Home assignment # 3

(Suggested solutions)

1. Decide if the following claims are true or false (and explain why):

   (a) When we add an omitted variable into a regression model, the coefficients of the remaining variables can change, but they cannot lose significance.

   (b) Compared with the unconstrained regression, estimation of a least squares regression under a constraint (say $\beta_2 = \beta_3$) will result in a higher R^2 if the constraint is true and a lower R^2 if it is false.

   (c) Since x^2 is an exact function of x, we will be faced with the exact multicollinearity if we attempt to use both x and x^2 as regressors.

   (d) If we reject the null hypothesis of the RESET test, we conclude that the our model is correctly specified.

   Solution:

   (a) FALSE. If we add an omitted variable into a regression model, the coefficients of variables that are correlated with that variable will change (they suffered from omitted variable bias before we added that variable), and they can loose significance as well. For example, a variable can loose significance, because its true effect on the dependent variable is zero, but the coefficient in the model with omitted variable was biased towards positive effect (it was positive and significant, and then it lost its significance when the omitted variable was added).

   (b) FALSE. Estimation of a constrained model will always result in lower R^2, no matter if the constraint is true or false. If the constraint is false, the drop in R^2 should be larger than if it is true.

   (c) FALSE. Exact multicollinearity problem arises only if the function linking the two variables is linear.

   (d) FALSE. The null hypothesis of the RESET test is that the model is correctly specified, because all important variables are included in the model and therefore the residuals are a white noise. If we reject the null hypothesis of the RESET test, the conclusion is that the model is misspecified and some important variables are not included in the model.

2. 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]

```
      Source |       SS       df       MS              Number of obs =    1726
-------------+------------------------------           F(  4,  1721) =    9.70
       Model |  12848047.5     4  3212011.87           Prob > F      =  0.0000
    Residual |   570003184  1721  331204.639           R-squared     =  0.0220
-------------+------------------------------           Adj R-squared =  0.0198
       Total |   582851231  1725  337884.772           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).

[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 deviations 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              Number of obs =    1720
-------------+------------------------------           F(  6,  1713) =    8.25
       Model |  16270165.8     6  2711694.3            Prob > F      =  0.0000
    Residual |   563258231  1713  328813.912           R-squared     =  0.0281
-------------+------------------------------           Adj R-squared =  0.0247
       Total |   579528396  1719  337131.121           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              Number of obs =    1726
-------------+------------------------------           F(  5,  1720) =    8.75
       Model |  14451685.6     5  2890337.13           Prob > F      =  0.0000
    Residual |   568399545  1720  330464.852           R-squared     =  0.0248
-------------+------------------------------           Adj R-squared =  0.0220
       Total |   582851231  1725  337884.772           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              Number of obs =    1622
-------------+------------------------------           F(  6,  1615) =    7.49
       Model |  14560828.9     6  2426804.81           Prob > F      =  0.0000
    Residual |   523281374  1615  324013.235           R-squared     =  0.0271
-------------+------------------------------           Adj R-squared =  0.0235
       Total |   537842203  1621  331796.547           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.

Solution:

(a) i. 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. 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. Such hypothesis can be stated as

$$H_0: \ \beta_4 = 100 \ \text{ vs } \ H_A: \ \beta_4 \neq 100 \ .$$

The test statistic is

$$t = \frac{\widehat{\beta}_4 - 100}{s.e.(\widehat{\beta}_4)} \sim t_{n-k} \ ,$$

where $k = 5$ and $n = 1726$ in this case. When we compute this test statistic, we have to compare its absolute value to the critical value $t_{1721,0.975}$, since the test is two-sided. The statistic is

$$t = \frac{\widehat{\beta}_4 - 100}{s.e.(\widehat{\beta}_4)} = \frac{76.69243 - 100}{27.76083} = -0.839584767 \ .$$

The corresponding critical value is $t_{\infty,0.975} = 1.96$, and we see that the absolute value of the statistic is smaller than this critical value. Hence, we cannot reject the null hypothesis and we conclude that newborn boys weight by 100 grams more than newborn girls at 95% confidence level.

(b) i. 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. Let us introduce the following notation for the model from this part:

$$bwght = \beta_0 + \beta_1 \, npvis + \beta_2 \, npvis^2 + \beta_3 \, monpre + \beta_4 \, male + \beta_5 mage + \beta_6 fage + \varepsilon \ .$$

Using this notation, the hypothesis that both $mage$ and $fage$ are jointly significant can be stated as

$$H_0: \ \begin{array}{c} \beta_5 = 0 \\ \beta_6 = 0 \end{array} \ \text{ vs } \ H_A: \ \begin{array}{c} \beta_5 \neq 0 \\ \beta_6 \neq 0 \end{array} \ or \ .$$

When we incorporate the restrictions, we see that the restricted model is exactly the same as the one estimated in part (a). Hence, we can use the SSE from part (a) as SSE_R and the SSE from part (b) as SSE_U and construct the test statistic of the F-test (note that we cannot use the R^2 version of the F-test, because $SST_R \neq SST_U$ as the number of observations in the two models are different):

$$F = \frac{(SSE_R - SSE_U)/J}{(SSE_U)/(n-k)} \sim F_{J,n-k} \ ,$$

where $J = 2$, $N = 1720$ and $k = 7$ in this case. The test statistic is equal to

$$F = \frac{(570003184 - 563258231)/2}{(563258231)/(1720 - 7)} = 10.26$$

and it is larger than the corresponding critical value $F_{2,\infty;0.95} = 3.00$. Hence, we can reject the null hypothesis and we conclude that $mage$ and $fage$ are jointly significant.

iii. 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) 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) i. 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. 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. 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 sepately 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.

3. Suppose that you have a sample of n individuals who apart from their mother tongue (Czech) can speak English, German, or are trlingual (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.

Solution:

(a) If we included the dummy for people who are trilingual, we would have the complete set of dummies in the model (describing all three possible options - German speaker, English speaker, both foreign languages). Since we have the intercept in the model, this would lead to perfect multicollinearity.

(b) For women, the dummy DM is equal to 0 and the model stands as follows:

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

For men, the dummy DM is equal to 1 and the model stands as follows:

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

Therefore, ceteris paribus, the difference between the wage of men and the wage of women is equal to β_4. If this coefficient is positive, then men earn more than women. Hence, our hypothesis to be tested is

$$H_0: \ \beta_4 \leq 0 \ \text{ vs } \ H_A: \ \beta_4 > 0 \quad .$$

This leads to a one-sided t-test with the test statistic

$$t = \frac{\widehat{\beta}_4}{s.e.(\widehat{\beta}_4)} \sim t_{n-k} \quad ,$$

where $k = 7$ in this case. When we compute this test statistic, we compare it to the critical value $t_{n-7,0.95}$. If the test statistic is larger than this critical value, then we reject the H_0 at 95% confidence level and we conclude that there is discrimination against females.

(c) The payoff of a trilingual person is

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

the payoff of a German speaking person is

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

and the payoff of an English speaking person is

$$wage = \beta_0 + \beta_1 educ + \beta_2 IQ + \beta_3 exper + \beta_4 DM + \beta_6 + \varepsilon \quad .$$

Hence, by becoming trilingual, a person who can already speak English gains β_5 and a person who can already speak German gains β_6. If we assume that both coefficients are positive, this payoff should be positive.

(d) To allow the on-the-job experience to be greater for males than for females, we have to define a slope coefficient on *exper* that would be different for males and for females. We can do so using the following model:

$$wage = \beta_0 + \beta_1 educ + \beta_2 IQ + \beta_3 exper + \beta_4 DM + \beta_5 Germ + \beta_6 Engl + \beta_7 exper \cdot DM + \varepsilon \quad .$$

In this case, the impact of on the on-the-job experience on wage would be β_3 for females and $\beta_3 + \beta_7$ for males. Hence, if β_7 is positive, then men gain more from experience than women. Hence, our hypothesis to be tested is

$$H_0: \ \beta_7 \leq 0 \ \text{ vs } \ H_A: \ \beta_7 > 0 \quad .$$

This leads to a one-sided t-test with the test statistic

$$t = \frac{\widehat{\beta}_7}{s.e.(\widehat{\beta}_7)} \sim t_{n-k} \quad ,$$

where $k = 8$ in this case. When we compute this test statistic, we compare it to the critical value $t_{n-8,0.95}$. If the test statistic is larger than this critical value, then we reject the H_0 at 95% confidence level and we conclude that the influence of on-the-job experience is greater for males than for females.