Applied Econometrics Project 3

Prachi Shah

College of Business, The University of Akron

3250:325 Applied Econometrics I  
Dr. Amanda Weinstein  
May 2, 2022

**Relationship between Educational Attainment and Wages:**

Wages earned by individuals is generally assumed to be correlated with the years of education obtained by him/her. To begin my analysis for Wages and Educational Attainment, I extracted two different data sets such as Average Earnings Per Job (millions) in 2019 and Educational attainment across states in 2019 from Bureau of Economic Analysis and U.S. Census Bureau respectively.

Following tables 1 and 2 show the top 10 states according to average earnings and educational attainment (%) respectively.

**Table 1:**

|  |  |
| --- | --- |
| Top 10 states in Average Earnings per job (2019) | |
|
| States | 2019 |
| New York | $ 92,061.00 |
| Massachusetts | $ 89,258.00 |
| California | $ 87,226.00 |
| Connecticut | $ 83,968.00 |
| Washington | $ 83,904.00 |
| New Jersey | $ 82,047.00 |
| Alaska | $ 81,213.00 |
| Maryland | $ 80,157.00 |
| Illinois | $ 77,248.00 |
| Virginia | $ 75,465.00 |
| *Source*: Bureau of Economic Analysis | |

To calculate the % of educated people, I divided total number of individuals in a state by individuals with bachelor’s degree from ages 25-44 within the state. This would certainly create bias because individuals with high school degree, diploma, etc. are not included in the analysis but they would be considered educated.

**Table 2:**

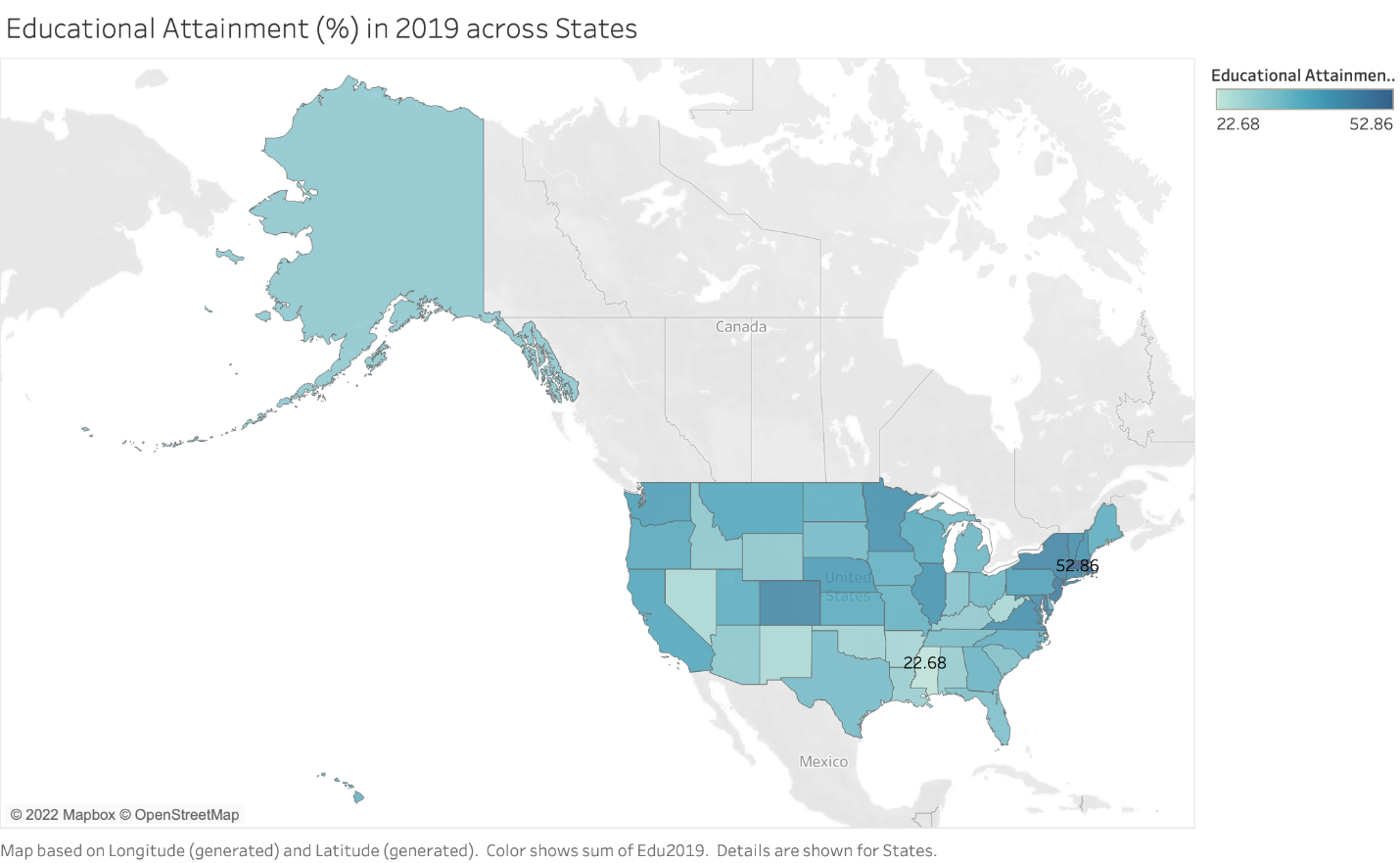
|  |  |
| --- | --- |
| Top 10 states in Educational Attainment in 2019 (%) | |
|
| States | 2019 |
| Massachusetts | 52.9 |
| New Jersey | 48.1 |
| New York | 45.6 |
| Colorado | 45.1 |
| Connecticut | 44.6 |
| Maryland | 44.2 |
| Vermont | 43.8 |
| Virginia | 43.5 |
| Minnesota | 43.5 |
| Illinois | 42.4 |
| *Source*: U.S. Census Bureau | |

**Figure 1:**

*Source*: U.S. Census Bureau (2021), Bureau of Economic Analysis (2021)

Figure 1 shows the relationship between average earnings and % of educated people across states in 2019. As shown in the scatterplot, both variables seem to have a positive correlation, i.e., as more percentage of people get educated in a state, average earnings of the state increase respectively. As per the equation, with 1% increase in number of educated people, average earnings of a state increase by $977.35.

**Figure 2:**



*Source*: U.S. Census Bureau

Figure 2 shows the educational attainment (%) across states in 2019. The state with lowest percent of educated people is Mississippi with 22.7% and the state with highest percent of educated people is Massachusetts with 52.9%.

The darker shades of blue determine more % of educated individuals in the respective state in the following visualization.

To better understand the omitted variables in the analysis, multiple other variables such as hours, employment status, age, gender, race, industry, occupation, and marital status were extracted from IPUMS USA.

To get a better idea of hourly wages, throughout the paper log(HourlyWages) is used which would help de-scaling the effect of large numbers or huge outliers.

**Figure 3:**

Chart, histogram

Description automatically generated

*Source*: IPUMS USA

Figure 3 shows the relationship between log of hourly wages and average of educational attainment across all states. On the x-axis is the average educational attainment between 22.7% and 52.9%. The state with highest ln(HourlyWages) is California with 38.2% educational attainment, and the state with lowest ln(HourlyWages) is Wyoming with 29.81% educational attainment. This graph shows that it is not mandatory for the average earnings of individuals in a state to be higher or lower if more or less percentage of people are educated respectively. There are certainly some omitted variables which would explain the correlation in a better way.

It answers the question that returns to education do not necessarily differ if an individual resides in a state with higher educational attainment. Other factors such as years of experience in fieldwork, more updated skillset, adaptability, years of experience etc. also affect when it comes to making decision for returns to education. According to Figure 3, state educational attainment does not directly affect wages.

**Table 3: Descriptive Statistics**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Number of Observations** | **Mean** | **Standard Deviation** | **Minimum** | **Maximum** |
| **State Educational Attainment (%) in 2019** | 48676 | 38.0149446 | 6.3670573 | 22.6839728 | 52.8590041 |
| **Years of Education** | 48872 | 16.2860329 | 2.0997536 | 0 | 18 |
| **Log (Hourly Wages)** | 48872 | 3.1036429 | 0.9023783 | -3.152736 | 9.8240234 |
| **Experience** | 48872 | 24.3759822 | 15.8515527 | 2 | 94 |
| **Experience\*2** | 48872 | 845.455087 | 906.7525892 | 4 | 8836 |

*Source*: IPUMS USA and U.S. Census Bureau

Table 3 shows the descriptive statistics for the important variables used in this paper.

Educational attainment across states in 2019 shows the number of individuals between age 25-44 with bachelor’s degree.

Years of Education = Calculated with the vague idea of how much an individual has studied.

Log (HourlyWages) is calculated first by calculating hourly wages and then taking the log of it.

Experience = Age – Years of education

Experiencesq = Experience^2

**Table 4: Regression Results**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Dependent Variable – Log(HourlyWages)** | | | | | |
| **Variables** | **Model 1** | **Model 2** | **Model 3** | **Model 4** | **Model 5** |
| **Intercept** | 1.42238\*\*\*  (0.00375) | 0.6706\*\*\*  (0.00415) | 0.24109\*\*\*  (0.00551) | 0.73082\*\*\*  (0.0064) | 1.16843\*\*\*  (0.02321) |
| **eduyr**  (Years of education) | 0.11013\*\*\*  (0.00025) | 0.11053\*\*\*  (0.000246) | 0.10778\*\*\*  (0.0002463) | 0.10294\*\*\*  (0.0002467) | 0.07159\*\*\*  (0.0015) |
| **Experience** | - | 0.04727\*\*\*  (0.00017) | 0.04766\*\*\*  (0.0001697) | 0.03963 \*\*\*  (0.0001771) | 0.03953\*\*\*  (0.000177) |
| **Experiencesq** | - | -0.00058\*\*\*  (2.76E-06) | -0.00059\*\*\*  (2.75E-06) | -0.00049\*\*\*  (2.81E-06) | -0.00049\*\*\*  (2.81E-06) |
| **edu2019**  (Educational Attainment % in 2019) | - | - | 0.01245\*\*\*  (0.000105) | 0.01256\*\*\*  (0.0001042) | 0.00003512  (0.0005978) |
| **educyr19**  (eduyr \* edu2019) | - | - | - | - | 0.00084367\*\*\*  (3.963E-05) |
| **MaritalStatus** | - | - | - | -0.20373\*\*\*  (0.00139) | -0.20365\*\*\*  (0.00139) |
| **eduind**  (7870 – Education Industry) | - | - | - | - | 0.02861\*\*\*  (0.00351) |
| **Observations** | 1450390 | 1450390 | 1446590 | 1446590 | 1446590 |
| **R-Square** | 0.1176 | 0.1911 | 0.1986 | 0.2104 | 0.2107 |

Standard errors in parentheses \*\*\* p < 0.01, \*\* p < 0.05, \* p < 0.1  
*Source*: IPUMS USA and U.S. Census Bureau

**Regression Model 1**

Equation: : *log (HourlyWages)* = 1.42238 + 0.11013*educyr* + *u*

Regression Model 1 equation shows the results from first regression model where log of hourly wage is the dependent variable and years of education is the independent variable. The regression model is shown in Table 4, which shows that wages would increase by 11.01% for every additional year of education. I do not believe that this is a good model, because there are a lot of important variables which are not considered while running this regression. The R-square of this model is 0.1176 which explains 11.76% of total variation in log of hourly wages, and the years of education variable is significant at 1% level according to their p-values.

**Regression Model 2**

Equation: *log (HourlyWages)* = 0.6706 + 0.11053*educyr* + 0.04727*experience* – 0.00058*experiencesq* + *u*

Regression Model 2 equation shows the results from second regression model where log of hourly wage is the dependent variable and years of education, experience, and experiencesq are the independent variables. The regression model is shown in Table 4, which shows that wages would increase by 11.05% for every additional year of education, 4.73% for every additional year of experience, and decrease by 0.058% for every experiencesq. Experiencesq seems to be having a negative relationship with wages. The R-square of this model is 0.1911 which explains 19.11% of total variation in log of hourly wages which shows that it is a better model than the previous one in with Model 1 equation, but it still has a lot of omitted variables which can have a huge importance in determining how exactly are wages affected considering multiple variables. The variables years of education, experience, and experiencesq are significant at 1% level according to their p-values.

**Regression Model 3**

Equation: *log (HourlyWages)* = 0.24109+ 0.10778*educyr* + 0.04766*experience* – 0.00059*experiencesq* + 0.01245*edu2019* + *u*

Regression Model 3 equation shows the results from third regression model where log of hourly wage is the dependent variable and years of education, experience, and experiencesq, edu2019 are the independent variables. The regression model is shown in Table 4, indicating that wages would increase by 10.78% for every additional year of education, 4.77% for every additional year of experience, decrease by 0.06% for every experiencesq, and increase by 1.25% for every additional number of people educated (between 25-44 with bachelor’s degree). The R-square of this model is 0.1986 which explains 19.86% of total variation in log of hourly wages which shows that it is a better model than the previous one with Model 1 and 2 equations, but educational attainment doesn’t seem to have much effect on wages because the R-square only increase from 0.1911 to 0.1958, and this model can have a lot of relevant variables determining how wages are affected. The variables years of education, experience, experiencesq, and educational attainment in 2019 are significant at 1% level according to their p-values.

State educational attainment has little to no influence on log (hourly wages) even after controlling for the person’s own educational attainment.

**Regression Model 4**

Equation: *log (HourlyWages)* = 0.73082+ 0.10294*educyr* + 0.03963*experience* -0.00049*experiencesq* + 0.01256*edu2019* – 0.20373*maritalstatus* + *u*

Regression Model 4 equation shows the results from the fourth regression model where log of hourly wage is the dependent variable and years of education, experience, experiencesq, edu2019, and marital status are the independent variables. I chose Marital status as one of the variables to see how being married or unmarried affects an individual’s wages. The regression model is shown in Table 4, which shows that wages would increase by 10.29% for every additional year of education, 3.96% for every additional year of experience, decrease by 0.05% for every experiencesq, increase by 1.26% for every additional number of people educated (between 25-44 with bachelor’s degree), and decrease by 20.4% for being married. The R-square of this model is 0.2104 which explains 21.04% of total variation in log of hourly wages which shows that it is a better model than the all the previously discussed models. Apparently, marital status plays a big role in determining wages because as seen in the gender wage gap project last semester, married women tend to be underpaid due to discrimination, fewer skill set, uncertainty of taking maternity leave, etc. It can be close to a good model to show the variation in wages. The variables years of education, experience, experiencesq, educational attainment in 2019, and marital status are significant at 1% level according to their p-values.

**Regression Model 5**

Equation: *log (HourlyWages)* = 1.16843+ 0.07159*educyr* + 0.03953*experience* – 0.00049*experiencesq* + 0.00003512*edu2019* + 0.00084367educyr19 – 0.20365*maritalstatus* + 0.02861eduind + *u*

To get the interaction effect between education and state educational attainment, I multiplied educyr with edu2019 and got a new variable named educyr19 which shows the interaction effect between variables in 2019.

To include an industry control, I chose to Education industry (7870). First, I converted the qualitative data into quantitative data indicating individuals in Education industry would be denoted with number 0 and all other individuals with number 1. The quantitative variable is named eduind to show the control of one particular industry.

Regression Model 5 equation shows the results from the fifth regression model where log of hourly wage is the dependent variable and years of education, experience, experiencesq, edu2019, educyr19, marital status, and eduind are the independent variables. I created variables educyr19 and eduind to put see how interaction effect and industry control changes the regression results respectively. The regression model is shown in Table 4, which shows that wages would increase by 7.15% for every additional year of education, 3.96% for every additional year of experience, decrease by 0.05% for every experiencesq, increase by 0.004% for every additional number of people educated (between 25-44 with bachelor’s degree), increase by 0.08% for every additional interaction between years of education & state educational attainment, decrease by 20.36% for being married, and increase by 2.86% for being in education industry controlling for all other industries. The R-square of this model is 0.2107 which explains 21.07% of total variation in log of hourly wages which shows that it is a better model than the all the previously discussed models but there is not much difference in the model even after adding educyr19 and eduind variable. All the variables are significant at 1% level according to their p-values except for educational attainment in 2019.

**MLR.1 (Linear in parameters)**

For the MLR.1 assumption to hold true, all the parameters in the equation must be linear. As per equation model 5, all the variables included are liner. Even though we have used Experiencesq as one of the variables, that won’t change the assumption because a squared variable can be included to produce U-shaped curve.

**MLR.2 (Random Sampling)**

All the data extracted from IPUMS USA, U.S. Census Bureau and Bureau of Economic Analysis is completely random. For the aspect of this project, individuals with bachelor’s degree between ages 25 and 44 are analyzed as % of educated people within a state. Data cleaning helped in narrowing down and getting rid of unnecessary variables.

**MLR.3 (No perfect collinearity)**

In the given sample, none of the independent variables are constant. Therefore, there are no exact linear relationships among the independent variables.

Even though we include Experiencesq which is a squared term for Experience, MLR.3 assumption is not violated because Experiencesq is not an exact linear function on Experience.

The regression to check tolerance and variance inflation shows that VIF for educyr19 (interaction term – educyr \* edu2019) has the highest variance inflation and one of the lowest tolerances which shows that there is certain multicollinearity in the regression model.

The variable causing more multicollinearity can be dropped from the model to reduce it but that would lead to omitted variable bias.

**MLR.4 (Zero conditional mean)**

Following conditions would make MLR.4 assumption biased:

1. Forgetting to include the quadratic term in the equation to estimate the model
2. If the dependent variable is not correctly used in the true model and regression analysis
3. Omitting an important factor that is correlated with any of the variables used in the regression

According to the five regressions showed in this paper, MLR.4 assumptions does not seem to be biased because it does not have either of the above conditions mentioned. There might be a slight possibility that some other variable can be added in the regression analysis which would seem to be strongly correlated but with the given data, there seems no bias. To hold the MLR.4 assumption true, Experience and Experiencesq both variables are supposed to be used or else it the model would be mis specified.

**MLR.5 (Homoskedasticity)**

To hold MLR.5 assumption true, the value of the explanatory variables must contain no information about the variance of the unobserved factors. As seen in Figure 1, Homoscedasticity is observed between educational attainment (%) in 2019 and average earnings per job (millions) in 2019. There are not many outliers which would impose bias on the graph.

Although to check for heteroskedasticity in the model 5 regression, we can run a regression with ‘acov’ at the end and that shows that parameters of the variables are not changed but standard errors and p-values are slightly changed to fix the heteroskedasticity in the regression model.

**MLR.6 (Normality)**

Normality can be achieved by transforming dependent variables sometimes, i.e., using log(wage) instead of wage. In all the regression models, log9HourlyWages) is used as the dependent variable which holds MLR.6 assumption true. If instead of log(HourlyWages), Hourly Wages were used then the model would have been mis specified.

If the variables included in a regression model are strongly correlated, sampling variance of the estimated effects will be large, and as the increases, the sampling variance increases. of all five models is seen in Table 4 and as more variables are added into the model, the increases which shows that the variables are strongly correlated with the dependent variable.

**Hypothesis Test**

For the hypothesis test, (null hypothesis) is that years of education have no significance on log(HourlyWages) and (alternate hypothesis) is that years of education have quite some significant effects on log(HourlyWages). It is tested using a 0.05 significance level and the test is run in SAS. After running the test, p-value is < 0.0001 which shows that the model is significant at 1% and the model rejects the null hypothesis in favor of alternate hypothesis test. This explains that years of education have a great significance on logged variable of hourly wages. In conclusion, it means that years of education and hourly wages are correlated.

After running five models of regressions, it seems like every time a new variable is added increases by some percentage, which shows that more variables would be better able to explain the model. Although, the change from Model 4 to Model 5 is very minimal which gives an idea that there is possibly no omitted variables bias shown in the analysis.

Works Cited

“Bureau of Economic Analysis.” *U.S. Department of Commerce*, 23 Feb. 2022, https://www.commerce.gov/bureaus-and-offices/bea.

Team, MPC UX/UI. “U.S. Census Data for Social, Economic, and Health Research.” *IPUMS USA*, https://usa.ipums.org/usa/.

“U.S. Census Bureau.” *USAGov*, https://www.usa.gov/federal-agencies/u-s-census-bureau.