and regressions in our research, we have come to the conclusion that educational attainment is indeed affected by most of the variables we have chosen. Our state-level cross sectional data was best explained through a non-linear OLS model. The three variables that have less of an effect than anticipated are the high-school graduation rate, percentage of the population that is Asian, and the ratio of males to females within a state. The first two are hard to explain away -- why they had a negative effect on educational attainment is something that needs to be examined more carefully. The most likely explanations for these anomalies is that our data may not have been manipulated properly, or perhaps our variables were not relevant to our study in how we applied them. In regards to the percentage of females, the data error was certainly not anticipated, but proved largely irrelevant, since the difference in ratios between states is not significant enough to determine an impact.

While our hypothesis was not correct in predicting what would significantly decrease the level of bachelor's degree attainment in the state, we did end up with some convincing data that a state's median income, population density, and unemployment rate do indeed have a significant effect on an individual's opportunity to get a college degree. Specifically, median income and population density have a non-linear effect and demonstrated economical significance that should be taken into consideration as local and state level government personnel look to effectively decrease the impact of national economic stagnation on their communities. Human capital, which is inarguably improved by furthering education, is at the heart of innovation, growth, and overall success.

Works Cited

American FactFinder. (2015, July 1). American FactFinder - Results. Retrieved November 10, 2016, from http://factfinder.census.gov/faces/tableservices/jsf/pages/productview .xhtml?src=bkmk

Bureau of Labor Statistics. (2016, February 26). Local Area Unemployment Statistics: Unemployment Rates for States. Retrieved November 9, 2016, from http://www.bls.gov/lau/lastrk14.htm

Jensen, & Rushton. (2005, April 26). Race differences in average IQ are largely genetic. Retrieved December 1, 2016, from http://www.news-medical.net/news/2005/04/26/Race-differences-in-average-IQ-are-largely-genetic.aspx

List of U.S. states by population density. (2016, March 5). Retrieved November 10, 2016, from https://en.wikipedia.org/wiki/List_of_U.S._states_by_population_density

National Center for Education Statistics. (2016, January). Public High School 4-year ACGR, by Selected Student Characteristics and State: 2010-11 through 2013-14. Retrieved November 9, 2016, from http://nces.ed.gov/programs/digest/d15/tables/dt15_219.46.asp

U.S. Census Bureau. (2016). Educational Attainment: 2010-2014 American Community Survey 5-Year Estimates. Retrieved November 9, 2016, from http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=CF

U.S. Census Bureau (2015, July 1). Historical Income Tables: Households. Retrieved November 10, 2016, from http://census.gov/data/tables/time-series/demo/income-poverty/historical-income-households.html

U.S. Census Bureau (2015, July 1). Population Estimates. Retrieved November 10, 2016, from https://www.census.gov/popest/data/state/asrh/2015/index.html