IE4

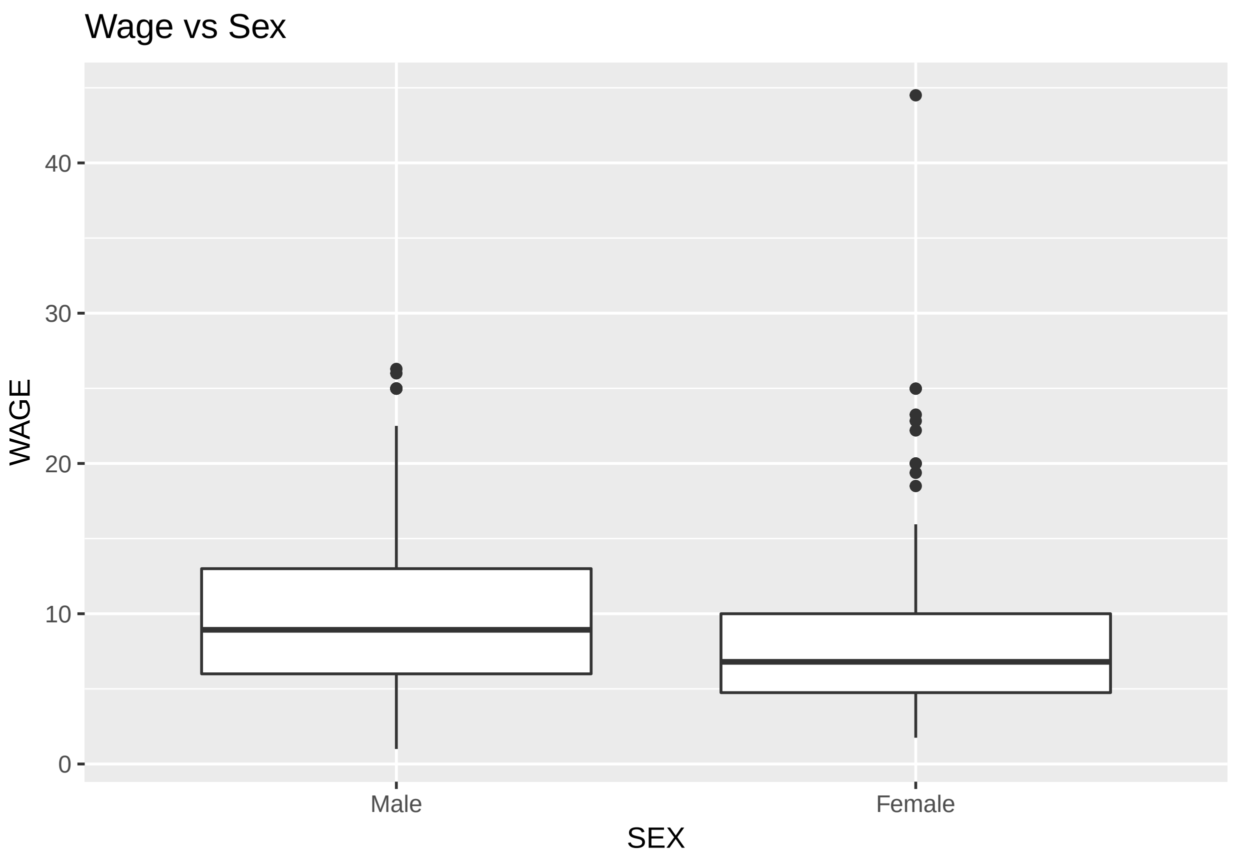
Introduction

The Goal of our project is to see whether there is a wage gap in American society between males and females. We would like to study the different reasons on why there is a wage gap and would hope to find a answer to this wage gap. The data we used was provided by Dae Young of University of Massachusetts: Amherst Mathematics Associate Professor. The data was collected in 1985 and consist of a random sample of 534 people in the work force. The predictors that we used are

- *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

Checking if there is a wage gap

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



In this paper we woulud like to see if there is a wage gap between male and female. Using a box plot one would see that the averages for male and female are different. Using a large sample hypothesis testing we are able to confirm whether there is a difference between the average pay for males compared to female using a 95% confidence.

```r
#nrow(subset(wages,SEX=='Female'))
#nrow(subset(wages,SEX=='Male'))
t.test(subset(wages,SEX=='Male')$WAGE, subset(wages,SEX=='Female')$WAGE, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  subset(wages, SEX == "Male")$WAGE and subset(wages, SEX == "Female")$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
```

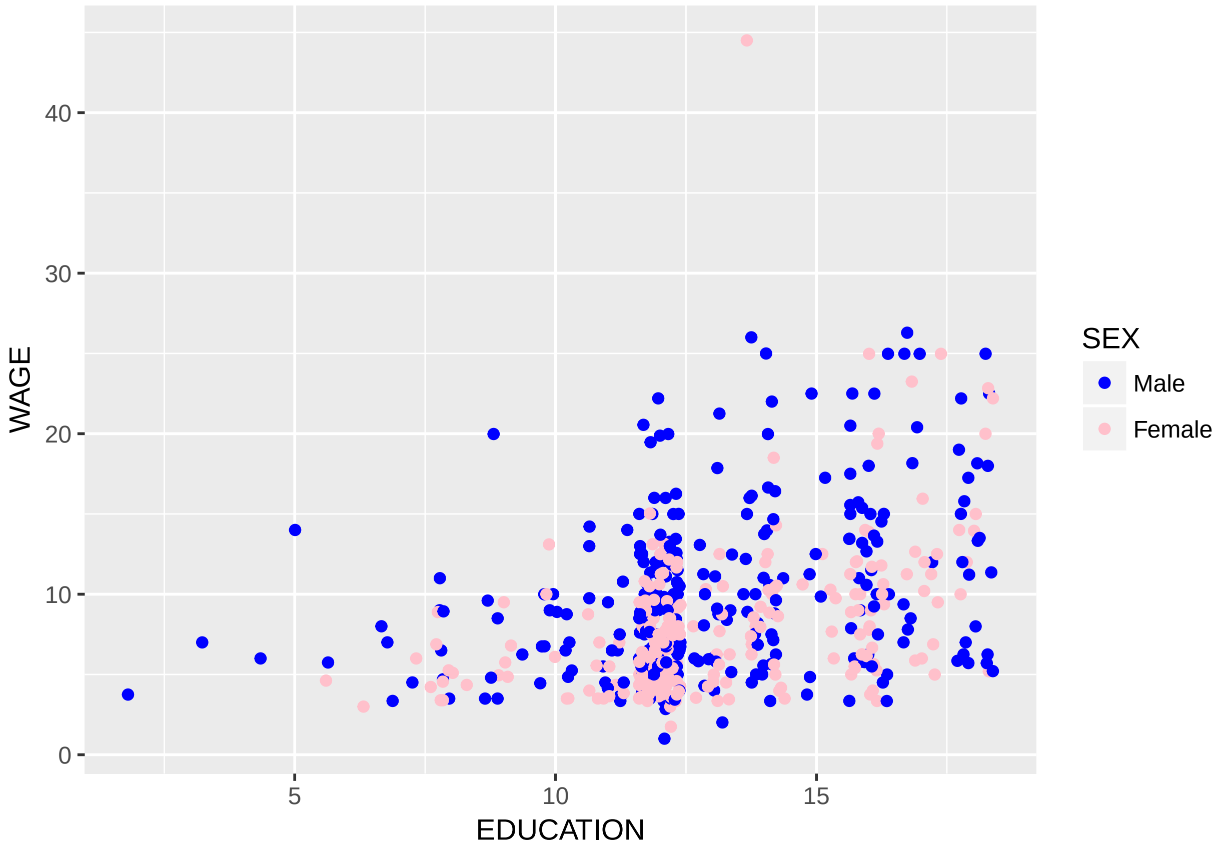
Based on the results from using the T-test with the hypothesis test being

$$H_o : \mu_{male} = \mu_{female} H_1 : \mu_{male} \neq \mu_{female}$$

we would be to see that there is a wage gap as we reject our initial hypothesis and accept our alternative hypthoesis. This motivates us to move further in order to investigate why a wage gap exist and maybe solutions on how to fix this.

Education and Wage Gap

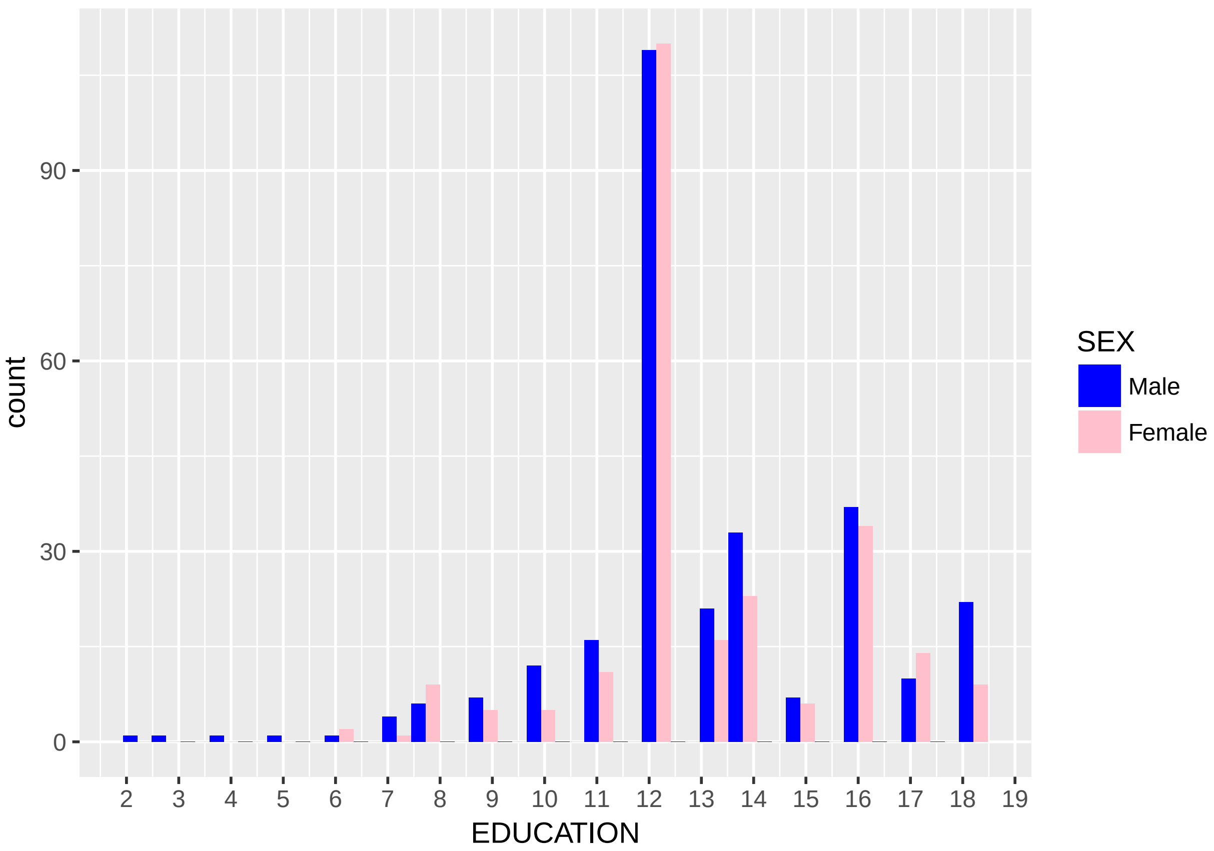
```
p_ed_wg = ggplot( wages, aes ( x = EDUCATION, y = WAGE) )
p_ed_wg + geom_jitter(aes(color=SEX)) + scale_color_manual(name = 'SEX', values = c("blue", "pink"))
```



Using a scatter plot we were able to see a lot of data overlapping in one area. We then switched to using a histogram which allowed us to see more clearly whether this was true or not.

```
ggplot(wages, aes(EDUCATION, fill = SEX)) + geom_histogram(position = 'dodge') + scale_fill_manual(valu
```

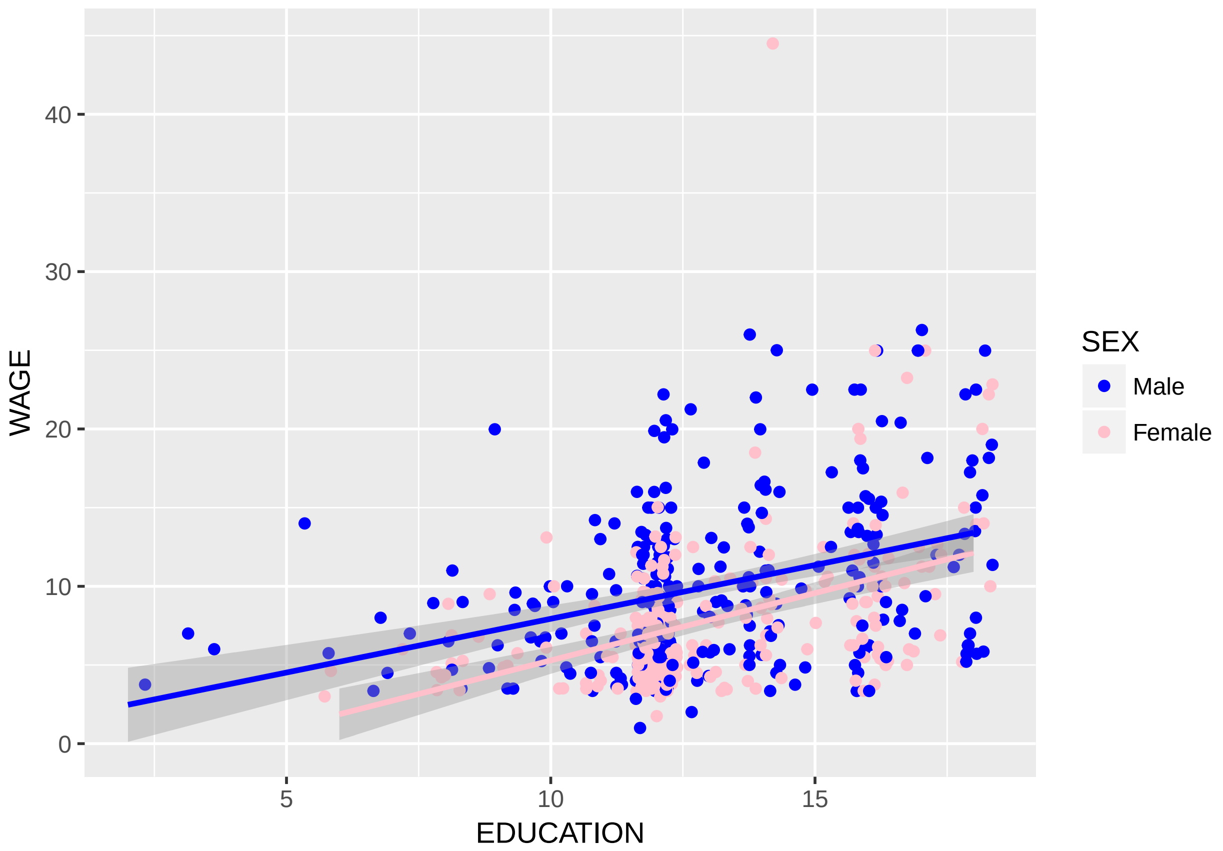
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The histogram above shows the highest level of education a male or female have in our data set. One can see that on average male and females mostly have 12 years of education which translates to finishing high school. Another observation is males having more high levels of education compared to females. Based on everyday life we would expect that having more education would increase one's wage.

Our second hypthesis is people are paided proportional to their level of education which explains the reason for a wage gape between males and females. So if females and males have the same levels of education then a wage gap would not exist.

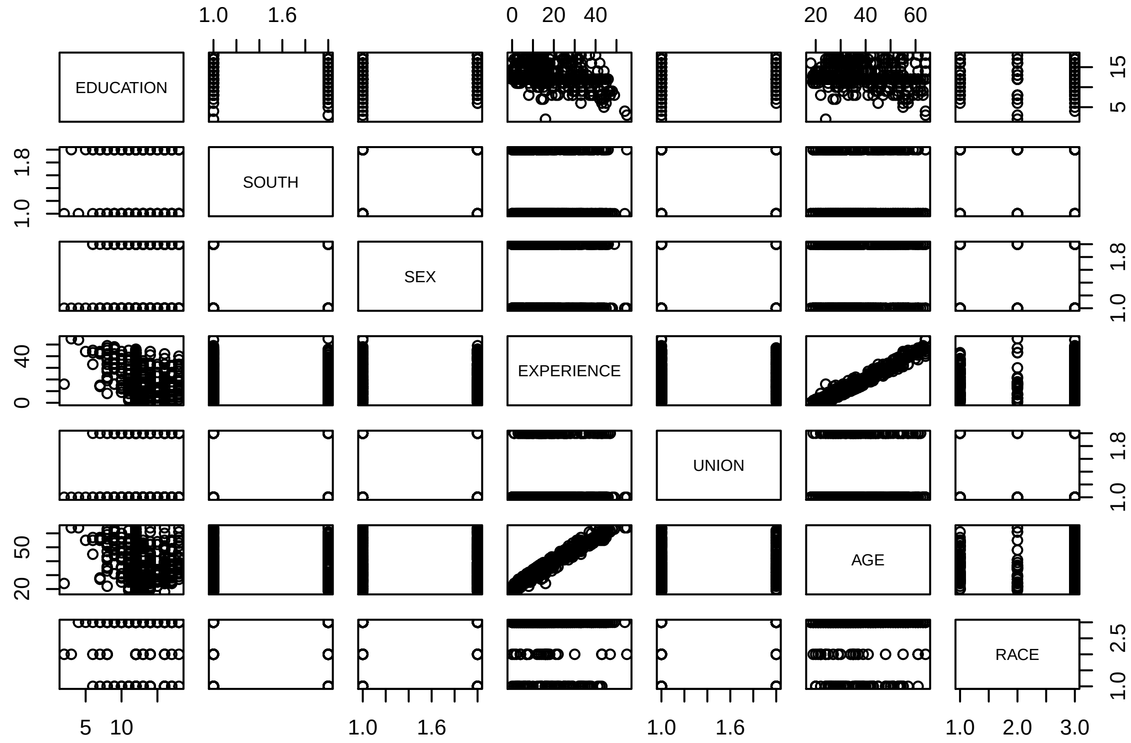
```
p_ed_wg + geom_jitter(aes(color=SEX)) +
  geom_smooth(data=women, method='lm',formula=y~x,color="pink") +
  geom_smooth(data=men, method='lm',formula=y~x,color="blue") +
  scale_color_manual(name = 'SEX', values = c("blue", "pink"))
```



Looking at the same graph above and adding regression lines we are able to see that there are other factors other than education which contributes to the wage gap. This helps to re-enforce a wage gap between male and female workers based on experience.

Looking for Collinearity

```
pairs(wages[c(1,2,3,4,5,6,7)])
```



From using the pairs function we can see that experience and age have a high R^2 value of 0.95 which shows that they are collinear. This allows us to know before hand that creating a linear model with experience and age is not recommended and that we should transform our data before creating a linear model with both experience and age.

Comparing Linear Models

$$M1 : E(Wage|Education, Sex) = \beta_0 + \beta_1 Education + \beta_2 Sex \tag{1}$$

$$M2 : E(Wage|Education, Sex) = \beta_0 + \beta_1 Education + \beta_2 Sex + \beta_3 Sex * Education \tag{2}$$

$$\tag{3}$$

```
M1=lm(WAGE~EDUCATION+SEX, data=wages)
M2=lm(WAGE~EDUCATION+SEX+SEX:EDUCATION, data=wages)

summary(M1)
```

```
##
## 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
```

```r
summary(M2)
```

```
## 
## 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
```

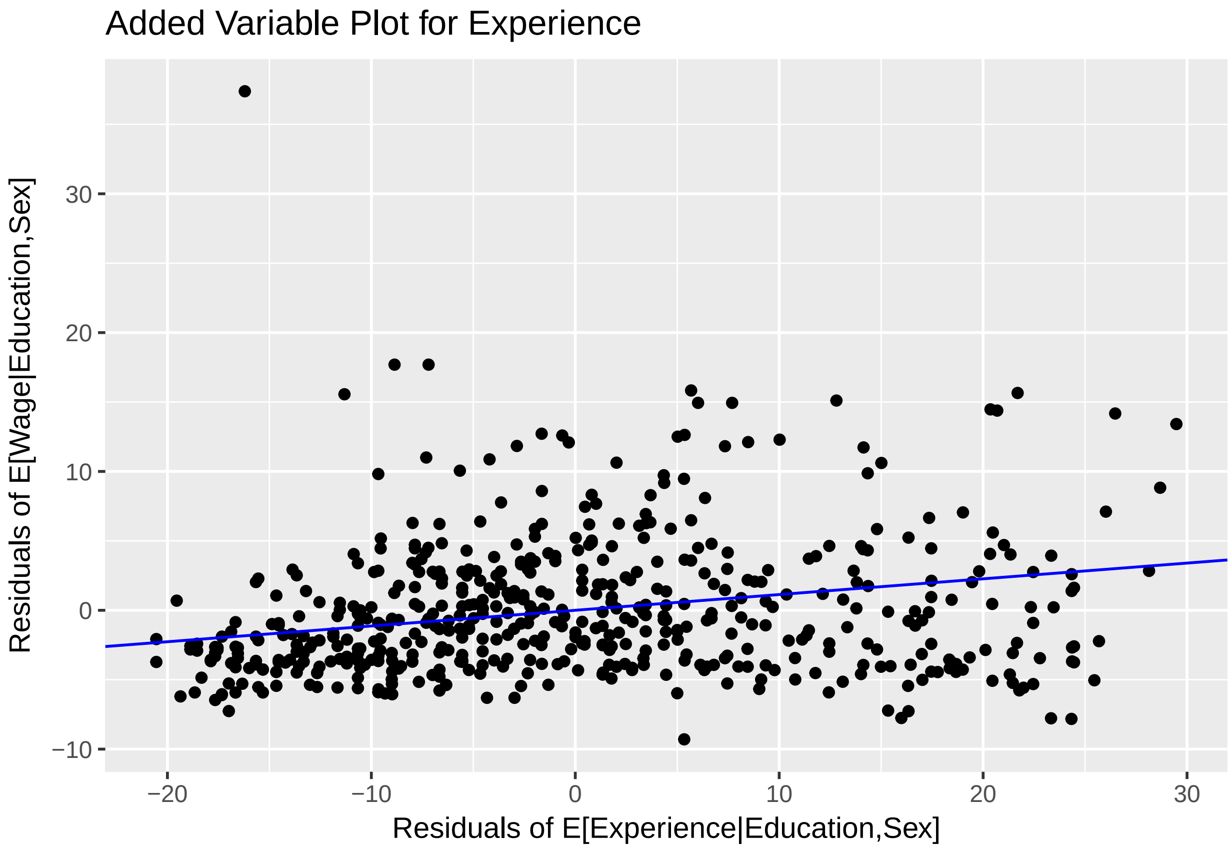
From the results above we can see that Mod1 compared to Mod2 have a very similar residual standard error and from that we can conclude that fitting two different slopes are not justified. From mod1 to mod2 we can also see a slight bump in R^2 value, but not significant enough to draw a conclusion from.

```r
M3 <- lm( EXPERIENCE ~ EDUCATION + SEX, data = wages)
M4 <- lm( WAGE ~ EDUCATION + SEX + EXPERIENCE, data = wages)
M5 <- lm( WAGE ~ SEX + EXPERIENCE, data = wages)

resid_wg_edu <- resid(M1)
resid_exp_edu <- resid(M3)
resid_wg_exp = resid(M5)



avp_ed_ex <- data.frame(resid_wg_edu, resid_exp_edu)
temp <- lm(resid_wg_edu ~ resid_exp_edu, data = avp_ed_ex)
p_added_v_edu_exp <- ggplot( avp_ed_ex, aes ( x = resid_exp_edu, y = resid_wg_exp) )
p_added_v_edu_exp + geom_point() +
  #geom_smooth(data=wages, method='lm',formula=y~x,color="pink") +
  geom_abline ( intercept = coef(temp)[1], slope = coef(temp)[2], color="blue" ) +
  ggtitle("Added Variable Plot for Experience") +
  xlab("Residuals of E[Experience|Education,Sex]") +
  ylab("Residuals of E[Wage|Education,Sex]")
```



```
summary(temp)$r.squared
```

```
## [1] 0.07983473
```

```
summary(M4)
```

```
##
## 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
```

Based on the results from the added variable plot we see that 7% of experience explains wages in our model. This is a large percentage as there are many other factors which can explain wages in our dataset. This gives us a reason to add experience into our linear models.