

Lecture Note 5: INSTRUMENTAL VARIABLE (IV) ESTIMATION

WANCHUAN LIN

11/02/2007

Outline:

1. Math Review
 - (1). Rules for Limiting Distribution
 - (2). Central Limit Theorem
 - (3). Delta Method
 - (4). GLS
2. OLS with Endogenous Variables
 - (1) Classical Linear Regression Model
 - (2) Least Square Estimator with Endogenous Regressors
3. Examples of Endogenous Variables
 - (1) Measurement Errors
 - (2) Omitted Variables
 - (3) Simultaneous Equations
4. Estimation of Regression with Endogenous Variables
 - (1). IV for Just-Identified
 - a. IV Estimator
 - b. Method of Moments
 - c. Asy. Properties of IV
 - (2). IV for Over-Identified
 - a. Instrumental Variables
 - b. Weighted IV Estimator
 - d. Asy. Properties of Weighed IV Estimator
 - e. Choice of the Optimal Weighting Matrix
 - (3) GMM Representation of Weighed IV
 - (4) GLS Representation of Weighted IV
 - (5) 2SLS Representation of Weighted IV
 - (6) 3SLS (Feasible 2SLS)

1 Math Review¹

1.1 Definitions on Convergence

- **Convergence in Probability**

Definition 1-1: x_n converges in probability to a constant c if and only if $\forall \varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \Pr(\omega : \|x_n(\omega) - x(\omega)\| > \varepsilon) = 0$$

or equivalently

$$\lim_{n \rightarrow \infty} \Pr(\omega : \|x_n(\omega) - x(\omega)\| \leq \varepsilon) = 1$$

we then write

$$x_n \xrightarrow{p} c$$

$$\text{or } p \lim_{n \rightarrow \infty} x_n = c$$

Another equivalent definition is

Definition 1-2: A sequence of random variables x_n converges to a constant μ in probability if for all $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \Pr(|x_n - \mu| \geq \varepsilon) = 0$$

- **Convergence in Distribution**

Definition 2-1: x_n converges in distribution (or in law or weakly) to a random variable x with cdf $F(x)$ if and only if $F_n(x)$ converges to $F(x)$, i.e., $\lim_{n \rightarrow \infty} |F_n(x) - F(x)| = 0$ at each continuity point of $F(x)$. we then write

$$x_n \xrightarrow{d} x$$

Definition 2-2: Let the cdf $F_n(x)$ of the random variable x_n depend on n , $n = 1, 2, \dots$. If $F(y)$ is a distribution function and if $\lim_{n \rightarrow \infty} F_n(x) = F(x)$ for every continuity point of $F(x)$, then the sequence of random variables x_1, x_2, \dots is said to converge in distribution to a random variable with distribution function $F(x)$. We sometimes call $F(x)$ the limiting distribution.

Definition 2-3: A sequence of random variables x_1, x_2, \dots converges in distribution to a random variable x if at all continuity points of $F_x(x)$,

$$\lim_{n \rightarrow \infty} F_{x_n}(x) = F_x(x)$$

1.2 Rules for Limiting Distribution

1. if $x_n \xrightarrow{d} x$ and $p \lim y_n = y$, then

$$x_n y_n \xrightarrow{d} c x$$

$$x_n + y_n \xrightarrow{d} x + c$$

$$x_n / y_n \xrightarrow{d} x / c, \text{ if } c \neq 0$$

¹See Chapter 3 of *Econometric Analysis* (Fourth Edition) by W. H. Greene in detail.

2. Slutsky Theorem

If x_n converges in distribution to a random variable x and y_n converges in probability to a constant c , for a continuous function $g(\cdot)$, then

$$g(x_n) \xrightarrow{d} g(x)$$
$$p \lim g(y_n) = g(p \lim y_n) = g(y)$$

1.3 Central Limit Theorem (CLT)

Let x_1, x_2, \dots be a sequence of independent and identically distributed (iid) random variables with common finite mean μ and common finite variance σ^2 , then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (x_i - \mu) \xrightarrow{d} N(0, \sigma^2)$$
$$\sqrt{n}(\bar{x}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

where $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

1.4 Delta Method

If a sequence of random variables X_n satisfies

$$\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

then for any continuity differentiable function $g(\cdot)$ with derivative $g'(\cdot)$,

$$\sqrt{n}(g(X_n) - g(\mu)) \xrightarrow{d} N(0, (g'(\mu))^2 \sigma^2)$$

1.5 GLS Review

The original regression equation is

$$y = X\beta + \varepsilon$$

where

$$E(\mu) = 0 \quad \text{and} \quad E(\mu\mu^T) = \sigma^2\Omega$$

and σ^2 is un known and Ω is known. If $\Omega \neq I$ (The I matrix means the errors have constant variance and are non-correlated). Suppose that Ω is a positive definition symmetric matrix, then it can be factored out

$$\Omega = C\Lambda C^T$$

where the columns of C are the characteristic vectors of Ω and the characteristic roots of Ω are arrayed into the diagonal matrix Λ .

Generalized least squares minimizes

$$(y - X\beta)^T \Omega^{-1} (y - X\beta)$$

which is solved by

$$S(X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y$$

Since we can write $P^T = C\Lambda^{-1/2}$ so that $\Omega^{-1} = P^T P$, where P is a triangular matrix using the Choleski Decomposition, we have

$$\begin{aligned}(y - X\beta)^T \Omega^{-1} (y - X\beta) &= (y - X\beta)^T P^T P (y - X\beta) \\ &= (Py - PX\beta)^T (Py - PX\beta)\end{aligned}$$

So GLS is like regressing PX on Py . Furthermore

$$\begin{aligned}y &= X\beta + \varepsilon \\ Py &= PX\beta + P\varepsilon \\ y^* &= X^*\beta + \varepsilon^*\end{aligned}$$

So we have a new regression equation $y^* = X^*\beta + \varepsilon^*$ where if we examine the variance of the new errors ε^* , we have

$$Var(\varepsilon^*) = Var(P\varepsilon) = PVar(\varepsilon)P^T = P\sigma^2\Omega P^T = \sigma^2 I$$

Hence, the classical regression model applies to this transformed model. In the classical model, OLS is efficient.

$$\begin{aligned}\hat{\beta}_{OLS}^* &= (X^{*T}X^*)^{-1} X^{*T}y^* \\ &= (X^T P^T P X)^{-1} X^T P^T P y \\ &= (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y \\ &= \hat{\beta}_{GLS}\end{aligned}$$

So you can view OLS as one kind of GLS which uses a weighting matrix I instead of P .

Properties of GLS:

- $E(\varepsilon^* | X^*) = 0$, $\hat{\beta}_{GLS}$ is unbiased
- $\hat{\beta}_{GLS}$ is consistent
- $\hat{\beta}_{GLS}$ is efficient. The GLS is the minimum variance unbiased estimator in the generalized regression.

2 OLS with Endogenous Variables

2.1 Classical Linear Regression Model

The regression equation is

$$y = X\beta + \varepsilon$$

and satisfies the following four assumptions of Classical Regression Model

1. X is $n \times K$ with rank K
2. X is a nonstochastic matrix
3. $E(\varepsilon|X) = 0$
4. $Var(\varepsilon|X) = \sigma^2 I$

Based on the assumptions, the estimator $\hat{\beta}$ of β is

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y$$

The Properties of the Least Squares Estimator

Property 1: $\hat{\beta}_{OLS}$ is an unbiased estimator for β

$$\begin{aligned} E(\hat{\beta}_{OLS}|X) &= E((X^T X)^{-1} X^T y|X) \\ &= (X^T X)^{-1} X^T E(y|X) \\ &= (X^T X)^{-1} X^T X \beta \\ &= \beta \end{aligned}$$

Property 2:

$$Var(\hat{\beta}_{OLS}|X) = \sigma^2 (X^T X)^{-1}$$

$$\begin{aligned} Var(\hat{\beta}_{OLS}|X) &= Var((X^T X)^{-1} X^T y|X) \\ &= (X^T X)^{-1} X^T Var(y|X) X (X^T X)^{-1} \\ &= (X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

Note that the variance-covariance matrix is $K \times K$.

Property 3: The OLS estimator is the **Best Linear Unbiased Estimator** of β (**BLUE**) conditionally on X , i.e., OLS is efficient in the class of linear unbiased estimators.

2.2 Least Square Estimator with Endogenous Regressors

When the regressors are endogenous, we have the 3rd assumption departure from Classical Linear Regression Model.

$$y = X\beta + \varepsilon, \text{ but } E(\varepsilon X) \neq 0 \Rightarrow E(\varepsilon|X) \neq 0$$

then the Property of **BLUE** will not happen.

With the endogenous regressors,

$$\begin{aligned}\widehat{\beta}_{OLS} &= (X^T X)^{-1} X^T y \\ &= (X^T X)^{-1} X^T (X\beta + \varepsilon) \\ &= \beta + (X^T X)^{-1} X^T \varepsilon\end{aligned}$$

$$p \lim_{n \rightarrow n} \widehat{\beta}_{OLS} = \beta + \left(p \lim_{n \rightarrow n} \frac{1}{n} X^T X \right)^{-1} \left(p \lim_{n \rightarrow n} \frac{1}{n} X^T \varepsilon \right)^{-1}$$

By **WLLN**,

$$\begin{aligned}p \lim_{n \rightarrow n} \frac{1}{n} X^T X &= p \lim_{n \rightarrow n} \frac{1}{n} \sum_{i=1}^n x_i x_i^T = E(x_i x_i^T) = \sum_{XX} \\ p \lim_{n \rightarrow n} \frac{1}{n} X^T \varepsilon &= p \lim_{n \rightarrow n} \frac{1}{n} \sum_{i=1}^n x_i \varepsilon_i = E(x_i \varepsilon_i) = \sum_{X\varepsilon}\end{aligned}$$

where x_i is the i th column of the matrix X . But $E(x_i \varepsilon_i) \neq 0$, therefore

$$p \lim_{n \rightarrow n} \widehat{\beta}_{OLS} = \beta + \sum_{XX}^{-1} \sum_{X\varepsilon} \neq \beta$$

$\widehat{\beta}_{OLS}$ is not consistent for β .

Method of Moments Interpretation of the Least-Square Problem

The Least-Square solves:

$$0 = \frac{1}{n} X^T (y - X\widehat{\beta}_{OLS}) = \frac{1}{n} \sum_{i=1}^n x_i (y_i - x_i^T \widehat{\beta}_{OLS})$$

But in the population

$$E\left(x_i (y_i - x_i^T \widehat{\beta}_{OLS})\right) = E(x_i \varepsilon_i) \neq 0$$

so it is as if we were using the wrong moments conditions. That is, the analogy principle does not hold. The sample moments do not have population counterparts.

3 Examples of Endogenous Variables

1. Measurement Errors;

True Model:

$$y_i = \alpha + \beta z_i + u_i$$

but z_i is not observed. Instead, we observe x_i , where

$$x_i = z_i + v_i$$

Assume that $(u_i, v_i) \sim iid(0, \Sigma)$, i.e.,

$$\begin{aligned} Var(u_i) &= \sigma_{11} \\ Var(v_i) &= \sigma_{22} \\ Cov(u_i, v_i) &= \sigma_{12} \end{aligned}$$

Also, $E(u_i) = E(v_i) = 0$ and (u_i, v_i) is independent of z .

Substituting x_i for z_i in the first stage gives

$$\begin{aligned} y_i &= \alpha_i + \beta(x_i - v_i) + u_i \\ &= \alpha_i + \beta x_i + (u_i - \beta v_i) \\ &= \alpha_i + \beta x_i + \varepsilon_i \end{aligned}$$

where $\varepsilon_i = u_i - \beta v_i$.

It follows that

$$\begin{aligned} E(x_i \varepsilon_i) &= E((z_i + v_i)(u_i - \beta v_i)) \\ &= E(z_i u_i - z_i \beta v_i + v_i u_i - \beta v_i^2) \\ &= E(v_i u_i) - \beta E(v_i^2) \\ &= \sigma_{12} - \beta \sigma_{22} \neq 0, \text{ in general.} \end{aligned}$$

2. Omitted Variables;

True Model:

$$y_i = x_i^T \beta + z_i^T \gamma + \varepsilon_i, \quad \varepsilon \sim iid(0, \sigma^2)$$

where ε_i is independent of x_i and z_i .

Suppose that z_i is unobserved, so that we consider the model (short regression)

$$y_i = x_i^T \beta + u_i, \quad u_i = \varepsilon_i + z_i^T \gamma$$

Hence,

$$\begin{aligned} E(x_i u_i) &= E(x_i (\varepsilon_i + z_i^T \gamma)) \\ &= E(x_i \varepsilon_i) + E(x_i z_i^T \gamma) \\ &= 0 + E(x_i z_i^T) \gamma \\ &\neq 0, \text{ in general.} \end{aligned}$$

3. Simultaneous Equations (Two quantities are jointly determined by one another's value. Here, joint determination of Price and Quantity)

Demand equation:

$$q_t = x_t^T \beta + \alpha p_t + \varepsilon_t \quad (\alpha < 0)$$

Inverse Supply equation:

$$p_t = z_t^T \gamma + \delta q_t + \eta_t \quad (\delta > 0)$$

Assume that $E(\varepsilon_t) = E(\eta_t) = 0$ and (ε_t, η_t) are independent of x_t and z_t .

For the demand equation:

$$\begin{aligned}
 E(p_t \varepsilon_t) &= E((z_t^T \gamma + \delta q_t + \eta_t) \varepsilon_t) \\
 &= \delta E(q_t \varepsilon_t) + E(\eta_t \varepsilon_t) \\
 &= \delta E((x_t^T \beta + \alpha p_t + \varepsilon_t) \varepsilon_t) + E(\eta_t \varepsilon_t) \\
 &= \delta \alpha E(p_t \varepsilon_t) + \delta \sigma_\varepsilon^2 + \sigma_{12} \\
 &\Rightarrow E(p_t \varepsilon_t) = \frac{\delta \sigma_\varepsilon^2 + \sigma_{12}}{1 - \delta \alpha} \neq 0
 \end{aligned}$$

4 Estimation of Regression with Endogenous Regressors

We first introduce IV estimation, then weighted IV estimation and represent weighted IV estimator from different respects of kinds of regressions like 2SLS, GLS or 3SLS, e.t.c.

The number of instruments needs to be at least as large as the number of endogenous variables.

- (1) IV for Just-Identified
 - a. IV Estimator for Just Identified Case
 - b. Method of Moments
 - c. Asy. Properties of IV
- (2). IV for Over-Identified
 - a. Instrumental Variables
 - b. Weighted IV Estimator
 - c. Asy. Properties of Weighed IV Estimator
 - d. Choice of the Optimal Weighting Matrix
- (3) GMM Representation of Weighed IV
- (4) GLS Representation of Weighted IV
- (5) 2SLS Representation of Weighted IV
- (6) 3SLS (Feasible 2SLS)

(1) Just Identified Case

What is the just means the number of instruments equals to the number of endogenous variables. Given the regression equation $y = X\beta + \varepsilon$, Let z_i is for one individual and we have n individuals. Suppose that there exists some $Z = (z_1, z_2, \dots, z_n)^T$ such that:

- $p \lim_n \frac{1}{n} Z^T X = p \lim_n \frac{1}{n} \sum_{i=1}^n z_i x_i^T = E(z_i x_i^T) \triangleq \sum_{ZX}$

where \sum_{ZX} is a $K \times K$ matrix with $\det(\sum_{ZX}) \neq 0$

- $p \lim_n \frac{1}{n} Z^T \varepsilon = p \lim_n \frac{1}{n} \sum_{i=1}^n z_i \varepsilon_i = E(z_i \varepsilon_i) \triangleq \sum_{Z\varepsilon} = 0$

where $\sum_{Z\varepsilon}$ is a $K \times 1$ vector and 0 is a $K \times 1$ vector of zero ($E(z_i \varepsilon_i) = 0, \forall i$).

Then the variables z_i are called instrumental variables.

a. IV Estimator

We can easily get the IV estimator $\widehat{\beta}_{IV}$ of β :

$$\begin{aligned}\widehat{\beta}_{IV} &= (Z^T X)^{-1} Z^T y \\ &= (Z^T X)^{-1} Z^T (X\beta + \varepsilon) \\ &= \beta + \left(\frac{1}{n} Z^T X\right)^{-1} \left(\frac{1}{n} Z^T \varepsilon\right)\end{aligned}$$

So

$$p \lim_n \widehat{\beta}_{IV} = \beta + \left(p \lim_n \frac{1}{n} Z^T X\right)^{-1} \left(p \lim_n \frac{1}{n} Z^T \varepsilon\right)$$

By **WLLN**,

$$\begin{aligned}p \lim_n \frac{1}{n} Z^T X &= p \lim_n \frac{1}{n} \sum_i z_i x_i^T = E(z_i x_i^T) = \sum_{ZX} \\ p \lim_n \frac{1}{n} Z^T \varepsilon &= p \lim_n \frac{1}{n} \sum_i z_i \varepsilon_i = E(z_i \varepsilon_i) = \sum_{Z\varepsilon} = 0\end{aligned}$$

Therefore,

$$p \lim_n \widehat{\beta}_{IV} = \beta$$

So IV estimator is a consistent estimator of β .

b. Method of Moments Interpretation of IV estimator

From the proof of the consistent of IV estimator, the key property is

$$E(z_i \varepsilon_i) = \sum_{Z\varepsilon} = 0.$$

The property of IV solves the problem exactly and $\widehat{\beta}_{IV}$ makes $\frac{1}{n} \sum_i z_i \widehat{\varepsilon}_i = \frac{1}{n} \sum_i z_i (y_i - x_i^T \widehat{\beta}_{IV}) = 0$ hold, where $\widehat{\varepsilon}_i = y_i - x_i^T \widehat{\beta}_{IV}$.

c. Asymptotic Properties of IV Estimators

Suppose that $E(z_i \varepsilon_i) = 0$, then by **CLT**

$$\frac{1}{\sqrt{n}} Z^T \varepsilon = \frac{1}{\sqrt{n}} \sum_i z_i \varepsilon_i \xrightarrow{d} N(0, \Phi)$$

where $\Phi = E(\varepsilon_i^2 z_i z_i^T)$

By Slutsky Theorem, the product $\left(\frac{1}{\sqrt{n}} \sum_i z_i x_i^T\right)^{-1} \frac{1}{\sqrt{n}} \sum_i z_i \varepsilon_i$ converges to the product of two limits, i.e.,

$$\begin{aligned}\sqrt{n} (\widehat{\beta}_{IV} - \beta) &= \left(\frac{1}{n} Z^T X\right)^{-1} \left(\frac{1}{\sqrt{n}} Z^T \varepsilon\right) \\ &= \left(\frac{1}{n} \sum_i z_i x_i^T\right)^{-1} \frac{1}{\sqrt{n}} \sum_i z_i \varepsilon_i \\ &\xrightarrow{d} N\left(0, \sum_{ZX}^{-1} \Phi \sum_{ZX}^{-1}\right)\end{aligned}$$

- Consistent estimator for \sum_{ZX}

$$\widehat{\sum}_{ZX} = \frac{1}{n} \sum_i z_i x_i^T$$

- Consistent estimator for $\Phi = E(\varepsilon_i^2 z_i z_i^T)$.

$$\widehat{\Phi} = \frac{1}{n} \sum_i \widehat{\varepsilon}_i^2 z_i z_i^T \xrightarrow{p} \Phi$$

This is known as Heteroskedasticity-Robust (Robust, White or Eicker-White) Estimator.

Under conditional homoskedasticity, i.e., $E(\varepsilon_i^2 | z_i) = \sigma^2$, then $\widehat{\Phi}$ simplified to

$$\widehat{\Phi} = \widehat{\sigma}^2 \left(\frac{1}{n} Z^T Z \right)$$

where $\widehat{\sigma}^2$ is any consistent estimator of σ^2 , e.g., the unbiased estimator $\frac{1}{n-K} \sum_i \widehat{\varepsilon}_i^2$ or the biased one $\frac{1}{n} \sum_i \widehat{\varepsilon}_i^2$.

(2) Over-Identified Case: Weighted IV

The over-identified points that the number of instruments is more than the number of endogenous variables.

a. Instrumental Variables

Let z_i be an $l \times 1$ vector with

$$E(z_i \varepsilon_i) = 0, \quad l > K$$

and

$$E(z_i x_i^T) = \sum_{ZX}$$

where \sum_{ZX} is a matrix with $\text{rank}(\sum_{ZX}) = K < l$, \sum_{ZX} is not invertible. Because of the over-identified, the rank will be considered.

Note that by the **CLT**,

$$\frac{1}{\sqrt{n}} \sum_i z_i \varepsilon_i \xrightarrow{d} N(0, V_0)$$

$$V_0 = E(\varepsilon_i^2 z_i z_i^T)$$

Rank Review:

- The column rank of a matrix A is the maximal number of linearly independent column of A . Likewise, the row rank is the maximal number of linearly independent rows of A .

Since the column rank and the row rank are always equal, they are simply called the rank of A ;

- The rank of an $m \times n$ matrix is at most $\min(m, n)$;
- A matrix that has as large rank as possible is said to have **full rank**.

b. Weighted IV Estimator

1.

$$\widehat{\beta}_{IV}(\widehat{\pi}) = \left(\widehat{\pi}^T Z^T X \right)^{-1} \widehat{\pi}^T Z^T y$$

where $\widehat{\pi}$ is an $l \times K$ matrix with $\widehat{\pi} \xrightarrow{P} \pi_0$

2.

$$\begin{aligned} \widehat{\beta}_{OLS} &= (X^T X)^{-1} X^T y \\ \widehat{\beta}_{IV} &= \left(\widehat{\pi}^T Z^T X \right)^{-1} \widehat{\pi}^T Z^T y \end{aligned}$$

Comparing $\widehat{\beta}_{OLS}$ with $\widehat{\beta}_{IV}(\widehat{\pi})$, it is like instead of using X^T in the , and we are now using $\widehat{\pi}^T Z^T$ for the $\widehat{\beta}_{IV}$, where $\widehat{\pi}$ is a weighting matrix with rank K .

The weighting matrix can select a subset of K instrumental variables from l instrumental variables, or might form K linear combinations of these l instruments. The only requirement for the weighting matrix is it has rank K .

3. The idea of the weighting matrix is to approximate $\widehat{\pi}^T Z^T \varepsilon$ by zero. Note that we have K unknown but l moment conditions. Therefore, not all l moments can be exactly satisfied.
4. Not all l moments can be satisfied. So a criterion function that weight them appropriately is used to improve the efficiency of the estimator.

C. Asymptotic Distribution of Weighted IV Estimator

Naturally, the asymptotic distribution of $\widehat{\beta}_{IV}(\widehat{\pi})$ depends on π_0 .

$$\sqrt{n} \left(\widehat{\beta}_{IV} - \beta \right) \xrightarrow{d} N(0, \Lambda)$$

where

$$\Lambda = \left(\pi_0^T \sum_{ZX} \right)^{-1} \pi_0^T V_0 \pi_0 \left(\pi_0^T \sum_{ZX} \right)^{-1}$$

d. Choice of an Optimal Weighting Matrix

What is the optimal weighting matrix π_0 ?

$$\begin{aligned} \pi_0 &\equiv \arg \min \Lambda \\ &= \arg \min \left[\left(\pi_0^T \sum_{ZX} \right)^{-1} \pi_0^T V_0 \pi_0 \left(\pi_0^T \sum_{ZX} \right)^{-1} \right] \end{aligned}$$

That is, the optimal π_0 minimize the asymptotic variance of $\widehat{\beta}_{IV}$
The optimal π_0^* is

$$\begin{aligned} \pi_0^* &= \widehat{V}_0^{-1} \left(\frac{1}{n} Z^T X \right) \\ \widehat{V}_0^{-1} &= \frac{1}{n} \sum_i \widehat{\varepsilon}_i^2 z_i z_i^T \end{aligned}$$

This gives the IV estimator that has the smallest asymptotic variance among those that could be formed from the instruments Z and a weighting matrix π .

(3) The GMM Representation of Weighted IV

The Generalized Method of moment estimator is the best estimator which solves the sample analogue of the population moments

$$0 = E(\pi^T z_i (y_i - x_i^T \beta))$$

(4) GLS Representation of Weighted IV

$$\begin{aligned} y &= X\beta + \varepsilon \\ \Rightarrow Z^T y &= Z^T X\beta + Z^T \varepsilon \\ \Rightarrow \frac{1}{\sqrt{n}} Z^T y &= \frac{1}{\sqrt{n}} Z^T X\beta + \frac{1}{\sqrt{n}} Z^T \varepsilon \end{aligned}$$

or the transformed model

$$\tilde{y} = \tilde{X}\beta + \tilde{\varepsilon}$$

$$\begin{aligned} \text{where } \tilde{y} &= \frac{1}{\sqrt{n}} Z^T y \\ \tilde{X} &= \frac{1}{\sqrt{n}} Z^T X \\ \tilde{\varepsilon} &= \frac{1}{\sqrt{n}} Z^T \varepsilon \end{aligned}$$

Note that $\tilde{\varepsilon} \xrightarrow{d} N(0, V_0)$ by **CLT**. Then,

$$\begin{aligned} \hat{\beta}_{GLS} &= (\tilde{X}^T \hat{V}_0^{-1} \tilde{X})^{-1} \tilde{X}^T \hat{V}_0^{-1} \tilde{Y} \\ &= \left(\frac{1}{n} X^T Z \hat{V}_0^{-1} Z^T X \right)^{-1} \left(\frac{1}{n} X^T Z \hat{V}_0^{-1} \right) Z^T y \end{aligned}$$

Comparing $\hat{\beta}_{GLS}$ with $\hat{\beta}_{IV}$ (weighted),

$$\begin{aligned} \hat{\beta}_{GLS} &= \left(\frac{1}{n} X^T Z \hat{V}_0^{-1} Z^T X \right)^{-1} \left(\frac{1}{n} X^T Z \hat{V}_0^{-1} \right) Z^T y \\ \hat{\beta}_{IV} &= (\hat{\pi}^T Z^T X)^{-1} \hat{\pi}^T Z^T y \end{aligned}$$

The difference is instead of $\frac{1}{n} X^T Z \hat{V}_0^{-1}$ using $\hat{\pi}^T$, or $\hat{V}_0^{-1} (\frac{1}{n} Z^T X)$ using $\hat{\pi}$. This also suggests that the optimal weighting matrix $\hat{\pi}$ should be $\hat{\pi}^* = \hat{V}_0^{-1} (\frac{1}{n} Z^T X)$, where

$$\hat{\pi}^* \xrightarrow{p} \pi^* = V_0^{-1} \sum_{ZX}$$

and $\hat{\beta}_{GLS} = \hat{\beta}_{IV}$ under the optimal weighting matrix $\hat{\pi}^*$.

(5) 2SLS Interpretation of the (Optimal) Weighted IV Estimator

1. Calculate $\hat{\beta}_{2SLS}$

$$\begin{aligned} y &= X\beta + \varepsilon \\ X &= Z\pi + V \end{aligned}$$

First Stage:

Regress X on Z using OLS, and we can get

$$\begin{aligned} \hat{\pi} &= (Z^T Z)^{-1} Z^T X \\ \hat{X} &= Z\hat{\pi} \\ &= Z(Z^T Z)^{-1} Z^T X \\ &= H_Z X \end{aligned}$$

where $H_Z = Z(Z^T Z)^{-1} Z^T$ and $E(\varepsilon|Z) = 0$

Second Stage

Regress y on \hat{X} , i.e.,

$$\begin{aligned} y_i &= x_i\beta + \varepsilon_i \\ \varepsilon_i^* &= \varepsilon_i + (x_i - \hat{x}_i)^T \beta \end{aligned}$$

and by construction $E(\varepsilon_i^*|\hat{x}_i) = 0$

Then we have

$$\begin{aligned} \hat{\beta}_{2SLS} &= (\hat{X}^T \hat{X})^{-1} \hat{X}^T y \\ &= (X^T Z (Z^T Z)^{-1} Z^T X)^{-1} X^T Z (Z^T Z)^{-1} Z^T y \\ &= (X^T H_Z X)^{-1} X^T H_Z y \end{aligned}$$

since H_Z is idempotent matrix.

2. Interpretation:

(a) we can write the 2SLS estimator as

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

That is, \hat{X} is used as an instrument for X .

(b) Alternatively, we can write as before

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

i.e., LS of \hat{X} on y .

(c) we can write

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

where $\hat{y} = H_Z y$, that is, residual regression of \hat{y} on \hat{X} .

3. Asymptotic Distribution of 2SLS Estimator.

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

where:

(a) $\Lambda_{2SLS} = \left(\pi_0^T \sum_{ZX} \right)^{-1} \pi_0^T V_0 \pi_0 \left(\sum_{ZX}^T \pi_0 \right)^{-1}$

(b) V_0 is the asymptotic variance of $\frac{1}{\sqrt{n}} Z^T \varepsilon$

$$\frac{1}{\sqrt{n}} Z^T \varepsilon \xrightarrow{d} N(0, V_0)$$

(c) $\pi_0 = p \lim_n (Z^T Z)^{-1} Z^T X = \left(p \lim_n \frac{1}{n} Z^T Z \right)^{-1} \left(p \lim_n \frac{1}{n} Z^T X \right) = \sum_{ZZ}^{-1} \sum_{ZX}$

So the asy. covariance matrix of a 2SLS estimator is

$$\Lambda_{2SLS} = \left(\sum_{ZZ}^T \sum_{ZZ}^{-1} \sum_{ZZ} \right)^{-1} \sum_{ZZ}^T \sum_{ZZ}^{-1} V_0 \sum_{ZZ} \sum_{ZX} \left(\sum_{ZZ}^T \sum_{ZZ}^{-1} \sum_{ZZ} \right)^{-1}$$

Under conditional homokedasticity disturbance, $V_0 = \sigma^2 \sum_{ZZ} = \sigma^2 Z^T Z$, Thus

$$\Lambda_{2SLS} = \sigma^2 \left(\sum_{ZX}^T \sum_{ZZ}^{-1} \sum_{ZX} \right)^{-1}$$

and

$$\hat{\sigma}^2 \equiv \frac{1}{n} \left(y - x \hat{\beta} \right)^T \left(y - x \hat{\beta} \right) \xrightarrow{p} \sigma^2$$

For the estimation of σ^2 we use X , NOT \hat{X} . We do use \hat{X} only in $\left(\hat{X}^T \hat{X} \right)^{-1}$.

Note that

$$\frac{1}{n} \hat{X}^T \hat{X} \xrightarrow{p} \left(\sum_{ZX}^T \sum_{ZZ}^{-1} \sum_{ZX} \right)$$

Thus,

$$\hat{\beta}_{2SLS} \overset{A}{\sim} N \left(\beta, \hat{\sigma}^2 \left(\hat{X}^T \hat{X} \right)^{-1} \right)$$

4. We know that

$$\begin{aligned} \hat{\beta}_{2SLS} &= \left(X^T Z (Z^T Z)^{-1} Z^T X \right)^{-1} \left(X^T Z (Z^T Z)^{-1} \right) Z^T y \\ \hat{\beta}_{GLS} &= \left(X^T Z \hat{V}_0^{-1} Z^T X \right)^{-1} \left(X^T Z \hat{V}_0^{-1} \right) Z^T y \end{aligned}$$

The 2SLS formula is very similar to GLS, just replace $(Z^T Z)^{-1}$ in the 2SLS formula with V_0^{-1}

5. In the homoskedastic case where $V_0 = \sigma^2 \sum_{ZZ}$ we have asy. variance of $\widehat{\beta}_{2SLS}$ equals to the asy. variance of $\widehat{\beta}_{GLS}$. But, in general,

$$Var\left(\widehat{\beta}_{GLS}\right) \leq Var\left(\widehat{\beta}_{2SLS}\right)$$

if

$$\frac{1}{\sqrt{n}} \sum_i z_i \varepsilon_i \xrightarrow{d} N(0, V_0)$$

$$V_0 = E\left(\varepsilon_i^2 z_i z_i^T\right) = \sigma^2 E\left(z_i z_i^T\right) = \sigma^2 \sum_{ZZ}$$

that is to say, under heterokedasity, GSL estimator is more efficient than 2SLS estimator even though they are both unbiased.

6. the **optimal weighted IV estimator** is

$$\begin{aligned} \widehat{\beta}_{OptimalIV} &= \left(\widehat{\pi}^{*T} Z^T X\right)^{-1} \widehat{\pi}^{*T} Z^T y \\ \pi_0^* &= V_0^{-1} \left(\frac{1}{n} Z^T X\right) \end{aligned}$$

if one makes the conditional homoskedasticity assumption, i.e.,

$$V_0 = \sigma^2 \sum_{ZZ} = \sigma^2 Z^T Z$$

the optimal IV estimator is

$$\begin{aligned} \widehat{\beta}_{OptimalIV} &= \left(\frac{1}{n} X^T Z \left(\sigma^2 \sum_{ZZ}\right)^{-1} Z^T X\right)^{-1} \left(\frac{1}{n} X^T Z \left(\sigma^2 \sum_{ZZ}\right)^{-1}\right) Z^T y \\ &= \left(X^T Z \left(\sum_{ZZ}\right)^{-1} Z^T X\right)^{-1} \left(X^T Z \left(\sum_{ZZ}\right)^{-1}\right) Z^T y \\ &= \left(X^T Z \left(Z^T Z\right)^{-1} Z^T X\right)^{-1} \left(X^T Z \left(Z^T Z\right)^{-1}\right) Z^T y \\ &= \widehat{\beta}_{2SLS} \end{aligned}$$

Under conditional homoskedasticity,

$$\widehat{\beta}_{OptimalIV} = \widehat{\beta}_{2SLS} = \widehat{\beta}_{GLS}$$

As **6.** showing above, the 2SLS estimator will no longer be best when the scalar covariance matrix assumption $E\left(\varepsilon_i \varepsilon_i^T\right) = \sigma^2 I$ fails. In general, $Var\left(\widehat{\beta}_{GLS}\right) \leq Var\left(\widehat{\beta}_{2SLS}\right)$. But under fairly general conditions it will remain consistent.

(6) 3SLS (Feasible 2SLS)

In practice, to obtain GLS estimator, one needs to have a consistent estimator \widehat{V} of V_0 . This can be done in 2 steps.

STEP I

Estimate $\hat{\beta}_{2SLS}$ and retrieve the residuals

$$\hat{\varepsilon}_i = y_i - X_i \hat{\beta}_{2SLS},$$

then use $\hat{\varepsilon}_i$ from 2SLS to form

$$\hat{V} = \frac{1}{n} \sum_i \hat{\varepsilon}_i^2 z_i z_i^T$$

STEP II

Find a Cholesky transformation P satisfying $P\hat{V}P^T = I$, then make the transformations

$$\tilde{y} = Py, \quad \tilde{X} = PX, \quad \tilde{Z} = (P^T)^{-1} Z$$

and do a 2SLS regression of \tilde{y} on \tilde{X} using \tilde{Z} as instruments.

$$\hat{\beta}_{3SLS} = \left(X^T Z \hat{V}^{-1} Z^T X \right)^{-1} X^T Z \hat{V}^{-1} Z^T y$$

$\hat{\beta}_{3SLS}$ is also called $\hat{\beta}$ feasible 2SLS.