Exam
Econometrics
31.01.2014

1. Exam takes 90 min.
2. This exam is a closed book exam.
3. Everybody is required to sign on the list.
4. The solution of exercise should be written on the sheet on which the exercise was printed or on the additional sheets on the back of the exam.
5. All the pages with solutions should be signed. If additional sheet is used it is very important to put the number of the exercise on the top of it.
6. Only one exercise should be solved on one sheet.
7. The minimum to obtain the pass grade is to answer two theoretical questions and to solve one exercise.

Theoretical questions

1	2	$\sum$

1. Write down the Box-Cox transformation and explain how it is used in econometrics.
2. Why discrete explanatory variable should be recoded into appropriate number of dummy variables before being included in the regression equation?

Theoretical questions cont.

3	4	$\sum$

3. Derive the variance matrix of the OLS estimator given that Classical Linear Regression Model assumptions are valid. Interpret the elements of this matrix.

4. Explain what are the advantages and disadvantages of imposing restrictions on the parameters of the model.

1	2	$\sum$

EXERCISE 1 We consider the following ,model

$$y_i = \beta x_i + \varepsilon_i, \quad \varepsilon_i \sim N\left(0, \sigma^2\right), \quad i = 1, \ldots, N$$

and we assume that x_i is nonrandom.

1. Show that estimator $\widetilde{\beta} = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}$ and estimator $\widehat{\beta} = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}$ are both unbiased.
2. Derive the variance of the estimator $\widetilde{\beta}$ and $\widehat{\beta}$ and show that the variance of $\widehat{\beta}$ is smaller that the variance of $\widetilde{\beta}$.

1	2a	2b	3	4	5	$\sum$

EXERCISE 2 We are using data set which includes following variables: the number of hours of training ($tothours$), average annual salary in USD ($avgsal$), logarithm of $avgsal$ ($lavgsal$), annual income from sales in USD ($sales$), logarithm of sales ($lsales$), dummy variable for year 1987 (dy1: 1 for year 1987, 0 otherwise), dummy variable for year 1988 (dy2: 1 for year 1988, 0 otherwise), dummy variable for year 1989 (dy3: 1 for year 1989, 0 otherwise), firm's size (large: 1 for large firm, 0 otherwise). Descriptive statistics for these variables are reported below.
Note: Significance level to be used in testing $\alpha = 0.1$. Results of the tests should be justified with respective p-values.

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      tothrs |       304    33.39474    52.42831         0        320
      avgsal |       304    18723.52    6967.911      4237      42583
     lavgsal |       304    9.773076    .3590829  8.351611   10.65921
       sales |       304     6413918     7899873    110000   4.90e+07
      lsales |       304    15.07133    1.126338  11.60824   17.70733
-------------+--------------------------------------------------------
         dy1 |       304    .3190789    .4668882         0          1
         dy2 |       304    .3322368     .471792         0          1
         dy3 |       304    .3486842    .4773396         0          1
       large |       304    .2138158    .4106743         0          1
```

1. Researcher used the random sample of 304 observations collected for years 1987, 1988 and 1989 to estimate regression of $tothours$ on the logarithm of $avgsal$ and the logarithm of sales.

```
      Source |       SS       df       MS              Number of obs =     304
-------------+------------------------------           F(  2,   301) =   10.14
       Model |  52584.9232     2  26292.4616            Prob > F      =  0.0001
    Residual |  780279.708   301  2592.29139            R-squared     =  0.0631
-------------+------------------------------           Adj R-squared =  0.0569
       Total |  832864.632   303  2748.72816            Root MSE      =  50.915

------------------------------------------------------------------------------
      tothrs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lavgsal |   17.08952   8.31468     2.06   0.041     .7272393    33.4518
      lsales |  -11.5002    2.65077    -4.34   0.000    -16.71659   -6.283811
       _cons |   39.70086  83.09774     0.48   0.633    -123.8252    203.227
------------------------------------------------------------------------------
```

   Interpret the estimated coefficient for variable $lavgsal$

2. In the next step researcher added variables $dy2$ and $dy3$ to the set of explanatory variables and has obtained the following regression results.

```
  Source |       SS       df       MS                    Number of obs =     304
---------+------------------------------                 F( 4,  299)   =    5.46
   Model |  56680.5896     4  14170.1474                 Prob > F      =  0.0003
Residual |  776184.042   299  2595.93325                 R-squared     =  0.0681
---------+------------------------------                 Adj R-squared =  0.0556
   Total |  832864.632   303 2748.72816                 Root MSE      =   50.95

------------------------------------------------------------------------------
  tothrs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
 lavgsal |   15.59959   8.405549     1.86   0.064    -.9419353    32.14112
  lsales |  -11.66302   2.655821    -4.39   0.000    -16.88949   -6.436551
     dy2 |   4.221359   7.271943     0.58   0.562    -10.08931    18.53203
     dy3 |   9.090815   7.254422     1.25   0.211    -5.185378    23.36701
   _cons |   52.14363   83.74939     0.62   0.534    -112.6693    216.9565
------------------------------------------------------------------------------
```

(a) Suggest the way the results above can be used to test the hypothesis that the number of hours of training does not depend on year. .

(b) Interprete the estimated value of the coefficient for $dy3$ (ignore the significance of this variable)

3. Researcher did the regression of the residuals from the last regression ($uhat$) and residual squared ($uhat2$) on the explanatory variables of the model and obtained the following results.

```
  Source |       SS       df       MS                    Number of obs =     304
---------+------------------------------                 F( 4,  299)   =    0.00
   Model |  1.1642e-10     4  2.9104e-11                 Prob > F      =  1.0000
Residual |  776184.046   299  2595.93326                 R-squared     =  0.0000
---------+------------------------------                 Adj R-squared = -0.0134
   Total |  776184.046   303 2561.66352                 Root MSE      =   50.95

------------------------------------------------------------------------------
    uhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
 lavgsal |   1.92e-07   8.405549     0.00   1.000    -16.54153    16.54153
  lsales |  -6.72e-08   2.655821    -0.00   1.000    -5.226469    5.226469
     dy2 |  -2.04e-07   7.271943    -0.00   1.000    -14.31067    14.31067
     dy3 |  -1.75e-07   7.254422    -0.00   1.000    -14.27619    14.27619
   _cons |  -7.52e-07   83.74939    -0.00   1.000    -164.8129    164.8129

  Source |       SS       df       MS                    Number of obs =     304
---------+------------------------------                 F( 4,  299)   =    3.82
   Model |  1.0166e+09     4  254151538                  R-squared     =  0.0486
Residual |  1.9889e+10   299 66516778.3                 Adj R-squared =  0.0359
---------+------------------------------                 Root MSE      =  8155.8
   Total |  2.0905e+10   303 68993804.9
------------------------------------------------------------------------------
   uhat2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
 lavgsal |   232.6349   1345.504     0.17   0.863    -2415.222    2880.492
  lsales |  -1631.846    425.126    -3.84   0.000    -2468.464   -795.2283
     dy2 |  -143.6411   1164.044    -0.12   0.902    -2432.397    2147.115
     dy3 |     534.56   1161.239     0.46   0.646    -1750.677    2819.797
   _cons |    24735.1   13406.04     1.85   0.006    -1647.041    51117.25
------------------------------------------------------------------------------
```

Is it possible to use the results from these regressions to test for heteroscedasticity of the error term? Justify your answer. Is it possible to use these regressions to identify the variable which is responsible for heteroscedasticity?

4. In the next regression researcher has added interaction between variables $large$ and $lsales$ ($large\_lsale = large * lsale$)and interaction between $large$ and $lavgsal$ ($large\_lavgsal = large * lavgsal$). How the results obtained from this regression can be used to verify the hypothesis that the model have the same parameters for large and small firms.

```
  Source |       SS       df       MS                    Number of obs =     304
-------------+------------------------------           F( 5,  298)   =    5.08
   Model | 65423.7882     5  13084.7576                  Prob > F      =  0.0002
Residual | 767440.843   298  2575.30484                  R-squared     =  0.0786
-------------+------------------------------           Adj R-squared =  0.0631
   Total | 832864.632   303 2748.72816                  Root MSE      =  50.747

------------------------------------------------------------------------------
      tothrs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lavgsal |   22.45324   8.665497     2.59   0.010     5.399921    39.50656
      lsales |  -9.971127   3.131076    -3.18   0.002    -16.13295   -3.809305
       large |   582.7495   273.2314     2.13   0.034     45.04194    1120.457
large_lavgsal|  -59.82844   30.77553    -1.94   0.053    -120.3933    .7364569
large_lsales |  -.0325676   9.070745    -0.00   0.997     -17.8834    17.81826
       _cons |  -34.48482   90.38981    -0.38   0.703     -212.368    143.3984
------------------------------------------------------------------------------
```

5. In the next regression the interaction between variable $large$ and variables $lavgsal$ was taken into account. New variable related to this interaction was defined as follows .

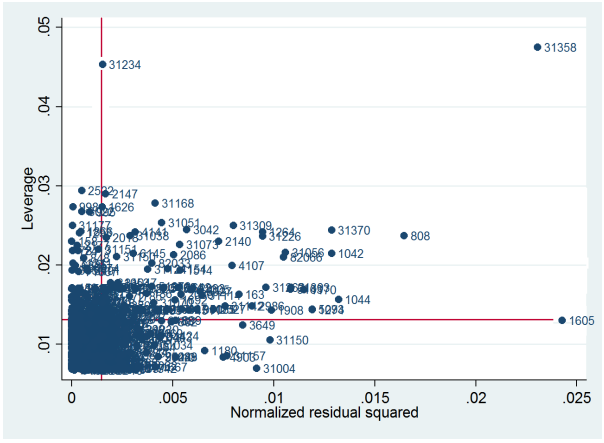
```
  Source |       SS       df       MS                    Number of obs =     304
-------------+------------------------------           F( 4,  299)   =    6.37
   Model |  65423.755     4   16355.9388                  Prob > F      =  0.0001
Residual | 767440.877   299   2566.69189                  R-squared     =  0.0786
-------------+------------------------------           Adj R-squared =  0.0662
   Total | 832864.632   303 2748.72816                  Root MSE      =  50.663

------------------------------------------------------------------------------
      tothrs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     lavgsal |   22.45481   8.640065     2.60   0.010     5.451766    39.45785
      lsales |  -9.975008   2.933707    -3.40   0.001    -15.74834   -4.20168
       large |   582.6985   272.4064     2.14   0.033     46.62183    1118.775
large_lavgsal|   -59.8758   27.75878    -2.16   0.032    -114.5031   -5.248465
       _cons |  -34.44257    89.4705    -0.38   0.701    -210.5142    141.6291
------------------------------------------------------------------------------
```

Interpret the estimated value of the coefficient for the interaction between variable $large$ and $lavgsal$.

1	2	3	$\sum$

EXERCISE 3 Using the data from Polish Labor Force survey for year 2006 the wage equation for students has been estimated. The dependent variable is the logarithm of wage ($lwage$) and the explanatory variables are as follows: age (age), place of residence (large_city: 0 city/village with population smaller than 100,000; 1 city with population larger than 100,000), type of studies ($stype$: 0 - full-time studies, 1 - evening studies, 2 - part time studies), sex (sex: 0 male, 1 female), type of employment contract (contract: 0 - fixed term employment contract, 1 - employment contract for an indefinite period), ownership (private 0 - state owned company, 1 private company) and interaction between sex and the type of the employment contract. Results of the regression is reported below.
Note: Significance level to be used in testing $\alpha = 0.1$. Results of the tests should be justified with respective p-values.

1. After the estimation of the model researcher has got the following graph representing the relationship between leverage and the standardized residual squared.



nietypowe.png

   Researcher has also calculated standardized residuals ($resid\_st$), leverage (leverage) and Cook distance. Observations for which these statistics are the largest are reported in the table below. Using the graph and the table explain which observations are suspect and why.

2. Researcher added to the model dummy variables related to the size of the city in which respondent live ($city\_1$, $city\_2$, $city\_3$) and the length of the job search ($time$) as well as the square of search time ($time2 = time^2$). Then the researcher calculated VIF statistics for variables of the model. The results of the regression and the VIF statistics are reported below.

```
Variable      |       VIF       1/VIF
--------------+----------------------
        czas  |      12.35    0.080972
       czas2  |      11.30    0.088496
sexXcontract  |       2.45    0.408996
    contract  |       2.35    0.425209
      type_1  |       1.85    0.541783
      type_2  |       1.65    0.605326
         age  |       1.63    0.615275
         sex  |       1.50    0.665691
     private  |       1.35    0.738736
  large_city  |       1.28    0.778591
--------------+----------------------
    Mean VIF  |       3.77
```

   Check whether the problem of strong multicollinearity is present in this model. How this problem can be solved in the context of this model? How the multicollinearity does influence the properties of the OLS estimator?

Variable	VIF	1/VIF
czas	12.35	0.080972
czas2	11.30	0.088496
sexXcontract	2.45	0.408996
contract	2.35	0.425209
type_1	1.85	0.541783
type_2	1.65	0.605326
age	1.63	0.615275
sex	1.50	0.665691
private	1.35	0.738736
large_city	1.28	0.778591
Mean VIF	3.77	

3. Assume that the number of hours worked influences wages. This variable is omitted from the regressions above. How this problem can influence the properties of the OLS estimator?

EXRECISE.....

NAME.....

EXRECISE.....

NAME.....