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* : Categorical 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* : Categorical variable indiciting ones race
    1. Other
    2. Hispanic
    3. White
- *OCCUPATION* : Categorical variable indiciting ones occupation
    1. Management
    2. Sales
    3. Clerical
    4. Service
    5. Professional
    6. Other
- *SECTOR* : Categorical variable indiciting ones sector
    0. Other
    1. Manufacturing
    2. Construction
- *MARRIED* : Categorical 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
```

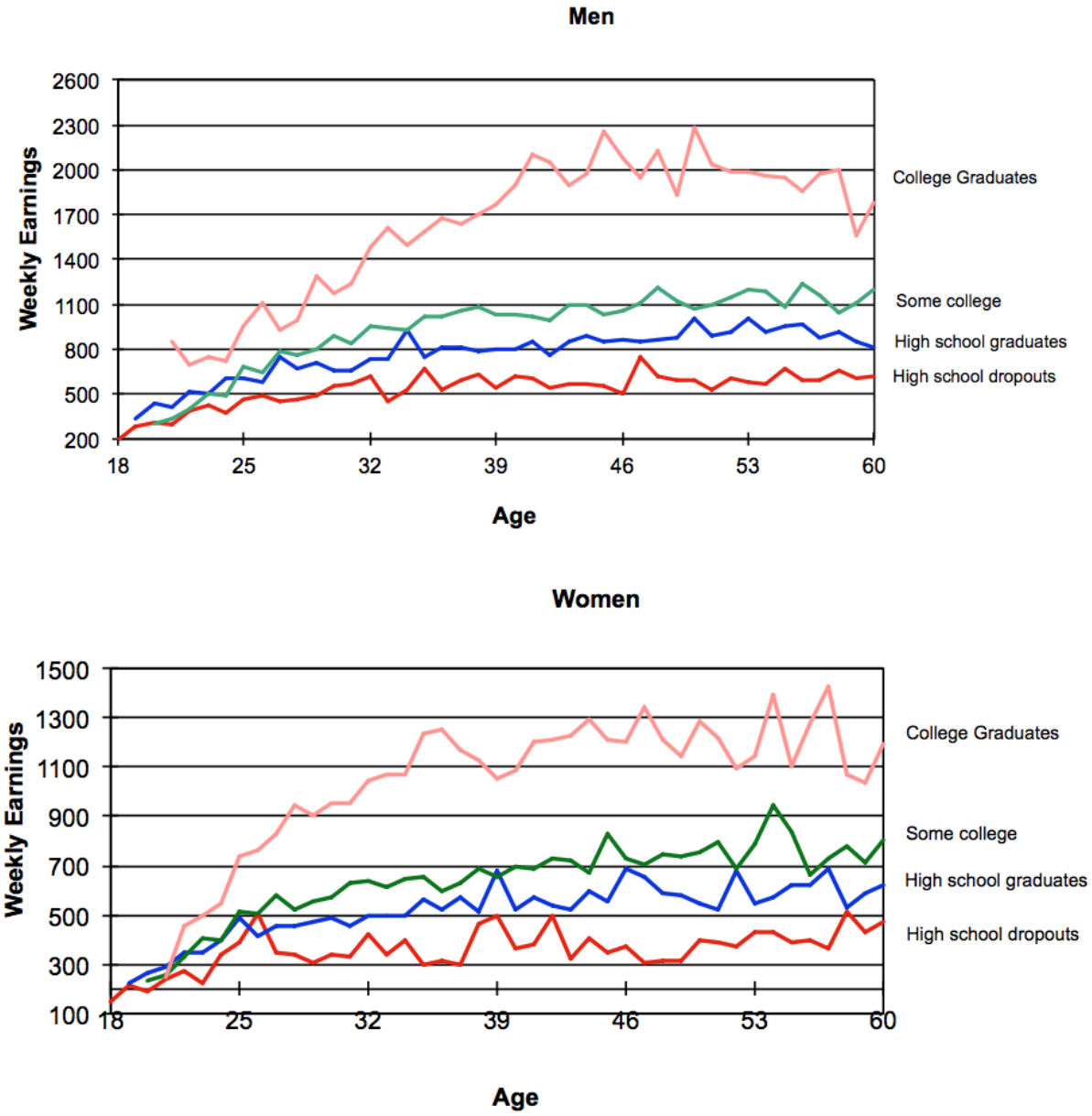


Figure 1: picture

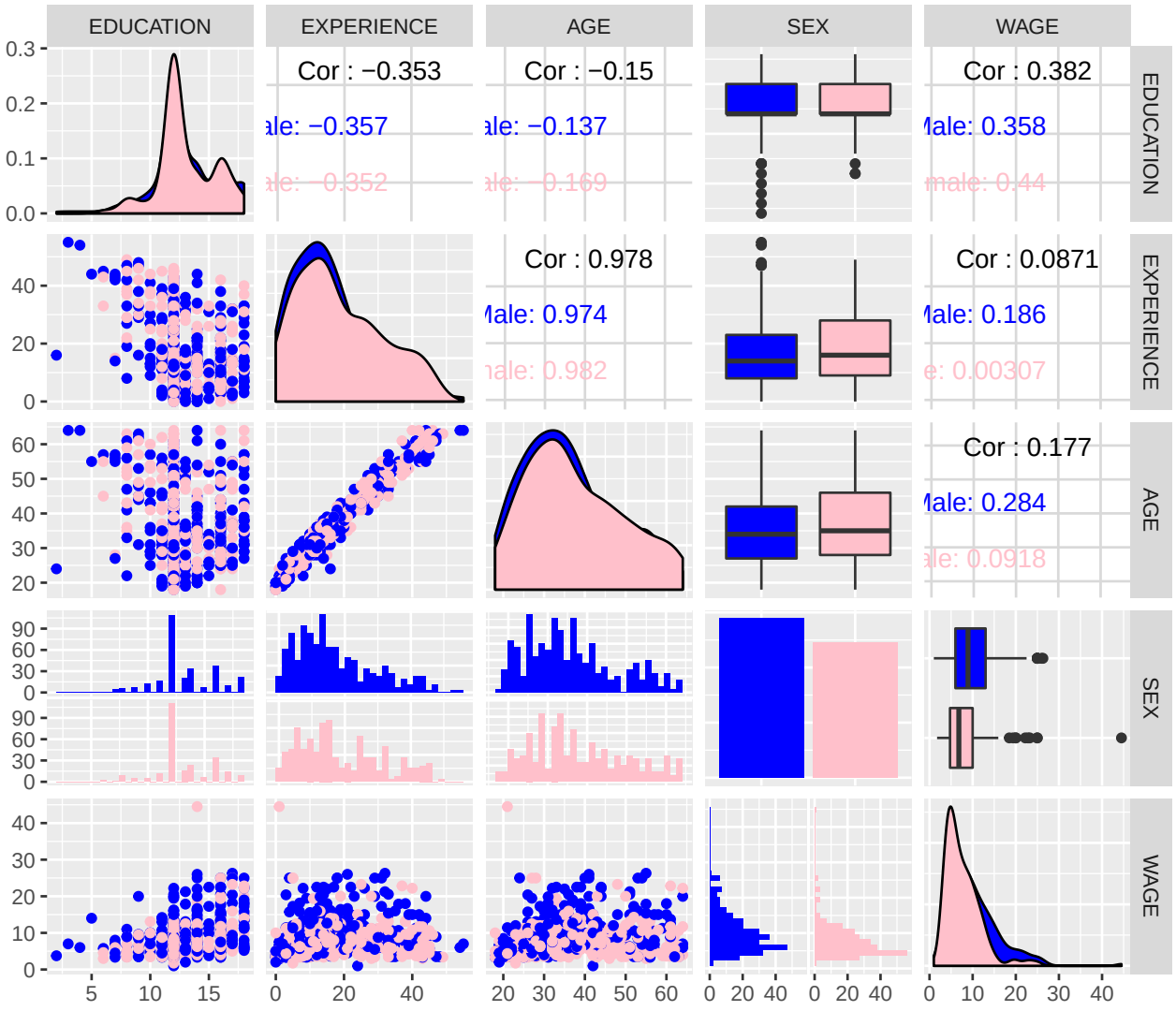


Figure 2: Scatterplot Matrix

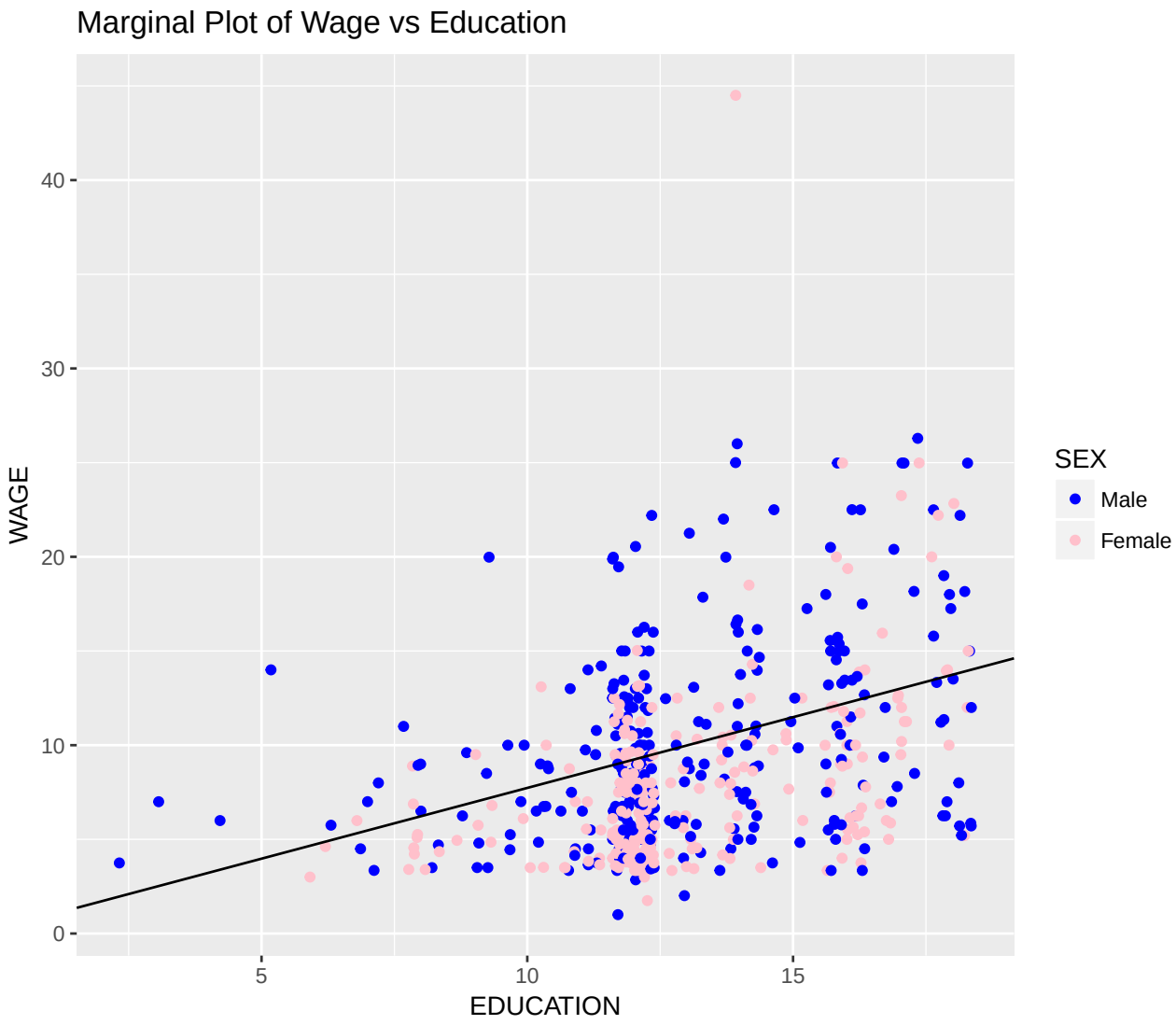


Figure 3: Fitted Line of M1

```
## 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
```

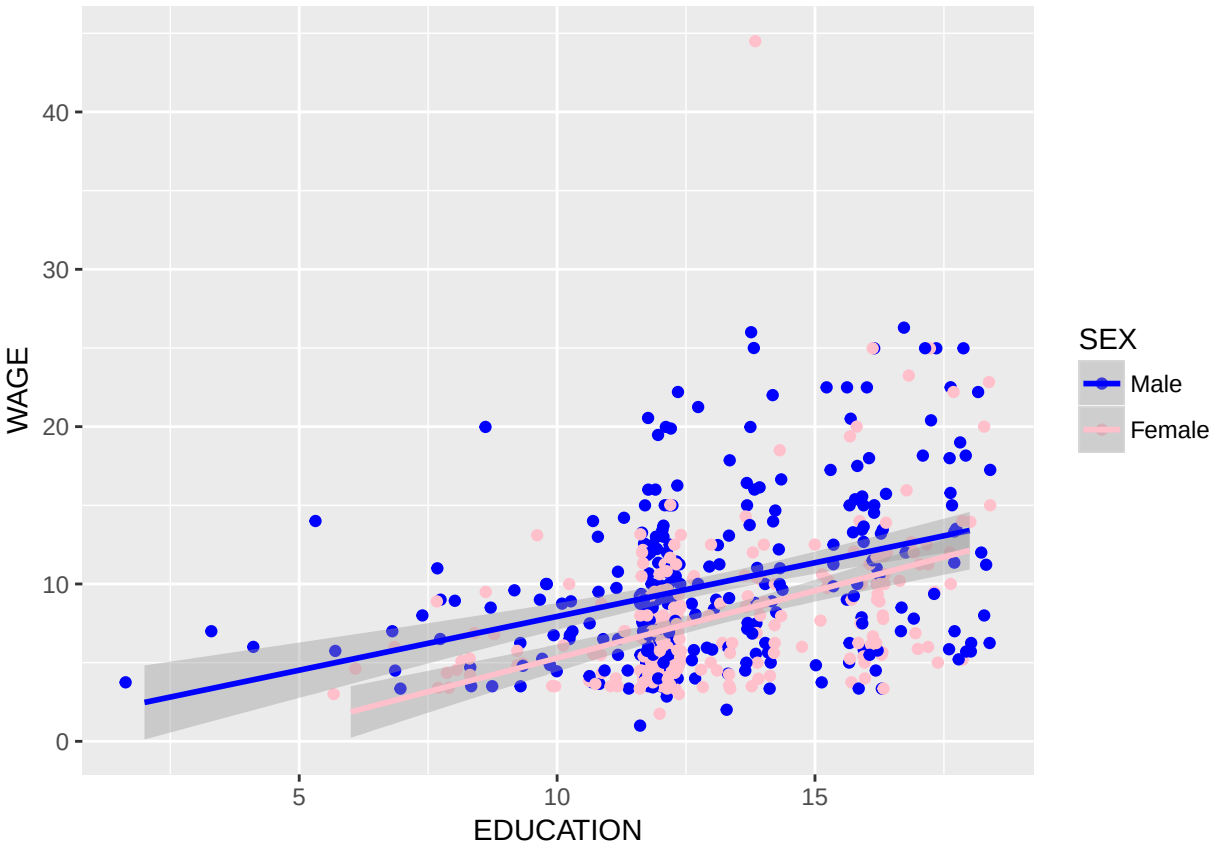
As we see in Figure 3 the model describes the positive relationship between *Wage* and *Education* as shown by the slope of the line, $\beta_{1,1} = 0.75128 > 0$. We also note that the confidence interval for this parameter is roughly (0.6, 0.9) which does not contain zero meaning that this parameter 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*.

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

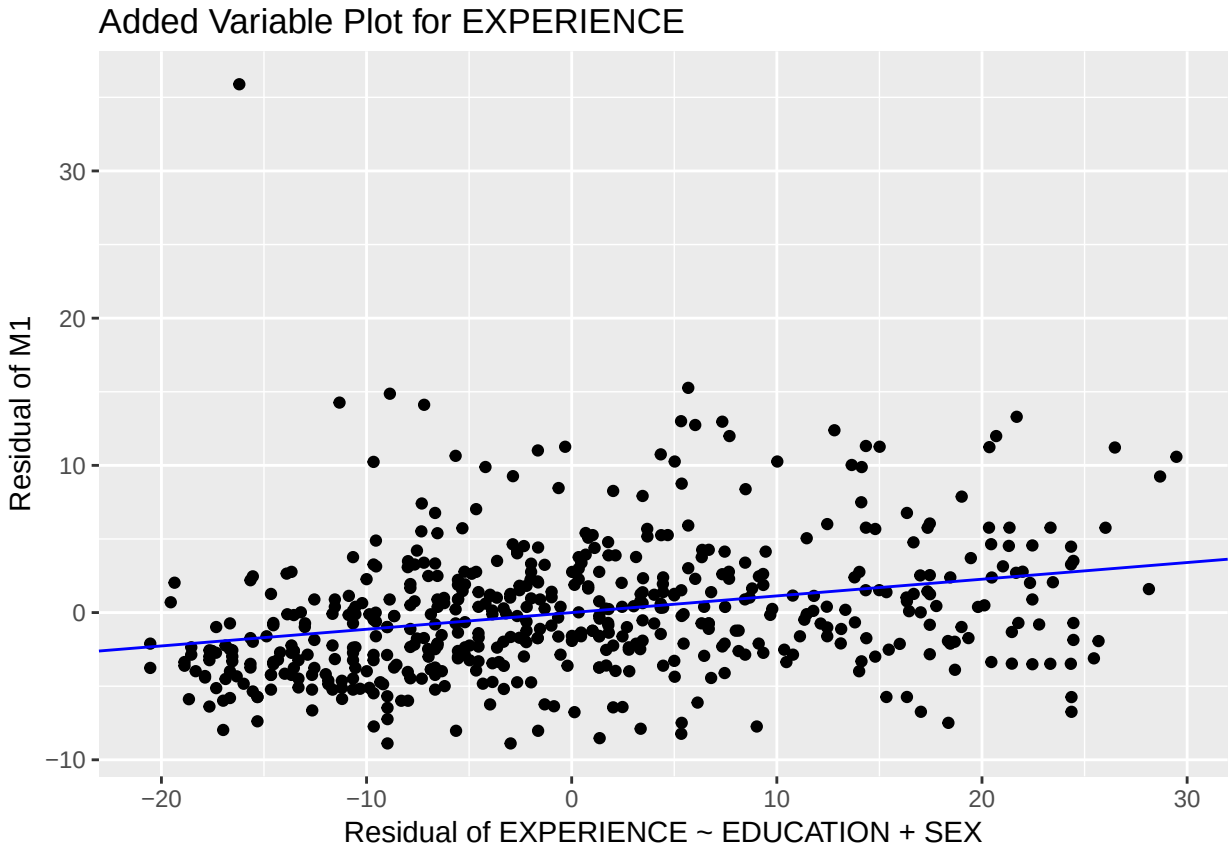
```
M2<-lm(WAGE~EDUCATION+SEX+SEX:EDUCATION, data=wages)

p <- ggplot(aes(x=EDUCATION, y=WAGE),data=wages)
p + geom_jitter(data=wages, aes(color=SEX)) +
  scale_color_manual(values=c("blue","pink")) +
  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

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

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.

```
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
```

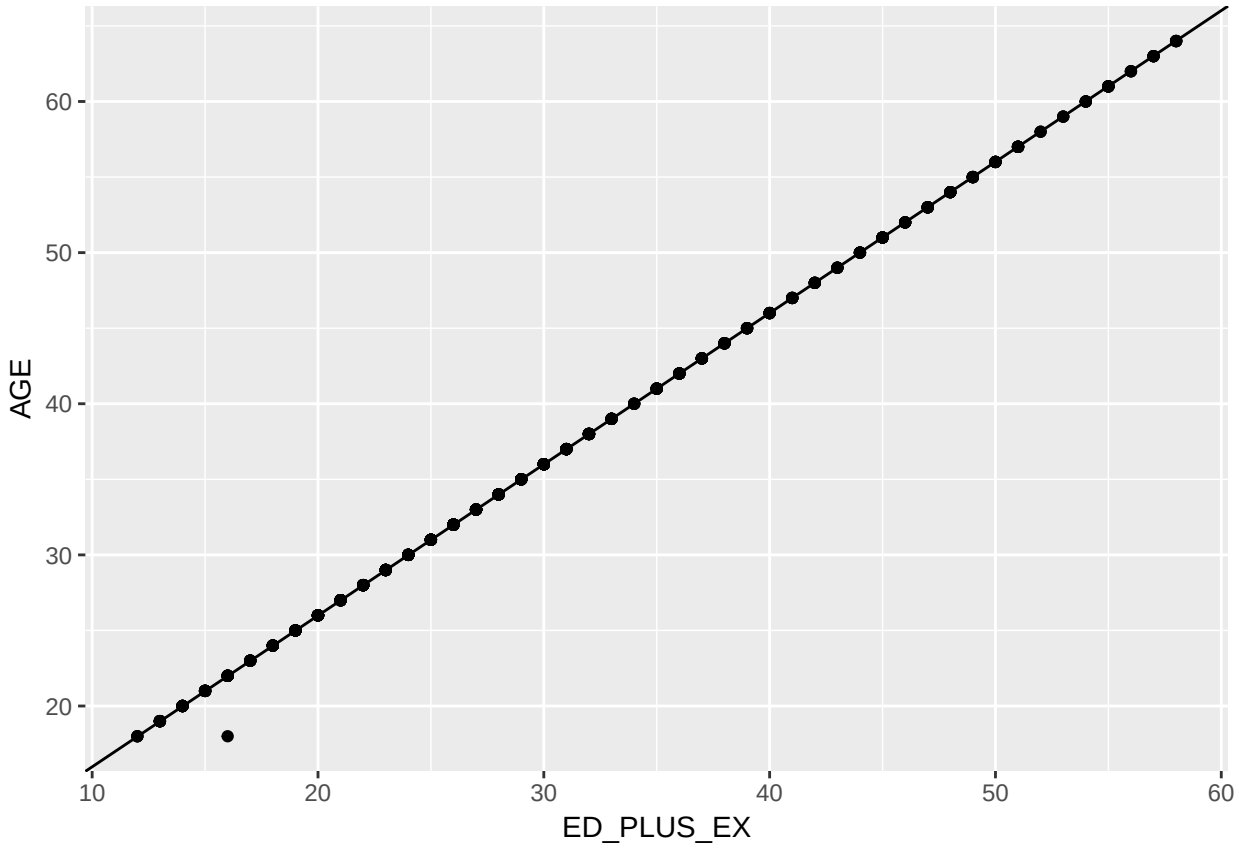
As we can see from the added variable plot adding *Age* to our model is not very helpful because a line might describe the data well if the outlier did not exist but the slope would infinite. In this case we see that the fitted line does not in fact describe the data well because of the outlier. The 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.

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

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

```
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.

```
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.

Quickly Reaffirming Assumptions Made Before Power Transformations

The first thing we tried to justify in our study is that there was non-significant interaction between Education and Sex in M2. This may have not been the case had we used log(Wage) as our response. Here we have made a model which models the log(Wage) by Education and Sex and the interaction between Sex and Education and find that the interaction term is not really significant. This means that in our future models which use log(Wage) as the response we don't have to include this interaction term because the 95% confindence interval on the interaction term contains zero (-0.000732, 0.062341). Visually we can see this in a graph similar to

shown below; the slopes of the regression lines for men and women are about the same.

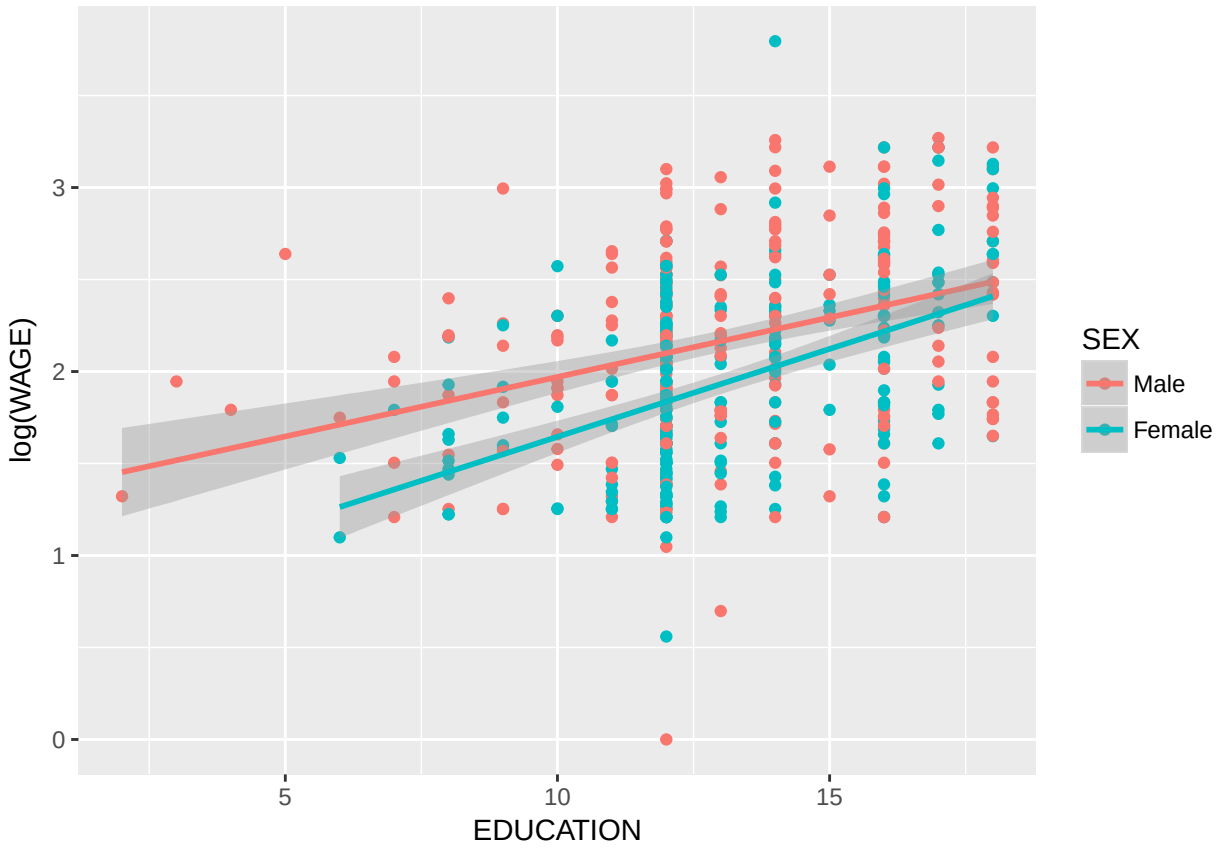
```
M2_2 <- lm( log(WAGE) ~ EDUCATION + SEX + SEX:EDUCATION, data = wages)
summary(M2_2)
```

```
##
## Call:
## lm(formula = log(WAGE) ~ EDUCATION + SEX + SEX:EDUCATION, data = wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.09971 -0.36313  0.03421  0.33156  1.76830
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.32352    0.13422   9.861  < 2e-16 ***
## EDUCATION            0.06468    0.01009   6.411 3.19e-10 ***
## SEXFemale           -0.63315    0.21303  -2.972  0.00309 **
## EDUCATION:SEXFemale  0.03080    0.01605   1.919  0.05554 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4738 on 530 degrees of freedom
## Multiple R-squared:  0.1984, Adjusted R-squared:  0.1938
## F-statistic: 43.72 on 3 and 530 DF,  p-value: < 2.2e-16
```

```
confint(M2_2)
```

```
##                             2.5 %      97.5 %
## (Intercept)          1.0598493402  1.58718228
## EDUCATION            0.0448638894  0.08450144
## SEXFemale           -1.0516463316 -0.21465386
## EDUCATION:SEXFemale -0.0007320259  0.06234104
```

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

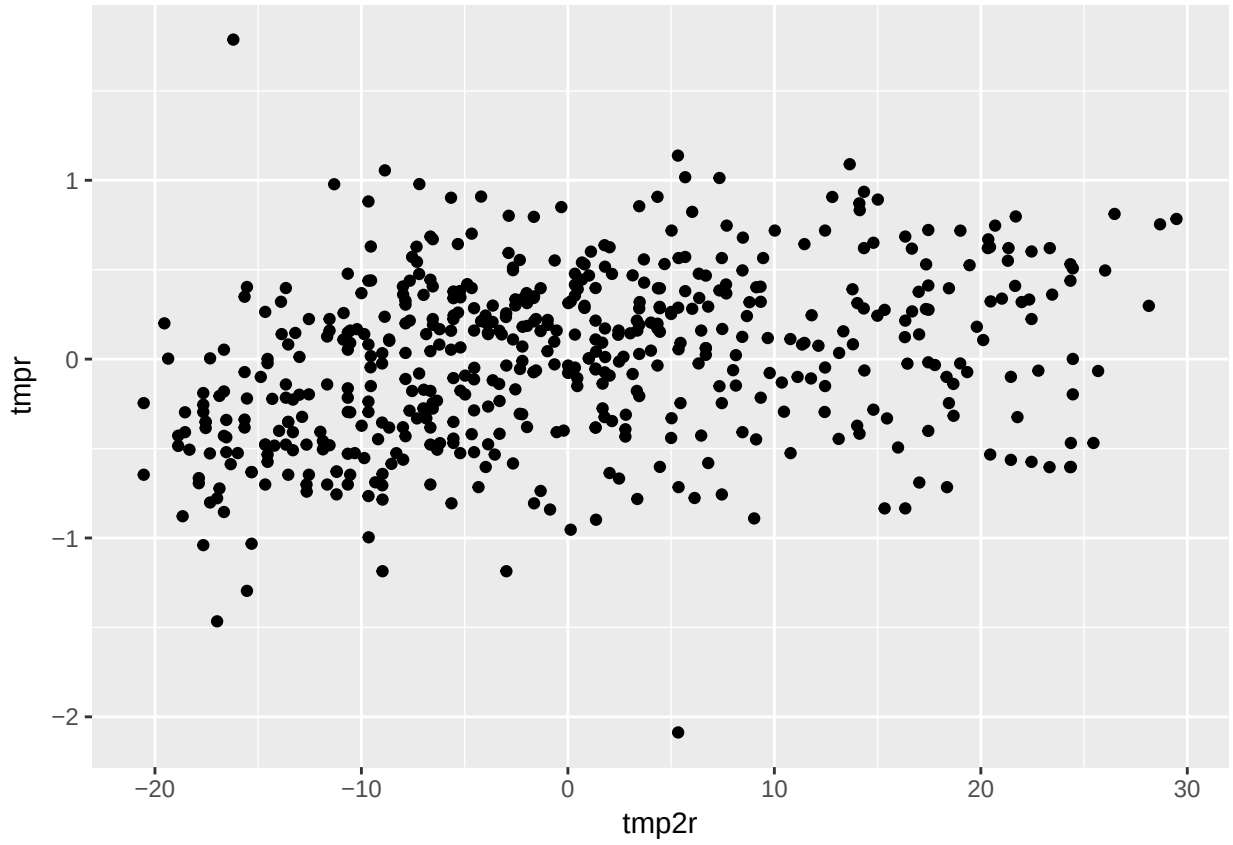


The other procedure we conducted was the choice to add Experience to our model. Here we will show that even when using the log(Wage) as our response we are justified in adding it to our model.

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

```
##
## Call:
## lm(formula = WAGE ~ EDUCATION + SEX + SEX:EDUCATION, data = wages)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.685 -3.048 -0.624  2.153 35.786
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.10457    1.31366   0.841   0.4008
## EDUCATION            0.68315    0.09874   6.918 1.32e-11 ***
## SEXFemale           -4.37045    2.08506  -2.096   0.0365 *
## EDUCATION:SEXFemale  0.17253    0.15712   1.098   0.2727
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.638 on 530 degrees of freedom
## Multiple R-squared:  0.1902, Adjusted R-squared:  0.1856
## F-statistic:  41.5 on 3 and 530 DF,  p-value: < 2.2e-16
```

```
tmp <- lm( log(WAGE) ~ EDUCATION + SEX, data = wages)
wages$tmpr <- resid(tmp)

tmp2 <- lm( EXPERIENCE ~ EDUCATION + SEX, data = wages)
wages$tmp2r <- resid(tmp2)

p <- ggplot( wages, aes ( x = tmp2r, y = tmpr) )
p + geom_point()
```




```
tmp3 <- lm( tmpr ~ tmp2r, data = wages)
summary(tmp3)
```

```
##
## Call:
## lm(formula = tmpr ~ tmp2r, data = wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15504 -0.30764  0.00708  0.30955  1.99176
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.534e-17  1.954e-02    0.00        1
## tmp2r       1.267e-02  1.694e-03    7.48 3.09e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.4514 on 532 degrees of freedom
## Multiple R-squared:  0.09517,    Adjusted R-squared:  0.09347
## F-statistic: 55.95 on 1 and 532 DF,  p-value: 3.086e-13
```

The figure above is an added variable plot for Experience added to M1 where the response for M1 is log(Wage) instead of Wage. We find that the remaining variation explained is 9.517% which is an increase from roughly 8% from our findings in our initial added variable plot.

Model 4: log(WAGE) ~ EDUCATION + SEX + sqrt(EXPERIENCE+1)

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

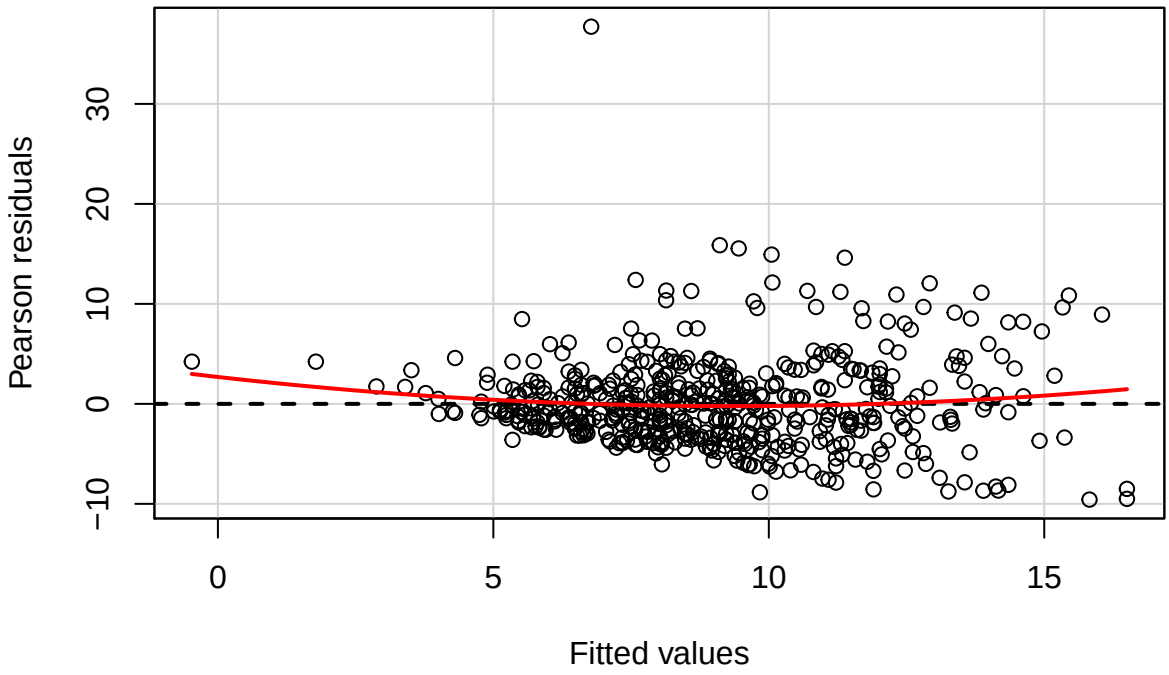
```
##
## Call:
## lm(formula = log(WAGE) ~ tranEXPERIENCE + EDUCATION + SEX, data = wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.18596 -0.29069  0.01133  0.30256  2.09292
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.426097   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.

While testing our models we found that in M4 we have finally solved our problem of non-constant variance (heteroscedasticity) which violates an assumption of linear models. We see this the plots of residuals versus fitted values. In residuals versus fitted values plot for model 3 we see an example of a right-opening megaphone:

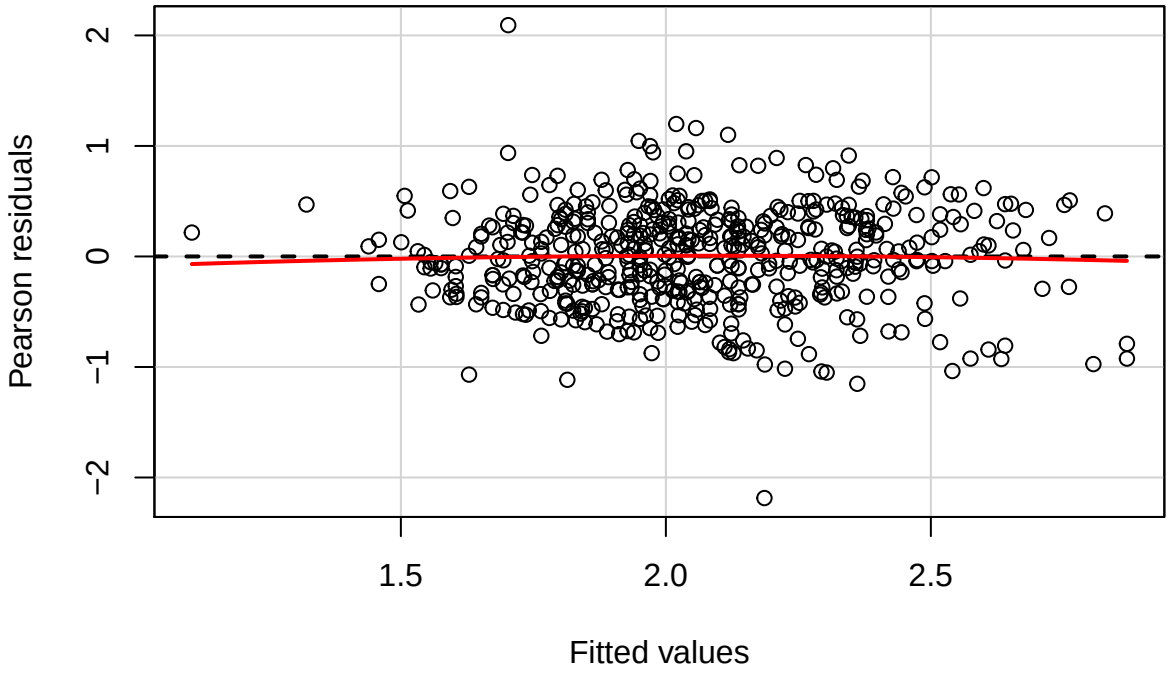
```
residualPlots(M3,terms = ~ 1)
```




```
##             Test stat Pr(>|t|)
## Tukey test      1.621    0.105
```

However, in model 4 we find there is constant variance which means our model is valid.

```
residualPlots(M4,terms = ~ 1)
```




```
##             Test stat Pr(>|t|)
## Tukey test     -0.436    0.663
```

Mincer Earnings Function

At this point our model looks very similar to a prominent econometric model called the Mincer Earnings Function. The model is defined roughly as follows in [1].

$$MEE : E(Wage|Education, Sex, Experience) = log(w_0) + aEducation + bExperience - cExperience^2 + \text{Other Variables} \tag{5}$$

In this case w_0 is the average wage a person in the reference group would expect to make. The coefficient a is interpretted as the average rate of return to schooling assuming workers are of equal ability. The coefficients b and c estimate the growth in earnings from an additional year of experience and are only greater than zero if it is assumed a worker invests in on-the-job-training. In short, since this isn't an economics paper, the squared term for experience comes from the idea that real post school investment in human capital decreases (assumed to be linearly decreasing) over the lifespan of the worker. This and the derivation of the Mincer Earnings Function is further described in [2].

For our purposes we will add Sex to our model.

```
MEF<-lm(log(WAGE)~ EDUCATION + EXPERIENCE + I(EXPERIENCE^2) + 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
```

```
summary(MEF)
```

```
##
## Call:
## lm(formula = log(WAGE) ~ EDUCATION + EXPERIENCE + I(EXPERIENCE^2) +
##     SEX, data = wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.24980 -0.29235  0.01609  0.29184  2.13816
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)      0.6007445  0.1194927   5.027 6.81e-07 ***
## EDUCATION        0.0912936  0.0080049  11.405  < 2e-16 ***
## EXPERIENCE       0.0360522  0.0054352   6.633 8.14e-11 ***
## I(EXPERIENCE^2) -0.0005412  0.0001197  -4.520 7.64e-06 ***
## SEXFemale       -0.2570355  0.0387066  -6.641 7.77e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4442 on 529 degrees of freedom
## Multiple R-squared:  0.2968, Adjusted R-squared:  0.2915
## F-statistic: 55.81 on 4 and 529 DF,  p-value: < 2.2e-16
```

What we find is that our R^2 hardly increases compared to our M4, the power transformed model which uses the square root of experience, from 0.2894 to 0.2968. What we also find is that the coefficient on the experience squared term is significant and explains about 3% of the remaining variance compared to just using experience alone. The main difference between M4 and the MEF is the interpretation of the effect of experience on wages. If we look at a graph the change in hourly wage due to experience versus the amount of years of expereince the shape of the these graphs differs between these models as shown in the figure below. What ends up happening is that the rate of return to schooling remains about the same at 0.097740 in M4 and 0.0912936 in MEF and the wage gap remains almost exactly the same at about -0.25.

Since our models are very similar and the justification for the Mincer Earnings Function founded by us we will use M4 as the base model for expansion in the rest of this study.

Model 5: log(WAGE) ~ EDUCATION + SEX + OCCUPATION + sqrt(EXPERIENCE+1)

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. The justification for adding occupation into our model is that even after you account for all other variables in our model, a person's occupation may influence their wage. One way to think about this is that two jobs may require the same levels of education and experience but may incur different levels of risk of physical harm on the worker. Economic theory would suggest that the person who incurs more risk should make proportionally more money. This is just an example of how the type of occupation one takes can influence wage while controlling for our other variables. In M5 we explore the main effects model.

```
##
## Call:
## lm(formula = log(WAGE) ~ EDUCATION + tranEXPERIENCE + SEX + OCCUPATION,
##     data = wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34865 -0.29338  0.00324  0.30504  1.91604
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.736920   0.146396   5.034 6.62e-07 ***
## EDUCATION                0.076052   0.009793   7.766 4.28e-14 ***
## tranEXPERIENCE           0.109849   0.013648   8.048 5.64e-15 ***
## SEXFemale               -0.227408   0.041741  -5.448 7.84e-08 ***
## OCCUPATIONManagement     0.149861   0.075271   1.991  0.04700 *
## OCCUPATIONSales         -0.201693   0.081177  -2.485  0.01328 *
## OCCUPATIONClerical      -0.050531   0.063321  -0.798  0.42522
```

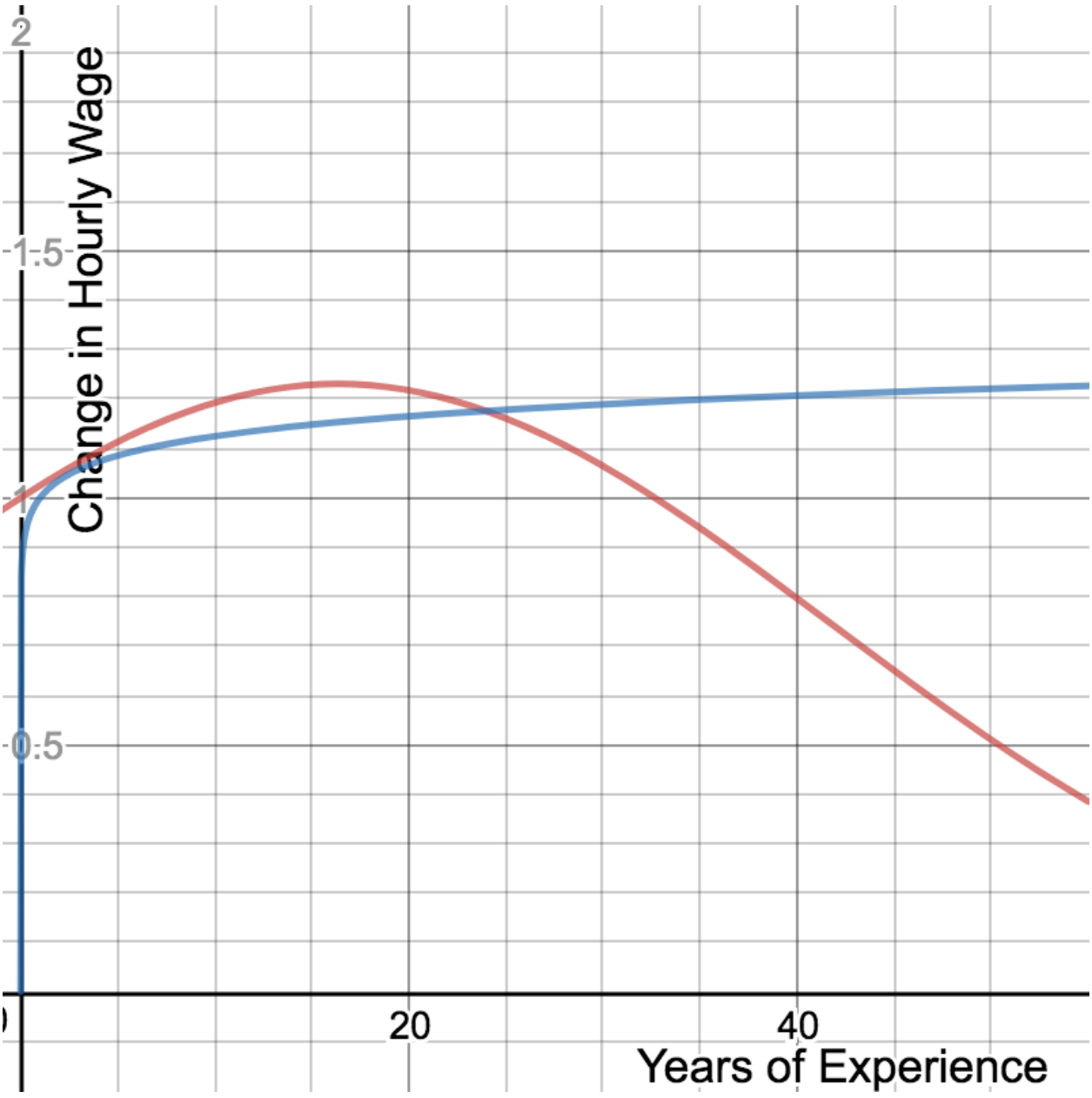


Figure 4: picture

```
## OCCUPATIONService      -0.200988   0.061294  -3.279  0.00111 **
## OCCUPATIONProfessional  0.140695   0.069451   2.026  0.04329 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4338 on 525 degrees of freedom
## Multiple R-squared:  0.3346, Adjusted R-squared:  0.3245
## F-statistic:    33 on 8 and 525 DF,  p-value: < 2.2e-16
```

Under this model we find that women make -22.7408% less than men while controlling for all of these factors. We also note that our R^2 increased to 0.3346 and our RSE slightly decreased to 0.4338.

Model 6: log(WAGE) ~ EDUCATION + SEX * OCCUPATION + sqrt(EXPERIENCE+1)

```
##
## Call:
## lm(formula = log(WAGE) ~ EDUCATION + tranEXPERIENCE + SEX * OCCUPATION,
##     data = wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.31132 -0.27486  0.02628  0.28322  1.85385
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      0.781732   0.146889   5.322 1.53e-07 ***
## EDUCATION                        0.075839   0.009755   7.774 4.10e-14 ***
## tranEXPERIENCE                   0.109303   0.013648   8.009 7.64e-15 ***
## SEXFemale                       -0.436061   0.087486  -4.984 8.48e-07 ***
## OCCUPATIONManagement             0.073001   0.089412   0.816  0.41461
## OCCUPATIONSales                 -0.123597   0.103332  -1.196  0.23219
## OCCUPATIONClerical              -0.219403   0.102572  -2.139  0.03290 *
## OCCUPATIONService               -0.251238   0.083113  -3.023  0.00263 **
## OCCUPATIONProfessional           0.040795   0.081992   0.498  0.61901
## SEXFemale:OCCUPATIONManagement   0.306638   0.148265   2.068  0.03912 *
## SEXFemale:OCCUPATIONSales       -0.054784   0.165409  -0.331  0.74062
## SEXFemale:OCCUPATIONClerical     0.373323   0.137398   2.717  0.00681 **
## SEXFemale:OCCUPATIONService      0.226000   0.130008   1.738  0.08274 .
## SEXFemale:OCCUPATIONProfessional 0.330856   0.121175   2.730  0.00654 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4299 on 520 degrees of freedom
## Multiple R-squared:  0.3526, Adjusted R-squared:  0.3364
## F-statistic: 21.79 on 13 and 520 DF,  p-value: < 2.2e-16
```

In M6 we explore the possiblity that the effect of you occupation on your wage is different for men and women. We can try to see this in a couple ways. If we look at the analysis of variance table for this model we find that the the interaction term between Sex and Occupation is significant.

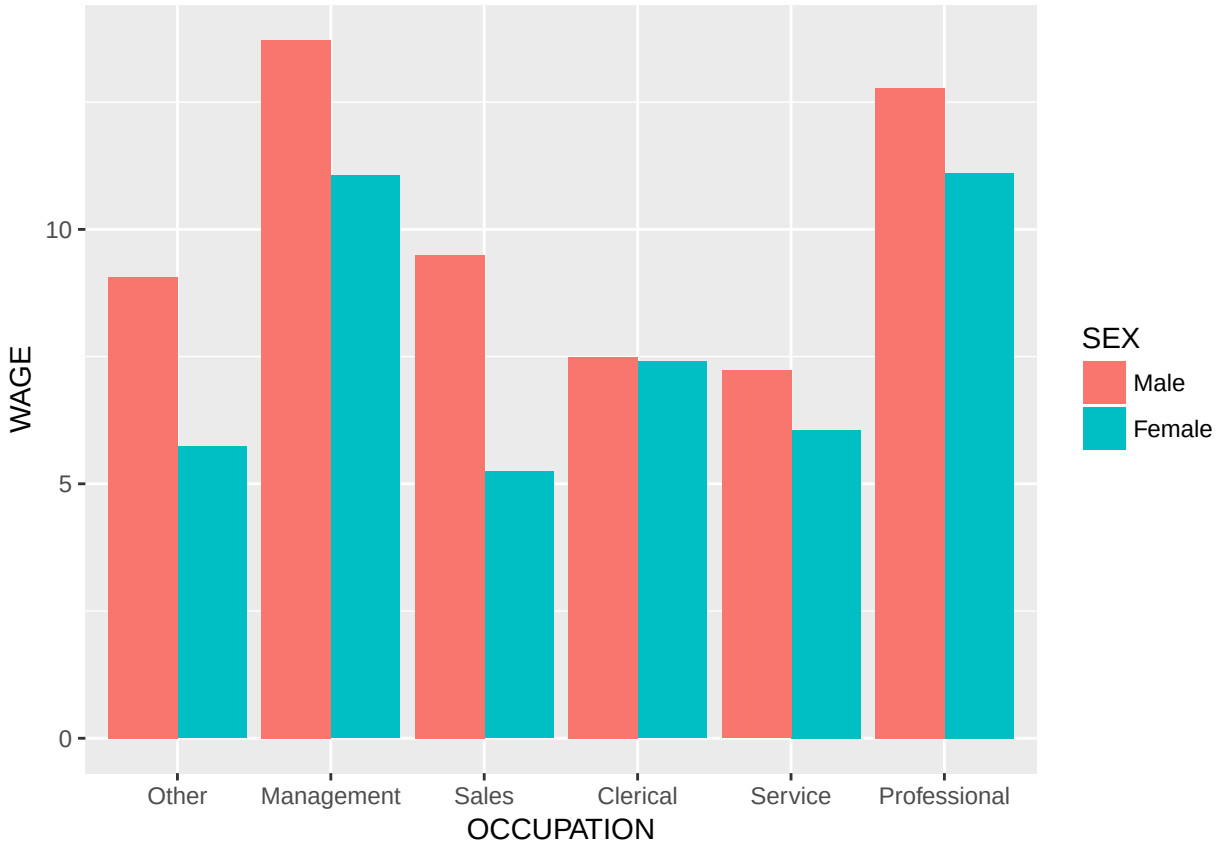
```
anova(M6)
```

```
## Analysis of Variance Table
```

```
##
## Response: log(WAGE)
##                 Df Sum Sq Mean Sq  F value    Pr(>F)
## EDUCATION        1 21.481 21.4807 116.2308 < 2.2e-16 ***
## tranEXPERIENCE   1 12.713 12.7133  68.7912 9.436e-16 ***
## SEX              1  8.765  8.7654  47.4291 1.651e-11 ***
## OCCUPATION       5  6.711  1.3421   7.2620 1.373e-06 ***
## SEX:OCCUPATION   5  2.675  0.5351   2.8953    0.0137 *
## Residuals      520 96.102  0.1848
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Another way to try to visualize this is by looking at the mean wages for men and women without controlling for Education and Experience.

```
ggplot(data=wages, aes(x=OCCUPATION, y=WAGE, fill=SEX)) +
geom_bar(stat="summary", fun.y = "mean", position=position_dodge())
```



When controlling for education and experience we find the following wage gaps between men and women. Negative percentages should be interpretted as the percentage amount of wage females make less than men while controlling for education and experience.

```
M6_summary <- summary(M6)
occs <- c(levels(wages$OCCUPATION))
occ_gaps <- percent(c(coef(M6_summary)[4],coef(M6_summary)[4] + coef(M6_summary)[10], coef(M6_summary)[4
occ_men <- c(126,34,21,21,34,53)
occ_women <- c(30,21,17,76,49,52)
occ_gap_df <- data.frame(occs,occ_gaps,occ_men,occ_women)
```

```
names(occ_gap_df) <- c("Occupation", "Male/Female Wage Gap", "# Men", "# Women")
occ_gap_df
```

```
##     Occupation Male/Female Wage Gap # Men # Women
## 1        Other               -43.6%   126      30
## 2   Management               -12.9%    34      21
## 3        Sales               -49.1%    21      17
## 4     Clerical                -6.3%    21      76
## 5      Service               -21.0%    34      49
## 6 Professional               -10.5%    53      52
```

```
kable(occ_gap_df, caption = "Wage Gaps for Females Compared to Males")
```

Table 1: Wage Gaps for Females Compared to Males

Occupation	Male/Female Wage Gap	# Men	# Women
Other	-43.6%	126	30
Management	-12.9%	34	21
Sales	-49.1%	21	17
Clerical	-6.3%	21	76
Service	-21.0%	34	49
Professional	-10.5%	53	52

The occupation specific wage gaps are interesting. As we would expect the wage gap in occupation Other is very large since it is as if we are not controlling for occupation at all. In management, professional and clerical jobs we find that the wage gap between men and women drastically decreases. In the service industry we find that the wage gap is at about the same level as before we controlled for occupation. However, in the sales occupation we found that the wage gap between men and women is massive, women make almost 50% less than men of the same level of education and experience.

Looking back at the results from the analysis of variance we found that overall the interaction between Sex and Occupation is signifcant enough to include in our model. When we looked at each of the significances and confidence intervals for each occupation-specific wage gap we found that they depend on the significance of the reference variable in our model to some extent. This is why we included the number of men and women in each occupation in our table so you can get a sense of what these wage gap numbers are calculated from.

Finally, to conclude model 6, we found that our R^2 increased to 0.3526 and our RSE decreased to 0.4299.

Model 7: log(WAGE) ~ EDUCATION + SEX + sqrt(EXPERIENCE+1) + OCCUPATION + UNION

In our final model we would like to account for union status. This is a main effects model which means we are going against model 6 because if we were to try to estimate the occupation- specific wage gap accounting for union status we would not have enough data in each occupation to make a reasonable claim. In the table below we can see that some occupations hardly have any unionized memebers. When we break it down by Sex to estimate the occupation-specific wage gap this problem only gets worse.

```
occ_un_df <-data.frame(wages %>%
  group_by(OCCUPATION, UNION) %>%
  summarise(no_rows = length(UNION)))
names(occ_un_df) <- c("Occupation", "Union Status", "Number of People")
kable(occ_un_df)
```

Occupation	Union Status	Number of People
Other	Non-Unionized	112
Other	Unionized	44
Management	Non-Unionized	52
Management	Unionized	3
Sales	Non-Unionized	37
Sales	Unionized	1
Clerical	Non-Unionized	89
Clerical	Unionized	8
Service	Non-Unionized	66
Service	Unionized	17
Professional	Non-Unionized	82
Professional	Unionized	23

```
M7 <- lm( log(WAGE) ~ EDUCATION + tranEXPERIENCE + SEX + OCCUPATION + UNION, data = wages)
summary(M7)
```

```
##
## Call:
## lm(formula = log(WAGE) ~ EDUCATION + tranEXPERIENCE + SEX + OCCUPATION +
##     UNION, data = wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.33044 -0.29265 -0.00479  0.30590  1.88917
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.747582   0.144225   5.183 3.11e-07 ***
## EDUCATION                0.072838   0.009678   7.526 2.29e-13 ***
## tranEXPERIENCE           0.100973   0.013614   7.417 4.87e-13 ***
## SEXFemale               -0.207725   0.041390  -5.019 7.14e-07 ***
## OCCUPATIONManagement     0.203929   0.075287   2.709  0.00698 **
## OCCUPATIONSales         -0.147715   0.081019  -1.823  0.06884 .
## OCCUPATIONClerical      -0.016381   0.062917  -0.260  0.79469
## OCCUPATIONService       -0.190808   0.060426  -3.158  0.00168 **
## OCCUPATIONProfessional   0.158982   0.068553   2.319  0.02077 *
## UNIONUnionized           0.208963   0.050538   4.135 4.14e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4273 on 524 degrees of freedom
## Multiple R-squared:  0.3556, Adjusted R-squared:  0.3446
## F-statistic: 32.13 on 9 and 524 DF,  p-value: < 2.2e-16
```

In model 7 we find that women on average earn 20.7% less than men when controlling for all the variables in model 7. This claim is significant because the confidence interval for the coefficient on Sex does not contain zero. Also, we find that unionized workers make on average 20.8% more than non-unionized workers when controlling for all other variables. This claim is also significant as supported by the confidence interval of the union coefficient not containing zero. Something to note about the "Union Gap" is that it is overestimated because it only considers your raw wage and not your wage net of union dues. Lastly, our R^2 increased to 0.3556 compared to model 5 and our RSE decreased to 0.4273.

Conclusion

To wrap up our investigation of this recently culturally relevant issue we have found evidence to support the existence of a male/female wage gap. In all of our models this wage gap has been both present and significant. We should also note that as we added more potential determinants of wage to our models, the wage gap seemed to decrease. Initially, the wage gap between men and women was around 23% not acccounting for any other variables. When we accounted for level of education, years of experience, occupation, and union status this number decreased to 20%. However, the largest changes occured when we accounted for differences in wage gaps in each occupation. Looking at occupation-specific wage gaps the male/female wage gap decreased to about 6% in clerical occupations and about 10% in management and professional jobs. The "outlier" occupation in this sense was sales which had a very large male/female wage gap of almost 50%. At this moment we can't offer an explaination as to why the wage gap exists beyond the parameters we examined and leave that for future research and interpretation.

Caveats

Here are some of our final thoughts we had while doing this project:

- Our R^2 is fairly low and our RSE is pretty high in all of our models. Since the focus of our project is to try estimate the wage gap rather than predict wages we can accept errors of about 40% in our model.
- In the future we might want to use the entire CPS wage dataset rather than a random sample.
- One major area in which this paper is lacking is in the application of causal inference. When we add variables to our model we aren't considering the effects of any potential mediator variables like occupation.