AFRE 835: Introductory Econometrics

Chapter 16: Simultaneous Equations Models

Spring 2017

Introduction

- We have looked into two key sources of *endogeneity* thus far: omitted variables and measurement error.
- This chapter focuses on another important source: *simultaneity*.

  "This arises when one or more of the explanatory variables is *jointly determined* with the dependent variable, typically through an equilibrium mechanism" (Wooldridge, p. 554).
- The interaction of supply and demand for a commodity is a good examples of this, where price and quantity are jointly determined.
- We will be looking at basic techniques for estimating simple simultaneous equation models (SEM's) drawing on IV techniques.
- We will not be covering either sections 16.5 (SEM's) with Time Series) or 16.6 (SEM's with Panel Data).
- We will, however, take a brief detour to discuss seemingly unrelated regression (SUR) models, not covered in Wooldridge.

Outline

1. Seemingly Unrelated Regression Models
2. Simultaneous Equations Models
   - The Nature of SEM's
   - Simultaneity Bias in OLS
   - Identification and Estimation a Structural Equation
   - Systems with More Than Two Equations

Seemingly Unrelated Regression Models

- This chapter focuses on *Simultaneous Equation Models*. (SEM's).
- These fall into the broader class: *systems of equations*.
- Another sub-class of models are *Seemingly Unrelated Regression (SUR)* models.
- The defining feature of SUR models is that one has a set of two or more equations, with exogenous regressors but potentially correlated errors.
- Consider, for example, a system of equations modeling the demand for beef (y_{1i}) and chicken (y_{2i}):

$$y_{1i} = \beta_{10} + \beta_{11}x_{11i} + \cdots + \beta_{1k_1}x_{1k_1i} + u_{1i} \quad (1)$$

$$y_{2i} = \beta_{20} + \beta_{21}x_{21i} + \cdots + \beta_{2k_2}x_{2k_2i} + u_{2i} \quad (2)$$

where $Cov(u_{1i}, u_{2i})$ need not be zero and $\boldsymbol{x}_{1i}$ and $\boldsymbol{x}_{2i}$ need not be the same or even have the same dimension.

Estimation via OLS

- One way to proceed would be to estimate the model separately for each equation.
  - The advantage of this approach is simplicity - OLS remains unbiased and consistent.
  - However, OLS is inefficient.
  - In many settings, there will be cross-equation constraints, making separate estimation impossible without combining the equations in some way.
- The problem of cross-equation constraints can be addressed by using **System OLS** - essentially stacking the two (or more) equations and applying OLS (See Wooldridge, 2010, Section 7.3).
  - Systems OLS is still inefficient and the estimated standard errors are biased.
  - However, one can obtain robust standard errors, allowing for both general cross-equation correlation and differences between $Var(u_1)$ and $Var(u_2)$.

SUR in Matrix Form

- We can write the SUR model in matrix form as

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix}$$

$$\text{or} \qquad \boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{u} \tag{3}$$

with

$$\boldsymbol{\Omega} = Cov(\boldsymbol{u}) = E(\boldsymbol{u}\boldsymbol{u}') = \begin{bmatrix} \sigma_1^2 \boldsymbol{I} & \sigma_{12}\boldsymbol{I} \\ \sigma_{12}\boldsymbol{I} & \sigma_1^1 \boldsymbol{I} \end{bmatrix} = \boldsymbol{\Sigma} \otimes \boldsymbol{I} \tag{4}$$

where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_1^1 \end{bmatrix} \tag{5}$$

- The model in (4) violates the standard model assumptions, with cross-observation correlation and heteroskedasticity.
- As a result, OLS will be inefficient and the traditional standard errors will be biased.

GLS

- If we knew Σ, and hence Ω, we could use a GLS-style transformation of the model that would satisfy the Gauss-Markov assumptions.
- Specifically, we need a nonsingular matrix $\boldsymbol{P}$ such that $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{\Omega}^{-1}$
- If we pre-multiply our model in (4) by $\boldsymbol{P}$ we get

$$\boldsymbol{Py} = \boldsymbol{PX}\beta + \boldsymbol{Pu}$$

or

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\beta + \tilde{\boldsymbol{u}}. \tag{6}$$

  where $\tilde{\boldsymbol{y}} = \boldsymbol{Py}$, $\tilde{\boldsymbol{X}} = \boldsymbol{PX}$, and $\tilde{\boldsymbol{u}} = \boldsymbol{Pu}$.
- The model in (6) satisfies the Gauss-Markov Theorem assumptions, since $Var(\tilde{\boldsymbol{u}}|\boldsymbol{X}) = Var(\boldsymbol{Pu}|\boldsymbol{X}) = \boldsymbol{P}E(\boldsymbol{uu}'|\boldsymbol{X})\boldsymbol{P}' = \sigma^2\boldsymbol{P\Omega P}' = \sigma^2\boldsymbol{P}(\boldsymbol{P}'\boldsymbol{P})^{-1}\boldsymbol{P}' = \sigma^2\boldsymbol{I}_n$.
- Constructing P in the case of the SUR model is straightforward, with $\boldsymbol{P} = \boldsymbol{R} \otimes \boldsymbol{I}$, where $\boldsymbol{R}'\boldsymbol{R} = \boldsymbol{\Sigma}^{-1}$.

Feasible GLS (FGLS)

- Feasible GLS involves first estimating the model via OLS and using the fitted residuals to estimate Σ.
- In the second stage, we transform the data and estimate the transformed model via OLS.
- If the model specification is correct, then FGLS will be asymptotically more efficient than OLS.
- There is no gain from GLS (or FGLS) if all of the explanatory variables are the same in every equation (i.e., $\boldsymbol{X}_1 = \boldsymbol{X}_2$).
- The command in stata:
  sureg (y1 x11 x12 x13) (y2 x21 x22)

The Nature of SEM's

- SUR models are not particularly problematic, as traditional OLS techniques yield consistent estimators.
- Moreover, robust standard errors can be used to adjust the standard errors and FGLS provide a potential improvement in efficiency.
- Simultaneous Equation Models (SEM's) cause more serious problems in that they induce endogeneities that, if not controlled for, will result in inconsistent parameter estimates.

Labor Market Example - Supply Side

- Wooldridge uses a simple agricultural labor market example to illustrate the problem.
- Let h_s denote the supply of labor in hours, with a supply function given by
$$h_s = \alpha_1 w + \beta_1 z_1 + u_1 \tag{7}$$
where, for simplicity, the intercept is assumed to be zero and
  - w denotes the wage rate;
  - z_1 denotes a *supply function shifter* (e.g., the wage rate in the manufacturing sector);
  - u_1 denotes the error term (*unobserved supply shifter*) with $E(u_1|z_1) = 0$.
- Equation (7) is the **structural equation** for labor supply.

Labor Market Example - Demand Side

- Distinct from the supply equation, there will be a demand for labor by the agricultural sector.
- Let h_d denote the demand of labor in hours, with a demand function given by

$$h_d = \alpha_2 w + \beta_2 z_2 + u_2 \tag{8}$$

where again, for simplicity, the intercept is assumed to be zero and

z_2 denotes a *demand function shifter* (e.g., the quantity of agricultural land);

u_2 denotes the error term (*unobserved demand shifter*) with $E(u_2|z_2) = 0$.

- Equation (8) is the **structural equation** for labor demand.

The Equilibrium

- Observed labor hours (h) and wage rate (w) are the equilibrium outcome of the interaction between supply and demand; i.e., $h = h_d = h_s$ and w solve (7) and(8) and are determined *simultaneously*.
- h and w are both *endogenous* variables.
- It is important to note that the supply and demand equations represent two distinct concepts and the actions of two distinct agents; i.e., workers and employers.
- Wooldridge gives two other examples
  - Appropriate SEM: Modeling the equilibrium outcome of murders per capita and police force per capita:
    1. Murders per capita as a function of police officers per capita (decided by murders) and
    2. Police officers per capita as a function of murders per capita (decided by city officials).
  - Inappropriate SEM: Housing and savings decided by the same agent.

The Reduced Form of a SEM

- To see the bias in OLS, consider a generic 2 equation SEM:

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1 \tag{9}$$
$$y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2 \tag{10}$$

- If we substitute (10) into (9) (and $\alpha_1\alpha_2 \neq 1$), we get

$$\begin{aligned} y_1 &= \alpha_1(\alpha_2 y_1 + \beta_2 z_2 + u_2) + \beta_1 z_1 + u_1 \\ \Rightarrow (1-\alpha_1\alpha_2)y_1 &= \alpha_1\beta_2 z_2 + \beta_1 z_1 + u_1 + \alpha_1 u_2 \\ \Rightarrow y_1 &= \frac{\beta_1}{(1-\alpha_1\alpha_2)} z_1 + \frac{\alpha_1\beta_2}{(1-\alpha_1\alpha_2)} z_2 + \frac{u_1 + \alpha_1 u_2}{(1-\alpha_1\alpha_2)} \\ &= \pi_{11} z_1 + \pi_{12} z_2 + v_1 \end{aligned} \tag{11}$$

- Equation (11) is known as the *reduced form equation* for y_1
- Clearly y_1 is correlated with both u_1 and u_2 $\Rightarrow$ OLS of equation (10) is biased. A similar result applies for y_2 and (9)

Indirect Least Squares

- One approach to estimating the structural parameters α_j and β_j is to estimate the reduced form parameters and solve back for the structural counterparts. This is sometimes referred to as **Indirect LS**.
- From the reduced form expressions for y_1 and y_2 we have:

$$\pi_{11} = \frac{\beta_1}{(1-\alpha_1\alpha_2)} \qquad \pi_{12} = \frac{\alpha_1\beta_2}{(1-\alpha_1\alpha_2)}$$
$$\pi_{21} = \frac{\alpha_2\beta_1}{(1-\alpha_1\alpha_2)} \qquad \pi_{22} = \frac{\beta_2}{(1-\alpha_1\alpha_2)}$$

We can the solve for our structural parameters using:

$$\alpha_1 = \frac{\pi_{12}}{\pi_{22}} \qquad \alpha_2 = \frac{\pi_{21}}{\pi_{11}}$$
$$\beta_1 = \pi_{11}(1-\alpha_1\alpha_2) \qquad \beta_2 = \pi_{22}(1-\alpha_1\alpha_2)$$

- OLS applied to reduced form equations yield consistent estimates of the π's because the z_j are exogenous in the structural equations.

Identification

- The Indirect Least Squares (ILS) approach suggests the source of *identification* in the SEM; i.e., what it is that allows us to estimate various parameter in our model.
- Notice what happens if we remove one of our exogenous variables from our structural equations.
- Specifically, suppose that $\beta_2 = 0$, leaving us with the structural equations:

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1 \tag{12}$$

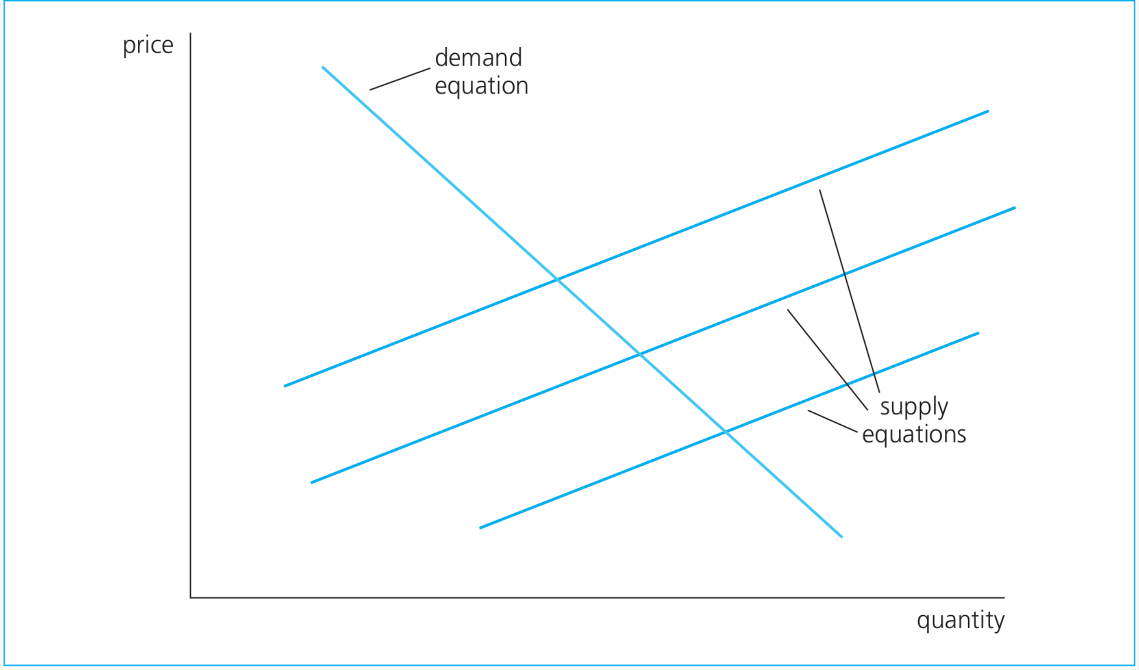
$$y_2 = \alpha_2 y_1 + u_2 \tag{13}$$

- The question is: which of our parameter can still be estimated?
- Considering the ILS equations on the previous slide, we now have $\pi_{12} = \pi_{22} = 0$, which precludes us from estimating α_1 and β_1, but we can still estimate $\alpha_2 = \pi_{21}/\pi_{11}$.

Supply and Demand Equation Context

- In the supply and demand equation context,
  - Suppose the equation without the observed exogenous shifter is demand (i.e., y_2), while the equation with the observed exogenous shifter is supply (i.e., y_1)
  - Identification of demand relies on having a source of *exogenous variation* in supply that doesn't also shift demand.
  - This allows us to trace out the demand curve.
- Another way of viewing it is that there is a valid instrument for the endogenous variable y_1 in equation (13) (i.e., z_1).
- There is, however, not a valid instrument for the endogenous variable y_2 in equation (12).

FIGURE 16.1 Shifting supply equations trace out the demand equation. Each supply equation is drawn for a different value of the exogenous variable, z_1.



Extending to the Case with More Regressors

- Consider the more general case, where

$$
\begin{aligned}
y_1 &= \beta_{10} + \alpha_1 y_2 + \beta_{11} z_{11} + \cdots + \beta_{1k_1} z_{1k_1} + u_1 \\
y_2 &= \beta_{20} + \alpha_2 y_1 + \beta_{21} z_{21} + \cdots + \beta_{2k_2} z_{2k_2} + u_2 \\
&\Rightarrow \\
y_1 &= \beta_{10} + \alpha_1 y_2 + \mathbf{z}_1 \boldsymbol{\beta_1} + u_1 \\
y_2 &= \beta_{20} + \alpha_2 y_1 + \mathbf{z}_2 \boldsymbol{\beta_2} + u_2
\end{aligned}
$$

where

$\mathbf{z}_j = (z_{j1}, \cdots, \ldots, z_{jk_1})$ $(j = 1, 2)$ is a row vector of k_j exogenous regressors for equation j.

and $\boldsymbol{\beta}_j = (\beta_{j1}, \cdots, \ldots, \boldsymbol{\beta}_{jk_1})'$ $(j = 1, 2)$ is the corresponding column vector of parameters associated with these regressors in equation j.

Extending to the Case with More Regressors (cont'd)

- **Order condition**: A necessary condition for the first equation to be identified is that at least one of all exogenous variables is excluded from this equation.
- **Rank condition**: 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.

  ... These are referred to as **exclusion restrictions**.

Example 16.3: Labor Supply of Married, Working Women

- This is using the MROZ.dta data set.
- The SEM structural equations are:

$$\begin{aligned} hours &= \beta_{10} + \alpha_1 log(wage) + \beta_{11} educ + \beta_{12} age \\ &\quad + \beta_{13} kidslt6 + \beta_{14} nwifeinc + u_1 \qquad (14) \\ log(wage) &= \beta_{20} + \alpha_2 hours + \beta_{21} educ + \beta_{22} exper \\ &\quad + \beta_{23} exper^2 + u_2 \qquad (15) \end{aligned}$$

- Equation (14) satisfies the
  - order condition because it does not include *exper* or $exper^2$.
  - rank condition if equation (15) includes *exper* and $exper^2$ with at least one of β_{22} or β_{23} being nonzero.
- The rank condition requires at least one of the coefficients on *exper* and $exper^2$ in the reduced form expression for $log(wage)$ be nonzero.
- This is the *relevance* condition we saw earlier in discussing IV's.
- Similar restrictions are being used in identifying (15)

Estimation by 2SLS

- Estimation of the individual equations can proceed using the 2SLS estimator presented in the previous chapter.
- All we have really done new here is
  - Explicitly specify a model for our endogenous variable;
  - Included in this is the possible simultaneous determination of the two endogenous variables.
- Tests for endogeneity in any given equation can proceed exactly as it did in Chapter 15.

Example 16.5: Labor Supply of Married Working Women

	Stage 1 RHS		Stage 2 RHS	
	log(wage)	hours	hours	logwage
educ	0.1011	30.6	-183.8	0.11033
	(0.0141)**	(13.1)*	(67.8)**	(0.01482)**
age	-0.0026	-28.4	-7.8	
	(0.0059)	(3.9)**	(10.5)	
kidslt6	-0.0532	-432.9	-198.2	
	(0.1048)	(55.3)**	(208.4)	
nwifeinc	0.0056	-3.6	-10.2	
	(0.0027)*	(2.2)	(5.3)	
exper	0.0419	66.8		0.03458
	(0.0151)**	(10.8)**		(0.01851)
expersq	-0.0008	-0.7		-0.00071
	(0.0004)	(0.4)		(0.00043)
lwage			1,639.6	
			(593.3)**	
hours				0.00013
				(0.00029)
_cons	-0.4472	1,165.7	2,225.7	-0.65573
	(0.2889)	(249.8)**	(603.1)**	(0.40977)
R^2	[illegible]	[illegible]	.	[illegible]
N	428	753	428	428

Systems with More Than Two Equations

- With more than two equations, the identification requirements become more complicated, but the intuition is essentially the same;
- We need a source of independent variation for each of the endogenous variables;
- For a given equation, we need the number of *excluded* exogenous variables to be at least as large as the number of endogenous right-hand side variables.
- Equations can be **unidentified**, **just identified**, or **overidentified**.
- Simultaneous estimation of the system of equations can be more efficient, using, for example Three-Stage Least Squares (3SLS).