Chapter 16: Simultaneous Equations Models

Introductory Econometrics: A Modern Approach

Outline

- Simultaneous equations – definition
- Simultaneity bias
- 2SLS estimation for simultaneous equations • Testing for rank condition

Introduction

An important form of endogeneity of explanatory variables is **simultaneity**.

- This arises when one or more of the explanatory variables is **jointly determined** with the dependent variable
- The most frequent case is when the two variables are the outcome of a system in equilibrium.
- Examples:
  - supply and demand equations with price and quantity jointly determined
  - labor supply and demand with hours worked and wage jointly determined

Simultaneity is another form of endogeneity.

Solution: instrumental variable

16.1 The Nature of Simultaneous Equations Models

Example: Labor demand and supply in agriculture

labor supply elasticity

Observed supply shifters e.g. manufacturing wage

Annual labor hours supplied by workers in a given county if the average hourly wage offered to such workers is w

$$h_s = \alpha_1 w + \beta_1 z_1 + u_1$$

Unobserved supply shifters e.g. immigration flows

Annual hours demanded by employers in a given county if the average hourly wage paid to workers is w

$$h_d = \alpha_2 w + \beta_2 z_2 + u_2$$

Unobserved demand shifters e.g. food market shocks

labor demand elasticity

Observed demand shifters e.g. agricultural land area

Example: Labor demand and supply in agriculture

- Competition on the labor market in each county i will lead to a county wage w_i
- The total number of hours his supplied $h_i s$ by workers in this county equals the total number of hours $h_i d$ demanded by agricultural employers:

$$h_{is} = h_{id} \Rightarrow (h_i, w_i) \quad (= \text{observed equilibrium outcomes in each county})$$

Simultaneous equations model (SEM):

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

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

Endogenous variables

Exogenous variables

Structural error terms (uncorrelated with the exogenous variables)

Note: Without separate exogenous variables in each equation, the two equations could never be distiguished/separately identified

Example: Murder rates and the size of the police force

Murders per capita | Police officers per capita | Income per capita

"Behavioral equation" of murderer population $\rightarrow$

$$murdpc = \alpha_1\, polpc + \beta_{10} + \beta_{11} incpc + u_1$$

"Behavioral equation" of city government $\rightarrow$

$$polpc = \alpha_2 murdpc + \beta_{20} + other\ factors$$

- `polpc` will not be exogenous because the number of police officers will dependent on how high the murder rate is **("reverse causation")**.
- The interesting equation for policy purposes is the first one. City governments will want to know by how much the murder rate decreases if the number of police officers is exogenously increased.
- This will be hard to measure because the number of police officers is not exogenously chosen (it depends on how much crime there is in the city, see second equation).

Going Further 16.1

A standard model of advertising for monopolistic firms has firms choosing profit maximizing levels of price and advertising expenditures. Does this mean we should use an SEM to model these variables at the firm level?

- Probably not. Firms choose price and advertising expenditures jointly but we are not interested in the experiment where advertising changes exogenously and we want to know the effect on price.
- Instead, we would model price and advertising each as a function of demand and cost variables (economic theory).

16.2 Simultaneity Bias in OLS

Variable y_2 is correlated with the error u_1 because u_1 is indirectly a part of y_2. OLS applied to this equation will be therefore be inconsistent.

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

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

Insert the first equation into the second

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

Renaming the coefficients, the reduced form equation for y_2

$$y_2 = \pi_{21} z_1 + \pi_{22} z_2 + v_2$$

The structural equation will produce biased and inconsistent estimators, but the reduced form equation will produce unbiased and consistent estimators because z_1 and z_2 are not correlated with the new error term.

2SLS estimation for simultaneous equations

- The simultaneous equations can be consistently estimated by 2SLS (two stage least squares).
- In the first stage, the endogenous variable is regressed on the exogenous variables and instruments from the other equation.
- In the second stage, the endogenous variables are replaced by the predicted values from the first stage, and the equations are estimatedby OLS.

2SLS estimation for simultaneous equations

Structural equations $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$

$y_2 = \alpha_0 + \alpha_1 y_1 + \alpha_2 z_2 + u_2$

- z_2 is a good instrument for y_2 because z_2 is not in the y_1 equation and z_2 is related to y_2.
- z_1 is a good instrument for y_1 because z_1 is not in the y_2 equation and z_1 is related to y_1.

2SLS estimation for simultaneous equations

2SLS, first stage: reduced form, regress endogenous variables on all exogenous variables and get predicted values

First stage for eq 1: endogenous variable y_2 on instrument z_2 and other exogenous variable z_1

$$\hat{y}_2 = \hat{\delta}_0 + \hat{\delta}_1 z_1 + \hat{\delta}_2 z_2$$

First stage for eq 2: endogenous variable y_1 on instrument z_1 and other exogenous variable z_2

$$\hat{y}_1 = \hat{\gamma}_0 + \hat{\gamma}_1 z_1 + \hat{\gamma}_2 z_2$$

2SLS, second stage: estimate the equations by replacing the predicted values from first stage for the endogenous variables

$$y_1 = \beta_0 + \beta_1 \hat{y}_2 + \beta_2 z_1 + u_1$$

$$y_2 = \alpha_0 + \alpha_1 \hat{y}_1 + \alpha_2 z_2 + u_2$$

Identifying and Estimating a Structural Equation

OLS is biased and inconsistent when applied to a structural equation in a simultaneous equations system

- The mechanics of 2SLS are similar to those in Chapter 15.
- The difference is that, because we specify a structural equation for each endogenous variable, we can immediately see whether sufficient IVs are available to estimate either equation.

a) Identification in a Two-Equation System

Example: Supply and demand system

supply of milk $\longrightarrow q = \alpha_1 p + \beta_1 z_1 + u_1$

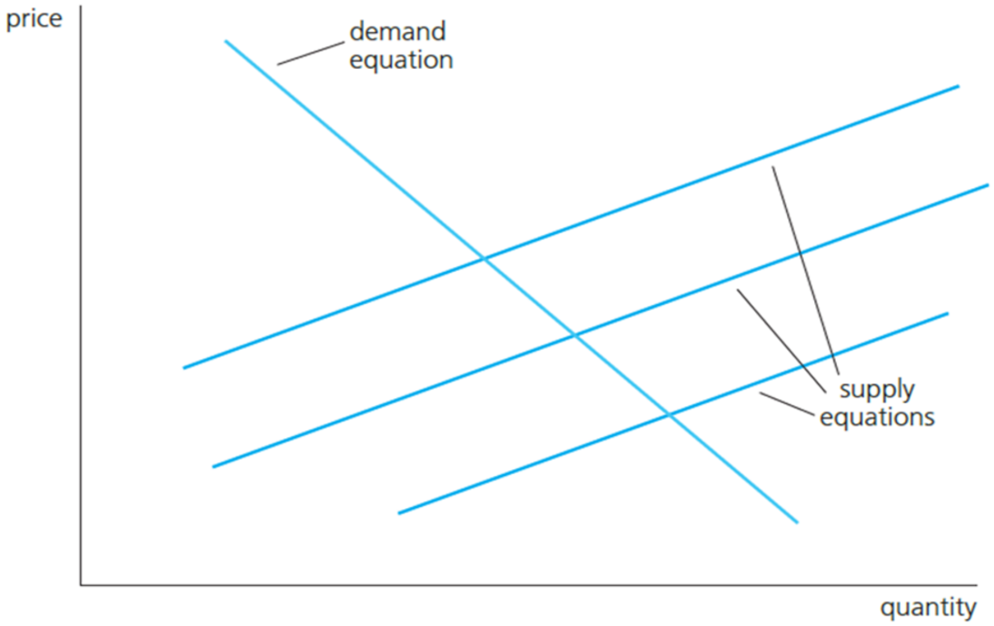
- where z_1 is the price of cattle feed

Demand for milk $\longrightarrow q = \alpha_2 p + u_2$

Which of the two equations is identified?

Supply $q = \alpha_1 p + \beta_1 z_1 + u_1$

Demand: $q = \alpha_2 p + u_2$



Which of the two equations is identified?

- The supply function cannot be consistently estimated because one of the regressors is endogenous and we do not have an instrument.
- The demand function can be consistently estimated because we can take z1 as an instrument for the endogenous price variable.

Example 16.3 Labor Supply for Married, Working Women

Labor supply and demand for working women.

- where `kidslt6` is number of kids under 6 years old, and `nwifeinc` is non-wife income

Labor supply of married, working women (hours worked):

$$hours = \beta_0 + \beta_1 lwage + \beta_2 educ + \beta_3 age + \beta_4 kidslt6 + \beta_5 nwifeinc + u_1$$

Labor demand for married, working women (wages offered):

$$lwage = \alpha_0 + \alpha_1 hours + \alpha_2 educ + \alpha_3 exper + \alpha_4 exper^2 + u_2$$

- `Age`, `number of young children`, and `non-wife income` is a determinant of the supply of labor but not the demand for labor (wages paid) and can be an instrument for `lwage`.
- `experience` is a determinat of `lwage` but not how many hours women work and can be used as an instrument for `hours`.

Example 16.3 Labor Supply for Married, Working Women

Structural equations

$\text{hours} = \beta_0 + \beta_1 \text{ lwage } + \beta_2 \text{ educ } + \beta_3 \text{ age } + \beta_4 \text{ kidslt } 6 + \beta_5 \text{ nwifeinc } + u_1$

$\text{lwage} = \alpha_0 + \alpha_1 \text{ hours } + \alpha_2 \text{ educ } + \alpha_3 \text{ exper } + \alpha_4 \text{ exper }^2 + u_2$

`exper` and $exper^2$ are instruments for `lwage` (`exper` and $exper^2$ are not in the `hours` equation, and `exper` and $exper^2$ are related to `lwage`)

same for `age`, `kidslt6`, and `nwifeinc` being good instruments for `hours`

Example 16.3 Labor Supply for Married, Working Women

2SLS, first stage: reduced form, regress endogenous variables on all exogenous variables and get predicted values

First stage for eq 1: endogenous variable `lwage` on instruments `exper` and $exper^2$ and exogenous variables

$$lwage = \delta_0 + \delta_1 exper + \delta_2 exper^2 + \delta_3 educ + \delta_4 age + \delta_5 kidslt6 + \delta_6 nwifeinc + e_1$$

2SLS, second stage: estimate the equations by using the predicted values from first stage for endogenous variables

$$hours = \beta_0 + \beta_1 \widehat{lwage} + \beta_2 educ + \beta_3 age + \beta_4 kidslt6 + \beta_5 nwifeinc + u_1$$

$$lwage = \alpha_0 + \alpha_1 \widehat{hours} + \alpha_2 educ + \alpha_3 exper + \alpha_4 exper^2 + u_2$$

```r
data(mroz, package = "wooldridge")
mroz %<>% filter(inlf == 1) # keep only working women
# Regression for hours using OLS estimation
model1 <- feols(hours ~ lwage + educ + age + kidslt6 + nwifeinc, data=mroz, se = 'hetero')

# Regression for hours using 2SLS estimation
model2 <-feols(hours ~ educ + age + kidslt6 + nwifeinc | lwage ~ exper+exper^2 , data = mroz) # lwa
msummary(list(model1,model2), stars = TRUE, gof_omit = 'AIC|BIC|Lik|F|R2')
```

	Model 1	Model 2
(Intercept)	1523.775***	2225.662***
	(309.423)	(574.564)
lwage	−2.047	
	(82.023)	
educ	−6.622	−183.751**
	(18.438)	(59.100)
age	0.562	−7.806
	(5.361)	(9.378)
kidslt6	−328.858**	−198.154
	(126.681)	(182.929)
nwifeinc	−5.918+	−10.170
	(3.385)	(6.615)
fit_lwage		1639.556***
		(470.576)
Num.Obs.	428	428

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```r
# Regression for lwage using OLS estimation
model3 <- feols(lwage ~ hours + educ + exper + expersq, mroz)

# Regression for lwage using 2SLS estimation
# hours is instrumented by variables from the other equation
model4 <- ivreg(lwage ~   educ + exper + expersq |
                hours ~ age + kidslt6 + nwifeinc,
                data = mroz)
msummary(list(model3,model4), diagnostics = TRUE)
```

	Model 1	Model 2
(Intercept)	−0.462	−0.656
	(0.204)	(0.338)
hours	0.000	0.000
	(0.000)	(0.000)
educ	0.106	0.110
	(0.014)	(0.016)
exper	0.045	0.035
	(0.013)	(0.019)
expersq	−0.001	−0.001
	(0.000)	(0.000)
Num.Obs.	428	428
R2	0.160	0.126
R2 Adj.	0.152	0.117
AIC	873.5	
BIC	897.9	

Example 16.3 Labor Supply for Married, Working Women

In the OLS model, the effect of wage on hours worked is not significant.

In the 2SLS, there is a significant effect of wage on hours worked.

The coefficient on `lwage` in the hours equation is 1,640.

- *For each 1% increase in wages, hours worked increase by 16.40 hours.*
- The magnitude and significance of the coefficients change using OLS vs 2SLS.
- The coefficients on the instruments `exper`, and `expersq` are individually significant.
- An F-test shows that these coefficients are jointly significant.

In the OLS model, the effect of hours on wage is not significant.

In the 2SLS, effect of hours on wage is also not significant.

The OLS results are very similar to the 2SLS results for this equation.

Two of the three coefficients on the instruments `age`, `kidslt6`, and `nwifeinc` are individually significant.

An F-test shows that these coefficients are jointly significant.

Order and rank conditions for identification

The **order condition** states that an equation is identified if at least one of the exogenous variables is excluded from this equation.

The **rank condition** states that an equation is identified if and only if the other equation includes at least one exogenous variable that is excluded from this equation.

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

$$y_2 = \alpha_0 + \alpha_1 y_1 + \alpha_2 z_2 + u_2$$

For the equation for y_1 to be identified, z_2 need to be excluded from this equation and included in the other equation of y_2.

These are the properties for z_2 to serve as an instrument for y_2 (excluded from the equation for y_1 and included in the equation for y_2).

Testing for the rank condition

The rank condition states that the exogenous variables that are excluded from the equation are included in the other equation. This can be tested using the reduced form equation.

For the equation for y_1 to be identified $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$, there needs to be at least one good instrument z_2 for y_2.

Estimate the reduced form equation: $y_2 = \delta_0 + \delta_1 z_1 + \delta_2 z_2 + e_1$, and test whether the coefficient on the instrument variable z_2 is significant.

$\mathrm{H}_0 : \delta_2 = 0$ (equation for y_1 is not identified)

$\mathrm{H}_0 : \delta_2 \neq 0$ (equation for y_1 is identified)

Testing for the rank condition

Testing for the rank condition: for the equation for `hours` to be identified, there needs to be at least one exogenous variable that is excluded from the equation for `hours` and included in the equation for `lwage`

Estimate the reduced form equation for `lwage`:

lwage $= \delta_0 + \delta_1$ exper $+ \delta_2$ exper $^2 + \delta_3$ educ $+ \delta_4$ age $+ \delta_5$ kidslt $6 + \delta_6$ nwifeinc $+ e_1$

- `exper` and `expersq` are the instruments for `lwage`.

Test if coefficients δ_1 and δ_2 on exper and exper ^{2} are jointly significantly different from zero.

$\mathrm{H}_0 : \delta_1 = 0$ and $\delta_2 = 0$ (equation for hours is not identified)

$\mathrm{H}_0 : \delta_1 \neq 0$ or $\delta_2 \neq 0$ (equation for hours is identified)

F-statistic $= 9.33$ and p-value $= 0.0001$. The coefficients are jointly significant.

Rank condition is satisfied and the equation for hours is identified.

Testing for the rank condition

Testing for rank condition: for the equation for `lwage` to be identified, there needs to be at least one exogenous variable that is excluded from the equation for `lwage` and included in the equation for `hours`.

Estimate the reduced form equation for `hours`:

$\text{hours} = \gamma_0 + \gamma_1 \text{ age} + \gamma_2 \text{ kidslt } 6 + \gamma_3 \text{ nwifeinc} + \gamma_4 \text{ educ} + \gamma_5 \text{ exper} + \gamma_6 \text{ exper}^2 + e_2$

- `age`, `kidslt6`, and `nwifeinc` are the instruments for `hours`.

Test if coefficients γ_1 and γ_2 and γ_3 on `age`, `kidslt6`, and `nwifeinc` are jointly significantly different from zero.

$\mathrm{H}_0 : \gamma_1 = 0$ and $\gamma_2 = 0$ and $\gamma_3 = 0$ (equation for `lwage` is not identified)

$\mathrm{H}_0 : \gamma_1 \neq 0$ or $\gamma_2 \neq 0$ or $\gamma_3 \neq 0$ (equation for `lwage` is identified)

F-statistic $= 4.46$ and p-value $= 0.0043$. The coefficients are jointly significant.

Rank condition is satisfied and the equation for lwage is identified.

Testing for rank condition

```r
# Testing for rank condition involves estimating the reduced form equation
# and testing for significance of the instrument variables.

# Reduced form equation for lwage, identifying equation for hours
model5 <- lm(lwage ~ educ + age + kidslt6 + nwifeinc + exper + expersq, mroz)
summary(model5)
linearHypothesis(model5, c("exper = 0", "expersq = 0"))
```

Testing for rank condition

```
## Linear hypothesis test
##
## Hypothesis:
## exper = 0
## expersq = 0
##
## Model 1: restricted model
## Model 2: lwage ~ educ + age + kidslt6 + nwifeinc + exper + expersq
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    423 195.14
## 2    421 186.86  2    8.2815 9.3293 0.0001085 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing for rank condition

```
# Testing for rank condition involves estimating the reduced form equation
# and testing for significance of the instrument variables.

# Reduced form equation for hours, identifying equation for lwage
model6 <- lm(hours ~ educ + age + kidslt6 + nwifeinc + exper + expersq, mroz)
summary(model6)
linearHypothesis(model6, c("age = 0", "kidslt6 = 0", "nwifeinc = 0"))
```

Testing for rank condition

```
## Linear hypothesis test
##
## Hypothesis:
## age = 0
## kidslt6 = 0
## nwifeinc = 0
##
## Model 1: restricted model
## Model 2: hours ~ educ + age + kidslt6 + nwifeinc + exper + expersq
##
##   Res.Df       RSS Df Sum of Sq      F   Pr(>F)
## 1    424 231321286
## 2    421 224200428  3   7120858 4.4571 0.004265 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```