Influence of Educational Attainment on Wages: A Regional Analysis of the United States of America

Abigail Hoffman

May 10, 2021

Abstract

This paper uses ordinary least squares and logistic regression techniques to analyze how educational attainment influences wages earned. The modeling of the data sets is divided into six regions to encompass the fifty states in the United States of America. The regions are as followed: Pacific (West), Rocky Mountain, Southwest, Midwest, Southeast and Northeast Regions. This regional pattern follows the same technique and division as regulated by the federal government agencies during this time frame studied. The years are from 2010-2019 and the data is obtained from the American Community Survey by the Integrated Public Use Microdata Series (IPUMS). The goal of the paper is to understand if some regions have higher payoffs for educational attainment compared to other regions.

1 Introduction

Education in the United States has served as a critical feature in history that remains to be an ever-present debate in national and state-wide politics. Early federal laws prevented students from attending schools based on sex, race and other socioeconomic factors. These outdated laws continue to have a lasting effect on present state economies. State economies and laws have caused a rise in the consumer price index (CPI) portion of education for American families nationwide. The Consumer Price Index (CPI) is a measure that examines the weighted average of prices of a basket of consumer goods and services, such as transportation, food, and medical care.It is calculated by taking price changes for each item in the predetermined basket of goods and averaging them. Changes in the CPI are used to assess price changes associated with the cost of living. American family's CPI has been increasing steadily over the past decade. (See Figure 1) (U.S. Bureau of Labor Statistics, 2021). With the cost of living rising for families in the United States, there is a greater preference on obtaining jobs in regions with higher wages. This means that despite paying more over time for education there is now a greater demand for attaining higher education with better returns on wages.

For this research, I am utilizing data that spans 2010 through 2019. During this recent decade, there have been three presidential elections which amounted to discordant topics over the importance of education by type and attainment. Proposed legislation from each dominant political party in the United States government is impacting not only primary and secondary education institutions but ,also, higher education institutions as well. Ideas like student loan forgiveness, tuition free colleges, school choice initiatives and charter schools have emerged in the discussion of education.

The six regions I am investigating are the Pacific (West), Rocky Mountain, Southwest, Midwest, Southeast and Northeast regions. I chose this regional division of the states as it is a common practice among federal government agencies to use these terms and state regions to serve better serve the public as well. Migration after higher educational attainment is a factor many students face when thinking about post-graduate plans. Significant plans for graduates are a daunting task as to where to accept job positions or where to look for their best returns for their education on their wages earned finishing their degree programs. The variables I have chosen to investigate are educational attainment, race, gender, age, age$^{\mathbf{2}}$, marital status and metropolitan area status. My goal with research into this area of labor economics is to determine if regionally university graduates will be able to see where their educational achievements that gain the greatest return on investment through wages earned post-graduation. I hypothesize that areas like the Pacific (West) and Northeast regions will have a greater return on educational attainment compared to the other four regions being investigated.

2 Literature Review

Labor economics seeks to understand the functioning dynamics of the markets for wage labor. Because labor is a commodity supplied by laborers in exchange for wages earned, this sector of economics exist as parts of the social, institutional, and political system within the examined economic market structure. A popular topic within this field is the examination of a direct relationship between educational attainment and wages earned in the market. This portion of economics is studied by labor economists and those specializing in education economics. There is a proven positive relationship that exists between schooling and wages which has led to my choice to add my own research into the literature realm (Dickson and Harmon, 2011).

Research within this area has continued as different elements of education like higher level of degrees have transitioned into a necessity for earning livable wages. As advanced levels of educational attainment increases the literature discussion has expanded to consider if higher education is a winners take all within labor markets (Elman and O'Rand, 2004). Research done by Elman considers two possibilities of earnings. Adult wages reflect either the human capital obtained through learning (measured by years of education in the traditional model) or the credentials preferred by employers (Elman and O'Rand, 2004). Other research elements utilize the natural logarithm of wages. Wages earned are positive, thus using the natural logarithm better generates easily understandable regression analysis for research purposes.

Issues of equity, fairness, and representation continue to be significant challenges in the market for public and private sector businesses. Educational attainment has been considered the great equalizer to anyone in a disadvantageous position resulting from previous discriminatory laws. The great equalizer argument indicates that labor markets for college-educated workers are meritocratic and thus the expansion of higher education is expected to promote greater inter-generational social mobility (Long, 2010). Factors like gender, race and other socio-economic factors are critical elements worth considering when determining education and wages. While researchers have made considerable progress over the last several decades, both in developing normative theories and improving the rigor of empirical analysis to guide practitioners and improve reform efforts, much work remains to be done (Rabovsky and Lee, 2018). Utilizing the findings of previous researchers in the field illuminates gaps that new research can help fill.

With education assisting as an equalizer of social status mobility, people are better able to self-select migration. This leads to a better selection on where to live after obtaining their educational degree(s) and build a career in their respective field as well. According to Dahl (2002), "if workers chose where to live and work based on comparative advantage, then the estimated returns to college in any given state could be biased upward or downward (Dahl, 2002)." With advancements in technology, U.S. workers are highly mobile meanwhile, labor markets vary widely despite this observation of migration opportunities. Dahl finds an upward bias in the returns to schooling. This is the result of individuals responding to above-average earnings shocks but states that other immeasurable non-wage variables would play pivotal roles in decisions of individuals with varying education levels (Dahl, 2002). From similar framework by Dahl in 2002, Ransom uses a Roy model analysis model of college major, occupational choice, and location choice with an expansion of the model to include monetary and non-monetary factors of influence for an individual's decision and more than two alternatives in the choice set analysis (Ransom, 2020). Ransom finds that According to Ransom, the results found "underscore the importance of appropriately measuring returns to human capital investment. It is well established that students choose majors in part because of earnings differences (Ransom, 2020)." This still leaves the elusive question is if students understand that their earning outcomes could differ geographically across the untied states. This helps motivate my research on location preferences based on education attainment.

3 Data

3.1 Source of data

I extracted my data from IPUMS-USA using sampling from the American Community Survey (ACS) results. This data is verified and authorized by the Minnesota Population Center with an affiliation to the University of Minnesota. The Public Use Micro-data Set from the American Community Survey uses data from the U.S. Census to easily identify critical variables and manipulate data for research purposes (Steven Ruggles, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler and Matthew Sobek, 2021). Data extracted focuses on the decade of 2010 through 2019 across the fifty United States and the District of Columbia. Each state was then organized into 6 data sets to represent 6 regions similar to how federal government agencies divide the states across the nation. The first region is the Pacific (West) with the following states: Washington, Oregon, California, Hawaii and Alaska. This data set has 677,420 observations. The Rocky Mountain region encompasses the following states: Montana, Idaho, Utah, Nevada, Wyoming and Colorado.This regional data set has 166,375 observations. The Southwest region has Arizona, New Mexico, Oklahoma and Texas with 409,548 observations. The Midwest region contains the following states: North Dakota, South Dakota, Nebraska, Kansas, Minnesota, Iowa, Missouri, Illinois, Indiana, Ohio, Michigan and Wisconsin. The Midwest region has 737,003 observations in the data set. The Southeast region contains the following states: Arkansas, Louisiana, Mississippi, Alabama, Georgia, Florida, South Carolina, North Carolina, Kentucky, Tennessee, West Virginia, and Virginia. Southeast region's data sets contains 857,030 observations. The last region is Northeast region containing the states as followed: Maine, New Hampshire, Maryland, New York, Pennsylvania, Rhode Island, Connecticut, New Jersey, Delaware, Vermont and Massachusetts. This region contains 969,553 data set observations. My audience is targeted towards those pursuing higher education and the potential to find areas within the United States that have a higher return on wages earned from educational attainment.

3.2 Description of Selected Variables

The explained variable of wage was mutated to natural logarithm of individual's wages, log(incwage). Converting wage to a natural logarithm allows researchers a greater ability to test and explain hypotheses when all numbers in the variable's data are positive like wages. Furthermore, the natural logarithm of wage given by the variable, logincwage, allows the independent variable to be interpreted as a percentage change in income wage earned as the respective dependent variable increases by one.

The primary explanatory variable being tested in my hypothesis is educational attainment in years which is represented by the variable coded as EDUCYRS. Additional dependent variables are age represented by AGE, a mutated variable AGE^2 to demonstrate labor market experience, gender of person in survey which coded as SEX, race of person either white or non-white is represented by RACENW, metropolitan status as METROSTATUS for either in the metropolitan area or not within the metropolitan limits, and MARTSD indicating whether the individual is married or single.

When interpreting the data found in the regressions, consider these unique coding scale.

- EDUCYRS:

00 = No education attainment,

01 = 1st grade/ 1 year of education,

02 = 2nd grade/2 years of education,

03 = 3rd grade/3 years of education,

04 = 4th grade/4 years of education,

05 = 5th grade/5years of education,

06 = 6th grade/6 years of education,

07 = 7th grade/7 years of education,

08 = 8th grade/8 years of education,

09 = 9th grade/9 years of education,

10 = 10th grade/10 years of education,

11 = 11th grade/11years of education,

12 = High School Diploma or GED acquired/12 years of education,

12.5 = less than one year of college/ approximately 12.5 years of education,

13 = 1 year of college/13 years of education,

14 = 2 years of college includes those who obtained Associates Degree/14 years of education,

15 = 3 years of college/ 15 years of education,

16 = 4 years of college and includes those with a Bachelor's Degree/ 16 years of education,

17 = 5 years of college/17 years of education,

18 = Master's Degree/ 6 years of college/ 18 years of education,

19 = 7 or more years of college/ 19 years of education,

20 = PhD achieved and further professional school that was 8 years of college/ 20 years of education

- SEX: male =0, female =1

- RACENW: white = 0, non-white (includes all races and ethnicities) =1
- MARST: Married = 0, Single = 1
- METROSTATUS: in metro = 0, out of metro = 1

3.3 Verifying Gauss-Markov Assumption

1. Linear in Parameters — The model consists of a multiple linear regression that is linear in parameters.
2. Random Sampling — The data includes approximately 7 million observations across the United States of America with a wide range of demographics. The sampling is to be assumed as random.
3. No perfect collinearity — There is no perfect collinearity between the variables in the model. No correlation between x variables. This was verified using the variance-inflation test in R studio.
4. Error term is independently distributed and not correlated — Gauss Markov Assumption 4 is likely to hold in this random sampling data. There are many observations along with the inclusion of relevant independent variables attests to the zero-conditional mean. X's are exogenous explanatory variables. If omitted variable bias arises in the data, this held assumption will fail.
5. Homoskedasticity — The model and variables constructed from the data are assumed to have constant variance of u's over x variables. This was verified using the ordinary least squares with robust standard errors test in R studio to correct heteroskedasticity.

4 Methods

Two modeling techniques were utilized for this research project: logistic regression with the utilization of ordinary least squares with robust standard errors which fits a linear model providing a variety of options for robust standard errors and conducts coefficient tests. This enabled me to run regression analysis to understand how educational attainment along with my other variables used impacted the logarithm of wages. The logarithm of wage was used for better modeling practices because wages take on only positive values. Utilizing these modeling methods allowed for each of the six data regional data sets to provide results and findings for the hypothesis presented.

4.1 Simple Regression

The equation for the first regression is a simple regression. I focus solely on the natural logarithm of wage influence by education attainment for each of the six regions.

Equation listed below.

$$y = \beta 0 + \beta 1 * x1 + u \tag{1}$$

$$log(incwage) = \beta 0 + \beta 1 * EDUCYRS + u \tag{2}$$

4.2 Multiple Regression Model with inclusion of Dummy Variables

This regression model focuses on the two numeric dependent variable's effects on natural logarithm of wage. The variable AGE^2 denotes labor market experience to better interpret whether the variable AGE has increasing or diminishing marginal returns to wage in the labor market. The regression, also, includes the addition of multiple dummy variables, sex and racenw. For the dummy variable SEX, women are coded as 1 while men are coded as 0. For the race variable RACENW, those identifying as white are coded as 0 and non-white (which includes all races and ethnicities) is coded as 1. The variable MARSTD determines marital status, where being married is coded as 0 and single is coded as 1. The final variable METROSTATUS determines the status of metropolitan living is coded as 0 and not in the metro is 1.

Equation listed below.

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \beta 3x3 + \beta 4x4 + \beta 5x5 + \beta 6x6 + \beta 7x7 + u \tag{3}$$

$$\begin{aligned} log(incwage) = \\ & \beta 0 + \beta 1 * EDUCYRS + \\ & \beta 2 * AGE + \beta 3 * AGE^2 + \\ & \beta 4 * SEX + \beta 5 * RACENW + \\ & \beta 6 * MARSTD + \beta 7 * METROSTATUS + u \end{aligned} \tag{4}$$

5 Findings

The regression models allowed for me to demonstrate a direct relationship between years of educational attainment effects on the natural logarithm of wages in each region during the time period 2010-2019 with the inclusion of additional numerical and dummy variables to help correct omitted variable bias in my research results. All variables in each of the 12 regressions ran were found statistically significant at the 95 percent confidence interval. Running the variance inflation factor (VIF) test on my multiple regression models demonstrated that there was not collinearity, strong correlation between two or more predictor variables, between my selected variables. Most variables returned data that was close to 1. The exception to this case was AGE and AGE^2 which was expected as they are utilizing the same data but for intended research purposes this was considered an acceptable correlation.

The most problematic factor of the regressions are the critically low R-square values. For the simple regressions the regressions R squared was 1 percent or below. For my multiple regression models, the R squared values returned below the 20 percent threshold. Despite having a large random sample size across a ten year time frame, this reveals a critical issue of omitted variable bias among exogenous variables in the regression. This hurts the validity of my assumption that Gauss-Markov assumption four would hold in my research findings.

5.1 Regional Regressions Discussion

Overall, a one percent increase in education in years results in at approximately a 6 percent increase in wages earned when additional variables were added in all six regions. The Pacific (West) resulted in the highest return at 7.3 percent and the Southwest demonstrated the lowest return of only 5.6 percent. Between all of the models, there was not one region that significantly gave a grander increase in wages to indicate mobility preference for increased educational attainment.

Each region discussed averaged about a 12 percent return on wages earned when age, serving as an experience variable, increased by one percent.AGE^2 was negative in all multiple regression models. This means returns to wage regarding age is increasing at a diminishing rate, approximately 0.1 percent, which is example of diminishing marginal returns in the models. As one's labor market experience increases, the person will earn higher wages at a diminishing marginal rate.

For women, the Pacific (West) and Northeast regions had the lowest gender wage gap compared to the other regions. Statistically, the worst regions for women to work were the Rocky Mountain region and the Southwest. Areas like the Pacific (West) and Northeast demonstrated a smaller gender wage gap of 75 cents on the male dollar compared to the calculated national average of 69 cents on the male dollar earned. Whereas, the areas of the Rocky Mountain region and Southwest region gave a more accurate depiction of the national gender wage gap at approximately 70 cents on the male dollar versus 69 cents on the male dollar. In the long-term this gender inequality gap will really hinder not only women's spending power in the economy but will be a detriment to the United States' gross domestic product (GDP).

For people of color, any race that is defined as non-white, the lowest racial gaps were found in the Pacific (West) and Northeast regions. At only a 2 to 3 percent gap, these would be the most advantageous regions for non-white workers. The worst regions to work for non-white workers were found in the Rocky Mountain and Southeast regions. Non-white workers experienced a little over 9 percent gap in earnings compared to their white counterparts.This means any person of color's purchasing power from wages earned is not as efficient as white counterparts.

6 Conclusion

My goal when first conducting this project was to test years of educational attainment on wages earned regionally across the United States for the most recent decade spanning from 2010-2019. I wanted to build upon the previous literature in understanding if there is value to relocating with the increased mobility of graduates and workers in specific regions across the United States. All of my models demonstrate that years of education are an important factor in wages earned.

Overall, my hypothesis was rejected. The region that did have the greatest return on education from wages was the Pacific (West) region, but none of the other five regions were significantly different from the Pacific (West). Overall, there is not one region that would significantly better wages from educational attainment. This leads to my suggestion that where to live is better chosen by personal preference and values than wages potentially earned. Despite the educational incentive to move not being a present factor, there are areas that offer less detrimental wage gaps for women and people of color. This could incentive minority groups that have been previously discriminated against based on gender and race to move to areas like the Pacific West and Northeast.

In this research project, it is worthy to note that Gender and Racial Bias are still likely to be prevalent, but due to omitted variable bias more variables need to be included to determine to what extent each region is suffering from wage bias for these variables. Continued empirical work into the Labor Economics field would benefit from additional research into issues of bias represented in my variables along with additional variables to deter omitted variable bias. This research topic could benefit from additional models like the Roy's model, Monte Carlo simulations and integration of fellow techniques from machine learning methods. Machine learning places a greater preference on $\hat{y}$ versus $\hat{beta}$ in econometric analysis which could better serve the investigation into this topic as complimentary sources from one another.

References

Dahl, Gordon B. 2002. "Mobility and the return to education: Testing a Roy model with multiple markets." *Econometrica* 70 (6):2367–2420.

Dickson, Matt and Colm Harmon. 2011. "Economic returns to education: What we know, what we don't know, and where we are going—some brief pointers." *Economics of education review* 30 (6):1118–1122.

Elman, Cheryl and Angela M O'Rand. 2004. "The race is to the swift: Socioeconomic origins, adult education, and wage attainment." *American Journal of Sociology* 110 (1):123–160.

Long, Mark C. 2010. "Changes in the returns to education and college quality." *Economics of Education review* 29 (3):338–347.

Rabovsky, Thomas and Hongseok Lee. 2018. "Exploring the antecedents of the gender pay gap in US higher education." *Public Administration Review* 78 (3):375–385.

Ransom, Tyler. 2020. "Selective migration, occupational choice, and the wage returns to college majors." .

Steven Ruggles, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler and Matthew Sobek. 2021. *IPUMS USA: Version 11.0 [dataset]*. University of Minnesota, Minneapolis, MN. URL https://www.ipums.org/.

U.S. Bureau of Labor Statistics. 2021. *U.S. Bureau of Labor Statistics, Consumer Price Index for All Urban Consumers: Education in U.S. City Average [CUSR0000SAE1]*. FRED, Federal Reserve Bank of St. Louis, St.Louis,MO. URL https://fred.stlouisfed.org/series/CUSR0000SAE1.

Figures and Tables

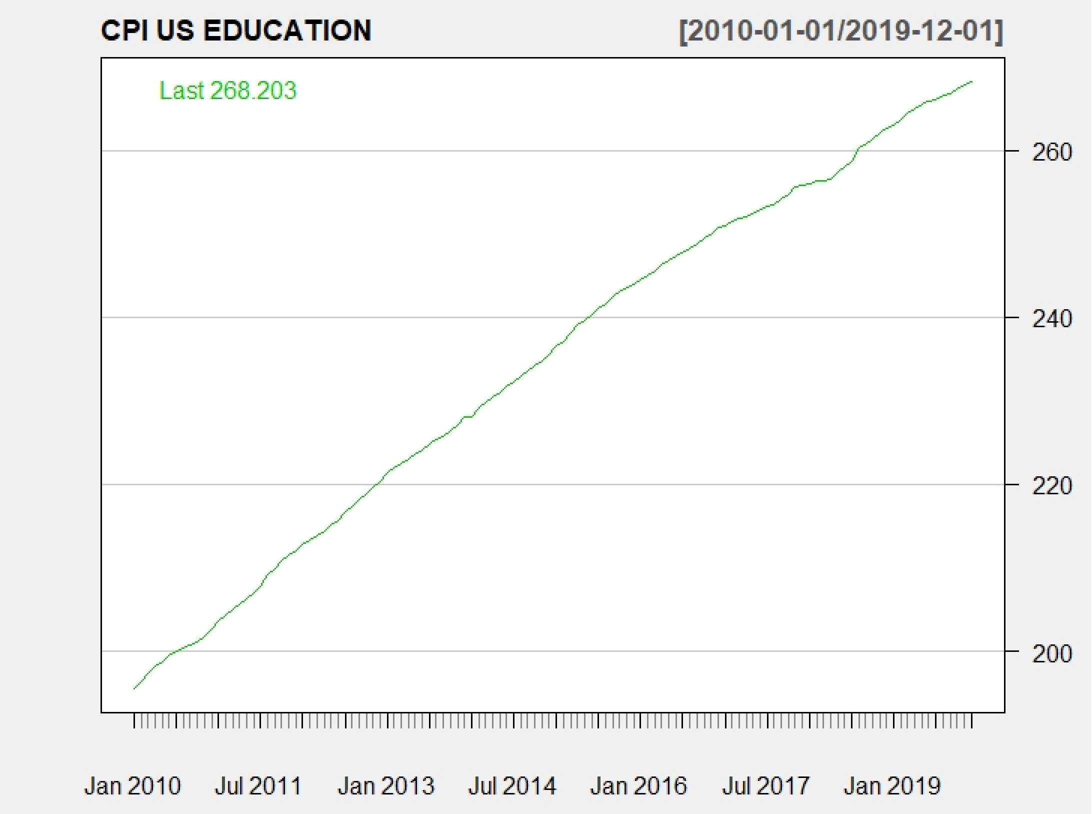

Figure 1: Average Consumer Price Index of US Education

Table 1: Pacific West Region

	Model 1	Model 2
(Intercept)	9.554	7.075
	(0.014)	(0.019)
EDUCYRS	0.085	0.073
	(0.001)	(0.001)
AGE		0.122
		(0.001)
AGE^2		-0.001
		(0.000)
SEX		-0.252
		(0.002)
RACENW		0.002
		(0.002)
MARSTD		-0.128
		(0.002)
METROSTATUS		0.339
		(0.006)
Num.Obs.	677420	677420
R2	0.015	0.156
R2 Adj.	0.015	0.156
se_type	HC2	HC2

Table 2: Rocky Mountain Region

	Model 1	Model 2
(Intercept)	9.620	6.993
	(0.030)	(0.037)
EDUCYRS	0.072	0.065
	(0.002)	(0.002)
AGE		0.134
		(0.001)
AGE^2		-0.001
		(0.000)
SEX		-0.291
		(0.004)
RACENW		-0.090
		(0.005)
MARSTD		-0.107
		(0.004)
METROSTATUS		0.208
		(0.006)
Num.Obs.	166375	166375
R2	0.010	0.179
R2 Adj.	0.010	0.179
se_type	HC2	HC2

Table 3: Southwest Region

	Model 1	Model 2
(Intercept)	9.877	7.458
	(0.019)	(0.024)
EDUCYRS	0.060	0.056
	(0.001)	(0.001)
AGE		0.120
		(0.001)
AGE^2		-0.001
		(0.000)
SEX		-0.299
		(0.002)
RACENW		-0.063
		(0.003)
MARSTD		-0.122
		(0.003)
METROSTATUS		0.248
		(0.005)
Num.Obs.	409548	409548
R2	0.007	0.167
R2 Adj.	0.007	0.167
se_type	HC2	HC2

Table 4: Midwest Region

	Model 1	Model 2
(Intercept)	9.727	7.007
	(0.014)	(0.018)
EDUCYRS	0.065	0.062
	(0.001)	(0.001)
AGE		0.136
		(0.001)
AGE^2		-0.002
		(0.000)
SEX		-0.277
		(0.002)
RACENW		-0.061
		(0.002)
MARSTD		-0.115
		(0.002)
METROSTATUS		0.257
		(0.002)
Num.Obs.	737003	737003
R2	0.008	0.199
R2 Adj.	0.008	0.199
se_type	HC2	HC2

Table 5: Southeast Region

	Model 1	Model 2
(Intercept)	9.695	7.198
	(0.013)	(0.016)
EDUCYRS	0.067	0.067
	(0.001)	(0.001)
AGE		0.122
		(0.001)
AGE^2		-0.001
		(0.000)
SEX		-0.284
		(0.002)
RACENW		-0.094
		(0.002)
MARSTD		-0.116
		(0.002)
METROSTATUS		0.228
		(0.003)
Num.Obs.	857030	857030
R2	0.010	0.171
R2 Adj.	0.010	0.171
se_type	HC2	HC2

Table 6: Northeast Region

	Model 1	Model 2
(Intercept)	9.693	7.101
	(0.012)	(0.015)
EDUCYRS	0.075	0.069
	(0.001)	(0.001)
AGE		0.125
		(0.001)
AGE^2		-0.001
		(0.000)
SEX		-0.252
		(0.002)
RACENW		-0.029
		(0.002)
MARSTD		-0.112
		(0.002)
METROSTATUS		0.333
		(0.004)
Num.Obs.	969553	969553
R2	0.012	0.168
R2 Adj.	0.012	0.168
se_type	HC2	HC2

Table 7: Side-by-Side Simple Regression Comparison of Regions

	Pacific (West)	Rocky Mountain	Southwest	Midwest	Southeast	Northeast
(Intercept)	9.554	9.620	9.877	9.727	9.695	9.693
	(0.014)	(0.030)	(0.019)	(0.014)	(0.013)	(0.012)
EDUCYRS	0.085	0.072	0.060	0.065	0.067	0.075
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
Num.Obs.	677420	166375	409548	737003	857030	969553
R2	0.015	0.010	0.007	0.008	0.010	0.012
R2 Adj.	0.015	0.010	0.007	0.008	0.010	0.012
se_type	HC2	HC2	HC2	HC2	HC2	HC2

Table 8: Side-by-Side Multiple Regression Comparison of Regions

	Pacific (West)	Rocky Mountain	Southwest	Midwest	Southeast	Northeast
(Intercept)	7.075	6.993	7.458	7.007	7.198	7.101
	(0.019)	(0.037)	(0.024)	(0.018)	(0.016)	(0.015)
EDUCYRS	0.073	0.065	0.056	0.062	0.067	0.069
	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)
AGE	0.122	0.134	0.120	0.136	0.122	0.125
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
AGE^2	-0.001	-0.001	-0.001	-0.002	-0.001	-0.001
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
SEX	-0.252	-0.291	-0.299	-0.277	-0.284	-0.252
	(0.002)	(0.004)	(0.002)	(0.002)	(0.002)	(0.002)
RACENW	0.002	-0.090	-0.063	-0.061	-0.094	-0.029
	(0.002)	(0.005)	(0.003)	(0.002)	(0.002)	(0.002)
MARSTD	-0.128	-0.107	-0.122	-0.115	-0.116	-0.112
	(0.002)	(0.004)	(0.003)	(0.002)	(0.002)	(0.002)
METROSTATUS	0.339	0.208	0.248	0.257	0.228	0.333
	(0.006)	(0.006)	(0.005)	(0.002)	(0.003)	(0.004)
Num.Obs.	677420	166375	409548	737003	857030	969553
R2	0.156	0.179	0.167	0.199	0.171	0.168
R2 Adj.	0.156	0.179	0.167	0.199	0.171	0.168
se_type	HC2	HC2	HC2	HC2	HC2	HC2