UNIVERSITY OF AMSTERDAM

Applied (Financial) Econometrics
Computer exercise 2: Instrumental Variables

Part 0 – Open Stata, and make your own do-file

- Start Stata through the start menu button
- In de white command window type `doedit` to start de do-file editor. Place the Stata screen on the left and the do file editor on the right such that you can easily switch between the two.
- Save the empty do-file as a new do-file under an applicable name such as `ectrcs_w2.do` in a directory that you want to use for this course, for example `H:\ectrcs`
- In the first two lines of the do-file type

```
cls                             //this clears the screen
clear all                       //this clears the memory
cd "H:/ectrcs"                  //this is your path
```

- If you want to use your own computer type instead:

```
cd "~/ectrcs"                   //this is your MAC path
cd "c:/ectrcs"                  //this is your PC path
```

Part 1 – Macroeconomic equilibrium in income and consumption

This computer exercise is about the effects of endogenous regressors on OLS and instrumental variables. In the first part of this exercise we are going to conduct a simulation experiment in order to learn about least squares and instrumental variables in case of endogeneity. You will create your own artificial data, hence the population regression model is known and the accuracy of both estimation methods can be compared.

We take a very simple version of a Keynesian model as starting point for generating data. The population regression model consists of two equations:

$$\begin{aligned} C_i &= \beta_0 + \beta_1 Y_i + u_i \\ Y_i &= C_i + I_i \end{aligned}$$

The first equation is a consumption function describing how consumption (C) varies with income (Y). Parameter of interest is the marginal propensity to consume β_1. The second equation is an identity describing that income is determined by adding up consumption and investments (I). The model above is an example of a simultaneous equations model. In such a model we distinguish endogenous (C and Y) from exogenous (I and the constant term) variables. It is clear that when we want to estimate the consumption function we are faced with an endogeneity issue. From the second equation we learn that C influences Y, hence it can no longer be argued that in the first equation Y is uncorrelated with the error term. This simultaneous causality will have detrimental effects on the OLS estimator of our parameter of interest β_1 (and also β_0).

To generate artificial data from the model above we rewrite it as:

$$C_i = \frac{\beta_0}{1-\beta_1} + \frac{\beta_1}{1-\beta_1} I_i + \frac{1}{1-\beta_1} u_i$$

$$Y_i = \frac{\beta_0}{1-\beta_1} + \frac{1}{1-\beta_1} I_i + \frac{1}{1-\beta_1} u_i$$

Now endogenous variables have been expressed solely in terms of exogenous variables, coefficients and error terms. We call this representation the reduced form, while the original structure is labeled structural form. Two things become clear from the reduced form: (1) Y is correlated with u; (2) I is a valid instrument for Y. We will now use the reduced form of the model to generate our artificial data.

1. Open Stata. Type `set seed 4444`,[1] and then `set obs 100`. We are going to generate our own data file.
2. We are going to generate artificial data in Stata by assuming that both I and u come from standard normal distributions, hence `gen u=invnorm(uniform())` and `gen i=invnorm(uniform())`. Next generate C and Y according to the reduced form equations above. We will take $\beta_0 = 5$ and $\beta_1 = 0.5$ as values for the structural parameters, hence the reduced form is $C = 10 + I + 2u$ and $Y = 10 + 2I + 2u$ (so type `gen c=10+i+2*u` and `gen y=10+2*i+2*u`)
3. It is clear that C and Y are endogenous, hence OLS in the consumption equation will be biased and inconsistent as Y is correlated with u. Verify this by `reg c y, robust` and comparing the true and estimated parameters.
4. Also the t test will be biased and inconsistent due to the endogeneity of Y. Using a significance level of 5% test the null hypothesis that the coefficient of Y is equal to 0.5 (its true value). Do you reject? Given the results above do you expect that on average 1 out of 20 students rejects?
5. It can be shown that (for this model) the probability limit of the OLS estimator of $\hat{\beta}_1 \xrightarrow{p} \beta_1 + (1-\beta_1)\frac{var(u_i)}{var(I_i)+var(u_i)}$. Substituting the true values for β_1 and the two variances we find that the probability limit is equal to 0.75, hence OLS will be centered around this value instead of the true value 0.5. Repeat questions 2, 3, and 4 several times (first type `drop _all`, then `set obs 200`, and then draw new random numbers for I and u, generate C and Y according to the reduced form, and estimate β_0 and β_1 with OLS; You can use PAGEUP, to scroll back to previous commands) to verify this.
6. In addition, use also n=1000 to get a better understanding of the persistence of the bias and inconsistency of the OLS estimator and corresponding t-test. First, type `clear` to remove data from memory. Next, repeat questions 1 to 4, but now with `set obs 1000` in question 1.
7. IV/TSLS using I as instrument for Y will be consistent as I is uncorrelated with u and correlated with Y. Type `ivregress 2sls c (y=i), robust`. Compare true and estimated parameters.
8. The t-test based on TSLS estimates will be consistent. Using a significance level of 5% test the null hypothesis that the coefficient of Y is equal to 0.5 (its true value). Do you

[1] When Stata starts, it chooses a random seed. If you want Stata to generate the same random numbers in different sessions, you can tell it to start at a given seed.

reject? Given the results above do you expect that still, on average, 1 out of 20 students rejects?

9. Perform TSLS now manually by first `reg y i, robust`, then `predict yfit`, and, finally, applying OLS to the second stage regression by `reg c yfit, robust`. Your estimate should be identical to that of question 7. Note that the reported estimated standard errors, however, of this manual procedure are wrong.
10. End this part of the exercise by `clear`. This removes all data from memory.

Part 2 – The returns to schooling in the Netherlands

In the second part of this exercise you are going to analyze a controversial issue in the empirical literature on returns to schooling, i.e. the possible endogeneity of schooling with respect to wages. One often mentioned argument is that of self-selection. Individuals who believe to have relatively high returns to schooling and expect to earn a high wage will choose for more schooling. Hence, the level of schooling is not exogenous and may depend on wages.

11. On the Blackboard site of this course you can download a Stata file named `w2_brabant.dta`. Make sure you put it in your project folder `H:\ectrcs`. After you downloaded it, open it in Stata by typing `use w2_brabant.dta, replace`. For a description of the data see the file named `w2_brabant.pdf`.
12. Estimate with OLS the Mincerian wage equation, i.e. `reg lwage educ c.lexp##c.lexp, robust`, and interpret the coefficient on `educ` (Note: the `c.` is needed to tell stata `lexp` is continuous and the option `robust` generates heteroscedastisity robust variance estimates). Type `est sto ols` to store the results such that we can use them later in the session.
13. Estimate the Mincerian wage equation also with TSLS by `ivregress 2sls lwage c.lexp##c.lexp (educ = faed mark ssoc fhigh fint fself), robust`. This Stata code implies that for the endogenous variable `educ` the instruments `faed`, `mark`, `ssoc`, `fhigh`, `fint`, and `fself` have been used. Why is it not possible to include `flow` as well as an instrument? Do you think it is reasonable to assume that these instruments are exogenous (i.e. are only correlated with children's wages via their schooling)? Type `est sto iv` to store the results such that we can use them later in the session.
14. The TSLS consists of two stages both estimated by OLS. In the first stage, the endogenous variable schooling is regressed on all instruments. Perform this regression, i.e. `regress educ faed mark ssoc fhigh fint fself c.lexp##c.lexp, robust`. Test the weakness of instruments by judging the appropriate *F*-statistic from this regression (*Hint*: with `testparm` you can quickly perform a joint test on multiple variables). Do you think the instruments capture enough variation of the endogenous variable `educ` to circumvent the weak instrument problem?
15. To perform the first stage *F*-test test quickly, Stata has a built-in postestimation command for `ivregress`. Type `est restore iv` to restore the IV estimates, and

then `estat firststage`. Verify that the F-statistic that you obtained manually is correct. Also note that the instruments explain about 19% of the variation in `educ`.

16. When writing a paper it is informative to show how OLS and IV results compare, and discuss their differences. A nicer tabulation package than stata's standard `est tab` command can be obtained by `ssc install estout, replace`. Then you can display the results by `esttab ols iv, b(%6.3f) se`. Does the OLS estimate seem to be biased, assuming that the instruments are valid? Can you explain the direction of the bias? Unfortunately, the difference between both `educ` point estimates cannot be tested statistically using the table. Stata has a built in command, however, to test whether OLS is rejected assuming the IV estimates are consistent. Type `estat endogenous` to get the p-values for this test (use the second one, which is based on a regression based test developed by Wooldridge (1995) that also allows for clustering within groups). What do you conclude?

    Comparison of the OLS and IV columns from the estimates table also shows that estimated SE of IV is much larger. This loss in precision stems from the fact that the instruments only *partially* explain the variation in `educ` (i.e. 19%). In large samples the IV standard error will be inflated by $\frac{1}{\sqrt{R^2_{partial}}}$, which is 2.3 in this case. The IV s.e. increases further in small samples due to the estimation error of the first stage, leading to a total increase by a factor 3 in this dataset (0.0054 → 0.0162).

17. Stata also provides a simple command to perform the overidentification J-test (see p449[485] of S&W). Type `estat overid` to get Woodridge's (1995) robust test statistic, which is also valid under heteroscedastic errors (as opposed to the J-test in Key Concept 12.6, which is valid under homoscedastic errors only). The test statistic follows a χ^2 distribution with $m - k = 6 - 1 = 5$ degrees of freedom, and we see that it is rejected at the 10% level.

    The inquisitive student should realize, however, that rejection of the J-test is not always informative. It tells us that there is statistical evidence to believe that the instruments contradict each other. We still do not know which of the instruments is invalid, however, and it could even be that they are all valid but contradict because they estimate different causal effects due to non-linearity [not discussed] or heterogeneity [Lecture 2]. Also, even when we do not reject the J-test, all the instruments may still be wrong but simply not contradict. For these reasons outcomes from the J-test should be taken with care in applied economic research.

References

Wooldridge, J. M. 1995. Score diagnostics for linear models estimated by two stage least squares. In Advances in Econometrics and Quantitative Economics: Essays in Honor of Professor C. R. Rao, ed. G. S. Maddala, P. C. B. Phillips, and T. N. Srinivasan, 66–87. Oxford: Blackwell