Project

## Shawn Huynh

## u0906730

# I. Introduction

## In this study, our primary objective is to see what would happen to an individual's wage if they were given an additional year of education. This is first observed through the use of graphs, where education is our dependent varialbe, and wage is our dependent variable. We also use a regression model to find out what is the return (gain in wage) to education, and is this information enough for us to conclude that there is some sort of correlation here. We also measure other characteristics in the regression model such as gender, whether they are male or female, race, in terms of being White, Black, or Asian, and respective work experience prior. What we are trying to figure out here is if are any differences in the categories of these characteristics, gender, race, experience. The last thing we try to find is if we encounter a homoskedastic problem in our data. At the time the data set was taken, there were actually many more observations technically, but they were not reasonably observational given that certain observations did not fit the criteria for being elligible for education, (i.e. those who were too young). In this data set, we took 4,344 reasonable observations with 12 variables. In the first table that displays the descriptive statistics of our data set, we get a glance of the variables that were measured, as well as their definitions.

# II. Conceptual Framework

## Our mindset behind choosing the variables that we wanted to look at was that we wanted to see if there was any difference for a white male vs. a white female, or a white male vs. a black male. We also wanted to look and see if gender had any sort of impact on our results.

# III. Data

library(readr)  
cps <- read\_csv("~/Desktop/Econometrics/cps.csv", col\_types = cols(X1 = col\_skip()))

## Warning: Missing column names filled in: 'X1' [1]

library(stargazer); stargazer(as.data.frame(cps), title="Descriptive Statistics", type = "text")

##   
## Please cite as:

## Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2. http://CRAN.R-project.org/package=stargazer

##   
## Descriptive Statistics  
## =====================================================  
## Statistic N Mean St. Dev. Min Max   
## -----------------------------------------------------  
## wage 4,344 18.734 10.545 1.000 99.000   
## income 4,344 45,439.900 43,168.320 0 1,099,999  
## age 4,344 42.077 13.039 17 85   
## educ 4,344 13.143 2.312 0.000 21.000   
## male 4,344 0.548 0.498 0 1   
## female 4,344 0.452 0.498 0 1   
## white 4,344 0.808 0.394 0 1   
## black 4,344 0.114 0.318 0 1   
## asian 4,344 0.038 0.192 0 1   
## others 4,344 0.040 0.196 0 1   
## union 4,344 0.125 0.331 0 1   
## exper 4,344 22.934 13.213 0.000 67.000   
## -----------------------------------------------------

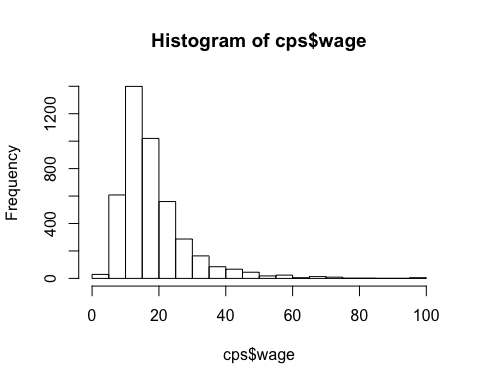
## wage: hourly wage in dollars in 2016, income: wage and salary

## income in 2016,age: age in yearseduc: educational attainment in ##yearsfemale: gender, coded 1 for female, 0 for malemale: gender, ##coded 1 for male, 0 for femalewhite: race, coded 1 for white ##workers, 0 for nonwhite workersblack: race, coded 1 for black ##workers, 0 for nonblack workersasian: race, coded 1 for asian ##workers, 0 for nonasian workersothers: ##race,coded1forotherthanwhite,black,andasianworkers,0forwhite,black,andasianworkers

## union: union status, coded 1 if workers are a member of a labor union, 0 otherwise

## exper: potential work experience (in years), defined as age minus years of schooling minus 6. (It is assumed that schooling starts at age 6).

hist(cps$wage)



plot(cps$educ, cps$wage, main = "Education vs. Wage",xlab = "Education", ylab = "Wage")

 ## The reason we measure wage for both graphs is because it is easier to look at the values for wage. If we used income, the values would not be as easy to look at.

correlation <- cor(cps)  
library(stargazer); stargazer(correlation, title="Correlation Matrix", type="text")

##   
## Correlation Matrix  
## ==========================================================================================  
## wage income age educ male female white black asian others union exper   
## ------------------------------------------------------------------------------------------  
## wage 1 0.339 0.195 0.333 0.117 -0.117 0.067 -0.075 0.024 -0.037 0.158 0.134   
## income 0.339 1 0.130 0.238 0.117 -0.117 0.057 -0.054 0.010 -0.036 0.089 0.087   
## age 0.195 0.130 1 0.013 -0.041 0.041 0.023 -0.001 0.010 -0.055 0.088 0.985   
## educ 0.333 0.238 0.013 1 -0.106 0.106 -0.016 0.005 0.059 -0.033 0.055 -0.162  
## male 0.117 0.117 -0.041 -0.106 1 -1 0.077 -0.084 -0.011 -0.007 0.088 -0.022  
## female -0.117 -0.117 0.041 0.106 -1 1 -0.077 0.084 0.011 0.007 -0.088 0.022   
## white 0.067 0.057 0.023 -0.016 0.077 -0.077 1 -0.735 -0.410 -0.417 -0.030 0.025   
## black -0.075 -0.054 -0.001 0.005 -0.084 0.084 -0.735 1 -0.072 -0.073 0.032 -0.002  
## asian 0.024 0.010 0.010 0.059 -0.011 0.011 -0.410 -0.072 1 -0.041 0.011 -0.001  
## others -0.037 -0.036 -0.055 -0.033 -0.007 0.007 -0.417 -0.073 -0.041 1 -0.003 -0.048  
## union 0.158 0.089 0.088 0.055 0.088 -0.088 -0.030 0.032 0.011 -0.003 1 0.077   
## exper 0.134 0.087 0.985 -0.162 -0.022 0.022 0.025 -0.002 -0.001 -0.048 0.077 1   
## ------------------------------------------------------------------------------------------

## Here we ran a correlation matrix to look at the kinds of independent variables that can be used when we build our model.

model1<-lm(wage~educ, data=cps)  
model2<-lm(wage~educ+male+white+exper, data=cps)  
model3<-lm(wage~educ+female+white+exper, data=cps)  
model4<-lm(wage~educ+male+black+exper, data=cps)  
model5<-lm(wage~educ+female+asian+exper, data=cps)  
stargazer(model1, model2, model3, model4, model5, title="Regression Results", type="text", out="reg.txt")

##   
## Regression Results  
## =====================================================================================================================================================  
## Dependent variable:   
## ---------------------------------------------------------------------------------------------------------------------------------  
## wage   
## (1) (2) (3) (4) (5)   
## -----------------------------------------------------------------------------------------------------------------------------------------------------  
## educ 1.521\*\*\* 1.747\*\*\* 1.747\*\*\* 1.745\*\*\* 1.745\*\*\*   
## (0.065) (0.064) (0.064) (0.064) (0.064)   
##   
## male 3.336\*\*\* 3.314\*\*\*   
## (0.295) (0.295)   
##   
## female -3.336\*\*\* -3.428\*\*\*   
## (0.295) (0.294)   
##   
## white 1.498\*\*\* 1.498\*\*\*   
## (0.370) (0.370)   
##   
## black -2.094\*\*\*   
## (0.458)   
##   
## asian 0.205   
## (0.758)   
##   
## exper 0.158\*\*\* 0.158\*\*\* 0.159\*\*\* 0.160\*\*\*   
## (0.011) (0.011) (0.011) (0.011)   
##   
## Constant -1.251 -10.904\*\*\* -7.568\*\*\* -9.433\*\*\* -6.325\*\*\*   
## (0.871) (1.003) (0.974) (0.964) (0.928)   
##   
## -----------------------------------------------------------------------------------------------------------------------------------------------------  
## Observations 4,344 4,344 4,344 4,344 4,344   
## R2 0.111 0.177 0.177 0.177 0.174   
## Adjusted R2 0.111 0.176 0.176 0.177 0.173   
## Residual Std. Error 9.942 (df = 4342) 9.573 (df = 4339) 9.573 (df = 4339) 9.568 (df = 4339) 9.591 (df = 4339)   
## F Statistic 543.152\*\*\* (df = 1; 4342) 232.692\*\*\* (df = 4; 4339) 232.692\*\*\* (df = 4; 4339) 234.041\*\*\* (df = 4; 4339) 227.744\*\*\* (df = 4; 4339)  
## =====================================================================================================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## Here we have our findings of our regression models. We use a total of 5 regression models, each with different variables added. The first thing we point to in our model is being a male. For white males, wage will increase by 3.336. If you are a black male, your wage will increase by 3.314. An interesting thing to point out is for the variables of female and black. First off, both of these variables when added into the model tell us that if you are black or female, your wage would decrease based on these variables. For white females, it is -3.336. For Asian females, it is -3.428. And for blacks, regardless of gender, your wage will decrease by -2.094.

library(sandwich)

## Warning: package 'sandwich' was built under R version 3.4.1

cov1<-vcovHC(model1, type="HC3"); cov2<-vcovHC(model2, type="HC3"); cov3<-vcovHC(model3, type="HC3"); cov4<-vcovHC(model4, type="HC3"); cov5<-vcovHC(model5, type="HC3")  
robust1<-sqrt(diag(cov1)); robust2<-sqrt(diag(cov2)); robust3<-sqrt(diag(cov3)); robust4<-sqrt(diag(cov4)); robust5<-sqrt(diag(cov5))  
library(stargazer)  
stargazer(model1, model2, model3, model4, model5, se=list(robust1, robust2, robust3, robust4, robust5), title="Regression Results", type= "text")

##   
## Regression Results  
## =====================================================================================================================================================  
## Dependent variable:   
## ---------------------------------------------------------------------------------------------------------------------------------  
## wage   
## (1) (2) (3) (4) (5)   
## -----------------------------------------------------------------------------------------------------------------------------------------------------  
## educ 1.521\*\*\* 1.747\*\*\* 1.747\*\*\* 1.745\*\*\* 1.745\*\*\*   
## (0.079) (0.081) (0.081) (0.081) (0.081)   
##   
## male 3.336\*\*\* 3.314\*\*\*   
## (0.293) (0.293)   
##   
## female -3.336\*\*\* -3.428\*\*\*   
## (0.293) (0.293)   
##   
## white 1.498\*\*\* 1.498\*\*\*   
## (0.336) (0.336)   
##   
## black -2.094\*\*\*   
## (0.395)   
##   
## asian 0.205   
## (0.786)   
##   
## exper 0.158\*\*\* 0.158\*\*\* 0.159\*\*\* 0.160\*\*\*   
## (0.012) (0.012) (0.012) (0.012)   
##   
## Constant -1.251 -10.904\*\*\* -7.568\*\*\* -9.433\*\*\* -6.325\*\*\*   
## (0.984) (1.187) (1.134) (1.139) (1.085)   
##   
## -----------------------------------------------------------------------------------------------------------------------------------------------------  
## Observations 4,344 4,344 4,344 4,344 4,344   
## R2 0.111 0.177 0.177 0.177 0.174   
## Adjusted R2 0.111 0.176 0.176 0.177 0.173   
## Residual Std. Error 9.942 (df = 4342) 9.573 (df = 4339) 9.573 (df = 4339) 9.568 (df = 4339) 9.591 (df = 4339)   
## F Statistic 543.152\*\*\* (df = 1; 4342) 232.692\*\*\* (df = 4; 4339) 232.692\*\*\* (df = 4; 4339) 234.041\*\*\* (df = 4; 4339) 227.744\*\*\* (df = 4; 4339)  
## =====================================================================================================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## This is the same data set with robust standard errors included. We see a difference in the value of the standard error for some of the variables in comparison to the 1st table.

res1<-lm(wage~educ+male+white+union+exper, data=cps)  
library(lmtest); library(sandwich)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

bptest(res1)

##   
## studentized Breusch-Pagan test  
##   
## data: res1  
## BP = 157.6, df = 5, p-value < 2.2e-16

res1<-lm(wage~educ+male+white+union+exper, data=cps)  
library(car);library(sandwich)#load required packages

## Warning: package 'car' was built under R version 3.4.3

linearHypothesis(res1,c("union=0","exper=0"),vcov=vcovHC(res1,type="HC3"))

## Linear hypothesis test  
##   
## Hypothesis:  
## union = 0  
## exper = 0  
##   
## Model 1: restricted model  
## Model 2: wage ~ educ + male + white + union + exper  
##   
## Note: Coefficient covariance matrix supplied.  
##   
## Res.Df Df F Pr(>F)   
## 1 4340   
## 2 4338 2 125.38 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

res1<-lm(wage~educ+male+white+union+exper, data=cps)  
library(car);library(sandwich)#load required packages  
linearHypothesis(res1,c("union=0","exper=0"),vcov=vcovHC(res1,type="HC3"))

## Linear hypothesis test  
##   
## Hypothesis:  
## union = 0  
## exper = 0  
##   
## Model 1: restricted model  
## Model 2: wage ~ educ + male + white + union + exper  
##   
## Note: Coefficient covariance matrix supplied.  
##   
## Res.Df Df F Pr(>F)   
## 1 4340   
## 2 4338 2 125.38 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

robust1<-coeftest(res1, vcov=vcovHC(res1, type="HC3")); robust1

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -10.585473 1.189879 -8.8963 < 2.2e-16 \*\*\*  
## educ 1.706448 0.081811 20.8584 < 2.2e-16 \*\*\*  
## male 3.093999 0.295503 10.4703 < 2.2e-16 \*\*\*  
## white 1.615939 0.338437 4.7747 1.859e-06 \*\*\*  
## union 3.577866 0.494397 7.2368 5.405e-13 \*\*\*  
## exper 0.150061 0.011569 12.9708 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Here, we ran a BP(Breusch-Pagan) test in order to find any sort of heteroskedasticity. We also ran a heteroskedastic-robust F-test to test joint hypothesis that neither of the variables "union" and "exper" had any effect at a 5% significance level. What we first found from the BP test was the F-statistic at 2125.38, at a p-value of 2.2e^-16. And the heteroskedastic-robust F-test was we also see the same p-value of 2.2e^-16. This means that we can reject the null hypothesis for the BP test, where the error variance is constant, in other words, there is no homoskedasticity because the p-value is very low. We can also reject the null hypothesis that both variables "union" and "exper" are equal to 0 at a 5% significance level. The reason we test a joint hypothesis and BP test is to see if there are any sort of errors in our data set that would skew results in one way or the other.

# Results

## There values of "union" and "exper" are not statistically significant at a 5% significance level. There is a observational difference between white males and all other variables. Blacks and females are also disadvantaged statistically.

# Conclusion

## It is evident based on our regression model that an additional year of education alone can have an affect on wage where we will see an increase in wage by 1.747. We also notice from our regression model that when you account for other variables or characteristics such as gender, and even race, can decrease your wage based on the factors.