Dealing With Endogeneity

Junhui Qian

December 22, 2014

Outline

- Introduction
- Instrumental Variable
- Instrumental Variable Estimation
- Two-Stage Least Square Estimation
- Panel Data

Endogeneity in Econometrics

- In a multiple linear regression, if at least one of the regressors is correlated with the residual, then the exogeneity assumption ($\mathbb{E}(u|x) = 0$) is violated. We say that the regression suffers from endogeneity problem.
- The endogeneity problem occurs when
  - there is an omitted variable that is correlated with some regressors.
  - the dependent variable and at least one of the independent variables are determined simultaneously in a system.
  - there is measurement error in at least one of the regressors.
- When there is endogeneity problem, OLS estimates are biased and inconsistent.

Example: When There Is An Omitted Variable

Consider our favorite example,

$$\log(\textit{income}) = \beta_0 + \beta_1 \textit{edu} + \beta_2 \textit{expr} + u,$$

where we omit innate *ability* from the equation due to data availability. Since higher *ability* contributes both to higher *edu* (indirectly higher income) and directly to income, omitting *ability* would result in correlation between *edu* and u, the residual that contains the influence of *ability*.

Example: When There Is Simultaneity

Consider an imagined regression of employment level (L) on the average wage (W) and the foreign exchange rate (X),

$$L = \beta_0 + \beta_1 W + \beta_2 X + u.$$

Here it is arguable that W and L are determined simultaneously from the equilibrium of the labor market, which is constantly perturbed by shocks from both supply side (e.g., migration, epidemic) and demand side (e.g., productivity, energy and commodity price). Since the residual u also contains shocks from both supply and demand sides, W would be correlated with u.

Why OLS Fails

Consider a simple linear regression,

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

- To estimate β_1 by OLS, we rely on the assumption that $\mathbb{E}(u|x) = 0$, which implies $\mathrm{cov}(x, u) = 0$. This assumes that when x changes, u would remain zero in average. Only under this assumption, the change in y is useful for inferring β_1 (and β_0).
- If $\mathrm{cov}(x, u) \neq 0$, when x changes, u would change accordingly, then the change in y is not useful information.
- To consistently estimate the regression when $\mathrm{cov}(x, u) \neq 0$, we obviously need more information. Instrumental variables bring such information.

Instrument Variable

Consider a simple linear regression,

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

where $\mathrm{cov}(x, u) \neq 0$.

- An instrument variable (IV) for x is a random variable w satisfying

$$\mathrm{cov}(x, w) \neq 0 \quad \text{and} \quad \mathrm{cov}(w, u) = 0.$$

- An IV must be correlated with the endogenous variable, and at the same time uncorrelated with unobserved factors that affect y.
- When the IV changes, x changes accordingly, u remains zero in average, so the change in y would be useful information for inferring β's.

Looking for IV

- An instrument variable (IV) for x must satisfy

$$\text{(a) } \operatorname{cov}(x, w) \neq 0 \quad \text{and} \quad \text{(b) } \operatorname{cov}(w, u) = 0.$$

- It is usually easy to find w that satisfies (a) or (b). But It is challenging to find one that satisfies both.
- In particular, (b) is generally not verifiable, since u is unobserved. We must *argue* for the validity of (b) on the ground of economic intuition.
- To verify (a), we may run a simple linear regression, $x = \gamma_0 + \gamma_1 w + v$, and test $\mathrm{H}_0 : \gamma_1 = 0$ against the two-sided alternative.

Example: Looking for IV

Consider an omitted-variable example:

$$\log(\textit{wage}) = \beta_0 + \beta_1 \textit{edu} + u,$$

where we omitted *ability*.

- Suppose that *ability* has a none-zero partial effect on $\log(\textit{wage})$ (that is, if we run $\log(\textit{wage}) = \beta_0 + \beta_1 \textit{edu} + \beta_2 \textit{ability} + u$, we would obtain a $\hat{\beta}_2$ that is statistically significant.). Since *edu* and *ability* are correlated, omitting *ability* would result in the endogeneity of *edu*.
- It is easy to find variables that are correlated with *edu*, for example, mother's education attainment, family income. But it is difficult to argue for the case that these are not related with *ability*. Mother's education attainment, for example, may be positively correlated with children's innate ability, since the well-educated mother may have a better approach to raising children.

Example Continued

- It is also easy to find variables that are uncorrelated with u, for example, the last digit of ID number. But finding one that is also correlated with *edu* is tough.
- Usually it takes some imagination and creativity to come up with a good instrument.
- Can you think of any instrument for *edu* in our example?

IV in Multiple Linear Regression

Consider a multiple linear regression,

$$y = \beta_0 + \beta_1 x + \beta_2 z + u, \tag{1}$$

where $\mathrm{cov}(x, u) \neq 0$ and z is exogenous.

- An instrument variable (IV) for x is a random variable w satisfying

$$\text{(a) } \mathrm{cov}(x, w|z) \neq 0 \quad \text{and} \quad \text{(b) } \mathrm{cov}(w, u) = 0.$$

- The condition (a) is different from the simple regression case. It states that an IV must be correlated with the endogenous variable after partially out the effect of z.
- Equivalently, γ_1 in the following equation should be nonzero.

$$x = \gamma_0 + \gamma_1 w + \gamma_2 z + v. \tag{2}$$

- The equation (1) is often called structural equation, while (2) is called the reduced-form equation.

Instrumental Variable Estimation

Consider a regression in matrix form,

$$Y = X\beta + u,$$

where at least one of the regressors are endogenous. Suppose that in X, some regressors are endogenous and others exogenous. We represent X by $X = [X^{ex}\ X^{en}]$, where columns in X^{ex} correspond to exogenous variables, and those in X^{en} correspond to endogenous variables. Suppose we find a set of IV for X^{en}, say Z, then we define $W = [X^{ex}\ Z]$. The IV estimator for β is given by

$$\hat{\beta}_{iv} = (W'X)^{-1}W'Y.$$

Note that

- Z should have the same number of columns as X^{en}.
- If there is a constant, then the column of 1's is in X^{ex}.
- We may consider X^{ex} as instruments for themselves.

Deriving IV Estimator

The IV estimator can be obtained from

- Method of moments. From the moment condition $\mathbb{E}(w_i u_i) = 0$, we solve for β from

$$\frac{1}{n}\sum_{i=1}^{n} w_i(y_i - x_i'\beta) = 0,$$

and obtain $\hat{\beta} = (\sum_{i=1}^{n} w_i x_i')^{-1} (\sum_{i=1}^{n} w_i y_i')$.

- Non-orthogonal projection. To project Y on $\mathcal{R}(X)$ along the direction of u, we need an instrument W that is orthogonal to u. That is

$$W'(Y - X\beta) = 0.$$

Solving for β obtains $\hat{\beta} = (W'X)^{-1}W'Y$. Thus the projection of Y on $\mathcal{R}(X)$ along the direction of u is $X(W'X)^{-1}W'Y$ and the non-orthogonal projection is $P = X(W'X)^{-1}W'$.

IV Estimation for Simple Regression

For a simple linear regression $y = \beta_0 + \beta_1 x + u$ with an instrument w for x, the IV estimator is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(w_i - \bar{w})(y_i - \bar{y})}{\sum_{i=1}^{n}(w_i - \bar{w})(x_i - \bar{x})},$$

and

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1.$$

Note that

$$\beta_1 = \frac{\mathrm{cov}(w, y)}{\mathrm{cov}(w, x)},$$

$\hat{\beta}_1$ replaces population covariances by their sample analogs.

Properties of IV Estimation

- The IV estimator is generally biased.

$$\mathbb{E}\left(\hat{\beta}_{iv} - \beta\right) = \mathbb{E}(W'X)^{-1}W'u \neq 0$$

- The IV estimator is consistent. As $n \to \infty$,

$$\hat{\beta}_{iv} - \beta = \left(\frac{1}{n}\sum_{i=1}^{n} w_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} w_i u_i\right) \to_p 0.$$

- The IV estimator is asymptotically normal. As $n \to \infty$,

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_{iv} - \beta\right) &= \left(\frac{1}{n}\sum_{i=1}^{n} w_i x_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i u_i\right) \\ &\to_d N\left(0, \sigma^2(\mathbb{E}wx')^{-1}(\mathbb{E}ww')(\mathbb{E}xw')^{-1}\right).\end{aligned}$$

Asymptotic Covariance Matrix of the IV Estimator

Assuming homoscedasticity, $\mathbb{E}(u^2|w) = \sigma^2$.

- The asymptotic covariance matrix of $\hat{\beta}_{iv}$ is given by

$$\Sigma_{iv} = \frac{\sigma^2}{n}(\mathbb{E}wx')^{-1}(\mathbb{E}ww')(\mathbb{E}xw')^{-1}.$$

- If x is exogenous and $w = x$, the above expression reduces to $\frac{\sigma^2}{n}(\mathbb{E}xx')^{-1}$, the asymptotic covariance matrix for OLS estimator.

Asymptotic Covariance Matrix of the IV Estimator

If x is exogenous, then both OLS and IV give consistent estimate. Which is better, in terms of asymptotic efficiency?

- Consider a special case: suppose x and w are scalar zero-mean variables, then

$$\mathrm{var}(\hat{\beta}_{iv}) = \frac{\sigma^2}{n}\frac{\mathrm{var}(w)}{\mathrm{cov}(w,x)^2} = \frac{\sigma^2}{n}\frac{1}{\mathrm{var}(x)}\frac{1}{\mathrm{corr}(w,x)^2},$$

  where $\mathrm{corr}(w,x)$ is correlation coefficient between w and x.
- Three messages:
  - The more correlated between w and x, the more accurate is the IV estimator.
  - Since $|\mathrm{corr}(w,x)| \leq 1$, $\mathrm{var}(\hat{\beta}_{iv}) \geq \frac{\sigma^2}{n}\frac{1}{\mathrm{var}(x)} = \mathrm{var}(\hat{\beta}_{ols})$. Hence, when x is exogenous, OLS has a smaller asymptotic variance. We say that in this case, OLS is asymptotically more efficient than IV.
  - A larger sample size helps reduce asymptotic variance.
- These messages carry to the general case.

Asymptotic Bias of IV Estimator

If there is slight correlation between the instrument and the residual, the IV estimator would be asymptotically biased. The bias may be severe if the correlation between the instrument and the endogenous variable is low.

- As $n \to \infty$,

$$\hat{\beta}_{iv} \to_p \beta + \left(\mathbb{E} w_i x_i'\right)^{-1} \left(\mathbb{E} w_i u_i\right).$$

- Consider the special case where we assume x and w are scalar non-zero variables,

$$\text{Asymp. Bias} = \frac{\text{corr}(w, u)}{\text{corr}(w, x)} \frac{\sigma_u}{\sigma_x}.$$

- To avoid large asymptotic bias, we should choose instruments that are more correlated with the endogenous variable.

Student t test with IV Estimator

- To construct Student t statistic, we need an estimate of the covariance matrix of $\hat{\beta}_{iv}$,

$$\Sigma_{iv} = \frac{\sigma^2}{n}(\mathbb{E}wx')^{-1}(\mathbb{E}ww')(\mathbb{E}xw')^{-1}$$

- σ^2 is estimated by

$$s^2 = \frac{1}{n-1-k}\sum_{i=1}^{n}\hat{u}_i^2,$$

and we estimate Σ_{iv} by

$$\hat{\Sigma}_{iv} = s^2(W'X)^{-1}(W'W)(X'W)^{-1}.$$

- Taking square root of the diagonal of $\hat{\Sigma}_{iv}$, we obtain standard errors for the IV estimators.
- The t statistic for, say, $\mathrm{H}_0 : \beta_j = b$, is given by $t_{\beta_j} = \frac{\hat{\beta}_j - b}{\mathrm{se}(\hat{\beta}_j)}$, which is distributed as $N(0,1)$ asymptotically.

When There Are More Instruments

Let w_i be a q-dimensional vector of IV for a p-dimensional regressor x_i. Some elements of x_i are endogenous. This is why we need IV. And some elements of w_i are identical to exogenous variables in x_i. (Almost always, there is the constant 1 in both x_i and w_i, allowing for a constant term in the regression.) We have discussed IV estimation which assumes $q = p$.

What if $q > p$? That is, how do we proceed if the number of instruments is bigger than the number of endogenous variables (over-identification)?

Two-State Least Square

The TSLS projects the columns of X onto the range of W, $\mathcal{R}(W)$. This results in a new set of instruments with identical number of columns with X,

$$\hat{X} = W(W'W)^{-1}W'X = P_W X. \tag{3}$$

Then we use this new instrument in IV estimation and obtain

$$\hat{\beta}_{2sls} = (\hat{X}'X)^{-1}\hat{X}'Y = \left(X'P_W X\right)^{-1} X'P_W' Y. \tag{4}$$

This is called two-stage least square because we can write (4) equivalently as

$$\hat{\beta}_{2sls} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y.$$

This is the second least square, of Y on $\hat{X}$. (The first one is X on W in (3).)

GMM Approach

The TSLS is a special case of the GMM (Generalized Method of Moments) estimation, which solves

$$\min_{\beta} \left[W'(Y - X\beta)\right]' \Omega \left[W'(Y - X\beta)\right],$$

where Ω is a symmetric positive definite matrix. Solving for β obtains

$$\hat{\beta} = \left(X'W\Omega W'X\right)^{-1} X'W\Omega W'Y.$$

Let $\Omega = (W'W)^{-1}$, the above estimator reduces to the TSLS estimator. It can be shown that if homoscedasticity holds, TSLS is the most efficient GMM estimator.

Panel Data

- A panel data contain information on the same group of individuals (persons, households, firms, provinces, countries, etc.) over a period of time.
- If a panel data is available, we may deal with endogeneity problems without resorting to IV, at least to some extent.
- See an example of a panel data set next page.

An Example of Panel Data

Person	Year	Wage	Gender	Age
1	2001	4000	0	22
1	2002	5000	0	23
1	2003	6000	0	24
2	2001	7000	1	27
2	2002	7500	1	28
2	2003	8000	1	29
3	2001	1500	0	19
3	2002	1600	0	20
3	2003	1650	0	21
		⋮		

A Panel Data Model for Endogeneity Problem

- Suppose that we regress y on x. If some of the elements in x is endogenous, then OLS of $y_i = \beta_0 + x_i'\beta + u_i$ using cross-section data would result in inconsistent estimates. Panel data, with more information on x and y, may help.
- We may write the following panel data model,

$$y_{it} = x_{it}'\beta + u_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T,$$

  where $u_{it} = \mu_i + v_{it}$,
  - μ_i is a time-invariant individual effect for individual i that may be correlated with x_{it}.
  - v_{it} is iid N(0,σ_v^2), independent of x and z. v_{it} is called idiosyncratic error.
- This model is often called "fixed-effect model". If, in addition, we assume that $\mu_i \sim \text{iid } \mathrm{N}(0, \sigma_\mu^2)$ is independent from x_{it} and v_{it}, then the model is often called "random-effect model".

Estimating Fixed-Effect Panel Data Model: I

- An obvious approach is to get rid of μ_i by taking first difference of the equation for each individual. Let $\Delta y_{it} \equiv y_{it} - y_{i,t-1}$, we have

$$\Delta y_{it} = \Delta x_{it}' \beta + e_{it},$$

  where $e_{it} = \Delta v_{it}$.
- Now we can estimate β by OLS.
- e_{it} is serially correlated, so OLS would be inefficient.

Estimating Fixed-Effect Panel Data Model: II

- A second approach is to get rid of μ_i by subtracting individual means from each observations. Specifically, let $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ and similarly for other variables. In terms of individual means, the model is

$$\bar{y}_i = \bar{x}_i'\beta + \mu_i + \bar{v}_i.$$

  Subtracting the individual means from the original model, we obtain

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)'\beta + (v_{it} - \bar{v}_i).$$

- Now OLS estimates β efficiently.

Estimating Fixed-Effect Panel Data Model: II

It can be shown that the second approach is, in effect, to treat individual effects as coefficients on dummy variables and run least square (LSDV). Specifically, let (y_i, X_i) be the T observations on the i-th individual. We can rewrite our model as

$$y_i = X_i\beta + \iota\mu_i + v_i, \ \ i = 1, \ldots, N.$$

Or in matrix form,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \beta + \begin{pmatrix} \iota & 0 & \cdots & 0 \\ 0 & \iota & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \iota \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} + \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}.$$

Assessing Fixed-Effect Panel Data Model

- Fixed-effects panel data model offers a solution to the endogeneity problem without resorting to IV. Instead, it relies on longer span of data collection on the same individual.
- Fixed-effects model can be consistently estimated as long as the idiosyncratic errors are uncorrelated with the regressors.
- Time-invariant regressors are absorbed by the fixed effects. Thus the effects of time-invariant regressors are unidentified in fixed-effects panel data models. In estimation, it is clear that any time-invariant regressor (e.g., gender, education) would disappear after the first-differencing or de-mean transformation.

The Random-Effect Panel Data Model

If the individual effects are not correlated with any regressors, i.e., there is no endogeneity problem, then we may use the random-effect panel data model,

$$y_{it} = x_{it}'\beta + u_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T,$$

where $u_{it} = \mu_i + v_{it}$,

- $\mu_i \sim \text{iid } \mathrm{N}(0, \sigma_\mu^2)$ is independent from x_{it} and v_{it}
- v_{it} is iid $\mathrm{N}(0,\sigma_v^2)$, independent of x and z.

The random-effect model can be consistently estimated by OLS, or, more efficiently, GLS.

Estimating Random-Effect Panel Data Model

Note that the covariance matrix of $u = (u_1', \ldots, u_n')'$ has a particular structure,

$$\Omega = \begin{pmatrix} \Sigma & 0 & \cdots & 0 \\ 0 & \Sigma & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \Sigma \end{pmatrix},$$

where

$$\Sigma = \begin{pmatrix} \sigma_\mu^2 + \sigma_v^2 & \sigma_\mu^2 & \cdots & \sigma_\mu^2 \\ \sigma_\mu^2 & \sigma_\mu^2 + \sigma_v^2 & \cdots & \sigma_\mu^2 \\ & & \vdots & \\ \sigma_\mu^2 & \sigma_\mu^2 & \cdots & \sigma_\mu^2 + \sigma_v^2 \end{pmatrix}$$

Assessing Random-Effect Panel Data Model

- In the random-effect model, time-invariant regressors are no longer absorbed by the fixed effects. Thus the effects of time-invariant regressors are identified in random-effects panel data models.
- When the random-effect assumptions hold, the random-effect approach is more efficient. However, if there is correlation between individual effects and any regressor, then the random-effect approach would yield inconsistent estimation.
- In practice, we use Hausman-Wu test to check whether the random-effect approach can be employed.