Simultaneous Equations Models

2025/2026, Semester 1

Ralf Wilke

Copenhagen Business School

Simultaneous Equations Models

Textbook: Wooldridge (2025), Introductory Econometrics , Chapter 16

Motivation

- Endogeneity can be for several reasons.
- In the case of omitted variables we can use the method of instrumental variables to obtain consistent estimates.
- Another important form of endogeneity of explanatory variables is **simultaneity**.
  - If explanatory variables are jointly determined with the dependent variable, typically through an equilibrium mechanism.
- We will see how we can estimate simple simultaneous equations models (SEM).
- The method for estimating is again the method of IV.
- However, crafting and interpretation of SEMs is challenging.

The Nature of SEMs

- We have a system of equations
- Each equation should have a causal interpretation
- Since we only observe the outcomes in equilibrium, we are required to use counterfactual reasoning.
- The classical example in Economics:
  - Demand and Supply equations for some commodity or input to production (such as labour or capital)
  - Let us focus on the demand for and supply of a retail good.
  - A simple supply function is:

$$q_s = \alpha_1 p + \beta_1 z_1 + u_1$$

with q: quantity supplied in a regional unit (e.g. country), *p*: retail price, z_1: some other observed variable affecting quantity supplied.

$$q_s = \alpha_1 p + \beta_1 z_1 + u_1$$

- This is an example of a **structural equation**.
  - As it is derivable from economic theory, it has a causal interpretation. It can be obtained from profit maximization of firms.
  - If *q* and *p* are in logarithmic form, α_1 is the elasticity of supply and we would expect $\alpha_1 \geq 0$.
  - A change in z_1 and u_1 shifts supply with the difference that u_1 is not observed.
  - z_1 is therefore an observed supply shifter and u_1 is an unobserved supply shifter.
  - What is then the issue here compared to equations in the previous lectures?
  - Even though the equation holds for all values of *p*, the *p* in the data cannot be considered as exogenously varying in the data.
    - This is because price is not randomly determined. It is a result of an interaction between product demand and product supply.

- Equilibrium prices and quantities are jointly determined by supply and demand. The latter is given by:

$$q_d = \alpha_2 p + \beta_2 z_2 + u_2$$

where q_d is the quantity of products demanded.

- z_2 is an observable demand shifter.
- u_2 is an unobservable demand shifter.
- The demand equation is also a **structural equation**.
- It can be obtained from utility maximization of individuals.
- If q_d and *p* are in logarithmic terms, α_2 is the demand elasticity with respect to changes in retail prices.
- Economic theory (normally) suggests: $\alpha_2 < 0$.

- Supply is an equation for firms, demand is a behavioural equation for individuals.
- Each has its own interpretation and stands for its own.
- They become linked because observed retail prices and quantities are determined by the intersection of demand and supply:

$$q_{is} = q_{id}$$

  where *i* is for example a county.
- q_i is then the equilibrium quantity for county *i*. This is what we observe in the data.
- The equilibrium condition then implies:

$$\begin{aligned} q_i &= \alpha_1 p_i + \beta_1 z_{i1} + u_{i1} \\ q_i &= \alpha_2 p_i + \beta_2 z_{i2} + u_{i2} \end{aligned}$$

where q_i and p_i are equilibrium observed values for region *i*.

$$
\begin{aligned}
q_i &= \alpha_1 p_i + \beta_1 z_{i1} + u_{i1} \\
q_i &= \alpha_2 p_i + \beta_2 z_{i2} + u_{i2}
\end{aligned}
$$

- These two equations constitute a **simultaneous equations model** (SEM).
- z_{i1} and z_{i2} are exogenous variables as they are determined outside the system.
  - We assume that they are not correlated with the structural errors u_{i1} and u_{i2}.
- q_i and p_i are endogenous variables as they are determined by the system.

- We also need to impose identifying assumptions on the model.
- For example: in absence of z_{i1} and z_{i2}, the demand and supply side of equations are not distinguishable.
- If z_{i1} and z_{i2} are identical, the equations look also identical and it will not be possible to identify the system.

- <u>Example</u>: Theft Rates and Size of the Private Security Force
  - Retailers normally have contracts with private security providers to ensure that shops are save (vandalism, theft, social disorder).
  - By how much do additional security officers decrease theft rates?
  - A simple shop cross section model is:

$$theftps = \alpha_1 offps + \beta_{10} + \beta_{11} incpc + u_1$$

    with *theftps* is thefts per shop, *offps* is number of security officers per shop, *incpc* is income per capita at shop's location.
  - *incpc* is exogenous (and one could also include more exogenous variables such as age, and education distribution.)
  - We want to answer: if an outlet exogenously increases its security force, will that increase, on average, lower theft rates?
  - If *offps* is exogenous, we could estimate it by OLS.
  - As spending on law enforcement is partly determined by expected crime, *offps* cannot be considered as exogenous.

- To reflect this, we can postulate a second relationship:

$$offps = \alpha_2 theftps + \beta_{20} + otherfactors$$

and we suppose $\alpha_2 > 0$.

- Other things being equal, shops with higher (expected) theft rates will have more officers per shop.
- If we fully specify the second equation, we have a simultaneous equation model.
- Although, we are interested in the first equation only we will be also required to properly state the second.

- Also note, each equation has a proper interpretation and stands on its own:
  - The first describes the action of potential thieves, while the second describes behaviour of the retailer.
- This makes the system an appropriate SEM.

- In a SEM it does not make sense to specify two equations where each is a function of the other.
  - Example: two variables are simultaneously determined by one economic agent:
    - Expenditures and savings by a household (in response to a change in income).
    - Weekly hours spent studying and weekly hours working by a student.
  - For a SEM to make sense, each equation has to have a clear ceteris paribus interpretation in isolation of the other equation.

Simultaneity Bias in OLS

- If an explanatory variable is determined simultaneously with the dependent variable, it is generally correlated with the error terms.
- In this case OLS is biased and inconsistent.
- As an example we consider two equations without an intercept:

$$\begin{aligned} y_1 &= \alpha_1 y_2 + \beta_1 z_1 + u_1 \\ y_2 &= \alpha_2 y_1 + \beta_2 z_2 + u_2 \end{aligned}$$

  the variables z_1 and z_2 are exogenous. We focus on estimation of the first equation.

$$
\begin{aligned}
y_1 &= \alpha_1 y_2 + \beta_1 z_1 + u_1 \\
y_2 &= \alpha_2 y_1 + \beta_2 z_2 + u_2
\end{aligned}
$$

- To show that the dependent variables are generally correlated with the error terms (e.g. y_2 with u_1), we plug in the first equation for y_2 into the second:

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

  or

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

  In order to solve for y_2 we have to assume $\alpha_1\alpha_2 \neq 1.$
- It depends on the application whether this is restrictive. In the supply and demand example it was very reasonable as $\alpha_1 \leq 0, \alpha_2 \geq 0.$

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

Can be rewritten as:

$$y_2 = \pi_{21} z_1 + \pi_{22} z_2 + \nu_2$$

where

$$\begin{aligned} \pi_{21} &= \alpha_2\beta_1/(1-\alpha_2\alpha_1) \\ \pi_{22} &= \beta_2/(1-\alpha_2\alpha_1) \\ \nu_2 &= (\alpha_2 u_1 + u_2)/(1-\alpha_2\alpha_1) \end{aligned}$$

This is a reduced form equation for y_2.

- π_{21}, π_{22} are again reduced form parameters.
- ν_2 is linear in $u_1 \text{ and } u_2$. For this reason, it is uncorrelated with $z_1 \text{ and } z_2$. We can apply OLS to estimate π_{21}, π_{22}.
- There is an equivalent reduced form equation for y_1.

$$
\begin{aligned}
y_2 &= \pi_{21} z_1 + \pi_{22} z_2 + \nu_2 \\
y_2 &= \pi_{21} z_1 + \pi_{22} z_2 + (\alpha_2 u_1 + u_2)/(1 - \alpha_2 \alpha_1)
\end{aligned}
$$

- We can use this equation to show that OLS estimation of the structural equations will generally result in biased and inconsistent estimates for α_j and β_j :

$$
y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1
$$

- We need that y_2 and u_1 are uncorrelated.
- From the reduced form equation, we see that y_2 and u_1 are correlated if ν_2 and u_1 are correlated. Since ν_2 is linear in u_1, it is generally correlated with u_1 .
- When is it not correlated?
  - If $\alpha_2 = 0$ and if u_1 and u_2 are uncorrelated.
  - In this case y_2 is not simultaneously determined with y_1 . Moreover, the error in determining y_2 is not correlated with y_1 .

- When y_2 is correlated with u_1 because of simultaneity, we say OLS suffers from simultaneity bias and it is inconsistent.

- Obtaining the direction of the bias is generally complicated. Simple expressions of the bias can be derived under additional assumptions but this is not covered here.

Identifying and Estimating a Structural Equation

- OLS is biased because an explanatory variable is endogenous.

- We will see that the two stage least squares (2SLS) can be applied to SEM's.
  - If we have an instrumental variable we can still obtain consistent estimates.

- We will also work out identification conditions for SEM's.

Identification

- Simple demand supply example for illustration.

$$
\begin{aligned}
q &= \alpha_1 p + \beta_1 z_1 + u_1 \\
q &= \alpha_2 p + u_2
\end{aligned}
$$

with $q_d = q_s = q$ is per capita milk consumption at the county level.

p is the average price of a pint of milk in the county.

z_1 is the price of cattle feed.

  - The first equation is therefore the supply equation (why?) and we assume $\beta_1 < 0$.
  - Given a random sample on (q, p, z_1) which of the above two equations can be estimated? Which is an identified equation?
  - It turns out, the demand equation (the second). Why?

- As in IV estimation, we use z_1 as an instrument for the estimation of the demand equation.

- Since there is no exogenous independent variable in the demand equation, we cannot trace out the supply equation, which is not identified in this example.

- If there were observed exogenous demand shifters in the demand equation, then we could also identify the supply equation.

- The presence of an exogenous variable in the supply equation allows us to estimate the demand equation.

- More generally,

$$\begin{aligned} y_1 &= \beta_{10} + \alpha_1 y_2 + z_1' \boldsymbol{\beta_1} + u_1 \\ y_2 &= \beta_{20} + \alpha_2 y_1 + z_2' \boldsymbol{\beta_2} + u_2 \end{aligned}$$

where z_1 and β_1 are $(k_1 \times 1)$ and z_2 and β_2 are $(k_2 \times 1)$.

- In many cases, the sets of exogenous variables z_1 and z_2 will overlap.
- We have to impose **exclusion restrictions**:
  - We assume that certain exogenous variables do not appear in the first equation and some do not appear in the second equation.
  - Then we can distinguish between the two structural equations.
- Moreover, when can we solve these two equations for y_1 and y_2? If $\alpha_1 \alpha_2 \neq 1$ (as shown earlier, sl.12).

- The question is now: when is this system identified?
  - There is a **rank condition** for identification.
- The first equation in a two equation SEM is identified if and only if the second equation contains at least one exogenous variable (with a nonzero coefficient) that is excluded from the first equation.
  - This is the necessary and sufficient condition for the first equation to be identified.
- In the second equation, we need at least one nonzero population coefficient on an exogenous variable omitted in the first model because it ensures that this variable actually appears in the reduced form of y_2.
- We can test this using a t or an F test.

Example: Labour supply of married Women

- Illustrate Identification
- Women already in the workforce.
- The two structural equations **labour supply** and **wage offer** are:

$$
\begin{aligned}
hours &= \alpha_1 log(wage) + \beta_{10} + \beta_{11} educ + \beta_{12} age + \beta_{13} kidslt6 + \beta_{14} nwifeinc + u_1 \\
log(wage) &= \alpha_2 hours + \beta_{20} + \beta_{21} educ + \beta_{22} exper + \beta_{23} exper^2 + u_2
\end{aligned}
$$

With *kidslt6*: number of kids under 6, *nwifeinc*: woman's nonwage income.

- All variables excepts *log(wage)* and *hours* are assumed to be exogenous. Then $y_1 = hours$ and $y_2 = log(wage)$.
- The supply equation satisfies the **order condition** as there are two exclusion restrictions. Which?
- The **rank conditions** required for the identification of the first equation also requires that $\beta_{22} \neq 0$ or $\beta_{23} \neq 0$.

- We can state the rank condition for identification equivalently in terms of the reduced form for log(wage):

$$\begin{aligned} log(wage) &= \pi_{20} + \pi_{21} educ + \pi_{22} age + \pi_{23} kidslt6 \\ &\quad + \pi_{24} nwifeinc + \pi_{25} exper + \pi_{26} exper^2 + \nu_2 \end{aligned}$$

  We need: $\pi_{25} \neq 0$ or $\pi_{26} \neq 0$ something that we can test using a standard F test.

- The wage offer equation is identified if at least one of *age*, *kidslt6*, or *nwifeinc* have a nonzero coefficient in the reduced form for hours:

$$\begin{aligned} hours &= \pi_{10} + \pi_{11} educ + \pi_{12} age + \pi_{13} kidslt6 \\ &\quad + \pi_{14} nwifeinc + \pi_{15} exper + \pi_{16} exper^2 + \nu_1 \end{aligned}$$

Estimation by 2SLS

- When we have determined that an equation is identified, we can estimate it by two stage least squares.
- The instrumental variables consist of the exogenous variables appearing in either equation.
- Example: (cont.)
  - The full set of instruments includes *educ, age, kidslt6, nwifeinc, exper, exper^2*.
  - Data: MROZ.dta
  - Estimate the labour supply and wage offer equation by 2SLS.
  - The latter yields:

$$
\begin{aligned}
\widehat{log(wage)} \quad = \quad & -0.656 + 0.00013 hours + 0.110 educ \\
& (0.338) \quad (0.00025) \qquad\qquad (0.016) \\
& +0.035 exper - 0.00071 exper^2, n = 428 \\
& (0.019) \qquad\quad (0.00045)
\end{aligned}
$$

- Compared to previous models, *hours* is now included:

$$
\begin{aligned}
\widehat{log(wage)} \quad &= \quad -0.656 + 0.00013 hours + 0.110 educ \\
&\quad\quad (0.338) \quad (0.00025) \qquad\qquad (0.016) \\
&\quad\quad +0.035 exper - 0.00071 exper^2, n = 428 \\
&\quad\quad (0.019) \qquad\quad (0.00045)
\end{aligned}
$$

- The coefficient on *hours* is practically zero and statistically insignificant. Other coefficients are similar as before.
- This suggest, one may better drop *hours* and estimate the model by OLS.

Systems with more than two equations

- It is difficult to derive general identification of this model (extension of the rank conditions).
- If it is identified, it can be estimated by 2SLS.
- Consider three equations without intercepts:

$$\begin{aligned} y_1 &= \alpha_{12} y_2 + \alpha_{13} y_3 + \beta_{11} z_1 + u_1 \\ y_2 &= \alpha_{21} y_1 + \beta_{21} z_1 + \beta_{22} z_2 + \beta_{23} z_3 + u_2 \\ y_3 &= \alpha_{32} y_2 + \beta_{31} z_1 + \beta_{32} z_2 + \beta_{33} z_3 + \beta_{34} z_4 + u_3 \end{aligned}$$

  - Which of these equations can be estimated?
  - The third equation is not identified because every exogenous variable appears in this equation.
  - The first equation has two endogenous variables and three excluded exogenous variables: it passes the order condition. (we have three potential IVs for two endogenous variables).

- Order condition for Identification: An equation in any SEM 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 sided endogenous variables.

- In applications, simply check the order conditions as it is often difficult to show identification (rank conditions).

- An equation is **overidentified** if there are more IVs than endogenous right-hand side variables.
- An equation is **just identified** if the number of IVs equals the endogenous right-hand side variables.
- Otherwise an equation is **unidentified**.

- Each identified equation in a SEM can be estimated by 2SLS.
- The IVs for a particular equation consist of the exogenous variables anywhere in the system.
- See IV estimation for tests for endogeneity, heteroscedasticity etc.

- System estimation methods are generally more efficient than estimating each equation by 2SLS. The most common system estimation method in this context is three stage least squares (not presented).

Summary

- SEM are appropriate if each equation in the system has a ceteris paribus interpretation on its own.
- By fully specifying the system, it is clear which variables are exogenous and which are endogenous.
- Check identification by verifying the order condition and the rank condition for each equation.
- OLS is biased and inconsistent. Use 2SLS.
- Systems with more than two equations are more complex (Identification).