R TUTORIAL 10

Siria Angino
Federica Romei

1. Suppose you wish to measure the impact of smoking on the weight of newborns. You are planning to use the following model:

$$log(bw_i) = \beta_0 + \beta_1 male_i + \beta_2 order_i + \beta_3 ln(y_i) + \beta_4 cig_i + u_i$$

where bw is the birth weight, male is a dummy variable assuming the value 1 if the baby is a boy and 0 otherwise, *order* is the birth order of the child, y is the income of the family and *cig* is the amount of cigarettes per day smoked during pregnancy.

(a) What could be the problem in using OLS to estimate the above model?

Solution:

Mothers who smoke during pregnancy might as well be less careful with other health issues that affect birth weight and that are not controlled for in the regression.

(b) Suppose you have data on the average price of cigarettes in the state of residence. Would this information help to identify the true parameters of the model?

Solution:

If people choose the state of residence independently from the price of cigarettes (which seems a plausible assumption), then the price of cigarettes should be uncorrelated with birth weight through ways other than the amount of cigarettes smoked. Moreover, the price of cigarettes should be correlated with the number of cigarettes smoked by the mother during pregnancy. Since both relevance and exogeneity assumptions, from a *a priori* point of view, hold, the price of cigarettes seems to be a nice instrument for *cigs*.

(c) Use data on *BirthWeight.dta* to estimate the model above. Use OLS and 2SLS. Discuss the results.

Solution:

Create the log variables *lbw* and *ly* and run the two regressions:

```
lbw<-log(bw)
ly<-log(y)

summary(lm(lbw~cig+male+order+ly))

Coefficients:  Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.675618   0.021881 213.681  < 2e-16
cig           -0.083728   0.017121  -4.890 1.12e-06
male           0.026241   0.010089   2.601  0.00940
order          0.014729   0.005665   2.600  0.00942
ly             0.018050   0.005584   3.233  0.00126

Residual standard error: 0.1876 on 1383 degrees of freedom
Multiple R-squared: 0.03504, Adjusted R-squared: 0.03225
F-statistic: 12.55 on 4 and 1383 DF,  p-value: 4.905e-10


summary(tsls(lbw~cig+male+order+ly, ~cigprice+male+order+ly,data=birthw))

2SLS Estimates
Model Formula: lbw ~ cig + male + order + ly
Instruments: ~cigprice + male + order + ly

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.467861    0.25883 17.2618  0.00000
cig          0.797106    1.08628  0.7338  0.46320
male         0.029821    0.01778  1.6773  0.09371
order       -0.001239    0.02193 -0.0565  0.95496
ly           0.063646    0.05701  1.1163  0.26447
Residual standard error: 0.3202 on 1383 degrees of freedom
```

The OLS results show that the amount of cigarettes smoked while pregnant significantly

reduces birth weight by 8%. However, the IV estimates show a non-significant effect for the same variable. If the IV estimates are correct, the impact seems to come from how smoking is related to other health issues or behaviours that affects birth weight, not through smoking directly. However, the IV estimates show a worrying feature: all coefficients become non significant (only the coefficient on *male* is significant at 10% level). IV estimates have generally a higher variance than the OLS ones, but maybe the problem is that the instrument is only weakly correlated with the endogenous explanatory variables. If this is so, the IV estimator becomes inconsistent and its variance becomes very large.

(d) Estimate the reduced form for *cig*. Discuss.

Solution:

The reduced form is nothing more than the first stage of the TSLS regression:

```
summary(lm(cig~cigprice+male+order+ly))
Call: lm(formula = cig ~ cigprice + male + order + ly)

Coefficients:      Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.1374075  0.1040005   1.321   0.1866
cigprice          0.0007770  0.0007763   1.001   0.3171
male             -0.0047261  0.0158539  -0.298   0.7657
order             0.0181491  0.0088802   2.044   0.0412
ly               -0.0526374  0.0086991  -6.051 1.85e-09

Residual standard error: 0.2945 on 1383 degrees of freedom
Multiple R-squared: 0.03045,Adjusted R-squared: 0.02765
F-statistic: 10.86 on 4 and 1383 DF,  p-value: 1.137e-08
```

What we predicted in the last question turns out to be true. The reduced for model for *cig* shows that the instrument *cigprice* is weak: the F-statistic is <10.

2. The data in fertil2_.dta includes, for women in Botswana during 1988, information on the number of children, years of education, age, and religious and economic status variables.

(a) Estimate this model using OLS:

$$children_i = \beta_0 + \beta_1 educ_i + \beta_2 age + \beta_3 age^2 + u_i$$

and interpret the estimates. In particular, holding age fixed, what is the estimated effect of another year of education on fertility? If 100 women receive another year of education, how many fewer children are they expected to have?

Solution:

First create the variable age^2, then run the regression:

```
age2=age^2
summary(lm(children~educ+age+age2))

Coefficients:         Estimate Std. Error t value Pr(>|t|)
(Intercept)        -4.1383066  0.2405942 -17.200   <2e-16
educ               -0.0905755  0.0059207 -15.298   <2e-16
age                 0.3324486  0.0165495  20.088   <2e-16
age2               -0.0026308  0.0002726  -9.651   <2e-16

Residual standard error: 1.46 on 4357 degrees of freedom
Multiple R-squared: 0.5687,Adjusted R-squared: 0.5684
F-statistic:  1915 on 3 and 4357 DF,  p-value: < 2.2e-16
```

Another year of education, holding age fixed, results in about .091 fewer children. In other words, for a group of 100 women, if each gets another of education, they collectively are predicted to have about nine children less.

(b) *Frsthalf* is a dummy variable equal to 1 if the woman was born during the first six months of the year. Assuming that *frsthalf* is uncorrelated with the error term, show that it is a reasonable IV candidate for *educ*. (Hint: run a regression.)

Solution:

The reduced form for *educ* is
$educ = \pi_0 + \pi_1 age + \pi_2 age2 + \pi_3 frsthalf + v_i$

```
summary(lm(educ~frsthalf+age+age2))

Coefficients:               Estimate Std. Error t value Pr(>|t|)
(Intercept)              9.6928643  0.5980686  16.207  < 2e-16
frsthalf                -0.8522854  0.1128296  -7.554 5.12e-14
```

```
age                     -0.1079504  0.0420402  -2.568   0.0103
age2                    -0.0005056  0.0006929  -0.730   0.4657

Residual standard error: 3.711 on 4357 degrees of freedom
Multiple R-squared: 0.1077, Adjusted R-squared: 0.107
F-statistic: 175.2 on 3 and 4357 DF,  p-value: < 2.2e-16
```

The F-statistic is greater than 10, so *frsthalf* is not a weak isntrument. Women born in the first half of the year are predicted to have almost one year less of education.

(c) Estimate the model from question (a) using *frsthalf* as an IV for *educ*. Compare the estimated effect of education with the OLS estimates from part (a).

Solution:

```
summary(tsls(children~educ+age+age2,~frsthalf+age+age2))

2SLS Estimates
Model Formula: children ~ educ + age + age2
Instruments: ~frsthalf + age + age2

                     Estimate Std. Error t value  Pr(>|t|)
(Intercept)          -3.387805  0.5481502  -6.180 6.978e-10
educ                 -0.171499  0.0531796  -3.225 1.269e-03
age                   0.323605  0.0178596  18.119 0.000e+00
age2                 -0.002672  0.0002797  -9.555 0.000e+00
Residual standard error: 1.4907 on 4357 degrees of freedom
```

The estimated effect of education on fertility is now much larger. Naturally, the standard errors for the IV estimate are also bigger (about nine times). This produces a fairly wide 95% CI for β_1.

(d) Add the binary variables *electric*, *tv* and *bicycle* to the model and assume these are exogenous. Estimate the equation by OLS and TSLS and compare the estimated coefficients on *educ*. Interpret the coefficient on *tv* and explain why television ownership has a negative effect on fertility.

Solution:

```
summary(lm(children~educ+age+age2+electric+tv+bicycle))

Coefficients:         Estimate Std. Error t value Pr(>|t|)
(Intercept)         -4.3897837  0.2403173 -18.267  < 2e-16
educ                -0.0767093  0.0063526 -12.075  < 2e-16
age                  0.3402038  0.0164417  20.692  < 2e-16
age2                -0.0027081  0.0002706 -10.010  < 2e-16
electric            -0.3027293  0.0761869  -3.974 7.20e-05
tv                  -0.2531443  0.0914374  -2.768  0.00566
bicycle              0.3178950  0.0493661   6.440 1.33e-10

Residual standard error: 1.448 on 4349 degrees of freedom
(5 observations deleted due to missingness)
Multiple R-squared: 0.5761, Adjusted R-squared: 0.5755
F-statistic: 984.9 on 6 and 4349 DF,  p-value: < 2.2e-16

summary(tsls(children~educ+age+age2++electric+tv+bicycle,
~frsthalf+age+age2+electric+tv+bicycle))

2SLS Estimates
Model Formula: children ~ educ + age + age2 + +electric + tv + bicycle
Instruments: ~frsthalf + age + age2 + electric + tv + bicycle

                    Estimate Std. Error  t value  Pr(>|t|)
(Intercept)        -3.591332  0.6450889 -5.56719 2.745e-08
educ               -0.163981  0.0655269 -2.50251 1.237e-02
age                 0.328145  0.0190587 17.21760 0.000e+00
age2               -0.002722  0.0002766 -9.84296 0.000e+00
electric           -0.106531  0.1659650 -0.64189 5.210e-01
tv                 -0.002555  0.2092301 -0.01221 9.903e-01
bicycle             0.332072  0.0515264  6.44470 1.284e-10

Residual standard error: 1.4789 on 4349 degrees of freedom
```

Adding *electric*, *tv* and *bycicle* to the model reduces the estimated effect of *educ* in both cases, but not too much. In OLS estimates, the coefficient on *tv* implies that, *ceteris paribus*, families that own a television will have about one child less than those without a TV. Television ownership can be a proxy for different things, including income and perhaps geographic location. A causal interpretation is that TV provides an alternative

form of recreation. Interestingly, the effect of TV ownership is practically and statistically non significant in the equation estimated by IV (even though we are not using an IV for *tv*). The coefficient on *electric* is also greatly reduced in magnitude in the IV estimation. The substantial drops in the absolute value of these coefficients suggest that a linear model might not be the right functional form, which is surprising since children is a count variable.

3. In this exercise you will study the effect of fertility on labor supply. Suppose you want to know how a woman's labor supply changes when she has an additional child. The dataset *fertility.dta* contains information on married women aged 21-35 with two or more children:

   - *morekids* =1 if woman had more than 2 children, 0 otherwise;
   - *samesex* =1 if the first two children have the same sex, 0 otherwise;
   - *age*: woman's age;
   - *black* =1 if woman is black, 0 otherwise;
   - *hispan* =1 if woman is hispanic, 0 otherwise;
   - *othrace* =1 if woman in not black, hispanic or white, 0 otherwise;
   - *weeks*: number of weeks woman has worked in 1979.

   (a) Regress weeks on the binary variable morekids using OLS. On average, do women with more than two children work less than women with two children? How much less?

   Solution:

```
summary(lm(weeks~morekids))

Coefficients:             Estimate Std. Error t value Pr(>|t|)
(Intercept)               21.4782     0.1592   135.0   <2e-16
morekids                  -6.0082     0.2590   -23.2   <2e-16

Residual standard error: 21.75 on 29998 degrees of freedom
Multiple R-squared: 0.01762, Adjusted R-squared: 0.01759
F-statistic: 538.2 on 1 and 29998 DF,  p-value: < 2.2e-16
```

   On average, women with more than two children work 6 weeks less than women with two children.

(b) Explain why OLS regression estimated in (a) is inappropriate for estimating the casual effect of fertility (morekids) on labor supply (weeks).

Solution:

The independent variable *morekids* might be correlated with other factors which affect labor supply, for example education. In this case, the OLS estimate will be biased and inconsistent.

(c) The dataset contains variable *samesex*, which is equal to 1 if the first two children are of the same sex (boy-boy or girl-girl) and equal to 0 otherwise. Are couples whose first two children are of the same sex more likely to have a third child? (Hint: use OLS to answer this question). Is the effect large? Is it statistically significant?

Solution:

```
summary(lm(morekids~samesex))

Coefficients:             Estimate Std. Error t value Pr(>|t|)
(Intercept)              0.343979   0.003962   86.83   <2e-16
samesex                  0.066820   0.005585   11.96   <2e-16

Residual standard error: 0.4837 on 29998 degrees of freedom
Multiple R-squared: 0.004749,Adjusted R-squared: 0.004716
F-statistic: 143.1 on 1 and 29998 DF,  p-value: < 2.2e-16
```

Regressing *morekids* on *samesex* we see that couples whose first two children are of the same sex are more likely to have a third child. The effect is not very large, but statistically significant (as you may see from the high t-statistic).

(d) Explain why samesex is a valid instrument for the instrumental variable regression of weeks on morekids, i.e. explain why the two necessary conditions (instrument relevance and instrument exogeneity) are likely to be satisfied.

Solution:

The variation in *samesex* is driven by genetic differences in the population and is not correlated with the factors which affect labor supply. Also, it is positively correlated with

morekids, as we have seen in (c). Therefore it is a valid instrument and can be used in the instrumental variable regression.

(e) Estimate the regression of *weeks* on *morekids* using *samesex* as an instrument. How large is the effect of fertility on labor supply? Is it statistically significant? Add age and race dummies to this IV regression. Does inclusion of these new variables change the effect of *morekids* on *weeks*? Compare your results to those obtained in (a).

Solution:

The commands are:

```
summary(tsls(weeks~morekids, ~samesex))
summary(tsls(weeks~morekids+age+black+hispan+othrace,
~samesex+age+black+hispan+othrace))
```