Multiple Regression Review

Antonio Pacifico
Federica Romei

October 22, 2012

1. We will use dataset Smoke.dta. This dataset contains 808 observations for the following variables:

   - *cigs*: cigarettes smoked per day;
   - *educ*: years of schooling;
   - *cigpric*: state cigarette price, cents per pack;
   - *age*: it is in years;
   - *income*: annual income in US Dollars;
   - *white*: dummy variable
     - =1 if white;
     - =0 otherwise.

   This dataset has been used in order to understand which are the variables that influence cigarettes consumption.

   (a) Regress *cigs* on *white*. Explain the coefficients and their significance.

   Solution: Under the assumptions:

   - *cigs* and *white* are i.i.d;
   - $E(u|white) = 0$
   - Fourth moments are well defined

   we can write the linear model:

   $$cigs_i = \beta_0 + \beta_1 white_i + u_i.$$

In this case

$$\beta_0 = E(cigs|non\ white)$$

and

$$\beta_1 = E(cigs|white) - E(cigs|non; white).$$

Hence their sample counterpart will be:

$$\hat{\beta}_0 = \overline{cigs_{non\ white}}$$

and

$$\hat{\beta}_1 = \overline{cigs_{white}} - \overline{cigs_{non\ white}}$$

where $\overline{cigs_{non\ white}}$ is the sample average of the smoked cigarettes by the non white and $\overline{cigs_{white}}$ is the sample average of the smoked cigarettes by the white.
Stata output is:

```
reg cigs white, r

Linear regression                               Number of obs =     807
                                                F(  1,   805) =    0.05
                                                Prob > F      =  0.8198
                                                R-squared     =  0.0001
                                                Root MSE      =   13.73

------------------------------------------------------------------------------
             |               Robust
        cigs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |   .3168015   1.389959     0.23   0.820     -2.41157    3.045173
       _cons |   8.408163   1.288837     6.52   0.000     5.878286    10.93804
------------------------------------------------------------------------------
```

β_0 is 8.4, which means that a non white person smokes on average 8 cigarettes per day.
β_1 is 0.31 which means that white individuals smoke more than non white ones.
As we expected, β_0 is significantly different from zero while β_1 is not significant at 5% level from zero. Indeed $|t_{\beta_1}| < 1.96$, $p - value_{\beta_1} > 0.05$ and the $C.I._{\beta_1}$ contains zero.
R^2 is very close to zero.
We can infer that white and non white individuals smoke the same number of cigarettes.

(b) Regress *cigs* on *income*. Explain the coefficient and their significance. What does it happen to β_0 and β_1 if we change the *income* from dollars to pounds and *cigs* from cigarettes to

pack[1]?

Solution: Under the assumptions:

- *cigs* and *income* are i.i.d;
- $E(u|income) = 0$
- Fourth moments are well defined

we can write the linear model as:

$$cigs_i = \beta_0 + \beta_1 income_i + u_i.$$

In this case

$$\beta_1 = \frac{\Delta cigs}{\Delta income}_{|\Delta income=1}$$

while β_0 has not a clear meaning because it is meaningless think to an income=0.
Running the regression on Stata we will see:

```
reg cigs white, r

reg cigs income,r

Linear regression                               Number of obs =     807
                                                F(  1,   805) =    2.57
                                                Prob > F      =  0.1095
                                                R-squared     =  0.0028
                                                Root MSE      =  13.711

------------------------------------------------------------------------------
             |               Robust
        cigs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |   .0000799   .0000498     1.60   0.110     -.000018    .0001777
       _cons |   7.144685   .9933633     7.19   0.000     5.194797    9.094573
------------------------------------------------------------------------------
```

We focus on β_1. It is very close to zero. His $|t| < 1.96$, his $p - value > 0.05$ and his Confidence interval contains the zero, which means that β_1 is not significantly different from zero at 5% level.
R^2 is close to zero, hence this variable alone doesn't explain much cigarettes consumption.

[1]We assume 20 cigarettes per pack and Dollar= 0.63 pounds.

We can now change the variable in this way:

$$pack = cigs/20$$

and

$$Income_{pounds} = income_{dollars} * 0.63$$

This means that the new β_1 will be:

$$\begin{aligned}\beta_{1,new} &= \frac{Cov(pack, ; income_p)}{Var(income_p)} \\ &\frac{Cov(cigs/20; income_d * 0.63)}{Var(income_d * 0.63)} \\ &\frac{Cov(cigs\ income_d)}{Var(income_d) * 20 * 0.63} \\ &\frac{1}{0.63 * 20}\beta_{1,old}\end{aligned}$$

Moreover $\beta_{0,new}$ will be

$$\begin{aligned}\beta_{0,new} &= \bar{pack} - \beta_{1,new}\bar{income_p} \\ &= \frac{\bar{cigs}}{20} - \frac{\beta_{1,old}}{0.63 * 20}\bar{income_d} * 0.63 \\ &= \frac{1}{20}\beta_{0,old}\end{aligned}$$

To be sure about the result we create the variables $income_p$ and $packs$ on Stata and run the new regression. This is the result:

```
 gen pack=cigs/20
(1 missing value generated)

. gen incomep=income*0.63
(1 missing value generated)

. reg pack incomep, r

Linear regression                                      Number of obs =     807
                                                       F(  1,   805) =    2.57
                                                       Prob > F      =  0.1095
                                                       R-squared     =  0.0028
                                                       Root MSE      =  .68553

------------------------------------------------------------------------------
             |               Robust
        pack |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     incomep |   6.34e-06   3.96e-06     1.60   0.110    -1.43e-06    .0000141
       _cons |   .3572342   .0496682     7.19   0.000     .2597398    .4547286
------------------------------------------------------------------------------
```

As you can see $\beta_0 = \frac{7.14}{20}$.

(c) Create a new variable *cigsan* as $cigs * 365$[2] Regress *cigsan* on *cigpric*, *white*, *income* and *age*. Describe the coefficients and their significance.

Solution: Under the assumptions:

- *cigsan*, *cigpric*, *white*, *income* and *age*, are i.i.d;
- $E(u|cigpric, income, white, age) = 0$
- Fourth moments are well defined;

we can write the linear model as:

$$cigsan_i = \beta_0 + \beta_1 white_i + \beta_2 income_i + \beta_3 age_i + \beta_4 cigpric_i + u_i.$$

Running the regression on Stata we will have:

[2]These are the cigarettes smoked per year. You can create on Stata writing `gen cigsan=cigs*365`.

```
reg cigsan white income age cigpric,r

Linear regression                                      Number of obs =     807
                                                       F(  4,   802) =    1.19
                                                       Prob > F      =  0.3119
                                                       R-squared     =  0.0045
                                                       Root MSE      =  5009.5

------------------------------------------------------------------------------
             |               Robust
      cigsan |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |   99.02365   510.9778     0.19   0.846    -903.9882    1102.035
      income |   .0282061   .0182423     1.55   0.122    -.0076022    .0640145
         age |  -11.08889   8.495644    -1.31   0.192    -27.76521    5.587436
     cigpric |  -13.06148   39.51218    -0.33   0.741    -90.62098    64.49802
       _cons |   3783.952   2558.308     1.48   0.140    -1237.818    8805.722
------------------------------------------------------------------------------
```

Now we have to interpret the coefficients:

- $\hat{\beta}_1$ is 99. Take 2 identical individuals, Sandra and Monika. They are similar in all the characteristic (same income, same age and their price costs the same). Monika is white while Sandra is not. Then Monika will smoke 99 cigarettes more in a year. Hence $\hat{\beta}_1$ is the difference between the cigarettes smoked in a year by a white individual minus the cigarettes smoked in a year by a non white individual, other things being equal.
- $\hat{\beta}_2$ is 0.02. Take 2 identical individual Paul and Samuel. They are both non white, have the same age and their cigarettes cost the same. Paul has 1 dollar more than Samuel. Then Paul will smoke 0.02 cigarettes more. This seems meaningless. If you say that Paul has 100 dollars more than Samuel you can say that Paul will smoke 2 cigarettes more in a year. Then $\hat{\beta}_2$ is the change in cigarettes smoked due to a unitary change in income, other things being equal.
- $\hat{\beta}_3$ is -11. As before, take 2 identical individuals, Michael and John. John is 1 year older than Michael, then John will smoke 11 cigarettes less than Michael. $\hat{\beta}_3$ is the change in cigarettes smoked per year due to a unitary change in age, other things being equal.

- $\hat{\beta}_4$ is -13. Suppose to have 2 identical individuals, Mark and Anthony. They are both white, have the same income, the same age and same features. The Mark's cigarettes cost 1 cents more than the Anthony's cigarettes. Then Mark will smokes 13 cigarettes less than Anthony in a year. $\hat{\beta}_4$ is the change in cigarettes smoked per year due to a unitary change in price, other things being equal.

As we can easily see none of these coefficients is significant at 5% level.
R^2 is very low. this is not a good model to predict cigarettes consumption .

(d) Add to the (c) regression *educ*. Does something change?

Solution: If we add the variable *educ* we will have:

```
 Linear regression                                      Number of obs =     807
                                                        F(  5,   801) =    1.87
                                                        Prob > F      =  0.0975
                                                        R-squared     =  0.0106
                                                        Root MSE      =  4997.4

------------------------------------------------------------------------------
             |               Robust
      cigsan |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       white |   97.04046   509.9962     0.19   0.849    -904.0465    1098.127
      income |   .0430486   .0195781     2.20   0.028     .0046181    .0814791
         age |  -15.04847   8.704165    -1.73   0.084    -32.13414    2.037195
     cigpric |  -11.28301   39.24039    -0.29   0.774    -88.30914    65.74312
        educ |  -137.3821   60.37493    -2.28   0.023    -255.8939   -18.87032
       _cons |   5268.482   2654.667     1.98   0.048     57.55565    10479.41
------------------------------------------------------------------------------
```

The coefficient are slightly changed with respect the previous case. $\hat{\beta}_5$ is -137. Assume to have 2 identical individual, Sam and Kenneth. Sam and Kenneth did the same college and same university but Sam did a master, while Kenneth didn't (this means 1 year of education more). Then Sam will smoke 137 cigarettes less than Kenneth.
$|t_{\beta_5}| > 1.96$, $p - value_{\beta_5} < 0.05$ and $0 \notin CI_{\beta_5}$, then β_5 is significant different from zero at 5% level.
R^2 is increased with respect to the previous case. Hence we can say that *educ* is a good variable to explain cigarettes consumption.

2. We use the dataset affairs.dta which contains cross section data from a survey conducted by Psychology Today in 1969. This dataset contains 601 observations on 9 variables:

   - $affairs$: how often engaged in extramarital sexual intercourse during the past years?
   - *gender*: dummy variable:
     - =1 if male;
     - =0 if female.
   - *age* variable coding years:
     - 17.5 if under 20;
     - 22 if 20-24;
     - 27 if 25-29;
     - 32 if 30-34;
     - 37 if 35-39;
     - 42 if 40-44;
     - 47 if 45-49;
     - 52 if 50-54;
     - 57 if over 55.
   - *yearmarried*: variable coding number of years married:
     - 0.125 if 3 months or less;
     - 0.417 if 4-6 months;
     - 0.75 if 6 months-1year;
     - 1.5 if 1-2 years;
     - 4 if 3-5 years;
     - 7 if 6-8 years;
     - 10 if 9-11 years;
     - 15 if 12 or more.
   - *children*: dummy variable:
     - =1 if you have children in the marriage;
     - =0 otherwise.
   - *religiousness*: variable coding religiousness:
     - 1= anti;
     - 2=not at all;
     - 3= slightly;
     - 4=somewhat;

  - 5=very.
- *education* : variable coding level of education:
  - 9= grad school;
  - 12=high school graduate;
  - 14= some college;
  - 16= graduate college;
  - 17 = some graduate work;
  - 18= master's degree;
  - 20= Ph.D, M.D or other advanced degree.
- *occupation* variable coding occupation according to Hollingshead classification;
- *rating*: variable coding self rating of marriage:
  - 1=very unhappy;
  - 2=somewhat unhappy;
  - 3= average;
  - 4= happier than the average;
  - very happy.

(a) Use this dataset and make a regression of $affairs$ on what you want (it should be a **multiple** regression). Explain to the class your results.

(b) Whatever x_1[3] you use, do you think that there is a **causal effect** between $affairs$ and your x_1?

Solution: These questions have not a proper answer. It depends on the regressor choose by you.

[3]With x_1 i mean the first regressor you use.