Lecture 16 EEP118

In [1]:
```
#Lecture16.R
#LECTURE 16



# Load the 'pacman' package
library(pacman)
#packages to use load them now using the pacman "manager"
p_load(dplyr, haven, readr)
#Another great feature of p_load(): if you try to load a package that is not
p_load(ggplot2)

pacman::p_load(lfe, lmtest, haven, sandwich, tidyverse)
# lfe for running fixed effects regression
# lmtest for displaying robust SE in output table
# haven for loading in dta files
# sandwich for producing robust Var-Cov matrix
# tidyverse for manipulating data and producing plots
```

In [2]:
```
#set scientific display off, thank you Roy
options(scipen=999)

#read in a Stata dataset
mydata <- read_dta("Lecture16.dta")
head(mydata)
```

A tibble: 6 × 8

wage	educ	exper	female	west	services	profocc	cateduc
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
3.75	2	39	0	0	0	0	1
2.92	3	51	0	0	1	0	1
3.51	4	39	0	1	0	0	1
3.00	4	48	0	0	0	0	1
3.00	4	36	0	0	0	0	1
5.20	6	47	0	0	0	0	1

In [3]:
```
#generate category education variable
mydata$cateduc=1
mydata$cateduc[mydata$educ==12]<-2
mydata$cateduc[mydata$educ>12]<-3

#summary stats variables
summary(mydata)
```

```
     wage             educ           exper           female      
 Min.   : 0.530   Min.   : 0.00   Min.   : 1.00   Min.   :0.0000  
 1st Qu.: 3.330   1st Qu.:12.00   1st Qu.: 5.00   1st Qu.:0.0000  
 Median : 4.650   Median :12.00   Median :13.50   Median :0.0000  
 Mean   : 5.896   Mean   :12.56   Mean   :17.02   Mean   :0.4791  
 3rd Qu.: 6.880   3rd Qu.:14.00   3rd Qu.:26.00   3rd Qu.:1.0000  
 Max.   :24.980   Max.   :18.00   Max.   :51.00   Max.   :1.0000  
     west           services         profocc          cateduc     
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :1.000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:2.000  
 Median :0.0000   Median :0.0000   Median :0.0000   Median :2.000  
 Mean   :0.1692   Mean   :0.1008   Mean   :0.3669   Mean   :2.183  
 3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:3.000  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :3.000  
```

*1. ESTIMATE DIFFERENCE IN MEANS BETWEEN MALE AND FEMALE

In [4]:
```
#*1. ESTIMATE DIFFERENCE IN MEANS BETWEEN MALE AND FEMALE

summary(mydata$wage[which(mydata$female==1)])
summary(mydata$wage[which(mydata$female==0)])

#standard errors of the data
sdfemale<-sd(mydata$wage[which(mydata$female==1)])
sdmale<-sd(mydata$wage[which(mydata$female==0)])
sdfemale
sdmale

#number of observations for female=1 and female=0

mydata %>% count(female)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.530   3.000   3.750   4.588   5.510  21.630 
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  1.500   4.143   6.000   7.099   8.765  24.980 
```

2.52936310395604

4.16085751149977

A tibble: 2 × 2

female	n
<dbl>	<int>
0	274
1	252

In [5]:
```
# t test difference in means

t.test(mydata$wage~mydata$female)
```

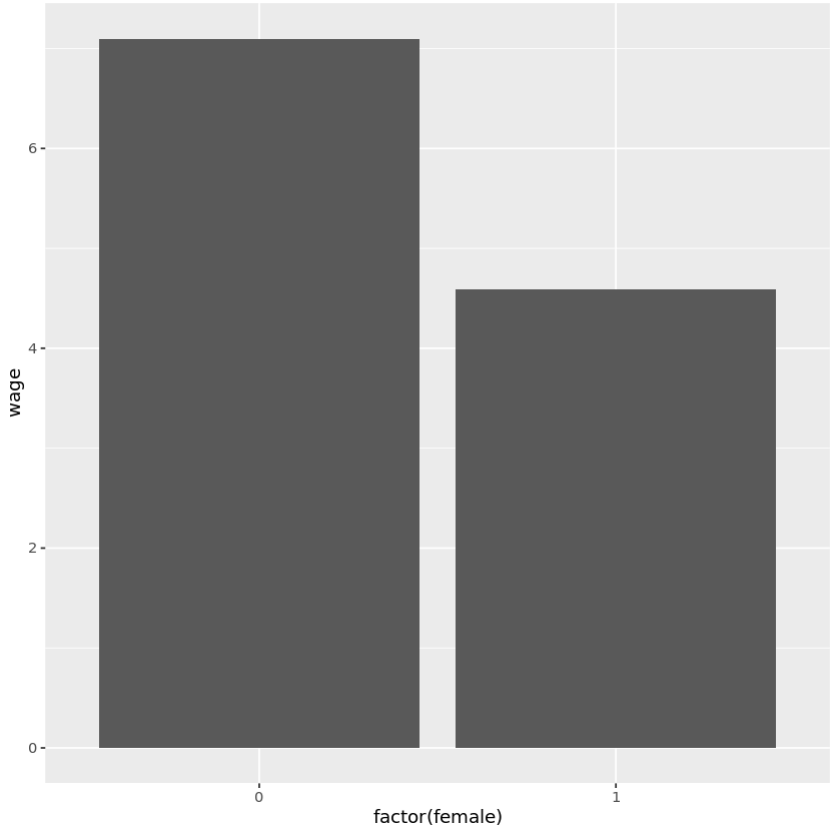
```
        Welch Two Sample t-test

data:  mydata$wage by mydata$female
t = 8.44, df = 456.33, p-value = 0.0000000000000004243
alternative hypothesis: true difference in means between group 0 and group 1
is not equal to 0
95 percent confidence interval:
 1.926971 3.096690
sample estimates:
mean in group 0 mean in group 1
       7.099489        4.587659
```

graph of wage by female indicator

In [6]:
```
ggplot(mydata, aes(x = factor(female), y = wage))+
  stat_summary(fun.y="mean", geom="bar")
```

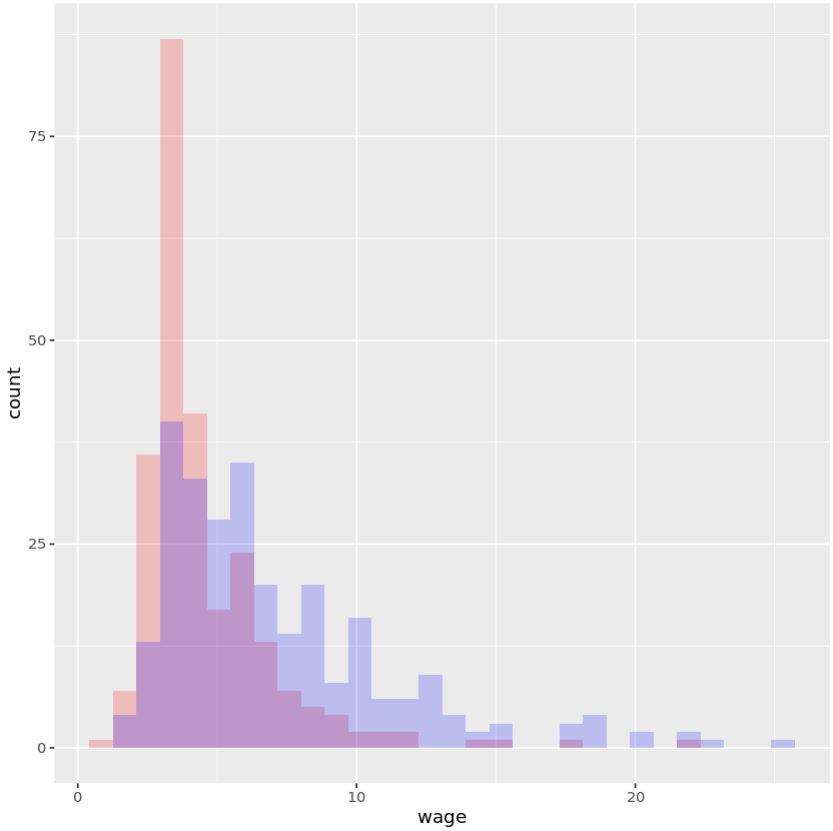
```
Warning message:
“The `fun.y` argument of `stat_summary()` is deprecated as of ggplot2 3.3.0.
i Please use the `fun` argument instead.”
```



*2. DISTRIBUTION OF WAGE BY GENDER

In [7]:
```
ggplot(mydata,aes(x=wage)) +
  geom_histogram(data=subset(mydata,female == 1),fill = "red", alpha = 0.2)
  geom_histogram(data=subset(mydata,female == 0),fill = "blue", alpha = 0.2)
```

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



In [8]:
```
#linear model
two<-lm(wage~female,mydata)
summary(two)
```

```
Call:
lm(formula = wage ~ female, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.5995 -1.8495 -0.9877  1.4260 17.8805 

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)    
(Intercept)   7.0995     0.2100  33.806 < 0.0000000000000002 ***
female       -2.5118     0.3034  -8.279  0.00000000000000104 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.476 on 524 degrees of freedom
Multiple R-squared:  0.1157,    Adjusted R-squared:  0.114 
F-statistic: 68.54 on 1 and 524 DF,  p-value: 0.000000000000001042
```

3. DO WOMEN HAVE DIFFERENT CHARACTERISTICS THAT MATTER FOR WAGE?

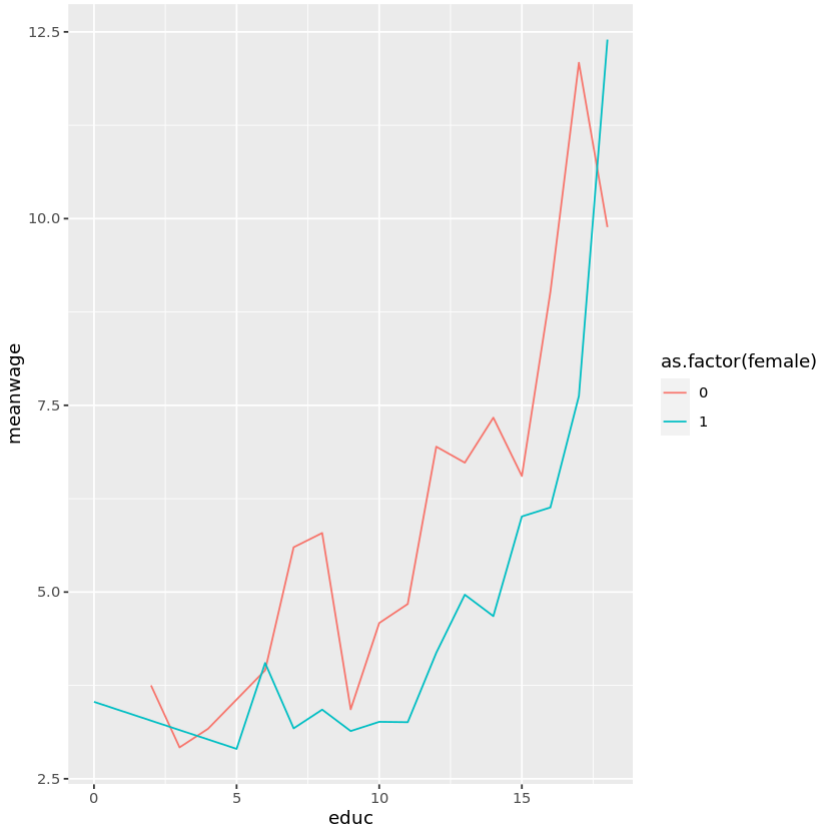
In [9]:
```
###graph average wages for different educaiton levels

mydata$wage_female=mydata$wage
mydata$wage_female[mydata$female==0]<-0
```

```
mydata$wage_male=mydata$wage
mydata$wage_male[mydata$female==1]<-0

mydata <- mydata %>%
  group_by(educ,female) %>%
  mutate(meanwage = mean(wage))


ggplot(mydata,aes(y = meanwage,x = educ,color = as.factor(female))) +
  geom_line()
```



#Discrimination, even after controlling for difference in characteristics:

In [10]:
```
#Discrimination, even after controlling for difference in characteristics:
#Additive female effect

three<-lm(wage~female+educ+exper+services+profocc, mydata)
summary(three)

four<-lm(wage~female, mydata)
summary(four)
```

```
Call:
lm(formula = wage ~ female + educ + exper + services + profocc,
    data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9982 -1.7264 -0.4366  1.0280 13.9494

Coefficients:
            Estimate Std. Error t value          Pr(>|t|)
(Intercept) -0.19250    0.77755  -0.248             0.8046
female      -1.89038    0.26606  -7.105 0.000000000003984 ***
educ         0.43773    0.05768   7.588 0.000000000000151 ***
exper        0.05571    0.01023   5.445 0.000000080193529 ***
services    -0.86378    0.43786  -1.973             0.0491 *
profocc      1.72822    0.32070   5.389 0.000000107676490 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.985 on 520 degrees of freedom
Multiple R-squared:  0.3531,    Adjusted R-squared:  0.3469
F-statistic: 56.77 on 5 and 520 DF,  p-value: < 0.00000000000000022
Call:
lm(formula = wage ~ female, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5995 -1.8495 -0.9877  1.4260 17.8805

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)   7.0995     0.2100  33.806 < 0.0000000000000002 ***
female       -2.5118     0.3034  -8.279  0.00000000000000104 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.476 on 524 degrees of freedom
Multiple R-squared:  0.1157,    Adjusted R-squared:  0.114
F-statistic: 68.54 on 1 and 524 DF,  p-value: 0.000000000000001042
```

*4. ARE THERE DIFFERENTIAL RETURNS TO EDUCATION FOR MEN AND WOMEN?

In [11]:
```r
#generate interaction
mydata$femeduc=mydata$female*mydata$educ

five<-lm(wage~female+educ+femeduc, mydata)
summary(five)
```

```
Call:
lm(formula = wage ~ female + educ + femeduc, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1611 -1.8028 -0.6367  1.0054 15.5258

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept)  0.20050    0.84356   0.238                0.812
female      -1.19852    1.32504  -0.905                0.366
educ         0.53948    0.06422   8.400 0.000000000000000424 ***
femeduc     -0.08600    0.10364  -0.830                0.407
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.186 on 522 degrees of freedom
Multiple R-squared:  0.2598,    Adjusted R-squared:  0.2555
F-statistic: 61.07 on 3 and 522 DF,  p-value: < 0.00000000000000022
```

In [12]:
```
#generate predictions
mydata$wagehat<-predict(lm(wage~female+educ+femeduc, mydata))
```

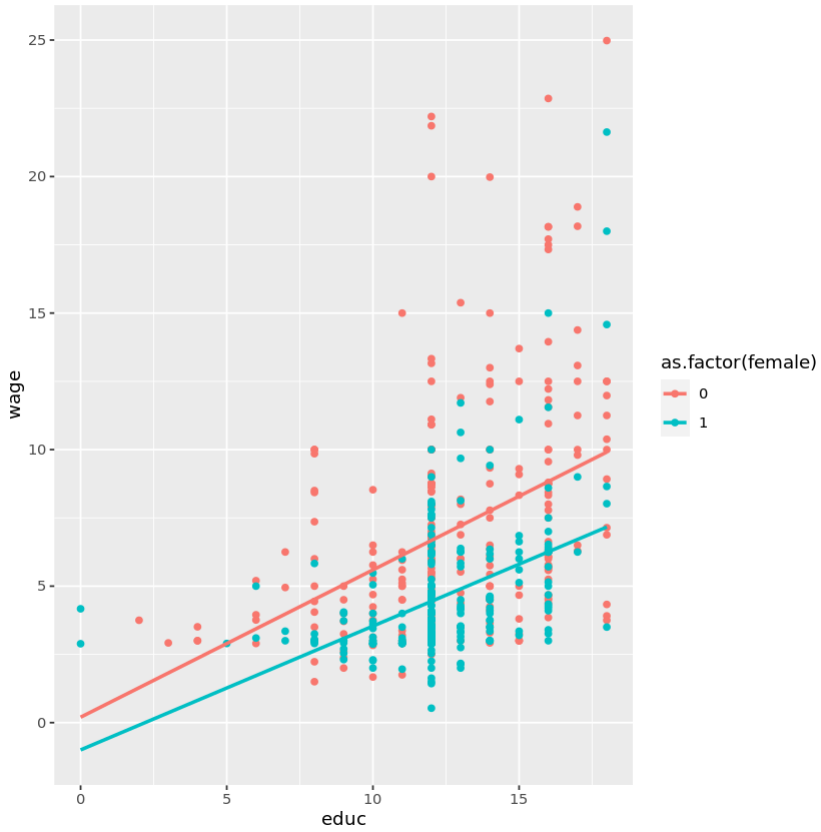
plot them the wages for female and non female and also the predictions

In [13]:
```
#mydata$trfem=mydata$wagehat
#mydata$trfem[mydata$female==0]<-0

#mydata$trmale=mydata$wagehat
#mydata$trmale[mydata$female==1]<-0

ggplot(mydata, aes(x=educ, y=wage, color=as.factor(female))) +
  geom_point() +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE)
```

```
`geom_smooth()` using formula = 'y ~ x'
```



#* Finally, is there a WEST COAST DIFFERENTIAL EFFECT ON MEN' AND WOMEN'S WAGES?

In [14]:
```
#* Finally, is there a WEST COAST DIFFERENTIAL EFFECT ON MEN' AND WOMEN'S WA
summary(mydata$west)

mydata$femwest=mydata$female*mydata$west

five<-lm(wage~female+west+femwest+educ+exper,mydata)
summary(five)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.1692  0.0000  1.0000 
```

```
Call:
lm(formula = wage ~ female + west + femwest + educ + exper, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
-6.189 -1.915 -0.484  1.136 15.020

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) -1.89416    0.75096  -2.522              0.01196 *
female      -2.07371    0.29462  -7.039     0.00000000000616 ***
west         1.37729    0.51892   2.654              0.00819 **
femwest     -0.73218    0.71485  -1.024              0.30619
educ         0.59808    0.05084  11.765 < 0.0000000000000002 ***
exper        0.06488    0.01034   6.274     0.00000000074037 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.058 on 520 degrees of freedom
Multiple R-squared:  0.3208,    Adjusted R-squared:  0.3142
F-statistic: 49.12 on 5 and 520 DF,  p-value: < 0.00000000000000022
```

In [ ]: