Yifei Guo (yg2496)
GU4205 Final Project
December 9, 2018

Introduction

Inequality is always a hot topic that worth studying. Although it is hard to completely eliminate inequalities in our society, efforts to understand current conditions and to make improvements are always needed. Inequalities have different types. One of the most prevail inequalities is the racial inequality, that people with different racial backgrounds are treated differently especially at workplaces and even at schools. Therefore, it is worth doing exploratory observational studies using statistical techniques to study inequalities in wages across different race classes.

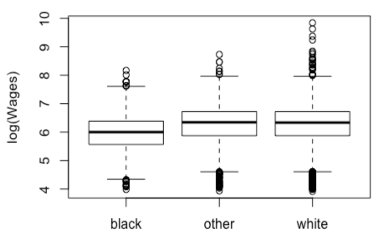
For this project, we are able to explore a data set containing information of roughly 25,000 full time male workers between the age of 18 and 70. Since there are a lot of variables that contribute to the differences in wages, our data set contains a variety of variables: *wage* (weekly wages in dollars), *edu* (years of education), *exp* (years of job experience), *city* (working in or near a city, yes or no), *reg* (US region, midwest, northeast, south, west), *race* (African America, Caucasian, Other), *deg* (college graduate, yes or no), *com* (commuting distance) and *emp* (number of employees in a company).

Our goal is to test using statistical techniques whether African American males have statistically different wages than Caucasian males and whether African American males have statistically different wages than all other males. To find out answers to these research questions, let's first briefly explore our data. The following is table including the average in wages for different race class. The mean wage for African American males is much lower than that of Caucasian males and of males from other race classes.

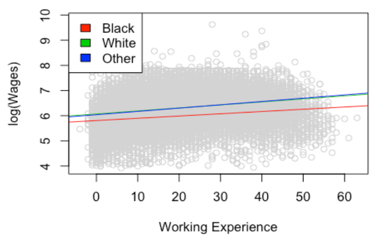
```
## African.American  Caucasian.Males           Other
##         472.2070         652.2706        649.9659
```

Figure.1 is a box plot with *log(wage)* on the y-axis and race classes on the x-axis. As we can see, the line representing the mean wage for African American males is below the lines representing the mean wages for Caucasian males and others. Another interesting feature shown by the graph is that there are more outliers for Caucasian males and the richest person in our data is Caucasian. Figure.2 is an interaction graph with three fitted lines representing relations between working experience and wages for different race classes. The two lines for Caucasian males and others almost coincide with each other but the fitted line for African American males is much below the other two lines, meaning that with the same amount of working experience, African American males get less salaries than males from other two race classes. Exploring the data set suggests that doing statistical analysis using our data set further can provide important evidences regarding inequality in workplaces and help answer the two research questions.

Wage by race Figure.1



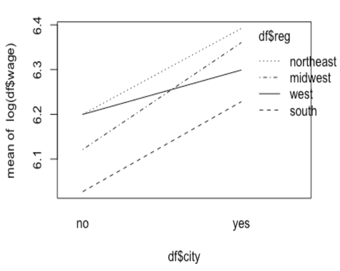
Race vs. Experience Figure.2



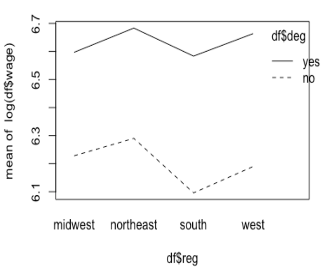
```
##             edu          exp          com          emp
## edu  1.000000000 -0.281948384 -0.003549225  0.018678815
## exp -0.281948384  1.000000000  0.001709886  0.005802539
## com -0.003549225  0.001709886  1.000000000 -0.002238374
## emp  0.018678815  0.005802539 -0.002238374  1.000000000
```

Then, we need to consider if adding any other interaction term is necessary. I decide starting my investigation by looking at interaction plots. If the slopes of fitted lines are different from each other (not parallel), then an interaction term between the two predictors included in the graph might be needed. The following 12 graphs are interaction graphs for all possible interactions. The fitted lines are very parallel to each other in Figure.14, Figure.18, Figure.19, Figure.20, Figure.21, Figure.22, Figure.24. Therefore, interactions between *reg* and *deg*, *reg* and *edu*, *city* and *emp, deg* and *emp, emp* and *reg, exp* and *city, exp* and *reg* are not needed. Candidates left are *reg* and *city* (Figure.13), *city* and *deg* (Figure.15), *city* and *edu* (Figure.16), *deg* and *edu* (Figure.17), *exp* and *deg* (Figure.23).

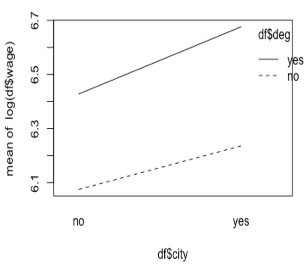
City vs. Region Figure.13



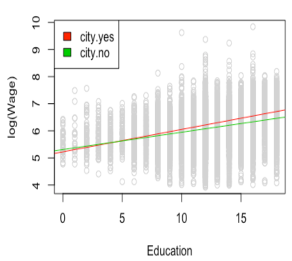
Region vs. Degree Figure.14



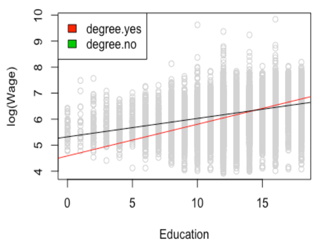
City vs. Degree Figure.15



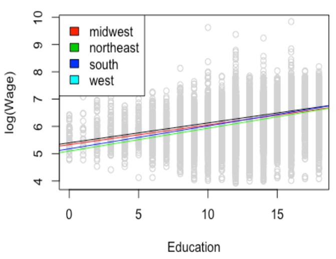
City vs. Education Figure.16



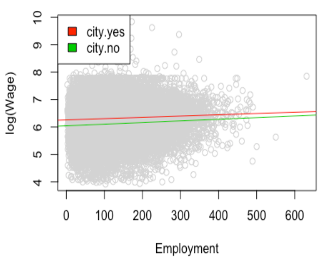
Degree vs. Education Figure.17



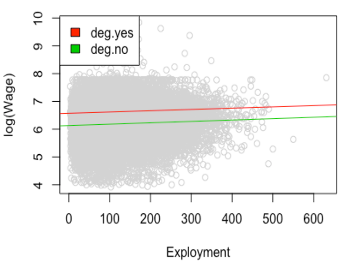
Region vs. Education Figure.18



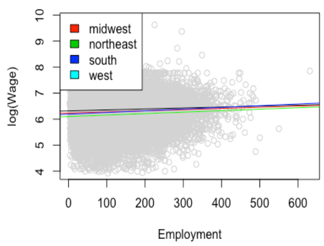
Employment vs. City Figure.19



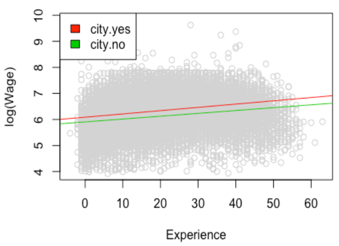
Employment vs. Degree Figure.20



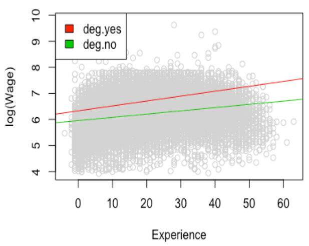
Employment vs. Region Figure.21



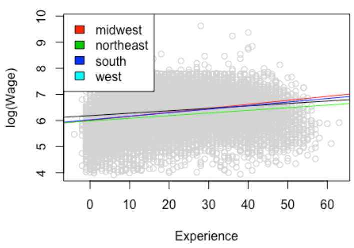
Experience vs. City Figure.22



Experience vs. Degree Figure.23



Experience vs. Region Figure.24



Adding too many interaction terms increases the risk of overfitting because the second interaction term added might have similar effects in mitigating multicollinearity as the first interaction term added. Therefore, it is better to examine the reduction in AIC by adding one interaction term at a time to our current Model 5. In Figure 13, only the slope of *west* is very different from the other three fitted-lines. Therefore, add an interaction between city and a dummy variable *west* is sufficient. The following table shows the reduction in AIC by adding one interaction term at a time.