I.V. Review, Probit and Logit

Siria Angino
Federica Romei

December 2, 2013

1. The debate on inflation and openness degree was quite important around the 1980. Some authors, like Romer, point out that if a country was quite open he will have less inflation. Indeed if the Central Bank decide to decrease the interest rate in order to increase the output above the natural level we have two effect:

   - Consumers will have more money and they will spend part of them in the imported goods and the domestic currency will depreciate with respect to the foreign one;
   - The imported goods become more costly than the domestic and the inflation will raise.

   Then the Central Bank has less incentive to create inflation in an open country than in a closed one.
   We will use the Romer dataset openness.dta that contains 114 observation on:

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

(a) Do a OLS regression of inf on *open* and *oil*. Do you think that there is a causal effect between inf and *open*?

Solution: If you run the regression your output will be

```
 Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.48928    4.81306   5.296 6.05e-07
open        -0.21189    0.08064  -2.628  0.00982
oil         -6.00470    2.35503  -2.550  0.01214
---
Heteroskadasticity robust standard errors used

Residual standard error: 23.61 on 111 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.0489,Adjusted R-squared:  0.03176
F-statistic:  9.08 on 2 and Inf DF,  p-value: 0.0001139
```

The degree of openness of a country seems to be correlated in a negative way with the inflation. This result is in favor of the Romer view.
"It is possible, for example, that countries that adopt protectionist policies also adopt other policies benefiting particular interest groups, and that this in turn leads to large budget deficits and hence to high rates of inflation to generate seignorage revenues. If so, a negative correlation between openness and inflation could arise through this channel rather than through the impact of openness on policy-makers' incentives to pursue expansionary policies."

(b) *open* can be endogenous, then Romer suggests to use *log*(*land*) as instrument. Explain weather *log*(*land*) is good instrument?

Solution: A instrument should be exogenous and relevant. We can argue that:

- $E(log(land), u|oil) = 0$ because the dimension of a country should be irrelevant for the determinants of the inflation rate;
- $E(log(land), open|oil) \neq 0$ because we can state that smaller is the country higher will be the level of openness.

To test if *land* is better than *log*(*land*) we can do a regression of:

$$open = \beta_0 + \beta_1\ oil + \beta_2\ log(land) + u$$

and see which one is the best predictor for *open*. If we run the regression our output will be:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  121.684     14.651   8.305 2.70e-13
oil            1.135      5.302   0.214    0.831
lland         -7.611      1.224  -6.216 9.18e-09
---
Heteroskadasticity robust standard errors used

Residual standard error: 17.8 on 111 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4481,Adjusted R-squared:  0.4382
F-statistic: 42.54 on 2 and Inf DF,  p-value: < 2.2e-16
```

log(*land*) seems to be a non weak instrument and a good predictor for the degree of openness of a country - try to figure out why.

(c) Run the regression where you use *log*(*land*) as instrument for *open* and explain the results.

Solution: We write

```
summary(tsls(inf open+oil, oil+lland,data=op))
```

and we got:

```
 reg open oil lland, r
Model Formula: inf ~ open + oil

Instruments: ~oil + lland

Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-21.340 -10.200  -5.076   0.000   2.097 184.500

              Estimate Std. Error  t value   Pr(>|t|)
(Intercept) 29.7630385  5.6738755  5.24563 7.5334e-07 ***
open        -0.3281007  0.1410597 -2.32597   0.021836 *
oil         -5.4289964  9.3022184 -0.58362   0.560657
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 23.7764047 on 111 degrees of freedom
.
```

If the share of import over GDP increases by one percentage point then inflation will decrease by .32 percentage points everything been equal. The marginal effect of the share of imports over GDP seems to be stronger (in absolute value) under tsls regression than under standard OLS. This means that before we were under estimated the impact of the openness degree on the inflation - TRY TO THINK WHY.

It has been conjectured that workplace smoking bans induce smokers to quit by reducing their opportunities to smoke. In this assignment you will estimate the effect of workplace smoking bans on smoking using data on a sample of 10000 U.S. indoor workers from 1991-1993. The dataset contains information on whether individuals were or were not subject to a workplace smoking ban, whether individuals smoked, and other individual characteristics.

Smoking contains 10000 observations on:

- $smokers$: dummy variable:
  - =1 if current smokers
  - =0 otherwise
- $smkban$: dummy variable:
  - =1 if there is a work area smoking ban
  - =0 otherwise
- age: age in years;
- $hsdrop$:dummy variable:
  - =1 if high school drop out
  - =0 otherwise
- $hsgrad$:dummy variable:
  - =1 if high school graduate
  - =0 otherwise
- $colsome$:dummy variable:
  - =1 if some college
  - =0 otherwise
- $black$:dummy variable:
  - =1 black
  - =0 otherwise
- $hispanic$:dummy variable:
  - =1 if hispanic
  - =0 otherwise
- $female$:dummy variable:
  - =1 if female
  - =0 otherwise

Suppose you are working at the Health Department and your boss want to know if the smoking ban decreases the probability to smoke.

(a) Estimate the probability of smoking for (*i*) all workers, (*ii*) workers affected by workplace smoking bans and (*iii*) workers not affected by workplace smoking bans.

Solution: To estimate the probability that a worker is a smoker we can compute the mean of the variable *smoker* and to estimate what is the effect of the smoking ban we can do a regression of *smoker* on *smkban*. This is the output:

```
lm(formula = smoker ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.2423 -0.2423 -0.2423 -0.2423  0.7577 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.242300   0.004285   56.55   <2e-16 ***
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 0.4285 on 9999 degrees of freedom
. summary_rob(lm(smoker~smkban))


> summary_rob(lm(smoker~smkban))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.289595   0.007263  39.873   <2e-16
smkban      -0.077558   0.008953  -8.663   <2e-16
---
Heteroskadasticity robust standard errors used

Residual standard error: 0.4268 on 9998 degrees of freedom
Multiple R-squared:  0.007796,Adjusted R-squared:  0.007697 
F-statistic: 75.04 on 1 and Inf DF,  p-value: < 2.2e-16
```

The probability that a worker is a smoker is 0.24, while the probability that a worker is a smoker in a place in which there is not a smoking ban is 0.28. The probability that a worker is a smoker in a place in which is enforced the smoking ban is 0.21. (You are supposed to know why!)

(b) What is the difference in probability of smoking between workers affected by a workplace smoking ban and workers not affected by a workplace smoking ban?

Solution: The difference is -0.07 and it's significantly different from zero.

(c) Estimate a linear probability model with *smoker* as dependent variables and the following regressor: *smkban*, *female*, *age*, age^2, *hsdrop*, *hsgrad*, *black* and *hispanic*. Compare the estimated effect of a smoking ban from this regression with your answer from (2). Suggest a reason, based on substance of this regression, explaining the change in the estimated effect of smoking ban between (2) and (3).

Solution: The output will be:

```
 > summary_rob(lm(smoker~smkban + age + age2+hsdrop+hsgrad+colsome+colgrad+female+black+hispani

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.411e-02  4.146e-02  -0.340 0.733645
smkban      -4.724e-02  8.971e-03  -5.266 1.43e-07
age          9.674e-03  1.898e-03   5.098 3.50e-07
age2        -1.318e-04  2.193e-05  -6.009 1.94e-09
hsdrop       3.227e-01  1.951e-02  16.539  < 2e-16
hsgrad       2.327e-01  1.260e-02  18.472  < 2e-16
colsome      1.643e-01  1.263e-02  13.006  < 2e-16
colgrad      4.480e-02  1.205e-02   3.717 0.000203
female      -3.326e-02  8.573e-03  -3.879 0.000105
black       -2.757e-02  1.610e-02  -1.712 0.086932
hispanic    -1.048e-01  1.399e-02  -7.492 7.37e-14
---
Heteroskadasticity robust standard errors used

Residual standard error: 0.4163 on 9989 degrees of freedom
Multiple R-squared:  0.05699,Adjusted R-squared:  0.05605
F-statistic: 686.5 on 10 and Inf DF,  p-value: < 2.2e-16
```

In this case the effect of smoking ban is reduced. This happens because the previous model doesn't take in consideration all the other variable. Indeed usually the highest education individual smokes less and usually they work in place in which the smoking ban is enforced. Then before we were over estimating the effect of smoking ban.

This model is not optimal. Indeed we know that *smoker* is a probability, then we don't want that it goes over one or under zero. Moreover the effect of the regressors on the dependent variable is not completely linear.

(d) Estimate a probit model using the same regressors as in point a and c.

Solution: in order to estimate a probit model we have to write:

```
myprob<- glm(smoker smkban,family=binomial(link="probit"),data=ba)
                          summary(myprob)
```

and we will have this output:

```
Call:
glm(formula = smoker ~ smkban, family = binomial(link = "probit"),
    data = ba)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8269  -0.8269  -0.6904  -0.6904   1.7612

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.55457    0.02123 -26.126   <2e-16 ***
smkban      -0.24481    0.02787  -8.784   <2e-16 ***
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11074  on 9999  degrees of freedom
Residual deviance: 10997  on 9998  degrees of freedom
AIC: 11001
```

This means that if the smoking ban is not enforced the worker will have a probability to be a smoker of 0.3. Indeed you have to compute $\Phi(-0.55)$, that is $Prob(z \leq -0.55) = 0.3$. Indeed the probability to be a smoker given that the smoking ban is enforced is 0.22. Again, you have to compute $\Phi(-0.79) = Prob(z \leq -0.79) = 0.22$.

Now we estimate the probit model with all the regressors except for $female$:

```
Call:
glm(formula = smoker ~ smkban + age + age2 + hsdrop + hsgrad +
    colsome + colgrad + black + hispanic, family = binomial(link = "probit"),
    data = ba)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1751  -0.8215  -0.6000  -0.3491   2.6452

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.765e+00  1.521e-01 -11.601  < 2e-16 ***
smkban      -1.725e-01  2.878e-02  -5.993 2.06e-09 ***
age          3.419e-02  6.915e-03   4.944 7.65e-07 ***
age2        -4.656e-04  8.266e-05  -5.632 1.78e-08 ***
hsdrop       1.134e+00  7.214e-02  15.718  < 2e-16 ***
hsgrad       8.608e-01  5.954e-02  14.457  < 2e-16 ***
colsome      6.592e-01  6.079e-02  10.843  < 2e-16 ***
colgrad      2.261e-01  6.502e-02   3.476 0.000508 ***
black       -8.949e-02  5.274e-02  -1.697 0.089713 .
hispanic    -3.318e-01  4.803e-02  -6.909 4.88e-12 ***
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11074  on 9999  degrees of freedom
Residual deviance: 10487  on 9990  degrees of freedom
AIC: 10507

Number of Fisher Scoring iterations: 4
```

Here is more complicated the evaluation of the model.

(e) Mr A is white, non Hispanic, 20 years old, and a high school dropout. Using the probit regression and assuming that he is not subject to a smoking ban, calculate the probability that Mr A smokes. Carry out the calculation again assuming that he is subject to a workplace smoking ban. What is the effect of smoking ban on the probability of smoking?

Solution: In the first case we have:

$$\textit{smoker} = -1.7 + 1.13 + 0.68 - 0.16 = -0.05$$

This means that the probability that Mr A is a smoker is $\Phi(-0.05) = 0.49$
If Mr A works in a place in which is enforced the smoking ban:

$$\textit{smoker} = -1.7 + 1.13 + 0.68 - 0.16 - 0.17 = -0.22$$

This means that the probability that Mr A is a smoker is $\Phi(-0.22) = 0.41$. Then smoking ban reduce the probability to smoke of 8% for Mr A.

(f) Estimate a logit model using the same regression of point a and c and estimate again point e.

Solution: In order to estimate a logit model you have to write:

```
mylog<- glm(smoker smkban,family="binomial",data=ba)
                summary(mylog)
```

and you will get this output:

```
Call:
glm(formula = smoker ~ smkban, family = "binomial", data = ba)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8269  -0.8269  -0.6904  -0.6904   1.7612

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.89735    0.03529 -25.425   <2e-16 ***
smkban      -0.41534    0.04719  -8.801   <2e-16 ***
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11074  on 9999  degrees of freedom
Residual deviance: 10997  on 9998  degrees of freedom
AIC: 11001

Number of Fisher Scoring iterations: 4
```

In this case if the smoking ban is not enforced the probability to be a smoker will be:

$$\frac{1}{1+exp(0.89)} = 0.29$$

If instead the ban is enforced you get:

$$\frac{1}{1+exp(1.3)} = 0.21$$

The second regression is:

```
Isummary(mylog)

Call:
glm(formula = smoker ˜ smkban + age + age2 + hsdrop + hsgrad +
    colsome + colgrad + black + hispanic, family = "binomial",
    data = ba)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1930  -0.8199  -0.5946  -0.3598   2.5575

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.0456823  0.2667659 -11.417  < 2e-16 ***
smkban      -0.2856655  0.0489002  -5.842 5.16e-09 ***
age          0.0593161  0.0119782   4.952 7.34e-07 ***
age2        -0.0008146  0.0001443  -5.646 1.64e-08 ***
hsdrop       2.0026276  0.1322591  15.142  < 2e-16 ***
hsgrad       1.5389034  0.1142371  13.471  < 2e-16 ***
colsome      1.1978599  0.1164945  10.283  < 2e-16 ***
colgrad      0.4289648  0.1257726   3.411 0.000648 ***
black       -0.1653757  0.0900340  -1.837 0.066237 .
hispanic    -0.5877989  0.0832579  -7.060 1.67e-12 ***
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11074  on 9999  degrees of freedom
Residual deviance: 10483  on 9990  degrees of freedom
AIC: 10503

Number of Fisher Scoring iterations: 4
```

In the non ban situation Mr A will have:

$$smoker = -3 + 2 + 1.18 - 0.324 = -0.144$$

Hence the probability that Mr A is a smoker is:

$$\frac{1}{1 + exp(-smoker)} = 0.46.$$

If instead he works in a place in which the smoking ban is enforced, he will have:

$$smoker = -3 + 2 + 1.18 - 0.324 - 0.28 = -0.424$$

Then the probability that he is a smoker will be:

$$\frac{1}{1 + exp(-smoker)} = 0.39$$

Then the smoking ban decreases the probability of 7%.