IE4

Introduction

In this paper we will be focusing around the impact of a worker's sex on their hourly wage. The dataset we are using for this project is a random sample of 534 people from the 1985 Current Population Survey (CPS) which contains cross-sectional data of potential determinants of wages. In this assignment we will use several statistical techniques we have acquired from our study combined with some real-world knowledge to try and answer some questions one may have about this dataset.

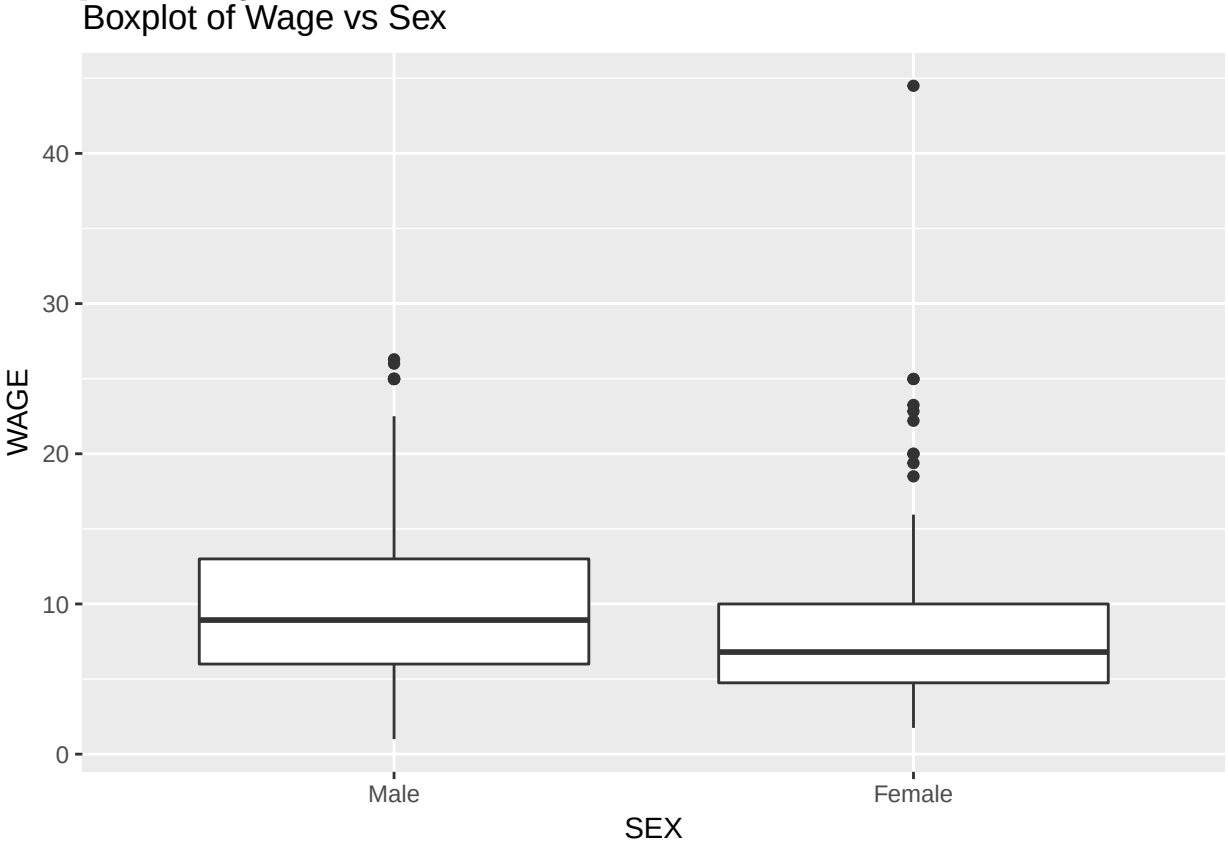
Data Definitions

- *EDUCATION* : Number of years of education
- *SOUTH* : Indictor variable if a person is from the south or not
  0. Not From South
  1. From South
- *SEX* : Catagorical variable indiciting whether one was male or female
  0. Male
  1. Female
- *EXPERIENCE* : Number of years of work experience
- *UNION* :
  0. Not in a Union
  1. In a Union
- *WAGE* : Wage earned per hour
- *AGE* : Age in year
- *RACE* : Catagorical variable indiciting ones race
  1. Other
  2. Hispanic
  3. White
- *OCCUPATION* : Catagorical variable indiciting ones occupation
  1. Management
  2. Sales
  3. Clerical
  4. Service
  5. Professional
  6. Other
- *SECTOR* : Catagorical variable indiciting ones sector
  0. Other
  1. Manufacturing
  2. Construction
- *MARRIED* : Catagorical variable indiciting ones marriage status
  0. Unmarried
  1. Married

Does a "Wage Gap" Exist between Men and Women?

The first question we will attempt to answer is whether or not a wage gap exists between men and women. We can try to visualize the difference in male and female wages by using a box plot. From the figure below we can clearly see that the mean wage for men and women, denoted by the black line in each respective box,

is unequal; on average men earn more than women.



A natural question to ask is, "Is this difference significant?" In order to answer this we must construct a statistical test. Since we have a relatively large sample, 245 women and 289 men, we will conduct a large sample hypothesis t-test with a 95% confidence interval.

```
## Number of women: 245
## Number of men: 289
```

$$H_o : \mu_{male} = \mu_{female} H_a : \mu_{male} \geq \mu_{female}$$

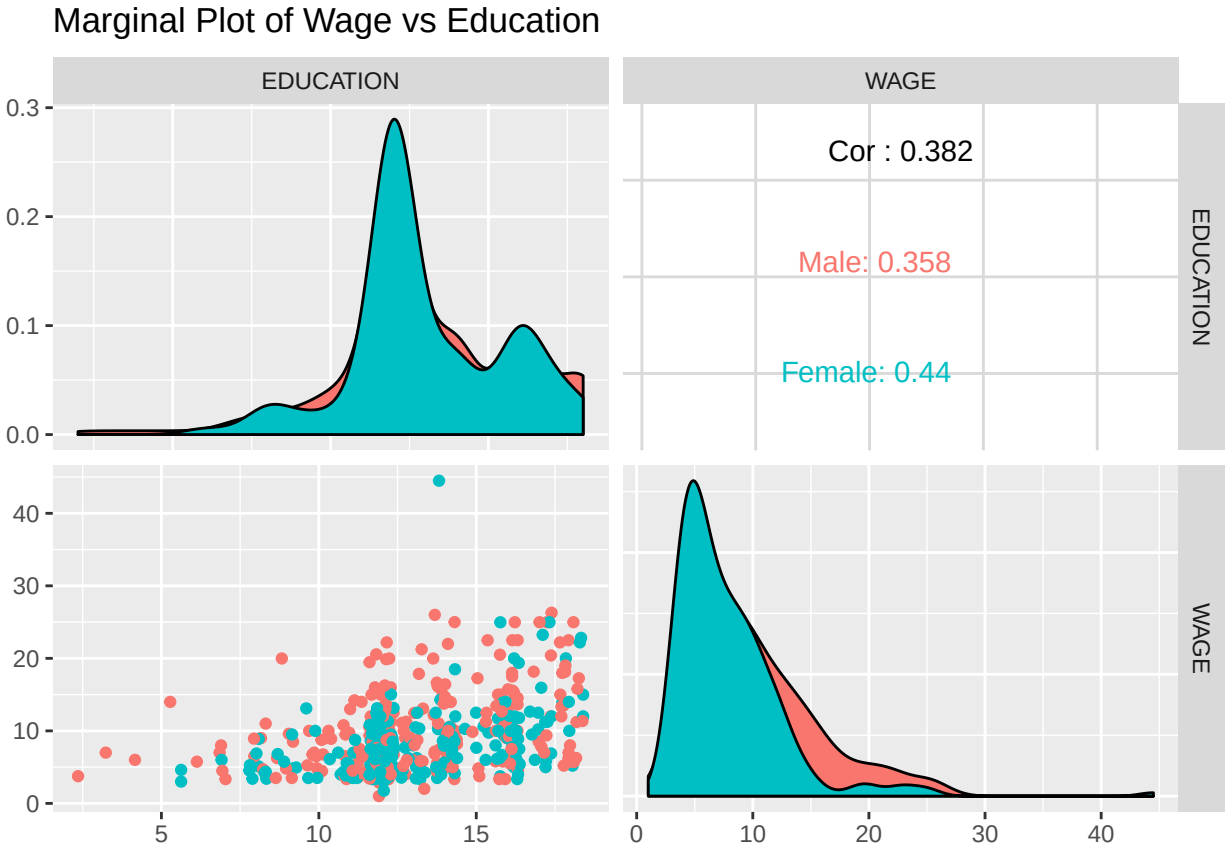
```
##
##  Welch Two Sample t-test
##
## data:  men$WAGE and women$WAGE
## t = 4.8853, df = 530.55, p-value = 6.847e-07
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.402348      Inf
## sample estimates:
## mean of x mean of y
##  9.994913  7.878857
```

Based on the results from our t-test we must reject the null hypothesis and accept the alternative hypothesis. In other words, the difference in the average wages between men and women is statistically significant. Men make ~$10/hr on average and women make ~$7.87 in 1985 dollars, in 2018 dollars this would be $23.12/hr and $18.21/hr respectively. Now that we have established that a statistically significant wage gap exists this motivates us to investigate why a wage gap between men and women exists.

Education and Wage Gap

Something that comes to mind when thinking of the determinants of wage is a person's level of education. Loosely speaking, one would expect that the more higher your level of education, the more money you will make. There is some empirical evidence behind this claim. If we look at data from the US BLS Annual Demographic Supplement of the CPS from 2013 we find that this is generally true for those above the age of 25. Also, interestingly, if we pay close attention to the scales of each axis, we find that the wage gap is still present.

In this study we will not be looking numerically at the data shown above, instead we will use it qualitatively for some intuition and partial justification for adding education into our multiple linear regression model. Before just wildly throwing a term into our model we can look at the marginal plot of wage versus number of years of education.



The focus of on this plot should be on the lower triangular elements. In the bottom left we have a scatterplot with jittered points with wage on the y-axis and number of years of education on the x-axis. In this scatterplot it is difficult to fully understand how many people have 12 years of education. The density plot in the top left roughly shows the percentage of the data for each level of education. To get an even better idea of this skew we found that 109 men and 110 women had exactly 12 years of education; combined, this accounts for a large part of our dataset, roughly 41%. In the bottom right we have a density plot for wages which isn't as interesting to us right now but it is worthy to note the skew in the wage distribution in this dataset.