IE5

Ben Kaufman
Vic Chan
Ziwei Zhang

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 explore our 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

Background Information on the 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 higher your level of education, the higher your hourly wage should be. 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 on average, controlling for education and sex, men make more than women on average.

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.

Scatterplot Matrix

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

The dataset we chose has many factors and only a couple of continuous variables. Above, in Figure 2 we only use a few of the variables from our entire dataset. A few things pop out from the scatterplot matrix. Starting from the top left cell and looking down the main diagonal we see that men and women have similar distributions of levels of education, years of experience and age. We also see that our dataset contains about the same amount of men and women, 289 and 245 respectively. Looking at the bottom right cell we find that men and women have dissimilar wage distributions. In the box plot above this cell we clearly see that the median wage for men is higher than that of women. When looking at the marginal plots in the lower triangular region of the scatterplot matrix the two plots that stick out the most are those comparing *Age* vs. *Experience* and *Wage* vs. *Education.* We will explore these relationships more deeply and create models using statistical reasoning combined with real-world knowledge in attempt to explain wage gaps between men and women.

Model 1: WAGE ~ EDUCATION + SEX

First, we would like to explore the existence of a wage gap when controlling for years of education.

$$M1 : E(Wage|Education, Sex) = \beta_{0,1} + \beta_{1,1}Education + \beta_{2,1}Sex \tag{1}$$

```
##
## Call:
## lm(formula = WAGE ~ EDUCATION + SEX, data = wages)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.888 -2.997 -0.709  2.255 35.888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.21783    1.03632   0.210    0.834
```

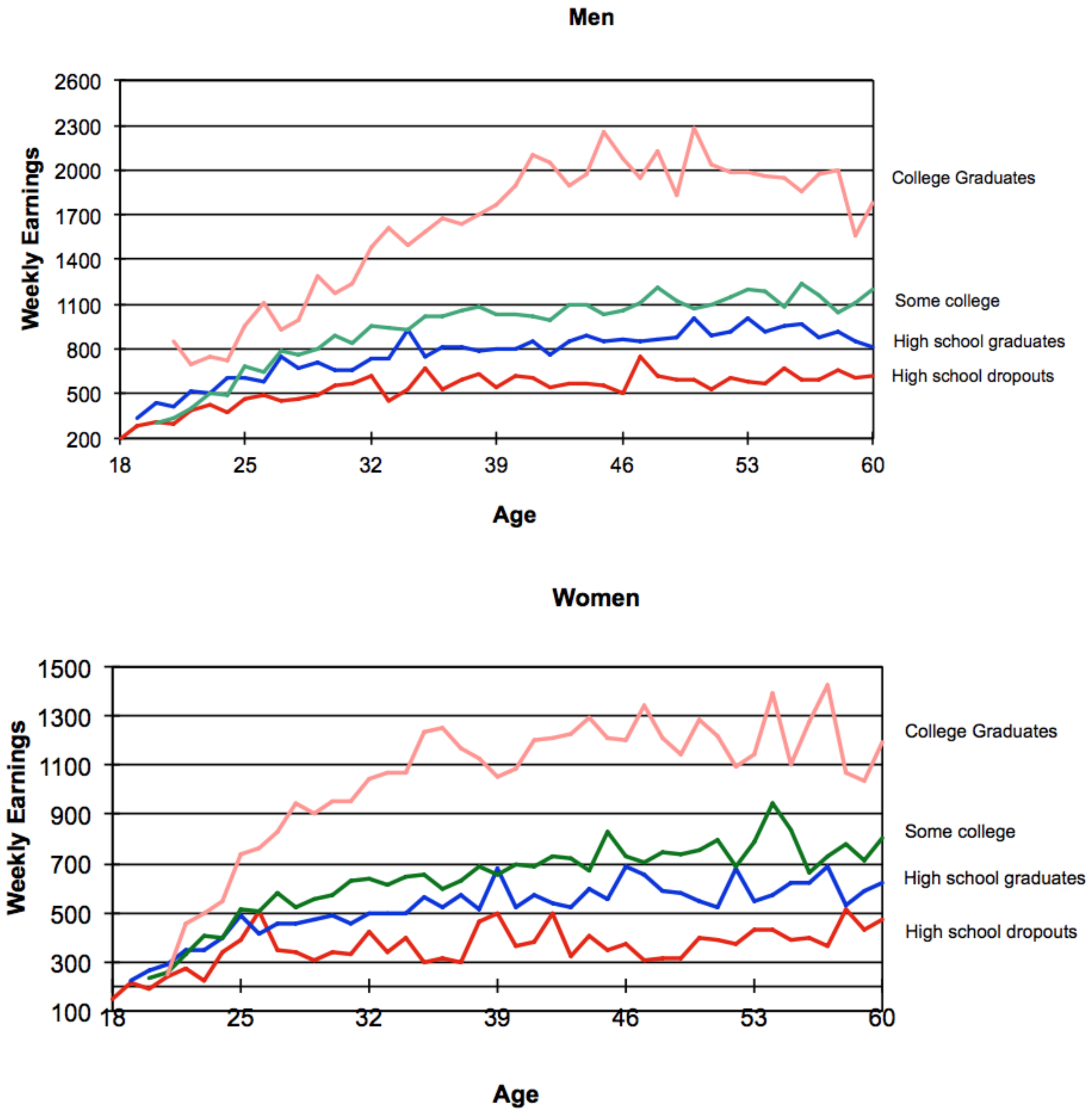


Figure 1: picture

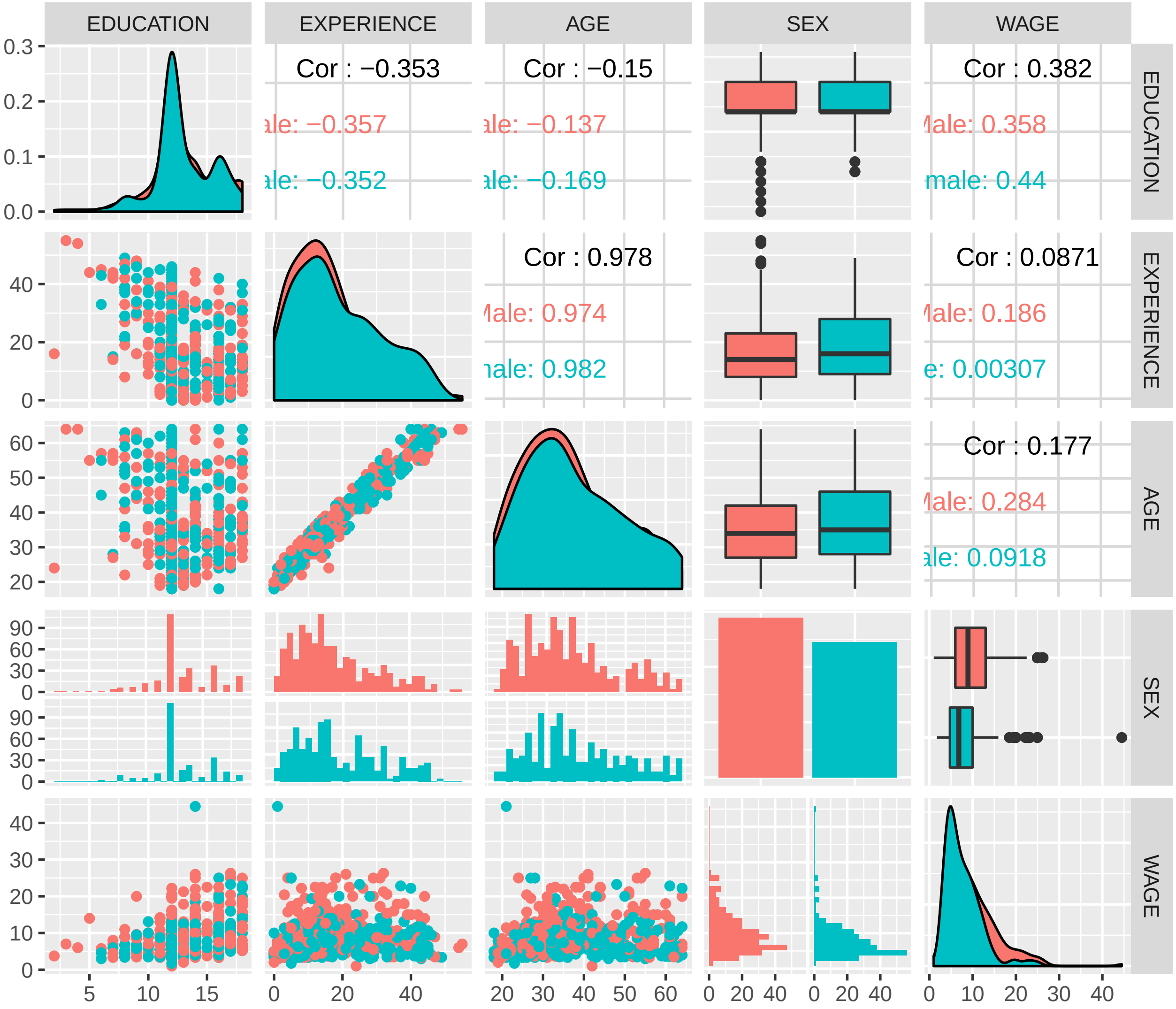


Figure 2: Scatterplot Matrix

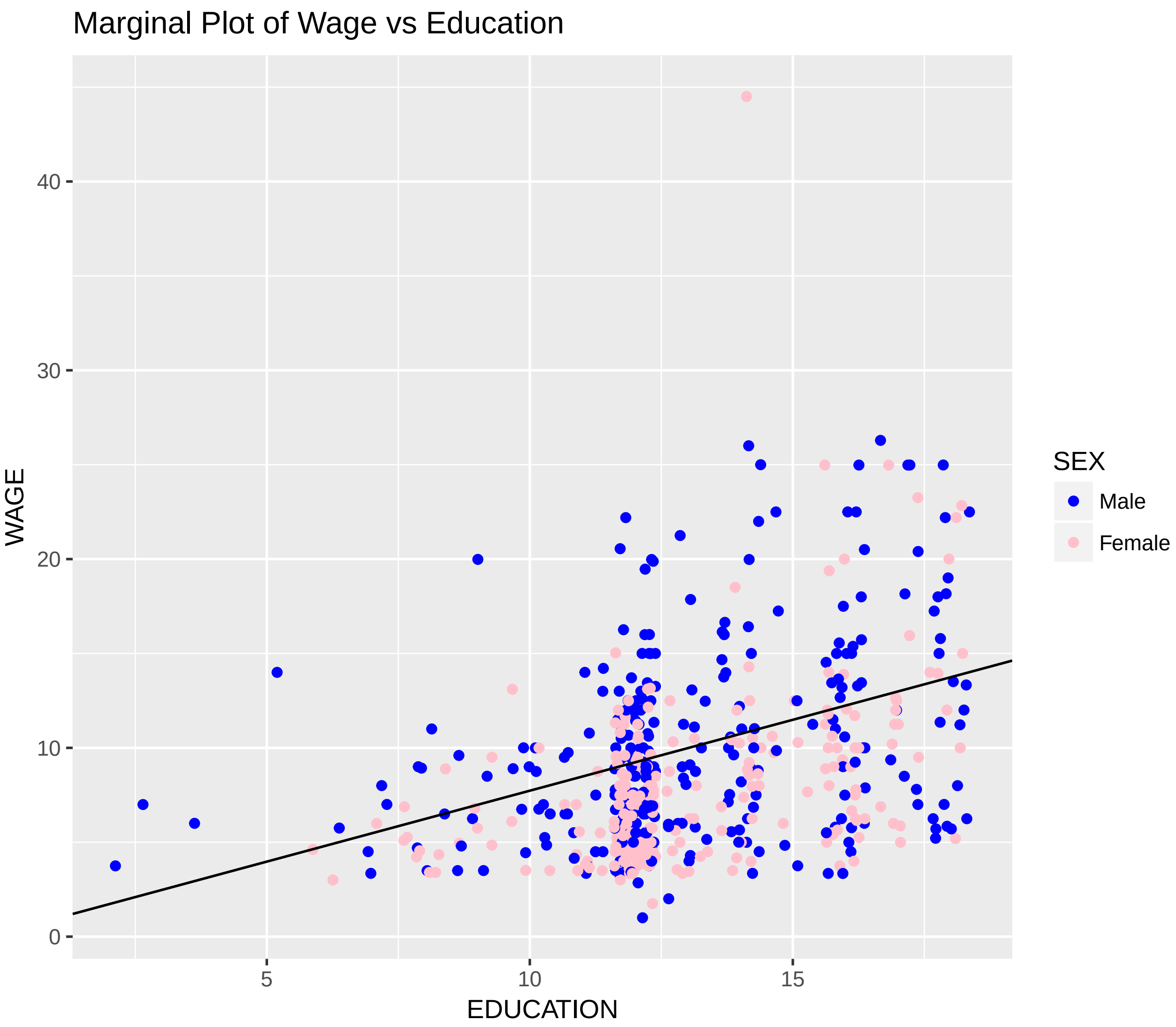


Figure 3: Fitted Line of M1

```
## EDUCATION    0.75128    0.07682   9.779  < 2e-16 ***
## SEXFemale   -2.12406    0.40283  -5.273 1.96e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.639 on 531 degrees of freedom
## Multiple R-squared:  0.1884, Adjusted R-squared:  0.1853
## F-statistic: 61.62 on 2 and 531 DF,  p-value: < 2.2e-16
```

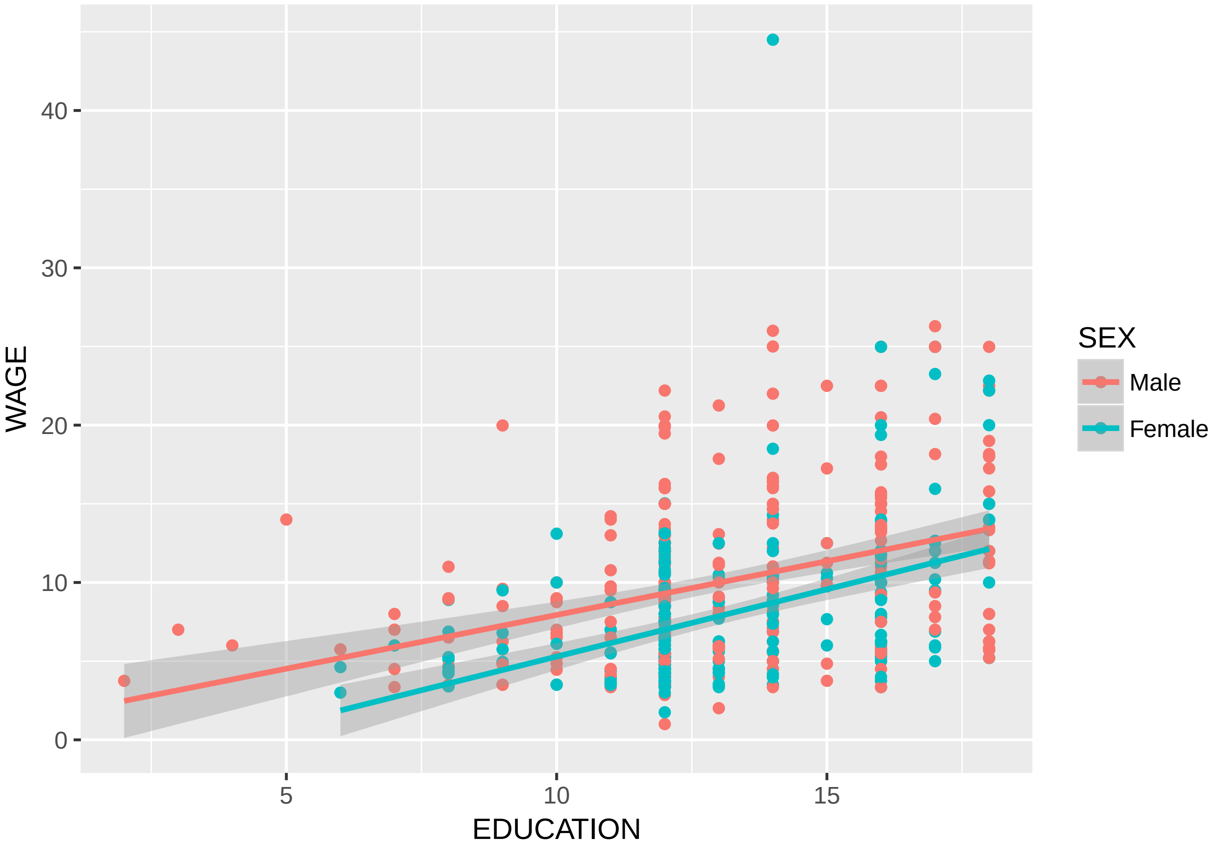
As we see in Figure 3 the model describes the positive relationship between *Wage* and *Education* as shown by the slope of the line, $\beta_{1,1} = 0.75128 > 0$. We also note that the confidence interval for this parameter is roughly (0.6, 0.9) which does not contain zero meaning that this parameter is significant. Under this model we also see that a wage gap between men and women is present. It is estimated that a women of will make \$2.12 ($\beta_{2,1} = -2.124$) per hour less than a man of similar education level. We note that the confidence interval for this estimate is (-2.915, -1.333) which again, does not contain zero and is significant. To wrap up this section, we will remember for the purposes of comparison later on that the R^2 for our "baseline" model is 0.1884.

Model 2: WAGE ~ EDUCATION + SEX + SEX:EDUCATION

A question that seems reasonable is, does the wage gap between men and women decrease as the level of education for both groups increases? In order to explore this we create a new model with an interaction term between *Sex* and *Education*.

$$M2 : E(Wage|Education, Sex) = \beta_{0,2} + \beta_{1,2}Education + \beta_{2,2}Sex + \beta_{3,2}Sex : Education \tag{2}$$

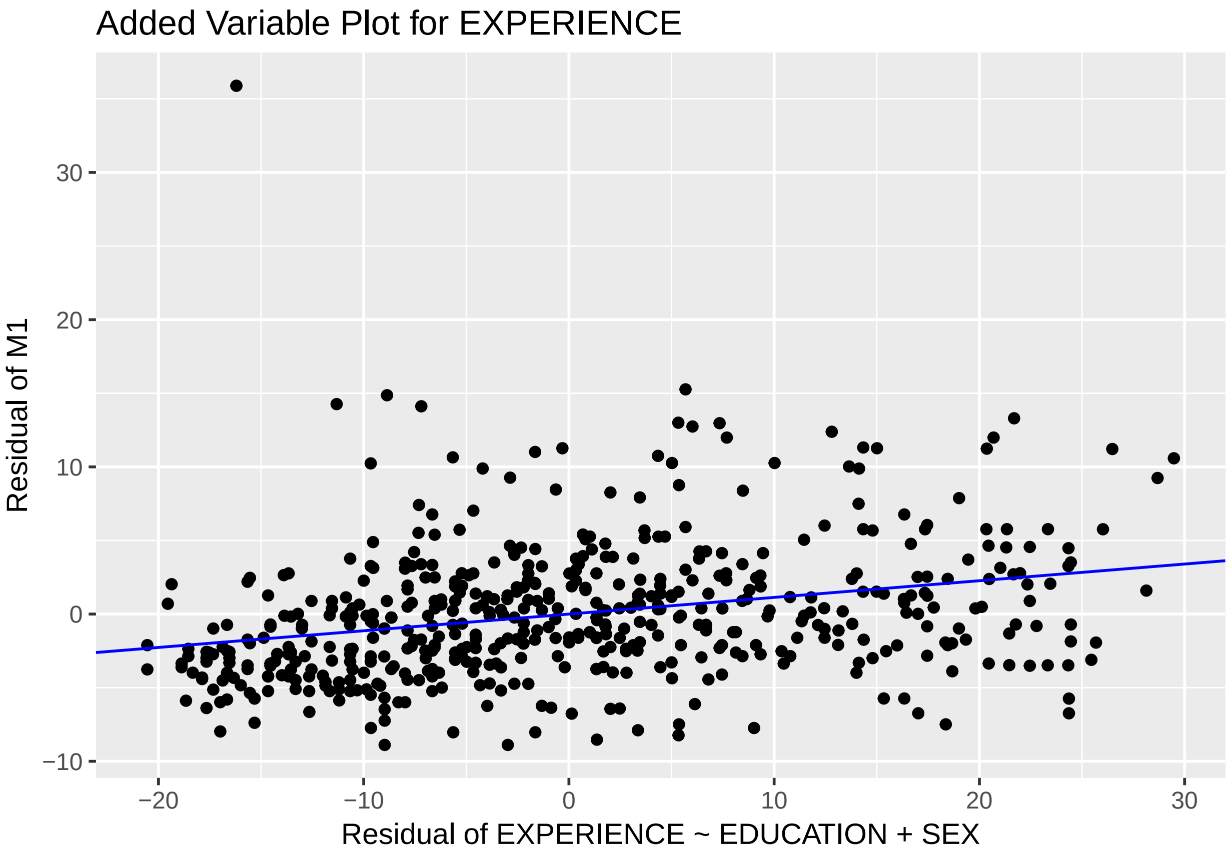
```r
M2<-lm(WAGE~EDUCATION+SEX+SEX:EDUCATION, data=wages)

p <- ggplot(aes(x=EDUCATION, y=WAGE),data=wages)
p + geom_point(data=wages, aes(color=SEX)) + geom_smooth(method='lm', aes(color=SEX), data=wages)
```



In Figure 4 we can see the the fitted lines between men and women seem to have a different intercept but have roughly the same slope. To hammer this point down we find that the estimated coefficient for the interaction term has a confindence interval of (-0.136, 0.481) which contains zero and cannot be effectively relied upon. All of this information supports the argument that wage gap between men and women **does not** decrease as the number of years of education increases.

Added Variable Plot: M1 and Experience

As it stands our model only uses sex and education as predictors. As we know from the real world, your wage might be related to the amount of experience you have had in a job. For example, more exclusive jobs normally pay more and the reason they may be exclusive is because they may require x years of experience. This allows us to hypothesize that your wage will depend upon the amount of experience you've had, more specifically, we would believe that your wage should increase with your level of experience. In order to justify adding this variable to our model we will create an added variable plot.



```
##
## Call:
## lm(formula = resid_M1 ~ resid_AVP1_x, data = wages)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.571 -2.746 -0.653  1.893 37.724
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.739e-17  1.924e-01   0.000        1
## resid_AVP1_x 1.133e-01  1.668e-02   6.794 2.93e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.445 on 532 degrees of freedom
## Multiple R-squared:  0.07983,    Adjusted R-squared:  0.07811
## F-statistic: 46.16 on 1 and 532 DF,  p-value: 2.931e-11
```

```
##                    2.5 %    97.5 %
## (Intercept)  -0.3778989 0.3778989
## resid_AVP1_x  0.0805399 0.1460608
```

In the above added variable plot we see that there is a solid linear fit with a non-zero slope. The slope has a 95% CI of (0.081, 0.146) which doesn't contain zero and we also see that slope is non-zero by inspection of the graph. Also, by looking at the R^2 value of an OLS model fitted to this data we find that almost 8% of the remaining variance in our model is explained by adding experience to our existing model. For our purposes this is a considerable amount because it increases our total explained variability by about 50%.

This leads to actually defining our new model which includes experience:

Model 3: WAGE ~ EDUCATION + SEX + EXPERIENCE

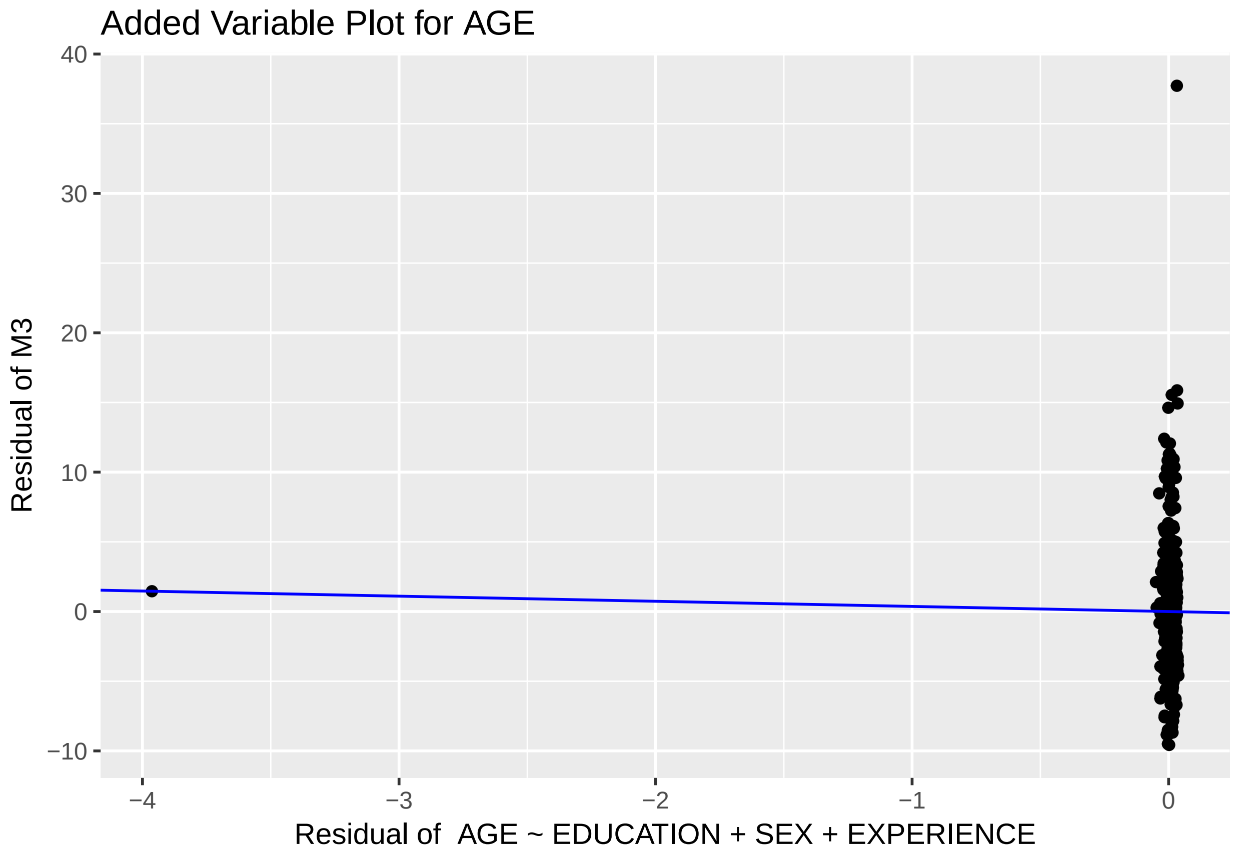
$$\begin{align} M3 : E(Wage|Education, Sex, Experience) = \beta_{0,3} + \beta_{1,3}Education + \beta_{2,3}Sex + \beta_{3,3}Experience \tag{3} \\ \tag{4} \end{align}$$

Looking at our new model we look back to our original question which is, does the wage gap between men and women still exist? According to M3 we found that the wage gap between men and women increased after we accounted for experience from \$2.12 per hour to \$2.34 per hour which opposed our initial assumptions. As usual, we found that the confidence interval for $\beta_{2,3}$ is (-3.0999,-1.5753) which does not contain zero and the p-value is $<< 0$. Lastly, to compare for the purposes of comparison to previous and future models the R^2 of this model is 0.2532.

Added Variable Plot: M3 and Age

Another variable in our data set we may want to consider is *Age*. *Age* may not directly influence your wage but it may influence other variables which directly influence your wage. This intuition combined with our observation of high correlation between *Age* and *Experience* from Figure 2 suggests that this variable may be collinear variables that already exist in our model. To further examine this we create an added variable plot.

```r
wages$resid_M3 = resid(M3)

avp2_mod <- lm(AGE ~ EDUCATION + SEX + EXPERIENCE, data=wages)
wages$resid_AVP2_x = resid(avp2_mod)
avp2_2 <- lm(resid_M3 ~ resid_AVP2_x, data=wages)

p <- ggplot(aes(x=resid_AVP2_x, y=resid_M3), data=wages)
p + geom_point() +
  ggtitle("Added Variable Plot for AGE") +
  xlab('Residual of  AGE ~ EDUCATION + SEX + EXPERIENCE') +
  ylab('Residual of M3') +
  geom_abline ( intercept = coef(avp2_2)[1], slope = coef(avp2_2)[2], color="blue" )
```



```
summary(avp2_2)
```

```
##
## Call:
## lm(formula = resid_M3 ~ resid_AVP2_x, data = wages)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.571 -2.745 -0.652  1.894 37.736
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.849e-16  1.924e-01   0.000    1.000
## resid_AVP2_x -3.675e-01  1.116e+00  -0.329    0.742
##
## Residual standard error: 4.445 on 532 degrees of freedom
## Multiple R-squared:  0.0002036,  Adjusted R-squared:  -0.001676
## F-statistic: 0.1084 on 1 and 532 DF,  p-value: 0.7422
```

```
confint(avp2_2)
```

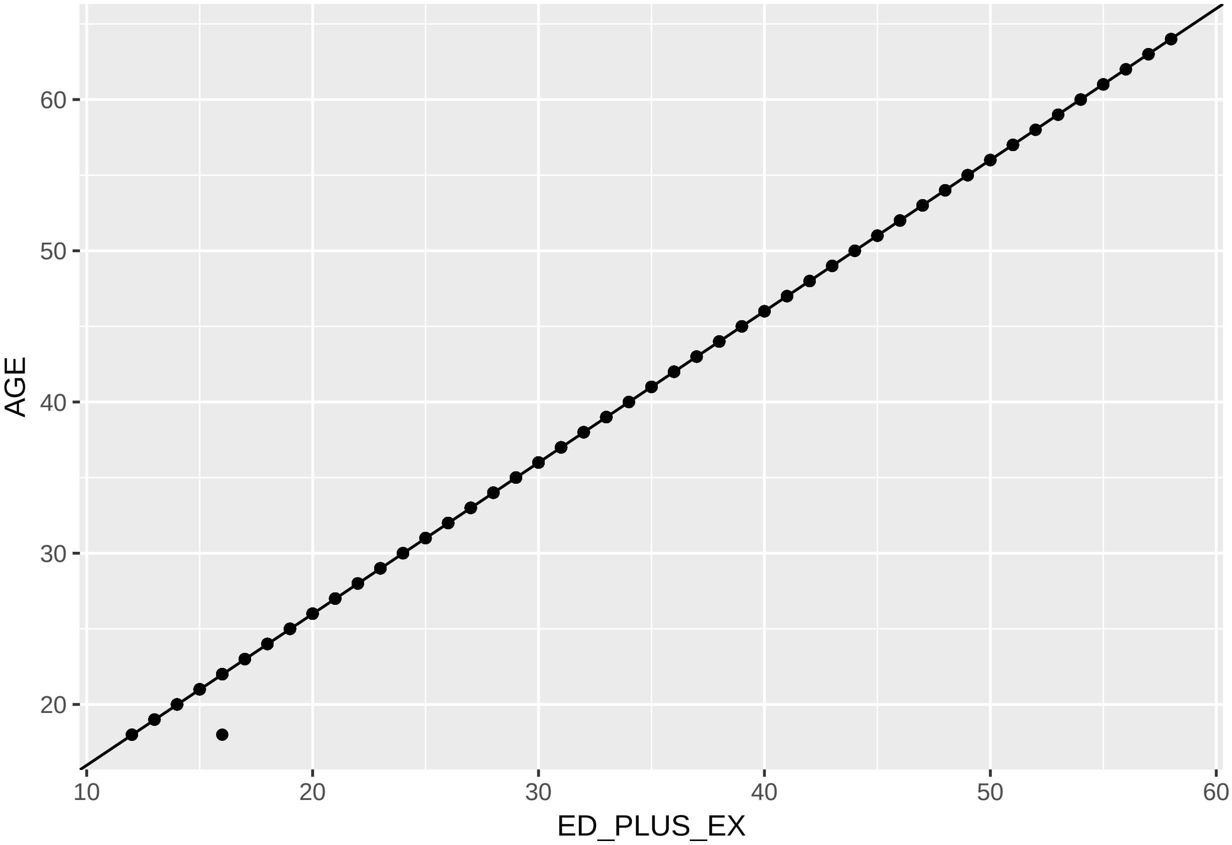
```
##                   2.5 %    97.5 %
## (Intercept)  -0.3778604 0.3778604
## resid_AVP2_x -2.5605558 1.8255822
```

As we can see from the added variable plot adding *Age* to our model is not very helpful because a line might describe the data well if the outlier did not exist but the slope would infinite. In this case we see that the fitted line does not in fact describe the data well because of the outlier. The confidence interval for the slope is (-2.56,1.83) which contains zero so we cannot assert that the slope is non-zero. Futhermore, our R^2 is 0.0002036 which is nearly zero meaning that variance that age would explain is for the most part already explained by variables in the model.

Collinearity

From Figure 2 the only variables that seemed to be particularly collinear were *Age* and *Experience*. The previous added variable plot suggests there is collinearity between the variables *Education*, *Experience*, *Sex*, and *Age*. To nail down this relationship we define a new variable *ED_PLUS_EX* which is the number of years of education plus the number of years of experience for each person in our dataset. If we look at the marginal plot of *Age* vs. *ED_PLUS_EX* and fit a model we see that these variables are nearly perfectly collinear apart from one outlier. Intersetingly, the outlier seems to be a person who is 18 years old and has already finished four years of college.

```
collin_mod <- lm(AGE ~ ED_PLUS_EX, wages)
p <- ggplot(data = wages , aes ( x = ED_PLUS_EX, y = AGE) )
p + geom_point() +
  geom_abline ( intercept = coef(collin_mod)[1], slope = coef(collin_mod)[2] ,color="black")
```



If we look at the fitted model we essentially see that $Age = 6 + 1 * (Education + Experience)$ which tells us something we should already know in the US. In the US most people begin their schooling, enter first grade, at around age 6. What we find is that for the rest of people's lives in our dataset people are either in school or working which makes sense because most people retire at 65 and the oldest person is 64 years old. To conclude our discussion of *Age* we have decided to not use it in our model because it would interfere with our existing variables in our model. However, we will note that since we have made this choice our estimate of the wage gap between genders may be underestimated because the amount of experience you have may be influenced by your gender; in this case women may have less experience than men because they may have to take time off from work to raise children. Since this is a subtle point and there is no data to support time off between genders in this case further discussion will be in the caveats section.

Power Transformations

In order to find a better fit to the data we can try to add some non-linearity to our model without making our model non-linear. We do this using power transformations on the predictors and then on the response.

```r
wages$EXPERIENCE_P1 = wages$EXPERIENCE + 1
coef(powerTransform(wages[,c("EDUCATION","EXPERIENCE_P1")]), round=TRUE)
```

```
##     EDUCATION EXPERIENCE_P1 
##           1.0           0.5
```

```r
summary(powerTransform(wages[,c("EDUCATION","EXPERIENCE_P1")]))
```

```
## bcPower Transformations to Multinormality 
##               Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd
## EDUCATION        1.1750         1.0       0.9181       1.4319
## EXPERIENCE_P1    0.4694         0.5       0.3722       0.5666
## 
## Likelihood ratio tests about transformation parameters
##                              LRT df pval
## LR test, lambda = (0 0) 225.64946  2    0
## LR test, lambda = (1 1)  99.97956  2    0
```

The results from the power transformation suggests that we should not transform *Education* and that it is advisable to take the square root of *Experience* once we have adjusted the domain of *Experience* to be strictly greater than zero. In the real world this may be justified by diminishing returns to experience (see caveats). We can then use the Box-Cox Transformation to transform our response variable.

```r
wages$tranEXPERIENCE = wages$EXPERIENCE_P1^(1/2)
coef(powerTransform(WAGE~tranEXPERIENCE + EDUCATION + SEX, data=wages), round=TRUE)
```

```
## Y1 
##  0
```

Since our suggested value of λ from the transformation is zero we will take the log of our response variable in our new model.

M4: log(WAGE) ~ EDUCATION + SEX + EXPERIENCE

```r
M4<-lm(log(WAGE)~tranEXPERIENCE + EDUCATION + SEX, data=wages)
summary(M4)
```

```
## 
## Call:
## lm(formula = log(WAGE) ~ tranEXPERIENCE + EDUCATION + SEX, data = wages)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.18596 -0.29069  0.01133  0.30256  2.09292 
## 
```

```
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.426097   0.132353   3.219  0.00136 **
## tranEXPERIENCE   0.117397   0.013832   8.488  < 2e-16 ***
## EDUCATION        0.097740   0.007788  12.550  < 2e-16 ***
## SEXFemale       -0.257906   0.038863  -6.636 7.97e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4461 on 530 degrees of freedom
## Multiple R-squared:  0.2894, Adjusted R-squared:  0.2854
## F-statistic: 71.95 on 3 and 530 DF,  p-value: < 2.2e-16
```

```
#RSS_M4 <- sum((wages$WAGE - exp(fitted(M4)))^2)
#SYY <- sum((wages$WAGE - mean(wages$WAGE))^2)
#R2_M4 <- 1 - (RSS_M4 / SYY)
#R2_M4
```

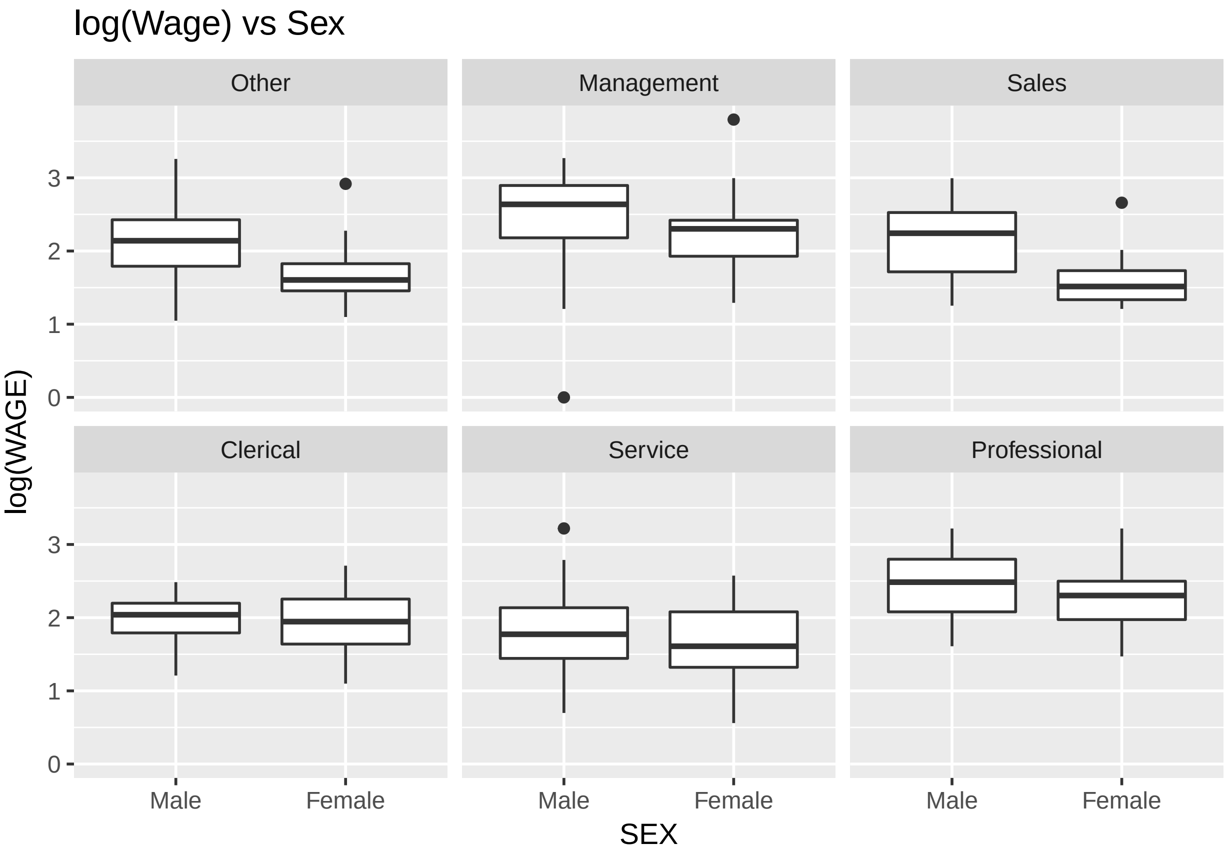
In our new transformed model our definition of the wage gap between men and women changes. Our interpretation of the wage gap now becomes the percentage difference in wage between men and women adjusting for all other variables. In M4 we find that adjusting for *Education* and our new interpretation of *Experience* women make 25.79% less than men. As usual, for reference our R^2 for M4 is 0.2894 however we cannot compared this directly to the R^2 values of our previous models since we have taken the log.

M5: log(WAGE) ~ EDUCATION + SEX*OCCUPATION + EXPERIENCE

Now that we are accounting for a person's education and experience we can now try to explore the effect of a person's occupation on wage.

We get some intuition for this without controlling for Education and Experience

```
p <- ggplot( wages, aes ( x = SEX, y = log(WAGE)) )
p + geom_boxplot() + ggtitle("log(Wage) vs Sex") + facet_wrap(~OCCUPATION)
```



Notice that the coefficients on the interaction terms are the differences in the gaps with respect to the gap in occupation other. Notice that the only other bigger gap is in Sales, this is denoted by negative sign on the SexFemale:OccupationSales coefficient

Caveats and Final Notes