Econometrics

Chapter 15: Instrumental Variables Estimation and Two Stage Least Squares

In Choi

Sogang University

Why Use Instrumental Variables?

- Instrumental Variables (IV) estimation is used when your model has endogenous $x\prime$s. That is, whenever $Cov(x, u) \neq 0$.
- Thus, IV can be used to address the problem of omitted variable bias.
- Additionally, IV can be used to solve the classic errors-in-variables problem.

Example

Simultaneous equations

Let C_t : consumption at time t
Y_t : income at time t
I_t : investment at time t

The Keynesian consumption function is

$$C_t = \alpha + \beta Y_t + u_t.$$

But $Y_t = C_t + I_t$. Using these two equations, we have

$$Y_t = \alpha + \beta Y_t + u_t + I_t \Rightarrow Y_t = \frac{1}{1-\beta}\left(\alpha + u_t + I_t\right).$$

Thus Y_t and u_t are correlated.

Examples

Example

Measurement error

Let the true regression model be

$$y_i = \alpha + \beta x_i + u_i.$$

Suppose that we observe

$$x_i^* = x_i + w_i \ \ \left(w_i \sim iid\left(0, \sigma_w^2\right)\right)$$

instead of x_i due to measurement error.

Examples

Example

(continued) Then, the regression model we use will be

$$
\begin{aligned}
y_i &= \alpha + \beta \left(x_i^* - w_i \right) + u_i \\
&= \alpha + \beta x_i^* + u_i - \beta w_i .
\end{aligned}
$$

Obviously, x_i^* and the error terms are correlated.

What Is an Instrumental Variable?

- In order for a variable, z, to serve as a valid instrument for x, the following must be true.

1. The instrument must be exogenous. That is, $Cov(z, u) = 0$.
2. The instrument must be correlated with the endogenous variable x. That is, $Cov(z, x) \neq 0$.

More on Valid Instruments

- We have to use common sense and economic theory to decide if it makes sense to assume $Cov(z, u) = 0$.
- We can test if $Cov(z, x) \neq 0$.
  Just test $H_0 : \pi_1 = 0$ in $x = \pi_0 + \pi_1 z + v$
- Sometimes refer to this regression as the first-stage regression

IV Estimation in the Simple Regression Case

- For

$$y_t = \beta_0 + \beta_1 x_t + u_t,$$

and given our assumptions

$$Cov(z_t, y_t) = \beta_1 Cov(z_t, x_t) + Cov(z_t, u_t).$$

So

$$\beta_1 = Cov(z_t, y_t) / Cov(z_t, x_t).$$

- Then the IV estimator for β_1 is

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n} (z_t - \bar{z})(y_t - \bar{y})}{\sum_{t=1}^{n} (z_t - \bar{z})(x_t - \bar{x})}.$$

IV Estimation in the Simple Regression Case

- Since

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{t=1}^{n}(z_t - \bar{z})u_t}{\sum_{t=1}^{n}(z_t - \bar{z})(x_t - \bar{x})},$$

approximate variance of $\hat{\beta}_1$ (note that $E(\hat{\beta}_1) \neq \beta_1$) is

$$\begin{aligned}
& \frac{E\left[\left(\sum_{t=1}^{n}(z_t - \bar{z})u_t\right)^2 \mid given\ all\ z\right]}{\left(\sum_{t=1}^{n}(z_t - \bar{z})(x_t - \bar{x})\right)^2} \\
= \quad & \frac{\sum_{t=1}^{n}(z_t - \bar{z})^2\sigma^2}{\left(\sum_{t=1}^{n}(z_t - \bar{z})(x_t - \bar{x})\right)^2} \\
= \quad & \frac{\sigma^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2 \frac{\left(\sum_{t=1}^{n}(z_t - \bar{z})(x_t - \bar{x})\right)^2 / \sum_{t=1}^{n}(z_t - \bar{z})^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2}} \\
= \quad & \frac{\sigma^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2 R_{x,z}^2},
\end{aligned}$$

where $R_{x,z}^2$ is the R-square from regressing x on z.

Inference with IV Estimation

- The homoskedasticity assumption in this case is

$$E(u_t^2 | all\ z) = \sigma^2.$$

- As in the OLS case, given the asymptotic variance, we can estimate the standard error

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2 R_{x,z}^2}}.$$

IV versus OLS estimation

- Since $R^2 < 1$, IV standard errors are larger.
- However, IV is consistent, while OLS is inconsistent when $Cov(x, u) \neq 0$.
- The stronger the correlation between z and x, the smaller the IV standard errors.

Two Stage Least Squares (2SLS)

- Structural model: A model based on economic theory.
- One or more of the variables in structural models may be endogenous. We need an instrument for each endogenous variable.
- Write the structural model as

$$y_{1t} = \beta_1 y_{2t} + \beta_2 z_{1t} + u_t,$$

where y_{2t} is endogenous and z_{1t} is exogenous.

- Assume the reduced form relations

$$\begin{aligned} y_{1t} &= \pi_1 z_{1t} + \pi_2 z_{2t} + w_t, \\ y_{2t} &= \psi_1 z_{1t} + \psi_2 z_{2t} + v_t, \end{aligned}$$

where z_{1t} and z_{2t} are exogenous, $w_t \sim iid(0, \sigma_w^2)$ and $v_t \sim iid(0, \sigma_v^2)$.

Then,

$$\begin{aligned} y_{1t} &= \beta_1 y_{2t} + \beta_2 z_{1t} + u_t \\ &= \beta_1 \left(\psi_1 z_{1t} + \psi_2 z_{2t} + v_t \right) + \beta_2 z_{1t} + u_t \\ &= \left(\beta_1 \psi_1 + \beta_2 \right) z_{1t} + \beta_1 \psi_2 z_{2t} + \beta_1 v_t + u_t \\ &= \pi_1 z_{1t} + \pi_2 z_{2t} + w_t. \end{aligned}$$

Thus, $u_t = w_t - \beta_1 v_t$, which shows that $\{u_t\}$ and $\{v_t\}$ are related if $\beta_1 \neq 0$. In addition,

$$\begin{aligned} \pi_1 &= \beta_1 \psi_1 + \beta_2 \\ \pi_2 &= \beta_1 \psi_2. \end{aligned}$$

Two Stage Least Squares (2SLS)

- Note that y_{2t} and u_t are correlated due to the presence of v_t in y_{2t}.
- Thus, substitute $\hat{y}_{2t}$ for y_{2t} in the structural model ($\hat{y}_{2t}$ is the part of y_t that is free of v_t) and obtain the OLS coefficient estimates. This is the 2SLS estimation.
- The standard errors of 2SLS are different from those of OLS.
- If $\psi_2 = 0$, we have a multicollinearity problem.

Addressing Errors-in-Variables with IV Estimation

- Remember the classical errors-in-variables problem where we observe x_1^* instead of x_1 where $x_1^* = x_1 + w_1$, and w_1 is uncorrelated with x_1.
- If there is a z such that $Cov(z, u) = 0$ and $Cov(z, x_1^*) \neq 0$, then IV will remove the bias.

Testing for Endogeneity

- Since OLS is preferred to IV if we do not have an endogeneity problem, we wish to be able to test for endogeneity.
- If we do not have endogeneity, both OLS and IV are consistent. Idea of the Durbin-Wu-Hausman test is to see if the estimates from OLS and IV are different.
- See Durbin (1954, Review of the International Statistical Institute), Wu (1973, Econometrica), and Hausman (1978, Econometrica).

Testing for Endogeneity

- The null hypothesis for the DWH test is

$$H_0 : Cov\,(x, u) = 0$$

- Under H_0, OLS and IV are both consistent. If H_0 is violated, only IV is consistent. Thus, the DWH test is based on the difference of IV and OLS.