## Economics 696F, Causal Inference and Program Evaluation

## Lecture Note 9: Instrumental Variables Part II

### Constant treatment effects and linear IV model

Next, we want to look at the case where treatments can be multivalued or continuous. Suppose that treatment $T_i$ can take values in some set $\mathcal{T} \subset \mathbb{R}$. Let the potential outcomes be:

$$Y_i(t), \quad t \in \mathcal{T}.$$

Let's assume that

$$Y_i(t) = Y_i(0) + t \cdot \alpha.$$

(Note: a treatment level of $t = 0$ might not be relevant; what is key is that the slope is constant across treatment levels and across individuals.) Then we can write

$$Y_i = \gamma + T_i\alpha + U_i,$$

where $\gamma := E[Y_i(0)]$ and $U_i := Y_i(0) - \gamma$.

We might further extend this to allow for conditioning variables $W_i$:

$$Y_i = \gamma + T_i\alpha + W_i'\eta + U_i.$$

Here, we will assume that $E[W_iU_i] = 0$.

To simplify the notation, define $X_i := (1, T_i, W_i')$, $\theta := (\gamma, \alpha, \eta')'$, so we can write

$$Y_i = X_i'\theta + U_i.$$

Since $T_i$ is not assumed to be independent of the potential outcomes, we expect that $T_i$ and $U_i$ will be correlated, and hence

$$E[X_iU_i] \neq 0,$$

so that least squares will not consistently estimate $\theta$.

Suppose we have an instrumental variable $\zeta_i$ such that $E[\zeta_iU_i] = 0$. Then, if we define

$$Z_i := (1, \zeta_i, W_i')',$$

we have the orthogonality condition

$$E[Z_iU_i] = E[Z_i(Y_i - X_i'\theta)] = 0.$$

Notice that this is actually a set of orthogonality conditions:

$$\begin{aligned} E[(Y_i - X_i'\theta)] &= 0 \\ E[\zeta_i(Y_i - X_i'\theta)] &= 0 \\ E[W_{i1}(Y_i - X_i'\theta)] &= 0 \\ &\vdots \\ E[W_{ik}(Y_i - X_i'\theta)] &= 0 \end{aligned}$$

A natural estimator takes the sample version of this orthogonality condition and tries to choose an estimate $\hat{\theta}$ to solve it:

$$\frac{1}{n}\sum_{i=1}^{n} Z_i(Y_i - X_i'\hat{\theta}) = 0.$$

We can write

$$\frac{1}{n}\sum_{i=1}^{n} Z_i(Y_i - X_i'\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} Z_iY_i - \left[\frac{1}{n}\sum_{i=1}^{n} Z_iX_i'\right]\hat{\theta} = 0,$$

so that, if the inverse exists,

$$\hat{\theta} = \left[\sum_{i=1}^{n} Z_iX_i'\right]^{-1}\sum_{i=1}^{n} Z_iY_i.$$

This is just the usual linear IV estimator, in the "just-identified" case.

### Brief aside on asymptotic properties of linear IV estimators

note: I am switching notation a bit here - be a little careful.

The IV estimator: Exactly Identified Case

Consider a general equation of the form

$$y_i = x_i'\beta + u_i. \tag{1}$$

Here $x_i$ is a vector of explanatory variables. The "disturbance" $u_i$, however, is not necessarily assumed to satisfy $E(x_iu_i) = 0$. Let the dimension of $x_i$ be $K \times 1$. Loosely speaking, an *instrumental variable* (or simply *instrument*) is any variable that is orthogonal to the structural disturbance $u_i$, and correlated with $x_i$. Let us suppose that we have a $K$-vector of instruments $z_i = (z_{i1}, \ldots, z_{iK})'$ which satisfy $E(z_iu_i) = 0$, and for which $E(z_ix_i')$ has full rank. Some or all of the elements of $x_i$ could be valid instruments and thus included in $z_i$.

Write out the orthogonality condition as

$$E\left(z_i(y_i - x_i'\beta)\right) = 0.$$

This can be regarded as a set of $K$ equations in $\beta$:

$$\begin{aligned} E(z_{i1}(y_i - x_i'\beta)) &= 0 \\ &\vdots \\ E(z_{iK}(y_i - x_i'\beta)) &= 0 \end{aligned}$$

So if we knew the joint distribution of $(x_i, z_i, y_i)$, we could form these expectations and solve for $\beta$. We do not know the joint distribution, but we can apply the sample analogy principle, and replace the population expectation with

$$\frac{1}{n}\sum_{i=1}^{n} z_i(y_i - x_i'\beta) = \frac{1}{n}Z'(y - X\beta),$$

where

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad Z = \begin{pmatrix} z_1' \\ \vdots \\ z_n' \end{pmatrix} \qquad X = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix}.$$

To get our estimate for the unknown $\beta$, we will try to set this sample expectation equal to 0:

$$\frac{1}{n}Z'(y - X\hat{\beta}) = 0$$

The solution is given by

$$\hat{\beta} = (Z'X)^{-1}Z'y$$

This requires that $(Z'X)$ be square and nonsingular.

**Example: OLS**

Suppose that the standard classical regression model assumptions hold, so that $E(x_iu_i) = 0$. Then we can simply set $z_i = x_i$. This leads to

$$\hat{\beta} = (X'X)^{-1}X'y$$

which amounts to running a least squares regression of $y$ on $X$. So OLS is a special case of IV.

Returning to the more general IV setup, let us consider the asymptotic properties of the IV estimator. We could just obtain these as a special case of the general M-estimator, but it is instructive to derive its asymptotic properties more directly.

To show consistency, write

$$\begin{aligned}\hat{\beta} &= (Z'X)^{-1}Z'y \\ &= (Z'X)^{-1}Z'(X\beta + u) \\ &= \beta + \left(\frac{Z'X}{n}\right)^{-1}\frac{Z'u}{n}\end{aligned}$$

Let us assume that $(x_i, z_i, u_i)$ are i.i.d. with finite fourth moments, and that $Q = E(z_i x_i')$ is a nonsingular matrix and $E(z_i u_i) = 0$. Then by the weak law of large numbers

$$\frac{1}{n}Z'X \xrightarrow{p} Q,$$

and

$$\frac{1}{n}Z'u \xrightarrow{p} 0,$$

so it follows that $\hat{\beta}$ will be consistent.

For the distribution theory, write

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n}Z'X\right)^{-1}\left(\frac{1}{\sqrt{n}}Z'u\right).$$

The first term converges to $Q^{-1}$ in probability (and hence in distribution), while the second term will satisfy a central limit theorem:

$$\frac{1}{\sqrt{n}}Z'u \xrightarrow{d} N(0, \Omega),$$

where

$$\Omega = E(u_i^2 z_i z_i').$$

Applying Slutsky's Lemma,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Delta),$$

where

$$\Delta = Q^{-1}\Omega Q'^{-1} = E(z_i x_i')^{-1}E(u_i^2 z_i z_i')E(x_i z_i')^{-1}.$$

Suppose that we assume

$$E(u_i|z_i) = 0,$$

$$E(u_i^2|z_i) = \sigma^2$$

That is, we assume that the structural disturbance is mean-independent of $z_i$ and homoskedastic. Then

$$\Delta = \sigma^2 E(z_i x_i')^{-1} E(z_i z_i') E(x_i z_i')^{-1}$$

We can estimate $\Delta$ consistently by taking sample analogs for $\sigma^2$, $E(z_i x_i')$, and $E(z_i z_i')$. However, the standard errors that result will not be appropriate under heteroskedasticity. In the more general case where $u_i$ can be heteroskedastic, we need to directly estimate $E(u_i^2 z_i z_i')$, which we can again do by the analogy principle, by using

$$\frac{1}{n}\sum_i e_i^2 z_i z_i', \qquad \text{where } e_i = (y_i - x_i'\hat{\beta}).$$

Return to the special case of the OLS estimator. Under homoskedasticity, the asymptotic variance formula simplifies to

$$\Delta = \sigma^2 E(x_i x_i')^{-1}$$

Thus the approximate variance of the OLS estimator is

$$\frac{1}{n}\Delta = \sigma^2 E(n x_i x_i')^{-1}$$

which can be estimated by

$$s^2(X'X)^{-1}$$

as usual. If the errors are not assumed to be homoskedastic, then the appropriate variance estimate is given by

$$\hat{V(b)} = n(X'X)^{-1}\left(\frac{1}{n}\sum_i e_i^2 x_i x_i'\right)(X'X)^{-1}$$

which is the Eicker-White formula for the variance of the OLS estimator under general heteroskedasticity. The standard errors obtained from this formula are somtimes called "robust" standard errors.

Overidentified Case: Generalized IV Estimator

In the previous section we assumed that there were the same number of instruments as explanatory variables. In some cases, we may have more instruments than explanatory variables. Suppose that $z_i$ has dimension $K^*$, where $K^* > K$. Then the instrumental variables estimator would try to solve

$$
\begin{aligned}
\frac{1}{n}\sum_i z_{i1}(y_i - x_i'\beta) &= 0, \\
&\vdots \\
\frac{1}{n}\sum_i z_{i,K^*}(y_i - x_i'\beta) &= 0
\end{aligned}
$$

If $K^* > K$, there will typically be no solution to this system of equations, since there are more equations than unknown elements of $\beta$. We say the system of equations is *overidentified*; this is in contrast to the *just-identified* case of the previous section.

GMM Estimator

One aproach is to try to choose an estimate for $\beta$ in order to set all of the $K^*$ equations "close" to 0. We will do this by solving the minimization problem

$$
\min_\beta \left(\frac{1}{n}\sum z_i(y_i - x_i'\beta)\right)' A_n^{-1} \left(\frac{1}{n}\sum z_i(y_i - x_i'\beta)\right),
$$

where $A_n^{-1}$ is a $K^* \times K^*$ symmetric and positive definite matrix. (The "$n$" subscript indicates that the weight matrix could be a function of the data.)

Unlike many GMM problems, this one has a fairly simple explicit solution (no need for numerical optimization techniques!). The matrix first order conditions associated with this minimization problem are

$$
-2X'ZA_n^{-1}Z'y + 2X'ZA_n^{-1}Z'X\beta = 0
$$

which leads to the following equation:

$$
X'ZA_n^{-1}Z'(y - X\beta) = 0. \tag{2}
$$

Define

$$
W_n' = X'ZA_n^{-1}Z'.
$$

Then equation (2) can be written as

$$
W_n'(y - X\beta) = 0
$$

which leads to the solution

$$
\hat{\beta} = (W_n'X)^{-1}W_n'y.
$$

So $W_n$ acts like a matrix of instruments in the just-identified case.

Example: Two-Stage Least Squares

Suppose we set $A_n^{-1} = \left(\frac{Z'Z}{n}\right)^{-1}$. Then

$$W_n = ZA_n^{-1}Z'X = Z\left(\frac{Z'Z}{n}\right)^{-1}Z'X = nZ(Z'Z)^{-1}Z'X.$$

The term $Z(Z'Z)^{-1}Z'X$ can be interpreted as the fitted values in a regression of each column of $X$ on the matrix $Z$. To see this, write

$$X = (x^1, \ldots, x^K)$$

where $x^j$ is the $j$th column of $X$. Notice that $x^j$ is the $n \times 1$ vector of observations on variable $x_{ij}$. Let

$$c^j = (Z'Z)^{-1}Z'x^j$$

be the least squares coefficients in the regression of $x^j$ on $Z$. This can be written compactly as

$$C = (c^1, \ldots, c^K) = (Z'Z)^{-1}Z'(x^1, \ldots, x^K) = (Z'Z)^{-1}Z'X.$$

The fitted values are

$$\hat{x}^j = Zc^j = Z(Z'Z)^{-1}Z'x^j.$$

These can be collected into a matrix

$$\hat{X} = (\hat{x}^1, \ldots, \hat{x}^K) = Z(Z'Z)^{-1}Z'X = P_Z X,$$

where $P_Z = Z(Z'Z)^{-1}Z'$. Recall from standard least squares theory that $P_Z$ is symmetric and idempotent.

The IV estimator is then given by

$$\begin{aligned}
\hat{\beta} &= (W_n'X)^{-1}W_n'y \\
&= (X'P_Z X)^{-1}X'P_Z y \\
&= (X'P_Z P_Z X)^{-1}X'P_Z y \\
&= (\hat{X}'\hat{X})^{-1}\hat{X}'y.
\end{aligned}$$

This is called the *two-stage least squares estimator*, or 2SLS, because it can be done by two applications of least squares:

1. Regress $X$ on $Z$ to get the fitted values $\hat{X}$.
2. Regress $y$ on $\hat{X}$.

So 2SLS is a special case of IV, where the $W_n$ instruments are just the fitted values from the first stage regressions. Further, we will show below that this choice for the weight matrix leads to the smallest variance among IV estimators under homoskedasticity.

Asymptotic Properties of the Generalized IV Estimator

Asymptotic properties of the IV estimator in the overidentified case are similar to the just-identified case. It's easiest to write the estimator as

$$\begin{aligned}\hat{\beta} &= (W_n'X)^{-1}W_n'y \\ &= (X'ZA_n^{-1}Z'X)^{-1}X'ZA_n^{-1}Z'y \\ &= \beta + (X'ZA_n^{-1}Z'X)^{-1}X'ZA_n^{-1}Z'u.\end{aligned}$$

Assume that $A^{-1} = \text{plim}(A_n^{-1})$, is a finite positive definite matrix, and $Q = \text{plim}\left(\frac{Z'X}{n}\right)$ is a $K^* \times K$ matrix with rank $K$. As before, $\frac{Z'u}{n} \xrightarrow{p} 0$, so $\hat{\beta} \xrightarrow{p} \beta$. Likewise, it can be shown that

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Delta\Omega\Delta'),$$

where

$$\Omega = E(u_i^2 z_i z_i'), \qquad \Delta = (Q'A^{-1}Q)^{-1}Q'A^{-1}$$

The variance can be shown to be minimized by setting $A = \Omega$, or any constant multiple of $\Omega$. In practice, this means that it is optimal to choose $A_n$ to be a consistent estimator of $\Omega$. This leads to an asymptotic variance of

$$(Q'\Omega^{-1}Q)^{-1}.$$

In the homoskedastic case, where $\Omega = \sigma^2 E(z_i z_i')$, it is enough to have $A = E(z_i z_i')$, leading to the optimal asymptotic variance of

$$\sigma^2(Q'E(z_i z_i')^{-1}Q)^{-1}.$$

We can operationalize this by choosing the weight matrix to be

$$A_n^{-1} = \left(\frac{Z'Z}{n}\right)^{-1}$$

since $A_n \xrightarrow{p} E(z_i z_i')$. But this is just 2SLS. This shows that 2SLS is optimal in the class of IV estimators under homoskedasticity. Under general heteroskedascity this optimality no longer holds.