Analyzing Mid-Atlantic Unemployment

DCI 202 Final

Jenna Marvet

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ISLR)
library(dplyr)
library(ggplot2)
```

Introduction

The data we are using is a collection of census survey response data from 3000 men in the Mid Atlantic region of the United States. Our dataset is included in the ISLR package and was manually assembled by Steve Miller of Open BI from the March 2011 Supplement to Current Population Survey data from the United States Census Bureau. This dataset has a total of 3000 observations and 11 variables.

In our research, we are utilizing the variables of wage, education, race and health. Wage is a factor with the person's raw wage in thousands. Race is a factor with levels 1. White 2. Black 3. Asian and 4. Other indicating race. Education is a factor with levels 1. < HS Grad 2. HS Grad 3. Some College 4. College Grad and 5. Advanced Degree indicating education level. Health is a factor with levels 1. <=Good and 2. >=Very Good indicating health level of worker.

Our research explores the question of whether or not and to what extent education level has an effect on wage rate for men in the Mid Atlantic region of the United States. We hypothesized that a higher level of education leads to higher wages. Therefore, our null hypothesis is that level of education has no effect on wage rate.

Methods

We found that the median wage across all levels of education is 104.92 and the mean was 111.70. Also, the minimum wage is 20.09 while the maximum wage is 318.34 in the sample. Additionally, the InterQuartile Range is 43.3. According to the IQR, there is high variability

within the data for wages which means the data on either side of the mean differs greatly. This means that wages vary greatly within the pool of wages.

Furthermore, we calculated the standard deviation for each of the levels of education. This is to show how much a wage deviated from the mean wage of each education level. For those with less than a high school degree the standard deviation is 21.57805, for a high school degree it is 28.57805, for some college it is 32.47473, for a college degree it is 41.18907 and for an Advanced degree it is 53.90421. We observed that, as the level of education increased, the standard deviation increased as well. Within the same degree group, people wages with more prestigious degrees deviated more from the mean than people with lower levels of education. This could mean that people with more advanced degrees have more job options. Therefore, people with more education can accept or deny different types of jobs with varying wage levels while people with less educational experiences do not have opportunities with much variety.

Standard deviation of each group of education level

```
Less_than_HS <- Wage[Wage$education == "1. < HS Grad",]
sd(Less_than_HS$wage)
```

```
## [1] 21.57805
```

```
High_School <- Wage[Wage$education == "2. HS Grad",]
sd(High_School$wage)
```

```
## [1] 28.56756
```

```
Some_College <- Wage[Wage$education == "3. Some College",]
sd(Some_College$wage)
```

```
## [1] 32.47473
```

```
College_Grad <- Wage[Wage$education == "4. College Grad",]
sd(College_Grad$wage)
```

```
## [1] 41.18907
```

```
Advanced_Degree <- Wage[Wage$education == "5. Advanced Degree",]
sd(Advanced_Degree$wage)
```

```
## [1] 53.90421
```

Mean and Median of each group of education level

```
mean(Less_than_HS$wage)
```

```
## [1] 84.10441
```

```
mean(High_School$wage)
```

```
## [1] 95.78335
```

```
mean(Some_College$wage)
```

```
## [1] 107.7556
```

```
mean(College_Grad$wage)
```

```
## [1] 124.4279
```

```
mean(Advanced_Degree$wage)
```

```
## [1] 150.9178
```

```
median(Less_than_HS$wage)
```

```
## [1] 81.28325
```

```
median(High_School$wage)
```

```
## [1] 94.07271
```

```
median(Some_College$wage)
```

```
## [1] 104.9215
```

```
median(College_Grad$wage)
```

```
## [1] 118.8844
```

```
median(Advanced_Degree$wage)
```

```
## [1] 141.7752
```

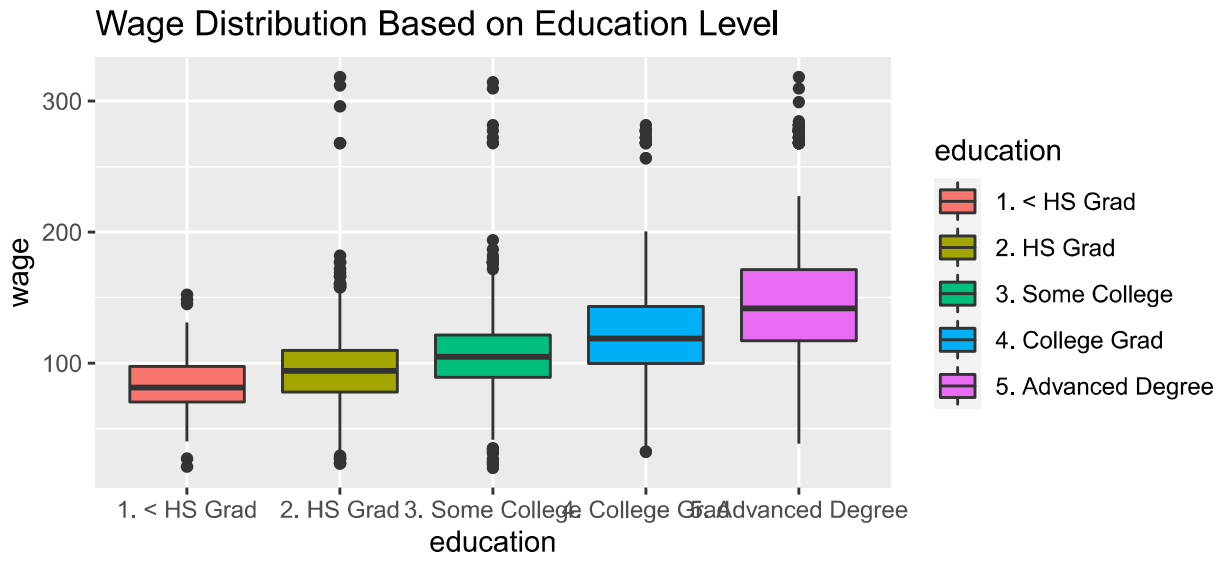
We observed that the median and mean wages increase as the level of educational experience increases. Median and mean are both measures of central tendency. Measures of central tendency help give an idea of the average or center of the data set. Considering both median and mode wage increase as education level increases, we constructed our alternative hypothesis around this trend.

Distribution plot of wage

```
qplot(wage, data=Wage, geom="density", alpha=I(.5), main="Distribution of Wage", xlab= "Wage Value", yl
```

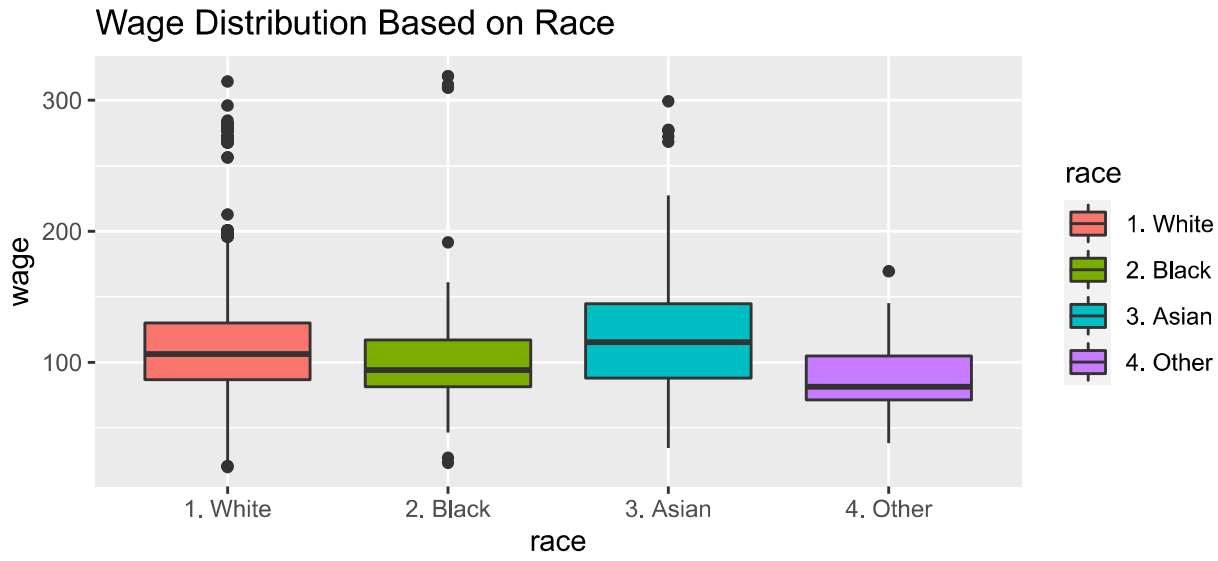
Box plot of wage and education

```
qplot(x=education, y=wage, data= Wage, geom=c("boxplot"), fill=education) +
ggtitle("Wage Distribution Based on Education Level")
```



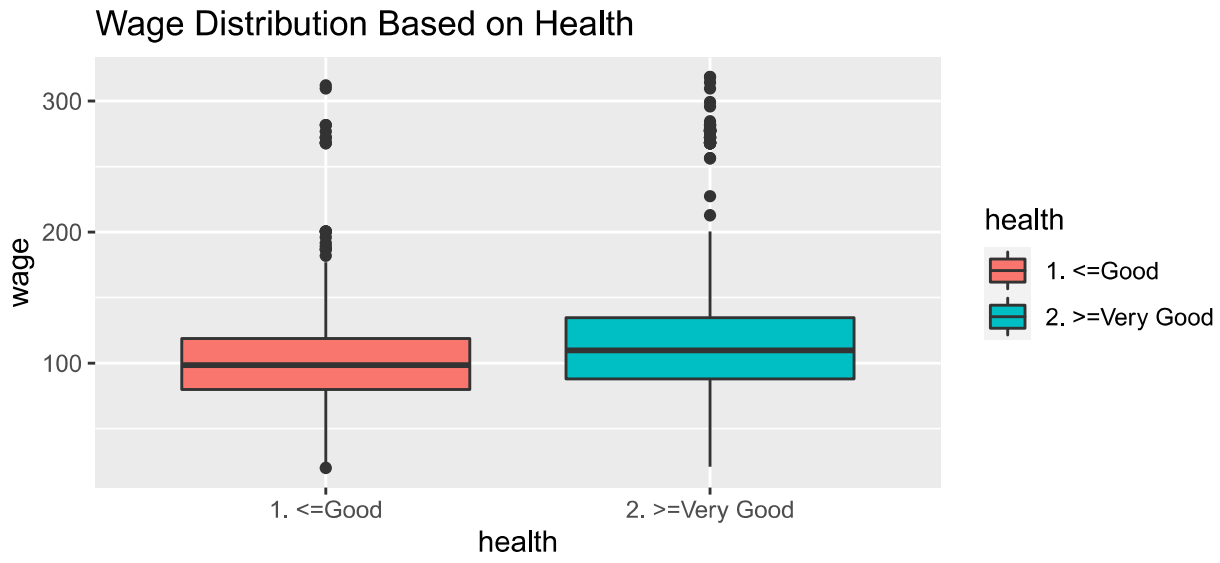
Box plot of wage and race

```
qplot(x=race, y=wage, data= Wage, geom=c("boxplot"), fill=race ) +
ggtitle("Wage Distribution Based on Race")
```



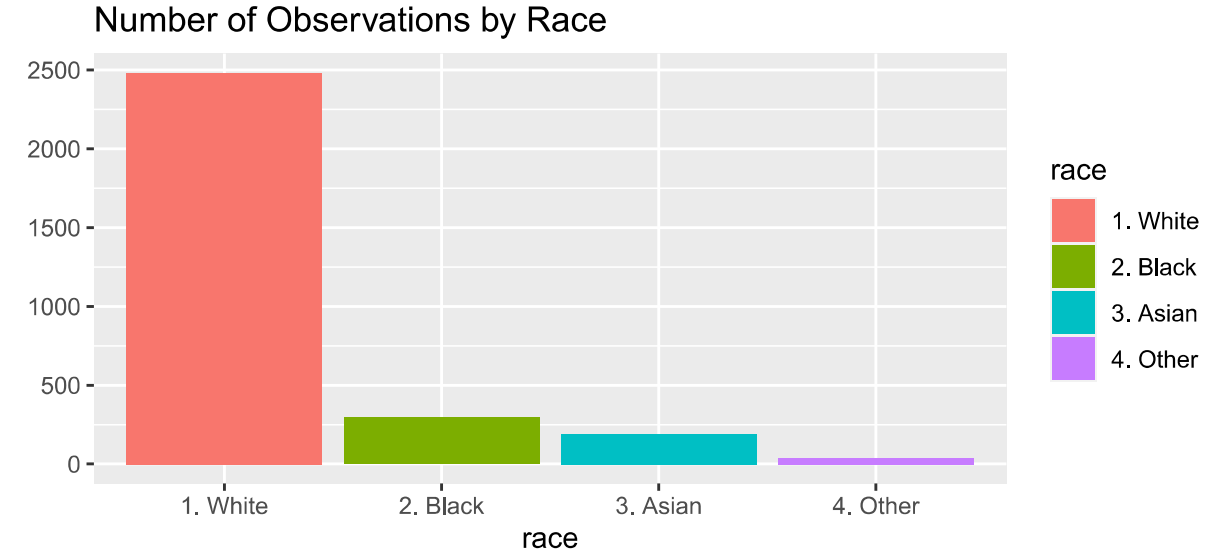
Box plot of wage and health

```
qplot(x=health, y=wage, data= Wage, geom=c("boxplot"), fill=health) +
ggtitle("Wage Distribution Based on Health")
```



Bar graph of number of observations by race

```
qplot(x=race,  data=Wage, geom="bar", fill=race) +
ggtitle("Number of Observations by Race")
```



Results

Bivariate Regression model of wage and education

```
Wage_Ed_Model <- lm(wage~education, data = Wage)
summary(Wage_Ed_Model)
```

```
##
## Call:
## lm(formula = wage ~ education, data = Wage)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -112.31  -19.94   -3.09   15.33  222.56
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    84.104      2.231  37.695  < 2e-16 ***
## education2. HS Grad            11.679      2.520   4.634 3.74e-06 ***
## education3. Some College       23.651      2.652   8.920  < 2e-16 ***
## education4. College Grad       40.323      2.632  15.322  < 2e-16 ***
## education5. Advanced Degree    66.813      2.848  23.462  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.53 on 2995 degrees of freedom
## Multiple R-squared:  0.2348, Adjusted R-squared:  0.2338
## F-statistic: 229.8 on 4 and 2995 DF,  p-value: < 2.2e-16
```

This regression uses a linear model to estimate the relationship between education level and wage. The coefficients for education level are being contrasted with any education less than a high school education.

The coefficient for high school diploma=11.679, some college=23.651, college diploma=40.323, advanced degree=66.813. These results are significant due to the associated p-values. The results of the model show that wage earnings increase significantly with each educational attainment. On average, people with a high school degree make $11,679 more than those without a high school degree. A college degree increases that difference to $40,323 and an advanced degree to $66,813. The most significant wage earning difference is between a college and advanced degree, with a difference of over $26,000.

Based on the results of this model, we can reject our null hypothesis and conclude that, in this population, education does affect wage earnings.

Multiple regression model of wage, education and race

```
Wage_Ed_Race_Model <- lm(wage~education + race, data = Wage)
summary(Wage_Ed_Race_Model)
```

```
##
## Call:
## lm(formula = wage ~ education + race, data = Wage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -113.309  -20.424   -3.502   14.975  227.966
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    85.535      2.265  37.762  < 2e-16 ***
## education2. HS Grad            11.219      2.523   4.447 9.04e-06 ***
## education3. Some College       23.375      2.655   8.805  < 2e-16 ***
## education4. College Grad       39.705      2.642  15.026  < 2e-16 ***
## education5. Advanced Degree    66.380      2.864  23.180  < 2e-16 ***
## race2. Black                   -6.378      2.264  -2.817  0.00488 **
```

```
## race3. Asian                  -4.046      2.782  -1.455  0.14587
## race4. Other                 -11.365      6.064  -1.874  0.06101 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.47 on 2992 degrees of freedom
## Multiple R-squared:  0.238,  Adjusted R-squared:  0.2362
## F-statistic: 133.5 on 7 and 2992 DF,  p-value: < 2.2e-16
```

To further investigate the relationship between wage and education, this model includes the variable of race. The base value being compared to is that for white people. Black has a coefficient of -6.378 and Asian a coefficient of -4.046. Holding all variables equal including educational attainment, black people earn $6,378 less than whites. Asian people earn $4,046 less than whites. Therefore, wage disparity is not only correlated with education. Race has a significant effect on wages on average, even when holding education levels constant.

Multiple regression model of wage, education and health

```
Wage_Ed_Health_Model <- lm(wage~education + health, data = Wage)
summary(Wage_Ed_Health_Model)
```

```
##
## Call:
## lm(formula = wage ~ education + health, data = Wage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -113.404  -20.914   -3.622   15.940  220.416
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   80.136      2.396  33.439  < 2e-16 ***
## education2. HS Grad           11.383      2.513   4.529 6.16e-06 ***
## education3. Some College      22.885      2.649   8.639  < 2e-16 ***
## education4. College Grad      39.052      2.639  14.798  < 2e-16 ***
## education5. Advanced Degree   65.227      2.861  22.797  < 2e-16 ***
## health2. >=Very Good           6.648      1.494   4.449 8.96e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.41 on 2994 degrees of freedom
## Multiple R-squared:  0.2399, Adjusted R-squared:  0.2386
## F-statistic:   189 on 5 and 2994 DF,  p-value: < 2.2e-16
```

We also looked at the variable of health in relation to education and wage. The base comparison group is those with health that is less than or equal to "good." The coefficient for very good health is 6.648. Compared to this group, those with very good health make $6,648 more on average than those with good health. This further shows that a number of variables impact wage beyond education level.

Conclusion

From our initial observations of the data, we were able to see interesting base trends indicating there may be a relationship between education level and wage. From there, we decided to focus on those variables to determine to what extent education level has an effect on wage. From our modeling, we found a clear positive and significant relationship between education level and wage: with each attainment level, workers were earning more than the previous on average.

Still, we wanted to look further into other variables which may affect wage level. We created a model which would model wage, education and race. We found that, with all other variables held constant, there was a negative correlation between race and wage for black and asian people compared to white people. Even with the same education level, race can influence how much a person is paid for their work. This indicates a level of racial discrimination in the labor market.

The final variable we investigated was health and we produced a model of wage, education and health. The data had limited health information, separating the survey respondents into only less than or equal to good health and less than or equal to very good health. However, we were still able to find a significant positive correlation between health quality and wage. This may mean that those with better health are more likely to be employed in better-paying jobs. It may also indicate that those in lower-paying jobs are often doing manual labor which influences their health more negatively than that of higher-paying jobs. Either way, health is a significant variable in the determination of wages.

Overall, we found that education had a significant effect on wage level. There was a correlation between educational attainment and wage which was positive as education level increased. However, we also determined that education was not the only variable which affects wage level. Factors such as race and health also have significant impacts on how much a person earns, and there are likely a number of other areas which can affect wage.