Economics 113 Fall 2014
Professor Spearot

Homework #1 (Solution Key)

1. **Using the Wage data from the course website, and the computer program Stata, summarize the variables *wage*, *education*, *iq*, and *age*. Report the mean, median, max, min, and standard deviation.**

. summarize wage educ iq age

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        wage |       935    957.9455    404.3608        115       3078
        educ |       935    13.46845    2.196654          9         18
          iq |       935    101.2824    15.05264         50        145
         age |       935    33.08021    3.107803         28         38
```

The above command and table report the mean, max, min and standard deviation of the variables of interest. Median is not reported, so we need to use another command.

. summarize wage educ iq age, detail

Note: I trimmed and rearranged the output table of Stata. The original table has much more information but it is not relevant for us.

	wage	educ	IQ	age
50% Percentile (or Median)	905	12	102	33

2. **Please regress *wage* on *education* and *iq*. Interpret the coefficient on *education*. Is it statistically different from zero? Please use a two---sided test, and a 95% level of confidence.**

. reg wage educ iq

```
      Source |       SS       df       MS              Number of obs =     935
-------------+------------------------------           F(  2,   932) =   72.02
       Model |  20441576.8     2  10220788.4           Prob > F      =  0.0000
    Residual |   132274591   932  141925.527           R-squared     =  0.1339
-------------+------------------------------           Adj R-squared =  0.1320
       Total |   152716168   934  163507.675           Root MSE      =  376.73

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   42.05762   6.549836     6.42   0.000     29.20348    54.91175
          iq |   5.137958   .9558274     5.38   0.000     3.262134    7.013781
       _cons |  -128.8899   92.18232    -1.40   0.162    -309.7988    52.01908
------------------------------------------------------------------------------
```

The coefficient on education is written in red, $\widehat{\beta_{educ}} = 42.05762$.

Holding *iq* constant, one additional year of education is expected to increase the monthly wage by $42.05672.

Now we want to know if β_{educ} statistically different from zero. We want to test if we can reject the following null hypothesis in favor of the alternative hypothesis.

The null hypothesis is H_0: $\beta_{educ}=0$

The alternative hypothesis is H_A: $\beta_{educ} \neq 0$

The t-value for this test is always given in regression outputs. The specific one we are looking for is for the variable *educ* and is highlighted on green in the table above (t-value=6.42).

The probability of being wrong if we reject the null hypothesis is also given above. It is highlighted in blue. This probability is almost equal to 0. Note that the p-value given in the stata output corresponds to two-sided test, not one-sided. Thus, at a 95% level of confidence, we reject the null hypothesis. In other words, β_{educ} is significantly different from zero at 95% level of confidence. Or we can say, education has a significant effect on wage.

Note that we could have rejected the null hypothesis even at 99% level of confidence since the probability of being wrong is still less 1%.

3. Please remove *iq* from the regression in problem 2, and re---run the regression. Comment on what happens to the coefficient on *educ*. This is an example of what problem? Why?

. reg 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
```

I wrote the new coefficient of education in red ($\widehat{\beta_{educ}} = 60.21428$). We observe that this coefficient increased when we remove the variable *iq* from the regression.

This could be an example of positive bias in our estimate $\widehat{\beta_{educ}}$ when we omit a variable. If the variable *iq* is correlated with both the independent (educ) and dependent variable (wage), then the conditional mean zero assumption (E(u/x)=0) is violated and we have a bias in our estimated coefficient.

We can check the direction (positive or negative) of the bias with our intuition. It would be reasonable to assume that *iq* is positively correlated with both education and wages (unless you go to academia). That would imply a positive bias for the coefficient in front of education when we omit *iq* in the regression. This is indeed in line with our regression results as $\widehat{\beta_{educ}} = 60.21428$ in the regression without *iq* is greater than $\widehat{\beta_{educ}} = 42.05762$ in the regression including *iq*.

4. Now using the wage dataset again, regress *wage* on *educ*, *iq*, and *age*. Suppose that I conclude that *age* has a significant impact on *wage*. What is the probability that I'm wrong?

. reg wage educ iq age

```
      Source |       SS       df       MS              Number of obs =     935
-------------+------------------------------           F(  3,   931) =   60.03
       Model |  24753918.5     3  8251306.18           Prob > F      =  0.0000
    Residual |   127962250   931  137446.025           R-squared     =  0.1621
-------------+------------------------------           Adj R-squared =  0.1594
       Total |   152716168   934  163507.675           Root MSE      =  370.74

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |      41.623    6.44611     6.46   0.000     28.97241    54.27359
          iq |    5.368319    .941521     5.70   0.000      3.52057    7.216069
         age |    21.88653   3.907387     5.60   0.000     14.21822    29.55484
       _cons |    -870.379    160.478    -5.42   0.000     -1185.32   -555.4385
```

Concluding that age has a significant impact on wage is the same thing as rejecting that the coefficient in front of age ($\widehat{\beta_{age}}$) is equal to zero.

If we reject the following null hypothesis in favor of the alternative hypothesis with some level of confidence,

H_0: $\beta_{age} = 0$ (Age does not have a significant impact on wage)

H_A: $\beta_{age} \neq 0$ (Age has a significant impact on wage)

Then we can say age has a significant effect on wage with some confidence. But we can never be 100% sure of our statement/conclusion because there is always a probability we wrongfully reject the null hypothesis. The probability that we are wrong to reject the null hypothesis (p-value) is highlighted in

blue. The probability that we are wrong is so small that even three decimal points is not enough to differentiate it from zero. Bear in mind that the probability that we are wrong can never be exactly zero.