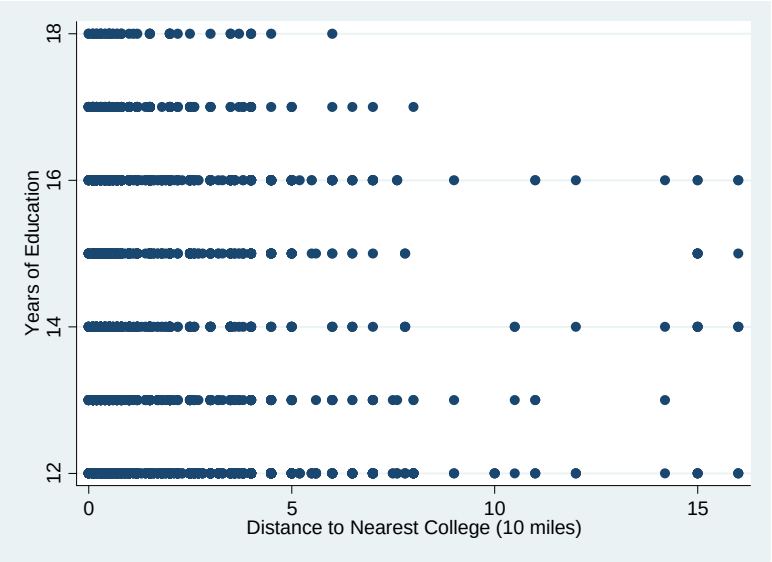
PUBPOL 639: ASSIGNMENT 2 - Solutions
Winter 2011

This assignment aims to get you comfortable with running and interpreting the output from bivariate regression. Be sure to copy your do and log files into the back of your solutions. Courier 8 or 9 pt font works well. Your log file will be long, so don't worry about cleaning/formatting it.

PROBLEM 1 – School Construction as Economic Development

Governor Snyder has decided to make extending access to higher education a central component of his plan to revitalize Michigan's economy. He's considering either providing subsidies for students from low-income families or allowing some rural community colleges to begin offering bachelor's degrees. Underlying this latter initiative is the belief that geographic accessibility is an important determinant of college attendance. He asks you to assess the importance of family income and distance to the nearest Bachelors-granting institution to educational attainment. In this exercise you will investigate the relationship between the completed schooling of adults, the distance of their childhood homes to the nearest college, and family income. You will examine this relationship using data on a random sample of 1982 high school seniors who were re-interviewed six years after high school to determine how many years of education they had completed. The data and it's documentation can be found on the CTools site (Resources > Data).

1. **Describe the data**
    a. Summary Statistics
        i. How does completed schooling vary by family income?
        Family income and completed schooling are positively related. Children from families with high income have more years of completed schooling.
        ii. How does completed schooling vary by whether a student's mother graduated college? What about their father?
        Completed schooling is also positively correlated with parental education. Children whose mother or father went to college complete more schooling themselves.
        iii. Describe the variation in distance to college. What is the mean, median, min, max, 25^{th} percentile, 75^{th} percentile?
        The average child in the sample lives 17.2 miles from a college. The 25^{th} and 75^{th} percentiles (also called the interquartile range) are 4 miles and 25 miles, respectively. The minimum is 0 miles and the maximum is 160 miles.
        iv. The distance data are top-coded. What is the top code value? What is the largest fraction of the observations that could be top-coded?
        The top code value is 160 miles. This means that any children living more than 160 miles from a college were assigned the value of 160. At most, 0.24% of the sample is top-coded.
    b. Graphical analysis
        i. Graph completed schooling against distance to nearest college using a scatterplot. Include the graph in your answers.



ii. What do you see? Does there appear to be a relationship between distance to college and years of completed education?
From the scatterplot above, it appears that years of schooling and distance to the nearest college are negatively correlated. That is, students living further from a college appear to have less schooling on average.

2. Regression analysis I

a. Regress completed schooling on distance to the nearest college

```
. reg ed dist, robust;

Linear regression                               Number of obs =    3796
                                                F(  1,  3794) =   29.83
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0074
                                                Root MSE      =  1.8074

------------------------------------------------------------------------------
             |               Robust
          ed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0733727   .0134334    -5.46   0.000    -.0997101   -.0470353
       _cons |   13.95586   .0378112   369.09   0.000     13.88172    14.02999
------------------------------------------------------------------------------
```

b. Write out the equation that corresponds to this regression.
I didn't specify population or sample regression, so you could have given either:

$education_i = \beta_0 + \beta_1 Distance_i + u_i$ (population)

OR

$education_i = \hat{\beta}_0 + \hat{\beta}_1 Distance_i + \hat{u}_i$ (sample)

c. Find and label the following on your Stata output and interpret these statistics.

i. $\hat{\beta}_1$ *= - 0.07. A ten mile increase in distance from the nearest college is associated with a 0.07 year decrease in years of completed schooling.*

ii. Standard error of $\hat{\beta}_1$: *SE(* $\hat{\beta}_1$*) = 0.013. This is the variation in our estimate of* $\hat{\beta}_1$ *due to sampling variability. If we drew a bunch of random samples and ran this regression on each of these samples, we would get a range of values for* $\hat{\beta}_1$*. This is our estimate of the standard deviation of these values of* $\hat{\beta}_1$*.*

iii. t-test for null hypotheses that $\beta_1 = 0$. *The t-statistic for the test of this null hypothesis is -5.46, which is greater (in absolute value) than the critical value of 1.96. Therefore we can reject the null hypotheses.*

iv. *confidence interval for* β_1 *: CI = [-0.010 to -0.047]. If we were to run this regression on a hundred different random samples, 95% of the time the estimate would fall in this range. We can be sure, with 95% certainty, that the true (population) value for* β_1 falls in this range.

v. $\hat{\beta}_0$ *= 13.96. This is the predicted years of schooling for someone who lives right next to a college (zero miles from one).*

vi. *Number of observations: N = 3796.*

vii. R^2 *= 0.0074. This means that 0.7% of the variation in educational attainment can be accounted for by variation in the distance to nearest college.*

d. Based on this regression, what is the predicted schooling of a person who lives ten miles from the nearest college? Fifty miles? Show your work.

$\widehat{education}_i = \hat{\beta}_0 + \hat{\beta}_1 Distance_i$

$= 13.96 - 0.07(Distance_i)$

$= 13.89\ for\ Distance = 1$

$= 13.61\ for\ Distance = 5$

(note: you may get slightly different numbers due to rounding error)

3. Regression analysis II

a. Regress completed schooling on the variable indicating whether the child comes from a high income family.

```
. reg ed incomehi, robust;

Linear regression                               Number of obs =    3796
                                                F(  1,  3794) =  179.82
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0470
                                                Root MSE      =  1.7711

------------------------------------------------------------------------------
             |               Robust
          ed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    incomehi |   .8695741   .0648458    13.41   0.000     .7424381    .9967101
       _cons |   13.58029   .0335697   404.54   0.000     13.51447     13.6461
------------------------------------------------------------------------------
```

b. Write out the equation that corresponds to this regression.

I didn't specify population or sample regression, so you could have given either:

$education_i = \beta_0 + \beta_1 HighIncome_i + u_i$ (population)

OR

$education_i = \hat{\beta}_0 + \hat{\beta}_1 HighIncome_i + \hat{u}_i$ (sample)

c. What is the "omitted" category in this regression?

The omitted category is children from low income families (assuming there are only two groups: high and low).

d. Find and label the following on your Stata output and interpret these statistics.
   i. $\hat{\beta}_1$ *= 0.87. Children from high income families have 0.87 more years of completed schooling than children from low income families..*
   ii. Standard error of $\hat{\beta}_1$: *SE(* $\hat{\beta}_1$ *) = 0.065. This is the variation in our estimate of* $\hat{\beta}_1$ *due to sampling variability. If we drew a bunch of random samples and ran this regression on each of these samples, we would get a range of values for* $\hat{\beta}_1$ *. This is our estimate of the standard deviation of these values of* $\hat{\beta}_1$ *.*
   iii. t-test for null hypotheses that $\beta_1 = 0$. *The t-statistic for the test of this null hypothesis is 13.41, which is greater (in absolute value) than the critical value of 1.96. Therefore we can reject the null hypotheses that children from low and high income families have the same years of completed schooling.*
   iv. *confidence interval for* β_1 *: CI = [0.742 to 0.997]. If we were to run this regression on a hundred different random samples, 95% of the time the estimate would fall in this range. We can be sure, with 95% certainty, that the true (population) value for* β_1 falls in this range.
   v. $\hat{\beta}_0$ *= 13.58. This is the average years of schooling for children from low income families.*
   vi. *Number of observations: N = 3796.*
   vii. R^2 *= 0.047. This means that 4.7% of the variation in educational attainment can be accounted for by variation in family income category.*

e. Based on this regression, what is the predicted schooling of a person who comes from a high income family? What about one that comes from a low income family? Show your work.

$\widehat{education}_i = \hat{\beta}_0 + \hat{\beta}_1 HighIncome_i$

For High Income:

$\widehat{education}_i = 13.58 + (0.87)(1)$

=14.45

For Low Income:

$\widehat{education}_i = 13.58 + (0.87)(0)$

=13.58

4. Model fit

Which factor – distance to college or family income – explains more of the variation in educational attainment in this data?

Family income explains more of the variation in educational attainment in this data. You can see this by comparing the R-squared from the two regressions.

PROBLEM 2 – Bed nets

The purpose of this exercise is to help you learn the mechanics of ordinary least squares (OLS) regression. First you will calculate the regression "by hand" then you will use Stata to confirm the calculation.

The table below contains data on the fraction of children enrolled in primary school and the fraction of children under five sleeping under insecticide-treated bed nets for five countries in 2007. Bed nets have been shown to be the most cost-effective prevention method against malaria and are part of UN's Millennium Development Goals (which is where this data comes from).

	Primary school enrollment rate	Fraction of children sleeping under insecticide-treated bed nets
Ethiopia	72.3	33.1
Indonesia	98	3.3
Namibia	88.1	10.5
Swaziland	87.2	0.6
Zambia	95.4	28.6

1. Using the appropriate formulas, show how to calculate each of the following. Note: "show how to calculate" means (1) write the appropriate formula; (2) plug in the appropriate values; and (3) show the computed answer. You do not need to show the intermediate calculations between steps 2 and 3. You may use Excel to do the calculations as long as you show the formulas used and you do not use the built-in regression function.

	(1)	(2)	(3)	(4)	(5)	(6)		(7)	(8)
	S	**S-Sbar**	**B**	**B-Bbar**	**(S-Sbar)*(B-Bbar)**	**(B-Bbar)^2**		**Shat**	**uhat**
Ethiopia	72.3	-15.9	33.1	17.88	-284.292	319.6944		82.293	-9.993
Indonesia	98	9.8	3.3	-11.92	-116.816	142.0864		92.138	5.862
Namibia	88.1	-0.1	10.5	-4.72	0.472	22.2784		89.759	-1.659
Swaziland	87.2	-1	0.6	-14.62	14.62	213.7444		93.030	-5.830
Zambia	95.4	7.2	28.6	13.38	96.336	179.0244		83.780	11.620
Average	88.2		15.22						
Sum					-289.68	876.828			
Beta1hat =	Sum[(S-Sbar)*(B-Bbar)] / Sum[(B-Bbar)^2]				=	-0.3304			
Beta0hat =	Sbar - Beta1hat*Bbar		=88.2 - (-.3304)*(15.22)		=	93.228			

a. $\hat{\beta}_1$ - the estimated slope coefficient from the regression of enrollment rate on fraction sleeping under bed nets

$$\hat{\beta}_1 = \frac{\sum(B-\bar{B})(S-\bar{S})}{\sum(B-\bar{B})^2}$$

$$where\ \bar{B} = \frac{\sum bednet}{5} = 15.22, \bar{S} = \frac{\sum enrollrate}{5} = 88.22$$

$\hat{\beta}_1 = -0.3304$

b. $\hat{\beta}_0$ - the estimated intercept from the same regression

$$\hat{\beta}_0 = \bar{S} - \hat{\beta}_1 \bar{B} = (88.22) - (-.3304) * (15.22) = 93.228$$

c. $\hat{Y}_i$ - the predicted values for the five countries

$\hat{S}_i = \hat{\beta}_0 + \hat{\beta}_1 B_i$

See table above for answers.

d. $\hat{u}_i$ - the OLS residuals for the five countries

$\hat{u}_i = S_i - \hat{S}_i$

See table above for answers.

2. In a concise paragraph drawing on the numbers you calculated above, describe the relationship between bed net utilization and the primary school enrollment rate as precisely as you can. Indicate the direction and magnitude of the relationship based on this sample of five countries. *In this sample of five countries, there is a negative relationship between the fraction of children enrolled in primary school and the fraction of children under five sleeping under insecticide-treated bed nets. According to our OLS estimates, a one percentage point increase in the fraction sleeping under bed nets (a seven percent increase at the average of 15%) is associated with a 0.33 percentage point reduction in the fraction of children enrolled in primary school (a 0.4 % decrease at the average of 88%).*

3. Now you will see how the same regression is produced by Stata. Open Stata and type "edit," which brings up something that looks like a spreadsheet. Enter the country name, enrollment, bed net entries in the first three columns. Double-click the column headers to enter variable names ("country", "enrollrate", "bednet"). Close the editor window when you are done. Type "list" to be sure you have typed in the numbers correctly, and type "sum" to inspect the variable means. Regress the enrollment rate against the bed net utilization rate. On the Stata output, find and label $\hat{\beta}_0$ and $\hat{\beta}_1$.

```
. reg enrollrate bednets, robust

Linear regression                               Number of obs =         5
                                                F(  1,     3) =      0.74
                                                Prob > F      =    0.4537
                                                R-squared     =    0.2382
                                                Root MSE      =    10.099

------------------------------------------------------------------------------
             |               Robust
  enrollrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     bednets |  -.3303725   .3848078    -0.86   0.454    -1.555003    .8942578
       _cons |   93.22827   4.875774    19.12   0.000     77.71138    108.7452
------------------------------------------------------------------------------
```

4. Depending on the sign of the association you find, answer the appropriate question below (i.e. only answer a or b):

   a. If you found a positive association, should we interpret your finding as evidence that bed nets improved primary school enrollment? Why or why not.

b. If instead you found a negative association, should we interpret your findings as evidence that bed nets decrease primary school enrollment (or at least don't help)? Why or why not.

We found a negative association between bed nets and primary school enrollment, but this association is unlikely to be causal. There may be other differences between the five countries that influence both schooling and the use of bed nets, such as poverty or geography. Also, it is likely that bed nets have been deployed in areas hardest-hit by Malaria, which reduces educational attainment directly. Malaria thus influences both bed net utilization and school enrollment, inducing a spurious negative correlation between these two variables. A separate issue is that the relationship we found was statistically insignificant (not surprising given our five data points). But even if the relationship was statistically significant, the concern about the omitted variable bias caused by Malaria would still prevent us from interpreting the relationship as causal.

Do file

```
clear all
#delimit ;
capture log close;
set mem 100m;
set more off;

log using assignment2.log, text replace;

* ================================================================;
* Public Policy 639 Assignment 2 Solutions
* ================================================================;
* ================================================================;
* Problem 1 - School Construction and Economic Development
* ================================================================;
use assignment2.dta, clear;

tab incomehi, sum(ed);

tab dadcoll, sum(ed);

tab momcoll, sum(ed);

sum dist, detail;

gen topcode = 1 if dist==16;
replace topcode = 0 if dist != 16;
tab topcode;

twoway scatter ed dist,
xtitle("Distance to Nearest College (10 miles)")
ytitle("Years of Education");
graph export assign2_1.wmf, replace;

reg ed dist, robust;

reg ed incomehi, robust;

* ================================================================;
* Problem 2 - Bednets
* ================================================================;
use bednet, clear;
desc;
```

```
list;
reg enrollrate bednets, robust;

log close;
```

Log file

```
       log:  assignment2.log
  log type:  text
. * ================================================================;
. * Public Policy 639 Assignment 2 Solutions
> * ================================================================;
. * ================================================================;
. * Problem 1 - School Construction and Economic Development
> * ================================================================;
. use assignment2.dta, clear;

. tab incomehi, sum(ed);

            |            Summary of ed
   incomehi |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          0 |   13.580288   1.7471001        2709
          1 |   14.449862    1.829525        1087
------------+------------------------------------
      Total |   13.829294   1.8139688        3796

. tab dadcoll, sum(ed);

            |            Summary of ed
    dadcoll |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          0 |   13.561902   1.7384877        3029
          1 |   14.885267   1.7191532         767
------------+------------------------------------
      Total |   13.829294   1.8139688        3796

. tab momcoll, sum(ed);

            |            Summary of ed
    momcoll |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          0 |   13.659014   1.7698348        3267
          1 |   14.880907   1.7284918         529
------------+------------------------------------
      Total |   13.829294   1.8139688        3796

. sum dist, detail;

                           dist
-------------------------------------------------------------
      Percentiles      Smallest
 1%           0              0
 5%          .1              0
10%          .1              0       Obs                3796
25%          .4              0       Sum of Wgt.        3796

50%           1                      Mean           1.724921
                        Largest       Std. Dev.      2.133836
75%         2.5             16
90%           4             16       Variance       4.553255
95%         5.2             16       Skewness       2.904585
99%          11             16       Kurtosis       15.48146

. gen topcode = 1 if dist==16;
(3787 missing values generated)
```

```
. replace topcode = 0 if dist != 16;
(3787 real changes made)

. tab topcode;

    topcode |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      3,787       99.76       99.76
          1 |          9        0.24      100.00
------------+-----------------------------------
      Total |      3,796      100.00

. twoway scatter ed dist,
> xtitle("Distance to Nearest College (10 miles)")
> ytitle("Years of Education");

. graph export assign2_1.wmf, replace;

. reg ed dist, robust;

Linear regression                                      Number of obs =    3796
                                                       F(  1,  3794) =   29.83
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0074
                                                       Root MSE      =  1.8074

------------------------------------------------------------------------------
             |               Robust
          ed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dist |  -.0733727   .0134334    -5.46   0.000    -.0997101   -.0470353
       _cons |   13.95586   .0378112   369.09   0.000     13.88172    14.02999
------------------------------------------------------------------------------

. reg ed incomehi, robust;

Linear regression                                      Number of obs =    3796
                                                       F(  1,  3794) =  179.82
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0470
                                                       Root MSE      =  1.7711

------------------------------------------------------------------------------
             |               Robust
          ed |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    incomehi |   .8695741   .0648458    13.41   0.000     .7424381    .9967101
       _cons |   13.58029   .0335697   404.54   0.000     13.51447     13.6461
------------------------------------------------------------------------------

. * ====================================================================;
. * Problem 2 - Bednets
> * ====================================================================;
. use bednet, clear;

. desc;

Contains data from bednet.dta
  obs:             5
 vars:             3                          21 Feb 2010 16:41
 size:           100 (99.9% of memory free)
------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
enrollrate      float  %8.0g
```

```
bednets         float  %8.0g
country         str8   %9s
-------------------------------------------------------------------------------
Sorted by:

. list;

     +------------------------------+
     | enroll~e   bednets    country |
     |------------------------------|
  1. |     72.3      33.1   Ethiopia |
  2. |       98       3.3        Ind |
  3. |     88.1      10.5        Nam |
  4. |     87.2        .6        Swa |
  5. |     95.4      28.6        Zam |
     +------------------------------+

. reg enrollrate bednets, robust;

Linear regression                                      Number of obs =       5
                                                       F(  1,     3) =    0.74
                                                       Prob > F      =  0.4537
                                                       R-squared     =  0.2382
                                                       Root MSE      =  10.099

------------------------------------------------------------------------------
             |               Robust
  enrollrate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     bednets |  -.3303725   .3848078    -0.86   0.454    -1.555003    .8942578
       _cons |   93.22827   4.875774    19.12   0.000     77.71138    108.7452
------------------------------------------------------------------------------

. log close;
      name:  <unnamed>
       log:  assignment2.log
  log type:  text
-------------------------------------------------------------------------------
```