ECMA 31100: Intro to Empirical Analysis II

Instrumental Variables

Joe Hardwick

University of Chicago

Winter 2022

Selection on observables

y_1 – Outcome under treatment $D=1$

y_0 – Outcome under treatment $D=0$

$ATE = E(y_1 - y_0)$ $\quad ATT = E(y_1 - y_0 | D=1)$

- Last week: Used unconfoundedness and overlap condition/linear conditional mean assumption to identify ATE. Unconfoundedness:

$$y_0, y_1 \perp D | w.$$

- Allows average treatment effects to vary by covariate value w:

$$\begin{aligned} ATE &= \mathrm{E}\left(ATE\left(w\right)\right) \\ &= \mathrm{E}\left(\mathrm{E}\left(y_1 - y_0 | w\right)\right) \\ &= \mathrm{E}\left(\mathrm{E}\left(Y | D=1, w\right) - \mathrm{E}\left(Y | D=0, w\right)\right). \end{aligned}$$

$\rightarrow E(E(y_1 - y_0 | w))$.

$Y = D y_1 + (1-D) y_0$

$E(y_1 | w) = E(y_1 | D_1 = w)$

$= E(Y | D=1, w)$.

Selection on observables

- Did not need to assume
  - Constant (homogeneous) treatment effects (perhaps conditional on w): $\mathrm{P}\left(y_1 - y_0 = c\left(w'\right) | w = w'\right) = 1$, or
  - Constant conditional average treatment effects: $\mathrm{E}\left(y_1 - y_0 | w\right) = c$,
- Such assumptions may simplify identification arguments e.g. in combination with linear regression.
- Several representations of ATE which suggest different estimation strategies.

Constant coefficients

Instrumental Variables

$y = x'\beta + u$ $\quad E(xu) \neq 0$

$y = x'\beta + u$ $\quad E(zu) = 0.$

β ↘ Random coefficient. (Heterogeneous treatment effects).

- Selection on observables not credible if there are unobserved factors affecting treatment choice and outcome.
- IV methods use an observable random variable z which satisfies:
  - Relevance: Instrument is correlated with the treatment (conditional on covariates).
  - Exclusion: Instrument has no direct effect on outcome.
  - Exogeneity: Instrument is independent of potential outcomes (conditional on covariates).
- Consequence: Covariance between instrument and observed outcome occurs only through changes in treatment, so we can measure the causal effect by adjusting for how much the treatment moves with the instrument.

Instrumental Variables

$Cov(D, Z) \approx 0$.

- We will explore:
  - Identification and estimation of homoegenous treatment effects in the linear model; empirical issues such as weak instruments.
  - Interpretation of IV estimates when treatment effects are heterogeneous.
- In contrast to selection on observables, heterogeneous treatment effects combined with IV assumptions greatly alter our interpretation of IV estimates.

Instrumental Variables: Potential Outcomes

- We observe outcomes y, covariates w, treatment status D and instrument z.
- Potential outcomes now a function of treatment d, (e.g. years of education) and instrument z:

$$Y = y(D, z).$$

- Exclusion: For all d, z', z'':

$$y(d, z') = y(d, z'').$$

- Exogeneity (conditional on covariates): For all d, z':

$$y(d, z') \perp z | w.$$

Instrumental Variables: Potential Outcomes

- Henceforth write $y(d) \equiv y(d, z')$ due to exclusion assumption.
- Exogeneity like selection on observables, but for the instrument. Allows for treatment to be correlated with unobserved determinants of outcome.
- Now assume constant linear treatment effects:

$$y(d) = x(w, d)'\beta + u$$

differenced out when we compute $y(d) - y(d')$.

where x is a known function, β is a vector of unknown constants and u is an unobserved random variable satisfying $\mathrm{E}(u|w) = 0$.

$E(y(d) \mid w) = x(w,d)'\beta$.

Instrumental Variables: Potential Outcomes

- Linear means 'linear in parameters': x can be nonlinear: e.g. effect of education on wage need not be equal for all education levels.
- However, treatment effects are constant conditional on covariates w:

$$y(d) - y(d') = \left[x(w,d) - x(w,d')\right]' \beta.$$

  e.g. for two individuals with the same level of experience (w), the effect of an additional year of education ($y(d) - y(d+1)$) is the same.
- Allows for interactions: e.g.

$$y(d) = \beta_0 + w'\beta_1 + d \cdot \beta_2 + d \cdot w'\beta_3 + u.$$

Instrumental Variables: Potential Outcomes

$y = \beta_0 + \beta_1 d + w'\beta_2 + u.$

$E(wu) = 0.$

- Model:

$$y(d) = x(w, d)'\beta + u; \quad \mathrm{E}(u|w) = 0.$$

- Assumption $\mathrm{E}(u|w) = 0$ equivalent to:

$$\mathrm{E}(y(d)|w) = x(w, d)'\beta.$$

- Assumed $\mathrm{E}(y(d)|w)$ is a linear (in β) conditional mean. Don't need to take a stance on whether w causally impacts $y(d)$ or is correlated with another variable that does.

$y_0, y_1 | D$ Reg. D on X - should it find correlation bt D, X.

$y = g(X) + U.$ $g(X) = X'\beta.$

Instrumental Variables: Potential Outcomes

$E(U|X) = 0$.

$E(Y|X) = g(X)$.

$y = g(X) + U$ $\quad E(U|X) \neq 0$.

$E(U|Z) = 0$.

Estimate g using "non-parametric IV".

$g(X) = X'\beta$.

- Finally, let the set of possible treatments be $\mathcal{D}$. Then

$$Y = \sum_{d \in \mathcal{D}} y(d)\,\mathbf{1}(D = d) \equiv y(D),$$

$\mathbf{1}(A) = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{otherwise.} \end{cases}$

yielding

$$Y = x(w, D)'\beta + u \equiv x'\beta + u,$$

where $x \equiv x(w, D)$.

$y(d) = x(w,d)'\beta + u$.

$$Y = \sum_{d \in \mathcal{D}} [x(w,d)'\beta + u]\,\mathbf{1}(D = d)$$

$$= x(w,D)'\beta + u.$$

Implication of exclusion assumption

- Note that for any d: $u = y(d) - x(w,d)'\beta$, so by exogeneity assumption:

$$y(d) \perp z|w$$

we have

$$\begin{aligned} \mathrm{E}(u|z,w) &= \mathrm{E}\left(y(d) - x(w,d)'\beta|z,w\right) \\ &= \mathrm{E}(y(d)|w) - x(w,d)'\beta \\ &= 0. \end{aligned}$$

In particular: $\mathrm{E}(zu) = 0$.

$E(y|d)|z,w) = E(y|d)|w)$

$= x(w,d)'\beta$

Exogeneity and exclusion often bundled together:

$y = x'\beta + u \; ; \; E(zu) = 0.$

Introduction to IV

$y = x'\beta + u.$

- Let (y, x, u) be a random vector such that y and u are scalar random variables and $x \in \mathbb{R}^{k+1}$.
- Assume the first component of x equals 1:

$$x = (x_0, x_1 \ldots, x_k),$$

  where $x_0 = 1$.
- Let $\beta = (\beta_0, \ldots, \beta_k) \in \mathbb{R}^{k+1}$ be a constant vector of unknown parameters such that

$$y = x'\beta + u.$$

Introduction to IV

x contains functions of D (endogenous)
functions of D, w (endogenous)
functions of w. (exogenous).

- If $\mathrm{E}(ux_j) = 0$ for some j, x_j is exogenous.
- If $\mathrm{E}(ux_j) \neq 0$ for some j, x_j is endogenous.
- x_0 can always be made exogenous by shifting β_0 such that $\mathrm{E}(x_0 u) = \mathrm{E}(u) = 0$.
- Multiply the model by x and take expectations:

$$\mathrm{E}(xy) = \mathrm{E}\left(xx'\right)\beta + \mathrm{E}(xu).$$

$E(u|w) = 0$.
$\Rightarrow E(u) = 0$.

$y = x'\beta + u$
$xy = xx'\beta + xu$
$E(xy) = E(xx')\beta + E(xu)$
$\downarrow$
$\neq 0$.

Introduction to IV

$E(xy) = E(xx')\beta + E(xu)$

$\hat{\beta}^{OLS} \to^p E(xx')^{-1} E(xy) = \beta + E(xx')^{-1} E(xu).$

$\hat{\beta} = \left(\frac{1}{n}\sum x_i x_i'\right)^{-1} \frac{1}{n}\sum x_i y_i$

- It follows that

$$\mathrm{E}\left(xx'\right)^{-1}\mathrm{E}\left(xy\right) = \beta + \mathrm{E}\left(xx'\right)^{-1}\mathrm{E}\left(xu\right).$$

- Therefore,

$$\hat{\beta}_n^{OLS} = \left(\frac{X'X}{n}\right)^{-1}\frac{X'Y}{n} \xrightarrow{p} \beta + \mathrm{E}\left(xx'\right)^{-1}\mathrm{E}\left(xu\right) \neq \beta.$$

- The OLS estimator is now an inconsistent estimator of β under endogeneity.

Instrumental Variables

Combining exclusion and exogeneity to justify validity.

- Our goal is to use a random vector $z \in \mathbb{R}^{l+1}$ such that $\mathrm{E}(zu) = 0$ to identify β.
- The condition $\mathrm{E}(zu) = 0$ is called instrument validity. (Multivariate version of $Cov(z, u) = 0$)
- First, note that any exogenous component of x is included in z. These components of x are called included instruments.
- The constant 1 is included, since we can always set $\mathrm{E}(u) = 0$. So, letting $z_0 = 1$:

$$z = (z_0, z_1, \ldots, z_l) \in \mathbb{R}^{l+1}.$$

$E(u)=0 \Rightarrow 1$ is an included instrument.

Instrumental Variables

- How to get β as a function of quantities we can estimate? Model
$$y = x'\beta + u.$$
- Pre-multiply by z:
$$zy = zx'\beta + zu.$$
- Take expectations:
$$\begin{aligned} \mathrm{E}(zy) &= \mathrm{E}(zx')\,\beta + \mathrm{E}(zu) \\ &= \mathrm{E}(zx')\,\beta. \end{aligned}$$
$\mathrm{E}(zu) = 0$ because of validity.
- If $l = k$ (exactly as many instruments as regressors), $\mathrm{E}(zx')$ is square, so
$$\beta = \left[\mathrm{E}(zx')\right]^{-1}\mathrm{E}(zy).$$

Instrumental Variables

$E(zy) = E(zx')\beta$

↑ Need this matrix to be full rank.

- The components of z are called instrumental variables.
- We further assume that $\mathrm{E}(zx')$ has rank $k+1$. (Instrument relevance/rank condition) (Multivariate version of $Cov(z,x) \neq 0$).
- Finally, we assume $\mathrm{E}(zz') < \infty$ and that there is no perfect collinearity in z.
- A necessary condition for the rank condition is $l \geq k$. This is called the order condition. In other words, we must have as many valid instruments as we have endogenous regressors.

Instrumental Variables: Order Condition

- If $l = k$, the system is exactly identified.
- If $l > k$, the system is overidentified, since we have more instruments than we need to identify β.
- Notice: If x_j is endogenous, it is not an included instrument.
- Given the order condition holds, the rank condition is necessary and sufficient to uniquely determine β.
- Later: What to do with extra instruments? Could throw them out and get an IV estimate, but this is inefficient.

$E(Z'X)$ has more rows than columns.
Sol: Pre-multiply by some C to obtain an invertible square matrix.

IV Estimator

$\ell = k$

- We showed under validity and relevance assumptions:

$$\beta = \mathrm{E}\left(zx'\right)^{-1}\mathrm{E}\left(zy\right).$$

- The sample analog principle yields

$$\frac{1}{n}\sum_{i=1}^{n} z_i\left(y_i - x_i'\hat{\beta}_{IV}\right) = 0,$$

or

$$\begin{aligned}\hat{\beta}_{IV} &= \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i y_i\right) \\ &\xrightarrow{p} \mathrm{E}\left(zx'\right)^{-1}\mathrm{E}\left(zy\right) = \beta.\end{aligned}$$

using the LLN and continuous mapping theorem, so the IV estimator is consistent.

IV Estimator

Matrix of observations of instruments.

$$Z = \begin{bmatrix} z_1' \\ \vdots \\ z_n' \end{bmatrix}.$$

- Stack the observations so that

$$Z' = (z_1, \ldots, z_n) \in \mathbb{R}^{(l+1)\times n},$$
$$X' = (x_1, \ldots, x_n) \in \mathbb{R}^{(k+1)\times n},$$
$$Y = (y_1, \ldots, y_n)' \in \mathbb{R}^{n}.$$

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

- Then:

$$\hat{\beta}_{IV} = \left(Z'X\right)^{-1} Z'Y.$$

$$= \left(\frac{1}{n}\sum z_i x_i'\right)^{-1}\left(\frac{1}{n}\sum z_i y_i\right)$$

GMM

$E(zy) = E(zx')\beta$.

- If $l > k$, the moment condition

$$\mathrm{E}\left(zu\right) = \mathrm{E}\left(z\left[y - x'\beta\right]\right) = 0$$

  has a solution by the model specification and the validity assumption, but its sample analog may not have a solution!

- That is, we cannot guarantee there exists $\hat{\beta}$ such that

$$\frac{1}{n}\sum_{i=1}^{n} z_i y_i = \frac{1}{n}\sum_{i=1}^{n} z_i x_i'\hat{\beta}.$$

  This would require that the $(l+1)\times 1$ vector on the LHS is a linear combination of the $k+1 < l+1$ columns of

$$\frac{1}{n}\sum_{i=1}^{n} z_i x_i.$$

GMM

$$C\frac{1}{n}\sum z_i y_i = C\frac{1}{n}\sum z_i x_i' \cdot \hat{\beta}.$$

$$C \in \mathbb{R}^{(k+1)\times(l+1)} \qquad \hat{\beta} = \left(C\tfrac{1}{n}\textstyle\sum z_i x_i'\right)^{-1}$$

- To obtain a unique solution, we must effectively reduce the $\left(C\frac{1}{n}\sum z_i y_i\right)$ number of rows in this equation to $k+1$.
- One (bad) option is to just discard extra instruments to yield a unique $\hat{\beta}$.
- This approach is not optimal because it discards information in the additional instruments that may improve our estimate of $\hat{\beta}$. It also doesn't provide us a way to decide which instruments to discard.

IV strategy: Observables $Y = Y(D)$ is observed outcome

D - treatment. Covariates - W.

Z - instrumental variables.

Potential Outcomes: $y(d)$: $Y(D) = \sum_{d \in \mathcal{D}} y(d)\, 1(D=d)$.

Conditions on TE's: (1) $y(d) = x(d,w)'\beta + u$

$\Rightarrow$ $y(d') - y(d)$ is constant after conditioning on w.

$y(d) = x(d,w)'\beta + u_d$ $E(u_d|w)=0$. $\leftrightarrow$ (2) $E(y(d)|w) = x(d,w)'\beta \Rightarrow E(u|w)=0$.

$y(d') = x(d',w)'\beta + u_{d'}$ $E(u_{d'}|w)=0$.

Conditions on instrument: (1) $y(d,z') = y(d,z'')$ (effect on outcome) for any choices of z', z''. Exclusion. Inst. has no direct effect on outcome.

Exogeneity. (2) $y(d) \perp\!\!\!\perp z | w$ (selection)

Relevance. (3) $E(zx')$ is full column rank.

Con TEs + Instruments: $y = x'\beta + u$: $E(zu) = 0$.

$\dim(x) = \dim(z)$: $E(zx')$ is square. $E(zy) = E(zx')\beta + E(zu)$

$\Rightarrow$ (Relevance) $\beta = E(zx')^{-1}E(zy)$

$\dim(x) < \dim(z)$: Choose C such that $CE(zx')$ is square.

$E(Czy) = E(Czx')\beta + E(Czu)$.

$\hookrightarrow CE(zu) = 0$

$\beta = E(Czx')^{-1}E(Czy)$

Note: Must choose C carefully to ensure $CE(zx')$ is full rank.

SAP: $\hat{\beta} = \left(\hat{C}\frac{1}{n}\sum z_i x_i'\right)^{-1}\left(\hat{C}\frac{1}{n}\sum z_i y_i\right)$ $\hat{C} \to^p C$.

GMM

$\hat{\beta} \to^p \beta$ $\quad \sqrt{n}(\hat{\beta} - \beta) \to^d N(0, V_C)$.

How to choose C to "minimize" V_C.

- Start with the overdetermined system

$$Z'Y = Z'X\hat{\beta},$$

$E(zy) = E(zx')\beta$

which may not have a solution. We first choose how to weight these sample moments by pre-multiplying by some full rank $(k+1) \times (l+1)$ matrix C, so

$$CZ'Y = CZ'X\hat{\beta},$$

then solve to give a GMM estimator:

$$\hat{\beta} = \left(CZ'X\right)^{-1} CZ'Y.$$

- We will see that the optimal C can be consistently estimated.

Questions?

The Rank Condition

- The assumption that $\mathrm{E}(zx')$ is full rank holds if and only if

$$\mathrm{E}(zz')^{-1}\mathrm{E}(zx')$$

  is full rank. To see this, note that if $\mathrm{E}(zz')^{-1}\mathrm{E}(zx')$ is full rank, then for any $c \in \mathbb{R}^{k+1} \setminus \{0\}$,

$$\mathrm{E}(zz')^{-1}\mathrm{E}(zx')\, c \neq 0,$$

  which implies $\mathrm{E}(zx')\, c \neq 0$.

- For the reverse implication, let $c \in \mathbb{R}^{k+1} \setminus \{0\}$ and note that if $c \neq 0$, then with $v = \mathrm{E}(zx')\, c$:

$$\mathrm{E}(zz')^{-1}\mathrm{E}(zx')\, c = \mathrm{E}(zz')^{-1} v \neq 0$$

  because $\mathrm{E}(zz')$ is full rank also.

The Rank Condition

- The matrix $\mathrm{E}\left(zz'\right)^{-1}\mathrm{E}\left(zx'\right)$ is a collection of coefficents of the best linear predictors of each x_j given z. if we let

$$x_j = z'\gamma_j + v_j, \quad \mathrm{E}\left(zv_j\right) = 0$$

then

$$\mathrm{E}\left(zz'\right)^{-1}\mathrm{E}\left(zx'\right) = \begin{bmatrix} | & | & | & | \\ \gamma_0 & \gamma_1 & \cdots & \gamma_k \\ | & | & | & | \end{bmatrix}.$$

The Rank Condition

- If there is a single endogenous regressor, x_k, and $k = l$ then

$$\mathrm{E}\left(zz'\right)^{-1}\mathrm{E}\left(zx'\right) = \begin{bmatrix} 1 & 0 & \cdots & 0 & \gamma_{k,0} \\ 0 & 1 & \cdots & 0 & \gamma_{k,1} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & \gamma_{k,l-1} \\ 0 & 0 & \cdots & 0 & \gamma_{k,l} \end{bmatrix}.$$

  This matrix has full rank iff $\gamma_{k,l} \neq 0$.

- In other words, with a single endogenous regressor and an exactly identified system, the rank condition holds if and only if a regression of x_k on the other $x's$ and the excluded instrument z_l produces a non-zero coefficient on z_l.
- x_k must be correlated with z_l "after controlling for $x_0, \ldots, x_{k-1}$."

Example: Returns to Schooling

- Suppose x_1 and x_2 are scalar random variables, and

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

  where $\mathrm{E}(u) = \mathrm{E}(x_1 u) = \mathrm{E}(x_2 u) = 0$, and $y = \ln(\textit{wage})$, $x_1 =$ years of schooling.
- Interpretation: Holding x_2 and other determinants of wage (u) fixed, each additional year of schooling leads to a $(100\beta_1)\%$ change in wage.
- Suppose we do not observe x_2 and rewrite the above as

$$y = \beta_0 + \beta_1 x_1 + v,$$

  where $v = \beta_2 x_2 + u$.

Omitted Variables bias causing endogeneity

Example: Returns to Schooling

$E[zu] = 0$

$E[xu] \neq 0$

- If students with greater x_2 generally opt for more years of schooling, $Cov(x_1, x_2) \neq 0$.
- So if $\beta_2 \neq 0$, $Cov(x_1, v) = \beta_2 Cov(x_1, x_2) \neq 0$.
- Besides the included instrument, $x_0 = 1$, we need a random variable z which is uncorrelated with x_2 and u. (Valid Instrument)
- Instrument relevance requires that $\gamma_1 \neq 0$ in the following regression

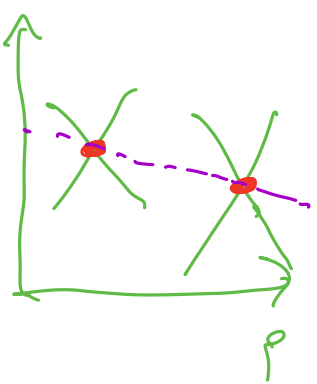
$$x_1 = \gamma_0 + \gamma_1 z + \epsilon,$$

  interpreted as the best linear predictor of x_1 given z. This condition holds iff $Cov(x_1, z_1) \neq 0$.
- One instrument suggested is presence of nearby college. Rationale is that living closer to a college will reduce cost of attendance while being unrelated to unobserved determinants of wage.

Simultaneous Equations

$y = x'\beta + u \quad E(xu) = 0.$

$q_D(p') = \beta_0 + \beta_1 p' + u.$



- Consider the following supply and demand system:

$$q_D = \beta_0 + \beta_1 p + u; \quad \mathrm{E}(u) = 0,$$
$$q_S = \gamma_0 + \gamma_1 p + v; \quad \mathrm{E}(v) = 0.$$

- Suppose also that $\mathrm{E}(uv) = 0$. We only observe supply and demand in equilibrium: $q_D = q_S$ when market clears. So:

$$\beta_0 + \beta_1 p + u = \gamma_0 + \gamma_1 p + v,$$
$$\implies p = \frac{1}{\beta_1 - \gamma_1}(\gamma_0 - \beta_0 + v - u).$$

- Is it reasonable to assume $\beta_1 \neq \gamma_1$?

Simultaneity Bias

$E[pu] = Cov(p,u) \neq 0.$

- It follows that p is endogenous in the equations

$$q = \beta_0 + \beta_1 p + u,$$
$$q = \gamma_0 + \gamma_1 p + v,$$

because

$$Cov(p,u) = Cov\left(\frac{1}{\beta_1 - \gamma_1}(\gamma_0 - \beta_0 + v - u), u\right) = -\frac{Var(u)}{\beta_1 - \gamma_1}$$
$$Cov(p,v) = Cov\left(\frac{1}{\beta_1 - \gamma_1}(\gamma_0 - \beta_0 + v - u), v\right) = \frac{Var(v)}{\beta_1 - \gamma_1}.$$

Exclusion Restrictions

- Now suppose the model is in fact given by

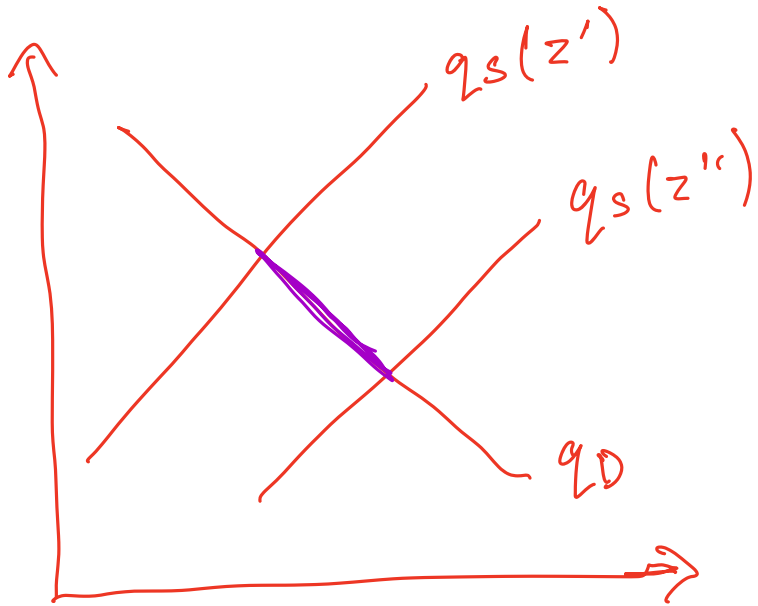
$$q_D = \beta_0 + \beta_1 p + u; \quad \mathrm{E}(u) = 0,$$
$$q_S = \gamma_0 + \gamma_1 p + \gamma_2 z + v; \quad \mathrm{E}(v) = \mathrm{E}(vz) = 0.$$

  where z is an exogenous "supply shifter", so $\mathrm{E}(zu) = 0$ also. Solving for the equilibrium price now yields

$$p = \frac{1}{\beta_1 - \gamma_1}(\gamma_0 - \beta_0 + \gamma_2 z + v - u).$$

Exclusion Restrictions

- Since $Cov(z, u) = 0$, can think of shifting z while holding u (and hence demand curve) fixed:



If we could alter raw materials costs holding all else fixed, we would observe two intersections of supply with the same demand curve, so we can identify the slope of demand curve.

Exclusion Restrictions

- The variable z (e.g. change in price of raw materials) affects supply but not demand. It is therefore excluded from the demand equation.
- The parameters of the demand equation

$$q = \beta_0 + \beta_1 p + u; \quad \mathrm{E}(u) = 0,$$

can now be estimated consistently, because z is a valid instrument for p

- Relevance holds if $\gamma_2 \neq 0$, since

$$Cov(p, z) = Cov\left(\frac{1}{\beta_1 - \gamma_1}(\gamma_0 - \beta_0 + \gamma_2 z + v - u), z\right)$$
$$= \frac{\gamma_2 Var(z)}{\beta_1 - \gamma_1}.$$

$E(zx') = E\left(\begin{pmatrix}1\\ z\end{pmatrix}(1 \; p)\right)$ must be full column rank.

Provided z is not constant this holds $\Leftrightarrow Cov(p, z) \neq 0$.

This occurs if $\gamma_2 \neq 0$

Questions?

and $\mathrm{Var}(z) \neq 0$. z not constant.

Bias of IV/GMM estimators

- IV/GMM estimators are typically biased. plugging in $Y = X\beta + U$ to the GMM estimator gives

$$\begin{aligned}\hat{\beta}_{GMM} &= \left(CZ'X\right)^{-1} CZ'Y \\ &= \beta + \left(CZ'X\right)^{-1} CZ'U,\end{aligned}$$

  and so in general

$$\mathrm{E}\left(\hat{\beta}_{GMM}|X,Z\right) = \beta + \left(CZ'X\right)^{-1} CZ'\mathrm{E}\left(U|X,Z\right) \neq \beta.$$

- The problem is that $\mathrm{E}\left(U|X\right) \neq 0$ because of endogeneity, so $\mathrm{E}\left(U|X,Z\right) \neq 0$.

Consistency of GMM estimators

$$\hat{\beta} = \left(\hat{C}\frac{1}{n}\sum z_i x_i'\right)^{-1}\hat{C}\frac{1}{n}\sum z_i y_i$$

$$\xrightarrow{p} \left(C\,E(z_i x_i')\right)^{-1} C\,E(z_i y_i) = \beta.$$

- Let $\hat{C} \xrightarrow{p} C$. The estimator based on $\hat{C}$ is consistent:

$$\begin{aligned}
\hat{\beta} &= \left(\hat{C}Z'X\right)^{-1}\hat{C}Z'Y \\
&= \beta + \left(\hat{C}\frac{Z'X}{n}\right)^{-1}\hat{C}\frac{Z'U}{n} \\
&\xrightarrow{p} \beta + \left(C\mathrm{E}\left(z_i x_i'\right)\right)^{-1} C\mathrm{E}\left(z_i u_i\right) \\
&= \beta.
\end{aligned}$$

Asymptotic normality of GMM estimators

- Rewrite

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

$$=\left(\hat{C}\frac{1}{n}\sum_{i=1}^{n}z_i x_i'\right)^{-1}\hat{C}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}z_i u_i$$

$$\xrightarrow{d}\left(C\mathrm{E}\left(z_i x_i'\right)\right)^{-1}C\times\mathcal{N}\left(0,\mathrm{E}\left(u_i^2 z_i z_i'\right)\right)$$

$$=\mathcal{N}\left(0,V\right),$$

where

$$V=\left(C\mathrm{E}\left(z_i x_i'\right)\right)^{-1}C\Omega C'\left(\mathrm{E}\left(z_i x_i'\right)'C'\right)^{-1},$$

$$\Omega=\mathrm{E}\left(u_i^2 z_i z_i'\right).$$

Optimal choice of C

$C = E(z_i x_i')' \left(E(u_i^2 z_i z_i')^{-1}\right)$

- Assume $\Omega = \mathrm{E}\left(u_i^2 z_i z_i'\right)$ is invertible and let $Q = \mathrm{E}\left(z_i x_i'\right)$.
- We now show that $C_{OGMM} = Q'\Omega^{-1}$ minimizes the variance.
- Plug $C_{OGMM} = Q'\Omega^{-1}$ into V:

$$
\begin{aligned}
V_{OGMM} &= \left(C_{OGMM} Q\right)^{-1} C_{OGMM} \Omega C_{OGMM}' \left(Q' C_{OGMM}'\right)^{-1} \\
&= \left(Q'\Omega^{-1} Q\right)^{-1} Q\Omega^{-1}\Omega\Omega^{-1} Q' \left(Q'\Omega^{-1} Q\right)^{-1} \\
&= \left(Q'\Omega^{-1} Q\right)^{-1}.
\end{aligned}
$$

Optimal choice of C

- Now show that $(CQ)^{-1} C\Omega C' (Q'C')^{-1} - \left(Q'\Omega^{-1}Q\right)^{-1}$ is positive semidefinite.
- To do this we will write $\left(Q'\Omega^{-1}Q\right)^{-1}$ in a sandwich form to relate it to $(CQ)^{-1} C\Omega C' (Q'C')^{-1}$.
- Note that since Ω is positive definite and symmetric, $\Omega^{1/2}$ exists, and we can write

$$\begin{aligned}\left(Q'\Omega^{-1}Q\right)^{-1} &= (CQ)^{-1} C\Omega^{1/2} \\ &\quad \times \left(\Omega^{-1/2} Q \left(Q'\Omega^{-1}Q\right)^{-1} Q'\Omega^{-1/2}\right) \\ &\quad \times \Omega^{1/2} C' \left(Q'C'\right)^{-1}.\end{aligned}$$

$$(CQ)^{-1} C\Omega C' \left(Q'C'\right)^{-1} = (CQ)^{-1} C\Omega^{1/2} \times \Omega^{1/2} C' \left(Q'C'\right)^{-1}$$

Optimal choice of C

- Letting $R = \Omega^{-1/2} Q$ yields that

$$\begin{aligned}
&(CQ)^{-1} C\Omega C' \left(Q'C'\right)^{-1} - \left(Q'\Omega^{-1}Q\right)^{-1} \\
=&(CQ)^{-1} C\Omega^{1/2} \left(I_{l+1} - R\left(R'R\right)^{-1} R'\right) \Omega^{1/2} C' \left(Q'C'\right)^{-1} \\
=&(CQ)^{-1} C\Omega^{1/2} M_R \Omega^{1/2} C' \left(Q'C'\right)^{-1} \geq 0
\end{aligned}$$

  since M_R is positive semidefinite.

- In summary, the asymptotically optimal linear combination of moments is found by setting

$$\hat{\beta} = \left(\hat{C} Z'X\right)^{-1} \hat{C} Z'Y,$$

  where $\hat{C}$ is a consistent estimator of $\mathrm{E}\left(x_i z_i'\right) \Omega^{-1}$.

GMM

$C = E(z_i x_i')' \left(E(u_i^2 z_i z_i')\right)^{-1}$

$\hat{\beta}_{OGMM}$

$\frac{1}{n}\sum z_i x_i'$ $\quad$ $\frac{1}{n}\sum \hat{u}_i^2 z_i z_i'$

Residuals from 2SLS estimation of β.

$\hat{u}_i = y_i - x_i'\hat{\beta}$

$\hat{\beta}$ is a consistent estimator of β.

- If $\hat{\Omega} \xrightarrow{p} \Omega$, we say

$$\hat{\beta}_{OGMM} = \left(X'Z\hat{\Omega}^{-1}Z'X\right)^{-1} X'Z\hat{\Omega}^{-1}Z'Y$$

is a (feasible) optimal GMM estimator. It follows that

$$\sqrt{n}\left(\hat{\beta}_{OGMM} - \beta\right) \xrightarrow{d} \mathcal{N}\left(0, \left(Q'\Omega^{-1}Q\right)^{-1}\right)$$

- Remaining question is how to get a consistent estimate of $\Omega = \mathrm{E}\left(u_i^2 z_i z_i'\right)$.

Questions?

GMM under conditional homoskedasticity

- Conditional homoskedasticity: $\mathrm{E}\left(u_i^2|z_i\right) = \mathrm{E}\left(u_i^2\right) = \sigma^2$.
- In this case,

$$\Omega = \mathrm{E}\left(u_i^2 z_i z_i'\right) = \mathrm{E}\left(\mathrm{E}\left(u_i^2|z_i\right) z_i z_i'\right) = \sigma^2 \mathrm{E}\left(z_i z_i'\right).$$

- In this case, a feasible optimal GMM estimator is given by

$$\begin{aligned}\hat{\beta}_{OGMM} &= \left(X'Z\left[\sigma^2 Z'Z\right]^{-1} Z'X\right)^{-1} X'Z\left[\sigma^2 Z'Z\right]^{-1} Z'Y \\ &= \left(X'P_Z X\right)^{-1} X'P_Z Y.\end{aligned}$$

Two Stage Least Squares

① Regress X on Z $\quad X = Z\Pi + V \quad \Pi \in \mathbb{R}^{(l+1)\times(k+1)}$

$\hat{X} = Z\hat{\Pi} = Z(Z'Z)^{-1}Z'X = P_Z X.$

② $Y = X\beta + U$ Estimate β by regressing Y on $\hat{X}$ using OLS

- This is called the two-stage least squares estimator, because it performs the previous task of reducing the number of moments by first regressing the columns of X on Z using OLS. Let

$$X = Z\Pi + V,$$

$\hat{\beta} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y.$
$= (X'P_Z X)^{-1}X'P_Z Y.$

  where Π is an $(l+1)\times(k+1)$ matrix of parameters.
- This is often called the "first stage regression". It finds the $k+1$ linear combinations of the $l+1$ instruments that are closest to X in the Euclidean norm.

Two Stage Least Squares

- The projection of each column of X onto Z is given by

$$P_Z X = Z\hat{\Pi}.$$

- Notice that for the included instruments, X_j, $P_Z X_j = X_j$ because X_j is one of the columns of Z.
- In the "Second Stage", the exogenous and endogenous regressors X are replaced by the exogenous regressors and the projection of the endogenous regressors onto Z. The original regression model is

$$Y = X\beta + U.$$

Two Stage Least Squares

- The model we actually estimate is

$$Y = P_Z X\bar{\beta} + \epsilon.$$

- Estimating this second stage regression by OLS produces

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

- Notice that if $l = k$, then $Z'Z$ and $X'Z$ are in fact square, and $\hat{\beta}_{2SLS}$ reduces to $\hat{\beta}_{IV}$.

Asymptotic distribution of 2SLS

Under homoskedasticity. Can do this even with heteroskedasticity. but limiting variance changes.

- The asymptotic distribution of the 2SLS estimator is given by

$$\sqrt{n}\left(\hat{\beta}_{2SLS}-\beta\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\left(Q'\mathrm{E}\left(z_i z_i'\right)^{-1} Q\right)^{-1}\right).$$

$\sigma^2 = \mathrm{var}(u|z)$.

- Let $\hat{U} = Y - X\hat{\beta}_{2SLS}$. A consistent estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{\hat{U}'\hat{U}}{n}. \quad = \quad \frac{SSR}{n}$$

- To see this, note that $\hat{U} = U - X'\left(\hat{\beta}_{2SLS}-\beta\right)$, and so

$$\frac{\hat{U}'\hat{U}}{n} = \frac{U'U}{n} + o_p(1).$$

Stuff which converges in probability to zero.

$\to p\ \sigma^2$.

$\mathrm{E}(u^2) = \mathrm{var}(u)$.

Inference with 2SLS

$\hat{V}_{hom} = \hat{\sigma}^2 \left(\frac{1}{n} \sum X_i Z_i' \cdot \frac{1}{n} \sum Z_i Z_i' \cdot \frac{1}{n} \sum Z_i X_i' \right)^{-1}$

- In summary, under homoskedasticity, $\hat{\beta}_{2SLS}$ is an asymptotically optimal GMM estimator, and

$$\sqrt{n}\left(\hat{\beta}_{2SLS} - \beta\right) \xrightarrow{d} \mathcal{N}\left(0, V_{hom}\right),$$

where $V_{hom} = \sigma^2 \left(Q' \mathrm{E}\left(z_i z_i'\right)^{-1} Q \right)^{-1}$ can be consistently estimated by

$$\hat{V}_{hom} = n\hat{\sigma}^2 \left(X' P_Z X\right)^{-1}.$$

$\beta_j = 0$ vs. $\beta_j \neq 0$.

- A confidence set for β_j may be found by noting that

$r = [0, 0, \ldots, \underset{j+1\text{'st position}}{1}, 0, 0]$

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

$$\frac{\sqrt{n} r'\left(\hat{\beta}_{2SLS} - \beta\right)}{\sqrt{r' \hat{V}_{hom} r}} \xrightarrow{d} \mathcal{N}(0,1),$$

Standard errors for linear combination of $\hat{\beta}_j$'s.

$r' \hat{\beta} = \hat{\beta}_j$

for any constant $(k+1) \times 1$ vector r, by Slutsky's Theorem.

GMM under Heteroskedasticity

- Under heteroskedasticity, the variance does not simplify. A consistent estimate of Ω is given by:

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 z_i z_i',$$

where $\hat{u}_i = y_i - x_i'\hat{\beta}_{2SLS}$.

$\hat{\beta}_{OGMM}$ depends on $\hat{\Omega}$ which is not necessarily a precise estimate of Ω. Asymptotics ignore this completely

- The proof of consistency is identical to the heteroskedastic case when considering OLS estimation. The result follows because $\hat{\beta}_{2SLS}$ is a $\sqrt{n}-$consistent estimator of β.
- Although $\hat{\beta}_{2SLS}$ is not asymptotically optimal, it does allow for consistent estimation of Ω because it depends only on Z, X, Y. Its finite sample performance is also not affected by the need to estimate Ω.

Inference with GMM

- Under heteroskedasticity, the optimal GMM estimator is $\hat{\beta}_{OGMM} = \left(X'Z\hat{\Omega}^{-1}Z'X\right)^{-1} X'Z\hat{\Omega}^{-1}Z'Y$, and

$$\sqrt{n}\left(\hat{\beta}_{OGMM} - \beta\right) \xrightarrow{d} \mathcal{N}\left(0, V_{het}\right),$$

  where $V_{het} = \left(Q'\Omega^{-1}Q\right)^{-1}$, which is consistently estimated by

$$\hat{V}_{het} = \left(\frac{X'Z}{n}\hat{\Omega}^{-1}\frac{Z'X}{n}\right)^{-1}.$$

- A confidence interval for β_j may be found in the same manner as with the 2SLS estimator.

Questions?

2SLS vs. GMM

- In PSET 1, you are asked to compute confidence intervals for regression parameters using both 2SLS and GMM under the assumptions of homoskedasticity and heteroskedasticity respectively.
- It is valid, and often done in practice, to use 2SLS with heteroskedasticity robust standard errors, despite possible loss of efficiency.
- Next we ask what quantity 2SLS/IV estimates if the coefficients of our model are in fact random.

Causal parameters from estimators

- Return to the model

$$y_i = \beta_0 + \beta_1 d_i + u_i,$$

  where y_i is the observed outcome of individual i and $d_i = 1$ if individual i is treated and 0 otherwise.
- A causal interpretation of this model implies that the ceteris paribus effect of the treatment, β_1, is the same for everybody.
- On the other hand, since d_i is binary we can always write

$$y_i = b_0 + b_1 d_i + \epsilon_i; \quad \mathrm{E}(\epsilon_i | d_i) = 0.$$

- The latter model does not necessarily have a causal interpretation, but the parameter b_1 may be of interest.
- The error terms u_i and ϵ_i are interpreted differently: u_i contains unobserved determinants of y_i, ϵ_i is the projection residual. It is not necessarily true that $\mathrm{E}(u_i | d_i) = 0$.

Causal parameters from estimators

- We may write

$$y_i = y_{i1} d_i + y_{i0} (1 - d_i) = y_{i0} + d_i (y_{i1} - y_{i0}),$$

  where y_{i0}, y_{i1} are the potential outcomes for individual i.
- y_{i0} is the outcome we would observe if the individual is not treated, and y_{i1} the outcome if the individual is treated.
- The difference $y_{i1} - y_{i0}$ is the individual treatment effect.
- The average across all individuals, $\mathrm{E}(y_{i1} - y_{i0})$, is the average treatment effect (ATE).

Causal parameters from estimators

- So far we have argued that if treatment is randomly assigned then the OLS estimate gives the ATE. Adding the assumption that $y_{i1} - y_{i0} = c$ allowed us to estimate c and to interpret the regression as a causal model.
- One reason to be interested in 2SLS is that while we may ask for a parameter of interest (such as the ATE) and then look at how to estimate it, we can go the other way around: Use the estimation method anyway and then ask what parameter it identifies.
- For 2SLS/IV the answer (under some assumptions) is a Local Average Treatment Effect.

Causal parameters from estimators

- Suppose we allow for heterogeneous treatment effects, so that $y_{i1} - y_{i0} = \beta_{i1}$, and the baseline outcome $\beta_{i0} = y_{i0}$ may also differ across individuals. Then

$$y_i = \beta_{i0} + \beta_{i1} d_i.$$

- Such a model is called a random coefficients model, because the fact we are sampling individuals at random implies that the actual value of β_{i0}, β_{i1} drawn is random also.

Causal parameters from estimators

- The assumption of random assignment to treatment, $(y_{i0}, y_{i1}) \perp\!\!\!\perp d_i$, implies that in the model

$$y_i = b_0 + b_1 d_i + \epsilon_i, \quad \mathrm{E}(\epsilon_i | d_i) = 0,$$

  we have

$$\begin{aligned} b_1 &= \mathrm{E}(y_i | d_i = 1) - \mathrm{E}(y_i | d_i = 0) \\ &= \mathrm{E}(y_{i1} | d_i = 1) - \mathrm{E}(y_{i0} | d_i = 0) \\ &= \mathrm{E}(y_{i1} - y_{i0}). \end{aligned}$$

- Random assignment implies OLS estimate of b_1 is a consistent estimate of the ATE.

Causal parameters from estimators

- If, on the other hand, there is selection into treatment, we obtain merely

$$b_1 = \mathrm{E}\left(y_{i1}|d_i = 1\right) - \mathrm{E}\left(y_{i0}|d_i = 0\right).$$

- Want to identify some feature of the distribution of $y_{i1} - y_{i0}$.
- Suppose we have a binary instrument $z \in \{0, 1\}$. For example, y might be worker output, $d = 1$ if the worker elects job training and 0 otherwise, and $z = 1$ if the worker is offered job training and 0 otherwise. z may be independent of (y_0, y_1) even if d isn't.

Causal parameters from estimators

- Now we not only have potential outcomes but also potential treatments:

$$d_i = d_{i1} z_i + d_{i0} (1 - z_i),$$

  where the potential outcome $d_{i1} = 1$ if individual i would take job training when they are offered it and 0 otherwise. $d_{i0} = 1$ if individual i would take job training when they are not offered it, and 0 otherwise.
- Question: "OLS produces $b_1 = \mathrm{E}(y_{i1}|d_i = 1) - \mathrm{E}(y_{i0}|d_i = 0)$, which is not a feature of $y_{i1} - y_{i0}$, but what if I do 2SLS/IV with my instrument?"

Causal parameters from estimators

- First we will make some relevance and validity assumptions analogous to those in the linear model.
- Validity: $(y_0, y_1, d_0, d_1) \perp\!\!\!\perp z$.
- Relevance: Two assumptions
  - Monotonicity assumption: $\mathrm{P}(d_1 \geq d_0) = 1$. This implies

$$\begin{aligned} Cov(d, z) &= \mathrm{E}(dz) - \mathrm{E}(d)\mathrm{E}(z) \\ &= \mathrm{P}(d = 1, z = 1) - \mathrm{P}(d = 1)\mathrm{P}(z = 1) \\ &= [\mathrm{P}(d = 1|z = 1) - \mathrm{P}(d = 1|z = 0)]\mathrm{P}(z = 1)\mathrm{P}(z = 0) \\ &= [\mathrm{P}(d_1 = 1) - \mathrm{P}(d_0 = 1)]\, Var(z) \\ &= \mathrm{P}(d_1 > d_0)\, Var(z) \geq 0. \end{aligned}$$

  - $\mathrm{P}(d_0 \neq d_1) > 0$. This says that the instrument must alter treatment choice for some positive fraction of individuals. Together with monotonicity, this implies $\mathrm{P}(d_1 > d_0) > 0$.

Causal parameters from estimators

- Note that the first stage regression produces fitted values

$$\hat{d}_i = \hat{\gamma}_0 + \hat{\gamma}_1 z_i,$$

where

$$\hat{\gamma}_1 = \frac{\hat{Cov}(d, z)}{\hat{Var}(z)} \xrightarrow{p} \mathrm{P}(d_1 > d_0) > 0.$$

- The second stage regression runs OLS on

$$y_i = \alpha + \beta \hat{d}_i + v_i.$$

- This gives:

$$\hat{\beta}_{2SLS} = \frac{\hat{Cov}(y, z) / \hat{Var}(z)}{\hat{Cov}(d, z) / \hat{Var}(z)} \xrightarrow{p} \frac{Cov(y, z)}{Cov(d, z)}.$$

Causal parameters from estimators

- A similar argument shows that

$$
\begin{aligned}
Cov\left(y,z\right) &= \left[\mathrm{E}\left(y|z=1\right)-\mathrm{E}\left(y|z=0\right)\right]Var\left(z\right)\\
&= \mathrm{E}\left(y_1 d_1 + y_0\left(1-d_1\right)|z=1\right)Var\left(z\right)\\
&- \mathrm{E}\left(y_1 d_0 + y_0\left(1-d_0\right)|z=0\right)Var\left(z\right)\\
&= \mathrm{E}\left(y_1 d_1 + y_0\left(1-d_1\right)\right)Var\left(z\right)\\
&- \mathrm{E}\left(y_1 d_0 + y_0\left(1-d_0\right)\right)Var\left(z\right)\\
&= \mathrm{E}\left(\left[y_1-y_0\right]\left[d_1-d_0\right]\right)Var\left(z\right)\\
&= \mathrm{E}\left(y_1-y_0|d_1>d_0\right)\mathrm{P}\left(d_1>d_0\right)Var\left(z\right).
\end{aligned}
$$

Causal parameters from estimators

- It follows that

$$\hat{\beta}_{2SLS} \xrightarrow{p} \frac{Cov\left(y,z\right)}{Cov\left(d,z\right)} = \mathrm{E}\left(y_1 - y_0 | d_1 > d_0\right).$$

- This parameter is the average treatment effect among those who would be switched from no treatment to treatment if the value of the instrument changed.
- In the job training example, it is the average effect among those who would elect training if offered it and not elect it otherwise.
- This parameter may or may not be of interest, and this interpretation depends on the monotonicity assumption.
- The interpretation changes if the instrument is changed (unless, of course, $y_1 - y_0$ is constant).

Questions?