No. 144 (82-11)

An Economic Approach to the Best Subset in Regression Analysis

by

Haruo ONISHI

March 1982

An Economic Approach to the Best Subset in Regression Analysis+

Haruo Onishi

Institute of Socio-Economic Planning
University of Tsukuba

This paper proposes a variable selection procedure to search the economically and statistically best subset in the feasible set for estimation which consists of economically meaningful subsets. In general, explanatory variable candidates are classified into eight basic groups through economic information. They are absolutely important, optionally important, gradually important, exclusively important, gradually optional, exclusively optional, completely optional, and fixed groups. Variable classification leads to the establishment of the feasible set for estimation. The economically and statistically best subset can be found through economic and statistical criteria in the feasible set, if it exists. This procedure can save much time, labor, and other resources, especially on the estimation of a large-scale economic model.

KEY WORDS

Best Subset
Linear Regression
Variable Selection Procedure
Variable Classification

1. INTRODUCTION

Stepwise regression procedure, forward selection procedure, backward elimination procedure, stagewise regression procedure [4,5,7,16] have been developed.

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

Efficient methods for generating all possible regressions (8,22) have been proposed. Furthermore, several criteria such as Akaike's information criterion, Mallow's c_p statistic and Allen's mean square error of prediction (1,2,9,10,11,15) have been discussed. Computer program packages for these variable selection procedures (6,13,17, 20) have been developed. In spite of these efforts, economists do not use these procedures much in their research, because of lack of economic consideration. If estimation is repeated equation by equation until a reasonable equation is obtained, it is time-consuming, labor-consuming, and other-resource-consuming (output paper, electricity, etc.). It is often observed that the statistically best subset efficiently estimated by these variable selection procedures is not necessarily adopted from the viewpoint of economics. For instance, if the sign of the coefficient of real income variable in a beef demand equation which is the statistically best is negative, it contradicts with our rational economic behavior hypothesis or economic theory based on the hypothesis. Hence, an economist would not accept this beef demand equation as the best from the viewpoint of economics.

Usually, an economist carefully selects explanatory variables, estimates the equation, checks whether or not the signs and/or effective ranges of all or some of the coefficients are consistent with rational behavior of 'economic men' in an advanced economy, applies T- or F-test and/or Durbin-Watson statistic test to the coefficients,

compare estimates with the corresponding data, and examines the coefficient of determination adjusted by the number of explanatory variables. He repeats this procedure or part of it until he can obtain a reasonable estimated equation. In an economic problem, the signs and/or magnitudes of coefficients of some explanatory variables are often known through field surveys or economic theories before estimation.

The purpose of this article is to show a procedure for finding the statistically as well as economically best subset among all economically meaningful subsets generated from a set of explanatory variable candidates (word 'candidate' is used for 'explanatory variable candidate' from here on) specified for an explained variable in question. The procedure is embodied in Computer Package OEPP[1/] and has been proven to work well.

2. GENERATION OF ECONOMICALLY MEANINGFUL EQUATIONS

If basic assumptions about estimation are met, an equation can be estimated. When an equation is estimated

1/ OEPP has been developed for education as well as research by the author. It deals with data management, econometric estimation, economic simulation, and input-output analysis and consists of about 23,000 steps in total in Fortran, composed of the main program and 181 subroutines. OEPP can automatically handle trial and error processes not only in estimation but also in simulation. The author is writing a manual at present.

with a set of loaded explanatory variables, an economist carefully selectes these explanatory variables and judges whether or not the estimated equation is reasonable by using his knowledge. His knowledge has been obtained through economic theories, human rational behavior hypothesis, field surveys, empirical studies, psychology, statistics, applied mathematics, etc. It is quite important to let him fully utilize his knowledge which is here called a priori information.

A priori information is used in two ways. Careful selection of appropriate candidates requires a priori information about the characteristics of candidates. Furthermore, correct evaluation or judgement for the selection of the best among many estimated subsets for an explained variable in question may need a priori information about the coefficients of candidates. If it is possible to embed in a package the algorithms dealing with these kinds of a priori information, continuous estimation and evaluation will shorten the research period, save labor and other resources , reduce the research cost, and even improve the quality of research.

Let us assume that an explained variable and its candidates are loaded by a functional format under the specification of linearity (or non-linearity) and all possible functional formats leading to all possible subsets or equations are derived from a set of candidates specified in the functional format. For instance, suppose that explained variable HART is estimated in a linear form with

In general, if N non-constant candidates are considered to be equally relevant to an explained variable, then 2^N-1 subsets are estimated in all possible regressions. However, if some candidates are considered to have stronger relevance to the explained variable by a priori information, the subsets which are consistent with this a priori information must be chosen among 2^N-1 subsets.

By a priori information about candidates we can basically classify candidates having different degrees of relevance to the explained variable into eight groups. These eight groups are those of (a) absolutely important, (b) optionally important, (c) gradually important, (d) exclusively important, (e) gradually optional, (f) exclusively optional, (g) completely optional or optional, and (h) fixed candidates (See Table 1). Absolutely important candidates are often called core candidates. Let us call symbols /, <, >, <+, +>, </, />, <-, ->, <*, *>, (, and) variable classifiers. Symbols (and) can be used as variable classifiers as well as time lag indicators.

First of all, an absolutely important candidate, if any, is such that any subset which does not include it does not make sense from the viewpoint of economics. Therefore, if H non-constant candidates are considered to be absolutely important for an explained variable, these H candidates must always be selected. If a subset does not include all H absolutely important candidates, it is

regarded as meaningless. Absolutely important candidates are entered between a pair of slashes like /...../ to distinguish them from other candidates. Only one pair of slashes can be used for ordinary least squares.

Optionally important candidates, if any, are such that (a) they are considered to explain an aspect of the movement of an explained variable and (b) a subset which includes none of them does not make sense from the viewpoint of economics. In other words, if I candidates are considered to be optionally important for an explained variable, at least one of them must be selected. A subset which includes none of the I optionally important candidates is regarded as meaningless. In order to distinguish optionally important candidates from other candidates, let us put them between a pair of less-than and greater-than signs like <.....>. It is possible that there is more than one group of optionally important candidates. In this case, a group of optionally important candidates must be considered to explain a different aspect of the movement of an explained variable from other groups of optionally important candidates and also from other candidates.

Gradually important candidates are such that (a) they must be gradually and increasingly selected beginning with the most important candidate among them or they must be gradually and decreasingly eliminated beginning with the least important candidate among them and (b) a subset which does not include the most important candidate among them

does not make sense from the viewpoint of economics. Concretely speaking, if J candidates X1, X2, ..., XJ are gradually important in the sense that X1 is more important than X2 more important than X3 ...more important than XJ, then they must be selected as one of the following J (sub-)subsets: subset 1 (X1); subset 2 (X1, X2); subset 3 (X1, X2, X3);...; subset J (X1, X2, ..., XJ). To distinguish gradually important candidates from other candidates, let us enter them into a pair of less-than sign followed by plus and greater-than sign preceeded by plus like <+.....+> and postulate that the most important candidate among them is entered in the left-most position of <+.....+>, the second most important being entered in the second left-most position, ..., and the least important being entered in the right-most position. In the above example, <+X1,X2,...,XJ+> is appropriate. It is possible that there is more than one group of gradually important candidates. In this case, a group of gradually important candidates must be considered to explain a different aspect of the movement of an explained variable from other groups of gradually important candidates and also from other candidates. The idea of gradually important candidates can be used for polynomial regression and Almon distributed lag method.

Exclusively important candidates, if any, are such that (a) they are considered to explain an aspect of the movement of an explained variable and (b) a subset which includes none or more than one of them does not make sense

from the viewpoint of economics. In other words, if K candidates are considered to be exclusively important for an explained variable, only one of them must be selected. A subset which includes none or more than one of the K exclusively important candidates is regarded as meaningless. To distinguish exclusively important candidates from other candidates, they are entered between a pair of less-than sign followed by slash and greater-than sign preceeded by slash like </...../>. It is possible that there is more than one group of exclusively important candidates. In this case, a group of exclusively important candidates must be considered to explain a different aspect of the movement of an explained variable from other groups of exclusively important candidates and also from other candidates.

Gradually optional candidates are such that (a) they are gradually and increasingly selected from the most important among them or they are gradually and decreasingly eliminated from the least important candidate among them, once they are selected and (b) any subset which does not include the most important candidate among them does not lose sense from the viewpoint of economics. It is obvious that gradually optional candidates are the extension of gradually important candidates and they are not so decisively important as gradually important candidates. If M candidates Y1, Y2, ..., YM are gradually optional in the sense that Y1 is more important than Y2 more important than Y3 .... and YM is the least important, then they are

selected as one of M+1 (sub-)subsets: subset 1 (empty or none of them is selected); subset 2 (Y1); subset 3 (Y1, Y2); subset 4 (Y1, Y2, Y3);...; subset M+1 (Y1, Y2, Y3, ..., YM). Let us enter gradually optional candidates into a pair of less-than sign followed by minus and greater-than sign preceeded by minus like <-.....-> and postulate that the most important candidate among them is entered in the left-most position of <-.....->, the second most important being entered in the second left-most position, ..., and the least important being entered in the right-most position. The above example is expressed as <-Y1,Y2, ...,YM->. It is possible that there is more than one group of gradually optional candidates. In this case, a group of gradually optional candidates must be considered to explain a different aspect of the movement of an explained variable from other groups of gradually optional candidates and also from other candidates.

Exclusively optional candidates are such that (a) they are considered to explain an aspect of the movement of an explained variable and (b) a subset which includes more than one of them does not make sense from the viewpoint of economics. In other words, if L candidates are considered to be exclusively optional for an explained variable, at most one of them (or none or only one of them) must be selected. A subset which includes more than one of the L exclusively optional candidates is regarded as meaningless. It shows that exclusively optional candidates are the extension of exclusively important candidates and they are

not so decisively important as exclusively important candidates. A group of exclusively optional candidates is entered between a pair of less-than sign followed by asterisk and greater-than sign preceeded by asterisk like <*.....*>. It is possible that there is more than one group of exclusively optional candidates. In this case, a group of exclusively optional candidates must be considered to explain a different aspect of the movement of an explained variable from other groups of exclusively optional candidates and also from other candidates.

Completely optional or optional candidates, if any, are such that (a) it is expected that they may explain some aspects of the movement of an explained variable, (b) whether or not they are included in subsets does not matter, and (c) their importance is examined after estimation. Any number of completely optional candidates can be selected. In other words, if there are N completely optional candidates for an explained variable in question, at most N completely optional candidates can be selected. It is clear that completely optional candidates are the extension of optionally important candidates. If none of the absolutely important, optionally important, gradually important, and exclusively important candidates is included in a loaded functional format, completely optional candidates are selected in 2^N-1 subsets. On the other hand, if at least one of the absolutely important, optionally important, gradually important, and exclusively important candidates appears in

a functional format, completely optional candidates are selected in 2^N subsets. Most variable selection procedures have discussed completely optional candidates. Completely optional candidates are freely entered in a functional format. If a sub-group of candidates is equivalent to a single candidate in a classified group, the candidates in this sub-group are called fixed. Fixed candidates are always selected as a fixed set but not separately. Fixed candidates are always optionally important, gradually important, exclusively important, gradually optional, exclusively optional or completely optional. Let us enter a sub-group of candidates between parentheses such as (.....). It is possible that there is more than one sub-group of fixed candidates. For instance, if candidates Z1, Z2, Z3, and Z4 are exclusively important and candidates Z2 and Z3 are fixed, then </Z1(Z2,Z3)Z4/> equivalent to <Z1,(Z2,Z3), Z4> generates (sub-)subset 1 (Z1); subset 2 (Z2, Z3); and subset 3 (Z4). If they are completely optional, then Z1(Z2,Z3)Z4 equivalent to Z1,(Z2,Z3),Z4 generates (sub-)subset 1 (Z1, Z2, Z3, Z4); subset 2 (Z1, Z2, Z3); subset 3 (Z2, Z3, Z4); subset 4 (Z1, Z4); subset 5 (Z1); subset 6 (Z2, Z3); subset 7 (Z4); and subset 8 (empty).

Whenever a constant term is included in a loaded functional format, it must always be selected. If not, a constant term should not be generated.

The equations which are consistent with the a priori information about candidates are here called meaningful

Table 1. 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.

ones. In economics, it is of great importance to classify candidates and generate only meaningful equations before estimation, whenever a priori information about candidates is available. In the above example, if candidate LIN is absolutely important and candidates WIN(-1) and KEVN(-2) are optionally important, the equations of Formats (1), (3), and (4) are meaningful and the remaining are meaningless. Formats (1), (3), and (4) can be derived from the following format:

```
HART=F($C/LIN/<WIN(-1),KEVN(-2)>)                    (9)
```

If Formats (3), (4) and (6) are meaningful, the following format can be utilized:

```
HART=F($C/LIN/<*WIN(-1),KEVN(-2)*>)                  (10)
```

If Formats (1) and (4) are meaningful, the following format can be entered:

```
HART=F($C/LIN/<+KEVN(-2),WIN(-1)+>)                  (11)
```

Whenever absolutely important candidates are considered, it is suggested that to minimize CPU time they be entered after the constant term which is always adopted.

Table 1 demonstrates variable classification and meaningful subsets of candidates in a three-candidate case.

3. NUMBER OF MEANINGFUL SUBSETS

It is clear that the number of absolutely important candidates does not change the number of meaningful subsets to be generated. We regard a sub-group of fixed candidates as a single candidate in counting the number of meaningful equations. Let us assume that there are I groups of optionally important candidates, J groups of gradually important candidates, K groups of exclusively important candidates, L groups of gradually optional candidates, M groups of exclusively optional candidates, and N completely optional candidates for an explained variable in quesion, where $I \geqq 0$, $J \geqq 0$, $K \geqq 0$, $L \geqq 0$, $M \geqq 0$, $N \geqq 0$, and $I+J+K+L+M+N > 0$. Furthermore, we assume that if $I > 0$, $J > 0$, $K > 0$, $L > 0$, $M > 0$, and $N > 0$, there are i_p optionally important candidates in the i-th pair of $\langle \ldots\ldots \rangle$ for $1 \leqq i \leqq I$, j_Q gradually important candidates in the j-th pair of $\langle +\ldots\ldots+ \rangle$ for $1 \leqq j \leqq J$, k_R exclusively important candidates in the k-th pair of $\langle /\ldots\ldots/ \rangle$ for $1 \leqq k \leqq K$, ℓ_S gradually optional candidates in the ℓ-th pair of $\langle -\ldots\ldots- \rangle$ for $1 \leqq \ell \leqq L$, and m_T exclusively optional candidates in the m-th pair of $\langle *\ldots\ldots* \rangle$ for $1 \leqq m \leqq N$. For instance, I=0 implies that there is no group of optionally important candidates, while I=2 implies that a functional format includes two groups of optionally important candidates for the explained variable such as

$$\langle U_{1_1}, U_{1_2}, \ldots, U_{1_p} \rangle \text{ and } \langle U_{2_1}, U_{2_2}, \ldots, U_{2_p} \rangle$$

where U_{i_p} stands for an optionally important candidate or a sub-group of fixed and optionally important candidates. Cases J=0, K=0, L=0, and M=0 can be understood like case I=0. Let us assume a normal situation in which the number of observations exceeds the maximum number of candidates to be adopted in a meaningful subset. The formula for calculating the number of meaningful subsets is given as follows:

$$\prod_{i=1}^{I}(2^{i_P}-1)*\prod_{j=1}^{J}j_Q*\prod_{k=1}^{K}k_R*\prod_{\ell=1}^{L}(\ell_S+1)*\prod_{m=1}^{M}(m_T+1)*2^N-g \qquad (12)$$

where g=1 if I=0, J=0 and K=0 and g=0 if I>0, J>0 or K>0. Other cases can be calculated by removing from (12) the term corresponding to I, J, K, L, M, or N, if I=0, J=0, L=0, K=0, M=0, or N=0. For instance, the following functional format

```
HLWK=F($C/AA,AB/<BA,BA(-1)><CA,CB></DA,(DB,DC),DD/>
        <*EA,EB*>FA,FB,FC)                          (13)
```

implies that I=2, J=0, K=1, L=0, M=1, N=3, g=0, 1_P=2, 2_P=2, 1_R=3, and 1_T=2 and generates $(2^2-1)*(2^2-1)*3*(2+1)*2^3$=648 meaningful subsets. There are 2^{15}-1=32,767 possible subsets derived from (13).

4. CHOICE OF STATISTICALLY AS WELL AS ECONOMICALLY BEST SUBSET

Since many criteria to evaluate an estimated equation have been proposed, a grand equation evaluation function or a lexicographic ordering of criteria is needed to find the theoretically best subset from the viewpoints of not only economics but also statistics. Unfortunately, there is no such function or ordering so far. Hence, it is impossible to define at present the theoretically best subset. Here, a practical method for choosing the statistically as well as economically best subset among many meaningful subsets which have the same functional form like linearity or log-linearity is proposed. The method consists of economic and statistical criteria. Economic criteria are related to the signs and magnitudes of coefficients of a meaningful subset. On the other hand, statistical criteria are related to T-test, Durbin-Watson statistic, relative absolute error, turning point error percentage, and coefficient of determination adjusted by the number of candidates adopted. Let us call the following consideration fitting criterion: the higher the coefficient of determination adjusted by the number of candidates in a meaningful subset, the better is the subset, ceteris pulibas. All economic and statistical criteria except for the fitting criterion are treated as discrete like 'pass or fail'. However, the fitting criterion is treated as a continuous one between 0 and 1. It is convenient to make

all criteria optional in a program or package.

We are now in a position to define the statistically as well as economically best subset as the one which passes all discrete criteria applied and has the highest adjusted coefficient of determination. An economist does not necessarily apply all discrete criteria, depending on the nature of his problem. For instance, Durbin-Watson statistic criterion cannot be applied to the case in which the number of observations is less than 15 or cross-sectional or pooling data are used. It may happen that he has a unique criterion peculiar to his problem. In this case, there must be a way that he can apply the unique criterion to meaningful subsets which pass all or some discrete criteria and have the adjusted coefficient of determination higher than some level specified. In the proposed method, the fitting criterion must always be applied to order the meaningful subsets which pass all discrete criteria applied.

4.1 SIGN CRITERION

This criterion is quite convenient to find meaningful subsets suitble for the hypothesis about rational behavior of an economic unit or for the technical relationship between output and inputs. Sometimes, it is possible to know a priori the signs of some candidates by means of

field surveys, experiments, empirical studies, or economic theories. In this case, an economist can hardly accept a subset whose coefficient shows the sign opposite to the one known a priori, except for a case where he discovers new findings. The criterion checks whether or not the signs of estimated coefficients are consistent with the ones known a priori, if applied.

By indicating P (positive), N (negative), or F (free, implying undetermined) for each of all candidates, he can make the computer sieve satisfactory meaningful subsets from unsatisfactory meaningful subsets. The surviving meaningful subsets are, furthermore, sieved by other criteria to be discussed later.

4.2 MAGNITUDE CRITERION

From economic theory on stability, the coefficient of some candidate may be expected to be within a certain range. In this case, the magnitude criterion can be applied to the coefficients of surviving meaningful subsets. Meaningful subsets which pass the sign criterion but do not pass the magnitude criterion are regarded as unsatisfactory and dropped. U (upper bound) and an upper bound value, L (lower bound) and a lower bound value, R (range) and upper and lower bound values, or F (free,

implying undetermined) is specified for each of all candidates, if this criterion is utilized.

4.3 T-TEST CRITERION

After economic criteria are applied, statistical criteria can be applied to the coefficients or estimates of surviving meaningful subsets. By this criterion, coefficients are examined to see whether or not they are significant at a specified level. Unless at least one coefficient is significant at a specified level, such a meaningful subset will be regarded as unsatisfactory and dropped. In order to allow for the flexibility of this criterion, it is convenient to let an economist be able to specify candidates to which T-test is applied at a significance level specified by him. This criterion is related to the sign criterion mentioned before. When the sign criterion is used, the following three T-tests are performed:

(i) If P (positive sign) is specified for coefficient a, the null hypothesis H_0: a=0 and the alternative hypothesis H_1: a>0 are assumed.

(ii) If N (negative sign) is specified for coefficient a, the null hypothesis H_0: a=0 and the alernative hypothesis H_1: a<0 are assumed.

(iii) If F (free or undetermined sign) is specified for coefficient a, the null hypothesis H_0: a=0 and the alternative hypothesis H_1: a≠0 are assumed.

Needless to say, if the sign test is not required but T-test is required, case (iii) is assumed.

4.4 DURBIN-WATSON STATISTIC CRITERION

As far as time series data are concerned, the Durbin-Watson statistic test is useful to check whether or not the disturbance term has autocorrelation or it has the minimum variances. This criterion cannot be applied in the cases where cross-sectional or pooling data are used, the number of observations is less than 15, or the constant term is not included. Of course, a significance level is specified by an economist.

4.5 RELATIVE ABSOLUTE ERROR CRITERION

In reality, it is not so easy to find outliers in the data of many candidates before estimation. In an equation in which time series or pooling data are used and a lagged candidate appears, it may not make sense to remove outliers. For instance, suppose that the current consumption is affected by the previous time's real income, which is absolutely important, and some other candidates representing real prices and eating habits. If the real income in 1973 looks like an outlier and is removed together with the data of other candidates and explained variable (consumption) in the same year, the consumption in 1974 will be explained by the real income and other candidates in 1972. This implies that there is no consistent consumption behavior over time. Accordingly, it is rather difficult or impossible to remove outliers.

In order to reject a meaningful subset which tracks well most of the observations of an explained variable except for one or two, this criterion can be used. This criterion checks whether or not an estimate concerning area i at time t is within a specified acceptance range centered on the corresponding observation. Let I and T stand for the numbers of areas and observation times, respectively. When criterion value w is given, the acceptance range is given as follows:

$$100*|(Y_i(t)-EY_i(t))/Y_i(t)| \leq w \text{ for area } i \text{ and time } t \quad (14)$$

where $Y_i(t)$ and $EY_i(t)$ stand for the observation and estimate of an explained variable in area i at time t for $1 \leq i \leq I$ and $1 \leq t \leq T$. Case I=1 and T>1 implies time series, case I>1 and T=1 implies cross-sectional, and case I>1 and T>1 implies pooling.

4.6 TURNING POINT CRITERION

It is often heard that it is rather easy to find a meaningful subset which tracks well a monotone movement of an explained variable, but it is not easy to obtain a meaningful subset which tracks well turning points implying changes in economic phases such as from booming to recession or vice versa. A turning point is indicated by the observations of an explained variable at three contiguous times. Business cycle and pig cycle are typical examples. A meaningful subset which tracks well turning points is important, especially for economic policy makers.

The condition to guarantee the existence of a V- or A-shape turning point at time t and in area i can be written as follows:

$$(Y_i(t)-Y_i(t-1))*((Y_i(t+1)-Y_i(t))<0 \text{ for each } i \text{ and } t \geq 3 \quad (15)$$

However, from the practical point of view, it would be

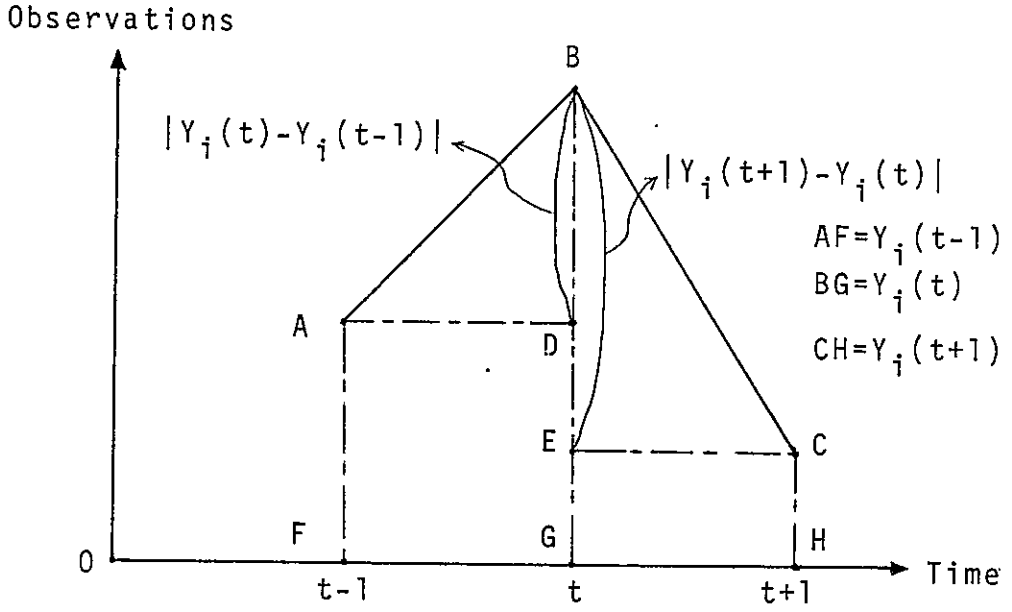

Figure 1. Definition of a Turning Point with respect to Area i and Time t

convenient to introduce the following condition in addition to (15):

$$\text{Min}(|(Y_i(t)-Y_i(t-1))/Y_i(t)|,|(Y_i(t+1)-Y_i(t))/Y_i(t)|)\geq u/100 \quad \text{for each } i \text{ and } t\geq 3 \qquad (16)$$

where u is a positive real number specified by the economist. He can make the computer ignore rather flat turning points by specifying a proper value (%) to u. For example, three observations $Y_i(t-1)=18,000,000$, $Y_i(t)=17,999,998$, $Y_i(t+1)=18,000,010$ shows a V-shape turning point at time t and in area i. Howver, this turning point is quite flat. If he specifiies 1 (%) to u, this turning point is ignored from the turning point error percentage criterion.

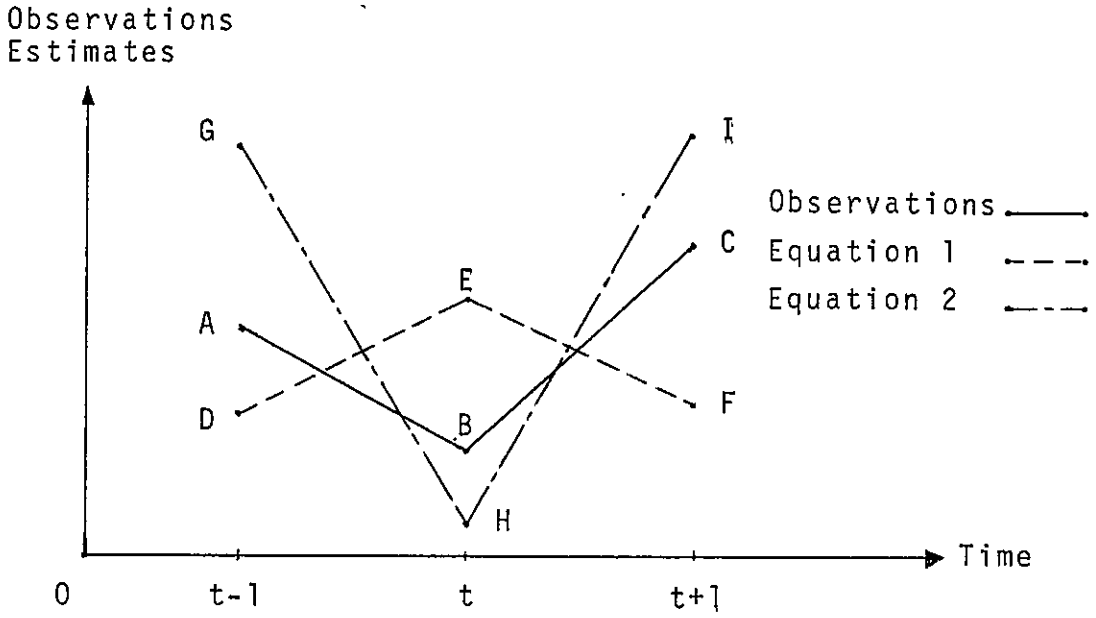

Figure 2. Tracking a V-shape Turning Point

In Figure 2., observation point B is a turning point showing V-shape. Estimate points D, E and F of equation 1 do not show V-shape, while estimate points G, H and I of equation 2 show V-shape. Therefore, equation 2 passes this criterion at time t, but equation 1 does not. On the other hand, estimate points G, H and I track observation points A, B and C more roughly than estimate points D, E and F, as far as times t-1, t and t+1 are concerned.

4.7 FITTING CRITERION

The criteria mentioned above are treated as discrete

like 'pass or fail'. On the other hand, this criterion is treated as continuous. The degree of fitting is measured by the coefficient of determination adjusted by the number of candidates, RR, defined by (18) as follows:

$$R=\sum_{t=1}^{T}\sum_{i=1}^{I}(EY_i(t)-\bar{Y})^2/\sum_{t=1}^{T}\sum_{i=1}^{I}(Y_i(t)-\bar{Y})^2 \qquad (17)$$

$$RR=1-(1-R)*(T*I-1)/(T*I-K) \qquad (18)$$

where K and $\bar{Y}$ stand for a number of candidates including a constant term, if any, and the mean of all observations of an explained variable, respectively. The surviving meaningful subset which has the highest adjusted coefficient of determination is regarded as the statistically and economically best.

5. EXAMPLES

Let us estimate an agricultural production function of Cobb-Douglas type by using the data of Japan from 1965 to 1979. Labor, capital, land, intermediate goods and services and technical progress are production factors taken into consideration. We assume that an agricultural production function must have the production factors of

labor, capital and land, but intermediate goods and services and technical progress are not decisively important production factors. Let us assume that LX=LOG(X), LL=LOG(L), LKA=LOG(KA), LKP=LOG(KP), LKM=LOG(KM), LRKM=LOG(R*KM), LK=LOG(KA+KP+KM), LRK=LOG(KA+KP+R*KM), LAX=LOG(AX), LCAX=LOG(C*AX), LC=LOG(C), LQ=LOG(Q) and LT=LOG(T)
where X=output, L=labor, KA=animal caiptal (cows, steers, pigs, etc.), KP=plant capital (apple trees, orange tress, tea tress, etc.), KM=machine capital, R=use rate of machine capital, AX=total cultivated acreage, C=weather index of rice, Q=intermediate goods and services and T=time trend. It must be remembered that agricultural capital consists of animal, plant and machine capital. Animal and plant caiptal are assumed to be used fully. However, machine capital is not necessarily fully used. Since the data of harvested acreages of all crops are not available, the weather index of rice is introduced to adjust total cultivated acreage.

We introduce the following functional format:

```
LX=F($C/LL/</LK,LRK,(LKA,LKP,LKM),(LKA,LKP,LRKM)/>
      </LAX,LCAX,(LC,LAX)/>,<*LT,T*>LQ)                    (19)
```

/LL/ implies that labor candidate is absolutely important and always selected. </LK...LRKM)/> implies that candidates representing capital are (1) LK, (2) LRK, (3)

LKA,LKP,LKM and (4) LKA,LKP,LRKM. </LAX...LAX)/> implies that candidates representing land are (1) LAX, (2) LCAX and (3) LC,LAX. <*LT,T*> implies that (1) LT, (2) T or (3) no techical progress variable is selected. LQ is treated as a completely optional candidate. (19) leads to $4*3*(2+1)*2=72$ meaningful subsets, while there are $2^{13}-1=8,191$ possible subsets, where there are 13 different non-constant candidates in (19). Less than 1 % of all possible subsets is meaningful. By loading sign test (requiring positive signs for all non-constant candidates), 5 % T-test for all nonconstant candidates, 5 % Durbin-Watson test, 5 % relative error test and 3 % turning point test, the following is selected as the economically as well as statistically best subset:

LX=2.59366+0.26624*LL+0.11125*LK+0.43510*LCAX+0.01966*T
(3.80787)(3.07854) (3.22452) (6.60864) (4.04693)
R=0.9674, RR=0.9544, SD=0.01383, FA=-0.2978, DW=2.490 (20)

where numbers in parentheses, R, RR, SD, FA and DW stand for T-ratios, coeff. of determ., adjusted coeff. of determ., standard deviation of disturbance terms, first-order autocorrelation coeff. and Durbin-Watson statistic, respectively.
In (20), serial correlation was undetermined by 5 % Durbin-Watson test.

6. CONCLUSION

Variable selection methods have been discussed mainly concerning completely optional candidates. However, in the field of economics, we often encounter absolutely important, optionally important, gradually important, exclusively important, gradually optional, and exclusively optional candidates in addition to completely optional candidates. It is of great importance to classify candidates by a priori information about candidates and generate only meaningful subsets which are worth being estimated. The classification of candidates reduces the number of (meaningful) subsets drastically from 2^N-1 subsets in all possible regressions, if the information is available. Meaningful subsets are estimated and evaluated by economic and statistical criteria. Sign, magnitude, T-test, Durbin-Watson statistic, relative absolute error, and turning point error percentage critria are treated as discrete like pass or fail, while fitting criterion is treated as continuous. The meaningful subset which passes all discrete criteria and possesses the highest coefficient of determination adjusted by the number of candidates adopted can be regarded as the statistically as well as economically best.

The method proposed shortens the research period, saves labor and other resources, reduces research cost, and

even improves the quality of research. Thus, it may be useful for business science, policy science, sociometrics, politometrics, psychometrics, and even 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.