No.181 (83-4)

VARIABLE CHOICE FOR CONSTRAINED ORDINARY LEAST SQUARES IN ECONOMIC ANALYSIS

by

Haruo ONISHI

March 1983

Variable Choice for Constrained Ordinary Least Squares in Economic Analysis

University of Tsukuba

Haruo ONISHI

ABSTRACT

Variable classification used for ordinary least squares can be extended to linearly-constrained ordinary least squares. Variable classification based on knowledge of a research field and the imposition of constraints on explanatory variable candidates enables the computer to generate, estimate and evaluate all possible meaningful subsets. It is possible to automatically choose the best subset which passes various criteria or tests, satisfies the constraints and has the highest coefficient of determination adjusted for the degree of freedom.

1. INTRODUCTION

In economic growth theories, a production function with constant returns to scale is often discussed. A method has been requested to solve a variable choice problem for constrained ordinary least squares (COLS) method. Onishi pointed out that, in general, explanatory variable candidates (called candidates from here on) can be classified into 8 basic groups with knowledge of a research field [5]. These are absolutely important (or forced or core), optionally important, gradually important, exclusively important, gradually optional, exclusively optional, completely optional and fixed groups. Variable

classification distinguishes meaningful from meaningless subsets in all possible subsets of a given set of all possible candidates. The generation of all possible meaningful subsets is decisively important, especially for social sciences like economics in which experiments are almost impossible. When we regard a set of all possible subsets as the domain for estimation, we can regard a set of all possible meaningful subsets as the feasible set for estimation. The best subset, if it exists, must be in the feasible set. Even if a meaningless subset fits almost perfectly the observations (or data) of an explained variable and passes all criteria applied, we cannot accept it. For instance, suppose that an equation is estimated for a beef demand function which includes the variables representing the wealth factor and some other minor factors like inertia effect of eating habit but it does not include any variables (e.g. beef/pork relative price, beef/poultry relative price, beef/fish relative price) representing the relative price factor. Even if it fits better than any alternative equations, it cannot be accepted by an economist as a beef demand equation, because the relative price factor affects consumers' decision of whether or not to buy beef.

In application, various criteria or tests are performed for estimated coefficients and estimated values to search for the best subset. Sign test for estimated coefficients is quite important. It is difficult to defend a beef demand equation in which the coefficient of a variable (e.g. real income or real income plus liquid assets) representing the wealth factor shows a negative sign,

although it passes all other criteria and fits well the observations of demand quantities for beef. Since beef is a superior good with respect to the wealth factor, a positive sign is expected for the coefficient of a variable for the wealth factor in the beef demand equation. Thus, it is quite difficult to efficiently search the best subset only with the criteria of T-test (or F-test) and goodness of fitting. Hence, economists load, estimate and evaluate all possible meaningful subsets, one by one, to search for the best subset. In order to reduce labor and cost of applied economic research and, at the same time, improve the quality, Onishi developed computer package OEPP which can handle variable choice problems for various estimation methods of regression analysis and econometrics [4].

2. VARIABLE CLASSIFICATION FOR CONSTRAINED ORDINARY LEAST SQUARES METHOD

Let X and X_i be a set of all possible candidates x_i (and their data) and the i-th meaningful subset of X. Suppose that $X=\{x_0,x_1,x_2,x_3\}$ where x_0 stands for a constant term. All possible subsets of X are (a) $X_1=\{x_0,x_1,x_2,x_3\}$, (b) $X_2=\{x_0,x_1,x_2\}$, (c) $X_3=\{x_0,x_1,x_3\}$, (d) $X_4=\{x_0,x_2,x_3\}$, (e) $X_5=\{x_0,x_1\}$, (f) $X_6=\{x_0,x_2\}$ and (g) $X_7=\{x_0,x_3\}$. If any subset can be chosen as the best with respect to candidates, all possible subsets become meaningful from the viewpoint of economics. Let X_* be a set of all possible meaningful subsets so that we have $X_*=\{X_1,X_2,X_3,X_4,X_5,X_6,X_7\}$. In general, if the number of all possible

non-constant candidates for an explained variable in question is N, 2^N-1 possible subsets exist. However, it is often seen that some subsets are meaningful from the viewpoint of economics but the remaining are meaningless. In this case, it is of crucial importance to distinguish meaningful from meaningless subsets by economic knowledge. This distinction prevents a meaningless subset from being chosen as the best. If it is a priori known in the above example that candidate x_2 is decisively important and candidates x_1 and x_3 are not necessarily important from the viewpoint of economics, candidate x_2 must always be selected and candidates x_1 and x_3 can be selected optionally to generate all possible meaningful subsets. In this case, subsets X_1, X_2, X_4 and X_6 include candidate x_2, so that they can be regarded as the meaningful subsets. On the other hand, all other subsets X_3, X_5 and X_7 are regarded as meaningless. The candidate of the best subset must be X_1, X_2, X_4 or X_6. Hence, meaningful set X_* consisting of only meaningful subsets becomes $X_*=\{X_1, X_2, X_4, X_6\}$.

In applied economic research, it is the first task of regression analysis to correctly classify candidates through economic theories, empirical studies, field surveys and/or hypotheses in order to generate only economically meaningful subsets. If an economist can generate only economically meaningful subsets, he can proceed to estimation and evaluation of all possible meaningful subsets by COLS method.

Onishi's variable classification for ordinary least squares method can be easily extended to the variable

classification for COLS method. If the candidates on which linear constraints are imposed are absolutely important, optionally important, gradually important and/or exclusively important from the viewpoint of economics, the constraints can be imposed on these candidates classified only with economic knowledge, because it is guaranteed that the candidates necessary for COLS method are automatically selected in all possible meaningful subsets. However, if the candidates on which the constraints are imposed are gradually optional, exclusively optional and/or completely optional from the viewpoint of economics, they must be treated as absolutely important, optionally important, gradually important and/or exclusively important candidates, from the viewpoint of estimation techniques, in such a way that the constraints can be imposed on all possible meaningful subsets. Even if some candidates on which the constraints are imposed are important but others are optional from the viewpoint of economics, the optional candidates ought to be treated just like the important candidates, when COLS method is utilized. Hence, it is possible to generate all possible meaningful subsets, for which COLS method can continuously be used, from a given set of all possible candidates by variable classification. A parsimonious variable choice method is desirable, but it must be reasonable in the sense that it can precisely choose at least the candidates indispensable for research at hand. Now, we are in a position to define the variable choice problem for COLS method.

3. VARIABLE CHOICE PROBLEM FOR CONSTRAINED ORDINARY LEAST SQUARES METHOD

I would like to introduce the following best subset problem to lead to the solution of a variable choice problem:

Find X_i and corresponding A_i which
(A) belongs to X_* generated from X,
(B) minimizes $(Y-X_iA_i)'(Y-X_iA_i)$ with respect to A_i subject to $D_i=C_iA_i$,
(C) clears sign test, magnitude test, T-test (or F-test), Durbin-Watson test ([1], [2], [3]), relative absolute error test [5] and turning point error test [5], and
(D) has the highest coefficient of determination adjusted for the degree of freedom,
where Y, X, X_*, X_i, A_i, C_i and D_i stand for an explained variable (and its data), a set of all possible candidates (and their data), a set of all possible meaningful subsets, the i-th meaningful subset in X_*, a vector of coefficients of candidates in X_i, a coefficient matrix of constraints imposed on A_i, and a vector of constants of the constraints imposed on A_i, respectively.

The statistically as well as economically best subset obtained in this problem is a sort of the practically best subset. Unless there are any other criteria to evaluate subsets, the above problem is regarded as a variable choice problem. However, if an economist has his unique criteria in addition to the set criteria in the above best subset problem, the best subset may not be most satisfactory from

the viewpoint of a variable choice problem. In this situation, I would like to suggest that an economist solve the best to the j-th best subset problem and choose the most satisfactory subset among the best j subsets. As an analogy, the j-th best subset problem is defined as the problem which possesses Conditions (A), (B) and (C) but in which Condition (D) is replaced with the j-th highest adjusted coefficient of determination in the above problem. In OEPP, he can choose, for instance, the best 10 subsets by specifying 10 on EQUA card with the other criteria if they exist and can obtain the most satisfactory subset among the best 10 subsets by checking them with his own criteria.

Let us explain how to solve the best subset problem. First of all, we make the following matrix and vector using all possible candidates and the corresponding constraints. From these matrix and vector, we select the sub-matrix and sub-vector for the i-th meaningful subset by the rule of variable classification and estimate the coefficients A_i and Lagrangean multipliers L_i for constraint set $D_i=C_iA_i$.

Matrix and vector of all possible candidates $\longrightarrow$ Sub-matrix and sub-vector of the candidates of X_i in X_*

$$\begin{bmatrix} X'X & C' \\ C & 0 \end{bmatrix} \text{ and } \begin{bmatrix} X'Y \\ D \end{bmatrix} \quad \begin{matrix} \text{Selection} \\ \text{by Variable} \\ \text{Classification} \end{matrix} \quad \begin{bmatrix} X_i'X_i & C_i' \\ C_i & 0 \end{bmatrix} \text{ and } \begin{bmatrix} X_i'Y \\ D_i \end{bmatrix}$$

Then, we can calculate A_i and L_i as follows:

$$\begin{bmatrix} A_i \\ L_i \end{bmatrix} = \begin{bmatrix} X_i'X_i & C_i \\ C_i & 0 \end{bmatrix}^{-1} \begin{bmatrix} X_i'Y \\ D_i \end{bmatrix} \qquad (1)$$

All or some of the following tests: sign test, magnitude test, T-test, Durbin-Watson test, relative absolute error test and turning point error test, are performed for the estimated coefficients and values as discrete or pass-or-fail tests. Finally, we choose as the best subset the meaningful subset which passes all discrete tests and has the highest adjusted coefficient of determination.

For the calculation of an adjusted coefficient of determination, we would like to utilize the following RR:

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

$$RR=1-(1-R)x(IxT-1)/(IxT-K-M) \qquad (3)$$

where $Y_i(t)$, $Y_i(t)$, $\bar{Y}$, I, T, K and M stand for the observation of an explained variable in area i at time t, the estimated value, the mean of the explained variable, number of areas (sectors, classes, companies, etc.), number of times, number of candidates selected in a meaningful subset and number of linear constraints, respectively.

4. TURNING POINT ERROR TEST FOR A DYNAMIC MODEL

When a dynamic simultaneous equation model (called model from here on) which includes lagged endogenous variables is built with time series or pooling data, turning point error test may be useful. Economic forecasting is often done with simulation by a model. It is often seen that all equations in a model show rather good fittings in partial tests or extrapolations but divergence occurs in the final

test or interpolation. One of the causes bringing divergence in the final test may be that an equation which has a rather high adjusted coefficient of determination, say more than 0.92, but does not track well the turning points of the observations (or the kinked behavior) of an explained (or left-hand side endogenous) variable in the stage of partial test is used in a model. When an economist cannot find a reasonable equation with a high adjusted coefficient of determination, he is inclined to introduce a one-time-period-lagged explained variable as a new candidate without sure economic knowledge. Then, he may be able to obtain an equation with a high adjusted coefficient of determination. However, this is a trap in many cases. This error cannot be discovered only by examining the estimated coefficients and T-ratios of the equation. Reasonable candidates are chosen. Signs and magnitudes of the estimated coefficnets are reasonable. T-ratios are quite high. Fitting is rather good. However, if this high adjusted coefficient of determination is brought about mainly by the introduction of the one-time-period-lagged explained variable candidate, a model including this equation may easily diverge in the final test. Onishi proposed two conditions for the definition of a turning point [5]. A turning point $Y_i(t)$ is a point which satisfies the following conditions (4) and (5):

$$\{Y_i(t)-Y_i(t-1)\}\times\{Y_i(t+1)-Y_i(t)\}<0 \qquad (4)$$

and

$$100 \times \mathrm{Min}[|\{Y_i(t)-Y_i(t-1)\}/Y_i(t)|, |\{Y_i(t+1)-Y_i(t)\}/Y_i(t)|] \geq u \quad (5)$$

where $Y_i(t)$ and u (%) stand for the observation of an explained variable in area i at time t and a definition value of a turning point specified by an economist (user), respectively.

Condition (4) is clear. Condition (5) depends on the nature of an economic problem. If an appropriate positive number is specified for u, the turning points which ought to be checked can be examined.

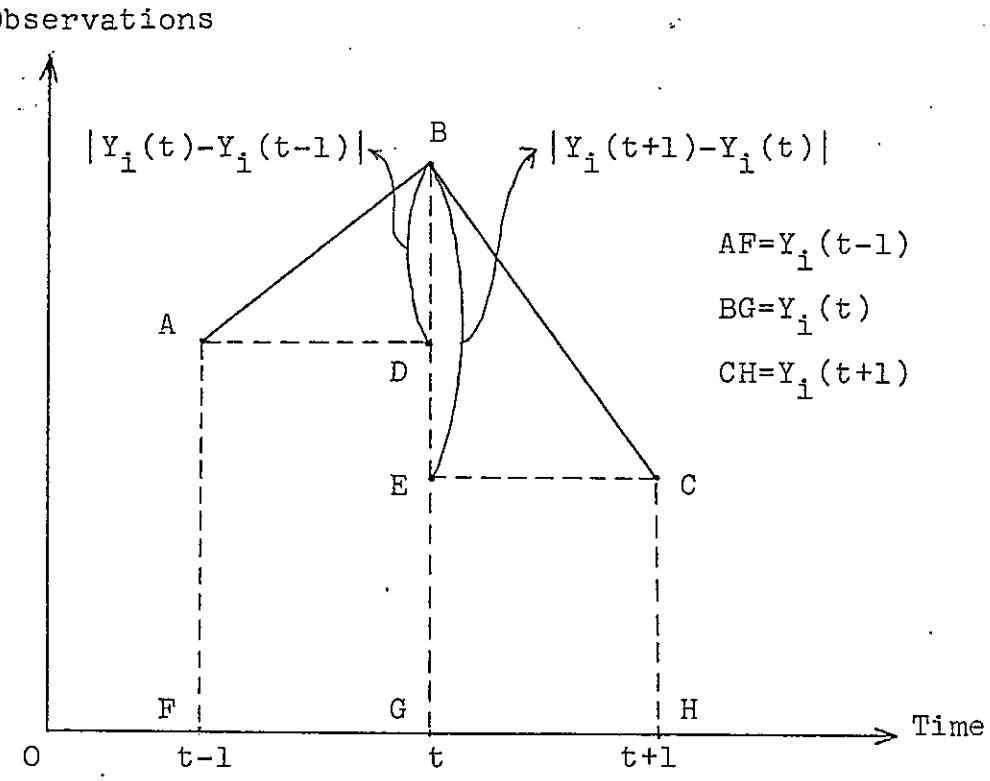

Figure 1. Definition of a Turning Point with respect to Area i and Time t

Suppose that investments (dollars) in area i are $Y_i(t-1)=1,000,000$, $Y_i(t)=1,000,001$ and $Y_i(t+1)=1,000,000$. These figures satisfy condition (4), because (1,000,001

-1,000,000)x(1,000,000-1,000,001)=-1<0. But this turning point is so flat that it should not be seriously checked, when the best subset is searched. By specifying, say, 1 (%) for u, we can avoid counting this $Y_i(t)$ as a turning point. Suppose as the second example that unemployment rates (%) in area i are $U_i(t-1)=2$, $U_i(t)=3$ and $U_i(t+1)=2$. Condition (4) is satisfied as (3-2)x(2-3)=-1<0. This turning point is rather steep and cannot be neglected. The estimated values corresponding to $U_i(t-1)$, $U_i(t)$ and $U_i(t+1)$ should track well the shape of this turning point, especially when unemployment rate is an endogenous variable in a model.

5. AN EXAMPLE

I would like to illustrate a variable choice method for COLS by a macro agricultural production function of Cobb-Douglas type by using the data for Japanese agriculture from 1965 to 1979. Conceivable factors for agricultural production are labor, land, capital, intermediate goods (and services) and technical progress. The factors labor, land and capital are so important that candidates representing them must be in an agricultural production function. We regard the factors intermediate goods and technical progress as optional factors. Here, even if any candidates representing these optional factors are not chosen in an agricultural production function, such a function is acceptable when the candidates representing the important factors are correctly chosen. We assume that the sum of the coefficients of the candidates which

represent the important factors labor, land and capital and the optional factor intermediate goods must be equal to 1, implying that constant returns to scale prevail with respect to amounts of inputs in agriculture. Concretely speaking, if a candidate for the optional factor intermediate goods is selected in a meaningful subset, the sum of the candidates representing the factors labor, land, capital and intermedidate goods must be 1. On the other hand, unless a candidate for the optional factor intermediate goods is selected in a meaningful subset, the sum of the coefficients of the candidates representing the factors labor, land and capital must be 1.

The explained variable and candidates are as follows: LY=LOG(Y)=LOG(total output), LL=LOG(L)=LOG(labor), LK=LOG(KA+KP+KM)=LOG(animal capital+plant capital+machinery capital), LRK=LOG(KA+KP+KMxR)=LOG(animal capital+plant capital+machinery capital multiplied by use rate), LKA=LOG(KA), LKP=LOG(KP), LKM=LOG(KM), LRKM=LOG(KMxR), LQ=LOG(Q)=LOG(intermediate goods), LAX=LOG(A-X)=LOG(planted acreage-abandoned acreage), LCAX=LOG((A-X) partially adjusted by rice cropping index), LT=LOG(T)= LOG(time trend) and $C=constant term. Since there are no data of total harvested acreage of all crops, the data of planted acreage minus abandoned acreage partially adjusted or unadjusted by rice cropping index are used.

We load the following format according to OEPP:

```
LY=F($C/LL/</LAX,LCAX/></LK,LRK,(LKA,LKP,LKM),(LKA,LKP,
     LRKM)/><*LT,T*>LQ)                                    (6)
```

/LL/ indicates that LL is an absolutely important candidate so that LL must be in a meaningful subset. </LAX,LCAX/> indicates that LAX and LCAX are exclusively important candidates in a sense that only LAX or LCAX must be in a meaningful subset. </LK, ...,LRKM)/> indicates that (a) LK, (b) LRK, (c) LKA, LKP and LKM and (d) LKA, LKP and LRKM are exclusively important candidates and only one of (a), (b), (c) or (d) must be in a meaningful subset. A group of candidates LKA, LKP and LKM or a group of LKA, LKP and LRKM are not only exclusively important but also fixed candidates. <*LT,T*> indicates that LT and T are exclusively optional candidates and only LT or T or none of them can be selected in a meaningful subset. LQ is a completely optional candidate.

Let us denote the coefficients of $C, LL, LAX, ..., LQ in Format (6) by a_0, a_1, a_2, ..., a_{14}. Then, the constraint for all possible candidates can be written as follows:

$$1=a_1+a_2+a_3+a_4+a_5+a_6+a_7+a_8+a_9+a_{10}+a_{11}+a_{14} \qquad (7)$$

The coefficients of this constraint can be expressed in matrix form as follows:

$$D=(1) \text{ and } C=(0,1,1,1,1,1,1,1,1,1,1,1,0,0,1) \qquad (8)$$

The order of the 0's and 1's in C of (8) corresponds to that of the candidates in Format (6). Only the coefficients of D and C are loaded in OEPP.

We include constant term $C in X, a set of all possible candidates, so that we have

$$X=\{\$C,LL,LAX,LCAX,LK,LRK,LKA,LKP,LKM,LKA,LKP,LRKM,\ LT,T,LQ\} \quad (9)$$

Although candidates LKA and LKP appear doubly in (9), they are never doubly selected in generating meaningful subsets by variable classification. Since there are 12 different non-constant candidates in X, $2^{12}-1=4,095$ possible subsets exist. It is known by variable classification that $2x4x3x2=48$ subsets are meaningful.

Let the first meaningful subset X_1 be

$$X_1=\{\$C,LL,LAX,LK,LT,LQ\} \quad (10)$$

where a constant term and candidates for the factors labor, land, capital, technical progress and intermediate goods are selected. It should be noted that total agricultural capital is selected in X_1.

When X_1 is selected, the constraint suitable for X_1 is

$$1=a_1+a_2+a_4+a_{14} \quad (11)$$

which is characterized by

$$D_1=\{1\} \text{ and } C_1=\{0,1,1,1,0,1\} \quad (12)$$

Furthermore, let the second meaningful subset X_2 be

$$X_2=\{\$C,LL,LCAX,LKA,LKP,LKM\} \quad (13)$$

where a constant term and candidates for the factors labor,

land and capital are selected. Three different items (animal capital, plant capital and machinery capital unadjusted by rate of use) of agricultural capital are selected in X_2.

Then, the corresponding constraint is expressed as follows:

$$1=a_1+a_3+a_6+a_7+a_8 \tag{14}$$

which leads to

$$D_2=\{1\} \text{ and } C_2=\{0,1,1,1,1,1\} \tag{15}$$

C_1 and C_2 can be easily selected, just as $X_1'X_1$ and $X_2'X_2$ are picked from X'X.

Let us introduce sign test, magnitude test, 5 % T-test, 5 % Durbin-Watson test, 5 % relative absolute error test, 1 % turning point error test and the minimum adjusted coefficient of determination 0.8 and find the best 5 subsets, if they exist. Positive signs are reasonable for all non-constant candidates, while either a positive or negative sign is acceptable for the constant term. Subsequently, 5 % one-tailed T-test is performed for all non-constant candidates in 48 meaningful subsets. In order to avoid unusually small (close to 0) or large (close to 1) coefficients of the important candidates which are significant at 5 %, we introduce the following magnitude tests:

$$0.5\geqq a_1>0.1,\ 0.5\geqq a_2+a_3>0.1,\ 0.5\geqq a_{14}>0.1 \text{ and}$$
$$0.5\geqq a_4+a_5+a_6+a_7+a_8+a_9+a_{10}+a_{11}>0.1 \tag{16}$$

It must be noted that $0.5\geqq a_2+a_3>0.1$ is actually used as $0.5\geqq a_2>0.1$ or $0.5\geqq a_3>0.1$, because </LAX,LCAX/> in Format (6) indicates that LAX is never selected together with LCAX in a meaningful subset. Similarly, $0.5\geqq a_4+a_5+\ldots+a_{11}>0.1$ implies that it is actually used as $0.5\geqq a_4>0.1$, $0.5\geqq a_5>0.1$, $0.5\geqq a_6+a_7+a_8>0.1$ or $0.5\geqq a_9+a_{10}+a_{11}>0.1$, because </LK,LRK, (LKA,LKP,LKM,)(LKA,LKP,LRKM)/> in Format (6). $0.5\geqq a_{14}>0.1$ is used only when completely optional candidate LQ is selected in a meaningful subset. In OEPP, the candidates corresponding to a_i's in (16) instead of a_i's themselves are loaded. Concretely speaking, the following are loaded:

0.5)LL>0.1, 0.5)LAX+LCAX>0.1, 0.5)LQ>0.1 and
0.5)LK+LRK+LKA+LKP+LKM+LKA+LKP+LRKM>0.1 (17)

where ")" is used for "$\geqq$".

Then, the following best equation was obtained in the CPU time of about 11 seconds at the cost of less than 24 US cents by FACOM-M200:

LY=1.19054+0.42384*LL+0.41682*LCAX+0.15934*LK+0.02512*T
(4.0828) (2.3857) (2.2623) (2.2876) (2.7461)
R=0.8624, RR=0.8249, SD=0.02709, FA=-0.1430, DW=2.241 (18)

where the numbers in parentheses, R, RR, SD, FA and DW stand for T-ratios, unadjusted coefficient of determination, adjusted coefficient of determination, standard deviation of disturbance term, first-order autocorrelation coefficient and Durbin-Watson statistic, respectively.

Although the best five subsets were requested, only one meaningful subset passed the above sign test, magnitude test, 5 % T-test, 5 % relative absolute error test and 1 % turning point error test and had the adjusted coefficient of determination which is greater than or equal to 0.8. Because 0.42384+0.41682+0.15934=1, Equation (18) satisfies the constraint. Then, we can choose Equation (18) for a linearly-constrained macro agricultural production function.

6. CONCLUSION

I proposed a variable choice method for linearly-constrained ordinary least squares. The method is available in computer package OEPP and an example was shown. With less time, labor and cost than before, a variable choice problem can be solved for linearly-constrained ordinary least squares method. As a result, it has become easier to build an econometric model and forecast by it. However, researchers in applied economics must keep in mind that all possible candidates for an explained variable must be introduced and classified in such a way that they express established economic theories, hypotheses of behaviors of economic units, hypotheses of institutions or technical relationships, economic thought or economic philosophy.

REFERENCES

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

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

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

[4] H. Onishi, OEPP(Institute of Socio-Economic Planning, University of Tsukuba, Ibaraki, Japan,1982)

[5] H. Onishi, An economic approach to the best subset in regression analysis, Computational Statistics and Data Analysis,forthcoming.