Problem Set 2 Empirical Methods

Patrick Glettig (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 β_1, β_0 we search for the partial derivatives:

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

We take the second partial derivative:

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

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

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

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

Using equation (a) in the second part:

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

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

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

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

(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 $R^2 = \frac{Var(\hat{y})}{Var(y)}$. The denominator will not change as y is not affected. $Var(\hat{y})$ 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.

Adding other regressors reduces the residual variance (as we minimize RSS) . 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).

$y = X\beta + \epsilon$ gives residuals e and R_1^2 $y = X_z\beta_z + z\gamma + \upsilon$ gives residuals u and R_2^2

It is possible to rewrite $y = X_z\beta_z + \upsilon$ as $y = (X\beta + z\gamma) + \upsilon$

Adding variables like $Z\gamma$ increases R^2 so we will have that $R_2^2 > R_1^2$ Thus we have that:

$$1 - \frac{e'e}{\tilde{y}'\tilde{y}} < 1 - \frac{u'u}{\tilde{y}'\tilde{y}}$$

And we have that:

$$\begin{aligned} e &= y - X\hat{\beta} \\ &= y - X(X'X)^{-1}X'y \\ &= (I_N - X(X'X)^{-1}X')y \\ &= I_N - P_X y = M_x y \end{aligned}$$

and we can demonstrate as well that:

$$\begin{aligned} u &= y - (X\widehat{\beta} + z\widehat{\gamma}) \\ &= y - X(X'X)^{-1}X'y - z\widehat{\gamma} \\ &= y - P_X y - z\widehat{\gamma} \\ &= e - z\widehat{\gamma} \end{aligned}$$

Given that:

$$X(X'X)^{-1}X = P_X$$

We know that $z\widehat{\gamma}$ is postive and non zero because:

$$\widehat{\gamma} = (z'M_k z)^{-1}z'M_k y = (\sum z^2)\sum zy$$

At this point we can clearly see that:

$$u'u < e'e$$

The demonstration of $\widehat{\gamma} = (z'M_k z)^{-1} z'M_k y$ is possible if we compute it using partitioned regression. The demonstration can be found in the last page of this document.

(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.

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

The difference between men and women is β_2. For the graphical representation see Figure 1.

(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?

$$Wage_i = \alpha + \beta_1 Edu_i + \beta_2 Male_i + \beta_3 Edu_i \times Male_i + \epsilon_i$$

(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

$$E(Wage_i|Edu_i, Male_i) = \begin{cases} \alpha + \beta_1 Edu_i, & \text{if } Male_i = 0 \\ \alpha + \beta_1 Edu_i + \beta_2 + \beta_3 Edu_i & \text{if } Male_i = 1 \end{cases}$$

The difference between men and women is $\beta_2 + \beta_3 Edu_i$. For the graphical representation see Figure 2.

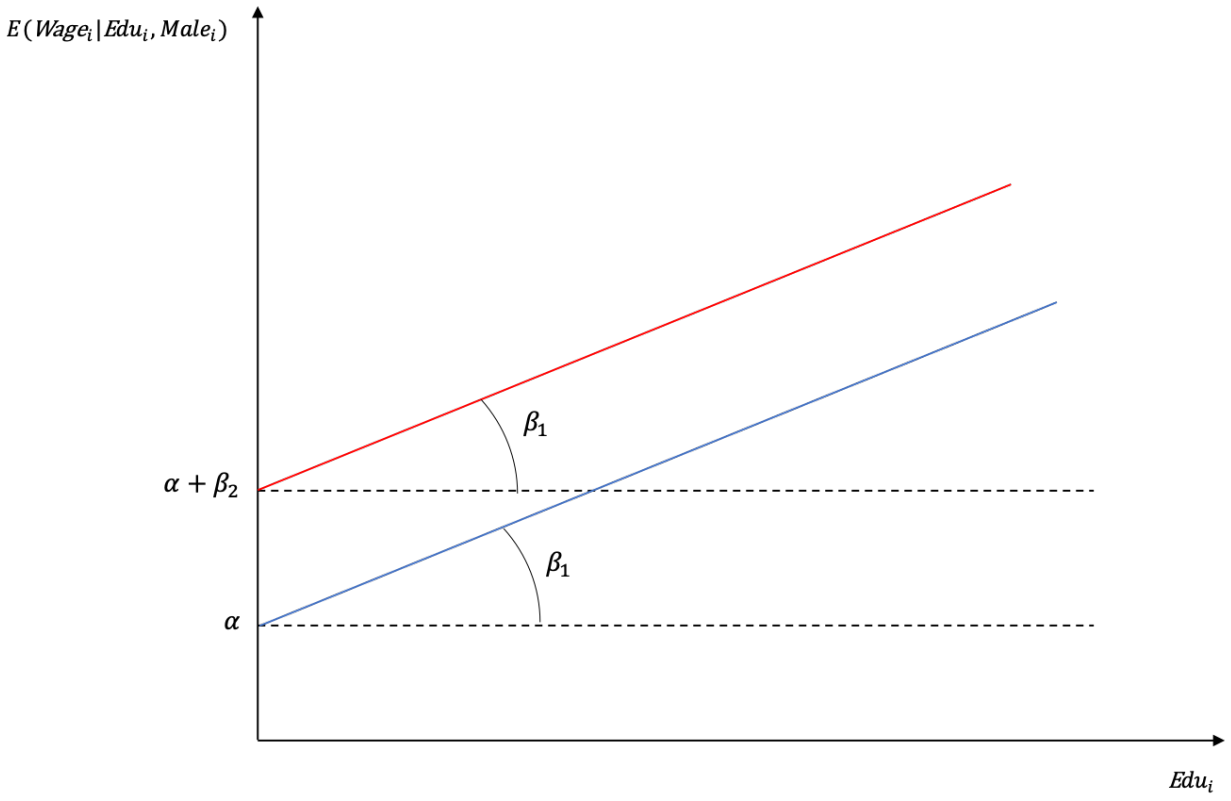

Figure 1: Conditional expectation function

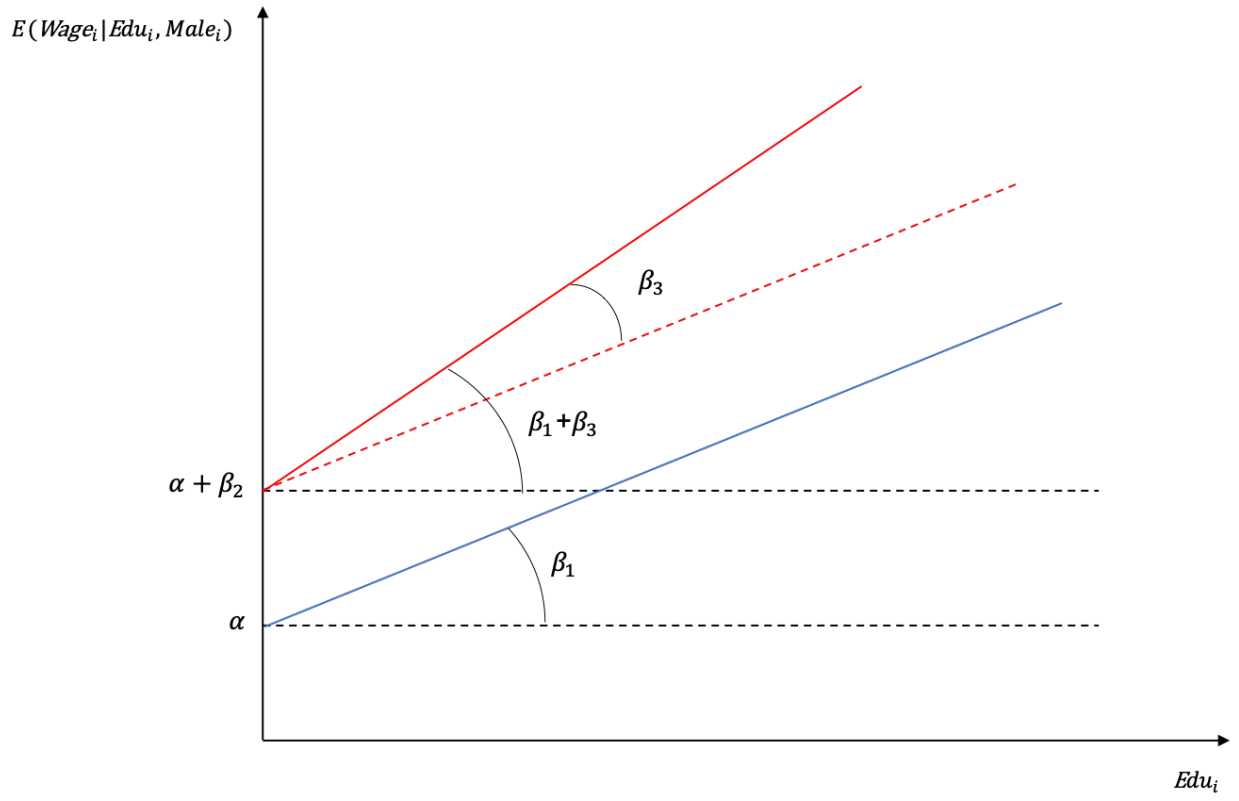

Figure 2: Conditional expectation function different marginal effect

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. 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
dim(USCPS)  #Check length
summary(USCPS)  #Explore the data
```

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

```
head(USCPS[rowSums(is.na(USCPS)) > 0, ])
```

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* ≥ 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)
```

Table 1: Regression Results Wage Education University

	Dependent variable:	
	Wage	
	(1)	(2)
Education	7.074***	4.975***
	(0.028)	(0.045)
University		15.180***
		(0.254)
Constant	−58.037***	−31.605***
	(0.446)	(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 (Table 1). 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 wheras 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)
```

Table 2: Regression Results Wage Education Age

	Dependent variable:	
	Wage	Log Wage
	(1)	(2)
Education	7.170***	0.120***
	(0.027)	(0.0004)
Age	1.461***	0.025***
	(0.009)	(0.0001)
Constant	−115.916***	0.713***
	(0.554)	(0.009)
Observations	561,076	561,076
R^2	0.147	0.158
Adjusted R^2	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<0.1; **p<0.05; ***p<0.01	

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 (See Table 2).

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%.

Comparison: Model 1 measures in level-level terms whereas model 2 measures in log-level terms. While they both measure the effect of education and age on wage, one measures the effect of education/age on wage in absolute terms and the other in percentage points.

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

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

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)
```

Table 3: Regression Results Education Age Gender

	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^2	0.206
Adjusted R^2	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

Females earn on average 44.25% less than males with the same education and age (see Table 3). This is economically significant, by using the average wage and multiplying with the female coefficient we get 23.88, 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 β_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:

$$\frac{\hat{\beta_4} - \beta_4}{stderr(\hat{\beta_4})} = \tilde{t} = \frac{-0.4427 - 0}{0.0024} = -184,46$$

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_0 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 β_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)
```

Table 4: Female on Education and Wage

	Dependent variable:
	Female
Education	0.015^{***}
	(0.0002)
Age	-0.001^{***}
	(0.0001)
Constant	0.229^{***}
	(0.005)
Observations	561,076
R^2	0.008
Adjusted R^2	0.008
Residual Std. Error	0.493 (df = 561073)
F Statistic	$2{,}255.955^{***}$ (df = 2; 561073)
Note:	*p<0.1; **p<0.05; ***p<0.01

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
```

This demonstrates 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)
```

Table 5: Log Wage on Female Residuals

	Dependent variable:
	Log Wage
Female Residuals	-0.443^{***}
	(0.003)
Constant	3.559^{***}
	(0.001)
Observations	561,076
R^2	0.047
Adjusted R^2	0.047
Residual Std. Error	0.979 (df = 561074)
F Statistic	$27{,}933.880^{***}$ (df = 1; 561074)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
print(Beta_hat_1)  # calculated constant
## [1] 0.8145355
print(summary(lm5)$coefficients[1, 1])  # 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:

$$\begin{aligned} lw_i = &\beta_1 + \beta_2 educ_i + \beta_3 age_i + \beta_4 female_i + \\ &\beta_5 female_i \times educ_i + \beta_6 female_i \times age_i + \epsilon_i \end{aligned}$$

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.*

```
Model8 = lw ~ educ + age + female + female:educ + female:age
lm8 <- lm(Model8, data = USCPS, x = TRUE)
```

$\beta_4, \beta_5, \beta_6$ are significantly different from zero with a p-value of well below 5% (see Table 6).

For the joint significance test, we first formulate our hypothesis:

$$\begin{aligned} H_0 &: \beta_4 + \beta_5 + \beta_6 = 0 \\ H_1 &: \beta_4 + \beta_5 + \beta_6 \neq 0 \end{aligned}$$

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

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

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.

Table 6: Wage on Education, Age and Gender

	Dependent variable:
	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^2	0.208
Adjusted R^2	0.208
Residual Std. Error	0.893 (df = 561070)
F Statistic	$29{,}436.260^{***}$ (df = 5; 561070)
Note:	*p<0.1; **p<0.05; ***p<0.01

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

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

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

(g)Run again the model in (1c) separately for males and females. How do the coefficients for educ and age in the males regression compare to the coefficient estimates in part (1f)? How do the coefficients for educ and age in the females regression compare to the coefficient estimates in part (1f)? What does this tell you about the impact of interacting a dummy variable with all the other variables (including the constant) in a regression?

Note that we exclude the gender dummy from the model because it become singular.

```
lm9 <- lm(Model5, data = dplyr::filter(USCPS, female == 0), x = TRUE)
lm10 <- lm(Model5, data = dplyr::filter(USCPS, female == 1),
    x = TRUE)
```

The coefficients nicely add up to each other. For male they are the same because we isolated female effects (see Table 7). To arrive at the female coefficients, we simply add education plus $Female \times Education$, same goes for wage. This means that by using an interaction variable, we can combine more information in one regression instead of creating two single regressions and comparing them. Interacting a dummy on all variables isolates the effect on the dummy subsample.

(h) Generate a dummy for each occupation category. Can you include all of them in your model? Why or why not?

Table 7: Regression Results for gender-separated samples

	Dependent variable:		
	Log Wage		
	Combined	Male	Female
	(1)	(2)	(3)
Education	0.120***	0.120***	0.136***
	(0.001)	(0.001)	(0.001)
Age	0.028***	0.028***	0.019***
	(0.0002)	(0.0002)	(0.0002)
Female	−0.342***		
	(0.018)		
Female x Education	0.017***		
	(0.001)		
Female x Age	−0.009***		
	(0.0003)		
Constant	0.759***	0.759***	0.417***
	(0.011)	(0.011)	(0.014)
Note:			*p<0.1; **p<0.05; ***p<0.01

```
USCPS <- USCPS %>% to_dummy(occupation, var.name = "label", suffix = "label") %>%
    bind_cols(USCPS)
```

We cannot include all dummies in the model because then we would fall into the "dummy trap": by using all dummies we would include perfect multicollinearity as the value of one occupation is inherently defined by all other occupations. Hence we need to drop one dummy.

(i) Now test the model in part (1c) for each occupational subsample (i.e. perform the regression in part (1c) each occupation at a time). Comment on the pattern of your wage gap estimates across occupations. Is the gender wage gap statistically different across occupations? Provide support for your conclusions.

```
lm11 <- lm(Model5, data = dplyr::filter(USCPS, occupation_other ==
    1), x = TRUE)
lm12 <- lm(Model5, data = dplyr::filter(USCPS, occupation_healthcare ==
    1), x = TRUE)
lm13 <- lm(Model5, data = dplyr::filter(USCPS, occupation_technology ==
    1), x = TRUE)
lm14 <- lm(Model5, data = dplyr::filter(USCPS, occupation_business ==
    1), x = TRUE)
lm15 <- lm(Model5, data = dplyr::filter(USCPS, occupation_science ==
    1), x = TRUE)
```

It is not possible to compare the coefficients of different sub-samples and perform a hypothesis test (see Table 8). What we need is a combined model that interacts the female and the occupation dummy (careful to avoid the dummy trap):

```
Model10 <- lw ~ educ + age + female + occupation_healthcare:female +
    occupation_technology:female + occupation_business:female +
    occupation_science:female
lm16 <- lm(Model10, data = USCPS, x = TRUE)
```

We want to prove that the gender gap is different across occupations, in order to do this we need to take a look at the p-values of the various interaction terms that states if change in wage of women in that field is

Table 8: Regression Results for different Occupations

	Dependent variable:				
	Other	Healthcare	Log Wage Technology	Business	Science
	(1)	(2)	(3)	(4)	(5)
Education	0.124***	0.136***	0.104***	0.061***	0.092***
	(0.002)	(0.002)	(0.0005)	(0.006)	(0.002)
Age	0.028***	0.022***	0.023***	0.031***	0.021***
	(0.0004)	(0.0004)	(0.0002)	(0.002)	(0.0004)
Female	−0.361***	−0.358***	−0.463***	−0.140***	−0.232***
	(0.007)	(0.009)	(0.003)	(0.026)	(0.009)
Constant	1.088***	0.941***	1.131***	1.938***	1.851***
	(0.031)	(0.036)	(0.010)	(0.123)	(0.036)
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 9: Combined Model for occupations

	Dependent variable:
	Log Wage
Education	0.118***
	$p = 0.000$
Age	0.024***
	$p = 0.000$
Female	0.035***
	$p = 0.00000$
Female x Healthcare	−0.188***
	$p = 0.000$
Female x Technology	−0.587***
	$p = 0.000$
Female x Business	−0.061**
	$p = 0.032$
Female x Science	0.038***
	$p = 0.006$
Constant	0.956***
	$p = 0.000$
Note:	*p<0.1; **p<0.05; ***p<0.01

significantly different from 0. Using $\alpha = 5\%$, we can state that the wage gap is not significant in business and science. Hence the wagegap mostly comes from *healthcare*, *technology* and *other* occupations (see Table 9).

(j) Drop all the males from your dataset.

```
USCPS <- dplyr::filter(USCPS, female == 1)
```

i) Regress log wages on educ, age, and childrenly. Test in R/Stata H1 : childrenly < 0 for workers in technology. Is the effect negative in every occupation? Provide support for your conclusion.

```
Model11 <- lw ~ educ + age + childrenly
lm11Female <- lm(Model11, data = dplyr::filter(USCPS, occupation_other ==
    1), x = TRUE)
lm12Female <- lm(Model11, data = dplyr::filter(USCPS, occupation_healthcare ==
    1), x = TRUE)
lm13Female <- lm(Model11, data = dplyr::filter(USCPS, occupation_technology ==
    1), x = TRUE)
lm14Female <- lm(Model11, data = dplyr::filter(USCPS, occupation_business ==
    1), x = TRUE)
lm15Female <- lm(Model11, data = dplyr::filter(USCPS, occupation_science ==
    1), x = TRUE)
```

Table 10: Regression Results for different Occupations on only Females

	Dependent variable:				
	Log Wage				
	Other	Healthcare	Technology	Business	Science
	(1)	(2)	(3)	(4)	(5)
Education	0.126^{***}	0.125^{***}	0.113^{***}	0.034^{***}	0.104^{***}
	$t = 52.118$	$t = 64.867$	$t = 135.511$	$t = 4.140$	$t = 23.647$
Age	0.022^{***}	0.016^{***}	0.018^{***}	0.025^{***}	0.018^{***}
	$t = 34.086$	$t = 33.373$	$t = 70.343$	$t = 11.527$	$t = 18.394$
Children	-0.052^{*}	-0.010	-0.071^{***}	0.232^{***}	-0.020
	$t = -1.877$	$t = -0.572$	$t = -6.198$	$t = 2.773$	$t = -0.401$
Constant	0.943^{***}	0.994^{***}	0.735^{***}	2.537^{***}	1.509^{***}
	$t = 19.319$	$t = 25.216$	$t = 43.493$	$t = 14.397$	$t = 16.605$
Note:					*p<0.1; **p<0.05; ***p<0.01

With $H_1 : \beta_{children} < 0$, $H_0 : \beta_{children} > 0$. In a one-sided test critical value is 1.64, $\tilde{t} < 1.64$ cause us to reject H_0. By looking at the reported t-statistics, we can find that we can reject H_0 in all occupations other than business" (see Table 10). We can confirm H_1 in technology, but it cannot be confirmed in all occupations.

ii) Regress log wages on educ, age, childrenly, and occupation dummies (exclude the dummy for "other"). Following the "p-value" path, test whether the gender wage gap is the same in business and science (i.e. test $\beta_{business} = \beta_{science}$). Do the test both in R/Stata and "by hand". How does your answer compare to Stata's/R's?

```
Model12 <- lw ~ educ + age + childrenly + occupation_healthcare +
    occupation_technology + occupation_business + occupation_science
lm18 <- lm(Model12, data = USCPS, x = TRUE)
```

Test $\beta_{business} = \beta_{science}$ in R:

```
linearHypothesis(lm18, c("occupation_business=occupation_science"))
## Linear hypothesis test
```

Table 11: Occupation Dummies with Children

	Dependent variable:
	Log Wage
Education	0.114^{***}
	(0.001)
Age	0.018^{***}
	(0.0002)
Children	-0.057^{***}
	(0.009)
Healthcare	-0.194^{***}
	(0.008)
Technology	-0.601^{***}
	(0.007)
Business	-0.067^{**}
	(0.029)
Science	0.038^{***}
	(0.014)
Constant	1.305^{***}
	(0.017)
Observations	242,199
R^2	0.194
Adjusted R^2	0.194
Residual Std. Error	0.900 (df = 242191)
F Statistic	$8{,}347.067^{***}$ (df = 7; 242191)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
##
## Hypothesis:
## occupation_business - occupation_science = 0
##
## Model 1: restricted model
## Model 2: lw ~ educ + age + childrenly + occupation_healthcare + occupation_technology +
##     occupation_business + occupation_science
##
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1 242192 196238
## 2 242191 196228  1    9.4718 11.69 0.0006283 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By hand:

$$H_0 : \beta_{business} - \beta_{science} = 0 \text{ and } H_1 : \beta_{business} - \beta_{science} \neq 0$$

$$t = \frac{\hat{\beta}_{Business} - \hat{\beta}_{Science} - 0}{Var(\hat{\beta}_{Business}) + Var(\hat{\beta}_{Science}) - 2 * Cov(\hat{\beta}_{Business}, \hat{\beta}_{Business})}$$

We can extract these values by the coefficients and the variance-covariance Matrix (command vcov()) of our model lm18 (see Table 11) and plug in:

```
Beta_business <- lm18$coefficients[7]
Beta_science <- lm18$coefficients[8]
Variance_Beta_business <- vcov(lm18)[7, 7]
Variance_Beta_science <- vcov(lm18)[8, 8]  #Variance Beta_Science
Covariance_Business_Science <- vcov(lm18)[8, 7]  #Covariance Beta_Business Beta_Science
t <- (Beta_business - Beta_science)/(Variance_Beta_business +
    Variance_Beta_science - 2 * Covariance_Business_Science)
```

Beta_business	Beta_science	Variance_Beta_business	Variance_Beta_science	Covariance_Business_Science	t
-0.0671991	0.0375701	0.0008378	0.0001896	4.42e-05	-111.5822

We go to a t-table to find that the p-value is 0% (as we could only find a t-table that sets DF=∞ with the number of observations that are in the table). This aligns with the result R calculated, which is marginally above zero.

iii. How do occupations' dummies compare to point (1i)?

The results from the dummies nicely illustrate why we should not compare different models to each other in (1i). I.e. the average change in earnings for females in the healthcare sample was $\sim 1.30 - 0.19 = 1.11$ while it was $\sim 0.94 - 0.36 = 0.58$ in the model including all female observations and the occupation dummies. The model from the subsample may suffer from unexplained variance and thus biased estimators. Also, the dummy modely confirms our proposed solution to check if the gender wage gap is different across industries using interaction variables: the estimates are approximately the same.

(k) Throughout this question we have assumed that Assumption 2 holds. What do you think about this assumption? Can you think about other factors we did not take into consideration in our model that could bias the conclusion that we are measuring the true gender wage gap?

Assumption 2 is one of the two "key assumptions". If it is violated, our estimates might be biased because we are also measuring some relation within the error (covariance between x_i and ϵ_i). So it is vital to verify that assumption 2 holds first.

The main sources of this bias are: omitted variable and/or measurment error. The first one could be due to the fact the we don't have a variable for location and the second given by the fact that people self report the wages and thus they are introducing noise in the model.

$X_z = (z \; X_k)$ $\quad \beta_z = \begin{pmatrix} \gamma \\ \beta_k \end{pmatrix}$

$X_z' y - X_z' X_z \hat{\beta}_z = 0 \rightarrow$ first order condition of RSS min problem.

$\hookrightarrow \begin{pmatrix} z \\ X_k \end{pmatrix} y - \begin{pmatrix} z'z & z'X_k \\ X_k z & X_k' X_k \end{pmatrix} \begin{pmatrix} \hat{\gamma} \\ \hat{\beta}_k \end{pmatrix} = 0$

$\Rightarrow \begin{cases} zy - z'z\hat{\gamma} - z'X_k\hat{\beta}_k = 0 \quad (1) \\ X_k'y - X_k'z\hat{\gamma} - X_k'X_k\hat{\beta}_k = 0 \quad (2) \end{cases}$

$\hookrightarrow \hat{\beta}_k = (X_k'X_k)^{-1} X_k (y - z\hat{\gamma})$

$\Rightarrow \hat{\beta}_k = (X_k'X_k)^{-1} X_k' y \overset{(3)}{\Rightarrow}$ Given that z is uncorrelate with X_k

[by symmetry: $\hat{\gamma} = (z'z)^{-1} z y$]

plug (3) in (1) $\Rightarrow z'y - z'z\hat{\gamma} - z'\underbrace{X_k(X_k'X_k)^{-1}X_k}_{P_k}(y - z\hat{\gamma}) = 0$

$z'y - z'z\hat{\gamma} - z'P_k(y - z\hat{\gamma})$ $\qquad \boxed{P_k = X_k(X_k'X_k)^{-1}X_k}$

Add identity matrix $\Rightarrow z'Iy - z'Iz\hat{\gamma} - z'P_k y + z'P_k\hat{\gamma} = 0$

$\Rightarrow z'(I - P_k)y - z'(I - P_k)z\hat{\gamma} = 0$

$\hookrightarrow z'M_k y - z'M_k z\hat{\gamma} = 0$

$\hat{\gamma} = (z'M_k z)^{-1} z'M_k y$

by Symmetry $\Rightarrow \hat{\beta}_k = (X_k'M_z X_k)^{-1} X_k M_z y$

$\hookrightarrow \hat{\beta}_k = (X_k'X_k)^{-1} X_k (y - z\hat{\gamma})$

Figure 3: Additional proof of how to compute the gamma coefficient 1.e