PROBLEM SET ONE
SOLUTIONS

Antonio Pacifico

Federica Romei

15/03/2012

1. Test the null hypothesis that birth weight in ounces is equal if the babies are white and non white against the alternative hypothesis that white babies weight more using a significance level of 5% and 10%. Comment.

SOLUTION

→ Open dataset typing

use bwght.dta, clear

→ Now, writing

mean bwght if white==1

and

mean bwght if white==0

you should be able to visualize these two output on STATA :

```
mean  bwght if  white==1

Mean estimation                     Number of obs    =    1089

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       bwght |   120.0533   .6024968      118.8711    121.2354
--------------------------------------------------------------
```

```
mean  bwght if  white==0

Mean estimation                         Number of obs    =     299

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
       bwght |   113.7692   1.232045      111.3446    116.1938
--------------------------------------------------------------
```

→ You have to test the following hypothesis:

$$H_0 : \Delta\left(\overline{bwght}\right) = \overline{bwght_W} - \overline{bwght_{NW}} = 0$$
$$H_1 : \Delta\left(\overline{bwght}\right) = \overline{bwght_W} - \overline{bwght_{NW}} > 0$$

To perform this hypothesis test you have to compute the t-statistic. By CLT, when sample size is large, the t-statistic is well approximated by the standard normal distribution; thus :

$$t_{value} = \Delta_{\widehat{bwght}} = \frac{\left(\overline{bwght_W} - \overline{bwght_{NW}}\right) - d_0}{\sqrt{SE_W^2 + SE_{NW}^2}} \xrightarrow{d} N(0,1)$$

In STATA you can compute it typing

display (120.05-113.76)/sqrt(0.6024^2+1.2320^2)

hence you should obtain that :

$$t_{value} = \Delta_{\widehat{bwght}} = 4.58$$

→ You are testing the Null hypothesis against the one sided Alternative hypothesis at 5% and 10% between two population means. It means that:

- You would not reject the Null hypothesis if $t_{value} < t_\alpha$
- You would accept the Alternative hypothesis if $t_{value} \geq t_\alpha$

In the first case, with a significance level of 5%, you have that :

$$4.58 > 1.64$$

hence you can reject the Null hypothesis that white babies weight as non white babies.

Instead, considering a significance level of 10%, you have that :

$$4.58 > 1.28$$

then you accept the Alternative hypothesis that white babies weight more.

2. **Compute the mean of the birth weight in pounds if the mother has less than 12 years of education and test the hypothesis:**

$$\boldsymbol{H_0} : \boldsymbol{birth\ weight} = \mathbf{7}$$
$$\boldsymbol{H_1} : \boldsymbol{birth\ weight} > \mathbf{7}$$

at 5% and 10% significant level. Compute the p-value (you should be able to do it without table). Comment.

<u>**SOLUTION**</u>

→ Writing

mean bwghtlbs if mother < 12

you obtain the output :

```
mean  bwghtlbs if  motheduc<12

Mean estimation                    Number of obs    =     222

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    bwghtlbs |   7.133164   .0812032      6.973133    7.293196
--------------------------------------------------------------
```

→ The t-statistic performing this new hypothesis test is :

$$t_{value} = \Delta_{\widehat{bwghtlbs}} = \frac{\left(\overline{bwghtlbs}\right) - d_0}{\sqrt{SE^2_{\overline{bwghtlbs}}}} \xrightarrow{d} N(0,1)$$

In STATA :

display (7.13-7)/sqrt(0.081^2)

hence

$$t_{value} = \Delta_{\widehat{bwghtlbs}} = 1.64$$

→ Testing the Null hypothesis against the one sided Alternative hypothesis at 5%, you have

$$1.64 = 1.64$$

hence you reject the Null hypothesis; in fact the sample mean (*7.13*) is bigger than hypothesized value d_0 (*7*).

Considering a significance level of 10%, you have

$$1.64 > 1.28$$

hence you cannot reject the Alternative hypothesis.

→ The p-value, also called calculated probability, is the probability of rejecting the Null hypothesis of a statistical study assuming the Null hypothesis is correct (with a value ranging from zero to one).

Often H_0 is rejected when p-value is less than the significance level. However, if p-value is small, it is reasonable to assume that the Null hypothesis is not true.

In this case, the computed value belongs to set lying on right tail of distribution; thus the p-value is less than α and you would reject the Null hypothesis.

3. **Test at 5% and 10% significant level if male betrays on average as female against the alternative Hypothesis that male betrays more. Comment.**

<u>SOLUTION</u>

→ Open dataset typing

use affairs.dta, clear

→ You have to test the following hypothesis :

$$H_0 : \Delta\left(\overline{betrayal}\right) = \overline{betrayal_M} - \overline{betrayal_F} = 0$$
$$H_1 : \Delta\left(\overline{betrayal}\right) = \overline{betrayal_M} - \overline{betrayal_F} > 0$$

→ Writing

mean affairs if gender==2

mean affairs if gender==1

you can visualize these two output on STATA :

```
mean  affairs if  gender==2

Mean estimation                          Number of obs    =     286

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
     affairs |   1.496503   .1946877      1.113295    1.879712
--------------------------------------------------------------
```

```
mean affairs if  gender==1

Mean estimation                          Number of obs    =     315

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
     affairs |   1.419048   .1864559      1.052187    1.785909
--------------------------------------------------------------
```

→ The t-statistic performing this hypothesis test is :

$$t_{value} = \Delta_{\widehat{betrayal}} = \frac{\left(\overline{betrayal_M} - \overline{betrayal_F}\right) - d_0}{\sqrt{SE_M^2 + SE_F^2}} \xrightarrow{d} N(0,1)$$

In STATA :

display (1.4965-1.4190)/sqrt(0.1947^2+0.1865^2)

hence

$$t_{value} = \Delta_{\widehat{betrayal}} = 0.3$$

→ Testing the Null hypothesis against the one sided Alternative Hypothesis at 5%, you have:

$$0.3 < 1.64$$

therefore you reject the Alternative hypothesis in which the male betrays more.

If you consider a significance level of 10%, you have:

$$0.3 < 1.28$$

thus you would accept the Null hypothesis.

4. Test at 5% and 10% if people with high level of education (above the mean) betrays as people with low level of education (below the mean) against the hypothesis that high education people betrays differently. Comment.

<u>SOLUTION</u>

→ You have to test the following hypothesis :

$$H_0 : \Delta\left(\overline{betrayal}\right) = \overline{betrayal_H} - \overline{betrayal_L} = 0$$
$$H_1 : \Delta\left(\overline{betrayal}\right) = \overline{betrayal_H} - \overline{betrayal_L} \neq 0$$

→ To compute level of education you have to consider its mean :

summ education

You obtain the following output :

```
summ  education

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   education |        601    16.16639    2.402555          9         20
```

Then you can write

mean affairs if education > 16.16

mean affairs if education < 16.16

hence you should be able to visualize :

```
mean  affairs if  education >16.16639

Mean estimation                    Number of obs    =     281

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
     affairs |   1.758007   .2088395      1.346912    2.169102
--------------------------------------------------------------
```

```
mean  affairs if  education <16.16639

Mean estimation                    Number of obs    =     320

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
     affairs |   1.190625   .1728363       .850582    1.530668
--------------------------------------------------------------
```

→ The t-statistic performing the hypothesis test is :

$$t_{value} = \Delta_{\widehat{betrayal}} = \frac{\left(\overline{betrayal_H} - \overline{betraya_L}\right) - d_0}{\sqrt{SE_H^2 + SE_L^2}} \xrightarrow{d} N(0,1)$$

In STATA :

display (1.758-1.1906)/sqrt(0.2088^2+0.1728^2)

hence

$$t_{value} = \Delta_{\widehat{betrayal}} = 2.13$$

→ You are testing the Null hypothesis against the two sided Alternative hypothesis at 5% and 10% between two population means. It means that:

- You would not reject the Null hypothesis if $|t_{value}| < +t\alpha/_2$
- You would accept the Alternative hypothesis if $|t_{value}| \geq +t\alpha/_2$

In the first case, with a significance level of 5%, you have that :

$$|2.13| > +1.96$$

hence you reject the Null hypothesis that high educated people betrays as low educated people.

Instead, considering a significance level of 10%, you have that :

$$|2.13| > +1.64$$

then you accept the Alternative hypothesis that high educated people betrays more. Why? Have you a plausible explanation?

5.

A. **Regress *packs variable* on *price variable* and explain the coefficients. Does the regression provide a statistically significant evidence that *price* affects *packs*? Do you think there is a correlation? Can you see if there is correlation? Comment.**

B. **Suppose you are working in an Health Department and you are interested in the consumption of cigarettes. Your boss want to know if the price have some effect in the reduction of smoked cigarettes. By previous point, do you advice to your boss to increase the price of packs? Comment.**

C. **Make a plot of *packs* against *price*. Can you prove there is correlation? Comment.**

SOLUTION

A.

→ If you type

reg packs price, r

you should be able to visualize :

```
reg  packs price, r

Linear regression                                      Number of obs =      46
                                                       F(  1,    44) =   15.30
                                                       Prob > F      =  0.0003
                                                       R-squared     =  0.3031
                                                       Root MSE      =  21.563

------------------------------------------------------------------------------
             |               Robust
       packs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |  -132.8751    33.9734    -3.91   0.000    -201.3439   -64.40617
       _cons |   293.5823   43.13554     6.81   0.000     206.6484    380.5163
------------------------------------------------------------------------------
```

A unitary increase in price decreases of 132 the smoked packs in a year. Indeed if you type

summ packs

you obtain the following output :

```
summ  packs

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       packs |         46    129.7996    25.54154   82.15354    216.818
```

hence, you would see that the sample average of packs is 129.

This means that if you increase the price of 1 dollar we will solve the problem of cigarettes.

In this case β_0 is difficult to interpret, you can say that if it should be possible to have cigarettes for free the people will smoke 293 packs per year.

B.

→ The t_{value} proves that both β_0 and β_1 are significant at 5% level.

The regression R^2 is quite high if you consider that there is only one regressor.

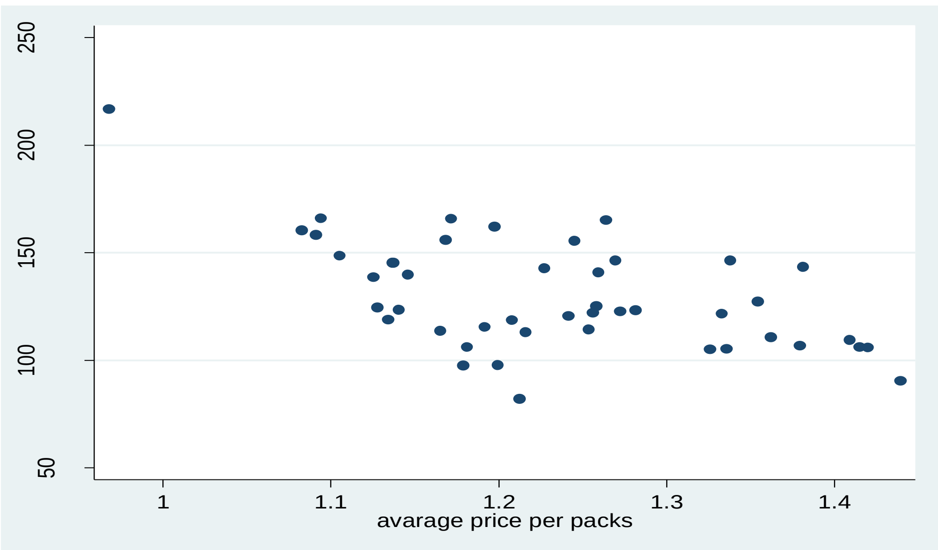
You would say to boss to increase the pack price even if you have to alert him/her about the problem of this regression.

C.

→ Writing

scatter packs price

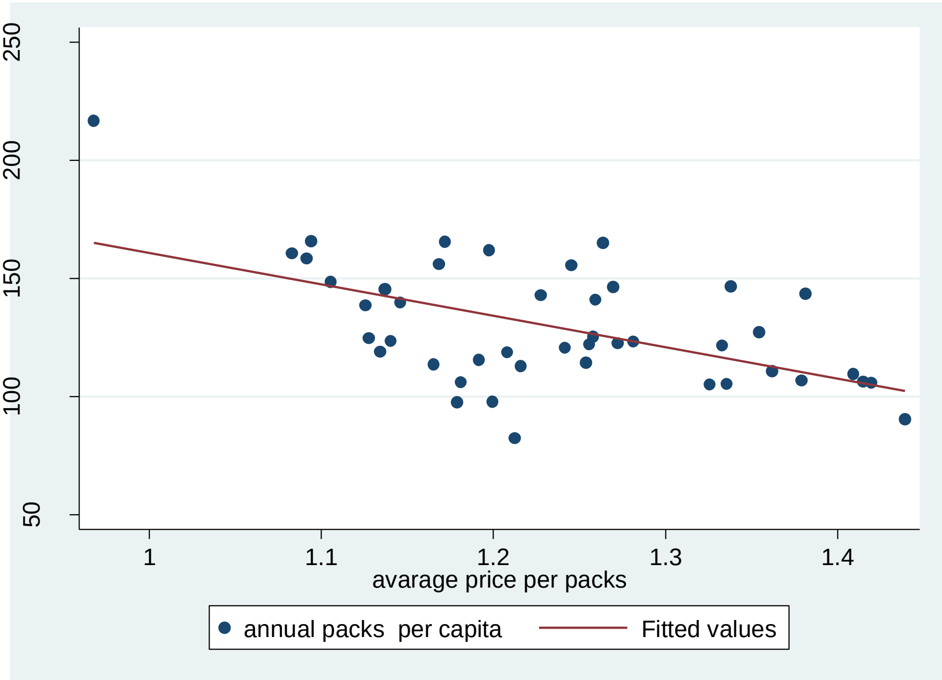
you should be able to visualize the following graph :



Moreovere, writing :

twoway (scatter packs price) (lfit packs price)

you should be able to visualize this graph :



You can see it seems to be a negative relation between variables.