Simultaneous Equations with Error Components

Mike Bronner
Marko Ledic
Anja Breitwieser

PRESENTATION OUTLINE

Part I:

- Simultaneous equation models: overview
- Empirical example

Part II:

- Hausman and Taylor estimator
- Empirical example

INTRODUCTION

- A form of endogeneity of explanatory variables is simultaneity: one or more of the explanatory variables is jointly determined with the dependent variable, typically through an equilibrium mechanism.

- Important method for estimating simultaneous equations models (SEM) is the method of instrumental variables.

- Given a full system, we are able to determine which equations can be identified (that is, can be estimated).

- OLS estimation of an equation that contains an endogenous explanatory variable generally produces biased and inconsistent estimators.

- Instead, 2SLS can be used to estimate any identified equation in a system.

- SEM applications with panel data allow to control for unobserved heterogeneity while dealing with simultaneity.

NATURE OF SIMULTANEOUS EQUATION MODELS

- Classic example: supply and demand equation for some commodity or input to production (such as labor).

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

- Equilibrium condition:

$$h_{is} = h_{id}.$$

- SEM:

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

- Two equations determine labor and wages together → endogenous variables.
- z's → exogenous variables (uncorrelated with supply and demand errors).
- Identification problem: which equation is supply function, and which is demand function?

SIMULTANEITY BIAS IN OLS

- An explanatory variable determined simultaneously with dependant variable is generally correlated with error term → leads to bias and inconsistency in OLS.

- Consider following model (focus on estimating first equation):

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

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

- It follows that:

$$y_2 = \alpha_2(\alpha_1 y_2 + \beta_1 z_1 + u_1) + \beta_2 z_2 + u_2$$

$$(1 - \alpha_2\alpha_1)y_2 = \alpha_2\beta_1 z_1 + \beta_2 z_2 + \alpha_2 u_1 + u_2.$$

$$\alpha_2\alpha_1 \neq 1.$$

- Reduced form:

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

IDENTIFICATION AND ESTIMATION

- Method of two stage least squares (2SLS) can be used to solve the problem of endogenous explanatory variables → can be applied to SEMs.

- Estimate a model by OLS, the key identification condition is that each explanatory variable is uncorrelated with the error term → in general, this condition does not hold for SEMs.

- Instrumental variables can be used to identify (or consistently estimate) the parameters in an SEM equation.

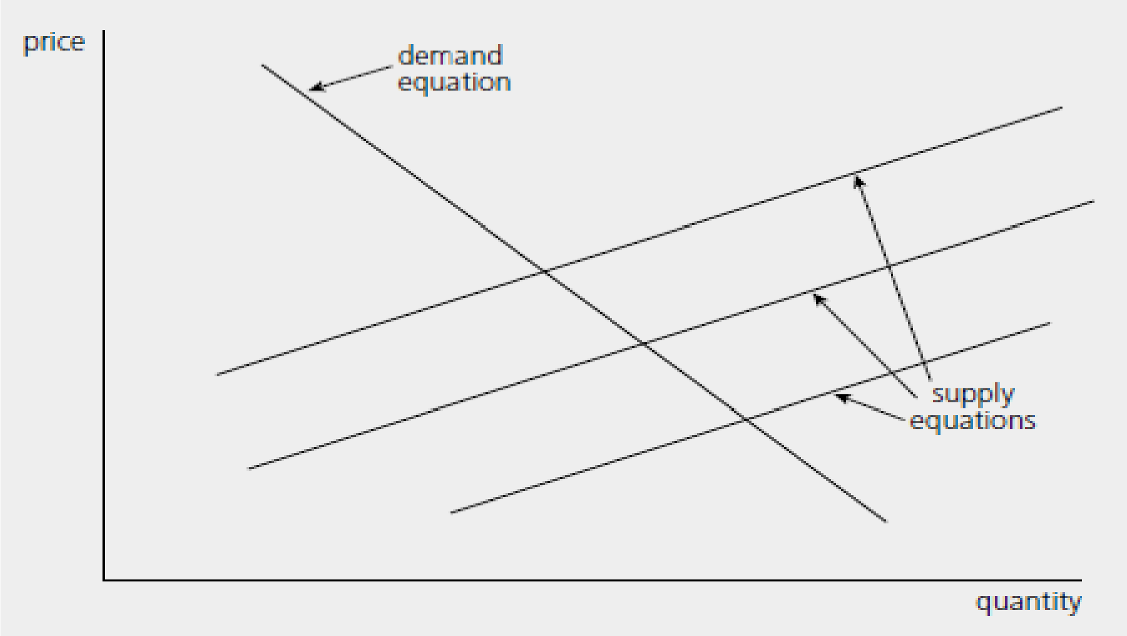
- Supply and demand example:

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

$$q = \alpha_2 p + u_2.$$

- Demand equation (second equation) identified; supply equation (first equation) not.

IDENTIFICATION AND ESTIMATION



Shifting supply equations trace out the demand equation.

- Extension to general two-equation model:

$$y_1 = \beta_{10} + \alpha_1 y_2 + z_1\beta_1 + u_1$$

$$y_2 = \beta_{20} + \alpha_2 y_1 + z_2\beta_2 + u_2,$$

where

$$z_1\beta_1 = \beta_{11}z_{11} + \beta_{12}z_{12} + \ldots + \beta_{1k_1}z_{1k_1}$$

$$z_2\beta_2 = \beta_{21}z_{21} + \beta_{22}z_{22} + \ldots + \beta_{2k_2}z_{2k_2};$$

SYSTEMS WITH MORE THAN TWO EQUATIONS

- Showing that an equation in an SEM with more than two equations is identified is generally difficult, but it is easy to see when certain equations are *not* identified.

- An equation in any SEM satisfies the order condition for identification if the number of excluded exogenous variables from the equation is at least as large as the number of right-hand side endogenous variables.

- But order condition is only necessary, not sufficient, for identification.

- To obtain sufficient conditions, need to extend the rank condition for identification in two-equation systems.

- In practice, one often simply assumes that an equation that satisfies the order condition is identified.

SYSTEMS WITH MORE THAN TWO EQUATIONS

$$y_1 = \alpha_{12}y_2 + \alpha_{13}y_3 + \beta_{11}z_1 + u_1$$

$$y_2 = \alpha_{21}y_1 + \beta_{21}z_1 + \beta_{22}z_2 + \beta_{23}z_3 + u_2$$

$$y_3 = \alpha_{32}y_2 + \beta_{31}z_1 + \beta_{32}z_2 + \beta_{33}z_3 + \beta_{34}z_4 + u_3,$$

- In terms of the order condition, first equation is overidentified.

  [One overidentifying restriction: total number of exogenous variables in system minus total number of explanatory variables in equation].

- Second equation is a just identified.

- Third equation is unidentified.

- Once an equation in a general system has been shown to be identified, it can be estimated by 2SLS.

SEMs WITH PANEL DATA

- For example, imagine estimating labor supply and wage offer equations for a group of people working over a given period of time.

- Write an SEM for panel data as:

$$y_{it1} = \alpha_1 y_{it2} + z_{it1}\beta_1 + a_{i1} + u_{it1}$$

$$y_{it2} = \alpha_2 y_{it1} + z_{it2}\beta_2 + a_{i2} + u_{it2},$$

  Suppose, interested in first equation → cannot estimate by OLS, as the composite error is potentially correlated with all explanatory variables.

- Two steps:

  (1) eliminate the unobserved effects from the equations of interest using the fixed effects transformation or first differencing.

  (2) find instrumental variables for the endogenous variables in the transformed equation.

SYSTEM OF LABOR SUPPLY AND DEMAND

- The data,taken from the National Longitudinal Survey, comprise a sample of full time working males(545 men) who have completed their schooling by 1980.
- Found in Vella F. and Verbeek M.,(1998); Whose wages do unions raise? A dynamic model of unionism and wage rate determination for young men. Journal of Applied Econometrics, 13, 163-183.

Variable label

```
. describe nr year agric black construc educ exper expersq hisp hours lwage married
  min rur union

                  storage  display     value
variable name       type   format      label      variable label
------------------------------------------------------------------------------------
nr                  int    %9.0g                  person identifier (545 men)
year                int    %9.0g                  1980 to 1987
agric               byte   %9.0g                  =1 if in agriculture (industry dummy)
black               byte   %9.0g                  =1 if black
construc            byte   %9.0g                  =1 if in construction(industry dummy)
educ                byte   %9.0g                  years of schooling
exper               byte   %9.0g                  labor mkt experience (age-6-school)
expersq             int    %9.0g                  exper^2
hisp                byte   %9.0g                  =1 if Hispanic
hours               int    %9.0g                  annual hours worked
lwage               float  %9.0g                  log(wage) (logarithm of hourly wage)
married             byte   %9.0g                  =1 if married
min                 byte   %9.0g                  =1 if mining (industry dummy)
rur                 byte   %9.0g                  =1 if live in rural area
union               byte   %9.0g                  =1 if in union
```

Structural simultaneous equations model

- Consider wage offer as a function of annual hours worked and productivity variables, i.e. education, and experience. Labor supply for men is a function of wage, education and binary variable indicating marital status. In addition to allowing for simultaneous determination of variables there is an unobserved effect in each equation.

- The equilibrium conditions for the wage offer and labor supply equations are:

$$\log(wage_{it}) = \beta_t + \beta_1 hours_{it} + \beta_2 \exp er_{it} + \beta_3 \exp ersq_{it} + \beta_4 educ_i + \alpha_i + \varepsilon_{it}$$

$$hours_{it} = \gamma_t + \gamma_1 (\log wage_{it}) + \gamma_2 married_{it} + \gamma_3 \, educ_i + \delta_i + \xi_{it}$$

- Suppose that we are interested in the labor demand equation and we estimate it by pooled OLS (therefore assuming that the fixed effects are uncorrelate with all the explanatory variables)

- Pooled OLS estimation using standard errors

```
. regress lwage hours exper expersq educ
      Source |       SS       df       MS              Number of obs =    4360
-------------+------------------------------           F(  4,  4355) =  195.59
       Model |  188.312683     4  47.0781707           Prob > F      =  0.0000
    Residual |  1048.21694  4355  .240692753           R-squared     =  0.1523
-------------+------------------------------           Adj R-squared =  0.1515
       Total |  1236.52962  4359  .283672774           Root MSE      =   .4906
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0000657   .0000136    -4.84   0.000    -.0000923   -.0000391
       exper |   .1138919   .0103125    11.04   0.000     .0936742    .1341097
     expersq |  -.0039977   .0007232    -5.53   0.000    -.0054155   -.0025799
        educ |   .1035051   .0046783    22.12   0.000     .0943332     .112677
       _cons |   .0347862   .0664896     0.52   0.601    -.0955674    .1651397
------------------------------------------------------------------------------
```

- With each additional hour of work wages decrease by 0,007% (ceteris paribus)
- If we assume that working week consists of 40 working hours then each additional week worked decreases wage by 0,26% ceteris paribus
- However: OLS standard errors are (suspiciously) small

- Pooled OLS estimation with robust standard errors

```
. reg lwage hours exper expersq educ,vce(cluster nr)
Linear regression                                      Number of obs =     4360
                                                       F(  4,    544) =    85.57
                                                       Prob > F      =   0.0000
                                                       R-squared     =   0.1523
                                                       Root MSE      =    .4906
                                   (Std. Err. adjusted for 545 clusters in nr)
------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0000657   .0000256    -2.57   0.011     -.000116   -.0000154
       exper |   .1138919   .0126997     8.97   0.000     .0889455    .1388384
     expersq |  -.0039977    .000891    -4.49   0.000    -.0057478   -.0022475
        educ |   .1035051   .0090161    11.48   0.000     .0857944    .1212158
       _cons |   .0347862   .1262752     0.28   0.783    -.2132606    .2828329
------------------------------------------------------------------------------
```

- Cluster robust standard errors require that N→ and those errors are independent over i. If the errors for individuals from the same household is correlated than we could use vce(cluster id)option.
- Robust standard errors (with cluster option) will be about twice as large as the usual standard errors. That is why we cannot estimate a wage offer equation by OLS (composite error is potentially correlated with all explanatory variables)

Individual effecs model

In order to remove the unobserved effects first difference (over time) can be applied:

$$\Delta \log(wage_{it}) = +\beta_1 \Delta hours_{it} + \beta_2 \Delta \exp er_{it} + \beta_3 \Delta \exp ersq_{it} + \Delta \varepsilon_{it}$$

- The fixed effect and first difference estimator provide consistent estimates of the population coefficients of the time varying regressors under a limited form of endogeneity of the regressors. Therefore regressor hours might be correlated with the fixed effects but not with error term

- Subsequently we assumed richer form of endogeneity where regressor hours is correlated with error term and instrument variable needs to be developed that is correlated with regressor hours but is uncorrelated with the error term

- Fixed effects estimation

```
. xtreg lwage hours exper expersq educ,fe
Fixed-effects (within) regression               Number of obs      =      4360
Group variable (i): nr                          Number of groups   =       545
R-sq:  within  = 0.1946                         Obs per group: min =         8
       between = 0.0068                                        avg =       8.0
       overall = 0.0465                                        max =         8
                                                F(3,3812)          =    307.08
corr(u_i, Xb)  = -0.1875                        Prob > F           =    0.0000
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0001363   .0000134   -10.19   0.000    -.0001625   -.0001101
       exper |   .1427165   .0083263    17.14   0.000     .1263919     .159041
     expersq |  -.0055127   .0006025    -9.15   0.000    -.0066939   -.0043315
        educ |  (dropped)
       _cons |   1.296076    .033438    38.76   0.000     1.230518    1.361634
-------------+----------------------------------------------------------------
     sigma_u |  .41360637
     sigma_e |  .34764688
         rho |  .58600027   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(544, 3812) =     8.94           Prob > F = 0.0000
```

- When comparing within estimation with pooled OLS estimation we see that standard errors for within estimators are bigger because only within variation of the data is being used.

- The coefficient on education is not identified because the data on education is time invariant.

Application of IV estimation with fixed effects

where exper and expersq are exogenous variables and educ is correlated with time invariant component of the error (for example through correlation of education level and cognitive ability that is unobservable) but is uncorrelated with time varying component of the error

- Given these assumptions we need to control for fixed effects and within estimator yields consistent estimator of coefficients on expersq ,expersq and hours
- The coefficient on educ will not be identified because it is time invariant regressor. Moreover assume that a regressor hour is correlated with time varying component of the error. Then the within estimator becomes inconsistent and we need IV for hours
- Assume that marital status is a valid IV for annual hours worked

Fixed effects estimation using IV

```
.xtivreg lwage (hours=married) exper expersq educ,fe
Fixed-effects (within) IV regression          Number of obs      =       4360
Group variable: nr                            Number of groups   =        545
R-sq:  within  =      .                       Obs per group: min =          8
       between = 0.0009                                      avg =        8.0
       overall = 0.0023                                      max =          8
                                              Wald chi2(3)       =    958.46
corr(u_i, Xb)  = -0.6908                      Prob > chi2        =    0.0000
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |   .0082144   .0319377     0.26   0.797    -.0543824    .0708112
       exper |  -1.110629   4.794188    -0.23   0.817    -10.50707    8.285807
     expersq |   .0551292   .2320068     0.24   0.812    -.3995958    .5098543
        educ |  (dropped)
       _cons |  -11.89521   50.45141    -0.24   0.814    -110.7782    86.98775
-------------+----------------------------------------------------------------
     sigma_u |  3.1546413
     sigma_e |  3.5320931
         rho |  .44373143   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F  test that all u_i=0:     F(544,3812) =     0.09         Prob > F    = 1.0000
------------------------------------------------------------------------------
Instrumented:   hours
Instruments:    exper expersq educ married
```

- The estimation on hours coefficient implies that for each additional working hour wages will increase by 0,8% (ceteris paribus)

- However the two tailed area under the standard normal distribution given an absolute z score |0,26| (i.e. two tailed probability from the absolute z score to infinity on both tails of distribution) is 0,797 for hours regression coefficient and therefore the estimated coefficient is not statistically significant. (Note: The same is the case for the other variables.)

Hausman and Taylor estimator (HT)

The model:

(1) $y_{it} = X_{it}\beta + Z_i\gamma + \mu_i + \nu_{it}$, $i=1,\dots,N;\ t=1,\dots,T$

$X = \left[X_1; X_2\right]$, X_1 is $n \times k_1$ and X_2 is $n \times k_2$; $n=NT$

$Z = \left[Z_1; Z_2\right]$, Z_1 is $n \times g_1$ and Z_2 is $n \times g_2$

where X_2 and Z_2 are assumed endogenous

$\rightarrow$ correlated with the μ_i (but not the ν_{it})

Hausman and Taylor estimator

- RE estimator ignores correlation with μ_i and is biased.
- FE estimator removes the μ_i and hence removes the bias. However the FE estimator does not allow any estimates of γ_i since the time-invariant Z_i are removed by the transformation as well.

- Hausman and Taylor (1981) suggested multiplying the model by $\Omega^{-½}$ and using the set of instruments A_0 =[Q, X_1, Z_1] with $Q = I_{NT} - P$ and $P = I_N \otimes \bar{J}_T$

- It can be shown that the projection of this set of instruments (A_0) is equivalent to the set of instruments A =[QX_1, QX_2, PX_1, Z_1]. X_1 is used twice, once as deviation from the averages and once as averages.
- HT estimator takes advantage of the panel structure by using instruments from within the model.

Hausman and Taylor estimator

Obtaining an estimate for γ:

Obtaining the within residuals and averaging them over time, yields:

$$\hat{d}_i = \bar{y}_{i.} - \bar{X}_{i.}\tilde{\beta}_W$$

Then by regressing $\hat{\mathrm{d}}_{\mathrm{i}}$ on $Z_{\mathbf{i}}$ using $X_{\mathbf{1}}$ and $Z_{\mathbf{1}}$ as instruments intermediate consistent estimates of γ are obtained:

$$\hat{\gamma}_{2SLS} = (Z'P_A Z)^{-1} Z'P_A \hat{\mathrm{d}}, \qquad P_A = A(A'A)^{-1}A'$$

Next variance-components estimates can be obtained and equation (1) can be premultiplied by $\hat{\Omega}^{-1/2}$.

Hausman and Taylor estimator

- Basically the HT estimator is a 2SLS estimation on:

$$\hat{\Omega}^{-1/2}y_{it} = \hat{\Omega}^{-1/2}X_{it}\beta + \hat{\Omega}^{-1/2}Z_i\gamma + \hat{\Omega}^{-1/2}u_{it}, \quad \text{with the set of instruments } A = [\tilde{X}, \bar{X}_1, Z_1]$$

,where $\widetilde{X} = QX_1 + QX_2$ and $\overline{X}_1 = PX_1$

if $k_1 < g_2$, $\hat{\gamma}_{HT}$ does not exist as the equation is underidentified ($\hat{\beta}_{HT} = \tilde{\beta}_W$)

if $k_1 = g_2$, the equation is just identified and $\hat{\gamma}_{HT} = \hat{\gamma}_{2SLS}$

if $k_1 > g_2$, the equation is overidentified and the Hausman and Taylor estimator is more efficient than FE

Example: Hausman and Taylor estimator

- Turning to the previous example we now can estimate a coefficient for the time-invariant variable educ.
- We assumed that education is the only variable that is correlated with the fixed effect. In order to have valid identification we need at least one time varying regressor that is uncorrelated with the fixed effect
- For the time invariant regressors, educ is endogenous while black and hisp are exogenous time invariant regressors

Hausman-Taylor estimation

```
. xthtaylor lwage agric construc min rur exper expersq hours married union black hisp educ,endog(exper
    expersq hours married union ed)
Hausman-Taylor estimation                       Number of obs      =      4360
Group variable: nr                              Number of groups   =       545
                                                Obs per group: min =         8
                                                               avg =         8
                                                               max =         8
Random effects u_i ~ i.i.d.                     Wald chi2(12)      =    991.23
                                                Prob > chi2        =    0.0000
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
TVexogenous  |
       agric |  -.0597505   .0413266    -1.45   0.148    -.1407492    .0212482
    construc |  -.0110166   .0301432    -0.37   0.715    -.0700962     .048063
         min |   .0584836   .0620396     0.94   0.346    -.0631117    .1800789
         rur |   .0254011   .0262761     0.97   0.334     -.026099    .0769013
TVendogenous |
       exper |   .1382003   .0085126    16.23   0.000     .1215159    .1548846
     expersq |  -.0053716   .0006034    -8.90   0.000    -.0065543   -.0041889
       hours |  -.0001366   .0000133   -10.30   0.000    -.0001626   -.0001106
     married |   .0441056   .0179522     2.46   0.014       .00892    .0792912
       union |   .0743242   .0189427     3.92   0.000     .0371972    .1114512
TIexogenous  |
       black |  -.1163288   .0573931    -2.03   0.043    -.2288171   -.0038404
        hisp |   .1153327   .0646563     1.78   0.074    -.0113912    .2420566
TIendogenous |
        educ |   .1842564   .0400236     4.60   0.000     .1058116    .2627011
             |
       _cons |  -.8945233   .4839347    -1.85   0.065    -1.843018    .0539713
-------------+----------------------------------------------------------------
     sigma_u |  .38692038
     sigma_e |  .34618513
         rho |   .5553943   (fraction of variance due to u_i)
```

- Compared with the pooled OLS estimation the coefficient on educ has increased from 0.1035 to 0.1842 and the standard error has increased from 0.0047 to 0.0400. We can see that now all variables except time varying exogenous regressors are significant at 10% significance level. Each additional year of education will increase wage by 18.42% holding other variables constant. The validity of this claim (due to large effect) is questionable.

- References:

Baltagi, B. H.: Econometric analysis of panel data . 3rd ed., Chichester : Wiley , 2007.

Cameron, A.C., P. K.. Trivedi: Microeconometrics. Methods and Applications. New York: Cambridge University Press, 2005.

Wooldridge, J. M.: Introductory Econometrics. A Modern Approach. 3rd ed., Mason, Ohio: Thomson, South-Western , 2006.