Notes on Instrumental Variables (IV) for data calibration project

Paige Miller

February 22, 2016

1 Instrumental Variables

Method for estimation that is used in statistics, econometrics, and epidemiology when correlations are suspected between explanatory variables and the error term (due to omitted variables, measurement error, or other sources of simultaneity bias)

2 Context

- The basic situation is when we have $y_i = x_i'\beta + u_i$ for i = 1,..., N where x_i are explanatory variables that are correlated with u_i which are error terms and β is the parameter we wish to estimate
- The method of IV seeks to replace the actual realized values of x_i (which are correlated with the error terms) by predicted values of x_i that are necessarily (1) related to the actual x_i and (2) uncorrelated with u_i
- Doing this allows us to obtain a consistent (i.e. with large enough number of samples, our estimate converges to the true population estimate) estimator of β

3 Requirements of IV

- related to the explanatory variable(s) – they are 'informative'
- uncorrelated with errors – they are 'valid'
- **biggest problem** with using instrumental vars is that they have to have *both* of these properties
- Note that sometimes (in multiple regression setting) some explanatory vars may be 'endogenous' or correlated with error terms while others are 'exogenous' or uncorrelated with error terms but all instruments are required

to have explanatory power for each endogenous variable *after* conditioning on all remaining exogenous explanatory variables

4 Example case: Single endogenous (x_i), and single IV (z_i)

- Assume both x_i and z_i have mean = 0 for simplicity and all vectors are Nx1
- Then for i=1,...N

$$\begin{aligned} E(u_i) = 0, E(x_i u_i) \neq 0 \\ y_i = x_i\beta + u_i \\ y = X\beta + u \end{aligned} \tag{1}$$

- The first stage regression (linear projection with IV) for i=1,...,N and all vectors are Nx1

$$\begin{aligned} x_i = z_i\pi + \gamma_i \\ X = Z\pi + \gamma \end{aligned} \tag{2}$$

- where E($z_i x_i$)=0 AND $\pi \neq 0$ so that the IV, z_i, is valid and informative, respectively

5 Controversy of Ordinary Least Squares (OLS) vs. Two Stage Least Squares (2SLS)

- OLS: goal is to minimize the differences between observed responses in data and responses predicted by the linear approximation of the data
- 2SLS: "regression analysis technique that is used in the analysis of structural equations. This technique is the extension of the OLS method" - statsolutions.com.
- motivation for using instrumental variables rather than OLS estimation is that we suspect the OLS estimates would be biased and inconsistent as a result of correlation between the error term and one or more of the explanatory variable(s)
- If our instrumental variables are both valid (i.e. uncorrelated with the error term) and informative (i.e. satisfying the requirements for identication), then we should expect the 2SLS parameter estimates to be quite different from the OLS parameter estimates
- Found some material on comparing the two estimates using a Hausman test. Will read more into this later.

6 Problems with IV for finite samples

- Overfitting: describes the situation where researcher may include too many IVs, resulting in inflated R^2 values
- Weak IVs: describes the situation where researcher uses "weak" instruments (ie they are poor predictors of explanatory vars). This problem could result in bias towards β_{OLS} if instrument is strictly valid or large inconsistency in β_{IV}
- but insturment strength can be directly assesed because endogenous covariates and instruments are observable

7 Questions

- What is the relation of these ideas to the case of SDM?
- What problems does this tackle for SDM?
- What questions can we answer using these ideas?
- How could this be a 'pre-processing' step for all data used in SDMs?

References

[1] http://www.nuff.ox.ac.uk/teaching/economics/bond/instrumental

[2] The latest version of the ivreg2 command in Stata now implements a number of these suggestions

[3] $https://en.wikipedia.org/wiki/Instrumental_variable$