A Variable Selection Procedure for Two Stage Least Squares Method

HARUO ONISHI*

The variable classifications based on the characteristics of the two stage least squares (2SLS) and scientific knowledge of the research field lead to the derivation of meaningful and just- or over-identifiable subsets from a set of all possible included predetermined, explanatory endogenous and excluded predetermined variable candidates specified for an explained endogenous variable. The j-th best subset is defined as the one which is meaningful for the research, is just- or over-identified, passes the criteria set for the signs and/or magnitudes (of values or absolute values calculated by linear functions) of estimated coefficients, the Basmann over-identifiability restriction test, the Durbin-Watson serial correlation test, the absolute relative error and the turning point tests for estimated observations, and has the j-th highest adjusted coefficient of determination. The first to the J-th best subset, for example, $J=10$, can be chosen from all possible subsets in one computer-run by the package OEPP.

1. Introduction

In general, if N non-constant explanatory variables for an explained variable are completely optional, the number of all possible subsets is 2^N-1. If N is large, 2^N-1 becomes a quite large number. The author [9] pointed out that before estimation, all possible subsets are divided into a group of meaningful subsets and a group of meaningless subsets from the viewpoint of a research field in question. A meaningful subset means one which (i) is derivable from a given set of variable candidates (called candidates from here on) specified for an explained variable, (ii) has only candidates necessary for an equation which is reasonable for the research but (iii) never has unnecessary candidates. The derivation of all possible meaningful subsets is quite important, especially for social sciences like economics. 2SLS introduced by H. Theil [12] and R.L. Basmann [1] requires the classification of included predetermined candidates (called included candidates from here on), explanatory endogenous candidates (called endogenous candidates from here on) and excluded predetermined candidates (called excluded candidates from here on) to derive all possible just- or over-identifiable (called identifiable from here on) subsets [7]. Accordingly, a variable selection procedure for 2SLS depends on the derivation of all possible meaningful and identifiable subsets from a given set of all possible included, endogenous and excluded candidates specified for an explained endogenous variable (called an explained variable from here on). The classification of all possible included, endogenous and excluded candidates based on the information from a research field and econometrics enables a computer to derive all possible meaningful and identifiable subsets.

The author would like to formulate the variable selection problem for 2SLS. The package OEPP [10] designed for socio-economic analysis and forecasting can handle the variable selection procedure to be proposed.

2. Derivation of All Possible Meaningful and Just-or Over-identifiable Subsets

The information necessary for variable classifications is usually obtained from theories, field surveys, experiments, empirical studies and so on. The author [9] pointed out that candidates can be classified into 8 basic groups in order to derive only meaningful subsets. The 8 basic groups 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 ones. Let us postulate that variables or candidates are loaded through a functional form like $y=F(X)$ into a computer and X is a set of explanatory and instrumental candidates which are classified on the basis of the information from a research field and econometrics.

Furthermore, we assume that absolutely important candidates are enclosed within / . . . /, optionally important candidates within 〈 . . . 〉, exclusively important candidates within〈/ . . . /〉, gradually important candidates within 〈+ + . . . + 〉, 〈+ . . . + 〉 or 〈+ . . . + + 〉, exclusively optional candidates within 〈* . . . *〉, gradually optional candidates within 〈− − . . . −〉, 〈− . . . −〉 or 〈− . . . − −〉, fixed candidates within (. . .) and completely optional candidates freely in a functional format. Nested variable classifications, which can be handled by the OEPP, are occasionally seen in applications but can be reduced to

*Institute of Socio-Economic Planning, University of Tsukuba.

the 8 basic variable classifications. Thus, the candidates in the X of $y=F(X)$ are separated from each other with these symbols in addition to a comma or a blank to derive meaningful subsets. The meanings of 8 kinds of basic variable classifications are summarized in Table 1.

In addition to the above variable classifications, the variable selection problem for 2SLS requires that all possible candidates be classified into (ix) uniquely included, (x) non-uniquely included, (xi) endogenous and (xii) excluded candidates for the derivation of all identifiable subsets. A uniquely included candidate is defined as one which could appear only in the equation at hand but never appears in any other equations and identities in a simultaneous equation model (called a model from here on). On the other hand, a non-uniquely included candidate is defined as one which could appear in the equation at hand and does appear in at least one of the other equations and identities in a model. Thus, a uniquely included candidate cannot be used as an excluded candidate, when it is not chosen as an included candidate in a subset. However, when a non-uniquely included candidate is not absolutely important, it can be used as an excluded candidate even if it is not chosen as an included candidate in a subset. Of course, if a non-uniquely included candidate is absolutely important, it can never be used as an excluded candidate. Therefore, we add non-absolutely-important and non-uniquely (abbreviated as NAI & NU from here on) included candidates to a group of all possible excluded candidates and enclose them within '. . .' to distinguish them from all possible excluded candidates. When we derive only meaningful and identifiable subsets, we ignore subsets which (i) possess the same candidate not only as an included candidate but also as an excluded candidate or (ii) do not possess at least one of the NAI & NU included candidates enclosed within '. . .'. We postulate in $y=F(X)$ that (i) all possible included and endogenous candidates are considered to explain the behavior or movement of an explained variable, (ii) a set X^1 of all possible included candidates which are classified is entered first in a functional format, (iii) a set Y of all possible endogenous candidates which are classified follows X^1, (iv) a combined set X^2 of all NAI & NU included candidates and all possible excluded candidates which are classified follows Y, and (v) X^1, Y and X^2 are separated by a colon (:). Then, $y=F(X)$ can be written as $y=F(X^1:Y:X^2)$.

Let us give a simple example. DB is an explained variable, @C, DB1 and Y are included variables, BPP, BPLP and BFP are endogenous variables and Y1, DP1, DP2, DPL1, FP, FP1, BP1, PP and PLP1 are excluded variables, where @C stands for a constant term. Then, the following format can be used for 2SLS:

DB=F(@C, DB1, Y: BPP, BPLP, BFP:
Y1, DP1, DP2, DPL1, FP, FP1, BP1, PP, PLP1) (1)

Table 1 Variable Classifications for the 2SLS 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 (7) C.
Exclusively Important	⟨/A, B, C/⟩	(1) A; (2) B; or (3) C.
Gradually Important	⟨++A, B, C+⟩ ⟨+A, B, C+⟩ or ⟨+C, B, A++⟩	(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−⟩ ⟨−A, B, C−⟩ or ⟨−C, B, A−−⟩	(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 and 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.

The estimated equation of format (1), which is over-identifiable, is expressed as follows:

$$DB=a_0+a_1DB1+a_2Y+a_3BPP+a_4BPLP+a_5BFP \quad (2)$$

where a_i's stand for coefficients. The excluded variables in format (1) do not appear in equation (2), although they are used for the calculation of coefficients. When all of the included, endogenous and excluded variables in format (1) are completely optional candidates and all of the included candidates are uniquely included ones, equation (2) is one of 14,020 possible identifiable subsets, where we have

$$2^2\sum_{l=1}^{3}\binom{3}{l}\sum_{m=l}^{9}\binom{9}{m}=14020.$$

However, it is rather rare to use excluded candidates in a combinatorial manner. Let us assume that (i) all excluded candidates must always be used for the calculation of meaningful and identifiable subsets, (ii) DB1 is a

completely optional and uniquely included candidate, (iii) Y and BPP are absolutely important candidates and (iv) BPLP and BFP are optionally important candidates. Then, the following format can carry the above assumptions:

DB=*F*(@C, DB1/Y: BPP/⟨BPLP, BFP⟩:
/Y1, DP1, DP2, DPL1, FP, FP1,
BP1, PP, PLP1/) (3)

All excluded candidates must be treated as absolutely important, from the viewpoint of econometrics, because they are always used for the calculation of coefficients. Format (3) can generate and estimate the following 5 equations in addition to equation (2) (by a command):

$$DB=b_0+b_1DB1+b_2Y+b_3BPP+b_4BPLP \quad (4)$$

$$DB=c_0+c_1DB1+c_2Y+c_3BPP+c_4BFP \quad (5)$$

$$DB=d_0+d_1Y+d_2BPP+d_3BPLP+d_4BFP \quad (6)$$

$$DB=e_0+e_1Y+e_2BPP+e_3BPLP \quad (7)$$

$$DB=f_0+f_1Y+f_2BPP+f_3BFP \quad (8)$$

where $b_i's$, $c_i's$, $d_i's$, $e_i's$ and $f_i's$ stand for coefficients and all excluded candidates Y1, DP1, . . ., PLP1 are used as instrumental candidates for each of equations (2) and (4) to (8) which are over-identified. Needless to say, the order of candidates in each of 3 econometrically-classified groups of format (3) can be changed. For instance, the following is equivalent to format (3):

DB=*F*(@C/Y/DB1: ⟨BPLP, BFP⟩/BPP:
Y1, DP1, DP2, DPL1, FP, FP1,
BP1, PP, PLP1/) (9)

Let us examine equations (6) to (8). Regardless of whether included candidate Y is a uniquely or non-uniquely included candidate, Y must not be included in the group of excluded candidates in format (3), because Y is absolutely important. However, if DB1 is a non-uniquely included candidate, equations (6) to (8) are estimated without excluded candidate DB1 which should have been used. To allow DB1 to be selected as an excluded candidate for equations (6) to (8), we add 'DB1' to the group of excluded candidates in format (3). The following format is suitable for this case:

DB=*F*(@C/Y/DB1: ⟨BPLP, BFP⟩/BPP/:
'DB1'/Y1, DBP1, DP2, DPL1, FP, FP1,
BP1, PP, PLP1/) (10)

Exactly the same equations as equations (2) and (4) to (8) derivable from format (3) can be estimated from format (10). However, equations (6) to (8) derivable from format (3) are estimated with excluded candidates Y1, DP1, . . ., PLP1 but without DB1, whereas the corresponding equations derivable from format (10) are estimated with excluded candidates Y1, DP1, . . ., PLP1 and DB1. If 'DB1' is replaced with DB1 in format (10), 9 equations are estimated, because DB1 is treated as a completely optional instrumental candidate. They are (i) equations (2), (4) and (5) whose excluded candidates are Y1, DP1, . . ., PLP1 without DB1, (ii) equations (6) to (8) whose excluded candidates are Y1, DP1, . . ., PLP1 without DB1 and (iii) equations (6) to (8) whose excluded candidates are Y1, DP1, . . ., PLP1 and DB1.

It should be kept in mind that all possible excluded candidates cannot always be treated as absolutely important ones. For instance, suppose that excluded candidates W1, W2 and W3 have the following relation: W3=W1+W2. If all of W1, W2 and W3 are treated as absolutely important candidates, all meaningful and identifiable subsets cannot be estimated, because of the multicollinearity among W1, W2 and W3. In this case, for instance, ⟨/(W1, W2)(W2, W3)(W1, W3)/⟩ can be used. The candidates in each of the three pairs are used as excluded candidates.

Thus, it is possible to derive all possible meaningful and identifiable subsets from a given set of all possible included, endogenous and excluded candidates for an explained variable, regardless of research fields.

3. The Variable Selection Problem for the Two Stage Least Squares Method

Here, we formulate the variable selection problem for 2SLS. We treat the case of N (cross-sectional) units (e.g., regions, firms, industries, classes or plots) and T observation times, where $N\geqq 1$, $T\geqq 1$ and $NT>1$. We assume that y stands for an $(NT\times 1)$-vector of an explained variable, X^1 stands for an $(NT\times K)$-matrix of all possible included candidates including a constant term, Y stands for an $(NT\times L)$-matrix of all possible endogenous candidates and X^2 stands for an $(NT\times M)$-matrix of all NAI & NU included candidates and all possible excluded candidates. y is expressed as $[y_1(1), y_2(1), \ldots, y_N(1), \ldots, y_1(T), y_2(T), \ldots, y_N(T)]'$, where $y_n(t)$ stands for the datum of an explained variable of unit n at time t for $1\leqq n\leqq N$ and $1\leqq t\leqq T$. Furthermore, let (X_i^1, Y_i, X_i^2) stand for the i-th meaningful and identifiable subset derivable from (X^1, Y, X^2) and satisfying $X_i^1\cap X_i^2=\emptyset$, A_i and B_i stand for the coefficient row vectors of X_i^1 and Y_i, respectively, and K_i, L_i and M_i stand for the numbers of candidates in X_i^1, Y_i and X_i^2, respectively. We can express the i-th meaningful and identifiable subset as follows:

$$y=X_i^1A_i'+Y_iB_i'+u \quad (11)$$

where $u\sim N(0, \sigma^2I)$ stands for a disturbance term and X_i^2 is used in the first stage of the estimation by 2SLS. Now, we formulate the j-th best subset problem for the variable selection problem of 2SLS as follows:

Find subset(X_i^1, Y_i, X_i^2) 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,

(III) subsets X_i^1 and X_i^2 satisfy $X_i^1 \cap X_i^2 = \emptyset$ and $X_i = (X_i^1, X_i^2)$ includes all predetermined candidates in a model or X_i^2 includes the excluded candidates which a researcher intends to use in the case where all predetermined candidates in a model cannot be used,

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

$$\begin{bmatrix} \bar{A}_i' \\ \bar{B}_i' \end{bmatrix} = \begin{bmatrix} X_i^{1\prime} X_i^1 & X_i^{1\prime} Y_i \\ Y_i' X_i^1 & Y_i' X_i (X_i' X_i)^{-1} X_i' Y_i \end{bmatrix}^{-1} \times \begin{bmatrix} X_i^{1\prime} y \\ Y_i' X_i (X_i' X_i)^{-1} X_i' y \end{bmatrix} \quad (12)$$

with the asymptotic variance-covariance matrix

$$\bar{s}_i^2 \begin{bmatrix} X_i^{1\prime} X_i^1 & X_i^{1\prime} Y_i \\ Y_i' X_i^1 & Y_i' X_i (X_i' X_i)^{-1} X_i' Y_i \end{bmatrix}^{-1}$$

where

$$\bar{y}_i = X_i^1 \bar{A}_i' + Y_i \bar{B}_i', \quad \bar{e}_i = y - \bar{y}_i,$$
$$\bar{s}_i^2 = \bar{e}_i' \bar{e}_i / (NT - K_i - L_i),$$

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

$$D_{hi}^1 \bar{C}_i' \pm |D_{hi}^2 \bar{C}_i'| \pm |D_{hi}^3 \bar{C}_i'| \geqq d_h^1,$$
$$D_{hi}^1 \bar{C}_i' \pm |D_{hi}^2 \bar{C}_i'| \pm |D_{hi}^3 \bar{C}_i'| \leqq d_h^2,$$

and/or

$$d_h^1 \leqq D_{hi}^1 \bar{C}_i' \pm |D_{hi}^2 \bar{C}_i'| \pm |D_{hi}^3 \bar{C}_i'| \leqq d_h^2$$
$$\text{for} \quad 1 \leqq h \leqq H \quad (13)$$

where D_{hi}^k for $k = 1, 2, 3$ stands for a known row vector, d_h^1 and d_h^2 stand for a lower bound and an upper bound, respectively, $|D_{hi}^k \bar{C}_i'|$ for $k = 2, 3$ stands for the absolute value of $D_{hi}^k \bar{C}_i'$, "$\pm$" stands for "$+$" or "$-$", and "$\geqq$" and "$\leqq$" stand for "$>$" or "$\geqq$" and "$<$" or "$\leqq$", respectively,

(VI) the following Basmann over-identifiability restriction statistic is not significant at a level of F test specified by the researcher [3]:

$$F_{M_i - L_i,\ NT - K_i - M_i} = \{(NT - K_i - M_i)/(M_i - L_i)\}\bar{g}_i \quad (14)$$

where

$$\bar{g}_i = \frac{(1, -\bar{B}_i)(y, Y_i)' \{I - X_i^1 \times (X_i^{1\prime} X_i^1)^{-1} X_i^{1\prime}\} (y, Y_i)(1, -\bar{B}_i)'}{(1, -\bar{B}_i)(y, Y_i)' \{I - X_i \times (X_i' X_i)^{-1} X_i'\} (y, Y_i)(1, -\bar{B}_i)'} - 1 \quad (15)$$

(VII) the following Durbin-Watson statistic DW_i is significant at a level specified by the researcher, only when $N = 1$, $T \geqq 6$ and $K_i + L_i - 1 \leqq 20$ ([5], [6], [13]): in a just-identifiable case,

$$DW_i = \sum_{t=2}^{T} \{\bar{e}_{i1}(t) - \bar{e}_{i1}(t-1)\}^2 \Big/ \sum_{t=1}^{T} \bar{e}_{i1}(t)^2 \quad (16)$$

where $\bar{e}_{i1}(t)'s$ are the elements of $\bar{e}_i$, i.e.,

$$\bar{e}_{in}(t) = y_n(t) - \bar{y}_{in}(t) \quad \text{for} \quad 1 \leqq n \leqq N$$
$$\text{and} \quad 1 \leqq t \leqq T \quad (17)$$

and

in an over-identifiable case,

$$DW_i = \sum_{t=2}^{T} \{\bar{e}_{i1}(t) - \bar{e}_{i1}(t-1)\}^2 \Big/ \sum_{t=1}^{T} \bar{e}_{i1}(t)^2 \quad (18)$$

where $\bar{e}_{i1}(t)'s$ are the elements of

$$\bar{e}_i = \{1 - X_i (X_i' X_i)^{-1} X_i'\}(y, Y_i)(1, -\bar{B}_i)' \quad (19)$$

(VIII) $\bar{y}_i$ with elements $\bar{y}_{in}(t)'s$ satisfies the following absolute relative error test, if necessary:

$$100 \times |\bar{e}_{in}(t)/y_n(t)| \leqq w_1 \quad \text{for} \quad y_n(t) \neq 0 \quad (20)$$

and

$$|\bar{y}_{in}(t)| \leqq w_2 \quad \text{for} \quad y_n(t) = 0 \quad (21)$$

for $1 \leqq n \leqq N$ and $1 \leqq t \leqq T$

where $w_1(\%)$ and w_2 are specified by the researcher,

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

if (i)

$$\{y_n(t) - y_n(t - t_i)\}\{y_n(t + t_i) - y_n(t)\} < 0 \quad (22)$$

and (ii-1)

$$100 \times \mathrm{Min}[|\{y_n(t) - y_n(t - t_i)\}/y_n(t)|, \ |\{y_n(t) - y_n(t + t_i)\}/y_n(t)|] \geqq v_1$$
$$\text{for} \quad y_n(t) \neq 0 \quad (23)$$

or (ii-2)

$$\mathrm{Min}[|y_n(t - t_i)|, \ |y_n(t + t_i)|] \geqq v_2$$
$$\text{for} \quad y_n(t) = 0 \quad (24)$$

then (iii)

$$\{y_n(t) - y_n(t - t_i)\}\{\bar{y}_{in}(t) - \bar{y}_{in}(t - t_i)\} > 0 \quad (25)$$

and (iv)

$$\{y_n(t + t_i) - y_n(t)\}\{\bar{y}_{in}(t + t_i) - \bar{y}_{in}(t)\} > 0 \quad (26)$$

(a) for $1 \leqq n \leqq N$, $3 \leqq T$, $2 \leqq t \leqq T - 1$ and $t_i = 1$ in the case where no lagged explained variables are included in X_i^1,

(b) for $1 \leqq n \leqq N$, $2T_i + 1 \leqq T$, $1 + T_i \leqq t \leqq T - T_i$ and $t_i = T_i$ in the case where only one lagged explained variable, whose time lag number is T_i, is included in X_i^1, or

(c) for $1 \leqq n \leqq N$, $2T_i + 1 \leqq T$, $1 + t_i \leqq t \leqq T - t_i$ for all $t_i = 1, 2, \ldots, T_i$ in the case where two or more lagged explained variables are included in X_i^1 and T_i stands for the maximum time lag

number among the time lag numbers of the lagged explained variables, where $v_1(\%)$ and v_2 are specified by the researcher and special attention should be paid to the type of time series or longitudinal data (annual, quarterly, monthly, etc.),

and

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

$$RR_i = \text{Max}\{0,\ 1-(1-R_i)(NT-1)/(NT-K_i-L_i)\} \tag{27}$$

where

$$R_i = 1-(y-\bar{y}_i)'(y-\bar{y}_i)/(y-\bar{y}E)'(y-\bar{y}E) \tag{28}$$

$\bar{y}=y'E/NT$ and $E=(1, 1, 1, \ldots, 1)'$ with dimension $(NT\times 1)$.

Let us explain the above problem briefly. Conditions (I) and (V) depend on whether or not information from a research field is available. Condition (II) implies a necessary condition $1\leqq L_i\leqq M_i$ for estimation by 2SLS. Condition (III) guarantees that subsets X_i^1 and X_i^2 do not have the same predetermined candidates and all available excluded candidates are used for the estimation by 2SLS. Condition (IV) shows formulae to calculate coefficients and their asymptotic variance-covariance matrix by 2SLS. Condition (VI) implies that if the over-identifiability restriction hypothesis of the i-th meaningful and over-identifiable subset is not rejected at a specified significance level, we accept that over-identifiability. Condition (VII) checks whether or not a disturbance term u has autocorrelation only when time series data are used. In the case of $y_n(t)\neq 0$, condition (VIII) checks whether or not each observation $y_n(t)$ is estimated within the tolerance interval $[(1-w_1)y_n(t), (1+w_1)y_n(t)]$ by $\bar{y}_{in}(t)$. Condition (IX) is a useful criterion, when a model having lagged endogenous variables is used for the final test or forecasting. Condition (X) measures a goodness of fit. Conditions (I), (V), (VI), (VII), (VIII) and (IX) are called discrete criteria or pass-or-fail criteria. However, the last condition (X) is a continuous criterion. Thus, it is possible to rank by $RR_i's$ the meaningful and identifiable subsets which pass all discrete criteria applied from conditions (V) to (IX).

Accordingly, if a researcher does not need to apply any additional criterion, the best subset is defined as the one which (i) is meaningful, identifiable and estimable with 2SLS, (ii) satisfies all discrete criteria applied from conditions (V) to (IX) and (iii) has the highest adjusted coefficient of determination. However, if a researcher needs to apply a new criterion or if he does not know appropriate criteria for the tests in the above problem, he should solve the first best subset problem to the J-th best subset problem (e.g., $J=10$) in one computer-run (by the OEPP) and then find the ultimately best subset among the best J subsets through the new criterion or through comparing them by himself. 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 2SLS.

4. An Example

We would like to demonstrate the proposed variable selection method by estimating an agricultural production function of Cobb-Douglas type with the data of Japanese agriculture from 1964 to 1979. Let us introduce the following variable notations: LY=log (agricultural outputs), LL=log (labor), LKA=log (KA)=log (animal capital), LKP=log (KP)=log (plant capital), KM=machinery capital, LK=log (agricultural capital)=log (KA+KP+KM), LKR=log (adjusted agricultural capital)=log (KA+KP+KM*R), R=an estimated use rate where $0\leqq R\leqq 1$, LQ=log (intermediate goods and services), LAX=log (A-X)=log (cultivated acreage minus abandoned and damaged acreage), LCAX=log ((A-X) adjusted by a cropping index of rice), LRFI=log (real farm income), LRRPP=log (real producer price of rice), LRWRPF=log (real wage rate of farming), LWIQ=log (wheat import quantity), DVCS=dummy variable where normal or hot summer=0 and cold summer=1, TT=time trend and @C=constant term.

We assume that (i) an explained variable is LY, (ii) @C is always needed, (iii) DVCS is a uniquely included and completely optional candidate, (iv) TT is a non-uniquely included and completely optional candidate, (v) LL, LK, LKR, LAX, LCAX and LQ are endogenous candidates, (vi) LL is absolutely important, (vii) LK and LKR are exclusively important, (viii) LAX and LCAX are exclusively important, (ix) LQ is completely optional, (x) LWIQ, LRFI, LRRPP, LRWRPF, LRFI(−1), LKA(−1) and LKP(−1) are excluded candidates and (xi) all excluded candidates are always used for the estimation, implying that they are treated as absolutely important, where candidates LRFI(−1), LKA(−1) and LKP(−1) indicate candidates LRFI, LKA and LKP with time lag number 1, respectively. The above assumptions concerned with candidates can be compactly expressed as follows:

$$\begin{aligned}&\text{LY}=F(\text{@C, DVCS, TT:}\\&\quad /\text{LL}/\langle/\text{LAX, LCAX}/\rangle\langle/\text{LK, LKR}/\rangle\text{LQ:}\\&\quad \text{'TT'}/\text{LWIQ, LRFI, LRRPP, LRWRPF,}\\&\quad \text{LRFI}(-1),\ \text{LKA}(-1),\ \text{LKP}(-1)/)\end{aligned} \tag{29}$$

There are 32 meaningful and over-identifiable subsets in format (29). No meaningful and just-identifiable subsets exist. Since there are 16 non-constant candidates, the number of all possible subsets is $2^{16}-1=65535$. Hence, 65,503 subsets are meaningless, under-identified or unnecessary for the research.

We set the following conditions in order to check for and avoid unusual subsets: (i) a free sign (implying that a sign is not determined before estimation) for

@C, positive signs for TT, LL, LAX, LCAX, LK, LKR and LQ, and a negative sign for DVCS, (ii) $0.1 < LL \leqq 0.5$, (iii) $0.1 < LAX + LCAX \leqq 0.6$, (iv) $0.1 < LK + LKR \leqq 0.5$, (v) $0.1 < LQ \leqq 0.3$, (vi) $0.9 \leqq LL + LAX + LCAX + LK + LKR + LQ < 1.1$, (vii) 5% Basmann over-identifiability restriction test, (viii) 5% Durbin-Watson serial correlation test (we accept an inconclusive case by a researcher's subjective judgement), (ix) 1% absolute relative error test, (x) at least 90% of all turning points defined by at least 1% slope must be adequately tracked, and (xi) the minimum adjusted coefficient of determination is 0.7, where the variable notations in (ii) to (vi) imply their coefficients. Although the sign conditions are redundant in this example, they are introduced to reduce the computer's checking time. The land factor is emphasized in (iii), but the intermediate goods and services factor is less emphasized in (v). Unusually decreasing or increasing returns to scale in the agricultural production is checked and removed by (vi). Statistical tests are (vii) and (viii). (ix) requires that estimated observations be within $\pm 1\%$ of actual observations. The 1% slope of (x) defines the less steep slope of a V-shape or $\wedge$-shape turning point. (xi) prevents a roughly fitted equation from becoming the best subset, even if it passes all discrete tests.

When we required the best 2 subsets, the following 2 equations were obtained in about 2 seconds CPU time by the FACOM M-380 (about 23 MIPS): as the first best subset

$$\begin{aligned} LY = {} & 1.02060 + 0.02050 * TT + 0.29453 * LL \\ & (1.5096) \quad (0.00670) \quad (0.13251) \\ & + 0.56189 * LCAX + 0.14859 * LKR \\ & (0.19503) \quad (0.04217) \\ & RR = 0.9295,\ SD = 0.0191,\ BS = 1.845, \\ & DW = 3.066,\ REV = 3,\ EPV = 7 \end{aligned} \tag{30}$$

and as the second best subset

$$\begin{aligned} LY = {} & 1.52587 + 0.02377 * TT + 0.38004 * LL \\ & (3.9021) \quad (0.01004) \quad (0.21431) \\ & + 0.43625 * LAX + 0.14625 * LKR \\ & (0.51614) \quad (0.06656) \\ & RR = 0.8553,\ SD = 0.0274,\ BS = 1.619, \\ & DW = 3.203,\ REV = 3,\ EPV = 7 \end{aligned} \tag{31}$$

where numbers in parentheses, RR, SD, BS, DW, REV and EPV stand for asymptotic standard derivations of coefficients, adjusted coefficient of determination, asymptotic standard deviation of a disturbance term, Basmann statistic, Durbin-Watson statistic, number of endogenous candidates and number of excluded candidates, respectively, and all available excluded candidates LWIQ, LRFI, LRRPP, LRWRPF, LRFI(−1), LKA(−1) and LKP(−1) are used for both (30) and (31). The sum, 1.00501, of the coefficients of candidates LL, LCAX and LKR in (30) implies almost constant returns to scale in the agricultural production, while the sum, 0.96254, of the coefficients of candidates LL, LAX and LKR in (31) implies slightly decreasing returns to scale. We cannot find any good reason that (31) is better than (30), so that we decided to choose (30) as an agricultural production function of Cobb-Douglas type estimated by 2SLS.

5. Summary

The proposed variable selection procedure, which utilizes both knowledges of econometrics and a research field, may be useful to find a solution to the variable selection problem for the two stage least squares method. The j-th best subset problem was formulated for the variable selection problem. The (ultimately) best subset of the first (to the J-th) best subset problem(s) is exactly the same as the equation obtained by loading, estimating and evaluating all possible equations one at a time until a researcher finds the best one. The procedure may be able to save time, labor and other resources like paper and electricity and improve the quality of applied research.

Acknowledgements

The author is grateful to the referees for their comments and suggestions on earlier drafts of this paper.

References

1. BASMANN, R. L. A generalized classical method of linear estimation of coefficients in a structural equation, Econometrica, **25** (1957) 77-83.
2. BASMANN, R. L. On the asymptotic distribution of generalized linear estimators, Econometrica, **28**, 1 (1960) 97-107.
3. BASMANN, R. L. On finite sample distributions of generalized classical linear identifiability test statistics, American Statistics Association, **55**, 292 (1960) 650-659.
4. CHRIST, C. F. Econometric Models and Methods, Wiley, NY (1966).
5. DURBIN, J. Testing for serial correlation in systems of simultaneous regression equations, Biometrika, **44**, 3 (1957) 370-377.
6. DURBIN, J. and WATSON, G. S. Testing for serial correlation in least squares regression I, Biometrica, 38 (1951) 159-178.
7. GOLDBERGER, A. S. Econometric Theory, 3rd ed., Wiley, NY (1966).
8. JUDGE, G., GRIFFITHS, W. E., HILL, R. and LEE, T.-C. The Theory and Practice of Econometrics, Wiley, NY (1980).
9. ONISHI, H. A variable selection procedure for econometric models, Computational Statistics and Data Analysis, 1 (1983) 85-95.
10. ONISHI, H. Computer Package OEPP for Socio-Economic Analysis and Forecasting, Institute of Socio-Economic Planning, University of Tsukuba (1985).
11. ONISHI, H. On a variable selection procedure for a modified limited information maximum likelihood method, J. of Japan Statistical Society, **15**, 1 (1985) 35-44.
12. THEIL, H. Economic Forecasts and Policy, 2nd ed., North-Holland, Amsterdam (1961).
13. SAVIN, N. E. and WHITE, K. J. The Durbin-Watson test for serial correlation with extreme sample sizes or many regressors, Econometrica, **45**, 8 (1977) 1989-1996.

(Received December 26, 1983; revised August 19, 1985)