Lecture 8 – Instrumental variables

Economics 2123
George Washington University

Instructor: Prof. Ben Williams

Linear IV

Weak instruments

Exogeneity condition

Roadmap

- Most of this lecture will roughly correspond to Sections 4.8 and 4.9 in Cameron and Trivedi and 4.1, 4.2, and 4.6.4 in Angrist and Pischke.
- Two important issues that we will leave out today:
  - nonlinearity
  - heterogeneous effects

Linear IV

Weak instruments

Exogeneity condition

The IV model

- Consider the regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

  where x_i is *endogenous*, meaning that $E(u_i x_i) \neq 0$.
- If
  - z_i is *exogenous*, meaning that $E(u_i z_i) = 0$,
  - and *relevant*, meaning that $Cov(z_i, x_i) \neq 0$,

The IV model

- Consider the regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

  where x_i is *endogenous*, meaning that $E(u_i x_i) \neq 0$.
- If
  - z_i is *exogenous*, meaning that $E(u_i z_i) = 0$,
  - and *relevant*, meaning that $Cov(z_i, x_i) \neq 0$,

  then

$$\frac{Cov(y_i, z_i)}{Cov(x_i, z_i)} = \beta_1$$

Examples

- Supply and Demand (Wright, 1928)
  - To estimate demand, z_i should be a “supply shifter”.
  - To estimate supply, z_i should be a “demand shifter”.

Examples

- KIPP charter school evaluation (Angrist et al., 2012)
    - instrument: dummy for lottery result

The general model

- Consider the regression model

$$y_i = \beta_0 + \beta_1' X_{i1} + \beta_2 X_{i2} + u_i$$

  where X_{1i} is *exogenous* but X_{2i} is endogenous.
- Let Z_i denote a vector of exogenous variables not included in X_{i1}.
- Let $X_i = (X_{i1}', X_{i2}')'$ and $\tilde{Z}_i = (X_{i1}', Z_i')'$ and let $\beta = (\beta_1', \beta_2')'$.

The general model

- Exogeneity of Z_i now means "exogeneous *conditional on* X_{1i}"
  - In the supply/demand example, X_{i1} should include common determinants of supply and demand.
  - In the KIPP example, X_{i1} should include indicators for different lotteries if multiple lotteries are pooled.
    - In that paper, they need to include year dummies, as they pool data across multiple years, and grade dummies, as there were separate lotteries for entry into different grades.

The general model

- This could be motivated by the following triangular model.

$$X_{i2} = \gamma_0 + \gamma_1' X_{i1} + \gamma_2' Z_i + \nu_i \qquad \text{(First stage)}$$
$$y_i = \beta_0 + \beta_1' X_{i1} + \beta_2 X_{i2} + u_i \qquad \text{(Second stage)}$$

- If multiple regressors are endogenous, this can be modeled using multiple first stage equations.

The general model

- This could be motivated by the following triangular model.

$$X_{i2} = \gamma_0 + \gamma_1' X_{i1} + \gamma_2' Z_i + \nu_i \qquad \text{(First stage)}$$
$$y_i = \beta_0 + \beta_1' X_{i1} + \beta_2 X_{i2} + u_i \qquad \text{(Second stage)}$$

- If multiple regressors are endogenous, this can be modeled using multiple first stage equations.
- The triangular model also arises from a simultaneous equations model of the general form:

$$Y_i = B_0 + B_1 X_i + B_2 Y_i + U_i$$

where identification relies on sufficient restrictions on B_1 and B_2.

IV estimators

- Two-stage least squares (2SLS):
  - Estimate the OLS regression of X_{i2} on $\tilde{Z}_i$, producing fitted values $\hat{X}_{i2}$
  - Estimate the OLS regression of y_i on X_{i1} and $\hat{X}_{i2}$
  - In matrix notation,

$$\hat{\beta}_{2SLS} = \left(X'\tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'X \right)^{-1} X'\tilde{Z}(\tilde{Z}'\tilde{Z})^{-1}\tilde{Z}'y$$

where y is the $N \times 1$ vector (y_i), X is the $N \times K$ matrix (X_i') and $\tilde{Z}$ is the $N \times r$ matrix $(\tilde{Z}_i')$.

IV estimators

- The 2SLS estimator can also be written as

$$\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$$

where $\hat{X} = (X'_{i1}, \hat{X}_{i2})$.

- If $dim(\tilde{Z}_i) = dim(X_i)$ (just-identified) then it simplifies to

$$\hat{\beta}_{2SLS} = \left(\tilde{Z}'X\right)^{-1}\tilde{Z}'y.$$

IV estimators

- In the just identified case with X_{i2} scalar, 2SLS is also equivalent to the following.
  - Let $\hat{\gamma}_2$ denote the coefficient on Z_i in the first stage regression.
  - Let $\hat{\alpha}_2$ denote the coefficient on Z_i in the *reduced form* regression,
  $$y_i = \alpha_0 + \alpha_1' X_{i1} + \alpha_2 Z_i + \varepsilon_i$$
  - Then take the ratio, $\hat{\beta}_2 = \frac{\hat{\alpha}_2}{\hat{\gamma}_2}$
- The simplest version of this, the *Wald estimator*, is when $\tilde{Z}_i$ and X_i are both scalars and Z_i is binary:

$$\frac{E(y_i \mid Z_i = 1) - E(y_i \mid Z_i = 0)}{E(X_i \mid Z_i = 1) - E(X_i \mid Z_i = 0)}$$

IV estimators

- Indirect least squares (ILS)
  - Let Z_i^* denote a vector with the same dimension as X_i and define

$$\hat{\beta}^{ILS} = (Z^{*\prime}X)^{-1}Z^{*\prime}y$$

  - If (i) plim $N^{-1}\sum_{i=1}^{N} Z_i^* u_i = 0$ and (ii) plim $N^{-1}\sum_{i=1}^{N} Z_i^* X_i'$ is full rank, then $\hat{\beta}^{ILS} \rightarrow_p \beta$.
  - Clearly, just identified 2SLS is an example of an ILS estimator.
  - Actually, 2SLS is always an ILS estimator with $Z^* = \hat{X}$.

IV estimators

- When is the exogeneity condition, plim $N^{-1}\sum_{i=1}^{N} Z_i^* u_i = 0$, satisfied?
  - Suppose $Z_i^* = \psi(X_{i1}, Z_i)$ and that $E(u_i \mid X_{i1}, Z_i) = 0$.
  - Then if, for example, the sample is iid and $Var(Z_i^* u_i) < \infty$, the WLLN $\Longrightarrow$

$$N^{-1}\sum_{i=1}^{N} Z_i^* u_i \to_p E(Z_i^* u_i) = E(Z_i^* E(u_i \mid X_{i1}, Z_i)) = 0$$

IV estimators

- In my notes on IV estimators, I also show that
  - If the $\hat{X}$ in the 2SLS estimator, $(\hat{X}'\hat{X})^{-1}\hat{X}'y$, is replaced with a consistent estimator for $E(X_i \mid \tilde{Z}_i)$ then we obtain a consistent estimator of β. However, this is not the standard 2SLS estimator.

IV estimators

- In my notes on IV estimators, I also show that
  - If the $\hat{X}$ in the 2SLS estimator, $(\hat{X}'\hat{X})^{-1}\hat{X}'y$, is replaced with a consistent estimator for $E(X_i \mid \tilde{Z}_i)$ then we obtain a consistent estimator of β. However, this is not the standard 2SLS estimator.
  - If the model used to estimate $\hat{X}$ is misspecified (so that $\hat{X}$ is not consistent for $E(X_i \mid \tilde{Z}_i)$) then generally the modified 2SLS estimator will not be consistent. The standard 2SLS estimator is the exception to this rule – the OLS first stage does not have to be a consistent estimator for $E(X_i \mid \tilde{Z}_i)$.

IV estimators

- In my notes on IV estimators, I also show that
  - If the $\hat{X}$ in the 2SLS estimator, $(\hat{X}'\hat{X})^{-1}\hat{X}'y$, is replaced with a consistent estimator for $E(X_i \mid \tilde{Z}_i)$ then we obtain a consistent estimator of β. However, this is not the standard 2SLS estimator.
  - If the model used to estimate $\hat{X}$ is misspecified (so that $\hat{X}$ is not consistent for $E(X_i \mid \tilde{Z}_i)$) then generally the modified 2SLS estimator will not be consistent. The standard 2SLS estimator is the exception to this rule – the OLS first stage does not have to be a consistent estimator for $E(X_i \mid \tilde{Z}_i)$.
- This is often referred to as a “forbidden” regression.
- Examples
  - Not using all of X_{i1} in the first stage.
  - Estimating a nonlinear (e.g., a probit for binary X_{i2}) first stage.

IV estimators

- So why does the standard 2SLS work, even when $E(X_i \mid \tilde{Z}_i)$ is not linear?
  - The OLS residuals are uncorrelated with the OLS predicted values by construction. So

$$
\begin{aligned}
\hat{X}'\hat{X} &= \hat{X}'(X - \hat{e}) \\
&= \hat{X}'X
\end{aligned}
$$

IV estimators

- So why does the standard 2SLS work, even when $E(X_i \mid \tilde{Z}_i)$ is not linear?
  - The OLS residuals are uncorrelated with the OLS predicted values by construction. So

$$\begin{aligned} \hat{X}'\hat{X} &= \hat{X}'(X - \hat{e}) \\ &= \hat{X}'X \end{aligned}$$

  - But the residuals from a nonlinear model are not generally uncorrelated with the predicted values.

IV estimators

- Another consistent estimator
  - Use a (possibly nonlinear) model to estimate $\hat{X} = \hat{\mu}(\tilde{Z})$.
  - Let $Z^* = \hat{X}$.
  - The ILS estimator is a just identified 2SLS estimator with $\hat{X}$ as the *instrument*,

$$\left(\hat{\mu}(\tilde{Z})'X\right)^{-1}\hat{\mu}(\tilde{Z})'y$$

IV estimators

- Another consistent estimator
  - Use a (possibly nonlinear) model to estimate $\hat{X} = \hat{\mu}(\tilde{Z})$.
  - Let $Z^* = \hat{X}$.
  - The ILS estimator is a just identified 2SLS estimator with $\hat{X}$ as the *instrument*,

$$\left(\hat{\mu}(\tilde{Z})'X\right)^{-1}\hat{\mu}(\tilde{Z})'y$$

not

$$\left(\hat{\mu}(\tilde{Z})'\hat{\mu}(\tilde{Z})\right)^{-1}\hat{\mu}(\tilde{Z})'y$$

IV estimators

- Another consistent estimator
  - Use a (possibly nonlinear) model to estimate $\hat{X} = \hat{\mu}(\tilde{Z})$.
  - Let $Z^* = \hat{X}$.
  - The ILS estimator is a just identified 2SLS estimator with $\hat{X}$ as the *instrument*,

$$\left(\hat{\mu}(\tilde{Z})'X\right)^{-1}\hat{\mu}(\tilde{Z})'y$$

not

$$\left(\hat{\mu}(\tilde{Z})'\hat{\mu}(\tilde{Z})\right)^{-1}\hat{\mu}(\tilde{Z})'y$$

- The first stage does *not* have to provide a consistent estimator of the conditional expectation.

IV estimators

- Another IV estimator is the generalized method of moments (GMM):
  - Based on moments: $E((y_i - \beta' X_i)\tilde{Z}_i) = 0$
  - The GMM estimator uses a weighting matrix W_N and minimizes

$$\left(n^{-1}\sum_{i=1}^{n}((y_i - \beta' X_i)\tilde{Z}_i)\right)' W_N \left(n^{-1}\sum_{i=1}^{n}((y_i - \beta' X_i)\tilde{Z}_i)\right)$$

IV estimators

- Another IV estimator is the generalized method of moments (GMM):
  - Based on moments: $E((y_i - \beta' X_i)\tilde{Z}_i) = 0$
  - The GMM estimator uses a weighting matrix W_N and minimizes

$$\left(n^{-1}\sum_{i=1}^{n}((y_i - \beta' X_i)\tilde{Z}_i)\right)' W_N \left(n^{-1}\sum_{i=1}^{n}((y_i - \beta' X_i)\tilde{Z}_i)\right)$$

  - In matrix notation,

$$\hat{\beta}_{GMM} = \left(X'\tilde{Z}W_N\tilde{Z}'X\right)^{-1} X'\tilde{Z}W_N\tilde{Z}'y$$

Comparison of these estimators

1. GMM can be written as an ILS estimator as well.

Comparison of these estimators

1. GMM can be written as an ILS estimator as well.
2. If $dim(X_i) = dim(\tilde{Z}_i)$ (just identified case), GMM corresponds to the 2SLS estimator.

Comparison of these estimators

1. GMM can be written as an ILS estimator as well.
2. If $dim(X_i) = dim(\tilde{Z}_i)$ (just identified case), GMM corresponds to the 2SLS estimator.
3. If $dim(X_i) > dim(\tilde{Z}_i)$ different weights for the GMM estimator lead to different results.

Comparison of these estimators

1. GMM can be written as an ILS estimator as well.
2. If $dim(X_i) = dim(\tilde{Z}_i)$ (just identified case), GMM corresponds to the 2SLS estimator.
3. If $dim(X_i) > dim(\tilde{Z}_i)$ different weights for the GMM estimator lead to different results.
4. If $dim(X_i) > dim(\tilde{Z}_i)$ and errors are "spherical" ($Var(u) = \sigma_u^2 I$) then the optimal GMM estimator with *optimal* W_N coincides with 2SLS.

Comparison of these estimators

1. GMM can be written as an ILS estimator as well.
2. If $dim(X_i) = dim(\tilde{Z}_i)$ (just identified case), GMM corresponds to the 2SLS estimator.
3. If $dim(X_i) > dim(\tilde{Z}_i)$ different weights for the GMM estimator lead to different results.
4. If $dim(X_i) > dim(\tilde{Z}_i)$ and errors are "spherical" ($Var(u) = \sigma_u^2 I$) then the optimal GMM estimator with *optimal* W_N coincides with 2SLS.
5. If $dim(X_i) > dim(\tilde{Z}_i)$ and errors are heteroskedastic or autocorrelated, optimal GMM is more efficient than 2SLS.

- Other estimators:
  - LIML, jacknife 2SLS, Fuller estimator (later)
  - iterated GMM and CUE (later)
  - 3SLS, systems GMM (see CT Section 6.9)

Linear IV

Weak instruments

Exogeneity condition

The problem

- The IV estimator is *always* biased.
- Typically this bias is negligible for large samples.
- Tests have the wrong size. *Separate* problems:
  - a weak instrument can exacerbate large sample bias
  - a weak instrument will increase standard errors

The problem

- The IV estimator is *always* biased.
- Typically this bias is negligible for large samples.
- Tests have the wrong size. *Separate* problems:
  - a weak instrument can exacerbate large sample bias
  - a weak instrument will increase standard errors

Stock and Yogo (2005)

- The model of Stock and Yogo (2005):

$$y = Y\beta + X\gamma + u$$
$$Y = Z\Pi + X\Phi + V$$

where Y consists of n endogenous regressors, X consists of K_1 exogenous regressors, and Z consists of $K_2 \geq n$ exogenous instruments.

- Let $\underline{Z} = [X, Z]$ (what we previously called $\tilde{Z}$).

When X_{2i} is scalar

- Suppose $n = 1$ (one endogenous regressor) and $K_1 = 0$ (no "controls").
  - define the concentration parameter

$$\mu^2 = \frac{\Pi' Z' Z \Pi}{Var(V)}$$

  - the bias is

$$E(\hat{\beta}^{2SLS} - \beta) \approx \frac{Cov(u, V)}{Var(V)} \left(\mu^2 / K_2 + 1\right)^{-1}$$

  - note that $\frac{Cov(u,V)}{Var(V)}$ is the bias of the OLS estimator!

Weak instrument tests

- The easiest approach is based on the first stage F statistic.
    - the F statistic for testing $H_0 : \Pi = 0$
- if $F < 10$ then instruments are weak
- this rule-of-thumb is still fairly common but we can do better
- this rule-of-thumb and the tests below are *not* tests of the null hypothesis that $\Pi = 0$

Weak instrument tests

- The idea is to test the null hypothesis that the bias is less than or equal to $x\%$ of the OLS bias.
  - This approach can be extended to the general model.
  - An alternative is to test the null that the size of the Wald test for significance of β is no larger than $r \geq \alpha$.
  - Two relevant issues here: (1) heteroskedastic errors and (2) number of endogenous regressors
  - The Stock and Yogo (2005) test allows for $n > 1$ but assumes homoskedasticity.
  - Montiel Olea-Pflueger (2013) allows for heteroskedasticity but is only for the $n = 1$ case

- Stock and Yogo (2005)
    - A different first stage regression for each endogenous regressor.
    - Construct a matrix analogue of the first stage F statistic.
    - Stock and Yogo (2005) provide critical values for a test (Cragg-Donald) based on the smallest eigenvalue of this matrix.
        - The Cragg-Donald test is a test for underidentification.
        - Stock and Yogo (2005) test is for weak identification: critical values more conservative

- Montiel Olea-Pflueger (2013)
  - The Kleibergen-Paap statistic is a robust version of the Cragg-Donald statistic.
  - Montiel Olea-Pflueger (2013) argue that this is not the right statistic to use for identifying weak instruments.
  - They provide an alternative that only works for the $n = 1$ case.

Robust estimators

- Robust estimators.
  - *k*-class estimators
    - let $Y^\perp$ and $Z^\perp$ denote projections onto X
    - for a given k define

$$\hat{\beta}(k) = (Y^{\perp\prime}(I - kM_{Z^\perp})Y^\perp)^{-1} Y^{\perp\prime}(I - kM_{Z^\perp})y^\perp$$

    - $\hat{\beta}^{2SLS} = \hat{\beta}(1)$
    - LIML, Fuller, and bias-adjusted 2SLS use different (data-dependent) values of k – all asymptotically equivalent under standard asymptotics
    - Fuller seems to be the best in weak IV situations
    - available as options to `ivreg2` in Stata

Robust inference on β

- Robust hypothesis tests.
  - The basic idea is that we can base inference on the reduced form regression:

$$y = Z\Pi\beta + X(\Phi\beta + \gamma) + V\beta + u$$

  - To test $H_0 : \beta = 0$ this is obvious...
  - To test $H_0 : \beta = b$, it takes a little more work.
  - Two issues: (1) which test is efficient under weak instruments (2) robust to heteroskedasticity?

Robust inference on β

- Anderson Rubin test - the best choice in the just identified case
- Conditional likelihood ratio (CLR) test - the best choice in the over-identified case if errors are homoskedastic
- Heteroskedasticity and over-identification? There are extensions of the CLR test...active research on this.

Conclusion

- AP's suggestions
  - report first stage results
  - report first stage F statistic and compare to 10
  - estimate just-identified model
  - compare 2SLS with LIML
  - look at reduced form outcome regression
- my take
  - checking first stage results and reduced form outcome results makes sense
  - use Stock and Yogo (2005) rather than $F > 10$
  - try AR and CLR tests for significance
  - try GMM/CUE for improvements under heteroskedasticity
  - See NBER session

Large sample bias of IV

- Consider the simple setup with:
    - model: $Y_i = \beta' X_i + u_i$
    - ILS estimator: $(Z^{*\prime} X)^{-1} Z^{*\prime} Y$
    - This is consistent if $n^{-1} Z^{*\prime} u \rightarrow_p 0$

Large sample bias of IV

- What if $n^{-1}Z^{*\prime}u \not\to_p 0$?
- Consider the case where X_i and Z_i^* are scalar:
  - Let ρ_{Z^*u} and ρ_{Xu} denote nonzero correlations between Z^* and u and X and u.
  - Then the large sample bias of this IV estimator relative to the large sample bias of OLS:

$$\frac{\hat{\beta}_{IV} - \beta}{\hat{\beta}_{OLS} - \beta} \to_p \frac{\rho_{Z^*u}}{\rho_{Xu}\rho_{Z^*X}}$$

where ρ_{Z^*X} is the correlation between Z^* and X

Large sample bias of IV

- Another useful formula is possible in the case where X_{i2} and Z_i are scalar.
  - In this case, the 2SLS estimator is equal to

$$\frac{Cov(e_{ZX_1,i}, Y_i)}{Cov(e_{ZX_1,i}, X_{i2})}$$

    where $e_{ZX_1,i}$ is the residual from a regression of Z_i on X_{i1}.
  - The relative bias is

$$\frac{\hat{\beta}_{IV} - \beta}{\hat{\beta}_{OLS} - \beta} \rightarrow_p \frac{\rho_{e_{ZX_1} u}}{\rho_{e_{X_2 X_1} u} \rho_{e_{ZX_1} e_{X_2 X_1}}}$$

Large sample bias of IV

- Lessons:
  1. IV is more biased when X and Z^* are "equally endogenous"
  2. Minimal endogeneity of Z^* can lead to relatively large bias if ρ_{Z^*X} is small.
  3. Adding controls (X_{1i}) can reduce $\rho_{e_{ZX_1}u}$ but may also reduce $\rho_{e_{ZX_1}e_{X_2X_1}}$.

Tests of this assumption

- The exogeneity assumption is not testable.
- Overidentification tests:
    - If there are more instruments than needed, tests based on this overidentification are possible.
    - The simplest version is a Hausman test that compares two just-identified estimators.
    - The Hansen-Sargan test is a generalization of this.
    - The null of these tests is that all instruments are exogenous.
        - Rejection means that at least one instrument is not exogenous.
        - When there are heterogeneous effects, implications are even less clear.