PROCEEDINGS OF THE

9th International Biometric Conference

Volume I

BOSTON
AUGUST 22-27, 1976

THE BIOMETRIC SOCIETY

PROCEEDINGS OF THE

9th International Biometric Conference

Invited Papers

Volume I

BOSTON
AUGUST 22-27, 1976

THE BIOMETRIC SOCIETY

ACKNOWLEDGMENTS

The Biometric Society expresses sincere appreciation to the following organizations and individuals for their support toward the costs of the meetings and the publication of the volumes of proceedings:

National Institutes of Health
General Medical Sciences, Bethesda, Maryland
National Cancer Institute, Bethesda, Maryland
National Eye Institute, Bethesda, Maryland
National Heart, Lung, and Blood Institute, Bethesda, Maryland
National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina
National Institute of Neurological Diseases and Stroke, Bethesda, Maryland
National Center for Health Services Research, Rockville, Maryland
The Biometric Society, Raleigh, North Carolina
Abbott Laboratories, North Chicago, Illinois
Arthur D. Little, Inc., Cambridge, Massachusetts
Bristol Laboratories, Division of Bristol-Myers Company, Syracuse, New York
Burroughs Wellcome Company, Research Triangle Park, North Carolina
Carter-Wallace Inc., Cranbury, New Jersey
Ciba-Geigy Corporation, Summit, New Jersey
Energy Research and Development Administration, Oak Ridge, Tennessee
Hoechst-Roussel Pharmaceuticals Inc., Somerville, New Jersey
Hoffman-La Roche Inc., Nutley, New Jersey
Mr. William R. Holt, Elmsford, New York
ICI United States Inc., Wilmington, Delaware
International Union of Biological Sciences, Paris, France
Lederle Laboratories, Pearl River, New York
Lilly Research Laboratories, Division of Eli Lilly and Company, Indianapolis, Indiana
McNeil Laboratories, Inc., Fort Washington, Pennsylvania
Mead Johnson Research Center, Evansville, Indiana
Merck Sharp and Dohme Research Laboratories, West Point, Pennsylvania
Merrell-National Laboratories, Cincinnati, Ohio
Norwich Pharmacal Company, Norwich, New York
Ortho Pharmaceutical Corporation, Raritan, New Jersey
Pfizer Pharmaceuticals, New York, New York
R.J. Reynolds Industries, Inc., Winston-Salem, North Carolina
A.H. Robins Company, Richmond, Virginia
Sandoz, Inc., East Hanover, New Jersey
Smith Kline and French Laboratories, Philadelphia, Pennsylvania
E.R. Squibb and Sons, Inc., Princeton, New Jersey
Syntex (USA) Inc., Palo Alto, California
The Upjohn Company, Kalamazoo, Michigan
USV Pharmaceutical Corporation, Tuckahoe, New York
Vick Chemical Company, Philadelphia, Pennsylvania
Warner-Lambert Research Institute, Morris Plains, New Jersey
Wyeth Laboratories, Philadelphia, Pennsylvania

OFFICERS OF THE BIOMETRIC SOCIETY

President	H.L. LeRoy (ROeS)
Vice-President	C.R. Rao (GInd)
Secretary	J.S. Williams (WNAR)
Treasurer	L.A. Nelson (ENAR)

PROGRAM COMMITTEE, IBC

E.A. Gehan (Chairperson) (ENAR)
L.D. Calvin (WNAR)
L.C.A. Corsten (ANed)
R.N. Curnow (BR)
K. Dietz (Advisor) (ROeS)
J.J. Gart (ENAR)
F.P. Gomes (RBras)
A. Kudô (GJAP)
J. Lellouch (RF)
E. Marubini (RItl)
D. Rasch (RGDR)
E.J. Williams (AR)

PROGRAM CHAIRMAN BIOMETRICS SECTION, ASA

J.S. Williams

PROGRAM COMMITTEE, WNAR

R. Elashoff (WNAR)

LOCAL ORGANIZING COMMITTEE

Y.M. Bishop (Chairperson)
W.G. Cochran
T. Colton
S. Hartz
R. Kent
T. Louis
S. McKinlay
I. Miller
J. Warram
J. Worcester

FINANCE COMMITTEE

W.F. Taylor (Chairperson) (ENAR)
C. Bennett (WNAR)
L.D. Calvin (WNAR)
W. Dixon (WNAR)
S.M. Free, Jr. (ENAR)
E.A. Gehan (ENAR)
D. Gosslee (ENAR)
M. Kastenbaum (ENAR)
L.A. Nelson (ENAR)
M.A. Schneiderman (ENAR)

TABLE OF CONTENTS

SUNDAY, 22 AUGUST, 1976, 5:00 P.M.

OPENING CEREMONY

Boston Science Museum

WELCOME	F.B. Cady, Cornell University
REMARKS	R.L. Anderson, University of Kentucky
	C.I. Bliss, Yale University
	G.M. Cox, North Carolina State University
	W.G. Cochran, Harvard University
OPENING OF CONFERENCE	H.L. LeRoy, President, Biometric Society

MONDAY, 23 AUGUST, 1976, 8:30 A.M.

REGRESSION METHODS

(Joint with Biometrics Section, ASA
and Biometric Society (ENAR & WNAR))

ORGANIZER AND CHAIRMAN A. Kudô (Japan)

SPEAKERS page

DISCUSSANTS C. Daniel (USA) and

R.L. Plackett (England)

THE ANALYSIS AND SELECTION OF VARIABLES IN LINEAR REGRESSION

R. R. Hocking
Department of Computer Science and Statistics
Mississippi State University
Mississippi State, Mississippi 39762

SUMMARY

The purpose of this paper is to review some of the concepts and methods associated with the analysis and selection of variables in linear regression. Included is a summary of the highlights of variable selection methods and of biased estimation procedures. The relative performance of these techniques is illustrated by examples.

1. INTRODUCTION

In the multiple linear regression problem, the analyst attempts to develop an equation, depending on a set of input variables, which is to be used to describe a response variable. The available "pool" of input variables may consist of actual measurements, functions of them and indicator variables to describe qualitative characteristics. This pool is often quite sizeable, including some variables which are not relevant and some which are highly intercorrelated. Typically, the analysis will include a variable elimination procedure to reduce the number of variables and, presumably, yield a good equation.

Computational procedures for evaluating subset regressions and criteria for choosing the "best" subset have been the subject of numerous papers. Usually, the analysis of the original and the subset equations has been based on least squares.

More recently, several alternatives to least squares estimation have been proposed under the general heading of biased estimators.

These procedures were motivated, primarily, by the problem of multicollinearities among the input variables and were not directly concerned with variable elimination. However, the deletion of variables with small standardized coefficients is a natural step.

The purpose of this paper is to summarize the essential ideas in variable selection and biased estimation. It is natural to ask if the quality of the equation which results by any of these procedures is essentially independent of the method used to obtain it. Some tentative conclusions are made based on the results for two examples.

To limit the scope of this discussion it is assumed that the data satisfies the usual linear model assumptions. This includes the assumption that the data is free of outliers. On this point, Marquardt, [1974], observes that, the instability of least squares in the presence of near-linear dependencies among the input variables, may lead to erroneous conclusions with regard to possible outliers. Thus, one should include outlier analysis as an integral part of the entire process of determining a good equation rather than just a preliminary step.

The complexity of the problem and the volume of literature on the topic forces us to restrict this paper to some of the highlights of the problem. A more detailed account and an extensive bibliography is provided by Hocking [1976].

2. NOTATION AND BASIC CONCEPTS

2.1. Notation. The usual linear model with n observations on t input variables is assumed. For convenience, all variables, including the dependent variable, are standardized to have mean zero and unit sum of squares. The model and the data are expressed in matrix notation as

$$Y = X\beta + e \qquad (1)$$

where X is nxt of rank t. The least squares estimator of β is given by solving the normal equations

$$X'X\hat{\beta} = X'Y \tag{2}$$

where in view of our assumptions, X'X and X'Y consist of sample correlation coefficients.

In the discussion of alternative estimators, it is convenient to transform to the space of orthogonal predictors. To achieve this, let Λ denote the diagonal matrix of eigenvalues, λ_i, of X'X and T, the corresponding orthogonal matrix of eigenvectors, T_i. Thus,

$$T'X'XT = \Lambda. \tag{3}$$

Under the transformation

$$Z = XT \quad , \quad \beta = T\gamma \tag{4}$$

the model, (1), becomes

$$Y = Z\gamma + e \tag{5}$$

and the least squares estimate of γ is given by solving the equation,

$$\Lambda\hat{\gamma} = Z'Y. \tag{6}$$

The relation between $\hat{\beta}$ and $\hat{\gamma}$ is given by (4). That is, $\hat{\beta} = T\hat{\gamma}$. The variables in (5) are labelled according to the magnitude of the λ_i, with λ_t denoting the smallest and λ_1 the largest.

2.2. The Consequences of Incorrect Model Specification. In addition to the economical and practical reasons for wanting to minimize the number of variables in the final equation, variable deletion may be desireable in terms of the statistical properties of the estimates and of the final equation. The consequences of incorrectly specifying the model either by retaining extraneous variables or deleting essential variables can be examined by writing the model (1) in partitioned form as

$$Y = X_p\beta_p + X_r\beta_r + e \tag{7}$$

where X_r denotes the extraneous variables or the variables to be deleted.

A detailed account of properties of the estimates obtained by fitting (7) or the model with X_r deleted is given by Hocking [1976]. Letting $\hat{\beta}_p$ and $\hat{\beta}_r$ denote the least squares estimates for (7) and $\tilde{\beta}_p$ the estimate of β_p for the reduced model, the motivation for variable elimination is contained in the following:

1. $VAR(\hat{\beta}_p) - VAR(\tilde{\beta}_p)$ is positive semi-definite and if $\beta_r' (VAR(\hat{\beta}_r))^{-1} \beta_r \leq \sigma^2$ then $VAR(\hat{\beta}_p) - MSE(\tilde{\beta}_p)$ is positive semi-definite.
2. For prediction at $x' = (x_p' \; x_r')$, let $\hat{y} = x'\hat{\beta}$ and $\tilde{y}_p = x'_p\tilde{\beta}_p$. Then $VARP(\hat{y}) \geq VARP(\tilde{y}_p)$ and if $\beta_r' (VAR(\hat{\beta}_r))^{-1}\beta_r \leq \sigma^2$ then $VARP(\hat{y}) \geq MSEP(\tilde{y}_p)$. (Here VARP and MSEP denote prediction variance and mean squared error.)

2.3. The Effects of Multicollinearity. The problems which arise because of near-linear dependencies among the input variables were recently reviewed by Mason et al. [1975]. A near-linear relation between two variables is indicated by the simple correlation coefficient. More complex linear relations are indicated by small eigenvalues of the correlation matrix. Alternatively, if R_i^2 denotes the squared multiple correlation coefficient for regressing variable i on the remaining inputs, then R_i^2 close to one indicates a dependency. The specific nature of the relation can be obtained from this regression or, alternatively, by examining the eigenvectors.

To see the effects of such a degeneracy in the data, let r_{ij} denote the components of $X'X$ and r^{ij} denote the components of $(X'X)^{-1}$. Then,

$$r^{ii} = (1-R_i^2)^{-1} \tag{8}$$

and, for $i \neq j$

$$r^{ij} = -r_{ij.x} \, r^{ii} \, r^{jj}. \tag{9}$$

Here, $r_{ij.x}$ denotes the partial correlation between variables i and j conditioned on the remaining inputs variables. Clearly, the elements of $(X'X)^{-1}$ corresponding to the variables involved in a near-linear dependency will be large, and the estimates of the associated β_i will be highly variable. Marquardt [1970] referred to this as variance inflation and defined the variance inflation factor as VIF = Max (r^{ii}).

Hoerl and Kennard [1970a] noted that the average squared distance from $\hat{\beta}$ to β is given by

$$\text{trace } (\text{VAR}(\hat{\beta})) = \sigma^2 \sum_{i=1}^{t} \lambda_i^{-1}. \tag{10}$$

Near-linear dependencies, characterized by small eigenvalues, will inflate this quantity.

To see the effect on $\hat{\beta}_i$, note that if r_{iy} denotes the simple correlation between x_i and y then,

$$\hat{\beta}_i = r^{ii} r_{iy} + \sum_{j \neq i}^{t} r^{ij} r_{jy}. \tag{11}$$

If variable i is involved in the degeneracy, the effect of r_{iy} will be inflated.

With respect to variable elimination, the reduction in the regression sum of squares caused by eliminating variable i is given by

$$\text{Red}_i = \hat{\beta}_i^2 / r^{ii} = \hat{\beta}_i^2 (1-R_i^2). \tag{12}$$

A small value of $1-R_i^2$ may well be offset by the inflated value of $\hat{\beta}_i$ in (12), hence, variables involved in the dependency may be retained.

Finally, consider the effect on the residual sum of squares. The essential quantity is $X\hat{\beta}$ whose variance may be written as,

$$V(X\hat{\beta}) = \sum_{i=1}^{t} (XT_i)(XT_i)' \lambda_i^{-1} \sigma^2. \tag{13}$$

Note that if λ_i is near zero then XT_i will be nearly the zero vector, hence the effect of the small eigenvalues is dampened. As a result, the

residual sum of squares may not reflect the presence of small eigenvalues.

3. COMPUTATIONAL PROCEDURES AND CRITERIA FOR SUBSET SELECTION

The objective in a regression study is to obtain a thorough understanding of the relations between the input and response variables and the relative roles of the inputs in describing the response. There are classic examples of situations in which an input is apparently unrelated to the response but, when used in combination with other variables, this input is important. Conversely, when X is nearly degenerate, it is not unusual that the deletion of a variable with a large, standardized coefficient has little effect on the performance of the resulting equation.

One means of studying the relative roles of the input variables is to examine their performance in all possible combinations with other variables. It seems reasonable to assume that a careful study of all 2^t such subsets would provide the necessary information. The computational problem of fitting all 2^t regressions with the necessary output to provide the proper indicators is formidable for even moderate values of t. Also, when t is large, there is a need for reducing the evaluation of the performance of a subset to the inspection of a small number of statistics.

The computing problem is conceptually the simplest. The literature on computational techniques has been devoted to the development of efficient methods for evaluating all possible subsets for small t and methods that will reveal most of the desired information while performing a modest amount of computing for large t. In the former category the papers by Garside [1965], Schatzoff et al. [1968], Furnival [1971]

and Morgan and Tatar [1972] are important and in the latter category, we mention the papers by Hocking and Leslie [1967], Beale et al. [1967], Kirton [1967], Beale [1970], Lamotte and Hocking [1970] and Furnival and Wilson [1974]. The efficiency of these procedures combined with the quality and quantity of information which is provided would appear to make the stepwise procedures inadmissible.

The "optimal" subset algorithms, in the second category mentioned above, will identify the subset of each size with smallest residual sum of squares (RSS) and a number of other subsets with small RSS while evaluating only a small fraction of the possible 2^t subsets. The minimum RSS subsets are frequently referred to as "best", but it is emphasized that they are best only in this restricted sense. The implied optimality may reflect only this sample and may not be typical of the population. In many cases, for reasons of economy and convenience, and, as we shall see, variance inflation, other suboptimal subsets may be preferable.

Most of the criteria for subset evaluation are based on simple functions of RSS. If we let RSS_p denote RSS for a particular p-term equation, then some of the most common functions are,

1. $R_p^2 = 1 - RSS_p/Total$
2. $\bar{R}_p^2 = 1 - RMS_p(n-1)/Total$
3. $RMS_p = RSS_p/(n-p)$
4. $C_p = RSS_p/\hat{\sigma}^2 + 2p-n$
5. $S_p = RMS_p/(n-p-1)$.

A detailed account of these and other criteria functions is given by Hocking [1976].