ECMA31000: Introduction to Empirical Analysis
Hypothesis Testing; Instrumental Variables

Joe Hardwick

University of Chicago

Autumn 2021

Outline

- This Week:
  - Hypothesis testing in the linear regression model
  - IV as a solution to Endogeneity
  - Properties of IV Estimators

Large sample inference

- Joint normality allows us to obtain exact finite sample distributions for these statistics, but we may also appeal to asymptotic normality.
- A test is called asymptotically of size α if

$$\lim_{n\to\infty} \beta_n(\theta) \leq \alpha \quad \text{for all } \theta \in \Theta_0.$$

- Now suppose

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) \xrightarrow{d} \mathcal{N}(0, V),$$

where $V = \mathrm{E}(xx')^{-1}\,\mathrm{E}\left(u^2xx'\right)\mathrm{E}(xx')^{-1}$. Suppose also that V is non-singular and $\hat{V}_n \xrightarrow{p} V$ is a consistent estimator of V.

Testing a single linear restriction

- Consider testing

$$H_0 : r'\beta = c \quad vs. \quad H_1 : r'\beta \neq c,$$

where r is some specified vector in $\mathbb{R}^{k+1}$, and c is a scalar.

- By the CMT:

$$\sqrt{n}\left(r'\hat{\beta}_n - r'\beta\right) \xrightarrow{d} \mathcal{N}\left(0, r'Vr\right),$$

and so by Slutsky's theorem:

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

Testing a single linear restriction

- It follows that under H_0, the test statistic

$$T_n = \frac{\sqrt{n}\left(r'\hat{\beta}_n - c\right)}{\sqrt{r'\hat{V}_n r}} \xrightarrow{d} \mathcal{N}(0,1).$$

- The test we use is $\phi_n = \mathbf{1}\left(|T_n| > z_{1-\alpha/2}\right)$. It is of asymptotic size α, because under H_0 :

$$\begin{aligned} Pr\left(\text{Reject } H_0\right) &= \mathrm{P}\left(|T_n| > z_{1-\alpha/2}\right) \\ &= \mathrm{P}\left(T_n < -z_{1-\alpha/2}\right) + \mathrm{P}\left(T_n > z_{1-\alpha/2}\right) \\ &\to \Phi\left(-z_{1-\alpha/2}\right) + 1 - \Phi\left(z_{1-\alpha/2}\right) \\ &= \frac{\alpha}{2} + 1 - \left(1 - \frac{\alpha}{2}\right) = \alpha. \end{aligned}$$

- Use $\phi_n = \mathbf{1}\left(T_n > z_{1-\alpha}\right)$ for testing $H_0 : r'\beta \leq c$ vs. $H_1 : r'\beta > c$.

Asymptotic Confidence Set for $r'\beta$

- It follows by definition of convergence in distribution that for any value of $r'\beta$:

$$\lim_{n\to\infty} \mathrm{P}_{r'\beta}\left(\left|\frac{\sqrt{n}\left(r'\hat{\beta}_n - r'\beta\right)}{\sqrt{r'\hat{V}_n r}}\right| \le z_{1-\alpha/2}\right) = 1-\alpha.$$

- Since $z_{\alpha/2} = -z_{1-\alpha/2}$ by symmetry of the standard normal about 0, rearranging yields that

$$C_n = \left[r'\hat{\beta}_n - z_{1-\alpha/2}\sqrt{\frac{r'\hat{V}_n r}{n}}, r'\hat{\beta}_n + z_{1-\alpha/2}\sqrt{\frac{r'\hat{V}_n r}{n}}\right]$$

is an asymptotic $1-\alpha$ confidence interval for $r'\beta$.

Testing Multiple Linear Restrictions

- Consider testing

$$H_0 : R\beta = c \quad vs. \quad R\beta \neq c,$$

  where R is a $p \times (k+1)$ −dimensional matrix of full row rank and c is a $p \times 1$ vector.
- The full rank condition means none of our restrictions are redundant.
- By the CMT:

$$\sqrt{n}\left(R\hat{\beta}_n - R\beta\right) \xrightarrow{d} \mathcal{N}\left(0, RVR'\right),$$

  where RVR' is full rank (because R and V are), and hence positive definite (because V is).

Testing Multiple Linear Restrictions

- To see that RVR' is positive definite, note that if $a \neq 0$, $R'a \neq 0$, so
$$\left(R'a\right)' V \left(R'a\right) > 0,$$
because V is positive definite.
- A positive definite and symmetric matrix A has a square root $A^{1/2}$ with inverse $A^{-1/2} = \left(A^{-1}\right)^{1/2}$.
- It follows by Slutsky's Theorem that
$$\left(R\hat{V}_n R'\right)^{-1/2} \sqrt{n} \left(R\hat{\beta}_n - R\beta\right) \xrightarrow{d} \mathcal{N}\left(0, I_p\right).$$

Testing Multiple Linear Restrictions

- It follows that

$$n \cdot \left(R\hat{\beta}_n - R\beta\right)' \left(R\hat{V}_n R'\right)^{-1} \left(R\hat{\beta}_n - R\beta\right) \xrightarrow{d} \chi^2_p.$$

- Under H_0,

$$T_n = n \cdot \left(R\hat{\beta}_n - c\right)' \left(R\hat{V}_n R'\right)^{-1} \left(R\hat{\beta}_n - c\right) \xrightarrow{d} \chi^2_p,$$

and so we reject iff $T_n > \chi^2_{p,1-\alpha}$.

Asymptotic Confidence Set for $R\beta$

- It follows that

$$C_n = \left\{ c \in \mathbb{R}^p : n \cdot \left(R\hat{\beta}_n - c\right)' \left(R\hat{V}_n R'\right)^{-1} \left(R\hat{\beta}_n - c\right) \leq \chi^2_{p,1-\alpha} \right\}$$

  is an asymptotic $1-\alpha$ confidence set for $R\beta$.

- This set is an ellipsoid centered at $R\hat{\beta}_n$, and satisfies

$$\mathrm{P}_{R\beta}\left(R\beta \in C_n\right) \to 1-\alpha.$$

- Taking $R = I_{k+1}$ yields an asymptotic $1-\alpha$ confidence set for β.

Tests of Non-Linear Restrictions

- Finally, consider testing

$$H_0 : f(\beta) = 0 \quad vs. \quad H_1 : f(\beta) \neq 0,$$

where $f : \mathbb{R}^{k+1} \to \mathbb{R}^p$ is continuously differentiable at β.

- Let $D_\beta f(\beta)$ denote the $p \times (k+1)$ dimensional matrix of partial derivatives of f evaluated at β.
- The Delta Method implies

$$\sqrt{n}\left(f\left(\hat{\beta}_n\right) - f(\beta)\right) \xrightarrow{d} \mathcal{N}\left(0, D_\beta f(\beta) V D_\beta f(\beta)'\right)$$

Tests of Non-Linear Restrictions

- The continuous mapping theorem implies that

$$D_\beta f\left(\hat{\beta}_n\right)\hat{V}_n D_\beta f\left(\hat{\beta}_n\right)' \xrightarrow{p} D_\beta f(\beta)\, V D_\beta f(\beta)'.$$

- Now assume $D_\beta f(\beta)$ is full row rank. We can construct a statistic with asymptotic χ^2_p distribution as before.
- Note that $f(\beta) = R\beta$ yields our previous analysis as a special case, since $D_\beta f(\beta) = R$.

Questions?

Introduction to IV

- 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

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

- We *no longer* assume $\mathrm{E}(ux) = 0$, so β may not represent the best linear predictor, and therefore not the best predictor either.

Introduction to IV

- We are therefore interpreting this regression as a causal model.
- 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:

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

Introduction to IV

- 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,

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

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

Instrumental Variables

- 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 $z_0 = 1$:

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

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:
$$\begin{aligned} \mathrm{E}(zy) &= \mathrm{E}\left(zx'\right)\beta + \mathrm{E}(zu) \\ &= \mathrm{E}\left(zx'\right)\beta. \end{aligned}$$
- If $l = k$ (exactly as many instruments as regressors), $\mathrm{E}(zx')$ is square, so
$$\beta = \left[\mathrm{E}\left(zx'\right)\right]^{-1}\mathrm{E}(zy).$$

Instrumental Variables

- 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.

IV Estimator

- 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

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

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

IV Estimator

- Stack the observations so that

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

- Then:

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

GMM

- If $l > k$, the moment condition

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

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

- To obtain a unique solution, we must effectively reduce the 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.

GMM

- Start with the overdetermined system

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

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

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

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

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

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 $v = \mathrm{E}(zx')\,c$:

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

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

The Rank Condition

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

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

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 $y = \ln(wage)$, $x_1 =$ 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 $v = \beta_2 x_2 + u$.

Example: Returns to Schooling

- 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 ability 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.

Measurement Error

- Suppose x_1^* is a scalar random variable, and
$$y = \beta_0 + \beta_1 x_1^* + u,$$
where $\mathrm{E}(u) = \mathrm{E}(x_1^* u) = 0$.
- Suppose we observe a noisy signal of x_1^* given by
$$x_1 = x_1^* + v,$$
where $\mathrm{E}(x_1^* v) = \mathrm{E}(v) = \mathrm{E}(uv) = 0$.

Measurement Error

- Rewrite the true model as

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

  where $\epsilon = u - \beta_1 v$. Now $\mathrm{E}\left(\epsilon\right) = 0$, but $\mathrm{E}\left(x_1 \epsilon\right) = -\beta_1 Var\left(v\right)$.
- An instrument for x_1 may be another noisy measurement of x_1^*:

$$z_1 = x_1^* + w,$$

  where $\mathrm{E}\left(w\right) = 0$, $Cov\left(x_1^*, w\right) = 0$, $Cov\left(u, w\right) = 0$ and $Cov\left(w, v\right) = 0$. We have

$$Cov\left(z_1, \epsilon\right) = Cov\left(x_1^* + w, u - \beta_1 v\right) = 0.$$

Measurement Error

- To check relevance, note that we again require $\gamma_1 \neq 0$ in the regression

$$z_1 = \gamma_0 + \gamma_1 x_1 + \eta,$$

which holds iff $Cov\left(z_1, x_1\right) \neq 0$.

- But

$$Cov\left(z_1, x_1\right) = Cov\left(x_1^* + v, x_1^* + w\right) = Var\left(x_1^*\right),$$

which is nonzero provided x_1^* is not almost surely constant.

Simultaneous Equations

- 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

- It follows that p is endogenous in the equations

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

because

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

Exclusion Restrictions

- Now suppose the model is in fact given by

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

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

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

Exclusion Restrictions

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

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

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

  can now be estimated consistently, because z is a valid instrument for x.
- Relevance holds if $\gamma_2 \neq 0$, since

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

Questions?

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 $\mathrm{E}\left(U|X,Z\right) \neq 0$. (PSET 7 asks for an explicit example).

Consistency of GMM estimators

- 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

$$
\begin{aligned}
\sqrt{n}\left(\hat{\beta}-\beta\right) &= \left(\hat{C}\frac{Z'X}{n}\right)^{-1}\hat{C}\frac{Z'U}{\sqrt{n}} \\
&= \left(\hat{C}\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right)^{-1}\hat{C}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i u_i \\
&\xrightarrow{d} \left(C\mathrm{E}\left(z_i x_i'\right)\right)^{-1} C \times \mathcal{N}\left(0, \mathrm{E}\left(u_i^2 z_i z_i'\right)\right) \\
&= \mathcal{N}\left(0, V\right),
\end{aligned}
$$

where

$$
\begin{aligned}
V &= \left(C\mathrm{E}\left(z_i x_i'\right)\right)^{-1} C\Omega C'\left(\mathrm{E}\left(z_i x_i'\right)' C'\right)^{-1}, \\
\Omega &= \mathrm{E}\left(u_i^2 z_i z_i'\right).
\end{aligned}
$$

Optimal choice of C

- 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:

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

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

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

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

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 Weight Matrix

- Another method of choosing from $l+1$ moments minimizes

$$\left(Z'\left[Y-Xb\right]\right)'\hat{W}_n\left(Z'\left[Y-Xb\right]\right)$$

  over $b\in\mathbb{R}^{k+1}$, where $\hat{W}_n\xrightarrow{p}W$ is a weighting matrix. The solution is given by

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

- Comparing this formula with the previous slide reveals $\hat{W}_n\xrightarrow{p}\Omega^{-1}$ is asymptotically optimal.
- $W=\Omega^{-1}$ is called the optimal weight matrix.

GMM

- 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

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

- The only 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,

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

- 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

- 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

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

  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

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

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

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

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

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

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

Inference with 2SLS

- 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

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

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

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

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

Questions?