Assignment #1: Good practice example ***Labor and public economics, CERGE-EI***

Regression printouts outputs could have been shorter.

1. Creating artificial dataset:

When creating an artificial dataset, I was using real US wage data[1] as a benchmark for the plausible values, distributions and correlations of variables. My approach was to create larger dataset (consisting of 1000 observations) with matching distribution and correlation structure, then drop observation with values that were not plausible and at the end keep 200 observations serving as a basic dataset.

a. Generating RHS variables:

- **Age (`age`):** drawing from normal distribution (mean = 36, st.dev. = 12), only positive and integer values
- **Education (`edu`):** drawing from normal distribution (mean = 13, st.dev. = 4), integer values larger than 2 (I wanted to assure that a person can at least read and write, moreover, it was also minimal value in US dataset.[2]) , corr (age, edu) = `-.14` (again to account for real data feature, older people did not have the same access to higher education)
- **Error term (`e`):** drawing from normal distribution (mean = 0, st.dev. = 0.1) , correlation with other RHS variables set to 0 – orthogonality
- **Experience (exp, exp2):** I created exp = age – edu -6, so I have to assure that (age-edu)>=6; exp2 = exp^2 – this term should account for decreasing earnings profile in the higher age

b. Generating LHS variables:

For the creation of LHS variable, i.e. logy I have to set the parameter values in the basic model. I used following equation:

logY = 0.7 + 0.08*edu + 0.05*exp + 0.001*exp2 + e.

c. Summary statistics:

`sum age edu exp e logy`

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |       200      37.755    11.13711         18         75
         edu |       200      12.775     3.92702          3         23
         exp |       200       18.98    12.18078          0         59
        logy |       200    2.161799    .3423162   .8901643    2.92425
           e |       200   -.0013313    .1023772  -.2484918   .2920441
```

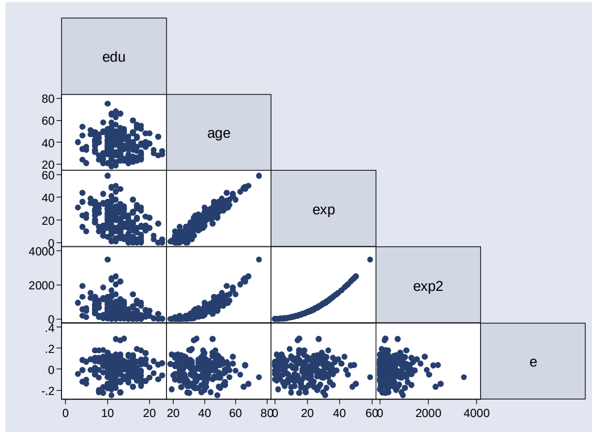
First, I present the summary statistics for all the RHS and also LHS variable. We see that RHS variables have approximately the values we have prescribed them to have (the lower variance of age can be explained by dropping observations with age<16). I also present the graphical illustration of relationship among LHS variables.

I also checked for the correlation structure of LHS variables. Note that age and education have negative relationship (although lower than I first specified) and that error term is practically uncorrelated with LHS variables (needed for unbiasedness of OLS).

---

[1] Available on www.economicswebinstitute.org/data/**wagesmicrodata**.xls .

[2] However, as for example in Slovakia school attendance is compulsory up to 10 years of study, we would have to account for this in data creation.




```
corr  age edu  e exp(obs=200)
             |      age      edu      exp     exp2        e
-------------+---------------------------------------------
         age |   1.0000
         edu |  -0.1019   1.0000
         exp |   0.9472  -0.4156   1.0000
        exp2 |   0.9069  -0.3522   0.9428   1.0000
           e |  -0.0464   0.0257  -0.0507  -0.0600   1.0000
```

2.

a. Estimating the underlying model by OLS

Underlying funct. form: `logY = a + b1*edu + c1*exp + c2*exp2 + e`

```
. reg logy edu exp exp2
      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  3,   196) =  668.04
       Model |  21.2415152     3  7.08050507           Prob > F      =  0.0000
    Residual |   2.0773738   196  .010598846           R-squared     =  0.9109
-------------+------------------------------           Adj R-squared =  0.9096
       Total |   23.318889   199  .117180347           Root MSE      =  .10295
------------------------------------------------------------------------------
        logy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |   .0802864   .0020608    38.96   0.000     .0762222    .0843506
         exp |   .0505131   .0018645    27.09   0.000      .046836    .0541902
        exp2 |  -.0010209   .0000396   -25.77   0.000    -.001099    -.0009428
       _cons |   .6958802   .0378364    18.39   0.000     .6212614     .770499
```

All the estimated coefficients are statistically significant (check p-value) and are consistent with our underlying model (`logY = 0.7 + 0.08*edu + 0.05*exp-0.001*exp2 + e`). The small differences in parameter estimates are caused by correlation of our randomly created error term and RHS variables (it is very small but still exists) resulting in a bias.

b. Omitted variables problem:

When excluding RHS variables, we basically create omitted variables problem. Thus, our estimates would be biased and the magnitude of this bias depends on the correlation with omitted variable.

```
.reg logy exp exp2 (excluding education)
      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  2,   197) =   27.95
       Model |  5.15476606     2  2.57738303           Prob > F      =  0.0000
    Residual |  18.1641229   197   .09220367           R-squared     =  0.2211
```

```
-------------+------------------------------           Adj R-squared =  0.2131
       Total |   23.318889   199  .117180347           Root MSE      =  .30365
------------------------------------------------------------------------------
        logy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   .0310744   .0052988     5.86   0.000     .0206248    .0415241
        exp2 |  -.0008196   .0001158    -7.08   0.000     -.001048   -.0005911
       _cons |   1.988245   .0536747    37.04   0.000     1.882395    2.094096
```

If we omit edu, it is contained in the error term and so we basically create endogenity (due to high correlation between edu and exp) and our OLS estimates are biased and inconsistent.

```
.reg logy edu exp2  (excluding experience)
      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  2,   197) =  134.54
       Model |  13.4623899     2  6.73119496           Prob > F      =  0.0000
    Residual |  9.85649909   197   .05003299           R-squared     =  0.5773
-------------+------------------------------           Adj R-squared =  0.5730
       Total |   23.318889   199  .117180347           Root MSE      =  .22368
------------------------------------------------------------------------------
        logy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |   .0653458   .0043142    15.15   0.000     .0568378    .0738537
        exp2 |   -.000017   .0000304    -0.56   0.577    -.0000769     .000043
       _cons |   1.335627   .0642313    20.79   0.000     1.208958    1.462297
. reg logy edu exp   (excluding experience squared)
      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  2,   197) =  153.42
       Model |  14.2011379     2  7.10056895           Prob > F      =  0.0000
    Residual |  9.11775111   197  .046283001           R-squared     =  0.6090
-------------+------------------------------           Adj R-squared =  0.6050
       Total |   23.318889   199  .117180347           Root MSE      =  .21513
------------------------------------------------------------------------------
        logy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |   .0733576   .0042696    17.18   0.000     .0649375    .0817776
         exp |   .0055572   .0013765     4.04   0.000     .0028426    .0082717
       _cons |   1.119181   .0712288    15.71   0.000     .9787118     1.25965
```

In this setting, we do not account for concave earnings- experience profile.

c. Estimation of the model using levels:

In this task we are basically estimating level – level model, while up to now we were estimating logs – level model. The main difference lies in the interpretation of the coefficients: while in the original regression the coefficient*100 were indicating the percentage change, now we are speaking about absolute changes.

Example: from the results of the log-level regression, for each additional year of education we could expect (`0.08*100)% = 8%` higher in wage, in the new specification one year of education brings additional `0.75` “units of currency” to the wage.

```
reg y edu exp exp2
      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  3,   196) =  443.32
       Model |  1691.31892     3  563.772975           Prob > F      =  0.0000
    Residual |    249.2551   196  1.27170969           R-squared     =  0.8716
-------------+------------------------------           Adj R-squared =  0.8696
       Total |  1940.57402   199  9.75162826           Root MSE      =  1.1277
```

```
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |   .7478083   .0225736    33.13   0.000     .7032899    .7923267
         exp |   .4280906   .0204236    20.96   0.000     .3878124    .4683689
        exp2 |   -.008208   .0004339   -18.92   0.000    -.0090637   -.0073523
       _cons |  -4.313831   .4144526   -10.41   0.000     -5.13119   -3.496471
------------------------------------------------------------------------------
```

d. Estimating experience of maximum earnings:

From the derivation of basic functional form log`Y = a + b1*edu + c1*exp – c2*exp2` with respect to `exp` we find that earnings are maximized at value `exp*=-c1/2*c2`. Given our underlying model , our `exp*=-0.05/2*0.001 = 25`. First, I test the difference of estimated exp* (`= 24.73978` years) from point value of 35 years:

```
testnl - (_b[exp]/(_b[exp2]*2))= 35
  (1)  - (_b[exp]/(_b[exp2]*2)) = 35
             F(1, 196) =      913.54;          Prob > F =        0.0000
```

I reject the H_0 => my estimated exp* is significantly different from 35.

Then I test the difference of estimated exp* from value given by our underlying model – 25 years.

```
testnl - (_b[exp]/(_b[exp2]*2))= 25
  (1)  - (_b[exp]/(_b[exp2]*2)) = 25
             F(1, 196) =        0.59;           Prob > F =        0.4443
```

I cannot reject the H_0 => my estimated exp* is significantly different from 35.

3.

a. Heteroskedasticity

I introduced heteroskedasticity into error term by putting `ehet=edu/4*e`. Note, that I did not change the mean, only the variance of error term by making it dependent on the value of education .

```
reg logyhet edu exp exp2
      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  3,   196) =   67.35
       Model |  21.8388399     3  7.27961331           Prob > F      =  0.0000
    Residual |  21.1837715   196  .108080467           R-squared     =  0.5076
-------------+------------------------------           Adj R-squared =  0.5001
       Total |  43.0226114   199  .216194027           Root MSE      =  .32876
------------------------------------------------------------------------------
     logyhet |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |   .0797948   .0065808    12.13   0.000     .0668165    .0927732
         exp |   .0509701    .005954     8.56   0.000     .0392279    .0627123
        exp2 |  -.0010598   .0001265    -8.38   0.000    -.0013092   -.0008103
       _cons |   .7128863   .1208243     5.90   0.000     .4746037     .951169
```

Let's test for heteroskedasticity:

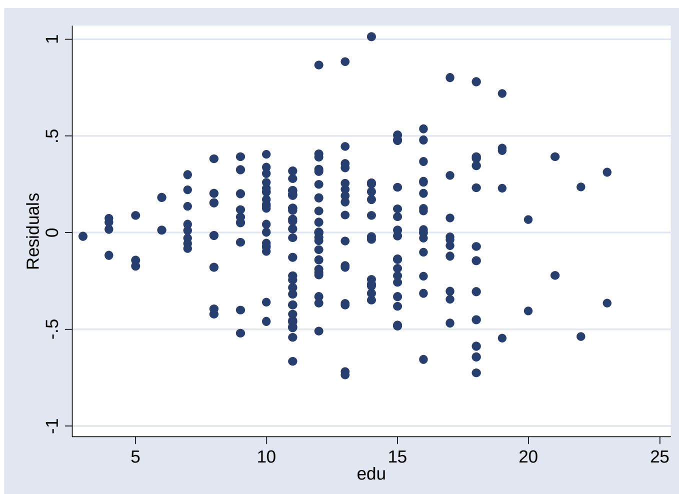
```
. hettest
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of logyhet
         chi2(1)      =    14.16
         Prob > chi2  =   0.0002
```

I reject the H_0 = > our residuals are heteroskedastic, resulting into inconsistent estimation of std. errors. We have to use White robust std. errors estimator. Apparently, the estimates of standard errors have changed.

```
Regression with robust standard errors                 Number of obs =     200
                                                       F(  3,   196) =   73.57
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.5076
                                                       Root MSE      =  .32876

------------------------------------------------------------------------------
             |               Robust
     logyhet |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |   .0797948   .0064395    12.39   0.000     .0670951    .0924945
         exp |   .0509701   .0055026     9.26   0.000     .0401181     .061822
        exp2 |  -.0010598   .0001062    -9.98   0.000    -.0012692   -.0008504
       _cons |   .7128863   .1145936     6.22   0.000     .4868916    .9388811
```

To illustrate the heteroskedasticity, we plot the residuals from regression against `edu`. We see that the variance of residuals is increasing with increasing education.



b. Measurement error in RHS variable

I introduced measurement error in the `edu` variable by creating new variable `EDUERR=edu+2.5*e1, where e1 is N(0,1).` I reestimated the basic model and obtained following results.

```
reg logy EDUERR exp exp2
      Source |       SS       df       MS              Number of obs =     911
-------------+------------------------------           F(  3,   907) =  619.36
       Model |  65.9621694     3  21.9873898           Prob > F      =  0.0000
    Residual |  32.1986838   907  .035500203           R-squared     =  0.6720
-------------+------------------------------           Adj R-squared =  0.6709
       Total |  98.1608533   910   .10786907           Root MSE      =  .18841

------------------------------------------------------------------------------
        logy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      EDUERR |   .0509627   .0014336    35.55   0.000     .0481491    .0537762
         exp |   .0466858   .0018258    25.57   0.000     .0431026     .050269
```

```
        exp2 |  -.0010121   .0000402   -25.19   0.000    -.0010909   -.0009332
       _cons |   1.144964   .0289347    39.57   0.000     1.088178    1.201751
```

See that coefficient by `EDUERR` is smaller than the true one and on the other hand coefficient by constant is much higher. Much bigger problem, however, is the endogeneity of `EDUERR` (see construction of `EDUERR`, it is now correlated with error term = e+e1). I tried to account for it by creating an instrumental variable `INSTR`, which is highly correlated with `edu` and has also similar correlation structure w.r.t. other RHS variables.

```
Instrumental variables (2SLS) regression
      Source |       SS       df       MS              Number of obs =     911
-------------+------------------------------           F(  3,   907) =  319.12
       Model |  53.9976389     3  17.999213            Prob > F      =  0.0000
    Residual |  44.1632143   907  .048691526           R-squared     =  0.5501
-------------+------------------------------           Adj R-squared =  0.5486
       Total |  98.1608533   910  .10786907            Root MSE      =  .22066
------------------------------------------------------------------------------
        logy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      EDUERR |    .077281   .0033759    22.89   0.000     .0706554    .0839065
         exp |   .0518549   .0022143    23.42   0.000     .0475092    .0562006
        exp2 |   -.001048   .0000472   -22.19   0.000    -.0011407   -.0009553
       _cons |   .7273676   .0575153    12.65   0.000     .6144891    .8402461
------------------------------------------------------------------------------
Instrumented:  EDUERR Instruments:   exp exp2 instr
```

Using instrumental variable INSTR we have achieved parameter estimates which are very similar to true parameter values. Moreover, we have solved the problem of endogeneity.

c. Measurement error in LHS variable

When introducing stochastic measurement error (uncorrelated with RHS variables) in LHS variable we basically increase the variance of this variable – in our case logY. Therefore, the parameter estimates does not change that much, but the standard errors are higher and R-squared lower than in the basic regression (as less of the variance in the data is explained).

```
. reg logYERR edu exp exp2
      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  3,   196) =  124.55
       Model |  19.7564252     3  6.58547508           Prob > F      =  0.0000
    Residual |  10.3635556   196  .052875284           R-squared     =  0.6559
-------------+------------------------------           Adj R-squared =  0.6507
       Total |  30.1199809   199  .151356688           Root MSE      =  .22995
------------------------------------------------------------------------------
     logYERR |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |   .0774717   .0046029    16.83   0.000     .0683941    .0865494
         exp |   .0494175   .0041645    11.87   0.000     .0412045    .0576305
        exp2 |  -.0009915   .0000885   -11.21   0.000     -.001166   -.0008171
       _cons |   .7298577   .0845098     8.64   0.000     .5631924     .896523
```

d. Including irrelevant variable:

We are considering the 3rd order polynomial of exp instead of 2nd order. The coefficient by exp3 turned out to be insignificant. In fact, we are including irrelevant variable, as we know that underlying model assumed only quadratic relation. By doing this, we are loosing efficiency.

```
reg logy edu exp exp2 exp3
      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  4,   195) =  500.11
       Model |  21.2476871     4  5.31192177           Prob > F      =  0.0000
    Residual |  2.07120192   195  .010621548           R-squared     =  0.9112
-------------+------------------------------           Adj R-squared =  0.9094
       Total |   23.318889   199  .117180347           Root MSE      =  .10306
------------------------------------------------------------------------------
        logy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |    .080246   .0020637    38.88   0.000      .076176     .084316
         exp |   .0478967    .003907    12.26   0.000     .0401914    .0556021
        exp2 |  -.0008853   .0001822    -4.86   0.000    -.0012446   -.0005261
        exp3 |  -1.82e-06   2.38e-06    -0.76   0.447    -6.51e-06    2.88e-06
       _cons |   .7062705   .0402549    17.54   0.000     .6268797    .7856613
```

e. Using 2nd order polynomial of age instead of exp:

As the correlation between age and exp is very high (namely 0.9472), we can use it instead of experience and obtain similar results as in original regression with respect to coefficients by age (exp) and age2 (exp2). It is basically the same system as using age as instrumental variable for edu.

```
reg logy edu age age2
      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  3,   196) =  409.81
       Model |  20.1124839     3  6.70416131           Prob > F      =  0.0000
    Residual |  3.20640509   196   .01635921           R-squared     =  0.8625
-------------+------------------------------           Adj R-squared =  0.8604
       Total |   23.318889   199  .117180347           Root MSE      =   .1279
------------------------------------------------------------------------------
        logy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         edu |    .066651   .0023217    28.71   0.000     .0620723    .0712297
         age |    .091958   .0046183    19.91   0.000       .08285     .101066
        age2 |  -.0010674   .0000562   -19.01   0.000    -.0011782   -.0009567
       _cons |  -.5082627   .0955759    -5.32   0.000    -.6967519   -.3197736
```

4. Method of splines:

I used linear spline with three knots at values 10,20 and 40 to approximate the earning-experience profile. It has brought approximately the same fit as the real = quadratic functional form (R-squared = 0.9076).

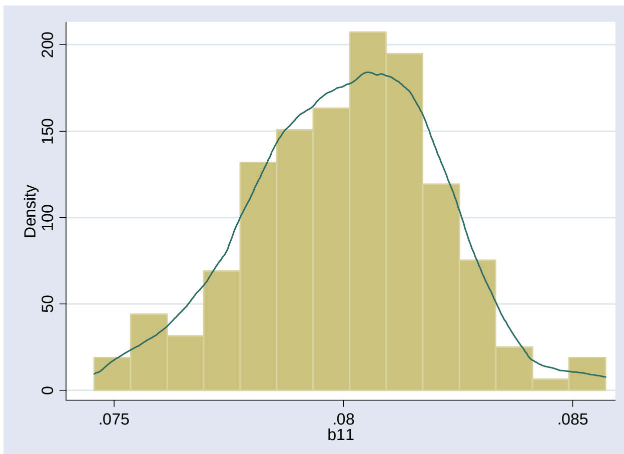
```
reg logy edu exp_1-exp_4
      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  5,   194) =  381.04
       Model |  21.1638663     5  4.23277327           Prob > F      =  0.0000
    Residual |  2.15502266   194  .011108364           R-squared     =  0.9076
-------------+------------------------------           Adj R-squared =  0.9052
       Total |   23.318889   199  .117180347           Root MSE      =   .1054
------------------------------------------------------------------------------
        logy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

edu	.0807019	.0021305	37.88	0.000	.0765	.0849038
exp_1	.0394724	.0036023	10.96	0.000	.0323676	.0465772
exp_2	.0217317	.0029518	7.36	0.000	.0159099	.0275534
exp_3	-.0058315	.0018432	-3.16	0.002	-.0094669	-.0021961
exp_4	-.0592381	.0046256	-12.81	0.000	-.068361	-.0501152
_cons	.7102411	.0428287	16.58	0.000	.6257714	.7947109

5. Mimicking the distribution of estimated coefficient b1:

We are repeating task #1 200 times using different seed for each run, saving estimated coefficient b1 from each run. We got following results:

Variable	Obs	Mean	Std. Dev.	Min	Max
b1	200	.0800368	.0020486	.0745673	.085711



As we see, the mean of the newly created variable b1 is `0.080` what is exactly the value b1 from our parameterized underlying model. In this exercise we are trying **to mimic the distribution of the estimator of b1** and we can say it is unbiased (as the mean = true value). We can also say that it is consistent and efficient, as this is the property of OLS estimators.