Department of Economics
Columbia University

UN3412
Fall 2021

SOLUTIONS TO
Problem Set 2
Introduction to Econometrics
Seyhan Erden and Tamrat Gashaw

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

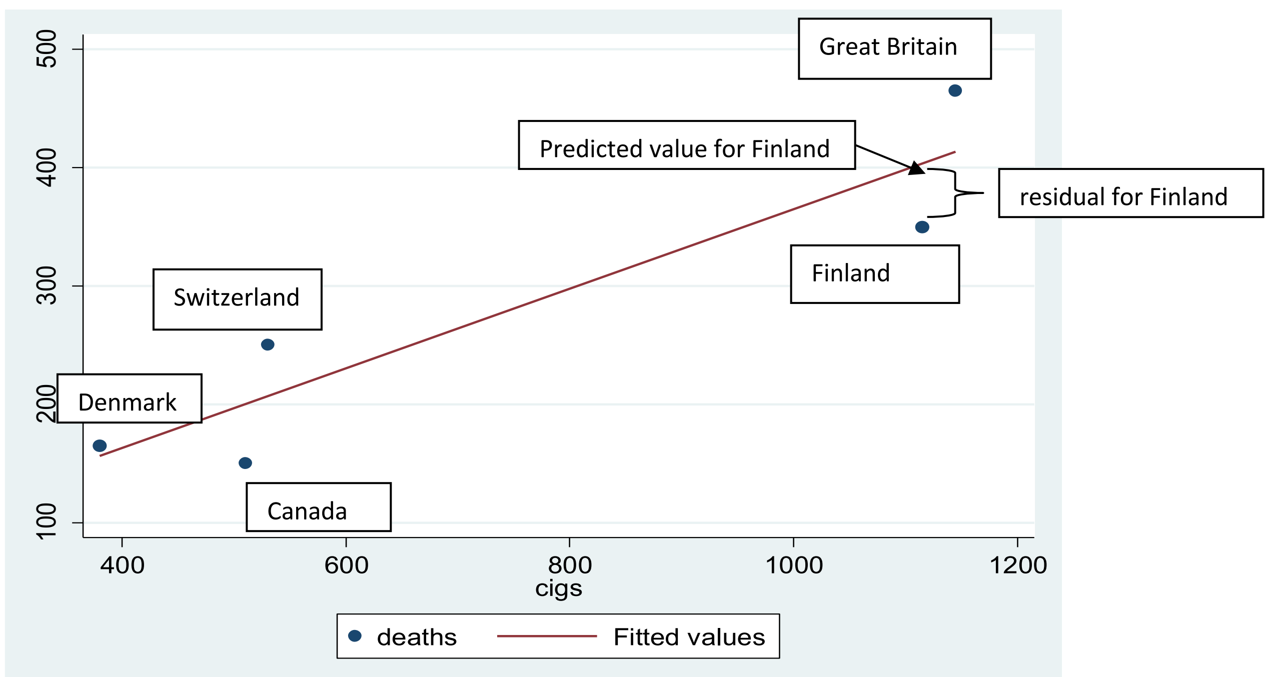
```
. predict dhat;
(option xb assumed; fitted values)

. 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
```

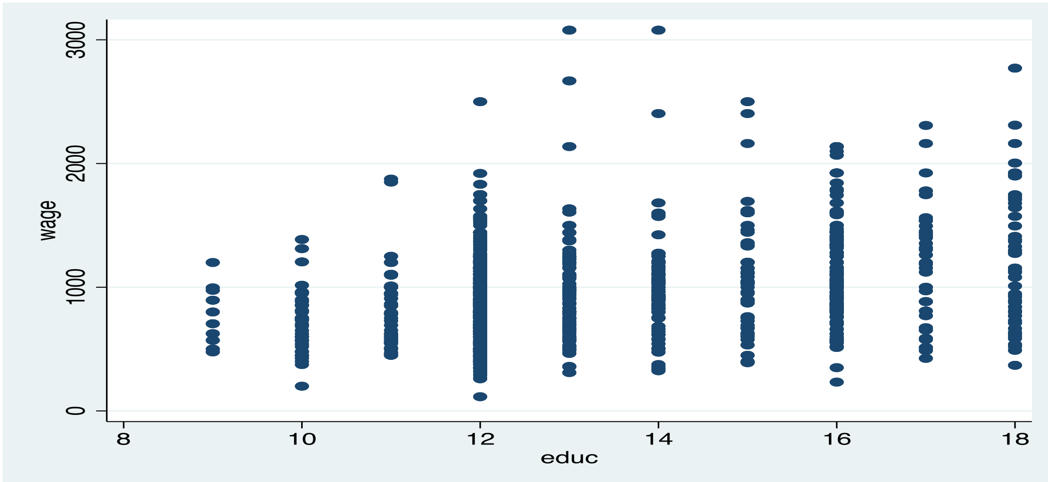
Variable	Obs	Mean	Std. Dev.	Min	Max
wage	935	957.9455	404.3608	115	3078
hours	935	43.92941	7.224256	20	80
IQ	935	101.2824	15.05264	50	145
KWW	935	35.74439	7.638788	12	56
educ	935	13.46845	2.196654	9	18
exper	935	11.56364	4.374586	1	23
tenure	935	7.234225	5.075206	0	22
age	935	33.08021	3.107803	28	38
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

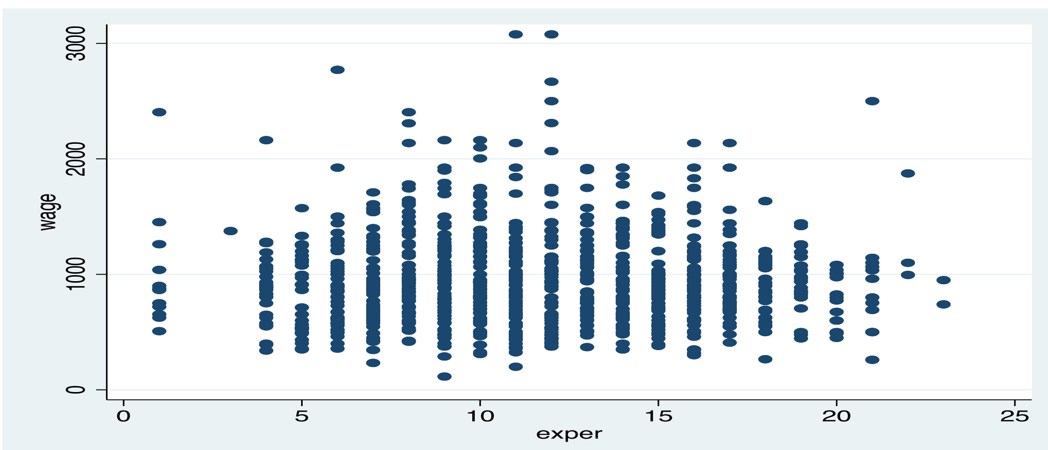
```
. twoway (scatter wage educ)
```



- 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

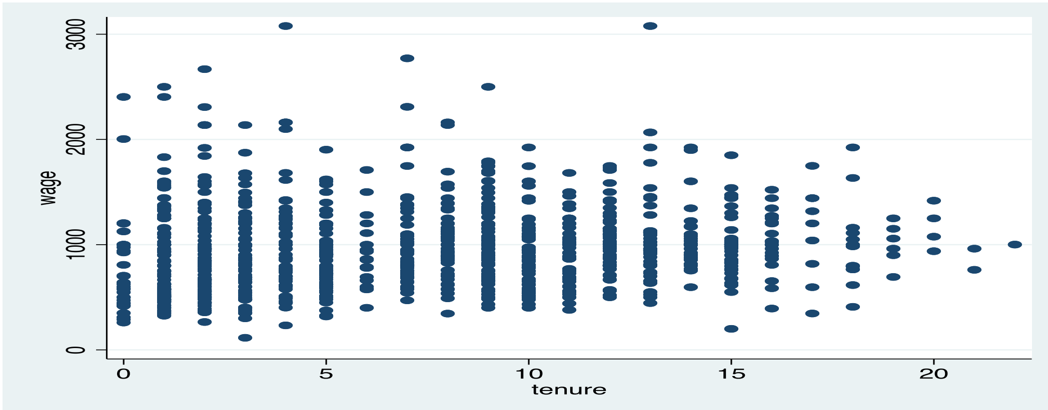
```
. twoway (scatter wage exper)
```



- 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
```

Following questions will not be graded, they are for you to practice and will be discussed at the recitation:

1. [Practice question, not graded] SW Problem 4.1

(a) The predicted average test score is

$$\widehat{Test\ Score} = 520.4 - 5.82\text{x}22 = 392.36$$

(b) The predicted decrease in the classroom average test score is

$$\Delta\widehat{TestScore} = (-5.82\text{x}19) - (-5.82\text{x}23) = 23.28$$

or the predicted change is

$$\Delta\widehat{TestScore} = (-5.82\text{x}23) - (-5.82\text{x}19) = -23.28$$

(c) Using the formula for $\hat{\beta}_0$, we know the sample average of the test scores across the 100 classroom is

$$\overline{TestScore} = \hat{\beta}_0 + \hat{\beta}_1\text{x}\overline{CS} = 520.4 - 5.82\text{x}21.4 = 395.85$$

(d) Use the formula for the standard error of the regression (SER) to get the sum of squared residuals:

$$SSR = (n-2)SER^2 = (100-2)\text{x}11.5^2 = 12961$$

Use the formula for R^2 to get the total sum of squares:

$$TSS = \frac{SSR}{1-R^2} = \frac{12961}{1-0.08} = 14088$$

The sample variance is $s_Y^2 = \frac{TSS}{n-1} = \frac{14088}{99} = 142.3$. Thus, the standard deviation is $s_Y = \sqrt{s_Y^2} = 11.9$

2. [Practice question, not graded] SW Problem 4.3

(a) The coefficient 9.6 shows the marginal effect of Age on AWE; that is, AWE is expected to increase by \$9.6 for each additional year of age. 696.7 is the intercept of the regression line. It determines the overall level of the line.

(b) SER is in the same units as the dependent variable (Y, or AWE in this example). Thus SER is measures in dollars per week.

(c) R^2 is unit free.

(d) (i) $696.7 + 9.6 \times 25 = \$936.7$;

(ii) $696.7 + 9.6 \times 45 = \$1{,}128.7$

(e) No. The oldest worker in the sample is 65 years old. 99 years is far outside the range of the sample data.

(f) No. The distribution of earning is positively skewed and has kurtosis larger than the normal.

(g) $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}$, so that $\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1\overline{X}$. Thus the sample mean of AWE is $696.7 + 9.6 \times 41.6 =$ \$1,096.06.

3. [Practice question, not graded] Let $KIDS$ denote the number of children born to a woman, and let $EDUC$ denote years of education for the woman. A simple model relating fertility to years of education is

$KIDS = a + b * EDUC + u,$

where u is the unobserved residual.

a) What kinds of factors are contained in u? Are these likely to be correlated with level of education?

Income, age, and family background (such as number of siblings) are just a few possibilities. It seems that each of these could be correlated with years of education.(Income and education are probably positively correlated; age and education may be negatively correlated because women in more recent cohorts have, on average, more education; and number of siblings and education are probably negatively correlated.)

b) Will simple regression of kids on *EDUC* uncover the ceteris paribus ('all else equal') effect of education on fertility? Explain.

Not if the factors we listed in part (i) are correlated with EDUC. Because we would like to hold these factors fixed, they are part of the error term. But if u is correlated with EDUC, then E(u|EDUC) is not zero, and thus OLS Assumption (A2) fails.

4. [Practice question, not graded] SW Problem 4.9

(a) With $\hat{\beta}_1 = 0$, $\hat{\beta}_0 = \overline{Y}$, and $\hat{Y}_i = \hat{\beta}_0 = \overline{Y}$. Thus ESS = 0 and $R^2 = 0$.

(b) If $R^2 = 0$, then ESS = 0, so that $\hat{Y}_i = \overline{Y}$ for all i. But $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, so that $\hat{Y}_i = \overline{Y}$ for all i, which implies that $\hat{\beta}_1 = 0$, or that X_i is constant for all i. If X_i is constant for all i, then $\sum_{i=1}^{n}(X_i - \overline{X})^2 = 0$ and $\hat{\beta}_1$ is undefined (see equation (4.7)).

5. [Practice question, not graded] SW Empirical Exercise 4.1

(a) $\widehat{AHE} = 1.08 + 0.60 \times Age$ Earnings increase, on average, by 0.60 dollars per hour when workers age by 1 year.

(b) Bob's predicted earnings = 1.08 + (0.60 × 26) = \$16.68 Alexis's predicted earnings = 1.08 + (0.60 × 30) = \$19.08

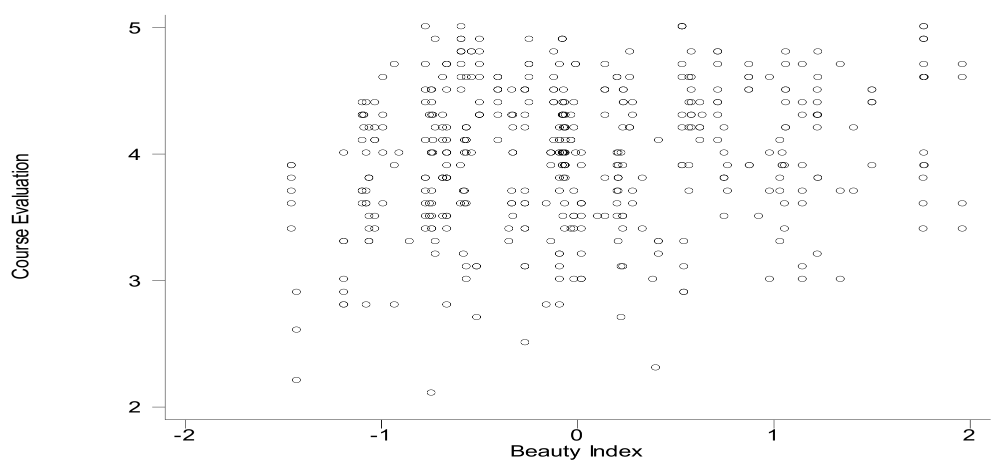
(c) The regression R^2 is 0.03.This means that age explains a small fraction of the variability in earnings across individuals.

.do file for q10:

```
use cps08.dta, clear // Load data and clear workspace
regress ahe age // Run regression to see effect of age on AHE
scalar b0 = _b[_cons] // save beta_0
scalar b1 = _b[age] // save beta_1
display b0+b1*26 // for 26 year-old
display b0+b1*30 // for 30 year-old
```

6. [Practice question, not graded] SW Empirical Exercises 4.2

(a)



There appears to be a weak positive relationship between course evaluation and the beauty index.

(b) $\widehat{Course\_Eval} = 4.00 + 0.133 \times Beauty$. The variable *Beauty* has a mean that is equal to 0; the estimated intercept is the mean of the dependent variable (*Course_Eval*) minus the estimated slope (0.133) times the mean of the regressor (*Beauty*). Thus, the estimated intercept is equal to the mean of *Course_Eval*.

(c) The standard deviation of *Beauty is* 0.789. Thus

Professor Watson's predicted course evaluations $= 4.00 + 0.133 \times 0 \times 0.789 = 4.00$

Professor Stock's predicted course evaluations $= 4.00 + 0.133 \times 1 \times 0.789 = 4.105$

(d) The standard deviation of course evaluations is 0.55 and the standard deviation of beauty is 0.789. A one standard deviation increase in beauty is expected to increase course evaluation by $0.133 \times 0.789 = 0.105$, or 1/5 of a standard deviation of course evaluations. The effect is small.

(e) The regression R^2 is 0.036, so that *Beauty* explains only 3.6% of the variance in course evaluations.

.do file for q 11:

```
use TeachingRatings, clear
scatter course_eval beauty
graph export beauty_effect.pdf, replace
reg course_eval beauty
scalar b0 = _b[_cons]
scalar b1 = _b[beauty]
sum beauty // The MACRO's `r(mean)' and `r(sd)' are loaded by this command
display b0+b1*`r(mean)' // Watson
display b0+b1*(`r(mean)'+`r(sd)') // Stock
sum course_eval
```