Exercise session 4

1) When estimating wage equations, we expect that young, inexperienced workers will have relatively low wages and that with additional experience their wages will rise, but then begin to decline after middle age, as the worker nears retirement. This lifecycle pattern of wages can be captured by introducing experience and experience squared to explain the level of wages. If we also include years of education, we have the equation:

$$Wage = \beta_0 + \beta_1 * Educ + \beta_2 * Exper + \beta_3 Exper^2 + u$$

a) What is the marginal effect of experience on wages? $\boldsymbol{\beta_2 + 2 * \beta_3 * Exper}$

b) What sign do you expect for each of the coefficients? Why? **$\boldsymbol{\beta_2}$ positive $\boldsymbol{\beta_3}$ negative, because there should be diminishing marginal increase in the wages with experience**

c) Estimate the model using data *cps_small.gdt*. Do the estimated coefficients have expecting signs?

genr exp2=exper^2

ols wage const educ exper exp2

Output:

```
Model 1: OLS, using observations 1-1000
Dependent variable: wage

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const      -9.81770      1.05496      -9.306    8.19e-020 ***
  educ        1.21007      0.0702378    17.23     2.04e-058 ***
  exper       0.340949     0.0514314     6.629    5.52e-011 ***
  exp2       -0.00509306   0.00119794   -4.252    2.32e-05  ***

Mean dependent var    10.21302   S.D. dependent var    6.246641
Sum squared resid     28420.08   S.E. of regression    5.341743
R-squared             0.270934   Adjusted R-squared    0.268738
F(3, 996)             123.3772   P-value(F)            5.98e-68
Log-likelihood       -3092.487   Akaike criterion      6192.973
Schwarz criterion     6212.604   Hannan-Quinn          6200.434
```

Yes

d) Test the hypothesis that education has no effect on wages. What do you conclude?

Test statistic for educ is very large 17.23, therefore we reject such hypothesis even without looking at critical values ☺

e) Test the hypothesis that the explanatory variables have no effect on wages. What do you conclude?

Here we are testing a joint hypothesis that $\boldsymbol{\beta_1, \beta_2\ and\ \beta_3 = 0}$, which we already have in GRETL output. See red circle in the GRETL output. The p-value is very small, therefore we reject $\mathbf{H_0}$

f) Include the dummy variable *black* in the regression. Interpret the coefficient and comment on its significance.

ols wage const educ exper exp2 black

	coefficient	std. error	t-ratio	p-value	
const	-9.55171	1.05516	-9.052	7.21e-019	***
educ	1.19881	0.0700907	17.10	1.08e-057	***
exper	0.346425	0.0512790	6.756	2.42e-011	***
exp2	-0.00523499	0.00119459	-4.382	1.30e-05	***
black	-1.71571	0.595372	-2.882	0.0040	***

Mean dependent var	10.21302	S.D. dependent var	6.246641
Sum squared resid	28184.85	S.E. of regression	5.322263
R-squared	0.276969	Adjusted R-squared	0.274062
F(4, 995)	95.28762	P-value(F)	1.18e-68
Log-likelihood	-3088.331	Akaike criterion	6186.662
Schwarz criterion	6211.200	Hannan-Quinn	6195.988

The coefficient on black is -1.71, which means that being black rather than white reduces your wages by 1.71 dollars per hour. The coefficient on black is statistically significant at the 1% level since test statistic is -2.882 and the critical value in the student table is -2.57. Also P-Value=0.004<0.01, meaning statistically significant at 1% level. Three stars in the end of variables are also indicators of statistical significance at 1% level.

g) Include the interaction term of *black* and *educ*. Interpret the coefficient and comment on its significance.

genr bleduc=black*educ

Model 3: OLS, using observations 1-1000
Dependent variable: wage

	coefficient	std. error	t-ratio	p-value	
const	-10.1179	1.08227	-9.349	5.68e-020	***
educ	1.23865	0.0721249	17.17	4.35e-058	***
exper	0.351995	0.0512321	6.871	1.13e-011	***
exp2	-0.00537840	0.00119380	-4.505	7.42e-06	***
black	6.30110	3.59031	1.755	0.0796	*
bleduc	-0.620954	0.274259	-2.264	0.0238	**

Mean dependent var	10.21302	S.D. dependent var	6.246641
Sum squared resid	28040.24	S.E. of regression	5.311261
R-squared	0.280678	Adjusted R-squared	0.277060
F(5, 994)	77.57147	P-value(F)	9.75e-69
Log-likelihood	-3085.759	Akaike criterion	6183.518
Schwarz criterion	6212.964	Hannan-Quinn	6194.709

?

Coefficient on bleduc implies that for each extra year of education blacks receive less wages than whites by 0.62. It is statistically significant at the 5% level (2 stars). Including this term also reduces significance of the black variable alone and strangely, changes its sign to positive.

h) Transform dependent variable in logarithmic form and estimate the equation. Interpret the coefficients.

genr lwage=log(wage)

ols lwage const educ exper exp2 black bleduc

```
Model 4: OLS, using observations 1-1000
Dependent variable: lwage

             coefficient    std. error    t-ratio    p-value
  -----------------------------------------------------------
  const       0.298229      0.0938407      3.178     0.0015    ***
  educ        0.110994      0.00625375    17.75      1.96e-061 ***
  exper       0.0371932     0.00444220     8.373     1.90e-016 ***
  exp2       -0.000602239   0.000103511   -5.818     8.02e-09  ***
  black       0.289908      0.311306       0.9313    0.3519
  bleduc     -0.0356783     0.0237802     -1.500     0.1338

Mean dependent var   2.166837   S.D. dependent var   0.552806
Sum squared resid    210.8106   S.E. of regression   0.460525
R-squared            0.309472   Adjusted R-squared   0.305998
F(5, 994)            89.09560   P-value(F)           1.72e-77
Log-likelihood      -640.5409   Akaike criterion     1293.082
Schwarz criterion    1322.528   Hannan-Quinn         1304.274

Log-likelihood for wage = -2807.38
```

Increasing educ by one year increases the wage by 11%

Increasing exper by one year increases the wage by 100*(0.03-0.0006*exper) percent

Black and bleduc do not have significant impact on logarithmic wages

1) Your aim is to estimate how the number of prenatal examinations and several other characteristics influence the birth weight of a baby. Your initial hypothesis is that more responsible pregnant women visit the doctor more often and this leads to healthier and thus also bigger babies.

(a) In your first specification, you run the following model:

$$bwght = \beta_0 + \beta_1\, npvis + \beta_2\, npvis^2 + \beta_3\, monpre + \beta_4\, male + \varepsilon\ ,$$

where *bwght* is birth weight of the baby (in grams), *npvis* is the number of prenatal doctor's visits, *monpre* is the month on pregnancy in which the prenatal care began and *male* is a dummy, equal to one if the baby is a boy and zero if it is a girl. You obtain the following results from Stata[1]:

Source	SS	df	MS
Model	12848047.5	4	3212011.87
RESIDUAL	570003184	1721	331204.639
TOTAL	582851231	1725	337884.772

Number of obs =	1726
F(4, 1721) =	9.70
Prob > F =	0.0000
R-SQUARED =	0.0220
Adj R-SQUARED =	0.0198
Root MSE =	575.5

bwght	Coef.	Std. Err.	t	P>|t|	[95% Conf. INTERVAL]	
npvis	53.50974	11.41313	4.69	0.000	31.12468	75.8948
npvissq	-1.173175	.3591552	-3.27	0.001	-1.877601	-.4687481
monpre	30.47033	12.40794	2.46	0.014	6.134091	54.80657
MALE	76.69243	27.76083	2.76	0.006	22.24391	131.141
_cons	2853.196	101.3073	28.16	0.000	2654.498	3051.895

i. Is there strong evidence that *npvissq* (stands for $npvis^2$) should be included in the model? **The *p*-value on the coefficient on *npvissq* is very small, and hence the variable is strongly significant and should be included in the model.**

ii. How do you interpret the negative coefficient of *npvissq*? **The negative coefficient on *npvissq* signals a concave form of the impact of the number of prenatal doctor's visits, meaning that there are decreasing returns to visiting the doctor. A possible explanation is that some number of visits is beneficiary for all pregnant women, but higher necessity of visits could mean that the pregnancy is risky for some reasons and the woman has to go to the doctor more often than usually. Such woman is also more likely to have smaller baby.**

iii. Holding *npvis* and *monpre* fixed, test the hypothesis that newborn boys weight by 100 grams more than newborn girls (at 95% confidence level).

Such hypothesis can be stated as

$$H_0: \beta_4 = 100 \;\; H_a: \beta_4 \neq 100$$

Test statistic $t = \frac{\widehat{\beta_4}-100}{SE(\widehat{\beta_4})} = \frac{76.69-100}{27.76} = -0.84 \sim t_{\infty,1721}$**=-1.96. Therefore, we failed to reject the null hypothesis that newborn boys weigh by 100 grams more than newborn girls at 95% confidence level.**

[1] Stata is a statistical software, which can be used to for econometric purposes. The Stata output is quite similar to the Gretl output you are familiar with. In particular, *Coef.* denotes the estimated coefficients, *Std.Err.* denotes the standard errors of these coefficients, *t* denotes the *t*-statistic of the test of significance of the coefficients, $P > |t|$ denotes the corresponding *p*-value.

b. A friend of yours, student of medicine, reminds you of the fact that the age of the parents (especially of the mother) might be a decisive factor for the health and for the weight of the baby. Therefore, in your second specification, you decide to include in your model also the age of the mother ($mage$) and of the father ($fage$). The results of your estimation are now the following:

Source	SS	df	MS
Model	16270165.8	6	2711694.3
Residual	563258231	1713	328813.912
Total	579528396	1719	337131.121

Number of obs =	1720
F(6, 1713) =	8.25
Prob > F =	0.0000
R-squared =	0.0281
Adj R-squared =	0.0247
Root MSE =	573.42

bwght	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
npvis	52.43859	11.40558	4.60	0.000	30.06826	74.80891
npvissq	-1.138545	.3585648	-3.18	0.002	-1.841816	-.4352743
monpre	34.35661	12.69477	2.71	0.007	9.457725	59.2555
male	74.45482	27.75247	2.68	0.007	20.02252	128.8871
mage	.5285275	4.218069	0.13	0.900	-7.744582	8.801637
fage	8.697342	3.465973	2.51	0.012	1.899357	15.49533
_cons	2592.813	139.6173	18.57	0.000	2318.974	2866.651

i. Comment on the significance of the coefficients on $mage$ and $fage$ separately: are they in line with your friend's claim?

When we look on the p-values of the corresponding coefficients, we see that whereas $fage$ is significant at 99% confidence level, $mage$ is insignificant. This is not in line with our friend's claim, who says that especially the age of the mother should be an important factor.

ii. Test the hypothesis that $mage$ and $fage$ are jointly significant (at 95% confidence level). Is the result in line with your friend's claim? **To test joint significance, we need restricted and unrestricted models. In the regression in part (b) we have included *mage* and *fage* while they are not included in the regression in part (a). Therefore, we can use SSR from both regression outputs in order to judge the joint significance of the *mage* and *fage* variables. According to output in part (a) SSR_r=570003184, According to output in part (b) SSR_{ur}=563258231. We construct F test based on the formula: $F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/df}$, where q is the number of restrictions in this case q=2 (mage and fage) and df is degrees of freedom.**

Df=n-k-1=1720-7

Therefore, $F = \frac{(570003184-563258231)/2}{563258231/1713} = 10.36$ in the F-table we will find a critical value at 5% it will be $F_{2,\infty} = 3.00$.

10.36>3, hence, we can reject the null hypothesis and we conclude that mage and fage are jointly significant.

iii. How can you reconcile you findings from the two previous questions?

The finding about the joint significance from the second question is not surprising,

since we know already from the first question that *fage* is individually significant. If a variable is significant, then the H_A of the test of the joint significance has to be valid and so the variables have to be jointly significant.

c) In your third specification, you decide to drop *fage* and you get the following results:

Source	SS	df	MS
Model	14451685.6	5	2890337.13
Residual	568399545	1720	330464.852
Total	582851231	1725	337884.772

Number of obs =	1726
F(5, 1720) =	8.75
Prob > F =	0.0000
R-squared =	0.0248
Adj R-squared =	0.0220
Root MSE =	574.86

bwght	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
npvis	52.27885	11.41406	4.58	0.000	29.89196	74.66575
npvissq	-1.142647	.3590214	-3.18	0.001	-1.846811	-.4384821
monpre	35.25912	12.58328	2.80	0.005	10.57898	59.93927
male	79.38175	27.75667	2.86	0.004	24.94136	133.8221
mage	-6.91257	3.137972	-2.20	0.028	-13.06721	-.757928
_cons	2648.851	137.2778	19.30	0.000	2379.602	2918.1

Comment on the significance of the coefficient on *mage*, compared to the results from part (b). Is your finding in line with your reasoning in part (b)? Does it confirm your friend's claim?

Now, the *p*-value of the coefficient on *mage* is very low and so the coefficient is strongly significant. When we compare this finding to part (b), we realize that the insignificance of this coefficient in that part was probably given by a strong correlation between *mage* and *fage*, leading to the multicollinearity problem, which increases the standard errors and decreases thus the significance of the coefficients. When we drop *fage*, the multicollinearity problem is solved and we see that our friend's claim was true.

d) Having regained trust in your friend, you consult your results once more with him. Together, you come up with an interesting question: whether smoking during pregnancy can affect the weight of the baby. Fortunately, you have at your disposition the variable *cigs*, standing for the average number of cigarettes each woman in your sample smokes per day during the pregnancy, and so you can include it in your model. However, your friend warns you that women who smoke during pregnancy are in general less responsible than those who do not smoke, and that these women also tend to visit the doctor less often. (In other words, the more the women smokes, the less prenatal doctor's visits she has). This is an important fact that you have to take into consideration while interpreting your final results, which are:

Source	SS	df	MS
Model	14560828.9	6	2426804.81
Residual	523281374	1615	324013.235
Total	537842203	1621	331796.547

Number of obs =	1622
F(6, 1615) =	7.49
Prob > F =	0.0000
R-squared =	0.0271
Adj R-squared =	0.0235
Root MSE =	569.22

bwght	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
npvis	42.43442	11.59582	3.66	0.000	19.68999	65.17885
npvissq	-.8948737	.3624432	-2.47	0.014	-1.605782	-.1839653
monpre	31.77658	12.78156	2.49	0.013	6.706395	56.84676
male	82.39438	28.34937	2.91	0.004	26.78897	137.9998
mage	-6.980738	3.227181	-2.16	0.031	-13.31064	-.6508356
cigs	-10.209	3.398309	-3.00	0.003	-16.87456	-3.54344
_cons	2748.856	141.868	19.38	0.000	2470.591	3027.12

i. Interpret the coefficient on *cigs*.

The coefficient on *cigs* tells us that with each additional cigarette smoked by the pregnant woman on average per day, the weight of the baby is smaller by 10 grams, ceteris paribus.

ii. What evidence do you find that *cigs* really should be included in the model? List at least two arguments.

We can see from the *p*-value that the coefficient on *cigs* is strongly significant. We can also see that the R^2 as well as the adjusted R^2 are higher than in the model without this variable (in part (c)). Moreover, we see that the coefficient on *npvis* has changed quite a lot once we included *cigs*, which is a signal of an omitted variable bias in part (c) and a proof that *cigs* indeed should be included in the model.

iii. Compare the coefficient on *npvis* with the one you obtained in part (c). Do you think there was a bias? If yes, explain where it came from and interpret its sign.

In part (c), the coefficient on *npvis* was approximatively equal to 52, now it is equal to 42. This shows there was a positive bias in part (c): the coefficient was overestimated there. We know that the sign of this bias is the sign of the product of two correlations: the correlation between the omitted variable *cigs* and the variable *npvis* and the correlation between *cigs* and the dependent variable *bwght*. The correlation between *cigs* and the dependent variable *bwght* is negative as we can see from the negative coefficient on *cigs* in the model estimated in part (d), the correlation between *cigs* and *npvis* is negative as we learn from our friend (women who smoke tend to visit the doctor less often). The product of these two correlations is thus positive and so is the bias in part (c).

Intuitively, we can say that when *cigs* was omitted, everything that could measure the degree of responsibility of pregnant women in our model was the variable *npvis*. Once we included *cigs*, we can measure separately the responsibility of going to the doctor and the responsibility of not smoking, and so the coefficient on *npvs* is reflecting only

the correct part of this influence and it is not overestimated.