L1: Dealing with Reverse Causation:

Simultaneous Equation Modelling

Prof Gwilym Pryce
AQIM Training

June 2006

Introduction

- Social Science Statistics I & II:
  - We have assumed only one dependent variable, Y, and any number of independent variables, X:

    $Y = a + b_1X_1 + b_2X_2$

- We have assumed that X_1, X_2 etc. cause Y.
  - In actual fact, causation is very difficult to prove empirically, but often our theory makes the direction of causation fairly clear.
    - E.g. *"your income at age 30 is partly determined by your gender"*
      - The causation is unlikely to run the other way:
      - If your income changes, your gender is unlikely to change.
    - E.g. *"your income is partly determined by your age"*
      - The causation is unlikely to run the other way.
      - If your income changes, your age will not change.

- Q1/ In your own research, what is the dependent variable? What are the determinants?
- Q2/ Is there scope for ‘reverse causation’:
  - I.e. one of your explanatory variables actually being affected by the dependent variable
- Q3/ Can you think of any other situations where you might have two or more variables being simultaneously determined by each other and by the other variables in the model?

2. Systems of Equations

- Where we have more than dependent variable, we need to write a system of equations:
  - These are called the "structural equations"
- For example,

$$Y_1 = b_0 + b_1X_1 + b_2X_2 + b_3Y_2$$

$$Y_2 = c_0 + c_1X_1 + c_2X_3 + c_3Y_1$$

  - Where:
    - Y_1, Y_2 are the dependent variables or "endogenous" (i.e. determined within the model).
    - X_1, X_2, X_3 are the independent variables, or "exogenous" variables (i.e. determined outside the model).
    - You 'model' is a multiple equation system.

- Q/ How might a theory in your own field be represented in this way?
  - I.e. in the two equation system,

$$Y_1 = b_0 + b_1X_1 + b_2X_2 + b_3Y_2$$

$$Y_2 = c_0 + c_1X_1 + c_2X_3 + c_3Y_1$$

  - replace the “Xs” and “Ys” with real variable names.

Employee Loyalty Example:

$$\text{Loyalty} = b_0 + b_1X_1 + b_2X_2 + b_3\text{Tenure} \quad [1]$$

$$\text{Tenure} = c_0 + c_1X_1 + c_2X_3 + c_3\text{Loyalty} \quad [2]$$

- Where X_1 = income, X_2 = gender, X_3 = education.
  - Might there be a case for arguing for a 3 equation system here?

3. What happens if we try to estimate the parameters directly?

- Suppose we are most interested in b_2, the effect of gender on employee loyalty.
- What happens if we try to estimate this relationship as a single equation system?
  - e.g. run a regression of L on X_1, X_2, X_3?

- If we try to run a regression on [1] without taking any account of [2]:
  - The coefficients we get from the regression output will actually be a mixture of all the other coefficients.
  - To see this we need to do some algebra:
- Q/ What do you get if you solve equation [1] in terms of L?
  - I.e. substitute [2] in [1] and collect terms.

Answer:

$$L = b_0 + b_1X_1 + b_2X_2 + b_3T \qquad [1]$$

$$T = c_0 + c_1X_1 + c_2X_3 + c_3L \qquad [2]$$

Substitute expression for T from equation [2] into [1]:

$$L = b_0 + b_1X_1 + b_2X_2 + b_3(c_0 + c_1X_1 + c_2X_3 + c_3L)$$

Expand the term on the RHS:

$$L = b_0 + b_1X_1 + b_2X_2 + b_3c_0 + b_3c_1X_1 + b_3c_2X_3 + b_3\ c_3L$$

Collect terms on the RHS:

$$L = (b_0+b_3c_0) + (b_1+b_3c_1)X_1 + b_2X_2 + b_3c_2X_3 + b_3\ c_3L$$

Now write in terms of L:

$$(1- b_3\ c_3)L = (b_0+b_3c_0) + (b_1+b_3c_1)X_1 + b_2X_2 + b_3c_2X_3$$

$$L = {}^{(b0+b3c0)}/_{(1-\ b3\ c3)} + {}^{(b1+b3c1)}/_{(1-\ b3\ c3)}\ X_1 + {}^{b2}/_{(1-\ b3\ c3)}\ X_2 + {}^{b3c2}/_{(1-\ b3\ c3)}\ X_3$$

This is called the Reduced Form equation for L:

I.e. Endogenous variable written as a function of all the exogenous variables in the system:

$$L = g_0 + g_1X_1 + g_2X_2 + g_3X_3$$

Where:

$g_0 = {}^{(b0+b3c0)}/_{(1- b3\ c3)}$

$g_1 = {}^{(b1+b3c1)}/_{(1- b3\ c3)}$ X1

$g_2 = {}^{b2}{}_{/\ (1- b3\ c3)}\ X_2$

$g_3 = {}^{b3c2}{}_{/\ (1- b3\ c3)}\ X_3$

- So, if we run a regression of L on X_1, X_2, X_3, the second coefficient would not give an estimate of b_2:

  $L = b_0 + b_1X_1 + b_2X_2 + b_3T$ [1]

- but of g_2:

  $g_2 = {}^{b2}/_{(1 - b3\ c3)}\ X_2$

- I.e. our estimate would be a mixture of the effects from gender, tenure, income and education.
  - The results would be meaningless…
  - Simply adding in T as an extra explanatory variable would confuse things even further.

Identification problem:

- This is called the *identification problem*
- It arises when our regression results do not allow us to identify the value of the parameter we are seeking to estimate
  - E.g. the impact of gender on employee loyalty.

4. Solution:

- There are two things we need to do to make estimate sure our system is ‘identified’:
  - **[A] make sure we have set up the structural equations properly**
    - I.e. we need the right balance of exogenous and endogous variables in each structural equation
  - **[B] apply an appropriate estimation technique**
    - E.g. 2SLS, 3SLS, MLE.

[A] setting up the structural equations properly

- You need to check whether the parameters in your system can be identified.
- There are two tests for this:
  - ***Rank Condition***:
    - Tells us an equation is identified or not.
  - ***Order Condition***:
    - Tells us whether the equation is exactly identified or over-identified.
- Ideally, we want our equation to be exactly identified.
  - Often there is only one equation we are really interested in, so it doesn't matter if the other equations are not E.I.

Rank Condition:

- (i) Write out the equations:

$L = b_0 + b_1X_1 + b_2X_2 + b_3T$ [1]

$T = c_0 + c_1X_1 + c_2X_3 + c_3L$ [2]

- (ii) Construct a table of exog & endog vars:

Eq.	L	T	X_1	X_2	X_3
[1]	1	1	1	1	0
[2]	1	1	1	0	1

Eq.	L	T	X_1	X_2	X_3
[1]	1	1	1	1	0
[2]	1	1	1	0	1

- **Rank Condition for a particular equation:**
  a. Highlight the columns for which variables are missing from that equation
    - (I.e. highlight the columns where the zeros are on that row)

  b. Delete the row relating to the equation in question

  c. See if you can find (*g*-1) rows and columns that are not all zeros, where *g* is the number of endogenous variables.
    - If so, the equation is identified (the rank condition for id^n is satisfied).
    - If not, the equation is not identified (“ “ not satisfied)

Eq.	L	T	X_1	X_2	X_3
[1]	1	1	1	1	0
[2]	1	1	1	0	1

- **Rank Condition for a particular equation:**
  - **a. Highlight the columns for which variables are missing from that equation**
    - (I.e. highlight the columns where the zeros are on that row)
  - b. Delete the row relating to the equation in question
  - c. See if you can find (g-1) rows and columns that are not all zeros, where g is the number of endogenous variables.
    - If so, the equation is identified (the rank condition for id^n is satisfied).
    - If not, the equation is not identified (" " not satisfied)

Eq.	L	T	X_1	X_2	X_3
[1]	1	1	1	1	0
[2]	1	1	1	0	1

- **Rank Condition for a particular equation:**
  a. Highlight the columns for which variables are missing from that equation
    - (I.e. highlight the columns where the zeros are on that row)

  b. Delete the row relating to the equation in question

  c. See if you can find (g-1) rows and columns that are not all zeros, where g is the number of endogenous variables.
    - If so, the equation is identified (the rank condition for id^n is satisfied).
    - If not, the equation is not identified (" " not satisfied)

Eq.	L	T	X_1	X_2	X_3
[1]	1	1	1	1	0
[2]	1	1	1	0	1

- **Rank Condition for a particular equation:**
  - a. Highlight the columns for which variables are missing from that equation
    - (I.e. highlight the columns where the zeros are on that row)
  - b. Delete the row relating to the equation in question
  - **c. See if you can find (*g*-1) rows and columns that are not all zeros, where *g* is the number of endogenous variables.**
  - **$g = 2$, so $g - 1 = 1$. Of the highlighted columns, can we find at least 1 row and column that is not all zeros?**
    - **Yes, so equation [1] meets the rank condition for identification.**
- **Q/What about equation [2]?**

Eq.	L	T	X_1	X_2	X_3
[1]	1	1	1	1	0
[2]	1	1	1	0	1

- **Rank Condition for a particular equation:**
  - a. Highlight the columns for which variables are missing from that equation
    - (I.e. highlight the columns where the zeros are on that row)
  - b. Delete the row relating to the equation in question
  - c. See if you can find (*g*-1) rows and columns that are not all zeros, where *g* is the number of endogenous variables.
    - If so, the equation is identified (the rank condition for id^n is satisfied).
    - If not, the equation is not identified (“ “ not satisfied)

Order Condition:

- Let g be the number of endogenous variables
- Let k be the total number of variables (endogenous and exogenous) **missing** from the equation under consideration
- Then:
  - 1. If $k = g-1$, the equation is exactly identified
  - 2. If $k > g-1$, the equation is over-identified
  - 3. If $k < g-1$, the equation is under-identified.
- Q/ Establish whether the order condition is satisfied for equation [1] and for equation [2]:

$$L = b_0 + b_1X_1 + b_2X_2 + b_3T \qquad [1]$$

$$T = c_0 + c_1X_1 + c_2X_3 + c_3L \qquad [2]$$

Order Condition for Equation [1]:

g=2

$$L = b_0 + b_1X_1 + b_2X_2 + b_3T \quad [1]$$

$$T = c_0 + c_1X_1 + c_2X_3 + c_3L \quad [2]$$

- For eqution 1, k = no. of missing vars = 1
  - So k = 1= g – 1
    - I.e. equation [1] is *exactly identified*
- For equation [2], k = no. of missing vars = 1
  - So k = 1 = g-1
    - I.e. equation [2] is *exactly identified*

Solutions:

- Since equation [1] is exactly identified,
  - we can apply 2 Stage Least Squares to estimate b_2, the effect of gender on employee loyalty.
  - It doesn't matter whether equation [2] is identified since we are not interested in those parameters.

2SLS

- Stage 1:
  - Estimate the reduced form equations by OLS and obtain the predicted values for the endogenous variables.
- Stage 2:
  - Replace the right-hand-side endogenous variables with these predicted values and estimate the equation by OLS.

2SLS estimation of Equation [1]:

$$L = b_0 + b_1X_1 + b_2X_2 + b_3T \qquad [1]$$

$$T = c_0 + c_1X_1 + c_2X_3 + c_3L \qquad [2]$$

- *<u>Stage 1:</u>* Obtain T^{hat} from reduced form regression:

```
REGRESSION   /DEPENDENT T  /METHOD=ENTER X1 X2 X3     /SAVE PRED(That).
```

- *<u>Stage 2:</u>* Replace T with T^{hat}:

```
REGRESSION   /DEPENDENT L  /METHOD=ENTER X1 X2 That.
```

  – The coefficient from this regression for X_2 should be a reliable measure of b_2, the impact of gender on employee loyalty.

Other Solutions:

- There are more sophisticated solutions:
  - 3 stage least squares
  - Full information max likelihood
- But these methods don’t usually offer much of an improvement on 2SLS and are v. complicated.

Summary:

- First ask whether there is more than one dependent ("endogenous") variable
- If so, there are two things we need to do to make estimate sure our system is 'identified':
  - **[A] set up an appropriate system of structural equations**
    - I.e. we need the right balance of exogenous and endogous variables in each structural equation
    - Run the *Rank* and *Order* tests for identification.
  - **[B] apply an appropriate estimation technique**
    - 2SLS:
      - 1. Get predicted RHS endogenous variables from reduced form.
      - 2. Include these predicted values on RHS of the equation of interest.

Reading:

- Kennedy, P., ch. 10
- Maddala, G. S. (1992) “Introductory Econometrics”, ch. 9.
- Example:
  - Pryce, G. (1999) 'Construction Elasticities and Land Availability: A Two Stage Least Squares Model of Housing Supply Using the Variable Elasticity Approach', *Urban Studies*, 36(13), pp 2283-2304.