Dummy variables and Multiple Regression

Siria Angino
Federica Romei

October 14, 2013

1. For this exercise we are going to use the dataset Openness.dta that contains inflation and import share data from 1973.The dataset contains all non-centrally planned countries with flexible and fixed exchange rate listed by Summers and Heston for whom data on inflation and openness are available. Inflation is measured as average annual change in the log GDP or GNP deflator since 1973. Openness is measured as the average share of imports in GDP or GNP over the years beginning in1973. This dataset contains:

   - *open*: imports as % GDP, '73-'80
   - inf: avg. annual inflation, '73-
   - *oil*: dummy variable :
     - =1if major oil producer,
     - =0 otherwise;

   This dataset have been used by Romer in order to test if countries more open have lower level of inflation. The data are collecting from 1973 for a specific reason : before the '73, there was the Bretton Woods [1] system that limited the possibility to pursue independent monetary policy.

   You will not use other variables hence you don't need the description.

   (a) Construct a statistical procedure to test whether oil countries and non-oil countries have the same level of inflation, against the Alternative Hypothesis that they have different level.

[1] The chief features of the Bretton Woods system were an obligation for each country to adopt a monetary policy that maintained the exchange rate of its currency within a fixed valueÑplus or minus one percentÑin terms of gold and the ability of the International Monetary Funds to bridge temporary imbalances of payments. Then, on August 15, 1971 the United States unilaterally terminated convertibility of the dollar to gold. This action created the situation whereby the United States dollar became the sole backing of currencies and a reserve currency for the member states. In the face of increasing financial strain, the system collapsed in 1971.

Solution:

First of all, let's state the Null Hypothesis (H_0) and the alternative (H_1):

$$H_0 : E(inflation|oil = 1) = E(inflation|oil = 0),$$
$$\text{vs}$$
$$H_1 : E(inflation|oil = 1) \neq E(inflation|oil = 0).$$

We are considering a dual side alternative, but it could also be possible to entertain as alternative the situation in which test whether oil countries tend to have lower or higher inflation than non oil ones. Our variable of interest is

$$\Delta i = E(inflation|oil = 1) - E(inflation|oil = 0).$$

Unfortunately we don't have the real expected values, but under regularity condition we can replace Δi with his sample counterpart, that is an estimator. To be more precise if $inflation$ is i.i.d (independent means that level of inflation of the observation i doesn't depend on the level of inflation of another observation in the sample while identically distributed means that we are assume that all the observations have the same distribution) and $E(inflationl) < \infty$, we can apply the Law of Large Number, i.e.

$$\widehat{\Delta i} = \overline{inflation_{oil}} - \overline{inflation_{non\ oil}} \xrightarrow{p} \Delta i$$

where $\overline{inflation_{oil}}$ and $\overline{inflation_{non\ oil}}$ are the sample averages of the inflation level of oil and non oil countries, respectively. Moreover, if we assume that $Var(inflation) < \infty$, we can apply the Central Limit Theorem, which implies that under the null hypothesis,

$$t = \frac{\widehat{\Delta i}}{SE(\widehat{\Delta i})} \xrightarrow{D} N(0, 1)$$

where

$$SE(\widehat{\Delta i}) = \sqrt{SE(inflation_{oil})^2 + SE(Inflation_{non\ oil})}^2.$$

We will reject H_0 if $|t| > 1.96$ and not reject otherwise. From the data we get:

```
> Delta_inf= (mean(inf[oil==1] )-mean(inf[oil==0])) /
+   sqrt(sd(inf[oil==1])^2/length(inf[oil==1]) +sd(inf[oil==0])^2/length(inf[oil==0]))
> Delta_inf
[1] -2.582764
```

We have all the possible information to compute t, which is

$$t = \frac{10.64 - 17.69}{\sqrt{1.32^2 + 2.38^2}} = -2.59.$$

Thus, we can reject the null hypothesis at 5%.

(b) Regress inf on oil. How can you interpret β_1? And β_0? Do you notice some similarities with the point (a)?

Solution: Under the assumptions:

- inf and u are i.i.d.;
- $E(u|oil) = 0$;
- fourth moment of inf and oil are well defined;

we can build a model like:

$$inf_i = \beta_0 + \beta_1 oil_i + u_i.$$

We can now interpret the coefficients. β_0 and β_1 have a specific meaning in this type of regression, indeed :

$$\beta_0 = E(inf|oil = 0)$$

and

$$\beta_1 = E(inf|oil = 1) - E(inf|oil = 0).$$

Testing that $\beta_1 = 0$ is equivalent to test that $E(inf|oil = 1) - E(inf|oil = 0) = 0$, which means that $\hat{\beta}_1$ is $\widehat{\Delta i}$ of the previous exercise. While $\hat{\beta}_0$ corresponds to $\overline{inflation}_{non\ oil}$.

You can check easily writing on R

```
dum=lm(inf oil)
rob(dum,0.05)
```

```
 > dum=lm(inf~oil)
             Estimate Std. Error   t value     Pr(>|t|)  c.i. 5 % c.i. 95 %
(Intercept) 17.697206   2.399135  7.376495 3.029373e-11  12.99499 22.399423
oil         -7.054343   2.794080 -2.524746 1.297831e-02 -12.53064 -1.578047
```

$\hat{\beta}_1$ is -7.05 and means that oil producer countries have on average -7.05 points of inflation than non oil producer countries.

$\hat{\beta}_0$ is the average level of inflation of the non producer oil countries.

$t-statistic$ of β_1 is -2.52 and corresponds to the $t-statitstc$ of the previous exercise.

Pay Attention: t is not exactly the same because the variance in the previous exercise have been divided by n_oil and $n_non\ oil$ respectively. Now, indeed, they have been divided by $n-k$ where $n = n_oil + n_non\ oil$ and k is the number of regressors(included the constant).

We can reject the Null Hypothesis that β_1 is equal to zero at 5% significant level. This means that oil producer countries tend to have lower level of inflation than non oil producer countries.

$p - value$ is very low both for $\hat{\beta}_0$ and $\hat{\beta}_1$. However at 1% significant level we won't reject the Null Hypothesis that $\beta_1 = 0$ (remember that in order to reject $p-value < \alpha$).

Confidence intervals do not contain zero. We can reject the null hypothesis that β_0 is equal to zero as well as the null hypothesis that β_1 is equal to zero.

2. For this exercise we use Crime.dta file that contains data on arrests during the year 1986 and other information on 2725 men born in either 1960 or 1961 in California. Each man in the sample was arrested at least once prior to 1986. The variables of interest are:

   - $narr86$: number of times the man was arrested during 1986, it is zero for most of the sample (72.29%) and it varies from 0 to 12;
   - $pcnv$: is the proportion of arrests prior to 1986 that lead to conviction;
   - $avgsen$: it is the average sentence length served for prior convictions;
   - $ptime86$ is the months spent in prison in 1986;
   - $qemp86$: is the number of quarters during which the man was employed in 1986 (from zero to four).

   $pcnv$ should be interpreted as a proxy for the likelihood to be convicted of a crime while $avgsen$ is a measure of expected severity of punishment, if convicted. $ptime86$ captures the incarcerative effects of crime: if an individual is in prison, he cannot be arrested for a crime outside the prison. Labor market opportunities are crudely captured by $qemp86$.

   (a) Regress $narr86$ on the other variables except $avgsen$. Interpret the coefficients and their significance.

   Solution: Under the assumptions:

   - $pcnv$, $ptime86$, $qemp86$ and u are iid;
   - $E(u|pcvn, ptime86, qemp86) = 0$;
   - Fourth moments are well defined;

   a linear model explaining the arrest can be:

   $$narr86 = \beta_0 + \beta_1 pcnv + \beta_2 ptime86 + \beta_3 qemp86 + u.$$

   In Stata we can write

   ```
   > crim=lm(narr86 pcnv+ptime86 +qemp86)
   >robt(crim, 0,05)
   ```

   and visualize this output:

```
   > crim=lm(narr86~pcnv+ptime86 +qemp86)
>               Estimate  Std. Error   t value     Pr(>|t|)    c.i. 5 %   c.i. 95 %
(Intercept)  0.71177149 0.040931827 17.389194 2.630857e-64  0.63154658  0.79199639
pcnv        -0.14992737 0.034016986 -4.407427 1.086573e-05 -0.21659944 -0.08325531
ptime86     -0.03441992 0.005788535 -5.946223 3.094655e-09 -0.04576524 -0.02307460
qemp86      -0.10411303 0.011543861 -9.018908 3.498731e-19 -0.12673858 -0.08148747
```

As in the univariate regression case, the first column shows the variables you are using as regressors. The constant, β_0 is always in the last row.
The second column shows the coefficients. In the multivariate case the interpretation of the coefficients should be more accurate.

- β_1, is -0.14. Suppose to have 2 men that have been not employed and not in prison in 1986. The first man, Mr Brown, has $pcnv = 0$, which means that he has never been convicted once arrested and the other, Mr White, has been always convicted, once arrested, which means, $pcnv = 1$. Then Mr White should have been arrested -.14 less in 1986 than Mr. Brown. This may be unusual because the arrest cannot decrease of a fraction. Think to have 100 Mr Brown and 100 Mr White. Then the "Mr White" group should have been arrested 14 times less than "Mr Brown" group. This means that to be convicted after the arrest lowers the level of next crime.

- β_2 is -0.03. Assume to have 2 men who as $pcnv = qemp86 = 0$. Assume that the first, Mr Pink, spent 1 month in prison in 1986, while the second, Mr Red, spent 0 month in prison in 1986. Then Mr Pink should have been arrested -0.03 times than Mr Red in 1986. As before, arrests can't decrease of a fraction. Hence, as before, assume to have 100 Mr Pink and 100 Mr Red. Than in the group of Mr Pink there should be 3 arrest less than in Mr Red group.

- β_3 is -0.1. Now assume to have 2 "groups" made by 100 members. For both group $pcnv = ptime86 = 0$. The first group, Mr Green group, works for 1 quarter in 1986 while the second, Mr Black, didn't work. Then in the second group we will observe 10 arrest more in 1986.

The third column shows the Standard Error of the coefficients.

The fourth column shows the $t - statistic$, as in the univariate case. This t is built under the Null Hypothesis that each $\beta_i = 0,\ for\ i = 1, 2, 3$ (**We are performing 4 different tests, one for each β_i**). If $|t| > 1.96$, you will reject the Null Hypothesis at 5% and we can say that the coefficient is significant at 5% level. As you can see all coefficients are significant at 5%.

The fifth column shows the $p-value$. This $p-value$ is built under the Null Hypothesis that that each $\beta_i = 0$, $for\ i = 1, 2, 3$(**We are performing 4 different tests, one for each** β_i). If $p - value < \alpha$ you can reject the Null Hypothesis at $\alpha\%$ significant level. The sixth and seventh columns display the confidence interval at 5%.

(b) Now do the same regression of point (a) adding $avgsen$. Does this new variable help you in explaining $narr86$? Why?

Solution: Adding the new variable we should have this output:

	Estimate	Std. Error	t value	Pr(>|t|)	c.i. 5 %	c.i. 95 %
(Intercept)	0.706756443	0.041196323	17.155814	1.009583e-62	0.626013134	0.78749975
pcnv	-0.150831839	0.033931920	-4.445131	9.135587e-06	-0.217337179	-0.08432650
ptime86	-0.037390818	0.006157812	-6.072095	1.438494e-09	-0.049459907	-0.02532173
qemp86	-0.103341006	0.011575048	-8.927911	7.809221e-19	-0.126027683	-0.08065433
avgsen	0.007443114	0.005332276	1.395861	1.628704e-01	-0.003007955	0.01789418.

We have to interpret only the $\hat{\beta}_4$. Suppose, as before, we have 2 groups of individual. Both group has $pcvn = ptime86 = qemp86 = 0$ and are composed by 1000 individuals. The first group, Mr Magenta, has a sentence length for prior convictions equal to 1, while the second group, Mr Yellow, has a sentence length equal to zero. The regression predicts that in the first group there should be 7 arrests more. This is not the sign we expected.

$|t - statistic| < 1.96$, then you cannot reject the Null Hypothesis at 5% significant level.

$p - value > 0.05$ and you cannot reject at 5% significant level that β_4 is different from 0. You can also test at 10% and not reject H_0.

The Confidence Interval contains zero.

(c) Do you think that there is a causal effect between $narr86$ and $pcnv$[2]?

Solution: We can find some variables that have correlation with $narr86$ and $pcvn$. For example the income of your family can have a negative effect on both. Also the neighborhood or the district in which you live. This means that there is no a causal effect between $narr86$ and $pcvn$.

[2]Causal effect means that $E(pcnv, u) = 0$. There is no variables that has an effect on the proportion of arrests prior to 1986 and contemporary on $narr86$ except the variables we put in the regression