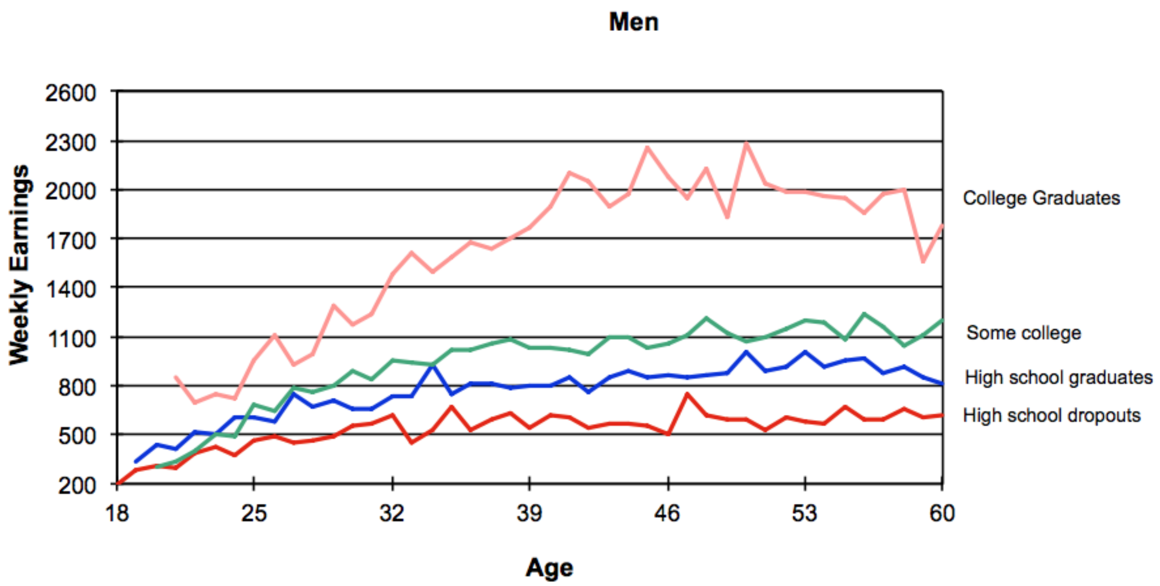
R Notebook

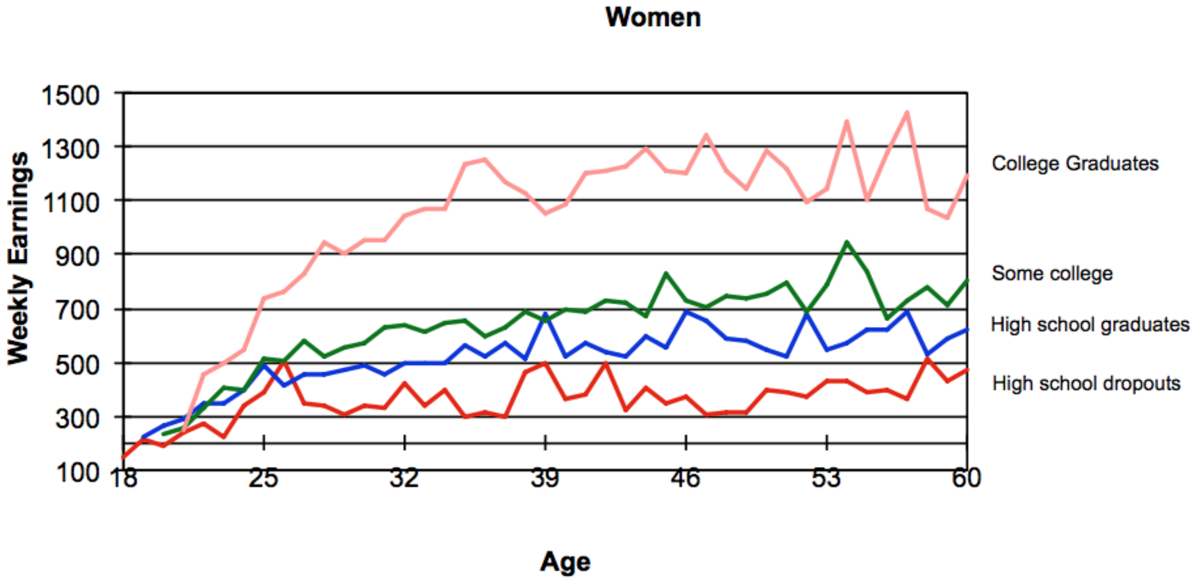
#Introduction In this paper we will be focusing around the impact of a worker's sex on their hourly wage. The dataset we are using for this project is a random sample of 534 people from the 1985 Current Population Survey (CPS) which contains cross-sectional data of potential determinants of wages. In 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

#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 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.

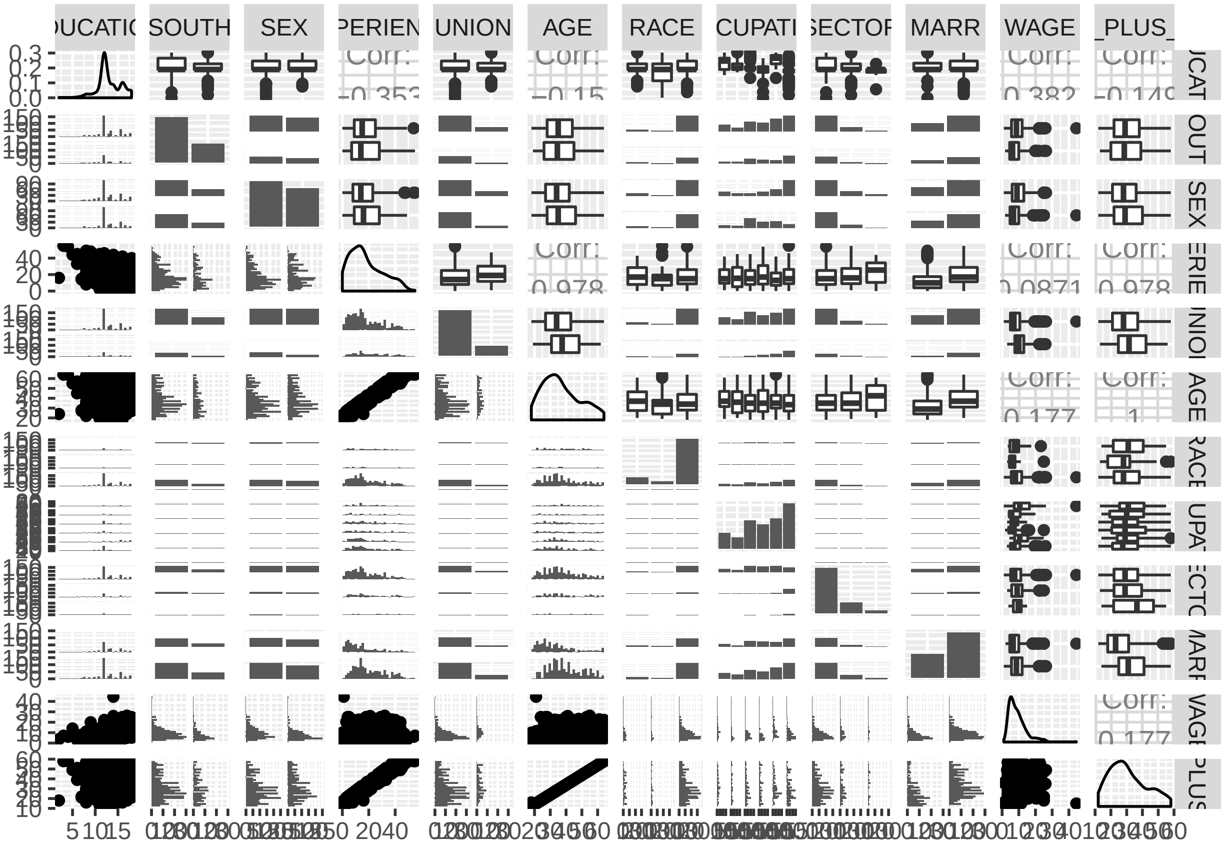




In this study we will not be looking numerically at the data shown above, instead we will use it qualitatively for some intuition and partial justification for adding education into our multiple linear regression model. Before just wildly throwing a term into our model we can look at the marginal plot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

#GGpairs plot We are going to use the ggpairs function in order to see if there is anything interesting in our data set

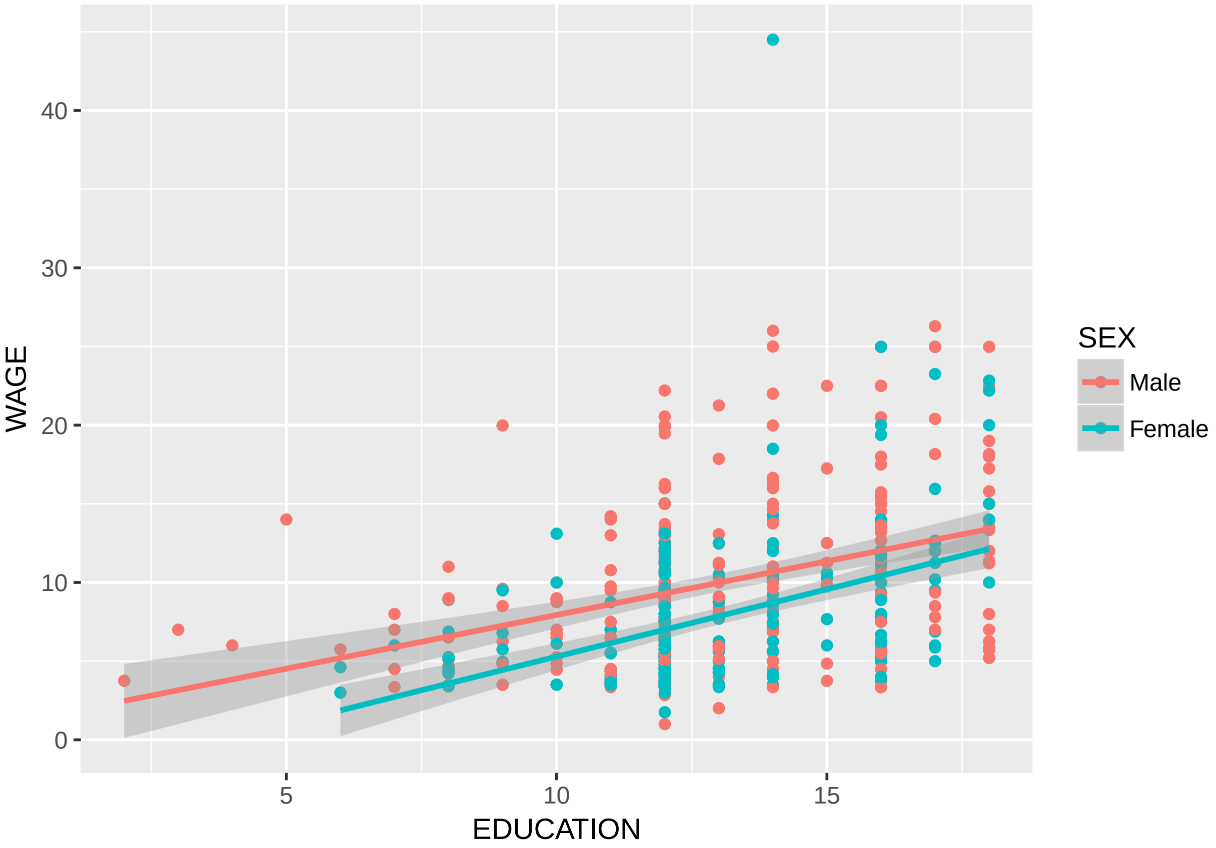
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `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`.
## `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`.
## `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`.
## `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`.
## `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`.
## `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`.
## `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`.
## `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. From the plot above we can see that Experience vs Age has a nice linear model and Wage vs Education also has a bit of a linear model. We are going explore more indepth into these models and using real life knowledge see if we are able to explain the reasoning for a wage gap between male and female.

#Model 1: WAGE~EDUCATION We first would like to explorer whether the wage gap does exist. In order to do so we first graphed the plot Education vs Wage and differentiated each point by the sex. We then created linear lines for Male and Female to see if there was anything interesting.

```
#We guess that education and sex will have to do something with wage

p = ggplot(aes(x=EDUCATION, y=WAGE),data=wages)
#p + geom_point(data=wages, aes(color=SEX)) #See that a lot of data is clumped together

p + geom_point(data=wages, aes(color=SEX)) + geom_smooth(method='lm', aes(color=SEX), data=wages)
```



```
#see that the slopes are very close to identical but have different intercept so there is wage gap
```

We are able to see that the slopes for male and female and quite similar but the intercept for them are different with male having a higher intercept than female. This tells us that male on average make more than females which tells us that there is indeed a wage gap.

```
mod = lm(WAGE~EDUCATION, data=wages)

mod1 = lm(WAGE~EDUCATION + SEX, data=wages)
summary(mod)
```

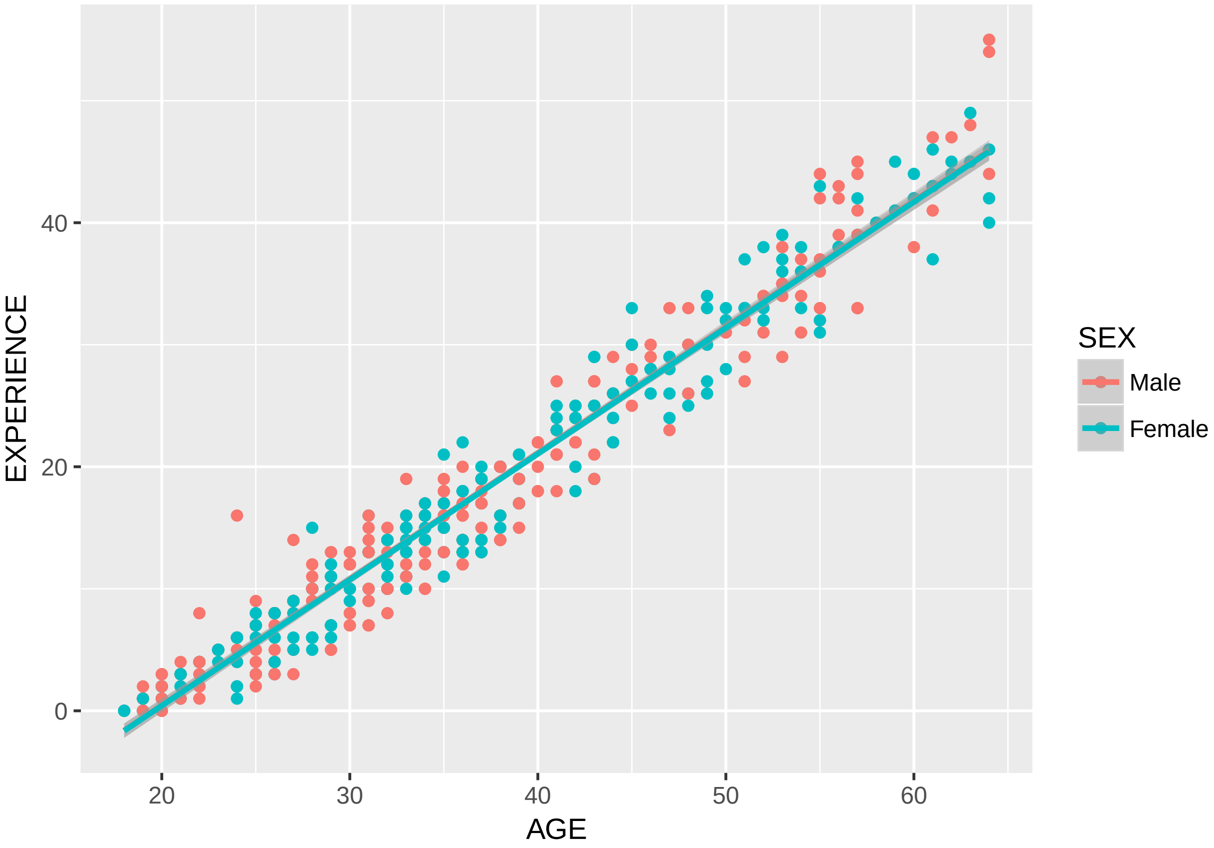
```
##
## Call:
## lm(formula = WAGE ~ EDUCATION, data = wages)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.911 -3.260 -0.760  2.240 34.740
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.74598    1.04545  -0.714    0.476
## EDUCATION    0.75046    0.07873   9.532   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.754 on 532 degrees of freedom
## Multiple R-squared:  0.1459, Adjusted R-squared:  0.1443
## F-statistic: 90.85 on 1 and 532 DF,  p-value: < 2.2e-16
```

```
summary(mod1) #see that we have low r^2 value and nothing is interesting lets try doing added variable
```

```
##
## 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
## 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
```

Comparing the a linear model with sex and without sex we can see that the linear model with education and sex had a slightly higher R^2 value and lower RSE. We added sex into the model as we know women and men have different levels of education as men are more likely to be in STEM fields which pay more compared to women. We chose to work with the linear model with sex and started to do added variable plots in order to find if anything else effected wages besides sex and education. #Model 2: Experience ~ Age As we saw from the ggpairs plot in the beginning Experience and Age had a high correlation. This is to be expected as Age is directly corelated with Experience. We still plotted this graph in order to make sure this is correct

```
p2 = ggplot(aes(x=AGE, y=EXPERIENCE), data=wages)
p2 + geom_point(aes(color=SEX), data=wages) + geom_smooth(method="lm", aes(color=SEX), data=wages)
```

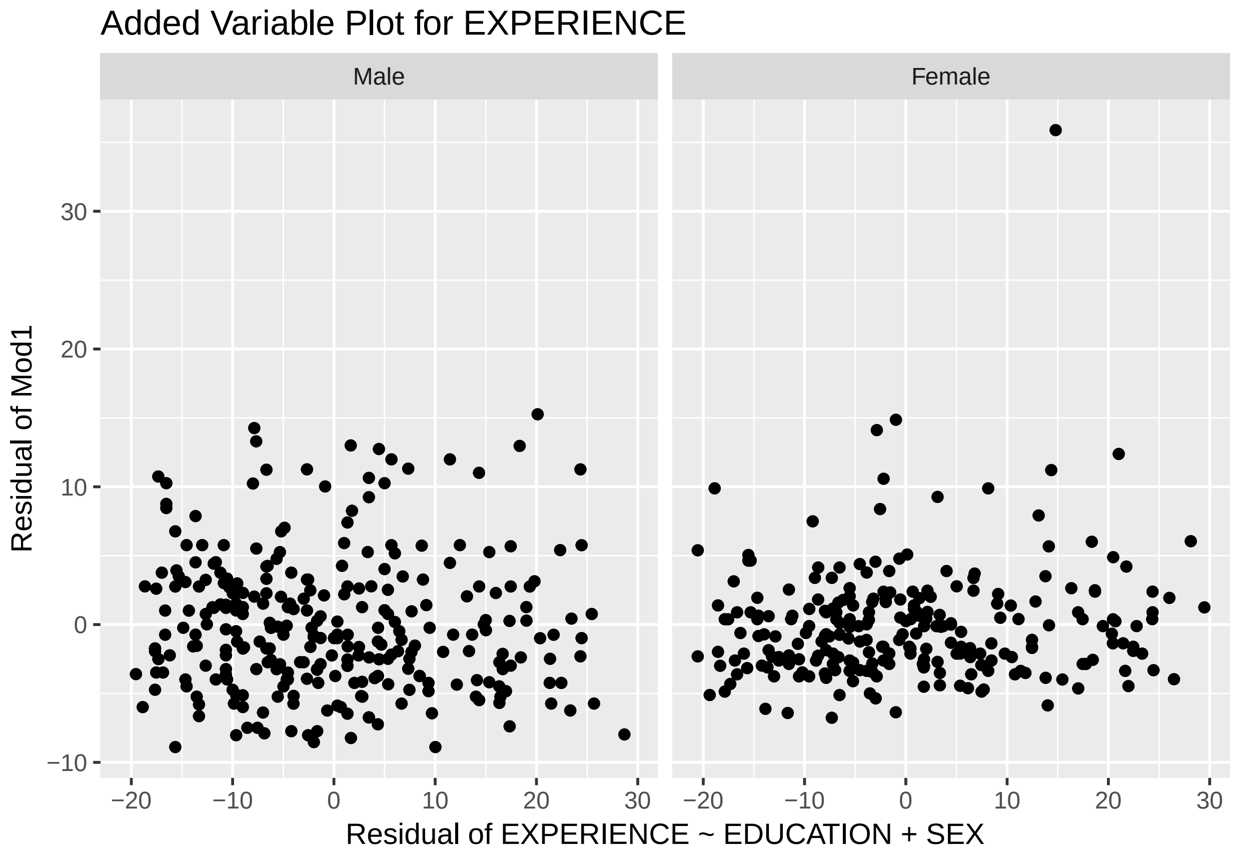


```
mod2 = lm(EXPERIENCE~AGE, data=wages)
```

As we see from the graph above male and female almost exactly the same line which allows us see that there is no sex descrimination when it comes to EXPERIENCE vs AGE. This also enforces our knowledge that EXPERIENCE is directly correlated to AGE.

#AVP 2: Added variable plot with Experience We create added variable plots for Experience and Age in order to see whether these two variables add anything new to our old model

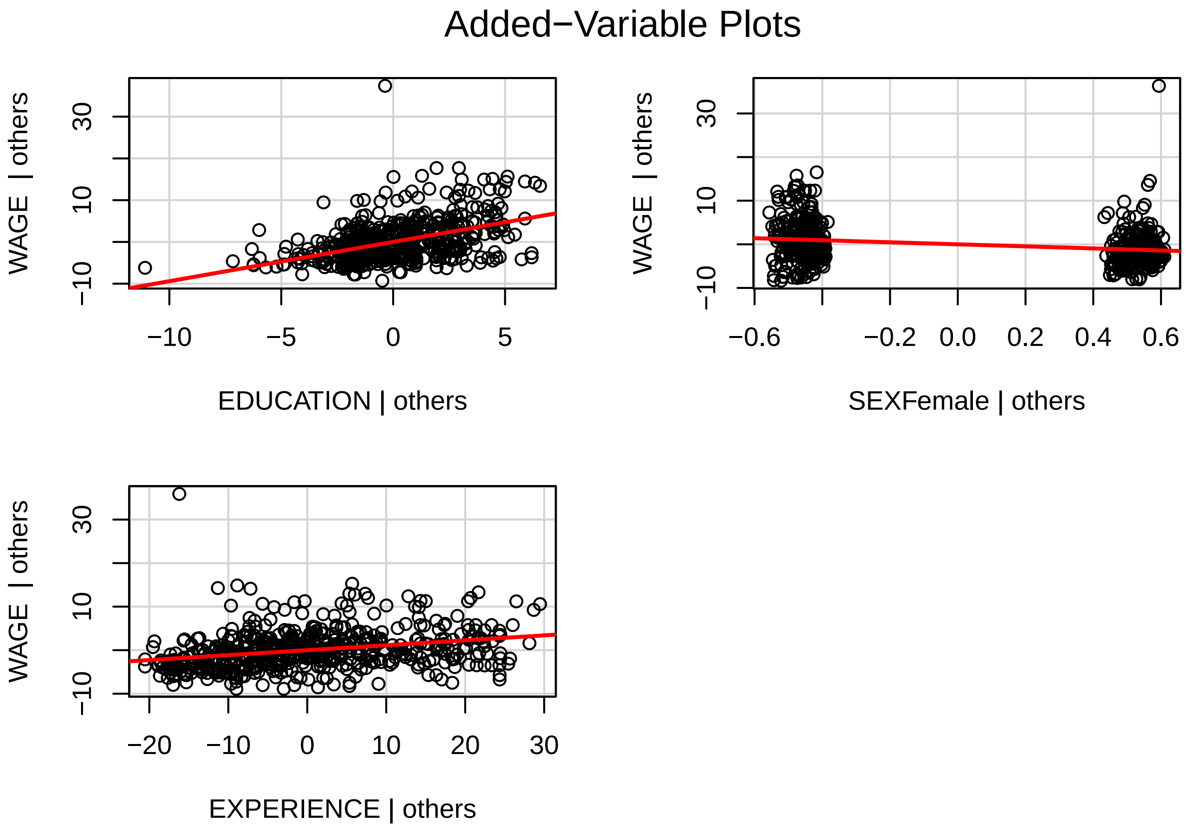
```
wages$resid1 = resid(mod1)

p2 = ggplot(aes(y=resid1), data=wages)
p2 + geom_point(aes(x=resid(lm(EXPERIENCE~EDUCATION + SEX, data=wages)))) + ggtitle("Added Variable Plo
```



```r
#has good linear model except for that one outlier
mod2 = lm(WAGE~EDUCATION + SEX + EXPERIENCE, data=wages)
```

We can see that the added variable plots have good linear models for both male and female. We create model2 by adding Experience to our existing model. Using avPlots function we are able to see more indepth how EDUCATION effects our other regressors.

```r
avPlots(mod2)
```



As we can see from the avPlots WAGE~EXPERIENCE and WAGE~EDUCATION both have nice linear lines which shows that adding EDUCATION into our plot is a good idea.

#Checking our new model results and comparing it with our old model

```
summary(mod2); summary(mod1)
```

```
##
## Call:
## lm(formula = WAGE ~ EDUCATION + SEX + EXPERIENCE, 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) -4.16687    1.18667  -3.511 0.000484 ***
## EDUCATION    0.94051    0.07886  11.926  < 2e-16 ***
## SEXFemale   -2.33763    0.38806  -6.024 3.19e-09 ***
## EXPERIENCE   0.11330    0.01671   6.781 3.19e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.454 on 530 degrees of freedom
## Multiple R-squared:  0.2532, Adjusted R-squared:  0.2489
## F-statistic: 59.88 on 3 and 530 DF,  p-value: < 2.2e-16
```

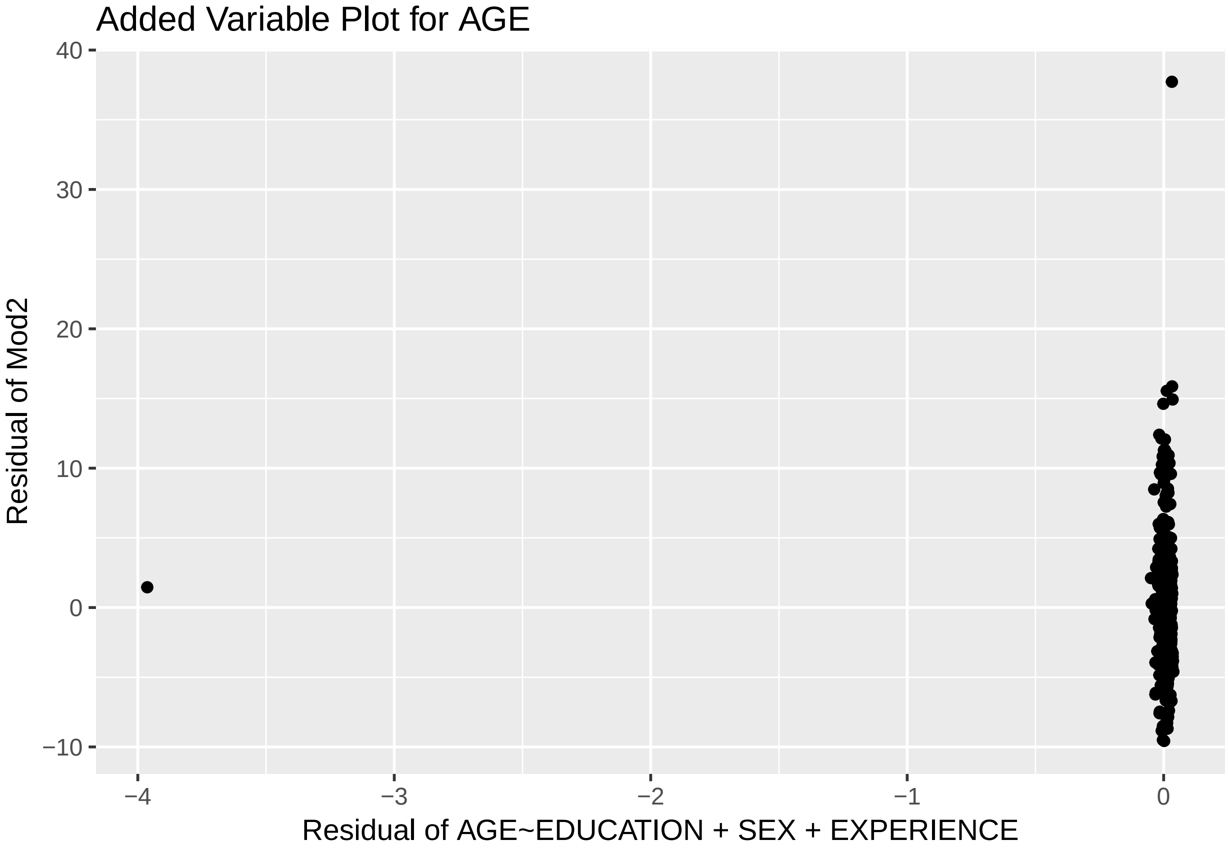
```
##
```

```
## 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
## 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 can see from the result the R^2 value increased from model1 to model2 and the Residual Standard Error(RSE) decreased slightly. Still the new model explains nothing interesting about the effects on wage.

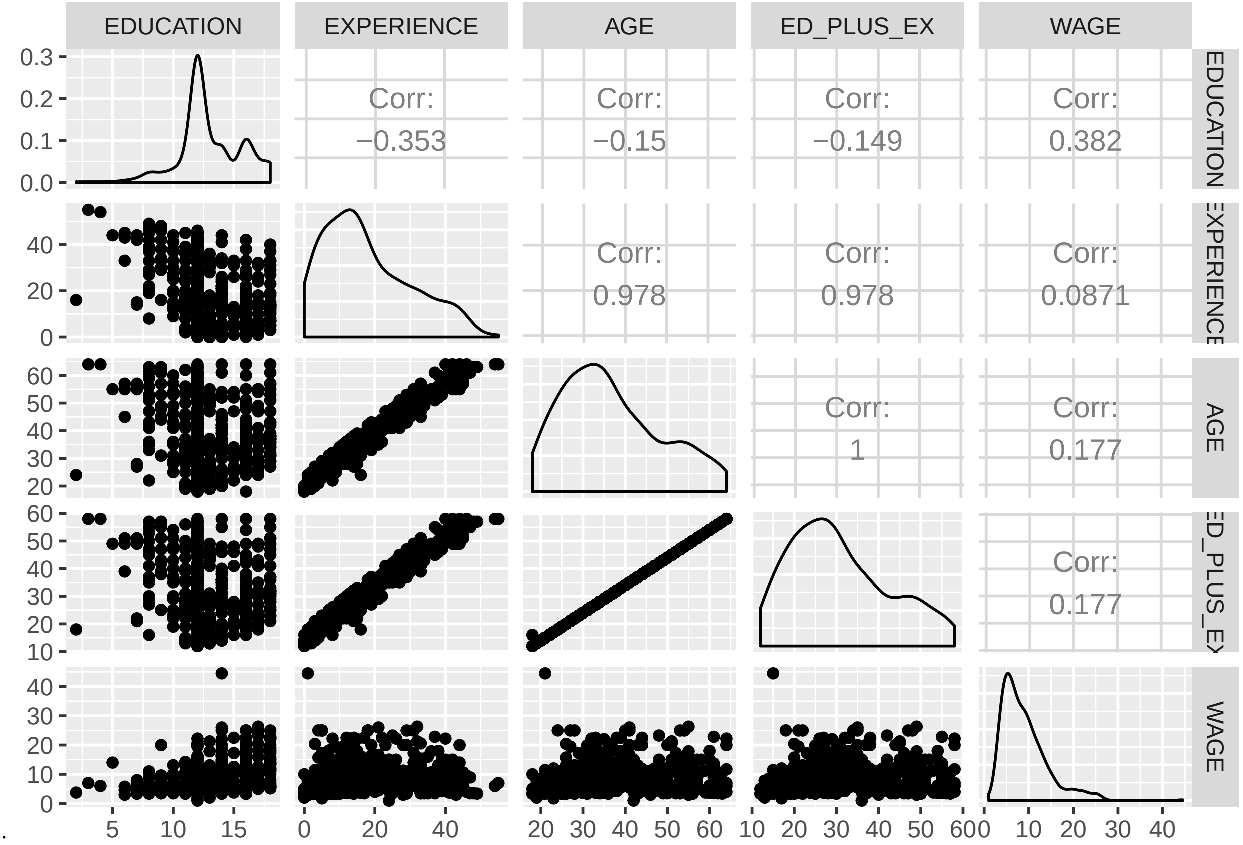
#Added variable plot2 We create a new added variable plot and see if age helps explain out dataset a bit better

```
wages$resid2 = resid(mod2)
p3 = ggplot(aes(y=resid2), data=wages)
p3 + geom_point(aes(x=resid(lm(AGE~EDUCATION + SEX + EXPERIENCE, data=wages)))) + ggtitle('Added Variabl
```



We see that all of the data is all clumped up together with two extreme outliers. We will see if transforming the data would help fix this issue or if there is anything else that can tell us if we should be using age in our model.

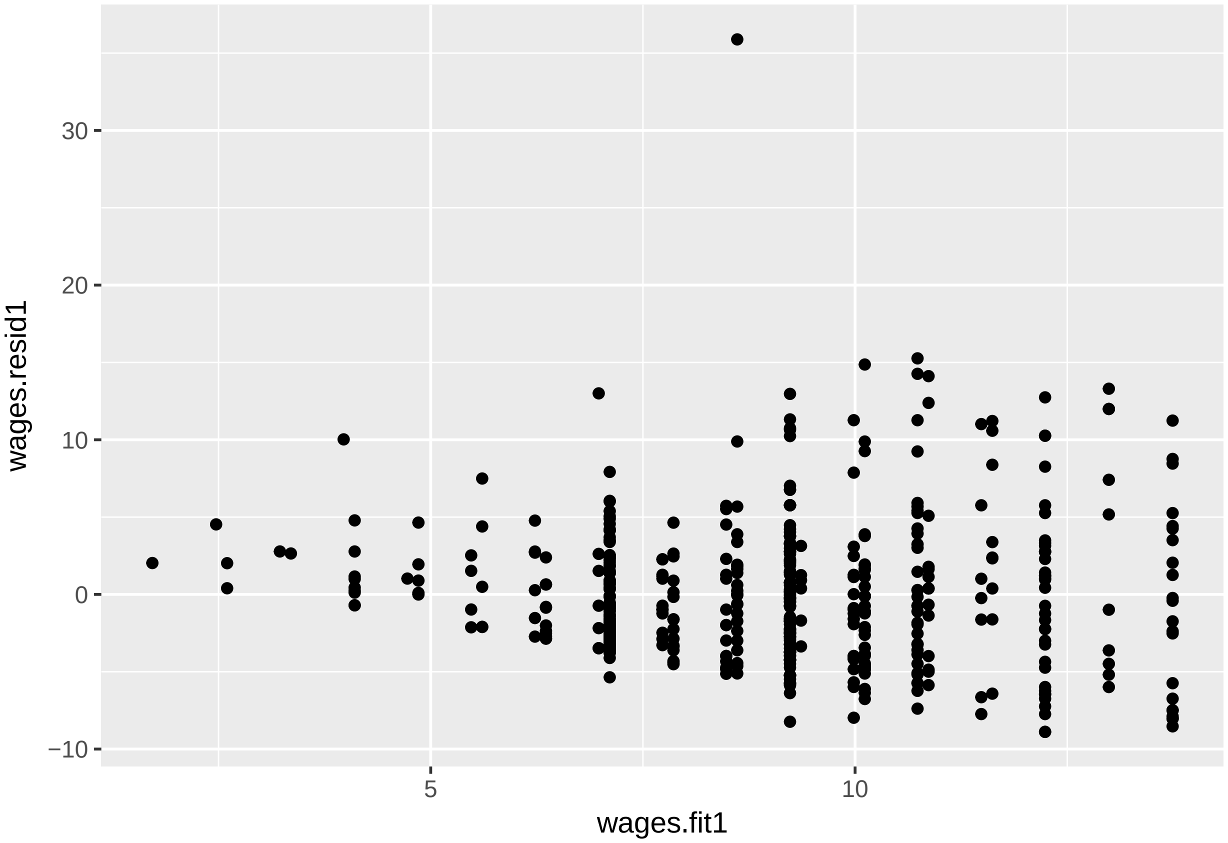
#Checking for co-linearity Will define a new variable EDU_PLUS_EXP which is defined by education plus experience.



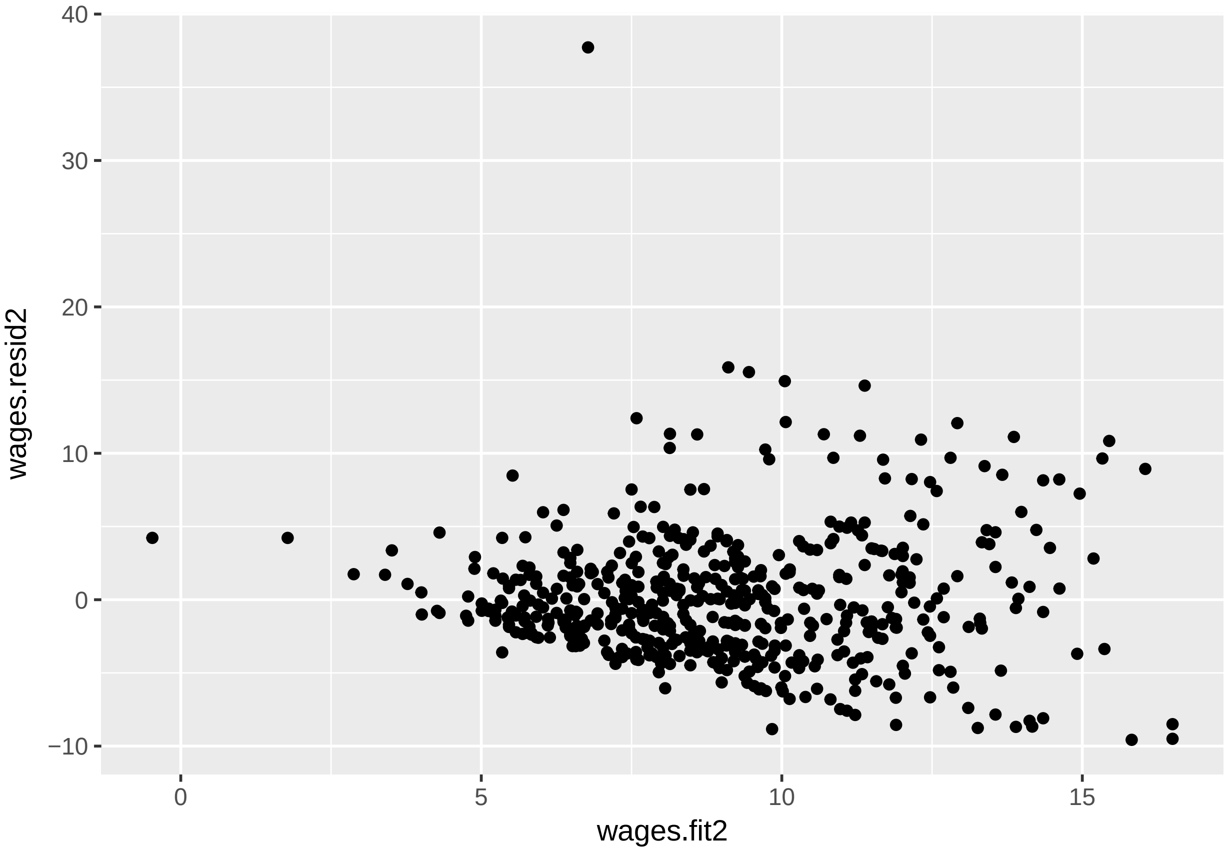
As you can see from the results ED_PLUS_EXP has a correlation of 1 with age. This shows us that age is not needed in our model as ED_PLUS_EXP with AGE would compete to try to explain WAGE which would create a inaccurate results and data.

#Checking our model using residual plot

```
wages.fit1 = fitted(mod1)
wages.resid1 = resid(mod1)

wages.fit2 = fitted(mod2)
wages.resid2 = resid(mod2)

p4 = ggplot(aes(x=wages.fit1, y=wages.resid1), data=wages)
p4 + geom_point()
```



```
p5 = ggplot(aes(x=wages.fit2, y=wages.resid2), data=wages)
p5 + geom_point()
```



We can see that since model 1 variance is non consistant then it shows that we do not have a good model. Doing the same thing to model 2 we can see a lot more constant variance but there still is a slight cone shape variance which shows that this is not the best model to use.

#Power Transformation We then used a power transformation in order to see whether we are able to get better linear models

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

```
##  EDUCATION EXPERIENCE
##        1.0        0.5
```

The results above show that EXPERIENCE should be square rooted.

```
wages$tranEXPERIENCE = wages$EXPERIENCE^(1/2)
tranMod2 = lm(WAGE~tranEXPERIENCE + EDUCATION + SEX, data=wages)
summary(tranMod2);summary(mod2)
```

```
##
## Call:
## lm(formula = WAGE ~ tranEXPERIENCE + EDUCATION + SEX, data = wages)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.710 -2.710 -0.505  1.799 38.523
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -6.1361     1.3120  -4.677 3.70e-06 ***
## tranEXPERIENCE   1.0093     0.1371   7.361 7.02e-13 ***
## EDUCATION        0.9309     0.0772  12.058  < 2e-16 ***
## SEXFemale       -2.3462     0.3852  -6.090 2.17e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.422 on 530 degrees of freedom
## Multiple R-squared:  0.2636, Adjusted R-squared:  0.2595
## F-statistic: 63.25 on 3 and 530 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = WAGE ~ EDUCATION + SEX + EXPERIENCE, 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) -4.16687    1.18667  -3.511 0.000484 ***
## EDUCATION    0.94051    0.07886  11.926  < 2e-16 ***
## SEXFemale   -2.33763    0.38806  -6.024 3.19e-09 ***
## EXPERIENCE   0.11330    0.01671   6.781 3.19e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.454 on 530 degrees of freedom
## Multiple R-squared:  0.2532, Adjusted R-squared:  0.2489
## F-statistic: 59.88 on 3 and 530 DF,  p-value: < 2.2e-16
```

We see that the R^2 value increases slightly and the RSE decreases slightly. This shows that the transformation of the models did not make that big of a difference. We then checked to see if the predictor needed to be transformed.

```
coef(powerTransform(WAGE~tranEXPERIENCE + EDUCATION + SEX, data=wages), round=TRUE)
```

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

We see that we should take the log of WAGE in order to get a more accurate linear model

```
tranMod2.1 = lm(log(WAGE)~tranEXPERIENCE + EDUCATION + SEX, data=wages)
summary(tranMod2.1);summary(tranMod2)
```

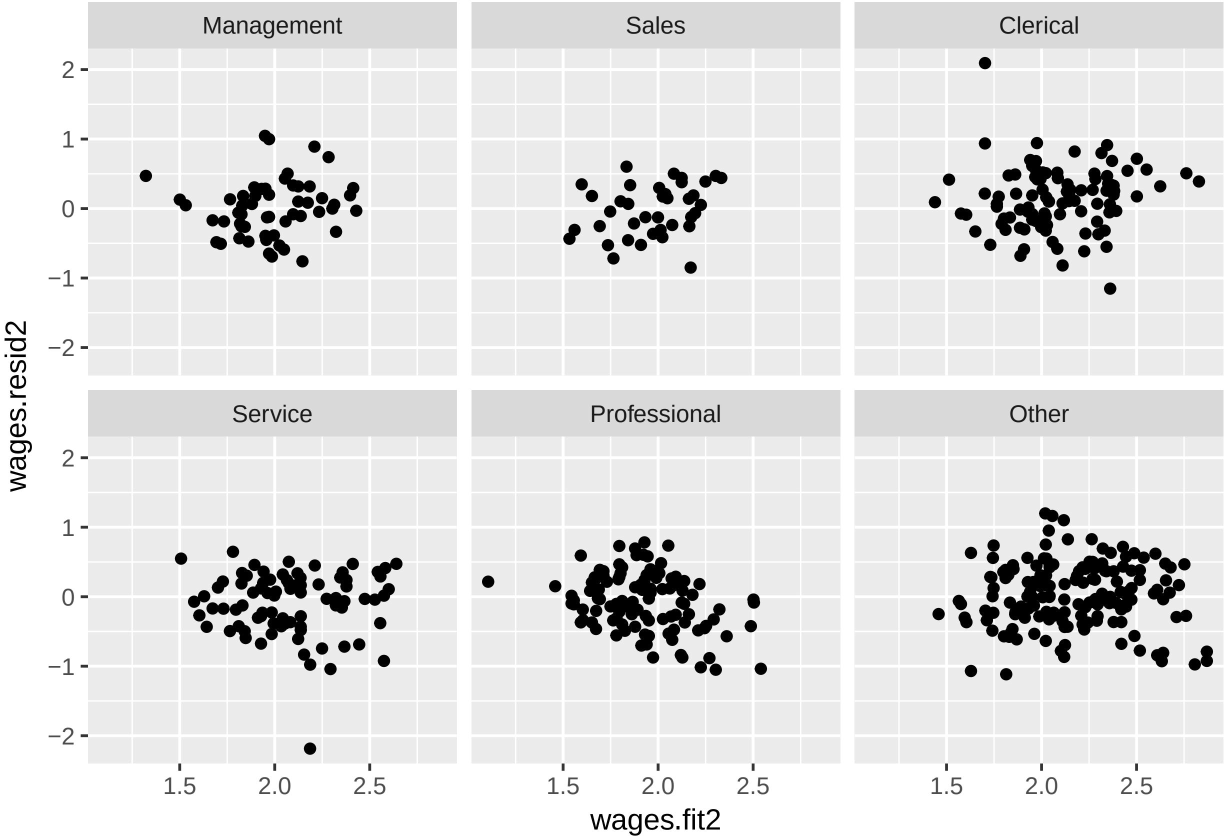
```
##
## Call:
## lm(formula = log(WAGE) ~ tranEXPERIENCE + EDUCATION + SEX, data = wages)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.18596 -0.29069  0.01133  0.30256  2.09292
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.426097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

```
##
## Call:
## lm(formula = WAGE ~ tranEXPERIENCE + EDUCATION + SEX, data = wages)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.710 -2.710 -0.505  1.799 38.523
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -6.1361     1.3120  -4.677 3.70e-06 ***
## tranEXPERIENCE    1.0093     0.1371   7.361 7.02e-13 ***
## EDUCATION         0.9309     0.0772  12.058  < 2e-16 ***
## SEXFemale        -2.3462     0.3852  -6.090 2.17e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.422 on 530 degrees of freedom
## Multiple R-squared:  0.2636, Adjusted R-squared:  0.2595
## F-statistic: 63.25 on 3 and 530 DF,  p-value: < 2.2e-16
```

We see that the R^2 value increased slightly once again and the RSE also decreased slightly. This shows that transforming our linear model was a correct choice.

#Checking Occupation with our Model

```r
wages.resid2 = resid(tranMod2.1)
wages.fit2 = fitted(tranMod2.1)

p6 = ggplot(aes(x=wages.fit2, y=wages.resid2), data=wages)
p6 + geom_point() + facet_wrap(~wages$OCCUPATION)
```



Base on the facet wrap we can see a constant variance for profession and other. This shows that our model is a good fit for those two occupations while it is a bad model for other occupations. #Conclusion Throughout the IE project we thought of many different reasons why a wage gap would exist through real life experiences. We throught it was because of women working less because of pregancy and taking care of children which is a reason for the wage gap. Unfortunately out