Econometrics

Week 9

Institute of Economic Studies
Faculty of Social Sciences
Charles University in Prague

Fall 2022

Recommended Reading

For today

- Simultaneous Equations Models
- Chapter 16

For next week

- Continuation of Simultaneous Equations models (Chapter 16)
- Introduction to limited dependent variables
- Revision of maximum likelihood estimation

Today's Talk

- Today, we will consider another important form of endogeneity:
  - **simultaneity**
- **Simultaneity** is a specific type of endogeneity occurring when the explanatory variable is determined jointly with the dependent variable.
- This usually happens when the dependent and explanatory variables are connected through an equilibrium system.
- We will show that this source of endogeneity can be also treated using the instrumental variable estimation.
- But to apply it correctly, we need to understand the data generating process very well.

Example of Simultaneity

Supply Equation

Consider labor supply function (intercept suppressed for simplicity):

$$h_i^s = a_1 w_i + b_1 z_{1i} + u_{1i},$$

where h_i^s is (annual) labor hours supplied by workers in agriculture, w_i average hourly wage in agriculture, and z_{1i} observed variable affecting supply (i.e. wage in manufacturing sector) in country i.

- We call supply equation a **structural equation** as it comes from the economic theory and has casual interpretation.
- a_1 measures how labor supply changes with change of wage.
- When h_i^s and w_i are in logarithms, a_1 is labor supply elasticity.

Example of Simultaneity cont.

$$h_i^s = a_1 w_i + b_1 z_{1i} + u_{1i},$$

- Note that both z_{1i} and u_{1i} shift supply, z_{1i} is observed, while u_{1i} is not.
- So how does this equation differ from what we have studied previously?
- Where does w_i come from?
- Wage w_i cannot be viewed as exogenous, because it does not vary randomly.
- We cannot simply regress h_i^s on w_i, as w_i is an **endogenous** variable.
- We have to understand that the data are best described by an equilibrium system of labor supply and demand.

Example of Simultaneity cont.

Demand Equation

Consider labor demand function (intercept suppressed for simplicity):

$$h_i^d = a_2 w_i + b_2 z_{2i} + u_{2i},$$

where h_i^d is (annual) labor hours demanded in agriculture, w_i average hourly wage in agriculture, and z_{2i} some observed variable affecting demand (i.e. the total agricultural land area) in country i.

- Demand equation is also **a structural equation**.
- These two equations are linked through intersection of supply and demand, which is equilibrium $h_i^s = h_i^d$.
- As we observe only equilibrium hours for each country i, we denote h_i observed hours.

Example of Simultaneity cont.

Labor Supply and Demand Equation

$$h_i = a_1 w_i + b_1 z_{1i} + u_{1i},$$

$$h_i = a_2 w_i + b_2 z_{2i} + u_{2i},$$

- We call this a **simultaneous equations model (SEM)**.
- h_i and w_i are **endogenous variables** determined within the system.
- z_{1i} and z_{2i} are **exogenous variables** determined outside of the system.
- Without z_{1i} and z_{2i} we can not recognize demand from supply $\Rightarrow$ they identify the equations.
- Without exogenous **observable** supply and demand shifters, we are not able to estimate the system.

Simultaneity Bias in OLS

Consider a general structural model

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1, \ (1)$$
$$y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2$$

- Variables z_1 and z_2 are exogenous ($Cov(z_1, u_1) = 0$, $Cov(z_2, u_2) = 0$).
- If we put first equation to the second one, we have:
  $y_2 = \alpha_2(\alpha_1 y_2 + \beta_1 z_1 + u_1) + \beta_2 z_2 + u_2$,
  $(1 - \alpha_1\alpha_2)y_2 = \alpha_2\beta_1 z_1 + \beta_2 z_2 + \alpha_2 u_1 + u_2$.

Reduced form for y_2

$$y_2 = \pi_{21} z_1 + \pi_{22} z_2 + \nu_2,$$
where $\pi_{21} = \alpha_2\beta_1/(1 - \alpha_2\alpha_1)$, $\pi_{22} = \beta_2/(1 - \alpha_2\alpha_1)$ and
$$\nu_2 = (\alpha_2 u_1 + u_2)/(1 - \alpha_2\alpha_1).$$

Simultaneity Bias in OLS cont.

$$y_2 = \underbrace{\frac{\alpha_2\beta_1}{(1-\alpha_2\alpha_1)}}_{\pi_{21}} z_1 + \underbrace{\frac{\beta_2}{(1-\alpha_2\alpha_1)}}_{\pi_{22}} z_2 + \underbrace{\frac{\alpha_2 u_1 + u_2}{(1-\alpha_2\alpha_1)}}_{\nu_2},$$

What can we learn from the reduced-form equation?

- y_2 and u_1 are correlated.
- Thus, OLS estimates $\hat{\alpha}_1^{OLS}$ and $\hat{\beta}_1^{OLS}$ from equation (1) are inconsistent.
- When y_2 and u_1 are correlated because of simultaneity, OLS estimates suffer from the **simultaneity bias**.
- ν_2 is a linear function of u_1 and u_2, u_1 and u_2 are uncorrelated with z_1 and $z_2 \Rightarrow \nu_2$ is also uncorrelated with z_1 and z_2.
- Thus OLS estimates $\hat{\pi}_{21}$ and $\hat{\pi}_{22}$ are consistent.

Simultaneity Bias in OLS cont.

Important

Estimating a structural equation in a simultaneous equations system by OLS results in biased and inconsistent estimates!

- We can solve this problem by using instrumental variables
  - Either by estimating the reduced form and recovering coefficients,
  - or by **two-stage least square estimator (2SLS)**.
- As we specify the structural equation for each endogenous variable, we can immediately see if sufficient number of instrumental variables is available to estimate the equation.
- We call this **identification problem.**

Identification in Two-Equation System

Example: Supply and Demand of labor

$$h_i = a_1 w_i + u_{1i}$$

$$h_i = a_2 w_i + b_2 land_i + u_{2i},$$

where h_i is total number of hours worked in agriculture in country i, w_i is wage in agriculture, $land_i$ is total agricultural land (exogenous to supply and demand of labor).

- Second equation is demand ($\uparrow land \Rightarrow \uparrow$ demand).
- Which equation can be estimated (which is **identified**)?
  - The supply equation (first), because we have an instrumental variable (IV) for it.
  - Recall, that the IV has to be correlated with the endogenous explanatory variable (here with wage), but at the same time excluded from the equation (i.e. exogenous).
  - $land_i$ satisfies these conditions for the first equation.
  - But we have no IV for wage in the demand equation!

Identification in Two-Equation System

Example: Supply and Demand of labor

$$h_i = a_1 w_i + u_{1i}$$

$$h_i = a_2 w_i + b_2 land_i + u_{2i},$$

where h_i is total number of hours worked in agriculture in country i, w_i is wage in agriculture, $land_i$ is total agricultural land (exogenous to supply and demand of labor).

- We can estimate the supply equation using 2SLS.
- Simply use the agricultural land area ($land_i$) as the instrumental variable for wage in the supply equation.

Identification in Two-Equation System cont.

General two-equation model

$$y_1 = \beta_{10} + \alpha_1 y_2 + \beta_{11} z_{11} + \ldots + \beta_{1k} z_{1k} + u_1$$

$$y_2 = \beta_{20} + \alpha_2 y_1 + \beta_{21} z_{21} + \ldots + \beta_{2k} z_{2k} + u_2$$

where y_1 and y_2 are endogenous variables and u_1, u_2 are structural error terms and we have k exogenous variables $\mathbf{z}$.

- Under what assumptions can we estimate the parameters in this model?
- This is what we call an identification issue.

Rank Condition for Identification

- A necessary and sufficient condition for one of the equations to be identified is:

Rank Condition

The first equation in a two-equation simultaneous equations model is identified if, and only if, the second equation contains at least one exogenous variable (with nonzero coefficient) that is excluded from the first equation.

- **Order condition** is necessary for the rank condition.

Order Condition

The first equation in a two-equation simultaneous equations model is identified if at least one exogenous variable is excluded from the first equation.

- Order condition is trivial to check.

Identification in Systems with More Equations

- Identification of systems with more equations requires more advanced matrix algebra.
- But, we can discuss some issues.

Three-equation system

$$
\begin{aligned}
y_1 &= \alpha_{13}y_3 + \beta_{11}z_1 + \beta_{12}z_2 + \beta_{13}z_3 + u_1 \\
y_2 &= \alpha_{21}y_1 + \alpha_{23}y_3 + \beta_{23}z_3 + u_2 \\
y_3 &= \alpha_{32}y_2 + \beta_{31}z_1 + \beta_{32}z_2 + \beta_{33}z_3 + \beta_{34}z_4 + u_3,
\end{aligned}
$$

where y's are endogenous and z's are exogenous.

- Which of these equations can be consistently estimated?
- Check the order condition first!
- We can see that third equation contains all exogenous variables, thus there is no IV for y_3 and this equation can not be estimated consistently.

Conditions for Identification

Order Condition

An equation in a system of equations model satisfies the order condition for identification if the number of *excluded* exogenous variables from the equation is at least as large as the number of right-hand side endogenous variables.

- In our three-equation system, first equation passes this condition $\Rightarrow$ there is z_4 excluded for y_2.
- Second equation also passes the order condition $\Rightarrow$ there are three excluded variables for y_1 and y_3: z_1, z_2 and z_4.
- BUT remember: order condition is only necessary, not sufficient condition for identification.
- Suppose $\beta_{34} = 0$. Then first equation is not identified.
- We need to extend rank condition (but you have to wait until advanced econometrics course).

Conditions for Identification

Rank Condition

An equation in a system of G-equations model (i.e. a model with G endogenous variables) is identified if the rank of the matrix of parameters of all the excluded variables (endogenous and exogenous) from that equation is equal or larger than (G-1).

- In our three-equation system, second equation passes this condition $\Rightarrow$ there is z_4 excluded for y_1.
- In more complicated models the rank condition requires advanced matrix algebra.

Conditions for Identification cont.

- If we have more excluded exogenous variables from the equation than included endogenous variables, equation is **overidentified**.
- The second equation from our example is overidentified.
- First equation is **just identified**.
- Third equation is **unidentified** and cannot be estimated.
- Each identified equation can be estimated by 2SLS.
- We also know methods which are more efficient, like **three-stage least squares (3SLS)**.
- But these are little more complicated and you have to wait until advanced econometrics course.

Thank you

Remember

- Read Chapter 17 (first part - binary response models)
- Revise maximum likelihood estimation from Statistics!

Home assignment 3

Due on Wednesday, December 7