Week 9: Simultaneous Equation Models and Miscellaneous Topics

Advanced Econometrics 4EK608

Vysoká škola ekonomická v Praze

Outline

1. Introduction
2. Simultaneity Bias
3. Identification problem
4. Identification conditions
5. Systems with more than two equations
6. Miscellaneous topics
   - Alternative approaches to econometric modeling
   - Monte Carlo studies
   - Data mining

Introduction

Simultaneity is another important form of endogeneity

Simultaneity occurs if at least two variables are jointly determined. A typical case is when observed outcomes are the result of separate behavioral mechanisms that are coordinated in an equilibrium.

Prototypical case: a system of demand and supply equations:

- $D(p)$ how high *would* demand be if the price was set to p?
- $S(p)$ how high *would* supply be if the price was set to p?

- Both mechanisms have a ceteris paribus interpretation.
- Observed quantity and price will be determined in equilibrium, where $D(p) = S(p)$.

Simultaneous equations systems can be estimated by 2SLS/IVR … Identification conditions apply.

Examples

Example 1: Labor supply and demand in agriculture

$$h_s = \alpha_1 w + \beta_1 z_1 + u_1$$
$$h_d = \alpha_2 w + \beta_2 z_2 + u_2$$

- Endogenous variables, exogenous variables, observed and unobserved supply shifter, observed and unobserved demand shifter
- We have n regions, market sets equilibrium price and quantity in each. We observe the equilibrium values only

$$h_{is} = h_{id} \Rightarrow (h_i, w_i)$$

Examples

Example 1: Labor supply and demand in agriculture contnd.

$$h_i = \alpha_1 w_i + \beta_1 z_{i1} + u_{i1}$$
$$h_i = \alpha_2 w_i + \beta_2 z_{i2} + u_{i2}$$

- If we have the same exogenous variables in each equation, we cannot identify (distinguish) equations.
- We assume independence between errors in structural equations & exogenous regressors.

Examples

Example 1: Labor supply and demand in agriculture contnd.

If we estimate the structural equation with OLS method, estimators will be biased – so called “simultaneity bias”.

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

$$y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2$$

y_2 is dependent on u_1
(substitute RHS of the 1^{st} equation for y_1 in the 2^{nd} eq.)

$$\Rightarrow y_2 = \left[\frac{\alpha_2\beta_1}{1-\alpha_2\alpha_1}\right] z_1 + \left[\frac{\beta_2}{1-\alpha_2\alpha_1}\right] z_2 + \left[\frac{\alpha_2 u_1 + u_2}{1-\alpha_2\alpha_1}\right]$$

Structural and reduced form equations, 2SLS method

Structural equations (example)

$$y_1 = \beta_{10} + \beta_{11} y_2 + \beta_{12} z_1 + u_1$$

$$y_2 = \beta_{20} + \beta_{21} y_1 + \beta_{22} z_2 + u_2$$

Reduced form equations

$$y_1 = \pi_{10} + \pi_{11} z_1 + \pi_{12} z_2 + \varepsilon_1 \quad \Rightarrow \quad \hat{y}_1 \text{ by OLS}$$

$$y_2 = \pi_{20} + \pi_{21} z_1 + \pi_{22} z_2 + \varepsilon_2 \quad \Rightarrow \quad \hat{y}_2 \text{ by OLS}$$

2SLS (a special case of IVR)

- 1^{st} stage: Estimate reduced forms, get $\hat{y}_1$ and $\hat{y}_2$.
- 2^{nd} stage: Replace endogenous regressors in structural equations by fitted values from 1^{st} stage, estimate by OLS.
- ... Identification of structural equations,
  ... Statistical inference in structural equations (2^{nd} stage).

Examples

Example 2: (Structural equations)

Estimation of murder rates

$$
\begin{aligned}
murdpc &= \alpha_1 polpc + \beta_{10} + \beta_{11} incpc + u_1 \\
polpc &= \alpha_2 murdpc + \beta_{20} + \boldsymbol{\beta}(other\ factors) + u_2
\end{aligned}
$$

- 1^{st} equation describes the behaviour of murderers, 2^{nd} one the behaviour of municipalities. Each one has its ceteris paribus interpretation.
- For the municipality policy, the 1^{st} equation is interesting: what is the impact of exogenous increase of police force on the murder rate?
- However, the number of police officers is not exogenous (simultaneity problem).

Identification problem

Example 3: (Identification)

Identification problem in a SEM

- Example: Supply and demand for milk

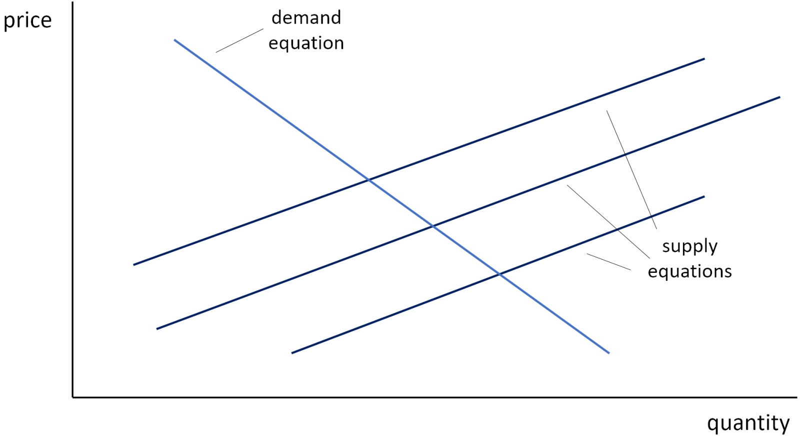
  Supply of milk: $q = \alpha_1 p + \beta_1 z_1 + u_1$
  Demand for milk: $q = \alpha_2 p + u_2$

- Supply of milk cannot be consistently estimated because we do not have (at least) one exogenous variable "available" to be used as instrument for p in the supply equation.

- Demand for milk can be consistently estimated because we can use exogenous variable z_1 as instrument for p in the demand equation.

Identification problem

- Ilustration



Identification conditions

Identification conditions for a sample 2-equation SEM (individual i subscripts omitted)

$$y_1 = \beta_{10} + \alpha_1 y_2 + \beta_{11} z_{11} + \beta_{12} z_{12} + \cdots + \beta_{1k} z_{1k} + u_1$$
$$y_2 = \beta_{20} + \alpha_2 y_1 + \beta_{21} z_{21} + \beta_{22} z_{22} + \cdots + \beta_{2k} z_{2k} + u_2$$

- Order condition (necessary): 1^{st} equation is identified if at least one exogenous variable z is excluded from 1^{st} equation (yet in the SEM) .
- Rank condition (necessary and sufficient): 1^{st} equation is identified if and only if the second equation includes at least one exogenous variable excluded from the first equation with a nonzero coefficient, so that it actually appears in the reduced form.
- For the second equation, the conditions are analogous.

Examples

Example 4: (Identification)

Labor supply of married working women

Supply:

$$\begin{aligned} hours = \alpha_1 \log(wage) + \beta_{10} + \beta_{11} educ + \beta_{12} age + \beta_{13} kidslt6 \\ + \beta_{14} nwifeinc + u_1 \end{aligned}$$

Demand:

$$\log(wage) = \alpha_2 hours + \beta_{20} + \beta_{21} educ + \beta_{22} exper + \beta_{23} exper^2 + u_2$$

Order condition is fulfilled in both equations.

Examples

Example 4: (Identification)
Labor supply of married working women contnd.

- To evaluate the rank condition for supply equation, we estimate the reduced form for $\log(wage)$ and test if we can reject the null hypothesis that coefficients for both coefficients for $exper$ and $exper^2$ are zero.
  If H_0 is rejected, the rank condition is fulfilled.
- We would do the evaluation of the rank condition for the demand equation analogically.

Estimation

- We can consistently estimate identified equations with the 2SLS method.
- In the 1^{st} stage, we regress each endogenous variable on all exogenous variables (“reduced forms”).
- In the 2^{nd} stage we put into the structural equations instead of endogenous variables their predictions from the 1^{st} stage and estimate with the OLS method.
- The reduced form can be always estimated (by OLS).
- In the 2^{nd} stage, we cannot estimate unidentified structural equations.
- With some additional assumptions, we can use a more efficient estimation method than 2SLS: 3SLS.

Systems with more than two equations

Example 5: Keynesian macroeconomic model

$$
\begin{aligned}
C_t &= \beta_0 + \beta_1 (Y_t - T_t) + \beta_2 r_t + u_{t1} \\
I_t &= \gamma_0 + \gamma_1 r_t + u_{t2} \\
Y_t &\equiv C_t + I_t + G_t
\end{aligned}
$$

Endogenous: C_t, I_t, Y_t

Exogenous: T_t, G_t, r_t

- Order condition for identification is the same as for two equations systems, rank condition is more complicated.
- There exist complicated models based on macroeconomic time series. There is a lot of problems with these models: series are usually not weakly dependent, it is difficult to find enough exogenous variables as instruments. Question is, if any macroeconomic variables are exogenous at all.

Identification in SEMs with more than two equations

$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{u}_i$ is the i-th equation of a SEM.

K - # of exogenous/predetermined variables in the SEM,
K_i - # of K in the i-th equation,
G_i - # of endogenous variables in the i-th equation.

Order condition for the i-th equation:
necessary, not sufficient condition for identification

$$K - K_i \geq G_i - 1$$

Condition evaluates as:

- $=$ Equation i is just-identified,
- $>$ Equation i is over-identified,
- $<$ Equation i is not identified,
  structural equation i cannot be estimated by 2SLS/IVR.

Identification in SEMs with more than two equations

Rank condition: based on matrix algebra & IV estimator

We begin explanation of rank condition as follows:

consider IVR for a just-identified i-th equation of SEM

$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{u}_i$

$\boldsymbol{X}_i$ is a $(n \times k)$ matrix, includes the intercept column,

$\boldsymbol{W}$ is a $(n \times k)$ matrix, includes the intercept column,
(endogenous regressors are replaced by the same number of IVs).

OLS $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}_i' \boldsymbol{X}_i)^{-1} \boldsymbol{X}_i' \boldsymbol{y}$

IVR $\hat{\boldsymbol{\beta}}_{IVR} = (\boldsymbol{W}' \boldsymbol{X}_i)^{-1} \boldsymbol{W}' \boldsymbol{y}$

MM $\boldsymbol{W}' \left(\boldsymbol{y} - \boldsymbol{X}_i \hat{\boldsymbol{\beta}} \right) = \boldsymbol{0}$ (moment conditions)

Identification in SEMs with more than two equations

Rank condition: based on matrix algebra & IV estimator (cont.)

- For over-identified SEM equations, $\hat{\boldsymbol{\beta}}_{IVR} = (\boldsymbol{W}'\boldsymbol{X}_i)^{-1}\boldsymbol{W}'\boldsymbol{y}$ cannot be calculated as dimensions of $\boldsymbol{W}$, $\boldsymbol{X}_i$ are not compatible with calculating the inverse $(\boldsymbol{W}'\boldsymbol{X}_i)^{-1}$.

- **Generalized IV estimator** (GIVE), based on a $(n \times n)$ projection matrix $\boldsymbol{P}_W$:

  $$\boldsymbol{P}_W = \boldsymbol{W}\left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1}\boldsymbol{W}'$$

GIVE $\hat{\boldsymbol{\beta}}_{GIVE} = (\boldsymbol{X}_i'\boldsymbol{P}_W\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\boldsymbol{P}_W\,\boldsymbol{y}$

MM $\boldsymbol{X}_i'\boldsymbol{P}_W\left(\boldsymbol{y} - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}\right) = \boldsymbol{0}$ (moment conditions)

Identification in SEMs with more than two equations

Rank condition: based on matrix algebra & IV estimator (cont.)

$\hat{\beta}_{GIVE} = (\boldsymbol{X}_i' \boldsymbol{P}_W \boldsymbol{X}_i)^{-1} \boldsymbol{X}_i' \boldsymbol{P}_W \, \boldsymbol{y}$

- **Order condition**: The necessary condition for the i-th equation to be identified (full rank of $\boldsymbol{P}_W \boldsymbol{X}_i$) is that the number instruments contained in $\boldsymbol{W}$ should be no less than the number of explanatory variables in $\boldsymbol{X}_i$.
- **Rank condition**: The necessary and sufficient condition for identification of the i-th equation is that $\boldsymbol{P}_W \boldsymbol{X}_i$ should have full column rank of $\boldsymbol{X}_i$.
  …ensures non-singularity of $\boldsymbol{X}_i' \boldsymbol{P}_W \boldsymbol{X}_i$,
  …and the existence of $(\boldsymbol{X}_i' \boldsymbol{P}_W \boldsymbol{X}_i)^{-1}$.

Identification in SEMs with more than two equations

Identification: recap & final remarks

- Reduced form equations can always be estimated.
- Structural equations can be estimated (IV/2SLS) only if identified: i.e. if rank condition is met.

- Checking rank condition for $\boldsymbol{P}_W \boldsymbol{X}_i$ is easy for any given (finite) dataset.
- Asymptotic identification may be "tricky": because some columns in $\boldsymbol{X}_i$ are endogenous,

  $\operatorname{plim} n^{-1} \boldsymbol{X}_i' \boldsymbol{P}_W \boldsymbol{X}_i$

  depends on the parameters of the DGP.

  ...see Davidson-MacKinnon (2009) Econometric theory and methods

Miscellaneous topics

Miscellaneous topics
not specifically related to SEMs

- Simple-to-general approach to econometric modeling
- General-to-specific approach to econometric modeling

- Monte Carlo studies

- Data mining

Alternative approaches to econometric modeling

Simple-to-general approach

- Traditional approach to econometric modeling
- Starts with formulation of the simplest model consistent with the relevant economic theory.
- If this initial model proves unsatisfactory, it is improved in some way – adding or changing variables, using different estimators etc.

Alternative approaches to econometric modeling

Criticism of the simple-to-general approach

- Revisions to the simple model are carried out arbitrarily and simply reflect investigator's prior beliefs: danger of always finding what you want to find.
- It is open to accusation of data mining: researchers usually presents just the final model (true significance level is problematic).

Alternative approaches to econometric modeling

General-to-specific approach

- Professor Hendry, London School of Economics started this approach in the 80ies.
- It starts with formulation of a very general and maybe quite complicated model.
- Starting model contains a series of simpler models, nested within it as special cases.
- These simpler models should represent all the alternative economic hypotheses that require consideration.

Alternative approaches to econometric modeling

General-to-specific approach

- General model must be able to explain existing data and be able to satisfy various tests of misspecification.
- What follows is simplification search (testing-down procedure). Through parameter restrictions, we test nested models against the containing model. If the nested model does not pass the tests, we can reject the whole branch of sub-nested models.
- If we find more non-nested models satisfying tests, we can compare them using e.g. F-test.

Alternative approaches to econometric modeling

Advantages of the general-to-specific approach

- "Data mining" present in this approach is transparent (for all to see) and it is carried out in a systematic manner that avoids worst data mining problems.
- Researcher usually reports both the initial general model and all steps involved so it is possible to get some idea about the true significance levels.
- Supporters of this approach stress the importance of both testing final models against new data and the ability of the model to provide adequate out-of-sample forecasts.

Monte Carlo studies

Simulation exercises designed to shed light on small-sample properties for a given estimation problem.

For many estimators, small-sample properties cannot (are hard to) be derived analytically.

Monte Carlo studies (in 4 steps):

1. Model the data generating process
2. Generate many sets of artificial data
3. Use the data and estimator to create repeated estimates
4. Use these estimates to gauge the sampling distribution properties (predictive properties . . .) of that estimator.

Data mining

We “torture” the data until we find some statistically significant relationship. It can be completely misleading – as following example shows.

Repetition:

t test: $H_0 : \beta_j = 0 \quad H_1 : \beta_j \neq 0$

Significance level:
probability of a type I error, i.e. probability of rejecting H_0 when it is in fact true, i.e. finding a regressor significant when -in fact- it does not influence the dependent variable.

Data mining

Example:

1. Suppose we have 20 "possible" regressors $x_1, x_2, \ldots, x_{20}$, but all are factually unrelated to the dependent variable y
2. Suppose we have computed 20 simple regressions of the form
$$\hat{y} = \hat{\beta}_{0p} + \hat{\beta}_{1p} x_p$$
3. If we use significance level 0.05, we can expect one of the 20 regressors to appear significant just by chance, even if none of them actually influences y.

Data mining

$$\Pr(X_1 \text{ appears significant by chance}) = 0.05$$
$$\Pr(X_1 \text{ does not appear significant}) = 0.95$$

$$\Pr(X_2 \text{ does not appear significant}) = 0.95$$

$$\Pr(\text{neither } X_1 \text{ nor } X_2 \text{ appear significant}) = 0.95 \times 0.95 = = 0.9025$$

$$\Pr(\text{at least one of } X_1, X_2 \text{ appear significant}) = 1 - 0.9025 = = 0.0975$$

true significant level
$\alpha^* = (1 - (1 - \alpha)^2)$

For c independent candidates: $\alpha^* = 1 - (1 - \alpha)^c$
If we want the true significance level to be 0.05, we must solve the equation $0.05 = 1 - (1 - \alpha)^c$ and do all t-tests on significance level α.

Data mining

Lovell (1983): rule of thumb for finding the true significance level in the case where k regressors are selected from c possible candidates.

$$\alpha^* = 1 - (1 - \alpha)^{\frac{c}{k}}$$