Causality in Economics

Topics on Instrumental Variable Regression Techniques

Alex Klein

April 2016

Motivation I

- Causality - a crucial issue in economics (maybe more than in other social sciences)
- **Non-experimental** nature of data as opposed to experiments such as laboratory experiments or randomized controlled trials
- Estimation techniques developed over the past 70 or so years to estimate a causal effect of variables on the outcome of interest
- Development of *'instrumental variable estimation techniques'* is an attempt to account for causality in non-experimental data

Basic set-up I

- Consider a basic regression $y = x_i \beta_i + u, i = 1, ..., K$
- *Key condition of consistency* of OLS estimator is that the error term is uncorrelated with each of the regressors: $cov\left(x_i, u\right) = 0, i = 1, ..., K$
- Sufficient condition for $cov\left(x_i, u\right) = 0$ is $E\left(u|x_i\right) = 0$
- An explanatory variable is endogenous if it is correlated with the error term which is caused by
  1. omitted variables
  2. measurement error
  3. simultaneity

Basic set-up I

- $\widehat{\beta_{OLS}} = (X'X)^{-1} X'y = (X'X)^{-1} X'X\beta + (X'X)^{-1} X'u = \beta + (X'X)^{-1} X'u$

- $\widehat{\beta_{OLS}} = \beta + \left(N^{-1}X'X\right)^{-1} N^{-1}X'u$ - renormalization to allow the use of large numbers to be applied to $X'X$
- plim $\widehat{\beta_{OLS}} = \beta + \left(\text{p}\lim N^{-1}X'X\right)^{-1} (\text{plim}\, N^{-1}X'u)$ (Slutsky's

  theorem)

$$\text{OLS is consistent if plim}\, N^{-1}X'u = 0$$

- a necessary condition for the above equality to hold is that $E\left[X'u\right] = 0$

Instrumental Variable Regression I

- To obtain consistent estimates of β when $cov\left(x_i, u\right) \neq 0$, we need to find a variable - call it z_i - which satisfies two conditions:

1. Instrument relevance: $cov\left(z_i, x_i\right) \neq 0$
2. Instrument exogeneity: $cov\left(z_i, u\right) = 0$

- failure of the first condition leads to weak instrumental variable problem, but we can deal with it (somehow)
- failure of the second condition is fatal and **we can't interpret the estimated relationship as causal** (only as a sophisticated correlation)

Instrumental Variable Regression I

- we will deal with a single equation model
- number of instruments can be the same as the number of endogenous variables (*just-identified* model) or larger (*overidentified* model)
- just-identified model:

$$\widehat{\beta_{IV}} = (Z'X)^{-1} Z'y = \beta + (Z'X)^{-1} Z'u = \left(N^{-1}Z'X\right)^{-1} N^{-1}Z'u$$

- consistency of IV estimator requires plim $N^{-1}Z'u = 0$ and plim $N^{-1}Z'X \neq 0$
- variance of $\widehat{\beta_{IV}}$: $\widehat{V}(\widehat{\beta_{IV}}) = (Z'X)^{-1}Z'\widehat{\Omega}Z(Z'X)^{-1}$ where $\widehat{\Omega} = Diag(\widehat{u_i}^2)$
- though consistent, IV estimators exhibit efficiency loss

Instrumental Variable Regression I

- over-identified model requires Two-Stage Least Square estimator (TSLS/2SLS)

$$\widehat{\beta_{2SLS}} = [X'Z(Z'Z)^{-1}Z'X]^{-1}[X'Z(Z'Z)^{-1}Z'y]$$

- in just-identified model *2SLS=IV*
- **Stage 1:** obtain predicted values of X from a regression of X on Z: $\widehat{X} = Z(Z'Z)^{-1}Z'X$
- **Stage 2:** run OLS with predicted values $\widehat{X}$
- again, 2SLS causes efficiency loss relative to OLS, but, it is effecienct estimator in the class of all instrumental variable estimators using instrument *linear* in z

Instrumental Variable Regression I

- Even though 2SLS is a consistent estimator when instruments satisfy the conditions of relevance and exogeneity, it is biased in finite samples
- In fact, we must rely on large sample analysis to derive the properties of 2SLS (mean of just-identified 2SLS does not even exist)
- When instruments are weak, 2SLS is biased even in very large sample
- Consider the 'degree of inconsistency' - there is some, though very mild, correlation between instruments are error terms
- When instruments are weak, the degree of inconsistency increases

Instrumental Variable Regression I

- consider a simple model with one endogenous variable: $Y_{1i} = \alpha_1 + \beta_1 Y_{2i} + \epsilon_i$ and $Y_{2i} = \alpha_2 + \beta_2 Z_i + \mu_i$
- assume that Var(ϵ_i)=1 and Var(μ_i)=1 =>cov(ϵ_i, μ_i)=ρ where ρ is the correlation coefficient
- if we assume that Z$_i$ is exogenous, then ρ measures the degree to which y_{2i} is correlated with ϵ_i
- Hahn and Hausman (2005) showed that in this simple case, the finite sample bias of 2SLS in overidentified case is, to a second-degree approximation

$$E(\beta_1^{2SLS}) - \beta_1 \approx \frac{l\rho(1-\widetilde{R}^2)}{n\widetilde{R}^2}$$

- l is the number of instruments, n is sample size, $\widetilde{R}^2$ is R^2 from the regression of Z$_i$ on Y$_{2i}$ and measures the strength of instruments

Instrumental Variable Regression I

- the bias of 2SLS in finite samples is toward inconsistent OLS
- a fundamental question arises: if a consistent 2SLS estimator is biased in finite samples toward inconsistent OLS, is 2SLS bias smaller or larger then that of OLS?
- Hahn and Hausman (2005) offer the following equation

$$\frac{Bias(\beta_1^{2SLS})}{Bias(\beta_1^{OLS})} \approx \frac{l}{n\widetilde{R}^2}$$

- as long as the denominator is larger than the nominator, 2SLS bias is smaller than OLS bias
- ceteris paribus, the bias of 2SLS grows with the number of instruments
- weak instruments (low $\widetilde{R}^2$) **increase the bias of 2SLS toward inconsistent OLS!!!**

Instrumental Variable Regression I

- weak instruments and 'mild inconsistency':

$$\text{plim } \widehat{\beta_{IV}} = \beta + \frac{cov(Z,u)}{cov(Z,X)} = \frac{\sigma_u}{\sigma_u}\left[\frac{corr(Z,u)}{corr(Z,X)}\right]$$

- relative inconsistency of 2SLS

$$\frac{\text{plim}\widehat{\beta_{2SLS}} - \beta}{\text{plim}\widehat{\beta OLS} - \beta} = \frac{corr(\widehat{X},u)}{corr(X,u)}\frac{1}{R_p^2}$$

- if instruments are weak **and** moderately correlated with error term (mildly endogenous), instrumental variable estimator *is even more inconsistent than OLS*

Instrumental Variable Regression I

- unless we have a perfect natural experiment of a perfectly exogenous instrument, weak instrument is more fatal than running a simple OLS even when a correlation between instrument and error term is very small
- this result is due to Bound, Jaeger and Baker (1995) and has not received much attention in the literature
- literature on weak instruments assumes that instruments satisfy exogeneity assumption and the only problem is their weak correlation with endogenous variables

Weak Instruments I

- how to detect it:
  1. Shea's partial R^2 from the first stage regression
  2. F-statistics from the first stage regression
- logic of R^2 from the first stage regressions: consider $y = \beta_1 x_1 + \beta_2 x_2 + u$ where x_1 is endogenous and x_2 exogenous, and let z be a vector of instruments (includes x_2)
- we need a measure of the correlation between z and x_1 which purges out x_2
- R^2measure adjusted for the presence of x_2 proposed by Bound, Jaeger, and Baker (1995)
- R^2measure adjusted for the presence of x_2 and another endogenous variables proposed Shea (1997)

Weak Instruments I

- F-statistics from the first-stage regression; the test statistics are **not** drawn from the standard F-distribution
- Stock and Yogo (2005) offer critical values which depend on the number of instruments and endogenous variables
- Null hypothesis: the bias in 2SLS is less than **some percentage of the bias of OLS**
- for example, for one endogenous variable and three instruments, and H0 stating the bias being less than 10%, the critical value of F-statistic is 9.08

Weak Instruments - Solution(s) I

- alternative estimators to 2SLS which exhibit better properties in the presence of weak instruments
- test statistics which are robust to weak-instrument problem

An Example - Housing Expenditures I

- the model allows for household fixed effects

$$
\begin{aligned}
d_{it} &= 1\left(\pi' x_{it} + \eta_i - u_{it} \geq 0\right) \\
y_{0it} &= \beta_0' x_{it} + \alpha_{0i} + \varepsilon_{0it} \text{ if } d_{it} = 0 \\
y_{1it} &= \beta_1' x_{it} + \alpha_{1i} + \varepsilon_{1it} \text{ if } d_{it} = 1
\end{aligned}
$$

- the selection variable d_{it} is a choice between owning a property $(d_{it} = 1)$ and renting a property $(d_{it} = 0)$
- x_{it} is a vector of explanatory variables (total expenditures, square of total expenditures, prices, household characteristics)
- y_{1it} and y_{0it} are budget shares spent on housing for renters and owners respectively
- α_{0i}, α_{1i}, η_i are unobservable household specific time-invariant effects

An Example - Housing Expenditures I

- x_i is decomposed into x_{ai} (log of total expend, square of total expend), x_{bi} (log of hh income, square of hh income), x_{di} (prices, hh characteristics), x_{ci} are exclusion restrictions
- selection equation includes x_{bi} and x_{di}, the budget equation x_{ai} and x_{ci}
- taking the difference between period t and τ yields:

$$y_{pit} - y_{pi\tau} = \beta'_{pa} \left(x_{ait} - x_{ai\tau} \right) + \beta'_{pc} \left(x_{cit} - x_{ci\tau} \right) + \left(\varepsilon_{pit} - \varepsilon_{pi\tau} \right) \text{ if}$$
$$d_{it} = d_{i\tau} = p, \text{ p=0,1}$$
$$d_{is} = 1 \left(\pi'_b x_{bit} + \pi'_d x_{dit} + \eta_i - u_{it} \geq 0 \right), \text{ s= } t, \tau$$

An Example - Housing Expenditures I

- we can rewrite the above equation as

$$y_{pit} - y_{pi\tau} = \beta'_{pa}\left(x_{ait} - x_{ai\tau}\right) + \beta'_{pc}\left(x_{cit} - x_{ci\tau}\right) + \\ g_{pt\tau}\left(x_{bit}, x_{bi\tau}, x_{dit}, x_{di\tau}\right) + \widetilde{\varepsilon}_{pit\tau}$$

- the function $g_{pt\tau}, p = 0, 1$ is given by

$$g_{pt\tau}\left(x_{bit}, x_{bi\tau}, x_{dit}, x_{di\tau}\right) = \\ E\left(\varepsilon_{pit} - \varepsilon_{pi\tau} | x_{bit}, x_{bi\tau}, x_{dit}, x_{di\tau}, d_{it} = d_{is} = p\right)$$

- and $\widetilde{\varepsilon}_{pit\tau}$ satisfies

$$E\left(\widetilde{\varepsilon}_{pit\tau} | x_{bit}, x_{bi\tau}, x_{dit}, x_{di\tau}, d_{it} = d_{is} = p\right) = 0, p = 0, 1$$

An Example - Housing Expenditures I

- we can assume no sample selection after differencing $=>$ $g_{pt\tau} = 0, p = 0, 1$ which is equivalent to saying that $\eta_i - u_{it}$ is independent of ε_{0it} and ε_{1it} for all t
- in other words, possible selection effect on budget shares operate only through correlation between α_i and (η_i, u_{it})