Department of Economics
Winter 2008

Economics 120B
Prof. Berman

Problem Set #2
Due Tuesday, February 5
Please hand in answers on this sheet and staple the output (.log) file to it.

1. Hypothesis Testing: The file cps06.dta contains information about wages and education for 82,228 observations from the Current Population Survey of 2006. It is available on the course website. In those data $\mu_y = E(Y) = 2.786$. Linear regression describes the relationship between log wages (y) and years of education (x) where the intercept of the regression line is β_0, the slope is β_1 and $\mu_y = E(Y) = 2.786$.

$$y = \beta_0 + \beta_1 x + \epsilon,$$
$$\text{with} \quad \beta_0 = 1.376\ , \ \beta_1 = 0.1030\ , \ Cov(x, \epsilon) = 0.$$

Treat these data as a population.
a) Use Stata to reproduce these three population "parameters." (Attach the output.)

b) Generate a sample of 40 observations from the population as in the Stata log file attached below. We are interested in the sampling variance of the least squares estimates of β_0 and β_1.

(In Stata, generating a random sample using the "bsample" command requires setting a "seed" value. Choose the seed to be some arbitrary large positive, odd number. <u>Don't</u> use the same number as any of your classmates. Identical seed values will be interpreted in the worst possible way and marks will be deducted.)

Calculate OLS estimates of μ_y β_0 and β_1 using your 40 observation sample. Report your sample estimates here.

c) Now pretend that you don't know anything about the population except for the information in the sample. Test the null hypothesis that $\beta_1 = 0.1030$ using the data you have in your 40 observation sample, using a two-tailed test and $\alpha = 0.05$.

Did you reject the null hypothesis? Yes / No

What was the probability of that happening?

d) Assuming that 100 of your classmates draw their own independent random samples and answer question (b) correctly. What's the probability that all 100 of them reject/accept as you did?

e) Did you use a normal distribution in your test in (c). Explain.

How can you justify using a normal distribution when the distributions of y and x are not normal?

e) Are these data experimental? Yes / No

Why (not)?

2. Least Squares.

You have a sample of N observations Y_1, Y_2, ... Y_N

You are interested in finding a number A which has the smallest average distance from the observations, where the measure of distance is the “error” term $e_i = (Y_i - A)$.

What’s the formula for the (minimand)A which minimizes the average of $(Y_i - A)^2$ over N observations?

Prove your claim.

```
. log using cps_example
-------------------------------------------------------------------------------
       log:  C:\work\120B\cps_example.smcl
  log type:  smcl
 opened on:  22 Jan 2007, 20:25:11

. * Example progam which treats the CPS from 2006 as a population
. use cps06

. desc

Contains data from cps06.dta
  obs:        82,228
 vars:             7                        22 Jan 2007 19:51
 size:     1,151,192 (94.5% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
age             byte   %19.0g      agelbl     Age
educ            byte   %38.0g      educ99lbl
                                              Educational attainment, 1990
fullpart        byte   %9.0g       fullpartlbl
                                              Worked full or part time last
                                                year
black           byte   %9.0g
asian           byte   %9.0g
hwage1          float  %9.0g                  annual earnings/annual hours
gender          byte   %9.0g                  female==1
-------------------------------------------------------------------------------
Sorted by:

. summ

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |     82228    43.38813    11.19701         25         85
        educ |     82228    13.68711    2.818516          0         21
    fullpart |     82228    1.130381    .3367246          1          2
       black |     82228    .1036995    .3048721          0          1
       asian |     82228    .0469426    .2115173          0          1
-------------+--------------------------------------------------------
      hwage1 |     82228    21.51597    28.03584   .0003698   2777.778
      gender |     82228     .477684    .4995048          0          1
```

```
. * create a new variable - the logarithm of hourly wages:
. generate lhwage=log(hwage1)

. summ lhwage

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      lhwage |     82228    2.786399    .7449476  -7.902487   7.929407



. * Calculate a simple linear regression of log hourly wage on education
. regress lhwage educ, robust

Regression with robust standard errors                 Number of obs =   82228
                                                       F(  1, 82226) =12512.02
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1520
                                                       Root MSE      =    .686

------------------------------------------------------------------------------
             |               Robust
      lhwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .103048   .0009212   111.86   0.000     .1012423    .1048536
       _cons |    1.37597   .0126578   108.71   0.000     1.351161    1.400779
------------------------------------------------------------------------------

. * So each year of education predicts an hourly wage increase of about 10.3% in
>  2006.
.
. * Now treat the full CPS as a population and draw a sample from it.
. * i.e.,  y = beta_0 + beta_1 x +  epsilon
. * we will sample from that population and estimate the population parameters b
> eta_0=1.376 and beta_1=0.103
. set seed 098709870198768761

. bsample 50
. summ

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |        50       42.06    10.66447         25         67
        educ |        50        13.7    2.358225          8         18
    fullpart |        50        1.08    .2740475          1          2
       black |        50         .12    .3282607          0          1
       asian |        50         .08    .2740475          0          1
-------------+--------------------------------------------------------
      hwage1 |        50    19.59099    12.06503      3.125       62.5
      gender |        50         .54    .5034574          0          1
      lhwage |        50    2.822957    .5557725   1.139434   4.135167
```

```
. * Note that the bsample command threw out all the data except for 50 randomly
> chosen observations
. * With those 50 we can estimate parameters beta_0 and beta_1
.
. regress lhwage educ, robust

Regression with robust standard errors                 Number of obs =      50
                                                       F(  1,    48) =   11.43
                                                       Prob > F      =  0.0014
                                                       R-squared     =  0.2129
                                                       Root MSE      =  .49818

------------------------------------------------------------------------------
             |               Robust
      lhwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1087445   .0321676     3.38   0.001     .0440672    .1734218
       _cons |   1.333158   .4618206     2.89   0.006     .4046051     2.26171
------------------------------------------------------------------------------

. * Our estimates in this 50 obs. sample are 1.33 for beta_0 and 0.11 for beta_1 .
.
. * You can try this on your own to check that every seed value gives a differen
> t sample, and different estimates.
```