Chapter3_Problem_Set

Tianwei Liu

9/22/2019

```
# Load Data
load("Ch3_Exercise3_Height_and_Wages_UK.RData")
load("Ch3_Exercise4_Divorce_rates_Men.RData")
load("Ch3_Exercise4_Divorce_rates_Women.RData")
```

3.

(a) Estimate a model where height at age 33 explains income at age 33. Explain Beta1 and Beta0.

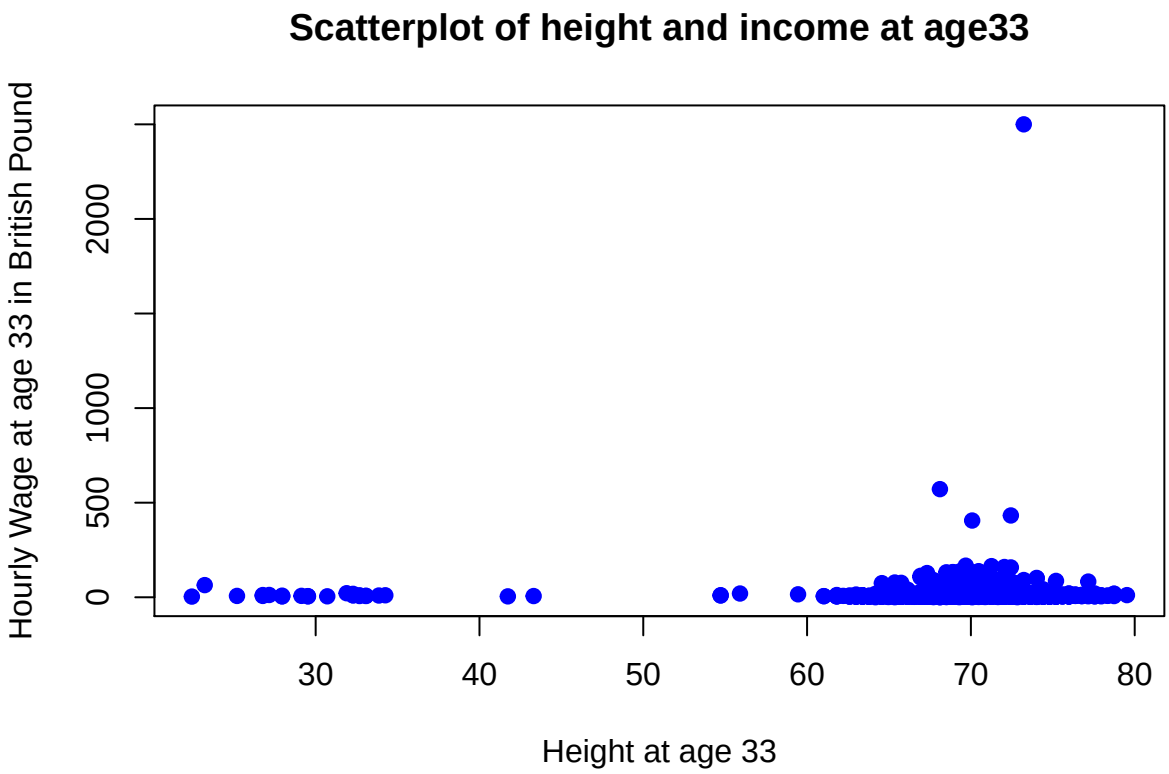
```
my_reg1 <- lm (gwage33 ~ height33, data=dta)
summary(my_reg1)
```

```
##
## Call:
## lm(formula = gwage33 ~ height33, data = dta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -10.68   -4.96   -3.11   -0.68 2488.68
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6.5994    12.2143  -0.540    0.589
## height33      0.2447     0.1757   1.393    0.164
##
## Residual standard error: 44.71 on 3694 degrees of freedom
## Multiple R-squared:  0.0005252,  Adjusted R-squared:  0.0002546
## F-statistic: 1.941 on 1 and 3694 DF,  p-value: 0.1636
```

Beta1 has a value of 0.2447. This means that an one-inch increase in the 33 years-old man group increases hourly wage by 0.2447 pound. Beta0 has a value of -6.5994. This means that when a man has 0 inch of height, his hourly wage is -6.5994.

(b) Create a scatterplot of height and income at age 33. Identify outliers.

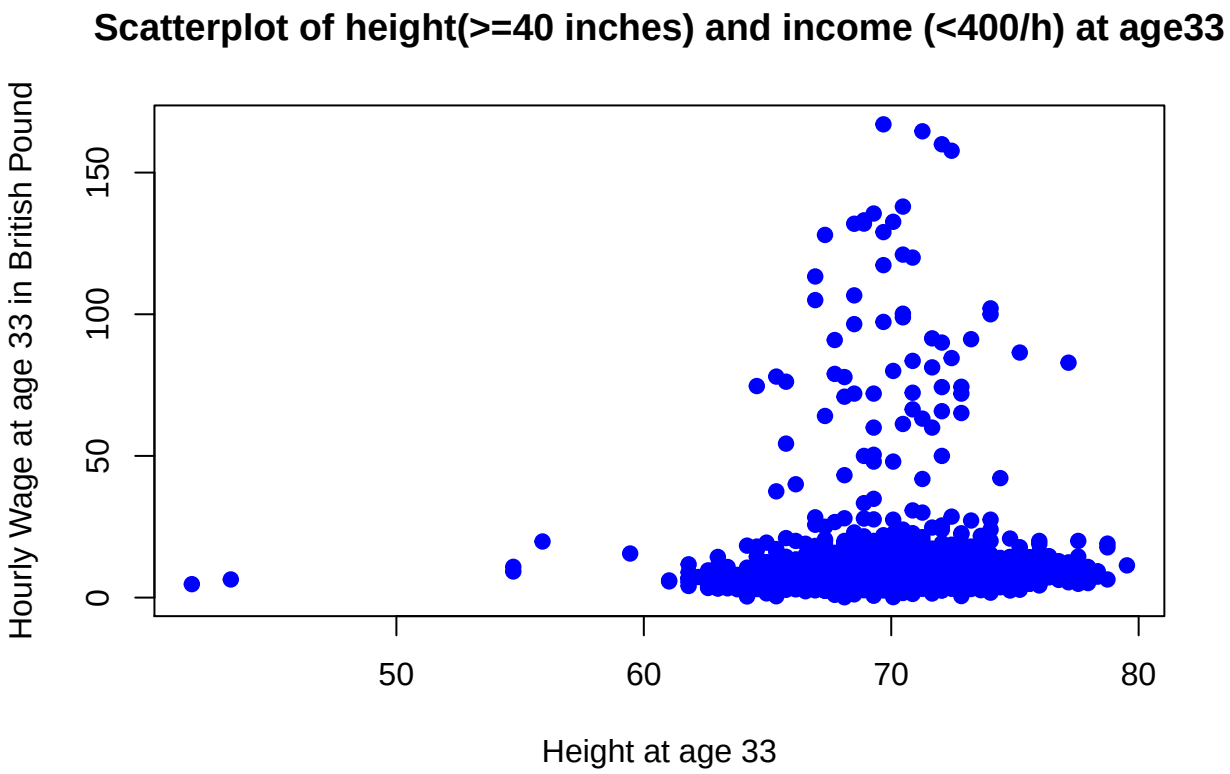
```
plot(dta$height33, dta$gwage33, pch = 19, col = "blue", xlab = "Height at age 33", ylab = "Hourly Wage a
```



There is one observation whose hourly wage is about 2500 British pounds and three observations whose wage is around 500 pounds per hour. These observations are outliers.

(c) Create a scatterplot of height and income at age 33, but exclude observations with wages per hour more than 400 British pounds and height less than 40 inches. Describe the difference from the earlier plot. Which plot seems the more reasonable basis for statistical analysis? Why?

```
plot(dta$height33[dta$height33>=40 & dta$gwage33<400], dta$gwage33[dta$height33>=40 & dta$gwage33<400],
```



In this graph, compared with the earlier one, we can see the variation among the 0-200 income group better. The second plot is a more reasonable basis for statistical analysis because it has moreved the unreasonably high income observations and therefore potentially corrected for mistakes as well.

(d) Reestimate the bivariate OLS model from part (a), but exclude four outliers with very high wages and outliers with height below 40 inches. Briefly compare results to earlier results.

```
## Prepare the data set with conditions imposed
dta2 <- subset(dta, (dta$height33>=40 & dta$gwage33<400))
## Run Reg on the new dataset with conditions
my_reg2 <- lm (gwage33 ~ height33, data=dta2)
summary(my_reg2)
```

```
##
## Call:
## lm(formula = gwage33 ~ height33, data = dta2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -9.632  -3.835  -2.014   0.356 157.690
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.34591    5.01799  -1.862 0.062616 .
## height33     0.26810    0.07199   3.724 0.000199 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.04 on 3667 degrees of freedom
## Multiple R-squared:  0.003768,   Adjusted R-squared:  0.003496
## F-statistic: 13.87 on 1 and 3667 DF,  p-value: 0.0001989
```

The biggest difference is that after removing outliers from our analysis, the estimate of beta1 coefficient is now statistically significant.

(e) What happens when the sample size is smaller? To answer this question, reestimate the bivariate OLS model from above (that excludes outliers), but limit the analysis to the first 800 observations. Which changes more from the results with the full sample: the estimated coefficient on height or the estimated standard error of the coefficient on height? Explain.

```
## Include only the first 800 Observations
my_reg3 <- lm (gwage33 ~ height33, data=dta2[1:800,])
summary(my_reg3)
```

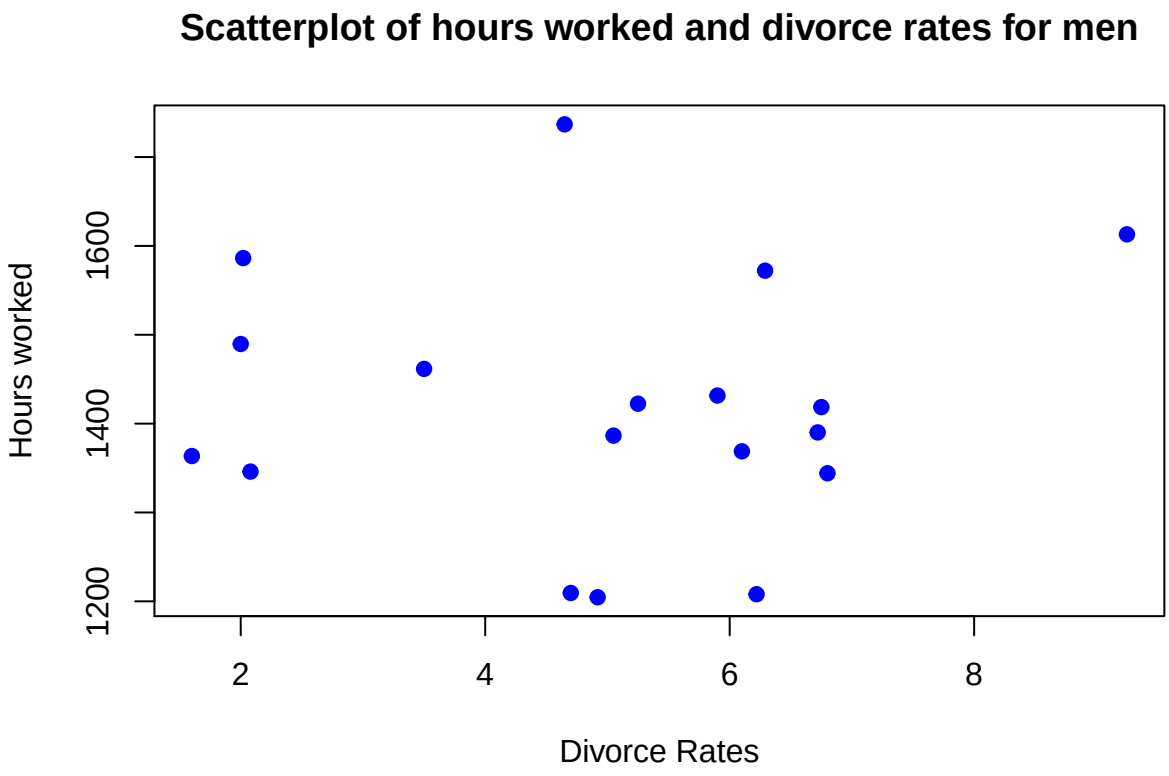
```
##
## Call:
## lm(formula = gwage33 ~ height33, data = dta2[1:800, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -8.340  -3.530  -1.770   0.493 148.070
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.0802     9.0750  -1.111   0.2670
## height33      0.2723     0.1307   2.083   0.0376 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.17 on 798 degrees of freedom
## Multiple R-squared:  0.005408,   Adjusted R-squared:  0.004161
## F-statistic: 4.339 on 1 and 798 DF,  p-value: 0.03757
```

Once we reduce the sample size, the results become less significant. The standard error changed more while the estimated coefficient on height are pretty similar between the two models. Because the standard error is the square root of variance, which is directly linked to sample size N, so standard error is more likely to be affected by a change in the number of observations.

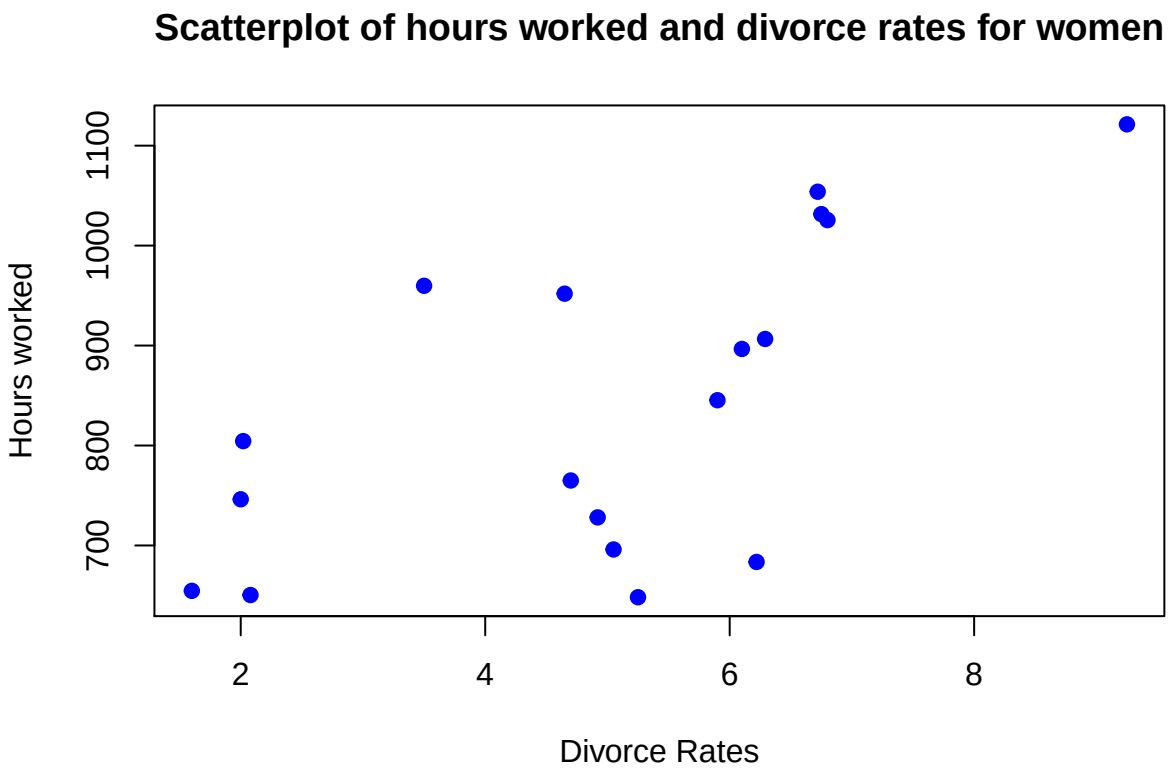
4.

(a) For each data set (for women and for men), create a scatterplot of hours worked on the Y-axis and divorce rates on the X-axis.

```
plot(Mdata$divorcerate, Mdata$hours, pch = 19, col = "blue", xlab = "Divorce Rates", ylab = "Hours work
```




```
plot(Wdata$divorcerate, Wdata$hours, pch = 19, col = "blue", xlab = "Divorce Rates", ylab = "Hours worke
```



(b) For each data set, estimate an OLS regression in which hours worked is regressed on divorce rates. Report the estimated regression equation, and interpret the coefficients. Explain any differences in coefficients.

```
my_regmen <- lm (hours~divorcerate, data=Mdata)
summary (my_regmen)
```

```
##
## Call:
## lm(formula = hours ~ divorcerate, data = Mdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -214.89  -64.49  -18.40   67.69  317.80
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1410.642     93.087  15.154 6.55e-11 ***
## divorcerate    1.798     17.302   0.104    0.919
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 147.8 on 16 degrees of freedom
## Multiple R-squared:  0.0006743,  Adjusted R-squared:  -0.06178
## F-statistic: 0.0108 on 1 and 16 DF,  p-value: 0.9185
```

Estimated Regression Equation: Hours_worked = 1410.642 + 0.000375*divorce_rate Intepretation: An one-percent increase in divorce rate is associated with a 1.798 increase in hours worked for men. In countries where divorce rate is 0, men work 1410.642 hours.

```
my_regwomen <- lm (hours~divorcerate, data=Wdata)
summary (my_regwomen)
```

```
##
## Call:
## lm(formula = hours ~ divorcerate, data = Wdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -218.645  -60.793    0.662  101.665  188.968
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   601.86      77.25   7.791  7.8e-07 ***
## divorcerate    48.28      14.36   3.362  0.00396 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 122.7 on 16 degrees of freedom
## Multiple R-squared:  0.414,  Adjusted R-squared:  0.3774
## F-statistic:  11.3 on 1 and 16 DF,  p-value: 0.003965
```

Estimated Regression Equation: Hours_Worked = 601.86 + 48.28*divorce_rate Intepretation: An one-percent increase in divorce rate is associated with 48.28 more hours worked for women. In a country where divorce rate is 0, women work for 601.86 hours.

In general, hours worked of women is more strongly correlated with divorce rate than that of men. An one-percent increase in divorce rate is associate with 1.798 more hours worked for men but 48.28 hours for women. Also, women work fewer hours than men on average.

(c) What are the fitted value and residual for men in Germany?

```
my_regmen$fitted.values[6]
```

```
##        6
## 1419.487
```

```
my_regmen$residuals[6]
```

```
##         6
## -214.8873
```

(d) What are the fitted value and residual for women in Spain?

```
my_regwomen$fitted.values[14]
```

```
##       14
## 702.2794
```

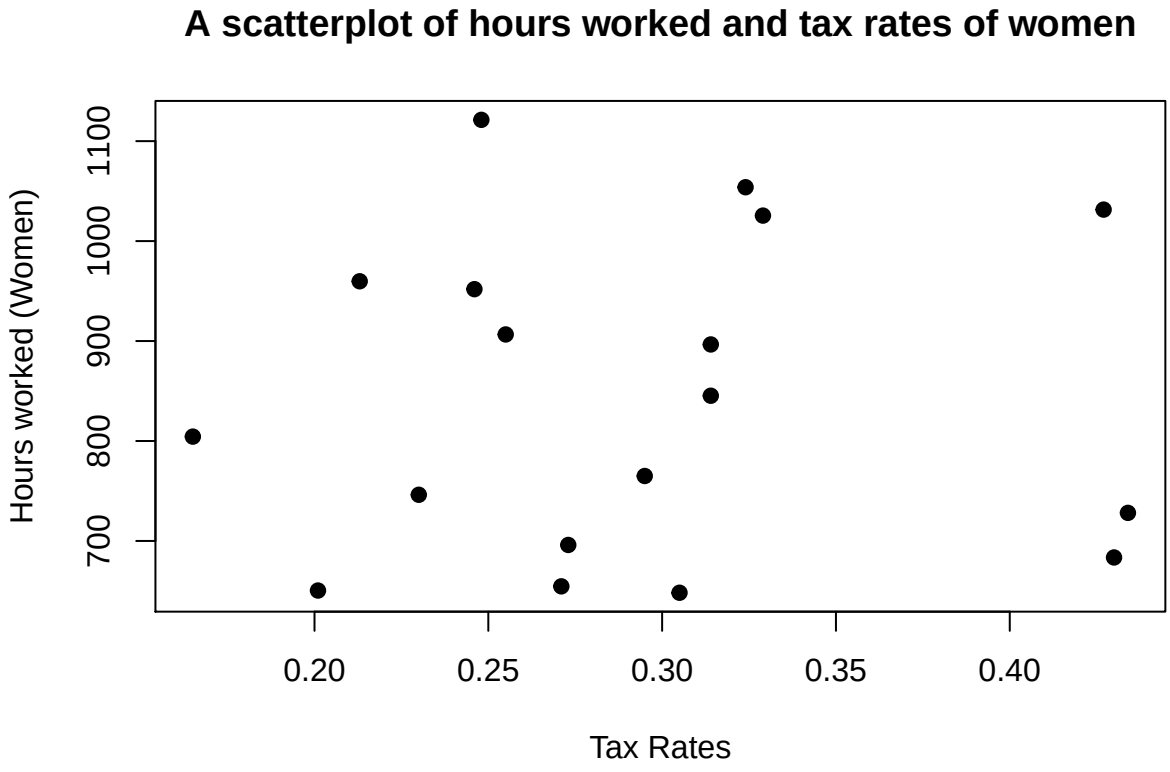
```
my_regwomen$residuals[14]
```

```
##        14
## -51.87941
```

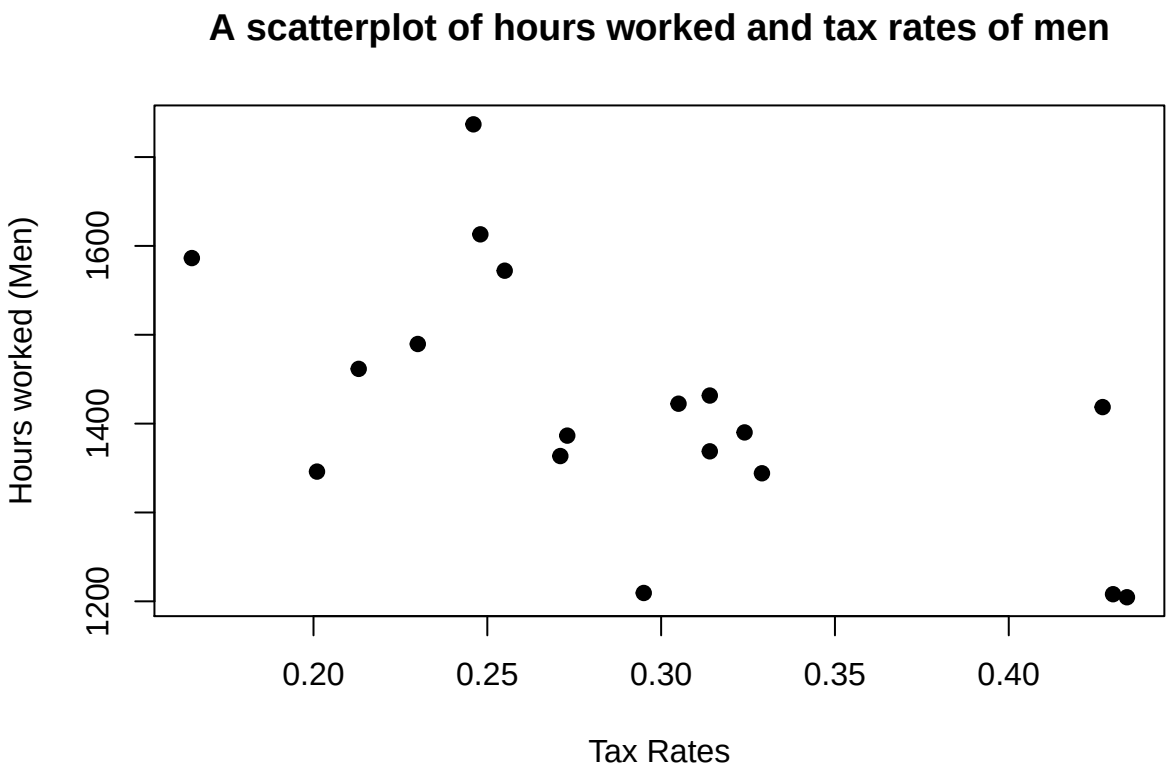
5.

(a) For each data set (for women and for men), create a scatterplot of hours worked on the Y-axis and tax rates on the X-axis.

```
plot (Wdata$taxrate, Wdata$hours, pch=19, xlab = "Tax Rates", ylab = "Hours worked (Women)", main = "A s
```




```
plot (Mdata$taxrate, Mdata$hours, pch=19, xlab = "Tax Rates", ylab = "Hours worked (Men)", main = "A sca
```



(b) For each data, set estimate an OLS regression in which hours worked is regressed on tax rates. Report the estimated regression equation, and interpret the coefficients. Explain any differences in coefficients.

```
my_regwomen1 <- lm (hours~taxrate, data=Wdata)
summary (my_regwomen1)
```

```
##
## Call:
## lm(formula = hours ~ taxrate, data = Wdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -195.2 -139.8  -15.0  119.0  281.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   827.05     151.97   5.442 5.43e-05 ***
## taxrate        53.46     502.42   0.106    0.917
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 160.2 on 16 degrees of freedom
## Multiple R-squared:  0.0007071,  Adjusted R-squared:  -0.06175
## F-statistic: 0.01132 on 1 and 16 DF,  p-value: 0.9166
```

Estimated Regression Equation: Hours_Worked = 827.05 + 53.45*tax_rate Inteperatation: An one-percent increase in tax rate increases hours worked for women by 0.53 hours (Note that tax rate is measured on a scale of 0-1). In a country where tax rate is 0, women work 827.05 hours.

```
my_regmen1 <- lm (hours~taxrate, data=Mdata)
summary (my_regmen1)
```

```
##
## Call:
## lm(formula = hours ~ taxrate, data = Mdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -207.97  -56.45  -13.98   32.42  264.44
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1748.5      111.7   15.65 4.02e-11 ***
## taxrate      -1122.4      369.2   -3.04   0.0078 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 117.7 on 16 degrees of freedom
## Multiple R-squared:  0.3661, Adjusted R-squared:  0.3265
## F-statistic: 9.241 on 1 and 16 DF,  p-value: 0.007803
```

Estimated Regression Equation: Hours_worked = 1748.5 + (-1122.4)*tax_rate Intepretation: An one-percent increase in tax rate decreases hours worked by men by 11.22 hours (tax rate on a sclae of 0-1). In a country where tax rate is 0, men work 1748.5 hours.

Differences: Men work less hours when tax rate increases while women work more hours.

(c) What are the fitted value and residual for men in the United States?

```
my_regmen1$fitted.values[18]
```

```
##      18
## 1470.12
```

```
my_regmen1$residuals[18]
```

```
##       18
## 142.8799
```

(d) What are the fitted value and residual for women in Italy?

```
my_regwomen1$fitted.values[9]
```

```
##       9
## 841.535
```

```r
my_regwomen1$residuals[9]
```

```
##        9
## -186.935
```