a) We restrict our attention to a subset of individuals and variables. Keep only the observations:

   (a) On weekly earnings, usual hours, sex, race, and age. The corresponding variable names are earnwke, uhourse, sex, age, and race.

   (b) For respondents aged 25-65

   (c) From Massachusetts

Answer in code, lines 4-5

b) Data cleaning: Drop all the NA data. Create a dummy variable equal to one if the worker worked more than 0 hours.

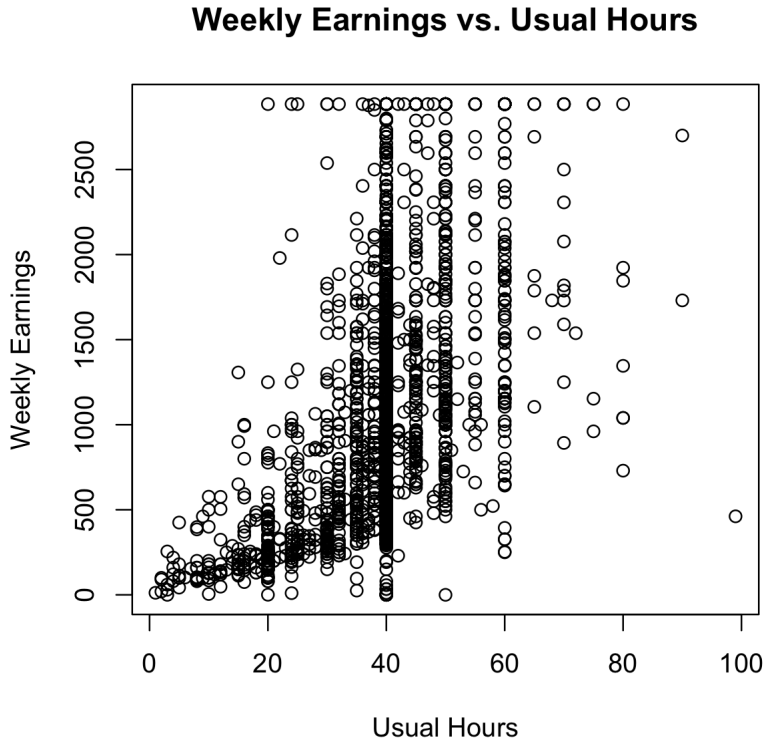
   (a) What is the share of workers working more than 0 hours?

   (b) What is the mean number of hours worked?

   Restrict to workers usually working positive hours and calculate the mean number of hours worked in the restricted sample. Compare the two values (restricted and non-restricted sample)

   Share of workers working more than 0 hours = 0.9476933
   Mean hours of work in the restricted sample = 39.50908
   Mean hours of work in the non-restricted sample = 37.23327

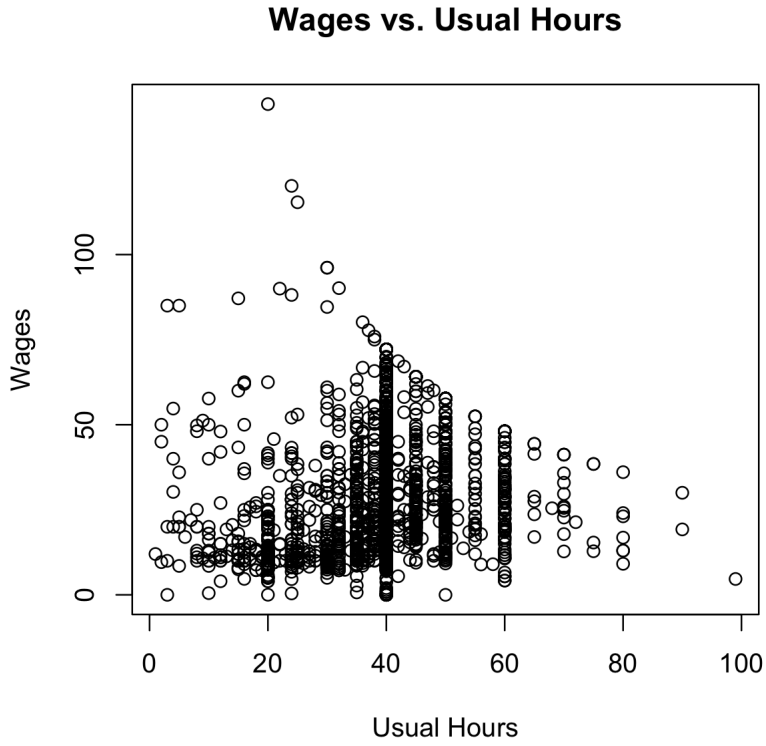
c) Plot the weekly earnings of individuals against the usual hours.



d) Create a measure of wages by dividing weekly earnings by usual hours. Create a new column called 'wages' for it.

Answer in code, lines 26-27

e) Plot our measure of wages against the usual hours worked and compare Plot 1 and Plot 2.



Comparing Plot 1 and Plot 2

1. **Scale of the Y-Axis**: The first plot presents "Weekly Earnings," which seems to have a larger range with values exceeding 2000. The second plot presents "Wages," which has a much lower range capped at around 100. The difference is trivial as the first one is weekly and the second one is hourly.

2. **Spread and Density of Data Points**:

- In the "Weekly Earnings" plot, there is a denser cluster of data points at the lower end of the usual hours, with the range of earnings widely spread out. This could indicate a wide variation in weekly earnings among individuals working fewer hours.

- In the "Wages" plot, the data points also cluster towards the lower end of usual hours, but the range of hourly wages is narrower, suggesting that hourly wages may vary less than total weekly earnings.

3. **Trend Analysis**:

- The "Weekly Earnings" plot might show that as the number of hours increases, the total weekly earnings increase, but not necessarily at a constant rate.

- The "Wages" plot could indicate that hourly wages may not significantly increase with more hours worked, as seen by the dense clustering of data points across the range of hours without a clear upward trend.

4. **Outliers**:

- There seem to be more noticeable outliers in the "Weekly Earnings" plot, where some individuals earn significantly more than others in the same usual hours range.

- The "Wages" plot seems to have fewer extreme outliers, possibly because the variation in hourly wages is naturally less extreme than the variation in total earnings.

5. **Relationship Between Hours and Earnings/Wages**:

- The first graph might be used to analyze factors such as overtime pay or the effect of working part-time versus full-time on total earnings.

- The second graph is likely more useful for evaluating whether hourly wages change with an increase in the number of hours worked, which can indicate if overtime is paid at the same rate as regular time or if there is a differential.

f) Plot our measure of wages against the usual hours worked for the identified male and the identified female separately. Overlay a regression line in your plot. Compare the relation for the identified male and female.

1. **Data Distribution**: Both male and female data points are scattered across the plot, with a concentration of points at the lower end of usual hours. This suggests that there are more observations of both genders working fewer hours per week.

2. **Wage Levels**: There is a wide range of wages for both genders. However, it appears that there are more high-wage outliers among the male-identified sample (gray points) than the female-identified sample (red points), as evidenced by some gray points extending towards the higher end of the wage axis.

3. **Regression Lines**: We can see that the Male regression line (in blue) is above the Female regression line (in magenta). This means that, on average, the Male sample had a higher wage than the female sample.

4. **Variance**: The Variance in wages, as can be seen by the spread of points around the regression line, is notable for both genders. This spread indicates that factors other than usual hours worked contribute to wage differences.

g) Run a regression with wages as dependent variable and age as independent variable for the sample of male. Do the same thing for the sample of female.

Here is a screenshot of the tables. The full ones will be in the R file (**lines 48-50** of "Econometrics Project.R").

Regression for the Sample of male

regression_male	list [12] (S3: lm)	List of length 12
coefficients	double [2]	17.933 0.306
residuals	double [1470]	21.732 12.709 -7.973 0.262 -14.344 6.775 ...
effects	double [1470]	-1.20e+03 1.38e+02 -8.83e+00 3.34e-03 -1.53e+01 5.97e+00 ...
rank	integer [1]	2
fitted.values	double [1470]	36.0 27.1 27.7 34.7 26.8 28.3 ...
assign	integer [2]	0 1
qr	list [5] (S3: qr)	List of length 5
df.residual	integer [1]	1468
xlevels	list [0]	List of length 0
call	language	lm(formula = wages ~ age, data = male_data)
terms	formula	wages ~ age
model	list [1470 x 2] (S3: data.frame	A data.frame with 1470 rows and 2 columns

We can see that the constant is 17.933 and the slope is 0.306

Regression for the sample of female

regression_female	list [12] (S3: lm)	List of length 12
coefficients	double [2]	21.098 0.118
residuals	double [1447]	-14.32 1.48 -7.64 -9.52 -9.87 -13.35 ...
effects	double [1447]	-999.10 -52.04 -7.31 -9.09 -9.50 -12.98 ...
rank	integer [1]	2
fitted.values	double [1447]	28.7 27.4 27.4 24.5 26.1 26.2 ...
assign	integer [2]	0 1
qr	list [5] (S3: qr)	List of length 5
df.residual	integer [1]	1445
xlevels	list [0]	List of length 0
call	language	lm(formula = wages ~ age, data = female_data)
terms	formula	wages ~ age
model	list [1447 x 2] (S3: data.frame	A data.frame with 1447 rows and 2 columns

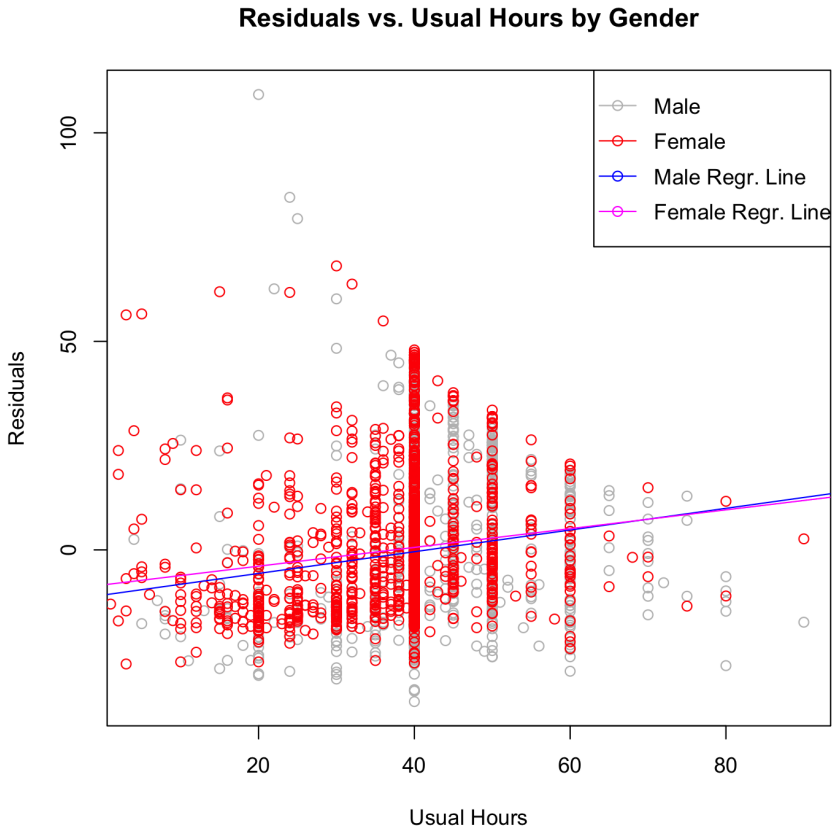
We can see that the constant is 21.098 and the slope is 0.118

h) Report your results of OLS by a formated table. Two columns. One for each regression results.

Formatted table. Regression male is 1 and Regression female is 2

	Dependent variable:	
	wages	
	(1)	(2)
age	0.306***	0.118***
	(0.038)	(0.035)
Constant	17.933***	21.098***
	(1.725)	(1.588)
Observations	1,470	1,447
R2	0.041	0.008
Adjusted R2	0.041	0.007
Residual Std. Error	17.389 (df = 1468)	15.457 (df = 1445)
F Statistic	63.289*** (df = 1; 1468)	11.333*** (df = 1; 1445)
Note:		*p<0.1; **p<0.05; ***p<0.01

i) Generate the residuals from the regression for the sample identified male and the sample of identified female. Run a regression of the residuals you just generated against the usual hours worked for the identified male and the identified female separately. Is there a difference between the plot in f and in this question.



The graph in this question and the one in question f are practically identical. Of

course, there are discrepancies in the y-axis values but the pattern is pretty similar. As Residual=Observed Value-Predicted Value, if the variance of wages is large compared to the size of the residuals, then the residuals plot may look similar because the residuals don't deviate much from the wages, visually speaking. Furthermore, we can see that the regression lines in part f are considerably flat, aiding more in this similitude between plots. The scale of the y-axis in both plots can affect this perception.

The slope of the regression line in i is slightly more pronounced than the one in part f

In addition, a difference between the plots is that we can see that the residuals in male and female data stay the same, in average, across the plot (this is represented by the regression lines). On the contrary, the regression lines in the plot from part f separate from each other as the hours increase.