PROBLEM SET TWO
SOLUTIONS

Antonio Pacifico

Federica Romei

22/03/2012

1. Use the dataset *birthweight.dta*.

A. Regress *bwghtlbs* (birth weight in pounds) on *cigs* (cigs smoked per day) and explain the model. Is there correlation between the variables? Is the regression statistically significant? How should you prove it? Comment.

B. Make a plot of *bwghtlbs* against *cigs*. Can you prove there is correlation? Comment.

C. Construct a statistical procedure to test that birth weight in pounds is equal whether babies are male or not against the alternative hypothesis that male babies weight more. Use a significance level of 5%. Comment.

SOLUTION

A.

→ Open dataset typing

use birthweight.dta, clear

→ Now, writing

reg bwghtlbs cigs, r

you should be able to visualize this output on STATA :

```
reg bwghtlbs cigs, r

Linear regression                                      Number of obs =     1388
                                                       F(  1,  1386) =    34.29
                                                       Prob > F      =   0.0000
                                                       R-squared     =   0.0227
                                                       Root MSE      =    1.258

------------------------------------------------------------------------------
             |               Robust
    bwghtlbs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.0321108   .0054833    -5.86   0.000    -.0428673   -.0213542
       _cons |   7.485744   .0359093   208.46   0.000     7.415301    7.556186
------------------------------------------------------------------------------
```

→ The model is :

$$\text{bwghtlbs} = \beta_0 + \beta_1 \text{cigs} + \upsilon$$

where :

- β_0 denotes the intercept of population regression line. It has not a clear meaning because it is equivalent to zero cigarettes smoked and does not help you in explaining model. The coefficient is significant and positive correlated with *bwghtlbs.*
- β_1 denotes the slope of population regression line. It is the coefficient of regressor and is negative correlated with *bwghtlbs*; that is it decreases of *-0.03* points for every unitary increase of *cigs*. You can write :

$$\beta_1 = \frac{\Delta_{\text{bwghtlbs}}}{\Delta_{\text{cigs}|\Delta_{\text{cigs}=1}|}}$$

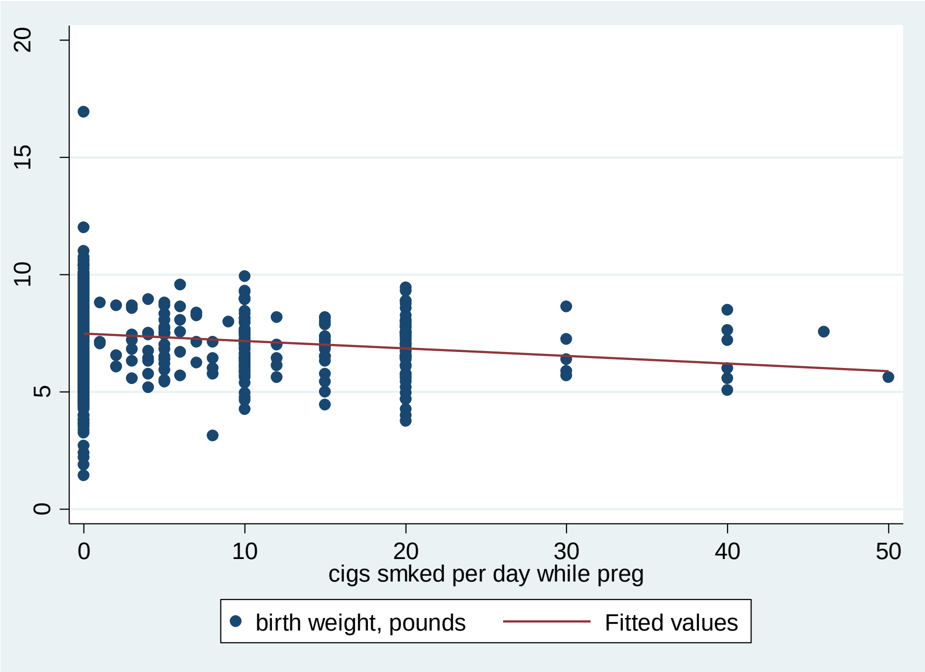
The coefficient is significant since p-value is lower than α and therefore it is significantly different from zero at 5%.

The regression R^2 is closed to zero (0.02), thus the proportion of variance in *bwghtlbs* which can be predicted from *cigs* is low. You can conclude that the model is significant at 5%, but you need to consider other variables in explaining *bwghtlbs.*

B.

→ Writing :

twoway (scatter bwghtlbs cigs) (lfit bwghtlbs cigs) you obtain the following graph :



Observing the graph you can easily prove the negative correlation between *bwghtlbs* and *cigs*.

C.

→ You have to test the following hypothesis :

$$H_0 : \Delta\left(\overline{bwghtlbs}\right) = \overline{bwghtlbs_M} - \overline{bwghtlbs_F} = 0$$
$$H_1 : \Delta\left(\overline{bwghtlbs}\right) = \overline{bwghtlbs_M} - \overline{bwghtlbs_F} > 0$$

→ 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.

Writing

mean bwghtlbs if male==1

and

mean bwghtlbs if male==0

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

```
mean bwghtlbs if male==1

Mean estimation                     Number of obs    =     723

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    bwghtlbs |   7.506829   .0471615      7.414239    7.599419
--------------------------------------------------------------
```

```
mean  bwghtlbs if male==0

Mean estimation                     Number of obs    =     665

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    bwghtlbs |   7.322932    .049268      7.226192    7.419672
--------------------------------------------------------------
```

→ Now, you can compute t-statistic as :

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

In STATA :

display (7.506829 – 7.322932)/sqrt(.0471615^2+.049268^2)

hence you should obtain that :

$$t_{value} = \Delta_{\widehat{bwghtlbs}} = 2.69$$

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

2.69 > +1.64

hence you cannot reject the Alternative hypothesis that male babies weight more.

2. Use the dataset *birthweight.dta*.

A. Do the same regression of point 1.A adding *male* and explain the new model. Does this new variable help you in explaining *bwghtlbs*? Why? Comment. { HINT : Try observing significance of model}

B. How can you interpret β_2 and β_0? Do you note some similarities with the point 1.C? Why? Comment. {HINT : Try testing that $\beta_2 = 0$ against the alternative hypothesis that $\beta_2 > 0$, hence compare the two t-statistics}

SOLUTION

A.

→ Open dataset typing

use birthweight.dta, clear

→ Now, writing

reg bwghtlbs cigs male, r

you can visualize :

```
reg  bwghtlbs cigs male, r

Linear regression                                      Number of obs =    1388
                                                       F(  2,  1385) =   22.51
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0279
                                                       Root MSE      =  1.2551

------------------------------------------------------------------------------
             |               Robust
    bwghtlbs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.0321032   .0054368    -5.90   0.000    -.0427685    -.021438
        male |   .1837089   .0674546     2.72   0.007     .0513847    .3160331
       _cons |   7.390035   .0508868   145.23   0.000     7.290212    7.489859
------------------------------------------------------------------------------
```

→ The model is :

$$\text{bwghtlbs} = \beta_0 + \beta_1\text{cigs} + \beta_2\text{male} + \upsilon$$

where :

β_2 denotes the coefficients of added new regressor. This latter is a dummy variable =1 if babies are male and =0 otherwise. It is positive correlated with dependent variable; to be more precise *bwghtlbs* increases of *+0.1837* units for every unitary increase of *male*.
The coefficient is significant since p-value is lower than α, hence it is significantly different from zero at 5%.

If you observe the negative correlation between *bwghtlbs* and *cigs* is about unvaried and the regression R^2 is not increased enough [from *0.0227* to *0.0279*].

Therefore you can conclude that the new regressor in not a good predictor of *bwghtlbs*.

B.

→ Interpreting β_0 and β_2

- The first coefficient [β_0] is the average level of birth weight in pounds when babies are not male, hence :

$$\beta_0 = E(bwghtlbs|male = 0)$$

You can easily compute β_0 by typing :

mean bwghtlbs if male==0

```
mean  bwghtlbs if male==0

Mean estimation                          Number of obs    =     665

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    bwghtlbs |   7.322932    .049268      7.226192    7.419672
--------------------------------------------------------------
```

thus : $\beta_0 = 7.32$

- The second coefficient [β_2] is the difference between expected value of birth weight in pounds when babies are male and not, hence :

$$\beta_2 = E(bwghtlbs|male = 1) - E(bwghtlbs|male = 0)$$

If you type :

mean bwghtlbs if male==1

```
mean bwghtlbs if male==1

Mean estimation                   Number of obs    =     723

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    bwghtlbs |   7.506829   .0471615      7.414239    7.599419
--------------------------------------------------------------
```

you can easily compute it by differentiating :

`7.506829-7.322932`

Writing :

display `7.506829-7.322932`

you obtain that :

$$\beta_2 = .183897$$

You can prove it making a simple linear regression, thus :

```
reg bwghtlbs male, r

Linear regression                                      Number of obs =    1388
                                                       F(  1,  1386) =    7.27
                                                       Prob > F      =  0.0071
                                                       R-squared     =  0.0052
                                                       Root MSE      =  1.2693

------------------------------------------------------------------------------
             |               Robust
    bwghtlbs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .1838969    .068202     2.70   0.007     .0501065    .3176872
       _cons |   7.322932   .0492664   148.64   0.000     7.226287    7.419577
------------------------------------------------------------------------------
```

→ By observing point **<u>1.C</u>**, you can observe that both two statistic-test are equal to *2.7*. Pay attention to distinguish them, because :

In the regression, lo *Standard Error* is *robust* [that is unbiased estimator of standard deviation] and equal to :

$SE = \frac{\sigma_y}{\sqrt{n-k}}$ <u>where</u> $k = number\ of\ regressors\ [including\ the\ constant]$
$n = number\ of\ observations$

In hypothesis test it is not unbiased and equal to :

$$SE = \frac{\sigma_y}{\sqrt{n}}$$

Therefore <u>they are equal but not the same.</u>

3. Use the dataset *cigs.dta*.

A. After you've made a regression (*packs* on *price*) and explained significance of coefficients, suppose that your boss wants the regression expressed in cigarettes smoked and euro. How will β_1 and β_0 change? Comment carefully alighting on every computation. {HINT : Remember that € = *1.39$* and each *pack* contains *20 cigs*}

B. You have been moved in the Tax Department and you have a new boss who wants to collect money from the taxation of cigarettes. He knows that you did a regression in a similar topic for the other Department. Assuming that your regression is reliable, do you advice to him to raise the taxation on cigarettes? Comment carefully.

<u>SOLUTION</u>

A.

→ Open dataset typing

use cigs.dta, clear

→ Now, writing

reg packs price, r

you obtain the following output on STATA :

```
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
------------------------------------------------------------------------------
```

→ The model is :

$$packs = \beta_0 + \beta_1 price + \upsilon$$

The coefficients are both significant since p-value is very low; to be more precise it is lower than α.

→ If your boss wants the regression expressed in cigarettes smoked and euro, you have to generate two new variables :

- *Euro = 1\$*/1.39 , which denotes the price in euro for each pack [or for 20 cigarettes smoked]
- *Cigtts = 20*1pack* , which denotes the cigarettes smoked contained in a pack

In STATA :

gen euro = price(1/1.39)*

*gen cigtts = packs*20*

The model becomes :

$$cigtts = \beta_0 + \beta_1 euro + \upsilon$$

thus you have to make a new regression writing :

reg cigtts euro, r

```
reg cigtts euro, r

Linear regression                                      Number of obs =      46
                                                       F(  1,    44) =   15.30
                                                       Prob > F      =  0.0003
                                                       R-squared     =  0.3031
                                                       Root MSE      =  431.25

------------------------------------------------------------------------------
             |               Robust
      cigtts |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        euro |  -3693.927   944.4604    -3.91   0.000    -5597.361   -1790.492
       _cons |   5871.647   862.7107     6.81   0.000     4132.967    7610.326
------------------------------------------------------------------------------
```

You can observe that the model is always significant at 5% since p-value is very low and closed to zero. The negative correlation between dependent variable and the only regressor is increased; *cigtts* decreases of *-3693.927* for every unitary increase of *euro*.
Pay attention to note that regression R^2 is unchanged because the significance of model does not depend on measure of coefficients.

→ You can obtain the new coefficients without making a new regression, hence :

$$\beta_{1,NEW} = \frac{Cov(euro|cigtts)}{Var(euro)} = \frac{Cov\left(\frac{1}{1.39} * price \middle| 20 * packs\right)}{Var\left(\frac{1}{1.39} * price\right)} = \frac{\frac{20}{1.39} Cov(price|packs)}{\frac{1}{1.39^2} Var(price)} =$$

$$= \frac{20}{\cancel{1.39}} * 1.39^{\cancel{2}} \frac{Cov(price|packs)}{Var(price)} = 20 * 1.39 * \beta_{1,OLD} =$$

$$= 20 * 1.39 * (-132.8751) = -3693.927$$

This way of reasoning is quite complicated, but you can simplify the thing in this way :

- An increase in 1$ decreases *packs* of $\beta_{1,OLD}$ unit
- An increase in 1$ decreases *cigtts* of $\beta_{1,OLD} * 20$
- An increase in 1.39$ correspond to an increase in 1€
- An increase in 1.39$ decreases *cigtts* of $\beta_{1,OLD} * 20 * 1.39$
- An increase in 1€ decreases *cigtts* of $\beta_{1,OLD} * 20 * 1.39$

$$\beta_{0,NEW} = \overline{cigtts} - \beta_{1,NEW} * \overline{euro} = \overline{cigtts} - \overline{euro} * 20 * 1.39 * \beta_{1,OLD} =$$

$$= 20 * packs - \frac{1}{\cancel{1.39}} * price * 20 * \cancel{1.39} * \beta_{1,OLD} = 20[packs - \beta_{1,OLD} * price] =$$

$$= 20 * \beta_{0,OLD} = 20 * 293.5823 = 5871.64$$

Notice that the change in β_0 depends on only change in the dependent variable. This is always true if and only if the change in independent variable is of this type :

$$x_{NEW} = b * x_{OLD}$$

When, instead, you have a change of this type :

$$x_{NEW} = a + b * x_{OLD}$$

then also independent variable has an effect on β_0.

B.

→ The answer is No. Being all economist you should be able to motivate by yourself.

4.

A. **Regress *narr86* on *black* and explain the model. How can you interpret the coefficients? Have a shot of constructing a statistical procedure. Comment.**

B. **Regress *narr86* on *hispan*. Explain the model and interpret the coefficients. Have a shot of constructing a statistical procedure. Comment.**

C. **Regress jointly *narr86* on *black* and *hispan* explaining the model. How will model change? Comment. {HINT : you have to comment carefully significance of coefficients}**

SOLUTION

A.

→ Open dataset typing

use crime.dta, clear

→ Now, writing

reg narr86 black, r

you can visualize :

```
reg narr86 black, r

Linear regression                                      Number of obs =    2725
                                                       F(  1,  2723) =   35.37
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0223
                                                       Root MSE      =   .8496

------------------------------------------------------------------------------
             |               Robust
      narr86 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |   .3488323    .058657     5.95   0.000     .2338156    .4638489
       _cons |   .3482065   .0160955    21.63   0.000     .3166458    .3797671
------------------------------------------------------------------------------
```

→ The model is :

$$\text{narr86} = \beta_0 + \beta_1 \text{black} + \upsilon$$

You should be able to prove that model is significant at 5%.

→ Interpreting β_0 and β_1 :

- The first coefficient [β_0] is the average level of number of arrests in 1986 when people are not black, hence :

  $$\beta_0 = E(narr86|black = 0)$$

  You can easily compute β_0 by typing :

  mean narr86 if black==0

```
mean  narr86 if  black==0

Mean estimation                    Number of obs    =    2286

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
      narr86 |   .3482065   .0160931      .3166478    .3797651
--------------------------------------------------------------
```

thus : $\beta_0 = .3482$

- The second coefficient [β_1] is the difference between expected value of number of arrests when people are black and not, hence :

$$\beta_1 = E(narr86|black = 1) - E(narr86|black = 0)$$

If you type :

mean narr86 if black==1

```
mean  narr86 if  black==1

Mean estimation                      Number of obs     =     439

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
      narr86 |   .6970387   .0564491       .586094    .8079834
--------------------------------------------------------------
```

you can easily compute it by differentiating :

`.6970387-.3482065`

Writing :

display `.6970387-.3482065`

you obtain that :

$$\beta_1 = .3488$$

→ You can test the following hypothesis :

$$\beta_1 = 0 \xRightarrow{implies\ that} H_0 : \Delta\,(\overline{narr86}) = \overline{narr86_B} - \overline{narr86_{NB}} = 0$$
$$\beta_1 \neq 0 \xRightarrow{implies\ that} H_1 : \Delta\,(\overline{narr86}) = \overline{narr86_B} - \overline{narr86_{NB}} \neq 0$$

The t-statistic is equal to *5.95.* In absolute value :

$$|5.95| > 1.96$$

therefore you would not reject the Alternative hypothesis.

B.

→ Writing

reg narr86 hispan, r

you can visualize :

```
reg narr86 hispan, r

Linear regression                                      Number of obs =     2725
                                                       F(  1,  2723) =     6.92
                                                       Prob > F      =   0.0086
                                                       R-squared     =   0.0028
                                                       Root MSE      =   .85803

------------------------------------------------------------------------------
             |               Robust
      narr86 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      hispan |   .1103311    .041952     2.63   0.009     .0280702    .1925921
       _cons |    .380394   .0181878    20.91   0.000     .3447308    .4160572
------------------------------------------------------------------------------
```

→ The model is :

$$narr86 = \beta_0 + \beta_1 hispan + \upsilon$$

You should be able to prove that model is significant at 5%.

→ Interpreting β_0 and β_1 :

- The first coefficient [β_0] is the average level of number of arrests in 1986 when people are not hispanic, hence :

$$\beta_0 = E(narr86|hispan = 0)$$

You can easily compute β_0 by typing :

mean narr86 if hispan==0

```
mean  narr86 if hispan==0

Mean estimation                     Number of obs    =    2132

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
      narr86 |    .380394   .0181854      .3447311    .4160569
--------------------------------------------------------------
```

thus : $\beta_0 = .380394$

- The second coefficient [β_1] is the difference between expected value of number of arrests when people are hispanic and not, hence :

$$\beta_1 = E(narr86|hispan = 1) - E(narr86|hispan = 0)$$

If you type :

mean narr86 if hispan==1

```
mean narr86 if hispan==1

Mean estimation                         Number of obs    =     593

--------------------------------------------------------------
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
      narr86 |   .4907251   .0378225      .4164426    .5650076
--------------------------------------------------------------
```

you can easily compute it by differentiating :

`.4907251-.380394`

Writing :

display `.4907251-.380394`

you obtain that :

$$\beta_1 = .11033$$

→ You can test the following hypothesis :

$$\beta_1 = 0 \xRightarrow{\textit{implies that}} H_0 : \Delta\left(\overline{narr86}\right) = \overline{narr86_H} - \overline{narr86_{NH}} = 0$$

$$\beta_1 \neq 0 \xRightarrow{\textit{implies that}} H_1 : \Delta\left(\overline{narr86}\right) = \overline{narr86_H} - \overline{narr86_{NH}} \neq 0$$

The t-statistic is equal to `2.63.` In absolute value :

$$|2.63| > 1.96$$

therefore you cannot reject the Alternative hypothesis.

C.

→ Writing :

reg narr86 black hispan, r

you can visualize :

```
reg narr86 black hispan, r

Linear regression                                      Number of obs =     2725
                                                       F(  2,  2722) =    30.33
                                                       Prob > F      =   0.0000
                                                       R-squared     =   0.0304
                                                       Root MSE      =   .84624

------------------------------------------------------------------------------
             |               Robust
      narr86 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |   .3987517    .058942     6.77   0.000      .283176    .5143273
      hispan |   .1924381   .0414864     4.64   0.000     .1110901    .2737861
       _cons |   .2982871   .0170711    17.47   0.000     .2648135    .3317606
------------------------------------------------------------------------------
```

→ The model is :

$$\text{narr86} = \beta_0 + \beta_1 \text{black} + \beta_2 \text{hispan} + \upsilon$$

You should be able to prove that model is significant at 5%. In this case the significance of model improves, but the regression R^2 is still very closed to zero.

You can conclude that variables are good predictors of *narr86*, nevertheless you need to consider other regressors in explaining the dependent variable. Have you an idea?