Single and Multiple Equation GMM: What do we do in linear model when orthogonality no longer holds

Motivation: In our previous linear model, the most important assumption we made is the orthogonality between error term and regressors (i.e. Strict exogeneity or predetermined regressors), without which the OLS estimator is not even consistent for the desired β (i.e. from our model $y_i = x_i' \beta + \varepsilon_i$) → **Endogeneity Bias**!

Since in economics the orthogonality condition is not satisfied, we develop methods here to deal with **endogenous regressors**, called Generalized Method of Moments (GMM), which includes OLS as a special case.

Example 1:

Single Equation GMM: Here we relax assumptions even further and take away the predetermined regressor assumption. Instead, we have orthogonality condition from instruments.

I. Assumptions:

3.1 **Linearity**: The data we observe comes from underlying RV's $\{ y_{i\ (1x1)}, \mathbf{x}_{i\ (1xd)}\}$ with
$y_i = \mathbf{x}_i'\delta + \varepsilon_i$ ($i = 1,2,\ldots,n$) (this is the equation we want to estimate)

3.2 **Ergodic Stationarity**:
Let $\mathbf{z}_i$ be a M-dimensional vector of instruments, and let $\mathbf{w}_i$ be the unique and nonconstant elements of ($y_i, \mathbf{x}_i, \mathbf{z}_i$).
$\{\mathbf{w}_i\}$ is jointly stationary and ergodic.

3.3 Orthogonality Condition
All the M variables in $\mathbf{z}_i$ are predetermined in the sense that they are all orthogonal to the current error term:
$E(z_{im}\varepsilon_i) = 0\ \forall i, \forall m \Leftrightarrow E[(y_i - \mathbf{x}_i'\beta) . \mathbf{z}_i] = 0 \Leftrightarrow E(\mathbf{g}_i) = \mathbf{0}$ where $\mathbf{g}_i = \mathbf{z}_i . (y_i - \mathbf{x}_i'\beta) = \mathbf{z}_i . \varepsilon_i$

Note on Moment Conditions: $E[(y_i - \mathbf{x}_i'\beta)_{1x1} . \mathbf{z}_{i\ Mx1}]_{Mx1} = 0$ → These are the **M moment conditions**

Note on Instruments vs. Regressors: Even though we denote regressors and instruments by $\mathbf{x}_i$ and $\mathbf{z}_i$, this does not mean that they do not share the same variables. Not true! Regressors that are predetermined are instruments, and regressors that are not predetermined are endogenous regressors.

Note on 1 as an Instrument: Typically we will include 1 as an instrument → $\mathbf{E(\varepsilon_i) = 0}$!

3.4 Rank Condition for Identification : Guarantees there's a unique solution to the system of equations[1]
The m x d matrix $E(\mathbf{z}_i\ \mathbf{x}_i')_{mxd}$ is of full column rank (or $E(\mathbf{x}_i\ \mathbf{z}_i')$ is of full row rank), **M $\geq$ d (# of equations $\geq$ # of unknowns)**.
We denote this matrix by $\mathbf{\Sigma_{ZX}}$.

3.5 **Martingale Difference with Finite Second Moments**: Assumption for Asymptotic Normality
$\mathbf{g}_i = \mathbf{x}_i . \varepsilon_i$ is a martingale difference sequence with finite second moments → **$\mathbf{g}_i$ is the sequence of moment conditions**!
$\{ \mathbf{g}_i \}$ is a martingale difference sequence (so $E(\mathbf{g}_i) = \mathbf{0}$ with $E(\mathbf{g}_i \mid \mathbf{g}_{i-1}, \mathbf{g}_{i-2},\ldots, \mathbf{g}_1) = \mathbf{0}$ for $i \geq 2$) → no serial correlation in $\mathbf{g}_i$
The KxK matrix of cross moments, $E(\mathbf{g}_i\ \mathbf{g}_i')$ is nonsingular.

→ so, $\bar{g}_{nx1} = \frac{1}{n}\sum_{i=1}^{n} g_i \xrightarrow{N} N(0, E(g_i g_i')),\ S \equiv A\text{var}(\bar{g}_{nx1}) = E(g_i g_i')$ by 3.2 (so g_i ergodic stationary) and Ergodic Differences CLT

Note: Again, if $\mathbf{z}_i$ includes a constant term → $\boldsymbol{\varepsilon_i}$ is MDS → no autocorrelation in $\boldsymbol{\varepsilon_i}$
Note: The same 4 comments here apply as in 2.5[2]

Add'l:

3.6 **Finite fourth moments for Regressors**: (For consistent estimation of **S**)
$E[(\mathbf{x}_{ik}\mathbf{z}_{ij})^2]$ exists and is finite for all k = 1,…, d,j = 1,2, …, m

3.7 **Conditional Homoskedasticity**
$E(\varepsilon_i^2 \mid \mathbf{z}_i) = \sigma^2$

---

[1] This is called the rank condition for identification for the following reason (see proof: the condition guarantees unique min)
We can rewrite the moment/orthogonality condition as a system of K simultaneous equations:
$E[\mathbf{g}(\mathbf{w}_i;\boldsymbol{\delta}_{\delta\xi 1})] = \mathbf{0}_{mx1}$ where $\mathbf{g}(\mathbf{w}_i;\boldsymbol{\delta}) = \mathbf{x}_i . (y_i - \mathbf{z}_i'\boldsymbol{\delta})$, $\mathbf{w}_i$ is the unique and nonconstant elements of ($y_i, \mathbf{x}_i, \mathbf{z}_i$), and $\boldsymbol{\delta}_{\delta\xi 1}$ is the parameter vector / coefficient vector

The moment condition means that the "true" value of the coefficient vector $\boldsymbol{\delta}_{\delta\xi 1}$ is ***a* solution** to this system of K simultaneous equations. Assumptions 3.1 – 3.3. guarantees that there exists a solution to the moment conditions, but if the coefficient vector (or the equation) is identified if **there is a unique solution to the moment condition**. A necessary and sufficient condition for a unique solution in the system of simultaneous equations is that $\mathbf{\Sigma_{XZ}}$ has full column rank.

Derivation: Want unique minimization of $E[\mathbf{g}(\mathbf{w}_i;\boldsymbol{\delta}_{\delta\xi 1})]$. Let β be a minimizer.
β *unique iff* $\forall\ b \neq \beta,\ E_P[g(w_i;b)] \neq E_P[g(w_i;\beta)]$ *where P is the underlying distribution that generated the data*
$\Leftrightarrow E_P[z_i \cdot (y_i - x_i'b)] \neq E_P[z_i \cdot (y_i - x_i'\beta)] \Leftrightarrow E_P[z_i \cdot (x_i'b)] \neq E_P[z_i \cdot (x_i'\beta)] \Leftrightarrow E_P[z_i \cdot (x_i'b)] - E_P[z_i \cdot (x_i'\beta)] \neq 0$
$\Leftrightarrow E_P[z_i \cdot (x_i'b - x_i'\beta)] \neq 0 \Leftrightarrow E_P[z_i \cdot x_i']_{mxd}(b - \beta) \neq 0 \Leftrightarrow E_P[z_i \cdot x_i']_{mxd} = \Sigma_{XZ}$ *has full column rank with* $m \geq d$
If not, then there exists a nonzero vector b' – β such that (b' – β) is in the $\text{Ker}(E_P[\mathbf{z}_i\mathbf{x}_i])$. i.e. The columns are linearly dependent, so there exists a non-trivial linear combination (b' – β) such that $E_P[\mathbf{z}_i\mathbf{x}_i]$ (b' – β) = 0 (Minimizer not unique!)

Order Condition for Identification - m $\geq$ d
A necessary condition (embedded in the proofs here) is that **m $\geq$ d (# of equations $\geq$ # of unknowns)→ order condition for identification**[1]
We can interpret this in 3 ways: #predetermined vars ≥ #regressors or #orthogonality conditions ≥ #parameters or # orthogonality conditions ≥ #parameters
If order condition is not satisfied, then the equation or parameter is not identified.

We say that the equation is…

1. **Overidentified** if the rank condition is satisfied and m > d
2. **Exactly identified / just identified** if the rank condition is satisfied and m = d
3. **Underidentified / not identified** if the rank condition is not satisfied or m < d

[2] IID is a special case. If iid, we only need Lindberg-Chevy CLT
If instruments include a constant, then the error term is a martingales sequence (and a fortiori serially uncorrelated)
The assumption is hard to interpret, so we interpret an easier/sufficient condition: $E(\varepsilon_i \mid \varepsilon_{i-1}, \varepsilon_{i-2},\ldots, \varepsilon_1, \mathbf{z}_z, \mathbf{z}_{i-1},\ldots, \mathbf{z}_1) = 0$
→ Besides being a martingales sequence and therefore uncorrelated with itself, the error term is orthogonal not only to the current but also to the past instruments
Since $\mathbf{g}_i\mathbf{g}_i' = \varepsilon_i^2\ \mathbf{z}_i\mathbf{z}_i'$, $\mathbf{S} = E(\mathbf{g}_i\mathbf{g}_i')$ is a matrix of 4th moments, so consistent estimation of **S** will require a 4th moment assumption (assumption 3.6)

II. What do Assumptions Imply about Properties of Instruments?

1. Orthogonality condition → Instruments are orthogonal to errors (and uncorrelated if 1 is included as an instrument, i.e. $E(\varepsilon_i) = 0$)
2. Rank Condition → (non-constant) Instruments are correlated with the endogenous variables[3]
3. Instruments uncorrelated with

Note on Orthogonality vs. Covariance:
We can always think of orthogonality between error and instrument as covariance provided that one of them is de-meaned.
Cov(x,e) = Cov(x,e-E(e)) = E(x(e-E(e)) and similarly we can rewrite model as: $y = b_0+E(e) + b_1\, x + [e – E(e)]$

III. Using 1 as an instrument: 2 Assumptions Made

It is common practice to use 1 as an instrument, however, doing so we are making 2 important assumptions:

1. $\mathbf{E(\varepsilon_i) = 0}$ (this comes from the orthogonality condition)
2. ε_i is an MDS → no serial correlation in ε_i

---

[3] *Suppose regressors are* $\begin{pmatrix}1 & x_i & z_i\end{pmatrix}$, x_i *endogenous, instruments are* $\begin{pmatrix}1 & z_{2i} & z_i\end{pmatrix}$

Then, moment condition is $E\left(\begin{bmatrix}1\\ z_{2i}\\ z_i\end{bmatrix}\begin{bmatrix}1 & x_i & z_i\end{bmatrix}\right) = \begin{pmatrix}1 & Ex_i & Ez_i\\ Ez_{2i} & Ez_{2i}x_i & Ez_{2i}z_i\\ Ez_i & Ez_ix_i & Ez_i^2\end{pmatrix} \equiv \mathbf{A}$

If $\operatorname{cov}(z_{2i}, x_i) = 0$, *then* 1*st column of* $\mathbf{A}$ *times* $E(x_i) =$ 2*nd column of* $\mathbf{A}$

IV. **Generalized Method of Moments Defined**: We'll show that IV estimator is a special GMM estimator (i.e. exactly identified system)

1. General Setup: The true parameter of interest is the solution to the moment conditions

$$\beta \text{ s.t. } E[g(W_i,b)]=0 \Leftrightarrow \beta = \arg\min E[g(W_i,b)]' W\, E[g(W_i,b)] \text{ for some p.d. weighting matrix}$$
$$\Leftrightarrow \beta = \arg\max - E[g(W_i,b)]' W\, E[g(W_i,b)]$$

By Analogy Principle, $\hat{b}_{GMM} = \arg\max_{b\in\Theta} -\left[\frac{1}{n}\sum_{i=1}^{n} g(W_i,b)\right]' W_n \left[\frac{1}{n}\sum_{i=1}^{n} g(W_i,b)\right]$

where sample moment conditions are defined to be $g_n(W_i;b) = \frac{1}{n}\sum_{i=1}^{n} g(W_i,b)$

2. Applied to Linear Model: Our model is $y_i = \mathbf{x}_i'\delta + \varepsilon_i$ with the moment condition $E[(y_i - \mathbf{x}_i'\beta) \,.\, \mathbf{z}_i] = 0$
   - Expression for sample moment condition in linear model

$$g_n(W_i,b) = \frac{1}{n}\sum_{i=1}^{n} z_i\cdot(y_i - x_i'b) = \frac{1}{n}\sum_{i=1}^{n} z_i\cdot y_i - \frac{1}{n}\sum_{i=1}^{n} z_i\cdot x_i'b = \frac{1}{n}\sum_{i=1}^{n} z_i\cdot y_i - \frac{1}{n}\sum_{i=1}^{n} z_{i(mx1)}\cdot x'_{i(1xm)}b = \frac{1}{n}\sum_{i=1}^{n} z_i\cdot y_i - \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right)b$$
$$\equiv s_{ZY(mx1)} - S_{ZX(mxd)}b_{(dx1)}$$

   - Method of Moments: If the equation is exactly identified m = d, and there exists a unique **b** such that $g_n(\mathbf{W}_i,\mathbf{b}) = 0$, and Σ_{zx} invertible, then for sufficiently large n, $S_{ZX} \to_p \Sigma_{zx}$ by ergodic theorem and is invertible with prob. 1. So, with large sample size the system of simultaneous equation has unique solution given by…

The MM estimator

$$\boxed{\hat{b}_{IV} = (S_{ZX})^{-1} s_{ZY} = \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i\cdot y_i\right)}$$
(**IV estimator** with $\mathbf{z}_i$ as instruments, def. for m = d)[4]

   - General Method of Moments:
     If the equation overly identified m > d, and there does not exists a unique **b** such that $g_n(\mathbf{W}_i,\mathbf{b}) = 0$ exactly, we choose **b** to minimize $g_n(\mathbf{W}_i,\mathbf{b})$ or equivalently $g_n(\mathbf{W}_i,\mathbf{b})' W_n\, g_n(\mathbf{W}_i,\mathbf{b})$ (for some choice of p.d. weighting matrix that converges in probability to some p.d. W)[5]

$$\boxed{\hat{b}_{GMM} = \arg\max_{b\in R^N} -\left[\frac{1}{n}\sum_{i=1}^{n} g(W_i,b)\right]' W_{n(mxm)} \left[\frac{1}{n}\sum_{i=1}^{n} g(W_i,b)\right]}$$

Note: If the equation is just identified, then regardless of the weighting matrix, the GMM estimator = the IV estimator numerically

3. GMM Estimator and Sampling Error
   - GMM Estimator:
     $\mathbf{b}_{GMM} = (\mathbf{Szx'Wn\,Szx})^{-1}\, \mathbf{Szx'W_n s_{ZY}}$ [6]
     (By 3.2 and 3.4, S_{xz} has full column rank for sufficiently large n with prob 1, then **Szx'Wn Szx** invertible) [7]
     If m = d, then **Szx** is a mxm p.d. square matrix and the GMM est. reduces to the IV estimator regardless of Wn![8]
     (P.d. by full column rank assumption → full rank) Also note that this means **OLS is just an IV estimator**!
   - Sampling Error: Our model is $y_i = \mathbf{x}_i'\delta + \varepsilon_i$
     $\mathbf{b}_{GMM} - \delta = (\mathbf{Szx'Wn\,Szx})^{-1}\, \mathbf{Szx'W_n s_{Z}\varepsilon}$ [9]

---

[4] The IV estimator is defined for the EXACTLY IDENTIFIED case (i.e. the case where there are as many instruments as endogenous regressors). If $\mathbf{z}_i = \mathbf{x}_i$, i.e. all the regressors are predetermined/orthogonal to the error contemporaneous term, then this boils down to the OLS estimator. So, OLS is a special case of MM estimator. And, IV and OLS are both special cases of GMM

[5] The quadratic form formulation $g_n(\mathbf{W}_i,\mathbf{b})' W_n\, g_n(\mathbf{W}_i,\mathbf{b})$ gives us a 1x1 real number over which we can define the minimization/maximization problem. Otherwise it would be impossible to minimize over a m-dimensional vector of moment conditions $g_n(\mathbf{W}_i,\mathbf{b})$zd

$$\hat{b}_{GMM} = \arg\max_{b\in R^n} -\frac{1}{2} g_n(W_i,b)' W_n\, g_n(W_i,b) = \arg\max_{b\in R^n} -\frac{1}{2}(s_{ZY} - S_{ZX}b)' W_n (s_{ZY} - S_{ZX}b)$$

[6] $FOC: \text{Assume interiority, } \frac{\partial}{\partial b} -\frac{1}{2}\left(s_{ZY(mx1)} - S_{ZX}\hat{b}\right) W_n \left(s_{ZY(mx1)} - S_{ZX}\hat{b}\right) = 0 \Rightarrow S_{ZX}' W_n \left(s_{ZY} - S_{ZX}\hat{b}\right) = 0 \Rightarrow S_{ZX}' W_n s_{ZY} = S_{ZX}' W_n S_{ZX}\hat{b}$

By assumption 3.2 *and* 3.4, S_{XZ} *is of full column rank for sufficiently big n with prob* 1, *then* $S_{XZ}' W_n S_{XZ}$ *invert.* (W_n *p.d.*)

$\therefore \hat{b}_{GMM} = (S_{ZX}' W_n S_{ZX})^{-1} S_{ZX}' W_n s_{ZY}$

[7] Claim: if A_{mxn} with m≥ n has full column rank, Wmxm p.d., then A'WA invertible.
Proof: Suppose not. Then there exists non-zero nx1 vector c s.t. c'A'WAc = 0 → there exists a mx1 vector d = Ac, d nonzero (since A has full column rank so there are no nontrivial linear combination of columns that give zero vector), and d'Wd = 0. Contradiction to the assumption that W is p.d.!

[8] $b_{GMM} = (S_{ZX}' W_n S_{ZX})^{-1} S_{ZX}' W_n s_{ZY} = (S_{ZX}^{-1} W_n^{-1} (S_{ZX}')^{-1}) S_{ZX}' W_n s_{ZY} = S_{ZX}^{-1} s_{ZY} = b_{IV}$ $\text{ Note}: (S_{ZX}' W_n S_{ZX})^{-1} = (S_{ZX}^{-1} W_n^{-1} (S_{ZX}')^{-1}) \text{ since } S_{ZX}' W_n S_{ZX} S_{ZX}^{-1} W_n^{-1} (S_{ZX}')^{-1} = I$

V. **Large Sample Properties of GMM and Efficient GMM**

(We established these results more generally before using the multivariate mean value theorem. Here we can be more specific now because we impose linearity – our parameter is linear in $\mathbf{x}_i$. Also note that we index the estimator by the weight matrix W_n to denote sample size)

1. Consistency[10]: Under assumptions 3.1 – 3.4, $\mathbf{b}_{GMM}(\mathbf{W}_n) \to_p \delta$

2. Asymptotic Normality[11]: Under assumptions 3.1 – 3.5,

$$\sqrt{n}\left(b_{GMM}(W_n)-\delta\right) \xrightarrow{D} N(0,V)$$
$$\text{where } V = A\text{var}\left(b_{GMM}(W_n)\right) = \left(\Sigma'_{XZ} W \Sigma_{XZ}\right)^{-1} \Sigma'_{XZ} W S W \Sigma_{XZ} \left(\Sigma'_{XZ} W \Sigma_{XZ}\right)^{-1}$$
$$= \left(E(x_i z_i')'W E(x_i z_i')\right)^{-1} E(x_i z_i')' W E(g_i g_i') W E(x_i z_i') \left(E(x_i z_i')'W E(x_i z_i')\right)^{-1} \quad \text{where } W = p\lim W_n$$

3. Consistent Estimate of Avar($\mathbf{b}_{GMM}(\mathbf{W}_n)$)[12]:
Suppose there exists a consistent estimator **S*** of $\mathbf{S}_{\mathbf{mxm}} = E(\mathbf{g}_i\mathbf{g}_i')$. Then, under 3.2, Avar($\mathbf{b}_{GMM}(\mathbf{W}_n)$) is consistently estimated by

$$\hat{A}\text{var}(\hat{b}_{GMM}) \equiv \left(S_{ZX}'W_n S_{ZX}\right)^{-1} S_{ZX}'W_n \hat{S} W_n S_{ZX} \left(S_{ZX}'W_n S_{ZX}\right)^{-1}$$

4. Consistency of s^2 (estimation of variance of "true" error is consistent)[13]:
For any consistent estimator $\hat{b}(W_n)$ *of* δ, *define* $\hat{\varepsilon}_i \equiv y_i - x_i'\hat{b}(W_n)$. *Under* 3.1, 3.2, *and assume* $E(x_i x_i')$ *exists and is finite*,

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \xrightarrow{P} E\left(\varepsilon_i^2\right) \text{ provided } E\left(\varepsilon_i^2\right) \text{ exists and is finite}$$

5. Consistent estimation of **S**: We've assumed S* exists thus far- How do we obtain consistent estimator $\hat{S}$ of S_{kxk} from the sample (***y***,***X***)? Suppose the coefficient estimate $\hat{b}$ used for calculating the residual $\hat{\varepsilon}_i$ for $\hat{S}$ is consistent for δ, and suppose $\mathbf{S} = E(\boldsymbol{g}_i\boldsymbol{g}_i')$ exists and is finite. Then, under assumptions 3.1, 3.2, and 3.6, $\hat{S} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 z_i z_i'$ is consistent for **S**.[14]

9

$$\hat{b}_{GMM} = \left(S_{ZX}'W_n S_{ZX}\right)^{-1} S_{ZX}'W_n s_{ZY} = \left(S_{ZX}'W_n S_{ZX}\right)^{-1} S_{ZX}'W_n \left(\frac{1}{n}\sum_{i=1}^{n} z_i \cdot y_i\right) = \left(S_{ZX}'W_n S_{ZX}\right)^{-1} S_{ZX}'W_n \left(\frac{1}{n}\sum_{i=1}^{n} z_i \cdot (x_i'\delta + \varepsilon_i)\right) \text{ by } 3.1$$
$$= \left(S_{ZX}'W_n S_{ZX}\right)^{-1} S_{ZX}'W_n \left(\frac{1}{n}\sum_{i=1}^{n} z_i \cdot (x_i'\delta)\right) + \left(S_{ZX}'W_n S_{ZX}\right)^{-1} S_{ZX}'W_n \left(\frac{1}{n}\sum_{i=1}^{n} z_i \cdot \varepsilon_i\right) = \delta + \left(S_{ZX}'W_n S_{ZX}\right)^{-1} S_{ZX}'W_n s_{z\varepsilon}$$

10 From above, since

$$\left(\frac{1}{n}\sum_{i=1}^{n} z_i \cdot \varepsilon_i\right) = \bar{g}_n \xrightarrow{P} E\left(z_i \cdot \varepsilon_i\right) = E(g_i) = 0 \text{ by ergodic theorem and } 3.3$$
$$\therefore \hat{b}_{GMM} \xrightarrow{P} \delta$$

11 Continuing from above,

$$\hat{b}_{GMM} - \delta = \left(S_{ZX}'W_n S_{ZX}\right)^{-1} S_{ZX}'W_n s_{x\varepsilon} \Rightarrow \sqrt{n}\left(\hat{b}_{GMM} - \delta\right) = \left(S_{ZX}'W_n S_{ZX}\right)^{-1} S_{ZX}'W_n \sqrt{n} s_{z\varepsilon}$$
$$\sqrt{n}s_{x\varepsilon} = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i \cdot \varepsilon_i\right) \xrightarrow{D} N(0, E(g_i g_i')) \text{ by Ergodic Martingale Differences CLT}, S_{ZX} \xrightarrow{P} \Sigma_{XZ} \text{ by Ergodic Theorem}, W_n \xrightarrow{P} W \text{ by construction}$$
$$\therefore \sqrt{n}\left(\hat{b}_{GMM} - \delta\right) = \left(S_{ZX}'W_n S_{ZX}\right)^{-1} S_{ZX}'W_n \sqrt{n}s_{z\varepsilon} \xrightarrow{D} N\left(0, \left(\Sigma_{ZX}'W_n \Sigma_{ZX}\right)^{-1} \Sigma_{ZX}'W\, E(g_i g_i')W'\Sigma_{ZX}\left(\Sigma_{ZX}'W_n\Sigma_{ZX}\right)^{-1}\right) \text{ by CMT and Slutsky's}$$

12 This follows from above. Standard asymptotic tools.

13 This proof is very similar to 3D from previous notes

$$\hat{\varepsilon}_i = y_i - x_i'\hat{b} = y_i - x_i'\hat{b} + x_i'\delta - x_i'\delta = \varepsilon_i + x_i'(\hat{b}-\delta) \Rightarrow e_i^2 = \varepsilon_i^2 + 2\varepsilon_i x_i'(\hat{b}-\delta) + (\hat{b}-\delta)'x_i x_i'(\hat{b}-\delta)$$
$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 + 2\varepsilon_i x_i'(\hat{b}-\delta) + (\hat{b}-\delta)'x_i x_i'(\hat{b}-\delta) = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 + 2(\hat{b}-\delta)\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i x_i'\right) + (\hat{b}-\delta)'\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)(\hat{b}-\delta) = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 + 2(\hat{b}-\beta)'s_{x\varepsilon} + (\hat{b}-\beta)'S_{XX}(\hat{b}-\beta)$$
$$2(\hat{b}-\beta)'s_{x\varepsilon} \xrightarrow{P} 0 \text{ since } s_{x\varepsilon} \xrightarrow{P} \text{some finite vector and } \hat{b} \xrightarrow{P} \beta$$
$$(\hat{b}-\beta)'S_{XX}(\hat{b}-\beta) \xrightarrow{P} 0 \text{ since } \hat{b}-\beta \xrightarrow{P} 0 \text{ and } S_{XX} \xrightarrow{P} \Sigma_{XX} \text{ finite by assumption}$$

14 (Proof is similar to 4 – multiple $z_i z_i'$ on both sides!)

VI. **Efficient GMM**: How do we choose W_n to minimize Avar($\mathbf{b}_{GMM}(\mathbf{W}_n)$)? Let $W_n = E(g_i g_i')^{-1}$ → inverse of variance of moment conditions

1. **Efficient weighting matrix** is given by $W_n^* = \mathbf{S}^{-1} = E(\mathbf{g}_i \mathbf{g}_i')^{-1}$

   For any weighting matrix $\mathbf{W}_n$, Avar($\mathbf{b}_{GMM}(\mathbf{W}_n)$) ≥ Avar($\mathbf{b}_{GMM}(\mathbf{W}_n^*)$)= $[\Sigma_{zx}' W_n^* \Sigma_{zx}]^{-1} = [\Sigma_{zx}' S^{-1} \Sigma_{zx}]^{-1} = [E(\mathbf{x}_i\mathbf{z}_i')' (E(\mathbf{g}_i\mathbf{g}_i')^{-1}) E(\mathbf{x}_i\mathbf{z}_i')]^{-1}$

$$\hat{b}_{GMM}^{eff}(\hat{S}^{-1}) = \left(\mathbf{S}_{ZX}'\hat{\mathbf{S}}^{-1}\mathbf{S}_{ZX}\right)^{-1}\left(\mathbf{S}_{ZX}'\hat{\mathbf{S}}^{-1}\mathbf{s}_{ZY}\right)$$

2. Large Sample Properties of Efficient GMM Estimator

   From above, efficient GMM is consistent, asymptotically normal with the following asymptotic variance and its consistent estimate:

$$A\text{var}\left(\hat{b}_{GMM}^{eff}(\hat{\mathbf{S}}^{-1})\right) = \left(\Sigma_{ZX}' S^{-1}\Sigma_{ZX}\right)^{-1}$$
$$\hat{A}\text{var}\left(\hat{b}_{GMM}^{eff}(\hat{\mathbf{S}}^{-1})\right) = \left(S_{ZX}'\hat{S}^{-1}S_{ZX}\right)^{-1}$$

3. Hypothesis Testing: Robust t-Ratio and Wald Statistic

   From below, the formulas for robust t and Wald statistics become:

$$t_l = \frac{\hat{b}(\hat{S}^{-1})_l - \overline{\beta}_l}{Robust\ SE_l^*} \quad where\ Robust\ SE_l^* = \sqrt{\frac{1}{n}\left(\left(S_{ZX}'\hat{S}^{-1}S_{ZX}\right)^{-1}\right)_{ll}}$$

$$W = n\cdot a\left(\hat{b}(\hat{S}^{-1})\right)'\left\{A\left(\hat{b}(\hat{S}^{-1})\right)\left(S_{ZX}'\hat{S}^{-1}S_{ZX}\right)^{-1}A\left(\hat{b}(\hat{S}^{-1})\right)'\right\}^{-1}a\left(\hat{b}(\hat{S}^{-1})\right)$$

4. **2-Step Efficient GMM Procedure**: How do we construct an Efficient GMM estimator?

   A. Pick some arbitrary weighting matrix (e.g. I) (that converges in probability to a symmetric p.d. W) and obtain a preliminary consistent GMM estimator $\mathbf{b}_{GMM} = (\mathbf{Szx'Wn\ Szx})^{-1}\ \mathbf{Szx'W_n s_{ZY}}$, which we will use to construct the optimal W

   Note: Usually we set $\mathbf{W}_n = \mathbf{S_{XX}}^{-1}$. Then $\mathbf{b}_{GMM}(\mathbf{S_{XX}}^{-1}) = (\mathbf{Szx'Szz^{-1}Szx})^{-1}\ \mathbf{Szx'}\ \mathbf{Szz^{-1}s_{ZY}}$ →This is the 2SLS

   Then, using the preliminary estimator $\mathbf{b}_{GMM}(\mathbf{S_{ZZ}}^{-1})$ we construct $\hat{S}^{-1}$ with $\hat{S} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 z_i z_i'$

   B. In the second step, the efficient GMM estimator is obtained as:

$$\hat{b}_{n,GMM}^{eff} = \left[S_{ZX}'\hat{S}^{-1}S_{ZX}\right]^{-1}S_{ZX}'\hat{S}^{-1}s_{ZY}\ \ with\ \hat{A}\text{var}\left(\hat{b}_{n,GMM}^{eff}\right) = \left(S_{ZX}'\hat{S}^{-1}S_{ZX}\right)^{-1}$$ *(FIX THE AVAR)*

   Or in matrix notation: $\hat{b}(\hat{S}^{-1}) = [X'Z(Z'BZ)^{-1}Z'X]X'Z(Z'BZ)^{-1}Z'y \qquad with\ \ B \equiv \begin{bmatrix}\hat{\varepsilon}_1^2 & & \\ & \ddots & \\ & & \hat{\varepsilon}_n^2\end{bmatrix}$ (in 2SLS, $B = \sigma^2 I$)

   Note: With this notation wan see that the efficient GMM is a GLS estimator!

5. Note on Small Sample Properties

   The efficient GMM estimator uses $\hat{S}^{-1}$, a function of estimated fourth moments, as weighting matrix. **Generally, it takes a substantially larger sample size to estimate fourth moments reliably** (compared to 1st and 2nd moments). Therefore, the **efficient GMM estimator has poorer small-sample properties than the GMM estimators that do not use fourth moments for $\mathbf{W_n}$.** <u>Equally weighted GMM estimator with W_n = I generally outperforms the efficient GMM in terms of the bias and variance in finite samples!</u>

VII. **Hypothesis Testing**

Prop: Robust T-Ratio and Wald Statistic (Testing Linear and NonLinear Restrictions)

Suppose Assumptions 3.1 – 3.5 hold, and suppose there is available a consistent estimate $\hat{S}$ of **S**.

Then, by above, $\hat{A}\text{var}(\hat{b}) \equiv \left(S_{ZX}^{'}\hat{W}S_{ZX}\right)^{-1} S_{ZX}^{'}\hat{W}\hat{S}\hat{W}S_{ZX}^{'}\left(S_{ZX}^{'}\hat{W}S_{ZX}\right)^{-1}$,

And…

(a) Under the null hypothesis $H_0 : \beta_k = \overline{\beta}_k$,

$$t_k \equiv \frac{\sqrt{n}\left(\hat{b}_k(\hat{W}) - \overline{\beta}_k\right)}{\sqrt{\hat{A}\text{var}(\hat{b}\ (\hat{W}))_{kk}}} = \frac{\hat{b}_k - \overline{\beta}_k}{\sqrt{\frac{1}{n}\hat{A}\text{var}(\hat{b}\ (\hat{W}))_{kk}}} = \frac{\hat{b}_k - \overline{\beta}_k}{Robust\ S.E.^*(\hat{b}_k)} \rightarrow_D N(0,1)$$ [15]

This t-ratio is the **robust t-ratio** because it uses the S.E. that is **robust to errors that can be conditionally heteroskedastic**.

(b) Under the null hypothesis, $H_0 : R_{\#rxK}\beta_{Kx1} = r_{\#rx1}$ where **R** is an #**r**xK matrix, #**r** < K, with full row rank (where #r is the number of restrictions on β)

$$W \equiv (R\hat{b}(\hat{W}) - r)\left(R[\hat{A}\text{var}(\hat{b}(\hat{W}))]R'\right)^{-1}(R\hat{b}(\hat{W}) - r) \rightarrow_D \chi^2(\#r)$$ [16]

(c) Under the null hypothesis with #**a** restrictions[17]

$H_0 : a(\beta) = 0$ *for some #a – dimensional vector – valued function with continuous first derivatives*

s.t. (first deriv evaluated at β*)* $A(\beta) = \frac{\partial a(\beta)}{\partial \beta_{Kx1}}$ *is* $\#a \times K$ *matrix of continuous derivatives with full row rank*

Then,

$$W \equiv n \cdot a(\hat{b}(\hat{W}))_{1x\#a}^{'}\left(A(\hat{b}(\hat{W}))_{\#axK}\hat{A}\text{var}(\hat{b}(\hat{W}))_{KxK}A(\hat{b}(\hat{W}))_{Kx\#a}^{'}\right)a(\hat{b}(\hat{W}))_{\#ax1} \rightarrow_D \chi^2(\#a)$$ [18]

Hypothesis Testing by the Likelihood Ratio Principle

15

By V.2 above $\sqrt{n}\left(\hat{b}_k - \overline{\beta}_k\right) \rightarrow_D N\left(0, A\text{var}(\hat{b}_k)\right)$, *by V.3 above,* $\hat{A}\text{var}(\hat{b}_k) \rightarrow_P A\text{var}(\hat{b}_k)$

Therefore, by Slutsky, $\frac{\sqrt{n}\left(\hat{b}_k - \overline{\beta}_k\right)}{\sqrt{\hat{A}\text{var}(\hat{b}_k)}} \rightarrow_D N(0,1)$

16

We can rewrite $W = \sqrt{n}\left(R\hat{b} - r\right)'\left(R[\hat{A}\text{var}(\hat{b})]R'\right)^{-1}\sqrt{n}\left(R\hat{b} - r\right) = c_n^{'}Q_n^{-1}c_n$

Under the null, $R\hat{b} = r \Rightarrow W = \sqrt{n}R\left(\hat{b} - \beta\right)'\left(R[\hat{A}\text{var}(\hat{b})]R'\right)^{-1}\sqrt{n}R\left(R\hat{b} - \beta\right)$

From (b) above $\sqrt{n}\left(\hat{b} - \beta\right) \rightarrow_D N(0, A\text{var}(\hat{b})) \Rightarrow R\sqrt{n}\left(\hat{b} - \beta\right) \rightarrow_D N\left(0, RA\text{var}(\hat{b})R'\right)$

From (c) above, $[\hat{A}\text{var}(\hat{b})] \rightarrow_P A\text{var}(\hat{b}) \Rightarrow R[\hat{A}\text{var}(\hat{b})]R' \rightarrow_D RA\text{var}(\hat{b})R'$

Recall, if $x_{mx1} \sim N_m(\mu, \Sigma_{mxm})$, $(x-\mu)\Sigma^{-1}(x-\mu)' \sim \chi^2(m)$

Here, $c_n \rightarrow c \sim N(RA\text{var}(\hat{b})R')$.

$c_n^{'}Q_n^{-1}c_n \rightarrow c'Q^{-1}c \sim \chi^2(\#r)$

17 The full row rank condition is there so that the hypothesis is well-defined. Thisis the generalization of the requirement for linear restrictions Rb = r that R is full row rank.

18

Under Null, $a(\beta) = 0 \Rightarrow \sqrt{n}a(\hat{b}) = \sqrt{n}\left(a(\hat{b}) - a(\beta)\right)$

By (b) above, $\sqrt{n}\left(\hat{b} - \beta\right) \xrightarrow{D} N(0, A\text{var}(\hat{b})) \Rightarrow$ *By Delta Method,* $\sqrt{n}\left(a(\hat{b}) - a(\beta)\right) \rightarrow_D c \;\; c \sim N(0, A(\hat{b})A\text{var}(\hat{b})A(\hat{b})')$

The rest follows same from proof of (b) above.

VIII. Test for Overidentifying Restrictions: If the equation is exactly identified, then it is possible to choose **b*** s.t. $\mathbf{g}_b(\mathbf{b}^*) = 0$ and $J(\delta,\mathbf{W}_n) = g_n(\mathbf{b}^*)'\mathbf{W}_n\, g_n(\mathbf{b}^*) = 0$ (We call **b*** the IV estimator). If the equation is overidentified, then the distance cannot be set to 0 exactly (since there is no correlation between the moment conditions), though we expect the minimized distance to be close to 0. If we choose the efficient weighting matrix W* s.t. plim W* = $\mathbf{S}^{-1}$, then the minimized distance is asymptotically chi-squared.

1. Hansen's Test of Overidentifying Restrictions:

   Suppose there is available a consistent estimator $\hat{S}$ of **S** (= E($\boldsymbol{g}_i\boldsymbol{g}_i'$)). Under assumptions 3.1 – 3.5,

$$\boxed{J\left(\hat{b}_{GMM}\left(\hat{S}^{-1}\right),\hat{S}^{-1}\right) = g_n\left(\hat{b}_{GMM}\left(\hat{S}^{-1}\right)\right)'\hat{S}^{-1} g_n\left(\hat{b}_{GMM}\left(\hat{S}^{-1}\right)\right) \xrightarrow{D} \chi^2(m-d)}$$

   Note:
   This says that the **objective function** evaluated at the estimator, i.e. the minimum distance, is asymptotically chi-squared.
   This is a **specification test**, testing whether all the restrictions of the model (i.e. 3.1 – 3.5) are satisfied. Given a large enough sample, if the J statistic is "surprisingly" large, then either the orthogonality condition (3.3) or the other assumptions (or both) are likely to be false.

2. Testing a subset of orthogonality conditions (Newey, Eichenbaum, Hansen, and Singleton)
   Suppose assumptions 3.1 – 3.5 hold. Let $\mathbf{z}_{i1}$ be a subvector of $\mathbf{z}_i$, and strengthen Assumption 3.4 by requiring that the rank condition for identification is satisfied for $\mathbf{z}_{i1}$ (so $E(\mathbf{x}_{i1}\mathbf{z}_i')$ is full column rank). Then, for any consistent estimator **S*** of **S**, and $\mathbf{S}_{11}^*$ of $\mathbf{S}_{11}$,

   $C \equiv J - J_1 \xrightarrow{D} \chi^2(m - m_1)$ where m = #$\mathbf{z}_i$ (dimension of $\mathbf{z}_i$), m_1 = #$\mathbf{z}_{i1}$ (dimension of $\mathbf{z}_{i1}$)

$$J(\delta,\mathbf{W}_n) = n\, g_n(\mathbf{b}^*)'\mathbf{S}^{*-1} g_n(\mathbf{b}^*)$$

$$J_1(\delta,\mathbf{W}_n) = n\, g_{1n}(\mathbf{b}^*)'\,\mathbf{S}_{11}^{*-1}\, g_{1n}(\mathbf{b}^*)$$

$$g_n(\hat{b})_{mx1} \equiv \begin{bmatrix} g_{1n}(\hat{b})_{d_1x1} \\ g_{2n}(\hat{b})_{d_2x1} \end{bmatrix},\quad S_{dxd} \equiv \begin{bmatrix} S_{11(d_1xd_1)} & S_{12(d_1xd_2)} \\ S_{21(d_2xd_1)} & S_{22(d_2xd_2)} \end{bmatrix}$$

IX. **Implications of Conditional Homoskedasticity**: Assumption 3.7 – $E(\varepsilon_i^2 | \mathbf{z}_i) = \sigma^2$.

A. $\mathbf{S} = E(\mathbf{g}_i \mathbf{g}_i') = E[\varepsilon_i^2 \mathbf{z}_i \mathbf{z}_i'] = \sigma^2 \Sigma_{ZZ}$ ($\Sigma_{ZZ} = E[\mathbf{z}_i \mathbf{z}_i']$)[19]

(as in chapter2, this decomposition has several implications)

- **S** nonsingular (by 3.5), the decomposition implies $\sigma^2\Sigma_{ZZ}$ nonsingular → σ^2 !=0 and Σ_{ZZ} nonsingular
- A consistent estimator **S*** of **S** is: $S^* = \hat{\sigma}^2 \frac{1}{n}\sum_{i=1}^{n} z_i z_i^{'} = \hat{\sigma}^2 S_{ZZ}$ where $\hat{\sigma}^2$ is some consistent estimator of σ^2

(By Ergodic stationarity, **S***$\to_{a.s.}$ **S** , we don't need 4th moment assumption!)

B. **Efficient GMM becomes 2SLS**:

GMM Estimator Under Conditional Homoskedasticity:

Setting $S^* = \hat{\sigma}^2 \frac{1}{n}\sum_{i=1}^{n} z_i z_i^{'} = \hat{\sigma}^2 S_{ZZ}$, GMM estimator becomes

$\mathbf{b}_{GMM}((\hat{\sigma}^2 \mathbf{S_{ZZ}})^{-1}) = (\mathbf{Szx}'(\hat{\sigma}\, \mathbf{Szz})^{-1}\mathbf{Szx})^{-1}\, \mathbf{Szx}'\, (\hat{\sigma}\, \mathbf{Szz})^{-1}\mathbf{s_{ZY}} = (\mathbf{Szx}'(\mathbf{Szz})^{-1}\mathbf{Szx})^{-1}\, \mathbf{Szx}'\, (\mathbf{Szz})^{-1}\mathbf{s_{ZY}} = \mathbf{b}_{GMM}(S_{ZZ}^{-1}) = \mathbf{b_{2SLS}}$

(Does not depend on $\hat{\sigma}^2$!)[20]

X. 2SLS: **2SLS is a (special case) GMM estimator, i.e. with a particular choice of weighting matrix - ($\hat{\sigma}^2$ $\mathbf{S_{ZZ}})^{-1}$.**
It's also the efficient GMM estimator obtained under conditional homoskedasticity.

A. Alternative Derivations of 2SLS: 2SLS as IV estimator and 2SLS as 2 Regressions

$$Let\ X_{nxd} = \begin{bmatrix} x_1^{'} \\ x_2^{'} \\ \vdots \\ x_n^{'} \end{bmatrix},\ Z_{nxm} = \begin{bmatrix} z_1^{'} \\ z_2^{'} \\ \vdots \\ z_n^{'} \end{bmatrix},\ y_{nx1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Then,

$$\hat{b}_{2SLS} = (S_{ZX}'(S_{ZZ})^{-1}S_{ZX})^{-1}S_{ZX}'(S_{ZZ})^{-1}s_{ZY}$$
$$= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y = (X'PX)^{-1}X'Py\ where\ P = Z(Z'Z)^{-1}Z'$$

$$\hat{A}Var(\hat{b}_{2SLS}) = n\hat{\sigma}^2\left[X'Z(Z'Z)^{-1}Z'X\right]^{-1} = n\hat{\sigma}^2\left[X'PX\right]^{-1}\ where\ \hat{\sigma} \equiv \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}\ and\ \hat{\varepsilon} \equiv y - X\,\hat{b}_{2SLS}$$

$$T-Statistic:\ t_l = \frac{\hat{b}_{2SLS,l} - \delta_l}{\sqrt{\left(\hat{\sigma}^2\left[X'PX\right]^{-1}\right)_{ll}}} \to_D N(0,1)$$

$$Wald-Statistic: W = \frac{a(\hat{b}_{2SLS})'\left\{A(\hat{b}_{2SLS})\left[X'Z(Z'Z)^{-1}Z'X\right]^{-1}A(\hat{b}_{2SLS})'\right\}^{-1}a(\hat{b}_{2SLS})}{\hat{\sigma}^2} \to_D ChiSq(\#a)$$

$$J-Statistic:\ J\left(\hat{b},(\hat{\sigma}^2S_{XX})^{-1}\right) = \frac{\left(y - X\hat{b}\right)^{'}P\left(y - X\hat{b}\right)}{\hat{\sigma}^2} \to_D ChiSq(m-d)$$

$$Sargan's-Statistic: \frac{\hat{\varepsilon}'P\hat{\varepsilon}}{\hat{\sigma}^2}$$

---

[19] $\mathbf{S} = E(\mathbf{g}_i\mathbf{g}_i') = E[(\mathbf{z}_i\,\varepsilon_i)(\mathbf{z}_i\,\varepsilon_i)'] = E[\varepsilon_i^2\,\mathbf{z}_i\,\mathbf{z}_i'] = E[E(\varepsilon_i^2\,\mathbf{z}_i\,\mathbf{z}_i' \mid \mathbf{z}_i)] = E[\mathbf{z}_i\,\mathbf{z}_i' E(\varepsilon_i^2 \mid \mathbf{z}_i)] = E[\sigma^2\,\mathbf{z}_i\,\mathbf{z}_i'] = \sigma^2\Sigma_{ZZ}$

[20]**Note**: In the efficient 2-step GMM estimation, the first step is to obtain a consistent estimator of **S**. Under conditional homoskedasticity, we don't need to perform the first step since from the assumption we immediately get the consistent estimator $S^* = \hat{\sigma}^2 \frac{1}{n}\sum_{i=1}^{n} z_i z_i^{'} = \hat{\sigma}^2 S_{ZZ}$. So, the second step estimator collapses to the GMM estimator with $\mathbf{S_{XX}}^{-1}$ as the weighting matrix. This estimator is called the 2SLS estimator because it can be estimated by 2 OLS regressions.

B. 2 Interpretations of 2SLS: $\hat{b}_{2SLS} = (X'PX)^{-1}X'Py$

(i) 2SLS as an IV Estimator

Let $\mathbf{x}^*_i$ (Dx1) be the vector of D instruments for the D regressors $\mathbf{x}_i$, these instruments will be generated from $\mathbf{z}_i$ (Mx1) as follows:

a. The *d*-th instrument is the fitted value from regressing the *d*-th regressor, $\boldsymbol{x}_{id}$, on $\mathbf{z}_i$:

We obtain $\boxed{\hat{x}_d = Z\left(Z'Z\right)^{-1}Z'x_d}$ and $\boxed{\hat{X}_{Dx1} = \left[Z\left(Z'Z\right)^{-1}Z'x_1,...,Z\left(Z'Z\right)^{-1}Z'x_D\right] = Z\left(Z'Z\right)^{-1}Z'X = PX}$

(Verify that these are indeed instruments: uncorrelated with errors and correlated with endogenous var)

b. Recall, if Z_{nxD} is the data matrix of D instruments for the D endogenous regressors X_{nxD}, then

$$b_{IV} = S_{ZX}^{-1}s_{ZY} = \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i y_i\right)$$

Here, the IV estimator is:

$$b_{IV} = S_{ZX}^{-1}s_{ZY} = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{x}_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{x}_i y_i\right) = \left(X'\hat{X}\right)^{-1}\left(\hat{X}'Y\right) = b_{2SLS}$$

(ii) 2SLS as 2 Regressions

Since P is symmetric and idempotent, $\hat{b}_{2SLS} = (X'PX)^{-1}X'Py = \left(X'P'PX\right)^{-1}X'Py = \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'y$

We can obtain this estimator by 2 OLS estimations:

a. Regress X on Z to obtain fitted values $\hat{x}_i$: i.e. obtain $\tilde{X} = PX$
   - Note: We only need to regress the endogenous variables on the instruments.
     The part of the vector $\boldsymbol{x}$ that is pre-determined it should be treated as an instrument, so projecting it onto the column space will get the same thing back.

b. Regress Y on $\tilde{X}$: i.e. obtain $\hat{b}_{2SLS} = \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'y \text{ where } P = Z(Z'Z)^{-1}Z'$

Note: OLS packages return 2SLS SE's based on the residual vector $\mathbf{y} - \hat{X}\hat{b}_{2SLS}$.

This is NOT the same $\boldsymbol{y} - X\hat{b}_{2SLS}$ (i.e. the true estimated residual of interest). Therefore the estimated asymptotic standard variance from the second stage cannot be used for statistical inference.

C. Asymptotic Properties of 2SLS: These results follow from the fact that 2SLS is special case of GMM with $W_n = (\mathbf{Szz})^{-1}$

a. Consistency[21]: Under Assumptions 3.1 – 3.4, the 2SLS estimator $\mathbf{b_{2SLS}} = (\mathbf{Szx}'(\mathbf{Szz})^{-1}\mathbf{Szx})^{-1}\,\mathbf{Szx}'\,(\mathbf{Szz})^{-1}\mathbf{s_{ZY}}$ is **consistent**.

b. Asymptotic Normality: If we add Assumption 3.5 to 3.1 – 3.4, then the 2SLS estimator is asymptotically normal

$$\sqrt{n}\left(b_{GMM}(S_{ZZ}^{-1}) - \delta\right) \xrightarrow{D} N(0,V)$$

$$\text{where } V = A\text{var}\left(b_{GMM}(S_{ZZ}^{-1})\right) = \left(\Sigma_{XZ}'\Sigma_{ZZ}^{-1}\Sigma_{XZ}\right)^{-1}\Sigma_{XZ}'\Sigma_{ZZ}^{-1}S\,\Sigma_{ZZ}^{-1}\Sigma_{XZ}\left(\Sigma_{XZ}'\Sigma_{ZZ}^{-1}\Sigma_{XZ}\right)^{-1}$$

$$= \left(E(x_iz_i')'E(z_iz_i')E(x_iz_i')\right)^{-1}E(x_iz_i')'\,E(z_iz_i')E(g_ig_i')E(z_iz_i')E(x_iz_i')\left(E(x_iz_i')'E(z_iz_i')E(x_iz_i')\right)^{-1}$$

$$bc\ S_{ZZ} = \frac{1}{n}\sum_{i=1}^{n} z_i z_i' \xrightarrow{P} E(z_iz_i') \text{ by Ergodic Theorem (since instruments are ergodic stationary)}$$

[21] Since 2SLS estimator is a special case of GMM, therefore consistency follows from the general case

c. Conditional Homoskedasticity: If we add Assumption 3.7 to 3.1 – 3.5, then the estimator is the efficient GMM estimator with the asymptotic variance given by:

GMM Estimator Under Conditional Homoskedasticity: Setting $S^* = \hat{\sigma}^2 \frac{1}{n}\sum_{i=1}^{n} z_i z_i^{'} = \hat{\sigma}^2 S_{ZZ}$, GMM estimator becomes

$\mathbf{b}_{GMM}((\hat{\sigma}\, \mathbf{S_{ZZ}})^{-1}) = (\mathbf{Szx}'(\hat{\sigma}\, \mathbf{Szz})^{-1}\mathbf{Szx})^{-1}\, \mathbf{Szx}'\, (\hat{\sigma}\, \mathbf{Szz})^{-1}\mathbf{s_{ZY}} = (\mathbf{Szx}'(\mathbf{Szz})^{-1}\mathbf{Szx})^{-1}\, \mathbf{Szx}'\, (\mathbf{Szz})^{-1}\mathbf{s_{ZY}} = \mathbf{b}_{GMM}(\mathbf{S}_{ZZ}^{-1}) = \mathbf{b_{2SLS}}$

$$A\,\text{var}(\hat{b}_{2SLS}) = A\,\text{var}\left\{\hat{b}_{GMM}^{eff}\left(\left(S^{-1}\right)^{-1}\right)\right\} = under\ cond\ \text{hom}o = \left\{\Sigma_{ZX}^{'}\left(\sigma^2\,\Sigma_{ZZ}\right)^{-1}\Sigma_{ZX}^{'}\right\}^{-1} = \sigma^2\left(\Sigma_{ZX}^{'}\Sigma_{ZZ}^{-1}\Sigma_{ZX}\right)^{-1}$$

A natural, consistent estimator of asymptotic variance is

$$\hat{A}\,\text{var}(\hat{b}_{2SLS}) = \hat{\sigma}^2\left(S_{ZX}^{'}S_{ZZ}^{-1}S_{ZX}\right)^{-1} \quad where \quad \hat{\sigma}^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i{}'\hat{b}_{2SLS}\right)^2$$

d. T-Statistic Converges in distribution to Standard Normal:

$$T_l = \frac{\hat{b}_{2SLS} - \delta_l}{Robust\ SE(\hat{b}_{2SLS})_l} \xrightarrow{D} N(0,1)$$

$$Robust\ SE_l = \sqrt{\frac{\hat{\sigma}^2}{n}\left[\left(S_{ZX}^{'}S_{ZZ}^{-1}S_{ZX}\right)^{-1}\right]_l} = \sqrt{\hat{\sigma}^2 X'Z(Z'Z)^{-1}ZX'} = \hat{\sigma}\sqrt{X'P_Z X'}$$

e. Wald-Statistic Converges in distribution to Chi-Square(#**r**) (where #r = the dimensionality of restrictions)

$$W = \frac{a(\hat{b}_{2SLS})'\left\{A(\hat{b}_{2SLS})\left[X'Z(Z'Z)^{-1}Z'X\right]^{-1}A(\hat{b}_{2SLS})'\right\}^{-1}a(\hat{b}_{2SLS})}{\hat{\sigma}^2} \xrightarrow{D} \chi^2(\#r)$$

f. The Sargan Statistic Converges to Chi-Sq($m - r$)

$$S\arg an's\ Stat = \frac{\hat{\varepsilon}'P\hat{\varepsilon}}{\hat{\sigma}^2} \xrightarrow{D} \chi^2(m-d)$$

D. **Note on Small Sample Properties of 2SLS**

There is research that shows that **if the R^2 in the first-stage regression is low, then we should suspect that the large sample approximation to the finite sample distribution of the SLS estimator to be poor**.

E. **When Regressors are Predetermined and Errors are Conditionally Homoskedastic: Efficient GMM = OLS!**
When all regressors are predetermined and errors are conditionally homoskedastic, the objective function (J statistic) for the efficient GMM estimator/2SLS is:

$$J\left(b_{2SLS},\left(\hat{\sigma}^2 S_{ZZ}\right)^{-1}\right)=\left[Z\left(y-Xb_{2SLS}\right)\right]'\left(\hat{\sigma}^2 S_{ZZ}\right)^{-1}\left[Z\left(y-Xb_{2SLS}\right)\right]=\frac{(y-Xb_{2SLS})'P(y-Xb_{2SLS})}{\hat{\sigma}^2}$$

$$=\frac{y'Py-b_{2SLS}'X'Py-y'PXb_{2SLS}+b_{2SLS}'X'PXb_{2SLS}}{\hat{\sigma}^2}=\frac{y'Py-2b_{2SLS}'X'Py+b_{2SLS}'X'PXb_{2SLS}}{\hat{\sigma}^2}$$

$$=\frac{y'Py-2b_{2SLS}'X'P'y+b_{2SLS}'X'PXb_{2SLS}}{\hat{\sigma}^2}\quad Since\ P=P'$$

$$=\frac{y'Py-2b_{2SLS}'X'y+b_{2SLS}'X'Xb_{2SLS}}{\hat{\sigma}^2}\quad Since\ PX=X\ when\ x_i\subseteq z_i\ (i.e.\ regressors\ are\ instruments)$$

$$=\frac{\left(y-Zb_{2SLS}\right)'\left(y-Zb_{2SLS}\right)}{\hat{\sigma}^2}-\frac{y'y-y'Py}{\hat{\sigma}^2}$$

$$=\frac{\left(y-Zb_{2SLS}\right)'\left(y-Zb_{2SLS}\right)}{\hat{\sigma}^2}-\frac{\left(y-\hat{y}\right)\left(y-\hat{y}\right)'}{\hat{\sigma}^2}\quad where\ \hat{y}\equiv Py$$

Since the last term does not depend on b, minimizing J amounts to minimizing the SSR = ($\mathbf{y}$ – $\mathbf{Z}$ $\mathbf{b}$)’ ($\mathbf{y}$ – $\mathbf{Z}$ $\mathbf{b}$).

Implication:

i. **Efficient GMM estimator is OLS** (This is true as long as $\mathbf{z}_i = \mathbf{x}_i$, ie. regressors are predetermined)

ii. The restricted efficient GMM estimator subject to constraints of the null hypothesis is the restricted OLS (whose objective function is not J but SSR

iii. Wald statistic, which is numerically equal to the LR statistic, can be calculated as the difference in SSR with and whtout the imposition of the null, normalized to $\hat{\sigma}^2$. (this confirms the derivation in 2.6 that the LR principle can derive the Wald)

(Note: **This is why we can fit OLS in the GMM framework, treating x’s as instruments**)

F. Limited Information Maximum Likelihood Estimator (LIML): This is the ML counterpart of 2SLS

They’re both k class estimators. 2SLS is a k class estimator with k = 1 (p. 541). So when the equation is just identified (k = 1), LIML = 2SLS numerically.

Multiple Equation GMM

Background: Having seen how to estimate 1 equation via GMM, now we can estimate a system of multiple equations as well. Going from simple to multiple equation is easy because the multiple-equation GMM estimator can be expressed as a single-equation GMM estimator by suitably specifying the matrices and vectors comprising the single-equation GMM formula.

Summary: Under conditional homoskedasticity, multiple equation GMM reduces to the full-information instrumental variable efficient estimator (**FIVE**), which reduces to the **3SLS** if the set of instruments is common to all equations. If we further assume that all regressors are predetermined, the 3SLS reduces to seemingly unrelated regressions (**SUR**), which in turn reduces to the **multivariate regression** when all the equations have the same regressors.

I. Assumptions: There are M equations, each of which is a linear equation like the one in GMM

4.1 **Linearity**: There are M linear equation,

$y_{im} = \mathbf{x}_{im}'\delta_m + \varepsilon_{im}$ ($m = 1, 2, \ldots, m$; $i = 1,2,\ldots,n$) (this is the system of equations we want to estimate)

($\mathbf{x}_{im}$ is the d_m-dim. vector of regressors, δ_m is the coefficient vector, and ε_{im} is the unobservable error term for the m-th equation)

Note on interequation correlation and cross-equation restriction[22]

Cross-Equation restrictions often occur in panel data models, where the same relationship can be estimated for different points in time. [23]

4.2 **Ergodic Stationarity**:

Let $\mathbf{w}_i$ be the unique and nonconstant elements of ($y_{i1},\ldots, y_{iM}, \mathbf{x}_{i1},\ldots,\mathbf{x}_{iM},\mathbf{z}_{i1},\ldots,\mathbf{z}_{iM}$). $\{\mathbf{w}_i\}$ is jointly stationary and ergodic.

Note: This is stronger than assuming ergodic stationarity is satisfied for each of the M equations in the system. Even if $\{y_{im}, \mathbf{z}_{im}, \mathbf{x}_{im}\}$ is stationary and ergodic for each m does not imply the whole system $\{\mathbf{w}_i\}$, i.e. the union of individual processes, is jointly stationary and ergodic.[24]

4.3 **Orthogonality Conditions**: Conditions for the M-equation system are just a collection conditions for each equation

For each equation *m*, the K variables in $\mathbf{z}_i$ are predetermined in the sense that they are all orthogonal to the current error term:

$$E(\bar{z}_{im}\varepsilon_{im}) = 0 \ \forall i, m = 1,2,\ldots,M \iff E(g_i)_{\left(\sum_{i=1}^{M} K_m\right)x1} \equiv E\begin{bmatrix} \bar{z}_{i1(K_1x1)} \cdot \left(y_{i1} - \bar{z}_{i1(1xd_1)}'\delta_{d_1x1}\right)_{1x1} \\ \vdots \\ \bar{z}_{i1(K_Mx1)} \cdot \left(y_{i1} - \bar{z}_{i1(1xd_M)}'\delta_{d_Mx1}\right)_{1x1} \end{bmatrix} = E\begin{bmatrix} \bar{z}_{i1(K_1x1)} \cdot \varepsilon_{i1} \\ \vdots \\ \bar{z}_{iM(K_Mx1)} \cdot \varepsilon_{i1} \end{bmatrix} = 0_{\left(\sum_{i=1}^{M} K_m\right)x1}$$

Note on Cross Orthogonalities: The model assumes no "cross" orthogonalities. e.g. $\mathbf{z}_{i1}$ and ε_{i2} do not have to be orthogonal. However, if a variable is included in both $\mathbf{z}_{i1}$ and $\mathbf{z}_{i2}$ (shared instrument), then 4.3 implies that the variable is orthogonal to both ε_{i1} and ε_{i2}.

---

[22] The model makes no assumptions about the interequation (or contemporaneous) correlation between errors ($\varepsilon_{i1},\ldots, \varepsilon_{iM}$). Also, there is no a priori restrictions on the coefficients from different equations: i.e. the model assumes no cross-equation restrictions on the coefficients.

Example: Suppose we want to estimate the wage equation (a la Grliches) and add to it the equation for KWW (score on the "Knowledge of the World Test")

$$LogWage_i = \phi_1 + \beta_1 S_i + \gamma_1 IQ_i + \pi EXPR_i + \varepsilon_{i1}$$

$$KWW_i = \phi_2 + \beta_2 S_i + \gamma_2 IQ_i + \varepsilon_{i2}$$

$$x_{i1} = (1\ S_i\ \ IQ_i\ \ EXPR_i)',\ x_{i2} = (1\ S_i\ \ IQ_i)',\ \delta_1 = (\phi_1, \beta_1, \gamma_1, \pi)',\ \delta_1 = (\phi_2, \beta_2, \gamma_2)'$$

Here, ε_{i1} and ε_{i2} can be correlated. Correlation arrises if, for example, there is unobservable individual characteristic that affects both wage rate and test score. There are no cross-equation restrictions such as $\beta_1 = \beta_2$

[23] Panel data Example: Suppose we have data for the log-wage exercise (a la Griliches) in 1969 and 1980. We can estimate 2 wage equations:

$$LW69_i = \phi_1 + \beta_1 S69_i + \gamma_1 IQ_i + \pi_1 EXPR69_i + \varepsilon_{i1}$$

$$LW80_i = \phi_2 + \beta_2 S80_i + \gamma_2 IQ_i + \pi_2 EXPR80_i + \varepsilon_{i2}$$

In this setup, $S69_i$ (education in 1969) and $S80_i$ (education in 1980) are 2 different variables. One possible set of restrictions is that the set of coefficients remain unchanged through time (i.e. same effect) → $\phi_1 = \phi_2$, $\beta_1 = \beta_2$, $\gamma_1 = \gamma_2$, $\pi_1 = \pi_2$

[24] Example: Element-wise vs. joint stationarity

Let $\{e_i\}$ (i = 1,2,…) be a scalar i.i.d. process. Create a 2-dimensional process $\{\mathbf{z}_i\}$ from it by defining $\mathbf{z}_{i1} = e_i$ and $\mathbf{z}_{i2} = e_1$.

Note: The first case is an example of iid sequence. The second is an example of a constant sequence (maximum serial dependence). Both are types of stationary processes.

Here, the scalar processes $\{\mathbf{z}_{i1}\}$ and $\{\mathbf{z}_{i2}\}$is are stationary. However, the vector process $\{\mathbf{z}_i\}$is not jointly stationary because the joint distribution of $\mathbf{z}_1 = (\varepsilon_1,\varepsilon_1)'$ differs from that of $\mathbf{z}_2 = (\varepsilon_2,\varepsilon_1)'$.

4.4 **Rank Condition for Identification**: Guarantees there's a unique solution to the system of equations[25]

For each of the m (=1,2,…,M) equations, the K_m x d_m matrix $E(\mathbf{z}_i\, \mathbf{x}_i')_{K_m x D_m}$ is of full column rank (or $E(\mathbf{x}_i\, \mathbf{z}_i')$ is of full row rank), $\mathbf{K_m \geq d_m}$ **(# of equations ≥ # of unknowns)** for all m.

4.5 $\mathbf{g_i}$ **is a Martingale Difference with Finite Second Moments**: Assumption for Asymptotic Normality

$\{\mathbf{g}_i\}$ is a joint martingale difference sequence with finite second moments

(so $E(\mathbf{g}_i) = \mathbf{0}$ with $E(\mathbf{g}_i \mid \mathbf{g}_{i-1}, \mathbf{g}_{i-2},\ldots, \mathbf{g}_1)= \mathbf{0}$ for $i \geq 2$) → no serial correlation in $\mathbf{g}_i$)

The matrix of cross moments, $S = E(\mathbf{g}_i\, \mathbf{g}_i') = S = E(g_i g_i')_{\left(\sum_{m=1}^{M} K_m\right)\times\left(\sum_{m=1}^{M} K_m\right)} = \begin{bmatrix} E(\varepsilon_{i1}\varepsilon_{i1}\bar{z}_{i1}\bar{z}_{i1}') & \ldots & E(\varepsilon_{i1}\varepsilon_{iM}\bar{z}_{i1}\bar{z}_{iM}') \\ \vdots & & \vdots \\ E(\varepsilon_{i1}\varepsilon_{i1}\bar{z}_{iM}\bar{z}_{i1}') & \ldots & E(\varepsilon_{iM}\varepsilon_{iM}\bar{z}_{iM}\bar{z}_{iM}') \end{bmatrix}$ is nonsingular.

Note: This is stronger than assuming $\mathbf{g}_{im} = z_{im}\cdot \varepsilon_{im}$ is MDS in each equation m.

Add'l:

4.6 **Finite fourth moments for Regressors**: (For consistent estimation of **S**)

$E[(\mathbf{z}_{imk}\mathbf{x}_{ihj})^2]$ exists and is finite for all k = 1,…, K_m, j = 1,2, …, D_h m,h = 1,2,…M, where z_{imk} is the kth element of $\mathbf{x}_{im}$ and z_{ihj} is the jth elemnt of $\mathbf{z}_{ih}$

4.7 **Conditional Homoskedasticity:** Constant Cross Moment

$E(\varepsilon_{im}\varepsilon_{ih} \mid \mathbf{z}_{im}, \mathbf{z}_{ih}) = \sigma_{mh}^2$ for all m, h = 1,2,…, M or $E(\varepsilon_i\varepsilon_i' \mid Z_i)= \Sigma$

Note on Complete System of Simultaneous Equations: the "Complete" system adds more assumptions to our model, assumptions which are unnecessary for development of ME GMM. They are covered in 8.5.

II. Multuple-Equation GMM Defined: This is same as Single Equation GMM but with re-defined matrices

1. General Setup

The parameter of interest $\delta_{GMM} = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_M \end{bmatrix}$ is defined implicitly as the solution to the moment conditions

$$E(g_i(w_i,\delta))_{\left(\sum_{i=1}^{M} K_m\right)x1} \equiv E\begin{bmatrix} \bar{z}_{i1(K_1x1)}\cdot\left(y_{i1}-\bar{x}_{i1(1xd_1)}{}'\delta_{d_1x1}\right)_{1x1} \\ \vdots \\ \bar{z}_{iM(K_Mx1)}\cdot\left(y_{i1}-\bar{x}_{iM(1xd_M)}{}'\delta_{d_Mx1}\right)_{1x1} \end{bmatrix} = \begin{bmatrix} E(\bar{z}_{i1(K_1x1)}\cdot y_{i1}) \\ \vdots \\ E(\bar{z}_{iM(K_Mx1)}\cdot y_{iM}) \end{bmatrix}_{\left(\sum_{i=1}^{M} K_m\right)x1} - \begin{bmatrix} E(\bar{z}_{i1(K_1x1)}\bar{x}_{i1(1xd_1)}'{}'\delta_{d_1x1}) \\ \vdots \\ E(\bar{z}_{i1(K_1x1)}\cdot\bar{x}_{iM(1xd_1)}'{}'\delta_{d_Mx1}) \end{bmatrix}_{\left(\sum_{i=1}^{M} K_m\right)x1}$$

$$= \begin{bmatrix} E(\bar{z}_{i1(K_1x1)}\cdot y_{i1}) \\ \vdots \\ E(\bar{z}_{iM(K_Mx1)}\cdot y_{iM}) \end{bmatrix} - \begin{bmatrix} E(\bar{z}_{i1(K_1x1)}\bar{x}_{i1(1xd_1)}'{}') & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & E(\bar{z}_{i1(K_1x1)}\cdot\bar{x}_{iM(1xd_1)}'{}') \end{bmatrix}_{\left(\sum_{i=1}^{M} K_m\right)x\left(\sum_{i=1}^{M} d_m\right)} \begin{bmatrix} \delta_{1(d_1x1)} \\ \vdots \\ \delta_{M(d_Mx1)} \end{bmatrix}_{\left(\sum_{i=1}^{M} d_m\right)x1}$$

$$\equiv \sigma_{ZY} - \Sigma_{ZX}\delta = 0$$

Where $\sigma_{ZY} = E(Z_iY_i)$ *with* $Z_i = \begin{bmatrix} \bar{z}_{i1(K_1x1)} & & \\ & \ddots & \\ & & \bar{z}_{iM(K_Mx1)} \end{bmatrix}$ *and* $\Sigma_{ZX} = E(Z_iX_i')$ *with* $X_i = \begin{bmatrix} \bar{x}_{i1(d_1x1)} & & \\ & \ddots & \\ & & \bar{x}_{iM(d_Mx1)} \end{bmatrix}$

So, $E(Z_iY_i) = E(Z_iX_i')\delta$

Sample Analogue:

$$g_n(w_i,\delta)_{\left(\sum_{i=1}^{M} K_m\right)x1} \equiv \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}\bar{z}_{i1(K_1x1)}\cdot y_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n}\bar{z}_{iM(K_Mx1)}\cdot y_{iM} \end{bmatrix} - \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}\bar{z}_{i1(K_1x1)}\bar{x}_{i1(1xd_1)}' & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{1}{n}\sum_{i=1}^{n}\bar{z}_{iM(K_1x1)}\bar{x}_{iM(1xd_1)}' \end{bmatrix}_{\left(\sum_{i=1}^{M} K_m\right)x\left(\sum_{i=1}^{M} d_m\right)} \begin{bmatrix} \delta_{1(d_1x1)} \\ \vdots \\ \delta_{M(d_Mx1)} \end{bmatrix}_{\left(\sum_{i=1}^{M} d_m\right)x1}$$

$$\equiv s_{ZY} - S_{ZX}\delta = 0$$

[25] We can uniquely determine all the coefficient vectors δ_1,..,δ_m iff each coefficient vector δ_m is uniquely determined, which occurs iff Assumption 3.4 holds for each equation. The rank condition is simple here because there are no cross-equation restrictions – when coefficients are assumed to be the same across all equations we will have different identification condition.

4 Special Features of M.E. GMM: We substitute these into Single Equation GMM and get same results!

i. s_{ZY} is a stacked vector

ii. S_{ZX} is a block diagonal matrix

iii. By ii, W_n is a $S_m K_m$ x $S_m K_m$ matrix

iv. $\bar{g} \equiv \frac{1}{n}\sum_{i=1}^{n} g_i = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} \vec{z}_{i1} \cdot \varepsilon_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} \vec{z}_{iM} \cdot \varepsilon_{iM} \end{bmatrix} = g_n(\delta)$ is a stacked vector

2. Applied to Linear Model: Our model is $y_i = \mathbf{x}_i'\delta + \varepsilon_i$ with the moment condition $E[(y_i - \mathbf{x}_i'\beta) . \mathbf{z}_i] = 0$
   - Expression for sample moment condition in linear model
     $g_n(W_i, b) = s_{ZY(mx1)} - S_{ZX(mxd)} b_{(dx1)}$
   - Method of Moments: If the equation is exactly identified $K_m = d_m$ for all m, and there exists a unique **b** such that $g_n(\mathbf{W}_i,\mathbf{b}) = 0$, and Σ_{zx} invertible, then for sufficiently large n, $S_{ZX} \rightarrow_p \Sigma_{zx}$ by ergodic theorem and is invertible with prob. 1. So, with large sample size the system of simultaneous equation has unique solution given by…

     The MM estimator (CHECK THIS!!!)

     $\hat{\delta}_{IV} = (S_{ZX})^{-1} s_{ZY}$ (**IV estimator** with $\mathbf{z}_i$ as instruments, def. for m = d)[26]

   - General Method of Moments:
     If the equation overly identified m > d, and there does not exists a unique **b** such that $g_n(\mathbf{W}_i,\mathbf{b}) = 0$ exactly, we choose $\boldsymbol{\delta}(\mathbf{W_n})$ to minimize $g_n(\mathbf{W}_i,\mathbf{b})$ or equivalently $g_n(\mathbf{W}_i,\mathbf{b})' W_n\, g_n(\mathbf{W}_i,\mathbf{b})$ (for some choice of p.d. weighting matrix that converges in probability to some p.d. W)[27]

$$\delta_{GMM} = \arg\max_{b \in R^N} -\left[E(g_i(W_i,b))\right]' \; W \; \left[E(g_i(W_i,b))\right] = -\left[E(Z_i(Y_i - X_i'b))\right] W \left[E(Z_i(Y_i - X_i'b))\right]$$

$$\hat{\delta}_{GMM} = \arg\max_{b \in R^N} -\left[\frac{1}{n}\sum_{i=1}^{n} g_i(W_i,b)\right]' \; W_{n(mxm)} \; \left[\frac{1}{n}\sum_{i=1}^{n} g_i(W_i,b)\right]$$

$$\delta_{GMM} = \left[E(Z_i X_i')' W E(Z_i X_i')\right]^{-1} E(Z_i X_i')\, W\; E(Z_i Y_i)$$

$$\hat{\delta}_{GMM} = \left\{\left[\frac{1}{n}\sum_{i=1}^{n} Z_i' X_i\right]' W_{n(mxm)} \left[\frac{1}{n}\sum_{i=1}^{n} Z_i' X_i\right]\right\}^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} Z_i' X_i\right]' W_{n(mxm)} \left[\frac{1}{n}\sum_{i=1}^{n} Z_i Y_i\right]$$

---

[26] The IV estimator is defined for the EXACTLY IDENTIFIED case (i.e. the case where there are as many instruments as endogenous regressors). If $\mathbf{z}_i = \mathbf{x}_i$, i.e. all the regressors are predetermined/orthogonal to the error contemporaneous term, then this boils down to the OLS estimator. So, OLS is a special case of MM estimator. And, IV and OLS are both special cases of GMM

[27] The quadratic form formulation $g_n(\mathbf{W}_i,\mathbf{b})' W_n\, g_n(\mathbf{W}_i,\mathbf{b})$ gives us a 1x1 real number over which we can define the minimization/maximization problem. Otherwise it would be impossible to minimize over a m-dimensional vector of moment conditions $g_n(\mathbf{W}_i,\mathbf{b})$zd

XI. ME GMM Estimator and Sampling Error

- GMM Estimator:

  $\delta_{GMM}(W_n) = (\mathbf{Szx'Wn\ Szx})^{-1}\ \mathbf{Szx'W_n s_{ZY}}$ [28]

  (By 4.2 and 4.4, S_{xz} has full column rank for sufficiently large n with prob 1, then **Szx'Wn Szx** invertible)

  Note: If m = d, then **Szx** is a mxm square matrix and the GMM estimator reduces to the IV estimator if we pick Wn =

- Sampling Error:

  $\delta_{GMM}(W_n) - \delta = (\mathbf{Szx'Wn\ Szx})^{-1}\ \mathbf{Szx'W_n s_Z\varepsilon}$

28

$$\hat{\delta}_{GMM}(\hat{W}) = \begin{bmatrix} \hat{\delta}_1(\hat{W}) \\ \vdots \\ \hat{\delta}_M(\hat{W}) \end{bmatrix} = \left(S_{ZX}'\hat{W}S_{ZX}\right)^{-1} S_{ZX}'\hat{W}s_{ZY}$$

$$= \left(\begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}_{i1}\bar{z}_{i1}'\right)_{d_1xK_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}_{iM}\bar{z}_{iM}'\right)_{d_MxK_M} \end{bmatrix} \begin{bmatrix} \hat{W}_{11(K_1xK_1)} & \cdots & \hat{W}_{1M(K_1xK_M)} \\ \vdots & & \vdots \\ \hat{W}_{M1(K_MxK_1)} & \cdots & \hat{W}_{MM(K_MxK_M)} \end{bmatrix} \begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^{n}\bar{z}_{i1}\bar{x}_{i1}'\right)_{K_1xd_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \left(\frac{1}{n}\sum_{i=1}^{n}\bar{z}_{iM}\bar{x}_{iM}'\right)_{K_Mxd_M} \end{bmatrix} \right)^{-1}$$

$$\begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}_{i1}\bar{z}_{i1}'\right)_{d_1xK_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}_{iM}\bar{z}_{iM}'\right)_{d_MxK_M} \end{bmatrix} \begin{bmatrix} \hat{W}_{11(K_1xK_1)} & \cdots & \hat{W}_{1M(K_1xK_M)} \\ \vdots & & \vdots \\ \hat{W}_{M1(K_MxK_1)} & \cdots & \hat{W}_{MM(K_MxK_M)} \end{bmatrix} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}\bar{z}_{i1(K_1x1)}\cdot y_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n}\bar{z}_{iM(K_Mx1)}\cdot y_{iM} \end{bmatrix}$$

$$= \left(\begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}_{i1}\bar{z}_{i1}'\right)_{d_1xK_1} \hat{W}_{11(K_1xK_1)} \left(\frac{1}{n}\sum_{i=1}^{n}\bar{z}_{i1}\bar{x}_{i1}'\right)_{K_1xd_1} & \cdots & \left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}_{i1}\bar{z}_{i1}'\right)_{d_1xK_1} \hat{W}_{1M(K_1xK_M)} \left(\frac{1}{n}\sum_{i=1}^{n}\bar{z}_{iM}\bar{x}_{iM}'\right)_{K_Mxd_M} \\ \vdots & & \vdots \\ \left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}_{iM}\bar{z}_{iM}'\right)_{d_MxK_M} \hat{W}_{M1(K_MxK_1)} \left(\frac{1}{n}\sum_{i=1}^{n}\bar{z}_{i1}\bar{x}_{i1}'\right)_{K_1xd_1} & \cdots & \left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}_{iM}\bar{z}_{iM}'\right)_{d_MxK_M} \hat{W}_{MM(K_MxK_M)} \left(\frac{1}{n}\sum_{i=1}^{n}\bar{z}_{iM}\bar{x}_{iM}'\right)_{K_Mxd_M} \end{bmatrix} \right)^{-1}$$

$$\begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}_{i1}\bar{z}_{i1}'\right)_{d_1xK_1} \hat{W}_{11(K_1xK_1)} \left(\frac{1}{n}\sum_{i=1}^{n}\bar{z}_{i1(K_1x1)}\cdot y_{i1}\right)_{K_1xd_1} & +\cdots+ & \left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}_{i1}\bar{z}_{i1}'\right)_{d_1xK_1} \hat{W}_{1M(K_1xK_M)} \left(\frac{1}{n}\sum_{i=1}^{n}\bar{z}_{iM(K_Mx1)}\cdot y_{iM}\right)_{K_Mxd_M} \\ \vdots & \vdots & \vdots \\ \left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}_{iM}\bar{z}_{iM}'\right)_{d_MxK_M} \hat{W}_{M1(K_MxK_1)} \left(\frac{1}{n}\sum_{i=1}^{n}\bar{z}_{i1(K_1x1)}\cdot y_{i1}\right)_{K_1xd_1} & +\cdots+ & \left(\frac{1}{n}\sum_{i=1}^{n}\bar{x}_{iM}\bar{z}_{iM}'\right)_{d_MxK_M} \hat{W}_{MM(K_MxK_M)} \left(\frac{1}{n}\sum_{i=1}^{n}\bar{z}_{iM(K_Mx1)}\cdot y_{iM}\right)_{K_Mxd_M} \end{bmatrix}$$

Where $W_{m,h}$ ($K_m x K_h$) is the (m,h) block of W_n.

Recall, for Block matrices:

1. $$\begin{bmatrix} A_{11}' & & \\ & \ddots & \\ & & A_{MM}' \end{bmatrix} \begin{bmatrix} B_{11} & \cdots & B_{1M} \\ \vdots & & \vdots \\ B_{M1} & \cdots & B_{MM} \end{bmatrix} \begin{bmatrix} A_{11} & & \\ & \ddots & \\ & & A_{MM} \end{bmatrix} = \begin{bmatrix} A_{11}'B_{11}A_{11} & \cdots & A_{11}'B_{1M}A_{MM} \\ \vdots & & \vdots \\ A_{MM}'B_{M1}A_{11} & \cdots & A_{MM}'B_{MM}A_{MM} \end{bmatrix}$$

2. $$\begin{bmatrix} A_{11}' & & \\ & \ddots & \\ & & A_{MM}' \end{bmatrix} \begin{bmatrix} B_{11} & \cdots & B_{1M} \\ \vdots & & \vdots \\ B_{M1} & \cdots & B_{MM} \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_M \end{bmatrix} = \begin{bmatrix} A_{11}'B_{11}c_1 & +\cdots+ & A_{11}'B_{1M}c_M \\ \vdots & & \vdots \\ A_{MM}'Bc_1 & +\cdots+ & A_{MM}'B_{MM}c_M \end{bmatrix}$$

III. **Large Sample Theory**: As seen above, all the theory/formulas for the multiple-equation GMM is a matter of substitution of the newly defined matrices into the equations. (See summary)

IV. GMM Summary

Population Orthogonality Conditions	$g(w_i,\delta)\equiv E(Z_iY_i)-E(Z_iX_i')\delta=\sigma_{ZY}-\Sigma_{ZX}\delta=0$
Sample Analogue of Orthogonality Conditions	$g_n(w_i,\delta)\equiv s_{ZY}-S_{ZX}\delta=0$
IV Estimator (#moment conditions = # of parameters)	$\delta_{IV}\equiv(S_{ZX})^{-1}s_{ZY}$ since S_{ZX} square and invertible
GMM Estimator (#moment conditions > # of parameters → Can't set g_n exatly to 0)	$\delta_{GMM}(W_n) = (\mathbf{Szx'Wn\ Szx})^{-1}\ \mathbf{Szx'W_n s_{ZY}}$
GMM Estimator's Sampling Error	$\delta_{GMM}(W_n) - \delta = (\mathbf{Szx'Wn\ Szx})^{-1}\ \mathbf{Szx'W_n s_Z\varepsilon}$
GMM Estimator's Asymptotic Variance (Under Asymptotic Normality)	$A\text{var}(b_{GMM}(W_n))=\left(\Sigma_{XZ}'W\Sigma_{XZ}\right)^{-1}\Sigma_{XZ}'WSW\Sigma_{XZ}\left(\Sigma_{XZ}'W\Sigma_{XZ}\right)^{-1}$
Consistent Estimator of GMM Estimator's Asymptotic Variance	$\hat{A}\text{var}(\hat{b}_{GMM})\equiv(S_{ZX}'W_nS_{ZX})^{-1}S_{ZX}'W\ S^*W'S_{ZX}(S_{ZX}'W_nS_{ZX})^{-1}$
Optimal/Efficient Weighting Matrix	$\mathbf{S^{*-1}}$ **s.t.** $\mathbf{S^{*-1}\rightarrow_p S^{-1}=E(g_ig_i')^{-1}}$
Efficient GMM Estimator	$b_{n,GMM}(\hat{S}^{-1})=(S_{ZX}'\hat{S}^{-1}S_{ZX})^{-1}S_{ZX}'\hat{S}^{-1}s_{ZY}$
Asymptotic (Minimum) Variance of the Efficient/Optimal GMM	$A\text{var}(b_{n,GMM}(\hat{S}^{-1}))=(\Sigma_{ZX}'S^{-1}\Sigma_{ZX})^{-1}=(\Sigma_{ZX}'E(g_ig_i')^{-1}\Sigma_{ZX})^{-1}$
Asymptotic (Minimum) Variance of the Efficient/Optimal GMM	$\hat{A}\text{var}(\hat{b}_{GMM})\equiv\left(S_{ZX}'\hat{S}^{-1}S_{ZX}\right)^{-1}$
J-Statistic / Objective Function evaluated at Optimal Weighting Matrix	$J\left(\hat{b}_{GMM}\left(\hat{S}^{-1}\right),\hat{S}^{-1}\right)=n\cdot g_n\left(\hat{b}_{GMM}\left(\hat{S}^{-1}\right)\right)'\hat{S}^{-1}g_n\left(\hat{b}_{GMM}\left(\hat{S}^{-1}\right)\right)$

Definitions and Properties of Single Equation vs. Multiple Equation Estimators:

	Single – Equation GMM	Multiple Equation GMM
$\mathbf{g_i}$	$\mathbf{z}_{\mathbf{i}(KX1)} \cdot \varepsilon_i$	$g_i = \left[\vec{z}_{i1} \cdot \varepsilon_{i1} \quad \dots \quad \vec{z}_{iM} \cdot \varepsilon_{iM}\right]' \left(\sum_{m=1}^{M} K_m\right) x1$
δ	δ_{dx1}	$\delta = \left[\delta_{1(D_1 x1)},\dots,\delta_{1(D_M x1)}\right]' \left(\sum_{m=1}^{M} D_m\right) x1$
$\mathbf{s_{ZY}}$	$\left(\frac{1}{n}\sum_{i=1}^{n} z_i y_i\right)$	$\begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} \vec{z}_{i1(K_1 x1)} \cdot y_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} \vec{z}_{iM(K_M x1)} \cdot y_{iM} \end{bmatrix}$
$\mathbf{S_{ZX}}$	$\left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right)$	$\begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} \vec{z}_{i1(K_1 x1)} \vec{x}_{i1(1xd_1)}' & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{1}{n}\sum_{i=1}^{n} \vec{z}_{iM(K_1 x1)} \vec{x}_{iM(1xd_1)}' \end{bmatrix}_{\left(\sum_{i=1}^{M} K_m\right) x \left(\sum_{i=1}^{M} d_m\right)}$
Size of W	KXK	$\left(\sum_{i=1}^{M} K_m\right) x \left(\sum_{i=1}^{M} K_m\right)$
$\mathbf{\Sigma_{ZX}}$	$E(\mathbf{z}_i \, \mathbf{x}_i')$	$\begin{bmatrix} E(\vec{z}_{i1(K_1 x1)} \vec{x}_{i1(1xd_1)}') & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & E(\vec{z}_{i1(K_1 x1)} \cdot \vec{x}_{iM(1xd_1)}') \end{bmatrix}_{\left(\sum_{i=1}^{M} K_m\right) x \left(\sum_{i=1}^{M} d_m\right)}$
$\mathbf{S} = A\,\text{var}(\bar{g}) = E(g_i g_i')$	$S = E(g_i g_i') = E(\varepsilon_i^2 z_i z_i')$	$S = E(g_i g_i')_{\left(\sum_{m=1}^{M} K_m\right) \times \left(\sum_{m=1}^{M} K_m\right)} = \begin{bmatrix} E(\varepsilon_{i1}\varepsilon_{i1}\vec{z}_{i1}\vec{z}_{i1}') & \dots & E(\varepsilon_{i1}\varepsilon_{iM}\vec{z}_{i1}\vec{z}_{iM}') \\ \vdots & & \vdots \\ E(\varepsilon_{i1}\varepsilon_{i1}\vec{z}_{iM}\vec{z}_{i1}') & \dots & E(\varepsilon_{iM}\varepsilon_{iM}\vec{z}_{iM}\vec{z}_{iM}') \end{bmatrix}$
$\hat{S}$ (consistent estimator of **S**)	$\hat{S} = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 z_i z_i'$	$\hat{S} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_{i1}\hat{\varepsilon}_{i1}\vec{z}_{i1}\vec{z}_{i1}' & \dots & \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_{i1}\hat{\varepsilon}_{iM}\vec{z}_{i1}\vec{z}_{iM}' \\ \vdots & & \vdots \\ \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_{iM}\hat{\varepsilon}_{i1}\vec{z}_{iM}\vec{z}_{i1}' & \dots & \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_{MM}\hat{\varepsilon}_{MM}\vec{z}_{MM}\vec{z}_{MM}' \end{bmatrix}_{\left(\sum_{m=1}^{M} K_m\right) \times \left(\sum_{m=1}^{M} K_m\right)}$
Estimator consistent under assumptions…	3.1 – 3.4	4.1 – 4.4
Estimator asymptotically normal under assumptions…	3.1 – 3.4 + 3.5	4.1 – 4.4 + 4.5
$\hat{S} \rightarrow_P \mathbf{S}$ under assumptions…	3.1, 3.2, 3.6, E($\mathbf{g_i g_i}'$) finite	4.1, 4.2, 4.6, E($\mathbf{g_i g_i}'$) finite
D.F. of J-Statistic	K-D	$\Sigma_m (K_m - D_m)$

V. **Single Equation vs. Multiple Equation Estimation**

A. Equation by Equation GMM vs. Joint Estimation: An alternative to joint estimation is to apply single-equation GMM separately to each equation.

- **The equation-by-equation estimator is a particular M.E. GMM estimator** with particular choice of W_n
  Estimate the M equations via SE GMM s.t. the weighting matrix for the m-th equation is W^*_{mm} (K_m x K_m)
  Then, if we stack the equation-by-equation GMM estimator, it can be written as a M.E. GMM estimator with the block diagonal matrix whose m-th diagonal block is W_{mm}: $\hat{W} = \begin{bmatrix} \hat{W}_{11} & & \\ & \ddots & \\ & & \hat{W}_{11} \end{bmatrix}$ (this works bc S_{ZX} and W are block diagonal)
- **When are they Equivalent**
  i. If all equations are just identified, then the equation-by equation GMM and multiple-equation GMM are numerically the same and equal to the IV estimator (regardless of weighting matrix!)
  ii. If at least one equation is overidentified but the equations are "unrelated" in the sense that $E(e_{im}e_{ih}\mathbf{x}_{im}\mathbf{x}_{oh}) = 0$ for all m !=h, then the efficient equation-by-equation GMM and the efficient multiple-equation GMM are asymptotically equivalent in that $\delta(\hat{W}) - \delta(\hat{S}^{-1}) \xrightarrow{P} 0$
- **Joint Estimation Can Be Hazardous**
  Except for the above cases, joint estimation is asymptotically more efficient since it takes advantage of cross-equation correlations. Even if you are interested in estimation 1 particular equation, you can generally gain asymptotic efficiency by combining it with some other equations (though there are special cases where joint estimation entails no efficiency gain even if the added equations are not unrelated).
  Caveats:
  i. Small Sample Properties: Small-sample properties of the equation of interest might be better without joint estimation
  ii. Misspecification: Asymptotic result presumes that the model is **correctly specified**, i.e. model assumptions are satisfied. If the model is misspecified, neither the single-equation GMM nor the multiple equation GMM is guaranteed to be even consistent. Furthermore, chances of misspecification increase as you add equations to the system. (see p272 for example)

VI. **Special Cases of Multiple Equation GMM under Conditional Homoskedasticity: FIVE, 3SLS, and SUR**

0. S under conditional homoskedasticity:

$$S = E(g_i g_i') = \begin{bmatrix} E(\varepsilon_{i1}\varepsilon_{i1}\bar{z}_{i1}\bar{z}_{i1}') & \dots & E(\varepsilon_{i1}\varepsilon_{iM}\bar{z}_{i1}\bar{z}_{iM}') \\ \vdots & & \vdots \\ E(\varepsilon_{i1}\varepsilon_{i1}\bar{z}_{iM}\bar{z}_{i1}') & \dots & E(\varepsilon_{iM}\varepsilon_{iM}\bar{z}_{iM}\bar{z}_{iM}') \end{bmatrix} = \begin{bmatrix} \sigma_{11}E(\bar{z}_{i1}\bar{z}_{i1}') & \dots & \sigma_{1M}E(\bar{z}_{i1}\bar{z}_{iM}') \\ \vdots & & \vdots \\ \sigma_{M1}E(\bar{z}_{iM}\bar{z}_{i1}') & \dots & \sigma_{MM}E(\bar{z}_{iM}\bar{z}_{iM}') \end{bmatrix} = E(Z_i'\Sigma Z_i)$$ [29]

1. Full-Information Instrumental Variables Efficient (FIVE) Estimator
   - Simplification of **S**: This estimator exploits the above structure of 4th moments by using the following consistent estimator $\hat{S}$ of **S**:

$$\hat{S} = \begin{pmatrix} \hat{\sigma}_{11}\frac{1}{n}\sum_{i=1}^{n}\bar{z}_{i1}\bar{z}_{i1}' & \dots & \hat{\sigma}_{1M}\frac{1}{n}\sum_{i=1}^{n}\bar{z}_{i1}\bar{z}_{iM}' \\ \vdots & & \vdots \\ \hat{\sigma}_{M1}\frac{1}{n}\sum_{i=1}^{n}\bar{z}_{iM}\bar{z}_{i1}' & \dots & \hat{\sigma}_{MM}\frac{1}{n}\sum_{i=1}^{n}\bar{z}_{iM}\bar{z}_{iM}' \end{pmatrix} = \frac{1}{n}Z\hat{\Sigma}Z' \;^{30} \quad \text{where } \hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_{11} & \cdots & \hat{\sigma}_{1M} \\ \vdots & & \vdots \\ \hat{\sigma}_{M1} & \cdots & \hat{\sigma}_{MM} \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i\hat{\varepsilon}_i'$$

     ($\hat{\delta}_m$ is usually the 2SLS estimator for equation m. So for the cross moments we need 2 2SLS estimators)

   - Then, our ME efficient GMM estimator simplifies to:

     $\hat{\delta}_{FIVE} \equiv \hat{\delta}(\hat{S}^{-1}) = \left(s_{ZX}'\hat{S}^{-1}s_{ZX}\right)^{-1} s_{ZX}'\hat{S}^{-1}s_{ZY} = \left(XZ'(Z\hat{\Sigma}Z')^{-1}ZX'\right)^{-1} XZ'(Z\hat{\Sigma}Z')^{-1}ZY$ (the 1/n's cancel)

   Large Sample Properties of FIVE: (these follow from the large sample properties of M.E. GMM estimators)
   Suppose Assumptions 4.1 – 4.5 and 4.7 hold. Suppose further that E($\mathbf{z}_{ih}\mathbf{z}_{im}$) exists and is finite for all m, h = 1,2,…M. Let S and $\hat{S}$ be defined as above. Then,

   (a) $\hat{S} \to_P$ **S**

   (b) $\hat{\delta}_{FIVE} \equiv \hat{\delta}(\hat{S}^{-1})$ is consistent, asymptotically normal, and efficient with Avar($\hat{\delta}_{FIVE}$) = $(\Sigma_{ZX}'S^{-1}\Sigma_{ZX})^{-1}$

   (c) The estimated asymptotic variance Avar($\hat{A}\text{var}(\hat{\delta}_{FIVE}) = (S_{ZX}'\hat{S}^{-1}S_{ZX})^{-1}$ is consistent for Avar($\hat{\delta}_{FIVE}$)

   (d) Sargan's Statistic / J Statistic: $J(\hat{\delta}_{FIVE}, \hat{S}^{-1}) \equiv n \cdot g_n(\hat{\delta}_{FIVE})'\hat{S}^{-1}g_n(\hat{\delta}_{FIVE}) \xrightarrow{D} \chi^2\left(\sum_m (K_m - D_m)\right)$

---

[29] The (m,h) block of E($g_i g_i'$) = E($\varepsilon_{im}\varepsilon_{ih}\mathbf{x}_{im}\mathbf{x}_{ih}'$) = E(E($\varepsilon_{im}\varepsilon_{ih} \mid \mathbf{x}_{im}\mathbf{x}_{ih}$) $\mathbf{x}_{im}\mathbf{x}_{ih}'$) by Law of Iterated Exp and linearity of conditional exp
= E($\sigma_{mh}\mathbf{x}_{im}\mathbf{x}_{ih}'$) by conditional homoskedasticity = σ_{mh} E($\mathbf{x}_{im}\mathbf{x}_{ih}'$)

[30] $\hat{\sigma}_{mh} \equiv \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_{im}\hat{\varepsilon}_{ih}$, $\hat{\varepsilon}_{im} \equiv y_{im} - x_{im}'\hat{\delta}_m$ $(m,h = 1,2,...,M)$ *for some consistent estimator* $\hat{\delta}_m$ *of* δ_m

By Assumption 4.1, 4.2, and E($\mathbf{z}_{ih}\mathbf{z}_{im}$) exists , $\hat{\sigma}_{mh} \xrightarrow{P} \sigma_{mh}$ (Prop 4.1 p.269). By (joint) ergodic stationarity, $\frac{1}{n}\sum_{i=1}^{n}\bar{z}_{im}\bar{z}_{ih}' \xrightarrow{P} E(\bar{z}_{im}\bar{z}_{ih}')$ which exists and is finite by assumption. Therefore, $\hat{S} \to_P$ **S** without finite 4th moment assumption.

2. **Three-Stage Least Squares (3SLS):** When the set of instruments is same across equations, FIVE can be simplified to 3SLS

- Simplification of $\mathbf{g}_i$, $\mathbf{S}$, and $\hat{S}$: If $\mathbf{z}_i$ ($= \mathbf{z}_{i1} = \mathbf{z}_{i2} = \mathbf{z}_{i3} = \ldots = \mathbf{z}_{iM}$) is the common set of instruments (for all M equations) with dimension K, then $\mathbf{g}_i$, $\mathbf{S}$, and $\hat{S}$ can be written compactly using the **Kronecker product**[31] as follows:

$$\mathbf{g}_i = \begin{bmatrix} \vec{z}_{i1} \cdot \varepsilon_{i1} \\ \vdots \\ \vec{z}_{iM} \cdot \varepsilon_{iM} \end{bmatrix} = \begin{bmatrix} \vec{z}_{i} \cdot \varepsilon_{i1} \\ \vdots \\ \vec{z}_{i} \cdot \varepsilon_{iM} \end{bmatrix}_{MKx1} \quad \text{since instruments are common} = \boldsymbol{\varepsilon}_i \otimes \mathbf{z}_i$$

$$\mathbf{S} = E(g_i g_i') = \begin{bmatrix} E(\varepsilon_{i1}\varepsilon_{i1}\bar{z}_{i1}\bar{z}_{i1}) & \cdots & E(\varepsilon_{i1}\varepsilon_{iM}\bar{z}_{i1}\bar{z}_{iM}) \\ \vdots & & \vdots \\ E(\varepsilon_{iM}\varepsilon_{i1}\bar{z}_{iM}\bar{z}_{i1}) & \cdots & E(\varepsilon_{iM}\varepsilon_{iM}\bar{z}_{iM}\bar{z}_{iM}) \end{bmatrix} = \begin{bmatrix} \sigma_{11}E(\bar{z}_{i1}\bar{z}_{i1}) & \cdots & \sigma_{1M}E(\bar{z}_{i1}\bar{z}_{iM}) \\ \vdots & & \vdots \\ \sigma_{M1}E(\bar{z}_{iM}\bar{z}_{i1}) & \cdots & \sigma_{MM}E(\bar{z}_{iM}\bar{z}_{iM}) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11}E(\bar{z}_{i}\bar{z}_{i}) & \cdots & \sigma_{1M}E(\bar{z}_{i}\bar{z}_{i}) \\ \vdots & & \vdots \\ \sigma_{M1}E(\bar{z}_{i}\bar{z}_{i}) & \cdots & \sigma_{MM}E(\bar{z}_{i}\bar{z}_{i}) \end{bmatrix} \textit{ by common instruments}$$

$$= \Sigma \otimes E(\bar{z}_i\bar{z}_i)_{MK\times MK} \quad \textit{where } \Sigma \equiv \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1M} \\ \vdots & & \vdots \\ \sigma_{M1} & \cdots & \sigma_{MM} \end{bmatrix} = E(\bar{\varepsilon}_i\bar{\varepsilon}_i') \quad \textit{for } \bar{\varepsilon}_i = \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iM} \end{bmatrix}$$

$$\mathbf{S}^{-1} = \boldsymbol{\Sigma}^{-1} \otimes \mathrm{E}(\mathbf{z}_i\, \mathbf{z}_i')^{-1}$$

$$\hat{S} = \hat{\Sigma} \otimes \left(\frac{1}{n}\sum_{i=1}^{n} \bar{z}_i \bar{z}_i' \right) = \hat{\Sigma} \otimes \frac{1}{n} Z'Z = \frac{1}{n}(\hat{\Sigma} \otimes Z'Z) \qquad \textit{where } \hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_{11} & \cdots & \hat{\sigma}_{1M} \\ \vdots & & \vdots \\ \hat{\sigma}_{M1} & \cdots & \hat{\sigma}_{MM} \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} \hat{\bar{\varepsilon}}_i \hat{\bar{\varepsilon}}_i'$$ [32]

$$\hat{S}^{-1} = \hat{\Sigma}^{-1} \otimes \left(\frac{1}{n}\sum_{i=1}^{n} \bar{z}_i \bar{z}_i' \right)^{-1} = n \cdot \left[\hat{\Sigma}^{-1} \otimes (Z'Z)^{-1} \right]$$

- With above simplifications, our M. E. GMM estimator

$$\hat{\delta}_{3SLS} = \hat{\delta}_{FIVE}\left(\hat{S}^{-1}\right) = \left(S_{ZX}'\hat{S}^{-1}S_{ZX}\right)^{-1} S_{ZX}'\hat{S}^{-1}s_{ZY} = \left[X'\left(\hat{\Sigma}^{-1} \otimes Z(Z'Z)^{-1}Z'\right)X \right]^{-1} X'\left(\hat{\Sigma}^{-1} \otimes Z(Z'Z)^{-1}Z'\right)Y$$ [33]

$$= \left[X'\left(\hat{\Sigma}^{-1} \otimes P_Z\right)X \right]^{-1} X'\left(\hat{\Sigma}^{-1} \otimes P_Z\right)Y$$

---

[31] Recall: For general matrices $\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & & \vdots \\ a_{M1} & \cdots & a_{MN} \end{bmatrix}$ (MxN) and $\mathbf{B} = \begin{bmatrix} b_{11} & \cdots & b_{1L} \\ \vdots & & \vdots \\ b_{K1} & \cdots & b_{KL} \end{bmatrix}$ (KxL), $\mathbf{a} = \begin{bmatrix} a_{11} \\ \vdots \\ a_{M1} \end{bmatrix}$ (Mx1), $\mathbf{b} = \begin{bmatrix} b_{11} \\ \vdots \\ b_{N1} \end{bmatrix}$ (Nx1)

the Kronecker product is defined as: $\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}B & \cdots & a_{1N}B \\ \vdots & & \vdots \\ a_{M1}B & \cdots & a_{MN}B \end{bmatrix}_{MKxNL}$, $\mathbf{a} \otimes \mathbf{b} = \begin{bmatrix} a_{11} \\ \vdots \\ a_{M1} \end{bmatrix} \otimes \begin{bmatrix} b_{11} \\ \vdots \\ b_{N1} \end{bmatrix} = \begin{bmatrix} a_{11}b \\ \vdots \\ a_{M1}b \end{bmatrix}_{MNx1}$

Useful Properties:

$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$ (provided that **A** and **C** are conformable and **B** and **D** are conformable)

$(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$

$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$

[32] Just as before, $\hat{\sigma}_{mh} \equiv \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_{im}\hat{\varepsilon}_{ih}$, $\hat{\varepsilon}_{im} \equiv y_{im} - x_{im}'\hat{\delta}_m$ $(m, h = 1, 2, \ldots, M)$ *for some consistent estimator* $\hat{\delta}_m$ *of* δ_m

($\hat{\delta}_m$ is usually the 2SLS estimator for equation m. So for the cross moments we need 2 2SLS estimators)

[33] This is bc

Large Sample Properties of 3SLS: (SEE ANALYTICAL EXERCISES FOR CHAPTER 4)
Suppose Assumptions 4.1 – 4.5 and 4.7 hold, and suppose that there exists a common set of instruments for all equations ($\mathbf{z}_{im} = \mathbf{z}_i$). Suppose further that $E(\mathbf{z}_{ih}\mathbf{z}_{im})$ exists and is finite for all m, h = 1,2,…M. Let $\hat{\Sigma}$ be the MxM matrix of estimated error cross moments calculated using 2SLS residuals. Then,

(a) $\hat{\delta}_{3SLS} = \hat{\delta}_{FIVE}(\hat{S}^{-1})$ is consistent, asymptotically normal, and efficient

with Avar($\hat{\delta}_{3SLS}$) = $(\Sigma_{ZX}' S^{-1} \Sigma_{ZX})^{-1} = n\left(X'\left(\Sigma^{-1} \otimes Z(Z'Z)^{-1}Z'\right)X\right)^{-1}$

(b) The estimated asymptotic variance $\hat{A}\text{var}(\hat{\delta}_{3SLS}) = n\left(X'\left(\hat{\Sigma}^{-1} \otimes Z(Z'Z)^{-1}Z'\right)X\right)^{-1}$ is consistent for Avar($\hat{\delta}_{3SLS}$) (SEE 4.5.17)

(c) Sargan's Statistic / J Statistic: $J(\hat{\delta}_{3SLS}, \hat{S}^{-1}) \equiv n \cdot g_n(\hat{\delta}_{3SLS})' \hat{S}^{-1} g_n(\hat{\delta}_{3SLS}) \xrightarrow{D} \chi^2\left(MK - \sum_m D_m\right)$

3. Seemingly Unrelated Regressions: Suppose in addition to common instruments, $\mathbf{z}_i$ = union of ($\mathbf{x}_{i1}$,…, $\mathbf{x}_{i1}$) (all regressors are instruments!)
   - The **SUR "cross orthogonality"** condition is equivalent to $E(\mathbf{x}_{im} \cdot \varepsilon_{ih}) = \mathbf{0}$ (m, h = 1,2,…,M)
     Predetermined regressors satisfy cross orthogonalities: not only are the regressors predetermined in each equation $E(\mathbf{x}_{im} \cdot \varepsilon_{im})$, but also they are predetermined in other equations (**so regressors in any equation is an instrument for all equations**).
     This simplification produces the **SUR estimator** → WHAT IS THE IMPORTANCE OF CROSS ORTHOG?

   - With the above, (S and g still defined the same) we get that… (bc X_i are in the column space of P_Z)

$$\boxed{\hat{\delta}_{SUR} = \hat{\delta}_{3SLS} = \hat{\delta}_{FIVE}\left(\hat{S}^{-1}\right) = \left(s_{ZX}' \hat{S}^{-1} s_{ZX}\right)^{-1} s_{ZX}' \hat{S}^{-1} s_{ZY} = \left[X'\left(\hat{\Sigma}^{-1} \otimes I_n\right)X\right]^{-1} X'\left(\hat{\Sigma}^{-1} \otimes I_n\right)Y}$$ [34]

$$S_{ZX} = \frac{1}{n}\sum_{i=1}^{n} Z_i X_i' = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n} \vec{z}_{i1(K_1x1)}\vec{x}_{i1(1xd_1)}' & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sum_{i=1}^{n} \vec{z}_{iM(K_1x1)}\vec{x}_{iM(1xd_1)}' \end{bmatrix} = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n} \vec{z}_i\vec{x}_{i1(1xd_1)}' & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sum_{i=1}^{n} \vec{z}_i\vec{x}_{iM(1xd_1)}' \end{bmatrix} = \frac{1}{n}\left(I_M \otimes Z\right)' X$$

$$where\ Z = \begin{bmatrix} \vec{z}_{1(1xK)}' \\ \vdots \\ \vec{z}_{n(1xK)}' \end{bmatrix} = \begin{bmatrix} z_{11} & \cdots & z_{1K} \\ \vdots & & \vdots \\ z_{n1} & \cdots & z_{nK} \end{bmatrix}_{nxK} \quad and \quad X = \begin{bmatrix} X_{1(nxD_M)} & & \\ & \ddots & \\ & & X_{M(nxD_M)} \end{bmatrix}_{nMx\sum_m D_m}, \quad X_m = \begin{bmatrix} \vec{x}_{1m}' \\ \vdots \\ \vec{x}_{nm}' \end{bmatrix}_{nxD_m} \quad (data\ matrix\ for\ mth\ equation's\ regressors)$$

$$s_{ZY} = \frac{1}{n}\left(I_M \otimes Z\right)'Y \quad where\ Y = \begin{bmatrix} \vec{y}_1 \\ \vdots \\ \vec{y}_n \end{bmatrix}_{nMx1}, \ y_m = \begin{bmatrix} y_{m1} \\ \vdots \\ y_{mn} \end{bmatrix}_{nx1} \quad (data\ matrix\ for\ mth\ equation's\ dependent\ \text{var})$$

$$\therefore \left(s_{ZX}' \hat{S}^{-1} s_{ZX}\right)^{-1} = \left(\frac{1}{n}X'\left(I_M \otimes Z\right)\left[n\left(\hat{\Sigma}^{-1} \otimes (Z'Z)^{-1}\right)\right]\frac{1}{n}\left(I_M \otimes Z\right)'X\right)^{-1} = n\left(X'\left(\hat{\Sigma}^{-1} \otimes Z(Z'Z)^{-1}\right)\left(I_M \otimes Z\right)'X\right)^{-1} = n\left(X'\left(\hat{\Sigma}^{-1} \otimes Z(Z'Z)^{-1}Z\right)X\right)^{-1}$$

$$s_{ZX}' \hat{S}^{-1} s_{ZY} = \frac{1}{n}X'\left(I_M \otimes Z\right)\left[n\left(\hat{\Sigma}^{-1} \otimes (Z'Z)^{-1}\right)\right]\frac{1}{n}\left(I_M \otimes Z\right)'y = \frac{1}{n}X'\left(\hat{\Sigma}^{-1} \otimes Z(Z'Z)^{-1}\right)\left(I_M \otimes Z\right)'y = \frac{1}{n}X'\left(\hat{\Sigma}^{-1} \otimes Z(Z'Z)^{-1}Z'\right)Y$$

Large Sample Properties of SUR: (SEE ANALYTICAL EXERCISES FOR CHAPTER 4)

Suppose Assumptions 4.1 – 4.5 and 4.7 hold, and suppose that there exists a common set of instruments for all equations ($\mathbf{z}_{im} = \mathbf{z}_i$), and suppose $\mathbf{z}_i$ = union of ($\mathbf{x}_{i1},\ldots, \mathbf{x}_{i1}$). Suppose further that E($\mathbf{z}_{ih}\mathbf{z}_{im}$) exists and is finite for all m, h = 1,2,…M. Let $\hat{\Sigma}$ be the MxM matrix of estimated error cross moments calculated using **OLS residuals**. Then,

(a) $\hat{\delta}_{SUR} = \hat{\delta}_{3SLS} = \hat{\delta}_{FIVE}(\hat{S}^{-1})$ is consistent, asymptotically normal, and efficient

with Avar($\hat{\delta}_{SUR}$) = $(\Sigma_{ZX}' S^{-1}\Sigma_{ZX})^{-1} = n\left(X'\left(\Sigma^{-1}\otimes I_n\right)X\right)^{-1}$

(b) The estimated asymptotic variance $\hat{A}\text{var}(\hat{\delta}_{SUR}) = n\left(X'\left(\hat{\Sigma}^{-1}\otimes I_n\right)X\right)^{-1}$ is consistent for Avar($\hat{\delta}_{SUR}$) (SEE 4.5.17)

(c) Sargan's Statistic / J Statistic: $J(\hat{\delta}_{SUR},\hat{S}^{-1}) \equiv n\cdot g_n(\hat{\delta}_{SUR})'\hat{S}^{-1}\, g_n(\hat{\delta}_{SUR}) \xrightarrow{D} \chi^2\left(MK - \sum_m D_m\right)$

- 3SLS vs. SUR:

  SUR estimator is a special case of 3SLS. In 3SLS, the initial (consistent) estimator used to calculate $\hat{\Sigma}$ was 2SLS. But we know 2SLS when regressors are a subset of the instrument set (i.e. when regressors are predetermined) is OLS (Ch. 3.8 Review Question 7). So for SUR the initial estimator is the OLS estimator.

- SUR vs. OLS:

  Since the regressors are predetermined, the system can also be estimated by the equation-by-equation OLS.
  Then **why SUR over OLS?**

  The relationship between SUR and equation-by-equation OLS is strictly analogous to the relationship between M.E. GMM and equation-by-equation GMM: Under conditional homoskedasticity, the efficient M.E. GMM is FIVE, which is equal to SUR under the cross orthogonality condition. On the other hand, under conditional homoskedasticity, efficient single-equation GMM is SLS, which is OLS under cross orthogonality since it implies that regressors are predetermined.

  2 Cases to consider:

  (i) Each equation is just identified:

  Then, since common instrument set is the union of all regressors, this is possible only if the regressors are the same for all equations, i.e. $\mathbf{x}_{im} = \mathbf{z}_i$ for all m → Multivariate Regression → equation by equation OLS!

  (ii) At least one equation is overidentified:

  Then, SUR is more efficient than equation by equation OLS, unless the equations are "unrelated" to each other in the sense that E($e_{im}e_{ih}\mathbf{x}_{im}\mathbf{x}_{oh}$) = σ_{mh} E($\mathbf{x}_{im}\mathbf{x}_{oh}$) = 0 (first equality true by conditional homoskedasticity) for all m !=h (recall, in this case the ME GMM is asymptotically equivalent to SE GMM).

  Since E($\mathbf{x}_{im}\mathbf{x}_{oh}$) is assumed to be non-0 (by rank condition), σ_{mh} E($\mathbf{x}_{im}\mathbf{x}_{oh}$) = 0 iff σ_{mh} = 0 (cov of error terms = 0)

  Therefore, SUR is more efficient than OLS if σ_{mh} != 0 for some pair (m,h), and they are are asymptotically equivalent if σ_{mh} = 0 for all (m,h)

$$\left(S_{ZX}'\hat{S}^{-1}S_{ZX}\right)^{-1} = n\left(X'\left(\hat{\Sigma}^{-1}\otimes Z(Z'Z)^{-1}Z\right)X\right)^{-1} = n\left(X'\left(\hat{\Sigma}^{-1}\otimes P_Z\right)X\right)^{-1}$$

$$X'\left(\hat{\Sigma}^{-1}\otimes P_Z\right)X = \begin{pmatrix} X_{1(D_1xn)}' & & \\ & \ddots & \\ & & X_M' \end{pmatrix}\begin{pmatrix} \hat{\sigma}_{11}P_{Z(nxn)} & \cdots & \hat{\sigma}_{1M}P_Z \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{M1}P_Z & \cdots & \hat{\sigma}_{MM}P_Z \end{pmatrix}\begin{pmatrix} X_{1(nxD_1)} & & \\ & \ddots & \\ & & X_M \end{pmatrix}$$

$$= \begin{pmatrix} \hat{\sigma}_{11}X_{1(D_1xn)}'P_{Z(nxn)}'P_{Z(nxn)}X_{1(nxD_1)} & & \\ & \ddots & \\ & & \hat{\sigma}_{MM}X_M'P_Z'P_ZX_M \end{pmatrix} \quad \text{since } P_Z \text{ symm idempotent}$$

$$= \begin{pmatrix} \hat{\sigma}_{11}X_{1(D_1xn)}'X_{1(nxD_1)} & & \\ & \ddots & \\ & & \hat{\sigma}_{MM}X_M'X_M \end{pmatrix} \quad \text{since } X_m \text{ is in the column space of } P_Z \ (Z \text{ is projection onto space spanned by union of } \vec{x}_m\text{'s})$$

$$= \begin{pmatrix} X_{1(D_1xn)}' & & \\ & \ddots & \\ & & X_M' \end{pmatrix}\begin{pmatrix} \hat{\sigma}_{11}I_n & & \\ & \ddots & \\ & & \hat{\sigma}_{MM}I_n \end{pmatrix}\begin{pmatrix} X_1 & & \\ & \ddots & \\ & & X_M \end{pmatrix}$$

$$= X'\left(\hat{\Sigma}^{-1}\otimes I_n\right)X$$

Similarly,

$$S_{ZX}'\hat{S}^{-1}s_{ZY} = \frac{1}{n}X'\left(\hat{\Sigma}^{-1}\otimes Z(Z'Z)^{-1}Z'\right)Y = \frac{1}{n}X'\left(\hat{\Sigma}^{-1}\otimes I_n\right)Y$$

4. Multivariate Regression : Suppose in addition to common instruments, $\mathbf{z}_i$ = union of ($\mathbf{x}_{i1}, \ldots, \mathbf{x}_{i1}$), and all equations are identified.
   - This condition implies $\mathbf{x}_{im} = \mathbf{z}_i$ for all m (same regressors for all equations, which are all exogenous)[35]
   - Simplification of **X**:
     With the assumption, we get that $\mathbf{X} = \mathbf{I}_M \otimes \mathbf{Z}$
     - With the above, (S and g still defined the same) we get that…

$$\boxed{\hat{\delta}_{MVR} = \hat{\delta}_{SUR} = \hat{\delta}_{3SLS} = \hat{\delta}_{FIVE}\left(\hat{S}^{-1}\right) = \left(s_{ZX}'\hat{S}^{-1}s_{ZX}\right)^{-1} s_{ZX}'\hat{S}^{-1}s_{ZY} = \left[I_M \otimes (Z'Z)^{-1}Z'\right]Y}$$ [36]

   - Multivariate Regression Estimator = Equation-by-Equation OLS
     Recall, when equations are just identified, the ME GMM and efficient single-equation GMM are numerically equal to the IV estimator, and since the regressors are predetermined, the GMM estimator of the multivariate regression model is just equation-by-equation OLS.

   - Multivariate Regression Interpretation of SUR: We can think of SUR model as a multivariate regression model with a priori exclusion restrictions. (i.e. a system with same regressors but with restrictions on certain coefficients)

Multiple Equation GMM with Common Coefficients: We modify the ME GMM model to allow this restriction. We get RE and Pooled OLS.
   - **Background**:
     In many applications (e.g. panel data) we deal with a special case of ME GMM model where the number of regressors is the same across equations with the same coefficients. How do we apply ME GMM while imposing common coefficient restriction

Random Effects Estimator and Pooled OLS

---

[35] This is true bc $\mathbf{z}_i$ instruments all m equations. So, if not, $\mathbf{x}_{im}$ c $\mathbf{z}_i$ for some m, then dim($\mathbf{z}_i$) > dim($\mathbf{x}_{im}$) and the mth equation is overidentified.

[36]

$$X'\left(\hat{\Sigma}^{-1} \otimes I_n\right)X = (I_M \otimes Z)'\left(\hat{\Sigma}^{-1} \otimes I_n\right)(I_M \otimes Z) = \left(I_M' \otimes Z'\right)\left(\hat{\Sigma}^{-1} \otimes I_n\right)(I_M \otimes Z) = \left(\hat{\Sigma}^{-1} \otimes Z'\right)(I_M \otimes Z) = \left(\hat{\Sigma}^{-1} \otimes Z'Z\right)$$

$$X'\left(\hat{\Sigma}^{-1} \otimes I_n\right)Y = \left(I_M' \otimes Z'\right)\left(\hat{\Sigma}^{-1} \otimes I_n\right)Y = \left(\hat{\Sigma}^{-1} \otimes Z'\right)Y$$

$$\therefore \hat{\delta}_{MVR} = \left[X'\left(\hat{\Sigma}^{-1} \otimes I_n\right)X\right]^{-1} X'\left(\hat{\Sigma}^{-1} \otimes I_n\right)Y = \left[\hat{\Sigma}^{-1} \otimes Z'Z\right]^{-1}\left(\hat{\Sigma}^{-1} \otimes Z'\right)Y = \left(\hat{\Sigma} \otimes (Z'Z)^{-1}\right)\left(\hat{\Sigma}^{-1} \otimes Z'\right)Y = \left(I_m \otimes (Z'Z)^{-1}Z'\right)Y$$

Relationship between ME. Estimators

	Multiple Equation GMM	**FIVE**	**3SLS**	**SUR**	**Multivariate Regression**
The Model	Assumptions 4.1 – 4.6	Assumptions 4.1 – 4.5, 4.7 E($\mathbf{x}_{im}\mathbf{x}_{ih}$') finite Note: 4.6 (finite 4th moments) not needed under homoskedasticity, 4.7	Assumptions 4.1 – 4.5, 4.7, E($\mathbf{x}_{im}\mathbf{x}_{ih}$') finite $\mathbf{z}_{im} = \mathbf{z}_i$ for all m (same set of instruments across equations)	Assumptions 4.1 – 4.5, 4.7 E($\mathbf{x}_{im}\mathbf{x}_{ih}$') finite $\mathbf{z}_{im} = \mathbf{z}_i$ for all m $\mathbf{z}_i = \text{union}_m\, \mathbf{x}_{im}$ (This is the cross-equation orthogonality: $E(\mathbf{x}_{im} \,.\, \varepsilon_{ih}) = \mathbf{0}$ (m, h = 1,2,…,M))	Assumptions 4.1 – 4.5, 4.7 E($\mathbf{x}_{im}\mathbf{x}_{ih}$') finite $\mathbf{z}_{im} = \mathbf{z}_i$ for all m $\mathbf{x}_{im} = \mathbf{z}_i$ for all m
S	$\begin{bmatrix} E(\varepsilon_{i1}\varepsilon_{i1}\bar z_{i1}\bar z_{i1}') & \cdots & E(\varepsilon_{i1}\varepsilon_{iM}\bar z_{i1}\bar z_{iM}') \\ \vdots & & \vdots \\ E(\varepsilon_{i1}\varepsilon_{i1}\bar z_{iM}\bar z_{i1}') & \cdots & E(\varepsilon_{iM}\varepsilon_{iM}\bar z_{iM}\bar z_{iM}') \end{bmatrix}$	$\begin{bmatrix} E(\varepsilon_{i1}\varepsilon_{i1}\bar z_{i1}\bar z_{i1}') & \cdots & E(\varepsilon_{i1}\varepsilon_{iM}\bar z_{i1}\bar z_{iM}') \\ \vdots & & \vdots \\ E(\varepsilon_{i1}\varepsilon_{i1}\bar z_{iM}\bar z_{i1}') & \cdots & E(\varepsilon_{iM}\varepsilon_{iM}\bar z_{iM}\bar z_{iM}') \end{bmatrix}$ $= \begin{bmatrix} \sigma_{11}E(\bar z_{i1}\bar z_{i1}') & \cdots & \sigma_{1M}E(\bar z_{i1}\bar z_{iM}') \\ \vdots & & \vdots \\ \sigma_{M1}E(\bar z_{iM}\bar z_{i1}') & \cdots & \sigma_{MM}E(\bar z_{iM}\bar z_{iM}') \end{bmatrix}$ $= E(Z_i'\Sigma Z_i)$	$\begin{bmatrix} \sigma_{11}E(\bar z_i\bar z_i) & \cdots & \sigma_{1M}E(\bar z_i\bar z_i) \\ \vdots & & \vdots \\ \sigma_{M1}E(\bar z_i\bar z_i) & \cdots & \sigma_{MM}E(\bar z_i\bar z_i) \end{bmatrix}$ $= \Sigma\otimes E(\bar z_i\bar z_i)_{MK\times MK}$	$\begin{bmatrix} \sigma_{11}E(\bar z_i\bar z_i) & \cdots & \sigma_{1M}E(\bar z_i\bar z_i) \\ \vdots & & \vdots \\ \sigma_{M1}E(\bar z_i\bar z_i) & \cdots & \sigma_{MM}E(\bar z_i\bar z_i) \end{bmatrix}$ $= \Sigma\otimes E(\bar z_i\bar z_i)_{MK\times MK}$	Irrelevant
$\hat S$	$\begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}\hat\varepsilon_{i1}\hat\varepsilon_{i1}\bar z_{i1}\bar z_{i1}' & \cdots & \frac{1}{n}\sum_{i=1}^{n}\hat\varepsilon_{i1}\hat\varepsilon_{iM}\bar z_{i1}\bar z_{iM}' \\ \vdots & & \vdots \\ \frac{1}{n}\sum_{i=1}^{n}\hat\varepsilon_{iM}\hat\varepsilon_{i1}\bar z_{iM}\bar z_{i1}' & \cdots & \frac{1}{n}\sum_{i=1}^{n}\hat\varepsilon_{MM}\hat\varepsilon_{MM}\bar z_{MM}\bar z_{MM}' \end{bmatrix}$	$E(Z_i'\hat\Sigma Z_i)$	$\hat S = \hat\Sigma\otimes\left(\frac{1}{n}\sum_{i=1}^{n}\bar z_i\bar z_i'\right)$ $= \hat\Sigma\otimes\frac{1}{n}Z'Z$ $\hat\Sigma$ *from 2SLS residuals*	$\hat S = \hat\Sigma\otimes\left(\frac{1}{n}\sum_{i=1}^{n}\bar z_i\bar z_i'\right)$ $= \hat\Sigma\otimes\frac{1}{n}Z'Z$ $\hat\Sigma$ *from OLS residuals*	Irrelevant
$\hat\delta(\hat S^{-1})$	$\left(S_{ZX}'\hat S^{-1}S_{ZX}\right)^{-1}S_{ZX}'\hat S^{-1}s_{ZY}$	$\left(S_{ZX}'\hat S^{-1}S_{ZX}\right)^{-1}S_{ZX}'\hat S^{-1}s_{ZY}$ $=\left(XZ'(Z\hat\Sigma Z')^{-1}ZX'\right)^{-1}XZ'(Z\hat\Sigma Z')^{-1}ZY$	$\left(S_{ZX}'\hat S^{-1}S_{ZX}\right)^{-1}S_{ZX}'\hat S^{-1}s_{ZY}$ $=\left[X'\left(\hat\Sigma^{-1}\otimes P_Z\right)X\right]^{-1}X'\left(\hat\Sigma^{-1}\otimes P_Z\right)Y$	$\left(S_{ZX}'\hat S^{-1}S_{ZX}\right)^{-1}S_{ZX}'\hat S^{-1}s_{ZY}$ $=\left[X'\left(\hat\Sigma^{-1}\otimes I_n\right)X\right]^{-1}X'\left(\hat\Sigma^{-1}\otimes I_n\right)Y$	Equation-by-Equation OLS
Avar $\hat\delta(\hat S^{-1})$	$(\Sigma_{ZX}'S^{-1}\Sigma_{ZX})^{-1}$ with **S** defined above	$(\Sigma_{ZX}'S^{-1}\Sigma_{ZX})^{-1}$ with **S** defined above	$n\left(X'\left(\Sigma^{-1}\otimes Z(Z'Z)^{-1}Z'\right)X\right)^{-1}$	$n\left(X'\left(\hat\Sigma^{-1}\otimes I_n\right)X\right)^{-1}$	OLS Formula
$\widehat{\text{A}}$var $\hat\delta(\hat S^{-1})$	$\left(S_{ZX}'\hat S^{-1}S_{ZX}\right)^{-1}$ with **S** defined above	$\left(S_{ZX}'\hat S^{-1}S_{ZX}\right)^{-1}$ with **S** defined above	$n\left(X'\left(\hat\Sigma^{-1}\otimes Z(Z'Z)^{-1}Z'\right)X\right)^{-1}$	$n\left(X'\left(\hat\Sigma^{-1}\otimes I_n\right)X\right)^{-1}$	OLS Formula

Note:

$$Z = \begin{bmatrix} \bar z_{1(1xK)}' \\ \vdots \\ \bar z_{n(1xK)}' \end{bmatrix} = \begin{bmatrix} z_{11} & \cdots & z_{1K} \\ \vdots & & \vdots \\ z_{n1} & \cdots & z_{nK} \end{bmatrix}_{nxK}, \quad X = \begin{bmatrix} X_{1(nxD_M)} & & \\ & \ddots & \\ & & X_{M(nxD_M)} \end{bmatrix}_{nMx\sum_m D_m}, \quad X_m = \begin{bmatrix} \bar x_{1m}' \\ \vdots \\ \bar x_{nm}' \end{bmatrix}_{nxD_m}, \delta_{\sum_m D_m} = \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_M \end{bmatrix}, \; Y = \begin{bmatrix} y_{1(nx1)} \\ \vdots \\ y_{M(nx1)} \end{bmatrix}_{nMx1}, \; y_m = \begin{bmatrix} y_{1m} \\ \vdots \\ y_{nm} \end{bmatrix}_{nx1}, \; \varepsilon = \begin{bmatrix} \varepsilon_{1(nx1)} \\ \vdots \\ \varepsilon_{M(nx1)} \end{bmatrix}_{nMx1}, \; \varepsilon_m = \begin{bmatrix} \varepsilon_{1m} \\ \vdots \\ \varepsilon_{nm} \end{bmatrix}_{nx1}$$

VII. **Simultaneous Equations, FIML (ML Counterpart to 3SLS) and LIML (ML Counterpart to 2SLS)**

Background: Given that we're going to estimate simultaneous equations system (with same instruments across all equations) via maximum likelihood, we will assume iid and normality. But first, we need to "complete" the system (# of endogenous variables = # equations)

- Recall, there are M linear equation we want to estimate, $y_{im} = \mathbf{x}_{im}'\boldsymbol{\delta}_m + \varepsilon_{im}$ ($m = 1, 2, \ldots, m$; $i = 1,2,\ldots,n$)
- Common instruments will be used across all equations (this is the 3SLS assumption)

1. Full Information Maximum Likelihood (ML estimation of 3SLS)

- **Complete System of Simultaneous Equations**

Setup: In order for the M equations of the system to form a "complete" system of simultaneous equations, we need:

A. # Endogenous Variables = #Equations: This implies that we can write the M-Equation system in **structural form**

$$\Gamma_{0(M\times M)}\mathbf{y}_{t(M\times 1)} + \mathbf{B}_{0(M\times K)}\mathbf{x}_{t(K\times 1)} = \varepsilon_{t(M\times 1)}$$

where $\mathbf{y}_t$ is the vector of endogenous variables in the system and $\mathbf{x}_t$ is the vector of exogenous variables in the system[37]

Note: It is not possible to estimate an incomplete system by FIML unless we complete it by adding appropriate equations This is in contrast to GMM. An incomplete system can be estimated via 3SLS (or M.E. GMM if we don't assume conditional homoskedasticity), as long **as the rank condition for identification is satisfied**.

Note: To complete the system, we can add equations involving instruments (Endogenous variable and instruments).

B. The square matrix $\Gamma_{0(M\times M)}$ is nonsingular: This implies that the **structural equation can be solved for endog. var's**

$$\mathbf{y}_{t(M\times 1)} = -\Gamma_0^{-1}{}_{(M\times M)}\mathbf{B}_{0(M\times K)}\mathbf{x}_{t(K\times 1)} + \Gamma_0^{-1}\varepsilon_{t(M\times 1)} \equiv \Pi'_{0(M\times K)}\mathbf{x}_{t(K\times 1)} + \mathbf{v}_t$$

- **Regression Form is a Multivariate Regression Model**

Def: **Reduced Form is**: $\boxed{\mathbf{y}_{t(M\times 1)} = \Pi'_{0(M\times K)}\mathbf{x}_{t(K\times 1)} + \mathbf{v}_t}$

Def: The coefficients in $\Pi'_{0(M\times K)}$ are called **reduced-form coefficients**

Predetermined Condition is satisfied

Since we have put all the predetermined variables in $\mathbf{x}_t$, then by assumption (of predetermined), $E(\mathbf{x}_t\,\mathbf{v}_t') = 0$.

Therefore, we have a (ME GMM) Multiple Regression Model (Same regressors/instruments for all equations)

- **Estimating System of Simultaneous Equations by FIML**

2 Assumptions:

A. Structural Error $\boldsymbol{\varepsilon}_t$ vector is jointly normal conditional on $\mathbf{x}_t$: $\varepsilon_t \mid \mathbf{x}_t \sim N_M(0, \Sigma_0)$

B. $\{\mathbf{y}_t, \mathbf{x}_t\}$ *iid* (to Strength Assumption 4.2 and 4.5)

From here, we obtain:

C. Distribution of Reduced form Error : $\boxed{\mathbf{v}_t \mid \mathbf{x}_t = \Gamma_0^{-1}\varepsilon_t \mid \mathbf{x}_t \sim N(0, \Gamma_0^{-1}\Sigma_0(\Gamma_0^{-1})')}$

Distribution of endogenous $\mathbf{y}_t$: $\boxed{\mathbf{y}_t \mid \mathbf{x}_t = -\Gamma_0^{-1}\mathbf{B}_0\mathbf{x}_t + \mathbf{v}_t \sim N(-\Gamma_0^{-1}\mathbf{B}_0\mathbf{x}_t, \Gamma_0^{-1}\Sigma_0(\Gamma_0^{-1})')}$

D. Log-likelihood function for the sample (i.e. our objective function) (pg 531 - 532 Hayashi):

$$Q_n(\delta,\Sigma) = -\frac{M}{2}\log(2\pi) + \frac{1}{2}\log\left(|\Gamma|^2\right) - \frac{1}{2}\log\left(|\Gamma|\right) - \frac{1}{2n}\sum_{t=1}^{n}\left(\Gamma\mathbf{y}_t + \mathbf{B}\mathbf{x}_t\right)'\Sigma^{-1}\left(\Gamma\mathbf{y}_t + \mathbf{B}\mathbf{x}_t\right)$$

Therefore, the FIML estimate of (δ_0, Σ_0) (i.e. the coefficients in the $\mathbf{B}_0$ and Γ_0 matrix and the variance of errors) is the (δ, Σ) that maximizes the objective function.

---

[37] How does this relate to instruments? If it's not complete, then we can add equations that involve instruments that we take to be true, which are orthogonal to the error terms (See LIML example later).

- **Properties of FIML**
  A. **Identification of the FIML estimator**
  The identification condition is equivalent to the rank condition being satisfied for each equation:
  $$\boxed{E(\mathbf{z}_t\mathbf{x}_{tm}^{'})\ is\ \ full\ column\ rank\ \ for\ all\ m=1,2,...,M}$$
  B. **Invariance**: Since FIML is an ML estimator, the invariance property holds (see HW2 and HW3 on why useful)

  C. **Asymptotic Properties of FIML**
  Consider the M-Equation system: $y_{im} = \mathbf{x}_{im}'\boldsymbol{\delta}_m + \varepsilon_{im}$ (m = 1, 2, ... , m; i = 1,2,…,n)
  Let $\boldsymbol{\delta}_0$ be the stacked vector collecting all the coefficients in the M-equation system.
  Suppose the following assumptions are satisfied
  A1: Rank condition for identification is satisfied: $\boxed{E(\mathbf{z}_t\mathbf{x}_{tm}^{'})\ is\ \ full\ column\ rank\ \ for\ all\ m=1,2,...,M}$
  A2: $E\left(\mathbf{z}_t\mathbf{z}_t^{'}\right)$ nonsingular
  A3: M-Equation system can be written as a complete system of simultaneousequations with Γ_0 nonsingular
  A4: $\varepsilon_t \mid \mathbf{x}_t \sim N(0,\Sigma_0)\ \ for\ some\ \Sigma_0\ \ p.d.$
  A5: $\{\mathbf{y}_t, \mathbf{x}_t\}\ iid$
  A6: Parameter space for (δ_0, Σ_0) is compact with the "true" parameter vector in the interior

  Then,

  (a) The FIML estimator $(\hat{\delta},\hat{\Sigma})$ which maximizes the objective function is consistent and asymptotically normal
  (b) The asymptotic variance of $\hat{\delta}$ is $n\left(X'\left(\Sigma^{-1}\otimes Z(Z'Z)^{-1}Z'\right)X\right)^{-1}$ (same as 3SLS)
  A consistent estimator of the asymptotic variance is $n\left(X'\left(\hat{\Sigma}^{-1}\otimes Z(Z'Z)^{-1}Z'\right)X\right)^{-1}$ (same as 3SLS)
  (c) The likelihood ratio static for testing overidentifying restrictions is asymptotically $ChiSq\left(KM-\sum_{m=1}^{M}L_m\right)$
  Furthermore, these asymptotic results hold even without the normality assumption.

- LIML vs. 3SLS: They are asymptotically equivalent. However, LIML has invariance property that 3SLS (and 2SLS don't).

2. Limited Information Maximum Likelihood (ML Estimation of 2SLS)
Setup: The difference here is that we are only estimating 1 equation, instead of a whole system, and we have endogeneity problem. The only "trick" here is that we need to "complete" the system by adding 1 more equation relating the endogenous variable to the set of predetermined variables (just take something you know to be "true").

Example: Completing the System for LIML
e.g. Suppose the population model of interest is $Y_1 = \beta_0 + \beta_1 Y_2 + \varepsilon$ and we want to estimate (β_0, β_1)
where Y_1 and Y_2 are endogenous. In order to complete the system, suppose we KNOW the following relationship, $Y_2 = \pi_0 + \pi_1'\mathbf{Z} + u$.

Then, we can write the structural form of the system as:

$$\left.\begin{matrix} Y_1 = \beta_0 + \beta_1 Y_2 + \varepsilon \\ Y_2 = \pi_0 + \pi_1'\mathbf{Z} + u \end{matrix}\right\} = \left.\begin{matrix} Y_1 - \beta_2 Y_2 = \beta_0 + \varepsilon \\ Y_2 = -\pi_0 + \pi_1'\mathbf{Z}_{(mx1)} + u \end{matrix}\right\} \Rightarrow \begin{bmatrix} 1 & -\beta_2 \\ 0 & 1 \end{bmatrix}_{2x2} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}_{2x1} = \begin{bmatrix} \beta_0 & \mathbf{0} \\ \pi_0 & \pi_1^{'} \end{bmatrix}_{2xm} \begin{bmatrix} 1 \\ \mathbf{Z} \end{bmatrix}_{mx1} + \begin{bmatrix} \varepsilon \\ u \end{bmatrix}$$
$$\Rightarrow \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}_{2x1} = \begin{bmatrix} 1 & -\beta_2 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \beta_0 & \mathbf{0} \\ \pi_0 & \pi_1^{'} \end{bmatrix} \begin{bmatrix} 1 \\ \mathbf{Z} \end{bmatrix} + \begin{bmatrix} 1 & -\beta_2 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} \varepsilon \\ u \end{bmatrix}$$

LIML vs. 2SLS
- They're both k class estimators. 2SLS is a k class estimator with k = 1 (p. 541). So when the equation is just identified (k = 1), LIML = 2SLS numerically.
- LIML and 2SLS have same asymptotic distribution, so we cannot prefer one over the other on asymptotic grounds
- In finite sample, LIML has invariance property while 2SLS does not (which makes LIML more desirable)
  Other literature suggest that LIML should be preferred in finite sample over 2SLS

Example of LIML (271A PS2 Empirical Ex #2): Single Equation System, Completed by Equation Given

Suppose the population model of interest is: $Y_1 = \beta_0 + \beta_1 Y_2 + \varepsilon$

We suspect endogeneity, and we complete system with: $Y_2 = \pi_0 + \pi_1 Z + u$

Such that $E\left(\begin{bmatrix}1\\Z\end{bmatrix}\varepsilon\right) = 0, \begin{bmatrix}\varepsilon\\u\end{bmatrix} | Z \sim N(0,\Sigma) \text{ where } \Sigma = \begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{bmatrix}$, and we observe an iid sample of (Y_1,Y_2,Z)

Derive LIML Estimator (Use Invariance Property!):

$$Y_1 = \beta_0 + \beta_1 Y_2 + \varepsilon \Rightarrow Y_1 = \beta_0 + \beta_1(\pi_0 + \pi_1 Z + u) + \varepsilon = (\beta_0 + \beta_1\pi_0) + \beta_1\pi_1 Z + (\beta_1 u + \varepsilon) = \alpha_0 + \alpha_1 Z + v$$

$$E(Zv) = E\left(Z\left(\beta_1 u + \varepsilon\right)\right) = \beta_1 E(Zu) + E(Z\varepsilon) = 0$$

So, orthogonality condition holds. E(ZZ') assumed to be invertible. And we have iid and Gaussian errors, therefore, we can estimate this equation consistently by OLS → **But OLS with is the same as the MLE estimate (under Gaussian errors assumption)**
Similarly, we can estimate the second equation consistently via OLS to obtain MLE estimate (under Gaussian errors)

$$\text{We obtain MLE estimates } \hat{\alpha}_0, \hat{\alpha}_1, \hat{\pi}_0, \hat{\pi}_1 \text{ from OLS}$$

$$\text{We also know } \alpha_1 = \beta_1\pi_1 \Rightarrow \text{By inv. prop of MLE, } \hat{\beta}_1 = \frac{\hat{\alpha}_1}{\hat{\pi}_1}$$

$$\text{Similarly, } \alpha_0 = \beta_0 + \beta_1\pi_0 \Rightarrow \text{By inv. prop MLE, } \hat{\beta}_0 = \hat{\alpha}_0 - \hat{\beta}_1\hat{\pi}_0$$

$$(\hat{\beta}_0, \hat{\beta}_1) \text{ are the LIML Estimators!}$$

Example of FIML (271A PS3 #5): Multiple (2) Equation System, Completed by Equation Given

The structural model is:

$$Y_1 = \gamma_{12}Y_2 + \gamma_{13}Y_3 + \delta_{13}Z_3 + \delta_{14}Z_4 + u_1$$

$$Y_2 = \gamma_{22}Y_1 + \delta_{21}Z_1 + u_2$$

$$Y_3 = \delta_{31}Z_1 + \delta_{32}Z_2 + \delta_{33}Z_3 + \delta_{34}Z_4 + u_4$$

$$\text{where } Z_1 = 1,\ E(u_s) = 0 \text{ for } s = 1,2,3 \text{ and } E(Z_j u_s) = 0 \text{ for } j = 1,..,4,\ s = 1,2,3$$

$$\text{Assume in addition that } \delta_{13} + \delta_{14} = 1$$

1. 3 Cases of Endogeneity (Simultaneity, Errors in Variables/Measurement Error, Omitted Varaibes), Examples.

A. **Simultaneity Example**: (Working's) Simultaneous Equations Model for Market Equilibrium

Setup: The "true" relationship between demand and supply of coffee is modeled as follows

Demand Equation: $q_i^d = \alpha_0 + \alpha_1 p_i + u_i$ (u_i represents factors that influence coffee demand other than price)

Supply Equation: $q_i^s = \beta_0 + \beta_1 p_i + v_i$ (v_i represents factors that influence coffee supply other than price)

Market Equilibrium: $q_i^d = q_i^s$

Note: We assume $E(u_i) = 0$ and $E(v_i) = 0$ (if not, include nonzero means in the intercepts)

Endogeneity: Here, the regressor p_i is **endogenous/not predetermined**, i.e. not orthogonal to the (contemporaneous) error term, and therefore does not satisfy the orthogonality condition that
$E(p_i \,.\, u_i) = 0 \Leftrightarrow \text{cov}(p_i,u_i) = 0$ and $E(p_i \,.\, v_i) = 0 \Leftrightarrow \text{cov}(p_i,v_i) = 0$

The endogeneity in this example arises from the face that price is a function of both error terms u_i and v_i, which is a result of market equilibrium[38]. Therefore, $\text{cov}(p_i,u_i) = 0$ and $\text{cov}(p_i,v_i) = 0$ iff $\text{Var}(u_i) = 0$ and $\text{Var}(v_i) = 0$ respectively. Not possible (except in the extreme case when, for example, there are no other factors that shift demand, so $u_i = 0$)!

Problem with Endogeneity: Endogeneity Bias

When we regress observed quantity on a constant and price, we neither estimate the demand nor supply curve because price is endogenous in both equations. Recall that the OLS estimator is consistent for the least squares projection coefficients: in this case, the least squares projection of (true) q_i on a constant and (true) p_i gives a coefficient of p_i given by $\text{Cov}(p_i,q_i)/\text{Var}(p_i)$

Suppose we observe $\{q_i, p_i\}$ and we regress q_i on a constant and p_i, what is it that we estimate?

OLS estimate of the price coefficient $\hat{\alpha}_1$ (from the demand equation)

$$\xrightarrow{P} \frac{Cov(p_i,q_i)}{Var(p_i)} = \frac{Cov(p_i,\alpha_0+\alpha_1 p_i+u_i)}{Var(p_i)} = \frac{Cov(p_i,\alpha_1 p_i+u_i)}{Var(p_i)} = \frac{\alpha_1 Var(p_i)+Cov(u_i,p_i)}{Var(p_i)} = \alpha_1 + \frac{Cov(u_i,p_i)}{Var(p_i)}$$

$$\text{Asymptotic Bias} = \frac{Cov(u_i,p_i)}{Var(p_i)}$$

OLS estimate of the price coefficient $\hat{\beta}_1$ (from the supply equation)

$$\xrightarrow{P} \frac{Cov(p_i,q_i)}{Var(p_i)} = \frac{Cov(p_i,\beta_0+\beta_1 p_i+u_i)}{Var(p_i)} = \frac{Cov(p_i,\beta_1 p_i+v_i)}{Var(p_i)} = \frac{\beta_1 Var(p_i)+Cov(v_i,p_i)}{Var(p_i)} = \beta_1 + \frac{Cov(v_i,p_i)}{Var(p_i)}$$

$$\text{Asymptotic Bias} = \frac{Cov(v_i,p_i)}{Var(p_i)}$$

Since $Cov(p_i,u_i) \neq 0$ and $Cov(p_i,v_i) \neq 0$, therefore **endogeneity bias/simultaneous equation bias/simultaneity bias** exists! (bc regressor and error term are related to each other through a system of simultaneous equations).
So, **OLS estimator is not consistent for either α_1 or β_1.**

Solution: Instrumental Variables and 2 Stage Least Squares

The reason why demand curve nor supply curve can be consistently estimated because we cannot infer from the data whether the observed changes in price and quantity is due to a shift in demand or supply. Therefore, we might be able to estimate the demand/supply if some of the factors that shift the supply/demand curves are observable.

Def: A predetermined variable (predetermined in the system) that is correlated with the endogenous regressor is called an **instrumental variable** or **instrument**. Sometimes we call it a **valid instrument** to emphasize that the correlation with the endogenous regressor is not 0.

Observable Supply Shifters:

Given "appropriate" observable supply shifters (Instrument), we can estimate demand and supply!

Suppose the supply shifter v_i can be divided into an observable factor x_i and an unobservable factor ξ_I with $\text{Cov}(x_i.\xi_I)=0$[39]

---

[38] To see endogeneity, treat the 3 equations as a system of simultaneous equations and solve for p_i and q_i

$$q_i^d = q_i^s \Rightarrow \alpha_0 + \alpha_1 p_i + u_i = \beta_0 + \beta_1 p_i + v_i \Rightarrow (\alpha_1 - \beta_1)p_i = (\beta_0 - \alpha_0) + (v_i - u_i) \Rightarrow p_i = \frac{(\beta_0 - \alpha_0)}{(\alpha_1 - \beta_1)} + \frac{(v_i - u_i)}{(\alpha_1 - \beta_1)}$$

$$So,\ Cov(p_i,u_i) = Cov\left(\frac{(\beta_0 - \alpha_0)}{(\alpha_1 - \beta_1)} + \frac{(v_i - u_i)}{(\alpha_1 - \beta_1)}, u_i\right) = \frac{1}{(\alpha_1 - \beta_1)} Cov((v_i - u_i), u_i) = \frac{1}{(\alpha_1 - \beta_1)}(Cov(v_i,u_i) - Var(u_i)) = \frac{-Var(u_i)}{(\alpha_1 - \beta_1)}\ \ (Since\ Cov(v_i,u_i) = 0\ by\ assumption)$$

$$Cov(p_i,v_i) = Cov\left(\frac{(\beta_0 - \alpha_0)}{(\alpha_1 - \beta_1)} + \frac{(v_i - u_i)}{(\alpha_1 - \beta_1)}, v_i\right) = \frac{1}{(\alpha_1 - \beta_1)} Cov((v_i - u_i), v_i) = \frac{1}{(\alpha_1 - \beta_1)}(Var(v_i) - Cov(v_i,u_i)) = \frac{Var(v_i)}{(\alpha_1 - \beta_1)}\ \ (Since\ Cov(v_i,u_i) = 0\ by\ assumption)$$

→ Supply Equation: $q_i^s = \beta_0 + \beta_1 p_i + \beta_2 x_i + \xi_i$

Suppose further that the observed supply shifter x_i is **predetermined in the demand equation**, i.e. uncorrelated with the error term u_i (e.g. think of x_i is the temperature in coffee growing regions). If the temperature (x_i) is uncorrelated with the unobserved factors that shift demand (u_i), i.e. temperature (x_i) is an instrument (for the demand equation), it would be possible to extract from observed price movements a component that is related to the temperature (i.e. the observed supply shifter) but uncorrelated with the demand shifter. Then, we can estimate the demand curve by examining the relationship between coffee consumption and that component of price.

IV Estimator for α_1: We derive the IV estimator for α_1 below
We can re-express price as:

$$q_i^d = q_i^s \Rightarrow \alpha_0 + \alpha_1 p_i + u_i = \beta_0 + \beta_1 p_i + \beta_2 x_i + \varsigma_i \Rightarrow (\alpha_1 - \beta_1)p_i = (\beta_0 - \alpha_0) + \beta_2 x_i + (\varsigma_i - u_i)$$

$$\Rightarrow p_i = \frac{(\beta_0 - \alpha_0)}{(\alpha_1 - \beta_1)} + \frac{\beta_2}{(\alpha_1 - \beta_1)} x_i + \frac{(\varsigma_i - u_i)}{(\alpha_1 - \beta_1)}$$

$$\therefore Cov(p_i, x_i) = Cov\left(\frac{(\beta_0 - \alpha_0)}{(\alpha_1 - \beta_1)} + \frac{\beta_2}{(\alpha_1 - \beta_1)} x_i + \frac{(\varsigma_i - u_i)}{(\alpha_1 - \beta_1)}, x_i\right) = \frac{\beta_2}{(\alpha_1 - \beta_1)} Var(x_i) + \frac{1}{(\alpha_1 - \beta_1)}\left(Cov(\varsigma_i, x_i) - Cov(u_i, x_i)\right)$$

$$= \frac{\beta_2}{(\alpha_1 - \beta_1)} Var(x_i) \quad \text{since } Cov(\varsigma_i, x_i) = 0 \text{ by construction and } Cov(u_i, x_i) = 0 \text{ by assumption}$$

$$\neq 0 \quad (\text{So } x_i \text{ a valid instrument})$$

With a valid instrument, we can estimate the price coefficient α_1 of demand curve consistently.

$$Cov(q_i, x_i) = Cov(\alpha_0 + \alpha_1 p_i + u_i, x_i) = \alpha_1 Cov(p_i, x_i) + Cov(u_i, x_i) = \alpha_1 Cov(p_i, x_i) \text{ since } Cov(u_i, x_i) = 0 \text{ by assumption}$$

$$\therefore \alpha_1 = \frac{Cov(q_i, x_i)}{Cov(p_i, x_i)}$$

If we observe an iid sample (q_i, p_i, x_i), then by the analogy principle, the natural (consistent) estimator is:
(**we say the endougenous regressor p_i is instrumente by x_i**)

$$\hat{\alpha}_{1,IV} = \frac{Sample\ \text{cov}\ bt\ x_i \text{ and } q_i}{Sample\ \text{cov}\ bt\ x_i \text{ and } p_i} = \frac{\sum_i x_i q_i}{\sum_i p_i x_i} = \frac{\sum_i x_i(\alpha_0 + \alpha_1 p_i + \varepsilon_i)}{\sum_i p_i x_i} = \alpha_0 \frac{\sum_i x_i}{\sum_i p_i x_i} + \alpha_1 \frac{\sum_i p_i x_i}{\sum_i p_i x_i} + \frac{\sum_i \varepsilon_i x_i}{\sum_i p_i x_i}$$ [40]

→ IV estimator is consistent for α_1

An instrumental variable is one that is correlated with the independent variable but not with the error term. The estimator is

$$\widehat{\beta}_{IV} = \frac{\sum_i z_i y_i}{\sum_i z_i x_i} = \frac{\sum_i z_i(x_i\beta + \epsilon_i)}{\sum_i z_i x_i} = \beta + \frac{\sum_i z_i \epsilon_i}{\sum_i z_i x_i}.$$

When z and ε are uncorrelated, the final term vanishes in the limit providing a consistent estimator. Note that when x is uncorrelated with the error term, x is itself an instrument. In that case the OLS estimator is a type of IV estimator.

The approach above generalizes in a straightforward way to a regression with multiple explanatory variables. Suppose X is the $T \times K$ matrix of explanatory variables resulting from T observations on K variables. Let Z be a $T \times K$ matrix of instruments. Then,

$$\widehat{\beta}_{IV} = (Z'X)^{-1}Z'Y = (Z'X)^{-1}Z'(X\beta + \epsilon) = \beta + (Z'X)^{-1}Z'\epsilon.$$

---

[39] This decomposition is always possible by the projection theorem. v_i can be expressed as the projection onto the space spanned by x_i and the orthogonal complement (remember, v_i includes all factors that affect supply, so by definition has at least as many dimensions than x_i). i.e. If the least squares projection of v_i on a constant and x_i is $E^*(v_i \mid 1\ x_i) = \gamma_0 + \beta_2 x_i$. Define $\xi_i = v_i - \gamma_0 + \beta_2 x_i$. By definition, ξ_i is orthogonal to x_i and $E(\xi_i) = 0$, therefore ξ_i, x_i uncorrelated. Substituting this into the original supply equation, and combining the intercept terms we get the resulting expression.

[40] Recall, sample covariance can be expressed as

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i (x_i y_i - \bar{x} y_i - x_i \bar{y} + \overline{xy}) = \sum_i (x_i y_i) - \bar{x}\sum_i (y_i) - \bar{y}\sum_i (x_i) + \sum_i (\overline{xy}) = \sum_i (x_i y_i) - n\overline{xy} - n\overline{xy} + n\overline{xy} = \sum_i (x_i y_i) - n\overline{xy}$$

Here, average of

One computational method often used for implementing the technique is two-stage least-squares (2SLS). One advantage of this approach is that it can efficiently combine information from multiple instruments for *over-identified* regressions: where there are fewer covariates than instruments. Under the 2SLS approach, in a first stage, each endogenous covariate (predictor variable) is regressed on all valid instruments, including the full set of exogenous covariates in the main regression. Since the instruments are exogenous, these approximations of the endogenous covariates will not be correlated with the error term. So, intuitively they provide a way to analyze the relationship between the outcome variable and the endogenous covariates. In the second stage, the regression of interest is estimated as usual, *except* that in this each endogenous covariate is replaced with its approximation estimated in the first stage. The slope estimator thus obtained is consistent. A small correction must be made to the sum-of-squared residuals in the second-stage fitted model in order that the associated standard errors be computed correctly.

Stage 1: $\widehat{X} = Z(Z'Z)^{-1}Z'X$

Stage 2: $\widehat{B}_{\mathrm{IV}} = (\widehat{X}'\widehat{X})^{-1}\widehat{X}'Y$

Mathematically, this estimator is identical to the single stage estimator presented above when the number of instruments is the same as the number of covariates.

Two-Stage Least Squares (2SLS) Estimator for α_1: This is another procedure for consistently estimating α_1 which is named thusly because the procedure consists of running 2 least squares (OLS) regressions.

- **First Stage**: Endogenous regressor p_i is regressed on a constant and the predetermined variable x_i to obtain fitted values $\hat{p}_i$.
  (OLS coeff. for x_i is Sample Cov bt p_i and x_i / sample variance of x_i).
- **Second Stage**: Regress dependent variable q_i on a constant and $\hat{p}_i$.
  (OLS coeff. for x_i is Sample Cov bt $\hat{p}_i$ and x_i / sample variance of x_i).

The 2nd stage estimates the equation (bracketed term is error): $q_i = \alpha_0 + \alpha_1 \hat{p}_i + [u_i + \alpha_1 (p_i - \hat{p}_i)]$

The 2SLS Estimator is consistent: Why???

Here, IV and 2SLS estimators are numerically the same (generally this will be true – see later).
Generally, 2SLS estimator can be written as an IV estimator with an appropriate choice of instruments, and the IV estimator is a particular GMM estimator.

B. **Errors-in-Variables/Measurement Error Example**:
This is the phenomenon that an otherwise predetermined regressor becomes endogenous when measured with error

C.

The point is, if you have endogeneity problems and you can find an instruments that are not correlated with error terms (i.e. meets orthogonality condition) but is correlated with the endogenous terms (and correlated with dependent variable only through the endogenous terms), then we can find a consistent estimator for the underlying parameter of interest.