Exercise 2

Suggested Solutions

Alexandre Jenni and Emanuele Dicarlo

01/11/2019

Theory

The gender wage gap

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

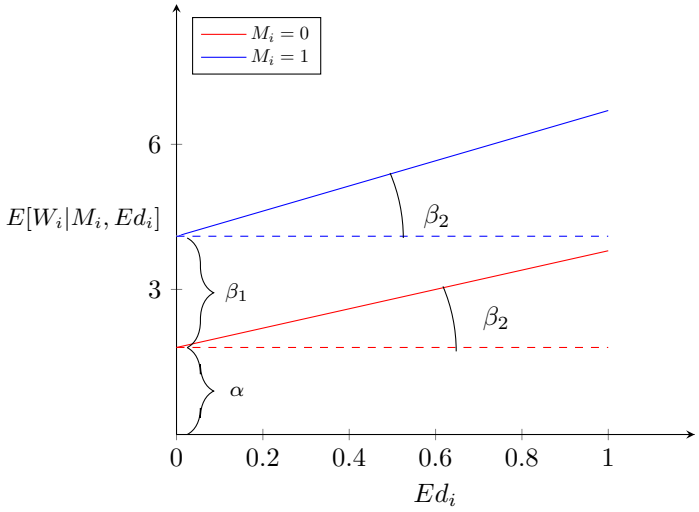
- *monthly wage*
- *gender*
- *years of education*

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

We can test whether $\beta_1 = 0$ *using our estimate* $\hat{\beta}_1$ *from the regression*

$$wage_i = \alpha + \beta_1 M_i + \beta_2 Ed_i + \epsilon_i. \tag{1}$$

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



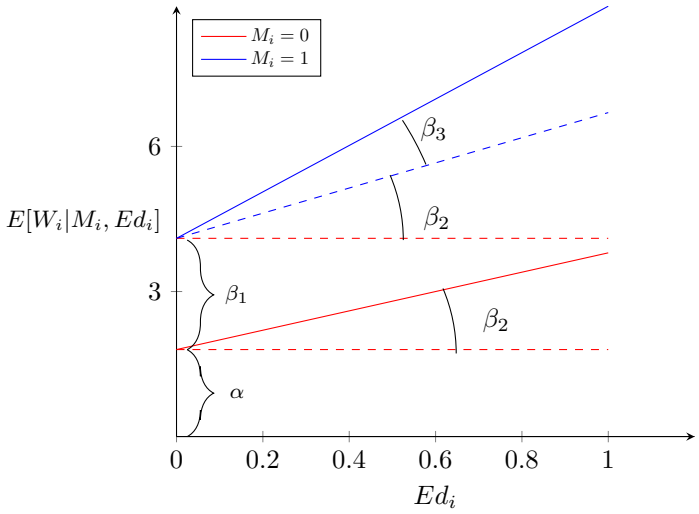
The conditional expectation function (CEF) of wage for males is a vertical shift of the CEF for females. The direction and size of the vertical shift is given by β_1.

(c) You consider now that *years of education* might have a different marginal effect on women compared to men. How would you modify your regression model to account for this differential?

We need to to add an interaction term between the dummy for gender and years of education:

$$wage_i = \alpha + \beta_1 M_i + \beta_2 Ed_i + \beta_3 M_i * Ed_i + \epsilon_i \tag{2}$$

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



With this model, the CEF of wage for males differs from the CEF for females both in terms of the intercept and the slope. The difference between the intercepts is given by β_1, while β_3 is the difference in slopes. We do not force the CEF for males and females to share any common parameter.

Empirical Application

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 toward 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 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.[1]

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
- *educ*, 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

```
wagegap <- read.csv("./sampleUScens2015.csv")

#install.packages("dummies") if you don't have it installed yet
library(dummies)

# Create auxiliary dataset with dummies for each occupation
occupations <- data.frame(dummy(wagegap$occupation, sep = "_"), nrow = nrow(wagegap))
# Drop last column, as it is useless
occupations$nrow <- NULL
# Rename these variables
colnames(occupations) <- c("business", "healthcare", "other", "science", "technology")
# Merge with main dataset
wagegap <- cbind(wagegap,occupations)

wagegap$female <- as.numeric(wagegap$female)
wagegap$age <- as.numeric(wagegap$age)
```

(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?

```
wage <- wagegap$incwage/1000
lw <- log(wage)
university<-ifelse(wagegap$educ>=16, 1, 0)
wagegap <- cbind(wagegap,wage,lw, university)
model1 <- lm(wage ~ 1 + educ, data = wagegap)
```

[1] Note that this is a very strong assumption that is unlikely to hold, but we want to focus on other aspects of econometrics for the moment.

```
model2 <- lm(wage ~ 1 + educ + university, data = wagegap)
stargazer(model1, model2, header = FALSE, keep.stat=c("n","rsq"))
```

Table 1:

	Dependent variable:	
	wage	
	(1)	(2)
educ	7.074***	4.975***
	(0.028)	(0.045)
university		15.180***
		(0.254)
Constant	−58.037***	−31.605***
	(0.446)	(0.628)
Observations	561,076	561,076
R^2	0.104	0.110
Note:	*p<0.1; **p<0.05; ***p<0.01	

In the first column of Table 1, we see that one extra year of education increases yearly wage by \$7,074 on average. When the univerity dummy is included (column 2), an extra year of education increases yearly wage by \$4,975. In this case, the coefficient of education measures the impact on wage of one more year of education conditional on the graduating status. The estimated effect is lower than in the absence of the university dummy, which suggests that the expected wage increases discontinuously at 16 years of education (called a credential effect), or at least that the effect of years of education on wage is nonlinear.

(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.]

```
wagegap$university <- NULL
model3.a <- lm(wage ~ 1 + educ + age, data = wagegap)
model3.b <- lm(lw ~ 1 + educ + age, data = wagegap)
stargazer(model3.a, model3.b, header = FALSE,
          keep.stat=c("n","rsq"))
```

Looking at the results from model 3.a in Table 2 column 1, an extra year of education increases the expected wage by about \$7,000. Because the dependent variable in model 3.b is in log, we concude that an extra year of education increases the expected yearly wage by 12% (column 2). Mean wages are approximately \$54,000. Hence, the person at the mean of the income distribution would increase her wage by \$54,000 × 0.12 = \$6,080 for an additional year of education. We see that the results from the two models are roughly consistent with each other.

(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.

Table 2:

	Dependent variable:	
	wage	lw
	(1)	(2)
educ	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
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 3:

	Dependent variable:
	lw
educ	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
F Statistic	$48{,}385.030^{***}$
Note:	*p<0.1; **p<0.05; ***p<0.01

We obtain that for a given age and level of education $\widehat{\log w^f} = \widehat{\log w^m} - 0.44$ *(Table 3). This can be (approximately) rewritten*[2] *as* $(\widehat{w^f} - \widehat{w^m})/\widehat{w^h} = e^{-0.44} - 1$*=-0.36. Hence, females earn on average about 5.6% less than men.*

```
# Testing for a 0 coefficient "manually"
beta4 <- coef(model4)[4]
se.beta4 <- summary(model4)$coefficients[4,2]
t.beta4 <- round(beta4/se.beta4, digits=4)
print(paste("The t-stat is ",t.beta4))
```

```
## [1] "The t-stat is  -183.011"
```

It is significant at the 1 percent confidence level. Let's do it by hand. Our degrees of freedom are $N - K = 561,076 - 4 = 561,072$*. For such a large number of degress of freedom, the distribution of the t statistic is indistinguishable from a standard normal. Since we want to test the presence of discrimination against female, a one sided test against* $\beta_4 < 0$ *is more appropriate. The critical value of our t (actually z statistic) is then* $z_{-0.95} = -1.64$ *at the 5% level or* $z_{-0.99} = -2.33$ *at the 1% level. Our t-statistic is: -183 which is clearly lower than both critical values. So we can reject the null hypothesis* $H_0 : \beta_4 >= 0$ *at any significance level.*

```
# Using R's F-test
lh <- linearHypothesis(model4, "female")
F.beta4 <- round(lh$F[2], digits=4)
print(paste("The F-stat is ", F.beta4))
```

```
## [1] "The F-stat is  33493.0426"
```

When testing a single hypothesis, we can recover the t-statistic by taking take the square root of the F statistic, so $\pm\sqrt{F} = \pm\sqrt{33493} = \pm 183.011$*. And the sign of the t-statistic will be given by the sign of the numerator of the t-statistic.*

(d). Use R/stata to get $\hat{\beta}_4$ (the coefficient of $female$) using partitioned regression.

```
# Regression of lw on resid of female_on_else
partition1 <- lm(female ~ 1 + educ + age, data = wagegap, na.action = "na.exclude")
e1 <- resid(partition1)
# ensure that same sample of non missing is used:
wagegap.f <- filter(wagegap, is.na(educ)==FALSE, is.na(age)==FALSE)
part.reg <- lm(wagegap$lw ~ e1)
beta4 <- round(coef(part.reg)[2], digits=4)
print(paste("$\beta_4$ is ", beta4))
```

[1] "β_4 is -0.4427" *This corresponds to the value obtained in the previous regression (this is just an application of the partitioned regression result).*

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

```
y.bar <- mean(wagegap$lw, na.rm=TRUE)
educ.bar <- mean(wagegap$educ, na.rm=TRUE)
age.bar <- mean(wagegap$age, na.rm=TRUE)
female.bar <- mean(wagegap$female, na.rm=TRUE)
beta2 <- coef(model4)[2]
beta3 <- coef(model4)[3]
beta4 <- coef(model4)[4]
beta1 <- y.bar - beta2*educ.bar - beta3*age.bar - beta4*female.bar
beta1 <- round(beta1, digits=4)
print(paste("$\beta_1$ is ", beta1))
```

[2] This is an approximation since in general $E(exp\{\epsilon\}|x) \neq 1$ even if $E(\epsilon|X) = 0$.

[1] "β_1 is 0.8145"

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

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

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.

```
model5 <- lm(lw ~ 1 + educ + age + female*(educ+age), data = wagegap, na.rm=TRUE)
stargazer(model5, header=FALSE, keep.stat=c("n","rsq", "f"))
```

Table 4:

	Dependent variable:
	lw
educ	0.120***
	(0.001)
age	0.028***
	(0.0002)
female	−0.342***
	(0.018)
educ:female	0.017***
	(0.001)
age:female	−0.009***
	(0.0003)
Constant	0.759***
	(0.011)
Observations	561,076
R^2	0.208
F Statistic	29,436.260*** (df = 5; 561070)
Note:	*p<0.1; **p<0.05; ***p<0.01

β_4, β_5, and β_6 are all different from 0 according to the p-values of their coefficients. However, this is a test for three different null hypotheses (that each coefficient individually is different from 0).

```
# Test the joint restrictions
lh <- linearHypothesis(model5, matchCoefs(model5, "female"))
f.stat <- round(lh$F[2], digits=4)
print(paste("The F-stat is ", f.stat))
```

```
## [1] "The F-stat is  11733.0668"
```

In order to test the joint hypothesis, we need to compute the F-statistic that these coefficients are jointly 0. The F test statistics is > 10,000. We have 3 restrictions and 561,070 d.o.f. ($\approx \infty$). The 99th percentile of

the $F_{3,K}$ distribution for K very large is smaller than 5 and thus much lower than our F statistic. Hence we reject the null hypothesis that these coefficients are jointly 0.

Alternatively, we can use the R^2-based formula for the F statistic to compute its value per hand. The restricted model (Model3 where the female dummy does not enter) shows a R^2 of 0.158, while the unrestricted model (model5) $R^2 = 0.208$ so that $F_{(3,\ 561{,}072)} = \frac{(0.208-0.158)/3}{(1-0.208)/561'072} \approx 11,820$. This is different from the F test computed by R only because of our approximations.

(g). Run again the model in (c) **separately** for males and females. What are the coefficients of *educ* and *age*. How do they compare to point (f)?

```
wagegap.males <- subset(wagegap, female == 0)
wagegap.females <- subset(wagegap, female == 1)
model5a <- lm(lw ~ 1 + educ + age, data = wagegap.males)
model5b <- lm(lw ~ 1 + educ + age, data = wagegap.females)
stargazer(model5, model5a, model5b, header = FALSE,
          column.labels=c("All","Males","Females"),
          font.size = "small", keep.stat=c("n","rsq"))
```

Table 5:

	Dependent variable:		
	lw		
	All	Males	Females
	(1)	(2)	(3)
educ	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)		
educ:female	0.017^{***}		
	(0.001)		
age:female	-0.009^{***}		
	(0.0003)		
Constant	0.759^{***}	0.759^{***}	0.417^{***}
	(0.011)	(0.011)	(0.014)
Observations	561,076	318,877	242,199
R^2	0.208	0.199	0.152
Note:		*p<0.1; **p<0.05; ***p<0.01	

By comparing Table 5 and Table 4, we observe that including dummies and fully interacting them is the same as running two separate models: $\beta_{educ}^{female} = \beta_{educ}^{males} + \beta_5$ and $\beta_{age}^{female} = \beta_{age}^{males} + \beta_6$.

(h). Generate a dummy for each occupation category. Can you include all of them in your model? Why?

Unless you omit the constant, you cannot include all the occupation categories because of multicollinearity. The reason is that all the occupation dummies always sum up to 1 for each observation (which is also the value of the constant term). Intuitively, once we know that a person is not active in all but the last sector, we alreay know that she works in the last sector. Hence the last dummy does not provide any independent information.

(i). Now test the model in part (c) for each occupational subsample (i.e. perform the regression in part (c) 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.

```
wagegap.business <- subset(wagegap, occupation == "business")
wagegap.healthcare <- subset(wagegap, occupation == "healthcare")
wagegap.other <- subset(wagegap, occupation == "other")
wagegap.science <- subset(wagegap, occupation == "science")
wagegap.technology <- subset(wagegap, occupation == "technology")

model6a <-  lm(lw ~ 1 + educ + age + female, data = wagegap.business)
model6b <-  lm(lw ~ 1 + educ + age + female, data = wagegap.healthcare)
model6c <-  lm(lw ~ 1 + educ + age + female, data = wagegap.other)
model6d <-  lm(lw ~ 1 + educ + age + female, data = wagegap.science)
model6e <-  lm(lw ~ 1 + educ + age + female, data = wagegap.technology)
stargazer(model6a, model6b, model6c, model6d, model6e,
          header = FALSE, no.space = TRUE, font.size = "small",
          keep.stat=c("n","rsq"),
          column.labels = c("Bus.", "Health.", "Other", "Sci.", "Techn."))
```

Table 6:

	Dependent variable:				
	lw				
	Bus.	Health.	Other	Sci.	Techn.
	(1)	(2)	(3)	(4)	(5)
educ	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)
Observations	45,466	40,998	441,507	2,440	30,665
R^2	0.244	0.227	0.163	0.212	0.153
Note:					*p<0.1; **p<0.05; ***p<0.01

First of all it is important to see if the wage gap is different from zero in every occupation. Which is evident from the p-value of each regression. An interesting thing would be to test whether the wage gap is different in all the occupations or if it is the same. I.e. we want to test if β_4 is the same in models 6a to 6e (columns 1 to 5 in the table above). To do so we have to rewrite the model (where m is the number of occupations) as:

$$lw_i = \beta_1 + \beta_2 educ_i + \beta_3 age_i + \beta_4 female_i + \sum_{j=1}^{m-1} \gamma_j occup_j + \sum_{j=1}^{m-1} \delta_j female_i * occup_j + \epsilon_i \tag{3}$$

and we want to test: $\delta_1 = \delta_2 = \delta_3 = \delta_4 = 0$

```
# Unrestricted model
model7u <- lm(lw ~ 1 +educ + age + female*(business + healthcare +
              science + technology), data = wagegap)
# Restricted model
model7r <- lm(lw ~ 1 +educ + age + female+ business + healthcare +
              science + technology, data = wagegap)

#stargazer(model7u, model7r, header = FALSE, keep.stat=c("rsq","n"))
```

```
# Construct and compute F stat
r2u <- summary(model7u)$r.squared
r2r <- summary(model7r)$r.squared
f.stat <- round(((r2u-r2r)/4)/((1-r2u)/561064), digits=4)
print(paste("The F-stat is ", f.stat))
```

[1] "The F-stat is 125.4037"

The F statistic is: $F_{(3,\ 561{,}065)} = \frac{(0.241-0.240)/4}{(1-0.241)/561{,}065} \approx 125.4$. *Hence, we can reject the null hypothesis that the wage gap is the same in all industries. This means that there is at least one industry where the wage gap is statistically significant from the wage gap in the others industry.*

(j). Drop all the males from your dataset.

(j.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. Test this individually.

```
model8 <- lm(lw ~ 1 + educ + age + childrenly + business + healthcare +
                 science + technology, data = wagegap.females)

wagegap.f.tec <- subset(wagegap.females, occupation == "technology")
model8sub.1 <- lm(lw ~ 1 + educ + age + childrenly, data = wagegap.f.tec)
wagegap.f.bus <- subset(wagegap.females, occupation == "business")
model8sub.2 <- lm(lw ~ 1 + educ + age + childrenly, data = wagegap.f.bus)
wagegap.f.health <- subset(wagegap.females, occupation == "healthcare")
model8sub.3 <- lm(lw ~ 1 + educ + age + childrenly, data = wagegap.f.health)
wagegap.f.oth <- subset(wagegap.females, occupation == "other")
model8sub.4 <- lm(lw ~ 1 + educ + age + childrenly, data = wagegap.f.oth)
wagegap.f.sci <- subset(wagegap.females, occupation == "science")
model8sub.5 <- lm(lw ~ 1 + educ + age + childrenly, data = wagegap.f.sci)
```

```
stargazer(model8sub.1, model8sub.2, model8sub.3, model8sub.5, header = FALSE,
          keep.stat=c("n","rsq"), column.labels = c("Tech.","Bus.","Health.","Sci."))
```

Childrenly is negative and (barely) significant in business. It is positive (and very significant!) in science. Not different from zero in healthcare or technology.

Alternatively one could run a fully interacted model and test the linear hypotheses that the coefficient for childrenly plus that of the interaction term is significantly different from zero.

```
model8.unr <- lm(lw ~ 1 + educ*(business + healthcare + science + technology) +
                 age*(business + healthcare + science + technology) +
                 childrenly*(business + healthcare + science + technology),
                 data = wagegap.females)
```

```
lh <- linearHypothesis(model8.unr,"childrenly+business:childrenly=0")
pval <- round(lh$`Pr(>F)`[2], digits=4)
print(paste("P-value is ", pval))
```

```
## [1] "P-value is  0.1235"
```

```
lh <- linearHypothesis(model8.unr,"childrenly+healthcare:childrenly=0")
pval <- round(lh$`Pr(>F)`[2], digits=4)
print(paste("P-value is ", pval))
```

```
## [1] "P-value is  0.6502"
```

Table 7:

	Dependent variable:			
	lw			
	Tech.	Bus.	Health.	Sci.
	(1)	(2)	(3)	(4)
educ	0.104***	0.126***	0.125***	0.034***
	(0.004)	(0.002)	(0.002)	(0.008)
age	0.018***	0.022***	0.016***	0.025***
	(0.001)	(0.001)	(0.0005)	(0.002)
childrenly	−0.020	−0.052*	−0.010	0.232***
	(0.050)	(0.028)	(0.018)	(0.084)
Constant	1.509***	0.943***	0.994***	2.537***
	(0.091)	(0.049)	(0.039)	(0.176)
Observations	5,549	18,689	31,298	1,025
R^2	0.128	0.173	0.141	0.137
Note:				*p<0.1; **p<0.05; ***p<0.01

```
lh <- linearHypothesis(model8.unr,"childrenly+science:childrenly=0")
pval <- round(lh$`Pr(>F)`[2], digits=4)
print(paste("P-value is ", pval))
```

```
## [1] "P-value is  0.0838"
```

```
lh <- linearHypothesis(model8.unr,"childrenly+technology:childrenly=0")
pval <- round(lh$`Pr(>F)`[2], digits=4)
print(paste("P-value is ", pval))
```

```
## [1] "P-value is  0.7701"
```

These p-values in the fully interacted model are consistent (though larger) with those obtained running separate regressions.

(j.ii). Regress log wages on *educ*, *age*, *childrenly*, and occupation dummies (exclude the dummy for "other"). Following the"p-value" path, test whether women's wages are 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?

```
model9 <- lm(lw ~ 1 + educ + age + childrenly + business + healthcare +
             science + technology, data = wagegap.females)
stargazer(model9, header = FALSE, keep.stat=c("n","rsq"))
```

```
beta.business <- coef(model9)[5]
beta.science <- coef(model9)[7]
v.beta.business <- vcov(model9)[5,5]
v.beta.science <- vcov(model9)[7,7]
cov.bus.sci <- vcov(model9)[5,7]
t.stat <- (beta.business - beta.science)/sqrt(v.beta.business +
            v.beta.science -2*cov.bus.sci)
```

Table 8:

	Dependent variable:
	lw
educ	0.114^{***}
	(0.001)
age	0.018^{***}
	(0.0002)
childrenly	-0.057^{***}
	(0.009)
business	0.601^{***}
	(0.007)
healthcare	0.407^{***}
	(0.006)
science	0.534^{***}
	(0.028)
technology	0.639^{***}
	(0.012)
Constant	0.704^{***}
	(0.014)
Observations	242,199
R^2	0.194
Note:	*p<0.1; **p<0.05; ***p<0.01

```r
t.stat <- round(t.stat, digits=4)
print(paste("The t-stat is ", t.stat))
```

```
## [1] "The t-stat is  2.3216"
```

```
## [1] "The F-stat is  5.39"
```

```
## [1] "The p-value is  0.0203"
```

Given the t statistic we computed by hand (> 2), and the outcome of the test (p-value< 0.05), we reject that they are the same at a confidence level $\alpha = 0.05$.

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

In point (i), we estimate a different model for every occupation. This allows not only the constant, but also the parameters to vary by occupation. With occupations' dummies, we only allow the constant to depend on the occupation. This is why we obtain different intercepts if we add up the constant and the dummy coefficient for each of the occupations in Table 8 than the constants estimated in Table 7.

(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?

If we want to interpret the coefficients on female as the impact of changing a worker's gender but holding everything else constant, then the estimates are likely to be biased. For example, women are still more likely to take some time out of the labor force to raise children than men, this will translate into fewer years of experience for women of a given age than men. Hence our coefficients on female will be negatively biased since they will incorporate some of this variation in experience. Neverthelss, it is always tricky to think about gender differences in causal terms. Given that gender cannot be manipulated, some statisticians argue that we cannot talk about causal impact in this context ("no causation without manipulation")