1989

JAARLIKSE KONFERENSIE

VAN DIE

LANDBOU-EKONOMIEVERENIGING

VAN SUIDER-AFRIKA

VERRIGTINGE/PROCEEDINGS

ANNUAL CONFERENCE

OF THE

AGRICULTURAL ECONOMICS

ASSOCIATION OF SOUTHERN AFRICA

25 - 27 SEPTEMBER 1989 BLOEMFONTEIN

ISBN 0 620 14741 5

METHODS OF ESTIMATING SIMULTANEOUS-EQUATIONS MODELS - OLS, 2SLS AND 3SLS

Darroch M[1]

Ordinary Least Squares (OLS), Two-stage Least Squares (2SLS) and Three-stage Least Squares (3SLS) are compared as methods of estimating parameters for simultaneous-equations models.

OLS produces biased and inconsistent parameter estimates due to correlation between the endogenous explanatory variables and the error terms. 2SLS and 3SLS both produce consistent parameter estimates, but the 3SLS estimates are asymptotically more efficient under specified conditions.

Introduction

Much empirical work in agricultural economics studies economic relationships using single-equation regression models estimated by Ordinary Least Squares (OLS). Such models assume (1) the explanatory variables (X's) are exogenous (predetermined) and (2) the cause-and-effect relationship between the endogenous (Y) variable and the X's is one-way: the X's are the cause and Y is the effect (Gujarati:335-336).

There are, however, economic situations where a two-way, or **simultaneous**, relationship exists between Y and some of the X's: Y is not only determined by the X's, but some of the X's are in turn determined by Y. Two-way causation implies that the relationship cannot be treated as a single equation model, but rather belongs to a wider system of equations which describes relationships amongst all relevant variables.

This multi-equation model would have separate equations in which Y and the relevant X's appear as endogenous variables, although they might appear as explanatory variables in other equations of the model. Such a model describing the joint dependence of variables is termed a **simultaneous-equations model** (Koutsoyiannis:331). Unlike single-equation models, this model estimates equation parameters taking account of information provided by other equations in the system.

This paper presents a basic introduction to methods of estimating equation parameters in simultaneous-equations models. Using a simple two-equation market demand-supply model, it shows why two-stage (2SLS) and three-stage (3SLS) least squares are more appropriate estimating methods than OLS. Differences between the methods are highlighted without

1. *Lecturer, Department of Agricultural Economics, University of Natal, Pietermaritzburg.*

recourse to advanced mathematics so that readers can acquire an intuitive feel for the techniques.

1. Why OLS is inappropriate

Consider the following simple food demand-supply model taken from Kmenta:605:

$$Q_t^d = {}_0 + a_1P_t + a_2D_t + u_{1t} \quad \text{(demand)} \quad (1)$$

$$Q_t^s = b_0 + b_1P_t + b_2F_t + b_3A_t + u_{2t} \quad \text{(supply)} \quad (2)$$

$$Q_t^d = Q_t^s \quad \text{(equilibrium condition)} \quad (3)$$

where Q_t = per capita quantity demanded

Q_t^s = per capita quantity supplied

P_t = ration of food prices to general consumer prices

D_t = per capita disposable income

F_t = ratio of previous year's prices received by farmers for products to general consumer prices

A_t = time in years•

$a_0 \ldots\ldots a_2$ = demand equation parameters

$b_0 \ldots\ldots b_3$ = supply equation parameters

u_{1t}, u_{2t} = error terms

This is a simultaneous-equations model because Q and P are jointly dependent (endogenous), being simultaneously determined (assuming the market clears) by the intersection of the (linear) demand and supply curves. As a result, P, which is also an explanatory variable in both equations, is correlated with both u_{1t} and u_{2t}. To illustrate, assume that the market is in equilibrium (demand equals supply):

$$a_0 + a_1P_t + a_2D_t + u_{1t} = b_0 + b_1P_t + b_2F_t + b_3A_t + u_{2t} \quad (4)$$

$$a_1P_t - b_1P_t = b_0 + b_2F_t + b_3A_t - a_0 - a_2D_t - u_{1t} + u_{2t}$$

$$P_t(a_1 - b_1) = b_0 - a_0 + b_2F_t + b_3A_t - a_2D_t - u_{1t} + u_{2t}$$

$$P_t = \frac{(b_0-a_0+b_2)}{(a_1-b_1)} F_t + \frac{b_3}{(a_1-b_1)} A_t - \frac{a_2}{(a_1-b_1)} D_t - \frac{u_{1t}+u_{2t}}{(a_1-b_1)} \quad (5)$$

From equation (5), P_t is obviously dependent on both u_{1t} and u_{2t}. An OLS regression of Q on P would therefore violate the OLS assumption of no correlation between the explanatory variable(s) and the error term. Hence, estimation of the demand and supply equations individually by OLS gives parameter estimates that are not only biased but also inconsistent (do not converge on the true population) values even as sample size is increased). See Gujarati:342-344 or Koutsoyiannis:333 for mathematical proofs of this simultaneous equations bias.

The 2SLS and 3SLS methods are specifically designed to estimate consistent parameters for simultaneous-equations models.

2. Two-stage least squares

This is a method of estimating consistent structural parameters for exactly identified or overidentified equations in a simultaneous-equations system. An equation is identified (exactly or over) if its parameters can be estimated meaningfully. The simultaneous-equations model as a whole is identified if each equation is identified. Two rules - the order and rank conditions - must be met for an equation to be identified. These conditions are derived from a matrix algebra approach to simultaneous-equations (see Kmenta:662-664). The derivations are not given here, but to understand the conditions let:

M = number of endogenous variables in the model

m = number of endogenous variables in a given equation

K = number of exogenous (predetermined or determined outside of the model) variables in the model

k = number of exogenous variables in a given equation.

2.1 Order Condition for Identification

This is a necessary but not sufficient condition for identifying an equation. By the order condition an equation is identified if:

the number of exogenous variables excluded from the equation, (K-k), is not less than the number of endogenous variables minus one, (m-1).

If K-k = m-1, the equation is exactly identified.

If K-k > m-1, the equation is overidentified.

An equation in a model of simultaneous equations may thus be identified if it excludes one or more variables which appear elsewhere in the model. This is called the exclusion (of variables) criterion or zero restrictions criterion

(coefficients of variables not appearing in the equation are assumed to have zero values). In effect, information provided by the excluded variable(s) helps to identify the equation (Gujarati:360-362).

To illustrate the order condition, consider the food demand-supply model in section 1. It has M = 2 (Q and P) endogenous variables and K = 3 (D, F and A) exogenous variables.

For the demand equation:

m = 2 (P and Q) and k = 1 (D), so that

K-k = 3-1 = 2 and m-1 = 2-1 = 1, and as

K-k > m-1, the equation is overidentified.

For the supply equation:

m = 2 (P and Q) and k = 2 (F and A), so that

K-k = 3-2 = 1 and m-1 = 2-1 = 1, and as

K-k = m-1, the equation is exactly identified.

Parameters of both equations can thus be estimated by 2SLS as both are identified.

The order condition for identification is necessary but not sufficient. While it is likely to be a satisfactory rule of thumb for identification, it may fail when the exogenous variables excluded from an equation but present in the model may not all be independent (Gujarati:362).

2.2 Rank Condition for Identification

This is both a necessary and sufficient condition for identifying an equation. By the rank condition an equation is identified if:

in a system of G equations, at least one non-zero determinant of order (G-1) can be constructed from the coefficients of the variables (both endogenous and exogenous) excluded from that particular equation, but included in the other equations of the model.

To apply the rank condition:

1. Rewrite the food demand-supply model as

$$Q_t^d - a_0 - a_1P_t - a_2D_t = u_{1t} \quad (6)$$

$$Q_t^s - b_0 - b_1P_t - b_2F_t - b_3A_t = u_{2t} \quad (7)$$

$$Q_t^d - Q_t^s = 0 \quad (8)$$

2. Write the above system in tabular form as in Table 1. Given that G = 3, Table 1 must provide at least one non-zero determinant of order G-1 = 2 from the coefficients of the variables excluded from each equation.

Table 1: Rank condition

Equation	Q_t^d	Q_t^s	Pt	Dt	Ft	At
(6)	1	0	-a1	-a2	0	0
(7)	0	1	-b1	0	-b2	-b3
(8)	1	-1	0	0	0	0

3. Strike out coefficients of the row in which the equation under consideration appears.

4. Also strike out columns in which a non-zero coefficient of the equation being examined occurs.

5. Entries left in the table are the coefficients of variables included in the system but not in the equation under consideration.

For the demand equation (6), steps 1 to 5 produce a non-zero determinant of order 2 given by:

$$\begin{vmatrix} 1 & -b_2 \\ -1 & 0 \end{vmatrix} = (1)\,(0) - (-1)\,(-b_2) = -b_2$$

For the supply equation (7), steps 1 to 5 produce a non-zero determinant of order 2 given by:

$$\begin{vmatrix} 1 & -a_2 \\ 1 & 0 \end{vmatrix} = (1)\,(0) - (1)\,(-a_2) = a_2$$

Both equations, and hence the model, are therefore identified by the necessary and sufficient rank condition. Parameters of the demand and supply equations can thus be estimated by 2SLS.

2.3 Estimation Procedure

The aim of 2SLS is to replace endogenous explanatory variables (correlated with the error terms) with 'proxy' explanatory variables which resemble them, but are uncorrelated with the error terms. This enables consistent parameters of the system equations to be estimated.

2SLS involves the application of OLS in two stages:

Stage 1

Regress each endogenous variable on all the exogenous variables of the system by OLS. Substitute actual values of the exogenous variables into the reduced-form equations so derived, to obtain estimated (proxy) values of the endogenous variables.

Stage 2

Replace the endogenous variables in the original structural equations by their 'proxy' (reduced-form) values and apply OLS to each equation. This produces consistent parameter estimates because the 'proxy' variables are uncorrelated with the error terms.

Note, however, that the estimates may not satisfy small sample properties such as unbiasedness and minimum variance. Therefore 2SLS results from small samples should be interpreted with due caution (Gujarati:376-382).

To illustrate 2SLS, the demand-supply model is estimated using data obtained from Kmenta:687, which are listed in Appendix 1.

Stage 1

Regressing the endogenous explanatory P on the exogenous variables D, F and A gives:

$$\hat{P}_t = 90{,}2678 + 0{,}6632D_t - 0{,}4884F_t - 0{,}7370A_t \qquad (9)$$

$$\begin{array}{lllll} & (3{,}2993) & (0{,}0414) & (0{,}3802) & (0{,}0753) \\ t = & (27{,}3596) & (16{,}0107) & (-12{,}8471) & (-9{,}7923) \end{array}$$

$$\bar{r}_2 = 0{,}93 \qquad df = 16$$

where figures in the first set of parentheses are the estimated standard errors of the regression coefficients.

Stage 2

Estimating the demand and supply equations replacing P_t by $\hat{P}_t$ gives:

$$Q_t^d = 94,6333 - 0,2436\hat{P}_t + 0,3140D_t \qquad (10)$$

$$(7,9208) \quad (0,0965) \quad (0,0469)$$

$$t = (11,9474) \quad (-2,5243) \quad (6,6887)$$

$$\bar{r}_2 = 0,73 \qquad df = 17$$

$$Q_t^s = 49,5324 + 0,2401\hat{P}_t + 0,2557F_t + 0,2529A_t \qquad (11)$$

$$(12,0105) \quad (0,0999) \quad (0,0473) \quad (0,0997)$$

$$t = (4,1241) \quad (2,4023) \quad (5,4096) \quad (2,5380)$$

$$\bar{r}_2 = 0,57 \qquad df = 16$$

For comparison, the (inconsistent) OLS estimates of the demand and supply equations without 'purging' the correlation between P_t and u_{1t} and u_{2t} are:

$$Q_t^d = 99,8954 - 0,3163P_t + 0,3347D_t \qquad (12)$$

$$(7,5194) \quad (0,0907) \quad (0,0454)$$

$$t = (13,2851) \quad (-3,4882) \quad (7,3673)$$

$$\bar{r}_2 = 0,74 \qquad df = 17$$

$$Q_t^s = 58,2754 + 0,1604P_t + 0,2481F_t + 0,2483A_t \qquad (13)$$

$$(11,4629) \quad (0,0949) \quad (0,0462) \quad (0,0975)$$

$$t = (5,0838) \quad (1,6901) \quad (5,3723) \quad (2,5462)$$

$$\bar{r}_2 = 0,59 \qquad df = 16$$

The OLS estimates compare quite well with the 2SLS estimates. This is expected due to the very high r_2 value in equation (9), which shows P_t is almost identical to $\hat{P}_t$. The OLS estimators tend to have smaller standard errors than the 2SLS estimators. Signs for both the OLS and 2SLS parameter estimates accord with a priori expectations.

The efficiency of parameter estimates may be improved (lower standard errors) by using 3SLS.

3. Three-stage least squares

This method leads to more efficient parameter estimates than 2SLS by estimating all equations of the system simultaneously. It uses more information than 2SLS, since it also takes into account the correlation between the error terms of the structural equations.

3SLS involves the application of OLS in three stages. The first two stages are the same as 2SLS. Once 2SLS parameter estimates are calculated, the error terms of each equation are used to estimate the cross-equation variances and covariances of the error terms. In stage three, these variances and covariances are used to transform the original equation variables. OLS is then applied to estimate all parameters. By accounting for correlation of the error terms across equations, 3SLS estimates are **consistent** and **asymptotically more efficient** than 2SLS estimates, provided that (Koutsoyiannis:475):

1. Complete specification of the entire system is correctly known.

2. The error term of each equation is serially independent (no auto-correlation).

3. Error terms of the system equations are contemporaneously dependent. If the same less important variables are omitted from a number of system equations, it is inevitable that the error terms which absorb their influence will be correlated. The use of 3SLS will hence be appropriate (see Kmenta:695-700 for a mathematical proof).

The major problem with using 3SLS is that a serious specification error in one equation can affect parameter estimates in all equations - since by construction, estimates are sensitive to specification of the whole model system. Thus the decision to use 3SLS entails a trade-off between the gain in efficiency and potential costs of specification error (Pindyck and Rubinfeld:339).

To illustrate the possible efficiency gains from 3SLS, the demand-supply model 3SLS parameter estimates are:

$$Q_t^d = 94{,}6333 - 0{,}2436\hat{P}_t + 0{,}3140D_t \qquad (14)$$

$$(7{,}3027) \quad (0{,}0890) \quad (0{,}0433)$$

$$t = (12{,}9588) \quad (-2{,}7380) \quad (7{,}2549)$$

$$\bar{r}_2 = 0{,}73 \qquad df = 17$$

$$Q_t^S = 52{,}1176 + 0{,}2289\hat{P}_t + 0{,}2289F_t + 0{,}3579A_t \qquad (15)$$

$$\begin{array}{lllll} & (10{,}6378) & (0{,}0892) & (0{,}0393) & (0{,}0652) \\ t = & (4{,}8993) & (2{,}5679) & (5{,}8191) & (5{,}4899) \end{array}$$

$$\bar{r}_2 = 0{,}53 \qquad df = 16$$

Gains in efficiency (standard error reduction) range from 8% for the demand equation (14) parameters to between 11% and 35% for the supply equation (15) parameters.

4. Conclusion

OLS is an inappropriate method of estimating parameters for econometric models in which a two-way, or simultaneous relationship exists between the dependent (endogenous) variable and some of the explanatory (exogenous) variables. Correlation between these explanatory variables and the error terms results in biased and inconsistent OLS parameters.

2SLS produces unbiased and consistent parameter estimates by purging the correlation between the relevant explanatory variables and the error terms. By taking account of cross-equation correlation between the error terms, 3SLS gives parameter estimates which are consistent and asymptotically more efficient than 2SLS estimates under specified conditions.

---oOo---

Appendix 1

Data for demand-supply model estimation

Q_t	P_t	D_t	F_t	A_t
98,485	100,323	87,4	98,0	1
99,187	104,264	97,6	99,1	2
102,163	103,435	96,7	99,1	3
101,504	104,506	98,2	98,1	4
104,240	98,001	99,8	110,8	5
103,243	99,456	100,5	108,2	6
103,993	101,066	103,2	105,6	7
99,900	104,763	107,8	109,8	8
100,350	96,446	96,6	108,7	9
102,820	91,228	88,9	100,6	10
95,435	93,085	75,1	81,0	11
92,424	98,801	76,9	68,6	12
94,535	102,908	84,6	70,9	13
98,757	98,756	90,6	81,4	14
105,797	95,119	103,1	102,3	15
100,225	98,451	105,1	105,0	16
103,522	86,498	96,4	110,5	17
99,929	104,016	104,4	92,5	18
105,223	105,769	110,7	89,3	19
106,232	113,490	127,1	93,0	20

References

Gujarati D 1979. *Basic Econometrics*. McGraw-Hill, Inc, London. 1st Edition.

Kmenta J 1986. *Elements of Econometrics*. MacMillan Publishing Co, New York. 2nd Edition.

Koutsoyiannis A 1977. *Theory of Econometrics*. MacMillan Publishing Co, New York, 2nd Edition.

Pindyck RS and Rubinfeld DL , 1981. *Econometric Models and Economic Forecasts*. McGraw-Hill, Inc, London. 2nd Edition.