<p style="text-align:center;">**SOLUTIONS TO**
Problem Set 2
Introduction to Econometrics
Seyhan Erden and Tamrat Gashaw</p>

1. **[25 P]** In Problem Set 1, last week, you have calculated intercept and slope of the sample regression of *lung cancer deaths in 1950* on *cigarettes consumed per capita in 1930* for five countries given below:

Observation #	Country	Cigarettes consumed per capita in 1930 (X)	Lung cancer deaths per million people in 1950 (Y)
1	Switzerland	530	250
2	Finland	1115	350
3	Great Britain	1145	465
4	Canada	510	150
5	Denmark	380	165

This week, please calculate the same statistics using STATA. On the STATA output file, find and label the items.

i) The sample means of X and Y, $\bar{X}$ and $\bar{Y}$.

ii) The standard deviations of X and Y, s_X and s_Y.

iii) The correlation coefficient, r, between X and Y

iv) $\hat{\beta}_1$, the OLS estimated slope coefficient from the regression $Y_i = \beta_0 + \beta_1 X_i + u_i$

v) $\hat{\beta}_0$, the OLS estimated intercept term from the same regression

vi) $\hat{Y}_i$, $i = 1,\ldots, n$, the predicted values for each country from the regression

vii) $\hat{u}_i$, the OLS residual for each country.

STATA HINTS: First load STATA and type "edit," which brings up something that looks like a spreadsheet. Enter the smoking and cancer values in the first two columns. Double-click the column headers to enter variable names (e.g. "smoke", "death"). Close the editor window when you are done. The following commands will be useful:

list	lists the data (to be sure you typed it in correctly)
summarize	computes sample means and standard deviations (the option ",detail" gives additional statistics, including the sample variance)
correlate	produces correlation coefficients (with the option ", covariance" this command produces covariances)
regress	estimates regression by OLS
predict	compute OLS predicted values and residuals

Note that STATA has on-line help.

Do not be concerned if you do not yet understand all the statistics shown in the output – we will discuss them in class in due course.

Answers:

    a) Listing of the data:

```
     +------------------------+
     | country   cigs  deaths |
     |------------------------|
  1. |   Switz    530     250 |
  2. | Finland   1115     350 |
  3. | Britain   1145     465 |
  4. |  Canada    510     150 |
  5. | Denmark    380     165 |
     +------------------------+
```

    b) Mean and standard deviation:

```
. summarize cigs deaths;

    Variable |       Obs        Mean    Std. Dev.        Min        Max
-------------+--------------------------------------------------------
        cigs |         5         736     364.4071        380       1145
      deaths |         5         276     132.3537        150        465
```

    c) Correlation coefficient:

```
. * ----- compute correlation -----;
. correlate cigs deaths;
(obs=5)
             |     cigs    deaths
-------------+------------------
        cigs |   1.0000
      deaths |   0.9263    1.0000
```

    d) OLS Regression:

```
. regress deaths cigs;

      Source |       SS       df       MS              Number of obs =       5
-------------+------------------------------           F(  1,     3) =   18.12
       Model |  60116.1644     1  60116.1644           Prob > F      =  0.0238
    Residual |  9953.83564     3  3317.94521           R-squared     =  0.8579
-------------+------------------------------           Adj R-squared =  0.8106
       Total |       70070     4     17517.5           Root MSE      =  57.602


-----------------------------------------------------------------------------
      deaths |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+---------------------------------------------------------------
        cigs |   .3364177   .0790347     4.26   0.024     .084894    .5879414
       _cons |   28.39656   63.61827     0.45   0.686   -174.0652   230.8583
-----------------------------------------------------------------------------
```

$\hat{\beta}_0$ = **28.39656**

$\hat{\beta}_1$ = **.3364177**

e) Predicted values and residuals

```
. predict dhat;
(option xb assumed; fitted values)

. generate uhat = deaths - dhat;

. list deaths dhat uhat;

      +-------------------------------+
      | deaths       dhat        uhat |
      |-------------------------------|
  1.  |    250    206.698    43.30205 |
  2.  |    350   403.5023   -53.50232 |
  3.  |    465   413.5948    51.40515 |
  4.  |    150   199.9696   -49.96959 |
  5.  |    165   156.2353    8.764709 |
      +-------------------------------+
```
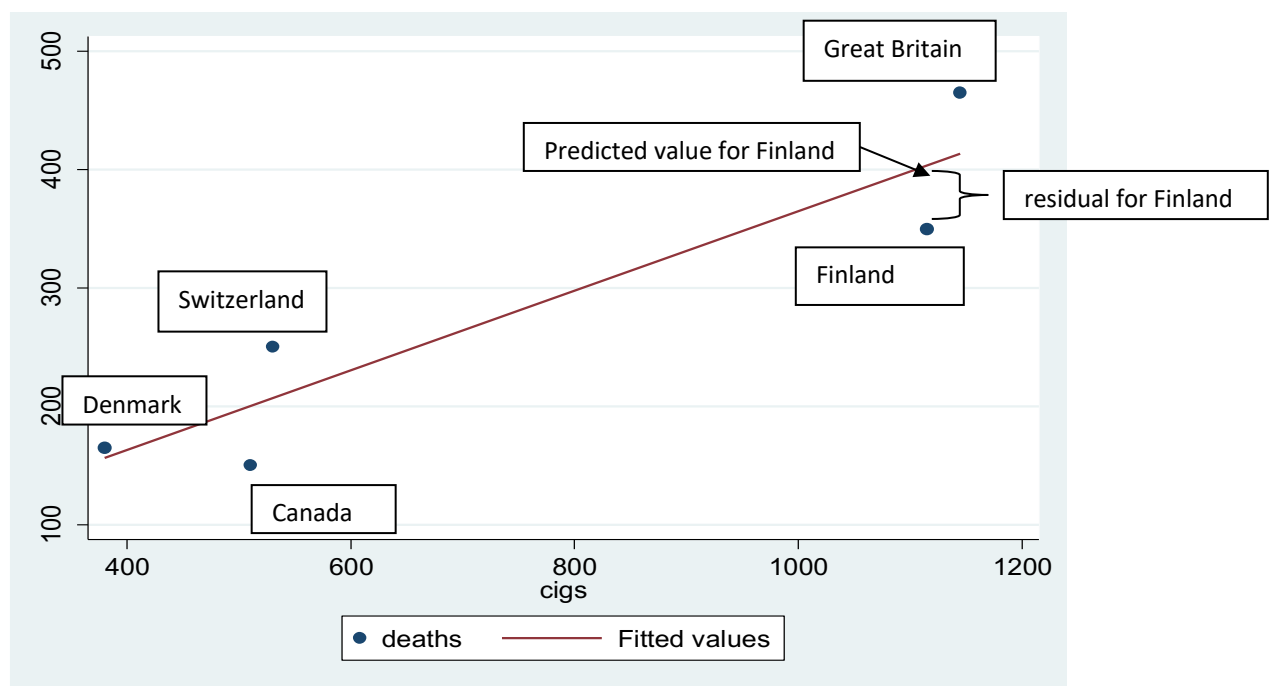
In this table, the predicted values are dhat and the residuals are uhat.

2. **[25 P]** Using "graph twoway" command in STATA, graph the scatterplot of the five data points and the regression line. Interpret sample slope and sample intercept.

Answers:

Once we run the cigar.do file the graph is generated, the command for it is

Graph twoway (scatter deaths cigs) (lfit deaths cigs)

The estimated intercept, $\hat{\beta}_0 = 28.4$, is the value at which the regression line intercepts the vertical axis. The slope of the regression line is 0.336, so an increase of one cigarette per capita is associated with an increase in the death rate of 0.336 lung cancer deaths per million.

cigar.do file

```
clear all
************************************************************
* PS2-cigar.do
* STATA calculations for W3412, problem set #2
************************************************************
log using PS2-cigar.log, replace
set more 1
************************************************************
* read in data
input str8 country cigs deaths
"Switz" 530 250
"Finland" 1115 350
"Britain" 1145 465
"Canada" 510 150
"Denmark" 380 165
end
*
list
* ---- compute mean and variance -----
summarize cigs deaths
* ----- compute correlation -----
correlate cigs deaths
* ----- regression of death rate on cigarettes per capita -----
regress deaths cigs
* ----- compute predicted values and residuals -----
predict dhat
generate uhat = deaths - dhat
list deaths dhat uhat
* ------ scatterplot and regression line -----
Graph twoway (scatter deaths cigs) (lfit deaths cigs)
log close
clear
exit
```

3. **[25 P]** Using the **WAGE** data that is posted on Coursework, answer the following questions by doing the required data analysis in STATA and report the results.

(a) [5 P] Import the data into STATA and conduct descriptive statistics analysis of the data set.

(b) **[5 P]** Graph the scatterplot for {wage, education}; {wage, experience}, and {wage, tenure} using the dataset. Say a few words about the relationship in the graphs.

(c) **[5 P]** Run separate simple regressions of wage on education; wage on experience, and wage on tenure. Interpret your results.

(d) **[5 P]** Construct a 99% confidence interval for your slope coefficient of all the three regressions. Test the null hypothesis if the slope coefficient is zero against the alternative that it is not.

Answer:

```
     name:  <unnamed>
      log:  /Volumes/CUF2018/ECON3412 FALL 2019/WAGE.RAW LOG.log
 log type:  text
opened on:  25 Sep 2019, 11:28:13
```

**(a)** Descriptive Statistics:

```
. sum

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        wage |       935    957.9455    404.3608        115       3078
       hours |       935    43.92941    7.224256         20         80
          IQ |       935    101.2824    15.05264         50        145
         KWW |       935    35.74439    7.638788         12         56
        educ |       935    13.46845    2.196654          9         18
-------------+--------------------------------------------------------
       exper |       935    11.56364    4.374586          1         23
      tenure |       935    7.234225    5.075206          0         22
         age |       935    33.08021    3.107803         28         38
     married |       935    .8930481    .3092174          0          1
       black |       935    .1283422    .3346495          0          1
-------------+--------------------------------------------------------
       south |       935    .3411765    .4743582          0          1
       urban |       935    .7176471    .4503851          0          1
        sibs |       935    2.941176    2.306254          0         14
     brthord |       852    2.276995    1.595613          1         10
       meduc |       857    10.68261    2.849756          0         18
-------------+--------------------------------------------------------
       feduc |       741    10.21727      3.3007          0         18
       lwage |       935    6.779004    .4211439    4.744932    8.032035
```

- The mean value of wage is 957.95 units, its SD = 404.36, with min value of 115 and max value of 3078.
- Do the same for the other model variables.
- If they want, they can add and interpret correlation coefficients for the variables in the data set.

(b) Graphs
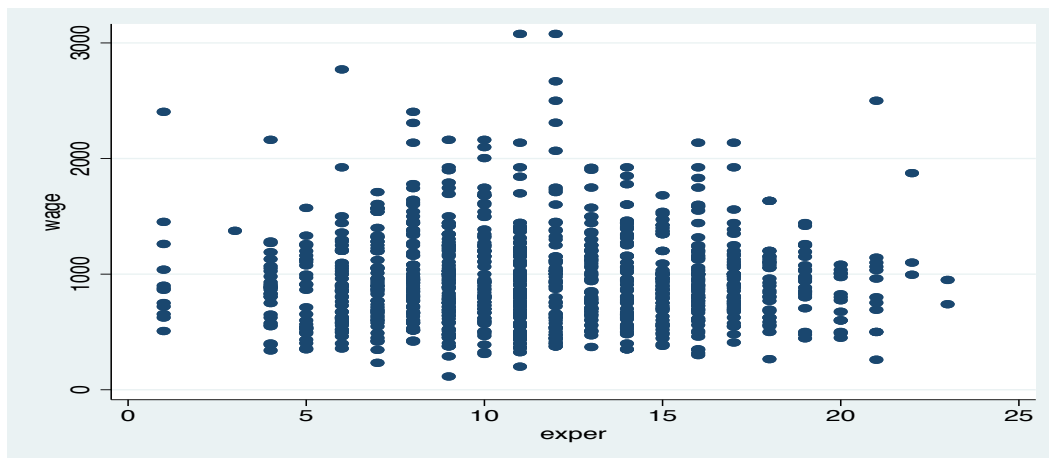
1. Wage vs Education

```
. twoway (scatter wage educ)
```
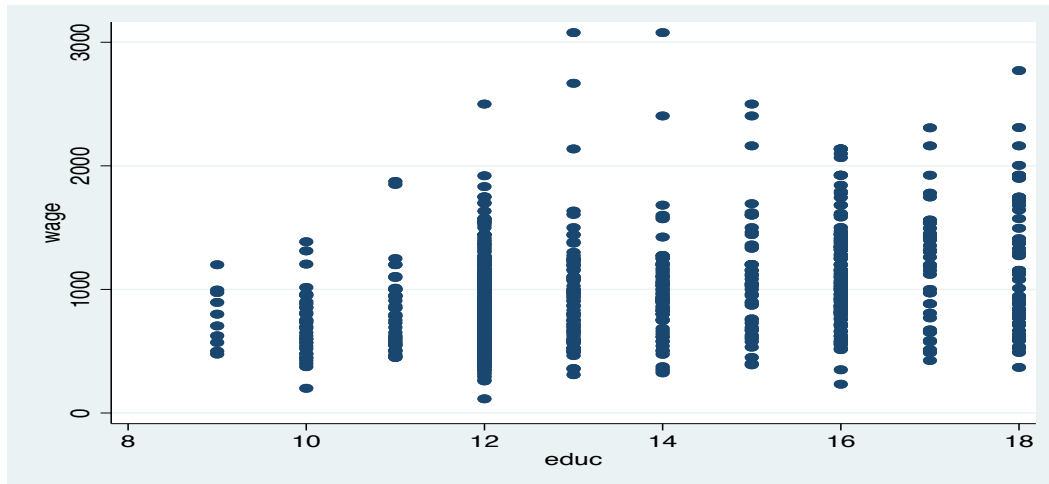
- It looks like that as the number of years of education increases, wage tends to increase.
- It seems that they are positively correlated (although it is not clear).
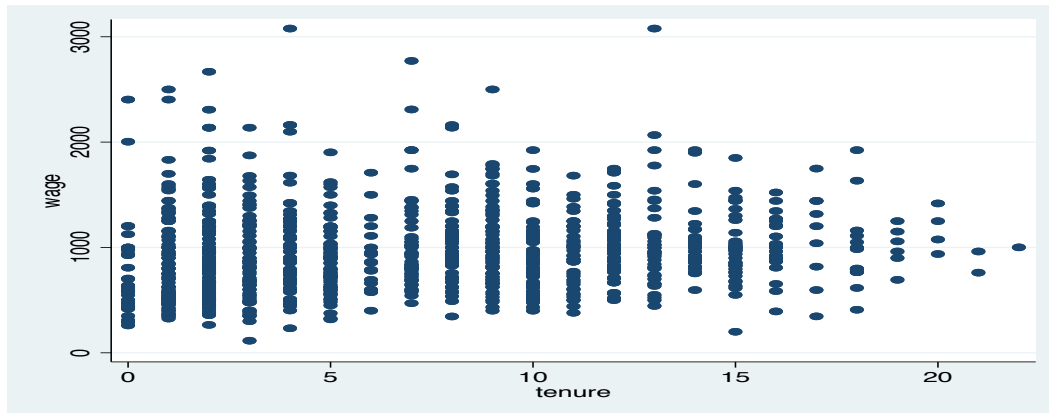
2. **Wage vs Experience**

`. twoway (scatter wage exper)`

- It looks like that as the number of years of experience increases; wage tends to increase first and starts to decrease after a certain threshold level of years of experience.
- It seems that they are not linearly correlated.

3. **Wage vs Tenure**

`. twoway (scatter wage tenure)`

- There seems that there is no meaningful or clear positive or negative relationship between these two variables by visual inspection of this plot.
- It seems that they may have a slightly positively linearly correlated.

(c) Simple Linear Regression Model
1. Wage vs Education

```
. regress wage educ

. regress wage educ

      Source |       SS           df       MS            Number of obs   =       935
-------------+----------------------------------         F(1, 933)       =    111.79
       Model |  16340644.5         1   16340644.5         Prob > F        =    0.0000
    Residual |   136375524       933   146168.836         R-squared       =    0.1070
-------------+----------------------------------         Adj R-squared   =    0.1060
       Total |   152716168       934   163507.675         Root MSE        =    382.32


--------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+------------------------------------------------------------------
        educ |   60.21428   5.694982    10.57   0.000     49.03783    71.39074
       _cons |   146.9524   77.71496     1.89   0.059     -5.56393    299.4688
--------------------------------------------------------------------------------
```

Interpretation of the result from this table:
- In this regression, the slope coefficient is statistically significant as t=10.57 (i.e., p<0.001) but the intercept is not.
- This implies that as the number of years of education increases by one unit, earnings (wage) tends to increase by 60.21 units.
- About 10.6% of the variation in wage is explained by our explanatory variable-years of education.
- The 95% confidence interval for the slope is (49.04, 71.39). This interval doesn't contain zero and hence, we can easily reject a null of zero slope coefficient. This is also the same for the intercept term as the confidence interval for the intercept doesn't contain zero in it and hence, we reject a null of zero intercept.

2. Wage vs Experience

```
. regress wage   exper

      Source |       SS           df       MS            Number of obs   =       935
-------------+----------------------------------         F(1, 933)       =      0.00
       Model |  732.242855         1   732.242855        Prob > F        =    0.9467
    Residual |   152715436       933   163682.139        R-squared       =    0.0000
-------------+----------------------------------         Adj R-squared   =   -0.0011
       Total |   152716168       934   163507.675        Root MSE        =    404.58


        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .2024031   3.026148     0.07   0.947    -5.736443    6.141249
       _cons |   955.6049    37.4111    25.54   0.000     882.1853    1029.025
```

Interpretation of the result from this table:
- In this regression, the slope coefficient is statistically significant as t=0.07 (i.e., p>0.94) and the intercept is significant at 1%.
- This implies that as the number of years of experience increases by one unit, earnings (wage) tends not to respond significantly.
- Only 1% (it is odd that it is negative. WHY?) of the variation in wage is explained by our explanatory variable-years of experience. This suggests that experience only doesn't explain much of the variation in wage.
- The 95% confidence interval for the slope is (-5.736443   6.141249). This interval does contain zero and hence, we cannot reject a null of zero slope coefficient. However, the confidence interval for the intercept doesn't contains zero in it and hence we can reject a null of zero intercept.

3. Wage vs Tenure

```
. regress wage    tenure


      Source |       SS           df       MS            Number of obs   =       935
-------------+----------------------------------         F(1, 933)       =     15.61
       Model |   2512527.2         1    2512527.2        Prob > F        =    0.0001
    Residual |   150203641       933   160989.969        R-squared       =    0.0165
-------------+----------------------------------         Adj R-squared   =    0.0154
       Total |   152716168       934   163507.675        Root MSE        =    401.24


        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      tenure |   10.21947   2.586856     3.95   0.000     5.142737     15.2962
       _cons |   884.0155   22.85589    38.68   0.000     839.1606    928.8704
```

Interpretation of the result from this table:
- In this regression, the slope coefficient is statistically significant as t=3.95 (i.e., p<0.001) at 1% and also the intercept at 1%.
- This implies that when we move from being non-tenured to being tenured, earnings (wage) tends to increase by 10.22 units.

- About 17% of the variation in wage is explained by our explanatory variable-being tenured.
- The 95% confidence interval for the slope is (5.142737   15.2962). This interval doesn't contain zero and hence, we can easily reject a null of zero slope coefficient. The same is true for the intercept term.

(d) The 99% confidence interval

To construct a 99% confidence interval, we use:

$$\beta_i \pm 2.58 \times SE(\beta_i)$$

We can also do it in STATA by adding the confidence interval level at the end of the regression command as shown below.

1. Wage vs Education

```
. regress wage educ, level(99)
```

Source	SS	df	MS		Number of obs	=	935
					F(1, 933)	=	111.79
Model	16340644.5	1	16340644.5		Prob > F	=	0.0000
Residual	136375524	933	146168.836		R-squared	=	0.1070
					Adj R-squared	=	0.1060
Total	152716168	934	163507.675		Root MSE	=	382.32

wage	Coef.	Std. Err.	t	P>|t|	[99% Conf. Interval]	
educ	60.21428	5.694982	10.57	0.000	45.51491	74.91365
_cons	146.9524	77.71496	1.89	0.059	-53.63834	347.5432

2. Wage vs Experience

```
. regress wage  exper, level(99)
```

Source	SS	df	MS		Number of obs	=	935
					F(1, 933)	=	0.00
Model	732.242855	1	732.242855		Prob > F	=	0.9467
Residual	152715436	933	163682.139		R-squared	=	0.0000
					Adj R-squared	=	-0.0011
Total	152716168	934	163507.675		Root MSE	=	404.58

wage	Coef.	Std. Err.	t	P>|t|	[99% Conf. Interval]	
exper	.2024031	3.026148	0.07	0.947	-7.608416	8.013222
_cons	955.6049	37.4111	25.54	0.000	859.0428	1052.167

3. Wage vs Tenure

```
. regress wage    tenure, level(99)

      Source |       SS           df       MS            Number of obs   =       935
-------------+----------------------------------        F(1, 933)       =     15.61
       Model |   2512527.2            1   2512527.2      Prob > F        =    0.0001
    Residual |   150203641          933  160989.969      R-squared       =    0.0165
-------------+----------------------------------        Adj R-squared   =    0.0154
       Total |   152716168          934  163507.675      Root MSE        =    401.24


------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [99% Conf. Interval]
-------------+----------------------------------------------------------------
      tenure |   10.21947   2.586856     3.95   0.000      3.54251    16.89643
       _cons |   884.0155   22.85589    38.68   0.000      825.022    943.0091
------------------------------------------------------------------------------

. log close
```