MOEC0021 - Empirical Methods

Group BlancSchneiderMazidi

Fabienne Blanc (15-732-142)

Flavio Schneider (15-716-202)

Manuel Mazidi (15-704-984)

Course

Empirical Methods

Prof. Greg Crawford

University of Zurich

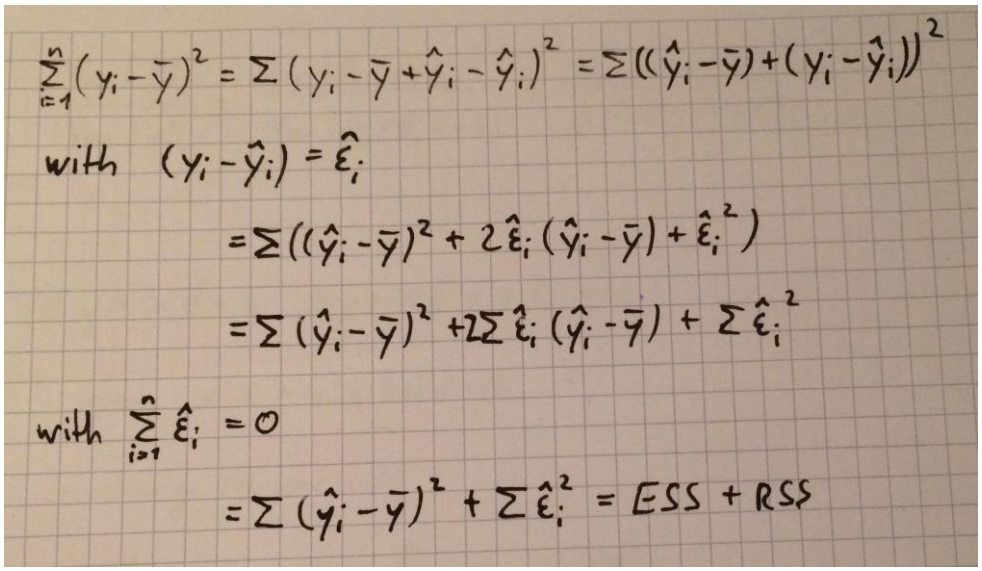
Submitted on November 5, 2018

1 Pencil and Paper Questions

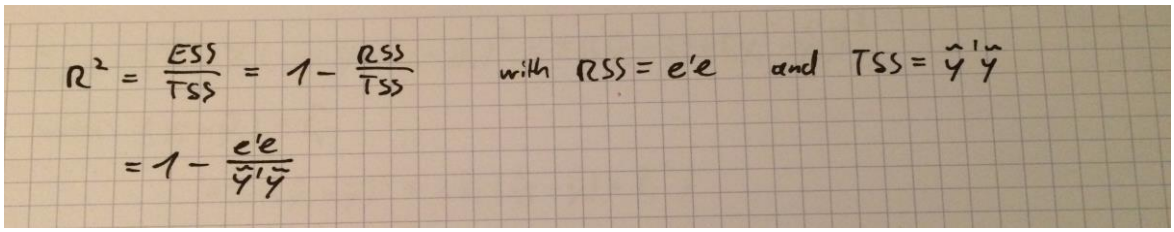
Exercise 1

(a) Given that our model includes a constant, we can write the total sum of squared (TSS) as follows: $TSS = \sum(y_i - \bar{y})^2$

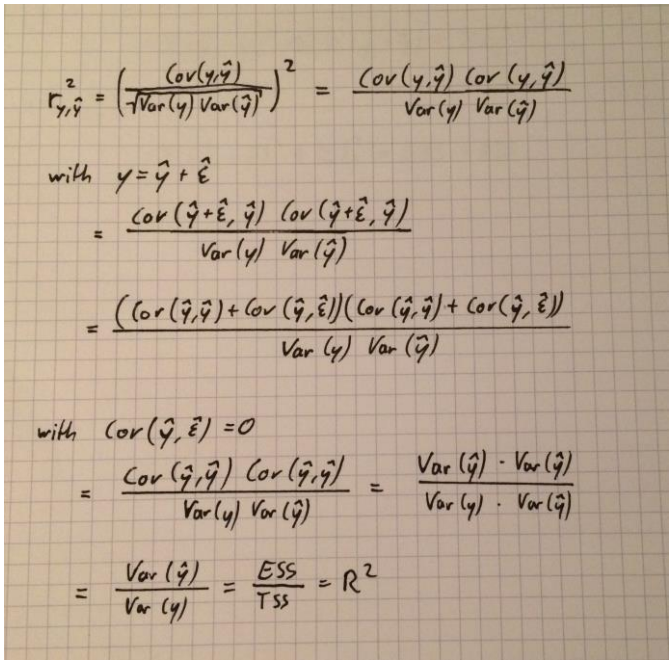
Therefore:



To proof that $R^2 = \frac{e'e}{\tilde{y}'\tilde{y}}$:



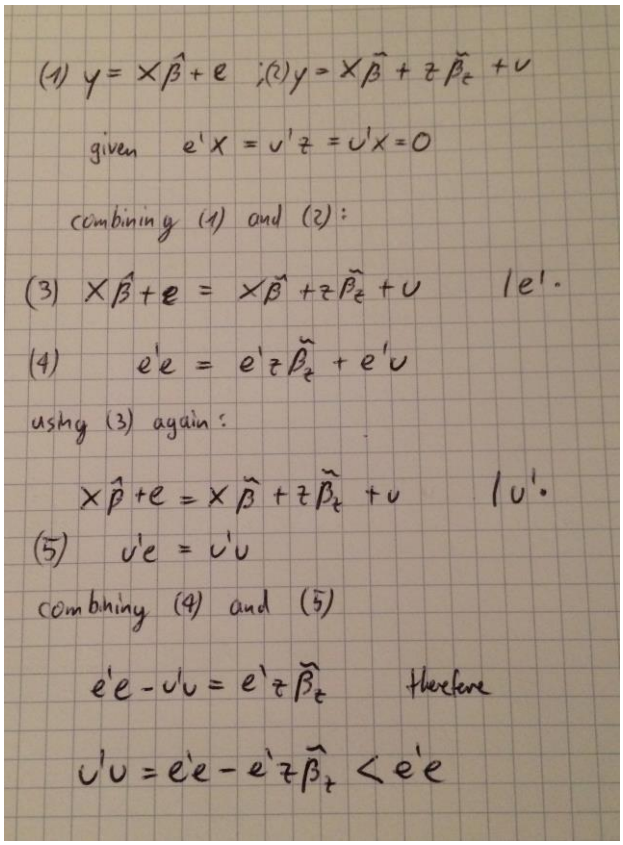
(b) We can interpret R^2 intuitively as the squared correlation coefficient between the true values of y_i and our estimated values $\hat{y}_i$. Since the R^2 is a measurement of how good the estimated values explain the true values, this intuitively translates to the correlation between these two variables. The higher R^2 or the squared correlation ($r^2_{y,\hat{y}}$) the better we can predict the true values with our model:



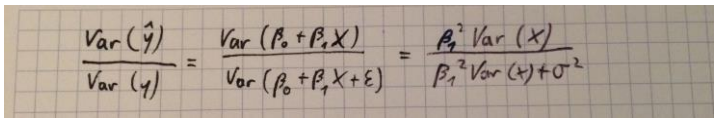
(c) By transforming our X variables linearly into different units, we cannot gain more information out of our variables than before. For each datapoint the true value of y_i and the estimated value of $\hat{y}_i$ do not change. Since we showed before, that $R^2 = r^2_{y,\hat{y}}$ and we see that the correlation between y_i and $\hat{y}_i$ isn't affected either, the calculated R^2 remains the same.

(d) Intuitively the residual sum squared always decreases if we add another regressor into our model. With this new regressor we might gather more information from our data, since the estimated values may be predicted more precisely. Only in an extremely unlikely case that the added regressor is perfectly correlated with an already existing one of our model, the RSS would stay the same, since we cannot gather more information by adding the new one.

(e) Proof:



(f) R^2 isn't necessarily an adequate measure for how good a model is in general. Adjusting the formula from b) considering a very basic model we can show that:



This means that we may arbitrarily lower R^2 even though the model is correct, when Var(X) is small or the standard deviation of the error-terms high. On the other hand, R^2 may also be very close to 1 even though our model is wrong, by just increasing Var(X) or decreasing standard deviation of the error-terms.

Furthermore, adding more explanatory variables always increases R^2. This may lead one towards overfitting the model with too many variables. So, we can increase our R^2 even by just adding a large set of totally random predictors.

Exercise 2 – The gender wage gap.

(a) As we gather data on the whole working population in the country (on *monthly wage*, *gender*, and *years of education*), we can formulate a population regression function:

$$wage_i = \beta_1 + \beta_2 * Male_i + \beta_3 * Educ_i$$

Where

$$Male_i = \begin{cases} 1 & if\ i\ is\ a\ male \\ 0 & else \end{cases}$$

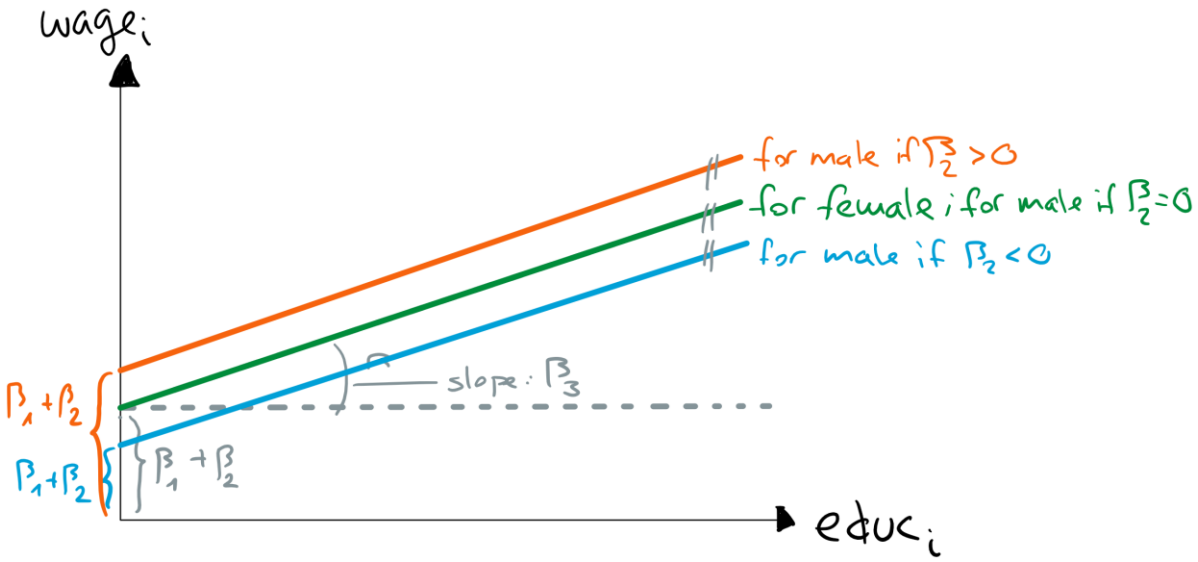
We want to test whether men and women have different salaries. Given that *years of education* (*Educ*) have the same effect on wages for both men and women, we propose the following simple regression model to test the hypothesis:

$$wage_i = \beta_1 + \beta_2 * Male_i + \beta_3 * Educ_i + \epsilon_i$$

$$H_0: \beta_2 = 0 \quad ; \quad H_1: \beta_2 \neq 0$$

If $\hat{\beta}_2$ is significantly different from zero, we can say with certain confidence that men and women have different salaries. I.e., β_2 represents the "gender gap".

(b) To provide a graphical representation of the conditional expectation function and show if and how it differs for men and women, we analyse the following three cases (different intercepts): $\beta_2 = 0$, $\beta_2 < 0$, $\beta_2 > 0$. The slope of all three cases is determined by β_3.



- $\beta_2 = 0$: Men and women do not seem to have different salaries (with same education).
- $\beta_2 < 0$: Men seem to have lower salaries than women (with same education).
- $\beta_2 > 0$: Men seem to have higher salaries than women (with same education).

(c) In retrospect, we decide that *years of education* might have a different marginal effect on men compared to women. Hence, we modify the regression model the following way:

$$wage_i = \beta_1 + \beta_2 * Male_i + \beta_3 * Educ_i + \beta_4 * Male_i * Educ_i + \epsilon_i$$

$$Male_i = \begin{cases} 1 & if\ i\ is\ a\ male \\ 0 & else \end{cases} \quad ; \quad H_0: \beta_2 = \beta_4 = 0 \quad ; \quad H_1: \beta_2 \neq 0\ or\ \beta_4 \neq 0$$

The interaction term takes the differential effect of *years of education* into account. The marginal effect of *years of education* is as follows:

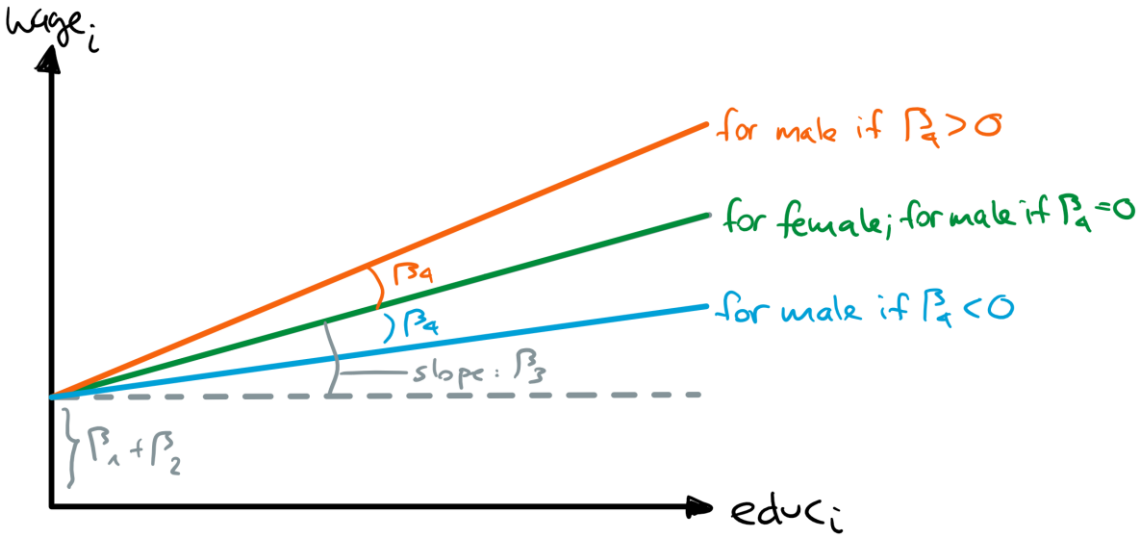
$$\frac{\partial wage_i}{\partial Educ_i} = \beta_3 + \beta_4 * Male_i$$

Therefore:

$$E[wage_i|educ_i, male_i] = \begin{cases} \beta_1 + \beta_2 * Male_i + \beta_3 * Educ_i + \beta_4 * Male_i * Educ_i & if\ male_i = 1 \\ \beta_1 + \beta_3 * Educ_i & if\ male_i = 0 \end{cases}$$

I.e., the "gender gap" is represented by $\beta_2 + \beta_4 * Educ_i$.

(d) To provide a graphical representation of the conditional expectation function and show if and how it differs for men and women, we assume that $\beta_2 = 0$ (same intercept). We analyse the following three cases: $\beta_4 = 0$, $\beta_4 < 0$, $\beta_4 > 0$. The slope of all three cases is determined by $\beta_3 + \beta_4$.



If $\beta_2 \neq 0$ the intercept would be different (cf. exercise (b)).

2 Computer Questions

Exercise 1

(a)

	Dependent variable:	
	wage	
	(1)	(2)
educ	7.074***	4.975***
	(0.028)	(0.045)
university1		15.180***
		(0.254)
Constant	-58.037***	-31.605***
	(0.446)	(0.628)
Observations	561,076	561,076
R2	0.104	0.110
Adjusted R2	0.104	0.110
Residual Std. Error	58.751 (df = 561074)	58.566 (df = 561073)
F Statistic	65,026.500*** (df = 1; 561074)	34,499.500*** (df = 2; 561073)
Note:		*p<0.1; **p<0.05; ***p<0.01

On average, in the first model in which we do not control for a university degree, an additional year of education increases yearly wages by roughly 7000 USD. This coefficient is significant on a 1%-significance level. If we include the generated dummy variable *university*, then the impact of an additional year of education decreases by roughly 2000 USD while being still significant on a 1%-significance level. The new dummy coefficient *university* is also significant and has a value of 15180. Hence, on average a person with a university degree earns 15180 USD more per year than a person without a university degree.

(b)

	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
R2	0.147	0.158
Adjusted R2	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	

Compared to the first model in exercise a), the coefficient *educ* slightly increased by roughly 100 USD whereas the coefficient *age* accounts on average for an additional increase of 1461 USD in annual wages, i.e. if age increases by one year, annual wages increase by 1461 USD. As a result, the constant has to be much smaller as in model 1 where we left out *age*, in fact, the smaller constant is seen in the table above. Hence, the results in the table above do not contradict the results in a) as the additional effect of age is correct by a smaller constant, yet one can say that an additional year of education has almost the same effect on annual wages as in a). All coefficients are significant on a 1%-significance level.

In the model where we regress log wages (*lw*) on *education* and *age*, the interpretation of the coefficients goes as follows: A unit change in *educ* changes yearly wages by 12.75% (exact computation) and a unit change in *age* changes yearly wages by roughly 2.5%. We use this approximated values as it is close to zero and therefore the exact value does not differ a lot from the approximated value.

(c)

```
================================================
                        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
R2                             0.206
Adjusted R2                    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
```

We look at the following model: $lw_i = \beta_1 + \beta_2 educ_i + \beta_3 age_i + \beta_4 female_i + \epsilon_i$

The coefficient β_4 for itself does not say much. It can be interpreted as the gender wage gap in the following way (as the difference is not close to zero):

$$Gender\ wage\ gap = \frac{\mathrm{E}(W_i \mid female_i = 1) - \mathrm{E}(W_i \mid female_i = 0)}{\mathrm{E}(W_i \mid female_i = 0)}$$

$$= \frac{e^{\beta_1 + \beta_2\, \mathrm{educ_i} + \beta_3 \mathrm{age_i} + \beta_4} - e^{\beta_1 + \beta_2\, \mathrm{educ_i} + \beta_3 \mathrm{age_i}}}{e^{\beta_1 + \beta_2\, \mathrm{educ_i} + \beta_3 \mathrm{age_i}}}$$

$$= e^{\beta_4} - 1 = e^{-0.443} - 1 \approx -35.77\%$$

The regression model therefore states that females are expected to earn 35.77% less than males if all other coefficients remain constant. This very high level implies a high economic relevance. The t-value (R-output) for of -183.01 leads to a very small p-value that lies below all customary confidence levels.

$$t_{score} = \frac{\hat{\beta}_4 - H_0}{SE_{\hat{\beta}_4}} \qquad H_0: \beta_4 = 0 \quad H_1: \beta_4 \neq 0$$

We test whether our estimate for $\hat{\beta}_4$ has <u>any</u> influence on our dependent variable at all or if it is zero. Thus, we construct a two-sided test where our Null hypothesis H_0 tests $\beta_4 = 0$ ($\beta_4 \neq 0$). Alternatively, a one-sided test could be performed. One-sided tests require strong conviction that a

discrimination towards female workers (i.e., lower wages) exists before performing the calculations. However, this is not the case for this exercise. Hence, we use a two-sided test.

$$t_{score} = \frac{-0.443}{0.00242} = -183.06$$

We calculate a (in absolute values) large t-value which indicates strong evidence against our Null hypothesis that our coefficient for *females* is indeed zero. In fact, our critical t-value for 99%-confidence and adjusted for the degrees of freedom is at -2.58 (see R-Code at the end) is smaller than our calculated t-score (again absolute values). We therefore reject this hypothesis as also our calculated p-value is nearly zero and can assess that there is strong evidence, that females earn indeed less than male.

(d)

	Dependent variable:		
	lw	female	residuals
	(1)	(2)	(3)
educ	0.120***	0.015***	
	(0.0004)	(0.0002)	
age	0.025***	-0.001***	
	(0.0001)	(0.0001)	
residuals			**-0.443*** **
			(0.002)
Constant	0.713***	0.229***	0.000
	(0.009)	(0.005)	(0.001)
Observations	561,076	561,076	561,076
R2	0.158	0.008	0.056
Adjusted R2	0.158	0.008	0.056
Residual Std. Error	0.920 (df = 561073)	0.493 (df = 561073)	0.894 (df = 561074)
F Statistic	52,686.040*** (df = 2; 561073)	2,255.955*** (df = 2; 561073)	33,493.160*** (df = 1; 561074)
Note:			*p<0.1; **p<0.05; ***p<0.01

Our goal with the partitioned regression is to get the same coefficient with a regression of the residuals as for the coefficient for *female* in the table in c). In a first step, we regress all explanatory variables on log-wage (*lw*). These results are summed up in the first regression in the table above. In a second step, we regress the same explanatory variables on our dummy variable *female* (regression 2 in the table above). We then regress the residuals from the second regression on the residuals of the first regression (simple linear regression model). This coefficient in the last regression will be the same as in the regression in c) where we included *female* as can be seen above: β_4 = -0.443.

(e) We want to show that

$$\hat{\beta}_1 = \bar{y} - \bar{X}'_{-1}\hat{\beta}_{-1}$$

For the sake of simplicity, we will show this in R using the model from exercise 2.1b) above. We used the mean of *age* and *educ* and multiplied them with their coefficients. When we subtract this from the mean of wage (our dependent variable), a value of -115.9171 results, which is also the value of the coefficient in the regression output in exercise b) (See R code at the end). The intuition for this is fairly simple. We know that our OLS regression line goes through $\bar{X}$ and $\bar{y}$. We subtract from $\bar{y}$ the amount of y that results by multiplying a (1 x K) vector of our averaged variables by the (K x 1) vector of our coefficients. What is left is exactly the amount of y which the coefficients do not account for i.e. is the amount of y that results when all other coefficients of our regression model are zero. Thus, the constant results.

(f)

	Dependent variable:	
	lw	
	(unrestricted)	(restricted)
educ	0.120***	0.120***
	(0.001)	(0.0004)
age	0.028***	0.025***
	(0.0002)	(0.0001)
female	-0.342***	
	(0.018)	
educ:female	0.017***	
	(0.001)	
age:female	-0.009***	
	(0.0003)	
Constant	0.759***	0.713***
	(0.011)	(0.009)
Observations	561,076	561,076
R2	0.208	0.158
Adjusted R2	0.208	0.158
Residual Std. Error	0.893 (df = 561070)	0.920 (df = 561073)
F Statistic	29,436.260*** (df = 5; 561070)	52,686.040*** (df = 2; 561073)
Note:		*p<0.1; **p<0.05; ***p<0.01

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Res.Df	2	561,071.500	2.121	561,070	561,070.8	561,072.2	561,073
RSS	2	460,969.000	19,827.080	**446,949.100**	453,959.100	467,978.900	**474,988.900**
Df	1	-3.000		-3.000	-3.000	-3.000	-3.000
Sum of Sq	1	-28,039.730		-28,039.730	-28,039.730	-28,039.730	-28,039.730
F	1	**11,733.070**		11,733.070	11,733.070	11,733.070	11,733.070
Pr(> F)	1	0.000		0.000	0.000	0.000	0.000

We compare the model from the first table above of the form

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

to the restricted model in exercise b) and want to know if $\beta_4 = \beta_5 = \beta_6 = 0$. This is done by using an F-statistics.

$$F_{stat} = \frac{N-K}{q} \frac{RSS_R - RSS_U}{RSS_U}$$

We set in these values from the tables above:

$$= \frac{561076 - 6}{3} \frac{474989 - 446949}{446949} = 11733.18$$

The F-value of 11733.18 lies above the critical F-value of 8.53 and thus, H_0: $\beta_4 = \beta_5 = \beta_6 = 0$ can be rejected and is in favour of the unrestricted model.

(g)

	Dependent variable:	
	lw	
	(male)	(female)
educ	0.120***	0.136***
	(0.001)	(0.001)
age	0.028***	0.019***
	(0.0002)	(0.0002)
Constant	0.759***	0.417***
	(0.011)	(0.014)
Observations	318,877	242,199
R2	0.199	0.152
Adjusted R2	0.199	0.152
Residual Std. Error	0.868 (df = 318874)	0.924 (df = 242196)
F Statistic	39,546.210*** (df = 2; 318874)	21,635.730*** (df = 2; 242196)
Note:		*p<0.1; **p<0.05; ***p<0.01

The coefficients for *educ* and *age* in the *males* regression are equal to the coefficient estimates in f). The coefficients for *educ* and *age* in the females regression are different from the model with the interaction terms. However, we receive the coefficients in the table above for the female sub-sample with the model with interaction in f) by assuming $female_i = 1$.

$$lw_i = \beta_1 + \beta_2 educ_i + \beta_3 age_i + \beta_4 * 1 + \beta_5 educ_i * 1 + \beta_6 age_i * 1$$

$$= (\beta_1 + \beta_4) + (\beta_2 + \beta_5) educ_i + (\beta_3 + \beta_6) age_i$$

$$= (0.759 + (-0.342) + (0.12 + 0.017) educ_i + (0.028 + (-0.009) age_i$$

$$= 0.417 + 0.137 educ_i + 0.019 age_i$$

Whereas the last formula above describes exactly the regression in the table above for *female*. Thus, we can conclude that including the interaction terms for a dummy variable is equivalent to running the regression in the two separate subsamples.

(h) It is not possible to include all the dummies in the regression, as this would lead to perfect multi-collinearity, i.e. one variable can be expressed as a linear combination of the others. A variable wit n categories thus is expressed by n-1 dummies. For example, the information regarding each five occupations can be expressed by including four dummies. The effect of the fifth variable, which is not included with a dummy, is expressed if all other dummies are simultaneously zero.

(i)

	Dependent variable:				
	lw				
	(health)	(science)	(tech)	(business)	(other)
educ	0.136***	0.061***	0.092***	0.124***	0.104***
	(0.002)	(0.006)	(0.002)	(0.002)	(0.0005)
age	0.022***	0.031***	0.021***	0.028***	0.023***
	(0.0004)	(0.002)	(0.0004)	(0.0004)	(0.0002)
female	-0.358***	-0.140***	-0.232***	-0.361***	-0.463***
	(0.009)	(0.026)	(0.009)	(0.007)	(0.003)
Constant	0.941***	1.938***	1.851***	1.088***	1.131***
	(0.036)	(0.123)	(0.036)	(0.031)	(0.010)
Observations	40,998	2,440	30,665	45,466	441,507
R2	0.227	0.212	0.153	0.244	0.163
Adjusted R2	0.227	0.211	0.153	0.244	0.163
Residual Std. Error	0.750 (df = 40994)	0.631 (df = 2436)	0.618 (df = 30661)	0.750 (df = 45462)	0.911 (df = 441503)
F Statistic	4,015.569*** (df = 3; 40994)	218.977*** (df = 3; 2436)	1,850.308*** (df = 3; 30661)	4,898.794*** (df = 3; 45462)	28,639.610*** (df = 3; 441503)
Note:					*p<0.1; **p<0.05; ***p<0.01

The table above shows, that there is a significant (on a 1%-level) gender wage gap across all occupations. Hence, the coefficient for *female* is statistically different across all occupations. The coefficient is the largest for *other* and the smallest in science-related jobs.

(j) (i)

	Dependent variable:				
	lw				
	(health)	(science)	(tech)	(business)	(other)
educ	0.125***	0.034***	0.104***	0.126***	0.113***
	(0.002)	(0.008)	(0.004)	(0.002)	(0.001)
age	0.016***	0.025***	0.018***	0.022***	0.018***
	(0.0005)	(0.002)	(0.001)	(0.001)	(0.0002)
childrenly	-0.010	0.232***	-0.020	-0.052*	-0.071***
	(0.018)	(0.084)	(0.050)	(0.028)	(0.012)
Constant	0.994***	2.537***	1.509***	0.943***	0.735***
	(0.039)	(0.176)	(0.091)	(0.049)	(0.017)
Observations	31,298	1,025	5,549	18,689	185,638
R2	0.141	0.137	0.128	0.173	0.107
Adjusted R2	0.141	0.135	0.128	0.173	0.107
Residual Std. Error	0.713 (df = 31294)	0.561 (df = 1021)	0.655 (df = 5545)	0.739 (df = 18685)	0.949 (df = 185634)
F Statistic	1,714.657*** (df = 3; 31294)	54.147*** (df = 3; 1021)	272.404*** (df = 3; 5545)	1,305.685*** (df = 3; 18685)	7,425.929*** (df = 3; 185634)
Note:					*p<0.1; **p<0.05; ***p<0.01

The table above sums up the effect of *childrenly* across occupations. We see that the effect for of giving birth to a child in the last year does not decrease estimates log-wages for all occupations. For women in science, the coefficient is even positive and significant. However, for most women (in *other* with n=185638), having the effect of giving birth to a child does indeed decrease estimated log-wages in the next year as the coefficient *childrenly* is negative and significant in the regression for *others.* For women in *technology childrenly* is slightly negative but not statistically different from zero. Therefore, we cannot conclude a reliable influence of *childrenly* for women in tech. The effect might be as well positive for a large proportion of women.

We calculate a T-test to test the following hypothesis:

$$H_0: \beta_{childrenly} > 0$$

$$t_{score} = \frac{-0.02}{0.05} = -0.4$$

whereas df are 5545.

The critical t-value for the one-sided t-test with 5545 df lies at 1.645 ant the p-value for the obtained t-statistic is 0.344586. We can therefore not reject the Null hypothesis and it is not possible to state that there is an effect of *childrenly* in the technology sector.

(ii) We want to test if $\beta_{business} = \beta_{science} \rightarrow \beta_{business} - \beta_{science} = 0$

```
================================================
                        Dependent variable:
                    ---------------------------
                               lw
------------------------------------------------
educ                        0.114***
                             (0.001)

age                         0.018***
                            (0.0002)

childrenly                  -0.057***
                             (0.009)

occupation_healthcare       0.407***
                             (0.006)

occupation_science          0.534***
                             (0.028)

occupation_technology       0.639***
                             (0.012)

occupation_business         0.601***
                             (0.007)

Constant                    0.704***
                             (0.014)
------------------------------------------------
Observations                 242,199
R2                            0.194
Adjusted R2                   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
```

From the table above we see that $\hat{\beta}_{business} = 0.601$ with standarderror $se(\hat{\beta}_{business}) = 0.00698$ and $\hat{\beta}_{science} = 0.534$ with $se(\hat{\beta}_{science}) = 0.02834$. From the R-output (see R-code at the end) we further calculate a covariance between $\hat{\beta}_{business}$ and $\hat{\beta}_{science}$ of 0.00267.

We now test the following Null hypothesis by using a t-test.

$$H_0: \beta_{business} = \beta_{science}$$

$$t_{statistic} = \frac{\hat{\beta}_{business} - \hat{\beta}_{science}}{se(\hat{\beta}_{business} - \hat{\beta}_{science})}$$

$$= \frac{\hat{\beta}_{business} - \hat{\beta}_{science}}{\left[var(\hat{\beta}_{business}) + var(\hat{\beta}_{science}) - 2cov(\hat{\beta}_{business}, \hat{\beta}_{science})\right]^{0.5}}$$

$$\frac{0.601 - 0.534}{0.007 + 0.028 - 2 * 0.00267}$$

$$= \frac{0.067}{0.02966} = 2.314757$$

The p-value for $t_{statistic} = 2.314757$ for a two-sided hypothesis test with 242191 df lies at approximately 2.06%. We therefore reject the Null hypothesis that the two coefficients for science and business are equal on a 95% confidence level. (Intuition: Given the observed betas, there is a 2.06% chance that the true betas are equal.)

We also calculate an F-statistic in R to test the above hypothesis.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Res.Df	2	242,191.500	0.707	242,191	242,191.2	242,191.8	242,192
RSS	2	196,230.300	3.088	196,228.200	196,229.300	196,231.400	196,232.500
Df	1	-1.000		-1.000	-1.000	-1.000	-1.000
Sum of Sq	1	-4.367		-4.367	-4.367	-4.367	-4.367
F	1	5.390		5.390	5.390	5.390	5.390
Pr(> F)	1	**0.020**		0.020	0.020	0.020	0.020

Here we tested two models. The first model is

$$\log(wage) = \beta_1 + \beta_2 educ + \beta_4 age + \beta_5 childrenly + \beta_6 healthcare + \beta_7 science + \beta_8 technology + \beta_9 business$$

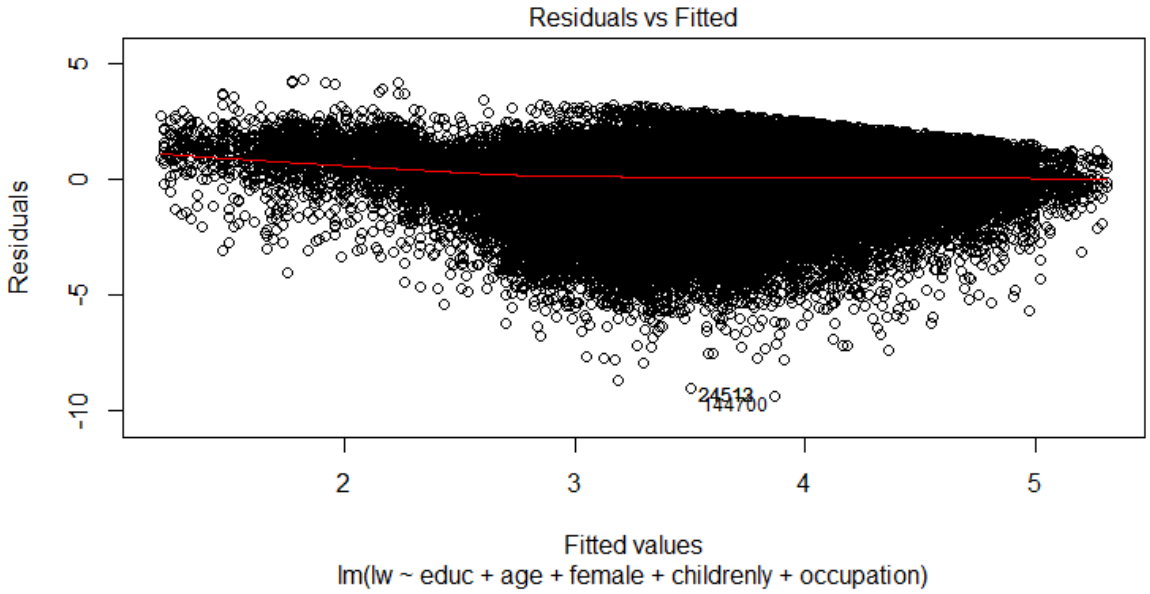
Which we test against the restricted model:

$$\log(wage) = \beta_1 + \beta_2 educ + \beta_4 age + \beta_5 childrenly + \beta_6 healthcare + \beta_7(science + business) + \beta_8 technology$$

We see that the p-value from R is very much the same as the one obtained by hand. The differences are due to the rounded input in the manually calculated t-statistic.

(iii) The results are comparable which the following examples shows: The constant for the separate regression of the occupation *other* is 1.131 – 0.463 = 0.668 in the table of 2.1.i). In the regression that includes all the dummies for the female subset, the constant is 0.704 (see table in 2.1.j.ii). The coefficients differ, as we have different predictor variables (namely *childrenyl*) as well as not included the interaction terms. If all the interaction terms where included, the separate regression should lead to the same results if the same predictive variables are used.

(k) One can test whether assumption two (mean zero errors) holds by looking at a residual vs. fitted plot:



We see that especially for the lower half of the fitted value, the residuals are systematically larger than zero. Thus, we cannot assume that assumption holds perfectly throughout our dataset. One possible explanation for this might be omitted variable bias. Hence, we did not include factors in our regression, which may also influence wage. For example, the professional position might be a very important factor for the wage differences that is omitted in our regression.

3 R Code

```
setwd("C:/Users/mmazid/Dropbox/Uni/Master/Herbstsemester 2018/Empirical Methods/My PS 2")

library(stargazer)

library(dummies)

# read csv file

dataPS2 = read.csv(file.choose(), header = T, sep = ",", dec = ".")

# Ex 2.1 a ----------------------

##create new variables

dataPS2$wage <- dataPS2$incwage/1000

dataPS2$lw <- log(dataPS2$wage)

dataPS2$university <- factor ( with ( dataPS2, ifelse ( ( educ <= 16 ), 0 , 1 ) ) )

##regression analysis

model1 <- lm(wage ~ educ, data= dataPS2)

model2 <- lm(wage ~ educ + university, data= dataPS2)

##generate output

stargazer(model1,model2, type= "text")

# Ex 2.1 b ----------------------

##regression analysis

model3 <- lm(wage ~ educ + age, data= dataPS2)

model3.5 <- lm(lw ~ educ + age, data= dataPS2)


##generate output
```

```
stargazer(model3, model3.5, type= "text")# Ex 2.1 c ------------------
----

##regression analysis

model4 <- lm(lw ~ educ + age + female, data = dataPS2)

##generate output

stargazer(model4, type= "text")

# get beta3 and stdv for manual calculation

sum_model4 <- summary(model4)

coef_model4 <- sum_model4$coefficients

beta3 <- coef_model4[c("female"),1]

beta3_stdv <- coef_model4[c("female"),2]

# calculation of gender gap

gen_gap <- exp(beta3)-1

# manual calculation of t stat

t <- beta3/beta3_stdv

# Find the p-value for the Student t distribution.

p_value <- pt(t, df=model4$df.residual[1])

# find critical t value.

t_crit <- qt(0.995, model4$df.residual[1])

# Ex 2.1 d ----------------------

# partitioned regression

# Step 1: regress y on all regrssors but female

model_part_1 <- lm(lw ~ educ + age, data= dataPS2)

# Step 2: regress female on all other regressors

model_part_2 <- lm(female ~ educ + age, data= dataPS2)
```

```
# Step 3: regress residuals on each other

model_part <- lm(model_part_1$residuals ~ model_part_2$residuals)

stargazer(model_part_1,model_part_2, model_part, type='text')

# Ex 2.1 e -----------------------

##calculation for the constant using model3

mean(dataPS2$wage)-((mean(dataPS2$educ,   na.rm   =   T)*model3$coeffi-
cients[2])+(mean(dataPS2$age, na.rm = T)*model3$coefficients[3]))

# Ex 2.2 f "interaction terms" ---------

# model with interaction terms

model_U <- lm(lw ~ educ + age + female + female*educ + female*age, data=
dataPS2)

stargazer(model_U, type="text")

# restricted model

model_R <- lm(lw ~ educ + age, data= dataPS2)

stargazer(model_U, model_R, type= "text")

# F test in R

anova(model_U,model_R)

stargazer(anova(model_U,model_R), type="text")

# manual calculation

df1 = 561070 # N-K

df2 = 3 # number of dropped regressors

# f stat

F_stat <- (df1/df2)*(474989-446949)/446949

# calculate critical f value

qf(0.95, df1, df2)
```

```
# Ex 2.2 g "separate reg -  male / female" ---------

model_male <- lm(lw ~ educ + age, data= subset(dataPS2, female=="0"))

model_female <- lm(lw ~ educ + age, data= subset(dataPS2, female=="1"))

stargazer(model_male,model_female, type="text")

# Ex 2.2 i "dummy var" ----------------

# create data set where the var "occupation" is encoded as 5 dummy variables.

dummy_data <- dummy.data.frame(dataPS2,names="occupation", sep= "_")

stargazer(dummy_data, type="text")

# dummy_model <- lm(lw ~ educ + age + female + occupation_healthcare +

#                     occupation_science + occupation_other +

#                   occupation_technology + occupation_business, data= dummy_data)

# last dummy var gets droped = in this model the dummies measure the effect of

# their industry compared to "business"

# regressions for all occupation subsamples

model_health <- lm(lw ~ educ + age + female, data= subset(dummy_data, occupation_healthcare=="1"))

model_science <- lm(lw ~ educ + age + female, data= subset(dummy_data, occupation_science=="1"))

model_tech <- lm(lw ~ educ + age + female, data= subset(dummy_data, occupation_technology=="1"))

model_business <- lm(lw ~ educ + age + female, data= subset(dummy_data, occupation_business=="1"))

model_other <- lm(lw ~ educ + age + female, data= subset(dummy_data, occupation_other=="1"))
```

```
stargazer(model_health, model_science, model_tech,

          model_business, model_other, type="text")

# 2.2 j "only female" ------

model_fhealth <- lm(lw ~ educ + age + childrenly, data= subset(dummy_data, female=="1" & occupation_healthcare=="1"))

model_fscience <- lm(lw ~ educ + age + childrenly, data= subset(dummy_data, female=="1" & occupation_science=="1"))

model_ftech <- lm(lw ~ educ + age + childrenly, data= subset(dummy_data, female=="1" & occupation_technology=="1"))

model_fbusiness <- lm(lw ~ educ + age + childrenly, data= subset(dummy_data, female=="1" & occupation_business=="1"))

model_fother <- lm(lw ~ educ + age + childrenly, data= subset(dummy_data, female=="1" & occupation_other=="1"))

stargazer(model_fhealth, model_fscience, model_ftech,

          model_fbusiness, model_fother, type="text")

t = -0.02/0.05

# Find the p-value for the Student t distribution (one sided)

p_value <- pt(t, df=model_ftech$df.residual[1])

# find critical t value.

t_crit <- qt(0.95, model_ftech$df.residual[1])

#full model for females with dummies for occupation

model_full <- lm(lw ~ educ + age + childrenly + occupation_healthcare +

                   occupation_science  +

                   occupation_technology + occupation_business,

                 data= subset(dummy_data, female=="1"))

stargazer(model_full, type="text")
```

```
covmatr <- vcov(model_full) # covariance matrix for estimated model

covmatr["occupation_business", "occupation_business"]^0.5 # SE for beta
business

covmatr["occupation_science", "occupation_science"]^0.5 # SE for beta
science

covmatr["occupation_science", "occupation_business"]^0.5 # sqr(Cov())

var <- covmatr["occupation_business", "occupation_business"] +
covmatr["occupation_science", "occupation_science"] - 2 * covmatr["oc-
cupation_science", "occupation_business"]

t_stat <- (0.601 - 0.534)/(var^0.5) # manually calculate t stat

p_value <- 2 * pt(t_stat, df=model_full$df.residual[1], lower=FALSE) #
get p value for two sided t test

# test beta_business = beta_science in R

model_test <- lm(lw ~ educ + age + childrenly + occupation_healthcare +

                   I(occupation_science  + occupation_business) + occu-
pation_technology,

                 data= subset(dummy_data, female=="1"))

anova(model_full,model_test) # f test for the two models = t-test as we
have q = 1

stargazer(anova(model_full,model_test), type = "text")

# 2.2 k E(e)=0 ----------------

model_large <- lm(lw ~ educ + age + female + childrenly + occupation,
data= dataPS2)

summary(model_large)

plot(model_large, which=1)
```