Home assignment # 2

(Deadline: Friday, November 10, 11:00 a.m., by email or a hard-copy in class, absolutely no late submissions will be accepted)

In this assignment, there is one computer exercise that is to be computed using Gretl. No other statistical software is allowed. You should present your results as a printout from the program (e.g. copy the output from Gretl to MS Word and print it out, or print it out directly from Gretl). When you are asked to comment your results, you can do so in the printout or on a separate sheet. Do not forget that when you are asked to test a hypothesis, it is not sufficient to present just the result of the test as it is presented in Gretl: you have to provide a clear conclusion whether you reject or not the null hypothesis, which has to be formally stated.

1. Your are given the following model

$$y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon$$

   Assume that you want to test the following set of restrictions:

   (a) $\beta_2 - \beta_3 = 1$

   (b) $\beta_4 = \beta_6$ and $\beta_5 = 0$

   Construct models that incorporate restrictions (a) and (b), separately and together. Describe what test you will use to test the restrictions, including its distribution and parameters (i.e., describe how would you test: the restriction (a), the restrictions (b), and all of them together).

2. Imagine you are interested in the determinants of the revenues in shoe stores in Prague. Suppose you have specified the following model:

$$Rev_t = \alpha + \beta Inc_t + \delta Price_t + \theta Popul_t + \eta Weekend_t + \varepsilon_t \quad ,$$

   where Rev_t denotes the amount of revenues in the Prague shoe stores on a particular day t, Inc_t is per capita income in Prague, $Price_t$ is a price index for shoes relative to other goods in Prague, $Popul_t$ is number of people living in Prague, and $Weekend_t$ is a dummy variable for weekend days.

   (a) This specification recognizes that people might go shopping for shoes more often on weekends than on working days. Explain how would you test for such a hypothesis.

   (b) What are the predicted revenues (in terms of the coefficients of the model) for weekends and for working days?

(c) Explain how would you alter the specification to account for the fact that people may buy more shoes during the sales period, which is in January and July.

(d) If people have higher income, they buy more shoes on weekends (i.e., the effect of per capita income on revenues is larger on weekends compared to working days). Is this incorporated in your specification? If not, how would you do it? How would you test for the hypothesis that if people have higher income, they buy more shoes on weekends?

3. Use data *ceosal2.gdt* for this exercise. Consider an equation to explain salaries of CEOs in terms of annual firm sales:

$$\ln(salary) = \beta_0 + \beta_1 \ln(sales) + \beta_2 roe + \beta_3 neg\_ros + \varepsilon \quad ,$$

where

salary	...	CEO's salary in thousands USD
sales	...	firm's sales in millions USD
roe	...	firm's return on equity
neg_ros	...	dummy, equal to 1 if return on firm's stock is negative

(a) Define the variables you need and estimate the equation.

(b) What is the interpretation of the coefficients β_1, β_2, and β_3?

(c) Test for the presence of a significant impact of firm's sales on CEO's salary by hand (using only the estimated coefficient and the standard error from the Gretl output) and then compare your results to the results of this test in Gretl. Define the null and alternative hypothesis, the test statistic, its distribution, and interpret the results of the test.

(d) You wonder if the impact of firm's return on equity on the CEO's salary is indeed linear. You decide to test for the presence of a non-linear relationship, which you approximate by a third order polynomial of roe (i.e., $\alpha_1 roe + \alpha_2 roe^2 + \alpha_3 roe^3$).

   i. Define the null and alternative hypothesis, the test statistics and its distribution. Describe all specifications you need to be able to conduct the test, construct the necessary variables, and estimate these specifications in Gretl.

   ii. Calculate the test statistics by hand, compare to the critical value at 99% significance level, and interpret the results.

   iii. Conduct the test in Gretl and compare the results.

4. Suppose that you have a sample of n individuals who apart from their mother tongue (Czech) can speak English, German, or are trilingual (i.e., all individuals in your

sample speak in addition to their mother tongue at least one foreign language). You estimate the following model:

$$wage = \beta_0 + \beta_1 educ + \beta_2 IQ + \beta_3 exper + \beta_4 DM + \beta_5 Germ + \beta_6 Engl + \varepsilon \quad ,$$

where

$educ$	...	years of education
IQ	...	IQ level
$exper$	...	years of on-the-job experience
DM	...	dummy, equal to one for males and zero for females
$Germ$	...	dummy, equal to one for German speakers and zero otherwise
$Engl$	...	dummy, equal to one for English speakers and zero otherwise

(a) Explain why a dummy equal to one for trilingual people and zero otherwise is not included in the model.

(b) Explain how you would test for discrimination against females (in the sense that *ceteris paribus* females earn less than males). Be specific: state the hypothesis, give the test statistic and its distribution.

(c) Explain how you would measure the payoff (in term of wage) to someone of becoming trilingual given that he can already speak (i) English, (ii) German.

(d) Explain how you would test if the influence of on-the-job experience is greater for males than for females. Be specific: specify the model, state the hypothesis, give the test statistic and its distribution.

5. 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:

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

   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 form Stata:[1]

[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.

Source	SS	df	MS		
Model	12848047.5	4	3212011.87	Number of obs =	1726
Residual	570003184	1721	331204.639	F(4, 1721) =	9.70
				Prob > F =	0.0000
Total	582851231	1725	337884.772	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?

ii. How do you interpret the negative coefficient of *npvissq*?

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

(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	Number of obs =	1720
Residual	563258231	1713	328813.912	F(6, 1713) =	8.25
				Prob > F =	0.0000
Total	579528396	1719	337131.121	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?

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?

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

(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?

(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*.

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

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.