Unit 3. Instrumental variables estimation of single-equation linear models

Block C. The generalized method of moments

Instructor: Henry Redondo

Dpto. de Fundamentos del Análisis Económico
Universidad de Alicante

Econometrics II · Academic Year 2026/2027

Outline of Block C

1. C. The generalized method of moments
   - The GMM estimator
   - Asymptotic properties of the GMM estimator
   - Optimal weighting matrix
   - Optimal instruments
   - Application: small firms, noisy gaps

GMM

- Consider the linear model

$$y = \mathbf{x}'\boldsymbol{\beta} + u \tag{30}$$

  where $\mathbf{x}$ is a $k \times 1$ vector whose first element is 1 and some of the explanatory variables are potentially endogenous.
- We assume that we have some other variables, the instruments, that are not in (30) and are uncorrelated with u (they are exogenous in (30)). Let $\mathbf{z}$ be the $L \times 1$ vector $(L \geq k)$ of all the exogenous variables (the exogenous regressors and the instruments) and assume that $2SLS1$ and $2SLS2$ are satisfied.
- Let $\{(y_i, \mathbf{x}_i', \mathbf{z}_i')' : i = 1, .., N\}$ be a random sample. If $L = k$, we know that the simple IV estimator is the solution to the following system of equations

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i(y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}) = 0$$

More moments than parameters

- When $L > k$ the system above has no solution, but we can choose $\widehat{\boldsymbol{\beta}}$ to minimize the distance from the $L \times 1$ vector $\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i(y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}})$ to zero using a certain criterion.
- For instance, if we use the Euclidean distance, $\widehat{\boldsymbol{\beta}}$ will be the solution to the following minimization problem

$$\min_{\mathbf{b}} \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i(y_i - \mathbf{x}_i'\mathbf{b}) \right)' \left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_i(y_i - \mathbf{x}_i'\mathbf{b}) \right)$$

  and it can be shown that $\widehat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$.
- However, there is no reason why we should use the identity matrix as a weighting in the quadratic form above.
- Instead, we could also use any positive semidefinite weighting matrix (with rank larger or equal to k) and we will obtain a consistent estimator.
- As we will show below, the asymptotic variance of the estimator will depend on our choice of the weighting matrix.

The GMM estimator

- Let $\widehat{\mathbf{W}}$ be any $L \times L$ symmetric and positive semidefinite matrix with rank larger or equal to k (the "$\widehat{\ }$" is included to emphasize that in many cases $\widehat{\mathbf{W}}$ will be an estimator).
- A *GMM* estimator of $\boldsymbol{\beta}$ is defined as the solution to the minimization problem

$$\begin{aligned} \min_{\mathbf{b}} S(\mathbf{b}) &= \min_{b} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_i (y_i - \mathbf{x}_i' \mathbf{b}) \right)' \widehat{\mathbf{W}} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_i (y_i - \mathbf{x}_i' \mathbf{b}) \right) \\ &= \min_{\mathbf{b}} (\mathbf{Z}'(\mathbf{y} - \mathbf{X}\mathbf{b}))' \widehat{\mathbf{W}} (\mathbf{Z}'(\mathbf{y} - \mathbf{X}\mathbf{b})) \end{aligned} \tag{31}$$

where

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1' \\ \vdots \\ \mathbf{z}_N' \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_N' \end{bmatrix}, \text{ and } \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

First-order conditions of GMM

- The partial derivatives of the objective function are:

$$\frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = -2\mathbf{X}'\mathbf{Z}\widehat{\mathbf{W}}(\mathbf{Z}'(\mathbf{y} - \mathbf{X}\mathbf{b}))$$

- and the FOC are

$$\begin{aligned} \mathbf{X}'\mathbf{Z}\widehat{\mathbf{W}}(\mathbf{Z}'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})) &= 0 \Leftrightarrow \\ \mathbf{X}'\mathbf{Z}\widehat{\mathbf{W}}\mathbf{Z}'\mathbf{X}\widehat{\boldsymbol{\beta}} &= \mathbf{X}'\mathbf{Z}\widehat{\mathbf{W}}\mathbf{Z}'\mathbf{y} \end{aligned}$$

- Assuming that $\mathbf{X}'\mathbf{Z}\widehat{\mathbf{W}}\mathbf{Z}'\mathbf{X}$ is non-singular, the *GMM* estimator is given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{Z}\widehat{\mathbf{W}}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\widehat{\mathbf{W}}\mathbf{Z}'\mathbf{y}$$

- We can obtain infinitely many *GMM* estimators using different weighting matrices.
- For instance the 2*SLS* estimator is the *GMM* estimator based on

$$\widehat{\mathbf{W}} = \left(\frac{\mathbf{Z}'\mathbf{Z}}{N}\right)^{-1}$$

Asymptotic properties of the GMM estimator

- In this section we study the asymptotic properties of the GMM estimator.
- We first study under which assumptions the GMM estimator is consistent.
- Next, we obtain the asymptotic distribution of the GMM estimator.
- Finally, we find the estimated variance matrix of the GMM estimator.

The proofs repeat the 2SLS steps with $\widehat{\mathbf{W}}$ in place of $(\mathbf{Z}'\mathbf{Z}/N)^{-1}$; $GMM4$ ($\widehat{\mathbf{W}} \to_p \mathbf{W}$) is what makes them go through.

Consistency of GMM

- Under assumptions $2SLS1$, $2SLS2$ and

$$\widehat{\mathbf{W}} \to_p \mathbf{W} \quad \text{where } \mathbf{W} \text{ is a non-random, symmetric and positive semidefinite matrix and } \operatorname{rank}(\mathbf{W}) \geq k \qquad GMM4$$

the GMM estimator defined as the solution of the minimization problem (31) is a consistent estimator of $\boldsymbol{\beta}$.

Proof of consistency of GMM

- *Proof*
  - We can write $\widehat{\boldsymbol{\beta}}$ as

$$\widehat{\boldsymbol{\beta}} = \left(\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{z}_i'\right)\widehat{\mathbf{W}}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i\mathbf{x}_i'\right)\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{z}_i'\right)\widehat{\mathbf{W}}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i y_i\right) =$$
$$\boldsymbol{\beta} + \left(\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{z}_i'\right)\widehat{\mathbf{W}}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i\mathbf{x}_i'\right)\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{z}_i'\right)\widehat{\mathbf{W}}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i u_i\right)$$

  - $\mathbf{x}_i\mathbf{z}_i'$ are iid, then, using the *LLN*,

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{z}_i' \rightarrow_p E(\mathbf{x}\mathbf{z}') \tag{32}$$

  - Using (32), GMM4 and the continuous function theorem, we have that

$$\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{z}_i'\right)\widehat{\mathbf{W}}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i\mathbf{x}_i'\right) \rightarrow_p E(\mathbf{x}\mathbf{z}')\mathbf{W}E(\mathbf{z}\mathbf{x}')$$

Proof of consistency of GMM, continued

- We know that $E(\mathbf{x}\mathbf{z}')$ has rank k (by $2SLS2b$) and $\text{rank}(\mathbf{W}) \geq k$ (by $GMM4$). Then, $E(\mathbf{x}\mathbf{z}')\mathbf{W}E(\mathbf{z}\mathbf{x}')$ has full rank, and using the continuous function theorem

$$\left(\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{z}_i'\right)\widehat{\mathbf{W}}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i\mathbf{x}_i'\right)\right)^{-1} \to_p \left(E(\mathbf{x}\mathbf{z}')\mathbf{W}E(\mathbf{z}\mathbf{x}')\right)^{-1} \tag{33}$$

- $\mathbf{z}_i u_i$ are iid and $E(\mathbf{z}_i u_i) = 0$ (by $2SLS1$), then, using the LLN

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i u_i \to_p 0 \tag{34}$$

- Then, using (34), $GMM4$, (32) and (33) we have that

$$\widehat{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}$$

Asymptotic normality of GMM

Theorem (asymptotic normality of GMM). If assumptions $2SLS1$, $2SLS2$ and $GMM4$ are satisfied and $E(u^2\mathbf{z}\mathbf{z}')$ exists, the GMM estimator of $\boldsymbol{\beta}$ based on a random sample is asymptotically normal and the asymptotic distribution is

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d N(0, \mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-1}) \tag{35}$$

where

$$\mathbf{C}^{-1} = (E(\mathbf{x}\mathbf{z}')\mathbf{W}E(\mathbf{z}\mathbf{x}'))^{-1}, \qquad \mathbf{A} = E(\mathbf{x}\mathbf{z}')\mathbf{W}E(u^2\mathbf{z}\mathbf{z}')\mathbf{W}E(\mathbf{z}\mathbf{x}')$$

- *Proof*
  - We have that

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = \left(\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{z}_i'\right)\widehat{\mathbf{W}}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i\mathbf{x}_i'\right)\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{z}_i'\right)\widehat{\mathbf{W}}\left(\frac{\sqrt{N}}{N}\sum_{i=1}^{N}\mathbf{z}_i u_i\right)$$

  - $\mathbf{x}_i\mathbf{z}_i'$ are iid, then, using the LLN,

$$\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{z}_i' \to_p E(\mathbf{x}\mathbf{z}') \tag{36}$$

Proof of asymptotic normality of GMM, continued

- *Proof (cont.)*
  - Using (36), GMM4 and the continuous function theorem

$$\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{z}_i'\right)\widehat{\mathbf{W}}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i\mathbf{x}_i'\right)\rightarrow_p E(\mathbf{x}\mathbf{z}')\mathbf{W}E(\mathbf{z}\mathbf{x}')$$

  - We know that $E(\mathbf{x}\mathbf{z}')$ has rank k (by $2SLS2b$) and $\text{rank}(\mathbf{W}) \geq k$ (by $GMM4$). Then, $E(\mathbf{x}\mathbf{z}')\mathbf{W}E(\mathbf{z}\mathbf{x}')$ has full rank, and using the continuous function theorem

$$\left(\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{z}_i'\right)\widehat{\mathbf{W}}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i\mathbf{x}_i'\right)\right)^{-1}\rightarrow_p \left(E(\mathbf{x}\mathbf{z}')\mathbf{W}E(\mathbf{z}\mathbf{x}')\right)^{-1}=\mathbf{C}^{-1} \quad (37)$$

  - $\mathbf{z}_i u_i$ are iid, $E(\mathbf{z}_i u_i)=0$ (by $2SLS1$) and $E(u^2\mathbf{z}\mathbf{z}')$ exists, then, using the *CLT*

$$\frac{\sqrt{N}}{N}\sum_{i=1}^{N}\mathbf{z}_i u_i \rightarrow_d N(0, E(u^2\mathbf{z}\mathbf{z}')) \quad (38)$$

Proof of asymptotic normality of GMM, continued

- *Proof (cont.)*
  - Then, using (38), *GMM*4, (36) and Slutsky's theorem, we have

$$\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\mathbf{z}_i'\right)\widehat{\mathbf{W}}\left(\frac{\sqrt{N}}{N}\sum_{i=1}^{N}\mathbf{z}_i u_i\right) \\ \rightarrow_d \quad N(0,\mathbf{A}) \\ \text{where } \mathbf{A} = E(\mathbf{x}\mathbf{z}')\mathbf{W}E(u^2\mathbf{z}\mathbf{z}')\mathbf{W}E(\mathbf{z}\mathbf{x}') \tag{39}$$

  - Finally, using (39), (37) and Slutsky's theorem, we have

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow_d N(0, \mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-1})$$

    where

$$\mathbf{C}^{-1} = (E(\mathbf{x}\mathbf{z}')\mathbf{W}E(\mathbf{z}\mathbf{x}'))^{-1}, \qquad \mathbf{A} = E(\mathbf{x}\mathbf{z}')\mathbf{W}E(u^2\mathbf{z}\mathbf{z}')\mathbf{W}E(\mathbf{z}\mathbf{x}')$$

Optimal weighting matrix

- We have shown that the asymptotic variance of the *GMM* estimator depends on the choice of the weighting matrix.
- Then, the question is, which is the optimal choice for $\widehat{\mathbf{W}}$?
- Consider $\widehat{\mathbf{W}}$ such that

$$\widehat{\mathbf{W}} \to_p \lambda E(u^2 \mathbf{z}\mathbf{z}')^{-1}$$

where λ is any non-random positive scalar.

Optimal weighting matrix

- We are going to compute the variance of the asymptotic distribution of the *GMM* estimator based on this choice for $\widehat{\mathbf{W}}$.
  - We have proved that the variance of the asymptotic distribution of the *GMM* estimator is $\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-1}$, where

$$\begin{aligned} \mathbf{A} &= E(\mathbf{x}\mathbf{z}')\mathbf{W}E(u^2\mathbf{z}\mathbf{z}')\mathbf{W}E(\mathbf{z}\mathbf{x}') \\ \mathbf{C}^{-1} &= (E(\mathbf{x}\mathbf{z}')\mathbf{W}E(\mathbf{z}\mathbf{x}'))^{-1} \end{aligned}$$

  - Then, when $\mathbf{W} = \lambda E(u^2\mathbf{z}\mathbf{z}')^{-1}$, we have that

$$\begin{aligned} \mathbf{A} &= E(\mathbf{x}\mathbf{z}')\lambda E(u^2\mathbf{z}\mathbf{z}')^{-1}E(u^2\mathbf{z}\mathbf{z}')\lambda E(u^2\mathbf{z}\mathbf{z}')^{-1}E(\mathbf{z}\mathbf{x}') \\ &= \lambda^2 E(\mathbf{x}\mathbf{z}')E(u^2\mathbf{z}\mathbf{z}')^{-1}E(\mathbf{z}\mathbf{x}') \\ \mathbf{C}^{-1} &= \frac{1}{\lambda}\left(E(\mathbf{x}\mathbf{z}')E(u^2\mathbf{z}\mathbf{z}')^{-1}E(\mathbf{z}\mathbf{x}')\right)^{-1} \end{aligned}$$

  - and the variance of the asymptotic distribution of this *GMM* estimator is

$$\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-1} = \left(E(\mathbf{x}\mathbf{z}')E(u^2\mathbf{z}\mathbf{z}')^{-1}E(\mathbf{z}\mathbf{x}')\right)^{-1}$$

Optimal weighting matrix

- It can be proved that $\left(E(\mathbf{x}\mathbf{z}')E(u^2\mathbf{z}\mathbf{z}')^{-1}E(\mathbf{z}\mathbf{x}')\right)^{-1}$ is smaller than or equal to the general expression in (35) for an arbitrary matrix $\mathbf{W}$.
- Therefore, the *GMM* estimator that uses as weighting matrix a consistent estimator of $E(u^2\mathbf{z}\mathbf{z}')^{-1}$ (or proportional to that) has the smallest possible variance among the class of *GMM* estimators based on z.
- Then, if $2SLS3$ is satisfied, $E(u^2\mathbf{z}\mathbf{z}') = \sigma^2 E(\mathbf{z}\mathbf{z}')$, the optimal choice for $\widehat{\mathbf{W}}$ is

$$\widehat{\mathbf{W}} = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i\mathbf{z}_i'\right)^{-1} = \left(\frac{\mathbf{Z}'\mathbf{Z}}{N}\right)^{-1}$$

  and the $2SLS$ estimator is optimal.

Optimal weighting matrix

- If assumption 2*SLS*3 is not satisfied the 2*SLS* estimator is no longer optimal.
- The optimal *GMM* is then obtained using as weighting matrix

$$\widehat{\mathbf{W}} = \left(\frac{1}{N} \sum_{i=1}^{N} \widehat{u}_i^2 \mathbf{z}_i \mathbf{z}_i' \right)^{-1}$$

  where $\widehat{u}_i = y_i - \mathbf{x}_i' \widehat{\boldsymbol{\beta}}$ are the residuals based on a consistent estimator like for instance 2*SLS*.
- The estimated variance of the optimal *GMM* is

$$\widehat{var(\widehat{\boldsymbol{\beta}}_{OGMM})} = \left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{z}_i' \left(\sum_{i=1}^{N} \widehat{u}_i^2 \mathbf{z}_i \mathbf{z}_i' \right)^{-1} \sum_{i=1}^{N} \mathbf{z}_i \mathbf{x}_i' \right)^{-1}$$

- We can do inference using the *GMM* estimator analogously as using 2*SLS*.

One family, several members

Estimator	Moments	Weighting matrix $\widehat{\mathbf{W}}$	Efficient among GMM on $\mathbf{z}$ when
Simple IV	$L = k$	irrelevant	always (all GMM estimators coincide)
2SLS	$L \geq k$	$\left(\frac{1}{N}\mathbf{Z}'\mathbf{Z}\right)^{-1}$	homoskedasticity ($2SLS3$) holds
Optimal GMM	$L \geq k$	$\left(\frac{1}{N}\sum_i \widehat{u}_i^2 \mathbf{z}_i \mathbf{z}_i'\right)^{-1}$	always

With $L > k$ we cannot set all L sample moments to zero. The weighting matrix decides which moments we try hardest to satisfy: optimal GMM gives more weight to the moments that are measured more precisely.

Beyond linear combinations of $\mathbf{z}$: with heteroskedasticity, even optimal GMM is not the best IV estimator; the optimal instruments are $E(\mathbf{x} \mid \mathbf{z})/\sigma^2(\mathbf{z})$.

Optimal instruments

- Consider the linear model
$$y = \mathbf{x}'\boldsymbol{\beta} + u \tag{40}$$
where $\mathbf{x}$ is a $k \times 1$ vector whose first element is 1.
- Suppose that
$$E(u \mid \mathbf{z}) = 0 \tag{41}$$
where $\mathbf{z}$ is an $L \times 1$ vector ($L \geq k$).
- We know that (41) implies that u is uncorrelated with any function of $\mathbf{z}$, that is
$$E(h(\mathbf{z})u) = E\left(h(\mathbf{z})(y - \mathbf{x}'\boldsymbol{\beta})\right) = 0$$
for any function h such that this expectation exists.
- Then, in general, there is an infinite number of orthogonality conditions (for different choices of h) we can use to identify $\boldsymbol{\beta}$.
- Notice that for each choice of h, $h(\mathbf{z})$ is a potential instrument, and therefore assumption (41) implies that we have an infinite number of available instruments.

Optimal instruments

- Notice that if $\mathbf{z}$ is discrete and takes only a finite number of different values, we only have a finite number of orthogonality conditions.
- In the general case when we have an infinite number of orthogonality conditions, the question is whether it is possible to find an optimal finite set of h functions.
- By an optimal finite set of h functions we mean that, the optimal GMM estimator based on the orthogonality conditions implied by this set of h functions is, asymptotically, at least as efficient as any other GMM estimator based on another set of h functions.
- Notice that we have written "the optimal GMM estimator..." and the word "optimal" in this sentence means "based on the optimal choice of the weighting matrix".
- We are first going to address the question of how to optimally choose the h functions when $OLS1'$ is satisfied.
  - Notice that if $OLS1'$ is satisfied, the OLS estimator is consistent.
  - What we are going to see is whether there is an IV estimator that uses functions of $\mathbf{x}$ as instruments that is asymptotically more efficient than OLS.

Optimal instruments

- Consider the linear model

$$y = \mathbf{x}'\boldsymbol{\beta} + u$$

where $\mathbf{x}$ is a $k \times 1$ vector whose first element is 1 and suppose that $OLS1'$ is satisfied, that is

$$E(u \mid \mathbf{x}) = 0$$

- Let $\mathbf{z} = \mathbf{z}(\mathbf{x})$ be a $p \times 1$ ($p \geq k$) vector of functions of $\mathbf{x}$. $OLS1'$ implies that $\mathbf{z}$ is a vector of valid instruments and we know that the variance of the asymptotic distribution of the optimal GMM estimator based on these instruments is

$$\left(E(\mathbf{x}\mathbf{z}')E(u^2\mathbf{z}\mathbf{z}')^{-1}E(\mathbf{z}\mathbf{x}')\right)^{-1}$$

- Let $\sigma^2(\mathbf{x}) = E(u^2 \mid \mathbf{x})$, if we choose

$$\mathbf{z} = \frac{1}{\sigma^2(\mathbf{x})}\mathbf{x} \tag{42}$$

Optimal instruments

- We have that, using the *LIE*,

$$E(u^2\mathbf{z}\mathbf{z}') = E\left(\frac{1}{(\sigma^2(\mathbf{x}))^2}u^2\mathbf{x}\mathbf{x}'\right) = E\left(E\left(\frac{1}{(\sigma^2(\mathbf{x}))^2}u^2\mathbf{x}\mathbf{x}' \mid \mathbf{x}\right)\right)$$
$$= E\left(\frac{1}{(\sigma^2(\mathbf{x}))^2}E\left(u^2 \mid \mathbf{x}\right)\mathbf{x}\mathbf{x}'\right) = E\left(\frac{1}{\sigma^2(\mathbf{x})}\mathbf{x}\mathbf{x}'\right)$$

- Therefore, the variance of the asymptotic distribution of the optimal *GMM* estimator based on this particular set of instruments is

$$\begin{aligned}
& \left(E(\mathbf{x}\mathbf{z}')E(u^2\mathbf{z}\mathbf{z}')^{-1}E(\mathbf{z}\mathbf{x}')\right)^{-1} \\
= & \left(E\left(\frac{1}{\sigma^2(\mathbf{x})}\mathbf{x}\mathbf{x}'\right)E\left(\frac{1}{\sigma^2(\mathbf{x})}\mathbf{x}\mathbf{x}'\right)^{-1}E\left(\frac{1}{\sigma^2(\mathbf{x})}\mathbf{x}\mathbf{x}'\right)\right)^{-1} \\
= & \left(E\left(\frac{1}{\sigma^2(\mathbf{x})}\mathbf{x}\mathbf{x}'\right)\right)^{-1}
\end{aligned}$$

Optimal instruments

- It can be shown that

$$\left(E(\mathbf{x}\mathbf{z}')E(u^2\mathbf{z}\mathbf{z}')^{-1}E(\mathbf{z}\mathbf{x}')\right)^{-1} - \left(E\left(\frac{1}{\sigma^2(\mathbf{x})}\mathbf{x}\mathbf{x}'\right)\right)^{-1} \geq 0$$

  and therefore the optimal choice of instruments is (42).
- Since $z = \frac{1}{\sigma^2(\mathbf{x})}\mathbf{x}$ is a $k \times 1$ vector, the optimal *GMM* is the simple *IV* estimator

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N}\mathbf{z}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{z}_i y_i\right) = \left(\sum_{i=1}^{N}\frac{1}{\sigma^2(\mathbf{x}_i)}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^{N}\frac{1}{\sigma^2(\mathbf{x}_i)}\mathbf{x}_i y_i\right)$$

  and this is the unfeasible *GLS* estimator.
- The feasible *GLS* will be obtained by replacing $\sigma^2(\mathbf{x}_i)$ by an estimate $\widehat{\sigma^2(\mathbf{x}_i)}$, and, under some assumptions, the feasible *GLS* will have the same asymptotic distribution as the unfeasible *GLS*.
- When *OLS*3′ is satisfied, that is when $E(u^2 \mid \mathbf{x}) = \sigma^2$, the optimal instruments are the x themselves and the *OLS* estimator is the optimal *IV* estimator that uses functions of $\mathbf{x}$ as instruments.

Optimal instruments

- In the general case when some of the regressors might be endogenous, if $\mathbf{z}$ is an $L \times 1$ ($L \geq k$) vector that satisfies

$$2SLS1': \qquad E(u \mid \mathbf{z}) = 0$$

- Under some regularity conditions, it can be shown that the optimal set of instruments is

$$\frac{1}{\sigma^2(\mathbf{z})} E(\mathbf{x} \mid \mathbf{z})$$

where $\sigma^2(\mathbf{z}) = E(u^2 \mid \mathbf{z})$.

- Then, the optimal *GMM* estimator is the unfeasible simple *IV* estimator

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \frac{1}{\sigma^2(\mathbf{z}_i)} E(\mathbf{x}_i \mid \mathbf{z}_i) \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^{N} \frac{1}{\sigma^2(\mathbf{z}_i)} E(\mathbf{x}_i \mid \mathbf{z}_i) y_i \right) \tag{43}$$

and the variance of its asymptotic distribution is

$$\left(E\left(\frac{1}{\sigma^2(\mathbf{z})} E(\mathbf{x} \mid \mathbf{z}) E(\mathbf{x} \mid \mathbf{z})' \right) \right)^{-1} \tag{44}$$

Optimal instruments

- The feasible optimal IV estimator requires estimating both $\sigma^2(\mathbf{z})$ and $E(\mathbf{x} \mid \mathbf{z})$ and this can be done non-parametrically.
- The other possibility is to specify a particular functional form for $\sigma^2(\mathbf{z})$ and $E(\mathbf{x} \mid \mathbf{z})$ and to estimate the parameters.
- We are now going to add the homoskedasticity assumption, that is $E(u^2 \mid \mathbf{z}) = \sigma^2$, and we are going to prove that, under homoskedasticity, if $E(\mathbf{x} \mid \mathbf{z})$ is linear in $\mathbf{z}$, the $2SLS$ estimator is the optimal IV estimator that uses functions of $\mathbf{z}$ as instruments.
- *Proof*
  - Since $E(\mathbf{x} \mid \mathbf{z})$ is linear in $\mathbf{z}$, we have that

$$E(\mathbf{x} \mid \mathbf{z}) = \begin{pmatrix} E(x_1 \mid \mathbf{z}) \\ E(x_2 \mid \mathbf{z}) \\ \vdots \\ E(x_k \mid \mathbf{z}) \end{pmatrix} = \mathbf{\Gamma z}$$

where the $j-th$ row of $\mathbf{\Gamma}$, $\boldsymbol{\gamma}_j'$, is the vector of coefficients of the LP of x_j on $\mathbf{z}$.

Optimal instruments

- *Proof (cont.)*
  - We know that
    $$\boldsymbol{\gamma}_j = E(\mathbf{z}\mathbf{z}')^{-1}E\left(\mathbf{z}x_j\right)$$
    and therefore
    $$\boldsymbol{\Gamma} = E(\mathbf{x}\mathbf{z}')E(\mathbf{z}\mathbf{z}')^{-1}$$
  - Substituting $E(\mathbf{x} \mid \mathbf{z}) = \boldsymbol{\Gamma}\mathbf{z} = E(\mathbf{x}\mathbf{z}')E(\mathbf{z}\mathbf{z}')^{-1}\mathbf{z}$ and $\sigma^2(\mathbf{z}) = \sigma^2$ in (44), we have that the variance of the asymptotic distribution of the optimal IV estimator is
    $$\begin{aligned} & \left(E\left(\frac{1}{\sigma^2(\mathbf{z})}E(\mathbf{x} \mid \mathbf{z})E(x \mid \mathbf{z})'\right)\right)^{-1} \\ = & \left(E\left(\frac{1}{\sigma^2}E(\mathbf{x}\mathbf{z}')E(\mathbf{z}\mathbf{z}')^{-1}\mathbf{z}\mathbf{z}'E(\mathbf{z}\mathbf{z}')^{-1}E(\mathbf{z}\mathbf{x}')\right)\right)^{-1} \\ = & \sigma^2\left(E(\mathbf{x}\mathbf{z}')E(\mathbf{z}\mathbf{z}')^{-1}E(\mathbf{z}\mathbf{x}')\right)^{-1} \end{aligned}$$
  - Thus, since this is the variance of the asymptotic distribution of the $2SLS$ estimator under homoskedasticity, we have that, when $E(u^2 \mid \mathbf{z}) = \sigma^2$ and $E(\mathbf{x} \mid \mathbf{z})$ is linear in $\mathbf{z}$, the $2SLS$ estimator is the optimal IV estimator that uses functions of $\mathbf{z}$ as instruments.

Optimal instruments

- When the errors are heteroskedastic, even if $E(\mathbf{x} \mid \mathbf{z})$ is linear in z, the optimal instruments are non-linear functions of $\mathbf{z}$.
- Then, since the “optimal" *GMM* defined in the previous section uses linear functions of $\mathbf{z}$ as instruments, it cannot be the optimal *IV* estimator that uses functions of z as instruments.
- The "optimal" *GMM* defined in the previous section is optimal in the sense that it combines optimally the instruments $\mathbf{z}$ but is non-optimal since it does not use the optimal set of instruments.
- Thus, we have that, if the errors are heteroskedastic, the “optimal" *GMM* estimator defined in the previous section is asymptotically more efficient than 2*SLS*, but inefficient relative to the unfeasible *IV* estimator in (43).

Optimal instruments in summary

- If $E(u \mid \mathbf{z}) = 0$, *any* function $h(\mathbf{z})$ is a valid instrument: there are infinitely many moment conditions.
- The best possible instruments are

$$\frac{1}{\sigma^2(\mathbf{z})} E(\mathbf{x} \mid \mathbf{z}), \qquad \sigma^2(\mathbf{z}) = E(u^2 \mid \mathbf{z}).$$

- Special cases you already know:
  - No endogeneity ($\mathbf{z} = \mathbf{x}$), heteroskedasticity: the optimal IV estimator is **GLS** (Unit 2).
  - Homoskedasticity and $E(\mathbf{x} \mid \mathbf{z})$ linear: the optimal IV estimator is **2SLS**.
- With heteroskedasticity, "optimal GMM" combines the linear instruments $\mathbf{z}$ optimally, but the fully optimal instruments are non-linear functions of $\mathbf{z}$.

$E(\mathbf{x} \mid \mathbf{z})$ is a prediction problem. This is where flexible first stages (and, with care, machine learning) enter IV.

More moments than parameters in our case

- With both arms, 2SLS uses these moment conditions for the one unknown β_1 (controls aside):

  $$E(letter \cdot u) = 0, \qquad E(software \cdot u) = 0.$$

- Two equations, one unknown: in a sample we cannot make both sample moments exactly zero.
- So we **minimise** a weighted distance of the sample moments from zero. That is a loss function, and choosing its weights is what GMM is about.

Exercise. Small firms measure their gap with very few workers, so their u is much noisier. Should every firm's moment count the same?

More moments than parameters in our case

- With both arms, 2SLS uses these moment conditions for the one unknown β_1 (controls aside):

$$E(letter \cdot u) = 0, \qquad E(software \cdot u) = 0.$$

- Two equations, one unknown: in a sample we cannot make both sample moments exactly zero.
- So we **minimise** a weighted distance of the sample moments from zero. That is a loss function, and choosing its weights is what GMM is about.

Exercise. Small firms measure their gap with very few workers, so their u is much noisier. Should every firm's moment count the same?

Optimal GMM weights each moment by its inverse variance: trust the precise moments more. The same logic as precision-weighted averaging and inverse-variance ensembles.

Analysis C1 · Small firms, noisy gaps

2SLS residuals by firm size

Employees	Firms	s.d. of $\widehat{u}$
10–25	965	0.082
26–50	1,901	0.067
51–100	1,428	0.055
more than 100	706	0.047

- The gap is an average over a firm's workers. With 5 women, one hire or departure moves it a lot.
- The residual s.d. is **1.8 times** larger in the smallest firms: textbook heteroskedasticity, $E(u^2\mathbf{z}\mathbf{z}') \neq \sigma^2 E(\mathbf{z}\mathbf{z}')$.
- So 2SLS is no longer the efficient GMM estimator, and robust standard errors are a must.

Analysis C2 · Two-step efficient GMM

Coefficient on `transp`*, instruments:* `letter software`

	$\widehat{\beta}_1$	s.e.
`ivreg`, robust	−0.0408	0.0098
`gmm` (two-step)	−0.0408	0.0098
Hansen $J = 0.14$, 1 d.f., $p = 0.71$		

The `gmm` package rebuilds the weight matrix at the final estimate: s.e. may differ in the 4th decimal and J is 61.7 instead of 61.3. Same conclusion.

J asks: do the two arms tell the same story? Here they do.

- Same estimate to the 4th decimal: **no efficiency gain here**.
- Why? The lottery makes the instruments unrelated to firm size, hence to the variance of u. Reweighting the moments changes almost nothing.
- Gains appear when the variance of u is related to the instruments. In small samples an estimated $\widehat{\mathbf{W}}$ can even hurt.

Analysis C3 · Adding a bad instrument

gmm(dgap ~ transp + X, ~ z + X), then specTest

Instruments	$\widehat{\beta}_1$	Hansen J
`letter software`	−0.041	0.14 ($p = 0.71$)
`letter software peer_share`	−0.083	**61.3** ($p < 10^{-13}$)

- Adding the invalid instrument doubles the estimated effect.
- But now the moments **contradict each other**: no β_1 can make all three close to zero, and J explodes.

With more instruments than regressors, part of the exogeneity assumption *can* be tested. The test only has power if at least one instrument is valid, and it cannot tell you *which* one is not.

Block C in R

```
library(gmm)
# C1: residuals by firm size
d$uhat <- resid(iv2)
d$size_class <- cut(d$size, c(0, 25, 50, 100, Inf))
aggregate(uhat ~ size_class, d, sd)

# C2: two-step efficient GMM
fy <- as.formula(paste("dgap ~ transp +", X))
g  <- gmm(fy, as.formula(paste("~", Z, "+", X)),
          data = d, vcov = "MDS",
          centeredVcov = FALSE)
summary(g)
specTest(g)                  # Hansen J

# C3: a bad instrument
g3 <- gmm(fy,
          as.formula(paste("~", Z, "+ peer_share +", X)),
          data = d, vcov = "MDS",
          centeredVcov = FALSE)
specTest(g3)
```

1. Compare `coef(g)` with `coef(iv2)`. They agree to four decimals: why is the efficiency gain zero here?
2. Replace `peer_share` by `parent_abroad` in C3. Does J catch it too? Why is it harder?

Summary of GMM

1. GMM minimises a weighted distance of the sample moments from zero; 2SLS is GMM with $\widehat{\mathbf{W}} = (\mathbf{Z}'\mathbf{Z}/N)^{-1}$.
2. The optimal weighting matrix is $E(u^2\mathbf{z}\mathbf{z}')^{-1}$: 2SLS is optimal under homoskedasticity, two-step GMM otherwise.
3. The efficiency gain can be negligible, as here with randomised instruments.
4. With $L > k$ the moments can contradict each other; the J statistic measures how much (Unit 4).