No. 212 (84-8)

# On a Variable Selection Procedure for Linearly Constrained Two Stage Least Squares Method

by

Haruo ONISHI

February 1984

# On a Variable Selection Procedure for Linearly Constrained Two Stage Least Squares Method

University of Tsukuba
Haruo ONISHI

Abstract: H. Theil [8] and R.L. Basmann [1] independently introduced the two stage least squares (2SLS) method for the estimation of equations in a simultaneous equation model. When linear constraints are imposed on the coefficients of an equation, the constrained two stage least squares (C2SLS) method is useful for the estimation of such an equation. The j-th best subset problem is defined for the variable selection problem of the C2SLS method. Then, a procedure is proposed to solve the first (to the J-th, e.g. J=10) best subset problem(s) in one computer-run and to regard the (ultimately) best subset (among the best J subsets) as a pragmatical solution to the variable selection problem for the C2SLS method. The best subset obtained thus (i) is derived from a given set of all possible included predetermined, explanatory endogenous and excluded predetermined variable candidates specified for an explained endogenous variable, (ii) is meaningful from the viewpoint of a research field in question, (iii) is just- or over-identified, (iv) satisfies the constraints, (v) satisfies the magnitude conditions (including the sign conditions), the Durbin-Watson test ([2], [3], [4]), the relative absolute error test and the turning point test [6], if applied, and (vi) shows the highest (or sufficiently high) adjusted coefficient of determination. The procedure can make estimation with the C2SLS method less time-consuming, labor-consuming and costly.

## 1. Introduction

No efficient variable selection procedures for the 2SLS or C2SLS method have been proposed so far in the literature. As a result, applied researchers have spent much time loading into computers, estimating by the computers and then evaluating by themselves many equations one at a time until finding a satisfactory one. Thus, much brain labor and other resources such as paper and electricity have been wasted in the world every year. We need a variable selection procedure to find the best subset estimated with the 2SLS or C2SLS method. The author would like to propose a variable selection procedure for the C2SLS method. Of course, unless constraints are imposed on coefficients of an equation, the proposed procedure can also be used for the variable selection problem for the 2SLS method. The package OEPP [7] developed for socio-economic analysis and forecasting by the author can deal with the proposed procedure.

## 2. Definition of a Meaningful and Just- or Over-identifiable Subset

The C2SLS method requires that variable candidates (called candidates from here on) specified for an explained endogeous variable (called an explained variable from here on) be classified into a set of included predetermined candidates (called included candidates from here on), a set of explanatory endogenous candidates (called endogenous candidates from here on) and a set of excluded predetermined candidates (called excluded candidates from here on)

[5]. In general, all possible subsets derivable from a given set of all possible included candidates and endogenous candidates specified for an explained variable are divided into two groups of meaningful and meaningless subsets, regardless of the research field. Researchers are interested only in a group of meaningful subsets. A meaningful subset is defined as the one which (i) contains included and endogenous candidates necessary for a reasonable equation and (ii) does not contain any unnecessary included and endogenous candidates. Constraints are imposed on included and/or endogenous candidates' coefficients of meaningful subsets. On the other hand, all possible subsets derivable from a given set of all possible endogenous and excluded candidates are divided into a group of just- or over-identifiable (called identifiable from here on) subsets and a group of under-identifiable subsets. Only identifiable subsets are estimable with the C2SLS method. Accordingly, it is of great importance to derive only meaningful as well as identifiable subsets from a given set of all possible included, endogenous and excluded candidates specified for an explained variable.

The author [6] pointed out that candidates specified for an explained variable estimated with the ordinary least squares method are classified on the basis of information from a research field into 8 groups. These are (i) absolutely important (or forced or core), (ii) optionally important, (iii) exclusively important, (iv) gradually important, (v) exclusively optional, (vi) gradually optional, (vii) completely optional and (viii) fixed groups. These basic variable classifications can be used for estimation

Table 1. Variable Classifications for the C2SLS Method by Functional Format $y=F(X^1:Y:X^2)$ in Case of Three Candidates A, B and C

| Names | Classifications | Selection of Candidates |
|---|---|---|
| Absolutely Important | /A,B,C/ | (1) A,B,C |
| Optionally Important | <A,B,C> | (1) A,B,C; (2) A,B; (3) A,C; (4) B,C; (5) A; (6) B; or or (7) C |
| Exclusively Important | </A,B,C/> | (1) A; (2) B; or (3) C |
| Gradually Important | <+A,B,C+> | (1) A,B,C; (2) A,B; or (3) A |
| Exclusively Optional | <*A,B,C*> | (1) A; (2) B; (3) C; or (4) Empty (no selection) |
| Gradually Optional | <-A,B,C-> | (1) A,B,C; (2) A,B; (3) A; or (4) Empty |
| Completely Optional | A,B,C | (1) A,B,C; (2) A,B; (3) A,C; (4) B,C; (5) A; (6) B; (7) C; or (8) Empty |
| Fixed | (D,E) | For instance, if B=(D,E) above, all B's must be replaced with D,E. |
| NAI & NU Included Predetermined | 'A,B,C' | If all or some of A, B and C are not selected as included candidates, they must be selected as excluded ones. |
| Constant Term | $C | Always selected, if any. |

Footnotes: (1) NAI & NU stands for non-absolutely-important and non-uniquely. (2) y=explained variable, $X^1$=set of included candidates, Y=set of endogenous candidates, $X^2$=set of excluded or NAI & NU included candidates. (3) $X^1$, Y and $X^2$ are classified by the above rules. (4) Candidates A, B and C can be expressed with at most 8 alphanumeric symbols (and minus time lag numbers in parentheses like HLWK(-2)). (5) Candidates in a functional format are separated from each other by a blank, a comma (,), /, <, >, </, />, <+, +>, <*, *>, <-, ->, (, ), : or '. (6) For example, AA=F($C/BC/D:<+E E(-1)+>FG</I(I1,I2)/>:<J,K>'D'/L,M,N,PQ/). (7) OEPP can handle nested variable classifications.

methods, including the C2SLS method, for a single equation (not for a sub-system) of a simultanous equation model (called a model from here on). Table 1 summarizes the meanings and characteristics of the basic variable classifications and how to select classified candidates.

To derive all possible identifiable subsets, we need to classify included candidates into a group of uniquely included ones and a group of non-uniquely included ones. A uniquely included candidate is defined as the one which (i) could appear in the equation at hand but (ii) never appears in the other equations and identities in a model. On the other hand, a non-uniquely included candidate is defined as the one which (i) could appear in the equation at hand and (ii) does appear in at least one of the other equations and/or identities in a model. Only when non-uniquely included candidates are not absolutely important, can they be selected as included or excluded ones. Although there are many good methods to select non-uniquely-included and non-absolutely-important candidates as excluded candidates in subsets which do not have them as included candidates, we follow the method embodied in the OEPP. We postulate that (i) an explained variable (and its data vector) y, a set (and data matrix), $X^1$, of all possible (uniquely or non-uniquely) included candidates $x_k^1$'s ($0 \leq k \leq K$), a set (and data matrix), Y, of all possible endogenous candidates $y_\ell$'s ($1 \leq \ell \leq L$) and a set (and data matrix), $X^2$, of all possible excluded candidates $x_m^2$'s ($1 \leq m \leq M$) and non-uniquely-included and non-absolutely-important candidates $x_k^1$'s (for some k's) are loaded into a computer through a functional format, (ii) sets $X^1$, Y and

$X^2$ are separated from each other by a colon (:) in the functional format, (iii) sets $X^1$, Y and $X^2$ are classified on the basis of information from a research field and econometrics, (iv) a group of non-uniquely- included and non-absolutely-important candidates in $X^2$ are enclosed within '...' and (v) a subset which has the same candidate as an included as well as excluded one must be ignored in the derivation of meaningful and identifiable subsets, where $x_0^1$ stands for a constant term (See Table 1). Then, a functional format can be expressed as $y=F(X^1:Y:X^2)$. Now we can derive only meaningful and identifiable subsets from a given set of all possible included, endogenous and excluded candidates specified for an explained variable.

Let us give an example. Suppose that a researcher loads the following entry into a computer:

```
ABCD=F($C/EE,F/<*(G1,G2)G3*>:</HA,IA/>/JAA,JBB/:'G3'
     /K,L,M/)                                          (1)
```

where $C stands for a constant term.

Then, he has the following 6 equations:

$$ABCD=a_0+a_1EE+a_2F+a_3G1+a_4G2+a_5HA+a_6JAA+a_7JBB \quad (2)$$

$$ABCD=b_0+b_1EE+b_2F+b_3G1+b_4G2+b_5IA+b_6JAA+b_7JBB \quad (3)$$

$$ABCD=c_0+c_1EE+c_2F+c_3G3+c_4HA+c_5JAA+c_6JBB \quad (4)$$

$$ABCD=d_0+d_1EE+d_2F+d_3G3+d_4IA+d_5JAA+d_6JBB \quad (5)$$

$$ABCD=e_0+e_1EE+e_2F+e_3HA+e_4JAA+e_5JBB \quad (6)$$

$$ABCD=f_0+f_1EE+f_2F+f_3IA+f_4JAA+f_5JBB \quad (7)$$

where $a_i$'s, $b_i$'s, $c_i$'s, $d_i$'s, $e_i$'s and $f_i$'s stand for coefficients. The 'G3' in format (1) shows that G3 is a non-uniquely included candidate, while G1 and G2 are uniquely included ones. Whether EE and F are uniquely or non-uniquely included candidates does not matter, because they are absolutely important. It must be noted that equations

(2), (3), (6) and (7) use excluded candidates G3, K, L and M, while equations (4) and (5) use excluded candidates K, L and M without G3. Therefore, it is known that equations (4) and (5) are just-identified and the others are over-identified.

How many equations are estimated, if 'G3' is replaced with G3 in format (1)? Because G3 can be regarded as a completely optional excluded candidate in equations which do not have G3, 4 more equations are estimated in addition to the above 6 equations. The additional equations have exactly the same forms as equations (2), (3), (6) and (7) but use excluded candidates K, L and M without G3. Needless to say, the additional equations corresponding to equations (2) and (3) are just-identified and those corresponding to equations (6) and (7) are over-identified. The estimated coefficients of these additional equations usually differ from the corresponding coefficients of the above equations (2), (3), (6) and (7). When constraints are imposed on the coefficients of equations (2) to (7), the C2SLS method is used.

## 3. Derivation of the Constraints Imposed on the i-th Meaningful and Just- or Over-identifiable Subset

To avoid confusing the coefficients of an equation with those of a constraint, we use variable notations for coefficients of an equation. Let us introduce the following constraint when $y=F(X^1:Y:X^2)$ is loaded into a computer:

$$\sum_{k=0}^{K} d_k^1 x_k^1 + \sum_{\ell=1}^{L} d_\ell y_\ell = d \qquad (8)$$

where $d_k^1$ and $d_\ell$ stand for the coefficients of a constraint

imposed on the coefficients of candidates $x_k^1$ and $y_\ell$, respectively, and d stands for the value of a constraint. We call constraint (8) which is expressed with only non-zero coefficients $d_k^1$ and $d_\ell$ a global constraint. A global constraint may not be effective, namely, it may not make sense as a constraint. However, all constraints derived for meaningful and identifiable subsets from a global constraint must be effective. From a technical viewpoint, we postulate that the value of a global constraint is entered on the right-hand side.

Let us give an example. Suppose that the following global constraint is loaded for format (1) into a computer:

1.2*EE-G1-G3+HA+IA=1.5 (9)

where an asterisk (*) is needed between a coefficient which does not assume 1 and -1 and the corresponding variable notation. Then, the following constraints are automatically derived for equations (2) to (7):

1.2*EE-G1+HA=1.5 (10)

1.2*EE-G1+IA=1.5 (11)

1.2*EE-G3+HA=1.5 (12)

1.2*EE-G3+IA=1.5 (13)

1.2*EE+HA=1.5 (14)

1.2*EE+IA=1.5 (15)

Derived constraint (10) is imposed on equation (2), derived constraint (11) is imposed on equation (3), and so on.

## 4. The j-th Best Subset Problem for the Constrained Two Stage Least Squares Method

We assume that there are N (cross-sectional) units like areas, classes, sectors and plots and T observation times

to allow for time series, cross-sectional and pooled data, $y_n(t)$ stands for the explained variable's datum in unit n at observation time t, where $1 \leq n \leq N$ and $1 \leq t \leq T$, and y as a vector stands for $\{y_1(1), y_2(1), \ldots, y_N(1), \ldots, y_1(T), y_2(T), \ldots, y_N(T)\}'$. Furthermore, we assume that $(X_i^1, Y_i, X_i^2)$ stands for the i-th meaningful and identifiable subset derivable from $y=F(X^1:Y:X^2)$ and satisfying $X_i^1 \cap X_i^2 = \phi$, $K_i$, $L_i$ and $M_i$ stand for the numbers of candidates in $X_i^1$, $Y_i$ and $X_i^2$, respectively, and $A_i$ and $B_i$ stand for the row coefficient vectors of $X_i^1$ and $Y_i$, respectively. Then, we can express the equation of the i-th meaningful and identifiable subset as follows:

$$y = X_i^1 A_i' + Y_i B_i' + u \tag{16}$$

where $u \sim N(0, \sigma^2 I)$ stands for a disturbance term and $X_i^2$ is used for the calculation of coefficients and their variance-covariance matrix.

We would like to formulate the j-th best subset problem for the C2SLS method as follows:

Find subset $(X_i^1, Y_i, X_i^2)$ derivable from a given set $(X^1, Y, X^2)$ of all possible included, endogenous, and excluded candidates specified for an explained variable y and estimate coefficient vector $(\bar{A}_i, \bar{B}_i)$ such that

(I) subset $(X_i^1, Y_i)$ is meaningful from the viewpoint of a research field in question,

(II) subset $(X_i^1, Y_i, X_i^2)$ is just- or over-identifiable, namely, $1 \leq L_i \leq M_i$,

(III) subsets $X_i^1$ and $X_i^2$ have no common candidates, namely, $X_i^1 \cap X_i^2 = \phi$,

(IV) $(\bar{A}_i, \bar{B}_i)$ must satisfy the following constraints:

$$(D^1_{gi}, D_{gi})(\tilde{A}_i, \tilde{B}_i)' = d_g \quad \text{for } 1 \leq g \leq G \tag{17}$$

where $D^1_{gi}$, $D_{gi}$ and $d_g$ stand for row vectors of coefficients and a value of the g-th linear constraint imposed on $(A_i, B_i)$,

(V) $(\tilde{A}_i, \tilde{B}_i)$ must be calculated by the following:

$$\begin{bmatrix} \tilde{A}_i' \\ \tilde{B}_i' \\ \tilde{V}_i' \end{bmatrix} = \begin{bmatrix} X^1_i{}'X^1_i & X^1_i{}'Y_i & D^1_i{}' \\ Y_i'X^1_i & Y_i'X_i(X_i'X_i)^{-1}X_i'Y_i & D_i' \\ D^1_i & D_i & 0 \end{bmatrix}^{-1} \begin{bmatrix} X^1_i{}'y \\ Y_i'X_i(X_i'X_i)^{-1}X_i'y \\ d \end{bmatrix}$$

with the asymptotic variance-covariance matrix

$$\tilde{s}^2_i \begin{bmatrix} X^1_i{}'X^1_i & X^1_i{}'Y_i & D^1_i{}' \\ Y_i'X^1_i & Y_i'X_i(X_i'X_i)^{-1}X_i'Y_i & D_i' \\ D^1_i & D_i & 0 \end{bmatrix}^{-1}$$

where

$\tilde{V}_i$ stands for a Lagrangean multiplier row vector corresponding to constraints $(D^1_i, D_i)(\tilde{A}_i, \tilde{B}_i)' = d$, $X_i = (X^1_i, X^2_i)$, $\tilde{y}_i = X^1_i\tilde{A}_i' + Y_i\tilde{B}_i'$, $\tilde{e}_i = y_i - \tilde{y}_i$, $\tilde{s}^2_i = \tilde{e}_i'\tilde{e}_i/(NT - K_i - L_i + G)$, $D^1_i = (D^1_{1i}{}', D^1_{2i}{}', \ldots, D^1_{Gi}{}')'$, $D_i = (D_{1i}', D_{2i}', \ldots, D_{Gi}')'$ and $d = (d_1, d_2, \ldots, d_G)'$

(VI) $\tilde{C}_i = (\tilde{A}_i, \tilde{B}_i)$ satisfies the following magnitude condition (including the sign condition), if necessary:

$$F^1_{hi}\tilde{C}_i' \pm |F^2_{hi}\tilde{C}_i'| \pm |F^3_{hi}\tilde{C}_i'| \geq f^1_h, \quad F^1_{hi}\tilde{C}_i' \pm |F^2_{hi}\tilde{C}_i'| \pm |F^3_{hi}\tilde{C}_i'| \leq f^2_h,$$

$$\text{and/or } f^1_h \leq F^1_{hi}\tilde{C}_i' \pm |F^2_{hi}\tilde{C}_i'| \pm |F^3_{hi}\tilde{C}_i'| \leq f^2_h \text{ for } 1 \leq h \leq H \tag{18}$$

where $F^k_{hi}$ for k=1,2,3, $f^1_h$ and $f^2_h$ stand for a row vector of coefficients, a lower bound and an upper bound of the h-th magnitude condition, respectively, $|F^k_{hi}\tilde{C}_i'|$ for k=2,3 stands for the absolute value of $F^k_{hi}\tilde{C}_i'$, "$\pm$" indicates "+" or "-", and "$\leq$" and "$\geq$" indicate "<" or "$\leq$" and ">" or "$\geq$", respectively,

(VII) the Durbin-Watson statistic defined below is significant at a level specified by a researcher, when $N=1$, $T\geqq 15$ and $K_i+L_i-G\leqq 5$:

$$DW_i=\sum_{t=2}^{T}\{\tilde{e}_{i1}(t)-\tilde{e}_{i1}(t-1)\}^2/\sum_{t=1}^{T}\tilde{e}_{i1}(t)^2 \qquad (19)$$

for $\tilde{e}_{i1}(t)=y_1(t)-\bar{y}_{i1}(t)$ for $1\leqq t\leqq T$,

(VIII) $\tilde{y}_i$ whose elements are denoted by $\bar{y}_{in}(t)$ satisfies the following relative absolute error test, if necessary:

$$100x|\{y_n(t)-\tilde{y}_{in}(t)\}/y_n(t)|\leqq w \text{ for } 1\leqq n\leqq N \text{ and } 1\leqq t\leqq T \qquad (20)$$

where w (%) is specified by a researcher,

(IX) $\tilde{y}_i$ satisfies the following turning point test, if necessary:

if

$$\{y_n(t)-y_n(t-1)\}\{y_n(t+1)-y_n(t)\}<0 \qquad (21)$$

and

$$100x\text{Min}[|\{y_n(t)-y_n(t-1)\}/y_n(t)|,|\{y_n(t)-y_n(t+1)\}/y_n(t)|]\geqq v \qquad (22)$$

then

$$\{y_n(t)-y_n(t-1)\}\{\bar{y}_{in}(t)-\tilde{y}_{in}(t-1)\}>0 \qquad (23)$$

and

$$\{y_n(t+1)-y_n(t)\}\{\bar{y}_{in}(t+1)-\tilde{y}_{in}(t)\}>0 \qquad (24)$$

for $1\leqq n\leqq N$, $2\leqq t\leqq T-1$ and $T\geqq 3$

where v (%) is specified by a researcher,

and

(X) $(\tilde{A}_i,\tilde{B}_i)$ shows the j-th highest adjusted coefficient of determination calculated by $RR_i$:

$$RR_i=\text{Max}\{0,1-(1-R_i)(NT-1)/(NT-K_i-L_i+G)\} \qquad (25)$$

where

$$R_i=1-\tilde{e}_i'\tilde{e}_i/(y-\bar{y}E)'(y-\bar{y}E),\ \bar{y}=\sum_{n=1}^{N}\sum_{t=1}^{T}y_n(t)/NT \text{ and}$$

$E=(1,1,1,\ldots,1)'$ with dimension $(NT\text{x}1)$.

Let us briefly explain the above problem. Conditions (I), (IV) and (VI) are necessary from the viewpoint of information from a research field in question. Conditions (II), (III), (V) and (VII) are related to statistics and/or econometrics. Conditions (VIII), (IX) and (X) are important criteria to evaluate meaningful and identifiable subsets through the comparison of estimated observations with actual ones. Condition (IX) is important, when an estimated model which contains lagged endogenous candidates is used for the final test and/or forecasting. The last condition measures a goodness of fit. Although many fitting criteria have been proposed, we adopted an adjusted coefficient of determination as a measure of goodness of fit, because it has been used so often in the literature of applied research and it can assume a value between 1 (for the perfect fitting) and 0 (for the worst fitting).

If a researcher does not have any additional criterion, the best subset is defined as the one which (i) is meaningful, identifiable and estimable with the C2SLS method, (ii) satisfies the constraints, (iii) satisfies all discrete (or pass-or-fail) criteria applied from (VII), (VIII) and (IX), and (ix) has the highest adjusted coefficient of determination. However, when a researcher needs to utilize a new criterion, the ultimately best subset is defined as the one which (i) is one of the J best subsets obtained by solving the first to the J-th best subset problems in one computer-run and (ii) satisfies the new criterion. Of course, he has to find by himself the ultimately best subset among the J best subsets through the new criterion. Finally, we regard the (ultimately) best subset as an approximate solution, which can be regarded as a pragmatically best

subset, to the variable selection problem for the C2SLS method. If it is impossible to derive all possible meaningful and identifiable subsets from a functional format of all possible included, endogenous and excluded candidates for an explained variable, all meaningful and identifiable subsets should be derived from several functional formats in which the same explained variable and subsets of possible included, endogenous and excluded candidates are entered and then the j-th best subset problem should be solved for each of the functional formats in one computer-run. Then, a researcher can find the ultimately best subset among some best subsets obtained by solving the first (to the J-th) best subset problem(s) for each of the functional formats.

## 5. An Example

Let us demonstrate the proposed procedure by using the data of Japanese agriculture from 1965 to 1979. We would like to find the most reasonable macro agricultural production function of Cobb-Douglas type by the proposed procedure for the C2SLS method. The following variables are used: LY=log(agricultural outputs), $C=constant term, LL=log(labor), LKA=log(KA)=log(animal capital), LKP=log(KP) =log(plant capital), LKM=log(KM)=log(machine capital), LK= log(KA+KP+KM), LKR=log(KA+KP+KM*R), R=an estimated use rate of machinery, LAX=log(A-X)=log(cultivated acreage minus damaged and abandoned acreage), LCAX=log((A-X)*Min(C,1)), C=a cropping index of rice, where C=1, C>1 and C<1 for average, rich and poor harvest, respectively, T=time trend

for technical progress, LT=log(T), DVCS=dummy variable for cold summer, where DVCS=1 for cold summer and DVCS=0 for normal or hot summer, LQ=log(intermediate goods and services), LWIQ=log(wheat import quantity), LRFI=log(real farm income), LRRPP= log(real rice producer price determined by the government), LRWRPF=log(real wage rate of part-time farming).

We assume that (i) LY is an explained variable, (ii) DVCS and LT are uniquely included candidates, (iii) T is a non-uniquely included candidate, (iv) DVCS is completely optional, (v) LT and T are exclusively optional, (vi) LL, LK, LKR, LAX, LCAX and LQ are endogenous, (vii) LL is absolutely important, (viii) LK and LKR are exclusively important, (ix) LAX and LCAX are exclusively important, (x) the sum of the coefficients (elasticities) of the candidates in a subset selected from LL, LK, LKR, LAX, LCAX and LQ must be 1, (xi) excluded candidates LWIQ, LRFI, LRWRPF, LKA(-1), LKP(-1) and LKM(-1) must be always used for all meaningful and identifiable subsets, where, for example, LKA(-1) stands for LKA with time lag number 1, (xii) T must be used as an excluded candidate only when T is not selected as an included candidate in a meaningful and identifiable subset. The following functional format carries the above assumptions except for (x) into the computer:

```
LY=F($C,DVCS<*LT,T*>:/LL/</LK,LKR/></LAX,LCAX/>LQ:'T'
/LWIQ,LRFI,LRRPP,LRWRPF,LKA(-1),LKP(-1),LKM(-1)/)  (26)
```

It is clear that y, $X^1$, Y and $X^2$ of the j-th best subset problem corresponding to the above format are as follows:

y={LY}, $X^1$={$C,DVCS<*LT,T*>},

Y={/LL/</LK,LKR/></LAX,LCAX/>LQ}, and

$X^2$={'T'/LWIQ,LRFI,LRRPP,LRWRPF,LKA(-1),LKP(-1),LKM(-1)/}

Since there are 17 non-constant candidates, the number of all possible subsets is $2^{17}$-1=131071. However, the number of all possible meaningful and identifiable subsets is only 48. Accordingly, the remaining 131,023 subsets are meaningless or cannot be estimated with the C2SLS method.

Assumption (x) shows the constraints imposed on the coefficients of all meaningful and identifiable subsets. These constraints can be derived from the following equation:

$$LL+LK+LKR+LAX+LCAX+LQ=1 \qquad (27)$$

where the variable notations imply their coefficients in the OEPP. Although global constraint (27) does not make sense as a constraint, the constraints derived for all meaningful and identifiable subsets from global constraint (27) according to the rule of the variable classifications must make sense.

In order to check for and avoid subsets which show unrealistic agricultural production, we would like to introduce the following conditions: (i) a free sign (a sign is not determined before estimation) for \$C, a negative sign for DVCS and positive signs for LT, T, LL, LK, LKR, LAX, LCAX and LQ, (ii) 0.1<LL(0.5, (iii) 0.1<LK+LKR(0.5, (iv) 0.1<LAX+LCAX(0.6, (v) 0.1<LQ(0.3, (vi) 5 % Durbin-Watson statistic test, (vii) 5 % relative absolute error test, (viii) 10 % turning point test and (ix) the minimum requirement of an adjusted coefficient of determination is 0.7, where the variable notations imply their coefficients and "(" stands for "≦". It can be seen that the land

factor is emphasized in (iv) but the intermediate goods and services factor is less emphasized in (v).

When we requested the best two subsets under these conditions, we obtained the following two equations in about 2 minutes 36 seconds CPU time by the FACOM M-200 (about 13 MIPS/CPU):

$$\begin{aligned} LY=&1.102936-0.018108*DVCS+0.023992*T+0.376300*LL \\ &(0.18968)(0.012802)\quad(0.006411)\ (0.120444) \\ &+0.143213*LKR+0.480487*LCAX \\ &(0.047151)\quad(0.120493) \end{aligned}$$

R=0.9429, RR=0.9201, SD=0.01830, FA=-0.0369, DW=2.073 (28)

and

$$\begin{aligned} LY=&1.058407-0.045829*DV.CS+0.025255*T+0.371945*LL \\ &(0.20780)(0.012645)\quad(0.006380)\ (0.125413) \\ &+0.120247*LKR+0.507808*LAX \\ &(0.049588)\quad(0.131799) \end{aligned}$$

R=0.9408, RR=0.9171, SD=0.01864, FA=-0.1051, DW=2.204 (29)

where numbers in parentheses, R, RR, SD, FA and DW stand for standard deviations of asymptotic variances of coefficients, coefficient of determination, adjusted coefficient of determination, standard deviation of asymptotic variance of a disturbance term, first-order autocorrelation coefficient and Durbin-Watson statistic, respectively.

The sum of the coefficients of candidates LL, LKR and LCAX in equation (28) is 1 because 0.376300+0.143213 +0.480487=1. On the other hand, the sum of the coefficients of candidates LL, LKR and LAX in equation (29) is also 1 because of 0.371945+0.120247+0.507808=1. Hence, both equations satisfy the respective constraints. Since we cannot find much difference between equations (28) and (29), we would like to adopt equation (28) as a macro agricultural production function of Cobb-Douglas type with

constant returns to scale because its adjusted coefficient of determination is slightly higher than that of equation (29).

## 6. Summary

Sometimes, a simultaneous equation model contains an equation on whose coefficients constraints are imposed. In this case, it takes much time to find a satisfactory equation estimated with the constrained two stage least squares method by trial and error. A variable selection procedure was proposed for the variable selection problem of this estimation method. The laborious work of estimation with this method may be drastically reduced by the proposed procedure available in the package OEPP.

References

[1] R.L. Basmann, A generalized classical method of linear estimation of coefficients in a structural equation, Econometrica 25 (1957) 77-83.

[2] J. Durbin and G.S. Watson, Testing for serial correlation in least squares regression I, Biometrica 37 (1950) 409-428.

[3] J. Durbin and G.S. Watson, Testing for serial correlation in least squares regression II, Biometrica 38 (1951) 159-178.

[4] J. Durbin and G.S. Watson, Testing for serial correlation in least squares regression III, Biometrica 58 (1971) 1-19.

[5] A.S. Goldberger, Econometric Theory (3rd ed., Wiley, N.Y., 1966).

[6] H. Onishi, A variable selection procedure for econometric models, Computational Statistics and Data Analysis 1 (2) (1983) 85-95.

[7] H. Onishi, Computer Package OEPP for Socio-Economic Analysis and Forecasting (Institute of Socio-Economic Planning, University of Tsukuba, Ibaraki, Japan, 1984).

[8] H. Theil, Economic Forecasts and Policy (2nd ed., North-Holland, Amsterdam, 1958).