INSTITUTE OF POLICY AND PLANNING SCIENCES

Discussion Paper Series

No. 961

A Knowledge-based Variable Selection Method for Box-Cox Transformation

by

Haruo Onishi

December 2001

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

A Knowledge-based Variable Selection Method for Box-Cox Transformation

Haruo ONISHI*

In actual applications of regression analysis, users face two difficult problems. One is to find the most appropriate functional form, while the other is to search for the best subset derivable from a given set of all possible explanatory variables. Variable selection for the Box-Cox transformation may be useful to concurrently solve both problems. The purpose of this paper is to (1) concretely formulate the j-th OLS-best subset problem for the Box-Cox transformation, (2) introduce a knowledge-based computational method to solve it and (3) propose a solution to the (first) OLS-best subset problem ($j = 1$) or one selected by a user among solutions to the first j OLS-best subset problems ($j > 1$) solved in a run of a computer as a solution to a variable selection problem for the Box-Cox transformation. The integer j, specified by the user, depends on his scientific knowledge, criteria for statistical and data-analytic tests and model-building experience.

Key Words and Phrases: Regression analysis; Box-Cox transformation; Variable selection; J-th best subset problem; Variable classification; Meaningful subset; Practically best regression equation; Intellectual Statistical System OEPP.

1 Introduction

The role of data analysis is eventually to let data tell the truth they veil. Unfortunately, they do not easily do so. When we set up an appropriate condition and environment for them in the same way that seeds of a crop germinate in an appropriately-moist, fertile and plowed soil with warm weather, they try to tell it. As the seeds never germinate healthily in a dry and hard soil and/or with cold weather, the data do not easily unveil the truth only through statistical tests. Data analysis is part of statistics. Statistics is an applied science but not one which surpasses all other sciences. It is independent of them and as important as well. The knowledge established in the science(s) related to research in question, including

*Institude of Policy and Planning Sciences, University of Tsukuba, Tsukuba, Japan 305-8573.
12:45, December 6 (Thursday), 2001, file: bcpaper.tex.

natural logic and correct common sense, is definitely needed to solve an applied statistical problem in addition to statistical knowledge. Let us call it **professional knowledge**. A statistical model constructed only with statistics is not so effective as expected to solve an actual problem, social or natural, in a society and frequently aggravates the situation, because it is not necessarily reasonable from the viewpoint of the science related to the research, even though it is statistically best. No mistakes and errors are allowed for users, especially business managers and policy makers. Their social responsibilities are serious. In applications of statistical methods, resources such as brain labor, paper, i.e., wood resources, electricity and toner should not be wasted. It is necessary to concretely formulate statistical problems which convince not only statisticians but also users in various academic, governmental and industrial fields and create knowledge-based computational methods to actually solve them. Generally speaking, so long as statistical methods fail to yield not only scientifically reasonable but also statistically and data-analytically best results, statistics cannot become an expansive science which attracts many wise students and collects sufficient research funds. As a result, the statistical society stagnates and even shrinks.

The purpose of this paper is to (i) concretely formulate the j-th OLS-best subset problem as a variable selection problem for the Box-Cox transformation (Box and Cox, 1964), (ii) introduce a knowledge-based computational method to solve it, (iii) propose a solution to the (first) OLS-best subset problem for the Box-Cox transformation ($j = 1$) or one selected by a user among solutions to the first j OLS-best subset problems for the Box-Cox transformation ($j > 1$) solved in a run of a computer as a solution to a variable selection problem for the Box-Cox transformation, (iv) install the program in the Intellectual Statistical System **OEPP**[1] and make it available and (v) demonstrate an application to civil servants in prefectural governments in Japan.

The method to be proposed searches for a **practically-best regression equation** which is defined as **not only scientifically reasonable but also statistically and data-analytically best** on the condition of the appropriate scientific, statistical and data-analytical criteria specified by a user. It is an informatic and computational method that referees of a journal in his academic field can accept, policy-makers can adopt or business excutives can employ. We define the following terms: a **meaningful subset** as a subset which includes all necessary explanatory variables for a dependent variable but excludes any unnecessary, redundant and/or contradiction-causing explanatory variables from the viewpoint of professional knowledge or, in other words, a subset which consists of only explanatory variables representing a behavioral, institutional, technical or natural-law-based relation to a dependent variable and a **regression subequation**

[1] The **OEPP** is software which can handle the **Onishi variable selection methods** for (C)OLS, (C)GLS, (C)ADLR, (C)BCT, (C)2SLS, (C)2SPC, LIML, LIPC, etc., which Kitagawa, professor emeritus of Kyushu University, first named in his book (Kitagawa, 1987). The "C" of (C) implies constrained.

as the regression equation of a meaningful subset when the j-th OLS-best subset problem is solved.

2 Box-Cox Transformation

When M dependent variable candidates, say, $Y^1, Y^2, \cdots, Y^M$, which represent M respective data sets, are regressed on the i-th meaningful subset $\boldsymbol{X}_i$ of a set $\boldsymbol{X}$ of all possible explanatory variables, the regression subequation for Y^m can be expressed as follows[2] :

$$Y^m = \boldsymbol{X}_i A_i^m + U \quad \text{for some } i \text{ and } m = 1, 2, \cdots, M \tag{1}$$

where Y^m consists of the data y_t^m's for the outcome; t ≡sample point number; A_i^m ≡column vector of true regression coefficients of $\boldsymbol{X}_i$; and U ≡disturbance term.

The Box-Cox transformation is often used when a true functional form of a regression equation is not known. Let Y be a variable, a data vector or a data set of the original data y_t's of an outcome where $y_t > 0$ for all t. M dependent variable candidates in the Box-Cox transformation for $M \geq 2$ are defined as (i) $Y^1 = Y$ which implies $y_t^1 = y_t$ for all t; (ii) $Y^m = \{(Y)^{\lambda_m} - 1\}/\lambda_m$ for $\lambda_m = (M-m)/(M-1)$ which implies $y_t^m = \{(y_t)^{\lambda_m} - 1\}/\lambda_m$ for $m = 2, 3, \cdots, M-1$ if $M \geq 3$ and for all t; and (iii) $Y^M = \lim_{m \to M}[\{(Y)^{(M-m)/(M-1)} - 1\}/\{(M-m)/(M-1)\}] = \lim_{\lambda_M \to 0}[\{(Y)^{\lambda_M} - 1\}/\lambda_M] = \ln(Y)$ by the L'Hôpital's rule (actually proved by Johann Bernoulli) which implies $y_t^M = \lim_{m \to M}[\{(y_t)^{(M-m)/(M-1)} - 1\}/\{(M-m)/(M-1)\}] = \lim_{\lambda_M \to 0}[\{(y_t)^{\lambda_M} - 1\}/\lambda_M] = \ln(y_t)$ for all t. Y^1 is the original dependent variable Y as a candidate, whereas $Y^2, \cdots, Y^{M-1}$ and Y^M are transformed dependent variable candidates concerned with Y. Concretely rewriting (1) in the Box-Cox transformation, we have

$$\begin{cases} y_t^1 \equiv y_t = \sum_{k_i=0}^{K_i} a_{k_i i}^1 x_t^{k_i i} + u_t \quad \text{for } m = 1, \\ y_t^m \equiv \dfrac{(y_t)^{\lambda_m} - 1}{\lambda_m} = \sum_{k_i=0}^{K_i} a_{k_i i}^m x_t^{k_i i} + u_t \quad \text{with } \lambda_m = \dfrac{M-m}{M-1} \\ \qquad \text{for } m = 2, 3, \cdots, M-1, \\ y_t^M \equiv \ln(y_t) = \sum_{k_i=0}^{K_i} a_{k_i i}^M x_t^{k_i i} + u_t \quad \text{for } m = M \end{cases} \tag{2}$$

where $a_{k_i i}^m$ ≡true but unknown regression coefficient; $x_t^{k_i i}$ ≡datum of an explanatory variable with $k_i = 0$ for a constant term; and u_t ≡disturbance.

[2] At present, the System **OEPP** can deal with the j-th OLS-best subset problems for cases in which the Box-Cox transformation is applied only for explanatory variables $\boldsymbol{X}$, leading to $Y = \boldsymbol{X}_i^m A_i^m + U$ for $m = 1, 2, \cdots, M$, and for both a dependent variable Y and explanatory variables $\boldsymbol{X}$, leading to $Y^m = \boldsymbol{X}_i^m A_i^m + U$ for $m = 1, 2, \cdots, M$ in addition to the present problem.

3 Variable Selection for Box-Cox Transformation

3.1 Notation

For simplicity, we assume that no constraints are imposed on regression coefficients and no lagged dependent variables are included in all possible explanatory variables.[3] We introduce the following notation and rules for the j-th OLS-best subset problem for the Box-Cox transformation, although the notation for criteria are introduced in Subsection 3.2 and the others in the j-th OLS-best subset problem in Subsection 3.4:

$T \equiv$ number of all samples each variable has where $t = 1, 2, \cdots, T$ for $T \gg 1$;

$Y \equiv$original dependent variable or $Y \equiv (y_1, y_2, \cdots, y_T)' = (Y \times 1)$-vector of its data y_t's observed or measured for the outcome at the t-th sample point where $y_t > 0$ for all t;

$M \equiv$number of all Box-Cox transformations;

$\lambda_m \equiv (M - m)/(M - 1)$ for $m = 2, 3, \cdots, M - 1$ for $M \geq 3$ where $0 < \lambda_m < 1$;

$\boldsymbol{\mathcal{M}} \equiv \{1, 2, \cdots, m, \cdots, M\}$ =set of all Box-Cox transformation numbers;

$y_t^m \equiv t$-th datum of the m-th Box-Cox-transformed dependent variable Y^m for $m \in \boldsymbol{\mathcal{M}}$ at sample point t where (i) $y_t^1 \equiv y_t$ for all t; (ii) $y_t^m \equiv \{(y_t)^{\lambda_m} - 1\}/\lambda_m$ for $m = 2, 3, \cdots, M-1$ and for all t; and (iii) $y_t^M \equiv \lim_{\lambda_M \to 0}(y_t^{\lambda_M} - 1)/\lambda_M = \ln(y_t)$ for all t;

$Y^m \equiv$ original dependent variable (for $m = 1$) or m-th Box-Cox-transformed dependent variable (for $m = 2, 3, \cdots, M$) or $Y^m \equiv (y_1^m, y_2^m, \cdots, y_T^m)' = (T \times 1)$-vector of its data y_t^m's for $m \in \boldsymbol{\mathcal{M}}$ and for all t;

$\boldsymbol{Y} \equiv \{Y^1, Y^2, \cdots, Y^M\}$ =set of all possible original $(m = 1)$ or Box-Cox transformed $(m = 2, 3, \cdots, M)$ dependent variables;

$X_0 \equiv$constant term or $X_0 \equiv (1, 1, \cdots, 1)' = (T \times 1)$-vector of its data 1's;

$\bar{y}^m \equiv X_0' Y^m / T = \sum_{t=1}^T y_t^m / T \equiv$average of Y^m for $m \in \boldsymbol{\mathcal{M}}$;

$K \equiv$number of all possible nonconstant explanatory variables where $K \geq 1$ and $k = 1, 2, \cdots, K$;

$X_k \equiv k$-th possible explanatory variable or $X_k \equiv (x_1^k, x_2^k, \cdots, x_T^k) = (T \times 1)$-vector of its data;

$\boldsymbol{X} \equiv \{X_0, X_1, X_2, \cdots, X_K\} = (K + 1)$-set of a constant term and all possible explanatory variables or $\boldsymbol{X} \equiv (X_0, X_1, X_2, \cdots, X_K) = \{T \times (K + 1)\}$-matrix of their data;

$u_t \equiv$ disturbance at sample point t;

$U \equiv$ disturbance term or $U \equiv (u_1, u_2, \cdots, u_T)' = (T \times 1)$-vector of disturbances u_t's;

[3] The System OEPP can deal with these cases.

$\sigma^2, \sigma \equiv$unknown variance and standard deviation (or standard error) of U, respectively;

$i \equiv$number assigned to each of all possible nonempty subsets or submatrices of $\boldsymbol{X}$ where $i = 1, 2, 3, \cdots, 2^K - 1$;

$\boldsymbol{X}_i \equiv \{X_0, X_{1i}, X_{2i}, \cdots, X_{K_i i}\} = i$-th $(K_i + 1)$-subset of $\boldsymbol{X}$ or $\boldsymbol{X}_i \equiv (X_0, X_{1i}, X_{2i}, \cdots, X_{K_i i}) = \{T \times (K_i + 1)\}$-submatrix of $\boldsymbol{X}$ where it is ruled that $X_0 \in \boldsymbol{X}_i$ for all i if $X_0 \in \boldsymbol{X}$;

$k_i \equiv$number assigned to the $(k_i + 1)$-st variable in $\boldsymbol{X}_i$ where $k_i = 0, 1, 2, \cdots, K_i$;

$X_{k_i i} \equiv (x_1^{k_i i}, x_2^{k_i i}, \cdots\cdots, x_T^{k_i i})' = (T \times 1)$-vector of its data $x_t^{k_i i}$'s;

$a_{k_i i}^m \equiv$true but unknown regression coefficient of the $(k_i + 1)$-st variable $X_{k_i i}$;

$A_i^m \equiv (a_{0i}^m, a_{1i}^m, a_{2i}^m, \cdots, a_{K_i i}^m)' = \{(K_i + 1) \times 1\}$-vector of true but unknown regression coefficients of $\boldsymbol{X}_i$ on which Y^m is regressed;

$\hat{m} \equiv$optimal Box-Cox transformation number where $\hat{m} \in \mathcal{M}$;

$\hat{\lambda}_{\hat{m}} \equiv$optimal value for λ_m for $m = 2, \cdots, M - 1$ where $\hat{\lambda}_{\hat{m}} = (M - \hat{m})/(M - 1)$;

$\hat{a}_{k_i i}^m \equiv$estimate of $a_{k_i i}^m$ based on Y^m and $\boldsymbol{X}_i$;

$\hat{A}_i^m \equiv (\hat{a}_{0i}^m, \hat{a}_{1i}^m, \hat{a}_{2i}^m, \cdots, \hat{a}_{K_i i}^m)' \equiv$estimate of A_i^m;

$\hat{\mathcal{C}}(\hat{A}_i^m) \equiv$estimated covariance matrix of $\hat{A}_i^m$;

$\hat{\mathcal{V}}(\hat{a}_{k_i i}^m) \equiv$estimated variance of $\hat{a}_{k_i i}^m$ which is the $(k_i + 1, k_i + 1)$-element of $\hat{\mathcal{C}}(\hat{A}_i^m)$ for all k_i;

$\hat{s}_{k_i i}^m \equiv$ estimated standard deviation of $\hat{a}_{k_i i}^m$ based on Y^m and $\boldsymbol{X}_i$;

$\hat{t}_{k_i i}^m = t$-ratio of $\hat{a}_{k_i i}^m$ based on Y^m and $\boldsymbol{X}_i$;

$\hat{y}_{it}^m \equiv$ (partial-test) estimate of y_t^m based on Y^m and $\boldsymbol{X}_i$;

$\hat{Y}_i^m \equiv (\hat{y}_{i1}^m, \hat{y}_{i2}^m, \cdots, \hat{y}_{iT}^m)' =$estimate of Y^m;

$\hat{\hat{y}}_{it}^m \equiv$(inversely-transformed) estimate of original datum y_t calculated by $\hat{\hat{y}}_{it}^1 \equiv \hat{y}_{it}^1$ for $m = 1$; $\hat{\hat{y}}_{it}^m \equiv (\lambda_m \hat{y}_{it}^m + 1)^{1/\lambda_m}$ if $\lambda_m \hat{y}_{it}^m > -1$ or $\hat{\hat{y}}_{it}^m = 0$ if $\lambda_m \hat{y}_{it}^m \leq -1$ for $m = 2, \cdots, M - 1$; and $\hat{\hat{y}}_{it}^M \equiv \exp(\hat{y}_{it}^M)$ for $m = M$ based on Y^m and $\boldsymbol{X}_i$;

$\hat{\hat{Y}}_i^m \equiv (\hat{\hat{y}}_{i1}^m, \hat{\hat{y}}_{i2}^m, \cdots, \hat{\hat{y}}_{iT}^m)' =$ estimate of Y based on Y^m and $\boldsymbol{X}_i$;

$\hat{e}_{it}^m \equiv y_t^m - \hat{y}_{it}^m =$ residual from y_t^m based on Y^m and $\boldsymbol{X}_i$;

$\hat{E}_i^m \equiv (\hat{e}_{i1}^m, \hat{e}_{i2}^m, \cdots, \hat{e}_{iT}^m)' = Y^m - \hat{Y}_i^m = (T \times 1)$-vector of residuals from Y^m;

$\hat{\hat{e}}_{it}^m \equiv y_t - \hat{\hat{y}}_{it}^m =$ residual from original datum y_t;

$\hat{\hat{E}}_i^m \equiv (\hat{\hat{e}}_{i1}^m, \hat{\hat{e}}_{i2}^m, \cdots, \hat{\hat{e}}_{iT}^m)' = Y - \hat{\hat{Y}}_i^m = (T \times 1)$-vector of residuals from Y;

$(\hat{R}_i^m)^2 \equiv$ (unadjusted) coefficient of determination of $\hat{Y}_i^m$ based on Y^m and $\boldsymbol{X}_i$;

$(\widehat{\mathcal{R}}_i^m)^2 \equiv$ (Theil) adjusted coefficient of determination of $\hat{Y}_i^m$ based on Y^m and $\boldsymbol{X}_i$;

$\mathbf{0}_n \equiv (n \times 1)$-zero vector;

$\mathbf{I}_n \equiv (n \times n)$-identity matrix;

$T_i \equiv T - K_i - 1 =$ number of degrees of freedom.

3.2 Criteria for Scientific Conditions, Statistical and Data-analytic Tests

A user has to specify the following criteria, if needed, for scientific conditions, statistical tests and data-analytic tests, depending on the type of data, the information on observations (for instance, structural changes caused by the oil crises) and the purpose of his research[4] :

α_h^1 = a priori known lower bound of the h-th magnitude condition based on the knowledge of the science related to the research at hand;

α_h^2 = a priori known upper bound of the h-th magnitude condition based on the knowledge of the science related to the research at hand;

β = significance level (100β %) of a one- or two-tailed t-test for regression coefficients ($0 < \beta \ll 1$);

γ = significance level (100γ %) of the Durbin-Watson serial correlation test ($0 < \gamma \ll 1$);

ε = tolerance level for standardized residual test;

η = significance level (100η %) of a χ^2-distribution for the Jarque-Bera normality test ($0 < \eta < 1$);

ν = significance level (100ν %) of a two-tailed t-test for a residual outlier ($0 < \nu \ll 1$);

ψ = significance level (100ψ %) of an F distribution for the Chow equal coefficients test ($0 < \psi < 1$);

ω = significance level (100ω %) of an F distribution for the Goldfeld-Quandt homoscedasticity test ($0 < \omega < 1$);

ζ = value (100ζ %) to define a turning point at t of y_t^m ($0 < \zeta \ll 1$);

θ = minimum tolerance level of an adjusted or unadjusted coefficient of determination where $0 \ll \theta < 1$;

$\mathcal{Q} = \{\alpha_h^1$'s, α_h^2's, β, γ, ε, ζ, η, θ, ν, ψ, $\omega\}$ = criterion set.

3.3 Assumptions

(1) A user must have not necessarily perfect but sufficient professional knowledge about the science related to his (or her) research at hand. He must be able to introduce the set $\boldsymbol{X}$ of all possible explanatory variables for the dependent variable Y through the professional knowledge and then classify $\boldsymbol{X}$ to generate only the subsets meaningful for his research from $\boldsymbol{X}$ (see Section 4).

[4] The percentiles of normal, χ^2, t-, F tests and the lower and upper limits of the Durbin-Watson serial correlation tests of appropriate degrees of freedom at specified significance levels are automatically calculated and compared with the corresponding test statistics in the **OEPP**.

Furthermore, he must have the professional knowledge about the signs and/or magnitude ranges of regression coefficients and the nature (for example, economy of scale in production) of the system focussed on, if any.

(2) The functional forms of all regression subequations are linear with respect to the regression coefficients of explanatory variables in the meaningful subsets.

(3) The sample size must exceed the number of the constant term and the possible explanatory variables in the smallest meaningful subset of $\boldsymbol{X}$. It is desirable that the sample size exceeds sufficiently the number of the constant term and the possible explanatory variables in the largest meaningful subset of $\boldsymbol{X}$.

(4) The disturbance term U is normally distributed as $U \sim \mathcal{N}(\mathbf{0}_T, \sigma^2 \mathrm{I}_T)$, regardless of m.

(5) $\mathrm{abs}(|\boldsymbol{X}_i' \boldsymbol{X}_i|) > \epsilon$ for at least one meaningful subset $\boldsymbol{X}_i$ with respect to a preset or user-specified inverse-matrix-existence criterion value ϵ.

(6) $\boldsymbol{X}$ is nonstochastic or independent of U if $\boldsymbol{X}$ is stochastic.

(7) The principle of minimizing the sum of squared errors with respect to regression coefficients is suitable for the research.

3.4 The J-th OLS-Best Subset Problem for Box-Cox Transformation

We formulate the j-th OLS-best subset problem for M Box-Cox transformations and define a solution to it as the j-th practically-best regression subequation, where integers j ($j \geq 1$) and M ($M \geq 2$) are specified by a user. If he knows all appropriate scientific, statistical and data-analytic criteria and has rich experience in model building, he can specify $j = 1$. Then, he should regard the (first) practically-best regression subequation of a solution to the (first) OLS-best subset problem as the practically-best regression equation. He should specify, for example, 3, 4 or 5 for j, otherwise. Then, he must select by himself the most satisfactory from among the first at most j practically-best regression subequations as the practically-best regression equation by using his own new criterion or comparing them with each other. It should be noted that the integer j appears in the last condition [**XI**].

The J-th OLS-Best Subset Problem for Box-Cox Transformation[5]

[5] For simplicity, it is assumed that no lagged dependent variable is explanatory and no constraint is imposed on regression coefficients, although the System OEPP can handle them. Accordingly, the Schur stability condition (for $m = 1$ or $m = M$), the Durbin h-test, generalized turning point test, final test, constrained estimation and adjustment of degrees of freedom are not referred to.

Obtain in a run of a computer the practically j-th OLS-best regression subequation $\widehat{Y}_i^{\hat{m}} = \boldsymbol{X}_i \widehat{A}_i^{\hat{m}}$ for $\hat{m} \in \mathcal{M}$ by (1) searching for a best dependent variable $Y^{\hat{m}}$ from the set $\boldsymbol{Y}$ of all possible dependent variables and a subset $\boldsymbol{X}_i$ from the set $\boldsymbol{X}$ of all possible explanatory variables specified for $Y^{\hat{m}}$, (2) estimating the true regression coefficient vector $A_i^{\hat{m}}$ of $\boldsymbol{X}_i$ and the true variance σ^2 and standard deviation σ of the disturbance term U and (3) calculating the standard deviations $\widehat{s}_{k_i i}^{\hat{m}}$'s and t-ratios $\widehat{t}_{k_i i}^{\hat{m}}$'s of $\widehat{A}_i^{\hat{m}}$, other important test statistics and the inversely-transformed estimate $\widehat{Y}_i^{\hat{m}}$ for Y if $\hat{m} = 2, 3, \cdots, M$ under the criterion set Q and M Box-Cox transformations such that

[I] $\boldsymbol{X}_i$ is meaningful for $Y^{\hat{m}}$ from the viewpoint of the science related to the research at hand[6] ;

[II] $\widehat{\lambda}_{\hat{m}}$ if $2 \leq \hat{m} \leq M-1$ and $\widehat{A}_i^{\hat{m}}$, $\widehat{C}(\widehat{A}_i^{\hat{m}})$, $\widehat{Y}_i^{\hat{m}}$, $\widehat{E}_i^{\hat{m}}$, $(\widehat{\sigma}_i^{\hat{m}})^2$, $\widehat{\sigma}_i^{\hat{m}}$, $\widehat{s}_{k_i i}^{\hat{m}}$ and $\widehat{t}_{k_i i}^{\hat{m}}$ for $\hat{m} \in \mathcal{M}$ are calculated as follows:

$$\text{(i)} \quad Y^{\hat{m}} = Y \text{ if } \hat{m} = 1,$$

$$\text{(ii)} \quad Y^{\hat{m}} = \{(Y)^{\widehat{\lambda}_{\hat{m}}} - 1\}/\lambda_{\hat{m}} \text{ with } \widehat{\lambda}_{\hat{m}} = \frac{M - \hat{m}}{M-1} \text{ if } m = 2, 3, \cdots, M-1,$$

$$\text{(iii)} \quad Y^{\hat{m}} = \ln Y \text{ if } \hat{m} = M,$$

$$\widehat{A}_i^{\hat{m}} = (\boldsymbol{X}_i' \boldsymbol{X}_i)^{-1} \boldsymbol{X}_i' Y^{\hat{m}}, \quad \widehat{C}(\widehat{A}_i^{\hat{m}}) = (\widehat{\sigma}_i^{\hat{m}})^2 (\boldsymbol{X}_i' \boldsymbol{X}_i)^{-1},$$

$$\widehat{Y}_i^{\hat{m}} = \boldsymbol{X}_i \widehat{A}_i^{\hat{m}}, \quad \widehat{E}_i^{\hat{m}} = Y^{\hat{m}} - \widehat{Y}_i^{\hat{m}}, \quad (\widehat{\sigma}_i^{\hat{m}})^2 = \frac{\widehat{E}_i^{\hat{m}\prime} \widehat{E}_i^{\hat{m}}}{T_i},$$

$$\widehat{\sigma}_i^{\hat{m}} = \sqrt{(\widehat{\sigma}_i^{\hat{m}})^2}, \quad \widehat{s}_{k_i i}^{\hat{m}} = \sqrt{\widehat{\mathcal{V}}(\widehat{a}_{k_i i}^{\hat{m}})} \quad \text{and} \quad \widehat{t}_{k_i i}^{\hat{m}} = \frac{\widehat{a}_{k_i i}^{\hat{m}}}{\widehat{s}_{k_i i}^{\hat{m}}}; \tag{3}$$

[III] (*i*) $\widehat{A}_i^{\hat{m}}$ must satisfy the following sign and/or magnitude conditions, if required from the viewpoints of the professional knowledge of the science related to the research:

$$\alpha_{h_1}^1 \leq f_{h_1}^1(\widehat{A}_i^{\hat{m}}) \quad \text{if } \hat{m} \in \mathcal{H}_{h_1}^1 \text{ for } h_1 = 1, 2, \cdots, H^1, \tag{4}$$

$$f_{h_2}^2(\widehat{A}_i^{\hat{m}}) \leq \alpha_{h_2}^2 \quad \text{if } \hat{m} \in \mathcal{H}_{h_2}^2 \text{ for } h_2 = 1, 2, \cdots, H^2, \tag{5}$$

and/or

$$\alpha_{h_3}^1 \leq f_{h_3}^3(\widehat{A}_i^{\hat{m}}) \leq \alpha_{h_3}^2 \quad \text{if } \hat{m} \in \mathcal{H}_{h_3}^3 \text{ for } h_3 = 1, 2, \cdots, H^3; \tag{6}$$

where each of $f_{h_1}^1(\widehat{A}_i^{\hat{m}})$, $f_{h_2}^2(\widehat{A}_i^{\hat{m}})$ and $f_{h_3}^3(\widehat{A}_i^{\hat{m}})$ which denotes a function of $\widehat{A}_i^{\hat{m}}$ (and is linear with respect to $\widehat{A}_i^{\hat{m}}$ in most cases); $H^1, H^2, H^3 \equiv$numbers

[6] To understand how important the condition [I] for meaningful subsets of possible explanatory variables and the condition [III] for signs and/or magnitudes about regression coefficients are, see the explanation by some examples shown on pp. 380 of Onishi (1995a).

of the sign and/or magnitude conditions in an upper bound case, a lower bound case and a range case, respectively; $\mathcal{H}^1_{d_1}, \mathcal{H}^2_{d_2}, \mathcal{H}^3_{d_3}$ ≡sets of the Box-Cox transformation numbers in which sign and/or magnitude conditions in an upper bound case, a lower bound case and a range case are employed, respectively; and $\mathcal{H}^1_{d_1}, \mathcal{H}^2_{d_2}, \mathcal{H}^3_{d_3} \subset \mathcal{M}$;

[IV] the following inequality for the Jarque-Bera normality test must hold to maintain the null hypothesis H_0 : U is normally distributed with the expectation $\mathcal{E}(U) = \mathbf{0}_T$ at a 100η % significance level of a χ^2 test[7] :

$$\widehat{JB}^{\hat{m}}_i = T\left\{\frac{(\widehat{\mathcal{S}}^{\hat{m}}_i)^2}{6} + \frac{(\widehat{\mathcal{K}}^{\hat{m}}_i - 3)^2}{24}\right\} \leq \chi^2_2(\eta) \tag{7}$$

for

$$(\widehat{\mathcal{S}}^{\hat{m}}_i)^2 = \frac{\{\sum_{t=1}^T (\hat{e}^{\hat{m}}_{it})^3/T\}^2}{(\widehat{E}^{\hat{m}\prime}_i \widehat{E}^{\hat{m}}_i/T)^3} \text{ and } \widehat{\mathcal{K}}^{\hat{m}}_i = \frac{\sum_{t=1}^T (\hat{e}^{\hat{m}}_{it})^4/T}{(\widehat{E}^{\hat{m}\prime}_i \widehat{E}^{\hat{m}}_i/T)^2}$$

where $\chi^2_2(\eta) = \eta$ percentile of a χ^2 distribution with 2 degrees of freedom;

[V] the following inequality must be satisfied to adopt the specified alternative hypothesis H_1 or maintain the specified null hypothesis H_0 at a 100β % significance level of a t-test, depending on the purpose of the research, if necessary:

(i) to adopt H_1 for $\mathrm{H}_0 : \mathrm{G}'_{id_1} A^{\hat{m}}_i = \mathrm{g}_{d_1}$ against $\mathrm{H}_1 : \mathrm{G}'_{id_1} A^{\hat{m}}_i \neq \mathrm{g}_{d_1}$,

$$\frac{|\mathrm{G}'_{id_1}\widehat{A}^{\hat{m}}_i - \mathrm{g}_{d_1}|}{\widehat{S}^{\hat{m}}_{id_1}} > t_{T_i}(\beta/2) \quad \text{if } \hat{m} \in \mathcal{D}^1_{d_1} \text{ for } d_1 = 1, 2, \cdots, D^1, \tag{8}$$

(ii) to adopt H_1 for $\mathrm{H}_0 : \mathrm{G}'_{id_2} A^{\hat{m}}_i = \mathrm{g}_{d_2}$ against $\mathrm{H}_1 : \mathrm{G}'_{id_2} A^{\hat{m}}_i > \mathrm{g}_{d_2}$,

$$\frac{\mathrm{G}'_{id_2}\widehat{A}^{\hat{m}}_i - \mathrm{g}_{d_2}}{\widehat{S}^{\hat{m}}_{id_2}} > t_{T_i}(\beta) \quad \text{if } \hat{m} \in \mathcal{D}^2_{d_2} \text{ for } d_2 = 1, 2, \cdots, D^2, \tag{9}$$

(iii) to adopt H_1 for $\mathrm{H}_0 : \mathrm{G}'_{id_3} A^{\hat{m}}_i = \mathrm{g}_{d_3}$ against $\mathrm{H}_1 : \mathrm{G}'_{id_3} A^{\hat{m}}_i < \mathrm{g}_{d_3}$,

$$\frac{\mathrm{g}_{d_3} - \mathrm{G}'_{id_3}\widehat{A}^{\hat{m}}_i}{\widehat{S}^{\hat{m}}_{id_3}} > t_{T_i}(\beta) \quad \text{if } \hat{m} \in \mathcal{D}^3_{d_3} \text{ for } d_3 = 1, 2, \cdots, D^3, \tag{10}$$

or

(iv) to maintain H_0 for $\mathrm{H}_0 : \mathrm{G}'_{id_4} A^{\hat{m}}_i = \mathrm{g}_{d_4}$ against $\mathrm{H}_1 : \mathrm{G}'_{id_4} A^{\hat{m}}_i \neq \mathrm{g}_{d_4}$,

$$\frac{|\mathrm{G}'_{id_4}\widehat{A}^{\hat{m}}_i - \mathrm{g}_{d_4}|}{\widehat{S}^{\hat{m}}_{id_4}} \leq t_{T_i}(\beta/2) \quad \text{if } \hat{m} \in \mathcal{D}^4_{d_4} \text{ for } d_4 = 1, 2, \cdots, D^4, \tag{11}$$

[7] $\widehat{\mathcal{S}}^{\hat{m}}_i$ and $\widehat{\mathcal{K}}^{\hat{m}}_i - 3$ correspond to skewness and kurtosis, respectively.

for

$$(\widehat{S}_{id_\ell}^{\hat{m}})^2 = \mathrm{G}'_{id_\ell}\widehat{\mathcal{C}}(\widehat{A}_i^{\hat{m}})\mathrm{G}_{id_\ell} \quad \text{and} \quad \widehat{S}_{id_\ell}^{\hat{m}} = \sqrt{(\widehat{S}_{id_\ell}^{\hat{m}})^2} \quad \text{for all } \ell = 1,2,3,4,$$

where $\mathrm{G}_{id_\ell} \equiv \{(K_i+1)\times 1\}$-vector of known coefficients of the d_ℓ-th hypothesis concerned with A_i^m for $\ell = 1,2,3,4$; $\mathrm{g}_{d_\ell} \equiv$ a priori known value of the d_ℓ-th hypothesis; $t_{T_i}(\beta/2) \equiv$ percentile of a two-tailed t-test of a 100β % significance level with T_i degrees of freedom; $t_{T_i}(\beta) \equiv$ percentile of a one-tailed t-test of a 100β % significance level with T_i degrees of freedom; $D^1, D^2, D^3, D^4 \equiv$numbers of hypothesis testings in (i), (ii), (iii) and (iv), respectively; $\boldsymbol{\mathcal{D}}_{d_1}^1, \boldsymbol{\mathcal{D}}_{d_2}^2, \boldsymbol{\mathcal{D}}_{d_3}^3, \boldsymbol{\mathcal{D}}_{d_4}^4$ $\equiv$sets of the Box-Cox transformation numbers in which the hypothesis testings (i), (ii), (iii) and (iv) are made, respectively; and $\boldsymbol{\mathcal{D}}_{d_1}^1, \boldsymbol{\mathcal{D}}_{d_2}^2, \boldsymbol{\mathcal{D}}_{d_3}^3, \boldsymbol{\mathcal{D}}_{d_4}^4 \subset \boldsymbol{\mathcal{M}}$;

[**VI**] the Durbin-Watson serial correlation test statistic $\widehat{DW}_i^{\hat{m}}$ defined below must satisfy the following inequality at a 100γ % significance level, if time series data are used (Durbin and Watson, 1950 and 1951, Wallis 1972): for $T > 6$

$$\widehat{DW}_i^{\hat{m}} > d_i(\gamma) \text{ if } \widehat{DW}_i^{\hat{m}} \leq 2 \text{ or } 4 - \widehat{DW}_i^{\hat{m}} \geq d_i(\gamma) \text{ if } \widehat{DW}_i^{\hat{m}} > 2 \qquad (12)$$

for annual data ($r = 1$) or for quarterly data ($r = 4$)

$$\widehat{DW}_i^{\hat{m}} = \frac{\sum_{t=1+r}^{T}(\widehat{e}_{it}^{\hat{m}} - \widehat{e}_{i,t-r}^{\hat{m}})^2}{\sum_{t=1}^{T}(\widehat{e}_{it}^{\hat{m}})^2} \qquad (13)$$

where $d_i(\gamma) = d_{T,K_i+1}^{ur}(\gamma)$ if an inconclusive case is regarded as subjectively unacceptable or $d_i(\gamma) = d_{T,K_i+1}^{\ell r}(\gamma)$ if an inconclusive case is regarded as subjectively acceptable; $d_{T,K_i+1}^{ur}(\gamma)$ = upper limit of the Durbin-Watson serial correlation test of a 100γ % significance level with (T, K_i+1) degrees of freedom for annual ($r = 1$) or quarterly ($r = 4$) data; $d_{T,K_i+1}^{\ell r}(\gamma)$ = its lower limit;

[**VII**] the Chow equal coefficients test statistic $\widehat{C}_i^{\hat{m}}$ must satisfy the following inequality at a 100ψ % significance level of an F test, if (i) it is considered that structural changes may have happened and affected $\widehat{A}_i^{\hat{m}}$ and (ii) the Chow test need be applied to maintain the null hypothesis H_0: $\widehat{A}_i^{\hat{m}1} = \widehat{A}_i^{\hat{m}2}$ against the alternative hypothesis H_1: $\widehat{A}_i^{\hat{m}1} \neq \widehat{A}_i^{\hat{m}2}$ ($\widehat{A}_i^{\hat{m}\ell}$ for $\ell = 1,2$ is expressed in column vectors) (Chow, 1960)[8] :

$$\widehat{C}_i^{\hat{m}} = \frac{\dfrac{\widehat{E}_i^{\hat{m}\prime}\widehat{E}_i^{\hat{m}} - \widehat{E}_i^{+\hat{m}\prime}\widehat{E}_i^{+\hat{m}}}{\sigma^2(K_i+1)}}{\dfrac{\widehat{E}_i^{+\hat{m}\prime}\widehat{E}_i^{+\hat{m}}}{\sigma^2\{T-2(K_i+1)\}}} = \frac{\dfrac{\widehat{E}_i^{\hat{m}\prime}\widehat{E}_i^{\hat{m}} - \widehat{E}_i^{+\hat{m}\prime}\widehat{E}_i^{+\hat{m}}}{K_i+1}}{\dfrac{\widehat{E}_i^{+\hat{m}\prime}\widehat{E}_i^{+\hat{m}}}{T-2(K_i+1)}} \leq F_{T-2(K_i+1)}^{K_i+1}(\psi) \qquad (14)$$

[8] If some explanatory variable like a 0-1 dummy variable is included in $\boldsymbol{X}_i$, then $|\boldsymbol{X}_i^{1\prime}\boldsymbol{X}_i^1| = 0$ or $|\boldsymbol{X}_i^{2\prime}\boldsymbol{X}_i^2| = 0$ may occur, even if $|\boldsymbol{X}_i'\boldsymbol{X}_i| \neq 0$. In this case, the Chow test must not be applied. In the System **OEPP**, a user must specify such a variable to notify a computer of it. If a meaningful subset includes such a variable, the Chow test will be automatically suspended for it. The same treatment is needed for the Goldfeld-Quandt homoscedasticity test in the condition [**VIII**].

where $(Y^{\hat{m}}, X_i)$ is partitioned into $\{T^1 \times (K_i+2)\}$-submatrix $(Y^{\hat{m}1}, X_i^1)$ and $\{T^2 \times (K_i+2)\}$-submatrix $(Y^{\hat{m}2}, X_i^2)$ for $T = T^1 + T^2$, $T^1 > 0$ and $T^2 > 0$ so that $Y^{\hat{m}} \equiv (Y^{\hat{m}1\prime}, Y^{\hat{m}2\prime})'$ and $X_i \equiv (X_i^{1\prime}, X_i^{2\prime})'$; $\widehat{E}_i^{+\hat{m}} \equiv (\widehat{E}_i^{\hat{m}1\prime}, \widehat{E}_i^{\hat{m}2\prime})'$; $\widehat{E}_i^{\hat{m}\ell} \equiv (T^\ell \times 1)$-vector of residuals resulted from regressing $Y^{\hat{m}\ell}$ on X_i^ℓ for $\ell = 1, 2$; and $F^{K_i+1}_{T-2(K_i+1)}(\psi) \equiv \psi$ percentile of an F distribution with $\{K_i+1, T-2(K_i+1)\}$ degrees of freedom;

[VIII] the Goldfeld-Quandt homoscedasticity test statistic $\widehat{GQ}_i^{\hat{m}}$ concerned with equal variance σ^2 of the disturbance term U must satisfy the following inequality at a 100ω % significance level of an F test, if (i) it is considered that structural changes may have happened and affected σ^2 and (ii) the Goldfeld-Quandt test need be applied to maintain the null hypothesis H_0: $\mathcal{V}(u_t) = \sigma^2$ for all t (and covariance $\mathcal{C}(u_t u_s) = 0$ for all $s = 1, 2, \cdots, T$ but $t \neq s$) against the alternative hypothesis H_1: $\mathcal{V}(u_t) \neq \mathcal{V}(u_s)$ for at least one of pairs $\{t, s\}$ (Goldfeld and Quandt, 1965)

$$\widehat{GQ}_i^{\hat{m}} = \frac{\dfrac{\widehat{E}_{1i}^{*\hat{m}\prime}\widehat{E}_{1i}^{*\hat{m}}}{\sigma^2(Q-K_i-1)}}{\dfrac{\widehat{E}_{2i}^{*\hat{m}\prime}\widehat{E}_{2i}^{*\hat{m}}}{\sigma^2(Q-K_i-1)}} = \frac{\widehat{E}_{1i}^{*\hat{m}\prime}\widehat{E}_{1i}^{*\hat{m}}}{\widehat{E}_{2i}^{*\hat{m}\prime}\widehat{E}_{2i}^{*\hat{m}}} \leq F^{Q-K_i-1}_{Q-K_i-1}(\omega) \tag{15}$$

where $(Y^{\hat{m}}, X_i)$ for $\hat{m} \in \mathcal{M}$ is rearranged into $(Y^{*\hat{m}}, X_i^*)$ by the order of the unknown but user-properly-guessed magnitudes of variances $\mathcal{V}(u_t)$'s and then partitioned into three submatrices $(Y_1^{*\hat{m}}, X_{1i}^*)$, $(Y_c^{*\hat{m}}, X_{ci}^*)$ and $(Y_2^{*\hat{m}}, X_{2i}^*)$ in such a way that the number Q of samples $Y_1^{*\hat{m}}$ is equal to that of $Y_2^{*\hat{m}}$ for user-specified Q such that $Q > K_i + 1$, $T = 2Q + Q_c$ and $Q_c \geq 0$ so that $Y^* = (Y_1^{*\hat{m}\prime}, Y_c^{*\hat{m}\prime}, Y_2^{*\hat{m}\prime})'$ and $X_i^* = (X_{1i}^{*\prime}, X_{ci}^{*\prime}, X_{2i}^{*\prime})'$; $\widehat{E}_{1i}^{*\hat{m}} \equiv (Q \times 1)$-vector of residuals resulted from regressing $Y_1^{*\hat{m}}$ on the first Q samples X_{1i}^* of X_i^*; $\widehat{E}_{2i}^{*\hat{m}} \equiv (Q \times 1)$-vector of residuals resulted from regressing $Y_2^{*\hat{m}}$ on the last Q samples X_{2i}^* of X_i^*; $(Y_c^{*\hat{m}}, X_{ci}^*)$ =central submatrix to be omitted; and $F^{Q-K_i-1}_{Q-K_i-1}(\omega) \equiv \omega$ percentile of an F distribution of $(Q-K_i-1, Q-K_i-1)$ degrees of freedom;

[IX] (*i*) $\widehat{E}_i^{\hat{m}}$ must satisfy the following t-test for a residual outlier at a 100ν % significance level for a rather small sample size T with $T_i \geq 1$, if necessary (Sawa, 1979)[9] :

$$\max_t \widehat{OT}_{it}^{\hat{m}} \leq t_{T_i-1}(\nu/2T) \quad \text{for some } t = 1, 2, \cdots, T \tag{16}$$

[9] If the sample size T is large, the outlier t-test is not effective, because $\nu/2T \to 0$ so that $t_{T_i-1}(\nu/2T) \to +\infty$. In regression analysis, an outlier test to identify an outlier in Y^m before estimation is unimportant or less important than an outlier test through residuals. If an outlier-like y_t^m is well tracked with outlier-like $x_t^{k_i i}$'s, it is not regarded as an outlier.

for

$$\widehat{OT}_{it}^{\hat{m}} = \frac{\frac{|\widehat{e}_{it}^{\hat{m}}|}{\widehat{s}_i^{\hat{m}}\sqrt{1-\widehat{q}_{it}^{\hat{m}}}}}{\frac{\sqrt{\widehat{E}_i^{\hat{m}\prime}\widehat{E}_i^{\hat{m}} - (\widehat{e}_{it}^{\hat{m}})^2/(1-\widehat{q}_{it}^{\hat{m}})}}{\widehat{s}_i^{\hat{m}}\sqrt{T_i-1}}} = \frac{\frac{|\widehat{e}_{it}^{\hat{m}}|}{\sqrt{1-\widehat{q}_{it}^{\hat{m}}}}}{\frac{\sqrt{\widehat{E}_i^{\hat{m}\prime}\widehat{E}_i^{\hat{m}} - (\widehat{e}_{it}^{\hat{m}})^2/(1-\widehat{q}_{it}^{\hat{m}})}}{\sqrt{T_i-1}}} \tag{17}$$

and/or

(*ii*) all standardized residuals $\widehat{\tilde{e}}_{it}^{\hat{m}}$'s defined below must not exceed the user-specified criterion value ε, if necessary[10] :

$$\max_t |\widehat{\tilde{e}}_{it}^{\hat{m}}| \leq \varepsilon \quad \text{for some } t = 1, 2, \cdots, T \tag{18}$$

for

$$\widehat{\tilde{e}}_{it}^{\hat{m}} \equiv \frac{\widehat{e}_{it}^{\hat{m}}}{\widehat{s}_i^{\hat{m}}\sqrt{1-\widehat{q}_{it}^{\hat{m}}}} \tag{19}$$

where $t_{T_i-1}(\nu/2T) \equiv \nu/2T$ percentile of a two-tailed t-test with T_i-1, i.e., $T-K_i-2$ degrees of freedom; and $\widehat{q}_{it}^{\hat{m}} \equiv (t,t)$-diagonal element of $\boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'$;

[**X**] $\widehat{Y}_i^{\hat{m}}$ must satisfy the following turning point test defined by the user-specified ζ, if (i) time series or longitudinal data are used ($T \geq 3$) and (ii) it is necessary: if

$$(y_t^{\hat{m}} - y_{t-1}^{\hat{m}})(y_{t+1}^{\hat{m}} - y_t^{\hat{m}}) < 0 \tag{20}$$

and

$$\min\left\{\left|1 - \frac{y_{t-1}^{\hat{m}}}{y_t^{\hat{m}}}\right|, \left|1 - \frac{y_{t+1}^{\hat{m}}}{y_t^{\hat{m}}}\right|\right\} \geq \zeta \quad \text{for } t = 2, 3, \cdots, T-1 \text{ and } m \in \mathcal{M}, \tag{21}$$

then

$$(y_t^{\hat{m}} - y_{t-1}^{\hat{m}})(\widehat{y}_{it}^{\hat{m}} - \widehat{y}_{i,t-1}^{\hat{m}}) > 0 \tag{22}$$

and

$$(y_{t+1}^{\hat{m}} - y_t^{\hat{m}})(\widehat{y}_{i,t+1}^{\hat{m}} - \widehat{y}_{it}^{\hat{m}}) > 0; \tag{23}$$

and

[**XI**] the specified adjusted coefficient of determination $(\widehat{\mathcal{R}}_i^{\hat{m}})^2$ of $\widehat{Y}_i^{\hat{m}}$ defined below is greater than or equal to θ and the $\underline{j\text{-th largest}}$ among the specified

[10] The standardized residual test may be applied with some flexibility, when cross-sectional data are used. $\Pr\{|\widehat{\tilde{e}}_{it}^{\hat{m}}| \leq 1\} \doteq 0.6827$; $\Pr\{|\widehat{\tilde{e}}_{it}^{\hat{m}}| \leq 1.6449\} \doteq 0.9000$; $\Pr\{|\widehat{\tilde{e}}_{it}^{\hat{m}}| \leq 1.9600\} \doteq 0.9500$; $\Pr\{|\widehat{\tilde{e}}_{it}^{\hat{m}}| \leq 2\} \doteq 0.9545$; and $\Pr\{|\widehat{\tilde{e}}_{it}^{\hat{m}}| \leq 3\} \doteq 0.9983$.

adjusted coefficients of determination of the subsets which satisfy (i) all those applied from among the conditions [I] to [X] and (ii) (24) in [XI][11] :

$$(\widehat{\mathcal{R}}_i^{\hat{m}})^2 = \max\left[1 - \frac{\{1-(\widehat{R}_i^{\hat{m}})^2\}(T-1)}{T_i}, 0\right] \geq \theta \tag{24}$$

for

$$(\widehat{R}_i^{\hat{m}})^2 = 1 - \frac{\widehat{E}_i^{\hat{m}\prime}\widehat{E}_i^{\hat{m}}}{(Y^{\hat{m}} - \bar{y}^{\hat{m}}X_0)'(Y^{\hat{m}} - \bar{y}^{\hat{m}}X_0)}.\square \tag{25}$$

3.5 Remarks on a Solution, Diagnosis and Prediction

When a user obtains an unsatisfactory regression equation as the best, he may be disappointed very much. Sooner or later, a distrust of statistics may be instilled in users' minds. Therefore, thorough scrutiny is needed to search for the practically-best regression equation. The first condition [I] in the j-th OLS-best subset problem for the Box-Cox transformation is not related to statistical computing but is a knowledge-based and decisively important condition for the research. The second condition [II] calculates the Box-Cox transformation parameter, the estimates of a dependent variable $Y^{\hat{m}}$, the variance and standard deviation of a disturbance term U, regression coefficients with their variances, standard deviations and t-ratios, etc. The third condition [III] deals with scientific conditions about regression coefficients which statistics can hardly handle at present but the professional knowledge of an applied field requires. The conditions [IV] to [IX] are concerned with statistical hypothesis testing. The tenth condition [X] is a data-analytic test. The last eleventh condition [XI] is concerned with fitting. The conditions [I] and [III] to [X] are pass or failure checks, whereas the condition [XI] is a continuous measure. All regression subequations which passed all conditions employed from among the conditions [I] and [III] to [X] are finally ranked by the condition [XI]. It must be noted that the practically j-th best regression subequation depends on $\mathcal{Q}$ and M specified by a user. If severe criteria are set at $\mathcal{Q}$, no solution may exist.

It is easy to make a computer print until which condition a meaningful subset has passed in the conditions [III] to [XI]. If all meaningful subsets failed to pass the sign and/or magnitude condition [III], there is no statistical and data-analytic diagnosis when the professional knowledge used is correct. However, if at most one meaningful subset passed the condition [III] but all meaningful subsets failed to pass one of statistical and data-analytic conditions [IV] to [XI], it is possible to

[11] If the Akaike information criterion (Akaike, 1973) defined as

$$\widehat{AIC}_i^{\hat{m}} = T\{\ln 2\pi + 1 + \ln(\widehat{E}_i^{\hat{m}\prime}\widehat{E}_i^{\hat{m}}/T_i)\} + 2(K_i+2)$$

is employed instead of $(\widehat{\mathcal{R}}_i^{\hat{m}})^2$ in the System **OEPP**, then (24) must be replaced with $\widehat{AIC}_i^{\hat{m}} \leq \theta$, the phrase "the j-th largest" must be replaced with "the j-th smallest" and an arbitrarily large positive number should be given to θ. $T(\ln 2\pi + 1) + 4$, which is constant, is not essential. The 2 of $(K_i + 2)$ implies that a constant term $a_{0i}^{\hat{m}}$ and variance σ^2 of a disturbance term are counted as unknowns. $-\infty < \widehat{AIC}_i^{\hat{m}} < +\infty$. The smaller $\widehat{AIC}_i^{\hat{m}}$ is, the better will $\widehat{Y}_i^{\hat{m}}$ fit to $Y^{\hat{m}}$.

provide the user such a diagnosis that a weaker criterion for that statistical or data-analytic test may yield a solution. Needless to say, a prediction must be made by each of the first at most j practically-best regression subequations, if requested. The **total significance level** ϕ for the practically best regression equation may be defined as $\phi = 1 - (1-\beta)(1-\eta)(1-\gamma)(1-\psi)(1-\omega)(1-\nu)$. If a statistical test is not applied, a zero value should be substituted into the corresponding significance level. Roughly speaking, the practically best regression equation is adopted, if it exists, with a 100ϕ % risk of committing a type I error or with a 100ϕ % risk that it was actually not best.

4 Processing of Professional Knowledge

4.1 A Priori Known Signs of Regression Coefficients

Here a computational method adopted in the System **OEPP** is explained to solve the j-th OLS-best subset problem for the Box-Cox transformation, although a computer programmer can develop his own computational techniques. The information about the signs of some regression coefficients is often available from the professional knowledge related to research at hand. A user may want to test hypotheses about the signs of some regression coefficients. In this situation, it is convenient to attach the a priori known or to-be-hypothetically-tested signs, + or −, to the fronts of such explanatory variables in loading a dependent variable together with all possible explanatory variables into a computer. Let $\Diamond$ denote +, − or nothing and braces { and } stand for a set or subset of variables. Therefore, $\Diamond X$ implies $+X$, $-X$ or X. $\{+X_1, -X_2, X_3\}$ implies a (sub)set of variables X_1, X_2 and X_3 the a priori known signs of whose regression coefficients are positive, negative and unavailable, respectively.

4.2 Basic Variable Classifications

Double variable classifications of (i) single or grouped variables and (ii) combinatorial or sequential variables must be applied for all possible nonconstant explanatory variables in order to solve the condition [I] in the j-th OLS-best subset problem for the Box-Cox transformation. Although we omit the proof, double variable classifications are necessary and sufficient for a computer to generate only all meaningful subsets in a run of a computer, no matter what research is conducted.

4.2.1 Single or Grouped Variables

A **single variable** is defined as one which has its meanings or role by itself in interpreting the regression equation. Most explanatory variables are usually treated as single variables. On the other hand, a **grouped variable** is defined as one which cannot have its clear meanings or role by itself but can have it only

when it is used together with the other appropriate variables. In most research, grouped variables represent (i) complementary relations such as a pair of patients' expenditures of medical doctors' services in hospitals and the prescribed medicines bought in pharmacies or a pair of computer hardware and software or (ii) a choice of an aggregate variable X or a cluster (or set) of all K componentwise variables $X_1, X_2, \cdots, X_K$ for $K = 2, 3, \cdots$, where $X = \sum_{k=1}^{K} \bar{\alpha}_k X_k$ where $\bar{\alpha}_k$ =known weight for X_k. It is understood that a cluster of grouped variables must be treated just like a single variable in variable selection.

In order to distinguish a single variable with/without the a priori known sign of the regression coefficient from a cluster of grouped variables with/without the a priori known signs of their regression coefficients, we postulate that the latter is enclosed within parentheses (and) like $(\Diamond X_1, \Diamond X_2, \cdots, \Diamond X_K)$ for some $K = 2, 3, \cdots$ and, furthermore, in variable selection, it is treated just like a single variable and the parentheses are ignored in the subsets which include the cluster of grouped variables.

Let $\mathcal{X}$ represent a single variable with/without the sign, $\Diamond X$, or a cluster of grouped variables with/without their signs, $(\Diamond X_1, \Diamond X_2, \cdots, \Diamond X_K)$. $\mathcal{X}$ is called a **condensed variable**. For instance, if $\mathcal{X} = +X$, then $\mathcal{X}$ means a single variable X whose regression coefficient must be positive, implying that it is a priori known or must be hypothetically tested that an increase (or a decrease) in X increases (or decreases) the dependent variable, ceteris paribus. Furthermore, if a user wants to apply a t-test, $+X$ requires a one-tailed t-test but not a two-tailed t-test, implying that an F test is inappropriate. If $\mathcal{X} = X$, then $\mathcal{X}$ means a single variable X whose regression coefficient can be positive or negative and requires a two-tailed t-test or an F test. If $\mathcal{X} = (+X_1, -X_2)$, then $\mathcal{X}$ means a cluster of grouped variables X_1 and X_2 whose regression coefficients must be positive and negative, respectively. The grouped variables X_1 and X_2 cannot be separately selected.

Example 1:

$$\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \mathcal{X}_4 = -X_1, (+X_2, -X_3), +X_4, X_5$$

implies that $\mathcal{X}_1 = -X_1$, $\mathcal{X}_2 = (+X_2, -X_3)$, $\mathcal{X}_3 = +X_4$ and $\mathcal{X}_4 = X_5$. There are 5 explanatory variables X_1, X_2, X_3, X_4 and X_5 with/without their signs, whereas there are 4 condensed variables $\mathcal{X}_1$, $\mathcal{X}_2$, $\mathcal{X}_3$ and $\mathcal{X}_4$.

4.2.2 Combinatorial or Sequential Variables

Only condensed variables are here focussed on. Let $\Omega_{0K} \equiv \{0, 1, 2, \cdots, K\}$ and $\Omega_{1K} \equiv \{1, 2, \cdots, K\}$.

- **Basic Combinatorial Variables**

Let $P \in \Omega_{0K}$, $Q \in \Omega_{0K}$, $P^* \equiv \min\{P, Q\}$, $Q^* \equiv \max\{P, Q\}$ and $K \equiv$number of condensed variables where $0 < P + Q$. We postulate that

$$< P < \mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_K > Q >, \tag{26}$$

(i) generates $\sum_{p=P^*}^{Q^*} \begin{pmatrix} K \\ p \end{pmatrix}$ meaningful subsets, $\emptyset$ if $P^* = 0$ and $\{\mathcal{X}_{p_1}, \mathcal{X}_{p_2}, \cdots, \mathcal{X}_{p_i}\}$ for $p_1, p_2, \cdots, p_i = 1, 2, \cdots, K$, $p_i \neq p_j$, $i \neq j$, $i, j = P^*, P^* + 1, \cdots, Q^*$, and, furthermore, (ii) investigates whether or not the estimated regression coefficients of the variables in all meaningful nonempty subsets derived from (26) meet the + or − signs, if indicated in $\mathcal{X}_k$'s for $k = 1, 2, \cdots, K$, when (26) is used for estimation. $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_K$ in (26) are called **a cluster of combinatorial condensed variables.**

A user introduces condensed variables $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_K$ and knows the appropriate integers for P and Q from the viewpoints of the professional knowledge. It is impossible for him to implement his research well, otherwise. It must be noted that the positions of $\mathcal{X}_k$'s within a pair of combinatorial variable classifiers $< P <$ and $> Q >$ do not matter. Needless to say, if $P = 0$ or $Q = 0$, then an empty subset becomes meaningful with respect to these condensed variables and is usually used together with other clusters of classified variables.

Example 2: case of $K = 4$, $P = 0$ and $Q = 2$,

$$< 0 < (+X_1, X_2), -X_3, X_4 > 2 >$$

which is regarded as $< 0 < \mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3 > 2 >$ for $\mathcal{X}_1 \equiv (+X_1, X_2)$, $\mathcal{X}_2 \equiv -X_3$ and $\mathcal{X}_3 \equiv X_4$, (i) generates the following 7 meaningful subsets with respect to $+X_1$, X_2, $-X_3$ and X_4: (i) $\emptyset$; (ii) $\{+X_1, X_2\}$; (iii) $\{-X_3\}$; (iv) $\{X_4\}$; (v) $\{+X_1, X_2, -X_3\}$; (vi) $\{+X_1, x_2, X_4\}$; and (vii) $\{-X_3, X_4\}$ and (ii) investigates the signs of the regression coefficients of these meaningful subsets, where $\begin{pmatrix} 3 \\ 0 \end{pmatrix} + \begin{pmatrix} 3 \\ 1 \end{pmatrix} + \begin{pmatrix} 3 \\ 2 \end{pmatrix} = 1 + 3 + 3 = 7$.

● Basic Sequential Variables

Let $P \in \Omega_{0K}$ or $P \in \Omega_{1K}$ and $K \equiv$number of condensed variables. We postulate that

$$< P \ < \mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_K >> \ \text{ or } \ << \mathcal{X}_K, \cdots, \mathcal{X}_2, \mathcal{X}_1 > \ P > \qquad (27)$$

(i-1) generates $(K - P + 1)$ meaningful subsets $\{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_p\}$ for $p = P, P + 1, \cdots, K$ f $P \in \Omega_{1K}$ or (i-2) generates $(K + 1)$ meaningful subsets $\emptyset$ and $\{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_p\}$ for $p = 1, 2, \cdots, K$ if $P = 0$ and, furthermore, (ii) investigates whether or not the estimated regression coefficients of the variables in all nonempty meaningful subsets derived from (27) meet the + or − signs, if included in $\mathcal{X}_k$'s for $k \in \Omega_{1K}$. $\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_K$ in (27) are called **a cluster of sequential condensed variables.** It must be kept in mind that the positions of sequential variables within a pair of sequential variable classifiers $< P <$ and $>>$ or $<<$ and $> P >$ are decisively important.

Example 3: case of $K = 5$ and $P = 2$

$$< 2 < +X_1, +X_2, -X_3, X_4, +X_5 >> \quad \text{or} \quad << +X_5, X_4, -X_3, +X_2, +X_1 > 2 >$$

(i) generates the following 4 meaningful subsets: (i) $\{+X_1, +X_2\}$; (ii) $\{+X_1, +X_2, -X_3\}$; (iii) $\{+X_1, +X_2, -X_3, X_4\}$; and (iv) $\{+X_1, +X_2, -X_3, X_4, +X_5\}$ and (ii) investigates the signs of the regression coefficients of these 4 meaningful subsets.

Example 4:

$$< 0 < +X_1, (-X_2, X_3), +X_4 >>, < 1 < X_5 > 1 >$$
or
$$< 1 < X_5 > 1 >, << +X_4, (-X_2, X_3), +X_1 > 0 >$$

(i) generates the following 4 meaningful subsets: (i) $\{X_5\}$; (ii) $\{+X_1, X_5\}$; (iii) $\{+X_1, -X_2, X_3, X_5\}$; and (iv) $\{+X_1, -X_2, X_3, +X_4, X_5\}$, which are obtained by the combinations of [1] the 4 partially meaningful subsets $\emptyset$, $\{\mathcal{X}_1\}$, $\{\mathcal{X}_1, \mathcal{X}_2\}$ and $\{\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3\}$ generated by $< 0 < \mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3 >>$ or $<< \mathcal{X}_3, \mathcal{X}_2, \mathcal{X}_1 > 0 >$ with respect to $+X_1$, $-X_2$, X_3 and $+X_4$ and [2] only one partially meaningful subset $\{X_5\}$ generated by $< 1 < X_5 > 1 >$ with respect to X_5, and, furthermore, (ii) investigates the signs of the estimated regression coefficients of the signed variables in these 4 meaningful subsets, where $\mathcal{X}_1 = +X_1$, $\mathcal{X}_2 = (-X_2, X_3)$ and $\mathcal{X}_3 = +X_4$. Needless to say, the following are equivalent to the above entries: $< 1 < X_5, +X_1, (-X_2, X_3), +X_4 >>$ and $<< +X_4, (X_3, -X_2), +X_1, X_5 > 1 >$.

4.3 Functional Form

All meaningful subsets generated through (26) and/or (27) are equivalent to each other before estimation and evaluation. In other words, it is not known before estimation and evaluation which meaningful subset becomes practically best. Let X_0 be a constant term. We postulate that X_0 is not enclosed within any symbols, a dependent variable is expressed as a function of a constant term, if needed, and a set of all possible explanatory variables which are classified and X_0 is automatically included in all derived meaningful subsets.

Let us give an example. Suppose that the Box-Cox transformation is utilized with the following functional form[12] : for M Box-Cox transformations

$$\begin{aligned} Y \;=\; & F(X_0, < 1 < +X_1 > 1 >, < 1 < -X_2(X_3, +X_4) > 1 >, \\ & < 0 < +X_5, +X_6, -X_7 >>) \end{aligned} \tag{28}$$

where Y =original dependent variable; X_0 =constant term; X_1, X_2 =single and combinatorial variables (X_1 =absolutely important, fixed or core variable in other

[12] A functional form need not be unique. If a computer can correctly identify signs, variables and classifiers, a comma or blank is not necessarily needed.

words); X_3, X_4 =grouped and combinatorial variables; and X_5, X_6, X_7 =single and sequential variables. The regression coefficients of X_1, X_4, X_5 and X_6 must be positive as (part of) scientific conditions, whereas those of X_2 and X_7 must be negative. It is clear that no a priori known signs are available for the regression coefficients of the constant term X_0 and the explanatory variable X_3. The functional form (i) generates the following $8\times M$ meaningful subsets: (i) $\{X_0, +X_1, -X_2\}$; (ii) $\{X_0, +X_1, X_3, +X_4\}$; (iii) $\{X_0, +X_1, -X_2, +X_5\}$; (iv) $\{X_0, +X_1, X_3, +X_4, +X_5\}$; (v) $\{X_0, +X_1, -X_2, +X_5, +X_6\}$; (vi) $\{X_0, +X_1, X_3, +X_4, +X_5, +X_6\}$; (vii) $\{X_0, +X_1, -X_2, +X_5, +X_6, -X_7\}$; and (viii) $\{X_0, +X_1, X_3, +X_4, +X_5, +X_6, -X_7\}$ for each of Y^m's for $m = 1, 2, \cdots, M$ and (ii) investigates whether or not the signs of the estimated regression coefficients $\widehat{a}^m_{k_i\ell}$'s of the following regression subequations except for those of the constant term and X_3 coincide with the specified signs:

$$Y^m = a^m_{01} + a^m_{11}X_1 + a^m_{21}X_2 + U, \tag{29}$$

$$Y^m = a^m_{02} + a^m_{12}X_1 + a^m_{22}X_3 + a^m_{32}X_4 + U, \tag{30}$$

$$Y^m = a^m_{03} + a^m_{13}X_1 + a^m_{23}X_2 + a^m_{33}X_5 + U, \tag{31}$$

$$Y^m = a^m_{04} + a^m_{14}X_1 + a^m_{24}X_3 + a^m_{34}X_4 + a^m_{44}X_5 + U, \tag{32}$$

$$Y^m = a^m_{05} + a^m_{15}X_1 + a^m_{25}X_2 + a^m_{35}X_5 + a^m_{45}X_6 + U, \tag{33}$$

$$Y^m = a^m_{06} + a^m_{16}X_1 + a^m_{26}X_3 + a^m_{36}X_4 + a^m_{46}X_5 + a^m_{56}X_6 + U, \tag{34}$$

$$Y^m = a^m_{07} + a^m_{17}X_1 + a^m_{27}X_2 + a^m_{37}X_5 + a^m_{47}X_6 + a^m_{57}X_7 + U, \tag{35}$$

and

$$Y^m = a^m_{08} + a^m_{18}X_1 + a^m_{28}X_3 + a^m_{38}X_4 + a^m_{48}X_5 + a^m_{58}X_6 + a^m_{68}X_7 + U, \tag{36}$$

where $Y^1 = Y$; $Y^m = \{(Y)^{\lambda_m} - 1\}/\lambda_m$ with $\lambda_m = (M - m)/(M - 1)$ for $m = 2, 3, \cdots, M - 1$; $Y^M = \ln(Y)$; and U =disturbance term.

For instance, by $+X_1$ introduced in the functional form, the estimated regression coefficients $\widehat{a}^m_{1\ell} > 0$ for all $\ell = 1, 2, 3, 4, 5, 6, 7, 8$ and all $m = 1, 2, \cdots, M$ are required as one of the preconditions for application of statistical and data-analytic tests. Only the regression subequations which passed all sign conditions are, furthermore, scrutinized by magnitude conditions for the regression coefficients and statistical and data-analytic tests, if required.

5 Computational Procedure

Let us illustrate the essential computational procedure for a computer. Suppose that L =number of all meaningful subsets $\boldsymbol{X}_\ell$'s derivable from $\boldsymbol{X}$, which can be easily calculated by variable classifications.

- Step 1: Identify a dependent variable Y, all possible explanatory variables $\boldsymbol{X}$ and variable classifications made on $\boldsymbol{X}$ in a functional form, the number M of all Box-Cox transformations[13] , scientific conditions, statistical and data-analytic criteria $\mathcal{Q}$ and the number j of the first j practically-best regression subequations.

- Step 2: Calculate $\boldsymbol{X}'\boldsymbol{X}$.

- Step 3: Initialize as $m = 1$ or increase m by 1 and calculate Y^m if $m > 1$ and $\boldsymbol{X}'Y^m$.

- Step 4: Initialize as $\ell = 1$ or increase ℓ by 1. Construct $\boldsymbol{X}_\ell'\boldsymbol{X}_\ell$ and $\boldsymbol{X}_\ell'Y^m$ from $\boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{X}'Y^m$, respectively, for a meaningful subset $\boldsymbol{X}_\ell$.

- Step 5: Calculate $\hat{A}_\ell^m = (\boldsymbol{X}_\ell'\boldsymbol{X}_\ell)^{-1}\boldsymbol{X}_\ell'Y^m$, $\hat{Y}_\ell^m = \boldsymbol{X}_\ell\hat{A}_\ell^m$, $\hat{E}_\ell^m = Y^m - \hat{Y}_\ell^m$, etc. and all test statistics.

- Step 6: Check all scientific conditions about $\hat{A}_\ell^m$, if employed. Go to Step 9 if one of the scientific conditions is unsatisfied.

- Step 7: Calculate all necessary percentitles of statistical tests employed and make all statistical and data-analytical tests employed. Go to Step 9 if one of the statistical and data-analytic tests fails.

- Step 8: Memorize $\hat{A}_\ell^m$, $\hat{Y}_\ell^m$, etc., if all first j practically-best regression subequations are not memorized yet or replace the j-th practically-best regression subequation with $\hat{A}_\ell^m$, $\hat{Y}_\ell^m$, etc., if the j-th practically-best regression subequation was already memorized. Upgrade the fitting criterion θ to the degree of fitting of the new j-th practically-best regression subequation.

- Step 9: Go back to Step 4, if $\ell < L$.

- Step 10: Go back to Step 3, if $\ell = L$ and $m < M$.

- Step 11: Print the first at most j practically-best regression subequations with important test statistics.

[13] In the System **OEPP**, only OLS is applied, if the first parameter OLS is inputted immediately after the command METHOD. Furthermore, the second and third parameters BC (or BCE or EBC) and the number M of the Box-Cox transformations are inputted for the command METHOD, then OLS for $m = 1$ and Box-Cox transformation for OLS for $m = 2, 3, \cdots, M$ are applied. BCE conducts estimation for the Box-Cox transformations only for explanatory variables, whereas EBC does so for the Box-Cox transformation extended to both a dependent variable and explanatory variables, where all data relevant to the Box-Cox transformation must be positive.

6 An Example

6.1 Variable Selection

Suppose that we have sufficient professional knowlede about the administrations of local governments and rich modelling experiences, plausible significance levels are subjectively selected for statistical tests and $j = 1$ is set. Let us demonstrate from the heuristic viewpoints how to use the proposed method in three steps together with two contingent steps by the System **OEPP**. OLS is directly used in the first and second steps and then Box-Cox transformation is applied in the third step as a result of the failures in the previous 2 steps. The purpose is to search for the scientifically reasonable and statistically best regression equation, i.e., the practically best regression equation, by trial and error, for the numbers of civil servants in the assembly and general affairs combined sectors in 46 prefectural governments, except for the Tokyo metropolitan government, in the research on an administrative reform for prefectural governments. Since Tokyo is the capital of Japan, the administrative functions of the Tokyo metropolitan government differ from those of the other 46 prefectural governments. Thus, the sample size is 46 and cross-sectional data in 1996 are used and listed in Appendix. We use the following variable notation:

Y =number of civil servants in assembly and general affairs combined sector (unit: persons); X_0 = constant term; X_1 =number of households (1,000 households); X_2 =number of citizens in all cities (1,000 persons); X_3 =number of residents in all towns and villages (1,000 persons); X_4 =habitable area (=administrated area minus mountainous and lacustrine areas) (km^2); X_5 =administrated areas (km^2); X_6 =number of citizens in ordinance-designated cities (shitei toshi in Japanese) whose populations are one million or more and to which various administrative rights of importance are transferred from prefectual governments (1,000 persons); X_7 =number of citizens in core or kernel cities (chuukaku toshi in Japanese) whose populations are less than one million but exceed 200 thousands and to which some administrative rights are transferred from prefectural governments, where no ordinance-designated cities are near (1,000 persons); X_8 =number of residents in high population-density areas (1,000 persons); X_9 =net population movements (100 persons); X_{10} =areas administrated by ordinance-designated cities (km^2); X_{11} =standard financial sizes (100 million yen); X_{12} =number of all towns and villages (towns or villages); X_{13} =dummy variable for the presence of the largest-in-the-Orient US military forces in Okinawa, which pushes the Okinawa government to hire more civil servants to deal with many problems on noise pollution by air planes, relocation of air force bases, criminal affairs, etc., whose regression coefficient is expected to be positive, if X_{13} is adopted.

We set the following statistical tests and criteria:

β = significance level 0.1 (10 %) of a one- or two-tailed t-test hypothesis testing for

the regression coefficients;
η = significance level 0.05 (5 %) of the Jarque-Bera normality test (χ^2 test);
ν = significance level 0.05 (5 %) of a two-tailed t-test for detecting a residual outlier;
ε = 2.5 standardized residual tolerance level with acceptance of up to 2 violations;
ψ =significance level 0.05 (5 %) of the Chow equal coefficients test (F test) for a group (Kantou or Eastern Japan) of cross-sectional unit numbers 1 to 23 and a group (Kansai or Western Japan) of cross-sectional unit numbers 24 to 46;
ω = significance level 0.05 (5 %) of the Goldfeld-Quandt homoscedasticity test (F test) for a group (snowy area) of cross-sectional unit numbers 1 to 15 and a group (typhoon-often-hit area) of cross-sectional unit numbers 32 to 46;
θ = 0.7 minimum tolerance level for an adjusted coefficient of determination which is used to determine the ordering of practically best regression subequation candidates.

The total significance level for the practically best regression equation defined as $1-(1-\beta)(1-\eta)(1-\nu)(1-\psi)(1-\omega)$ is 0.26694 or 26.694 %, when the dummy variable X_{13} is not selected, i.e., when the Chow equal coefficients test and the Goldfeld-Quandt homoscedasticity test are used. However, the total significance level for the practically best regression equation defined as $1-(1-\beta)(1-\eta)(1-\nu)$ is 0.145 or 14.5 %, when the dummy variable X_{13} is selected, i.e., when the Chow equal coefficients test and the Goldfeld-Quandt homoscedasticity test cannot be used due to the special data structure of X_{13}.

The following three steps are conducted by a notebook-type PC (VAIO, Sony).

Step 1 in which the professional knowledge both for variable classification of all possible explanatory variables and for scientific conditions on the signs (and magnitudes) of their regression coefficients is not used at all, although available here, and variances of a disturbance term are assumed to be constant (OLS without Box-Cox transformation is used for all possible regressions under the above statistical tests and criteria):

The number of all possible regressions of 13 possible explanatory variables is 8,191 due to $2^{13}-1=8191$. They are easily generated and estimated by the following functional form and evaluated with the above statistical criteria:

$$Y = F(X_0 < 1 < X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13} > 13 >). \tag{37}$$

Since the signs of the regression coefficients of all possible explanatory variables are not shown in the functional form (37), a two-tailed t-test is made for hypothesis testing for all regression coefficients. The statistically best regression subequation (38) was searched for among 8,191 possible regression subequations estimated with

OLS in less than 6 seconds of CPU time:

$$\begin{array}{lcccc}
\widehat{Y}_{1809} & = & 555.7625 & +0.3183152X_1 & +0.1475850X_3 \\
(S.DEV.) & & (64.63306) & (0.6345661 \times 10^{-1}) & (0.7611922 \times 10^{-1}) \\
(T\text{-}RATIO) & & (8.598734) & (5.016265) & (1.938866) \\
 & & +0.2875088 \times 10^{-1}X_5 & +0.2216401X_6 & +0.1271108X_7 \\
 & & (0.3647111 \times 10^{-2}) & (0.4292448 \times 10^{-1}) & (0.6428338 \times 10^{-1}) \\
 & & (7.883192) & (5.163491) & (1.977351) \\
 & & -0.3709955X_{10} & -1.858752X_{11} & +323.3830X_{13} \\
 & & (0.1064525) & (0.5589328) & (98.52871) \\
 & & (-3.485081) & (-3.325537) & (3.282120)
\end{array} \tag{38}$$

$$\widehat{R}^2 = 0.9608,\ \widehat{\mathcal{R}}^2 = 0.9523,\ \widehat{AIC} = 558.6327,\ \widehat{SD} = 96.1938,$$
$$\widehat{V} = 9253.25,\ \widehat{DF} = 37,\ \widehat{JB} = 0.4368,\ \widehat{OT} = 3.303,\ TSL = 0.145$$

where $\widehat{Y}_{1809}$ =estimate of Y by the 1,809-th meaningful subset; $(S.DEV.)$ = standard deviations (errors) of the estimated regression coefficients; $(T\text{-}RATIO) = t$-ratios of the estimated regression coefficients; $\widehat{R}^2$ =coefficient of determination; $\widehat{\mathcal{R}}^2$ = adjusted coefficient of determination; $\widehat{AIC}$ =Akaike information criterion; $\widehat{SD}, \widehat{V}$ =standard deviation and variance of a disturbance term, respectively; $\widehat{DF}$ = degrees of freedom; $\widehat{JB}$ = Jarque-Bera normality test statistic; $\widehat{OT}$ = t-test statistic for a residual outlier; and TSL =total significance level.

Eight explanatory variables are selected. Their t-ratios in absolute values are large enough, the number of residual outliers is 1 which is within the specified tolerance level, normality is guaranteed and the fitting is very good. The statistically best regression subequation (38) looks really good and useful. Let us examine the meanings or roles of selected explanatory variables and implications of the regression coefficients in detail from the viewpoints of prefectural administrations. First of all, it is clear that X_2 requires more civil servants than X_3. Unfortunately, X_2 was not selected but X_3 was. The selection of both X_1 and X_3 implies a sort of redundancy. X_6 assumes a positive regression coefficient. This implies that more civil servants are needed in spite of the fact that various administrative rights are transferred from prefectural governments to ordinace-designated cities and the related civil servants are transferred to other sectors or are not replaced after they retire. Civil servants have job security, if they behave well without committing serious crimes. The regression coefficient of X_{11} shows negative. However, the larger the standard financial sizes are, the more civil servants prefectural governments will be allowed to hire. The regression coefficients of X_6 and X_{11} contradict these facts. Thus, the accountability is very poor. As a result, the above statistically best regression subequation (38) cannot be adopted for policy making[14] .

[14] Nowadays, convenient and cheap communication systems like e-mail, internet, fax and tele-

Step 1′ in which the Box-Cox transformations for $M = 6$ are applied for the above Step 1:

The Box-Cox transformations for $M = 6$ (just for demonstration) were also applied for all possible regressions without any professional knowledge. Unfortunately, the same statistically best regression subequation as (38), which corresponds to $m = 1$, was obtained from among all 49,146 possible subsets in about 8 minutes 29 seconds of CPU time where 6×8191=49146. The Box-Cox transformations for $m = 2, 3, \cdots, 6$ were not effective.

Step 2 in which the professional knowledge for variable classification of all possible explanatory variables is available and actually used but that for scientific conditions on their regression coefficients is not used, although available here, and variances of a disturbance term are assumed to be constant (OLS without Box-Cox transformation is used for all meaningful subsets with no sign conditions):

Let us classify all 13 possible explanatory variables X_1 to X_{13} from the administrative viewpoints of prefectural governments. Since variable X_1 and a pair of X_2 and X_3 are alternatively important, $< 1 < X_1, (X_2, X_3) > 1 >$ is appropriate, because prefectural governments must definitely offer administrative services to the residents. Variables X_4 and X_5 are also alternatively important so that $< 1 < X_4, X_5 > 1 >$ is appropriate, because administrative services must be offered whether the administrated areas are large or small. Since variables X_6 to X_{13} are considered to be completely optional, $< 0 < X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13} > 8 >$ is proper. All regression coefficients are examined by a two-tailed t-test. We inputted the following functional form together with the same statistical criteria:

$$\begin{aligned} Y \;=\; & F(X_0 < 1 < X_1, (X_2, X_3) > 1 > \; < 1 < X_4, X_5 > 1 > \\ & < 0 < X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13} > 8 >). \end{aligned} \tag{39}$$

The number of all meaningful subsets of the functional form (39), which correspond to 1,024 primitive one-regression-at-a-time procedures, is 1,024 by $2 \times 2 \times 2^8 = 1024$. The following statistically best regression subequation was searched for in about 1

phone are available. We are in the age of information disclosure so that accountability is sought. If the statistically best but ineffective regression subequation (38) is employed in the local government administrative reform policy, the policy makers will be severely criticized and surely condemned by saying that they have been pursuing intellectual exercises and wasting residents' taxes.

second of CPU time:

$$\begin{array}{lcccc}
\widehat{Y}_{172} & = & 553.0272 & +0.3154005X_2 & +0.4618258X_3 \\
(S.DEV.) & & (64.32450) & (0.6303144\times10^{-1}) & (0.9829663\times10^{-1}) \\
(T\text{-}RATIO) & & (8.597457) & (5.003860) & (4.698287)
\end{array}$$

$$\begin{array}{ccc}
+0.2858288\times10^{-1}X_5 & +0.2212731X_6 & +0.1211306X_7 \\
(0.3625663\times10^{-2}) & (0.4297600\times10^{-1}) & (0.639346\times10^{-1}) \\
(7.883492) & (5.148760) & (1.894602)
\end{array}$$

$$\begin{array}{ccc}
-0.3706472X_{10} & -1.830162X_{11} & +323.1120X_{13} \\
(0.1065621) & (0.5547306) & (98.62510) \\
(-3.478225) & (-3.299190) & (3.276164)
\end{array} \qquad (40)$$

$$\widehat{R}^2 = 0.9607,\ \widehat{\overline{R}^2} = 0.9522,\ \widehat{AIC} = 558.7247,\ \widehat{SD} = 96.2901,$$
$$\widehat{V} = 9271.79,\ \widehat{DF} = 37,\ \widehat{JB} = 0.4063,\ \widehat{OT} = 3.305,\ TSL = 0.145.$$

The variable classification really reduced the calculation time. The statistically best regression subequation (40) is similar to but slightly worse than (38) in **Step 1**. X_2 is selected instead of X_1. The regression coefficients of X_6 and X_{11} are not considered to reflect the facts by the same reasons mentioned above. As a result, the statistically best regression subequation (40) cannot be adopted for policy making.

Step $2'$ in which the Box-Cox transformations for $M = 6$ are applied for the above Step 2:

The Box-Cox transformations for $M = 6$ were also applied for all meaningful subsets with no sign conditions. Unfortunately, the same statistically best regression subequation as (40), which corresponds to $m = 1$, was obtained from among all 6,144 meaningful subsets in about 1 minute 19 seconds of CPU time where 6×1024=6144. The Box-Cox transformations for $m = 2, 3, \cdots, 6$ were not effective.

So far the practically best regression subequation has not been found in the previous two steps of trial and error. Let us introduce scientific conditions on the regression coefficients in addition to the knowledge for variable classification and extend OLS to Box-Cox transformation.

Step 3 in which the professional knowledge both for variable classification of all possible explanatory variables and for scientific conditions on the signs of their regression coefficients is available and actually used and variances of a disturbance term are assumed to be constant with respect to the original data of a dependent variable (m =1) or become constant with respect to the Box-Cox transformed data (m =2,3,...,M) where OLS and Box-Cox transformation are used for all meaningful subsets with the sign conditions:

The increases in the data of variables X_k's for $k = 1, 2, 3, 4, 5, 11, 12$ press the prefectural governments to hire more civil servants, requiring that their regression coefficients assume the + signs and then leading to a one-tailed t-test hypothesis testing for these regression coefficients. On the other hand, the increases in the data of variables X_6 and X_{10} reduce the amount of administrative work so that the prefectural governments can reduce the number of civil servants, requiring that their regression coefficients assume the − signs and then leading to a one-tailed t-test hypothesis testing for these regression coefficients. The regression coefficients of X_7, X_8 and X_9 (and X_0) can assume either a positive or negative sign, leading to a two-tailed t-test hypothesis testing for these regression coefficients. Variable X_1 and a pair of X_2 and X_3 are specified as $< 1 < +X_1, (+X_2, +X_3) > 1 >$. Variables X_4 and X_5 should be set as $< 1 < +X_4, +X_5 > 1 >$. Finally, $< 0 < -X_6, X_7, X_8, X_9, -X_{10}, +X_{11}, +X_{12}, +X_{13} > 8 >$ is proper.

We set $M = 6$ for the Box-Cox transformation so that the total number of all possible regressions is 49,146, where 6×8191=49146, whereas the number of all meaningful subsets is 6,144, where $6 \times 1024 = 6144$. Hence, the remaining 43,002 nonempty subsets are regarded as meaningless from the administrative viewpoints. The following functional form was inputted together with the same statistical criteria as before:

$$\begin{aligned} Y \;=\; & F(X_0, < 1 < +X_1, (+X_2, +X_3) > 1 >, < 1 < +X_4, +X_5 > 1 >, \\ & < 0 < -X_6, X_7, X_8, X_9, -X_{10}, +X_{11}, +X_{12}, +X_{13} > 8 >). \end{aligned} \tag{41}$$

The following practically best regression subequation was obtained from among 6,144 meaningful subsets in less than 3 seconds of CPU time:

$$\begin{array}{lccc} \hat{Y}^4_{4094} & = & 27.35750 & +0.2521440 \times 10^{-2} X_1 \quad +0.1796585 \times 10^{-3} X_5 \\ (S.DEV.) & & (0.4637656) & (0.1390876 \times 10^{-3}) \quad (0.2421109 \times 10^{-4}) \\ (T\text{-}RATIO) & & (58.98992) & (18.12844) \quad (7.420507) \end{array}$$

$$+4.823529 X_{13} \qquad (42)$$
$$(1.912091)$$
$$(2.522646)$$

$$\hat{m} = 4,\ \hat{\lambda} = 0.4,\ (\hat{R}^4)^2 = 0.9137,\ (\widehat{\mathcal{R}^4})^2 = 0.9076,\ \widehat{AIC}^4 = 192.4663,$$
$$\widehat{SD}^4 = 1.88065,\ \hat{V}^4 = 3.53683,\ \widehat{DF}^4 = 42,\ \widehat{JB}^4 = 0.3923,\ \widehat{OT}^4 = 2.972,$$
$$TSL = 0.145$$

where $\hat{m} = 4$ implies the fourth Box-Cox transformation for Y and is used as a superscript; $\hat{Y}^4_{4094}$ =estimate of Y^4, i.e., $(Y^{0.4} - 1)/0.4$ by the 4,094-th meaningful subset.

Needless to say, the (inversely-transformed) estimates $\hat{\hat{Y}}^4_{4094}$, which correspond to

the original data Y, of the fourth Box-Cox transformed estimates $\hat{Y}^4_{4094}$ are automatically calculated and plotted in the diagram (not cited here). The CPU time was furthermore reduced. If a + or − sign is specified in front of an explanatory variable in a functional form, the CPU time is reduced, because the calculation of percentiles of statistical distributions is not needed, if the regression coefficient of a + or − signed explanatory variable is not consistent with the specified sign, implying that that regression subequation is already unsatisfactory. The Chow equal coefficients test and the Goldfeld-Quandt homoscedasticity test were not utilized because X_{13} was selected in (42). Accordingly, we subtract from y_{46} the number of civil servants hired for dealing with the work related to the US military forces in Okinawa ($t = 46$) characterized by the dummy variable X_{13} in (42), denote the new data set by Y_*, calculate $\{(Y_*)^{0.4} - 1\}/0.4$, denote it by Y^4_* and run $Y^4_* = F(X_0, X_1, X_5)$. Since the null hypotheses of these tests were maintained, the practically best regression subequation (42) was finally adopted for policy making with a 14.5 % total risk of committing a type I error or with a 14.5 % total risk that (42) was actually not best.

6.2 Normal Random Simulations

We regard the practically best regression subequation (42) as true and conduct 100 simulations by normal random numbers with $\mathcal{N}(0, 3.53683)$, which are assigned to a disturbance term, under the following 5 simulation environments.

Simulation 1 by normal random numbers:

Suppose that we do not have the professional knowledge about variable classification and the signs of regression coefficients. The functional form (37) was used under the same statistical tests and criteria as in **Step 1′** (with the normal random number command). Then, 0 out of 100 simulations (0 %) revealed the best subset $\{X_0, X_1, X_5, X_{13}\}$ of (42) in 9 minutes 627 milliseconds of CPU time. 819,100 possible subsets were examined.

Simulation 2 by normal random numbers:

Next, we assumed that we have the professional knowledge about variable classification but not about the signs of their regression coefficients. The functional form (39) was used under the same statistical tests and criteria as in **Step 2′**. Then, 2 out of 100 simulations (2 %) revealed the best subset $\{X_0, X_1, X_5, X_{13}\}$ in about 1 minute 24 seconds 972 milliseconds of CPU time. 102,400 meaningful subsets were examined.

Simulation 3 by normal random numbers:

Furthermore, we assumed that we have the professional knowledge about variable classification and the signs of their regression coefficients. The functional form (39) was used under the same statistical tests and criteria as in **Step 3**. Then, 31 out of 100 simulations (31 %) revealed the best subset $\{X_0, X_1, X_5, X_{13}\}$ in 43 seconds 452 milliseconds of CPU time. 102,400 meaningful subsets were examined.

Simulation 4 by normal random numbers:

We regard the practically best regression subequation (42) as true and conduct 100 simulations by normal random numbers, which are assigned to a disturbance term. Suppose that we do not have any professional knowledge about the signs of regression coefficients. The following functional form was inputted

$$Y^4 = F(X_0, < 3 < X_1, X_5, X_{13} > 3 >). \tag{42}$$

Then, 45 out of 100 simulations (45 %) revealed the best subset $\{X_0, X_1, X_5, X_{13}\}$ of (42) in 620 milliseconds of CPU time.

Simulation 5 by normal random numbers:

Next, the professional knowledge about the signs of their regression coefficients was used. The following functional form was inputted

$$Y^4 = F(X_0, < 3 < +X_1, +X_5, +X_{13} > 3 >). \tag{43}$$

Then, 64 out of 100 simulations (64 %) revealed the best subset $\{X_0, X_1, X_5, X_{13}\}$ in 460 milliseconds of CPU time.

We can say from the **Steps 1 to 3** and the **Simulations 1 to 5** that the correct professional knowledge for variable classification and the sure sign (and magnitude) conditions must be employed, if they are a priori known and available. An unaccountable regression subequation is easily selected as best, otherwise.

7 Concluding Remarks

Various statistical tests and estimation methods have been proposed in the literature. Nonetheless, any method to systematically utilize them has not been proposed to search for the best regression equation. It is impossible to define the perfect, optimal or impeccable regression equation, so long as the Neyman-Pearson approach is taken. In actual and responsible applications of regression analysis, the professional knowledge or information based on the science(s) related to the research in question

is usually needed in addition to statistical and data-analytic knowledge. The author has defined the practically best regression equation. Users of regression analysis have been waiting for a method to concretely solve a variable selection problem for the Box-Cox transformation. It is urgent to create such a method, because valuable resources have been wasted in the world every day.

To shed some light on the problem, the author concretely formulated the j-th OLS-best subset problem for the Box-Cox transformation, proposed a knowledge-based variable selection method to solve it and demonstrated how to solve it in three steps and five normal random simulations. The proposed variable selection method is quite resource-saving. The following cases are not here referred to: (i) lagged dependent variables are used as possible explanatory variables, (ii) the Box-Cox transformation is used for all or some possible explanatory variables and (iii) constraints are imposed on the regression coefficients of some explanatory variables. However, it is not difficult to handle them. The concepts and computational techniques proposed are quite effective for reducing research costs and for teaching beginners in statistics, including undergraduate students, up to advanced applied researchers, including business strategists and policy makers, regression analysis and its extensions. The author hopes that this will contribute to making statistics a more useful and pleasant science than at present.

Acknowledgements

The author deeply appreciates an anonymous referee for his valuable comments and Professor M. Koda for his useful suggestions which improved the quality of this paper. He also thanks Computer Technician Mr. H. Kikuchi for converting the System **OEPP** developed for a main frame machine into that for a PC and a UNIX machine.

Appendix

The data used for the demonstration in Section 6 are listed below. See the names of variables and the measuring units of the data in the same section[15] .

[15] Since the administrative functions of the Tokyo metropolitan government differ from those of the other 46 prefectural governments, the data of the Tokyo metropolitan government are not cited here. For instance, the land of the Diet of the Japanese government belongs to the Tokyo metropolitan government so that there is a special and strong pipe line between the central government and the Tokyo metropolitan government.

No.	Prefectures	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	Hokkaido	2651	5692	4356	1336	26752.2	83451.6	1768	0
2	Aomori	853	1508	965	543	3096.4	9605.6	0	0
3	Iwate	811	1430	865	565	3661.6	15277.8	0	0
4	Akita	894	2324	1513	811	3085.5	7284.6	957	0
5	Miyagi	781	1219	669	550	3142.9	11611.7	0	309
6	Yamagata	784	1253	896	357	2857.7	9323.3	0	0
7	Fukushima	987	2140	1366	774	4127.7	13782.5	0	323
8	Niigata	973	2975	1675	1300	3913.9	6093.8	0	0
9	Ibaraki	709	1989	1302	687	2886.4	6408.3	0	435
10	Tochigi	651	2005	1234	771	2257.5	6363.2	0	0
11	Gunma	1491	6766	5817	949	2539.8	3797.2	0	0
12	Chiba	1268	5807	5053	753	3449.5	5155.9	846	0
13	Saitama	2327	8217	7832	385	1434.8	2413.6	4487	0
14	Kanagawa	832	2491	1606	885	4562.5	12581.8	0	484
15	Yamanashi	600	1128	788	340	1844.8	4246.5	0	322
16	Nagano	621	1174	807	367	1386.3	4184.8	0	0
17	Toyama	551	827	556	271	1056.5	4188.4	0	0
18	Ishikawa	633	881	410	471	942.1	4465.4	0	0
19	Fukui	864	2194	1408	786	3285.8	13585.2	0	403
20	Shizuoka	982	2103	1346	757	2078.6	10598.2	0	436
21	Gifu	1141	3742	2912	830	2705.5	7779.0	0	1029
22	Aichi	1724	6801	5714	1088	2906.8	5150.5	2085	0
23	Shiga	869	1849	1251	598	1983.8	5773.7	0	0
24	Mie	674	1294	713	581	1290.0	4017.4	0	0
25	Kyoto	985	2555	2182	373	1129.5	4612.4	1390	0
26	Osaka	1615	8607	8376	232	1298.0	1892.1	2479	791
27	Nara	1441	5447	4591	855	2670.8	8386.6	1442	470
28	Wakayama	674	1441	1031	410	833.3	3691.1	0	0
29	Hyogo	842	1098	670	428	1088.6	4724.3	0	397
30	Tottori	467	619	370	249	882.7	3507.0	0	0
31	Okayama	663	771	453	318	1295.0	6706.7	0	0
32	Shimane	859	1954	1429	624	2196.5	7111.1	0	608
33	Hiroshima	804	2873	2242	631	2214.5	8474.8	1093	0
34	Yamaguchi	671	1548	1201	347	1703.7	6110.1	0	0
35	Tokushima	564	837	431	406	1003.8	4144.4	0	0
36	Kagawa	583	1034	554	480	981.4	1875.2	0	0
37	Ehime	754	1522	1084	438	1660.9	5675.2	0	0
38	Kochi	546	824	556	268	1161.1	7104.1	0	0
39	Fukuoka	1094	4920	3740	1180	2731.5	4967.6	2259	0
40	Saga	491	886	458	427	1354.7	2439.0	0	0
41	Nagasaki	812	1547	968	580	1638.7	4090.7	0	430
42	Oita	924	1868	1081	787	2665.9	7402.3	0	638
43	Kumamoto	713	1240	910	331	1770.4	6337.3	0	427
44	Miyazaki	612	1189	800	389	1828.2	7733.7	0	0
45	Kagoshima	891	1795	1030	765	3295.6	9186.0	0	541
46	Okinawa	922	1296	867	429	1112.3	2266.0	0	0

No.	Prefectures	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}
1	Hokkaido	3926	422	1121.1	1235	212	0
2	Aomori	646	-12	0	344	67	0
3	Iwate	406	34	0	356	59	0
4	Akita	1188	667	783.5	399	71	0
5	Miyagi	402	-132	0	312	69	0
6	Yamagata	496	-39	0	298	44	0
7	Fukushima	783	240	0	441	90	0
8	Niigata	902	814	0	495	85	0
9	Ibaraki	745	378	0	366	49	0
10	Tochigi	803	283	0	350	70	0
11	Gunma	4899	3008	0	830	92	0
12	Chiba	3844	1926	272.1	748	80	0
13	Saitama	7284	2150	578.3	970	37	0
14	Kanagawa	1097	128	0	523	112	0
15	Yamanashi	435	32	0	264	35	0
16	Nagano	551	123	0	266	41	0
17	Toyama	326	64	0	224	35	0
18	Ishikawa	288	183	0	221	64	0
19	Fukui	588	269	0	453	120	0
20	Shizuoka	824	252	0	387	99	0
21	Gifu	2023	474	0	588	74	0
22	Aichi	4693	1511	326.4	977	88	0
23	Shiga	723	347	0	351	69	0
24	Mie	412	571	0	262	50	0
25	Kyoto	2098	133	610.2	416	44	0
26	Osaka	8304	552	220.7	1184	44	0
27	Nara	3967	431	546.9	807	91	0
28	Wakayama	777	494	0	259	47	0
29	Hyogo	455	68	0	256	50	0
30	Tottori	175	3	0	190	39	0
31	Okayama	96	-81	0	247	59	0
32	Shimane	736	174	0	373	78	0
33	Hiroshima	1730	221	740.9	483	86	0
34	Yamaguchi	748	-139	0	324	56	0
35	Tokushima	248	-11	0	227	50	0
36	Kagawa	347	42	0	224	43	0
37	Ehime	714	-53	0	311	70	0
38	Kochi	337	-68	0	246	53	0
39	Fukuoka	3199	1199	820.5	707	97	0
40	Saga	234	44	0	224	49	0
41	Nagasaki	700	-141	0	337	79	0
42	Oita	708	167	0	367	94	0
43	Kumamoto	536	-22	0	292	58	0
44	Miyazaki	476	90	0	280	44	0
45	Kagoshima	754	9	0	396	96	0
46	Okinawa	743	479	0	268	53	1

References

[1] Akaike, H., Information theory and an extension of the maximum likelihood principle, in: B.N. Petrov and F. Csaki (Ed.), *Pro. 2nd inter. symp. information theory* (Akademiai Kiado, Budapest, 1973) 267-281.

[2] Box, G.E.P. and Cox, D.R., An analysis of transformations, *Royal Statist. Soc.* ser. B (1964) 211-252.

[3] Chow, G.C., Tests of equality between sets of coefficients in two linear regressions, *Econometrica*, **28** (1960) 591-605.

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

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

[6] Goldfeld, S.M. and Quandt, R.E., Some tests for homoscedasticity, *J. Amer. Statist. Assoc.*, **60** (1965) 539-547.

[7] Jarque, C.M. and A.K. Bera, Efficient tests for normality, homoscedasticity and serial independence of regression residuals, *Economics Letters*, **6** (1980) 255-259.

[8] Kitagawa, T., *Toukei Jyohou Ron, I, II* (in Japanese), (*Statistical Information Theory, I* and *II*) (Kyouritsu Shuppan, Tokyo, 1987).

[9] Onishi, H., A variable selection procedure for econometric models, *Computa. Statist. Data Anal.*, 1(2) (1983) 85-95.

[10] Onishi, H., An efficient method for building a prefectural government official reduction model by the Researcher System OEPP, *J. Japanese Soc. Computa. Statist.*, **7**(1) (1994) 119-139.

[11] Onishi, H., A user knowledge-based variable selection method for limited information maximum likelihood using principal components, *Computa. Statist. Data Anal.*, **19**(4) (1995) 379-399.

[12] Onishi, H., Generalized regression data analysis in the Researcher System OEPP, in: Scientific Program Committee (Ed.), *Pro. inter. conf. statistical methods and quality and productivity improvement*, Vol. 1 (Seoul, 1995) 337-346, 1995, invited and presented in Conference on Statistical Methods and Statistical Computing for Quality and Productivity Improvement, Seoul, Korea, August 17-19, 1995.

[13] Sawa, T., *Kaiki Bunseki* (in Japanese), (*Regression Analysis*) (Asakura, Tokyo, 1979).

[14] Theil, H., *Economic Forecast and Policy* (North-Holland, Amsterdam, 1961).

[15] Wallis, K.F., Testing for fourth order for autocorrelation in quarterly regression equations, *Econometrica*, **40** (1972) 617-636.