Intermediate Econometrics

Professor: Xiaojun Song

Timekeeper: Rui Zhou

Spring 2023

目录

10 IV Estimation & 2SLS

Instrumental variables estimation and two stage least squares are equivalent under some conditions. However, two stage least squares are more general compared to instrumental variables estimation.

The endogeneity problem is **endemic** in social sciences and economics.

- In many cases, important personal variables cannot be observed.
- These are often correlated with observed explanatory information.
- In addition, measurement error may also lead to endogeneity.
- Solutions to endogeneity problems considered so far:
  - Proxy variables method for omitted regressors
  - Fixed effects methods if
    - panel data is available,
    - endogeneity is time-constant and
    - regressors are not time-constant.

Instrumental Variables Method (IV) is now the most well-known and favored method to address endogeneity problems.

To begin with, first consider the simple linear regression model,

$$y = \beta_0 + \beta_1 x + u$$

where the "weak" requirement of $Cov(u, x) = 0$ makes sure that slope coefficient is consistent. If the requirement is met, OLS is the best method for estimation (i.e., BLUE); no need to find alternatives.

However, if $Cov(u, x) \neq 0$, endogeneity problem arises. Explanatory variable x is then called endogenous variable. (Note that, you do not need to struggle with the origin of endogeneity. You can simply think of it as omitting some important variables.) Moreover in MLR, $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$, as long as $Cov(x_j, u) \neq 0$ for some j, then x_j is called endogenous variable.

10.1 Instrumental Variable

Definition:

1. It does not appear in the regression.
   - It did not participate in explaining variation of y.
   - In mathematics, if an IV also appears in the structural equation, you will not have enough equation to estimate slope coefficients via method of moments; if you use 2SLS, multicollinearity problem will be quite severe.

2. It is highly correlated with the endogenous variable. (Instrument Relevance)

- This requirement ensures the induced instrument is a good "representative".
  - Otherwise, such IV is a *weak* one.
- This condition can be checked by regression.

3. It is uncorrelated with the error term. (Instrument Exogeneity)
   - However, the argument mostly comes down to a "belief" and *cannot* be checked.
   - Since x is endogenous, $\hat{\beta}_0, \hat{\beta}_1$ are not correctly estimated, so are residuals $\hat{u}_i$; impractical and impossible to check.

10.2 IV Estimation in SLR

10.2.1 Exogenous Case

Start from the simple linear regression,

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Identification of β_1 can be obtained from *zero conditional mean*.

$$\begin{aligned} & Cov(x_i, u_i) = 0 \\ \Longrightarrow & Cov(x_i, y_i - \beta_0 - \beta_1 x_i) = 0 \\ \Longrightarrow & Cov(x_i, y_i) - \beta_1 Var(x_i) = 0 \\ \Longrightarrow & \beta_1 = \frac{Cov(x_i, y_i)}{Var(x_i)} \end{aligned}$$

Replace "population parameter" with its sample-version:

$$\hat{\beta}_1 = \frac{\hat{Cov}(x_i, y_i)}{\hat{Var}(x_i)} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Also, $\hat{\beta}_1$ is consistent, i.e., $\hat{\beta}_1 \xrightarrow{p} \beta_1$, as long as exogeneity holds.

10.2.2 Endogenous Case

However, if $Cov(x_i, u_i) \neq 0$, x_i is endogenous. Luckily, z_i is found with $Cov(z_i, u_i) = 0$.

$$\begin{aligned} & Cov(z_i, u_i) = 0 \\ \Longrightarrow & Cov(z_i, y_i - \beta_0 - \beta_1 x_i) = 0 \\ \Longrightarrow & Cov(z_i, y_i) - \beta_1 Var(z_i) = 0 \\ \Longrightarrow & \beta_1 = \frac{Cov(z_i, y_i)}{Var(z_i)} \end{aligned}$$

The instrument z_i has to satisfy two important properties:

- Instrument Exogeneity: $Cov(z_i, u_i) = 0$
- Instrument Relevance: $Cov(z_i, x_i) \neq 0$
  - However, when $Cov(z_i, x_i) \to 0$, z_i is a weak instrument, and this will have a pratical influence on slope coefficients.

With sample-version statistics replacing population ones:

$$\hat{\beta}_{IV} = \frac{\widehat{Cov}(z_i, y_i)}{\widehat{Var}(z_i, x_i)} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$$

Note that even with endogeneity, $\bar{y} = \beta_0 + \beta_1 \bar{x}$ holds (as long as $E(u) = 0$). Intercept coefficient $\hat{\beta}_0$ can be obtained in this way.

If x_i is an exogenous variable, totally fine to set $z_i \equiv x_i$, and goes back to OLS estimation. Here we can get the sense that, any exogenous variable is the instrument for itself.

Note that **instrument must be biased, though consistent.** (Not covered here)

$$E(\hat{\beta}_{IV}) \neq \beta_1, \quad \hat{\beta}_{IV} \xrightarrow{p} \beta_1$$

10.2.3 Properties of IV with a POOR Instrumental Variable

IV may be much more inconsistent than OLS if the instrumental variable is not completely exogenous and only weakly related to x.

$$\begin{aligned}
\hat{\beta}_{1,IV} &= \frac{\widehat{Cov}(z_i, y_i)}{\widehat{Var}(z_i, x_i)} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} \\
&= \frac{\sum_{i=1}^{n}(z_i - \bar{z})(\beta_0 + \beta_1 x_i + u_i)}{\sum(z_i - \bar{z})(x_i - \bar{x})} \\
&= \beta_1 + \frac{\sum_{i=1}^{n}(z_i - \bar{z})u_i}{\sum(z_i - \bar{z})(x_i - \bar{x})} \\
\Longrightarrow &\begin{cases} \text{plim } \hat{\beta}_{1,OLS} & = \beta_1 + Corr(x, u) \cdot \frac{\sigma_u}{\sigma_x} \\ \text{plim } \hat{\beta}_{1,IV} & = \beta_1 + \frac{Corr(z,u)}{Corr(z,x)} \cdot \frac{\sigma_u}{\sigma_x} \end{cases}
\end{aligned}$$

IV is worse than OLS iff

$$\frac{Corr(z, u)}{Corr(z, x)} > Corr(x, u)$$

10.3 IV Estimation in MLR

Consider the *structural equation*

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + ... + \beta_k z_{k-1} + u_1$$

where y_2 is an endogenous variable, and all $z_j, \forall j = 1, 2, ..., k-1$ are exogenous variables. And note that such equation is called a structural equation because we are interested in β_j, but it does not mean the equation is a representation of causality.

We are to find an instrumental variable z_k satisfies that, z_j

1. Does not appear in regression equation above;
2. Is *UNcorrelated* with error term;
3. Is ***partially*** correlated with endogenous explanatory variable.

For the third requirement of partial correlation, consider the following equation

$$y_2 = \pi_0 + \pi_1 z_1 + ... + \pi_k z_{k-1} + \pi_k z_k + v_2$$

This is the so-called "*reduced form regression*". $z_1, ..., z_{k-1}$ are induced in case of any bias from omitting variables. The significance of π_k measures strength of correlation of y_2 and z_k. We are in the hope that null hypothesis $\pi_k = 0$ is rejected.

Traditionally, IV estimation can be realized by either method of moments, or Two Stage Least Squares.

10.3.1 Method of Moments

Consider the simple regression

$$y_{1i} = \beta_0 + \beta_1 y_{2i} + \beta_2 z_{1i} + u_i$$

where y_2 is endogenous, z_1 is exogenous, and we have found an IV z_2 for y_2. That is, with $E(u_i) = 0, Cov(z_{1i}, u_i) = 0$,and $Cov(z_{2i}, u_i) = 0$.

$$\begin{cases} E(u) = 0 \\ \quad \Longrightarrow \frac{1}{n}\sum_{i=1}^{n}(y_{1i} - \beta_0 - \beta_1 y_{2i} - \beta_2 z_{1i}) = 0 \\ Cov(z_{1i}, u_i) = 0 \Longrightarrow E(z_{1i}, u_i) = 0 \\ \quad \Longrightarrow \frac{1}{n}\sum_{i=1}^{n} z_{1i}(y_{1i} - \beta_0 - \beta_1 y_{2i} - \beta_2 z_{1i}) = 0 \\ Cov(z_{2i}, u_i) = 0 \Longrightarrow E(z_{2i}, u_i) = 0 \\ \quad \Longrightarrow \frac{1}{n}\sum_{i=1}^{n} z_{2i}(y_{1i} - \beta_0 - \beta_1 y_{2i} - \beta_2 z_{1i}) = 0 \end{cases}$$

We can find the solutions $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ for the three equations.

$$\begin{cases} \sum_{i=1}^{n}(y_{1i} - \hat{\beta}_0 - \hat{\beta}_1 y_{2i} - \hat{\beta}_2 z_{1i}) = 0 \\ \sum_{i=1}^{n} z_{1i}(y_{1i} - \hat{\beta}_0 - \hat{\beta}_1 y_{2i} - \hat{\beta}_2 z_{1i}) = 0 \\ \sum_{i=1}^{n} z_{2i}(y_{1i} - \hat{\beta}_0 - \hat{\beta}_1 y_{2i} - \hat{\beta}_2 z_{1i}) = 0 \end{cases}$$

If with more endogenous variables, more IV are needed. For instance, if you have two endogenous variables, you have to find at least two IVs. However, if you are lucky enough to find more than 2 IVs, say 3, method of moments would fail, since number of solutions binds one-to-one with number of equations. Therefore, if the number of IVs equates that of endogenous variables, this is a just-identified case. If more, over-identified; if less, under-identified. If with more IV than number of endogenous variables, it is really a pity that some potentially valuable cannot be utilized in method of moments.

10.4 Two Stage Least Squares Estimation

10.4.1 Approach

Consider a general case of multiple linear regression,

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + ... + \beta_k x_{k-1} + u$$

where y_2 is endogenous, and all other explanatory variables are exogenous. We have found an IV z_k for y_2.

- First Stage (Reduced Form Regression)
  - The endogenous explanatory variable y_2 is predicted using only exogenous information:

    $$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + ... + \hat{\pi}_{k-1} z_{k-1} + \hat{\pi}_k z_k$$

    In essence, you partial out endogenous part in y_2, and only its exdogenous part is kept. (Also, you should prove it to yourself that $\hat{\pi}_k$ is not zero. If there are more than one IV for y_2, prove it that the joint hypothesis that all slope coefficients for IVs are zero can be rejected!)
- Second Stage
  - OLS with y_2 replaced by its prediction $\hat{y}_2$ from the first stage.

$$y_1 = \beta_0 + \beta_1 \hat{y}_2 + \beta_2 z_1 + ... + \beta_k z_{k-1} + error$$

10.4.2 Principle of TSLS

10.4.2.1 From intuition

- All variables in the *second* stage regression are ***exogenous***, because y_2 was replaced by a prediction based on only exogenous information, and has been purged of its endogenous part (which is related to the error term).

- If there is one endogenous variable and one instrument, then 2SLS is equivalent to IV. Moreover, 2SLS estimation can also be used if there is more than one endogenous variables and at least as many instruments. (2SLS is more general than method of momenets.)

10.4.2.2 From Regression Back to the simplest case where the structural equation is $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u$. And we have found two exogenous variables z_2, z_3. z_2, z_3 did not appear in structural equation, and is each uncorrelated with u. Such assumption for z_2, z_3 is called exclusion restrictions. (equivalent to requirements for IV except instrument relevance)

Since exogenous in the structural equation is uncorrelated with u, and IV is also uncorrelated with u, any combination of them must also be uncorrelated with u. Therefore, any arbitrary linear combination of exogenous variables would construct a effective IV. To find the one that correlates with y_2 the most, we resort to the reduced form equation of y_2,

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v$$

where $E(v) = 0, Cov(z_1, v) = Cov(z_2, v) = Cov(z_3, v) = 0$.

The "best" IV for y_2 turns out to be the linear combination in the reduced form equation above. Denoted it as y_2^*,

$$y_2^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3$$

In order to make sure that such IV y_2^* is not completely correlated with z_1, or say instrument relevance should hold, we test the joint hypothesis that $z_2 = z_3 = 0$. If the joint hypothesis cannot be rejected, then the structural equation cannot be identified.

The reduced form of y_2 successfully take apart y_2 into its endogenous part with correlation to u, and the exogenous part uncorrelated with u. The former is the reason why y_2 may be endogenous.

Luckily, although we do not know the true value of π_j, we can estimate the reduced form equation via OLS.

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3$$

Once we have got $\hat{y}_2$, we can use it as IV for y_2. Note that regardless of how many IV you originally induced for y_2, after estimating the reduced form equation and get $\hat{y}_2$ (in essence, find a best linear combination of exogenous variables), $\hat{y}_2$ is the single IV for y_2. Using method of moments,

$$\begin{cases} \sum_{i=1}^n (y_{1i} - \hat{\beta}_0 - \hat{\beta}_1 y_{2i} - \hat{\beta}_2 z_{1i}) = 0 \\ \sum_{i=1}^n z_{1i}(y_{1i} - \hat{\beta}_0 - \hat{\beta}_1 y_{2i} - \hat{\beta}_2 z_{1i}) = 0 \\ \sum_{i=1}^n \hat{y}_{2i}(y_{1i} - \hat{\beta}_0 - \hat{\beta}_1 y_{2i} - \hat{\beta}_2 z_{1i}) = 0 \end{cases}$$

From such equations under method of moments, an intriguing discovery is that, this turns out to be equivalent to carry a Two Stage Least Squares. Especially in the second stage of regression, we can see ($y_2 = y_2^* + v$),

$$y_1 = \beta_0 + \beta_1 y_2^* + \beta_2 z_1 + (u + \beta_1 v)$$

where $E(u + \beta_1 v) = 0$, and $u + \beta_1 v$ is uncorrelated with both y_2^* and z_1. That is why the regression of y_1 on $\hat{y}_2, z_1$ is effective.

Along with this logic, we will in the meantime feel at ease to grasp the principle for endogeneity test. (Covered later)

10.4.3 More Endogenous Variables

Consider the case with 2 endogenous variables:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + u$$

And suppose we have found two IVs, z_2, z_3.

- In the first stage
  - Regress y_2 on z_1, z_2, z_3, and get $\hat{y}_2$.
  - Regress y_3 on z_1, z_2, z_3, and get $\hat{y}_3$.
- In the second stage
  - Regress y_1 on $\hat{y}_2, \hat{y}_3, z_1$.

10.4.4 Remarks for IV

- There will be the problem of multicollinearity if with more than one endogenous variables, since $\hat{y}_2, \hat{y}_3$ are both linear combinations of z_1, z_2, z_3. Nevertheless, if not perfect multicollinearity, totally fine to conduct TSLS.
- The standard errors from the OLS second stage regression are wrong, since $\hat{y}_2$ is itself not y_2. If you are to get the "right" standard errors, you have to introduce "uncertainty" from the first to the second stage. Note that the **magnitude** of slope coefficients are *correct*.

10.5 Weak Instrument

$$\hat{\beta}_{1,IV} = \frac{\widehat{Cov}(z_i, y_i)}{\widehat{Var}(z_i, x_i)} = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})}$$

If z is a weak IV, denominator $\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})$ is close to 0, the distribution of corresponding t statistic is heavily-tailed. If such problem is neglected, you would take more **risk** committing type I error before you even know it.

10.6 IV in Measurement Error

IV can not only help cope with issues of omitting variables, but also can address measurenment error. For example, consider the simple linear regression

$$y = \beta_0 + \beta_1 x^* + u$$

where x^* is unobservable, $Cov(x^*, u) = 0$. x is a feasible measurement for x^* with measurement error $x = x^* + e$. Under CEV assumption, $Cov(x^*, e) = 0 \Longrightarrow Cov(x, e) = \sigma_e^2 \neq 0$. The equation can be equivalently rewritten as

$$y = \beta_0 + \beta_1 x + (u - \beta e)$$

Since $(u - \beta e)$ is then correlated with x, x is an endogenous variable due to measurement error.

What we need is an IV for x_1, which should highly correlate with x_1, and uncorrelate with measurement error. One possibility is to get a second measurement of x^*, say z; and z has another version of measurement error, $z = x^* + \eta$. We must assume that e is uncorrelated with η. By nature, z must have high correlation with x (Correlate through their common correlation with x^*), and z is uncorrelated with u. Just take z as an instrument for x.

10.7 Testing for Endogeneity of Explanatory Variables

In the general case, consider structural equation as

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + ... + \beta_k z_{k-1} + u_1$$

where y_2 is suspected to be endogenous, i.e., $Cov(y_2, u) \neq 0$. All other explanatory variables, $z_1, ..., z_{k-1}$, are exogenous.

One simple and intuitive way to think about endogeneity is to compare the estimated coefficients under OLS and 2SLS. If statistically salient difference is detected, then it is safe to allege that y_2 Is endogenous. However, using regression whenever possible is more convenient. Consider the reduced form regression,

$$y_2 = \pi_0 + \pi_1 z_1 + ... + \pi_k z_k + v_2$$

where z_k satisfies exclusion restrictions.

As before, $\hat{y}_2$ is the cleaned version without endogeneity, with residual v_2 of such reduced form equation absorbs all potential endogeneity. Then, y_2 is exogenous if and only if v_2 is uncorrelated with u_1. Rewrite u_1 as linear function of v_2

$$u_1 = \delta_1 v_2 + e_1$$

That is, $\delta_1 = 0$ in such equation.

Use our sample to estimate the equations above. Use OLS to estimate reduced form equation of y_2, and get $\hat{v}_2$. Then, conduct a regression for the equation

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + ... + \beta_k z_{k-1} + \delta_1 \hat{v}_2 + error$$

Use t statistics to test the null hypothesis $H_0 : \delta_1 = 0$. If we successfully reject H_0, we can deduce from the correlation of v_2 and v_1 that y_2 is endogenous.

To sum up, the general approach for endogeneity test for a set of variable(s) is

1. Regress each variable in the set on all exogenous (including those in the structural equation and induced IV) to get the estimated reduced form equation, and keep the residual;
2. Plug in all residuals(s) to the original structural equation and conduct a OLS regression;
3. Test against the joint hypothesis that all slope coefficient(s) of residual(s) be 0.

10.8 Over-Identifying Restrictions

When the instruments we have are more than our need, we will be able to test whether part of them are correlated with the structural error term. If this is the case, then the estimation using each single IV would be statistically different. If all IVs are exogenous, apart from sampling error, the error of 2SLS should be uncorrelated with all IVs. The general approach works like as follows.

1. Use 2SLS to estimate the structural equation, and get residual $\hat{u}_1$;
2. Regress $\hat{u}_1$ on all exogenous variables, and get R^2;
3. Under the null hypothesis that all IVs are uncorrelated with u_1, $nR^2 \overset{a}{\sim} \chi^q_1$, where q is the difference between the number of IVs and that of endogenous variables. If nR^2 exceeds the critical value, then H_0 can be rejected, and we can say that at least a set of all the IVs is not exogenous.