Econometrics I

TA Session 14

Jukina HATAKEYAMA *

July 16, 2024

Contents

*E-mail: u868710a@ecs.osaka-u.ac.jp

1 Deriving the 2SLS Estimator

Consider the regression model such that:

$$y_i = x_i b + u_i \ (i = 1, 2, \cdots, n), \tag{1}$$

where $x_i = (1, x_{i2}, \cdots, x_{i_K})$ and $b \in \mathbb{R}^{1\times K}$. Assume that the instrumental variables in (1) are $z_{i1}, \cdots, z_{iM}$. In addition, the endogeneous variable is x_{iK}. Then, we have $1 \times (K + M)$ vector of the exogeneous variables $z_i = (1, x_{i1}, \cdots, x_{i(K-1)}, z_{i1}, \cdots, z_{iM})$. We can derive the two stage least square estimator by the following procedure.

1. We consider the stacked regression model $y = xb+u$. Here, we difine $z = (z_1', z_2', \cdots, z_n')' \in \mathbb{R}^{n\times(K+M)}$, $x = (x_1', x_2' \cdots, x_n')' \in \mathbb{R}^{n\times K}$, $y = (y_1, y_2, \cdots, y_n)'$ and $u = (u_1, u_2, \cdots, u_n)'$.

2. Regress z on x and derive the fitted values of x:

$$\hat{x} = z(z'z)^{-1}z'x = P_z x. \tag{2}$$

3. Regress $\hat{x}$ on y. Then, we can derive 2SLS estimator:

$$\hat{b}_{2SLS} = (\hat{x}'\hat{x})^{-1}\hat{x}'y. \tag{3}$$

This is why we call this estimation method 2SLS(: Two Stage Least Squer).
Remind that the 2SLS estimator equals to the IV estimator under some conditions. This will be explained later.

2 Properties of the 2SLS Estimator

2.1 Consistency

Now we review the consistency of $\hat{b}_{2SLS}$.

Theorem 2.1 (The Consistency of 2SLS estimator)**.** Suppose that we have the following assumptions:

(ASSUMPTION1) For some $1 \times L$ vector z_i, $\mathbb{E}[z_i'u_i] = 0$.

(ASSUMPTION2) $\text{rank}\,(\mathbb{E}[z_i'z_i]) = L,$ and $\text{rank}\,(\mathbb{E}[z_i'x_i]) = K$.

Under these assumptions, we have: $\hat{b}_{2SLS} \xrightarrow[n\to\infty]{p} b$.

Proof. By rewriting (3), we have

$$\begin{aligned}
\hat{b}_{2SLS} &= (\hat{x}'\hat{x})^{-1}\hat{x}'y \\
&= (x'P_z x)^{-1}x'P_z y \\
&= (x'P_z x)^{-1}x'P_z((xb+u)) \\
&= (x'P_z x)^{-1}x'P_z xb + (x'P_z x)^{-1}x'P_z u \\
&= b + (x'z(z'z)^{-1}z'x)^{-1}x'z(z'z)^{-1}z'u.
\end{aligned} \tag{4}$$

In the previous equation, remember that $P_z = z(z'z)^{-1}z'$.

$$\hat{b}_{2SLS} = b + \left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i'z_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} z_i'z_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i'x_i\right)\right]^{-1} \times \left(\frac{1}{n}\sum_{i=1}^{n} x_i'z_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} z_i'z_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} z_i'u_i\right) \quad (5)$$

Thus, we can apply WLLN and prove that $\hat{b}_{2SLS} \xrightarrow[n\to\infty]{p} b$ by (ASSUMPTION1). □

2.2 Asymptotic Normality

Using an additional assumption, we can derive the asymptotic normality of the 2SLS estimator.

Theorem 2.2 (Asymptotic Normality of 2SLS estimator)**.** Suppose that we have the following assumptions:

(ASSUMPTION1) For some $1 \times L$ vector z_i, $\mathbb{E}(z_i'u_i) = 0$.

(ASSUMPTION2) $\text{rank}\,(\mathbb{E}[z_i'z_i]) = L$, and $\text{rank}\,(\mathbb{E}[z_i'x_i]) = K$.

(ASSUMPTION3) $\mathbb{E}[u_i^2 z_i'z_i] = \sigma^2\mathbb{E}[z_i'z_i]$ with $\mathbb{E}[u_i^2] = \sigma^2$. (Note that $\mathbb{E}[u_i^2|z_i] = \sigma^2$ implies this assumption.)

Under (ASSUNPTION1–3), the limiting distribution of the 2SLS estimator is

$$\sqrt{n}(\hat{b}_{2SLS} - b) \xrightarrow[n\to\infty]{d} N(0, \sigma^2\{\mathbb{E}[x_i'z_i]\mathbb{E}[z_i'z_i]^{-1}\mathbb{E}[z_i'x_i]\}^{-1}). \quad (6)$$

Proof. By using (5), we can say

$$\sqrt{n}(\hat{b}_{2SLS} - b) = \left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i'z_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} z_i'z_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} z_i'x_i\right)\right]^{-1} \times \left(\frac{1}{n}\sum_{i=1}^{n} x_i'z_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} z_i'z_i\right)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i'u_i\right). \quad (7)$$

The WLLN and CLT imply that

$$\begin{aligned}
&\frac{1}{n}z'x = \frac{1}{n}\sum_{i=1}^{n} z_i'x_i \xrightarrow[n\to\infty]{p} \mathbb{E}[z_i'x_i] \\
&\frac{1}{n}x'z = \frac{1}{n}\sum_{i=1}^{n} x_i'z_i \xrightarrow[n\to\infty]{p} \mathbb{E}[x_i'z_i] \\
&\frac{1}{n}z'z = \frac{1}{n}\sum_{i=1}^{n} z_i'z_i \xrightarrow[n\to\infty]{p} \mathbb{E}[z_i'z_i] \\
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i'u_i = \sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} z_i'u_i - 0) \xrightarrow[n\to\infty]{d} N(0, \text{Var}[z'u]). \quad (8)
\end{aligned}$$

Remember that $\mathrm{Var}[z'u] = \sigma^2\mathbb{E}[z_i'z_i]$ because of the formula of the conditinal variance. By the continuous mapping theorem, we can can say the first term of the (RHS) in (8) converges to $[\mathbb{E}[x_i'z_i]\mathbb{E}[z_i'z_i]^{-1}\mathbb{E}[z_i'x_i]]^{-1}$ in probability. Therefore, we can apply the Slutsky's Theorem to the (RHS) in Eq.(8) and derive (6).

□

2.3 Asymptotic Efficiency

Now we explain the case that we can regard the method of 2SLS as the same one of the IV. This section is an advanced result and is devoted for students' information.

Theorem 2.3. Under ASSUMPTION1–3, the 2SLS estimator is efficient in the class of all instrumental variable estimators using instrument linear in z_i.

Proof. Let $\hat{b}_{2SLS}$ be the 2SLS estimator, and let $\tilde{b}$ be any other IV estimator using instruments linear in z_i, $1 \times L$ random vector. Let the instruments for $\tilde{b}$ be $\tilde{m}_i \equiv z_i\Gamma$, where Γ is an $L \times K$ nonstochastic matrix. We assume that rank condition holds for $\tilde{m}_i$. For 2SLS, the choice of IVs is effectively $m_i^* = z_i\Pi$, where $\Pi = [\mathbb{E}(z_i'z_i)]^{-1}\mathbb{E}(z_i'x_i) \equiv D^{-1}C$. (In both cases, we can replace Γ and Π with $\sqrt{n}$-consistent estimators without changing the asymptotic variance.) Now, under (ASSUMPTION1–3), we know the asymptotic variance of $\sqrt{n}(\hat{b}_{2SLS} - b)$ is $\sigma^2[\mathbb{E}(m_i^{*\prime}m_i^*)]^{-1}$, which is equal to Eq.(6). Additionally, recall that $\tilde{b} = (\sum_{i=1}^n \tilde{m}_i'x_i)^{-1}(\sum_{i=1} \tilde{m}_i'y_i)$ and we have

$$\sqrt{n}(\tilde{b} - b) = \left(\frac{1}{n}\tilde{m}'x\right)^{-1}\left(\frac{1}{\sqrt{n}}\tilde{m}'u\right),$$

when we define the stacked model. By the same procedure to derive (6), $\mathrm{var}[\sqrt{n}(\tilde{b} - b)] \xrightarrow[n\to\infty]{p}$ $\sigma^2[\mathbb{E}(\tilde{m}_i'x_i)]^{-1}[\mathbb{E}(\tilde{m}_i'\tilde{m}_i)][\mathbb{E}(x_i'\tilde{m}_i)]^{-1}$ is proven.

To show that $\mathrm{Avar}[\sqrt{n}(\tilde{b} - b)] - \mathrm{Avar}[\sqrt{n}(\hat{b}_{2SLS} - b)]$ is p.s.d., it suffices to show that $\mathbb{E}(m_i^{*\prime}m_i^*) - [\mathbb{E}(x_i'\tilde{m}_i)][\mathbb{E}(\tilde{m}_i'\tilde{m}_i)]^{-1}[\mathbb{E}(\tilde{m}_i'x_i)]$ is p.s.d. But $x_i = m_i^* + r_i$, where $\mathbb{E}(z_i'r_i) = 0$, and so $\mathbb{E}(\tilde{m}_i'r_i) = 0$. Here, x_i represents the data vector and r_i is exogeneous variable vector. It follows that $\mathbb{E}(\tilde{m}_i'x_i) = \mathbb{E}(\tilde{m}_i'm_i^*)$, and so

$$\begin{aligned}
&\mathbb{E}(m_i^{*\prime}m_i^*) - [\mathbb{E}(x_i'\tilde{m}_i)][\mathbb{E}(\tilde{m}_i'\tilde{m}_i)]^{-1}[\mathbb{E}(\tilde{m}_i'x_i)] \\
&= \mathbb{E}(m_i^{*\prime}m_i^*) - [\mathbb{E}(m_i^{*\prime}\tilde{m}_i)][\mathbb{E}(\tilde{m}_i'\tilde{m}_i)]^{-1}[\mathbb{E}(\tilde{m}_i'm_i^*)] \\
&= \mathbb{E}(s_i^{*\prime}s_i^*), \qquad (9)
\end{aligned}$$

where $s^* = m_i^* - L(m_i^*|\tilde{m}_i)$ is the population residual from the linear projection of m_i^* on $\tilde{m}_i$. Because $\mathbb{E}(s_i^{*\prime}s_i^*)$ is p.s.d, the proof is completed. □

2.4 Testing for Endogeneity

Consider the null hypothesis which implies all explanatory variables are exogeneous. In this case, we can test on the difference between the 2SLS estimator and the OLS estimator. We can use Durbin=Wu=Hausman (DWH) test statistic such as:

$$\mathrm{DWH} = (\hat{b}_{IV} - \hat{b}_{OLS})'[(\hat{x}'\hat{x})^{-1} - (x'x)^{-1}]^{-}(\hat{b}_{IV} - \hat{b}_{OLS})/\hat{\sigma}_{OLS}^2, \qquad (10)$$

where $[(\hat{x}'\hat{x})^{-1} - (x'x)^{-1}]^-$ is generalized inverse,[1] except in the usual case that all elements of x_i are allowed to be endogeneous under the alternative. Asymptotically, DWH $\sim \chi^2(d)$, where d is the rank of $\mathrm{Var}(\hat{b}_{IV}) - \mathrm{Var}(\hat{b}_{OLS})$.

[1] Remind that the rank of $(\hat{x}'\hat{x})^{-1}$ and that of $(x'x)^{-1}$ is not same.

3 R Excercise

Today, we estimate the wage function by using AER package. Suppose the following regression model such that:

$$\begin{aligned}\log(\text{Wage})_i = \alpha + \beta_1(\text{Educational Year})_i + \beta_2(\text{Test Score})_i \\ + \beta_3(\text{Unemployment Rate})_i + \beta_4(\text{Tuition})_i + u_i.\end{aligned} \tag{11}$$

In this estimation, the year of the education is regarded as the endogeneous variable. Remind that u_i implies the factors which are NOT represented as the explanatory variables in the previous equation (e.g. the study motivation). Therefore, we must apply the IV method to estimate above equation. The selected instrumental variable is the distance to the university.

```
rm(list=ls(all=TRUE))
library(AER)
library(lmtest)
library(sandwich)
library(stargazer)

data("CollegeDistance")
#fix(CollegeDistance)

unep<-CollegeDistance[,8]
wage<-CollegeDistance[,9]
score<-CollegeDistance[,3]
distance<-CollegeDistance[,10]
tuition<-CollegeDistance[,11]
year<-CollegeDistance[,12]

ivdata<-data.frame(wage,year,score,distance,unep,tuition)
fix(ivdata)

#Estimate IV by using exogeneous and instrumental variables.
#IV:ivreg function/ 2SLS:ivreg.fit function
ivest<-ivreg(log(wage) ~ year + score + unep + tuition |
 score + unep + tuition + distance)

#Checking the White Estimator (HCCME).
wivest<-coeftest(ivest, df = Inf, vcov = vcovHC(ivest, type = "HC0"))

#Score does NOT explain the wage.
stargazer(wivest, title="Wage Function",
style="all", type="latex")
```

Table 1: Wage Function

	Dependent variable:
year	0.042^{**}
	(0.018)
	z = 2.345
	p = 0.020
score	-0.003
	(0.002)
	z = -1.469
	p = 0.142
unep	0.011^{***}
	(0.001)
	z = 14.364
	p = 0.000
tuition	0.108^{***}
	(0.006)
	z = 19.208
	p = 0.000
Constant	1.619^{***}
	(0.163)
	z = 9.949
	p = 0.000
Note:	*p<0.1; **p<0.05; ***p<0.01