

# Problem Set 2 Empirical Methods

Patrick Gletting (13-252-143), Carmine Ragone (14-995-021)

05.11.2018

## Paper and Pencil Questions

(a) Show that  $TSS = ESS + RSS$ . Also, show you can write  $R^2 = 1 - \frac{e'e}{\tilde{y}'\tilde{y}}$  where  $\tilde{y}_i = y_i - \bar{y}$ .

$$\begin{aligned}\sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (y_i - \hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2\end{aligned}$$

We need to show that  $2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$ . By the regression equation we know that:

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

As we minimize SSE to find  $\beta_1, \beta_0$  we search for the partial derivatives:

$$\begin{aligned}\frac{\delta SSE}{\delta \beta_0} &= \sum_{i=1}^n 2(y_i - \beta_0 - \beta_1 x_i)(-1) \stackrel{!}{=} 0 \\ \sum_{i=1}^n \beta_0 &= \sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i \Rightarrow 0 = \sum_{i=1}^n (y_i - \hat{y}_i) \\ n\beta_0 &= \sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i \\ \beta_0 &= \frac{1}{n} \sum_{i=1}^n y_i - \beta_1 \frac{1}{n} \sum_{i=1}^n x_i\end{aligned} \tag{a}$$

We take the second partial derivative:

$$\frac{\delta SSE}{\delta \beta_1} = \sum_{i=1}^n 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) \stackrel{!}{=} 0 \sum_{i=1}^n 2(y_i - \beta_0 - \beta_1 x_i) = 0 \tag{1}$$

By the regression equation  $\hat{y}_i = \beta_0 + \beta_1 x_i$ , thus  $\frac{\hat{y}_i - \beta_0}{\beta_1}$  and so:

$$\sum_{i=1}^n x_i (y_i - \hat{y}_i) = 0 \sum_{i=1}^n \left( \frac{\hat{y}_i - \beta_0}{\beta_1} \right) (y_i - \hat{y}_i) = 0 \frac{1}{\beta_1} \sum_{i=1}^n \hat{y}_i (y_i - \hat{y}_i) - \frac{\beta_0}{\beta_1} \sum_{i=1}^n (y_i - \hat{y}_i) = 0 \tag{2}$$

Using equation (a) in the second part:

$$\sum_{i=1}^n \hat{y}_i (y_i - \hat{y}_i) = 0 \quad (b)$$

We needed to show that  $2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$ :

$$\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^n \hat{y}_i (y_i - \hat{y}_i) - \bar{y} \sum_{i=1}^n (y_i - \hat{y}_i)$$

Where the first part is 0 because of (b) and the second part because of (a).

For the second part of the question, we know that  $e'e = \sum e_i^2 = RSS$  and  $\tilde{y}'\tilde{y} = \sum y_i - \bar{y} = TSS$ . As  $ESS = TSS - RSS$  and  $TSS = RSS + ESS$  this becomes:

$$\begin{aligned} R^2 &= \frac{ESS}{TSS} \\ &= \frac{TSS - RSS}{TSS} \\ &= \frac{TSS}{TSS} - \frac{RSS}{TSS} \\ &= 1 - \frac{RSS}{TSS} \end{aligned}$$

(b) Show that  $R^2 = \text{corr}^2(y, \hat{y})$ . What is the intuition behind it?

$$\begin{aligned} \text{corr}^2(y, \hat{y}) &= \left( \frac{\text{Cov}(y, \hat{y})}{\sqrt{\text{Var}(y)\text{Var}(\hat{y})}} \right)^2 \\ &= \frac{(\text{Cov}(y, \hat{y}))^2}{\text{Var}(y)\text{Var}(\hat{y})} \\ &= \frac{(\text{Cov}(\hat{y} + e, \hat{y}))^2}{\text{Var}(y)\text{Var}(\hat{y})} \\ &= \frac{(\text{Cov}(\hat{y}, \hat{y}) + \text{Cov}(\hat{y}, e))^2}{\text{Var}(y)\text{Var}(\hat{y})} \\ &= \frac{(\text{Cov}(\hat{y}, \hat{y}))^2}{\text{Var}(y)\text{Var}(\hat{y})} \\ &= \frac{(\text{Var}(\hat{y}))^2}{\text{Var}(y)\text{Var}(\hat{y})} \\ &= \frac{\text{Var}(\hat{y})}{\text{Var}(y)} = \frac{ESS}{TSS} = R^2 \end{aligned}$$

The intuition behind this is how much of the total variance can be explained with our model that yields  $\text{Var}(\hat{y})$ .

(c) Suppose you decided to measure all of your  $X$  variables in different units such that your new  $X$  variable, call it  $\tilde{X}$ , is exactly double your old one, i.e.  $\tilde{X} = 2X$ . Suppose you run the regression of  $y$  on  $\tilde{X}$ ; call the resulting estimate  $\tilde{\beta}$ . You showed in Problem Set 1 that  $\tilde{\beta} = \frac{1}{2}\hat{\beta}$ . Is the  $R^2$  different in the two models? Provide an intuitive answer.

We have seen above that  $R^2 = \frac{\text{Var}(\hat{y})}{\text{Var}(y)}$ . The denominator will not change as  $y$  is not affected.  $\text{Var}(\hat{y})$  is given by the model  $\hat{y} = \beta_0 + \tilde{\beta}x$ . Using the hint from Problem Set 1,  $\hat{y} = \beta_0 + \frac{1}{2}\hat{\beta}2x$ , which yields the same model

as before. In other words, the new estimator  $\tilde{\beta}$  adjusts for the new units in  $X$ . Thus  $R^2$  is not going to change.

(d) *Intuitively discuss the fact that including another regressor in the linear model always decreases the  $RSS$ .*

Intuitively, if I keep adding another regressor in the linear model our model the variance left (which is approximated by the  $RSS$ ) will be less. More formally, by adding another variable, we add another covariance to the model which will mechanically always increase  $Var(\hat{y})$  and hence  $ESS$ . As  $RSS = TSS - ESS$ , this will always decrease  $RSS$ .

(e) *Provide a formal proof of point (d).*

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})}$$

$y, \bar{y}$  is not affected by a change in  $x$ . So we focus on  $\hat{y}$ :

$$\begin{aligned}\hat{y} &= \beta_0 + \tilde{\beta}x \\ \hat{y} &= \beta_0 + \frac{1}{2}\hat{\beta}_2x\end{aligned}$$

So  $\hat{y}$  remains unchanged and none of the variables in the equation for  $R^2$  changes.

(f) *Can you suggest a problem of interpreting the  $R^2$  as a measure of how “good” the model is? If you think the model might not be “good”, why might it nevertheless have a high  $R^2$ ?*

A good model identifies causal relationships between the dependent variable  $y$  and the independent variables  $x$ . By adding more independent variables, we increase  $R^2$  (seen above), but this tells us nothing about the causal relationship between  $y$  and  $x$ . To identify causal relationships, we need  $x$  variables with robust standard errors, however  $R^2$  is not affected by the errors in  $x$ . This means we can have a “bad” model which still has a high  $R^2$ . The problem is that  $R^2$  tells us how well our model fits the sample. If we only try to increase  $R^2$ , we have a tendency to overfit the model to the sample at hand, but might ignore the causal relationships of the true population.

## The gender wage gap

*Suppose you want to test whether in your country women are discriminated against relative to men in terms of wages. You decide that you want to test whether men and women have different salaries. Suppose you are able to gather data on the whole working population in your country. For each individual you have the following information.*

- *monthly wage*
- *gender*
- *years of education*

(a) *Suppose years of education have the same effect on wages for both men and women. Propose a simple regression model to test your hypothesis.*

$$Wage_i = \alpha + \beta_1 Edu_i + \beta_2 Male_i + \epsilon_i$$

Where  $Male$  is a dummy variable which takes value 0 for females and 1 for male observations.

(b) *Provide a graphical representation of the conditional expectation function (i.e. the part of wages that we can explain with our covariates) and show if and how it differs for men and women.*

$$E(Wage_i | Edu_i, Male_i) = \begin{cases} \alpha + \beta_1 Edu_i + \epsilon_i, & \text{if } Male_i = 0 \\ \alpha + \beta_1 Edu_i + \beta_2 + \epsilon_i, & \text{if } Male_i = 1 \end{cases}$$

(c) In retrospect, you decide that years of education might have a different marginal effect on men compared to women. How would you modify your regression model to account for this differential effect?

$$\text{Wage}_i = \alpha + \beta_1 \text{Edu}_i + \beta_2 \text{Male}_i + \beta_3 \text{Edu}_i \times \text{Male}_i + \epsilon_i$$

(d) Provide a graphical representation of the conditional expectation function (i.e. the part of wages that we can explain with our covariates) and show if and how it differs for men and women

$$E(\text{Wage}_i | \text{Edu}_i, \text{Male}_i) = \begin{cases} \alpha + \beta_1 \text{Edu}_i + \epsilon_i, & \text{if } \text{Male}_i = 0 \\ \alpha + \beta_1 \text{Edu}_i + \beta_2 + \beta_3 \text{Edu}_i + \epsilon_i, & \text{if } \text{Male}_i = 1 \end{cases}$$

## Empirical Application

**1. The Gender Wage Gap.** In this exercise we will try to explore some discrimination theories analyzing a subsample from the US CPS2015. Many politicians, institutional observers, and researchers still claim today the existence of discrimination against female workers in the labor market. They base their claims looking at the gender wage gap, i.e. the difference between men's and women's wages. As many other things in economics, this wage gap can be generated both from the demand side (employers who discriminate against women) and from the supply side (women having different skills or preferences for specific jobs or for entering the labor market at all). In this exercise we will try to learn more about the gender wage gap, while testing you on your econometric toolkit. For this question, assume that Assumption 2 (Mean-zero Error) holds so that you can make causal statements in your answers.

Download the dataset `sampleUScens2015.csv` from OLAT and import it into Stata or R. The dataset includes prime age individuals (i.e.  $\text{age} \in [25; 54]$ ) active in the labor market (i.e. either employed or looking for job), and working in the private sector. There are seven relevant variables:

- *age*, the age of the individual in 2015
- *education*, years of completed education
- *incwage*, income from wages in 2015 in USD
- *female*, dummy for female
- *childrenly*, dummy if had a children in the last year
- *degfield*, field of degree
- *occupation*, sector of occupation

At first we need to load the data in order to start our analysis:

```
USCPS <- fread("sampleUScens2015.CSV")
head(USCPS) # Check the format
##   age incwage female childrenly educ      firmsector occupation
## 1:  40  18200      0          0 10.0          other      other
## 2:  49  50000      0          0 14.0      manufacturing  other
## 3:  28  11300      1          0 15.0 professionalServices  other
## 4:  43  62000      0          0 14.0          other      other
## 5:  39  60000      0          0 14.5      transportation  other
## 6:  27  20000      0          0 13.0          other      other
##   degfield
## 1:   other
## 2:   other
## 3:   other
## 4:   other
## 5:   other
## 6:   other
dim(USCPS) #Check length
## [1] 561121      8
```

```
summary(USCPS) #Explore the data
##      age      incwage      female      childrenly
## Min.   :25.00   Min.    :    4   Min.    :0.0000   Min.    :0.00000
## 1st Qu.:31.00   1st Qu.: 22000   1st Qu.:0.0000   1st Qu.:0.00000
## Median :38.00   Median : 39000   Median :0.0000   Median :0.00000
## Mean   :38.59   Mean    : 53970   Mean    :0.4317   Mean    :0.01766
## 3rd Qu.:46.00   3rd Qu.: 65000   3rd Qu.:1.0000   3rd Qu.:0.00000
## Max.   :54.00   Max.    :658000   Max.    :1.0000   Max.    :1.00000
##
##      educ      firmsector      occupation      degfield
## Min.    : 0.00   Length:561121   Length:561121   Length:561121
## 1st Qu.:14.00   Class :character   Class :character   Class :character
## Median :15.00   Mode  :character   Mode  :character   Mode  :character
## Mean    :15.83
## 3rd Qu.:18.00
## Max.    :23.00
## NA's    :45
```

The summary tells us that there are some NAs in *educ*, we want to explore this:

```
head(USCPS[rowSums(is.na(USCPS)) > 0, ])
##      age incwage female childrenly educ      firmsector occupation
## 1:  26   12000      1          0   NA      wholesale      other
## 2:  48   50000      0          0   NA professionalServices  other
## 3:  48   50000      0          0   NA professionalServices  other
## 4:  48   50000      0          0   NA professionalServices  other
## 5:  48   50000      0          0   NA professionalServices  other
## 6:  47   42000      0          0   NA      other          other
##      degfield
## 1:   other
## 2:   other
## 3:   other
## 4:   other
## 5:   other
## 6:   other
```

The other columns seem fine, we could still keep these observations in the dataset. However, the data quality on these 45 observations seems poor anyway as *firmsector*, *occupation* and *degfield* are in most cases “other”, which is not really informative. Also, 45 observations are just a tiny fraction of the overall sample. As it is easier to work with a complete dataframe, we drop the observations containing NAs:

```
USCPS <- USCPS %>% drop_na()
```

(a) Generate a new variable called *wage* = *incwage*/1000. Also, generate *lw* taking the log of *wage*. Generate a dummy named *university* which is equal to 1 if education  $\geq 16$ . First regress *wage* on education, then regress *wage* on education and the *university* dummy. How does the coefficient on education change? How do you interpret it in both specifications?

Create the variables

```
USCPS$wage <- USCPS$incwage/1000 #create wage
USCPS$lw <- log(USCPS$wage) #create log wage
USCPS$university <- as.numeric(USCPS$educ >= 16) #create university dummy
```

Run the regressions

```

Model1 = wage ~ educ
lm1 <- lm(Model1, data = USCPS, x = TRUE)
Model2 = wage ~ educ + university
lm2 <- lm(Model2, data = USCPS, x = TRUE)

stargazer(lm1, lm2, title = "Regression Results", type = "latex",
  header = FALSE, align = TRUE, dep.var.labels = c("Wage"),
  covariate.labels = c("Education", "University"), omit.stat = c("all"),
  no.space = TRUE)

```

Table 1: Regression Results

	<i>Dependent variable:</i>	
	Wage	
	(1)	(2)
Education	7.074*** (0.028)	4.975*** (0.045)
University		15.180*** (0.254)
Constant	-58.037*** (0.446)	-31.605*** (0.628)

*Note:* \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Without the university dummy, a 1 year increase in education is reflected by a wage increase of 7.074 USD. By adding the dummy *university* we separate the effect of years of education from years of university education. This results in a decreased coefficient for *Education*, 4.975 whereas the dummy term tells that people who attended university earn on average 15.18 USD more. By the drop in *Education* we learn that 7 USD return for an additional year of education was at least partially driven by higher earnings of workers who attended university.

(b) Drop the university dummy. Now regress wage on education and age. Also, regress log wages (*lw*) on education and age. What are their coefficients? How do you interpret them? How do they compare? [Note: be sure you compare approximately equivalent objects from each specification.]

```

Model3 = wage ~ educ + age
lm3 <- lm(Model3, data = USCPS, x = TRUE)
Model4 = lw ~ educ + age
lm4 <- lm(Model4, data = USCPS, x = TRUE)

stargazer(lm3, lm4, title = "Regression Results", type = "latex",
  header = FALSE, align = TRUE, dep.var.labels = c("Wage",
    "Log Wage"), covariate.labels = c("Education", "Age"),
  no.space = TRUE)

```

**Interpretation Model 1:** One year increase in education controlling for age increases wage by 7.17USD. One year increase in age controlling for education increases wage by 1.461USD. **Interpretation Model 2:** One year increase in education controlling for age increases wage by 12%. One year increase in age controlling for education increases wage by 2.5%.

Although their interpretation is different, they are in fact the same model. We can see this by comparing the sum of the fitted values:

Table 2: Regression Results

	<i>Dependent variable:</i>	
	Wage	Log Wage
	(1)	(2)
Education	7.170*** (0.027)	0.120*** (0.0004)
Age	1.461*** (0.009)	0.025*** (0.0001)
Constant	-115.916*** (0.554)	0.713*** (0.009)
Observations	561,076	561,076
R <sup>2</sup>	0.147	0.158
Adjusted R <sup>2</sup>	0.147	0.158
Residual Std. Error (df = 561073)	57.312	0.920
F Statistic (df = 2; 561073)	48,428.710***	52,686.040***

Note:

\*p&lt;0.1; \*\*p&lt;0.05; \*\*\*p&lt;0.01

```
sum(fitted.values(lm3))
## [1] 30281173
sum(fitted.values(lm4))
## [1] 1997072
```

(c) Now regress log wages on education, age, and the female dummy. You get the following model:

$$lw_i = \beta_1 + \beta_2 educ_i + \beta_3 age_i + \beta_4 female_i + \epsilon_i$$

What is the coefficient on female? How do you interpret it? Is it economically significant in your opinion? Test both in R/Stata and “by hand” the hypothesis that  $\beta_4 = 0$ . Should you use a one-sided or two-sided test? Do the one you think most appropriate.

```
Model5 = lw ~ educ + age + female
lm5 <- lm(Model5, data = USCPS, x = TRUE)

stargazer(lm5, title = "Regression Results", type = "latex",
  header = FALSE, align = TRUE, dep.var.labels = c("Log Wage"),
  covariate.labels = c("Education", "Age", "Female"), no.space = TRUE,
  p.auto = TRUE)
```

Females earn on average 44.25% less than males with the same education and age. This is economically significant, by using the average wage and multiplying with the female coefficient we get 23.8816475, a substantial difference.

To test in R whether  $\beta_4 = 0$ , we can simply look at the p-value of the female coefficient and see that is well below 5%. Using  $\alpha = 5\%$ , we can state that  $\beta_4$  is significantly different from 0.

By hand:

$$H_0 : \beta_4 = 0 \text{ and } H_1 : \beta_4 \neq 0$$

Calculate degrees of freedom:

$$N - K = 561076 - 4 = 561072$$

Construct test statistic:

Table 3: Regression Results

	Dependent variable:
	Log Wage
Education	0.126*** (0.0004)
Age	0.024*** (0.0001)
Female	-0.443*** (0.002)
Constant	0.815*** (0.009)
Observations	561,076
R <sup>2</sup>	0.206
Adjusted R <sup>2</sup>	0.206
Residual Std. Error	0.894 (df = 561072)
F Statistic	48,385.030*** (df = 3; 561072)
Note:	*p<0.1; **p<0.05; ***p<0.01

$$\frac{\hat{\beta}_4 - \beta_4}{\text{stderr}(\hat{\beta}_4)} = \tilde{t} = \frac{-0.4427 - 0}{0.0024} = -184.46$$

Now we look up in a t-table to find the critical value to find that the critical value  $\bar{t}_{-184.46} = -1.967$  for  $\alpha = 5\%$ . As  $-184.46 < -1.967$ , we can reject  $H_1$  and state that the coefficient is significantly different from 0 at the 5% confidence level. We used a two-sided test here because the question was to identify if the effect is different from 0, which can go both ways.

(d) Use R/Stata to get  $\beta_4$  (the coefficient on female) using partitioned regression as we did in lecture. Given that:

$$\begin{aligned} Y &= \beta X + \epsilon_i \\ &= X_k \beta_k + X_{-k} \beta_{-k} + \epsilon_i \end{aligned}$$

The idea of partitionated regression is to show that The  $k^{th}$  coefficient in a multiple OLS regression is equivalent to the coefficient in a simple OLS regression of the y on the residual from a regression of  $X_k$  on all the other regressions. In this case, the first step is to regress the fact of being female on the education and age to see how much of the variance is explained by those.

```
Model6 = female ~ educ + age
lm6 <- lm(Model6, data = USCPS, x = TRUE)

stargazer(lm6, title = "Regression Results", type = "latex",
  header = FALSE, align = TRUE, dep.var.labels = c("Female"),
  covariate.labels = c("Education", "Age"), no.space = TRUE,
  p.auto = TRUE)
##
## \begin{table}[!htbp] \centering
## \caption{Regression Results}
## \label{}
```



```
## \begin{tabular}{@{\extracolsep{5pt}}lD{.}{.}{-3} }
## \[-1.8ex\]\hline
## \hline \[-1.8ex]
## & \multicolumn{1}{c}{\textit{Dependent variable:}} \\\
## \cline{2-2}
## \[-1.8ex] & \multicolumn{1}{c}{Female} \\\
## \hline \[-1.8ex]
## Education & 0.015^{***} \\\
## & (0.0002) \\\
## Age & -0.001^{***} \\\
## & (0.0001) \\\
## Constant & 0.229^{***} \\\
## & (0.005) \\\
## \hline \[-1.8ex]
## Observations & \multicolumn{1}{c}{561,076} \\\
## R^2 & \multicolumn{1}{c}{0.008} \\\
## Adjusted R^2 & \multicolumn{1}{c}{0.008} \\\
## Residual Std. Error & \multicolumn{1}{c}{0.493 (df = 561073)} \\\
## F Statistic & \multicolumn{1}{c}{2,255.955^{***}} (df = 2; 561073) \\\
## \hline
## \hline \[-1.8ex]
## \textit{Note:} & \multicolumn{1}{r}{*} p < $0.1; **} p < $0.05; ***} p < $0.01} \\\
## \end{tabular}
## \end{table}
```

Log wage is regressed on the residuals of the last regression and it is possible to see that in both Model 7 and Model 5 the coefficient of the female and of the residuals is the same.

```
fem_res <- residuals(lm6) # Save the female residual values

Model7 = lw ~ fem_res
lm7 <- lm(Model7, data = USCPS, x = TRUE)
stargazer(lm7, title = "Regression Results", type = "latex",
  header = FALSE, align = TRUE, dep.var.labels = c("Log Wage"),
  covariate.labels = c("Female Residuals"), no.space = TRUE,
  p.auto = TRUE)

##
## \begin{table}[!htbp] \centering
## \caption{Regression Results}
## \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lD{.}{.}{-3} }
## \[-1.8ex\]\hline
## \hline \[-1.8ex]
## & \multicolumn{1}{c}{\textit{Dependent variable:}} \\\
## \cline{2-2}
## \[-1.8ex] & \multicolumn{1}{c}{Log Wage} \\\
## \hline \[-1.8ex]
## Female Residuals & -0.443^{***} \\\
## & (0.003) \\\
## Constant & 3.559^{***} \\\
## & (0.001) \\\
## \hline \[-1.8ex]
## Observations & \multicolumn{1}{c}{561,076} \\\
## R^2 & \multicolumn{1}{c}{0.047} \\\
```

```
## Adjusted R2 & \multicolumn{1}{c}{0.047} \\
## Residual Std. Error & \multicolumn{1}{c}{0.979 (df = 561074)} \\
## F Statistic & \multicolumn{1}{c}{27,933.880***} (df = 1; 561074) \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:} & \multicolumn{1}{r}{*p<$0.1; **p<$0.05; ***p<$0.01} \\
## \end{tabular}
## \end{table}
```

This is the demonstration that the only variation left in the variable “female to identify its coefficient is the variation left after running a regression of female on education and age.

(e) Use R/Stata to show that  $\hat{\beta}_1 = \bar{y} - \bar{X}'_1 \hat{\beta}_{-1}$

```
Beta_hat <- c(summary(lm5)$coefficients[2, 1], summary(lm5)$coefficients[3,
  1], summary(lm5)$coefficients[4, 1])
X_bar <- c(mean(USCPS$educ), mean(USCPS$age), mean(USCPS$female))
Beta_hat_1 <- mean(USCPS$lw) - sum(Beta_hat * X_bar)
print(Beta_hat_1) #are calculated constant
## [1] 0.8145355
print(summary(lm5)$coefficients[1, 1]) #the constant from the model
## [1] 0.8145355
```

(f) Include in the model in (1c) the interaction between female and education, together with the interaction between female and age. So your model is now:

$$lw_i = \beta_1 + \beta_2 educ_i + \beta_3 age_i + \beta_4 female_i + \beta_5 female_i \times educ_i + \beta_6 female_i \times age_i + \epsilon_i$$

Test in R/Stata the individual hypotheses that  $\beta_4 = 0$ ,  $\beta_5 = 0$ , and  $\beta_6 = 0$ . Test “by hand” and in R/Stata the joint hypothesis that they are all zero.

```
Model18 = lw ~ educ + age + female + female:educ + female:age
lm8 <- lm(Model18, data = USCPS, x = TRUE)

stargazer(lm8, title = "Regression Results", type = "latex",
  header = FALSE, align = TRUE, dep.var.labels = c("Log Wage"),
  covariate.labels = c("Education", "Age", "Female", "Female x Education",
    "Female x Age"), no.space = TRUE, p.auto = TRUE)
```

$\beta_4, \beta_5, \beta_6$  are significantly different from zero with a p-value of well below 5%. For the joint significance test, we first formulate our hypothesis:

$$H_0 : \beta_4, \beta_5 \text{ and } \beta_6 = 0$$

$$H_1 : \beta_4 \text{ or } \beta_5 \text{ or } \beta_6 \neq 0$$

Restricted Model:  $lw_i = \beta_1 + \beta_2 educ_i + \beta_3 age_i + \epsilon_i$

Unrestricted Model:  $lw_i = \beta_1 + \beta_2 educ_i + \beta_3 age_i + \beta_4 female_i + \beta_5 female_i \times educ_i + \beta_6 female_i \times age_i + \epsilon_i$

$$F = \frac{(R_U^2 - R_R^2)/q}{(1 - R_U^2)(N - K)}$$

Table 4: Regression Results

	<i>Dependent variable:</i>
	Log Wage
Education	0.120*** (0.001)
Age	0.028*** (0.0002)
Female	-0.342*** (0.018)
Female x Education	0.017*** (0.001)
Female x Age	-0.009*** (0.0003)
Constant	0.759*** (0.011)
Observations	561,076
R <sup>2</sup>	0.208
Adjusted R <sup>2</sup>	0.208
Residual Std. Error	0.893 (df = 561070)
F Statistic	29,436.260*** (df = 5; 561070)
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

We identify the single elements:  $R_U^2$  is 0.2078094 from the summary statistics of lm6. Same for  $R_R^2$ , found in the summary statistics of lm4: 0.1581107.

$$DF_U = 561076 - 6 = 561070 \quad DF_R = 561076 - 3 = 561073 \quad q = DF_R - DF_U = 3$$

$$F = \frac{(0.207 - 0.158)/3}{(1 - 0.207)(561076 - 6)} = 11556.3$$

Looking up in a F-table, we find the critical value to be 1 at an  $\alpha = 5\%$ , we can reject the null hypothesis.

To do the same in R:

```
linearHypothesis(lm8, c("female + educ:female + age:female"))
## Linear hypothesis test
##
## Hypothesis:
## female + educ:female + age:female = 0
##
## Model 1: restricted model
## Model 2: lw ~ educ + age + female + female:educ + female:age
##
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1 561071 447260
## 2 561070 446949   1    310.94 390.33 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we look up in a t-table to find the critical value to find that the critical value  $\bar{t}_{-184.46} = -1.967$  for  $\alpha = 5\%$ . As  $-184.46 < -1.967$ , we can reject  $H_1$  and state that the coefficient is significantly different from

0 at the 5% confidence level. We used a two-sided test here because the question was to identify if the effect is different from 0, which can go both ways.