IV and Control Function Approaches

Econ 671

Purdue University

Linear Projection

Let's start out with a little detour to review linear projection.
Write (consider a representative observation, rather than the equation stacked over all observations)

$$y = \beta_0 + x\beta_1 + \nu,$$

where x is $1 \times k$ and

-

It follows that

$$E(\nu) = E(y) - \beta_0 - E(x)\beta_1 = E(y) - [E(y) - E(x)\beta_1] - E(x)\beta_1 = 0.$$

In addition, note (since ν is mean-zero):

-

Linear Projection

where $(x - E(x))$ is, again, $1 \times k$ so that

$$[x - E(x)]'[y - E(y)] = [x - E(x)]'[x - E(x)]\beta_1 + [x - E(x)]'\nu.$$

Taking expectations through, and noting again that ν is mean-zero, we obtain:

-

and given the definition of β_1 above, we obtain

-

Note that this decomposition is really definitional; however, this does not imply that ν and x are independent or that $E(\nu|x) = 0$. The linear projection should not be confused with the conditional expectation; when we assume $E(\nu|x) = 0$, we are assuming that the conditional expectation function is linear (and would, then, coincide with the linear projection).

2SLS Revisited

Consider a general regression model with an endogeneity problem:

-

where y_2 is considered endogenous, but Z_1 is exogneous, in a sense to be defined formally below. Here, we suppose that y_2 is a scalar random variable, but it need not be, and the arguments that follow generalize to the case of multiple endogenous variables (assuming the model is identified).
A set of instruments Z, (which is $n \times l$, and includes Z_1 as a strict subset) is available, and are assumed to satisfy the orthogonality condition:

-

Implementation of the 2SLS estimator first requires fitted values for y_2. To obtain these, we consider the first-stage model (which we think about as a linear projection of y_2 onto Z):

-

2SLS Revisited

leading to

$$\hat{y}_2 = Z\hat{\pi} = P_Z y_2, \quad \text{where} \quad P_Z \equiv Z(Z'Z)^{-1}Z'.$$

It follows that the 2SLS estimator, defined as a regression of y_1 on $[Z_1 \ \hat{y}_2]$, is calculated as:

-

noting that P_Z is symmetric and idempotent.

Control Functions

Now, consider an alternate way of estimating these parameters. To this end, first write the linear projection of u onto ν as:

-

We can then substitute this equation into our regression equation of interest to obtain:

-

This looks like a multiple regression where the η term is uncorrelated with the included regressors, since

-

where the last line follows since $E(Z'u) = 0$ (valid IVs) and $E(Z'\nu) = 0$ (linear projection).

Control Functions

So, it seems as if we could estimate β_1, β_2 and ρ from a regression of y_1 on Z_1, y_2 and ν. However, ν is unknown.
An intuitive idea is to replace ν with the estimated residuals from the first-stage:

-

We can then estimate the regression:

-

and expect that the added covariate, $\hat{\nu}$ will control for or correct the endogeneity problem.

Control Functions

We can take this further and investigate the resulting estimator (to be called the "control function" estimator) in detail. First, let

$$X_1 \equiv [Z_1 \ y_2] \quad \text{and} \quad X \equiv [X_1 \ \hat{v}] = [Z_1 \ y_2 \ M_Z y_2].$$

The matrix X then denotes the full covariate matrix, and we obtain

-

where "CF" is an abbreviation for "Control Function."

Control Functions

We are interested in the first two components of this vector in particular, to see how their estimated values compare to the 2SLS estimator. We can again use our Frisch-Waugh-Lovell / "Short vs. Long" / Partitioned Inverse result to obtain:

-

where

$$\begin{aligned} M &\equiv I_n - M_Z y_2 (y_2' M_Z' M_Z y_2)^{-1} (M_Z y_2)' \\ &= I_n - M_Z y_2 (y_2' M_Z y_2)^{-1} y_2' M_Z \end{aligned}$$

Control Functions

So, let's now take a closer look at the formula for the CF estimator. First, note:

$$X_1'M = \begin{bmatrix} Z_1' \\ y_2' \end{bmatrix} \left(I_n - M_Z y_2 (y_2' M_Z y_2)^{-1} y_2' M_Z\right).$$

Since $Z_1' M_Z = 0$ (because $Z' M_Z = 0$, and Z_1 is a subset of Z) This becomes:

$$\begin{aligned} X_1'M &= \begin{bmatrix} Z_1' \left(I_n - M_Z y_2 (y_2' M_Z y_2)^{-1} y_2' M_Z\right) \\ y_2' \left(I_n - M_Z y_2 (y_2' M_Z y_2)^{-1} y_2' M_Z\right) \end{bmatrix} \\ &= \begin{bmatrix} Z_1' \\ y_2' - y_2' M_Z y_2 (y_2' M_Z y_2)^{-1} y_2' M_Z \end{bmatrix} \\ &= \begin{bmatrix} Z_1' \\ y_2'(I_n - M_Z) \end{bmatrix} \\ &= \begin{bmatrix} Z_1' \\ y_2' P_Z \end{bmatrix}. \end{aligned}$$

Control Functions

Therefore,

-

which we recognize as the same inverse term involved in the 2SLS calculation.

Control Functions

Similarly,

-

which is, again, the same as 2SLS.

So, what we have shown is an equivalent way to calculate the 2SLS estimator - the control function approach.

However, note that the usual OLS standard errors will not be correct, as they will not correct for the fact that a regressor has been estimated / generated. You can calculate the 2SLS standard errors directly, use the bootstrap in this case (672) or apply a formal correction [e.g., Murphy and Topel (1985, *JBES*)].