No. 145 (82-12)

Variable Classification and Generation of Feasible Sets for Various Estimation Methods

by

Haruo ONISHI

March 1982

Variable Classification and Generation of Feasible Sets for Various Estimation Methods[+]

Haruo ONISHI

Institute of Socio-Economic Planning
University of Tsukuba

This paper proposes a method to generate meaningful and estimable subsets for various estimation methods. Variable candidates are classified into eight basic groups from the viewpoint of a research field in question and five groups from the viewpoint of econometrics. Nested variable classification can be reduced to the level of the eight basic groups but save bothersome labor of entering many variable candidates. Variable classification establishes the feasible set for estimation. When a variable selection procedure searches the best subset within the feasible set for estimation, the best subset could always be accepted.

KEY WORDS

Best Subset
Econometric Estimation
Variable Selection Procedure
Variable Classification
Feasible Set for Estimation

1. INTRODUCTION

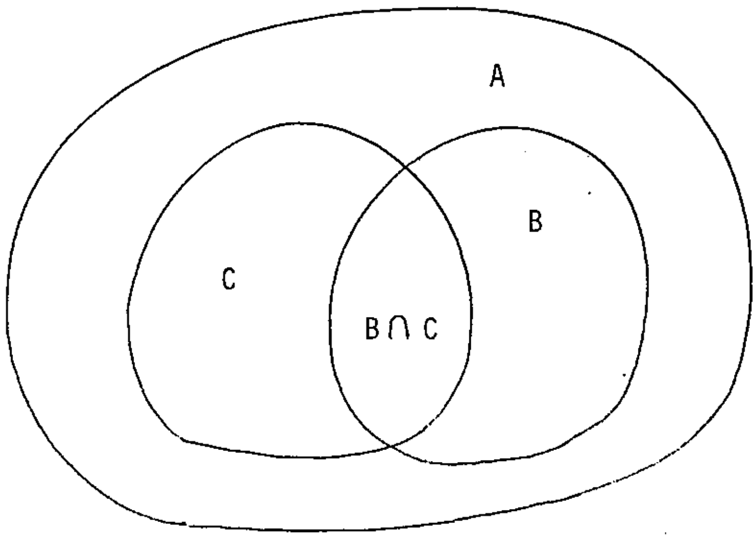
Various variable selection procedures have been proposed to automatically find the best subset. For example, all possible regression procedure, stepwise regression procedure, backward elimination procedure,

+ This research is partially supported by a Grant-in-Aid for Developmental Scientific Research from the Ministry of Education, Science and Culture, 1981.

forward selection procedure and stagewise regression procedure [4,7,8,22] are well known. Computer program packages [6,13,17,20] which can handle these variable selection procedures have been developed. In spite of these facts, it is rare in the field of economics to use these variable selection procedures. One reason is that most variable selection procedures so far developed are concerned only with completely optional and absolutely important (or core) explanatory variable candidates (the word 'candidate' is used for 'variable candidate' from here on). The second reason is that these variable selection procedures do not use the information available to a research field in which regression analysis or econometric estimation is made. The third reason is that these variable selection procedures can handle ordinary least squares but not other estimation methods. In applied econometric research, it is often observed that lagged candidates or endogenous candidates appear in economic models.

When an economist finds the best subset by estimating subset by subset or through trial and error processes, he takes economic meanings and characteristics of candidates into consideration and carefully avoids contradictory and meaingless subsets before estimation. Economic meaning and characteristics of candidates constitute a priori information about candidates. Thus, it is decisively important for a variable selection procedure to be able to generate only the subsets not only meaningful from the viewpoint of economics but also suitable for an estimation method to be utilized. When a variable selection procedure finds the best subset among the economically meaningful and econometrically estimable subsets, such a subset can be adopted as the economically and statistically best subset. Therefore, only economically meaningful and econometrically estimable subsets are worth being estimated and evaluated.

I would like to call the set of all 2^N-1 possible subsets the domain for estimation and the set of economically meaningful and econometrically estimable subsets the feasible set for estimation. Let us denote in Figure 1



A : Domain consisting of 2^N-1 possible subsets
B : Set of economically meaningful subsets
C : Set of econometrically estimable subsets
A ∩ B : Feasible set for estimation

Figure 1. Domain and Feasible Set for Estimation

the domain of all possible subsets by A, the set of economically meaningful subsets by B and the set of econometrically estimable subsets by C. The feasible set for estimation is expressed with $B \cap C$. When a variable selection procedure searches the best subset within $B \cap C$, such a subset is always accepted by an economist.
If $B \cap C = \emptyset$, there exists no economically and statistically best subset. Set A-B consists of economically meaningless subsets, while set A-C consists of econometrically unestimable subsets. For example, if subsets generated as a demand function for pork have no candidates representing income (including assets, if necessary) or relative price at all, an economist cannot accept them, because people cannot buy pork without money or they rationally buy pork by comparing the price of pork with those of the substitutes like beef, poultry, mutton, ham, sausage and/or fish.
Econometrically unestimable subsets are the subsets in which the numbers of selected candidates exceed the number of observations, selected candidates have multicollinearity or Almon-lag candidates are not selected continuously with respect to time lag (for example, A, A(-2), A(-5) are not selected continuously, because A(-1), A(-3) and A(-4) are missing where A(-i) stands for Almon-lag candidate with lag number i) and so on. When two stage least squares, constrained two stage least squares or limited information is used, underidentified subsets belong to set A-C. The case in which the number of selected candidates exceeds the number of observations often occurs in the field of

economics, when time series data are used.

The purpose of this paper is to show a method to generate only economically meaningful and econometrically estimable subsets or the feasible sets for various kinds of estimation methods.[1/]

2. BASIC CLASSIFICATION OF CANDIDATES

Let us introduce FORTRAN symbols /, <, >, <+, +>, </, />, <-, ->, <*, *>, (,), and : as variable classifiers and postulate that an explained variable and its candidates are loaded in a functional format such as Y=F(X), where Y and X stand for an explained variable and a set of candidates, respectively, a variable is expressed with a maximum of 8 FORTRAN characters except for the above variable classifiers, a lagged variable with time lag number i (>0) is expressed with its variable notation and negative i in parentheses, and $C denotes a constant term. For instance, variable HART is current, while variable HART(-2) is variable HART with time lag number 2. It is assumed that variable transformations are made before loading a

---

1/ Econometric Program Package OEPP, developed by the author, can generate meaningful and estimable subsets and find automatically the economically and statistically best subset which is called subequation in OEPP. The author is writing the manuals now.

functional format, if necessary. Candidates are classified by two kinds of a priori information about candidates.

First of all, candidates must be classified from the viewpoint of an estimation method to be applied. They are classified into the following five groups: (a) explanatory, (b) instrumental, (c) included predetermined, (d) right-hand side (RHS) endogenous, and (e) extraneous predetermined candidates.

Ordinary least squares (OLS), linearly constrained ordinary least squares (COLS), generalized least squares (GLS), linearly constrained generalized least squares (CGLS), Cochrane-Orcutt method (CO), Prais-Winston method (PW), Hildreth-Lu method (HL), ridge regression (R), Almon lag distribution method (A), Durbin-two-step methods for the first and second order of autocorrlation (DF and DS), and three pass least squares (3PLS) require explanatory candidates like Y=F(explanatory candidates).

Instrumental variable method (IV) requires the classification of candidates into explanatory and instrumental ones. Let us separate explanatory and instrumental candidates by a colon like Y=F(explanatory candidates:instrumental candidates). The econometric condition to generate a subset estimable with these methods is that the number of observations exceeds that of candidates selected.

Two stage least squares (2SLS), constrained two stage least squares (C2SL), and limited information method (LI) require the classification of candidates into included predetermined, RHS endogenous, and extraneous predetermined ones. 'Included predetermined candidates' here stand for exogenous or lagged endogenous ones included in an equation at hand. On the other hand, 'extraneous predetermined candidates' are here defined as exogenous or lagged endogenous ones which are not included in an equation at hand but included in at least one of the other equations and identities in a simultaneous equation model. Needless to say, an explained variable becomes a left-hand side (LHS) endogenous variable. Let us separate predetermined, RHS endogenous, and extraneous predetermined candidates with two colons like HART=F(included predetermined candidates:RHS endogenous candidates:extraneous predetermined candidates). The econometric conditions for the generation of estimable subsets are (i) the number of all included and extraneous predetermined candidates selected does not exceed that of observations, (ii) at least one RHS endogenous candidate is selected and (iii) the number of RHS endogenous candidates selected does not exceed that of extraneous predetermined candidates selected at the same time. A constant term is an included predetermined candidate, if any. Thus, included predetermined candidates are entered first and then RHS endogenous and extraneous predetermined candidates follow in a functional format (See Table 1).

Furthermore, candidates, whether explanatory, instrumental, included predetermined, RHS endogenous, or

Table 1. Variable Classification by A Priori Information Concerned with Econometrics

Methods	F.Formats	Exp. Var. and its Candidates
OLS, COLS, GLS, CGLS, CO, PW, R, A, DF, DS, 3PLS	Y=F(X)	Y=Explained Variable X=Set of Explanatory Candidates
IV	Y=F(X:V)	Y=Explained Variable X=Set of Explanatory Candidates V=Set of Instrumental Candidates
2SLS, C2SL, LI	Y=F(X:Z:W)	Y=LHS Endogenous Variable X=Set of Included Predetermined Candidates, if any Z=Set of RHS Endogenous Candidates W=Set of Extraneous Predetermined Candidates

Footnote (1) Candidates of X, candidates of Z and candidates of W are separated from each other with a blank, comma, "/", "<", ">", "</", "/>", "<+", "+>", "<*", "*>", "<-", "->", "(" or ")". (See Table 2) extraneous predetermined, need to be classified by a priori economic information into the following eight basic categories: (a) absolutely important (or core), (b) optionally important, (c) gradually important, (d) exclusively important, (e) gradually optional, (f) exclusively optional, (f) completely optional, and (g) fixed ones.

An absolutely important candidate is such that it must always be selected. Hence, all absolutely important candidates, if any, must be selected in all subsets. Let us put absolutely important candidates into /...../. One pair of slashes is enough for all estimation methods except for 2SLS, C2SL, and LI. However, 2SLS, C2SL, and LI need at most two pairs of slashes. The case in which two pairs of slashes are needed is that at least one included predetermined candidate and at least one extraneous predetermined candidate are absolutely important and at least one RHS endogenous candidate is not absolutely important. In this case, absolutely important extraneous predetermined candidates are not economically but econometrically meaningful. For instance, /X1,X2,...,XH/ implies that sub-subset X1,X2,...,XH are included in all economically meaningful subsets.

Optionally important candidates are such that at least one of them must be selected in order for a subset to become economically meaningful. Let us enter optionally important candidates into <.....>. If there are I optionally important candidates in <.....>, 2^N-1 (sub-)

subsets are economically meaningful. For instance, <X1,X2,X3> generates the following seven (sub-)subsets: (a) subset 1 (X1); subset 2 (X2); subset 3 (X3); subset 4 (X1,X2); subset 5 (X1,X3); subset 6 (X2,X3); and subset 7 (X1,X2,X3). It is possible to introduce more than one pair of <.....>.

Gradually important candidates are such that they must be gradually and increasingly selected beginning with the most important candidate among them. In other words, gradually important candidates have the ordering of their importance. It must be kept in mind in addition to the above characteristics that if none of them is selected in a subset, the subset loses economic meaning. To distinguish gradually important candidates from others, let us enter them into <+.....+> and postulate that the candidate in the left-most position is the most important among them, the candidate in the second left-most position is the second most important, ..., and then the candidate in the right-most position is the least important. For instance, <+X1,X2,...,XJ+> generates the following J (sub-)subsets: (1) subset 1 (X1); (2) subset 2 (X1,X2); (3) subset 3 (X1,X2,X3);, ..., (J) subset J (X1,X2,...,XJ). It is possible to introduce more than one pair of <+.....+>.

Exclusively important candidates are such that only one of them must be selected in order for a subset to become economically meaningful. In other words, the subsets which include none of them or more than one of them do not make economic sense. Let us put exclusively

Table 2. Variable Classification and Meaningful Sub-Subsets of Candidates in a Three-Candidate Case

Names	Classification	Meaningful Sub-Subsets 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; (7) C
Gradually Important	<+A,B,C+>	(1) A,B,C; (2) A,B; (3) A
Exclusively Important	</A,B,C/>	(1) A; (2) B; (3) C
Gradually Optional	<-A,B,C->	(1) A,B,C; (2) A,B; (3) A; (4) Empty (no selection)
Exclusively Optional	<*A,B,C*>	(1) A; (2) B; (3) C; (4) Empty (no selcetion)
Completely Optional	A,B,C	(1) A,B,C; (2) A,B; (3) A,C; (4) B,C; (5) A; (6) B; (7) C; (8) Empty (no selcetion)
Fixed	(D,E)	For instance, if B=(D,E), all B's above must be replaced with D,E.
Constant Term	$C	Always selected, if any

Footnotes: (1) Candidates A, B, C, D, and E can be expressed with more alphanumeric symbols with/without time lag numbers. (2) Any kinds of symbols can be used as variable classifiers in a new program or package, if there is no confusion between variable notations and classifiers. (3) A candidate can have two attributes like exclusively important and fixed one.

important candidates in </...../>. For instance, </X1,X2,...,XK/> generates the following K (sub-)subsets: (1) subset 1 (X1); (2) subset 2 (X2); ...; subset K (XK). It is possible to introduce more than one pair of </...../>.

The concept of gradually optional candidates is a simple extension of that of gradually important candidates. In addition to the characteristics of gradually important candidates, gradually optional candidates have the characteristic that a subset which includes none of them does not lose economic meaning. Let us put gradually optional candidates in a pair of less-than sign followed by minus and greater-than sign preceeded by minus like <-.....->. <-X1,X2,...,XL-> generates the following L+1 (sub-)subsets: (1) subset 1 (empty or no selection of candidates); (2) subset 2 (X1); (3) subset 3 (X1,X2);,..., (L+1) subset L+1 (X1,X2,...,XL). It is possible to introduce more than one pair of <-.....->.

Exclusively optional candidates correspond to exclusively important candidates. In addition to the characteristics of exclusively important candidates, exclusively optional candidates have the characteristic that a subset which includes none of them does not lose economic meaning. Let us enter exclusively optional candidates into <*.....*>. <*X1,X2,...,XM*> generates the following M+1 (sub-)subsets: (1) subset 1 (empty or no selection of candidates); (2) subset 2 (X1); (3) subset 3 (X2);,..., (M+1) subset M+1 (XM). It is possible to introduce more than one pair of <*.....*>.

Completely optional or optional candidates are an extension of optionally important candidates. In addition to the characteristics of optionally important candidates, completely optional candidates have the characteristic that a subset which includes none of them does not lose economic meaning. If there are N completely optional candidates, 2^N or 2^N-1 (sub-)subsets are economically meaningful, depending on the entries of non-completely-optional candidates. Since completely optional candidates are most often observed, we do not enter them in special brackets.

Fixed candidates are such that they are selected as a group but not individually. Let us enter fixed candidates into (.....). It is possible to introduce more than one (.....). A group of fixed candidates must be optionally important, gradually important, exclusively important, gradually optional, exclusively optional, or completely optional but not absolutely important.

Let us explain the relationship between an equation and a functional format in details. The following equation is to be estimated:

$$Y=a_0+a_1A+a_2B(-1)+a_3C+a_4D+a_5E \qquad (1)$$

where a_i's stand for coefficients.

If Equation (1) is estimated by one of these methods: OLS, COLS, GLS, CGLS, CO, PW, HL, R, A, or 3PLS, the following functional format is appropriate:

$$Y=F(\$C,A,B(-1),C,D,E) \tag{2}$$

Unless the estimation of all meaningful subsets is required, Format (2) yields only the coefficients of Equation (1) in a normal case where the inverse of the moment matrix exists. Suppose that A and B(-1) are absolutely important, C is completely optional, and D and E are optionally important. Then, the meaningful subsets are in addition to Equation (1) as follows:

$$Y=b_0+b_1A+b_2B(-1)+b_3C+b_4D \tag{3}$$

$$Y=c_0+c_1A+c_2B(-1)+c_3C+c_4E \tag{4}$$

$$Y=d_0+d_1A+d_2B(-1)+d_3D+d_4E \tag{5}$$

$$Y=e_0+e_1A+e_2B(-1)+e_3D \tag{6}$$

$$Y=f_0+f_1A+f_2B(-1)+f_3E \tag{7}$$

where b_i's, c_i's, d_i's, e_i's, and f_i's stand for coefficients. Equations (1) and (3) to (7) can be generated by the following format with a command for all possible meaningful subsets:

Y=F($C/A,B(-1)/C<D,E>) (8)

If Equation (1) is estimated by IV with instrumental candidates $C, Z1, Z2, Z3(-1), Z4, and Z5(-2) corresponding to explanatory candidates $C, A, B(-1), C, D, and E, respectively, the following format is required:

Y=F($C,A,B(-1),C,D,E:$C,Z1,Z2,Z3(-1),Z4,Z5(-2)) (9)

Equations (1) and (3) to (7) are generated and estimated with IV by the following format:

Y=F($C/A,B(-1)/C<D,E>:$C,Z1,Z2,Z3(-1),Z4,Z5(-2)) (10)

where the selection of instrumental candidates is determined by the classification of the corresponding explanatory candidates.

If we assume, without loss of generality, that $C, A, B(-1), and C are the included predetermined, D and E are the RHS endogenous, Z1, Z2, Z3, Z4(-1), Z5(-2), Z6, Z7, and Z8 are the extraneous predetermined candidates in Equation (1), then the following format is suitable:

Y=F($C,A,B(-1):C,D,E:Z1,Z2,Z3,Z4(-1),
Z5(-2),Z6,Z7,Z8) (11)

If the command for all possible meaningful and estimable subsets is used for Format (11), $2^{13}-1=8,191$ possible subsets become the domain for estimation with 2SLS, C2SL or LI. However, 6,304 subsets are meaningful and estimable. The number of meaningful and estimable subsets can be calculated as follows:

$$2^2*\left(\binom{3}{1}*\sum_{i=1}^{8}\binom{8}{i}+\binom{3}{2}*\sum_{i=2}^{8}\binom{8}{i}+\binom{3}{3}*\sum_{i=3}^{8}\binom{8}{i}\right)=6,304$$

where 2^2 comes from the selection of included predetermined candidates, $\binom{3}{i}$ for i=1,2,3 comes from the selection of RHS endogenous candidates and $\sum_{i=j}^{8}\binom{8}{j}$ for j=1,2,3 comes from the selection of extraneous predetermined candidates. Of course, unless non-constant candidates A, B(-1), and C are included in Equation (1), Format (11) becomes as follows:

Y=F($C:C,D,E:Z1,Z2,Z3,Z4(-1),Z5(-2),Z6,Z7,Z8) (12)

Suppose that all extraneous predetermined candidates are always used for the estimation of Equations (1) and (3) to (7). It is possible to expand the concept of an absolutely important candidate to extraneous predetermined candidates. From the viewpoint of estimation with 2SLS, C2SL or LI, all extraneous predetermined candidates can be treated as absolutely important, although they do not have economic meanings. The following format is available:

```
Y=F($C/A,B(-1)/:C<D,E>:/Z1,Z2,Z3,Z4(-1),
    Z5(-2),Z6,Z7,Z8/)                                   (13)
```

The first pair of slashes is introduced from the viewpoint of economics, while the second pair of slashes is introduced from the viewpoint of econometrics or an estimation method. The number of meaningful and estimable subsets derived from Format (13) is only 6. Format (13) is equivalent to Formats (11) and (14) to (18) without a command for all possible meaningful subsets.

```
Y=F($C,A,B(-1):D,E:Z1,Z2,Z3,...,Z8)                     (14)

Y=F($C,A,B(-1):C,D:Z1,Z2,Z3,...,Z8)                     (15)

Y=F($C,A,B(-1):C,E:Z1,Z2,Z3,...,Z8)                     (16)

Y=F($C,A,B(-1):D:Z1,Z2,Z3,...,Z8)                       (17)

Y=F($C,A,B(-1):E:Z1,Z2,Z3,...,Z8)                       (18)
```

If included predetermined candidate A and RHS endogenous candidate C are exclusively optional and included predetermined candidate B(-1) is absolutely important in Format (13), then the following format can be used:

```
Y=F($C/B(-1)/<*A:C*><D,E>:/Z1,Z2,Z3,Z4(-1),
    Z5(-2),Z6,Z7,Z8/)                                   (19)
```

Furthermore, I would like to classify an included predetermined candidate into a unique or non-unique one. A 'unique included predetermined' candidate is one which is included in an equation at hand but not in all other equations and identities in a simultaneous equation system. On the other hand, a 'non-unique included predetermined' candidate is one which is included in an equation at hand and also in at least one of the other equations and identities. If an included predetermined candidate is unique but not absolutely important and if extraneous predetermined candidates are treated as absolutely important, the number of extraneous predetermined candidates is never reduced in a meaningful and estimable subset. However, if an included predetermined candidate is not unique and absolutely important, the number of extraneous predetermined candidates is reduced by 1 in a meaningful and estimable subset in which the unique included predetermined candidate is not selected. Accordingly, it is suggested to add a non-unique included predetermined candidate to absolutely important extraneous predetermined candidates as a completely optional one and to treat as unestimable a subset in which the same candidate is selected both from a group of included predetermined candidates and from a group of extraneous predetermined ones. Thus, it is possible to avoid losing the data of a non-unique predetermined candidate by the above treatment.

If B(-1) is unique and completely optional and A is non-unique and comletely optional in Format (14), the following format can be loaded with the appropriate command:

Y=F($C/B(-1)/<*A:C*><D,E>:/Z1,Z2,Z3,Z4(-1),Z5(-2),
Z6,Z7,Z8/A) (20)

As far as 2SLS, C2SL or LI is concerned, it is suggested that OLS or COLS be applied to all equations in a simultaneous equation system first and then 2SLS, C2SL or LI be applied to them. An economist can roughly guess what the predetermined candidates in the system are.

3. NESTED VARIABLE CLASSIFICATION

Although there are many kinds of nested variable classifications which are theoretically interesting, an economist is interested only in economically reasonable nested variable classifications. Accordingly, I would like to refer only to economically possible nested classification. The following nested variable classifications are here referred to: (i) </.../>, <+...+>, <*...*> and <-...-> can be included in <...>, (ii) <...> and <+...+> can be included in </.../> and (iii) <...> and <-...-> can be included in <*...*>. Of course, (...) for

fixed candidates can be used anywhere.
Since it is quite rare that <+...+> and <-...-> need <...>, </.../> and <*...*> in it, I would like to omit referring to nested variable classifications concerning <+...+> and <-...->. It must be kept in mind that any kind of nested variable classification can be reduced to the level of basic variable classification.

3.1. NESTED VARIABLE CLASSIFICATION OF OPTIONALLY IMPORTANT CANDIDATES IN <...>

We postulate the following rules concerning outer pair of <...>:

(i) </.../>, <+...+>, <*...*>, <-...->, (...) and a single candidate are accepted to be entered into <...>.

(ii) The candidates in an inner pair of variable classifiers are treated like a single candidate from the viewpoint of outer <...>.

(iii) Inner </.../> and <+...+> must be always selected but others in outer <...> are optional. Inner <*...*> and <-...-> are treated like a completely optional candidate from the viewpoint of outer <...>.

(iv) At least one candidate must be selected due to outer <...>.

(v) Redundant meaningful subsets are ignored.

For instance, the following entries are accepted:

<A</B,C,D/><+E,F+><*G,H,I*><-J(K,L,M)->(N,O,P)> (21)

<<*A,B*><*C,D*>> (22)

<</A,B/></C,D/>E> (23)

We omit explaining Entry (21), because there are too many meaningful subsets. Entry (22) leads to <X,Y> where X=<*A,B*> and Y=<*C,D*>, which generates (1-1) X, (1-2) Y or (1-3) X,Y, i.e., (2-1) <*A,B*>, (2-2) <*C,D*> or (2-3) <*A,B*><C,D*>, i.e., (3-1) A, (3-2) B or (3-3) empty, (3-4) C, (3-5) D, (3-6) empty, (3-7) A,C, (3-8) A,D, (3-9) B,C, (3-10) B,D or (3-11) empty, i.e., (4-1) A, (4-2) B, (4-3) C, (4-4) D, (4-5) A,C, (4-6) A,D, (4-7) B,C or (4-8) B,D, because empty subsets must be excluded due to Rule (iv). The number of meaningful subsets is calculated as 3*3-1=8. 8 meaningful subsets are generated from Entry (22). <<*A,B*><*C,D*>> is equivalent to </A,B,C,D(A,C)(A,D)(B,C)(B,D)/>.

Entry (23) can be expanded as follows: <</A,B/></C,D/>E>= <X,Y,E> where X=</A,B/> and </C,D/> leads to (1-1) X,Y or (1-2) X,Y,E, i.e., (2-1) </A,B/></C,D/> or (2-2) </A,B/></C,D/>E, i.e., (3-1) A,C, (3-2) A,D, (3-3) B,C, (3-4) B,D, (3-5) A,C,E, (3-6) A,D,E, (3-7) B,C,E or (3-8) B,D,E. The number of meaningful subsets is calculated as 2*2*2=8. 8 meaningful subsets are generated from Entry (23). <</A,B/></C,D/>E> is equivalent to </(A,C)(A,D)(B,C)(B,D)(A,C,E)(A,D,E)(B,C,E)(B,D,E)/>.

Unless candidate E is included in Entry (23), nested variable classification is not needed in Entry (23), because <</A,B/></C,D/>>=</A,B/></C,D/>.

3.2. NESTED VARIABLE CLASSIFICATION OF EXCLUSIVELY IMPORTANT CANDIDATES IN </...../>

We postulate the following rules concerning outer </.../>:

(i) <...>, <-...->, (...) and a single candidate are accepted to be entered into </.../> which becomes outer </.../>.

(ii) Nested <...> can be entered into outer </.../>/

(iii) The candidates in an inner pair of variable classifiers are treated like a single candidate from the viewpoint of outer </.../>.

(iv) Only one candidate must be selected from the viewpoint of outer </.../>.

For example, the following entries are accepted:

</A<B,C>D/> (24)

</<</A,B/>C><</D,E/>F>/> (25)

</A<B(C,D,E)F><+G(H,I)+>J/> (26)

Entries (24) and (25) are explained as follows: </A<B,C>D/>=</A,X,D/> where X=<B,C> leads to the selections (1-1) A, (1-2) X or (1-3) D, i.e., (2-1) A, (2-2) <B,C> or (2-3) D, i.e., (3-1) A; (3-2) B, (3-3) C, (3-4) B,C or (3-5) D. The number of meaningful subsets is obtained by 1+3+1=5. 5 meaningful subsets (3-1) to (3-5) are generated from Entry (24). </A<B,C>D/> is equivalent to </A,B,C,D(B,C)/>.

</<</A,B/>C><</D,E/>F>/>=</X,Y/> where X=<</A,B/>C> and Y=<</D,E/>F> leads to the selections (1-1) X or (1-2) Y, i.e., (2-1) <V,D> where V=</A,B/> or (2-2) <W,F> where W=</D,E/>, i.e., (3-1) V, (3-2) V,D, (3-3) W or (3-4) W,F, i.e., (4-1) A, (4-2) B, (4-3) A,C, (4-4) B,C, (4-5) D, (4-6) E, (4-7) D,F or (4-8) E,F. The number of meaningful subsets is 2*2+2*2=8. Meaningful subsets (4-1) to (4-8) are generated from Entry (25). It is easily known that <<</A,B/>C><</D,E/>F>/> is equivalent to

</A,B,D,E(A,C)(B,C)(D,F)(E,F)/>. We omit explaining Entry (26).

3.3. NESTED VARIABLE CLASSIFICATION OF EXCLUSIVELY OPTIONAL CANDIDATES IN <*.....*>

We postulate the following rules concerning outer <*...*>:

(i) <...>, <-...->, (...) and a single candidate are allowed to be entered into <*...*> which becomes outer <...>.

(ii) Nested <...> can be entered into outer <*...*>.

(iii) The candidates in an inner pair of variable classifiers are treated like a single candidate from the viewpoint of outer <*...*>.

(iv) At most one candidate must be selected from the viewpoint of outer <*...*>.

For example, the following entries are accepted:

<*A<B,C(D,E)><-F,G->(I,J,K)*> (27)

<*<</A,B/>C><</D,E/>F>*> (28)

<*<<*A,B*>C><<*D,E*>F>*> (29)

Entry (28) is explained as follows: <*<</A,B/>C><</D,E/>F>*>=<*X,Y*> where X=<</A,B/>C> and Y=<</D,E/>F> leads to the selections (1-1) empty, (1-2) X, or (1-3) Y, i.e., (2-1) empty; (2-2) <V,C> where V=</A,B/> or (2-3) <W,F> where W=</D,E/>, i.e., (3-1) empty, (3-2) V, (3-3) V,C, (3-4) W or (3-5) W,F, i.e., (4-1) empty, (4-2) A, (4-3) B, (4-4) A,C, (4-5) B,C, (4-6) D, (4-7) E, (4-8) D,F or (4-9) E,F. The number of meaningful subsets is calculated as 2*2+2*2+1=9. 9 meaningful subsets (4-1) to (4-9) are generated from Entry (28). <*<</A,B/>C><</D,E/>F>*> is equivalent to <*A,B,D,E(A,C)(B,C)(D,F)(E,F)*>. We omit explaining Entries (27) and (29).

4. EXAMPLES IN ECONOMIC FIELD

Let us give some examples prevailing in the field of economics. Income variables Y, Y(-1) and Y(-2) are gradually important candidates for a demand function of pork and other income variables YY, YY(-1) and YY(-2) are also gradually important candidates, where YY=(Y+Y(-1))/2.

Candidates H, H(-1) and H(-2) of household expenditures are alternative candidates for two kinds of income variables Y's and YY's. In this case, <+Y,Y(-1),Y(-2)+>, <+YY, YY(-1),YY(-2)+> or <+H,H(-1),H(-2)+> must be used in a functional format for pork demand. In order to allow this case, we accept the following entry:

</<+Y,Y(-1),Y(-2)+><+YY,YY(-1),YY(-2)+><+H,H(-1),H(-2)+>/> (30)

When we put A=<+Y,Y(-1), Y(-2)+>, B=<+YY,YY(-1),YY(-2)+> and C=<+H,H(-1),H(-2)+>, Entry (30) is reduced to </A,B,C/> which leads to the selection of (1-1) A, (1-2) B or (1-3) C, i.e., (2-1) <+Y,Y(-1),Y(-2)+>, (2-2) <+YY,YY(-1), YY(-2)+> or (2-3) <+H,H(-1),H(-2)+>, i.e., (3-1) Y, (3-2) Y,Y(-1), (3-3) Y,Y(-1),Y(-2), (3-4) YY, (3-5) YY,YY(-1), (3-6) YY,YY(-1),YY(-2), (3-7) H, (3-8) H,H(-1) or (3-9) H,H(-1),H(-2). Exactly 9 economically meaningful subsets are generated from Entry (30). There is no contradiction in the above selection of candidates.

If the above candidates are regarded as optionally important, the following entry can be used :

</<Y,Y(1),Y(-2)><YY,YY(-1),YY(-2)><H,H(-1),H(-2)>/> (31)

By regarding each group of optionally important candidates as a single candidate like A=<Y,Y(-1),Y(-2)>, B=<YY,YY(-1), YY(-2)> and C=<H,H(-1),H(-2)>, Entry (31) is reduced to

</A,B,C/> which leads to the selection of (1-1) A, (1-2) B or (1-3) C, i.e., (2-1) <Y,Y(-1),Y(-2)>, (2-2) <YY,YY(-1), Y(-2)> or (2-3) <H,H(-1),H(-2)>, i.e., (3-1) Y, (3-2) Y(-1), (3-3) Y(-2), (3-4) Y,Y(-1), (3-5) Y,Y(-2), (3-6) Y(-1),Y(-2), (3-7) Y,Y(-1),Y(-2), (3-8) YY, (3-9) YY(-1), (3-10) YY(-2), (3-11) YY,YY(-1), (3-12) YY,YY(-2), (3-13) YY(-1),YY(-2), (3-14) YY,YY(-1),YY(-2), (3-15) H, (3-16) H(-1), (3-17) H(-2), (3-18) H,H(-1), (3-19) H,H(-2), (3-20) H(-1),H(-2) or (3-21) H,H(-1),H(-2). The number of meaningful subsets is calculated as $2^3-1+2^3-1+2^3-1=21$. 21 economically meaningful subsets can be generated by Entry (31).

Sometimes, there are no exact data available to a variable but there are a few kinds of alternative data which can be regarded as the data of the variable in question. For instance, there are no data for beef retail price, but there are the data of sirloin beef retail price and rib beef retail price. It is possible to generate new data by variable transformations of these prices such as the average of these prices. It happens that the researcher does not know which is most suitable for beef retail price. Let us assume that pork, poultry and fish retail prices are available and denoted as P, PL and F, respectively, while sirloin and rib beef retail prices are denoted as BS and BR, respectively. By introduing transformed candidates BSP=BS/P, BRP=BR/P, BSRP=(BS+BR)/(2*P), PLP=PL/P and FP=F/P, we can use the following entry for a pork demand function:

<<*BSP,BRP,BSRP*>PLP,FP> (32)

If we put A=<*BSP,BRP,BSRP*>, Entry (32) is reduced to <A,PLP,FP> which leads to the selection of (1-1) A, (1-2) PLP, (1-3) FP, (1-4) A,PLP, (1-5) A,FP, (1-6) PLP,FP or (1-7) A,PLP,FP, i.e., (2-1) <*BSP,BRP,BSRP*>, (2-2) PLP, (2-3) FP, (2-4) <*BSP,BRP, BSRP*>,PLP, (2-5) <*BSP,BRP, BSRP*>,FP, (2-6) PLP,FP or (2-7) <*BSP,BRP,BSRP*>,PLP,FP, i.e., (3-1) BSP, (3-2) BRP, (3-3) BSRP, (3-4) PLP, (3-5) FP, (3-6) BSP,PLP, (3-7) BRP,PLP, (3-8) BSRP,PLP, (3-9) BSP,FP, (3-10) BRP,FP, (3-11) BSRP,FP, (3-12) PLP,FP, (3-13) BSP,PLP,FP, (3-14) BRP,PLP,FP or (3-15) BSRP,PLP,FP. The number of meaningful subsets is calculated as $(3+1)*2^2-1=15$. 15 economically meaningful subsets can be derived from Entry (32).

If an economist has a priori information that beef is a strong substitute for pork so that relative retail price of beef and pork is definitely needed in a meaningful pork demand function, he can enter the following entry:

<</BSP,BRP,BSRP/>PLP,FP> (33)

By putting A=</BSP,BRP,BSRP/>, Entry (33) is reduced to <A,PLP,FP> which leads to the selection of (1-1) A, (1-2) A,PLP, (1-3) A,FP or (1-4) A,PL,FP, i.e., (2-1) BSP, (2-2) BRP, (2-3) BSRP, (2-4) BSP,PLP, (2-5) BRP,PLP, (2-6) BSRP,PLP, (2-7) BSP,FP, (2-8) BRP,FP, (2-9) BSRP,FP, (2-10) BSP,PLP,FP, (2-11) BRP,PLP,FP or (2-12) BSRP,PLP,FP. The number of meaningful subsets is calculated as $3*2^2=12$. 12 economically meaningful subsets can be derived from Entry (33).

If there are no exact data available for pork retail price but there are two kinds of alternative original data which can be regarded as the candidates of data for pork retail price, a more complicated entry is needed. Let us denote the sirloin and rib pork retail prices by PS and PR, respectively. For simplicity, we do not generate new data by variable transformations. By introducing transformed candidates BSPS=BS/PS, BRPS=BR/PS, BSRPS=BSR/PS, PLPS=PL/PS, FPS=F/PS, BSPR=BS/PR, BRPR=BR/PR, BSRPR=BSR/PR, PLPR=PL/PR and FPR=F/PR, we can write the following entry:

</<<*BSPS,BRPS,BSRPS*>PLPS,FPS><<*BSPR,BRPR,BSRPR*>PLPR, FPR>/> (34)

If we put A=<<*BSPS,BRPS,BSRPS*>PLPS,FPS>, B=<<*BSPR, BRPR,BSRPR*>PLPR,FPR>, C=<*BSPS,BRPS,BSRPS*> and D=<*BSPR, BRPR,BSRPR*>, we can reduce Entry (34) to </A,B/> which leads to the selection of (1-1) A or (1-2) B which are reduced to (2-1) <C,PLPS,FPS> or (2-2) <D,PLPR,FPR> which leads to the selection of (3-1) C, (3-2) PLPS, (3-3) FPS, (3-4) C,PLPS, (3-5) C,FPS, (3-6) PLPS,FPS, (3-7) C,PLPS,FPS, (3-8) D, (3-9) D,PLPR, (3-10) D,FPR, (3-11) D,PLPR, (3-12) D,FPR, (3-13) PLPF,FPR or (3-14) D,PLPR,FPR, i.e., (4-1) BSPS, (4-2) BRPS, (4-3) BSRPS, (4-4) PLPS, (4-5) FPS, (4-6) BSPS,PLPS, (4-7) BRPS,PLPS, (4-8)

BSRPS,PLPS, (4-9) BSPS,FPS, (4-10) BRPS,FPS, (4-11) BSRPS,FPS, (4-12) PLPS,FPS, (4-13) BSPS,PLPS,FPS, (4-14) BRPS,PLPS,FPS, (4-15) BSRPS,PLPS, FPS, (4-16) BSPR, (4-17) BRPR, (4-18) BSRPR, (4-19) PLPR, (4-20) FPR, (4-21) BSPR,PLPR, (4-22) BRPR,PLPR, (4-23) BSRPR,PLPR, (4-24) BSPR,FPR, (4-25) BRPR,FPR, (4-26) BSRPER,FPR, (4-27) BSPR,PLPR, FPR, (4-28) PLPR,FPR, (4-29) BRPR,PLPR,FPR or (4-30) BSRPR,PLPR,FPR. The number of meaningful subsets is calculated as $((3+1)*2^2-1)*2=30$. 30 economically meaningful subsets are generated from Entry (34).

If the researcher has a priori information that beef is a strong substutute for pork so that the relative retail price of beef and pork is important, he can use the following entry:

</<</BSPS,BRPS,BSRPS/>PLPS,FPS><</BSPR,BRPR,BSRPR/>PLPR, FPR>/> (35)

By putting A=<</BSPS,BRPS,BSRPS/>PLPS,FPS>, B=<</BSPR,BRPR, BSRPR/>PLPR,FPR>, C=</BSPS,BRPS,BSRPS/> and D=</BSPR,BRPR, BSRPR/>, Entry (35) is reduced to </A,B/> which leads to the selection of (1-1) A or (1-2) B, i.e., (2-1) <C,PLPS,FPS> or (2-2) <D,PLPR,FPR>, i.e., (3-1) C, (3-2) C,PLPS, (3-3) C,FPS, (3-4) C,PLPS,FPS, (3-5) D, (3-6) D,PLPR, (3-7) D,FPR or (3-8) D,PLPR,FPR, i.e., (4-1) BSPS, (4-2) BRPS, (4-3) BSRPS, (4-4) BSPS,PLPS, (4-5) BRPS,PLPS, (4-6) BSRPS,PLPS, (4-7) BSPS,FPS, (4-8) BRPS,FPS, (4-9) BSRPS,FPS, (4-10) BSPS,PLPS,FPS, (4-11) BRPS,PLPS,FPS,

(4-12) BSRPS,PLPS,FPS, (4-13) BSPR, (4-14) BRPR, (4-15) BSRPR, (4-16) BSPR,PLPR, (4-17) BRPR,PLPR, (4-18) BSRPR,PLPR, (4-19) BSPR,FPR, (4-20) BRPR,FPR, (4-21) BSRPR,FPR, (4-22) BSPR,PLPR,FPR, (4-23) BRPR,PLPR,FPR or (4-24) BSRPR,PLPR,FPR. The number of meaningful subsets is calculated as $3*2^2+3*2^2=24$. In this case, 24 economically meaningful subsets are generated from Entry (35). The a priori information that beef and pork are close substitutes for each other reduces the number of economically meaningful subsets from 30 of Entry (34) to 24 of Entry (35). Thus, certain a priori information about candidates is essentially important.

Let us load a pork demand function by introducing income factor, price factor and inertia effect factor of eating habits as follows:

PD=F(\$C</<+Y,Y(-1),Y(-2)+><+YY,YY(-1),YY(-2)+><+H,HH(-1),
H(-2)+>/></<</BSPS,BRPS,BSRPS/>PLPS,FPS><</BSPR,BRPR,
BSRPR/>PLPR,FPR>/><-PD(-1),PD(-2)->) (36)

where PD stands for demand for pork and <-PD(-1),PD(-2)-> implies inertia effect. Format (36) guarantees the selection of at least one candidate from income factor and at least one candidate from price factor. The number of economically meaningful subsets of Format (36) is 9*24*3=648. On the other hand, the number of all possible subsets is $2^{21}-1$=2,097,151. The feasible set for estimation consists of 648 subsets among the domain of

2,097,151 subsets in a normal case. The relative size of the feasible set to the domain for estimation is 0.0309 %!

5. VARIABLE CLASSIFICATION AND SOME ESTIMATION METHODS

5.1 LINEARLY CONSTRAINED ESTIMATION

It is likely that a linear constraint is imposed on the coefficients of candidates or a linear hypothesis is tested. It may not be difficult to handle linearly constrained estimation by using the ideas of absolutely important, optionally important, gradually important, and exclusively important candidates. Suppose that constant returns to scale prevail in the production of poultry which is determined by the amounts of labor and capital used. It is likely that some candidates represent labor, while others represent capital. For instance, the simple sum of working hours of men (farmers), women (their wives), children, and hired labor can be a labor candidate, but the weighted sum of their working hours is also another labor candidate, where the weights of 1.0 for men, 0.8 for women, 0.5 for children, and 1.0 for hired labor reflect their physical powers. On the other hand, the real value of cages and equipment for water and feed is a capital candidate, whereas another capital candidate includes the real value of trucks used for poultry production in addition to the real value of cages and equipment for water and feed. Let us denote the labor candidates by L1 and L2, the capital candidates by K1 and K2, production quantity by Q, and transform these candidates into LL1=LOG(L1), LL2=LOG(L2), LK1=LOG(K1), LK2=LOG(K2), and LQ=LOG(Q). Cobb-Douglas production function is assumed. Since LL1 is alternative for LL2 and LK1 is also alternative for LK2, four subsets LL1 and LK1, LL1 and LK2, LL2 and LK1, and LL2 and LK2 are economically meaningful. In this case, a linear constraint representing constant returns to scale is given by the following equations with respect to coefficients:

$$1=a_1+a_3,\ 1=a_1+a_4,\ 1=a_2+a_3,\ \text{or}\ 1=a_2+a_4 \qquad (37)$$

where a_1, a_2, a_3, and a_4 stand for the coefficients of LL1, LL2, LK1, and LK2, respectively. By loading coefficients 1, 0, 1, 1, 1, and 1 in such a way to correspond to LQ, \$C, LL1, LL2, LK1, and LK2, respectively in addition to Format (38) under which the coefficients of the constraint are written in a corresponding style, we can make the computer generate, estimate, and even evaluate all four meaningful subsets satisfying constant returns to scale.

```
LQ=F($C</LL1,LL2/></LK1,LK2/>)                    (38)
 1   0    1   1     1   1
```

The treatment of candidates as exclusively important does not make the constraint of constant returns to scale ineffective. Therefore, constradictory or meaningless subsets are not generated.

5.2 LAGGED REGRESSAND

A lagged regressand(s) is often included in an economic model such as a partial adjustment or expectation formation model. In this case, a lagged regressand(s) must be selected in all meaningful subsets to allow for a proper estimation method to be continuously applied. By treating a lagged regressand(s), for example Y(-1), as absolutely important, we can handle this problem. A format for a lagged regressand(s) is as follows:

Y=F($C/Y(-1)/X) (39)

where X stands for a set of other explanatory candidates. Cochrane-Orcutt method, Hildreth-Lu method, Durbin-two-step method for the first-order autocorrelation, or three pass least squares can be applied.

When Durbin-two-step method for the second-order autocorrelation is applied, the following format is a typical one:

Y=F($C/Y(-1),Y(-2)/X) (40)

5.3 ALMON LAG DISTRIBUTION METHOD

Let us denote Almon-lag candidates by X, X(-1), X(-2), ..., X(-T) and a set of non-Almon-lag candidates, if any, by Z, where $T>1$. Since both candidates are explanatory, it is important to separate them and notify the computer of the position of Almon-lag candidates in a functional format. Let us postulate that non-Almon-lag candidate Z is entered first and then Almon-lag candidates follow such as

Y=F($C,Z,X,X(-1),...,X(-T)) (41)

If the degree of a polynomial is S where $0<S<T$, meaningful subsets suitable for Almon lag distribution method must include one of the following sub-subsets: (a) X,X(-1),..., X(-(S+1)); (b) X,X(-1),...,X(-(S+2)); ...; (c) X,X(-1),..., X(-T). By loading Format (42) and selecting only the above sub-subsets, we can generate meaingful subsets for Almon lag distribution method.

Y=F($C,Z<+X,X(-1),X(-2),...,X(-T)+>) (42)

It is possible to automatically change the degree of a polynomial from min(S,t-1) to 1, where t stands for the maximum time lag number of Almon-lag candidates chosen in a meaningful subset and $2\leq t\leq T$. For instance, if $S=T-1$ and if non-Almon-lag candidates Z are not included in Format

(42), Almon lag distribution method must be applied T*(T-1)/2 times.

6. CONCLUSION

In the field of economics, the economically and statistically best subset is usually obtained through trial and error processes. The main reason is that variable selection procedures so far developed can not take economic information into consideration. This paper showed that candidates are basically classified into 8 categories from the viewpoint of economics and into 5 categories from the viewpoint of economterics. The classification of candidates enables the computer to generate only economically meaningful and economterically estimable subsets which compose the feasible set for estimation. If the best subset is searched within the feasible set for estimation, it can be accepted by an economist.

Although variable classification is discussed concerning economics, it can be utilized for regression analyses and econometric analyses in business science, politometrics, sociometrics, agriculture, medical science and engineering.

REFERENCES

[1] AKAIKE, H.(1973). Information theory and an extension of the maximum likelihood principle. The second international symposium on information theory. Budapest.

[2] ALLEN, D,M,(1971). Mean square error of prediction as a criterion for selecting variables. Technometric, 13, 469-475.

[3] BEALE, E.M.L., KENDALL, M.G. and MANN, D.W.(1967). The discarding of variables in multivariate analysis, Boimetrika,54,357-366.

[4] DANIEL, C. and WOOD, F.S.(1971). Fitting equations to data. John Wiley & Sons.

[5] DHRYMES, P.J., HOWREY, E.P., HYMANS, S.H., KMENTA, J., LEAMER, E.E., QUANDT, R.E, RAMSEY, J.B., SHAPIRO, H.T, and ZARNOWITZ, V.(1972). Criteria for evaluation of econometric model. Annals of economic and social measurements,291-324.

[6] DIXTON, W.J.(1977). BMDP. Health Sciences and Computing Facility, Department of Biomathematics, School of Medicine, University of California.

[7] DRAPER, N.R. and SMITH, H.(1980). Applied Regression Analysis. 2nd Edition. John Wiley & Sons.

[8] FURNIVAL, G.M.(1971). All possible regressions with less computation. Technometrics,13,403-408.

[9] GARSIDE, M.J.(1965). The best subset in multiple regression analysis. Applied Statistics,14,196-200.

[10] GORMAN, J.W. and TOMAN, R.J.(1966). Selection of variables for fitting equations to data. Technometric, 8,27-51.

[11] HOCKING, R.R. and LESLIE, R.J.(1967). Selection of the best subset in regression analysis. Technometrics,9, 531-540.

[12] HOCKING, R.R.(1976). The analysis and selection of variables in linear regressions. Biometrics,32,1-49.

(13) HOLWIG, J.T.(1980). SAS. SAS Institute, Raleigh, N.C.

(14) LA MOTTE, L.R. and HOCKING, R.R.(1970) Computational efficiency in the selection of regression variables. Technometrics,12,83-93.

(15) MALLOWS, C.L.(1973). Some comments on C_p. Technometrics,15,661-675.

(16) MANTEL, N.(1970). Why stepdown procedures in variable selection. Technometrics,12,621-625.

(17) NIE, N.H., HULL, C.H., JENKINS, J.C., STEINBRENNER, K. and BENT, D.H.(1980). SPSS. National Opinion Research Center, University of Chicago.

(18) ONISHI, H.(1980). A time-, labor-, and resource-saving as well as cost-reducing software method for estimating a large-scale simultaneous equations model. PP-80-12. International Institute for Applied Systems Analysis. Laxenburg, Austria.

(19) ONISHI., H.(1982). OEPP. Institute of Socio-Economic Planning, University of Tsukuba.

(20) RYAN, T.A.J., JOINER, B.L. and RYAN, B.F.(1980). MINITAB-80. Statistics Department, Pennsylvania State University.

(21) SEBER, G.A.F.(1977). Linear regression analysis. John Wiley & Sons.

(22) SCHATZOFF, M., TSAO, R. and FIENBERG, S.(1968). Efficient calculation of all possible regressions. Technometrics,10,769-779.