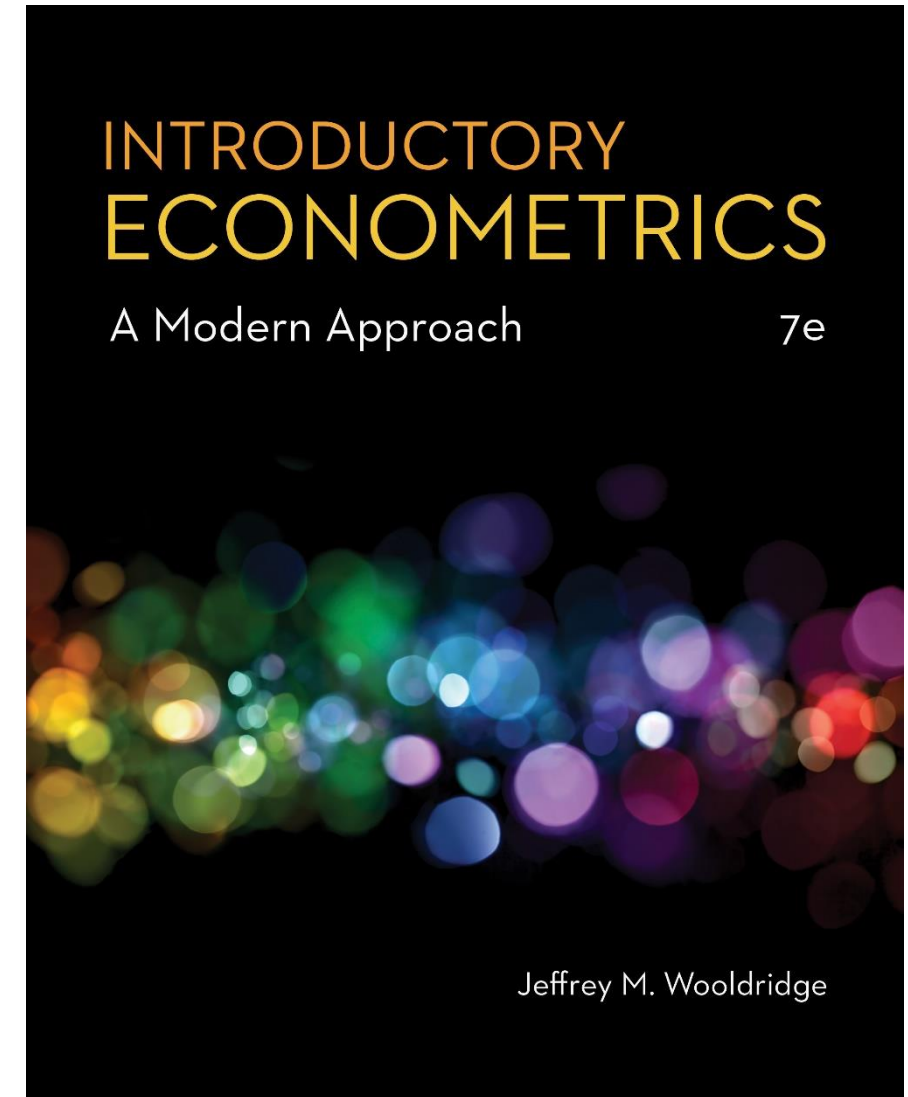


## Chapter 15

### Instrumental Variables Estimation and Two Stage Least Squares



# Instrumental Variables and Two Stage Least Squares (1 of 15)

- **The endogeneity problem is endemic in social sciences/economics**
  - In many cases important personal variables cannot be observed.
  - These are often correlated with observed explanatory information.
  - In addition, measurement error may also lead to endogeneity.
  - Solutions to endogeneity problems considered so far:
    - Proxy variables method for omitted regressors
    - Fixed effects methods if 1) panel data is available, 2) endogeneity is time-constant, and 3) regressors are not time-constant
- Instrumental variables method (IV)
  - IV is the most well-known method to address endogeneity problems.

# Instrumental Variables and Two Stage Least Squares (2 of 15)

- **Motivation: Omitted Variables in a Simple Regression Model**
- Example: Education in a wage equation

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + u_i$$

← Error terms contains factors such as innate ability which are correlated with education

- Definition of a instrumental variable:
  - 1) It does not appear in the regression
  - 2) It is highly correlated with the endogenous variable
  - 3) It is uncorrelated with the error term
- Reconsideration of OLS in a simple regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad \text{and assume} \quad Cov(x_i, u_i) = 0$$

# Instrumental Variables and Two Stage Least Squares (3 of 15)

- **A simple consistency proof for OLS under exogeneity:**

$$\text{Cov}(x_i, u_i) = 0 \quad (\text{Exogeneity})$$

$$\Leftrightarrow 0 = \text{Cov}(x_i, y_i - \beta_0 - \beta_1 x_i) = \text{Cov}(x_i, y_i) - \beta_1 \text{Var}(x_i)$$

$$\Leftrightarrow \beta_1 = \text{Cov}(x_i, y_i) / \text{Var}(x_i)$$

$$\Rightarrow \hat{\beta}_1 = \widehat{\text{Cov}}(x_i, y_i) / \widehat{\text{Var}}(x_i) \rightarrow \text{Cov}(x_i, y_i) / \text{Var}(x_i) = \beta_1$$

This holds as long as the data are such that sample variances and covariances converge to their theoretical counterparts as  $n$  goes to infinity; i.e. if the LLN holds. OLS will basically be consistent if, and only if, exogeneity holds.

# Instrumental Variables and Two Stage Least Squares (4 of 15)

- **Assume existence of an instrumental variable  $z$ :**

$$Cov(z_i, u_i) = 0 \quad (\text{but } Cov(x_i, u_i) \neq 0) \quad \leftarrow \text{The instrumental variable is uncorrelated with the error term.}$$

$$\Leftrightarrow 0 = Cov(z_i, y_i - \beta_0 - \beta_1 x_i) = Cov(z_i, y_i) - \beta_1 Cov(z_i, x_i)$$


$$\Leftrightarrow \beta_1 = Cov(z_i, y_i) / Cov(z_i, x_i)$$

$$\rightarrow \hat{\beta}_{IV} = \frac{\widehat{Cov}(z_i, y_i)}{\widehat{Cov}(z_i, x_i)} \quad \leftarrow \text{The instrumental variable is correlated with the explanatory variable}$$

$$\text{IV-estimator: } \hat{\beta}_{IV} = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x})}$$


# Instrumental Variables and Two Stage Least Squares (5 of 15)

## • Example: Father's education as an IV for education

OLS:  $\widehat{\log(wage)} = - .185 + .109 educ$   Return to education probably overestimated

(.185)    (.014)


$n = 428, R^2 = .118$


$\widehat{educ} = 10.24 + .269 fatheduc$   Is the education of the father a good IV?

(.28)    (.029)

$n = 428, R^2 = .173$

- 1) It doesn't appear as regressor
- 2) It is significantly correlated with educ
- 3) It is uncorrelated with the error (?)

IV:  $\widehat{\log(wage)} = .441 + .059 educ$   The estimated return to education decreases (which is to be expected)

(.446)    (.035)  It is also much less precisely estimated

$n = 428, R^2 = 1 - RSS_{IV}/TSS = .093$

## Instrumental Variables and Two Stage Least Squares (6 of 15)

- **Other IVs for education that have been used in the literature:**
- The number of siblings:
  - 1) No wage determinant, 2) Correlated with education because of resource constraints in hh, 3) Uncorrelated with innate ability
- College proximity when 16 years old:
  - 1) No wage determinant, 2) Correlated with education because more education if lived near college, 3) Uncorrelated with error (?)
- Month of birth:
  - 1) No wage determinant, 2) Correlated with education because of compulsory school attendance laws, 3) Uncorrelated with error

# Instrumental Variables and Two Stage Least Squares (7 of 15)

- **Properties of IV with a poor instrumental variable**
  - IV may be much more inconsistent than OLS if the instrumental variable is not completely exogenous and only weakly related to  $x$

$$\text{plim } \hat{\beta}_{1,OLS} = \beta_1 + \text{Corr}(x, u) \cdot \frac{\sigma_u}{\sigma_x}$$

$$\text{plim } \hat{\beta}_{1,IV} = \beta_1 + \frac{\text{Corr}(z, u)}{\text{Corr}(z, x)} \cdot \frac{\sigma_u}{\sigma_x}$$

There is no problem if the instrumental variable is really exogenous. If not, the asymptotic bias will be the larger the weaker the correlation with  $x$ .

IV worse than OLS if:  $\frac{\text{Corr}(z, u)}{\text{Corr}(z, x)} > \text{Corr}(x, u)$  e.g.  $\frac{0.03}{0.2} > 0.1$



# Instrumental Variables and Two Stage Least Squares (8 of 15)

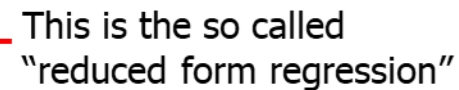
## • IV estimation in the multiple regression model


$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \dots + \beta_k z_{k-1} + u_1$$


 endogenous      exogenous variables

- Conditions for instrumental variable  $z_k$ :
  - 1) Does not appear in regression equation
  - 2) Is uncorrelated with error term
  - 3) Is partially correlated with endogenous explanatory variable

$$y_2 = \pi_0 + \pi_1 z_1 + \dots + \pi_k z_{k-1} + \pi_k z_k + v_2$$


 This is the so called  
"reduced form regression"


 In a regression of the endogenous explanatory variable on all exogenous variables, the instrumental variable must have a non-zero coefficient.

# Instrumental Variables and Two Stage Least Squares (9 of 15)

## • Computing IV estimates in the multiple regression case:

Exogeneity conditions:

$$Cov(z_j, u_1) = 0, \quad j = 1, \dots, k \text{ as well as } E(u_1) = 0$$

Use sample analogs of the exogeneity conditions:

$$n^{-1} \sum_{i=1}^n (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1} - \dots - \hat{\beta}_k z_{ik-1}) = n^{-1} \sum_{i_1}^n \hat{u}_{i1} = 0$$

$$n^{-1} \sum_{i=1}^n z_{ij} \hat{u}_{i1} = \widehat{Cov}(z_j, \hat{u}_1) = 0, \quad j = 1, \dots, k$$

This yields  $k+1$  equations from which the  $k+1$  estimates  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$  can be obtained.

# Instrumental Variables and Two Stage Least Squares (10 of 15)

## • Two Stage Least Squares (2SLS) estimation

- It turns out that the IV estimator is equivalent to the following procedure, which has a much more intuitive interpretation:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \dots + \beta_k z_{k-1} + u_1$$

- First stage (reduced form regression):
  - The endogenous explanatory variable  $y_2$  is predicted using only exogenous information

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \dots + \hat{\pi}_k z_{k-1} + \hat{\pi}_k z_k \leftarrow \text{Additional exogenous variable (instrument)}$$

- Second stage (OLS with  $y_2$  replaced by its prediction from the first stage)

$$y_1 = \beta_0 + \beta_1 \hat{y}_2 + \beta_2 z_1 + \dots + \beta_k z_{k-1} + \text{error}$$

# Instrumental Variables and Two Stage Least Squares (11 of 15)

- **Why does Two Stage Least Squares work?**
- All variables in the second stage regression are exogenous because  $y_2$  was replaced by a prediction based on only exogenous information.
- By using the prediction based on exogenous information,  $y_2$  is purged of its endogenous part (the part that is related to the error term).

# Instrumental Variables and Two Stage Least Squares (12 of 15)

- **Properties of Two Stage Least Squares**
- The standard errors from the OLS second stage regression are wrong. However, it is not difficult to compute correct standard errors.
- If there is one endogenous variable and one instrument then 2SLS = IV.
- The 2SLS estimation can also be used if there is more than one endogenous variable and at least as many instruments.

# Instrumental Variables and Two Stage Least Squares (13 of 15)

## • Example: 2SLS in a wage equation using two instruments

- First stage regression (regress educ on all exogenous variables):

$$\widehat{educ} = 8.37 + .085 \text{ exper} - .002 \text{ exper}^2 + .185 \text{ fatheduc} + .186 \text{ motheduc}$$

(.27)
(.026)
(.001)
(.024)
(.026)

← Education is significantly partially correlated with the education of the parents

- Two Stage Least Squares estimation results:

$$\widehat{\log(wage)} = .048 + .061 \text{ educ} + .044 \text{ exper} - .0009 \text{ exper}^2$$

(.400)
(.031)
(.013)
(.0004)

The return to education is much lower but also much more imprecise than with OLS

# Instrumental Variables and Two Stage Least Squares (14 of 15)

- **Using 2SLS/IV as a solution to errors-in-variables problems**
  - If a second measurement of the mismeasured variable is available, this can be used as an instrumental variable for the mismeasured variable.
- **Statistical properties of 2SLS/IV-estimation**
  - Under assumptions completely analogous to OLS, but conditioning on  $z_i$  rather than on  $x_i$ , 2SLS/IV is consistent and asymptotically normal.
- **Other features of 2SLS/IV-estimation**
  - 2SLS/IV is typically much less precise because there is more multicollinearity and less explanatory variation in the second stage regression.
  - Corrections for heteroskedasticity/serial correlation analogous to OLS.
  - 2SLS/IV easily extends to time series and panel data situations.

# Instrumental Variables and Two Stage Least Squares (15 of 15)

## • Testing for endogeneity of explanatory variables

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \dots + \beta_k z_{k-1} + u_1$$

 Variable that is suspected to be endogenous

Reduced form regression:


$$y_2 = \pi_0 + \pi_1 z_1 + \dots + \pi_k z_k + v_2$$


Variable  $y_2$  is exogenous if and only if  $v_2$  is uncorrelated with  $u_1$ , i.e. if the parameter  $\delta_1$  is zero in the regression:

$$u_1 = \delta_1 v_2 + e_1$$

Test equation:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \dots + \beta_k z_{k-1} + \delta_1 \hat{v}_2 + e_1$$

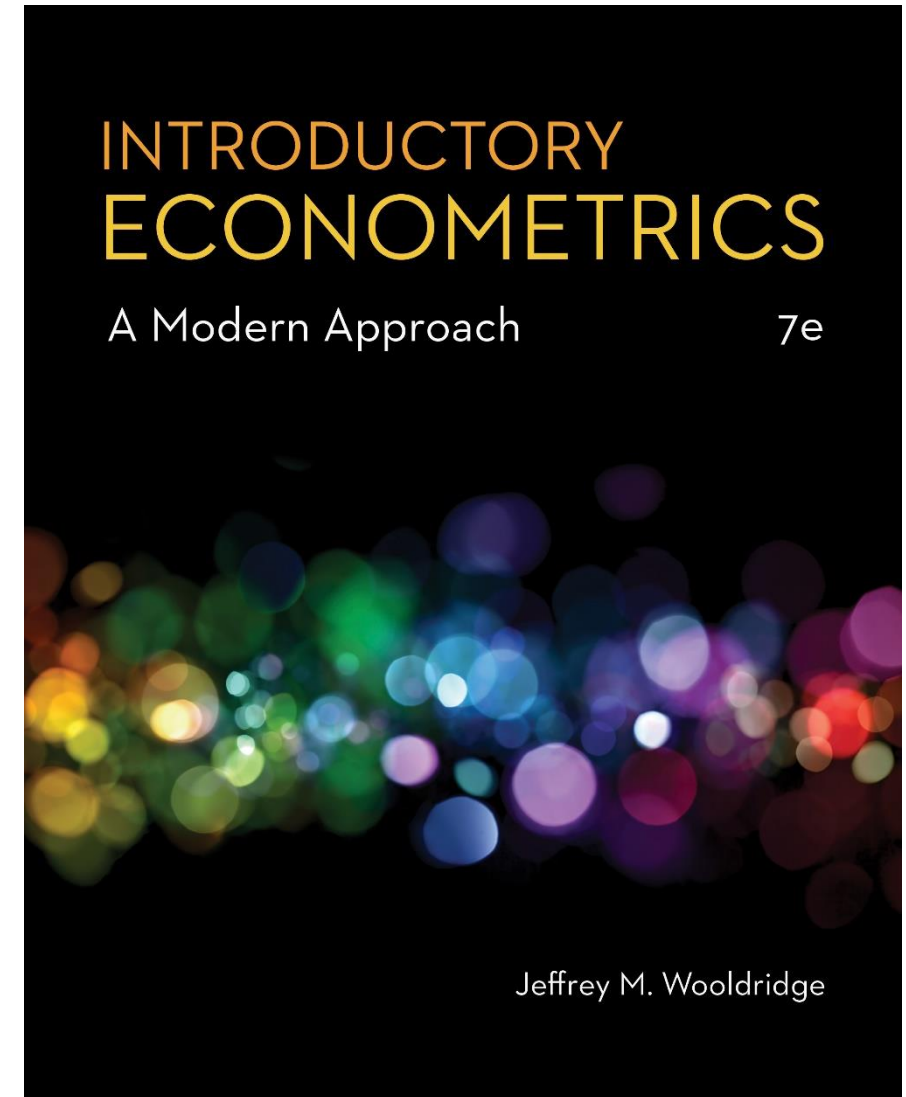
 The residuals from the first stage regression

 The null hypothesis of exogeneity of  $y_2$  is rejected, if in this regression the parameter  $\delta_1$  is significantly different from zero



## Chapter 16

### Simultaneous Equations Models



# Simultaneous Equations Models (1 of 14)

- **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.
- The prototypical case is 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.
- Simultaneous equations systems can be estimated by 2SLS/IV

# Simultaneous Equations Models (2 of 14)

## • Example: Labor demand and supply in agriculture

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

labor supply elasticity

Observed supply shifters e.g. manufacturing wage

Unobserved supply shifters e.g. immigration flows

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

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

labor demand elasticity

Observed demand shifters e.g. agricultural land area

Unobserved demand shifters e.g. food market shocks

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

# Simultaneous Equations Models (3 of 14)

## • Example: Labor demand and supply in agriculture (cont.)

Competition on the labor market in each county  $i$  will lead to a county wage  $w_i$  so that the total number of hours  $h_{is}$  supplied by workers in this county equals the total number of hours  $h_{id}$  demanded by agricultural employers in this county:

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

Simultaneous equations model (SEM):

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

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

Diagram illustrating the simultaneous equations model (SEM) for labor demand and supply in agriculture. The two equations are shown, with variables  $h_i$  and  $w_i$  appearing in both. Red arrows indicate the relationships between variables:

- Endogenous variables:**  $h_i$  and  $w_i$  are both endogenous, as they appear in both equations. Red arrows point from the  $h_i$  term in the second equation to the  $h_i$  term in the first equation, and from the  $w_i$  term in the first equation to the  $w_i$  term in the second equation.
- Exogenous variables:**  $z_{i1}$  and  $z_{i2}$  are exogenous, as they appear only in one equation each. Red arrows point from  $z_{i1}$  to the first equation and from  $z_{i2}$  to the second equation.
- Structural error terms:**  $u_{i1}$  and  $u_{i2}$  are structural error terms, which are uncorrelated with the exogenous variables. Red dashed arrows point from  $u_{i1}$  to the first equation and from  $u_{i2}$  to the second equation.

# Simultaneous Equations Models (4 of 14)

## • 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 →  $murdpc = \alpha_1 polpc + \beta_{10} + \beta_{11} incpc + u_1$

“Behavioral equation” of city government →  $polpc = \alpha_2 murdpc + \beta_{20} + other\ factors$

- $polpc$  will not be exogenous because the number of police officers will depend 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).

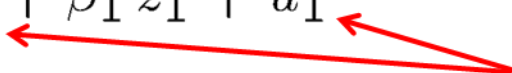
# Simultaneous Equations Models (5 of 14)

## • Simultaneity bias in OLS

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

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

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.



Insert the first equation into the second

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

$$\Leftrightarrow y_2 = \pi_{21} z_1 + \pi_{22} z_2 + v_2 \quad (\text{reduced form equation for } y_2)$$

# Simultaneous Equations Models (6 of 14)

- **Identification in simultaneous equations systems**
- Example: Supply and demand system

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

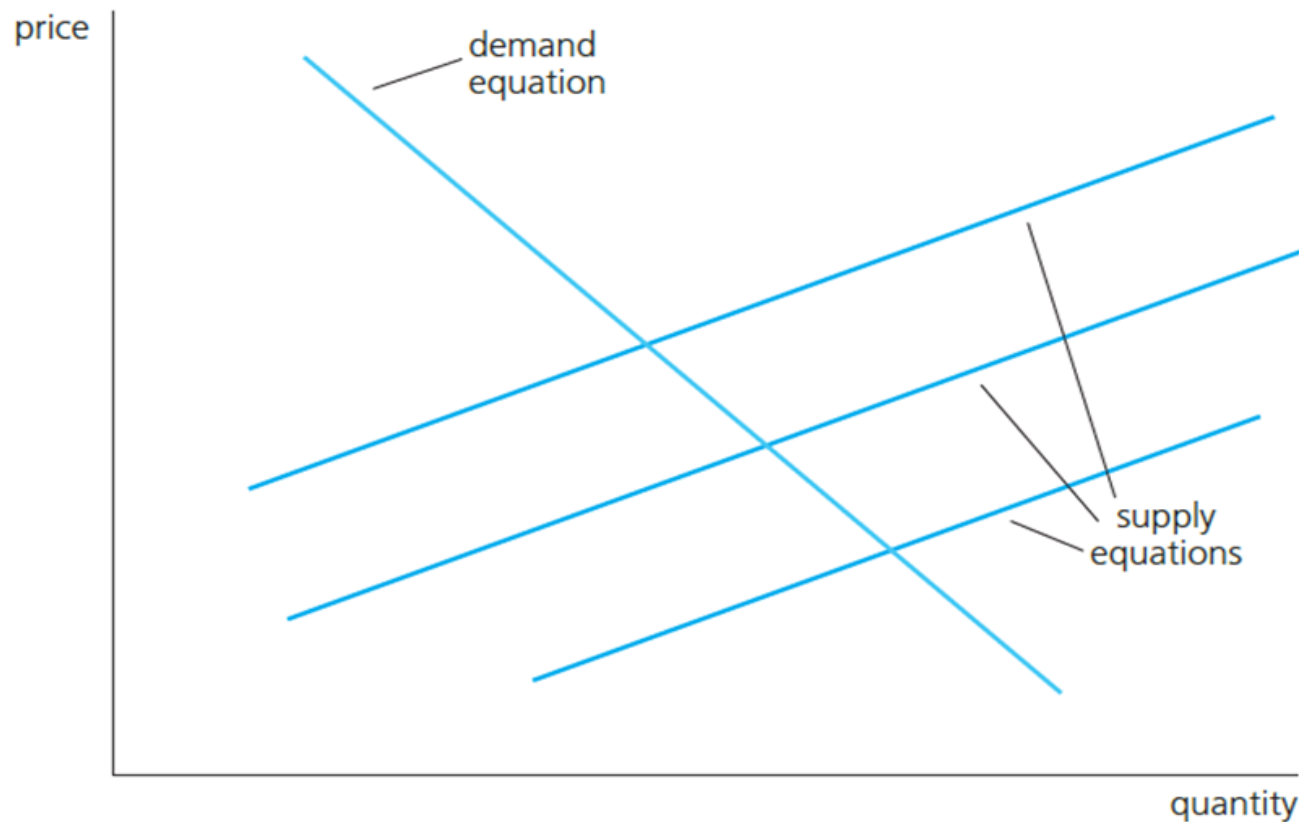
For example, price of cattle feed

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

- 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  $z_1$  as an instrument for the endogenous price variable.

# Simultaneous Equations Models (7 of 14)

- **Graphical illustration of identification problem**



- Intuitively, it is clear why the demand equation can be identified:
- We have an observed variable  $z_1$  that shifts the supply equation while not affecting the demand equation.
- In this way the demand equation can be traced out.



## Simultaneous Equations Models (8 of 14)

- **General rules for identification in simultaneous equation systems**

$$y_1 = \beta_{10} + \alpha_1 y_2 + \beta_{11} z_{11} + \beta_{12} z_{12} + \dots + \beta_{1k_1} z_{1k_1} + u_1$$

$$y_2 = \beta_{20} + \alpha_2 y_1 + \beta_{21} z_{21} + \beta_{22} z_{22} + \dots + \beta_{2k_2} z_{2k_2} + u_2$$

- **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 is identified if, and only if, the second equation contains at least one exogenous variable that is excluded from the first equation.

# Simultaneous Equations Models (9 of 14)

## • Example: Labor supply of married, working women

Supply equation:

Hours and wages are endogenous (market equilibrium)

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

← The supply equation is identified because it does not contain  $\text{exper}$  and  $\text{exper}^2$

Wage offer equation:

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

← The wage offer function is identified because it does not contain  $\text{age}$ ,  $\text{kidslt6}$ , and  $\text{nwifeinc}$

# Simultaneous Equations Models (10 of 14)

## • **Example: Labor supply of married, working women (cont.)**

The rank condition, i.e. the condition that exogenous variables excluded from the equation are included in the other equation can be tested using reduced form equations.

$$\begin{aligned}\log(wage) = & \pi_{20} + \pi_{21}educ + \pi_{22}age + \pi_{23}kidslt6 \\ & + \pi_{24}nwifeinc + \pi_{25}exper + \pi_{26}exper^2 + v_2\end{aligned}$$

Note that this equation can be consistently estimated by OLS as it contains only exogenous variables.

The labor supply function is identified if the hypothesis  $\pi_{25} = \pi_{26} = 0$  can be rejected. This is equivalent to rejecting  $\beta_{22} = \beta_{23} = 0$  in the system of equations.

The same argument applies to the identification of the wage offer function.

## Simultaneous Equations Models (11 of 14)

- **Estimation of simultaneous equation systems by 2SLS**

- Given the identification condition holds, the parameters of a simultaneous equations system can be consistently estimated by 2SLS.
- For this, in a first stage, each endogenous variable is regressed on the full list of exogenous variables (reduced form regressions).
- In a second stage, the system equations are estimated by OLS but with the endogenous regressors being replaced by predictions from stage one.
- If not all equations are identified, one can estimate only the identified ones.
- If certain additional conditions hold, one can also use more efficient system estimation methods (Three Stage Least Squares, 3SLS).

# Simultaneous Equations Models (12 of 14)

## • Example: Labor supply of married, working women using 2SLS

$$\widehat{hours} = 2,225.66 + 1,639.56 \log(wage) - 183.75 educ \\ (574.56) \quad (470.58) \quad (59.10) \\ - 7.81 age - 198.15 kidslt6 - 10.17 nwifeinc, n = 428 \\ (9.38) \quad (182.93) \quad (6.61)$$

$$\widehat{\log(wage)} = - .656 + .00013 hours + .110 educ \\ (.338) \quad (.00025) \quad (.016) \\ + .035 exper - .00071 exper^2, n = 428 \\ (.019) \quad (.00045)$$

# Simultaneous Equations Models (13 of 14)

- **Systems with more than two equations**
  - A necessary condition for identification of an equation is that there are more excluded exog. var. than endog. regressors (= order condition).
  - There is also a rank condition (but it is much more complicated).
- **Simultaneous equations models with time series**
  - Among the earliest applications of SEMs was the estimation of large systems of simultaneous equations for macroeconomic time series.
  - For a number of reasons, such systems are seldom estimated now.
  - The main problem is that most time series are not weakly dependent.
  - Another problem is the lack of enough exogenous variables.

# Simultaneous Equations Models (14 of 14)

## • Example: A simple Keynesian model of aggregate demand

Consumption  $\rightarrow C_t = \beta_0 + \beta_1(Y_t - T_t) + \beta_2 r_t + u_{t1}$

$\uparrow$  Taxes                       $\uparrow$  Interest rate

Investment  $\rightarrow I_t = \gamma_0 + \gamma_1 r_t + u_{t2}$

Income  $\rightarrow Y_t \equiv C_t + I_t + G_t \leftarrow$  Government spending

Endogenous:  $C_t, I_t, Y_t$       Exogenous:  $T_t, G_t, r_t \leftarrow$  The exogeneity of these variables is very questionable