A TIME-, LABOR-, AND RESOURCE-SAVING AS WELL AS COST-REDUCING SOFTWARE METHOD FOR ESTIMATING A LARGE-SCALE SIMULTANEOUS EQUATIONS MODEL

H. Onishi

October 1980
PP-80-12

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

PREFACE

Four years ago, I was asked to teach a computer-oriented econometrics course in the Ph.D. program of the Socio-Economic Planning Graduate School at the University of Tsukuba, Japan. At that time, the University was three years old and the small number of full and associate professors who were qualified to teach in the Ph.D. course were not very familiar with computers. I decided to make my own econometric computer program package to use for the course, because I was not aware of the many computer program packages available for purchase which handle econometric estimation and simulation. As economists insist on efficient use of resources, I wanted to make my package as efficient as possible. One year later, the University purchased some packages such as TSP, SPSS, MINITAB, and BMDP. Then, I discovered that as far as econometric jobs are concerned, these packages are unsatisfactory because they require too much expenditure of time, labor, and resources. This discovery lead me to concentrate on the development of a time-, labor-, and resource-saving package for input-output analysis, econometric estimation, and econometric simulation. The Onishi Econometric Program Package, OEPP, consisting of about 15,000 steps in Fortran, is the result of this research. The entire system of OEPP was designed solely by me. Although a great deal of my time was spent on teaching various courses, doing other research, and working on administrative jobs to establish our institute, undegraduate and graduate schools in our new University, I have finished the first version of OEPP. I would like to improve OEPP in any way possible and hope that OEPP can make some contribution

to the development of positive economics in the world.

This paper shows the software method for estimation, i.e., equation evaluation, which is one of the new ideas embodied in OEPP. In the near future, I would like to write another paper about the software method for econometric simulation, i.e., model evaluation used in OEPP.

The one-semester course in Introductory Computer Science I took in the University of Illinois gave me the basic knowledge on which I drew in developing OEPP. I am grateful to Professor E.R. Swanson, Professor T. Takayama, Professor G.G. Judge, Professor H.G. Halcrow, Professor D. Smith, Professor L. Feltner, Professor L.D. Hill, Professor J.T. Scott, Jr., Professor W.D. Seitz, Professor R.M. Leuthold and other professors in the Department of Agricultural Economics, at the University of Illinois.

I would also like to thank Professor S. Shishido, Professor T. Fukuchi, and Professor P.E. Katoh of the University of Tsukuba who understood the need to develop an efficient package and provided encouragement. Thanks are also due to the graduate students in my computer-oriented econometrics course, especially Mr. S. Tokunaga, who have been patient in using the early version of OEPP.

A research grant offered by the University of Tsukuba in 1979 partially contributed to the development of OEPP.

I would like to express my appreciation to the International Institute for Applied Systems Analysis for giving me the opportunity to join the Food and Agriculture Program and issue this paper. Finally, I am grateful to Ms. N. Matsubara for typing the User Reference Manual and Ms. T. Hubauer and Ms. B. Lopuch for typing this article.

Harty Haruo ONISHI
June 6, 1980

CONTENTS

A TIME-, LABOR-, AND RESOURCE-SAVING AS WELL AS COST-REDUCING SOFTWARE METHOD FOR ESTIMATING A LARGE-SCALE SIMULTANEOUS EQUATIONS MODEL

Haruo Onishi

INTRODUCTION

It is certain that nowadays applied econometricians can accomplish very little without the use of electronic computers. They spend time and labor for computer work and use various kinds of resources such as paper for output, carbon ribbons for printing, electricity, etc. in order to obtain satisfactory coefficients of the equations in their econometric models.

Every day a large number of applied econometricians, graduate and undergraduate students in the world seem to waste their time, labor, and resources with respect to their research and education, because all existing econometric computer program packages, as far as the author knows, cannot automatically handle the greater part of the trial and error processes. The total amount of waste in a day resulting from all research programs in the world is by no means small, and by the end of each year this waste accumulates to a vast total. Tremendous amounts of wood resources could be saved and more applied econometric research could be done with the same amount of labor, if an efficient method which can handle most of the trial and error computer work were developed.

The purpose of this paper is to offer the criterion by which econometric program packages should be evaluated and to show a time-, labor-, and resource-saving as well as cost-reducing method for the estimation of a large-scale simultaneous equations model. The method to be discussed here is not concerned with model evaluation but with equation evaluation in practice. The method to be presented has already been adopted in the Econometric Program Package, OEPP and been proven to work well.*

RESEARCH PERIOD CRITERION

A problem must urgently be solved. The assessment evaluation of some project is required as soon as possible. The world food problem and world energy problem are examples of problems which need to be solved soon. A project to construct a bridge between the mainland and an island requires an assessment evaluation as soon as possible. Without assessment evaluation, the project cannot actually be started, so that the islanders cannot enjoy the benefits stemming from the construction of the bridge.

Professors and researchers want to finish their research and publish their new findings earlier than others. They are usually evaluated by the number of publications during a certain time period, for example, a year. Of course, the quality of publications is an important factor of achievement evaluation, but it is quite difficult to properly evaluate the quality of publications. Consequently, the number of publications seems to be utilized as the most important factor of achievement evaluation. Only the one who publishes the original or new findings first can get the copyright, no matter how much he has spent for his research, if the same methodology is used. The number of publications is equivalent to the number of copyrights. Even if one used less research funds to get the same quality of research results but spent more time, even one more day, to finish his research than another, he would not get any credit for publishing his paper. Graduate and undergraduate students want to finish their term papers or homework as quickly as possible to spend more time for other subjects. This shows how important time is. It is quite important to reduce the time spent on research projects and even homework.

Let us focus on research. Research is composed of various activities. Collection and reading of journals, books, reports relevant to research, modelling the structure, collection of data, estimation of coefficients, simulation of the estimated model, and writing manuscripts for publication are common jobs of applied econometric research.

*OEPP stands for Onishi Econometric Program Package, developed by the author, at the Institute of Socio-Economic Planning, the University of Tsukuba, 1979, Japan, consisting of about 15,000 steps, written in Fortran. Data management, input-output analysis, econometric estimation, and econometric simulation by time series data, cross-sectional data, and pooling data can be handled. About 2,000 (sub)equations can be estimated and the best (sub)equation among them can be chosen at a cost of about 3 US dollars in less than 10 minutes (not CPU but real time) by the computer ACOS-800 installed at the University of Tsukuba, 1979.

The research period consists of time spent on (a) collecting and reading journals, books, and technical reports, (b) thinking of methodologies and analyzing tools appropriate to the research, (c) collecting data, (d) key-typing or punching in input data on the computer, (e) estimating the coefficients of equations in a model, (f) checking the output and re-thinking for the best estimated equations, (g) key-typing or punching in input data on the computer repeatedly, (h) estimating again and again until the best equations are obtained, (i) simulating an estimated model, (j) checking the output and re-thinking for the best model, and (k) writing manuscripts for a journal, book, or technical report or a resume for a conference or meeting.

From the viewpoint of applied software, the subtotal time period consisting of times (d) through (k) is important and relevant. The time period consisting of (a), (b), and (c) is assumed to be constant because applied software cannot contribute to this work.

Nowadays, the time spent on collecting journals, books, reports, and data necessary for the research can be considerably reduced due to the development of the computer library and data bank in some scientific fields.

In applied econometric research, a package is (or packages are) used many times during this subtotal time period. This subtotal time period is quite big. Since the first publication is decisively most important, econometric program packages must be evaluated by the duration of this subtotal time period. An econometric program package which reduces this subtotal time period drastically must be regarded as good. However, CPU time spent to estimate an equation or a certain set of equations in a run of the computer has been often used as the criterion to judge which package is better or the best. This criterion does not make sense if uncertainty is involved in the estimation. Especially for the estimation of large-scale simultaneous equations models, researchers have to spend much time for finding what explanatory variables are really needed to track the behavior (or data) of explained variables, what time lag structure is the most suitable for explanatory variables, what functional forms are appropriate for equations (linear, loglinear, etc.), what estimation methods are proper for equations, and so on. Only when all these questions are answered, can the comparison of CPU time lead to a proper criterion to evaluate econometric program packages. Unfortunately, these questions are answered by trial and error in most cases. The time, labor, and resources spent for answering these questions by trial and error are quite extensive. Why should we not let the computer handle most of this trial and error process? By letting the computer do so it is possible to reduce drastically the amount of time, labor, and resources spent. Existing econometric program packages cannot deal with this job and therefore force researchers to key-type or punch in repeatedly input data of each of all equations and judge whether or not the outcome is satisfactory.

Terminals are available so that it is possible to judge whether or not the outcome is satisfactory without having the printout printed by a line printer. However, if the output is large, it is difficult or impossible to compare the first and last parts of the outcome at the same time on the screen of the terminal. It is also impossible to carry away the outcome as one can with the printout printed by a line printer. It takes more time to get the printout after judging the outcome on the terminal. Hence, most researchers want to have the printout without seeing the outcome on the terminal. The fact is that many researchers do not like to spend extra time checking the outcome on the terminal and then obtaining the printout printed by a line printer, if the outcome is satisfactory. Thus, they do not mind spending more research funds for printing the outcome, whether it is satisfactory or not.

MINOR CRITERIA

In the case where the same research time period is required for the same results of the research by different packages, the comparison of CPU time required for the calculation of coefficients of an equation makes sense. A package requiring less CPU time can be regarded as better.

The convenience criterion is relatively minor and difficult to apply. It is connected with:

-- How easy is it to use popular estimation methods such as ordinary least squares method ?
-- How many sophisticated estimation methods are applicable ?
-- How many subsidiary functions useful to the research are available--for instance, printing estimated equations, tabulating and plotting the data (observations), estimated values, residuals, ex post forecasted values, and ex ante forecasted values, aggregation of the data, estimated values, residuals, etc. from non-annual to annual level, from regional to national level, and both, printing variable descriptions, calculating the simple correlation coefficient table and moment matrix of the data ?
-- Can inverse transformation of the data, estimated values, ex post forecasted values, and ex ante forecasted values of transformed explained variables be automatically made ?
-- Are important statistics calculated ?
-- Can card decks, files, and magnetic tapes be used ?
-- How attractive and easy to read is the printout ?
-- Can time series data, cross-sectional data, and pooling data be handled ?
-- How easily can existing data banks be utilized ?
-- Can the package automatically generate the data (and package-specifying) variable notation of all (or some specified) area dummy variables, periodical (or seasonal for quarterly data) dummy variables, time trend variable for time series data,

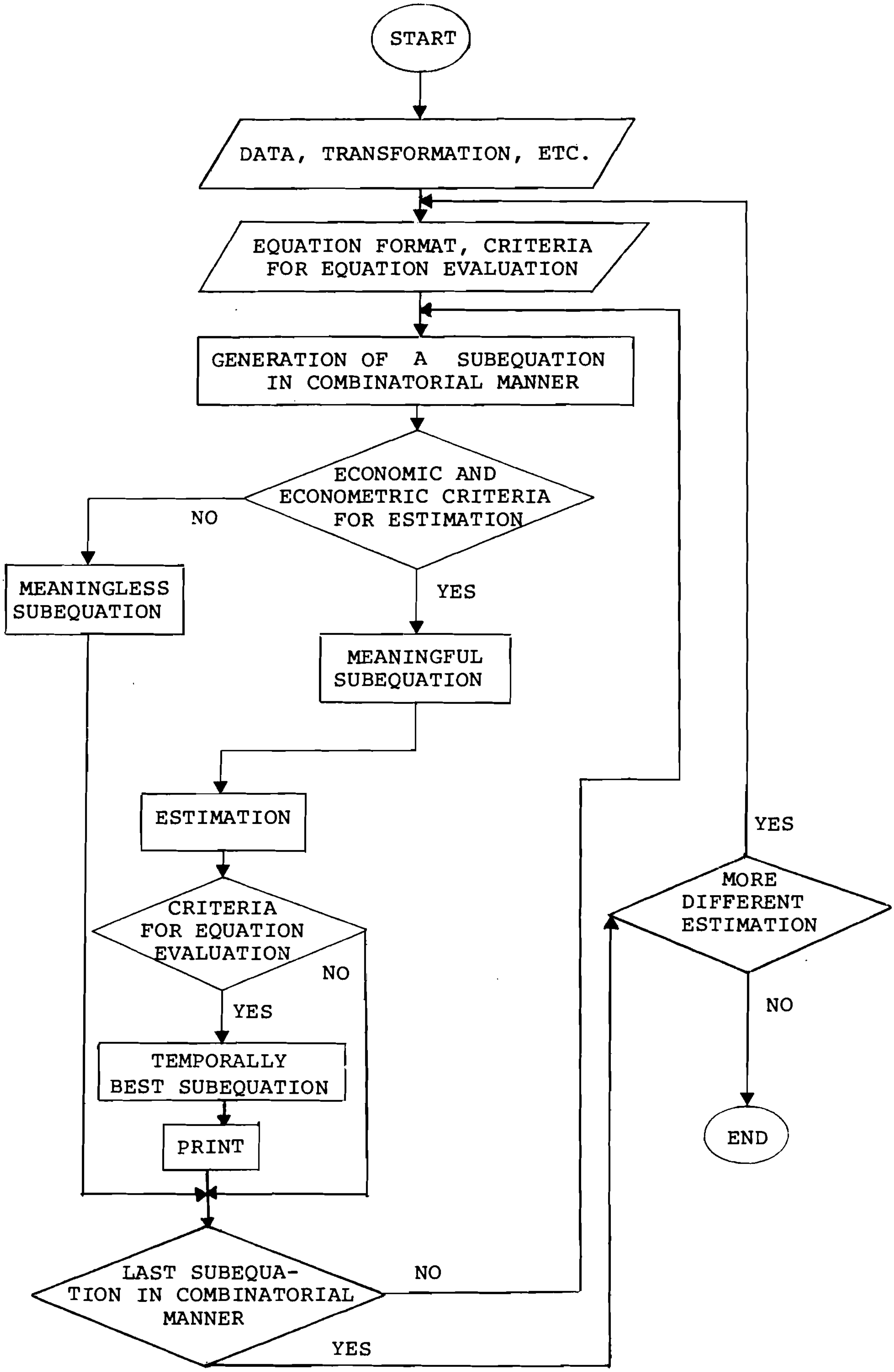
global time trend and all (or some specified) local time trend variables* for pooling data, and the ex ante forecasting data of policy variables which can be generated by functional forms of time trend variable ?

-- Is there a device which helps to find wrong data ?
-- How good is an error-finding system and how easily can the researcher fix an error ?
-- What kinds of variable transformations can be made and how easily are they made ?
-- What type of input-output analysis can be made ?
-- Is it possible to aggregate input-output data ?
-- What kinds of econometric simulation tests can be made ?
-- How many simultaneous equations models can automatically be dealt with for econometric simulation tests in a run of the computer ?
-- Can an econometric simulation test(s) be made immediately after the completion of the estimation of a simultaneous equations model(s) in a run of the computer ?

These optional functions can decrease the research time period. When researchers use computers and terminals, they easily make mistakes in carrying out the inverse transformation of the estimated values, residuals, growth rates, etc. of transformed variables by an electronic calculator, plotting the data, estimated values, ex post forecasted values, ex ante forecasted values, and residuals, aggregating non-annual- and/or regional-data-based results, etc. Hence, it is better to make the computer do these jobs. It is quite convenient if a package produces attractive printouts, because the (contracted) copies of the printouts can be used as manuscripts or resume. For instance, if variable description tables can be printed in the size which a researcher wants, he does not need to write variable notation used in the model, names, units, sources, issued dates, etc., of all original and transformed variables for the manuscripts or resume, because the (contracted) copies of variable description tables can be used as part of the manuscripts or resume. If estimated equations can be printed, the (contracted) copies of estimated equations and loaded identities of a simultaneous equations model reduce the time required to write the manuscripts or resume. Convenient functions should be included in an econometric program package as optional functions.

*For instance, suppose that there are 3 areas and 10 observation times. The data of global time trend variable are "1,1,1,2,2,2,...,10,10,10". On the other hand, the data of a local time trend variable, say, for the third area are "0,0,1,0,0,2,...,0,0,10".

Figure 1. Flow chart of combinatorial-variable-selection method for estimation.



A PRIORI INFORMATION AND MEANINGFUL SUBEQUATIONS

Let us define "subequations" as equations which can be derived or generated by selecting, in the combinatorial manner, variables from a set of non-constant explanatory variables given. For instance, suppose that variable A is the explained variable at hand, variables B, C(-1), and D(-2) are a set of non-constant explanatory variables, and variable A is expressed with a linear form of all or some of these non-constant variables B, C(-1), and D(-2) in addition to the constant term denoted by $C, where variables C(-1) and D(-2) imply variables C and D with time lag numbers 1 and 2, respectively. When we derive all subequations of variable A by selecting non-constant variables B, C(-1), and D(-2) in the combinatorial manner, the following are obtained in functional form F:

A=F($C,B,C(-1),D(-2)) (1)

A=F($C,B,C(-1)) (2)

A=F($C,C(-1),D(-2)) (3)

A=F($C,B,D(-2)) (4)

A=F($C,B) (5)

A=F($C,C(-1)) (6)

A=F($C,D(-2)). (7)

Equations (1) through (7) are here called subequations derived from non-constant variables B, C(-1), and D(-2) with the constant term $C.

The above equation format style* is used in OEPP so that we adopt the same style here and use it in the following chapters. Furthermore, (1) is the equation format by which explained variable A, constant term $C, and non-constant explanatory variables B, C(-1), and D(-2) are loaded. Therefore, if (1) is specified and the combinatorial-variable-selection command (parameter) is also specified with an estimation method, (1) through (7) must be automatically derived and estimated and then the best (and the second best, etc.) among all 7 estimated subequations is printed, if required. In general, if there are K non-constant explanatory variables considered possibly relevant to an explained variable, then $\sum_{k=1}^{K}\binom{K}{k} = 2^K - 1$ subequa-

*A blank(s) can be used in the place of a comma in the equation format of OEPP. For instance, (1) is equivalent to A=F($C B C(-1),D(-2)). If variable C exists, then variable C(-1) can be directly used as variable C with time lag number 1 without defining a new variable for C(-1).

tions can be derived. For instance, if 12 non-constant explanatory variables are considered as possibly relevant candidates, then $2^{12}-1 = 4,095$ subequations can be derived. Since 4,095 subequations are automatically derived, estimated, and judged from one equation format like (1), the user's labor becomes minimal and he can avoid many mistakes which could be easily made in entering 4,095 equation formats as existing packages require.

However, all subequations are not necessarily important. This means that a researcher can sometimes judge which subequation is important or unimportant for his research by utilizing his knowledge. This knowledge is here called a priori information. The point is how to make the computer recognize and use a priori information for the generation of subequations. If a priori information is available and actually used, all subequations are classified into two groups. One group consists of the subequations which are considered to be relevant to his research by the a priori information. The other group consists of the remaining subequations which are judged to be irrelevant to his research. The subequations which are considered relevant to his research are called "meaningful subequations". Only meaningful subequations need to be estimated and the best (and the second best, etc.) should be selected among all estimated meaningful subequations. Meaningless subequations do not need to be estimated. What is the a priori information used in the classification of subequations into meaningful and meaningless subequations? It is usually based on economic theory, rational behavior, and/or empirical studies. Researchers have some a priori information in most cases. Since a priori information has various degrees of certainty, importance, or relevance, it should be classified according to the degrees of certainty, importance, or relevance.

A priori information can be classified into four categories, which are those of (a) absolutely important, (b) optionally important, (c) exclusively important, and (d) exclusively optional variables. The number of absolutely important variables, if any, is just positive. But the number of optionally important variables, the number of exclusively important variables, and the number of exclusively optional variables, if any, are always greater than one, so that the variables in each of these three categories are equivalent to each other and can be used as substitute variables for each other. An absolutely important variable is defined as a variable such that any subequation which does not include it becomes meaningless. Hence, <u>all</u> of the absolutely important variables, if any, must be selected in each meaningful subequation. An optionally important variable is defined as a variable such that any subequation which does not include it or any of the variables equivalent to it becomes meaningless. Hence, <u>at least one</u> of the optionally important variables, if any, must be included in each meaningful subequation. An exclusively important variable is defined as a variable such that any subequation which does not include it or any of the variables equivalent to it or includes it with any number of its equivalent variables becomes meaningless. In other words, only one of the exclusively important variables must be included in each meaningful subequation. The variables of these three kinds mentioned above are "important" in the sense that if none of the variables each of which is an absolutely, optionally, or exclusively important variable is selected in a subequation, such a subequation always becomes meaningless. An exclusively optional variable is defined as a variable such that a subequation which does not include it does not become meaningless but once it is selected in a subequation, any variable equivalent to it cannot be selected in the subequation. In other words, none or only one of the exclusively optional variables must be included in each meaningful subequation. Non-constant explanatory variables for which no a priori information is available at all become optional (or completely optional). The selection of exclusively optional variables and (completely) optional variables depends on whether or not a priori information is available. If a priori information is available for important variables, namely if at least one variable can be classified as an absolutely important variable, an optionally important variable, or an exclusively important variable, then meaningful subequations do not necessarily include any of the exclusively optional variables and (completely) optional variables. Otherwise, meaningful subequations must include at least one of the exclusively optional variables or of the (completely) optional variables. Of course, a constant term, if any, must be included in each of all meaningful subequations.

Let us follow the OEPP notation and put absolutely important variables between a pair of "slashes", such as "/.../" , optionally important variables between a pair of "less-than and greater-than signs", such as "<...>" , exclusively important variables between a pair of "double less-than and greater-than signs", such as "<<...>>" , and exclusively optional variables between a pair of "less-than sign followed by asterisk" and "greater-than sign preceeded by asterisk", such as "<*...*>" .* For instance, absolutely important variables EEE, FF(-2), and G are distinguished from other variables by /EEE,FF(-2),G/ . Optionally important variables P(-1) , Q, and Q(-1) are represented by <P(-1),Q,Q(-1)> . Exclusively important variables RR, SS, TT, and UU are expressed as <<RR,SS,TT,UU>> . Finally, exclusively optional variables V and W(-1) are characterized by <*V,W(-1)*>.

Only one pair of slashes is needed for absolutely important variables, if any, when a non-simultaneous single equation estimation method is applied. On the other hand, at most two pairs of slashes are needed, when a simultaneous equation system single equation estimation method such as

Only alphanumeric symbols are allowed to be used for variable notation. Non-alphanumeric symbols such as "/" , "<" , ">" , "" , etc. cannot be used for variable notation, so that "/.../" , "<...>" , "<<...>>" , and "<*...*>" can classify a priori information.

Table 1. Classification of Explanatory Variable Candidates by A Priori Information.

Explanatory variable candidates	Distinguished by	Meaningful subequations include
Absolutely important variables, if any	/.../	All
Optionally important variables, if any	⟨...⟩	At least one
Exclusively important variables, if any	⟨⟨...⟩⟩	Only one
Exclusively optional variables, if any	⟨*...*⟩	None or only one
Optional (completely optional) variables, if any		None or at least one
Constant term, if any	$C	Always

Note: Any kind of symbols such as /.../ , ⟨...⟩ , [...] , {...} , ⌜...⌟ , ⌞...⌝ can be adopted, depending on the computer and terminal, if the proposed method is used in an econometric program package.

limited information maximum likelihood method or two stage least squares method is applied. However, it is possible to use more than one pair of less-than and greater-than signs, one pair of double less-than and greater-than signs, and one pair of less-than sign followed by asterisk and greater-than sign preceeded by asterisk regardless of an estimation method adopted.

Let us assume that explained variable A is expressed as a linear form of optional variables B, C(-1), and D(-2) in addition to the above absolutely important, optionally important, exclusively important, and exclusively optional variables so that the equation format for the equation including the above explanatory variable candidates is expressed as follows:

A=F($C,B,C(-1),D(-2)/EEE,FF(-2),G/<P(-1),Q,Q(-1)>

<<RR,SS,TT,UU>><*V,W(-1)*>). (8)

Since each of the optional variables and a set of variables belonging to a category can be put in any order, (8) is, of course, equivalent to

A=F($C/EEE,FF(-2),G/B<P(-1),Q,Q(-1)>C(-1)

<<RR,SS,TT,UU>>D(-2)<*V,W(-1)*>). (9)

There are many other styles of equation format which are equivalent to (8) and (9). In OEPP, (9), in which absolutely important variables are put immediately after a constant term, is a better equation format than (8), whenever absolutely important variables are included, because the time for searching meaningful subequations is held to a minimum. Even the variable notation of a constant term, $C, is not necessarily the first entry of the functional form in OEPP. But it is better to enter customarily the variable notation of a constant term, if any, first.

EXAMPLE

Let us give an example. Suppose that the demand quantity of beef denoted by BD is a linear function of all or some of the real disposable income Y, beef price relative to pork price BPP, beef price relative to poultry price BPLP, beef price relative to fish price BFP, beef price relative to mutton price BMP, beef price relative to egg price BEP, and beef price relative to dairy product price BDP. It is possible that beef price relative to horse meat price is important, but beef price relative to fish, egg, or dairy product price is not important to a certain country. Unless a priori information is available, the following equation format can be used:

BD=F($C,Y,BPP,BPLP,BFP,BMP,BEP,BDP). (10)

Consequently, $2^7-1=127$ subequations can be derived and estimated. Then, the best subequation must be selected among 127 estimated subequations. However, usually a researcher knows from demand theory that the real disposable income Y is always needed, beef price relative to pork price BPP is very important, and other relative prices are perhaps relevant. Because pork is a strong beef substitute and poultry, fish, mutton, eggs, and dairy products are weak beef substitutes in most countries, beef price relative to pork price is regarded as an absolutely important variable, and other relative prices are regarded as optionally important variables or optional variables. In the case where the relative beef prices other than beef price relative to pork price are optional, (10) can be changed into the following:

BD=F($C/Y,BPP/BPLP,BFP,BMP,BEP,BDP). (11)

Then, $2^5=32$ subequations become meaningful and must be estimated. The a priori information used is that any subequations which do not include both real disposable income Y and beef price relative to pork price BPP become meaningless. Accordingly, this a priori information reduces the number of meaningful subequations from 127 to 32.

Empirical studies by other researchers in the home country as well as in foreign countries may give some knowledge about real disposable income. Current real disposable income Y, equally-weighted average real disposable income YEW representing (Y+Y(-1))/2, arithmetically weighted average real disposable income putting more weight on the current but less on the previous period's real disposable income YUEW representing (2*Y+Y(-1))/3, or the maximum real disposable income YM3 among the current and the last 3 periods representing MAX(Y,Y(-1),Y(-2),Y(-3)) may be income variable candidates. Since these four income variable candidates are considered as exclusively important variables, the following equation format can be used:

BD=F($C/BPP/<<Y,YEW,YUEW,YM3>>BPLP,BFP,BMP,BEP,BDP). (12)

In this case, $4*2^5=128$ subequations become meaningful, so that they deserve to be estimated. If the a priori information that these income variable candidates are exclusively important and beef price relative to pork price is absolutely important is not used, $2^{10}-1=1,023$ subequations are derived and estimated. And then the best subequation is chosen among them. The treatment of these income variable candidates as exclusively important variables is based on the fact that all of these income variable candidates are related to the current real disposable income.

Another case may be as follows: real disposable incomes of the current and last two periods denoted by Y, Y(-1), and Y(-2) are income variable candidates. In this case, the income variable candidates can be regarded not as exclusively important but as optionally important. That is, at least one of Y, Y(-1), and Y(-2) must be an income variable. Then, the following can be entered:

BD=F($C/BPP/<Y,Y(-1),Y(-2)>BPLP,BFP,BMP,BEP,BDP). (13)

(13) leads to the derivation of $(2^3-1)*2^5=224$ meaningful subequations.

If income variables Y, YEW, YUEW, and YM3 are regarded as exclusively important variables in a income group, all relative beef prices are regarded as optionally important variables in a price group, and the previous period's beef demand BD(-1) which implies the inertia effect of consumption is regarded as optional, the following equation format can be used:

BD=F($C<<Y,YEW,YUEW,YM3>><BPP,BPLP,BFP,BMP,BEP,BDP>

BD(-1)). (14)

4*(2 -1)*2 =504 meaningful subequations are derived and estimated. How much time, labor, and resources can be saved by entering one equation format (14) by OEPP instead 504 equation formats by an existing econometric program package?

ABSOLUTELY IMPORTANT LAGGED EXPLAINED VARIABLE

When a lagged explained variable (lagged regressand) is considered to be a decisively important explanatory variable, it can be treated as an absolutely important variable. In this case, ordinary least squares method cannot be applied, so that Cochrane-Orcutt method, Prais-Winston method, Durbin 2 step method, or 3 pass least squares method is suitable for this estimation. By entering a lagged explained variable into a pair of slashes, we can make the computer easily find only the meaningful subequations which include the lagged explained variable. Hence, one of the estimation methods mentioned above can be automatically applied for all of the meaningful subequations.

MEANINGFUL SUBEQUATIONS FOR LIMITED INFORMATION MAXIMUM LIKELIHOOD METHOD AND TWO STAGE LEAST SQUARES METHOD

Let us call, for simplicity, predetermined variables not included in an equation at hand but included in the other equations or identities in a simultaneous equations model "extraneous predetermined variables", so that predetermined variables imply here the ones included in an equation at hand. When limited information maximum likelihood method or two stage least squares method is applied for the estimation of the coefficients of explanatory variables of an equation, the left-hand side endogenous variable (explained variable), predetermined variables, right-hand side endogenous variables, and extraneous predetermined variables must be identified. Let us divide a group of predetermined variables, a group of right-hand side endogenous variables, and a group of extraneous predetermined variables by a semi-colon ";" in an equation format and enter all variables (except for a left-hand side endogenous variable) in this order of three groups. For example, suppose that variables $C, BB(-1), and C are predetermined variables, variables E, FF, and GGG are right-hand side endogenous variables, and variables HH, I(-1), J(-2), KKK, LL(-1), MM, and N are extraneous predetermined variables for left-hand side endogenous variable HLWK. The equation format for this equation is expressed as follows:

HLWK=F($C,BB(-1),C;E,FF,GGG;HH,I(-1),J(-2),KKK,LL(-1), MM,N). (15)

In the above equation format, variables $C, BB(-1), C, E, and GGG are the explanatory variables for explained variable HLWK which are used for the calculation of their coefficients and appear in the estimated equation printed. However, extraneous predetermined variables HH, I(-1), J(-2), KKK, LL(-1), MM, and N are used for the calculation of the coefficients of the explanatory variables but never appear in the estimated equation printed. (15) is printed in a linear form as follows:

$$HLWK=a_0+a_1 BB(-1)+a_2 C+a_3 E+a_4 FF+a_5 GGG \qquad (16)$$

where a_0, a_1, a_2, a_3, a_4, and a_5 are estimated coefficients.

It is possible to apply the concepts of absolutely important variables, optionally important variables, exclusively important variables, exclusively optional variables, and optional vaiables to (15). Suppose that all extraneous predetermined variables are always used. This implies that all extraneous variables should be put between a pair of slashes. If all right-hand side endogenous variable candidates are optionally important, the equation format takes the following form:

HLWK=F($C,BB(-1),C;<E,FF,GG>;/HH,I(-1),J(-2),KKK,LL(-1),

MM,N/). (17)

If predetermined variable candidate BB(-1) and right-hand side endogenous variable candidate E are absolutely important but the remaining right-hand side endogenous variable candidates FF and GGG are optionally important, then the equation format can be written as follows:

HLWK=F($C,C/BB(-1);E/<FFF,G>;/HH,I(-1),J(-2),KKK,LL(-1),

MM,N/). (18)

From (18), the following meaningful subequations are automatically derived and estimated with either limited information maximum likelihood method or two stage least squares method adopted, only if they are just-identified or over-identified:

HLWK=F($C,C,BB(-1);E,FF,GGG;HH,I(-1),J(-2),KKK,

LL(-1),MM,N) (19)

```
HLWK=F($C,C,BB(-1);E,FF;HH,I(-1),J(-2),KKK,
     LL(-1),MM,N)                                              (20)
HLWK=F($C,C,BB(-1);E,GGG;HH,I(-1),J(-2),KKK,
     LL(-1),MM,N)                                              (21)
HLWK=F($C,BB(-1);E,FF,GGG;HH,I(-1),J(-2),KKK,
     LL(-1),MM,N)                                              (22)
HLWK=F($C,BB(-1);E,FF;HH,I(-1),J(-2),KKK,
     LL(-1),MM,N)                                              (23)
HLWK=F($C,BB(-1);E,GGG;HH,I(-1),J(-2),KKK,
     LL(-1),MM,N).                                             (24)
```

It is very important to distinguish under-identified cases from just-identified and over-identified cases with respect to meaningful subequations. It can be done by comparison of the numbers of right-hand side endogenous variables and extranenous predetermined variables adopted in a meaningful subequation. Unless the number of right-hand side endogenous variables adopted exceeds that of extraneous predetermined variables adopted in a meaningful subequation, such a meaningful subequation is just-identified or over-identified, so that it can be estimated. Meaningful subequations which are under-identified must be ignored.

If a package is well made in such a way that it can estimate with, for instance, ordinary least squares method the meaningful subequations which do not include any right-hand side endogenous variable candidates, the following equation format can be used with the combinatorial-variable-selection command:

```
HLWK=F($C,C,BB(-1);E,FF,GGG;/HH,I(-1),J(-2),KKK,
     LL(-1),MM,N/).                                            (25)
```

In this case, 2^5-1=31 (meaningful) subequations are derived.

Among 31 subequations, 2^2*(2^3-1)=28 subequations are estimated with either limited information maximum likelihood method or two stage least squares method and the remaining 3 subequations are estimated with ordinary least squares method. The last three subequations ignore all extraneous predetermined variables and are regarded as

```
HLWK=F($C,C,BB(-1))                                            (26)
HLWK=F($C,C)                                                   (27)
```

HLWK=F($C,BB(-1)). (28)

Let us define a "unique predetermined variable" as a predetermined variable which is included in an equation at hand but in none of the equations and identities remaining in a simultaneous equations model. Accordingly, a non-unique predetermined variable is a predetermined variable which is included not only in an equation at hand but also in at least one of the other equations or identities.

If all predetermined variable candidates in an equation at hand are unique in the sense defined above, all meaningful subequations, which include at least one of the right-hand side endogenous variable candidates, derivable by selecting the predetermined variable candidates and right-hand side endogenous variable candidates in the combinatorial manner under the selection of all extraneous predetermined variables, do not lose any of all the extraneous predetermined variables in a simultaneous equations model. Therefore, if predetermined variable candidates C and BB(-1) in (25) are not included in the other equations and identities in a simultaneous equations model, all meaningful (and just- or over-identified) subequations derivable from (25) do not lose any of all the extraneous predetermined variables. If all or some of the predetermined variable candidates in an equation at hand are non-unique, the meaningful subequations which include at least one right-hand side endogenous variable candidate but none of the non-unique predetermined variable candidates lose the data of the unselected non-unique predetermined variable candidates as the data of their extraneous predetermined variables. Next, we assume in (25) that predetermined variable candidate BB(-1) and all right-hand side endogenous variable candidates E, FF, and GGG are optional but predetermined variable candidate C is absolutely important, and furthermore optional predetermined variable candidate BB(-1) is non-unique. In this case, the meaningful subequations which include at least one of the right-hand side endogenous variable candidates but do not include predetermined variable BB(-1) lose the data of predetermined variable BB(-1) which can be employed as those of their extraneous predetermined variables. The lack of predetermined variable BB(-1) leads to the reduction of the information available for the estimation of these meaningful subequations. Hence, if a package can automatically check whether or not any redundant predetermined variables are included among the predetermined variables and extraneous predetermined variables adopted in a meaningful subequation and ignore the cases in which at least one identical predetermined variable is included among the whole set of the predetermined variables and extraneous predetermined variables selected, then the equation format including non-unique and non-absoluately-important predetermined variable BB(-1) in the set of extraneous predetermined variables can solve this problem. This equation format can be written as follows:

HLWK=F($C/C/BB(-1);E,FF,GGG;/HH,I(-1),J(-2),KKK,LL(-1), MM,N/BB(-1)). (29)

Furthermore, if predetermined variable C is non-unique and optional, the following equation format becomes appropriate:

HLWK=F($C,C,BB(-1);E,FF,GGG;/HH,I(-1),J(-2),KKK,LL(-1),

MM,N/BB(-1),C). (30)

So far, it has been asssumed that all extraneous predetermined variables must be selected by each meaningful subequation. However, it is possible to treat all or some of the extraneous predetermined variables as absolutely important, optionally important, exclusively important, exclusively optional, and/or (completely) optional variables. In this case, it is of great importance to make the computer print what extraneous predetermined variables are selected for the estimation of each meaningful just- or over-identified subequation, because the extraneous predetermined variables adopted do not appear in any of the meaningful just- or over-identified subequations printed.

As far as limited information maximum likelihood method and two stage least squares method are concerned, a package must have the following functions:

(a) Finding meaningful subequations,

(b) Distinguishing an under-identified meaningful subequation from a just-identified or over-identified meaningful subequation,

(c) Handling an abnormal case, if it occurs, which is explained in the next chapter,

(d) Estimating by ordinary least squares method or other appropriate estimation methods the meaningful subequations which do not include any right-hand side endogenous variables,

(e) Not losing any of the extraneous predetermined variables in a simultaneous equations model.

NORMAL AND ABNORMAL CASES

If there are many explanatory variable candidates in the case where a single equation estimation method is applied, the number of explanatory variable candidates may exceed the product of the number of observation times multiplied by the number of areas, i.e., the number of samples. This case is here called abnormal. If we denote J and T as the numbers of non-constant explanatory variable candidates and samples, respectively, a normal case is expressed as $1+J<T$ when a constant term is included in an equation.

On the other hand, an abnormal case is expressed as $1+J \geq T$.

Suppose that there are K pairs of less-than and greater-than signs, L pairs of double less-than and greater-than signs, M pairs of less-then sign followed by asterisk and greater-than sign preceeded by asterisk, K_k optionally important explanatory variable candidates between the k-th pair of less-than and greater-than signs for $1 \leq k \leq K$ with $K>0$, L_l exclusively important explanatory variable candidates between the l-th pair of double less-than and greater-than signs for $1 \leq l \leq L$ with $L>0$, M_m exclusively optional explanatory variable candidates between the m-th pair of less-than sign followed by asterisk and greater-than sign preceeded by asterisk for $1 \leq m \leq M$ with $M>0$, N optional explanatory variable candidates, and $J - \sum_k K_k - \sum_l L_l - \sum_m M_m - N (\geq 0)$ absolutely important explanatory variable candidates. In a normal case, the number of meaningful subequations is as follows:

(a) If $K = L = M = 0$, and $N > 0$,

$$2^N - 1 \tag{31}$$

(b) If $K = L = N = 0$, and $M > 0$,

$$\prod_{m=1}^{M} (M_m + 1) - 1 \tag{32}$$

(c) If $K = M = N = 0$, and $L > 0$,

$$\prod_{l=1}^{L} L_l \tag{33}$$

(d) If $L = M = N = 0$, and $K > 0$,

$$\prod_{k=1}^{K} (2^{K_k} - 1) \tag{34}$$

(e) If $K = L = 0$, $M > 0$, and $N > 0$,

$$2^N \cdot \prod_{m=1}^{M} (M_m + 1) - 1 \qquad (35)$$

(f) If $K = M = 0$, $L > 0$, and $N > 0$,

$$2^N \cdot \prod_{l=1}^{L} L_l \qquad (36)$$

(g) If $L = M = 0$, $K > 0$, and $N > 0$,

$$2^N \cdot \prod_{k=1}^{K} (2^{K_k} - 1) \qquad (37)$$

(h) If $K = N = 0$, $L > 0$, and $M > 0$,

$$\prod_{l=1}^{L} L_l \cdot \prod_{m=1}^{M} (M_m + 1) \qquad (38)$$

(i) If $L = N = 0$, $K > 0$, and $M > 0$,

$$\prod_{k=1}^{K} (2^{K_k} - 1) \cdot \prod_{m=1}^{M} (M_m + 1) \qquad (39)$$

(j) If $M = N = 0$, $K > 0$, and $L > 0$,

$$\prod_{k=1}^{K} (2^{K_k} - 1) \cdot \prod_{l=1}^{L} L_l \qquad (40)$$

(k) If $K = 0$, $L > 0$, $M > 0$, and $N > 0$,

$$2^N \cdot \prod_{l=1}^{L} L_l \cdot \prod_{m=1}^{M} (M_m + 1) \qquad (41)$$

(l) If L = 0 , K > 0 , M > 0 , and N > 0 ,

$$2^N \cdot \prod_{k=1}^{K} (2^{K_k} - 1) \cdot \prod_{m=1}^{M} (M_m + 1) \qquad (42)$$

(m) If M = 0 , K > 0 , L > 0 , and N > 0 ,

$$2^N \cdot \prod_{k=1}^{K} (2^{K_k} - 1) \cdot \prod_{l=1}^{L} L_l \qquad (43)$$

(n) If N = 0 , K > 0 , L > 0 , and M > 0 ,

$$\prod_{k=1}^{K} (2^{K_k} - 1) \cdot \prod_{l=1}^{L} L_l \cdot \prod_{m=1}^{M} (M_m + 1) \qquad (44)$$

(o) If K > 0 , L > 0 , M > 0 , and N > 0 ,

$$2^N \cdot \prod_{k=1}^{K} (2^{K_k} - 1) \cdot \prod_{l=1}^{L} L_l \cdot \prod_{m=1}^{M} (M_m + 1) \qquad (45)$$

In the case where $1+J \geq T$, the maximum number of explanatory variables selected for a meaningful subequation is T-1, so that if K=L=M=0 and N>0, $2^{T-1}-1$ meaningful subequations are generated. If K>0, L>0 and/or M>0, then the number of meaningful subequations is expressed with a complicated formula. Therefore, a package must be made in such a way that an abnormal case is perfectly handled.

The above can be applied to the cases of limited information maximum likelihood method and two stage least squares method.*

CHOICE OF THE BEST SUBEQUATION

What is the best subequation if thousands of subequations are estimated for an explained variable? If we know the characteristics of the best subequation, we can find the criteria the best subequation must satisfy. Usually researchers know what the criteria are for an explained

*OEPP can handle normal and abnormal cases for most estimation methods including limited information maximum likelihood method and two stage least squares method.

variable at hand, so that they check each estimated equation and then find the best equation, but they waste tremendous amounts of time, labor, and resources at the same time. It is natural that, if possible, a package should be used to find the best subequation among all estimated meaningful subequations in the place of a researcher. If a package in which a meaningful-subequation-deriving system is installed possessed a best-subequation-finding system, it would become really useful.

There are many equation evaluation criteria, which are independent of each other and equally important. Therefore, we need a grand equation evaluation function which is a function of criterion variables for equation evaluation and gives a scalar value for comparison if the evaluation values of the criterion variables are given. Unfortunately, we do not know whether or not such a function exists and which functional form it assumes, even if it exists. How about a theoretically consistent ranking of the equation evaluation criteria? For instance, which is the most important, 5% significance T-test, 5% Durbin-Watson statistic test, or 10% residual percentage criterion for time series data? We really do not know which the most important is.

Accordingly, some compromise is needed to find the best subequation at present. The compromise is reflected by the definition of the best subequation. We would like to define the best subequation as the one which (a) satisfies the sign criterion, magnitude criterion, T-test criterion, Durbin-Watson statistic criterion, residual percentage or (absolute) relative error criteria, depending on what kind of data is used, and (b) possesses the highest coefficient of determination adjusted by the number of explanatory variables. It must be kept in mind that the best subequation found by OEPP is not necessarily the theoretically best but the practically best under the condition that the researcher accepts the equation evaluation procedure installed in OEPP. The criterion values are specified by the researcher.

The equation evaluation criteria are classified into economic, statistical and mathematical criteria. The economic criteria are based on economic theory of production function, consumer choice, stability condition, etc., and are represented by sign criterion and magnitude criterion. The statistical criteria are based on the assumption that the disturbance term has normal distribution. They are represented by T-test (or F-test) and Durbin-Watson statistic test (and sometimes Chow test for structural change). The mathematical criteria are based on mathematical formulae and are represented by residual percentage criterion (relative error criterion), turning point error percentage criterion, and fitting criterion (root-mean-square error criterion, Theil's inequality coefficient criterion, simple correlation coefficient criterion, etc.)

It is quite convenient if a package can print the reason why a subequation is unsatisfactory and if it can print the subequations, even if unsatisfactory, which a researcher does want to see irrespective of the criteria adopted. Needless to say, all criteria must be optional.

SIGN CRITERION

The simple but important criterion is the sign criterion. The signs, positive or negative, of coefficients of some explanatory variable candidates to be estimated are a priori known through economic theory or empirical studies. If at least one of the estimated coefficients fails to meet the corresponding sign indicated by a researcher, it is impossible to justify such a subequation (except for the case in which completely new findings are discovered). Therefore, the coefficients of explanatory variables adopted in a meaningful subequation must meet the sign criterion, if applied. For instance, the positive and negative signs of coefficients of real disposable income Y and beef price relative to pork price BPP in beef demand function (11) are known through economic theory. Accordingly, any meaningful subequation which includes the negative sign of the estimated coefficient of real disposable income Y or the positive sign of the estimated coefficient of relative beef price BPP fails to meet the sign criterion, so that it should not be printed except for the case in which a researcher does want it to be printed, although it is unsatisfactory.

If the sign of an explanatory variable candidate is uncertain, the coefficient of such an explanatory variable candidate should be ignored. In OEPP, "P" (positive), "N" (negative), or "F" (free implying undetermined) is specified for each of all explanatory variable candidates including a constant term, if any, in exactly the same order as the order of all explanatory variable candidates entered in the equation format.

MAGNITUDE CRITERION

From economic theory on stability, the coefficient value of some explanatory variable is expected to be within a certain range. In this case, the sign criterion is not enough, so that magnitude criterion is of great use. For instance, values 0.75 and 1.25 are supposed to imply marginal propensity to consume. From the viewpoint of the sign criterion demanding a positive sign, both figures are acceptable. However, the marginal propensity to consume greater than one leads to the fact that people spend more than additional disposable income for consumption and such an economy collapses eventually if it is not confronted with an unusual situation during a short period. Hence, 1.25 is judged to be unsatisfactory or inappropriate.

A lagged regressand is sometimes used as an explanatory variable candidate which explains the inertia effect. The coefficient of such a lagged regressand is expected to fall within the range of -1 and 1. If not, the movement or behavior of such a regressand may diverge and lead to instability. Hence, checking the magnitude of the coefficient of some explanatory variable becomes important to find the best meaningful subequation. In OEPP, three kinds of magnitude criterion are available. They are (a) range, (b) lower bound, and (c) upper bound.

If the magnitude criterion is not applied to some explanatory variable candidate, it should be ignored with respect to the coefficient of such an explanatory variable candidate. "F" is specified for such an explanatory variable candidate in OEPP.

T-TEST CRITERION

After passing either or both of the economic criteria, the estimated coefficients of each meaningful subequation should be sieved with statistical criteria. The T-test is one of the statistical criteria. Instead of T-test, F-test can be used as a statistical criterion. T-test at some significance level should be made for all or some of the estimated coefficients of each meaningful subequation. In some cases, if T-test is rigorously made for all of the estimated coefficients, a researcher cannot find any satisfactory equation for some explained variable. Accordingly, he may want to apply T-test at some rigorous significance level for the estimated coefficients of some crucially important explanatory variable candiates and T-test at some loose significance level for the estimated coefficients of those remaining in a meaningful subequation. Hence, it is convenient that a package can optionally make T-test at a specified significance level (a) for the coefficients of all non-constant explanatory variable candidates and the constant term adopted in a meaningful subequation, (b) for the coefficients of non-constant explanatory variable candidates, and (c) for the coefficients of some non-constant explanatory variable candidates specified by a researcher and not for the coefficients of the remaining.

In OEPP, 1%, 5%, and 10% significance level T-test can be made optionally in the three ways mentioned above. It is possible to install different significance level T-tests in OEPP.

DURBIN-WATSON STATISTIC CRITERION

As far as time series data are concerned, the Durbin-Watson statistic criterion is useful to check the autocorrelation of the disturbance term when ordinary least squares method is applied. Even if the estimated coefficients of a meaningful subequation pass all of the sign, magnitude, and T-test criteria, the strong autocorrelation of the disturbance term leads a researcher to reject the adoption of such a meaningful subequation because the estimated coefficients have biases. The Durbin-Watson statistic test at some significance level should be applied.

RESIDUAL PERCENTAGE CRITERION OR RELATIVE ERROR CRITERION

In order to avoid the case in which most of the observations of an explained variable are well tracked but one or two observations are not really well tracked by an estimated subequation, residual percentage or relative error criterion, which can be regarded as a mathematical criterion, should be applied. Residual percentage or relative error is defined as

$$|100*(Y_i(t)-EY_i(t))/Y_i(t)| \quad \text{for each i and t} \qquad (46)$$

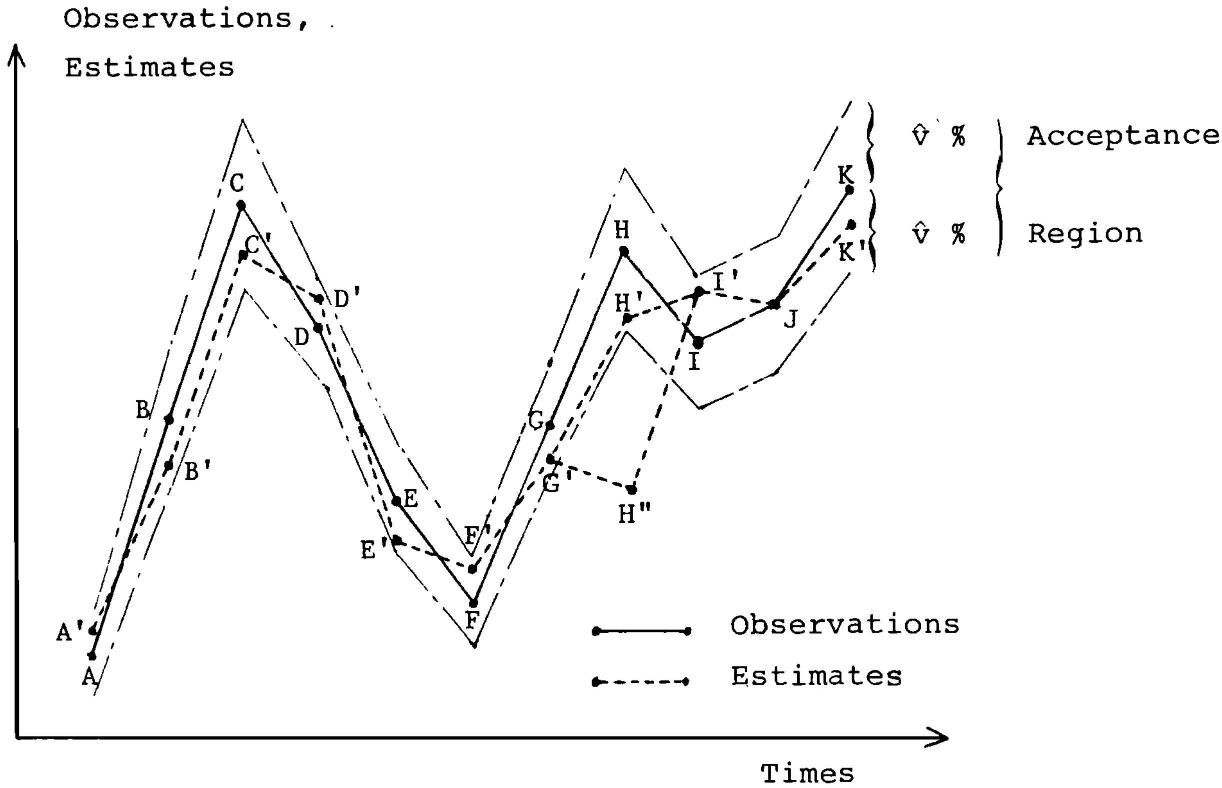
where $Y_i(t)$ and $EY_i(t)$ imply the observation (or transformed value of some kinds of observations for a transformed explained variable) and estimate of an explained variable in area i at time t, respectively. If all residual percentages of the estimation period fall short of some level, called residual percentage criterion value specified by a researcher, the estimated meaningful subequation can be considered to well track the behavior of the explained variable. The selection of residual percentage criterion value completely depends on a researcher.

When residual percentage criterion value, say $\hat{v}$, is given by a researcher, the acceptance region is derived as follows:

$$Y_i(t)(1-\hat{v}/100) \leq EY_i(t) \leq Y_i(t)(1+\hat{v}/100) \quad \text{for } \hat{v} > 0 \qquad (47)$$

(47) is represented by the acceptance region in Figure 2, where time series data are used so that i = 1. The meaningful subequation, even if it passes the economic and statistical criteria, is not considered good when it generates the estimates of the explained variable including estimate H" in Figure 2. Only one estimate H" leads to the rejection of such a subequation.

Figure 2. Residual Percentages (or Relative Errors) and Turning Points of Observations and Estimates by a Subequation in the Case of Time Series Data (i = 1).



TURNING POINT ERROR PERCENTAGE CRITERION

In some studies, the explanation of a sudden change in the upward or downward trend shown by the observations of an explained variable is quite important. For instance, investment falls down suddenly from the upward trend. This sudden change is called "a turning point" showing Λ-shape. Similarly, the opposite turning point showing V-shape is often seen. The best meaningful subequation must explain sufficiently why the situation is suddenly changed from upward to downward or vice versa. Even if the estimates of a meaningful subequation which passes all (or some) above criteria applied track well the observations showing an upward trend, such a meaningful subequation cannot be selected as the best when the estimate corresponding to the observation which drops from the upward trend still indicates an upward trend.

Let K_1 and K_2 denote the numbers of Λ-shape and V-shape turning points shown by the observations of an explained variable during an estimation period, respectively, and let K_3 and K_4 stand for the numbers of the Λ-shape and V-shape turning points not sufficiently explained by an estimated meaningful subequation, respectively. Then, turning point error percentage is defined as follows:

$$100*(K_3+K_4)/(K_1+K_2) \qquad \text{if } K_1+K_2 \neq 0. \tag{48}$$

Accordingly, the turning point error percentage criterion, which is regarded as a mathematical criterion, is that unless the turning point error percentage calculated by (48) exceeds a value $\hat{W}$ specified by a researcher, such a meaningful subequation is judged to be satisfactory, where $\hat{W}$ is called turning point error percentage criterion value. Hence, the turning point error percentage exceeding $\hat{W}$ leads such a meaningful subequation to be unsatisfactory.

For instance, suppose that a turning point error percentage criterion value is 20 and there exist 10 turning points, Λ-shape as well as V-shape, shown by the observations of an explained variable at hand. If all turning points except for one or two are well explained by a meaningful subequation for the explained variable, such a meaningful subequation is judged to be satisfactory. But, if more than two turning points are not well tracked by a meaningful subequation, such a meaningful subequation is regarded as unsatisfactory, because its turning point error percentage becomes more than 20%.

Needless to say, zero turning point error percentage is desirable. Unless any turning point is observed during an estimation period, the turning point error percentage criterion does not make sense.

In Figure 2, Λ-shape turning points are observed at points C and H, while V-shape turning points are observed at points F and I. Estimated point H" does not explain why a sudden change happened at point H. Estimated point I' does not explain well why a change from downward to upward happened at observation point I. If the estimates of a meaningful subequation are A' to G' and H' to K', the turning point error percentage is 25%. On the other hand, if the estimates are A' to G' and H" to K', the turning point error percentage is 50%.

FIXED, PARTIALLY UPGRADING OR GLOBALLY UPGRADING FITTING CRITERION

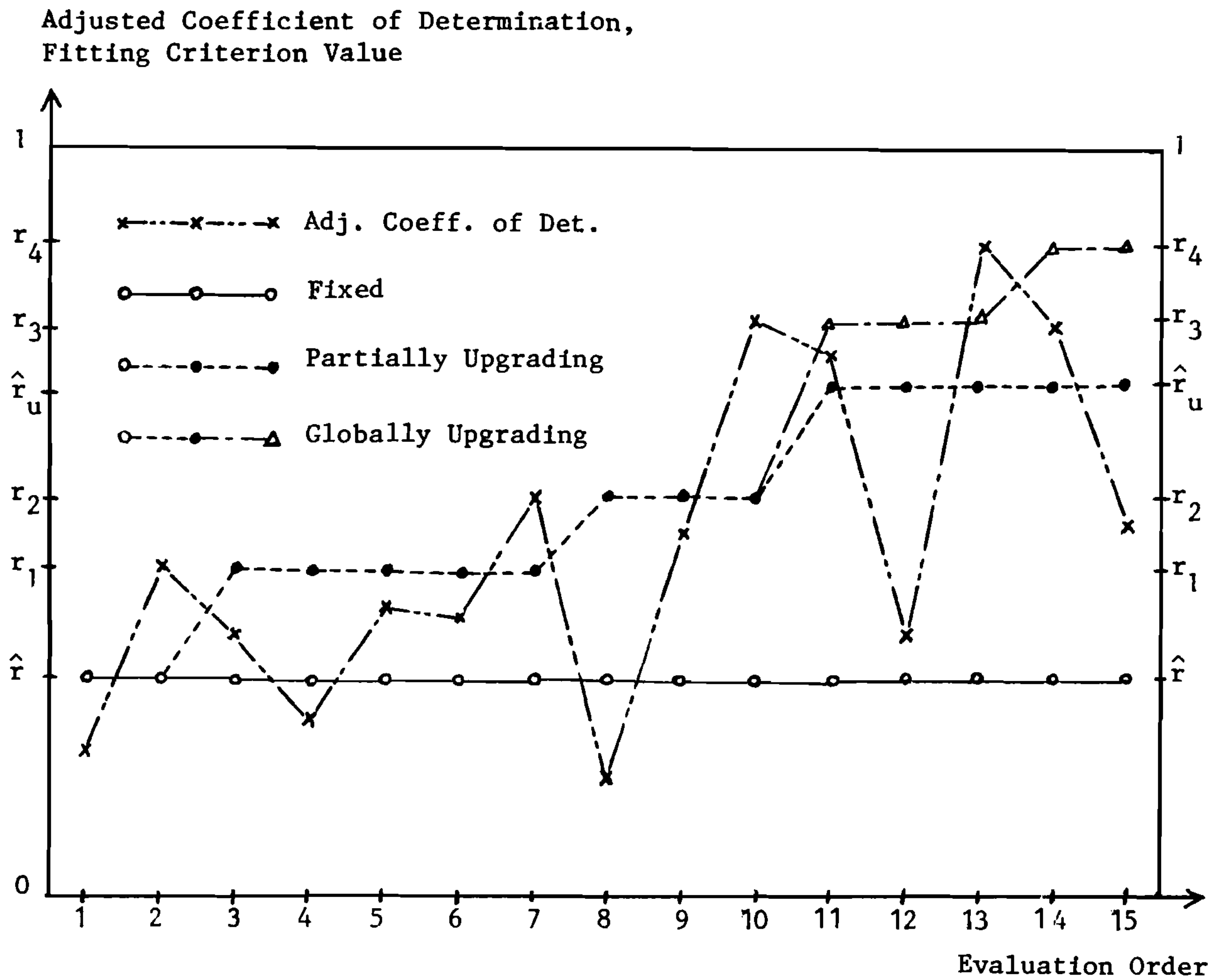
The last criterion to be applied is the fitting criterion. This criterion is based on the coefficient of determination adjusted by the number of explanatory variable candidates. The meaningful subequation with the highest coefficient of determination adjusted by the number of its explanatory variable candidates is regarded as the best, if it has already passed all (or some) above criteria which a researcher applies. If a fitting criterion value $\hat{r}$ satisfying $0<\hat{r}<1$ is specified by a researcher, only the meaningful subequations which satisfy all (or some) above applied criteria and possess the coefficients of determination adjusted by the numbers of their explanatory variable candidates equal to or greater than $\hat{r}$ are judged appropriate for printing. The case in which $\hat{r}$ is always fixed for all estimated meaningful subequations is called "fixed fitting criterion" in this article and in OEPP. It is possible that many subequations satisfy all (or some) above applied criteria and the fixed fitting criterion. In this case, the researcher has to find the best subequation by himself with the comparison of adjusted coefficients of determination and/or his own equation evaluation criterion (or criteria).

If a meaningful subequation clearing all (or some) above applied criteria indicates the adjusted coefficient of determination greater than $\hat{r}$, the fitting criterion value $\hat{r}$ can be replaced by the value of this adjusted coefficient of determination and a new fitting criterion value can be applied to the remaining meaningful subequations. If this process is repeated as many times as possible, the meaningful subequation printed last which possesses the highest adjusted coefficient of determination becomes the best. This case is here called "globally upgrading fitting criterion". In order to find the second best and third best meaningful subequations, the computer should stop raising a fitting criterion value after it reaches a certain level $\hat{r}$, where $\hat{r}<\hat{r}_u<1$, specified by a researcher who wants the second and third best. Then, not so many meaningful subequations clearing all (or some) above criteria applied and indicating the adjusted coefficients of determination equal to or greater than $\hat{r}_u$ are printed. In this case, the researcher must find the second or third best meaningful subequation by himself with the comparison of adjusted coefficients of determination and/or his own equation evaluaation criterion (or criteria). It is not difficult to do so because the number of printed meaningful subequations is not too large. Of course, if $\hat{r}_u$ is set too close to 1, only the best meaningful subequation is printed but no second and third best meaningful subequations are printed at all. Then, the researcher has to reset $\hat{r}_u$ at a value a little bit smaller than the best meaningful subequation's adjusted coefficient of determination. This case is here called "partially upgrading fitting criterion".

The specifications of the fixed, partially upgrading, and globally upgrading fitting criteria are: " $\hat{r}$, 0 ", " $\hat{r}$, $\hat{r}_u$ ", and " $\hat{r}$, 1", respectively, where $0 < \hat{r} < \hat{r}_u < 1$. Figure 3 explains in detail fixed, partially upgrading and globally upgrading fitting criteria.

It is possible to make the computer memorize the best (and second best, etc.) meaningful subequation(s) which is temporarily best (and second best, etc.) until the last meaningful subequation is estimated and then print only the best (and second best, etc.) meaningful subequation(s). The disadvantage is that the core size becomes larger than in the case mentioned above.

Figure 3: Fixed, Partially Upgrading, and Globally Upgrading Fitting Criteria in the Case Where It Is Assumed That 15 Meaningful Subequations Have Already Satisfied the Previous Applied Equation Evaluation Criteria.



Fixed		$\overline{P}$	$\overline{P}$		$\overline{P}$	$\overline{P}$	$\overline{P}$		$\overline{P}$	$\overline{P}$	$\overline{P}$	$\overline{P}$	$\overline{P}$	$\overline{P}$	$\overline{P}$
P.U.		P					P			$\overline{P}$	$\overline{P}$		$\overline{P}$	$\overline{P}$	
G.U.		P					P			P			$\hat{P}$		

where 1,2,3,...,15 imply the numbers of meaningful subequations which are assumed to have already satisfied the previous applied equation evaluation criteria, Fixed, P.U., and G.U. stand for fixed, partially upgrading, and globally upgrading fitting criterion, P implies the print of the subequation indicated by a number on the horizontal axis which was temporarily best and should be ignored, $\overline{P}$ implies the print of subequation which is a best subequation candidate, and $\hat{P}$ implies the print of the best subequation in the sense defined in this article and in OEPP.

CAN THE STEPWISE REGRESSION PROCEDURE ALWAYS FIND THE BEST SUBEQUATION?

Only when OLS can be used can the stepwise regression procedure efficiently find the statistically best subequation which possesses the highest coefficient of determination adjusted by the number of explanatory variables adopted in it and the coefficients significant at the level specified by a researcher. However, the statistically best subequation is not always acceptable, especially from the viewpoint of economics. A behavioral equation, for instance, has to have the coefficients of explanatory variables plausible from the point of view of human rational behavior. For instance, the signs and/or magnitudes of coefficients must be consistent with human rational behavior. Therefore, the stepwise regression procedure cannot always find the practically best subequation defined in OEPP except for a special case, in which all of the five conditions (a) to (e) to be explained below hold.

The stepwise regression procedure can be used if the following five conditions hold at the same time:

(a) OLS can be applied.

(b) All non-constant explanatory variable candidates are completely optional.

(c) The signs and/or magnitude constraints of the coefficients of any explanatory variable candidates are not available or not needed to evaluate estimated subequations, even if they are available.

(d) Minimum variance is not important.

(e) Residual percentage criterion and turning point error percentage criterion are not important.

If (b) holds in the example of beef demand mentioned above, a subequation which does not include any real income variable can be the best subequation from the viewpoint of the stepwise regression procedure. Can we accept such a best subequation as the demand function of beef? It is natural to reject such a best subequation. Thus, (b) does not hold in many cases.

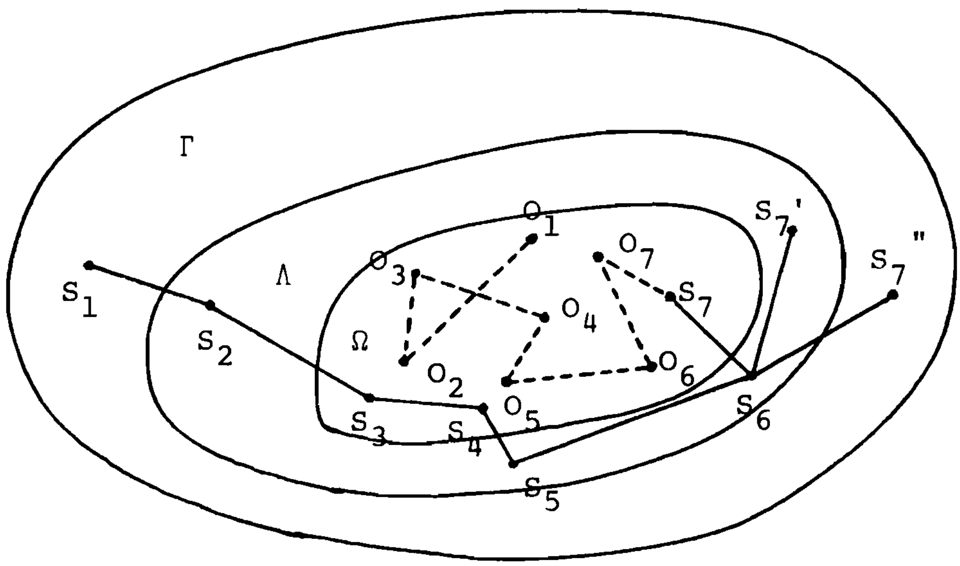
If (c) holds in the same example of beef demand mentioned above, a subequation which has the negative sign of the coefficient of real income variable and/or the positive sign of the coefficient of relative beef-pork price can be the best subequation from the viewpoint of the stepwise regression procedure. Do consumers want to buy more beef if real income is decreased or if beef price becomes relatively higher than pork price? It is rational that consumers reduce the purchase of beef if real income is decreased or if beef price becomes relatively higher than pork price, in the case where beef is not an inferior good. Thus, (c) does not hold in many cases, either.

If (d) holds for time series data, the coefficients of the best subequation obtained by the stepwise regression procedure do not necessarily reflect the best estimator with the minimum variance.

Residual percentage criterion checks smoothness of fitting, while fitting criterion cannot check it. The turning point error percentage criterion is entirely different from the fitting and T-test (or F-test) criteria.

Figure 4 illustrates the sequence of estimating (and printing) subequations by the stepwise regression procedure and the sequence of printing subequations by OEPP. If the subequation at the final step belongs to Ω as shown by S7, the stepwise regression procedure is most efficient. On the other hand, if the subequation at the final step belongs to $\Gamma - \Omega$ as shown by S7' or S7", the researcher cannot accept S7' to explain the behavioral (or institutional or technical) relation between the explained and explanatory variables, except for a case of new findings. The reason is that the coefficients are inconsistent with human rational behavior. Unless both exclusively important variables and exclusively optional variables are candidates, the stepwise regression procedure ends up with S7 or S7', once it estimates a subequation belonging to Λ as shown by S2. But it never ends up with S7". On the contrary, if exclusively important variables and/or exclusively optional variables are explanatory variable candidates, it is possible that the stepwise regression procedure ends up with S7", even after it finds a subequation belonging to Λ. If, and only if, $\Gamma=\Lambda=\Omega$, then the stepwise regression procedure is always the most efficient in the case where OLS can be applied. Unfortunately, the case in which $\Gamma=\Lambda=\Omega$ holds is quite rare in the field of behavioral science, especially applied economics.

Figure 4. OEPP Combinatorial-Variable-Selection Method and Stepwise Regression Procedure



●——●——● Sequence of subequations estimated by the stepwise regression procedure

●- - - -●- - - -● Sequence of subequations printed by OEPP combinatorial-variable-selection method

Γ : Set of all subequations derivable by selecting, in the combinatorial manner, the non-constant explanatory variable candidates given by a researcher.

Λ : Set of all meaningful subequations.

Γ-Λ: Set of all meaningless subequations.

Ω : Set of the meaningful subequations satisfying the sign (and magnitude) criterion, where other criteria are unimportant or ignored.

where S_1, S_2, ..., S_6 stand for the estimated subequations prior to the final step by the stepwise regression procedure, S_7, S_7', and S_7" stand for the three cases of the best subequation at the last (seventh) step, and O_1, O_2, ..., O_7, and S_7 stand for the subequations printed in this order in the printout by OEPP.

OEPP estimates all meaningful subequations belonging to Λ and prints (a) the subequation with the theoretically reasonable coefficients significant at a specified level which is first discovered to belong to Ω or (b) the subequation belonging to Ω and having the higher adjusted coefficient of determination with the coefficients significant at the specified level than the prior subequation belonging also to Ω.

Sequence 01, 02,...,07, S7 by OEPP stands for the sequence of printing the subequations upgraded in order. Thus, the last subequation printed in the output is the practically best subequation which the researcher can accept from the viewpoint of behavioral science and econometrics (or regression analysis).

Let us re-examine the stepwise regression procedure. Suppose, for simplicity, that beef demand quantity BD is regarded as a linear form of all or some of real income variable Y, beef-pork price ratio BPP, and beef-egg price ratio BEP, where eggs are considered to be a weak beef substitute from the viewpoint of animal protein and other relative beef prices are ignored. Thus, we can write the following equation format:

BD=F($C,Y,BPP,BEP) (49)

Suppose that the simple correlation coefficient matrix is as follows:

	BD	Y	BPP	BEP
BD	1			
Y	0.8	1		
BPP	-0.6	-0.4	1	
BEP	-0.85	-0.88	0.2	1

Since the absolute value of the simple correlation coefficient between BD and BEP is the highest, BEP is selected at the first step of the stepwise regression procedure. So the following is estimated:

BD=F($C,BEP) (50)

Since the absolute value of the simple correlation coefficient between Y and BEP is quite high, Y is not selected but BPP is selected at the last step. We have

BD=F($C,BEP,BPP) (51)

which gives the highest adjusted coefficient of determination. Who wants to accept (51) as the beef demand function? Nobody does! However, when an economist estimates equation candidates for beef demand one by one, he will estimate all or some of the following three equations:

BD=F($C,Y,BPP,BED) (52)

BD=F($C,Y,BEP) (53)

BD=F($C,Y,BPP) (54)

Then, the economist selects, for instance, (54) as the best equation for beef demand, if (54) is reasonable. This illustrates how dangerous the "blind" application of the stepwise regression procedure is. However, OEPP can estimate only (52), (53), and (54) by entering the following equation format:

BD=F($C/Y/<BPP,BEP>) (55)

and find (54) as the best subequation automatically if the necessary information about equation evaluation is entered together with equation format (55).

Needless to say, a priori information about explanatory variable candidates must be correctly used, whenever it is available. Hence it is possible for a researcher to use time and labor saved by OEPP on the essential work of research such as what variables affect human rational behavior, how institutional variables affect government decision-making, what input variables are key variables in agronomic research, and so on.

SOME FEATURES OF OEPP CONCERNING ESTIMATION

OEPP is designed to be useful for education as well as research in the field of input-output analysis, estimation, and economic simulation. Now, let us concentrate on only the features of OEPP concerning estimation. The estimation methods available in OEPP are as follows:

Estimation Methods Installed in OEPP*

1. Ordinary Least Squares (OLS)

2. Constrained Ordinary Least Squares (COLS)

3. Generalized Least Squares with diagonal Ω (GLSD) or Weighted Least Squares

*Nonlinear estimation method will be included in OEPP this year.

4. Generalized Least Squares with non-diagonal Ω (GLSN)

5. Deterministically-Constrained Generalized Least Squares with diagonal Ω (DCGD)

6. Deterministically-Constrained Generalized Least Squares with non-diagonal Ω (DCGN)

7. Probabilistically-Constrained Generalized Least Squares (Mixed Estimation) with diagonal Ω (PCGD)

8. Probabilistically-Constrained Generalized Least Squares (Mixed Estimation) with non-diagonal Ω (PCGN)

9. Instrumental Variable (IV)

10. Cochrane-Orcutt Method (CO)

11. Constrained Cochrane-Orcutt Method (CCO)

12. Hildrath-Leu Method (HL)

13. Constrained Hildrath-Leu Method (CHL)

14. Prais-Winston Method (PW)

15. Constrained Prais-Winston (CPW)

16. Ridge Regression (R)

17. Almon Lag Distribution Method (A)

18. Durbin First-Order Autocorrelation Two Step Method (DF)

19. Durbin Second-Order Autocorrelation Two Step Method (DS)

20. Three Pass Least Squares (3PLS)

21. Chow Test (C)

22. Indirect Least Squares (ILS)

23. Two stage Least Squares (2SLS)

24. Constrained Two Stage Least Squares (C2SL)

25. Limited Information Maximum Likelihood, Limited Information Least Generalized Variance, or Limited Information Least Variance Ratio (LI)

26. Three Stage Least Squares (3SLS)

27. Constrained Three Stage Least Squares (C3SL)

28. Iterative Three Stage Least Squares (I3SL)

29. Iterative Constrained Three Stage Least Squares (IC3S)

30. Totally-Constrained Three Stage Least Squares (TC3S)

31. Iterative Totally-Constrained Three Stage Least Squares (ITC3)

32. Globally-Constrained Three Stage Least Squares (GC3S)

33. Iterative Globally-Constrained Three Stage Least Squares (IGC3)

34. Zellner (Seemingly Unrelated Regression) Method (ZELL)

35. Constrained Zellner Method (CZ)

36. Iterative Zellner Method (IZ)

37. Iterative Constrained Zellner Method (ICZ)

38. Totally-Constrained Zellner Method (TCZ)

39. Iterative Totally-Constrained Zellner Method (ITCZ)

40. Globally-Constrained Zellner Method (GCZ)

41. Iterative Globally-Constrained Zellner Method (IGCZ)

where the combinatorial-variable-selection method cannot be used for 17, 21, 22, and 26 - 41, the symbol in parentheses stands for the command for an estimation method, "totally-constrained" implies that a linear constraint(s) is imposed only on coefficients of variables in different equations, e.g., integrability condition for net social quasi-welfare problem, and "globally-constrained" implies that a linear constraint(s) is imposed on the coefficients of variables in an equation(s) and, furthermore, a different linear constraint(s) is imposed on the coefficients of variables in different equations.

In the case where each undergraduate student in computer-oriented econometrics class does not have his own registered number for the computer, some undergraduate students may mistakenly or purposely put a lot of explanatory variable candidates in an equation format and run the computer with the combinatorial-variable-selection method. Millions of subequations may be estimated and printed in vain. The teacher cannot find who used up the educational budget allocated to his computer-oriented econometrics class and continue teaching without additional budget. Even if he finds and punishes someone who used up the budget, the budget does not come back. In order to prevent this case from occurring, OEPP requires a password if a user wants to put many explanatory variable candidates in an equation format and allows him to estimate subequations if the password is correct. Professors and graduate students working for a project must register their password and use it whenever they run a big job.

However, undergraduate students cannot use OEPP whenever the restrictions imposed on undergraduate students' use become effective, because no password is given to them.

They get only a warning or error message in this case. For example, the following are some of the present restrictions imposed on undergraduate students in my class:

(a) / absolutely important variables / → no restrictions

(b) < optionally important variables > → only one "<...>" and max. 6 variables in it

(c) << exclusively important variables >> → one "<<...>>" and max. 5 variables in it or two "<<...>>", "<<...>>" and max. 5 variables in them in total

(d) <* exclusively optional variables *> → only one "<*...*>" and max. 3 variables in it

(e) completely optional variables → the sum of the numbers of variables in (b) and completely optional variables must be equal to or less than 6.

(f) PLOT card → graphs of max. 10 variables' data

(g) SCAT card → scatter diagrams of max. 10 pairs of variables

ALMON LAG DISTRIBUTION AND ALL POSSIBLE DEGREES OF A POLYNOMIAL

Let us denote an explained variable, a non-Almon-lag variable, and an Almon-lag varieble by Y,Z, and X, respectively. If an Almon-lag variable has the maximum time lag number 4, the equation format must be as follows:

Y=F($C,Z,X,X(-1),X(-2),X(-3),X(-4)) (56)

where non-Almon-lag variable(s) must be entered prior to the Almon-lag variable and the Almon-lag variables with different time lag numbers 0, 1, 2, 3, and 4 follow non-Almon-lag variable(s) in this order.

In this case, the possible degrees of a polynomial for (56) are 1, 2, and 3. It is usual that it is not known a priori which degree of a polynomial is the most appropriate for (56). Therefore, OEPP can automatically estimate (56) with respect to each of all possible degrees of a polynomial, if a researcher wishes, so that three (sub)equations can be estimated in this case. Furthermore, OEPP can automatically estimate, with the indication of the command parameter requiring elimination of variables from the last variable in order,

Y=F($C,Z,X,X(-1),X(-2),X(-3)) (57)

with respect to degrees of a polynomial 1 and 2, and

Y=F($C,Z,X,X(-1),X(-2)) (58)

with respect to degree of a polynomial 1 in addition to (56) with respect to degrees of a polynomial 1,2, and 3. Accordingly, six subequations are estimated in total.

In general, $(J-1)\cdot J/2$ subequations can be automatically and continuously estimated and the best subequation is selected among them, if the maximum time lag number of Almon-lag variables is J where $J \geq 2$.

CONCLUSION

It is quite important and beneficial to develop an econometric program package which can deal with most of the trial and error computer work during a research period. In applied econometric research, the estimation of a large-scale simultaneous equations model takes a lot of research time. The best equation for each explained variable is usually found through a trial and error process, by an existing econometric program package, so that much waste of labor and resources is generated. In order to shorten the research period, save brain labor and resources, and reduce calculation and printing costs, an efficient method is required. The method proposed in this paper is to make the computer automatically generate and estimate "only meaningful subequations" and then find the best among them. Unless any a priori information is available for non-constant explanatory variable candidates, all subequations derived by selecting, in the combinatorial manner, the non-constant explanatory variable candidates, which are all meaningful, must be estimated and then the best among them must be found. However, if a priori information is available for all or some of the non-constant explanatory variable candidates, it "must" be classified into four categories of absolutely important, optionally important, exclusively important, and exclusively optional variables according to the degrees of importance and certainty. Then, only meaningful subequations must be chosen and estimated and then the best among them must be found. Four categories are here characterized by putting non-constant explanatory variables between a pair of slashes, a pair of less-than and greater-than signs, a pair of double less-than and greater-than signs, and a pair of less-than signs followed by asterisk and greater-than sign preceded by asterisk in an equation format.

To find the best meaningful subequation, the economic criteria, statistical criteria, and mathematical criteria are proposed. The economic criteria are sign and magnitude criteria. The statistical criteria are the T-test (or F-test) and Durbin-Watson statistic test. The mathematical criteria are residual percentage, turning point error percentage, and fitting criteria. Although Theil's inequality coefficient criterion, simple correlation coefficient criterion and the like are possible, they may be redundant because they are more or less similar to the fitting criterion. If Chow test needs to be applied, it can be applied to the best meaningful subequation by treating it as an equation. A researcher can apply all the proposed criteria or some of them to find the subequation which is best from his viewpoint. If this method is applied to, for instance, 20 explained variables in each of five runs of the computer a day, it is possible for an experienced applied econometrician who has a good data bank to finish estimating a simultaneous equations model which consists of 1,000 equations and identities in a few months.

Nowadays labor and resources such as paper for output and electricity are not cheap and in the future they may become quite dear. Any kind of computer package must be of time-, labor-, and resource-saving as well as cost-reducing type.

REFERENCES

Dixton, W.J. 1977. BMDP, Biomedical Computer Program, Health Sciences Computing Facility, Department of Biomathematics. School of Medicine, University of California, Los Angeles, U.S.A.: University of California Press.

Draper, N.R. and H. Smith. 1966. Applied Regression Analysis, John Wiley & Sons, New York.

Hall, B.H. 1977. TSP, Time Series Processor, 2nd Ed., Harvard Institute of Economics Research. Harvard University, Cambridge, Mass., U.S.A.

Holwig, J.T. 1978. SAS, Statistical Analysis System. Raleigh, N.C., U.S.A.: SAS Institute Inc.

Moriss, N. 1977. Software Package for Economic Modeling. RR-77-21. Laxenburg, Austria: International Institute for Applied Systems Analysis.

National Bureau of Economic Research, Inc. 1975. TROLL, Time-Shared Reactive On-Line Laboratory. MIT Information Processing Services, MIT, Cambridge, Mass., U.S.A.

Nie, N.H., C.H. Hull, J.C. Jenkins, K. Steinbrenner, and D.H. Bent. 1970. SPSS, Statistical Package for the Social Sciences, 2nd Ed. National Opinion Research Center, University of Chicago, Chicago, U.S.A.

Onishi, H.H. 1978. OEPP, Onishi Econometric Program Package. Institute of Socio-Economic Planning, University of

Tsukuba, Sakura, Ibaraki, Japan.

Plasser, K. 1978. IAS, Inter-Active Simulation System. Institute for Advanced Studies, Vienna, Austria.

Ryan, T.A.J., B.L. Joiner, and B.F. Ryan. 1980. MINITAB. Minitab Project, Statistics Department, The Pennsylvania State University, Pennsylvania, U.S.A.

There are many other computer program packages in Japan and in other countries which can deal with input-output analysis, econometric estimation, and econometric simulation. Most of the packages are offered as software services but cannot be bought.

Draper and Smith used the term "all possible regressions" instead of combinatorial-variable-selection method which I have used in this article as well as in the OEPP manual. Their book "Applied Regression Analysis" explains well all possible regressions, the backward elimination procedure, the forward selection procedure and the stepwise regression procedure in Chapter 6. However, it does not say much about a selection procedure in which the best (sub)equation selected in these regression procedures is always acceptable, especially from the viewpoint of behavioral sciences.