Monographs
on Statistics and
Applied Probability 40

Subset Selection in Regression

A. J. Miller

Springer-Science+Business Media, B.V.

MONOGRAPHS ON STATISTICS AND APPLIED PROBABILITY

General Editors

D.R. Cox, D.V. Hinkley, D. Rubin and B.W. Silverman

1 Stochastic Population Models in Ecology and Epidemiology *M.S. Bartlett* (1960)

2 Queues *D.R. Cox and W.L. Smith* (1961)

3 Monte Carlo Methods *J.M. Hammersley and D.C. Handscomb* (1964)

4 The Statistical Analysis of Series of Events *D.R. Cox and P.A.W. Lewis* (1966)

5 Population Genetics *W.J. Ewens* (1969)

6 Probability, Statistics and Time *M.S. Bartlett* (1975)

7 Statistical Inference *S.D. Silvey* (1975)

8 The Analysis of Contingency Tables *B.S. Everitt* (1977)

9 Multivariate Analysis in Behavioural Research *A.E. Maxwell* (1977)

10 Stochastic Abundance Models *S. Engen* (1978)

11 Some Basic Theory for Statistical Inference *E.J.G. Pitman* (1979)

12 Point Processes *D.R. Cox and V. Isham* (1980)

13 Identification of Outliers *D.M. Hawkins* (1980)

14 Optimal Design *S.D. Silvey* (1980)

15 Finite Mixture Distributions *B.S. Everitt and D.J. Hand* (1981)

16 Classification *A.D. Gordon* (1981)

17 Distribution-free Statistical Methods *J.S. Maritz* (1981)

18 Residuals and Influence in Regression *R.D. Cook and S. Weisberg* (1982)

19 Applications of Queueing Theory *G.F. Newell* (1982)

20 Risk Theory, 3rd edition *R.E. Beard, T. Pentikainen and E. Pesonen* (1984)

21 Analysis of Survival Data *D.R. Cox and D. Oakes* (1984)

22 An Introduction to Latent Variable Models *B.S. Everitt* (1984)

23 Bandit Problems *D.A. Berry and B. Fristedt* (1985)

24 Stochastic Modelling and Control *M.H.A. Davis and R. Vinter* (1985)

25 The Statistical Analysis of Compositional Data *J. Aitchison* (1986)

26 Density Estimation for Statistical and Data Analysis *B.W. Silverman* (1986)

27 Regression Analysis with Applications *G.B. Wetherill* (1986)

28 Sequential Methods in Statistics, 3rd edition *G.B. Wetherill* (1986)

29 Tensor Methods in Statistics *P. McCullagh* (1987)

30 Transformation and Weighting in Regression *R.J. Carroll and D. Ruppert* (1988)

31 Asymptotic Techniques for Use in Statistics *O.E. Barndoff-Nielsen and D.R. Cox* (1989)

32 Analysis of Binary Data, 2nd edition *D.R. Cox and E.J. Snell* (1989)

33 Analysis of Infectious Disease Data *N.G. Becker* (1989)

34 Design and Analysis of Cross-Over Trials *B. Jones and M.G. Kenward* (1989)

35 Empirical Bayes Methods, 2nd edition *J.S. Maritz and T. Lwin* (1989)

36 Symmetric Multivariate and Related Distributions *K-T. Fang, S. Kotz and K. Ng* (1989)

37 Generalized Linear Models, 2nd edition *P. McCullagh and J.A. Nelder* (1989)

38 Cyclic Designs *J.A. John* (1987)

39 Analog Estimation Methods in Econometrics *C.F. Manski* (1988)

40 Subset Selection in Regression *A.J. Miller* (1990)

41 Analysis of Repeated Measures *M.J. Crowder and D.J. Hand* (1990)

(Full details concerning this series are available from the publishers)

Subset Selection in Regression

A.J. MILLER
Senior Principal Research Scientist
CSIRO Division of Mathematics and Statistics
Melbourne, Australia

Springer-Science+Business Media, B.V.

Originally published by Chapman and Hall in 1990.
Softcover reprint of the hardcover 1st edition 1990

Typeset in 10/12 pt Times by
Thomson Press (India) Ltd, New Delhi

ISBN 978-0-412-35380-2 ISBN 978-1-4899-2939-6 (eBook)
DOI 10.1007/978-1-4899-2939-6

British Library Cataloguing in Publication Data

Miller, Alan J.
Subset selection in regression.
1. Multiple linear regression analysis
I. Title
519.536

Library of Congress Cataloging-in-Publication Data

Miller, Alan J.
Subset selection in regression/Alan J. Miller.
p. cm.—(Monographs on statistics and applied probability)
Includes bibliographical references.

1. Regression analysis. 2. Least squares. I. Title.
II. Series.
QA278.2.M56 1990
519.5'36—dc20

89-77350
CIP

Contents

Preface

Nearly all statistical packages, and many scientific computing libraries, contain facilities for the empirical choice of a model given a set of data and many variables or alternative models from which to select. There is an abundance of advice on how to perform the mechanics of choosing a model, much of which can only be described as folklore and some of which is quite contradictory. There is a dearth of respectable theory, or even of trustworthy advice, such as recommendations based upon adequate simulations. This monograph collects together what is known, and presents some new material on estimation. This relates almost entirely to multiple linear regression. The same problems apply to nonlinear regression, such as to the fitting of logistic regressions, to the fitting of autoregressive moving average models, or to any situation in which the same data are to be used both to choose a model and to fit it.

This monograph is not a cookbook of recommendations on how to carry out stepwise regression; anyone searching for such advice in its pages will be very disappointed. I hope that it will disturb many readers and awaken them to the dangers in using automatic packages which pick a model and then use least squares to estimate regression coefficients using the same data. My own awareness of these problems was brought home to me dramatically when fitting models for the prediction of meteorological variables such as temperature or rainfall. Many years of daily data were available, so we had very large sample sizes. We had the luxury of being able to fit different models for different seasons and to be able to use different parts of the data, chosen at random not systematically, for model selection, for estimation, and for testing the adequacy of the predictions. Selecting only those variables which were very highly 'significant', using 'F-to-enter' values of 8·0 or greater, it was found that some variables with 't-values' as large as 6 or even greater had their regression coefficients reversed in sign from the data subset

used for selection to that used for estimation. We were typically picking about 5 variables out of 150 available for selection.

Many statisticians and other scientists have long been aware that the so-called significance levels reported by subset selection packages are totally without foundation, but far fewer are aware of the substantial biases in the (least-squares or other) regression coefficients. This is one aspect of subset selection which is emphasized in this monograph.

The topic of subset selection in regression is one which is viewed by many statisticians as 'unclean' or 'distasteful'. Terms such as 'fishing expeditions', 'torturing the data until they confess', 'data mining', and others are used as descriptions of these practices. However, there are many situations in which it is difficult to recommend any alternative method and in which it is plainly not possible to collect further data to provide an independent estimate of regression coefficients, or to test the adequacy of fit of a prediction formula, yet there is very little theory to handle this very common problem. It is hoped that this monograph will provide the impetus for much badly needed research in this area.

It is a regret of mine that I have had to use textbook examples rather than those from my own consulting work within CSIRO. My experience from many seminars at conferences in Australia, North America and the UK, has been that as soon as one attempts to use 'real' examples, the audience complains that they are not 'typical', and secondly, there are always practical problems which are specific to each particular data set which distract attention from the main topic. I am sure that this applies very much to the textbook examples which I have used, and I am grateful that I do not know of these problems!

This is not in any sense a complete text on regression; there is no attempt to compete with the many hundreds of regression books. For instance, there is almost no mention of methods of examining residuals, of testing for outliers, or of the various diagnostic tests for independence, linearity, normality, etc. Very little is known of the properties of residuals and of other diagnostic statistics after model selection.

Many people must be thanked for their help in producing this monograph, which has taken more than a decade. The original impetus to develop computational algorithms came from John Maindonald and Bill Venables. John Best, John Connell (who

provided a real problem with 757 variables and 42 cases), Doug Shaw and Shane Youll tried the software I developed and found the bugs for me. It soon became obvious that the problems of inference and estimation were far more important than the computational ones. Joe Gani, then Chief of CSIRO Division of Mathematics and Statistics, arranged for me to spend a six-month sabbatical period at the University of Waterloo over the northern hemisphere winter of 1979/80. I am grateful to Jerry Lawless and others at Waterloo for the help and encouragement which they gave me. Hari Iyer is to be thanked for organizing a series of lectures which I gave at Colorado State University in early 1984, just prior to reading a paper on this subject to the Royal Statistical Society of London. The monograph was then almost completed at Griffith University (Brisbane) during a further sabbatical spell which Terry Speed generously allowed me from late 1985 to early 1987. The most important person to thank is Doug Ratcliff, who has been a constant source of encouragement, and has read all but the last version of the manuscript, and who still finds bugs in my software. I of course accept full responsibility for the errors remaining. I would also like to thank Sir David Cox for his support in bringing this monograph to publication.

Alan Miller

Melbourne

CHAPTER 1

Objectives

1.1 Prediction, explanation, elimination or what?

There are several fundamentally different situations in which it may be desired to select a subset from a larger number of variables. The situation with which this monograph is mainly concerned is that of predicting the value of one variable, which will be denoted by Y, from a number of other variables, which will usually denoted by X's. It may be required to do this because it is expensive to measure the variable Y and it is hoped to be able to predict it with sufficient accuracy from other variables which can be measured cheaply. A more common situation is that in which the X-variables measured at one time can be used to predict Y at some future time. In either case, unless the true form of the relationship between the X- and Y-variables is known, it will be necessary for the data used to select the variables and to calibrate the relationship to be representative of the conditions in which the relationship will be used for prediction. This last remark particularly applies when the prediction requires extrapolation, e.g. in time, beyond the range over which a relationship between the variables is believed to be an adequate approximation.

Some examples of applications are

1. the estimation of wool quality, which can be measured accurately using chemical techniques requiring considerable time and expense, from reflectances in the near infra-red region, which can be obtained quickly and relatively inexpensively;
2. the prediction of meteorological variables, e.g. rainfall or temperature, say 24 hours in advance, from current meteorological variables and variables predicted from mathematical models;
3. the prediction of tree heights at some future time from variables such as soil type, topography, tree spacing, rainfall, etc.

The emphasis here is upon the task of prediction not upon the explanation of the effects of the X-variables on the Y-variable, though the second problem will not be entirely ignored. The distinction between these two tasks is well spelt out by Cox and Snell (1974). However, for those whose objective is not prediction, Chapter 4 is devoted to testing inferences with respect to subsets of regression variables in the situation in which the alternative hypotheses to be tested have not been chosen *a priori*.

Also we will not be considering what is sometimes called the 'screening' problem, that is the problem of eliminating some variables (e.g. treatments or doses of drugs) so that effort can be concentrated upon the comparison of the effects of a smaller number of variables in future experimentation. The term 'screening' has been used for a variety of different meanings and, to avoid confusion, will not be used again in this monograph.

In prediction we are usually looking for a small subset of variables which gives adequate prediction accuracy for a reasonable cost of measurement. On the other hand, in trying to understand the effect of one variable on another, particularly when the only data available are observational or survey data rather than experimental data, it may be desirable to include many variables which are either known or believed to have an effect.

Sometimes the data for the predictor variables will be collected for other purposes and there will be no extra cost to include more predictors in the model. This is often the case with meterological data, or when using government-collected statistics in economic predictions. In other situations, there may be substantial extra cost so that the cost of data collection will need to be traded off against improved accuracy of prediction.

In general, we will assume that all predictors are available for inclusion or exclusion from the model, though this is not always the situation in practice. In many cases, the original set of measured variables will be augmented with other variables calculated from them. Such variables could include the squares of variables, to allow curvature in the relationship, or simple products of variables, to allow for the gradient of the regression of Y on say X_1 to change with the value of another variable, say X_2. Usually a quadratic term will only be included in the model if the linear term is also included; similarly a product of two variables will only be included if at least one of the two original variables is also included. Computational

methods will be described, in Chapter 3, for finding best-fitting linear models subject to the unrestricted selection of variables.

In some practical situations we will want to obtain a 'point estimate' of the Y-variable, that is a single value for it given the values of the predictor variables. In other situations we will want to predict a probability distribution for the response variable Y. For instance, rather than just predicting that tomorrow's rainfall will be 5 mm we may want to try to assign one probability that it will not rain at all and another probability that the rainfall will exceed say 20 mm. This kind of prediction requires a model for the distribution of the Y-variable about the regression line. In the case of rainfall, a lognormal or a gamma distribution is often assumed with parameters which are simple functions of the point estimate for the rainfall, though the distribution could be modelled in more detail. Attention in this monograph will be focused mainly on the point estimate problem.

All of the models which will be considered in this monograph will be linear, that is they will be linear in the regression coefficients. Though most of the ideas and problems carry over to the fitting of nonlinear models, the complexity is greatly increased. Also, though there are many ways of fitting regression lines, least squares (LS) will be almost exclusively used. Other types of model have been considered by Linhart and Zucchini (1986), while Boyce, Farhi and Weischedel (1974) consider the use of subset selection methods in optimal network algorithms.

There has been a small amount of work done on multivariate subset selection. The reader is referred to Seber (1984) and Sparks *et al.* (1985) for an introduction to this subject.

An entirely different form of empirical modelling is that of classification and regression trees. In this the data are split into two parts based upon the value of one variable, say X_1. This variable is chosen as that which minimizes the variation of the Y-variable within each part while maximizing the difference between the parts. Various measures of distance or similarity are used in different algorithms. After splitting on one variable, the separate parts are then split. Variable X_2 may be used to split one part, and perhaps X_3, or X_2 or even X_1 again, may be used to split the other part. Such methods are usually employed when the dependent variable is a categorical variable rather than a continuous one. This kind of modelling will not be considered here, but it suffers from the same problems of

over-fitting and biases in estimation as subset selection in multiple regression. For discussion of some of these clustering methods, see e.g. Everitt (1974), Hartigan (1975) or Breiman *et al.* (1984).

When the noise in the data is sufficiently small, or the quantity of data is sufficiently large, that the detailed shape of the relationship between the dependent variable and its predictors can be explored, the technique known as projection pursuit may be relevant. See e.g. Huber (1985), Friedman (1987), Jones and Sibson (1987) or Hall (1989).

1.2 How many variables in the prediction formula?

It is tempting to include in a prediction formula all of those variables which are known to affect or are believed to affect the variable to be predicted. Let us look closer at this idea. Suppose that the predictor variable, Y, is linearly related to the k predictor variables, $X_1, X_2, \ldots, X_k$ thus

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \varepsilon \tag{1.1}$$

where the residuals, ε, have zero mean and are independently sampled from the same distribution which has a finite variance σ^2. The coefficients $\beta_0, \beta_1, \ldots, \beta_k$ will usually be unknown, so let us estimate them using LS. The LS estimates of the regression coefficients, to be denoted by b's, are given in matrix notation by

$$\mathbf{b} = (X'X)^{-1}X'\mathbf{Y}$$

where

$$\mathbf{b}' = (b_0, b_1, \ldots, b_k),$$

X is an $n \times (k+1)$ matrix in which row i consists of a 1 followed by the values of variables $X_1, X_2, \ldots, X_k$ for the ith observation, and $\mathbf{Y}$ is a vector of length n containing the observed values of the variable to be predicted.

Now let us predict $\mathbf{Y}$ for a given vector $\mathbf{x}' = (1, x_1, \ldots, x_k)$ of the predictor variables, using

$$\begin{aligned}\hat{Y} &= \mathbf{x}'\mathbf{b} \\ &= b_0 + b_1 x_1 + \cdots + b_k x_k.\end{aligned}$$

Then from standard LS theory (see e.g. Seber, 1977 p. 364), we have

that

$$\text{var}(\mathbf{x}'\mathbf{b}) = \sigma^2 x'(X'X)^{-1}\mathbf{x}.$$

If we form the Cholesky factorization of $X'X$, i.e. we find a $(k+1) \times (k+1)$ upper-triangular matrix $\boldsymbol{R}$ such that

$$(X'X)^{-1} = \boldsymbol{R}^{-1}\boldsymbol{R}^{-T}$$

where the superscript $-T$ denotes the inverse of the transpose, then it follows that

$$\text{var}(\mathbf{x}'\mathbf{b}) = \sigma^2(\mathbf{x}'\boldsymbol{R}^{-1})(\mathbf{x}'\boldsymbol{R}^{-1})'. \tag{1.2}$$

Now $\mathbf{x}'\boldsymbol{R}^{-1}$ is a vector of length $(k+1)$ so that the variance of the predicted value of Y is the sum of squares of its elements. This is a suitable way in which to compute the variance of $\hat{Y}$, though we will recommend later that the Cholesky factorization, or a similar triangular factorization, should be obtained directly from the X-matrix without the intermediate step of forming the 'sum of squares and products' matrix $X'X$.

Now let us consider predicting Y using only the first p of the X-variables where $p < k$. Write

$$X = (X_A, X_B)$$

where X_A consists of the first $(p+1)$ columns of X, and X_B consists of the remaining $(k-p)$ columns. Then it is well known that if we form the Cholesky factorization

$$X_A'X_A = R_A'R_A$$

then $\boldsymbol{R}_A$ consists of the first $(p+1)$ rows and columns of $\boldsymbol{R}$, and also that the inverse $\boldsymbol{R}_A^{-1}$ is identical with the same rows and columns of $\boldsymbol{R}^{-1}$. The reader who is unfamiliar with these results can find them in such references as Rushton (1951) or Stewart (1973) though it is obvious to anyone who tries forming a Cholesky factorization and inverting it that the factorization down to row p and the inverse down to row p are independent of following rows. The Cholesky factorization of $X'X$ can be shown to exist and to be unique except for signs provided that $X'X$ is a positive-definite matrix.

Then if $\mathbf{x}_A$ consists of the first $(p+1)$ elements of $\mathbf{x}$ and $\mathbf{b}_A$ is the corresponding vector of LS regression coefficients for the model with only p variables, we have similarly to (1.2) that

$$\text{var}(\mathbf{x}_A'\mathbf{b}_A) = \sigma^2(\mathbf{x}_A'\boldsymbol{R}_A^{-1})(\mathbf{x}_A'\boldsymbol{R}_A^{-1})'. \tag{1.3}$$

That is, the variance of the predicted values of Y is the sum of squares of the first $(p+1)$ elements which were summed to obtain the variance of $\mathbf{x}'\mathbf{b}$, and hence

$$\text{var}(\mathbf{x}'\mathbf{b}) \geqslant \text{var}(\mathbf{x}'_A\mathbf{b}_A).$$

Thus the variance of the predicted values increases monotonically with the number of variables used in the prediction – or at least it does for linear models with the parameters fitted using least squares. This fairly well-known result is at first difficult to understand. Taken to its extremes it could appear that we get the best predictions with no variables in the model. If we always predict $Y = 7$ say, irrespective of the values of the X-variables, then our predictions have zero variance but probably have a very large bias.

If the true model is as given in (1.1) then

$$\mathbf{b}_A = (X'_AX_A)^{-1}X'_A\mathbf{Y}$$

and hence

$$\begin{aligned} E(\mathbf{b}_A) &= (X'_AX_A)^{-1}X'_AX\boldsymbol{\beta} \\ &= (X'_AX_A)^{-1}X'_A(X_A, X_B)\boldsymbol{\beta} \\ &= (X'_AX_A)^{-1}(X'_AX_A, X'_AX_B)\boldsymbol{\beta} \\ &= \boldsymbol{\beta}_A + (X'_AX_A)^{-1}X'_AX_B\boldsymbol{\beta}_B \end{aligned}$$

where $\boldsymbol{\beta}_A$, $\boldsymbol{\beta}_B$ consist of the first $(p+1)$ and last $(k-p)$ elements respectively of $\boldsymbol{\beta}$. The second term above is therefore the bias in the first $(p+1)$ regression coefficients arising from the omission of the last $(k-p)$ variables. The bias in estimating Y for a given $\mathbf{x}$ is then

$$\begin{aligned} \mathbf{x}'\boldsymbol{\beta} - E(\mathbf{x}'_A\mathbf{b}_A) &= \mathbf{x}'_A\boldsymbol{\beta}_A + \mathbf{x}'_B\boldsymbol{\beta}_B - \mathbf{x}'_A\boldsymbol{\beta}_A - \mathbf{x}'_A(X'_AX_A)^{-1}X'_AX_B\boldsymbol{\beta}_B \\ &= \{\mathbf{x}'_A - \mathbf{x}'_A(X'_AX_A)^{-1}X'_AX_B\}\boldsymbol{\beta}_B. \end{aligned} \tag{1.4}$$

As more variables are added to a model we are 'trading off' reduced bias against an increased variance. If a variable has no predictive value then adding that variable merely increases the variance. If the addition of a variable makes little difference to the biases then the increase in prediction variance may exceed the benefit from bias reduction. The question of how this trade-off should be handled is a central problem in this field, but its answer will not be attempted until Chapter 6 because of the very substantial problems of bias when the model has not been selected independently of the data. We note though that the addition of extra variables does not

generally reduce the bias for every vector **x**. Also, the best subset for prediction is a function of the range of vectors **x** for which we want to make predictions.

If the number of observations in the calibrating sample can be increased then the prediction variance given by (1.3) will usually be reduced. In most practical cases the prediction variance will be of the order n^{-1} while the biases from omitting variables will be of order 1 (that is, independent of n). Hence the number of variables in the best prediction subset will tend to increase with the size of the sample used to calibrate the model.

We note here that Thompson (1978) has discriminated between two prediction situations, one in which the X-variables are controllable, as for instance in an experimental situation, and the other in which the X-variables are random variables over which there is no control. In the latter case the biases caused by omitting variables can be considered as forming part of the residual variation and then the magnitude of the residual variance, σ^2, changes with the size of subset.

At this stage we should mention another kind of bias which is usually ignored. The mathematics given above is all for the case in which the subset of variables has been chosen independently of the data being used to estimate the regression coefficients. In practice the subset of variables is usually chosen from the same data as are used to estimate the regression coefficients. This introduces another kind of bias which we will call **selection bias**; the first kind of bias discussed above will be called **omission bias.** It is far more difficult to handle selection bias than the omission bias, and for this reason a whole chapter, Chapter 5, is devoted to this subject. Apart from a few notable exceptions, e.g. Kennedy and Bancroft (1971), this topic has been almost entirely neglected in the literature.

To illustrate the biases which can arise, consider the following simple artificial example. Let the population correlation matrix among three predictor variables, X_1, X_2, X_3, and a dependent variable, Y, be

$$\begin{pmatrix} 1 & 0{\cdot}5 & 0{\cdot}5 & 0{\cdot}5 \\ 0{\cdot}5 & 1 & 0{\cdot}5 & 0{\cdot}5 \\ 0{\cdot}5 & 0{\cdot}5 & 1 & 0{\cdot}5 \\ 0{\cdot}5 & 0{\cdot}5 & 0{\cdot}5 & 1 \end{pmatrix}.$$

Artificial data were generated from the multivariate normal distribution with this matrix as its covariance matrix, and with zero population means for all variables. Let us suppose that one and only one predictor is to be chosen. Thus we have a case of three predictors competing for selection. It is a case which is fairly typical of many which occur in practice in which once one variable has been selected, only a small improvement in the fit can be obtained by adding the others.

Using a sample of 20 cases, the best fitting predictor of the three was found. We know in this case that the three predictors are equally good, but we would not know that in a practical situation. The regression coefficient was estimated for the chosen variable, with only that variable in the model (excluding a constant for simplicity). The fraction R^2 of the sum of squares of the dependent variable 'explained' by the chosen predictor was also calculated. This exercise was repeated for 1000 samples. Figure 1.1 shows histograms of the regression coefficient, b, and of R^2 for the best-fitting predictor; these are the first and third histograms down. The other two histograms are the combined histograms for X_1, whether or not it was the best-fitting, for X_2 and similarly for X_3. The expected value of the regression coefficient of Y upon any one of the predictors is 0·5, and the sample mean for all three predictors considered one at a time was close to this (0·491), but the average regression coefficient for the best-fitting variable was 0·593. That is, the bias in the regression coefficient in this case is about 20%. About two-thirds of the sample regression coefficients were above the expected value of 0·5. There is a similar bias for R^2 for which the population value for any one predictor chosen *a priori* is 0·25. The sample average for the best-fitting variable was 0·374. Only 25% of the values of R^2 for the best-fitting predictor were less than 0·25.

Though it has long been known that standard LS theory does not apply when the model has not been chosen independently of the data being used for estimation, this has rarely been stated explicitly in the literature. Hocking (1976) warns the reader that this is the case, but gives the reader no indication of the magnitude of the over-fitting, or of the biases in regression coefficients. Copas (1983) states explicitly (p. 321) that 'x_j... is more likely to be selected if $|\hat{\beta}_j|$ overestimates...than if it underestimates. Thus the coefficients for a selected subset will be biased, as a result of which the usual measures of fit will be too optimistic, sometimes markedly so'. Miller (1984)

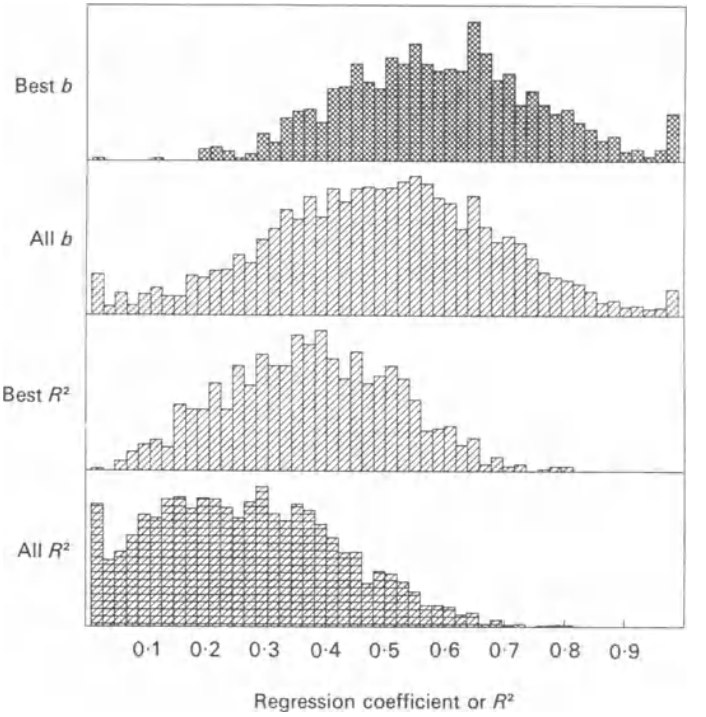


Fig. 1.1 *Histograms of values of the regression coefficient, b, and of R^2 for the best-fitting predictor out of three, and for all predictors taken one at a time.*

goes further and attempts to quantify the biases in a few simple cases.

The question of how many variables to include in the prediction equation, that is of deciding the 'stopping rule' in selection, is one which has developed along different lines in the multiple-regression context and in the context of fitting time series, though it is the same problem. In the time-series context, terms are often added in a predetermined order. For instance, if an autoregressive model is being fitted, the first term added is usually one with a lag of one time unit, the next is of two time units, etc. This closely parallels a common practice in fitting polynomial regressions, of first fitting a linear term, then a quadratic, etc. It is unusual in fitting time-series models to do a search for the best term to add next, though the

fitting of seasonal terms mixed with say monthly terms, plus the fitting of moving-average as well as autoregressive terms, means that the choice of alternative models considered can be moderately large.

In neither the multiple-regression nor the time-series case can an answer be given until selection bias is understood, except for the rare situation in which independent data sets are used for the selection of variables (or of the order of the model in fitting time series) and for the estimation of the regression coefficients. This topic will be discussed in detail in Chapter 6.

1.3 Alternatives to using subsets

The basic reason for not using all of the available predictor variables is that, unless we have sufficiently large data sets, some of the regression coefficients will be poorly determined and the predictions may be poor as a consequence. Two principal alternatives are available which use all of the variables, these are (i) using 'shrunken' estimators as in ridge regression, and (ii) using orthogonal (or nonorthogonal) linear combinations of the predictor variables.

The usual form in which the expression for the ridge regression coefficients is written is

$$\mathbf{b}(\theta) = (\boldsymbol{X}'\boldsymbol{X} + \theta \boldsymbol{I})^{-1}\boldsymbol{X}'\mathbf{Y}$$

where $\boldsymbol{I}$ is a $k \times k$ identity matrix, and θ is a scalar. In practice this is usually applied to predictor variables which have first been centred by subtracting the sample average and then scaled so that the diagonal elements of $\boldsymbol{X}'\boldsymbol{X}$ are all equal to 1. In this form the $\boldsymbol{X}'\boldsymbol{X}$-matrix is the sample correlation matrix of the original predictor variables. There is a very large literature on ridge regression and on the choice of the value for the ridge parameter θ, including several reviews (see e.g. Draper and van Nostrand, 1979; Smith and Campbell, 1980 and the discussion which follows this paper; Draper and Smith, 1981, p. 313–25).

The simplest shrunken estimator is that obtained by simply multiplying all of the LS regression coefficients by a constant between 0 and 1. Usually this shrinkage is not applied to the constant or intercept in the model (if there is one), and the fitted line is forced to pass through the centroid of the X- and Y-variables. The best known of these shrunken estimators is the so-called James–Stein

estimator (see Sclove, 1968). There have also been many other types of shrunken estimator proposed; see for instance the 57 varieties considered by Dempster, Schatzoff and Wermuth (1977).

All of these shrunken estimators yield biased estimates of the regression coefficients and hence usually of Y but, with a suitable choice of the parameter(s) controlling the amount of shrinkage, can produce estimates with smaller mean square errors of prediction than the LS estimator using all the predictor variables, over a range of the X-variables.

The use of orthogonal linear combinations of some or all of the predictor variables (called principal components in the statistical literature and empirical orthogonal functions in the geophysical sciences) is a way of reducing the number of variables. Usually only the combinations which correspond to the largest eigenvalues of $X'X$ (or, equivalently the largest singular values of X) are used. If there are say 100 predictor variables there will be 100 eigenvalues but perhaps only the linear combinations corresponding to the first 5 or 10 eigenvalues will be used. Often the selection of the subset of the new variables, i.e. the linear combinations, is done without reference to the values of the Y-variable, thus avoiding the problem of selection bias. It is usual practice to centre and scale the X-variables before calculating the eigenvalues or singular values, particularly if the predictor variables are of different orders of magnitude or have different dimensions. In some practical situations the first few eigenvectors may have some sensible interpretation, for instance, if the X-variables are the same quantity (such as pressure or rainfall) measured at different locations, the first eigenvector may represent a weighted average, the second may be east–west gradient, and the third a north–south gradient.

The derivation of principal components uses only the values of the predictor variables, not those of the variable to be predicted. In general there is no reason why the predictand should be highly correlated with the vectors corresponding to the largest eigenvalues, and it is quite common in practice to find that the vector which is the most highly correlated is one corresponding to one of the smaller eigenvalues. A valuable example has been given by Fearn (1983). This paper was rather controversial, and some of the enthusiasts for ridge regression have published alternative analyses of the data (see e.g. Hoerl, Kennard and Hoerl, 1985 and Naes, Irgens and Martens, 1986).

An alternative approach to principal component regression is to reject those components which make insignificant contributions to the fit to the predictor variable. This is the approach advocated by Jolliffe (1982). Mason and Gunst (1985) compare the advantages and biases of the two criteria for selecting principal components.

There is no need for the linear combinations to be orthogonal; any linear combinations could be used to reduce the dimensionality. The advantages are that the coefficients within the linear combinations are taken as known so that only a small number of parameters, that is the regression coefficients, have to be estimated from the data, and that if there is no selection from among the linear combinations based upon the Y-values then there is no selection bias. The principal disadvantages are that the linear combinations involve all the predictor variables so that they must still all be measured, and the Y-variable may not be well predicted by the chosen linear combinations though there may be some other linear combination which has been rejected which does yield good predictions.

1.4 'Black box' use of best-subsets techniques

The ready availability of computer software encourages the blind use of best-subsets methods. The high speed of computers coupled with the use of efficient algorithms means that it may be feasible to find say the subset of 10 variables out of 150 which gives the closest fit in the LS sense to a set of observed Y-values. This does not necessarily mean that the subset thus chosen will provide good predictions. Throwing in a few more variables produced using a random number generator or from the pages of a telephone directory could have a very salutary effect!

A number of derogatory phrases have been used in the past to describe the practices of subset selection, such as data grubbing, fishing expeditions, data mining (see Lovell, 1983), and torturing the data until they confess. Given a sufficiently exhaustive search, some apparent pattern can always be found, even if all the predictors have come from a random number generator. To the author's knowledge, none of the readily available computer packages at the time of writing makes any allowance for the over-fitting which undoubtedly occurs in these exercises.

Given a large number of variables and hence a very large number of possible subsets from which to choose, the best subset for

prediction may not be the one which gives the best fit to the sample data. In general, a number of the better-fitting subsets should be retained and examined in detail.

If possible, an independent sample should be obtained to test the adequacy of the prediction equation. Alternatively, the data set may be divided into three parts; one part to be used for model selection, the second for the calibration of parameters in the chosen model, and the last part for testing the adequacy of the predictions. In some scientific disciplines, it is the practice, or at least the advocated practice, to divide the data into two parts rather than three. If this is done, the calibration of parameters should not be done using the same part of the data as was used to choose the model. This is because of the substantial biases which can arise in the estimation of parameter values in such cases.

If splitting the data into parts is not possible, then cross-validation provides a poor substitute which can be used instead. These techniques will be discussed in Chapter 6.

The 'traditional' approach to empirical model building has been a progressive one of plotting and/or correlating the Y-variable against the predictor variables one at a time, possibly taking logarithms or using other transformations to obtain approximate linearity and homogeneity of variance, then selecting one variable on some basis, fitting it and repeating the process using the residuals. This type of approach may be feasible when the data are from a well-designed experiment, but can be difficult to apply to observational data when the predictor variables are highly correlated. This type of approach is essentially a form of forward selection, one of the procedures to be discussed in Chapter 3. Forward selection procedures are relatively cheap to apply and easy to use. While forward selection and similar procedures often uncover subsets which fit well, they can fail very badly. The Detroit homicide data used in Chapter 3 provides such a case (see Table 3.14) in which the best-fitting subset of three variables gives a residual sum of squares which is less than a third of that for the subset of three variables found by forward selection. The author has one set of data, not presented here, for which the same ratio is about 90:1. In general, it is gross optimism to hope that an *ad hoc* procedure of adding one variable at a time, and perhaps plotting residuals against everything which comes to mind, will find the best fitting subsets.

A sensible compromise between forward selection and the costly extreme of an exhaustive search is often required. If it is feasible to

carry out an exhaustive search for the best-fitting subsets of say five variables, then an examination of say the best 10 subsets each of three, four and five variables may show two variables which appear in all of them. Those two variables can then be forced into all future subsets and an exhaustive search carried out for the best-fitting subsets of up to seven variables including these two.

It will often not be possible to specify in advance a complete set of variables to be searched for best-fitting subsets, though large numbers of polynomial and interaction terms may be included in the initial set. For instance, if possible trends in time have to be considered then it may be sensible to include interactions between some of the variables and time as a way of allowing for regression coefficients to vary with time. Graphical examination may for instance show that polynomial terms and interactions are being included in most of the better-fitting subsets because the response variable has an asymptotic level, or a threshold level. In such cases a simple nonlinear model may provide a much better predictor, and by explaining a considerable amount of the noise in the data may show up the relationship with other variables more clearly.

When a small number of promising subsets of variables has been found, the fit of these subsets should be examined in detail. Such methods of examination are often graphical. This may show that there are outliers in the data, that some of the observations have far more influence than others, or that the residuals are highly autocorrelated in time. These methods are not described in detail in this monograph. Useful references on this subject are Atkinson (1985), Barnett and Lewis (1978), Baskerville and Toogood (1982), Belsley, Kuh and Welsch (1980), Cook and Weisberg (1982), Cox and Snell (1981), Gunst and Mason (1980), Hawkins (1980) and Weisberg (1980). The user should be cautious though as the techniques of empirical data analysis described in these references are for the case in which the model has been chosen independently of the data. Nothing appears to be known about the properties of LS residuals of best-fitting subsets.

CHAPTER 2

Least-squares computations

2.1 Using sums of squares and products (SSP) matrices

Most modern statistical packages use some form of orthogonal reduction for least-squares (LS) computations, yet such methods are still not widely known or taught. Most of the published algorithms for subset selection in regression use methods based upon sums of squares and products matrices, and so a brief introduction to such methods is given here. Most of this chapter though is devoted to orthogonal reduction methods, and to the properties of orthogonal projections.

Given a set of n observations of k variables, x_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$, the LS coefficients, b_j, $j = 1, 2, \ldots, k$, are obtained by solving the set of $(k + 1)$ 'normal' equations:

$$\begin{aligned}
\textstyle\sum y_i &= nb_0 + b_1 \textstyle\sum x_{i1} + \cdots + b_k \textstyle\sum x_{ik} \\
\textstyle\sum x_{i1} y_i &= b_0 \textstyle\sum x_{i1} + b_1 \textstyle\sum x_{i1}^2 + \cdots + b_k \textstyle\sum x_{i1} x_{ik} \\
&\cdots \qquad\quad \cdots \qquad\quad\quad \cdots \\
\textstyle\sum x_{ik} y_i &= b_0 \textstyle\sum x_{ik} + b_1 \textstyle\sum x_{ik} x_{i1} + \cdots + b_k \textstyle\sum x_{ik}^2
\end{aligned} \tag{2.1}$$

where all of the summations are over i, that is over the observations.

The oldest method for solving the normal equations is that of Gaussian elimination. The method is well known and is described in many elementary texts on numerical methods or on linear regression. If the normal equations must be solved, it is an efficient method provided that the number of equations, $k + 1$, is not large. For k greater than about 15–20, an iterative procedure of the Gauss–Seidel or over-relaxation type will usually be faster. Most modern LS algorithms do not use the normal equations because of the poor accuracy they can give.

It is instructive to look at the first stage of Gaussian elimination. Suppose that we start by eliminating the term b_0 from all except the first equation. For the $(j + 1)$st equation we do this by

subtracting $(\sum x_{ij}/n)$ times the first equation from it. This leaves this equation as

$$\sum x_{ij}y_i - \sum x_{ij}y_i/n = b_1(\sum x_{ij}x_{i1} - \sum x_{ij}\sum x_{i1}/n) + \cdots + b_k(\sum x_{ij}x_{ik} - \sum x_{ij}\sum x_{ik}/n). \quad (2.2)$$

Now it can be readily shown that

$$\sum x_{ij}x_{il} - \sum x_{ij}\sum x_{il}/n = \sum (x_{ij} - \bar{x}_j)(x_{il} - \bar{x}_l), \quad (2.3)$$

where $\bar{x}_j$, $\bar{x}_l$ are the means of the variables X_j and X_l. Hence (2.2) can be written as

$$\sum (x_{ij} - \bar{x}_j)(y_i - \bar{y}) = b_1 \sum (x_{ij} - \bar{x}_j)(x_{i1} - \bar{x}_1) + \cdots + b_k \sum (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k). \quad (2.4)$$

The set of k equations of the form (2.4) for $j = 1, 2, \ldots, k$ constitute what are known as the 'centred' or 'corrected' normal equations.

We note that regression programs have sometimes contained code to perform the centring operation on the left-hand side of (2.3) and then called a Gaussian elimination routine to solve the remaining k centred equations when the routine operating on the original $(k + 1)$ equations would have performed the same centring operations anyway!

The use of the right-hand side of (2.3) as the computational method for centring requires two passes through the data, the first to calculate the means and the second to calculate the cross-products. An alternative method uses progressively updated means as each new observation is added. If we let S_r be a partial cross-product calculating using the first r observations, then it can be shown that

$$S_{r+1} = \sum_{i=1}^{r+1} (x_i - \bar{x}_{r+1})(y_i - \bar{y}_{r+1}) = S_r + \delta_x \delta_y \cdot r/(r + 1), \quad (2.5)$$

where

$$\delta_x = x_{r+1} - \bar{x}_r$$
$$\delta_y = y_{r+1} - \bar{y}_r$$

and $\bar{x}_r$, $\bar{y}_r$, $\bar{x}_{r+1}$, $\bar{y}_{r+1}$ are the means of x and y from the first r or $r + 1$ observations. The means are updated using

$$\bar{x}_{r+1} = \bar{x}_r + \delta_x/(r + 1).$$

In the form (2.5), two multiplications are required for each observation for each pair of variables. This can be halved by using the deviation from the new mean for one of the two variables. Thus if

$$\delta_x^* = x_{r+1} - \bar{x}_{r+1}$$

then

$$S_{r+1} = S_r + \delta_x^* \delta_y.$$

This updating formula appears to be due to Jennrich (1977, equation 16 on p. 64). A review of other updating formulae has been given by Chan, Golub and LeVeque (1983), who give a method of their own, similar to the fast Fourier transform, which gives good accuracy when the sample size is an exact power of 2. See also Miller (1989).

The use of the right-hand side of (2.3) or of progressive updating (2.5) for calculating the centred equations can give very substantial improvements in accuracy, but only when fitting regressions which include an intercept. A classic example is that given by Longley (1967). He examined the regression of employment against six other variables for the years 1947–62. Table 2.1 gives the numbers of accurate digits in the coefficients of the centred normal equations as calculated in single precision using a computer which allocates 32 binary bits to floating-point numbers which are thus stored to an accuracy of about 7 decimal digits. The numbers of accurate digits were taken as $-\log_{10}|(e/x)|$, where x is the true value and e is the error. The number of accurate digits was taken as 7·2 when the error was zero or when the logarithm exceeded 7·2. In this case, the use of either the right-hand side of (2.3) or of progressive updating of the means (2.4) has resulted in a centred SSP-matrix with very little loss of accuracy. Regression coefficients calculated using the left-hand side of (2.3) yielded no accurate digits in any of the seven coefficients using Gaussian elimination, in fact five of the seven had

Table 2.1

Formula used	*Range of accurate digits*
Left-hand side of (2.3)	3·1–7·2
Right-hand side of (2.3)	6·7–7·2
Progressive updating (2.5)	6·0–7·2

the wrong signs. Using either of the other centring methods gave between 2·6 and 4·7 accurate digits for the different coefficients.

Writing the equations (2.1) in matrix form gives

$$X'\mathbf{Y} = X'X\mathbf{b} \tag{2.6}$$

where X is an $n \times (k+1)$ matrix with a column of 1's as its first column and such that the element in row i and column $(j+1)$ is the ith value of the variable X_j; $\mathbf{Y}$ is a vector of length n containing the observed values of the variable Y, and $\mathbf{b}$ is the vector of $(k+1)$ regression coefficients with the intercept b_0 as its first element. The matrix $X'X$ is then the SSP-matrix.

This formulation tempts the use of matrix inversion to obtain $\mathbf{b}$ using

$$\mathbf{b} = (X'X)^{-1}X'\mathbf{Y}. \tag{2.7}$$

If only the regression coefficients are required, this involves considerably more computation than using Gaussian elimination. However the covariance matrix of $\mathbf{b}$ is given by $\sigma^2(XX)^{-1}$ so that the inverse is usually needed.

The SSP-matrix is symmetric and positive definite, or at worst positive semi-definite, and there are extremely efficient ways to invert such matrices. Two of the popular methods are the so-called Gauss–Jordan method (though it was probably unknown to either Gauss or Jordan – it first appeared in a posthumous edition of a book by Jordan), and the Cholesky factorization method. The Cholesky factorization can also be used in a similar way to Gaussian elimination to obtain the regression coefficients by back-substitution. Both of these methods can be executed using the space required for either the upper or lower triangle of the matrix, with the inverse overwriting the original matrix if desired. Code for the Gauss–Jordan method is given in Wilkinson and Reinsch (1971, pp. 45–9), Garside (1971a) and Nash (1979, pp. 82–5). Code for the Cholesky factorization is given in Wilkinson and Reinsch (1971, pp. 17–21) and Healy (1968a, b), though it can easily be coded from Stewart (1973, algorithm 3·9 on p. 142). From error analyses, the Cholesky method should be slightly more accurate (see e.g. Wilkinson, 1965, pp. 244–5), but in an experiment to compare the Gauss–Jordan and Cholesky methods by Berk (1978a), the difference was barely detectable.

The Cholesky method requires the calculation of square roots. Square roots can be calculated very quickly in binary arithmetic

using a method which is often taught in schools for calculating square roots in the scale of ten. If we are trying to find the square root of a number y and have so far found a number x containing the first few digits or binary bits, then we attempt to find a δ such that

$$(x+\delta)^2 \approx y$$

whence

$$\delta \approx (y-x^2)/(2x+\delta).$$

That is, we divide the current remainder $(y-x^2)$ by twice x plus the next binary bit δ. The particular advantages in the scale of two are that the doubling operation simply requires a shift, and the next bit, δ, can only be 0 or 1 so that the division operation is simply a test of whether $(2x+\delta)$, when shifted the correct number of places, is greater than or equal to the remainder. As the divisor, $2x+\delta$, starts with only one binary bit and averages only half the number in the mantissa, the method is about twice as fast as a floating-point division. Unfortunately most computers use a different method to take advantage of hardware for division. This method usually uses two Newton–Raphson iterations and often gives errors in the last binary bit. This could explain Berk's findings on the relative accuracy of the Gauss–Jordan and Cholesky methods.

The Cholesky method uses the factorization

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{L}\boldsymbol{L}'$$

where $\boldsymbol{L}$ is a lower-triangular matrix. An alternative factorization, which is sometimes credited to Banachiewicz (1938), which avoids the calculation of square roots is

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}', \tag{2.8}$$

where $\boldsymbol{L}$ is lower-triangular with 1's on its diagonal, and $\boldsymbol{D}$ is a diagonal matrix. In computations, the elements on the diagonal of $\boldsymbol{D}$ are stored overwriting the diagonal elements of $\boldsymbol{L}$. This can be expected to be slightly more accurate than the Gauss–Jordan method and as efficient in terms of both speed and storage requirements. Code for forming the factorization (2.8) is given in Wilkinson and Reinsch (1971, pp. 21–4).

An advantage of the Cholesky or Banachiewicz methods over the Gauss–Jordan method in some situations is that the triangular factorizations can easily be updated when more data become

available. A method for doing this will be presented in section 2.2. If the Gauss–Jordan method has been used, the inverse matrix $(X'X)^{-1}$ can be updated using

$$A^* = A - (A\mathbf{x}\mathbf{x}'A)/(1 + \mathbf{x}'A\mathbf{x}), \tag{2.9}$$

where A is the old inverse and A^* is the updated inverse after an extra observation, x, has been added. Unless the inverse matrix is required after every new observation is added, this method is slow and can give poor accuracy. If only the regression coefficients are required after each observation is added, then the regression coefficients can be obtained quickly from a triangular factorization using back-substitution without the need for matrix inversion. The update formula (2.9) is usually credited to Plackett (1950) or Bartlett (1951), though it has been suggested by Kailath (1974) that the method was known to Gauss.

So far we have looked at some aspects of LS computations when all of the variables are to be included in the model. In selecting a subset of variables we will want to perform calculations for a number of different subsets, sometimes a very large number of them, to find those which give good fits to the data. Several ways of doing this will be described in Chapter 3. In choosing the computational procedure, we will need both speed and accuracy. Accuracy is particularly important for two reasons:

1. Subset selection procedures are often used when there is a choice among several fairly highly correlated variables which are attempting to measure the same attribute, and in such cases the normal equations can be badly ill-conditioned.
2. Some of the procedures for searching for best-fitting subsets require a very large number of arithmetic operations to be performed and we need to be sure that rounding errors accumulate as slowly as possible.

The emphasis will be upon calculating the residual sum of squares for each subset investigated, not upon regression coefficients which can be found later for the small number of subsets singled out for closer scrutiny. We will want to obtain the residual sum of squares for a subset with the smallest amount of additional computation from calculations already carried out for previous subsets. We shall see later that methods operating on the SSP- matrix (or equivalently on the correlation matrix) and its inverse sometimes have a slight

speed advantage over methods based around triangular factorization which will be described in section 2.2.

However, these latter methods have a very definite advantage in accuracy, particularly if the starting triangulation is computed accurately using either orthogonal reduction or extra precision. On many computers, two levels of precision are available, a single precision which often represents floating-point numbers to about 7–8 significant decimal digits, and a double precision which represents them to say 16–18 decimal digits. The accuracy advantage is such that it is often feasible to perform the search for best-fitting subsets in single precision using the methods based around triangular factorizations when double precision is necessary using SSP-matrices and Gauss–Jordan methods. When a floating-point processor is not available, double-precision calculations are usually several times slower.

2.2 Orthogonal reduction methods

The basic idea here is to find an orthogonal basis in which to express both the X- and Y-variables, to perform regression calculations in this basis, and then to transform back to obtain regression coefficients in the dimensions of the real problem. In most of the orthogonal reduction methods, the matrix $\boldsymbol{X}$ of n observations of each of k variables, $X_1, X_2, \ldots, X_k$ (where X_1 will be identically equal to one in most practical cases), is factored as

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}, \tag{2.10}$$

where either

1. $\boldsymbol{Q}$ is an $n \times k$ matrix and $\boldsymbol{R}$ is a $k \times k$ upper-triangular matrix, or
2. $\boldsymbol{Q}$ is an $n \times n$ matrix and $\boldsymbol{R}$ is an $n \times k$ matrix containing an upper-triangular matrix in its first k rows and zeros elsewhere.

In either case, the columns of $\boldsymbol{Q}$ are orthogonal and usually normalized so that

$$\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}$$

where $\boldsymbol{I}$ is either a $k \times k$ or an $n \times n$ identity matrix. The principal methods which use the factorization (2.10) are the Gram–Schmidt, Householder reduction, and various methods using planar rotations (also known as Jacobi or Givens rotations). A readable general introduction to these alternative methods is contained in

Seber (1977, Chapter 11). Code for the modified Gram–Schmidt method has been given by Farebrother (1974), Wampler (1979a, b) and Longley (1981), and for the Householder method by Lawson and Hanson (1974, algorithm HFTI on pp. 290–1), Lawson *et al.* (1979) and in the LINPACK package (Dongarra *et al.*, 1979).

An alternative type of orthogonal reduction method uses the singular-value decomposition (s.v.d.) in which $\boldsymbol{X}$ is factored as

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}', \tag{2.11}$$

where $\boldsymbol{U}$ is $n \times k$, $\boldsymbol{\Lambda}$ is a $k \times k$ diagonal matrix with the singular values along its diagonal, $\boldsymbol{V}$ is $k \times k$, and $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{V}'\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}' = \boldsymbol{I}$. Then

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{V}\boldsymbol{\Lambda}^2\boldsymbol{V}',$$

which provides a quick and accurate way of calculating principal components without first forming the SSP-matrix. In most statistical work the matrix $\boldsymbol{U}$ is not required, and in such cases $\boldsymbol{\Lambda}$ and $\boldsymbol{V}$ can conveniently be obtained starting from the factorization (2.10).

Principal components are usually formed after subtracting the means from each variable. If the orthogonal reduction (2.10) has been formed from a matrix $\boldsymbol{X}$ which had a column of 1's as its first column, then the means are conveniently removed simply by leaving out the top row of $\boldsymbol{R}$. Similarly, the X-variables can be scaled to have unit sum of squares about the mean by scaling each column of $\boldsymbol{R}$, after removing the top row, so that the sum of squares of elements in each column is 1. The modified version of $\boldsymbol{R}$ is then used in place of $\boldsymbol{X}$ in (2.11) to obtain the same principal components which are usually obtained (with poor accuracy) from the correlation matrix. As the matrix $\boldsymbol{R}$ has only $(k-1)$ rows, which is usually far fewer than the n rows of $\boldsymbol{X}$, the s.v.d. calculation is fairly fast; the bulk of the computational effort goes into calculating the orthogonal reduction (2.10). Ths s.v.d. can be used for multiple regression work but is usually somewhat slower than the other orthogonal reductions. Code for the s.v.d. is given in Chan (1982), Lawson and Hanson (1974, pp. 295–7), LINPACK (Dongarra *et al.*, 1979), Nash (1979, pp. 30–1) and Wilkinson and Reinsch (1971, pp. 134–51).

If we write $\boldsymbol{R} = \{r_{ij}\}$, then from (2.10) we have

$$\begin{aligned}
\boldsymbol{X}_1 &= r_{11}\mathbf{Q}_1 \\
\boldsymbol{X}_2 &= r_{12}\mathbf{Q}_1 + r_{22}\mathbf{Q}_2 \\
\boldsymbol{X}_3 &= r_{13}\mathbf{Q}_1 + r_{23}\mathbf{Q}_2 + r_{33}\mathbf{Q}_3, \text{ etc.}
\end{aligned}$$

Thus $\mathbf{Q}_1$ spans the space of X_1; $\mathbf{Q}_1$ and $\mathbf{Q}_2$ span the space of X_2, where $r_{22}\mathbf{Q}_2$ is that component of X_2 which is orthogonal to X_1, etc. We notice also that

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{R}'\boldsymbol{R},$$

that is that $\boldsymbol{R}$ is the upper triangle of the Cholesky factorization. SSP-matrices can thus be constructed from the $\boldsymbol{R}$-matrix if needed. If we omit the first row and column of $\boldsymbol{R}$, the remaining coefficients give the components of $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, etc. which are orthogonal to $\boldsymbol{X}_1$, in terms of the direction vectors $\mathbf{Q}_2, \mathbf{Q}_3$, etc. If X_1 is a column of 1's then $\boldsymbol{R}'_{-1}\boldsymbol{R}_{-1}$ gives the centred SSP-matrix, where $\boldsymbol{R}_{-1}$ is $\boldsymbol{R}$ after removal of the first row and column. Correlations between variables can then be calculated from the SSP-matrix. Similarly, by removing say the first three rows and columns of $\boldsymbol{R}$, the matrix of partial correlation among $X_4, X_5, \ldots, X_k$ can be obtained after regressing out X_1, X_2 and X_3.

The orthogonal reduction (2.10) can be achieved by multiplying $\boldsymbol{X}$ on the left by a series of orthonormal matrices, each one chosen to reduce one or more elements of $\boldsymbol{X}$ to zero. One such type of matrix is the planar rotation matrix

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix}$$

where c and s can be thought of as cosine and sine, as for the matrix to be orthonormal we require that $c^2 + s^2 = 1$. For instance, if we want to reduce the element y to zero by operating on the two rows of the matrix below, we choose $c = w/(w^2 + y^2)^{1/2}$ and $s = y/(w^2 + y^2)^{1/2}$, then

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix}\begin{pmatrix} w & x & \cdots \\ y & z & \cdots \end{pmatrix} = \begin{pmatrix} (w^2 + y^2)^{1/2} & cx + sz & \cdots \\ 0 & -sx + cz & \cdots \end{pmatrix}.$$

The full planar rotation matrix looks like

$$\begin{pmatrix} 1 & & & \\ & c & & s \\ & & 1 & \\ & -s & & c \end{pmatrix}$$

where the blanks denote zero values, for the rotation of rows 2 and 4 of a 4×4 matrix.

By applying planar rotations to one additional row of $\boldsymbol{X}$ at a time, the factorization (2.10) can be achieved requiring the storage of only one row of $\boldsymbol{X}$ and the k rows of $\boldsymbol{R}$ at any time. It is possible to eliminate the calculation of square roots by producing the Banachiewicz factorization instead of the Cholesky factorization. Efficient algorithms for this have been given by Gentleman (1973) and Hammarling (1974). Using this type of method, $\boldsymbol{Q}'$ is the product of the planar rotations as the product is such that $\boldsymbol{Q}'\boldsymbol{X} = \boldsymbol{R}$. The matrix $\boldsymbol{Q}$ is not usually formed explicitly, though the c's and s's can be stored if needed.

The linear transformations applied to the X-variables are simultaneously applied to the Y-variable giving a vector $\boldsymbol{Q}'\mathbf{Y}$. The vector of values of Y can be added to $\boldsymbol{X}$ as an extra column if wished, though that method will not be used here. If the orthogonal reduction is of the kind for which $\boldsymbol{Q}$ is an $n \times k$ matrix, e.g. the modified Gram–Schmidt method, then $\boldsymbol{Q}'\mathbf{Y}$ is of length k, and there is an associated residual vector which is orthogonal to the columns of $\boldsymbol{Q}$ and hence orthogonal to the X-space. If the reduction method is of the kind for which $\boldsymbol{Q}$ is an $n \times n$ matrix, e.g. Householder reduction or planar rotation methods, then $\mathbf{Q}'\mathbf{Y}$ is of length n. In either case, the first k elements of $\mathbf{Q}'\mathbf{Y}$ are the projections of Y in the directions $\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_k$.

The residuals in the last $(n-k)$ elements of $\mathbf{Q}'\mathbf{Y}$ from either the Householder reduction method or the planar rotation method can be shown to be uncorrelated and to be homogeneous, i.e. to have the same variance, if the true but unknown residuals from the model also have these properties. This is a property which LS residuals do not have, and it can be useful in model testing. The residuals from using Householder reduction are known as LUSH (linear unbiased with scalar, covariance Householder) and are discussed by Grossman and Styan (1972), Ward (1973) and Savin and White (1978). The residuals from using planar rotations have been shown by Farebrother (1978) to be identical to the 'recursive' residuals of Brown, Durbin and Evans (1975) and are much more readily calculated using planar rotations than by using the elaborate method given by them.

If we let r_{iy} denote the ith element of $\mathbf{Q}'\mathbf{Y}$, then

$$\mathbf{Y} = r_{1y}\mathbf{Q}_1 + r_{2y}\mathbf{Q}_2 + \cdots + r_{ky}\mathbf{Q}_k + \mathbf{e} \tag{2.12}$$

where $\mathbf{e}$ is the vector of residuals orthogonal to the directions

$\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_k$. The r_{iy}'s are therefore the LS regression coefficients of Y upon the $\mathbf{Q}$'s. Using (2.12), we can substitute for the $\mathbf{Q}$'s in terms of the variables of interest, that is the X's. In matrix notation, (2.12) is

$$\mathbf{Y} = \boldsymbol{Q}(\boldsymbol{Q}'\mathbf{Y}) + \mathbf{e}$$

where $\boldsymbol{Q}$ is an $n \times k$ matrix. Substituting for $\boldsymbol{Q}$ from (2.10) then gives

$$\begin{aligned}\mathbf{Y} &= \boldsymbol{X}\boldsymbol{R}^{-1}\boldsymbol{Q}'\mathbf{Y} + \mathbf{e} \\ &= \boldsymbol{X}\mathbf{b} + \mathbf{e}\end{aligned}$$

where $\mathbf{b}$ is the vector of regression coefficients of Y upon X. Hence $\mathbf{b}$ can be calculated from

$$\mathbf{b} = \boldsymbol{R}^{-1}\boldsymbol{Q}'\mathbf{Y}$$

or more usually by back-substitution in

$$\mathbf{R}\mathbf{b} = \boldsymbol{Q}'\mathbf{Y}. \tag{2.13}$$

The formula (2.13) can also be obtained by substitution in the 'usual' formula (2.7). By using only the first p equations we can quickly obtain the regression coefficients of $\mathbf{Y}$ upon the subset $X_1, X_2, \ldots, X_p$.

The break-up of the sum of squares is readily obtained from (2.12). The total sum of squares of Y is

$$\mathbf{Y}'\mathbf{Y} = r_{1y}^2 + r_{2y}^2 + \cdots + r_{ky}^2 + \mathbf{e}'\mathbf{e}.$$

If the first variable, X_1, is just a column of 1's then the total sum of squares of Y about its mean is

$$r_{2y}^2 + r_{3y}^2 + \cdots + r_{ky}^2 + \mathbf{e}'\mathbf{e}.$$

The residual sum of squares after regressing Y against $X_1, X_2, \ldots, X_p$ is

$$r_{p+1,y}^2 + \cdots + r_{ky}^2 + \mathbf{e}'\mathbf{e}.$$

Thus from the factorization (2.10), and the associated vector $\mathbf{Q}'\mathbf{Y}$, the set of k sequential regressions Y upon X_1; Y upon X_1 and X_2; Y upon $X_1, X_2, \ldots, X_k$ can be carried out very quickly. In the situation in which there is a hierarchical order for carrying out a set of regressions, such as when fitting polynomials, the various methods of orthogonal reduction provide a fast computational method.

Now suppose that we want to regress Y against a subset of the predictor variables such as X_1, X_3 and X_4 but excluding X_2. We can rearrange the columns of $\boldsymbol{X}$ and $\boldsymbol{Q}$ in (2.10), and both the rows and columns of $\boldsymbol{R}$. Unfortunately the rearranged matrix $\boldsymbol{R}$, let us call it $\boldsymbol{R}_T$, is no longer triangular. We have then

$$X_1X_3X_4X_2 = \mathbf{Q}_1\mathbf{Q}_3\mathbf{Q}_4\mathbf{Q}_2\boldsymbol{R}_T$$

where $\boldsymbol{R}_T$ looks like

$$\begin{matrix} \times & \times & \times & \times \\ & \times & \times & \\ & & \times & \\ & \times & \times & \times \end{matrix}$$

where an $\times$ denotes a nonzero element. A more serious problem is that the orthogonal directions Q_1, Q_3 and Q_4 do not form a basis for X_1, X_3 and X_4 as both X_3 and X_4 have components in the direction Q_2 which is orthogonal to this basis. A solution to this problem is to use planar rotations to restore $\boldsymbol{R}_T$ to upper-triangular form. By operating upon rows 2 and 4, the element in position (4, 2) can be reduced to zero. Other elements in these two rows will be changed, in particular a nonzero element will be introduced into position (2, 4). Another planar rotation applied to rows 3 and 4 removes the element in position (4, 3) and the matrix is in upper-triangular form. Let $\boldsymbol{P}$ be the product of these two planar rotations, so that $\boldsymbol{P}\boldsymbol{R}_T$ is the new triangular matrix. Then

$$X_1X_3X_4X_2 = (\mathbf{Q}_1\mathbf{Q}_3\mathbf{Q}_4\mathbf{Q}_2\boldsymbol{P}')(\boldsymbol{P}\boldsymbol{R}_T),$$

so that the first three columns of $(\mathbf{Q}_1\mathbf{Q}_3\mathbf{Q}_4\mathbf{Q}_2\boldsymbol{P}')$ form the new orthogonal basis. This method appears to have been used first by Elden (1972); it is described in more detail by Clarke (1980). The first publication of the method appears to be in Osborne (1976).

Software for changing the order of variables can be found in Osborne (1976), in the LINPACK package Dongarra *et al.* (1979), and in Clarke (1981).

It is possible to generate subset regressions from the s.v.d. as well as from triangular factorizations, but the process of adding or deleting a variable changes the entire factorization, not just one or two rows.

2.3 Gauss–Jordan v. orthogonal reduction methods

We have said that the orthogonal reduction methods, which lead to operations on triangular matrices, are much more accurate than methods which require the inversion of parts of an SSP-matrix. Why is this? Let e_i be used to denote the ith LS residual, then

$$e_i = (y_i - \bar{y}) - \sum_{j=1}^{p} b_j(x_{ij} - \bar{x}_j) \tag{2.14}$$

where the b_j's are the LS estimates of the regression coefficients and it is assumed that we are fitting models containing a constant; $\bar{x}_i, \bar{y}$ denote the sample means of the appropriate variables. Suppose that the e_i's are calculated using (2.14) with each quantity correctly rounded to t decimal digits. Now let us suppose that the e_i's are of a smaller order of magnitude than the $(y_i - \bar{y})$'s, which will be the case if the model fits the data closely. The order of magnitude could be defined as say the average absolute value of the quantities, or as their root-mean square, or as some such measure of their spread. Then if the e_i's are of order 10^{-d} times the order of the $(y_i - \bar{y})$'s, the e_i's will be accurate to about $(t - d)$ decimal digits, and the sum $\sum e_i^2$ will have a similar accuracy. This loss of d decimal digits is because of the cancellation errors which occur in performing the subtraction in (2.14). Suppose, for instance, that we are working to seven decimal digits, then the right-hand side of (2.14) for one observation might be

$$3333333. - 3311111.$$

giving a difference $e_i = 22222$. If the two quantities were correctly rounded then the maximum error in each is 0·5. Hence the maximum error in e_i is 1·0 or a possible error of one in the fifth decimal digit. This is the principal source of error in orthogonal reduction methods in which we work in the scale of the original variables, not in the scale of their squares as with SSP-matrices.

In using SSP-matrices, residual sums of squares are usually calculated as

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_j b_j \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(y_i - \bar{y}). \tag{2.15}$$

Now if $\sum e_i^2$ is of the order of 10^{-2d} times the order of $\sum (y_i - \bar{y})^2$, and all of the terms on the right-hand side of (2.15) are correctly rounded

to t decimal digits, then $2d$ digits in each of the terms on the right-hand side must cancel giving an accuracy of $(t-2d)$ digits for the residual sum of squares. Thus we can expect to lose about twice as many digits of accuracy using methods based upon SSP-matrices as we lose using orthogonal reduction methods.

The notation $(1-R^2)$ is commonly used for the ratio $\sum e_i^2/\sum(y_i-\bar{y})^2$. Then if $R^2=0{\cdot}9$ we can except to lose one decimal digit due to cancellation errors when using SSP-matrices. We can expect to lose two digits if $R^2=0{\cdot}99$, three if $R^2=0{\cdot}999$, etc. This assumes that $\sum(y_i-\bar{y})^2$ was calculated accurately in the first place and, as was shown in section 2.1, this can also be a source of serious error. These are lower limits to the loss in accuracy as they ignore errors in the b_j's, which can be substantial if there are high correlations among the X-variables.

Now let us look at how this applies to regressions involving subsets of variables. Let us denote by a_{ij} the element in row i and column j of the SSP-matrix. In adding variable number r into the regression we calculate new elements, a_{ij}^*, usually overwriting the old ones, using

$$a_{ij}^* = a_{ij} - a_{ir}a_{rj}/a_{rr} \qquad \text{for all } i, j \neq r,$$
$$a_{ir}^* = -a_{ir}/a_{rr} \qquad \text{for } i \neq r,$$

and

$$a_{rr}^* = 1/a_{rr}.$$

The elements, a_{ii}^*, along the diagonal are of particular interest. Initially a_{ii} is the sum of squares for variable X_i, which may have been centred if a constant is being fitted. (a_{ir}/a_{rr}) is the regression coefficient for the regression of variable X_i on variable X_r which we will denote by b_{ir}. Then

$$a_{ii}^* = a_{ii} - b_{ir}a_{ir} \qquad \text{for } i \neq r. \tag{2.16}$$

This has the same form as (2.15) as b_{ir} and a_{ir} have the same sign. Hence if the correlation coefficient between X_i and X_r is $\pm 0{\cdot}95$ (i.e. $R^2 \approx 0{\cdot}9$), one decimal digit will be lost in cancellation errors in calculating a_{ii}^*, two digits will be lost if the correlation is $\pm 0{\cdot}995$, etc. This means that high correlations among the predictor variables can lead to losses of one or more digits in calculating the second term on the right-hand side of (2.15) so that the residual sum of squares may be of low accuracy even when the Y-variable is not well predicted by the X-variables. Again these losses are roughly

halved when working in the scale of the X-variables, e.g. with triangular factorizations, rather than with sums of squares and products.

Planar rotations are stable transformations which lead to little build-up in error when used repeatedly. Suppose for instance that we need to calculate

$$x = uc + vs$$

where c, s are such that $c^2 + s^2 = 1$, and that the values of u and v have been calculated with errors δu and δv respectively. Then, neglecting errors in c and s, the error in the calculated x is

$$\delta x = c\,\delta u + s\,\delta v$$

plus any new rounding error in this calculation. The maximum value of δx is then $\{(\delta u)^2 + (\delta v)^2\}^{1/2}$ and occurs when $c = \pm\,\delta u/\{(\delta u)^2 + (\delta v)^2\}^{1/2}$, though in most cases it will be much smaller and often the two components of δx will have opposite signs and partially cancel each other. Thus the absolute size of the errors will remain small, though the relative errors may be large when the x resulting from calculations such as this is very small. Thus we can anticipate only little build-up of error when planar rotations are used repeatedly to change the order of variables in triangular factorizations. There is no such result limiting the size of build-up of errors in the inversion and reinversion of parts of matrices in the Gauss–Jordan method.

The above is a heuristic discussion of the errors in LS calculations. Detailed error analyses have been given for one-off LS calculations by Golub (1969, particularly pp. 382–5), Jennings and Osborne (1974), Gentleman, W.M. (1975) and Stewart (1977), among others. The basic building blocks for these analyses were given by Wilkinson (1965) on pp. 209–17 for Gaussian elimination, and on pp. 131–9 for planar rotations.

To investigate the build-up of errors in a subset selection procedure, the Garside (1971b, c) algorithm for generating all subsets of variables using a Gauss–Jordan method, and an equivalent routine using planar rotations were compared with respect to both speed and accuracy in single precision on machines which use 32 binary bits for floating-point numbers and hence store numbers with about seven significant decimal digits. In each case the appropriate starting matrix, the upper triangle of the SSP-matrix for the Garside algorithm and the upper triangular matrix from an orthogonal

reduction for the planar rotation algorithm, were calculated in double precision then rounded to single precision. Generating all possible subsets is usually not very sensible in practice; it was chosen here as an extreme example of a subset selection procedure requiring the maximum amount of computational effort and hence giving the slowest speeds and poorest accuracy.

The Garside algorithm was programmed in FORTRAN with the upper triangle of the SSP-matrix stored as a singly dimensioned array to minimize the effort required in evaluating storage addresses. The tests for singularities and the facility for the grouping of variables were omitted.

Many alternative ways of coding the planar rotation algorithm were considered. First, two alternative orders for producing the subsets were examined, namely the Hamiltonian cycle, as used by Garside (1965) and Schatzoff *et al.* (1968), and the binary sequence as used by Garside (1971b, c). Unless otherwise stated, the 'Garside algorithm' will be used to refer to his 1971 algorithm.

The Hamiltonian cycle can be thought of as a path joining all of the corners of a hypercube in which each step is to a neighbouring corner. If the corners of a three-dimensional cube with side of unit length are at $(0,0,0)$, $(0,0,1),\ldots,(1,1,1)$, then there are many Hamiltonian cycles of which one is

000 001 011 010 110 111 101 100

In applying this to subset selection, a '1' or '0' indicates whether a variable is in or out of the subset. The attractive feature of the Hamiltonian cycle is that only one variable is operated upon at each step. In the binary sequence, the string of 0's and 1's is treated as a binary number to which one is added at each step. Thus a sequence for three variables is

000 001 010 011 100 101 110 111

This sequence often requires several variables to be added or removed from the subset in one step. However, as the right-hand end digits change rapidly, these can be made to correspond to the bottom corner of the triangular matrix, in either the Garside or the planar rotation algorithm, where there are fewer elements to be operated on than in the upper rows. More details of these alternatives are contained in Appendix 2A, where it is shown that for a moderate number, k, of X-variables, the number of multiplications or divisions

per subset is about $(10 + k)$ for the Hamiltonian cycle and 15.75 per subset for the binary sequence. In contrast, Garside's algorithm requires 14 operations (multiplications or divisions) per subset (not 8 as stated by Garside). Both the Hamiltonian cycle and the binary sequence have been used with the planar rotation algorithm in the comparisons which follow.

The other major alternative considered was in the treatment of the response variable Y. It could either be left as the last variable or it could be allowed to move in the triangular matrix being placed immediately after the last variable included in the current subset. It was found to be much faster with the Hamiltonian cycle to leave it fixed as the last variable as otherwise two rows of the triangular matrix had to be swapped at each step instead of one. The column corresponding to the Y-variable, $\boldsymbol{Q}'\mathbf{Y}$ in our earlier notation, was stored as a separate vector as also were the progressive residual sums of squares, that is the quantities $\mathbf{e}'\mathbf{e}, \mathbf{e}'\mathbf{e} + r_{ky}^2, \mathbf{e}'\mathbf{e} + r_{ky}^2 + r_{k-1,y}^2$, etc.

Another alternative considered was whether to use the Gentleman (1973) algorithm or the Hammarling (1974) algorithm for the planar rotations. The Gentleman algorithm is usually slower but can be applied so that only small sub-triangles within the main triangular matrix need to be operated upon when a binary sequence is used to generate the subsets. The Hammarling algorithm has a disadvantage in that it requires occasional rescaling to prevent its row multipliers becoming too small. In the following comparisons, both the Hammarling and Gentleman algorithms have been used, the first with the Hamiltonian cycle and the second with the binary sequence.

For the comparisons of accuracy, five data sets were used. These are briefly summarized in Table 2.2. The WAMPLER data set is an artificial set which was deliberately constructed to be very ill-conditioned; the other data sets are all real and were chosen to give a range of numbers of variables and to give a range of ill-conditioning such as is often experienced in real problems.

Table 2.3 shows measures of the ill-conditioning of the SSP-matrices. If we denote the eigenvalues of the correlation matrix by λ_i's, then the ratio of the largest to the smallest, $\lambda_{\max}/\lambda_{\min}$, is often used as a measure of the ill-conditioning of the SSP-matrix. Berk (1978a) compared a number of measures of ill-conditioning and found that for matrix inversion the accuracy of the inverse matrix was most highly correlated with the trace of the inverse matrix,

Table 2.2 *Summary of data sets used*

Data set name	*Source*	*k = no. of X-variables*	*n = no. of observations*
WAMPLER	Wampler (1970) using his *Y*3	5	21
LONGLEY	Longley (1967), *Y* = total derived employment	6	16
STEAM	Draper and Smith (1981, p. 616)	9	25
DETROIT	Gunst and Mason (1980). Set *A*3 on p. 360. Origin, Fisher (1976)	11	13
POLLUTE	Gunst and Mason (1980). Set *B*1 on pp. 370–1. Origin, McDonald and Schwing (1973)	15	60

Table 2.3 *Measures of ill-conditioning of the test data sets*

	$\lambda_{max}/\lambda_{min}$		$\sum(1/\lambda_i)$	
Data set	*X only*	*X and Y*	*X only*	*X and Y*
WAMPLER	600 000	700 000	130 000	130 000
LONGLEY	12 000	6 000	3 100	2 200
STEAM	800	1 100	290	350
DETROIT	7 000	4 200	1 500	1 500
POLLUTE	900	1 000	260	280

which is $\sum(1/\lambda_i)$. Both of these measures are shown in the table, first for the matrix of correlations among the *X*-variables only and then with the *Y*-variable added.

Table 2.4 shows the lowest and average numbers of accurate decimal digits in the residual sums of squares for all subsets of variables for the Gauss–Jordan (GJ) and the two planar rotation algorithms. The calculations were performed using a CROMEMCO Z2-D microcomputer using Microsoft FORTRAN version 3.37 and the CDOS operating system. As mentioned earlier, the centred SSP-matrix and orthogonal reduction were calculated in double precision (equivalent to about 17 decimal digits) and then rounded to single precision. Some calculations were repeated on a PDP11-34 with identical results.

Table 2.4 *Lowest and average numbers of accurate decimal digits in the calculation of residual sums of squares for all subsets*

Data set	*Gauss–Jordan*		*Hamiltonian cycle*		*Binary sequence*	
	Lowest	*Average*	*Lowest*	*Average*	*Lowest*	*Average*
WAMPLER	1·6	3·5	4·9	6·6	5·0	6·4
LONGLEY	3·6	5·5	6·1	6·9	6·3	7·2
STEAM	4·4	5·5	5·4	6·6	5·6	6·6
DETROIT	2·0	5·0	4·9	6·1	4·5	6·2
POLLUTE	3·6	4·3	5·0	6·0	4·5	5·5

The performance of the planar rotation algorithms was very impressive in terms of accuracy. The LONGLEY data set is often used as a test of regression programs and we see that in the worst case only about one decimal digit of accuracy was lost; in fact the accuracy was calculated, using the logarithm to base 10 of the relative accuracy, to be at least 7·0 for 22 out of the 63 residual sums of squares using the Hamiltonian cycle and the Hammarling rotations, and even better for the binary sequence and Gentleman rotations. The POLLUTE data set required a large amount of computation yet there was very little sign of build-up of errors using planar rotations; for the last 100 subsets generated (out of 32 767), the lowest accuracy was 5·3 decimal digits and the average was 5·9 using the Hamiltonian cycle. In contrast, the Gauss–Jordan algorithm performed poorly on the very ill-conditioned WAMPLER data set, and on the moderately ill-conditioned DETROIT data set, and on the well-conditioned POLLUTE data set. On the POLLUTE data set a steady build-up of errors was apparent. The last 100 of the subsets generated contained none with an accuracy greater than 4·0 decimal digits.

How much accuracy is needed? A subset selection procedure is used to pick a small number of subsets of variables which give small residual sums of squares. Regression coefficients and other quantities can be calculated later using a saved, accurate copy of the appropriate matrix. In most practical cases an accuracy of two decimal digits, equivalent to errors of 1%, will be quite adequate as this is usually less than the difference needed between two or more subsets to be statistically significant. Hence, provided that the initial SSP-matrix is calculated accurately, the Gauss–Jordan algorithm is adequate

for all of these subsets except the artificial WAMPLER set, using a machine, such as the CROMEMCO or PDP11-34, which allocates 32 binary bits to floating-point numbers and performs correctly rounded arithmetic. However some computers perform truncated arithmetic. Limited experience with one such computer showed an alarming build-up of errors for the Gauss–Jordan algorithm but only slightly worse performance for the planar rotation algorithm.

The Gauss–Jordan method will usually give just acceptable accuracy in single precision and using 32-bit arithmetic provided that the initial SSP-matrix is calculated using greater precision, and using a computer which performs rounded arithmetic; either planar rotation method will almost always be adequate in such cases. How well do the planar rotation methods perform if the initial orthogonal reduction is also carried out in single precision? Table 2·5 shows the results obtained using a Hamiltonian cycle. Except for the LONGLEY data set, the accuracy is only about half a decimal digit worse than before. The poorer performance with the LONGLEY data is associated with one variable, the year. Subsets which included the year gave accuracies between 3·5 and 5·3 decimal digits, averaging 4·2 digits; the other subsets gave 5·1–6·7 accurate digits with an average of 5·6 digits. The value of this variable ranged from 1947 to 1962 so that the first two decimal digits were the same for each line of data, and there was not much variation in the third digit. This one variable causes a loss of about 2·5 decimal digits using orthogonal reduction and about 5 decimal digits using SSP-matrices with crude centring.

To compare the speeds of the planar rotation and Gauss–Jordan algorithms, artificial data were generated. Table 2·6 gives times in

Table 2.5 *Lowest and average numbers of accurate decimal digits in the residual sums of squares using planar rotations and an initial orthogonal reduction calculated in single precision*

Data set	*Lowest accuracy*	*Average accuracy*
WAMPLER	4·6	5·4
LONGLEY	3·5	4·9
STEAM	5·2	6·0
DETROIT	4·3	6·0
POLLUTE	5·0	5·6

Table 2.6 *Times in seconds, and their ratios, for calculating residual sums of squares for all subsets using planar rotation and Gauss–Jordan algorithms*

No. of variables	*Garside*	*Hamiltonian + Hammarling*	*Binary + Gentleman*	*Ratios*	
				Ham/GJ	*Bin/GJ*
10	3·8	5·5	4·7	1·42	1·23
11	7·7	12·1	9·8	1·56	1·27
12	15·4	25·3	19·1	1·64	1·24
13	31·	54·	38·	1·75	1·23
14	63·	113·	80·	1·80	1·27
15	129·	235·	152·	1·82	1·18
16	241·	494·	300·	2·05	1·25
17	484·	1037·	602·	2·14	1·24

seconds taken to calculate the residual sums of squares for all subsets on a PDP11-34 in FORTRAN IV under a UNIX operating system. No accuracy calculations were carried out during these speed tests and no use was made of the residual sums of squares which were not even stored. Averages of three runs were recorded for k up to 14; averages of two runs are recorded for larger k. We see that the Gauss–Jordan algorithm is faster for all values of k in the table, though its speed advantage over planar rotations using the binary sequence is not large. It is possible to increase the speed of this planar rotation algorithm by not operating upon the first row when a variable is deleted. This makes the algorithm fairly complex but increases both speed and accuracy. This has not been done by the author.

For a more detailed comparison of the merits of LS computations based upon SSP-matrices versus those based upon orthogonal reduction, see Maindonald (1984) or Farebrother (1988).

2.4 Interpretation of projections

The projections, r_{iy}, of the Y-variable on each of the orthogonal directions $\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_k$ are simple linear combinations of the values of the Y-variable, and hence have the same dimensions (e.g. length, mass, time, temperature, etc.) as the Y-variable. Similarly, the elements r_{ij} of matrix $\boldsymbol{R}$ in the orthogonal reduction (2.10) have the same units as variable X_j, i.e. the column variable. r_{ij} is simply the projection of variable X_j upon direction $\mathbf{Q}_i$, where $j \geqslant i$.

The size, r_{iy}, of a projection is dependent upon the ordering of the predictor variables, unless they are orthogonal. When the X-variables are correlated among themselves, changing their order often produces substantial changes in the projection of the Y-variable on the direction associated with a particular X-variable. Usually, the earlier that a variable occurs in the ordering, the larger will be the projection associated with it, but this is not a mathematical law and it is possible to construct examples for which the opposite applies. To illustrate the effect of ordering, consider the following artificial data:

X_1	40	41	42	43	44	45
Y	65 647	70 638	75 889	81 399	87 169	93 202
X_1	46	47	48	49	50	
Y	99 503	106 079	112 939	120 094	127 557	

If we construct two further variables X_2 and X_3 equal to the square and cube of X_1 respectively, then the LS regression equation relating Y to X_1, X_2 and X_3, including a constant, is

$$Y = 1 + X_1 + X_2 + X_3,$$

that is, all the regression coefficients should be equal to 1·0 exactly. This is a useful test of how well a regression package handles ill-conditioning. The projections of Y in the 'natural' and reverse order are

Constant	313 607	Cubic	320 267
Linear	64 856	Quadratic	−477
Quadratic	3 984	Linear	0.69
Cubic	78.6	Constant	0·000 87

The residual sum of squares is exactly 286. From the first ordering we see that if the cubic is omitted, the residual sum of squares is

$$286 + (78{\cdot}6)^2 = 6464$$

whereas, from the second ordering, if both the constant and linear

term are omitted, the residual sum of squares is only

$$286 + (0{\cdot}000\,87)^2 + (0{\cdot}69)^2 = 286{\cdot}5.$$

A useful statistical property of the projections is that if the true relationship between Y and the X-variables is linear with uncorrelated residuals which are normally distributed with homogeneous variance, σ^2, then the projections are also uncorrelated and normally distributed with variance σ^2. This can be demonstrated as follows. Let

$$Y = \mathrm{X}\beta + \varepsilon$$

where the ε are independent and $N(0, \sigma^2)$, then the projections are given by

$$\begin{aligned} \mathrm{Q}'Y &= \mathrm{Q}'\mathrm{X}\beta + \mathrm{Q}'\varepsilon \\ &= \begin{pmatrix} \mathrm{R} \\ 0 \end{pmatrix}\beta + \mathrm{Q}'\varepsilon. \end{aligned}$$

The vector $\mathrm{R}\beta$ contains the expected values of the first k projections, where k is the number of columns in X. The remaining $(n-k)$ projections have zero expected values. The stochastic part of the projections is $\mathrm{Q}'\varepsilon$. This vector has covariance matrix equal to

$$\begin{aligned} E(\mathrm{Q}'\varepsilon\varepsilon'\mathrm{Q}) &= \mathrm{Q}'\,E(\varepsilon\varepsilon')\,\mathrm{Q} \\ &= \sigma^2\mathrm{Q}'\mathrm{Q} \\ &= \sigma^2\mathrm{I}. \end{aligned}$$

That is, the elements of $\mathrm{Q}'\varepsilon$ are uncorrelated and all have variance σ^2. This part of the result is distribution-free. If the elements of ε are normally distributed, then the elements of $\mathrm{Q}'\varepsilon$, being linear combinations of normal variables, will also be normally distributed. Thus we can think of the projections as statistics with expected values and variances. This important property will be exploited in later chapters in explaining the nature of selection and stopping-rule biases. The result is due to Grossman and Styan (1972), though a much more readable account is given in Golub and Styan (1973).

Notice that the size of the elements of $\mathrm{Q}'\varepsilon$ is independent of both the sample size, n, and the number of predictors, k. As the sample size increases, if the X-predictors and Y continue to span roughly the same ranges of values, the elements of R will increase roughly in proportion to $\sqrt{n}$ so that the stochastic element in the projections decreases relative to $\mathrm{R}\beta$.

Appendix 2A Operations counts for all-subsets calculations

In the following derivations for Garside's Gauss–Jordan algorithm and for the planar rotations algorithms, the notation differs slightly from Garside's in that k is defined as the number of X-variables (excluding any variable representing a constant in the model). In Garside (1971b), the value of k is one larger than our value as it includes the Y-variable. The word 'operation' will be used to mean a multiplication or a division.

The derivations require the following two sums with various limits of summation and with $\alpha = 2$:

$$\sum_{r=1}^{k-1} r\alpha^{r-1} = \{(k-1)\alpha^k - k\alpha^{k-1} + 1\}/(1-\alpha)^2 \qquad (2A.1)$$

$$\sum_{r=1}^{k} r(r-1)\alpha^{r-2} = \{-k(k-1)\alpha^{k+1} + 2(k-1)(k+1)\alpha^k - k(k+1)\alpha^{k-1} + 2\}/(1-\alpha)^3. \qquad (2A.2)$$

2A.1 Garside's algorithm

Using a binary order for subset generation, the variable in position 1 is deleted only once, the variable in position 2 is deleted twice and reinstated once, and the variable in position i is deleted 2^{i-1} times and reinstated one less time. That is, variable number i is pivoted in or out a total of $(2^i - 1)$ times. In pivoting on row i, that row is left unchanged as variable i will be reinstated before any variable in a lower-numbered position is used as a pivot. Each higher-numbered row requires one operation to set up the calculations and then one operation per element in the row on and to the right of the diagonal. The setting-up operation also uses the reciprocal of the diagonal element in the pivot row. These setting-up operations, which were omitted from Garside's operations counts, are exactly those which would be calculated in operating upon the pivot row. In pivoting on row i, $(k+3-j)$ operations are performed on row j including the set-up operation for that row, where j ranges from $(i+1)$ to $(k+1)$. Thus the total number of operations required to pivot on row i is $(k+2-i)(k+3-i)/2$. Hence the total operations count for the Garside algorithm is

$$\sum_{i=1}^{k} (2^i - 1)(k+2-i)(k+3-i)/2 = 14(2^k) - (k+4)(k^2+8k+21)/6$$

As there are $(2^k - 1)$ subsets, the operations count is approximately 14 per subset for moderately large k. For small k, the numbers of operations per subset are as follows:

k	2	3	5	10	20
Operations per subset	5·0	7·0	10·29	13·56	13·998

2A.2 Planar rotations and a Hamiltonian cycle

The following simple algorithm generates a Hamiltonian cycle which is suitable for generating all subsets of k variables. Each time that step 3 is executed, a new subset of p variables is generated. The subsets of variables $1, 1\,2, 1\,2\,3, \ldots, 1\,2\,3 \ldots k$, are obtained from the initial ordering without any row swaps; the calculation of residual sums of squares for these subsets requires $2(k-1)$ operations.

1. For $i = 1$ to $k-1$, set index$(i) = i$.
2. Set $p = k - 1$.
3. Swap rows p and $p+1$.
4. Add 1 to index(p).

5(a). If index$(p) \leqslant k-1$, set index $(p+1)$ = index(p), add 1 to p, go to step 3.

5(b). Else, substract 1 from p,
If $p > 0$, go to step 3.
Else the end has been reached.

A new subset is generated each time that two rows are swapped. Hence row i and $(i+1)$ are swapped $({}^kC_i - 1)$ times. Using the Hammarling algorithm, it requires $10 + 2(k-i)$ operations to perform the swap. This count comprises 8 operations to set up the rotation and calculate the new elements in columns i and $(i+1)$, 2 operations on the $\boldsymbol{Q}'\mathbf{Y}$-vector, 2 operations to calculate the residual sum of squares for the new subset of i variables, and 1 operation for each remaining element in rows i and $(i+1)$. Hence the total count of operations is

$$\sum_{i=1}^{k-1} ({}^kC_i - 1)(10 + 2k - 2i) + 2(k-1) = (10+k)2^k - k^2 - 9k - 12$$

or about $(10 + k)$ operations per subset for moderately large k. For small k, the numbers of operations per subset are as follows:

k	2	3	5	10	20
Operations per subset	4·67	8·0	12·84	19·82	29·9995

At the end of this algorithm, the variables are in a scrambled order.

2A.3 Planar rotations and a binary sequence

To see how the algorithm works in this case, consider a case with six X-variables. Let us suppose that variable number 2 has just been deleted. The current order of the variables is 1 3 4 5 6 2. The binary code for this situation is 1 0 1 1 1 1, i.e. the only variable with a zero index is number 2. Subtracting 1 from the binary value gives 1 0 1 1 1 0, i.e. we drop variable number 6. This requires no change of order as the subset 1 3 4 5 is already in order. Subtracting another 1 gives 1 0 1 1 0 1. This requires the interchange of rows 4 and 5 which contain variables 5 and 6. After variable number 2 was deleted the triangular factorization looked like that shown in the table below, where X's and *'s denote nonzero elements. As variables are reinstated in exactly the reverse order in which they were introduced in the binary sequence, though not in the Hamiltonian cycle, we would hope to be able to avoid operating on elements other than those marked with an * in deleting variable number 5. Also, we would hope to be able to omit the column swaps which are a feature of the planar rotation algorithm. Unfortunately, if we use the

Variable represented in the row	*Variable represented in the column*					
	1	3	4	5	6	2
1	X	X	X	X	X	X
3		X	X	X	X	X
4			X	X	X	X
5				*	*	X
6					*	X
2						X

Hammarling algorithm, when variable 5 is later reinstated, its row multiplier is not the same as it was immediately before the variable was deleted. This means that the new row multiplier is not that which applied to the last elements in the rows for variables 5 and 6. If the Gentleman algorithm is used, this problem does not arise, nor does it arise with the so-called fast planar rotation given by the updating formula (9′) in Gentleman (1973). However, the latter is liable to give severe accuracy problems, which can be far worse than those associated with Gauss–Jordan methods.

In developing code to use the binary sequence with a planar rotation algorithm it is necessary to be able to find the appropriate variable to add to or delete from the current subset. Fortunately the variables in the subset are always represented in increasing order in the triangular factorization while those deleted are always in reverse numerical order. In the algorithm which follows, *nout*(*i*) stores the number of variables with number less than *i* which are currently out of the subset. This means that if variable number *i* is in the subset, its position is $i - nout(i)$. As the next-to-last variable is always reinstated immediately after being deleted, it is only necessary to calculate the value of the residual sum of squares when it is deleted; there is no need to calculate the new values for the three elements in the bottom corner of the triangular factorization. The array *ibin*( ) stores the 0–1 codes indicating whether a variable is in (1) or out (0) of the subset, k = the total number of variables, and *ifirst* is the number of the first variable which may be deleted (e.g. if there is a constant in the model which is to be in all subsets then $ifirst = 2$). The algorithm is

1. Calculate the initial residual sums of squares.
   Simulate the deletion of variable number $(k - 1)$.
2. For $i = 1$ to $k - 2$, set $ibin(i) = 1$ and $nout(i) = 0$.
   Set $last = k$.
3. Subtract 1 from position $k - 2$ in the binary value of the sequence.
   Set $p = k - 2$.
4. If $ibin(p) = 1$, go to step 6.
   Else set $ibin(p) = 1$.
   Set $ipos = p - nout(p)$
   Raise variable from position $(last + 1)$. to position *ipos*.
   Set $last = last + 1$.
5. Set $p = p - 1$.

If $p > 0$, go to step 4.
Else end has been reached.

6. 'Delete variable number p'
Set $ibin(p) = 0$.
Set $ipos = p - nout(p)$
Lower variable from row *ipos* to row *last*.
Set $last = last - 1$.
Calculate new residual sums of squares for rows *ipos* to *last*.
For $i = p + 1$ to $k - 2$, set $nout(i) = nout(p) + 1$.
Simulate the deletion of variable number $k - 1$ which is in row $last - 1$.
Go to step 3.

As for the Garside algorithm, variable i is operated upon $(2^i - 1)$ times, except that no calculations are required when $i = k$. In general, when variable number i is deleted, all of the higher-numbered variables are in the subset. Hence the variable must be rotated past variables numbered $i + 1,\ i + 2, \ldots, k$. Using the Gentleman algorithm (formula (9) in Gentleman, 1973), the number of operations required to swap variables i and $(i + 1)$ is $10 + 3(k - 1 + i)$. This is made up of 5 operations to set up the rotation, 3 operations on the Q'Y-vector, 2 operations to calculate the new residual sum of squares for the new subset of i variables, and 3 operations for each pair of remaining elements in the two rows up to and including column *last*. In the case of variable number $(k - 1)$, the residual sum of squares when it is deleted can be calculated in 7 operations. Hence the total count of operations is

$$\sum_{i=1}^{k-1} (2^i - 1)\{7 + 3(k - i)(k + 1 - i)/2\} + 7(2^k - 2) + 2(k - 1)$$

$$= (14 + 7/4)\cdot 2^k - (k^3 + 6k^2 + 27k + 22)/2$$

For moderately large k, this is about 15·75 operations per subset, or one-eighth more than for the Garside algorithm. For small k, the numbers of operations per subset are as follows:

k	2	4	5	10	20
Operations per subset	3·0	4·86	9·29	14·84	15·745

CHAPTER 3

Finding subsets which fit well

3.1 Objectives and limitations of this chapter

In this chapter we look at the problem of finding one or more subsets of variables which give models which fit a set of data fairly well. Though we will only be looking at models which fit well in the least-squares (LS) sense, similar ideas can be applied with other measures of goodness-of-fit. For instance, there have been considerable developments in the fitting of models to categorical data, see e.g. Goodman (1971), Brown (1976) and Benedetti and Brown (1978), in which the measure of goodness-of-fit is either a log-likelihood or a chi-square quantity. Other measures which have been used in subset selection have included that of minimizing the maximum deviation from the model, known simply as minimax fitting or as L_∞ fitting (e.g. Gentle and Kennedy 1978), and fitting by maximizing the sum of absolute deviations or L_1 fitting (e.g. Roodman, 1974; Gentle and Hanson, 1977; Narula and Wellington, 1979; Wellington and Narula, 1981).

We will also only be considering models involving linear combinations of variables, though some of the variables may themselves be functions of other variables. For instance, there may be four basic variables X_1, X_2, X_3 and X_4 from which other variables are derived such as $X_5 = X_1^2$, $X_6 = X_1 X_2$, $X_7 = X_4/X_3$, $X_8 = X_1 \log X_3$, etc.

The problem that we will then be looking at in this chapter is that, given a set of variables $X_1, X_2, \ldots, X_k$, we went to find a subset of $p < k$ variables $X_{(1)}, X_{(2)}, \ldots, X_{(p)}$ which minimizes or gives a suitably small value for

$$S = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} b_{(j)} x_{i,(j)} \right)^2, \tag{3.1}$$

where $x_{i,(j)}$, y_i are the ith observations, $i = 1, 2, \ldots, n$ of variables $X_{(j)}$

and Y, and $b_{(j)}$ is a LS regression coefficient. In most practical cases, the value of p is not predetermined and we will want to find good subsets for a range of values of p. We will often want to find not just one but perhaps the best 10 or more subsets of each size p. In some cases, the second, third or seventeenth-best may fit almost as well as the best, and may be preferable for future use for practical reasons, e.g. the predictors may be cheaper/easier to measure.

Problems of selection bias will be considered in Chapter 5, and the problem of deciding the best value of p to use, i.e. the so-called 'stopping rule' problem, will be considered in Chapter 6. This chapter is concerned with the mechanics of selecting subsets, not with their statistical properties.

In many practical cases, the minimization of (3.1) will be subject to constraints. One of these is that we may wish to force one or more variables to be in all subsets selected. For instance, most regression models include a constant which can be accommodated by making one of the variables, usually X_1, a dummy variable which always takes the 1 and which is forced into all subsets. There may also be constraints of the kind that one variable may only be included in a selected subset if another variable(s) is also included. For instance, it is often considered unreasonable to include a variable such as X_1^2 unless X_1 is also included. Dummy variables are often used to represent categorical variables. Thus if we have say five age groups, 0–16, 17–25, 26–40, 41–65, and over 65 years, we may introduce four dummy variables X_{17}, X_{18}, X_{19}, X_{20} with values assigned as in Table 3.1. In such cases it is often required that either all or none of the variables in such a group should be in the model.

It may be arguable whether such constraints should be applied. Here a distinction has to be made between a model which is intended

Table 3.1 *Use of dummy variables to represent a categorical variable (age)*

Age of subject (years)	X_{17}	X_{18}	X_{19}	X_{20}
0–16	0	0	0	0
17–25	1	0	0	0
26–40	0	1	0	0
41–65	0	0	1	0
>65	0	0	0	1

to be explanatory and meant to be an approximation to the real but unknown relationship between the variables, and a model which is intended to be used for prediction. The latter is our main objective in this monograph, though if the selected subset is to be used for prediction outside the range of the data used to select and calibrate the model, an explanatory model may be safer to use. A model which is 'plausible' is much more likely to be accepted by a client than one whose only merit is that it fits well.

In general we will expect that the number, p, of variables in the subset will be less than the number of observations, n, though the number of available predictors, k, often exceeds n. We will not require that the X-variables available for selection be linearly independent; for instance, it is often useful to include a difference $(X_1 - X_2)$ as another variable in addition to both X_1 and X_2.

3.2 Forward selection

In this procedure, the first variable selected is that variable X_j for which

$$S = \sum_{i=1}^{n} (y_i - b_j x_{ij})^2$$

is minimized, where b_j minimizes S for variable X_j. As the value of b_j is given by

$$b_j = \sum_{i=1}^{n} x_{ij} y_i \Big/ \sum_{i=1}^{n} x_{ij}^2$$

it follows that

$$S = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} x_{ij} y_i \right)^2 \Big/ \sum_{i=1}^{n} x_{ij}^2.$$

Hence the variable selected is that which maximizes

$$\left(\sum_{i=1}^{n} x_{ij} y_i \right)^2 \Big/ \sum_{i=1}^{n} x_{ij}^2. \tag{3.2}$$

If this expression is divided by $\sum_{i=1}^{n} y_i^2$, then we have the square of the cosine of the angle between vectors $\mathbf{X}_j$ and $\mathbf{Y}$. If the mean has been subtracted from each variable, then the cosine is the correlation between variables X_j and Y.

Let the first variable selected be denoted by $X_{(1)}$; this variable is then forced into all further subsets. The residuals $\mathbf{Y} - \mathbf{X}_{(1)}b_{(1)}$ are orthogonal to $X_{(1)}$, and so to reduce the sum of squares by adding further variables we must search the space orthogonal to $X_{(1)}$. From each variable X_j, other than the one already selected, we could form

$$\mathbf{X}_{j.(1)} = \mathbf{X}_j - b_{j.(1)}\mathbf{X}_{(1)}$$

where $b_{j.(1)}$ is the LS regression coefficient of X_j upon $X_{(1)}$. Now we find that variable, $X_{j.(1)}$, which maximizes expression (3.2) when $\mathbf{Y}$ is replaced with $\mathbf{Y} - \mathbf{X}_{(1)}b_{(1)}$ and $\mathbf{X}_j$ is replaced with $X_{j.(1)}$. The required sums of squares and products can be calculated directly from previous sums of squares and products without calculating these orthogonal components for each of the n observations, in fact the calculations are precisely those of a Gauss–Jordan pivoting out of the selected variable. If the mean had first been subtracted from each variable then the new variable selected is that which has the largest partial correlation in absolute value with Y after variable $X_{(1)}$ has been fitted.

Thus variables $X_{(1)}, X_{(2)}, \ldots, X_{(p)}$ are progressively added to the prediction equation, each variable being chosen because it minimizes the residual sum of squares when added to those already selected.

A computational method for forward selection using an orthogonal reduction is as follows. Let us suppose that we have reached the stage where r variables have been selected or forced into the subset, where r may be zero. Planar rotations are used to make these variables the first ones in the triangular factorization, that is they occupy the top r rows. Then the orthogonal reduction can be written as

$$(\boldsymbol{X}_A, \boldsymbol{X}_B) = (\boldsymbol{Q}_A, \boldsymbol{Q}_B)\begin{pmatrix} \boldsymbol{R}_A \\ \boldsymbol{R}_B \end{pmatrix}$$

$$\mathbf{Y} = (\boldsymbol{Q}_A, \boldsymbol{Q}_B)\begin{pmatrix} r_{yA} \\ r_{yB} \end{pmatrix} + \mathbf{e}$$

where $\boldsymbol{X}_A$ is an $n \times r$ matrix consisting of the values of the variables selected so far, $\boldsymbol{X}_B$ is an $n \times (k-r)$ matrix of the remaining variables, $\boldsymbol{Q}_A$ and $\boldsymbol{Q}_B$ have r and $(k-r)$ columns respectively, all of which are orthogonal, $\boldsymbol{R}_A$ consists of the top r rows of a $k \times k$ upper-triangular matrix and $\boldsymbol{R}_B$ consists of the last $(k-r)$ rows and hence is triangular, r_{yA} and r_{yB} consist of the first r and last $(k-r)$ projections of $\mathbf{Y}$ on

the directions given by the columns of $\boldsymbol{Q}_A$ and $\boldsymbol{Q}_B$, and $\mathbf{e}$ is a vector of n residuals. The information on the components of $\mathbf{Y}$ and of the remaining unselected X-variables which are orthogonal to the selected variables is then contained in r_{yB} and $\boldsymbol{R}_B$.

Let us write

$$\boldsymbol{R}_B = \begin{pmatrix} r_{11} & r_{12} & r_{13} & \cdots \\ & r_{22} & r_{23} & \cdots \\ & & r_{33} & \cdots \\ & & & \cdots \end{pmatrix} \qquad r_{yB} = \begin{pmatrix} r_{1y} \\ r_{2y} \\ r_{3y} \\ \cdots \end{pmatrix}$$

then if the variable in the top row of the sub-matrix $\boldsymbol{R}_B$ is added next, the reduction in the residual sum of squares (RSS) is r_{1y}^2. To find the reduction in RSS if the variable in the second row is added next instead, a planar rotation can be used to bring this variable into the top row and then the reduction in RSS is equal to the square of the value which is then in the top position of r_{yB}. There is no need to perform the full planar rotation of whole rows to calculate the effect of bringing a variable from row i to row 1 of $\boldsymbol{R}_B$. In swapping a pair of rows, only the elements on the diagonal and immediately next to it are needed to calculate the required planar rotations which are then applied only to calculate the new diagonal element for the upper row and the new element in r_{yB} for the upper row. Using Hammarling rotations, the reduction in RSS for the ith row of $\boldsymbol{R}_B$ can be calculated in $7(i-1)+2$ operations, and hence the total number of operations for all $(k-r)$ variables is $(7/2)(k-r)(k-r-1)+2(k-r)$. At the end, the variable selected to add next is rotated to the top row, if it is not already there, using the full planar rotations, adding a few more operations to the count.

An alternative and quicker way to calculate the reductions in RSS is to use (3.2) but with the x's and y's replaced by their components orthogonal to the already selected variables. The sums of squares of the x's are obtained from the diagonal elements of $\boldsymbol{R}_B'\boldsymbol{R}_B$, and the cross-products of x's and y's from $\boldsymbol{R}_B'r_{yB}$. Thus in the notation above, the sum of squares for the X-variable in the third row is $r_{13}^2+r_{23}^2+r_{33}^2$ and the cross-product is $r_{13}r_{1y}+r_{23}r_{2y}+r_{33}r_{3y}$. To calculate all $(k-r)$ reductions in RSS requires $(k-r)(k-r+3)$ operations if $\boldsymbol{R}_B$ is actually stored without the row multipliers used in the Hammarling and Gentleman algorithms, or about 50% more operations if row multipliers are being used. The selected variable

Table 3.2

Observation number	X_1	X_2	X_3	Y
1	1000	1002	0	−2
2	−1000	−999	−1	−1
3	−1000	−1001	1	1
4	1000	998	0	2

is then rotated into the top row of $\boldsymbol{R}_B$ and the process repeated to find the next variable with one row and column less in $\boldsymbol{R}_B$.

In general there is no reason why the subset of p variables which gives the smallest *RSS* should contain the subset of $(p-1)$ variables which gives the smallest *RSS* for $(p-1)$ variables. Table 3.11 provides an example in which the best-fitting subsets of three and two variables have no variables in common, though this is a rare situation in practice. Hence there is no guarantee that forward selection will find the best-fitting subsets of any size except for $p=1$ and $p=k$.

Consider the artificial example shown in Table 3.2. The correlations (cosines of angles) between Y and X_1, X_2, X_3 are 0·0, −0·0016 and 0·4472 respectively. Forward selection picks X_3 as the first variable. The partial correlations of Y upon X_1 and X_2 after being made orthogonal to X_3 are 0·0 and −0·0014 respectively. With X_3 selected, the subset of X_1 and X_2, which gives a perfect fit, $Y = X_1 - X_2$, cannot be obtained.

Examples similar to that above do occur in real life. The difference $(X_1 - X_2)$ may be a proxy for a rate of change in time or space, and in many situations such rates of change may be good predictors even when the values of the separate variables have little or no predictive value. In such cases, any subset-finding method which adds, deletes or replaces only one variable at a time may find very inferior subsets. If there is good reason to believe that differences may provide good predictors then they can be added to the subset of variables available for selection, though care must be taken to make sure that the software used can handle linear dependencies among the predictor variables.

3.3 Efroymson's algorithm

The name 'stepwise regression' is often used to mean an algorithm proposed by Efroymson (1960). This is a variation on forward

selection. After each variable (other than the first) is added to the set of selected variables, a test is made to see if any of the previously selected variables can be deleted without appreciably increasing the residual sum of squares. Efroymson's algorithm incorporates criteria for the addition and deletion of variables as follows.

(a) Addition

Let RSS_p denote the residual sum of squares with p variables and a constant in the model. Suppose the smallest RSS which can be obtained by adding another variable to the present set is RSS_{p+1}. The ratio

$$R = \frac{RSS_p - RSS_{p+1}}{RSS_{p+1}/(n-p-2)} \tag{3.3}$$

is calculated and compared with an 'F-to-enter' value, say F_e. If R is greater than F_e, the variable is added to the selected set.

(b) Deletion

With p variables and a constant in the selected subset, let RSS_{p-1} be the smallest RSS which can be obtained after deleting any variable from the previously selected variables. The ratio

$$R = \frac{RSS_{p-1} - RSS_p}{RSS_p/(n-p-1)} \tag{3.4}$$

is calculated and compared with an 'F-to-delete (or drop)' value, say F_d. If R is less than F_d, the variable is deleted from the selected set.

(c) Convergence of algorithm

From (3.3) it follows that when the criterion for adding a variable is satisfied

$$RSS_{p+1} \leqslant RSS_p/\{1 + F_e/(n-p-2)\},$$

while from (3.4) it follows that when the criterion for deletion of a variable is satisfied

$$RSS_p \leqslant RSS_{p+1}\{1 + F_d/(n-p-2)\}.$$

Hence when an addition is followed by a deletion, the new RSS, RSS_p^* say, is such that

$$RSS_p^* \leqslant RSS_p \cdot \frac{1 + F_d/(n-p-2)}{1 + F_e/(n-p-2)}. \tag{3.5}$$

The procedure stops when no further additions or deletions are possible which satisfy the criteria. As each RSS_p is bounded below by the smallest RSS for any subset of p variables, by ensuring that the RSS is reduced each time that a new subset of p variables is found, convergence is guaranteed. From (3.5) it follows that a sufficient condition for convergence is that $F_d < F_e$.

(d) True significance level

The use of the terms 'F-to-enter' and 'F-to-delete' suggests that the ratios R have an F-distribution under the null hypothesis, i.e. that the model is the true model, and subject to the true residuals being independently, identically and normally distributed. This is not so. Suppose that after p variables have been entered, these conditions are satisfied. If the value of R is calculated using (3.3) but using the value of RSS_{p+1} for one of the remaining variables chosen at random, then the distribution of R is the F-distribution. However, if we choose that variable which maximizes R, then the distribution is not an F-distribution or anything remotely like an F-distribution. This was pointed out by Draper *et al.* (1971) and by Pope and Webster (1972); both papers contain derivations of the distribution of the R-statistic for entering variables. Evaluation of the distribution requires multidimensional numerical integration. A rough approximation to the percentage points can be obtained by treating R as if it were the maximum of $(k-p)$ independent F-ratios. The R-value corresponding to a significance level of α is then the F-value which gives a significance level of α^* where

$$(1-\alpha^*)^{k-p} = 1-\alpha. \tag{3.6}$$

Pope and Webster suggest that if would be better to replace $RSS_{p+1}/(n-p-2)$ in the denominator of (3.3) with $RSS_k/(n-k-1)$, i.e. to use all of the available predictor variables in estimating the residual variance. Limited tables for the distribution of R have been given by Draper *et. al.* (1979); these show that the nominal 5% points for the maximum F-to-enter as incorrectly obtained from the F-distribution will often give true significance levels in excess of 50%. An early use of (3.6) to calculate F-to-enter values was by Miller (1962).

From (3.6) we can derive the rough Bonferroni bound that $\alpha < (k-p)\alpha^*$ for the true significance level, α, in terms of the value, α^*, read from tables of the F-distribution. Butler (1984) has given a

fairly tight lower bound for the selection of one more variable out of the remaining $(k-p)$ variables, provided that the information that none of them had been selected earlier can be neglected. This lower bound is

$$(k-p)\alpha^* - \sum_{i<j} p_{ij}$$

where p_{ij} is the probability that two of the remaining variables, X_i and X_j, satisfy the *a priori* condition for significance at the α^* level. Butler's derivation is in terms of partial correlations of the dependent variable with the predictors, though this could be translated into F-values. The joint probabilities, p_{ij}, can be expressed as bivariate integrals which must be evaluated numerically. An algorithm for doing this is described in Butler (1982).

As with forward selection, there is no guarantee that this algorithm will find the best-fitting subsets, though it often performs better than forward selection when some of the predictors are highly correlated. The algorithm incorporates its own built-in stopping rule; recommendations with respect to suitable values for F_e and F_d will be given in Chapter 6.

3.4 Backward elimination

In this procedure we start with all k variables, including a constant if there is one, in the selected set. Let RSS_k be the corresponding residual sum of squares. That variable is chosen for deletion which yields the smallest value of RSS_{k-1} after deletion. Then that variable from the remaining $(k-1)$ which yields the smallest RSS_{k-2} is deleted. The process continues until there is only one variable left, or until some stopping criterion is satisfied.

Backward elimination can be carried out fairly easily starting from an orthogonal reduction. If the variable to be deleted is in the last row of the triangular reduction, then the increase in RSS when it is deleted is simply r_{ky}^2. Hence each variable in turn can be moved to the bottom row to find which gives the smallest increase in RSS when deleted. The actual movement of rows can be simulated without altering any values in the triangular factorization. The values in the row to be moved are copied into a separate storage location and the effect of rotation past each lower row is calculated. Numbering rows from the bottom, the number of operations necessary to lower

the variable from row i to row 1, using Hammarling rotations, is $(i-1)(i+12)/2$ plus 2 operations to calculate the increase in *RSS*. Hence the total for all k variables is $k(k^2+18k-7)/6$. The selected variable is then moved to the last row and the process repeated on the $(k-1)\times(k-1)$ matrix obtained by omitting the last row and column.

There is an alternative method using SSP-matrices, as follows. It can be shown that the increase in *RSS* when variable i is deleted is b_i^2/c^{ii} where b_i is the LS regression coefficient with all variables in the model and c^{ii} is the ith diagonal element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. This formula can be verified very easily from the orthogonal reduction. Consider the case of the last variable. By back-substitution,

$$b_k = r_{ky}/r_{kk}.$$

Now as

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{R}^{-1}\boldsymbol{R}^{-T}$$

the element in the bottom right-hand corner of the inverse of the SSP-matrix is simply

$$c^{kk} = 1/r_{kk}^2.$$

Hence

$$b_k^2/c^{kk} = r_{ky}^2$$

which is the correct increase. Now any other variable can be moved into the last row of the triangular decomposition without changing the value of b_i^2/c^{ii} for that variable, and hence the formula holds for all variables. Alternative derivations of this well-known result are given in many books on linear regression.

Using this alternative method requires $k^2(k+1)/2$ operations for the inversion of the SSP-matrix. After the first variable has been deleted, the inverse matrix for the remaining variables can be obtained by pivoting out the selected variable using the usual Gauss–Jordan formulae. This requires only $k(k+1)/2$ operations. The method using SSP-matrices is usually much faster but the accuracy can be very poor if the SSP-matrix is ill-conditioned, and calculated *RSS*'s or increases in *RSS* can be negative unless great care is exercised.

Backward elimination is usually not feasible when there are more variables than observations. If we have 100 variables but only 50 observations then, provided that the 100 columns of the predictor variables have rank 50, the residual sum of squares will be zero. In

most cases 51 variables will have to be deleted before a nonzero *RSS* is obtained. The number of ways of selecting 51 out of 100 variables is approximately 9.9×10^{28}.

It has been argued by Mantel (1970) that in situations similar to that in the example in section 3.2 in which the variable Y is highly correlated with some linear combination of the X-variables, such as $X_1 - X_2$ or a second difference $X_1 - 2X_2 + X_3$, but where the correlations with the individual variables are small, backward elimination will tend to leave such groups of variables in the subset, whereas they would not enter in a forward selection until almost all variables have been included.

Beale (1970) countered with the common situation in which the X-variables are percentages, e.g. of chemical constituents, which sum to 100% or perhaps slightly less if there are small amounts of unidentified compounds present. Any one variable is then given either exactly or approximately by subtracting the others from 100%. It is then a matter of chance which variable is deleted first in the backward elimination; the first one could be the only one which is of any value for prediction from small subsets, and once a variable has been deleted it cannot be reintroduced in backward elimination. Example 3.3 in section 3.10 is a case in which the first variable deleted in backward elimination is the first one inserted in forward selection. Of course a backward analogue of the Efroymson procedure is possible.

Backward elimination always requires far more computation than forward selection. If for instance we have 50 variables available and expect to select a subset of less than 10 of them, in forward selection we would only proceed until about 10 or so variables have been included. In backward elimination we start with 50 variables, then 49, then 48, until eventually we reach the size of interest.

Both forward selection and backward elimination can fare arbitrarily badly in finding the best-fitting subsets. Berk (1978b) has shown that even when forward selection and backward elimination yield exactly the same subsets of all sizes, there can be much better-fitting subsets of some sizes.

3.5 Sequential replacement algorithms

The basic idea here is that once two or more variables have been selected, we see whether any of those variables can be replaced with

another which gives a smaller *RSS*. For instance, if we have 26 variables which are conveniently denoted by the letters of the alphabet, we may at some stage have selected the subset of four variables

ABCD.

Let us try replacing variable *A*. There may be several variables from the remaining 22 which give a smaller *RSS* when in a subset with *B*, *C* and *D*. Suppose that of these, variable *M* yields the smallest *RSS*. We can replace *A* with *M* giving the subset

MBCD.

Now we try replacing variable *B*, then variable *C*, then *D*, then back to *M*, etc. At some of these attempts there will be no variable which yields a reduction in the *RSS*, in which case we just move on to the next variable. Sometimes variables which have been replaced will return. The process continues until no further reduction is possible by replacing any variable.

The procedure must converge as each replacement reduces the *RSS* which is bounded below. In practice the procedure usually converges very rapidly.

Unfortunately this type of replacement algorithm does not guarantee convergence upon the best-fitting subset of the size being considered. In the above example, if we had started by trying to replace variable *B* instead of variable *A* then the procedure might have converged upon a different subset. Let us call these final subsets, **stationary subsets**.

Suppose that in our hypothetical example, the subset of four variables which gives the smallest *RSS* is

BEST.

If we start with the subset *PEST*, we are certain to reach the stationary subset *BEST* only if variable *P* is the first one to be replaced. Thus we are certain to reach the absolute minimum from only 23 out of the 14950 possible starting subsets of four variables, though in practice the best subset will usually be found from many more starting subsets. Even if we are lucky and find the best-fitting subset of this size, we have no way of knowing that it is the best one.

The following modification improves our chances of finding the best-fitting subset. Suppose again that we start with the subset *ABCD*

Table 3.3

	Variables	*RSS*
Initial subset	*ABCD*	100
Best replacement for *A*	*MBCD*	93
Best replacement for *B*	*AMCD*	91
Best replacement for *C*	*ABXD*	96
Best replacement for *D*	*ABCQ*	94

and that this gives an $RSS = 100$ say. We now look for the best replacement for variable A, but we do not make the replacement yet. Similarly, we try replacing variable B but with variable A still in the subset. Suppose we obtain the best replacements for each variable shown in Table 3.3. Replacing variable B gives the smallest RSS, so we make the replacement and then repeat the process. Notice that at the next stage we know that we cannot replace variable M with any variable which gives a smaller RSS so that we only consider replacing the other three at the next step.

Using this replacement algorithm there are now 92 out of the 14950 possible starting subsets from which we are guaranteed to find the best-fitting subset.

Either of the above replacement algorithms can be used in conjunction with any of the algorithms described in earlier sections of this chapter. Thus a sequential replacement algorithm can be obtained by taking the forward selection algorithm and applying a replacement procedure after each new variable is added.

Sequential replacement requires more computation than forward selection or the Efroymson algorithm but it is still feasible to apply to problems with several hundred variables when subsets of say up to 20–30 variables are required.

An alternative technique which can be used when there are large numbers of variables is to choose starting subsets randomly and then apply a replacement procedure. However, in one problem on which the author was involved in which there were 757 variables, 74 different stationary subsets of 6 variables were obtained from 100 random starts! Even then a subset which gave only about two-thirds of the RSS of the best of the 74, was later found by chance.

Replacing two variables at a time substantially reduces the maximum number of stationary subsets and means that there is a

much better chance of finding good subsets when, for instance, a difference between two variables is a good predictor but was not included in the available set of predictor variables. However, for subsets of p variables there are $p(p-1)/2$ pairs of variables to be considered for replacement and hence much more computation is required. Similarly, forward selection and backward elimination are possible two variables at a time.

3.6 Generating all subsets

We saw earlier in Chapter 2 that it is feasible to generate all subsets of variables provided that the number of predictor variables is not too large, say less than about 20, if only the *RSS* is calculated for each subset. After the complete search has been carried out, a small number of the more promising subsets can be examined in more detail. The obvious disadvantage of generating all subsets is cost. The number of possible subsets of one or more variables out of k is $(2^k - 1)$. Thus the computational cost roughly doubles with each additional variable.

In most cases, we are not interested in all subsets of all sizes. For instance, someone with 100 variables (and possibly only 50 observations) is unlikely to be interested in say subsets of 45 variables; subsets of 10 variables may be large enough. There are $(2^{100} - 1) = 1{\cdot}3 \times 10^{30}$ subsets of one or more variables out of 100, but only $1{\cdot}9 \times 10^{13}$ subsets of 10 or less. Even this last number is far too large for an exhaustive evaluation of all subsets to be feasible. However the device to be described in the next section will usually render this case feasible.

Many algorithms have been published for performing exhaustive evaluations including those of Garside (1965, 1971b, c), Schatzoff *et al.* (1968), Furnival (1971) and Morgan and Tatar (1972). These all consider all subsets of all sizes, and all require that the number of observations is at least as great as the number of predictor variables. Kudo and Tarumi (1974) have published an algorithm for searching for all subsets of p or less variables out of k. All of these algorithms use matrix inversion and SSP-matrices and present problems when the SSP-matrix is ill-conditioned or of less than full rank. In a series of papers, Narula and Wellington (1977a, b, 1979) and Wellington and Narula (1981) have presented an algorithm for finding the best-fitting subsets of p variables out of k variables when the criterion

used is that of minimizing the sum of absolute deviations, i.e. minimizing the L_1-norm.

Let us consider algorithms for generating all subsets of p variables out of k. Table 3.4 illustrates the generation of all subsets of three variables out of seven in what is known as lexicographic order. It should be read down the columns.

This would appear to be a good order of generation for software using SSP-matrices as the variable at the right-hand end is changing most rapidly and it is possible to perform the calculation of the *RSS*'s by operating only on sub-matrices of the SSP-matrix. In the case of the variable in position p, the calculation of the new *RSS* when this variable is changed only requires one multiplication and one division. However, the changes to variables in other positions require operations on almost every element in the SSP-matrix. An algorithm for generating the lexicographic order has been given by Gentleman J.F. (1975).

In using SSP-matrices, the order of the variables remains unchanged and the operations consist of inverting and reinverting part of the matrix so that the variables to be pivoted in or out are easily found. A lexicographic order of generation is not well suited to a planar rotation algorithm. Table 3.4 only shows the order of the first three variables; below we see what the full order of variables could look like. One of many possible orderings is shown in Table 3.5 for the first part of the lexicographic order above.

In the algorithm used to generate the start of the sequence, the next variable or variables needed were moved up from wherever they were in the complete ordering and the other variables kept their previous order but just moved down the appropriate number of places. As can be seen, the above ordering requires a large number

Table 3.4 *Lexicographic order of generation of all subsets of three variables out of seven*

123	136	167	247	356
124	137	234	256	357
125	145	235	257	367
126	146	236	267	456
127	147	237	345	457
134	156	245	346	467
135	157	246	347	567

Table 3.5 *Start of a possible sequence to generate the lexicographic order in Table 3.4*

123	4567
124	3567
125	4367
126	5437
127	6543
134	2765
135	4276
136	5427
137	6542
145	3762

of interchanges of variables. In particular, the move from 127 6543 to 134 2765 seems very wasteful. At this stage, all of the subsets of three variables including 12 have been exhausted so that either 1 or 2 must be dropped, but it would be much more efficient to introduce the 6 from the next position next. This can be done in the following algorithm which operates on whichever variables happen to be in the position from which a variable is to be moved irrespective of the number of that variable.

The basic idea of this combinatoric algorithm is that for each of the p positions to be filled there is a 'last' position to which the next variable taken from that position is moved. Initially it is position k for all of the p positions. The first variable moved from position i goes to position $last(i)$ among the last $(k-p)$. To make sure that the variable is not moved again with the current set of variables in positions $1, 2, \ldots, i-1$, one is subtracted from $last(i)$. Also, to make sure that all subsets of the variables in positions $p+1, \ldots,$ (new) $last(i)$ are used without moving the variable just moved from position i, the values of $last(j)$ for $j = i+1, \ldots, p$ are set equal to the new value of $last(i)$. The algorithm is as follows:

1. For $i = 1$ to p, set $last(i) = k$.
2. Set $ipt = p$. ipt is the pointer to the current active position among the first p positions.
3. Move the variable from position ipt to position $last(ipt)$.
4. Subtract 1 from $last(ipt)$.
5. If $last(ipt) = p$, go to step 7.
6. If $ipt = p$, then go to step 3.
   Else, for $i = ipt + 1, \ldots, p$ set $last(i) = last(ipt)$. Then go to step 2.

7. Set $ipt = ipt - 1$.
   If $ipt > 0$, for $i = ipt + 1, \ldots, p$ set $last(i) = last(ipt)$. Then go to step 2.
   Else, the end has been reached.

The above algorithm is in fact the same as the Gentleman (1975) algorithm except that instead of the index of the simulated nested DO-loops being incremented, the upper limit, i.e. $last(ipt)$, is decremented. It could equally well have been written in the same way as for the Gentleman algorithm, in fact the first exhaustive search algorithm written by the author was in this form. The algorithm does not generate a lexicographic order as the index of the DO-loops (or the upper limits here) are not used to determine the number but the position of a variable. The algorithm generates the sequence in Table 3.6 of subsets of three variables out of seven.

The algorithm above generates all subsets of p variables out of k; it does not in general generate all subsets of less than p variables. For instance, in Table 3.6 the pairs 23, 25, 27, 34 and 56 do not occur in the first two positions, and the single variables 2 and 5 do not appear in the first position. The algorithm can easily be modified to generate all subsets of p variables or less out of k as follows. The interchange in step 3 is only carried out if $ipt < last(ipt)$, and the condition in step 5 is changed to $last(ipt) = ipt$. With this modification to the algorithm, the generated sequence is as given in Table 3.7.

We noted earlier that the inner cycle of the lexicographic algorithm can be performed extremely rapidly using SSP-matrices. This gives it a major speed advantage over algorithms using planar rotations.

Table 3.6 *Ordering of generated subsets of three variables out of seven, indicating also the positions of excluded variables*

123 4567	134 5672	452 7361	735 2641
124 5673	135 6472	427 3651	752 6341
125 6743	136 5472	423 6751	756 2341
126 7543	165 4372	426 3751	762 5341
127 6543	164 5372	463 7251	625 3741
176 5432	145 6372	467 3251	623 5741
175 4362	456 3721	473 6251	635 2741
174 3562	453 7261	736 2541	352 6741
173 4562	457 2361	732 5641	

Table 3.7 *Ordering of generated subsets of three or less variables out of seven*

123 4567	136 5472	526 3741	362 4751
124 5673	165 4372	563 7241	326 4751
125 6743	164 5372	567 3241	264 7351
126 7543	145 6372	573 6241	267 4351
127 6543	154 6372	537 6241	274 6351
176 5432	546 3721	376 2451	247 6351
175 4362	543 7261	372 4651	476 2351
174 3562	547 2361	374 2651	467 2351
173 4562	542 7361	342 6751	674 2351
134 5672	527 3641	346 2751	764 2351
135 6472	523 6741		

However, the outer cycles are much slower using SSP-matrices as operations must be performed on all except a few rows at the top of the matrix, whereas planar rotation algorithms usually require operations on only a very small number of adjacent rows near the top of the matrix. The planar rotation algorithm can be made faster by using the trick employed in section 2.3 on forward selection. Instead of the slow inner cycle, an SSP-matrix is formed of those components of the variables in rows $p, p+1, \ldots, last(p)$ which are orthogonal to the variables in rows $1, 2, \ldots, p-1$. This simply means that we form

$$\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{R}'\boldsymbol{R}$$

using as $\boldsymbol{R}$ the triangular sub-matrix between rows and columns p and $last(p)$ inclusive of the triangular factorization which is current at that time. Further details will be given in the next section.

3.7 Using branch-and-bound techniques

Suppose that we are looking for the subset of 5 variables out of 26 which gives the smallest *RSS*. Let the variables again be denoted by the letters *A*–*Z*. We could proceed by dividing all the possible subsets into two 'branches', those which contain variable *A* and those which do not. Within each branch we can have sub-branches including and excluding variable *B*, etc. Now suppose that at some stage we have found a subset of five variables containing *A* or *B* or both

which gives $RSS = 100$. Let us suppose that we are about to start examining that sub-branch which excludes both A and B. A lower bound on the smallest RSS which can be obtained from this sub-branch is the RSS for all of the 24 variables C–Z. If this is say 108 then no subset of 5 variables from this sub-branch can do better than this, and as we have already found a smaller RSS, this whole sub-branch can be skipped.

This simple device appears to have been used first in subset selection by Beale, Kendall and Mann (1967), and by Hocking and Leslie (1967). It is further exploited by LaMotte and Hocking (1970). Using this device gives us the advantage of exhaustive searches that are guaranteed to find the best-fitting subsets, and at the same time the amount of computation is often reduced substantially.

The branch-and-bound device can similarly be applied with advantage with most other criteria of goodness-of-fit such as minimum sum of absolute deviations or maximum likelihood. One such application has been made by Edwards and Havranek (1987) to derive so-called minimal adequate sets.

Branch-and-bound can be applied in a number of ways. For instance, if we want to find the 10 best subsets of 5 variables, then the RSS for all the variables in the sub-branch is compared against the tenth best subset of 5 variables which has been found up to that stage.

Alternatively, consider the task of finding the best-fitting subsets of all sizes up to and including six variables. Suppose that we are about to start examining a sub-branch and that the smallest RSS's found up to this stage have been

No. of variables	1	2	3	4	5	6
RSS	503	368	251	148	93	71

Suppose that the RSS including all of the variables of this sub-branch is 105, then we cannot possibly find a better-fitting subset of 5 or 6 variables from this sub-branch, though there could be better-fitting subsets of 1, 2, 3 or 4 variables. Hence, until we complete searching this sub-branch, we look only for subsets of 4 or fewer variables.

Branch-and-bound is particularly useful when there are 'dominant' variables such that good-fitting subsets must include these variables. The device is of almost no value when there are more variables than observations as the lower bounds are nearly always zero.

An algorithmic trick which can be used in conjunction with branch-and-bound is that of saving the current state of matrices and arrays immediately before a variable is deleted. For instance, suppose we are looking for the best subset of 10 variables out of 100 and have reached the stage where the variable in position 7 is about to be deleted for the first time. Using orthogonal reduction methods, the variable is then moved from position 7 to position 100, an operation which requires a very large amount of computation. The *RSS* with the other 99 variables in the regression is then calculated. Suppose that this shows that with the variable from position 7 deleted we cannot possibly improve upon the best subset of four variables which we have found so far. The deleted variable must then be reinstated, that is moved all the way back to position 7. If a copy of the matrices and arrays had been kept then a considerable amount of computation is avoided.

The calculation of the bounds requires a similar amount of extra computation using methods based upon SSP-matrices. At the start of any branch, say that excluding variables A and B, the SSP-matrix will usually be available with none of the variables included in the model. All of the variables except A and B must then be pivoted on to obtain the bound. A way around this problem is to work with two matrices available, one being the SSP-matrix and the other being its complete inverse. One of these matrices is then used to calculate the *RSS*'s for small subsets while the other is used to obtain bounds. Furnival and Wilson (1974) have published an algorithm based upon SSP-matrices which uses this trick. It finds the best subsets of all sizes though and can be very slow. One important feature noted by Furnival and Wilson is that the amount of work needed to find say the 10 best-fitting subsets of each size is usually not much greater than that needed to find the best ones.

The Furnival and Wilson algorithm has been generalized by Narendra and Fukunaga (1977) to the minimization of a general quadratic form

$$\mathbf{x}_p' \boldsymbol{S}_p^{-1} \mathbf{x}_p$$

over all subsets of the predictors, where $\mathbf{x}$ is a vector of length k.

FORTRAN code for this is contained in Ridout (1988), though this does not use the Cholesky factorization of either $\boldsymbol{S}_p$ or its inverse. One way to use this would be to substitute for $\mathbf{x}$ the value of the predictors for which a prediction will be required, and to substitute for $\boldsymbol{S}^{-1}$ the covariance matrix of the LS estimates of the regression coefficients for the model using all the available predictors. This will be discussed further in Chapter 6.

If model building, rather than prediction, is the objective, then the algorithm of Edwards and Havranek (1987) may be used. This is similar to the Furnival and Wilson algorithm but attempts to find only what Aitkin (1974) has called minimal adequate (sub)sets. More details are given in section 4.2.

An equivalent algorithm is used by Narula and Wellington (1977b, 1979), Armstrong and Kung (1982) and Armstrong *et al* (1984) but using the weighted sum of absolute errors as the criterion instead of LS.

The equivalence of keeping an inverse matrix when using orthogonal reductions is to keep a second triangular factorization with the variables in reverse order. The bounds are then calculated from operations on variables which are in the lower rows of the triangular factorization. If the use of bounds results in variables frequently being reinstated immediately after deletion then this idea obviously has considerable merit. However, the author has usually used a replacement algorithm before calling an exhaustive search algorithm, and this usually establishes fairly good bounds and orders variables so that the 'best' variables are at the start and so are rarely deleted. Also, many of the author's problems have involved more variables than observations, in which case branch-and-bound is of no value until the final stages of the search.

An exhaustive search can be very time consuming if a large number of possible subsets have to be examined. Experience suggests that the maximum feasible size of problem is one for which the number of possible subsets, i.e. $\sum^k C_p$ where k is the number of available predictor variables and p ranges from 1 up to the maximum size of subset of interest, is of the order of 10^7. It is recommended that any software for exhaustive search should write out to disk from time to time the best subsets found up to that stage so that all is not lost if the computer run exceeds its time limit. If this information can then be read back in later, it can provide good bounds to save much of the computation during a future run.

3.8 Grouping variables

Gabriel and Pun (1979) have suggested that when there are too many variables for an exhaustive search to be feasible, it may be possible to break them down into groups within which an exhaustive search for best-fitting subsets is feasible. The grouping of variables would be such that if two variables X_i and X_j are in different groups, then their regression sums of squares are additive. That is, that if the reduction in *RSS* due to adding variable X_i to a previously selected subset of variables is S_i, and the corresponding reduction if variable X_j is added instead is S_j, then the reduction when both are added is $S_i + S_j$.

If we have a complete set of say 100 variables for which we want to find good subsets of 5 variables, we may be able to break the variables into groups containing say 21, 2, 31, 29 and 17 variables. Within each group, except the one with only 2 variables, the best subsets of 1, 2, 3, 4 and 5 variables would be found. All combinations of 5 variables made up from those selected from the separate groups would then be searched to find the best subsets. Thus a subset of 5 variables might be made up of the best subset of 3 variables from one group plus single variables from two other groups.

To do this we must first find when regression sums of squares are additive. Is it only when the variables are orthogonal? Let us consider two variables X_i and X_j. Their individual regression sums of squares are the squares of the lengths of the projections of $\mathbf{Y}$ on $\mathbf{X}_i$ and $\mathbf{X}_j$ respectively, i.e.

$$S_i = (\mathbf{X}_i'\mathbf{Y})^2/\mathbf{X}_i'\mathbf{X}_i$$
$$S_j = (\mathbf{X}_j'\mathbf{Y})^2/\mathbf{X}_j'\mathbf{X}_j.$$

Let $S_{i.j}$ be the incremental regression sum of squares when variable X_i is added after X_j. Then

$$S_{i.j} = (X_{i.j}'Y_{.j})^2/X_{i.j}'X_{i.j}$$

where $X_{i.j}$ is that part of $\mathbf{X}_i$ which is orthogonal to $\mathbf{X}_j$, i.e. the vector of residuals after X_i is regressed against X_j, i.e. $X_{i.j} = X_i - b_{ij}X_j$ where b_{ij} is the LS regression coefficient, and $\mathbf{Y}_{.j}$ is similarly that part of Y orthogonal to X_j, where b_{yj} is the regression coefficient. Then

$$S_{i.j} = \frac{[(\mathbf{X}_i - b_{ij}\mathbf{X}_j)'(\mathbf{Y} - b_{yj}\mathbf{X}_j)]^2}{(\mathbf{X}_i - b_{ij}\mathbf{X}_j)'(\mathbf{X}_i - b_{ij}\mathbf{X}_j)} = \frac{(\mathbf{X}_i'\mathbf{Y} - b_{ij}\mathbf{X}_j'\mathbf{Y})^2}{\mathbf{X}_i'\mathbf{X}_i - b_{ij}\mathbf{X}_j'\mathbf{X}_i}.$$

Now we introduce direction cosines r_{ij}, r_{iy}, r_{jy}, which are the same as correlation coefficients if the variables have zero means, where

$$r_{ij} = \mathbf{X}_i'\mathbf{X}_j/(\|\mathbf{X}_i\| \, \|\mathbf{X}_j\|)$$
$$r_{iy} = \mathbf{X}_i'\mathbf{Y}/(\|\mathbf{X}_i\| \, \|\mathbf{Y}\|)$$
$$r_{jy} = \mathbf{X}_j'\mathbf{Y}/(\|\mathbf{X}_j\| \, \|\mathbf{Y}\|)$$

where the norms are in the L_2-sense. Now we can write

$$S_{i.j} = S_i \frac{(1 - r_{ij}r_{jy}/r_{iy})^2}{1 - r_{ij}^2}. \tag{3.7}$$

Thus $S_{i.j} = S_i$, i.e. the regression sums of squares are additive, when

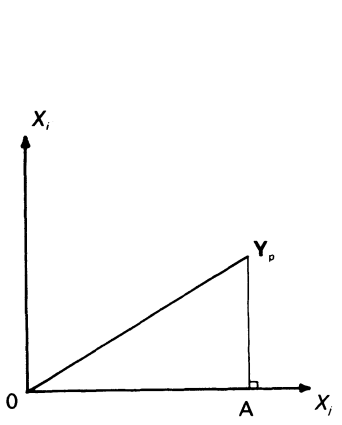
$$1 - r_{ij}^2 = (1 - r_{ij}r_{jy}/r_{iy})^2.$$

A little rearrangement shows that this condition is satisfied either when $r_{ij} = 0$, i.e. $\mathbf{X}_i$ and $\mathbf{X}_j$ are orthogonal, or when

$$r_{ij} = 2r_{iy}r_{jy}/(r_{iy}^2 + r_{jy}^2). \tag{3.8}$$

Now let us look at what this solution means graphically. In Fig. 3.1, $\boldsymbol{OY}_p$ represents the projection of vector $\mathbf{Y}$ on the plane defined by variables X_i and X_j. OA is the projection of $\mathbf{Y}_p$ on to direction $\mathbf{X}_j$ and hence its square is S_j. Similarly the square of the length of AY_p is the incremental sum of squares, $S_{i.j}$, when variable X_i is added after X_j. Now let us work out where X_i must be in this plane so that the projection of $\mathbf{Y}_p$ on to $\mathbf{X}_i$ is of length equal to AY_p. Draw a circle with centre at O and radius equal to the length of AY_p, then draw tangents to the circle from the point Y. There are two such tangents, the one to point B represents the solution (3.8), while the other shown by a broken line in Fig. 3.2 is the case in which $\mathbf{X}_i$ and $\mathbf{X}_j$ are orthogonal. Additional solutions can be obtained by reversing the directions of $\mathbf{X}_i$ or $\mathbf{X}_j$ or both.

We have only shown so far that if variables X_i and X_j satisfy (3.8) then their regression sums of squares are additive. Does this also apply to three variables X_i, X_j and X_k for which (3.8) is satisfied for all three pairs of variables? In other words, is the regression sum of squares for all three variables equal to the sum of their individual sums of squares? The answer is yes. This can be shown by using the methods above to derive say the incremental sum of squares for variable X_i after X_j and X_k, and then substituting the three conditions



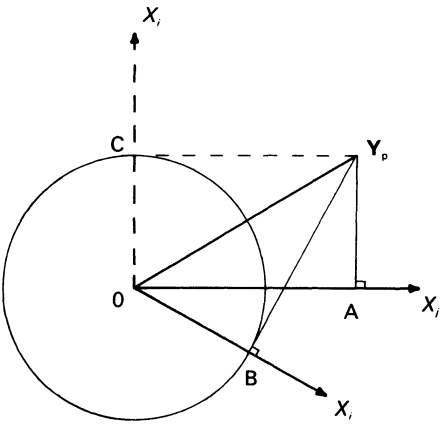


Fig. 3.1 (left) and 3.2 (right). *Illustrating situations in which regression sums of squares are additive.*

for r_{ij}, r_{ik} and r_{jk} obtained from (3.8) for pairwise additivity. This method of proof is extremely tedious and it would seem that there should be a simpler geometric argument which will allow the result to be extended recursively to any number of variables which either satisfy (3.8) or are orthogonal for all pairs. We note that in the two variables case illustrated in Fig.3.2, $\boldsymbol{OY}_p$ is the diameter of a circle on which A and B both lie. In the three-variable case, all of the points of projection lie on a sphere. Note also that if either X_i or X_j, but not both, is reflected about $\boldsymbol{OY}_p$ then the two **X**-vectors are orthogonal.

This method is as yet an untried suggestion for reducing the computational load when there are large numbers of variables from which to select. In practice satisfying (3.8) would have to be replaced with approximate satisfaction of the condition, and limits found for the deviation from perfect additivity of the regression sums of squares. The author is grateful to K. Ruben Gabriel for stimulating discussion of this idea, but leaves it to others to develop further.

3.9 Ridge regression and other alternatives

One technique which has attracted a considerable amount of interest is the ridge regression technique of Hoerl and Kennard (1970a, b). They suggested that, using all the available variables, biased

estimators, $b(d)$, of the regression of coefficients may be obtained using

$$b(d) = (X'X + dI)^{-1}X'Y \tag{3.9}$$

for a range of positive values of the scalar d. They recommended that the predictor variables should first be standardized to have zero mean and so that the sum of squares of elements in any column of X should be one, i.e. that $X'X$ should be replaced with the correlation matrix. $b(d)$ is then plotted against d; this plot was termed the 'ridge trace'. Visual examination of the trace usually shows some regression coefficients which are 'stable', that is they only change slowly, and others which either rapidly decrease or change sign. The latter variables are then deleted.

The basic justification for ridge regression is similar to that for subset selection of trading off bias against variance. For small values of d the amount of bias may be very small while the reduction in variance is very substantial. This applies particularly when the SSP-matrix is very ill-conditioned. To understand better what is happening when we use ridge estimation, it is useful to look at the singular-value decomposition (s.v.d.)

$$\begin{array}{cccccc} X & = & U & \Lambda & V \\ n \times k & & n \times k & k \times k & k \times k \end{array}$$

where the columns of U and V are orthogonal and normalized, that is $U'U = I$ and $V'V = I$, and Λ is a diagonal matrix with the singular values on the diagonal. Then

$$X'X = V'\Lambda^2 V.$$

In terms of the s.v.d., the biased estimates of the regression coefficients are

$$b(d) = V'(\Lambda^2 + dI)^{-1}\Lambda UY$$

with covariance matrix

$$\begin{aligned} &= \sigma^2 V'(\Lambda^2 + dI)^{-2}\Lambda^2 V \\ &= \sigma^2 V' \operatorname{diag}\{\lambda_i^2/(\lambda_i^2 + d)^2\} V, \end{aligned}$$

where λ_i, $i = 1, 2, \ldots, k$ are the singular values. The smallest singular values dominate the variance, but adding on d reduces their contribution substantially. For instance, for Jeffers' (1967) example on the compressive strength of pitprops, the singular values range from 2·05

to 0·20 so that for the LS estimator ($d = 0$), the covariance matrix of the standardized regression coefficient is

$$\sigma^2 V' \begin{bmatrix} 0{\cdot}237 & & & & \\ & 0{\cdot}421 & & & \\ & & \cdots & & \\ & & & 24{\cdot}4 & \\ & & & & 25{\cdot}6 \end{bmatrix} V,$$

whereas the covariance matrix for the ridge estimators when $d = 0{\cdot}05$ is

$$\sigma^2 V' \begin{bmatrix} 0{\cdot}232 & & & & \\ & 0{\cdot}403 & & & \\ & & \cdots & & \\ & & & 4{\cdot}95 & \\ & & & & 4{\cdot}02 \end{bmatrix} V.$$

There has been a very substantial amount of literature on the use of ridge regression (and on other biased estimators) when all k variables are retained. See for instance Lawless (1978), Hocking *et al.* (1976), Lawless and Wang (1976), Hoerl *et al.* (1975), McDonald and Galarneau (1975), Vinod and Ullah (1981), and the references given in section 1.3. In comparison, no attention has been paid to the idea of using the ridge trace to reduce the number of variables. We shall see later that the use of ridge regression and subset selection can not only reduce variance but can sometimes also reduce the bias which the selection introduces.

One situation in which ridge regression will not perform well when used for subset selection is when Y is strongly related to some linear combination of the predictor variables, such as $X_1 - X_2$, but has only small correlations with the individual variables. In the example given in section 3.2, ridge regression leads to the deletion of both X_1 and X_2. For this case, the squares of the singular values are 1·999 998 95, 1·0 and 0·000 001 05. The row of V corresponding to the smallest singular value is almost exactly equal to $(X_1 - X_2)/\sqrt{2}$, and once d exceeds about 10^{-5}, the regression coefficient for X_3 dominates.

If, instead of defining our biased estimators as in (3.9), we had used

$$\tilde{b}(d) = (1 + d)(\boldsymbol{X}'\boldsymbol{X} + d\boldsymbol{I})^{-1}\boldsymbol{X}'\mathbf{Y},$$

where the X- and Y-variables are normalized as before, then as $d \to \infty$, the individual elements of $\tilde{b}(d)$ tend to the regression coefficients which are obtained if Y is regressed separately against each X-variable one at a time. This means that ridge regression used as a selection technique will tend to select those variables which both (i) yield regression coefficients with the same signs in single variable regressions and with all variables in the regression, and (ii) which show up early in forward selection.

There are two popular choices for the value of d in (3.9). These are

$$d = k\hat{\sigma}/(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}) \tag{3.10}$$

and

$$d = k\hat{\sigma}^2/R^2 \tag{3.11}$$

where $\hat{\boldsymbol{\beta}}$ is the vector of LS regression coefficients, and $\hat{\sigma}^2$ is the estimated residual variance, when the X- and Y-variables have been scaled to have zero means and unit standard deviations, and R^2 is the usual coefficient of determination for LS regression. If the true values for σ and $\boldsymbol{\beta}$ are known, then (3.10) minimizes the mean squared error of $\tilde{\boldsymbol{\beta}}$ (Hoerl, Kennard and Baldwin, 1975). Lawless and Wang (1976) suggested estimator (3.11). An algorithm for choosing d which can use either (3.10) or (3.11) has been given by Lee (1987).

Suppose we write $\alpha = \boldsymbol{V}\boldsymbol{\beta}$ so that our regression model is

$$\mathrm{Y} = \boldsymbol{X}\boldsymbol{V}'\alpha + \varepsilon,$$

where the columns of $\boldsymbol{X}\boldsymbol{V}'$ are orthogonal. The ridge estimates of the elements of α are

$$\alpha_i(d) = \frac{\lambda_i^2}{\lambda_i^2 + d}\hat{\alpha}_i,$$

where $\hat{\alpha}_i$ is the LS estimate. This can be generalized by replacing d with d_i. Hocking, Speed and Lynn (1976) show that the values of d_i which minimize either the mean squared error

$$E(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})$$

or the mean squared error of prediction

$$E(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})$$

are given by

$$d_i = \sigma^2/\alpha_i^2. \tag{3.12}$$

Of course, both σ^2 and α_i^2 will be unknown in practice. If α_i is replaced first with the LS estimate, $\hat{\alpha}_i$, and then with successive ridge estimates, $\alpha_i(d_i)$, Hemmerle (1975) has shown that these estimates converge to

$$\alpha_i^* = \begin{cases} 0 & \text{if } t_i^2 \leqslant 4 \\ [1/2 + (1/4 - 1/t_i^2)^2 \hat{\alpha}_i & \text{otherwise} \end{cases} \tag{3.13}$$

where $t_i = \hat{\alpha}_i / \text{s.e.}(\hat{\alpha}_i)$. That is, t_i is the usual t-statistic in a multiple regression; in this case it is the t-value for the ith principal component of $\boldsymbol{X}$. Thus the application of an iterative ridge regression procedure can be equivalent to a subset selection procedure applied to principal components, followed by some shrinkage of the regression coefficients.

A further ridge estimator which is of this kind is

$$\alpha_i^* = \begin{cases} 0 & \text{if } t_i^2 \leqslant 1 \\ (1 - 1/t_i^2)\hat{\alpha}_i & \text{otherwise.} \end{cases} \tag{3.14}$$

This has been proposed by Obenchain (1975), Hemmerle and Brantle (1978) and Lawless (1978). Lawless (1981) has compared a number of these ridge estimators.

If Y is predicted well by a linear combination of a small number of eigenvectors, then these eigenvectors can be regarded as new variables. Unfortunately though, these eigenvectors are linear combinations of all the variables, which means that all of the variables will need to be measured to use the predictor. This may be defeating the purpose of selecting a subset.

3.10 Some examples and recommendations

Example 3.1

The data for this example are from a paper by Biondini, Simpson and Woodley (1977). In a cloud-seeding experiment, daily target rainfalls were measured as the dependent variable. An equation was required for predicting the target rainfall on the days on which clouds were seeded, by using equations developed from the data for days on which clouds were not seeded. If seeding affected rainfall then the actual rainfalls on the seeded days would be very different from those predicted. Five variables were available for use, but by includ-

Table 3.8 *Florida cloud-seeding data*

Date	X_1	X_2	X_3	X_4	X_5	Y
1 July 1971	2·0	0·041	2·70	2	12	0·32
15 July 1971	3·0	0·100	3·40	1	8	1·18
17 July 1973	3·0	0·607	3·60	1	12	1·93
9 Aug. 1973	23·0	0·058	3·60	2	8	2·67
9 Sept. 1973	1·0	0·026	3·55	0	10	0·16
25 June 1975	5·3	0·526	4·35	2	6	6·11
9 July 1975	4·6	0·307	2·30	1	8	0·47
16 July 1975	4·9	0·194	3·35	0	12	4·56
18 July 1975	12·1	0·751	4·85	2	8	6·35
24 July 1975	6·8	0·796	3·95	0	10	5·74
30 July 1975	11·3	0·398	4·00	0	12	4·45
16 Aug. 1975	2·2	0·230	3·80	0	8	1·16
28 Aug. 1975	2.6	0·136	3·15	0	12	0·82
12 Sept. 1975	7·4	0·168	4·65	0	10	0·28

ing linear, quadratic and interaction terms, the total number of variables was increased to 20. There were 58 observations in all, but these included those on seeded days. Also, the authors decided to subclassify the days according to radar echoes. Table 3.8 contains the data on the five variables and the rainfall for the largest subclassification of the data which contained 14 observations.

Table 3.9 shows the results obtained by using forward selection, sequential replacement and exhaustive search algorithms on this data. The variables X_6 to X_{20} were, in order, X_1^2 to X_5^2, X_1X_2, X_1X_3, X_1X_4, X_1X_5, X_2X_3, X_2X_4, X_2X_5, X_3X_4, X_3X_5, X_4X_5. In this case, sequential replacement has only found one subset which fits better than those found using forward selection. We see that neither forward selection nor sequential replacement found the subset of variables 9, 17 and 20 which gives a much better fit than variables 14, 15 and 17. In fact, the five best-fitting subsets of three variables (see Table 3.11) all fit much better than those found by forward selection and sequential replacement.

We note also that the variables numbered from 11 upwards all involve products of two different original variables, i.e. they are interactions. The first single variables selected by forward selection and sequential replacement do not appear until we reach subsets of five variables, and then only variable X_1 or its square; single variables

Table 3.9 *RSS's for subsets of variables for the data of Table 3.8. The numbers in brackets are the numbers of the selected variables*

No. of variables	*Forward selection*	*Sequential replacement*	*Exhaustive search*
Const.	72·29	72·29	72·29
1	26·87 (15)	26·87 (15)	26·87 (15)
2	21·56 (14, 15)	21·56 (14, 15)	21·56 (14, 15)
3	19·49 (14, 15, 17)	19·49 (14, 15, 17)	12·61 (9, 12, 20)
4	11·98 (12, 14, 15, 17)	11·98 (12, 14, 15, 17)	11·49 (9, 10, 17, 20)
5	9·05 (6, 12, 14, 15, 17)	8·70 (1, 12, 15, 17, 19)	6·61 (1, 2, 6, 12, 15)

start appearing from subsets of three variables in the best subsets found from the exhaustive search.

How well does the Efroymson algorithm perform? That depends upon the values used for F_d and F_e. If F_e is greater than 2·71 it stops with only one variable selected, that is variable 15. If F_e is between 1·07 and 2·71 then the two-variable subset of variables 14 and 15 is selected. If F_e is less than 1·07, the algorithm proceeds as for forward selection. To find any of the good subsets of three variables it would need to do some deletion. For instance, after the subset of variables 14, 15 and 17 has been found, a value of F_d greater than 2·74 is needed to delete variable 14. This would mean that $F_d > F_e$, and this results in indefinite cycling.

Now let us look at the Hoerl and Kennard method. The ridge trace for these data is shown in Fig. 3.3. We will try to apply the criteria of Hoerl and Kennard, though these were not precisely spelt out in their paper. They would probably retain variables 3, 6, 8 and 14, which have consistently large regression coefficients for all values of d. We would certainly not consider variables 10, 11 or 19, as their regression coefficients change sign. The above subset of four variables gives an $RSS = 42{\cdot}96$; there are many subsets of four variables which fit much better than this. Variables 18 and 20 are also interesting in that their regression coefficients are small initially but are the largest for large d.

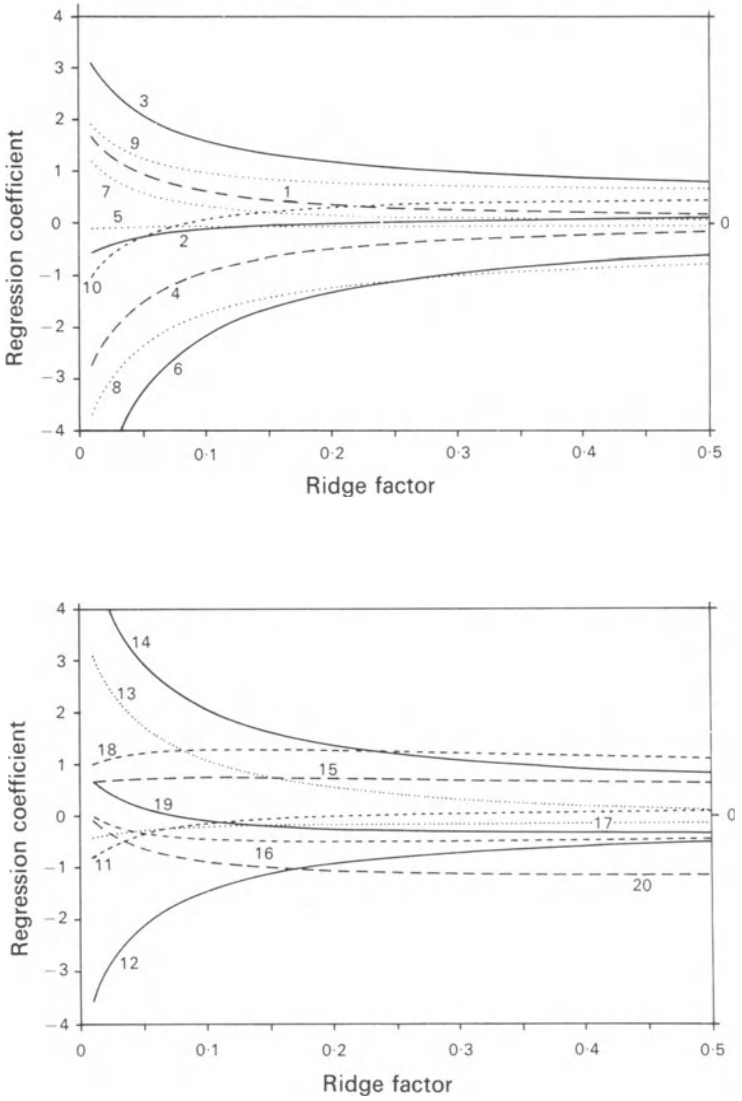


Fig. 3.3 *Ridge trace for the CLOUDS data set. Upper figure is for variables* X_1 *to* X_{10}*, lower figure for variables* X_{11} *to* X_{20}.

Table 3.10

Subset of 3 variables	*RSS*	*Subset of 4 variables*	*RSS*	*Frequency*
9 17 20	12·61	9 10 17 20	11·49	41
5 10 11	16·29	5 10 11 16	11·85	3
14 15 17	19·49	12 14 15 17	11·98	52
3 8 11	20·34	3 8 11 19	17·70	4

A method which is usually less expensive than the exhaustive search is that of using a replacement algorithm starting from randomly chosen subsets of variables. Using 100 random starting subsets of three variables gave the frequencies of subsets of three and four variables shown in Table 3.10. We note that the most frequently occurring stationary subset was not that which gave the best fit, though in this case the use of random starts has found the best-fitting subsets of both three and four variables.

So far we have only looked at the best subsets of each size found by these various methods. Table 3.11 lists the five best-fitting subsets of each size from one to five variables. We notice that in most cases there is very little separating the best from the second or even the fifth-best. In cases such as this we can expect very substantial selection bias if we were to calculate the LS regression coefficients. This will be examined in more detail in Chapter 5.

At this stage we will make only a few minor comments on these results as the topic of inference is to be treated in the next chapter. We notice that there is a very big drop in *RSS* from fitting a constant to fitting one variable, either $X_{15} = X_2X_3$ or $X_{11} = X_1X_2$, and only a gradual decline after that. This suggests that we may not be able to do better than use one of these two variables as the predictor of Y. A further practical comment is that rainfalls (variable Y is rainfall) have very skew distributions and it is common practice to use a transformation such a cube-root or logarithm. It looks as if such a transformation would give improved fits in this case.

Example 3.2

This is the STEAM data set of Chapter 2 (see Table 2.2 for the source). Table 3.12 shows how well forward selection, backward elimination and sequential replacement performed in relation to the

Table 3.11 *Five best-fitting subsets of sizes from one to five variables for the cloud-seeding data*

No. of variables	*RSS*	*Variables*				
1	26·87	15				
	27·20	11				
	32·18	2				
	34·01	7				
	42·99	17				
2	21·56	15	14			
	21·81	15	1			
	22·29	15	12			
	22·73	15	6			
	23·98	15	13			
3	12·61	9	17	20		
	15·56	2	9	20		
	16·12	9	15	20		
	16·29	5	10	11		
	17·24	7	9	20		
4	11·49	9	10	17	20	
	11·63	5	9	17	20	
	11·77	8	9	17	20	
	11·85	5	10	11	16	
	11·97	2	9	10	20	
5	6·61	1	2	6	12	15
	8·12	9	12	14	15	20
	8·44	1	2	12	13	15
	8·70	1	12	15	17	19
	8·82	1	3	6	8	13

best-fitting subsets of each size obtained by exhaustive search. We note that for this data set there are more observations than variables and backward elimination is feasible.

For this set of data, forward selection, backward elimination and sequential replacement have all performed fairly well, with sequential replacement having found the best-fitting subsets of all sizes. We notice that the *RSS* is almost constant for three or more variables for the best-fitting subsets of each size. This suggests that the best predictor is one containing not more than three variables. The *RSS* with all nine variables in the model is 4·87, and this has 15 degrees of freedom. Dividing the *RSS* by its number of degrees of freedom

Table 3.12 *RSS's for subsets of variables selected using various procedures for the STEAM data. The numbers in brackets are the numbers of the selected variables*

No. of variables	*Forward selection*	*Backward elimination*	*Sequential replacement*	*Exhaustive search*
Const.	63·82	63·82	63·82	63·82
1	18·22 (7)	18·22 (7)	18·22 (7)	18·22 (7)
2	8·93 (1, 7)	8·93 (1, 7)	8·93 (1, 7)	8·93 (1, 7)
3	7·68 (1, 5, 7)	7·68 (1, 5, 7)	7·34 (4, 5, 7)	7·34 (4, 5, 7)
4	6·80 (1, 4, 5, 7)	6·93 (1, 5, 7, 9)	6·80 (1, 4, 5, 7)	6·80 (1, 4, 5,7)
5	6·46 (1, 4, 5, 7, 9)	6·54 (1, 5, 7, 8, 9)	6·41 (1, 2, 5, 7, 9)	6·41 (1, 2, 5, 7, 9)

gives a residual variance estimate of 0·32 and this is of the same order of magnitude as the drops in *RSS* from three to four variables, and from four to five variables. If we regressed the LS residuals from fitting the first three variables against a variable which consisted of a column of random numbers, we would expect a drop in *RSS* of about 0·32.

The five best-fitting subsets of one, two and three variables are shown in Table 3.13. We see that there is no competition in the case of a single variable, that there are three close subsets of two variables, and at least five close subsets of three variables. If LS estimates of regression coefficients were used in a predictor, very little bias would result if the best subset of either one or two variables were used. There would be bias if the best subset of three variables were used.

Example 3.3

The data for this example are for the DETROIT data set (see Table 2.2 for the source). The data are of annual numbers of homicides in Detroit for the years 1961–73 inclusive, and hence contain 13 observations. As there are 11 predictor variables available, there is only one degree of freedom left for estimating the residual variance if a constant is included in the model. Table 3.14 shows the

Table 3.13 *Five best-fitting subsets of one, two and three variables for the STEAM data*

No. of variables	*RSS*	*Variables*
1	18·22	7
	37·62	6
	45·47	5
	49·46	3
	53·88	8
2	8·93	1 7
	9·63	5 7
	9·78	2 7
	15·60	4 7
	15·99	7 9
3	7·34	4 5 7
	7·68	1 5 7
	8·61	1 7 9
	8·69	1 4 7
	8·71	5 7 8

Table 3.14 *RSS's for subsets of variables for the DETROIT data set. Th numbers in brackets are the numbers of the selected variables*

No. of variables	*Forward selection*	*Backward elimination*	*Sequential replacement*	*Exhaustive search*
Const.	3221·8	3221·8	3221·8	3221·8
1	200·0 (6)	680·4 (11)	200·0 (6)	200·0 (6)
2	33·83 (4, 6)	134·0 (4, 11)	33·83 (4, 6)	33·83 (4, 6)
3	21·19 (4, 6, 10)	23·51 (3, 4, 11)	21·19 (4, 6, 10)	6·77 (2, 4, 11)
4	13·32 (1, 4, 6, 10)	10·67 (3, 4, 8, 11)	13·32 (1, 4, 6, 10)	3·79 (2, 4, 6, 11)
5	8·20 (1, 2, 4, 6, 10)	8·89 (3, 4, 7, 8, 11)	2·62 (1, 2, 4, 9, 11)	2·62 (1, 2, 4, 9, 11)
6	2·38 (1, 2, 4, 6, 10, 11)	6·91 (3, 4, 7, 8, 9, 11)	1·37 (1, 2, 4, 6, 7, 11)	1·37 (1, 2, 4, 6, 7, 11)

performance of forward selection, sequential replacement, backward elimination and exhaustive search on these data.

In this example, none of the 'cheap' methods has performed well, particularly in finding the best-fitting subsets of three and four variables. For larger subsets, sequential replacement has been successful. Backward elimination dropped variable number 6 first, and this appears in many of the better-fitting subsets, while it left in variables 3, 8 and 10, which appear in few of the best subsets, until a late stage. In the experience of the author, the 'cheap' methods of variable selection usually perform badly when the ratio of the number of observations to the number of variables is less than or close to 1. In such cases, the best-fitting subset of p variables often does not contain the best-fitting subset of $p-1$ variables, sometimes they have no variables in common, and methods which add or drop one variable at a time either cannot find the best-fitting subsets or have difficulty in finding them. One remarkable feature of this data set is that the first variable selected by forward selection, variable 6, was the first variable deleted in the backward elimination.

Table 3.15 shows the five best-fitting subsets of each size up to five variables. The subsets of three variables look extraordinary. Not merely is the subset of variables 2, 4 and 11 far superior to any other subset of three variables, but no subsets of one or two of these three variables appear in the five best-fitting subsets of one or two variables. Variable 2 has the lowest correlation in absolute value with the dependent variable; this correlation is 0.21, the next smallest is 0.55, and most of the others exceed 0.9 in absolute value. The *RSS*'s for subsets from these variables are

Variable	2	4	11	2,4	2,11	4,11
RSS	3080	1522	680	1158	652	134

These all compare very unfavourably with the better-fitting subsets of one and two variables. Is this just a case of remarkable over-fitting?

In view of the lack of competition from other subsets, the bias in the regression coefficients is likely to be very small if this subset of three variables is used for prediction with LS estimates for the parameters. The three variables in this case are % unemployed (variable 2), number of handgun licences per 100 000 population

Table 3.15 *Five best-fitting subsets of each size from one to five variables for the DETROIT data*

No. of variables	*RSS*	*Variables*				
1	200·0	6				
	227·4	1				
	264·6	9				
	277·7	8				
	298·7	7				
2	33·83	4	6			
	44·77	2	7			
	54·45	1	9			
	55·49	5	6			
	62·46	3	8			
3	6·77	2	4	11		
	21·19	4	6	10		
	23·05	1	4	6		
	23·51	3	4	11		
	25·04	4	6	11		
4	3·79	2	4	6	11	
	4·58	1	2	4	11	
	5·24	2	4	7	11	
	5·41	2	4	9	11	
	6·38	2	4	8	11	
5	2·62	1	2	4	9	11
	2·64	1	2	4	6	11
	2·75	1	2	4	7	11
	2·80	2	4	6	7	11
	3·12	2	4	6	9	11

(variable 4) and average weekly earnings (variable 11). All three regression coefficients are positive.

Example 3.4

The data for this example is the POLLUTE data set (see Table 2.2 for the source). The dependent variable is an age-adjusted mortality rate per 100 000 population. The data are for 60 metropolitan statistical areas in the USA. The predictor variables incude socio-economic, meteorological and pollution variables.

Table 3.16 shows the subsets of variables selected by forward selection, backward elimination, sequential replacement and

Table 3.16 *RSS's for subsets of variables for the POLLUTE data. The numbers in brackets are the numbers of the selected variables*

No. of variables	*Forward selection*	*Backward elimination*	*Sequential replacement*	*Exhaustive search*
Const.	228 308	228 308	228 308	228 308
1	133 695 (9)	133 695 (9)	133 695 (9)	133 695 (9)
2	99 841 (6, 9)	127 803 (9, 12)	99 841 (6, 9)	99 841 (6, 9)
3	82 389 (2, 6, 9)	91 777 (9, 12, 13)	82 389 (2, 6, 9)	82 389 (2, 6, 9)
4	72 250 (2, 6, 9, 14)	78 009 (6, 9, 12, 13)	69 154 (1, 2, 9, 14)	69 154 (1, 2, 9, 14)
5	64 634 (1, 2, 6, 9, 14)	69 136 (2, 6, 9, 12, 13)	64 634 (1, 2, 6, 9, 14)	64 634 (1, 2, 6, 9, 14)
6	60 539 (1, 2, 3, 6, 9, 14)	64 712 (2, 5, 6, 9, 12, 13)	60 539 (1, 2, 3, 6, 9, 14)	60 539 (1, 2, 3, 6, 9, 14)

exhaustive search. In this case, which is our only one with substantially more observations than variables, sequential replacement has found the best-fitting subsets of all sizes, while forward selection has only failed in the case of the subset of four variables where the subset selected is the second-best of that size. Backward elimination has performed poorly, having dropped variable 14 very early, in fact in going from 10 to 9 variables, and of course from all smaller subsets.

In this case we have a good estimate of the residual variance. The *RSS* with all 15 variables and a constant in the model is 53 680 which has 44 degrees of freedom giving a residual variance estimate of 1220. The drop in *RSS* from the best subset of four variables to the best subset of five variables is less than four times the residual variance, which is not very impressive, and suggests that the best subset for prediction should probably be that with only four variables.

From Table 3.17 showing the five best-fitting subsets of each size, we see that in this case there is one dominant variable, variable number 9. This variable is the percentage of the population which is nonwhite. Without this variable, the best fitting subsets are

Variables	6	1, 14	1, 4, 14	1, 4, 7, 14	1, 2, 4, 11, 14	1, 2, 3, 4, 11, 14
RSS	168 696	115 749	102 479	92 370	87 440	81 846

Table 3.17 *Five best-fitting subsets of each size from one to five variables for the POLLUTE data*

No. of variables	*RSS*	*Variables*				
1	133 695	9				
	168 696	6				
	169 041	1				
	186 716	7				
	186 896	14				
2	99 841	6	9			
	103 859	2	9			
	109 203	9	14			
	112 259	4	9			
	115 541	9	10			
3	82 389	2	6	9		
	83 335	1	9	14		
	85 242	6	9	14		
	88 543	2	9	14		
	88 920	6	9	11		
4	69 154	1	2	9	14	
	72 250	2	6	9	14	
	74 666	2	5	6	9	
	76 230	2	6	8	9	
	76 276	1	6	9	14	
5	64 634	1	2	6	9	14
	65 660	1	2	3	9	14
	66 555	1	2	8	9	14
	66 837	1	2	9	10	14
	67 622	2	4	6	9	14

These are all very inferior to the best subsets including variable 9. Some of the other variables most frequently selected are annual rainfall (variable 1), January (2) and July temperature (3), median number of years of education (6), and sulphur dioxide concentration (14).

As the values of the *RSS*'s are relatively close together for subsets of the same size, it could appear that there is close competition for selection. However, if we compare the differences between *RSS*'s with our residual variance estimate of 1220 we see that only a small number are close.

3.10.1 Conclusions and recommendations

It could be argued that the above examples, except for the POLLUTE data set, are not very typical in having low ratios of numbers of observations to variables; however these ratios are fairly representative of the author's experience as a consulting statistician. What is perhaps unrepresentative is the low number of available predictor variables in these examples. In meteorology, users of variable-selection procedures often have a choice from more than a hundred variables, while in using infra-red spectroscopy as a substitute for chemical analysis the author has occasionally encountered examples with over a thousand variables.

In our examples, the 'cheap' methods considered, forward selection, backward elimination and sequential replacement, have usually not found all of the best-fitting subsets and have sometimes fared poorly. Of these three methods, sequential replacement has been the most successful. A more extensive comparison has been made by Berk (1978b) who has examined nine published cases, though all had more observations than variables. In three of the nine cases, forward selection and backward elimination found the best-fitting subsets of all sizes. In most other cases, the differences between the *RSS*'s for the subsets selected by forward selection and backward elimination and the best subsets were very small.

An exhaustive search can take a very long while. The following times were recorded for the four principal procedures considered here on a set of data from the infra-red analysis of wood samples. There were 25 variables, which were reflectances at different wavelengths, and 72 observations. The best-fitting subsets of 6 variables plus a constant were sought and the best 5 subsets of each size were being found. The times were on a CROMEMCO Z-2D microcomputer based on a Z80A microprocessor. The procedures were used in the order listed which meant that the subsets found by the first three procedures were available as initial bounds for the exhaustive search:

Forward selection	20 sec.
Backward elimination	113 sec.
Sequential replacement	93 sec.
Exhaustive search	2 hr. 45 min.

When the procedures were repeated but with only the single

best-fitting subset of each size being sought, the exhaustive search took 2 hours 28 minutes. This is an extreme example in two regards. First, the first three procedures all performed very badly so that the bounds were of little benefit in cutting down the amount of computation in the exhaustive search.

In general, if it is feasible to carry out an exhaustive search then that is to be recommended. As the sequential replacement algorithm is fairly fast, that can always be used first to give some indication of the maximum size of subset which is likely to be of interest. If there are more observations than variables then the *RSS* with all of the variables in the model should be calculated first. This comes out of the orthogonal reduction as the data are input and so requires no additional computation. It is useful as a guide to determine the maximum size of subset to examine.

It is recommended that several subsets of each size should be recorded. The alternative subsets are useful both for testing whether the best-fitting subset is significantly better than others (see Spjøtvoll's test in Chapter 4), or for finding the subset which may give the best predictions for a specific set of future values of the predictor variables (see Chapter 6).

When it is not feasible to cary out the exhaustive search, the use of random starts followed by sequential replacement, or of two-at-a-time replacement, can be used though there can be no guarantee of finding the best-fitting subsets. A further alternative is that of grouping variables which was discussed in section 3.8 but which requires further research.

In all cases, graphical or other methods should be used to assess the adequacy of the fit obtained. These examinations often uncover patterns in the residuals which may indicate the suitability of using a transformation, or of using some kind of weighting, or of adding extra variables such as quadratic or interaction terms. Unfortunately inference becomes almost impossible if the total subset of available predictors is augmented subjectively in this way.

CHAPTER 4

Hypothesis testing

4.1 Is there any information in the remaining variables?

Suppose that by some method we have already picked p variables, where p may be zero, out of k variables available to include in our predictor subset. If the remaining variables contain no further information which is useful for predicting the response variable then we should certainly not include any more. But how do we know when the remaining variables contain no further information? In general, we do not; we can only apply tests and take gambles based upon the outcome of those tests.

The simplest such test known to the author is that of augmenting the set of predictor variables with one or more artificial variables whose values are produced using a random number generator. When the selection procedure first picks one of these artificial variables, the procedure is stopped and we go back to the last subset containing none of the artificial variables. Let us suppose that we have reached the stage in a selection procedure when there is in fact no useful information remaining (though we would not know this in a real case), and that there are 10 remaining variables plus one artificial variable. *A priori* the chance that the artificial variable will be selected next is then 1 in 11. Hence it is likely that several useless variables will be added before the artificial variable is chosen. For this method to be useful and cause the procedure to stop at about the right place we therefore need the number of artificial variables to be quite large, say of the same order as the number of real variables. This immediately makes the idea much less attractive as doubling the number of variables increases the amount of computation required by a much larger factor.

In Table 4.1 the *RSS*'s are shown for the five best-fitting subsets of 2, 3, 4 and 5 variables for the four data sets used in examples in section 3.10. The name 'CLOUDS' indicates the cloud-seeding data.

Table 4.1 *RSS's for the five best-fitting subsets of 2, 3, 4 and 5 variables with artificial variables added. The number of asterisks equals the number of artificial variables in the subset*

No. of variables	*Data set*			
	CLOUDS	*STEAM*	*DETROIT*	*POLLUTE*
2	21·56	8·93	33·83	99 841
	21·81	9·63	44·77	103 859
	22·29	9·78	54·46	109 203
	22·41*	15·39*	55·49	112 259
	22·42*	15·60	62·46	115 541
3	12·61	7·34	6·77	82 389
	15·56	7·68	21·19	83 335
	16·12	7·81*	23·05	85 242
	16·29	8·29*	23·51	87 365*
	16·95*	8·29*	25·01*	88 543
4	5·80**	6·41**	3·79	69 137*
	6·52**	6·72*	4·08*	69 154
	8·25**	6·80	4·58	72 250
	8·25**	6·93	5·24	74 666
	8·32*	7·02	5·38*	75 607*
5	3·69**	5·91**	1·33*	61 545*
	4·08**	5·93*	1·71*	62 494*
	4·09**	6·01*	2·15*	63 285**
	4·28**	6·05***	2·62	64 505*
	4·40**	6·11*	2·64	64 549*

The numbers of added variables were

CLOUDS 5, STEAM 9, DETROIT 11 and POLLUTE 10.

The values of the artificial variables were obtained using a uniform random number generator. Except for the DETROIT data set, the artificial variables indicate that we should stop at three variables. In the case of the DETROIT data, the separation between the best-fitting and the second-best of four variables is not large, and it looks doubtful whether the subset of four variables would have been the best if we had generated more artificial variables. It is of interest to notice that there is no challenge to the best subset of three variables for the DETROIT data; a subset which we commented in section 3.10

stood out very remarkably. In the case of the STEAM data we must have some doubts about whether the first two subsets of three variables are really an improvement over the best-fitting subsets of two variables; the next three subsets containing artificial variables are very close.

We saw in section 3.3 that the widely used 'F-to-enter' statistic does not have an F-distribution or anything remotely like one. However the quantity is a useful heuristic indicator and we propose to use it and to abuse it even further by using it to compare RSS's for different sizes of subsets when the larger subset does not necessarily contain the smaller one. Let us define

$$F\text{-to-enter} = \frac{RSS_p - RSS_{p+1}}{RSS_{p+1}/(n-p-1)}$$

where RSS_p and RSS_{p+1} are the RSS's for subsets of p and $p+1$ variables, and n is the number of observations. If the model includes a constant which has not been counted in the p variables, as is usually the case, then another 1 must be subtracted from the number of degrees of freedom. Comparing the best subsets of three and four variables in Table 4.1, the values of the F-to-enter are 10·6 (CLOUDS), 2·9 (STEAM), 6·3 (DETROIT) and 10·5 (POLLUTE) with 9, 20, 8 and 25 degrees of freedom respectively for the denominators. It is interesting to note that it is only the third largest of these which may be significant at say the 5% level, though we can only guess at this on the basis of the previous test using the addition of artificial variables.

An alternative test to the above is that of Forsythe *et al.* (1973). If the true model is linear in just p out of the k available predictor variables, these p variables have been correctly selected and are the first p variables in the triangular factorization, then the $(k-p)$ projections of the dependent variables, Y, on the remainder of the space spanned by the X-variables, are uncorrelated and have variance equal to the residual variance if the true residuals are uncorrelated and homoscedastic (i.e. they all have the same variance). This is easily shown as follows. We suppose that

$$\mathbf{Y} = X_A\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where X_A consists of the first p columns of X, the true residuals in vector $\boldsymbol{\varepsilon}$ have zero mean, the same variance, σ^2, and are uncorrelated,

i.e. $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \boldsymbol{I}$ where $\boldsymbol{I}$ is an $n \times n$ identity matrix with n equal to the number of observations. Let $\boldsymbol{X}_B$ consist of the remaining $(k-p)$ columns of X, and let the orthogonal reduction of $\boldsymbol{X}$ be

$$(\boldsymbol{X}_A, \boldsymbol{X}_B) = (\boldsymbol{Q}_A, \boldsymbol{Q}_B)\boldsymbol{R},$$

where the columns of $\boldsymbol{Q}_A$ and $\boldsymbol{Q}_B$ are mutually orthogonal and normalized. Then we have

$$\begin{pmatrix} \boldsymbol{Q}_A' \\ \boldsymbol{Q}_B' \end{pmatrix} \mathbf{Y} = \begin{pmatrix} \boldsymbol{Q}_A' \\ \boldsymbol{Q}_B' \end{pmatrix} \boldsymbol{X}_A \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{Q}_A' \\ \boldsymbol{Q}_B' \end{pmatrix} \boldsymbol{\varepsilon}$$

$$= \begin{pmatrix} \boldsymbol{I} \\ 0 \end{pmatrix} \boldsymbol{R}\boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{Q}_A' \\ \boldsymbol{Q}_B' \end{pmatrix} \boldsymbol{\varepsilon},$$

where $\boldsymbol{I}$ is a $p \times p$ identity matrix.

The last $(k-p)$ projections are given by

$$\boldsymbol{Q}_B' \mathbf{Y} = \boldsymbol{Q}_B' \boldsymbol{\varepsilon}.$$

Then

$$E(\boldsymbol{Q}_B' \mathbf{Y}) = 0$$

and

$$E(\boldsymbol{Q}_B' Y Y' \boldsymbol{Q}_B) = E(\boldsymbol{Q}_B' \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \boldsymbol{Q}_B)$$
$$= \sigma^2 \boldsymbol{Q}_B' \boldsymbol{Q}_B$$
$$= \sigma^2 \boldsymbol{I}$$

where $\boldsymbol{I}$ is a $(k-p) \times (k-p)$ identity matrix. Notice that there is no need for the matrix $\boldsymbol{R}$ to be triangular; any orthogonal reduction will suffice. Also the only distributional assumptions imposed on the true residuals are that they have finite variance, the same variance for each residual, and that the residuals are uncorrelated.

If we have selected too many variables but they include all of those in the true model for $\mathbf{Y}$, then the above results still apply as the model then has zeros for some of the elements of $\boldsymbol{\beta}$.

The projections in the last $(k-p)$ positions of $\boldsymbol{Q}'\mathbf{Y}$ can be used to test the hypothesis that there is no more useful information for predicting $\mathbf{Y}$ if we are prepared to accept that the residuals from the true model are uncorrelated and homoscedastic. Alternatively, they can be used for testing the properties of the true residuals if we are satisfied with the selected variables in the model. Forsythe *et al.* (1973) suggested a test for the former as follows. First, find the maximum reduction in *RSS* which can be obtained by adding one

further variable; call this S_{max}. Then permute the last $(k-p)$ elements of $\boldsymbol{Q}'\mathbf{Y}$ many times, each time calculating the maximum reduction in *RSS* which can be achieved by adding one variable to the p already selected, and counting the number of times that S_{max} is exceeded. This can be done quite quickly and can easily be incorporated into a forward selection routine. If the remaining $(k-p)$ variables contain no further useful information then S_{max} can be regarded as a random sample from the distribution of maximum reductions in *RSS*. Hence if we carry out say 1000 permutations, the number of times that S_{max} is exceeded is equally likely to take any value from 0 to 1000.

This test assumes that the last elements of $\boldsymbol{Q}'\mathbf{Y}$ are interchangeable, which requires that they are identically distributed. In general this will not be true unless the true residuals have a normal distribution. We have only shown that the elements have the same first two moments. However, as these elements are weighted linear combinations of the true but unknown residuals, by the central limit theorem they will usually have distributions which are close to normal, particularly if the number of observations is less than say double the number of variables. Using the elements of $\boldsymbol{Q}_B$ it is possible, though tedious, to calculate higher moments for each of the last elements of $\boldsymbol{Q}'\mathbf{Y}$ in terms of the moments of the true residuals. This exercise (or a similar one for LS residuals) shows that orthogonal reduction (or LS fitting) is very effective in making residuals look normal even when the true residuals are very nonnormal.

The above test has a disadvantage in that the significance level depends upon the order the occurrence of the last $(k-p)$ variables. It will not usually be practical to use all $(k-p)!$ permutations of these variables.

Table 4.2 shows the results obtained using the four sets of data used in the examples in section 3.10. In each case the best-fitting subset of three variables was used for the subset of selected variables.

Table 4.2 *Numbers of times that the maximum reduction in RSS from adding a fourth variable was exceeded when the projections for the remaining variables were permuted*

Data set	CLOUDS	STEAM	DETROIT	POLLUTE
No. of exceedances (out of 1000)	974	847	308	348

These results all support the belief that there is no further useful information in the remaining variables in these data sets. The very high number of exceedances for the cloud-seeding data suggests that there may have already been some over-fitting in selecting the subset of three variables.

In the case of the POLLUTE data set, the best single variable to add to the first three does not give the best-fitting subset of four variables. The best-fitting subset of three variables consists of variables 2, 6 and 9 which give an $RSS = 82\,389$. Adding variable 14 reduces the RSS to 72 250, but the subset of variables 1, 2, 9 and 14 gives an $RSS = 69\,154$. As the residual variance estimate with all 15 variables included is 1220, the difference of 3096 between these two RSS's is relatively large.

It is possible for the remaining variables to contain a pair of variables which together substantially improve the prediction of Y, but which are of very little value separately. This would not be detected by the Forsythe permutation test, though a 'two variables at a time' version of the test could be carried out if this were suspected. The test as presented by Forsythe *et al.* is probably better to use as a 'carry on' rule rather than as a 'stopping' rule as when S_{max} is only exceeded a small number of times then the remaining variables most probably contain additional information, whereas when the number of exceedances is larger, the situation is uncertain.

One important feature of this permutation test is that it can be used when the number of variables exceeds the number of observations, which is precisely the situation for which it was devised.

When there are more observations than variables, there is information in the residual variation which is not used by the Forsythe permutation test. A similar test is to use the maximum F-to-enter. Thus, after p variables have been selected, the maximum F-to-enter is found for the remaining $(k-p)$ variables. It is not difficult to write down the form of the distribution of this maximum F-to-enter making the usual assumptions that the true residuals are independently and identically normally distributed plus the added assumption that the selected p variables represent the true relationship, i.e. we have not already over-fitted by adding one or more variables which by chance had moderately high correlations with the variable to be predicted. This distribution depends upon the values of the X-variables so that it is not feasible to tabulate it except perhaps for the case in which the X-variables are uncorrelated.

However, the distribution can easily be simulated for the particular data set at hand by using a random normal generator for the last $(k-p)$ projections and a random gamma generator for the *RSS* with all the variables in the model. As the *F*-to-enter statistic is dimensionless, it is not necessary to estimate the residual variance, the random projections can have unit variance and the sum of squares of residuals can be sampled from a chi-square distribution without applying any scaling. However such a test is only of whether a single variable can significantly reduce the *RSS*, and will again fail to detect cases where two or more variables collectively can produce a significant reduction. The author can supply a FORTRAN routine for such a test, but it was not thought to have sufficient value to include here.

A simple alternative to the *F*-to-enter test is one which is often called the lack-of-fit test. If we have n observations and have fitted a linear model containing p out of k variables plus a constant, then the difference in *RSS* between fitting the p variables and fitting all k variables, $RSS_p - RSS_k$, can be compared with RSS_k giving the lack-of-fit statistic

$$\text{lack-of-fit } F = \frac{(RSS_p - RSS_k)/(k-p)}{RSS_k/(n-k-1)}.$$

Table 4.3 shows values of the lack-of-fit statistic after fitting three variables for each of our four examples. If the subset of three variables had been chosen *a priori*, the usual conditions of independence, constant variance and normality are satisfied, and there is no useful information in the remaining variables, then the lack-of-fit statistic is sampled from an *F*-distribution with the numbers of degrees of freedom shown in the table. None of these values is significant at

Table 4.3 *Lack-of-fit statistic after fitting the best-fitting subsets of three variables. Degs = degrees, Num. = numerator, Denom. = denominator*

Data set	*No. of variables in subset*	*Degs of freedom*		*Lack-of-fit*
		Num.	*Denom.*	*F*
CLOUDS	3	10	0	n.a.
STEAM	3	6	15	1.27
DETROIT	3	8	1	9.30
POLLUTE	3	12	44	1.96

the 5% level, though that for the POLLUTE data set is very close; it is at the 5.3% point in the tail of its F-distribution. The value of the lack-of-fit statistic for the DETROIT data set is large, but it needs to be very much larger to become significant as there is only one degree of freedom for the RSS with all 11 variables in the model.

A further statistic which can be used to test for no further information is the coefficient of determination, R^2. This has usually been employed though to test whether there is any predictive power in any of the variables. R^2 can be defined as

$$R^2 = 1 - RSS_p/RSS_1$$

provided that a constant has been included in the model as the first variable, so that RSS_1 is the sum of squares of Y about its mean. There is some ambiguity in the definition of R^2 when a constant is not fitted. Sometimes the definition above is used, in which case negative values are possible, while the total sum of squares of values of the variable is sometimes used instead of RSS_1.

The distribution of R^2 has been tabulated for a number of cases in which the response variable, Y, is normally distributed and independent of the X-variables. Diehr and Hoflin (1974), and Lawrence, Neumann and Caso (1975) have generated the distribution of R^2 using Monte Carlo methods for subset selection using respectively exhaustive search and forward selection, for uncorrelated normally distributed X-variables. Zurndorfer and Glahn (1977) and Rencher and Pun (1980) have also looked at the case of correlated X-variables using forward selection and the Efroymson algorithm respectively. It is clear from both of these last two studies that the value of R^2 tends to be higher when the X-variables are uncorrelated. Table 4.4 shows the values of the average and upper 95% points of the distribution of R^2, denoted by R_s^2 and R_{95}^2, obtained by Rencher and Pun for the case of uncorrelated X-variables. The number of X-variables, p, does not include the constant term which was fitted. Rencher and Pun do not state the values used for the F-to-enter and F-to-delete parameters in deriving their tables, but do say that they were reduced when necessary to force the required number of variables into subsets. The 95% points of R^2 found by Diehr and Hoflin tend to be a little higher than those found by Rencher and Pun, but this is to be expected as they used exhaustive search as their selection procedure.

Similar tables to Table 4.4 have also been produced by Wilkinson

Table 4.4 *Values of the average and upper 95% points of the distribution of R^2 for subsets selected using Efroymson's algorithm. n = sample size, k = number of available uncorrelated predictor variables. Reprinted from Rencher and Pun (1980) with permission of Rencher and the American Statistical Association.*

		Number of variables in selected subset									
		$p=2$		$p=4$		$p=6$		$p=8$		$p=10$	
n	k	R_s^2	R_{95}^2	R_s^2	R_{95}^2	R_s^2	R_{95}^2	R_s^2	R_{95}^2	R_s^2	R_{95}^2
5	5	0·784	0·981								
	10	0·900	0·991								
	20	0·952	0·995								
10	5	0·421	0·665	0·540	0·851						
	10	0·567	0·822	0·778	0·955	0·894	0·991				
	20	0·691	0·877	0·912	0·983	0·984	0·999	0·997	1·000		
	40	0·791	0·907	0·965	0·991	0·996	1·000	0·999	1·000		
20	10	0·299	0·510	0·423	0·658	0·488	0·726				
	20	0·391	0·562	0·585	0·771	0·701	0·858	0·786	0·920		
	40	0·469	0·618	0·700	0·819	0·835	0·920	0·916	0·969	0·963	0·989
30	10	0·202	0·337	0·285	0·451	0·326	0·511				
	20	0·264	0·388	0·405	0·546	0·496	0·648	0·561	0·716		
	40	0·331	0·456	0·514	0·640	0·639	0·759	0·733	0·846	0·803	0·901
40	10	0·147	0·260	0·206	0·339	0·233	0·374				
	20	0·203	0·319	0·309	0·445	0·376	0·527	0·424	0·583		
	40	0·251	0·349	0·397	0·507	0·502	0·626	0·584	0·704	0·651	0·775
50	30	0·184	0·267	0·291	0·397	0·367	0·484	0·424	0·559	0·469	0·612
	40	0·201	0·288	0·324	0·432	0·413	0·526	0·481	0·598	0·537	0·657
60	30	0·157	0·229	0·247	0·341	0·312	0·425	0·360	0·479	0·398	0·523
	40	0·169	0·249	0·272	0·372	0·348	0·457	0·406	0·515	0·454	0·569

and Dallal (1981) for R^2 for the case of forward selection, except that their tables are in terms of the number of available variables and the value used for F-to-enter.

The values of R^2 for the best-fitting subsets of three variables for our four examples are given in Table 4.5. From Table 4.4 we see that the values of R^2 for the STEAM, DETROIT and POLLUTE data sets are all significant at the 5% level. For the CLOUDS data, interpolation between four entries is necessary. It appears that the value of R^2 may be just short of the 5% point, though the tables are for the Efroymson algorithm, not for an exhaustive search.

Zirphile (1975) attempted to use extreme-value theory to derive the distribution of R^2. He makes the false assumption that the values of R^2 for the ${}^kC_p$ different subsets of p variables out of k are uncorrelated, and uses a normal distribution to approximate the distribution of R^2 for a randomly chosen subset of variables. The distribution of R^2 for a random subset when the Y-variable is uncorrelated with the X-variables is a beta distribution with

$$\text{prob}(R^2 < z) = \frac{1}{B(a,b)} \int_0^z t^{a-1}(1-t)^{b-1}\,dt$$

where $a = p/2$, $b = (n-p-1)/2$ if a constant has been included in the model but not counted in the p variable. Using the beta distribution and fitting constants to their tables, Rencher and Pun obtained the following formula for the upper $100(1-\gamma)\%$ point of the distribution of the maximum R^2 using the Efroymson algorithm as

$$R^2_\gamma = [1 + \log_e \gamma/(\log_e N)^{1.8N^{0.4}}]F^{-1}(\gamma) \tag{4.1}$$

where $N = {}^kC_p$ and $F^{-1}(\gamma)$ is the value of z such that $\text{prob}(R^2 < z) = \gamma$.

Table 4.5 *Values of R^2 for the best-fitting subsets of three variables*

Data set	n	k	p	R^2
CLOUDS	14	20	3	0·826
STEAM	25	9	3	0·885
DETROIT	13	11	3	0·998
POLLUTE	60	15	3	0·639

Values of $F^{-1}(\gamma)$ can be obtained from tables of the incomplete beta function, or from tables of the F-distribution as follows. Writing Reg_p to denote the regression sum of squares on p variables, we have

$$R^2 = Reg_p/(Reg_p + RSS_p).$$

Write

$$F = \frac{Reg_p/p}{RSS_p/(n-p-1)}$$

as the usual variance ratio for testing the significance of the subset of p variables if it had been chosen *a priori*. Then

$$R^2 = p/[p + (n-p-1)F]. \tag{4.2}$$

Thus the value of R^2 such that $\text{prob}(R^2 < z) = \gamma$ is the value of F, with p and $(n-p-1)$ degrees of freedom for the numerator and denominator respectively such that the upper tail area is γ. As the reciprocal of a variance ratio also has an F-distribution but with the degrees of freedom interchanged, because of the way in which the F-distribution is usually tabulated, we use the tables with $(n-p-1)$ and p degrees of freedom for numerator and denominator respectively and then take the reciprocal of the F-value read from the tables. The upper limit on R^2 is then obtained by substitution in (4.2), and then into (4.1).

In the case of the CLOUDS data set the above method gives $R^2_{95} = 0.878$ which means that our value is not significant at the 5% level. Bearing in mind the fact that the Rencher and Pun formula is for the Efroymson algorithm whereas we found the subset using an exhaustive search, our value for R^2 is even less significant.

4.2 Is one subset better than another?

In the first section of this chapter we were looking at ways of testing the hypothesis that $\beta_{p+1}, \ldots, \beta_k = 0$ where these β's are the regression coefficients of the variables which have not been selected. In some cases these tests were of dubious value as they require that a subset had been selected *a priori* even though the test is almost always applied using the same data as was used to select the subset. Those tests only tested whether one subset was better than another, in some sense, when one subset was completely contained in the other.

What is meant by saying that one subset is better than another?

There are many possible ways of answering this. One such way is that used by Spjøtvoll (1972a), and this is equivalent to that used by Borowiak (1981). In this, one subset is considered better than another if the regression sum of squares of the expected values of Y upon the subset of X-variables is larger for that subset; Borowiak used the complementary *RSS*. However, unlike Borowiak and most other workers, Spjøtvoll did not make the assumption that the linear model in one of the two subsets was the true model.

Let us suppose that $\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$ where the residuals, ε, are independently and identically normally distributed with zero mean and unknown variance σ^2. That is, that if we use all of the variable predictor variables, then we have the true model, though an unknown number of the $\boldsymbol{\beta}$'s may be zero. Then if we denote the LS estimate of vector β by $\hat{\boldsymbol{\beta}}$, we have from standard theory

$$P\{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leqslant ks^2 F_{\alpha,k,n-k}\} = 1 - \alpha \tag{4.3}$$

where s^2 is the sample estimate of the residual variance with all k variables in the model, that is $s^2 = RSS_k/(n-k)$, and $F_{\alpha,k,n-k}$ is the upper α-point (i.e. the tail area $= \alpha$) of the F-distribution with k and $(n-k)$ degrees of freedom for the numerator and denominator respectively. If the linear model including all of our predictor variables is not the true model then the equality in (4.3) should be replaced with $\geqslant$. Spjøtvoll refers readers to pp. 136–7 of Scheffe (1959) for a method of proof of this statement.

Let X_1, X_2 be two subsets of variables which we want to compare, with p_1 and p_2 variables respectively. For subset X_i the fitted values of Y for given values of the X-variables using the LS-fitted relationship are given by

$$\hat{\mathbf{Y}} = \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_i = \boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\mathbf{Y}.$$

If the expected values of $\mathbf{Y}$ are denoted by $\boldsymbol{\eta}$ then the sum of squares of derivations of the fitted values from the expected Y-values for a future set of data with the same values given for the X-variables is

$$\begin{aligned}&[\eta - \boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'(\boldsymbol{\eta} + \boldsymbol{\varepsilon})]'[\boldsymbol{\eta} - \boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'(\boldsymbol{\eta} + \boldsymbol{\varepsilon})]\\ &\quad = \boldsymbol{\eta}'\boldsymbol{\eta} - 2\boldsymbol{\eta}'\boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'(\boldsymbol{\eta} + \boldsymbol{\varepsilon}) + (\boldsymbol{\eta} + \boldsymbol{\varepsilon})'\boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'(\boldsymbol{\eta} + \boldsymbol{\varepsilon})\end{aligned}$$

which has expected value

$$= \boldsymbol{\eta}'\boldsymbol{\eta} - \boldsymbol{\eta}'\boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\boldsymbol{\eta} + \sigma^2 \operatorname{trace}[\boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i']. \tag{4.4}$$

We note that as $\text{trace}(\boldsymbol{AB}) = \text{trace}(\boldsymbol{BA})$, it can be shown that the trace above has value p_i provided that the columns of X_i have full rank.

Spjøtvoll suggested that the quantity $\boldsymbol{\eta}'\boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\boldsymbol{\eta}$ be used as the measure of goodness-of-fit of a regression function, with large values denoting a better fit. Replacing $\boldsymbol{\eta}$ with $\boldsymbol{X\beta}$, we want then to make inferences about the difference

$$\boldsymbol{\beta}'\boldsymbol{X}'[\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1' - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2']\boldsymbol{X\beta} = \boldsymbol{\beta}'\boldsymbol{C\beta} \quad (4.5)$$

which is a quadratic form in the unknown $\boldsymbol{\beta}$'s.

Now if the condition on the left-hand side of (4.3) is satisfied then we can find absolute values for the maximum and minimum of (4.5) as the condition means that the $\boldsymbol{\beta}$'s must be within the specified closeness of the known $\hat{\boldsymbol{\beta}}$. Only a summary of the results will be given here; more technical detail is given in Appendix 4A. Alternatively, the reader can refer to Spjøtvoll's paper, but this contains a very large number of printing errors and the notation differs slightly from ours.

Let $\boldsymbol{P}$ be a matrix such that both

$$\boldsymbol{P}'\boldsymbol{X}'\boldsymbol{XP} = \boldsymbol{I} \quad (4.6)$$

and

$$\boldsymbol{P}'\boldsymbol{CP} = \boldsymbol{D} \quad (4.7)$$

where $\boldsymbol{D}$ is a diagonal matrix. Such a matrix always exists, as will be clear from the method for finding $\boldsymbol{P}$ given in Appendix 4A. Let $\boldsymbol{\gamma} = \boldsymbol{P}^{-1}\boldsymbol{\beta}$ and $\hat{\boldsymbol{\gamma}} = \boldsymbol{P}^{-1}\hat{\boldsymbol{\beta}}$. (N.B. Spjøtvoll shows $\boldsymbol{P}'$ instead of $\boldsymbol{P}^{-1}$ which is not the same in general.) The condition (4.3) is now

$$\boldsymbol{P}\{(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})'(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}) \leqslant ks^2F\} = 1 - \alpha$$

while the quadratic form in (4.5) is now just $\sum d_i\gamma_i^2$ where d_i, γ_i are the ith elements of $\boldsymbol{D}$ and $\boldsymbol{\gamma}$ respectively. If the inequality on the left-hand side of (4.3) is satisfied, then

$$A_1 \leqslant \boldsymbol{\beta}'\boldsymbol{C\beta} \leqslant A_2$$

where

$$A_1 = a\left(\sum \frac{d_i\hat{\gamma}_i^2}{a + d_i} - ks^2F\right)$$

$$A_2 = b\left(\sum \frac{d_i\hat{\gamma}_i^2}{b - d_i} + ks^2F\right)$$

where $a = -\min(\min d_i, \lambda_{\min}), b = \max(\max d_i, \lambda_{\max})$, and $\lambda_{\min}, \lambda_{\max}$

are the minimum and maximum roots of

$$\sum \frac{d_i^2 \hat{\gamma}_i^2}{(d_i - \lambda)^2} + ks^2 F \tag{4.8}$$

except that $A_1 = 0$ if all of the d_i's are $\geqslant 0$ and $\sum_{d_i \neq 0} \hat{\gamma}_i^2 \leqslant ks^2 F$, and $A_2 = 0$ if all of the d_i's are $\leqslant 0$ and $\sum_{d_i \neq 0} \hat{\gamma}_i^2 \leqslant ks^2 F$. These results follow from Forsythe and Golub (1965) or Spjøtvoll (1972b). As the left-hand side of (4.8) tends to $+\infty$ from both sides as λ approaches the positive d_i's, and $-\infty$ for the negative d_i's, it is easy to see from a plot of (4.8) that $\lambda_{\min}$ is a little smaller than the smallest positive d_i. Similarly, $\lambda_{\max}$ is a little larger than the largest d_i. Any reasonable iterative method finds $\lambda_{\min}$ and $\lambda_{\max}$ very easily.

If A_1 and A_2 are both greater than zero then subset X_1 is significantly better than subset X_2 at the level used to obtain the value used for F. If the final term in (4.4) had been included somehow in the quadratic form $\boldsymbol{\beta}' \boldsymbol{C} \boldsymbol{\beta}$, then we would have been able to conclude that subset X_1 gave significantly better predictions for the particular X-values used. It does not appear to be easy to do this though a crude adjustment is simply to add $s^2(p_2 - p_1)$ to both A_1 and A_2 as though s^2 were a perfect, noise-free estimate of σ^2.

A particularly attractive feature of Spjøtvoll's method is that it gives simultaneous confidence limits and/or significance tests for any number of comparisons based upon the same data set. If the condition on the left-hand side of (4.3) is satisfied, and note that this condition is unrelated to the subsets being compared, then all of the confidence or significance statements based upon calculated values of A_1 and A_2 are simultaneously true.

A special case of the Spjøtvoll test is that in which one subset is completely contained in the other. This is the case for instance in both forward selection and backward elimination, and in carrying out the lack-of-fit test in which one 'subset' consists of all k available predictors. It is readily shown (see Appendix 4A), that the d_i's corresponding to variables which are common to both subsets are equal to zero, as also are those corresponding to variables which are in neither subset. Thus if there are p_0 variables common to both subsets, $(k - p_1 - p_2 + p_0)$ of the d_i's are equal to zero. If subset X_2 is completely contained in subset X_1, in which case $p_2 = p_0$, then the remaining $(p_1 - p_2)$ values of d_i are all equal to $+1$ and the corresponding γ_i's are just the projections of Y on the $(p_1 - p_2)$ X-variables in X_1 but not X_2, after making them orthogonal to the

p_0 common variables. (Conversely, if X_1 is contained in X_2 then there are $(p_2 - p_1)$ values of d_i equal to -1, all other d_i's being equal to zero.) In this case, (4.8) becomes

$$(1-\lambda)^{-2}\sum \hat{\gamma}_i^2 = ks^2F$$

so that $\lambda_{\max}$ and $\lambda_{\min} = 1 \pm (\sum \hat{\gamma}_i^2/ks^2F)^{1/2}$. Now $\sum d_i\hat{\gamma}_i^2$ is the difference in RSS between fitting subsets X_1 and X_2. In this case, it simplifies to $\sum \hat{\gamma}_i^2$. Denote this difference by ΔRSS, then

$$A_1 = \max\left[0, \{\Delta RSS^{1/2} - (ks^2F)^{1/2}\}^2\right]$$
$$A_2 = \{\Delta RSS^{1/2} + (ks^2F)^{1/2}\}^2.$$

Hence subset X_1 (which is the one with more variables) is significantly better than X_2 if $A_1 > 0$, that is if ΔRSS is greater than ks^2F. If it had been decided *a priori* to compare subsets X_1 and X_2 then the usual test would have been to compute the variance ratio $\Delta RSS/(p_1 - p_2)$ divided by s^2. Subset X_1 would then have been deemed significantly better if ΔRSS were greater than $(p_1 - p_2)s^2F$. The replacement of $(p_1 - p_2)$ by k in calculating the required difference in residual sums of squares, ΔRSS, is an indication of the degree of conservatism in the Spjøtvoll test.

This special case of Spjøtvoll's test has also been derived, using different arguments, by Aitkin (1974), McCabe (1978), McKay (1979) and Tarone (1976). Borowiak (1981) appears to have derived a similar result for the case in which the residual variance is assumed to be known.

The argument used by Aitkin is as follows. If we have decided, *a priori*, that we wanted to compare subset X_2 with the full model containing all of the variables in X, then we would have used the likelihood-ratio test which gives the variance ratio statistic:

$$F = \frac{(RSS_p - RSS_k)/(k-p)}{RSS_k/(n-k)} \tag{4.9}$$

where the counts of variables (p and k) include one degree of freedom for a constant if one is included in the models. Under the null hypothesis that none of the $(k - p)$ variables excluded from X_2 is in the 'true' model, this quantity is distributed as $F(k - p, n - k)$, subject of course to assumptions of independence, normality and homoscedacity of the residuals from the model. Aitkin then considers

the statistic:

$$U(X_2) = (k - p)F \tag{4.10}$$

The maximum value of U for all possible subsets (but including a constant) is then

$$U_{\max} = \frac{RSS_1 - RSS_k}{RSS_k/(n-k)}.$$

Hence a simultaneous $100\alpha\%$ test for all hypotheses $\beta_2 = 0$ for all subsets X_2 is obtained by testing that

$$U(X_2) = (k-1)F(\alpha, k-1, n-k) \tag{4.11}$$

Subsets satisfying (4.11) are called R^2-adequate sets. Aitkin expresses (4.9) in terms of R^2 instead of the residual sums of squares, and hence (4.10) can also be so expressed.

The term 'minimal adequate sets' is given to subsets which satisfy (4.11) but which are such that if any variable is removed from the subset, it fails to satisfy the condition. Edwards and Havranek (1987) give an algorithm for deriving minimal adequate sets.

4.2.1 *Applications of Spjøtvoll's method*

Spjøtvoll's method was applied to the best subsets found for the STEAM, DETROIT and POLLUTE data sets. Table 4.6 shows some of the comparisons for the STEAM data set. From this table we see that the best-fitting subsets of two variables are significantly better than the best-fitting single variable (number 7) at the 5% level, but that there is no significant improvement from adding further variables. Referring back to Table 3.13 we might anticipate that subsets (4, 7) and (7, 9) might be significantly worse than the best-three of three variables because of the big difference in *RSS*, but this is not so.

Notice that though we have obtained 90 and 98% confidence levels, we are quoting 5 and 1% significance levels. As the subsets which are being compared have been selected conditional upon their position among the best-fitting subsets, we know that a significant difference can occur in only one tail, and that tail is known before carrying out the test. It is appropriate then that a single-tail test is used.

Table 4.6 *Spjøtvoll's upper and lower confidence limits for the difference in regression sums of squares (or in RSS) for selected subset comparisons for the STEAM data set*

Subset X_1	*Subset* X_2	*Diff. in RSS* Sub. 2 − Sub. 1	*α-level* (%)	A_1	A_2
7	6	19·4	10	3·0	40·7
			2	−1·1	47·1
7	5	27·25	10	−8·8	69·0
7	3	31·2	10	7·1	63·7
			2	1·3	73·4
7	8	35·7	10	10·5	70·5
			2	4·8	81·0
7	1,7	−9·3	10	−31·7	−0·21
			2	−39·5	0
7	5,7	−8·6	10	−30·4	−0·12
			2	−38·1	0
7	2,7	−8·4	10	−30·1	−0·10
			2	−37·7	0
7	4,7	−2·6	10	−17·7	0
7	7,9	−2·2	10	−16·7	0
1,7	5,7	0·7	10	−13·6	15·5
1,7	2,7	0·9	10	−5·1	7·5
1,7	4,7	6·7	10	−7·8	27·0
1,7	7,9	7·1	10	−3·4	24·5
1,7	4,5,7	−1·6	10	−15·1	10·8
1,7	1,5,7	−1·2	10	−13·7	0
1,7	1,3,5,7,8,9	−3·6	10	−20·1	0

The comparisons among the single variables in Table 4.6 are interesting. At the 5% level, variable 7 fits significantly better than the second-, fourth- and fifth-best, but not the third-best. This is a situation which will often occur. If there are high correlations, positive or negative, between the variables in the two subsets, they span almost the same space and the upper and lower limits are relatively close together. In this case, the correlations between variable 7 and the others in the best five, together with the range between upper and lower 90% limits are

Variable	6	5	3	8
Correlation	−0·86	−0·21	−0·62	−0·54
Range	37·7	77·8	56·6	60·6

In the case of the DETROIT data set, the values of F are from the distribution with 12 degrees of freedom for the numerator but only 1 degree of freedom for the denominator. The 10 and 2% points in the tail are at 60·7 and 1526 respectively. If we had one year's data more then these values would be down to 9·41 and 49·4. The huge confidence ellipsoids for the regression coefficients are reflected in the big differences between the upper and lower limits, A_1 and A_2. Table 4.7 shows the results of a few comparisons of subsets for this data set. Only one of the comparisons in this table yields a significant difference; subset (4, 6) fits significantly better at the 5% level than variable 6 alone.

For the POLLUTE data set there was a moderately large number of degrees of freedom (44) for the residual variance so that much closer confidence limits and more powerful tests are possible. No significant differences at the 5% level were found when comparing subsets of the same size using only the five best-fitting subsets of each size as listed in Table 3.17. In most practical cases, more subsets of each size should be kept; the number was limited to five here to keep down the amount of space consumed by these tables. Table 4.8

Table 4.7 *Spjøtvoll's upper and lower confidence limits for the difference in regression sums of squares (or in RSS) for selected subset comparisons for the DETROIT data*

Subset X_1	*Subset* X_2	*Diff. in RSS Sub. 2 − Sub. 1*	*α-level (%)*	A_1	A_2
6	1	27	10	−176	231
6	9	65	10	−347	476
6	8	78	10	−337	495
6	7	98	10	−317	517
6	4,6	−166	10	−440	−23
			2	−2858	0
4,6	2,7	11	10	−143	250
4,6	1,9	21	10	−135	181
4,6	4,5	22	10	−94	148
4,6	3,8	29	10	−152	210
4,6	2,4,11	−27	10	−148	90
2,4,11	4,6,10	14	10	−93	132
4,6	1,2,4,6,7,11	−32	10	−190	0
2,4,11	1,2,4,6,7,11	−5	10	−109	0

Table 4.8 *Spjøtvoll's upper and lower confidence limits for the difference in regression sums of squares (or in RSS) for selected subset comparisons for the POLLUTE data*

Subset X_1	*Subset* X_2	*Diff. in RSS* *Sub. 2 − Sub. 1*	*α-level* *(%)*	A_1	A_2
9	6,9	−34000	10	−131000	−32
			2	−153000	0
6,9	2,6,9	−17000	10	−96000	0
6,9	1,2,9,14	−31000	10	−134000	51000
6,9	1,2,6,9,14	−35000	10	−134000	−86
			2	−156000	0
6,9	1,2,3,6,9,14	−39000	10	−142000	−395
			2	−164000	0
2,6,9	1,2,3,6,9,14	−22000	10	−106000	0

shows only the comparisons between the best-fitting subsets of each size.

In Table 4.8 we see that the subset of two variables, 6 and 9, fits just significantly better at the 5% level than variable 9 alone. The upper limit of −32 for this comparison is extremely small as the residual variance estimate for these data is 1220. In view of the term in (4.4) which was left out of the quadratic form used for the comparison of pairs of subsets, we cannot infer that the subset (6,9) will yield better predictions.

4.2.2 *Using other confidence ellipsoids*

Spjøtvoll has pointed out that in applying his method, the confidence ellipsoid (4.3) which was used as the starting point can be shrunk by reducing it with one in a smaller number of variables. For instance, if certain variables, such as the dummy variable representing the constant in the model, are to be forced into all models, then those variables can be removed from X and $\boldsymbol{\beta}$. If this leaves k^* variables then $ks^2F_{\alpha,k,n-k}$ should be replaced with $k^*s^2F_{\alpha,k^*,n-k}$. It must be emphasized that ellipsoids of the form (4.3), but for subsets of variables, are only valid if those subsets had been determined *a priori*; they are not valid if, for instance, they include only those variables which appear in certain best-fitting subsets, as the LS estimates of the regresion coefficients are then biased estimates of the coefficients

for that subset (see Chapter 5 for a detailed discussion of selection bias).

To assist in appreciating the difference between these confidence regions, Fig. 4.1 shows a hypothetical confidence ellipse for two regression coefficients, β_1 and β_2, and the confidence limits for the same confidence level for one of the coefficients. If the two X-variables have been standardized to have unit standard deviation, and the $\boldsymbol{X}'\boldsymbol{X}$-matrix, excluding the rows and columns for the constant in the model, is

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

then the elliptical confidence regions are given by

$$x^2 + 2\rho xy + y^2 \leqslant 2s^2F$$

where $x = (\beta_1 - \hat{\beta}_1), y = (\beta_2 - \hat{\beta}_2), s^2$ is the usual residual variance estimate with v degrees of freedom, and F is the appropriate percentage point of the F-distribution with 2 and v degrees of freedom respectively for the numerator and denominator. The most extreme points on the ellipses are then

$$\hat{\beta}_i \pm (s^2/(1-\rho^2))^{1/2}(2F)^{1/2}$$

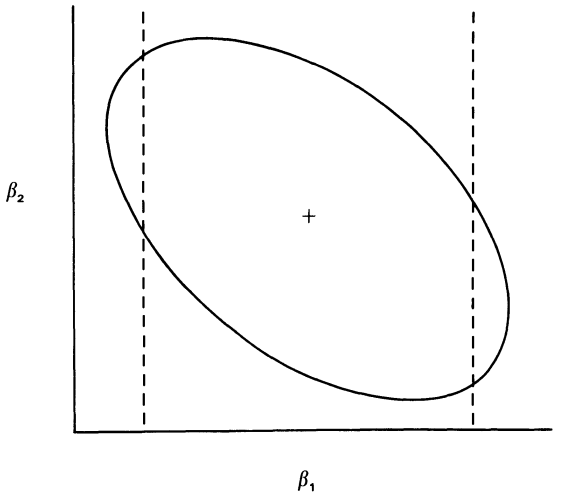


Fig. 4.1 *Confidence ellipse and confidence limits for a hypothetical case.*

compared with the corresponding confidence limits (i.e. for individual regression coefficients) which are

$$\hat{\beta}_i \pm (s^2/(1-\rho^2))^{1/2}\, t$$

where t is the appropriate percentage point from the t-distribution. For moderately large numbers of degrees of freedom, $t \approx 2$ and $F \approx 3$ so that the ratio of ranges of regression coefficients in this case (i.e. for two coefficients) is about 1.22:1.

In general, by reducing the number of β's, the ranges of those remaining are reduced but at the expense of allowing infinite ranges for the β's omitted.

4.2.3 Goodness-of-fit outside the calibration region

Unless there is sound scientific justification for a model, it is always very hazardous to attempt to use it with estimated coefficients outside the range over which it may have been found to fit tolerably well. If we have sufficient trust in the linear model which includes all of the predictor variables, then Spjøtvoll's method may be used to test whether one subset of variables can be expected to perform better than another in any specified region of the X-space. Let Z be an $m \times k$ array of values of the X-variables for which predictions are required. Then proceeding as before, we have that the sum of squares of deviations of the expected values of the dependent variables from those predicted using subset X_i is

$$[\boldsymbol{Z}\beta - \boldsymbol{Z}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'(\boldsymbol{X}\beta + \varepsilon)]'[\boldsymbol{Z}\beta - \boldsymbol{Z}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'(\boldsymbol{X}\boldsymbol{\beta} + \varepsilon)],$$

where $\boldsymbol{Z}_i$ refers to the values in $\boldsymbol{Z}$ for the subset of variables in $\boldsymbol{X}_i$. The equivalent goodness-of-fit measure to that used earlier is then

$$\boldsymbol{\beta}'[-2\boldsymbol{Z}'\boldsymbol{Z}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\boldsymbol{X} + \boldsymbol{X}'\boldsymbol{X}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{Z}_i'\boldsymbol{Z}_i(\boldsymbol{X}_i'\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i'\boldsymbol{X}]\boldsymbol{\beta}.$$

The new matrix $\boldsymbol{C}$ is the difference between the value of the expression inside the square brackets for subsets X_1 and X_2. A small amount of simplification results by replacing the various $\boldsymbol{X}$ and $\boldsymbol{Z}$ matrices by their orthogonal reductions.

Finally in this chapter we must emphasize that the methods described here will usually eliminate some subsets from consideration but will rarely leave a single subset which is indisputably the best. We must also emphasize that satisfying a significance test and providing good predictions are not synonymous.

Appendix 4A Spjøtvoll's method – detailed description

In applying Spjøtvoll's method, we want to find maximum and minimum values of the quadratic form

$$\boldsymbol{\beta}'\boldsymbol{X}'[\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1' - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2']\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}'\boldsymbol{C}\boldsymbol{\beta} \tag{4.5}$$

subject to β being close to the known vector of LS estimates $\hat{\boldsymbol{\beta}}$, where the degree of closeness is such that

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leqslant ks^2 F_{\alpha,k,n-k}.$$

The subscripts on F will be dropped. We will assume that $\boldsymbol{X}'\boldsymbol{X}$ is of full rank. First we attempt to find a matrix $\boldsymbol{P}$ such that

$$\boldsymbol{P}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{P} = \boldsymbol{I} \tag{4.6}$$

and

$$\boldsymbol{P}'\boldsymbol{C}\boldsymbol{P} = \boldsymbol{D} \tag{4.7}$$

where $\boldsymbol{D}$ is a diagonal matrix. If we can do this, then we have transformed the problem into that of finding the maximum and minimum of $\boldsymbol{\gamma}'\boldsymbol{D}\boldsymbol{\gamma}$ subject to

$$(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})'(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}) \leqslant ks^2 F$$

where $\boldsymbol{\gamma} = \boldsymbol{P}^{-1}\boldsymbol{\beta}$, and $\hat{\boldsymbol{\gamma}} = \boldsymbol{P}^{-1}\hat{\boldsymbol{\beta}}$.

First, let us form the Cholesky factorization $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{R}'\boldsymbol{R}$ where $\boldsymbol{R}$ is an upper-triangular matrix, though at this stage any nonsingular factorization will suffice. Then $\boldsymbol{P} = \boldsymbol{R}^{-1}$ satisfies (4.6). Now let us find a matrix $\boldsymbol{V}$ where columns are the normalized eigenvectors of $\boldsymbol{R}^{-T}\boldsymbol{C}\boldsymbol{R}^{-1}$, that is $\boldsymbol{V}$ satisfies

$$\boldsymbol{V}'(\boldsymbol{R}^{-T}\boldsymbol{C}\boldsymbol{R}^{-1})\boldsymbol{V} = \boldsymbol{D}$$

where $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}$ and $\boldsymbol{D}$ is a diagonal matrix with the eigenvalues of $\boldsymbol{R}^{-T}\boldsymbol{C}\boldsymbol{R}^{-1}$ on its diagonal. The matrix $\boldsymbol{P} = \boldsymbol{R}^{-1}\boldsymbol{V}$ then satisfies (4.6) and (4.7). We have then that $\hat{\boldsymbol{\gamma}} = \boldsymbol{V}'\boldsymbol{R}\hat{\boldsymbol{\beta}}$. Now if $\boldsymbol{R}$ were taken as the Cholesky factor of $\boldsymbol{X}'\boldsymbol{X}$, then $\boldsymbol{R}\hat{\boldsymbol{\beta}}$ is just the vector of projections of the dependent variable on the space spanned by the $\boldsymbol{X}$'s and so would normally be calculated in any LS calculations using orthogonal reduction.

In calculating the eigenstructure of the matrix $\boldsymbol{R}^{-T}\boldsymbol{C}\boldsymbol{R}^{-1}$, it is convenient to order the variables as follows (note: the order is not that used by Spjøtvoll):

1. the p_0 variables which are common to both X_1 and X_2;
2. the $(p_1 - p_0)$ variables in X_1 but not in X_2;
3. the $(p_2 - p_0)$ variables in X_2 but not in X_1;
4. the remaining $(k - p_1 - p_2 + p_0)$ variables in neither X_1 nor X_2.

In practice, some of the above groups will often be empty.

Then we form the orthogonal reduction

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{QR} \\ &= (\boldsymbol{Q}_1, \boldsymbol{Q}_2, \boldsymbol{Q}_3, \boldsymbol{Q}_4) \begin{pmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} & \boldsymbol{R}_{13} & \boldsymbol{R}_{14} \\ & \boldsymbol{R}_{22} & \boldsymbol{R}_{23} & \boldsymbol{R}_{24} \\ & & \boldsymbol{R}_{33} & \boldsymbol{R}_{34} \\ & & & \boldsymbol{R}_{44} \end{pmatrix} \end{aligned}$$

where the columns of the $\boldsymbol{Q}$ are orthonormal, that is that $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}$, and the subscripts refer to the groups of variables, some of which may be empty, as just described. The columns of $\boldsymbol{Q}$ are of length n equal to the number of observations and $\boldsymbol{R}$ is upper triangular with k rows and columns. In practice this orthogonal reduction will usually be obtained from another orthogonal reduction by rearranging the order of variables.

Now we can write

$$\boldsymbol{X}_1 = (\boldsymbol{Q}_1, \boldsymbol{Q}_2) \begin{pmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ & \boldsymbol{R}_{22} \end{pmatrix}$$

$$\boldsymbol{X}_2 = (\boldsymbol{Q}_1, \boldsymbol{Q}_2, \boldsymbol{Q}_3) \begin{pmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{13} \\ & \boldsymbol{R}_{23} \\ & \boldsymbol{R}_{33} \end{pmatrix}.$$

This gives us an upper-triangular factorization for $\boldsymbol{X}_1$ which is simply the first p_1 rows and columns of R. Unfortunately it does not give us anything simple for the factorization of $\boldsymbol{X}_2$. However by the use of planar rotations again, we can obtain a triangular factorization for $\boldsymbol{X}_2$ also. Let $\boldsymbol{P}_0$ be a product of the planar rotations thus required. Then $\boldsymbol{P}_0$ has the form

$$\boldsymbol{P}_0 = \begin{pmatrix} \boldsymbol{I} & & \\ & \boldsymbol{P}_1 & \boldsymbol{P}_2 \\ & \boldsymbol{P}_3 & \boldsymbol{P}_4 \end{pmatrix}$$

where $\boldsymbol{I}$ has p_0 rows and columns, $\boldsymbol{P}_1, \boldsymbol{P}_2$ have $(p_2 - p_0)$ rows, $\boldsymbol{P}_3, \boldsymbol{P}_4$ have $(p_1 - p_0)$ rows, and it will later be convenient to split the

columns so that P_1, P_3 have $(p_1 - p_0)$ columns, P_2, P_4 have $(p_2 - p_0)$ columns, and $P_0'P_0 = I$. Then

$$X_2 = (Q_1, Q_2, Q_3)P_0'P_0\begin{pmatrix} R_{11} & R_{13} \\ & R_{23} \\ & R_{23} \end{pmatrix}$$

$$= (Q_1, Q_3^*, Q_2^*)\begin{pmatrix} R_{11} & R_{13} \\ & R_{33}^* \\ & 0 \end{pmatrix}$$

$$= (Q_1, Q_3^*)\begin{pmatrix} R_{11} & R_{13} \\ & R_{33}^* \end{pmatrix}.$$

The matrix P_0 can be formed while Q_2 is being forced out of the orthogonal reduction, by applying the same planar rotations to another matrix which is initially an identity matrix with $(p_1 - p_0) + (p_2 - p_0)$ rows and columns.

By substitution of the appropriate orthogonal reduction, it is straightforward to show that

$$X_1(X_1'X_1)^{-1}X_1' = (Q_1, Q_2)\begin{pmatrix} Q_1' \\ Q_2' \end{pmatrix}$$

and

$$X_2(X_2'X_2)^{-1}X_2' = (Q_1, Q_3^*)\begin{pmatrix} Q_1' \\ Q_3^{*'} \end{pmatrix}.$$

We then find that the matrix of which we want the eigenstructure can be written as

$$R^{-T}CR^{-1} = \begin{pmatrix} Q_1' \\ Q_2' \\ Q_3' \\ Q_4' \end{pmatrix}\left[(Q_1, Q_2)\begin{pmatrix} Q_1' \\ Q_2' \end{pmatrix}\right.$$

$$\left. - (Q_1, Q_3^*)\begin{pmatrix} Q_1' \\ Q_3^{*'} \end{pmatrix}\right](Q_1, Q_2, Q_3, Q_4)$$

$$= \begin{pmatrix} I & 0 \\ 0 & I \\ 0 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \end{pmatrix} - \begin{pmatrix} I & 0 \\ 0 & A \\ 0 & B \\ 0 & 0 \end{pmatrix}\begin{pmatrix} I & 0 & 0 & 0 \\ 0 & A' & B' & 0 \end{pmatrix}$$

where

$$\begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} Q_2' \\ Q_3' \end{pmatrix} Q_3^*.$$

But

$$Q_3^* = (Q_2, Q_3)\begin{pmatrix} P_1' \\ P_2' \end{pmatrix}$$

and hence

$$\begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} P_1' \\ P_2' \end{pmatrix}.$$

In the general case, the matrix $R^{-T}CR^{-1}$ will have p_0 rows of zeros at the top and $(k - p_1 - p_2 + p_0)$ rows of zeros at the bottom, and has a zero eigenvalue for each such row. We are left then to find the eigenstructure of the symmetric inner matrix:

$$\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} P_1' \\ P_2' \end{pmatrix}(P_1, P_2) = \begin{pmatrix} I - P_1'P_1 & -P_1'P_2 \\ -P_2'P_1 & -P_2'P_2 \end{pmatrix} = Z \text{ say.}$$

The eigenvalues and eigenvectors of Z can be found using standard routines such as RS from EISPACK (Smith *et al.*, 1976). We notice that $P_0P_0' = I$ (as well as $P_0'P_0 = I$) and hence $P_1P_1' + P_2P_2' = I$.

Now the eigenvalues of Z satisfy $|Z - dI| = 0$ or

$$\begin{vmatrix} (1-d)I - P_1'P_1 & -P_1'P_2 \\ -P_2'P_1 & -P_2'P_2 - dI \end{vmatrix} = 0.$$

As the determinant of the product of two matrices equals the product of the determinants, we can multiply $(Z - dI)$ by any other matrix, being careful not to introduce any more zeros, and the determinant will still be zero. A convenient matrix to use is Z. Multiplying on the right by Z gives

$$\begin{vmatrix} (1-d)(I - P_1'P_1) & dP_1'P_2 \\ dP_2'P_1 & (1+d)P_2'P_2 \end{vmatrix} = 0$$

This matrix consists of the blocks of Z multiplied by different scalars. We now multiply on the right by

$$\begin{pmatrix} I & dP_1'P_2 \\ 0 & (1-d)P_2'P_2 \end{pmatrix}$$

noting that in the process we introduce $(p_2 - p_0)$ roots equal to 1, then the top right-hand side block is transformed to a matrix of zeros, and we can write

$$(1-d)^{p_1-p_0}|\boldsymbol{I}-\boldsymbol{P}_1'\boldsymbol{P}_1|\cdot|\boldsymbol{P}_2'[(\boldsymbol{I}-\boldsymbol{P}_1\boldsymbol{P}_1')-d^2\boldsymbol{I}]\boldsymbol{P}_2|=0.$$

If the eigenvalues of $(\boldsymbol{I}-\boldsymbol{P}_1\boldsymbol{P}_1')=\boldsymbol{P}_2\boldsymbol{P}_2'$ are λ_i, $i=1,\ldots,p_2-p_0$, then

$$d_i=\pm\lambda_i^{1/2}.$$

Our present equation has $(p_1-p_0)+2(p_2-p_0)$ roots, of which (p_2-p_0) equal to 1 were introduced. In general then, if $p_1 \geqslant p_2$ there will be (p_1-p_2) of the d_i's equal to $+1$ and (p_2-p_0) pairs of d_i's with opposite signs, while if $p_1 \leqslant p_2$ there will be (p_2-p_1) of the d_i's equal to -1 and (p_1-p_0) pairs of d_i's with opposite signs.

CHAPTER 5

Estimation of regression coefficients

5.1 Selection bias

In this chapter we look at some of the ways of estimating regression coefficients for a subset of variables when the data to be used for estimation are the same as those which were used to select the model. Most of these methods are based upon the biased least-squares (LS) regression coefficients, and require an estimate of the selection bias or depend upon properties of the selection bias. Such methods will be discussed in section 5.3, while selection bias in a very simple case will be discussed in section 5.2.

The term 'selection bias' was introduced in Chapter 1 but was not precisely defined at that stage. Suppose the true relationship between Y and the predictor variables is

$$\mathbf{Y} = X_A\boldsymbol{\beta}_A + X_B\boldsymbol{\beta}_B + \boldsymbol{\varepsilon}$$

where $X = (X_A, X_B)$ is a subdivision of the complete set of variables into two subsets A and B, and where the residuals have zero expected value. If we fit a model containing only the variables in subset A then the expected value of the vector, $\mathbf{b}_A$, of LS regression coefficients for subset A is

$$E(\mathbf{b}_A) = \boldsymbol{\beta}_A + (X'_A X_A)^{-1} X'_A X_B \boldsymbol{\beta}_B. \tag{5.1}$$

The second term on the right-hand side of (5.1) is what we have called the omission bias. Expression (5.1) is valid when the subset A has been chosen independently of the data. Let us suppose that some procedure for selecting a subset of variables has been prespecified. The definition of selection bias in the regression coefficients is then as the difference between the expected values when the data are such as to satisfy the conditions necessary for the

selection of the subset A, and the unconditional expected values given by (5.1), that is

$$\text{Selection bias} = E(\mathbf{b}_A | \text{Subset } A \text{ selected}) - E(\mathbf{b}_A).$$

That is, the first term on the right-hand side is the expected value of $\mathbf{b}_A$ over all possible $\mathbf{Y}$-vectors which would lead to subset A being selected, while the second term is the expected value over all Y irrespective of what subset is selected. The extent of the bias is therefore dependent upon the selection procedure (e.g. forward selection, sequential replacement, exhaustive search, etc.), and is also a function of the stopping rule used. It can usually be anticipated that the more extensive the search for the chosen model, the more extreme the data values must be to satisfy the conditions and hence the greater the selection bias. Thus in an exhaustive search, the subset is compared with all other subsets so that larger biases can be expected than for forward selection in which a much smaller number of comparisons is made.

There is a large literature on the subject of pre-test estimation, though in most cases only one or two tests are applied in selecting the model as contrasted with the large number usually carried out in subset selection. For the relatively small number (usually only two) of alternatives considered, explicit expressions can be derived for the biases in the regression coefficients, and hence the effect of these biases on prediction errors can be derived. There is almost no consideration of alternative estimates of regression coefficients other than ridge regression and the James–Stein/Sclove estimator estimators. Useful references on this topic are the surveys by Bancroft and Han (1977) and Wallace (1977), and the book by Judge and Bock (1978).

Though much of this chapter is concerned with understanding and trying to reduce bias, it should not be construed that bias elimination is a necessary or even sometimes a desirable objective. We have already discussed some biased estimators such as ridge estimators. Biased estimators are a standard part of the statistician's tool-kit. How many readers use unbiased estimates of standard deviations? NB The usual estimator $s = \{\sum(x - \bar{x})^2/(n-1)\}^{1/2}$ is biased, though most of the bias can be removed by using $(n - 3/2)$ instead of the $(n-1)$.

It is important though that we are aware of biases and have some idea of their magnitude. This is particularly true in the case of the

selection of subsets of regression variables when the biases can be substantial when there are many subsets which are close competitors for selection.

5.2 Choice between two variables

To illustrate the bias resulting from selection, let us consider a very simple example in which only two predictor variables, X_1 and X_2, are available and it has been decided *a priori* to select only one of them, that one being the one which gives the smaller residual sum of squares (*RSS*) when fitted to a set of data. For this case it is feasible to derive mathematically the properties of the LS (or other) estimate of the regression coefficient for the selected variable, the *RSS*, and other quantities of interest. However, before doing that, we present some simulation results.

The following example is constructed so that the expected *RSS* is the same whichever of the two variables is selected. For added simplicity, the fitted models do not include a constant. Let us define

$$X_1 = Z_1$$
$$X_2 = -3Z_1 + 4Z_2$$
$$Y = Z_1 + 2Z_2 + Z_3,$$

where Z_1, Z_2 and Z_3 are independently sampled from the standard normal distribution. If we take a sample of n independent observations and fit the two alternative models,

$$Y = \gamma_1 X_1 + \text{residual}$$
$$Y = \gamma_2 X_2 + \text{residual},$$

by LS, then the expected values of the sample regression coefficients, b_1 and b_2, are 1·0 and 0·2 respectively, and the expected *RSS* is $5(n-1)$ for both models. Note, the expected *RSS* with both variables in the model is $(n-2)$, so that if a similar case arose in practice, both variables would probably be included in the selected model. The simulation could have been modified to give a much larger *RSS* with both variables in the model simply by increasing the variance of the residual variation, Z_3.

Using samples of size $n = 12$, 200 sets of artificial data were generated. The solid histogram shown in Fig. 5.1 is that of the 200 sample values of b_1. These values averaged 0·991 which is close to

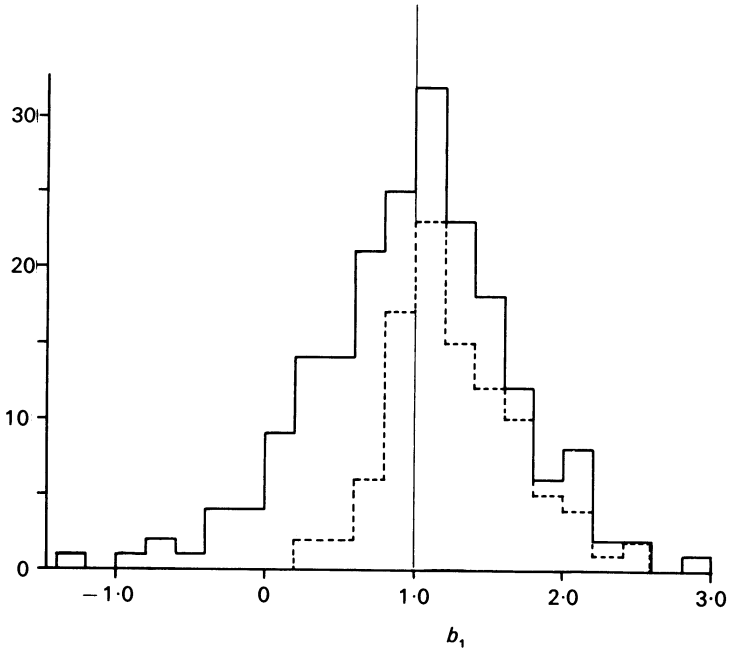


Fig. 5.1 *Histogram of values of b_1 from 200 sets of artificial data. The outer histogram is for all data sets, the inner one is for those sets for which variable X_1 gave the smaller RSS. The thin vertical line is at the expected value of b_1.*

the expected value; the average of the corresponding values of b_2 was 0·204. The inner, broken histogram shown in Fig. 5.1 is that of the values of b_1 when variable X_1 was selected. As can be seen, variable X_1 was usually selected when b_1 was above its expected value, but rarely selected when it was below. The corresponding histograms for b_2, which are not shown, look very similar. The average values of b_1 and b_2 for the data sets in which their corresponding variables were selected were 1·288 and 0·285. Variable X_1 was selected 99 times out of the 200 data sets.

Figure 5.2 shows the histograms of *RSS*'s. The solid histogram is that for all 400 *RSS*'s, that is it includes two *RSS*'s for each data set, one for each model. The inner, broken histogram is that of *RSS*'s for the selected models. The thin vertical line is at the expected value of the *RSS* (= 55). The sample average of all the *RSS*'s was 55·3, while the average for the selected models was 45·7.

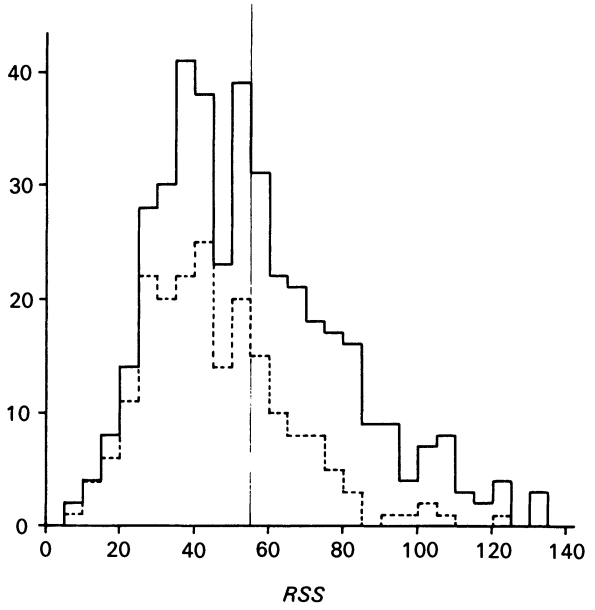


Fig. 5.2 *Histograms of values of the RSS from* 200 *artificial data sets. The outer histogram is of all values (two per data set), while the inner one is for the model which gave the smaller RSS for each data set. The thin vertical line is at the expected value of the RSS.*

Thus for one highly artificial example, we have found the estimated regression coefficients of selected variables to be biased on the high side and the *RSS* to be too small.

In practice we often have many more variables or subsets of variables competing for selection, and in such cases the biases are often far greater than here. When an exhaustive search has shown say several hundred subsets of 5 out of say 20 variables which fit a set of data about equally well, perhaps so that Spjøtvoll's test finds no significant differences between them at say the 5% level, then some of the regression coefficients for the best-fitting subset will probably be of the order of two to three standard errors from the expected values for the same subset if they are estimated from the same data as were used to select that subset. This assumes that the true standard error is known. If the usual LS estimates of the standard error, based upon the assumption that the model has been chosen independently of the data, are used, these estimates will be biased

on the low side, and the bias in the regression coefficients could easily be four or more estimated standard errors above their expected values for the subset. This will be demonstrated in the next section by splitting data sets.

Let us now derive more general results for the two-variable case. Let us suppose that the true model is

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon \tag{5.2}$$

where the residuals, ε, have zero mean and variance σ^2. Later we will also need to make assumptions about the shape of the distribution of the residuals. The LS estimate, b_1, of the regression coefficient for the simple regression of Y upon X_1 is then

$$\begin{aligned} b_1 &= \mathbf{X}_1'\mathbf{Y}/\mathbf{X}_1'\mathbf{X}_1 \\ &= (\beta_1 \mathbf{X}_1'\mathbf{X}_1 + \beta_2 \mathbf{X}_1'\mathbf{X}_2 + \mathbf{X}_1'\varepsilon)/\mathbf{X}_1'\mathbf{X}_1, \end{aligned}$$

and hence

$$\begin{aligned} E(b_1) &= \beta_1 + \beta_2 \mathbf{X}_1'\mathbf{X}_2/\mathbf{X}_1'\mathbf{X}_1 \\ &= \gamma_1. \end{aligned}$$

Similarly

$$\begin{aligned} E(b_2) &= \beta_2 + \beta_1 \mathbf{X}_2'\mathbf{X}_1/\mathbf{X}_2'\mathbf{X}_2 \\ &= \gamma_2. \end{aligned}$$

Note that these are the expected values over all samples, no selection has been considered so far. The difference between γ_1 and β_1, and similarly that between γ_2 and β_2, is what we called earlier the omission bias.

Now variable X_1 is selected when it gives the smaller *RSS*, or equivalently, when it gives the larger regression sum of squares. That is, X_1 is selected when

$$\mathbf{X}_1'\mathbf{X}_1 b_1^2 > \mathbf{X}_2'\mathbf{X}_2 b_2^2. \tag{5.3}$$

If we let $f(b_1, b_2)$ denote the joint probability density of b_1 and b_2, then the expected value of b_1 when variable X_1 is selected is

$$E(b_1 | E_1 \text{ selected}) = \frac{\int_R \int b_1 f(b_1, b_2)\, \mathrm{d}b_1\, \mathrm{d}b_2}{\int_R \int f(b_1, b_2)\, \mathrm{d}b_1\, \mathrm{d}b_2} \tag{5.4}$$

where the region R in the (b_1, b_2)-space is that in which condition (5.3) is satisfied. The denominator of the right-hand side of (5.4) is the probability that variable X_1 is selected. The region R can be re-expressed as that in which $|b_1| > C|b_2|$ where $C = (\mathbf{X}_2'\mathbf{X}_2/\mathbf{X}_1'\mathbf{X}_1)^{1/2}$. As the boundaries of this region are straight lines, it is relatively straightforward to evaluate (5.4) numerically for any assumed distribution of b_1 and b_2, given the sample values of $\mathbf{X}_1'\mathbf{X}_1$ and $\mathbf{X}_2'\mathbf{X}_2$. Similarly, by replacing the b_1 in the numerator of the right-hand side of (5.4) with b_1^r, we can calculate the rth moment of b_1 when X_1 is selected.

As b_1 and b_2 are both linear in the residuals, ε, it is feasible to calculate $f(b_1, b_2)$ for any distribution of the residuals. However, if the distribution of the residuals departs drastically from the normal, we should be using some other method of fitting regressions than LS, though in some cases, e.g. if the observations have a Poisson distribution or a gamma distribution with constant shape parameter, the maximum likelihood estimators of the regression coefficients are weighted LS estimators. If the residuals have a distribution which is close to normal then, by the central limit theorem, we can expect the distribution of b_1 and b_2 to be closer to normal, particularly if the sample size is large. The results which follow are for the normal distribution.

Given the values of $\mathbf{X}_1$ and $\mathbf{X}_2$, the covariance matrix of b_1, b_2 is

$$V = \sigma^2 \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} & \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1)^{-1}(\mathbf{X}_2'\mathbf{X}_2)^{-1} \\ \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1)^{-1}(\mathbf{X}_2'\mathbf{X}_2)^{-1} & (\mathbf{X}_2'\mathbf{X}_2)^{-1} \end{bmatrix}.$$

It will simplify the mathematical expressions if we scale $\mathbf{X}_1$ and $\mathbf{X}_2$ so that $\mathbf{X}_1'\mathbf{X}_1 = \mathbf{X}_2'\mathbf{X}_2 = 1$, that is, if we replace $\mathbf{X}_i$ with

$$\mathbf{X}_i^* = \mathbf{X}_i/(\mathbf{X}_i'\mathbf{X}_i)^{1/2},$$

and replace b_i with

$$b_i^* = b_i(\mathbf{X}_i'\mathbf{X}_i)^{1/2}.$$

We will assume that such a scaling has been carried out, and drop the use of the asterisks (*). The covariance matrix of the scaled b's is then

$$V = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

where $\rho = X_1' X_2$. The joint probability density of b_1 and b_2 is then

$$f(b_1, b_2) = \frac{\exp\{-(\mathbf{b}-\boldsymbol{\gamma})' V^{-1}(\mathbf{b}-\boldsymbol{\gamma})\}}{2\pi\sigma^2(1-\rho^2)^{1/2}},$$

where

$$V^{-1} = \frac{1}{\sigma^2(1-\rho^2)}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}.$$

The argument of the exponential is

$$\begin{aligned} &-\{(b_1-\gamma_1)^2 - 2\rho(b_1-\gamma_1)(b_2-\gamma_2) + (b_2-\gamma_2)^2\}/\{2\sigma^2(1-\rho^2)\} \\ &\quad = -\{[b_1-\mu(b_2)]^2 + (1-\rho^2)(b_2-\gamma_2)^2\}/\{2\sigma^2(1-\rho^2)\}, \end{aligned}$$

where $\mu(b_2) = \gamma_1 + \rho(b_2-\gamma_2)$. Hence we need to evaluate integrals of the form

$$I_r = \int_{-\infty}^{\infty} \frac{\exp[-(b_2-\gamma_2)^2/2\sigma^2]}{(2\pi\sigma^2)^{1/2}} \times \int_{R(b_2)} \frac{b_1^r \exp\{-[b_1-\mu(b_2)^2/[2\sigma^2(1-\rho^2)]\}}{\{2\pi\sigma^2(1-\rho^2)\}^{1/2}} \, db_1 \, db_2 \qquad (5.5)$$

where the region of integration for the inner integral are $b_1 > |b_2|$ and $b_1 < -|b_2|$. The inner integral can be evaluated easily for low moments, r, giving, for $r = 0$,

$$\Phi(z_1) + 1 - \Phi(z_2)$$

for $r = 1$,

$$\sigma(1-\rho^2)^{1/2}[\phi(z_2) - \phi(z_1)] + \mu(b_2)[\Phi(z_1) + 1 - \Phi(z_2)]$$

for $r = 2$,

$$\begin{aligned} &\sigma(1-\rho^2)^{1/2}\{[|b_2| + \mu(b_2)]\phi(z_2) + [|b_2| - \mu(b_2)]\phi(z_1)\} \\ &\quad + [\mu^2(b_2) + \sigma^2(1-\rho^2)][\Phi(z_1) + 1 - \Phi(z_2)], \end{aligned}$$

where ϕ and Φ are the probability density and distribution function of the standard normal distribution, and

$$\begin{aligned} z_1 &= [-|b_2| - \mu(b_2)]/[\sigma^2(1-\rho^2)]^{1/2} \\ z_2 &= [-|b_2| - \mu(b_2)]/[\sigma^2(1-\rho^2)]^{1/2}. \end{aligned}$$

Numerical integration can then be used to determine I_r. Unfortunately, none of the derivatives of the kernel of (5.5) is continuous at $b_2 = 0$, so that Hermite integration cannot be used.

Table 5.1 *Values of the expected value, $E(b_1|sel.)$, and standard deviation, st. dev.$(b_1|sel.)$, of b_1 when variable X_1 is selected, with $\gamma_1 = 1{\cdot}0$*

		$\sigma = 0{\cdot}3$		$\sigma = 0{\cdot}5$	
	γ_2	$E(b_1\|sel.)$	st. dev.$(b_1\|sel.)$	$E(b_1\|sel.)$	st. dev.$(b_1\|sel.)$
	0·0	1·02	0·28	1·11	0·43
$\rho = -0{\cdot}6$	0·5	1·08	0·25	1·21	0·39
	1·0	1·21	0·21	1·36	0·35
	1·5	1·39	0·18	1·53	0·32
	2·0	1·60	0·16	1·72	0·30
	0·0	1·01	0·29	1·10	0·45
$\rho = 0{\cdot}0$	0·5	1·05	0·28	1·15	0·44
	1·0	1·17	0·25	1·28	0·42
	1·5	1·35	0·23	1·46	0·40
	2·0	1·57	0·22	1·66	0·38
	0·0	1·02	0·28	1·11	0·43
$\rho = 0{\cdot}6$	0·5	1·01	0·29	1·09	0·46
	1·0	1·11	0·28	1·17	0·48
	1·5	1·30	0·27	1·34	0·51
	2·0	1·53	0·27	1·52	0·58

However the kernel is well behaved on each side of $b_2 = 0$ so that integration in two parts presents no problems. This can be done using half-Hermite integration for which tables of the weights and ordinates have been given by Steen, Byrne and Gelbard (1969) and Kahaner, Tietjen and Beckmann (1982).

Table 5.1 contains some values of the mean and standard deviation of b_1 when variable X_1 is selected. In this table the expected value of b_1 over all cases, i.e. whether or not variable X_1 is selected, is held at 1·0. To apply the table when the expected value of b_1, i.e. γ_1, is not equal to 1·0, the X-variables should be scaled as described earlier, and the Y-variable should be scaled by dividing by $\gamma_1(\mathbf{X}_1'\mathbf{X}_1)^{1/2}$. The residual standard deviation after fitting both variables, σ, should be divided by the same quantity, γ_2 should be multiplied by $(\mathbf{X}_2'\mathbf{X}_2/\mathbf{X}_1'\mathbf{X}_1)^{1/2}/\gamma_1$, and $\rho = \mathbf{X}_1'\mathbf{X}_2/(\mathbf{X}_1'\mathbf{X}_1\mathbf{X}_2'\mathbf{X}_2)^{1/2}$. Thus, for the simulation at the start of this section, we had

$$\gamma_1 = 1{\cdot}0, \quad \gamma_2 = 0{\cdot}2, \quad \mathbf{X}_1'\mathbf{X}_1 = 12, \quad \mathbf{X}_2'\mathbf{X}_2 = 300,$$
$$\mathbf{X}_1'\mathbf{X}_2 = -3, \quad \sigma = 1.$$

After scaling, and using asterisks as before to denote the scaled

values, we have

$$\gamma_1^* = 1{\cdot}0, \qquad \gamma_2^* = 1{\cdot}0, \qquad \rho = -0{\cdot}05, \qquad \sigma^* = 1/\sqrt{12} = 0{\cdot}289.$$

This is close to the entry in the table for $\rho = 0$, $\sigma = 0{\cdot}3$, for which the expected value of b_1 is 1·17 when X_1 is selected. However the average value in our simulations was 1·288, which is significantly larger than 1·17. The reason for the apparent discrepancy is that the theory above is for given values of X_1 and X_2 whereas X_1 and X_2 were random variables in our simulation taking different values in each artificial data set. As a check, the simulations were repeated with fixed values of X_1 and X_2 such that $\mathbf{X}_1'\mathbf{X}_1 = 12$, $\mathbf{X}_1'\mathbf{X}_2 = 0$, $\mathbf{X}_2'\mathbf{X}_2 = 300$ and $\sigma = 1$. The average value of b_1 for the 106 cases out of 200 in which variable X_1 was selected was 1·202.

Figure 5.3 is intended to give a geometric interpretation of selection bias. The ellipses are for two different cases, and are ellipses of constant probability density in (b_1, b_2) such that most pairs of values of (b_1, b_2) are contained within them. For this figure, it is assumed that X_1 and X_2 have both been scaled to unit length so that the regions in which X_1 and X_2 are selected are bounded by lines at 45 degrees to the axes. Thus X_1 is selected in regions to the left and right of the origin, and X_2 is selected if (b_1, b_2) is in the top or bottom regions.

Ellipse A represents a case in which b_1 is positive and b_2 is usually positive. The thin horizontal and vertical lines running from the centroid of the ellipse are at the unconditional expected values of b_1 and b_2. When X_2 is selected, (b_1, b_2) is in the small sliver to the top left of the ellipse or just above it. Most of the sliver is above the expected value of b_2, so that b_2 is biased substantially in those rare cases in which it is selected. As the few cases in which X_1 is not selected give values of b_1 less than its expected value, b_1 is biased slightly on the high side when X_1 is selected.

Ellipse B represents a case in which the principal axis of the ellipse is perpendicular to the nearest selection boundary. In this case, far more of the ellipse is on the 'wrong' side of the boundary and the biases in both b_1 and b_2, when their corresponding variables are selected, are relatively large.

In both case A and case B, the standard deviation of b_1 and b_2 when their variables are selected, are less than the unconditional standard deviations. This applies until the ellipses containing most of the joint distribution include the origin.

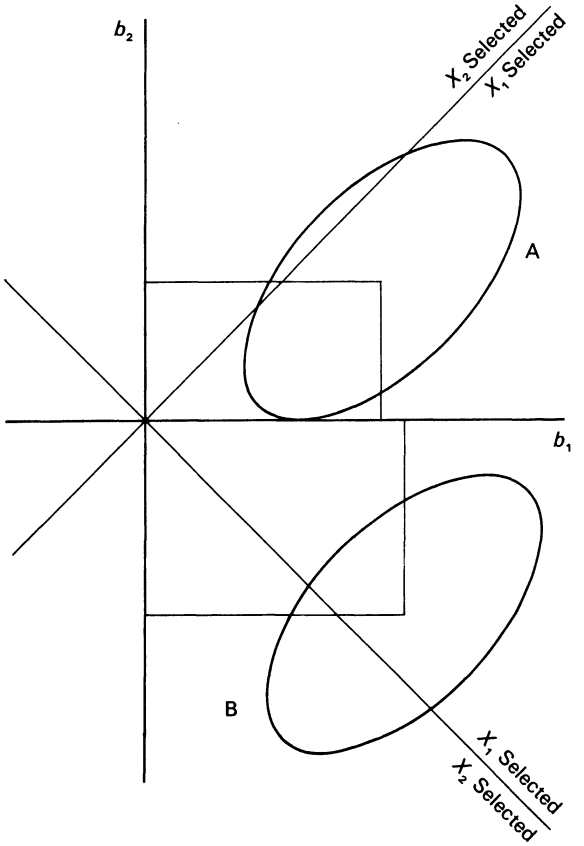


Fig. 5.3 *A figure to illustrate the size and direction of selection bias.*

Both ellipses shown here have $\rho > 0$; for $\rho < 0$ the directions of the major and minor axes of the ellipse are reversed. As ρ approaches $\pm 1{\cdot}0$, the ellipses become longer and narrower; for $\rho = 0$, the ellipses are circles. In can be seen that when γ_1 and γ_2 have the same signs and are well away from the origin, the biases are smallest when $\rho \gg 0$ and largest when $\rho \ll 0$. The case $\rho = 0$, that is when the predictor variables are orthogonal, gives an intermediate amount of bias. The popular belief that orthogonality gives protection against selection bias is fallacious; highly correlated variables will often give more protection.

The above derivations have been of the properties of the regression

coefficients. A similar exercise can be carried out using the joint distribution of the *RSS*'s for the two variables, to find the distribution of the minimum *RSS*. This is somewhat simpler as both *RSS*'s must be positive or zero, and the boundary is simply the straight line at which the two *RSS*'s are equal.

In the two-variable case it is possible to construct a function which approximates the selection bias, and then to use that function to eliminate much of the bias. From Table 5.1 it is obvious that the most important term in the function is that which measures the degree of competition between the variables. For σ sufficiently small that only one boundary crossing needs to be considered, the bias in b_1 is well approximated by a rational polynomial of the form

$$E(b_1 | X_1 \text{ selected}) = 1 + \sigma \cdot \frac{p_1 + p_2 x + p_3 x^2 + p_4 x^3}{1 + p_5 x + p_6 x^2} \tag{5.6}$$

where $x = (\gamma_2 - 1)/\sigma$, and the coefficients p_i are slowly changing functions of ρ. Using extensive tabulations, from which Table 5.1 was extracted, the (LS) fitted values of the parameters were as shown in Table 5.2 for three values of ρ, and for $x \geqslant 0$.

The discussion above has been for the case in which one and only one of the two predictors is selected. In a practical situation, both predictors or neither may be selected. This increases the number of alternative models to four. The regions of the (b_1, b_2)-space in which each model is selected are then much more complex than those shown in Fig. 5.3. This problem has been considered in detail by Sprevak (1976).

Table 5.2 *Coefficients in the rational polynomial* (5.6) *for three values of* ρ

Parameter	$\rho = -0{\cdot}6$	$\rho = 0$	$\rho = 0{\cdot}6$
p_1	0·71	0.56	0·36
p_2	0·43	0·40	0·34
p_3	0·090	0·094	0·103
p_4	0·0067	0·0075	0·0097
p_5	0·15	0·14	0·11
p_6	0·015	0·019	0·029

5.3 Selection bias in the general case, and its reduction

The type of approach used in the previous section is not easily extended. In the general case, there are many boundaries between the regions in which different variables or subsets of variables are selected, so that numerical integration rapidly ceases to be feasible. Also, the selection bias is a function of the selection method which has been used, and of the stopping rule.

If the predictor variables are orthogonal, as for instance when the data are from a designed experiment or when the user has constructed orthogonal variables from which to select, then we can easily derive upper limits for the selection bias. If we scale all the predictor variables to have unit length, then the worst case is when all of the regression coefficients have the same expected value (or strictly that the absolute values of the expected values are the same). If all of the regression coefficients have expected value equal to $\pm \beta$ with sample standard deviation equal to σ ($\sigma \ll \beta$), then if we pick just one variable, that with the largest regression coefficient in absolute value, the expected value of the absolute value of its regression coefficient is $\beta + \xi_i \sigma$, where ξ_1 is the first-order statistic for a random sample of k values from the standard normal distribution, where k is the number of variables available for selection. If we pick the three variables which give the best fit to a set of data, then the bias in the absolute values of the regression coefficients will have expected value equal to $\sigma(\xi_1 + \xi_2 + \xi_3)/3$, where ξ_i is the ith-order statistic. Thus, if we have say 25 available predictor variables, the bias in the regression coefficient of a single selected variable will be about 1·97 standard deviations.

The order-statistic argument gives only a rough guide to the likely size of the selection bias in general, though it does give an upper limit when the predictor variables are orthogonal. The selection bias can be higher than the order-statistic limit for correlated variables. In the author's experience, selection biases up to about three standard deviations are fairly common in real-life problems, particularly when an exhaustive search has been used to select the chosen subset of variables. In Chapter 6 we will see that the selection bias term is extremely important in estimating the magnitude of prediction errors, and in deciding upon a stopping rule.

To illustrate the extent of selection bias in practice, let us use the STEAM and POLLUTE data sets. We do not know the true

population regression coefficients. What we can do though is to split the data into two parts. The first part can then be used to select a subset and to estimate LS regression coefficients for that subset. The second part can then be used as an independent data set to give unbiased estimates for the same subset already selected.

The data sets were divided as nearly as possible into two equal parts. In the case of the STEAM data, which had 25 observations, 13 were used in the first part and 12 in the second. The two data sets were each split randomly into two parts, with the whole exercise repeated 100 times. An arbitrary decision was made to look at subsets of exactly three predictors plus a constant. Exhaustive searches were carried out to find the best-fitting subsets.

Table 5.3 shows the different subsets which were selected for the two data sets. We note that variable number 7 (average temperature) was selected 94 times out of 100 for the STEAM data, while variable number 9 (% nonwhite in the population) was selected 91 times out of 100 for the POLLUTE data. These two will be considered 'dominant' variables.

For the first splitting of the STEAM data, the regression coeffi-

Table 5.3 *Subsets of three variables which gave best fits to random halves of the STEAM and POLLUTE data sets*

STEAM				*POLLUTE*			
Subset			*Frequency*	*Subset*			*Frequency*
4	5	7	24	1	9	14	26
5	7	8	13	2	6	9	13
1	5	7	9	2	4	9	8
1	7	8	9	1	8	9	7
5	6	7	8	2	9	14	7
5	7	9	8	6	9	11	7
1	6	7	4	2	8	9	3
7	8	9	3	6	9	14	3
1	2	7	2	9	10	14	3
1	3	7	2	1	8	11	2
1	4	7	2	3	9	14	2
1	6	9	2	7	8	9	2
1	7	9	2	9	12	13	2
2	7	8	2				
Plus 10 others selected once				Plus 15 others selected once			

Table 5.4

	First half of data		*Second half*
	Regn coeff.	*Approx. s.e.*	*Regn coeff.*
Constant	3·34	1·96	1·99
Variable 4	0·19	0·05	0·60
Variable 5	0·48	0·27	−0·03
Variable 7	−0·080	0·011	−0·082

cients were as given in Table 5.4. The approximate standard errors shown are the usual LS estimates applicable when the model has been chosen independently of the data.

Ignoring the constant in the model, which was not subject to selection, the regression coefficient for variable 4 has increased by about eight standard errors from the data which selected it to the independent data, the regression coefficient for variable 5 has almost vanished, while that for variable 7 has remained steady.

The scale for each regression coefficient is different, so to combine the information on the shift of different regression coefficients from one half of the data to the other, the quantities

$$z_i = \frac{b_{2i} - b_{1i}}{s_{1i}} \operatorname{sign}(b_{1i})$$

were formed, where b_{1i}, b_{2i} are the LS regression coefficients for variable number i for each of the halves of the data, and s_{1i} is the estimated standard error of the ith regression coefficient calculated from the first half. Thus z_i is the shift, in standard errors, from the first half to the second half. The sign of z_i is positive if b_{2i} has the same sign as b_{1i} and is larger in magnitude, and negative if the regression coefficient shrank or changed sign.

Table 5.5 shows the frequency of shifts of the regression coefficients, that is of the z_i's, for the two data sets. In the majority of cases, the unbiased regression coefficients were smaller, with an average shift of just under one standard error.

Let us separate out the two 'dominant' variables. The average shift for variable 7 for the STEAM data was −0·08 of a standard error, while that for variable 9 for the POLLUTE data was +0·04 of a standard error. Thus in this case there appears to be very little

Table 5.5 *Frequency of shift of regression coefficients from data used to select model to those from independent data*

Shift (z) in std errors	*STEAM data*	*POLLUTE data*
< -5	10	9
-5 to -4	19	7
-4 to -3	29	22
-3 to -2	46	44
-2 to -1	59	71
-1 to 0	43	58
0 to 1	33	37
1 to 2	15	30
2 to 3	9	13
3 to 4	4	6
4 to 5	15	2
> 5	18	1
Average shift	$-0{\cdot}71$	$-0{\cdot}90$

Table 5.6 *Frequency of ratios of residual variance estimates for the data used for selection and for independent data*

Variance ratio	*STEAM data*	*POLLUTE data*
0·0–0·1	3	0
0·1–0·2	9	1
0·2–0·3	7	7
0·3–0·4	11	6
0·4–0·5	11	11
0·5–0·6	8	21
0·6–0·7	10	15
0·7–0·8	7	7
0·8–0·9	3	6
0·9–1·0	5	9
1·0–1·1	5	6
1·1–1·2	2	7
1·2–1·3	4	2
1·3–1·4	3	1
$> 1{\cdot}4$	12	1

overall average bias for the dominant variables, but an average bias of just over one standard error for the other variables.

Table 5.6 shows a histogram of the ratio s_1^2/s_2^2 where s_1^2, s_2^2 are the usual residual variance estimates for the two halves of the data for the subset of three variables which best fitted the first half of the data. We see that the average ratio was 0·76 for the STEAM data and 0·69 for the POLLUTE data. If s_1^2, s_2^2 had been independent estimates of the same variance, and the regression residuals have a normal distribution, then the expected value of this variance ratio is $\nu_2/(\nu_2 - 2)$ where ν_2 is the number of degrees of freedom of s_2^2. These numbers of degrees of freedom are $12 - 4 = 8$ and $25 - 4 = 21$ for the STEAM and POLLUTE data sets respectively. Thus the expected values of s_1^2/s_2^2 are 1·33 and 1·11 for unbiased estimates of s_1^2. This gives a rough estimate of the extent of over fitting which has occurred.

5.3.1 *Monte Carlo estimation of bias in forward selection*

The simplest selection rule is forward selection; let us see if we can estimate selection biases in this case. Suppose that we have selected the first $(p-1)$ variables and are considering which variable to add next. At this stage, the $(p-1)$ selected variables will be in the first $(p-1)$ rows of the triangular factorization. Let r_{iy} be the ith projection of Y, that is the ith element in the vector $\boldsymbol{Q}'\mathbf{Y}$. Now if the true relationship between Y and the complete set of predictor variables is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where the residuals, ε, are independently sampled from the normal distribution with zero mean and standard deviation σ, then

$$\begin{aligned}\boldsymbol{Q}'\mathbf{Y} &= \boldsymbol{R}\boldsymbol{\beta} + \boldsymbol{Q}'\boldsymbol{\varepsilon} \\ &= \boldsymbol{R}\boldsymbol{\beta} + \boldsymbol{\eta} \text{ say,}\end{aligned}$$

which means that the projections, r_{iy}, are normally distributed about their expected values, given by the appropriate element in $\boldsymbol{R}\boldsymbol{\beta}$, with standard deviation σ.

The reduction in *RSS* from adding next the variable in row p is r_{py}^2. Hence the variable in row p is more likely to be selected next if its deviation from its expected value, η_p, is large, say greater than

σ, and has the same sign as $\boldsymbol{R\beta}$. We can then use a Monte Carlo type of method to estimate the bias in the projections when the corresponding variable is selected. The following is an algorithm for doing this:

1. Rotate the next selected variable into row p if it is not already there.
2. Move the original projection, r_{py}, towards zero by a first guess of the bias, e.g. by $\hat{\sigma}$ where $\hat{\sigma}$ is an estimate of the residual standard deviation with all of the predictor variables in the model.
3. Generate pseudo-random normal vectors of values of η_i with zero mean and standard deviation σ, and add these to the projections r_{iy} for rows $p,\ldots,k$. Find whether the variable in row p is still selected with these adjusted projections.
4. Repeat step 3 many times and average the values of η_p for those cases in which the variable in row p is selected. Take this average as the new estimate of the bias in r_{py}. Repeat steps 2–4 until the estimate of the bias stabilizes.

The above technique was applied to the STEAM and POLLUTE data sets used in earlier chapters. It was not appropriate to apply it to either the CLOUDS or DETROIT and sets as the first had no estimate of the residual standard deviation, and the DETROIT data set has only one degree of freedom for its residual.

For the STEAM data, the first five variables selected in forward selection are those numbered 7, 1, 5, 4 and 9 in that order. With these variables in that order, the projections in vector $\boldsymbol{Q'Y}$ are as follows:

Variable	Const.	7	1	5	4	9	2	3	6	8
Projection	47·12	6·75	3·05	1·12	−0·94	0·59	−0·56	0·46	0·42	−0·94

The sample estimate of the residual standard deviation is 0·57 with 15 degrees of freedom. Comparing the projections with this residual standard deviation suggests that we are only justified in including two or possibly three variables, plus the constant, in our model. Applying the above algorithm to the selection of the first variable, that is variable number 7, after subtracting $\hat{\sigma} = 0{\cdot}57$ from the projection in row 2, the variable was selected 200 times out of 200. The estimate of bias obtained by strictly applying the method above was the sum of 200 pseudo-random normal deviates with zero mean and standard

Table 5.7

Iteration	*Starting bias estimate*	*Times out of 200 variable 1 selected*	*New bias estimate*
1	0·57	72	0·37
2	0·37	87	0·20
3	0·20	105	0·12
4	0·12	103	0·20

deviation $\hat{\sigma}$, and this turned out to be $+0.03$ with the random number generator used. Clearly there was no competition for selection in this case, and that the bias in the projection is zero for all practical purposes.

There was more competition for the next position. Consecutive iterations gave the data shown in Table 5.7. Using 0·20 as the bias estimate reduces the projection from 3·05 to 2·85.

For the third variable (row 4), the competition was greater. In this case the iterations gave the data shown in Table 5.8. Using 0·81 as the bias estimate reduces the projection from 1·12 to 0·31.

Using the adjusted projections and back-substitution, the fitted three-variable regression line changes from

$$Y = 8{\cdot}57 - 0{\cdot}0758X_7 + 0{\cdot}488X_1 + 0{\cdot}108X_5$$

to

$$Y = 9{\cdot}48 - 0{\cdot}0784X_7 + 0{\cdot}637X_1 + 0{\cdot}029X_5$$

and the *RSS* for this model increases from 7·68 to 10·40. Notice that reducing the absolute size of the projections does not necessarily reduce the sizes of all the regression coefficients. It always reduces the size of the coefficient of the last variable selected, but not necessarily the sizes of the others.

Table 5.8

Iteration	*Starting bias estimate*	*Times out of 200 variable 5 selected*	*New bias estimate*
1	0·57	50	0·64
2	0·64	33	0·81
3	0·81	23	0·81

Table 5.9

Variable number	*Original projection*	*Time variable selected in last iteration*	*Bias estimate*	*Adjusted projection*
Const.	7284·0			
9	307·6	176	11·1	296·5
6	−184·0	31	−41·1	−142·9
2	−132·1	28	−40·8	−91·3

For the POLLUTE data set, the residual standard deviation estimate is 34·9 with 44 degrees of freedom with all the variables in the model. The bias estimates again appeared to converge very rapidly. The results obtained are shown in Table 5.9. Using these projections, the fitted three-variable regression line changes from

$$Y = 1208{\cdot}1 + 5{\cdot}03X_9 - 23{\cdot}8X_6 - 1{\cdot}96X_2$$

to

$$Y = 1138{\cdot}7 + 4{\cdot}65X_9 - 18{\cdot}9X_6 - 1{\cdot}35X_2$$

and the *RSS* increases from 82 389 to 111 642.

This simple 'intuitive' method appears to produce good point estimates of regression coefficients, but has a number of shortcomings. First, when the first variable is selected, its sample projection may be appreciably larger in absolute value than its expected value. The method allows for that bias and for the fact that the variable may have been selected because some of the other projections deviated substantially from their expected values. However, any bias in these other projections was forgotten when we proceeded to select the next variable. This can easily be accommodated by estimating the biases of all the projections at each iteration instead of estimating only the bias for the projection of the selected variable. In most cases this will make very little difference.

Another objection to the method is that is provides only point estimates without confidence limits or approximate standard errors.

Implicit in the method just described is the notion that if there is apparently competition among variables for selection, then the variable selected must have a sample projection which is above its expected value.

The above method will not be developed further, though similar methods can be developed for other selection procedures, and it is possible at great computational expense to obtain confidence limits. In the next section, a method based upon conditional likelihood will be described, but before proceeding to that, let us look briefly at other alternative ways of tackling the selection bias problem.

5.3.2 *Shrinkage methods*

Figure 5.1 suggests that some kind of shrinkage should be applied. Two kinds of shrinkage have become popular; these are ridge regression and simple shrinkage of all the regression coefficients by the same factor.

The simplest form of shrinkage estimator is that suggested by James and Stein (1961). Let $T_i, i = 1, 2, \ldots, k$, be unbiased estimators of quantities μ_i, each having the same variance σ^2. The T_i are assumed to be uncorrelated. Consider the shrunken estimators

$$T_i^* = (1 - \alpha)T_i, \qquad 0 < \alpha < 1, \tag{5.7}$$

that is each estimate is shrunk towards zero by the same relative amount. The new estimates will be biased but have lower variances than the T_i's. Suppose we trade off the bias against the reduced variance by minimizing a loss function which is the expected square error:

$$\begin{aligned} \text{loss} &= E \sum_{i=1}^{k} (T_i^* - \mu_i)^2 \\ &= (\text{bias})^2 + \text{variance} \\ &= \alpha^2 \sum_{i=1}^{k} \mu_i^2 + k(1 - \alpha)^2 \sigma^2 \end{aligned}$$

Then

$$\frac{\mathrm{d}(\text{loss})}{\mathrm{d}\alpha} = 2\alpha \sum_{i=1}^{k} \mu_i^2 - 2k(1 - \alpha)\sigma^2.$$

Setting this equal to zero gives

$$\alpha = \frac{k\sigma^2}{\sum \mu_i^2 + k\sigma^2}.$$

Unfortunately this involves the unknown μ_i's. Now $\sum T_i^2$ is an unbiased estimator of $(\sum \mu_i^2 + k\sigma^2)$, and substituting this into (5.7) gives the estimator

$$T_i^* = \left(1 - \frac{k\sigma^2}{\sum T_i^2}\right) T_i.$$

Our derivation above assumed that α was not a random variable, yet we have now replaced it with a function of the T_i's, thus invalidating the derivation. If we allow for the variance of α, it can be shown (see James and Stein, 1961 for more details) that the estimator

$$T_i^* = \left(1 - \frac{(k-2)\sigma^2}{\sum T_i^2}\right) T_i \tag{5.8}$$

gives a smaller squared error for all $k > 2$. This is the James–Stein estimator.

In practice, σ^2 must be estimated. If the usual estimate

$$s^2 = \sum_{i=1}^{k} (T_i - \bar{T})^2/(k-1)$$

is used, where $\bar{T}$ is the sample mean of the T_i's, then Stein (1962) shows that the σ^2 in the above estimators should be replaced with $s^2(k-1)/(k+1)$.

Lindley (pp. 285–7 of Stein, 1962) suggested shrinkage towards the mean, rather than shrinkage towards zero. His estimator is

$$T_i^* = \bar{T} + \left(1 - \frac{(k-3)\sigma^2}{\sum (T - \bar{T})^2}\right)(T_i - \bar{T}).$$

The James–Stein estimator has been controversial. The following is a quote from Efron and Morris (1973):

> The James–Stein estimator seems to do the impossible. The estimator of each μ_i is made to depend not only on T_i but on the other T_j, whose distributions seemingly are unrelated to μ_i, and the result is an improvement over the maximum likelihood estimator no matter what the values of $\mu_1, \mu_2, \ldots, \mu_k$. Thus we have the 'speed of light' rhetorical question, 'Do you mean that if I want to estimate tea consumption in Taiwan I will do better to estimate simultaneously the speed of light and the weight of hogs in Montana?'

Of course the other T_i's (the speed of light and weight of hogs) are used to estimate the shrinkage factor α, and a critical assumption is that all the T_i's have the same variance. It seems extremely improbable that these three disparate estimates would have the same variance, even after appropriate scaling say by measuring the speed of light in knots and the weight of hogs in carats.

James–Stein shrinkage cannot be applied directly to regression coefficients as they do not in general have the same variance and are usually correlated. However, the LS projections, from which we calculate the regression coefficients, do have these properties.

Let t_i and τ_i denote the LS projections and their expected values, and let $\boldsymbol{R}$ denote the upper-triangular Cholesky factor of the design matrix $\boldsymbol{X}$. Then as

$$\boldsymbol{R}\mathbf{b} = \mathbf{t}$$
$$\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{\tau}$$

where $\mathbf{b}$ and $\boldsymbol{\beta}$ are the LS estimates of the regression coefficients and their expected values, then the loss function

$$\begin{aligned}\sum_{i=1}^{k} (t_i - \tau_i)^2 &= (t - \tau)'(t - \tau) \\ &= (\mathbf{b} - \beta)'\boldsymbol{R}'\boldsymbol{R}(\mathbf{b} - \boldsymbol{\beta}) \\ &= (\mathbf{b} - \beta)'\boldsymbol{X}'\boldsymbol{X}(\mathbf{b} - \boldsymbol{\beta}).\end{aligned}$$

This is the sum of squares of the elements of $\boldsymbol{X}(\mathbf{b} - \boldsymbol{\beta})$. As the elements of $\boldsymbol{X}\mathbf{b}$ and $\boldsymbol{X}\boldsymbol{\beta}$ are the LS fitted and expected values respectively of the dependent variable, the sum is the sum of squares of differences between the fitted and expected values. Minimizing this sum is then a reasonable objective function in many situations.

The factor α is then

$$\alpha = \frac{(k-2)\sigma^2}{\sum t_i^2}$$

The sum of squares of the LS projections in the denominator is the regression sum of squares (*regn SS*). Furthermore, if all the t_i's are reduced by the same factor, so are the regression coefficients derived from them, so that our new estimates, b_i^*, of the regression coefficients are

$$b_i^* = \left(1 - \frac{(k-2)\sigma^2}{regn\ SS}\right) b_i.$$

This estimator is due to Sclove (1968). Normally σ^2 will be replaced with an estimate $s^2 v/(v+2)$ where s^2 is the usual residual variance estimate and v is its number of degrees of freedom.

Other alternative derivations have been given by Copas (1983). In the first of these methods, he derives the James–Stein/Sclove predictor as the LS predictor of $\mathbf{Y}$ from the predicted LS values $\hat{\mathbf{Y}}$ for the given values of the predictor variables, when the prediction equation is based upon independent data. That is, one set of data has been used to obtain a regression equation which is then used to predict further values of $\mathbf{Y}$. The predictions, $\hat{\mathbf{Y}}$, are then used in a single-variable linear regression, but with the predictor being a random variable which is subject to error. Thus if we have a set of predictions $\hat{y}_i, i = 1, \ldots, m$ obtained from $\hat{\mathbf{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ where the vector of regression coefficients, $\hat{\boldsymbol{\beta}}$, has been obtained from an independent set of data, then the LS predictor of $\mathbf{Y}$ from $\hat{\mathbf{Y}}$ is say $(1-\alpha)\hat{\mathbf{Y}}$ where

$$(1-\alpha) = \frac{\mathbf{Y}'\hat{\mathbf{Y}}}{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}}.$$

Since

$$E(\mathbf{Y}'\hat{\mathbf{Y}}) = regn\ SS$$

and

$$E(\hat{\mathbf{Y}}'\hat{\mathbf{Y}}) = regn\ SS + k\sigma^2,$$

by substituting these expected values we derive the estimate:

$$\alpha = \frac{k\sigma^2}{regn\,SS + k\sigma^2},$$

which is the expression we derived earlier by minimizing the quadratic loss function (which is precisely what we have done here, though it may look somewhat different).

Notice that we have derived the amount of shrinkage to apply to unbiased LS regression coefficients. No allowance has been made for the bias introduced by selection. In most cases the amount of shrinkage using James–Stein estimates will be much less than is needed to overcome selection bias. For instance, for the STEAM data, $k = 9$ (excluding the constant), $s^2 = 0{\cdot}325$, and the regression sum of squares $= 58{\cdot}9$. Using s^2 for σ^2 gives $b_i^* = 0{\cdot}96 b_i$, i.e. the regression coefficients are shrunk by only 4%.

Ridge regression has been described in section 3.9. Much of the vast literature on ridge regression has focused upon mean squared

errors (MSE) of the regression coefficients, i.e.

$$MSE(\hat{\boldsymbol{\beta}}) = E(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}),$$

where $\hat{\boldsymbol{\beta}}$ is an estimate of the vector of regression coefficients, $\boldsymbol{\beta}$. In most cases in practice, the MSE's of prediction are likely to be of more interest. Lawless and Wang (1976) have particularly emphasized this distinction, and have shown that while ridge regression can produce huge reductions in the $MSE(\hat{\boldsymbol{\beta}})$ when the $X'X$-matrix is ill-conditioned, it produces far less reduction in the $MSE(X\hat{\boldsymbol{\beta}})$, that is in the MSE of prediction, and sometimes produces a small increase.

Very little attention has been paid to the ridge trace idea of Hoerl and Kennard (1970b) for variable selection, possibly because of its subjective nature. An explicit rule for deletion of variables has subsequently been given by Hoerl, Schuenemeyer and Hoerl (1986). They suggest using a modified t-statistic:

$$t = \hat{\beta}_i / s_i$$

where the estimates of the regression coefficients are given by

$$\hat{\boldsymbol{\beta}} = (X'X + dI)^{-1} X'\mathbf{Y}$$

after first shifting and scaling each X-predictor to have zero sample mean and unit sample standard deviation, and where the s_i's are the square roots of the diagonal elements of

$$\hat{\sigma}^2 (X'X + dI)^{-1} X'X (X'X + dI)^{-1}.$$

In their simulations, a range of significance levels was used, but those reported in their paper were for a nominal 20% level. These simulations showed that this ridge selection procedure gave good performance in terms of both $MSE(\hat{\boldsymbol{\beta}})$ and $MSE(X\hat{\boldsymbol{\beta}})$ when the Lawless and Wang (1976) value for d was used, i.e.

$$d = \frac{\text{residual mean square}}{\text{regression mean square}},$$

where the mean squares are evaluated using all the available predictors. The ridge regression estimator performed well with this value of d in any case without subset selection. When some of the true β_i's were zero, a moderate improvement was achieved using selection. In the case of the STEAM data, the value of $d =$

0·325/58·9 = 0·0055. That is the diagonal elements of $X'X$ are incremented by about half of 1%.

Both ridge regression and James–Stein shrinkage require a knowledge of the size of the selection bias to make the best choice of the amount of shrinkage to apply. As each method has only one parameter controlling the amount of shrinkage, it cannot be controlled to eliminate or reduce the bias simultaneously in all parameters. We believe that the method of conditional likelihood to be described in section 5.4 is a more satisfactory method of achieving this.

5.3.3 *Using the jack-knife*

A popular method of bias reduction is the so-called 'jack-knife' (Quenouille, 1956; Gray and Schucany, 1972; Miller, 1974). Suppose that T_n is a statistic based upon a sample of n observations, and that

$$E(T_n) = \theta + \frac{a}{n} + \frac{b}{n^2} + o(n^{-2}),$$

where θ is a parameter or vector of parameters which we want to estimate. Then

$$E\{nT_n - (n-1)T_{n-1}\} = \theta - \frac{b}{n(n-1)} + o(n^{-2}),$$

that is, the terms of order n^{-1} in the bias are eliminated, while those of order n^{-2} are reversed in sign and increased very slightly in magnitude.

The jack-knife could be applied to the estimation of regression coefficients or the *RSS* for a model. Suppose that T_n is the LS estimate of the regression coefficients for a subset of variables selected using a particular procedure and n observations. As the bias is due to the fact that the regression coefficient is being estimated conditional upon a certain subset being selected, T_{n-1} obtained from $(n-1)$ observations out of the n must be subject to the same condition. A sample of $(n-1)$ observations can be obtained in n different ways by deleting one of the n observations. Consider all n such samples, and apply the same selection procedure to each. This may be quite feasible if the procedure is forward selection, sequential replacement, or one of the other 'cheap' procedures, but will involve very substantial computational effort if the procedure is an exhaustive search.

Suppose that in m out of the n cases the subset of interest is selected, then we can use the m estimates of T_{n-1} for these cases in the jack-knife and average the results. In limited experiments by the author, the value of m has usually been close to or equal to n and rarely less than $n/2$, though there appears to be no reason why m must be greater than zero. These experiments suggested that the jack-knife may be fairly successful at removing bias but the variance of the jack-knife estimates was very large. There seems to be no reason for expecting selection bias to reduce roughly as n^{-1}, so that that part of the bias which is removed may be fairly small. Unless the term in n^{-1} accounts for most of the bias, the increased variance in the resulting estimates is too high a price to pay.

The order-statistic argument used earlier in this section leads us to anticipate that selection bias may be roughly proportional to $n^{-1/2}$ when the predictor variables are orthogonal and are all equally good choices. Also, substitution in (5.7) gives the leading term in the bias as proportional to $n^{-1/2}$. To eliminate this type of bias, the jack-knife statistic should be modified to

$$[n^{1/2}T_n - (n-1)^{1/2}T_{n-1}]/[n^{1/2} - (n-1)^{1/2}]. \tag{5.9}$$

If we write the jack-knife estimator as

$$(f_n T_n - f_{n-1}T_{n-1})/(f_n - f_{n-1})$$

where we have suggested that $f_n = \sqrt{n}$ is a suitable choice, then the estimator can be rewritten as

$$T_n + \frac{f_{n-1}}{f_n - f_{n-1}}(T_n - T_{n-1}).$$

Thus the initial estimate T_n is moved away from T_{n-1} by a substantial multiple of the difference between them. The use of a Taylor series expansion shows that the square-root jack-knife adjusts the biased estimate by about twice as much as the choice $f_n = n$.

Freedman, Navidi and Peters (1988) have applied the jack-knife to subset selection in regression, but not as described above. All n sets of data with one case deleted were used, with regression coefficients set to zero if a variable was not selected. This means that if a variable was selected for the full data set, then it was selected in perhaps 90% of the sets with one case deleted and rejected in the others. This results in typical values of T_{n-1} being of the order of

10% lower than T_n. This difference is then magnified with the result that the jack-knife appears to perform very poorly. They did not use the square-root variation of the jack-knife. These authors, and Dijkstra and Veldkamp (1988) in the same volume of conference proceedings, have also used a 'bootstrap' technique with very little success; the technique will not be described here. Platt (1982) had previously also advocated the use of the bootstrap after model selection.

5.3.4 *Independent data sets*

Selection bias can be completely eliminated by using independent data sets for the selection of the model and for estimating the regression coefficients. It is rarely sensible, however, to recommend this method in practice as it is inefficient in not using the information from the selection set of data in estimating the regression coefficients. In some cases, only the selected variables will have been measured for the second data set, though in other cases measurements of all variables will be available. If all of the variables have been measured for the second data set then it is tempting to see if our selected subset is one of the best-fitting subsets of its size. Suppose that it is not – what do we do now? We may well find something like the data in Table 5.10 for the best-fitting subsets of three variables. We notice that the best-fitting subset of three variables for the selection set of data does not appear among the best five for the set of data to be used for estimating the regression coefficients; let us suppose that it occurs much further down the list. We notice though that the second-best subset from the first data set occurs quite high up on the other list. It looks like a good choice. This is a crude way of looking for the best-fitting subset for the combined data set, so why

Table 5.10

Rank	*Selection data*	*Regn coeff. data*
Best	3, 7, 14	3, 10, 11
2nd	3, 7, 11	3, 10, 14
3rd	3, 4, 8	3, 7, 11
4th	3, 4, 7	3, 7, 10
5th	3, 8, 11	3, 11, 14

do we not do the search properly for the combined data set? But then we are back with the problem of selection bias if we use the whole data set both to select the model and estimate the regression coefficients.

5.4 Conditional likelihood estimation

A method which can usually be used to obtain parameter estimates, providing that we are prepared to make distributional assumptions, is maximum likelihood. We want to estimate parameters for a subset of variables after we have found that these variables give a better fit to a set of data than some other subsets. Taking the values of the X-variables as given, if we assume that the values of Y are normally distributed about expected values given by $X\boldsymbol{\beta}$, where for the moment X contains all of the predictor variables, with the same variance σ^2, and that the deviations of the Y's from $X\boldsymbol{\beta}$ are independent, then the unconditional likelihood for a sample of n observations is

$$\prod_{i=1}^{n} \phi\{(y_i - \sum \beta_j x_{ij})/\sigma\}$$

where ϕ is the standard normal probability density, that is

$$\phi(x) = (2\pi)^{-1/2} \exp(-\tfrac{1}{2}x^2).$$

Now given that a specific subset has been selected by some procedure (e.g. forward selection, sequential replacement, exhaustive search, etc.), many $\mathbf{Y}$-vectors are impossible as they would not lead to the selection of that subset. The conditional likelihood is then proportional to the above likelihood for acceptable $\mathbf{Y}$-vectors and zero elsewhere. Hence the likelihood of the sample values of Y, given X and that a certain selection procedure has selected a subset of variables, S, is

$$\frac{\displaystyle\prod_{i=1}^{n} \phi\{(y_i - \sum \beta_j x_{ij})/\sigma\}}{\displaystyle\int \cdots \int (\text{above density})\, dy_1 \cdots dy_n}$$

in a region R of the Y-space in which the procedure used selects subset S. The multidimensional integration is also over this region,

and the value of the integral is the *a priori* probability that S is selected given X. Substituting for ϕ, the logarithm of the conditional likelihood (LCL) over region R is then

$$\text{LCL} = -(n/2)\log_e(2\pi\sigma^2) - (2\sigma^2)^{-1}\sum_{i=1}^{n}(y_i - \sum \beta_j x_{ij})^2 - \log_e\{\text{prob } S \text{ is selected}\}. \tag{5.10}$$

The concept of conditioning a likelihood upon the occurrence of an event is one which many statisticians have found difficult to grasp. Let us look at what is being done in a slightly different way. Suppose that instead of having just one sample of n observations, we can go back and collect many more samples of the same size. Suppose that the selection procedure and stopping rule have been specified in advance. Suppose also for simplicity that the values of the predictor variables do not change from one sample to another. If we could take say 1000 samples of size n, it may be that the selection procedure picks subset S in only 50 cases, and that perhaps 150 different subsets will be chosen. We will only be estimating regression coefficients for the subset of variables in S in those 50 cases; we will be estimating different regression coefficients for the other samples. The conditional likelihood above is an unconditional likelihood for the sub-population of samples in which subset S is chosen.

The estimation problem has much in common with the problem of estimating parameters when the data have been censored or truncated in some known way, e.g. because values of Y greater than a certain size cannot be observed or measured.

Furthermore, the boundary of the region in which subset S is chosen is not a function of the parameters to be estimated; it is bounded by the intersection of a large number of quadratic forms in the values of Y and the predictor variables.

The difficulty in using this conditional likelihood is clearly in evaluating the probability of selection of subset S, which is a function of the parameters $\boldsymbol{\beta}$ and σ. In simple cases, such as when there are only two X-variables or in forward selection when the X-variables are orthogonal, the probability that subset S is selected can be evaluated explicitly. In general we need to evaluate the probability that the regression sum of squares for subset S is larger than those of others with which it was compared in the selection procedure used. These regression sums of squares are quadratic forms in the

y_i's so that the region in the Y-space in which one of them exceeds another is quite complex.

The probability of selection, to be denoted by P, can be estimated by Monte Carlo methods in a manner similar to that used in section 5.3. The expected values of the projections, $\boldsymbol{Q}'\mathbf{Y}$, are given by $\boldsymbol{R}\boldsymbol{\beta}$. By adding vectors $\boldsymbol{\eta}$ to $\boldsymbol{R}\boldsymbol{\beta}$, where the elements η_i of $\boldsymbol{\eta}$ are sampled from the $N(0, \sigma^2)$ distribution, random vectors of projections can be obtained. These can be subjected to the selection procedure which found subset S. The proportion of times in which subset S is selected then gives an estimate of P for the vector $\boldsymbol{\beta}$ used. This is a feasible method if one of the 'cheap' methods discussed in Chapter 3, such as forward selection, but it is not as practical in conjunction with an exhaustive search. An alternative method which it is feasible to use with an exhaustive search procedure is to consider only those subsets of variables which were found to be closely competitive with subset S. If say the best 10 or 20 subsets of each size which were found during the search for subset S were recorded, then these can be used. In the Monte Carlo simulations, the regression sum of squares for subset S can then be compared with these other 9 or 19 subsets. The probability that subset S fits better than these other subsets can then be used as an approximation to the required probability of selection.

Many ways of maximizing (5.10) are possible. For instance, a simplex method such as the Nelder and Mead (1965) algorithm could be used. Alternatively, the logarithm of P could be approximated by a quadratic form in $\boldsymbol{\beta}$ by evaluating it at $k(k+1)/2$ points and fitting a quadratic surface. Either of these methods requires a fairly large number of estimates of P and so requires a very substantial amount of computation.

Let us rewrite (5.10) in terms of an orthogonal reduction with sample projections $\mathbf{t} = \boldsymbol{Q}'\mathbf{Y}$, and their true but unknown expected values $\boldsymbol{\tau}$. The RSS using the true regression coefficients (or equivalently, using the projections) is

$$RSS_k + \sum_{j=1}^{k} (\tau_j - t_j)^2.$$

When the model has been chosen independently of the data, the LS estimate of τ_j is simply the sample projection t_j. If we substitute $\boldsymbol{\delta} = \boldsymbol{\tau} - \mathbf{t} = \boldsymbol{R}\boldsymbol{\beta} - \boldsymbol{Q}'\mathbf{Y}$, (5.10) can be rewritten as

$$LCL = \text{const.} - (2\sigma^2)^{-1}\left(RSS_k + \sum_j \delta_j^2\right) - \log_e P(\delta) \qquad (5.11)$$

where $P(\boldsymbol{\delta})$ = the probability of selection for given $\boldsymbol{\delta}$. We now maximize the *LCL* with respect to these deviations, $\boldsymbol{\delta}$, rather than with respect to $\boldsymbol{\beta}$. An alternative way of thinking of this method is as a transformation from the original X-variables to a set of orthogonal Q-variables. The regression coefficients with respect to which we are maximizing the *LCL* are the elements of $\boldsymbol{R\beta}$ which are the regression coefficients of Y upon the columns of the $\boldsymbol{Q}$-matrix.

Differentiating (5.11) we obtain

$$\frac{\mathrm{d}(LCL)}{\mathrm{d}\delta_j} = -\frac{\delta_j}{\sigma^2} - \frac{\mathrm{d}P/\mathrm{d}\delta_j}{P}. \tag{5.12}$$

By equating the left-hand side of (5.12) to zero, we obtain the following interative method for obtaining the rth estimate, $\delta_j^{(r)}$, from the preceding one

$$\delta_j^{(r)} = -\sigma^2 \frac{\mathrm{d}P/\mathrm{d}\delta_j}{P}$$

where the right-hand side is evaluated at $\delta_j^{(r-1)}$. The LS solution, which corresponds to $\delta_j = 0$, can be used as a starting point.

Appendix 5A describes in more detail an algorithm which has been developed to maximize the *LCL*. In this algorithm, Monte Carlo methods are used to estimate the probability of selection at points on an experimental design based around the LS projections. To avoid further Monte Carlo sampling each time that the estimates of the projections are updated, the initial sample is treated as a biased sample from the population of projections which would be generated with the new parameters.

The maximum likelihood estimates obtained for the biases $(-\delta_j)$ in the LS projections turn out to be simply the differences between the weighted averages of the projections when S is selected, and the weighted average for all cases. The process is iterative as the weights are functions of the estimated biases.

NB: The *LCL* is maximized with respect to all of the projections, not merely those for the variables in the selected subset, S. The variables in the orthogonal reduction are ordered so that those for the selected variables are first. After the maximum likelihood (ML) estimation of all of the projections, those for the variables not selected are dropped (set equal to zero) and the regression coefficients calculated using the first p projections.

The residual variance, σ^2, can also be estimated by maximizing

(5.10). Note that the 'constant' shown in (5.11) is a function of σ^2. In practice, the estimates thus obtained have been almost equal to those obtained by assuming that $P(\delta)$ is independent of σ^2. A simpler alternative is to use the estimate

$$\hat{\sigma}^2 = (RSS_k + \sum_j \delta_j^2)/(n-k). \tag{5.13}$$

This gives an unbiased estimate when there is no competition for selection, i.e. when the δ_j's $= 0$.

The usual estimate of the covariance matrix for the parameter estimates is minus the inverse of the (Fisher's) information matrix, that is as the inverse of minus the matrix of second derivatives of the *LCL*. Readers who are not familiar with this method and its assumptions are referred to Kendall and Stuart (1961 and later editions, Chapter 18), Silvey (1975, Chapter 4) or Cox and Hinkley (1974, Chapter 8). The second derivatives of (5.12) are

$$\frac{d^2(LCL)}{d\delta_i\, d\delta_j} = -\frac{\delta_{ij}}{\sigma^2} + \frac{(dP/d\delta_i)(dP/d\delta_j)}{P^2} - \frac{d^2P/(d\delta_i\, d\delta_j)}{P} \tag{5.14}$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

The use of ML in any nontrivial application always raises certain questions. First, does the likelihood have discontinuities, singularities or multiple maxima? Secondly, what are the properties of the estimates? Is the bias small? Is the approximate covariance matrix of any value? Are the estimators fairly robust against failures of the assumptions?

Our conditional likelihood has only the trivial singularity at $\sigma = 0$, and all of its derivatives are continuous everywhere else. For variables in subset S, the *LCL* is almost quadratic in δ_j as the variation in the sum of squares dominates the variation in $\log_e P$. However when δ_j is large for the projection of a variable not in S, the probability of selection is very small and then $\log_e P(\delta)$ changes very rapidly. For the standard normal distribution, the tail probability, often denoted by $Q(x)$, of an observation more than x standard deviations from the mean is approximately $\phi(x)/x$ for positive x (see e.g. Kendall and Stuart, 1961, Chapter 5, or Abramowitz and Stegun, 1964, formulae 26.2.12–14). If the variables

have been ordered so that those for subset S are first, then for large δ_j for a variable not in S, the probability of selection is approximately

$$(2\pi\sigma^2)^{-1/2}|(\delta_j-\mu)/\sigma|^{-1}\exp\{-\tfrac{1}{2}(\delta_j-\mu)^2/\sigma^2\}$$

for some location parameter μ. Substituting in (5.11) we find that the terms in δ_j^2 cancel and that the *LCL* is approximately linear in δ_j, with the value of μ determining whether it increases or decreases. Experience so far has been that the *LCL* has always decreased and the author conjectures that this will always be so. However, it is also clear that the *LCL* falls away much more slowly with δ_j on that side of the maximum on which $P(\delta)$ is small than it does on the other side. It appears that the *LCL* does not have multiple maxima, excepting its singularity at $\sigma^2=0$, but the author does not have a proof that this will always be the case.

Asymmetry of the *LCL* surface means that the covariance matrix obtained from the second derivatives (5.14) must be used with caution. If the probability of selection at the maximum-*LCL* point is fairly high, say more than 10%, then the *LCL* surface will be reasonably close to quadratic for most practical purposes unless the user wants say 99·9% confidence limits on parameter values. However, the probability of selection will often be very much smaller than this.

The probability of selection depends upon the distributional assumptions which have been made. The further we penetrate into the tail of a distribution, the more sensitive our methods are to distributional assumptions. Hence if the probability of selection at the maximum-*LCL* point is say 20% then we would have obtained much the same result if we had used some other distribution than the normal, though the mathematical manipulations would have been far more complex. If the probability of selection turns out to be less than 1% then the estimates would have been very different with other distributional assumptions.

At this stage we have estimates of the projections and the residual variance for the model containing all k variables. That is, our fitted model at this stage is

$$\mathbf{Y}=\sum_{j=1}^{k}(r_{jy}+\hat{\delta}_j)\mathbf{Q}_j,$$

where the r_{jy}'s are the elements of $\boldsymbol{Q}'\mathbf{Y}$, and the residuals are believed to have zero mean and variance $\hat{\sigma}^2$. If the first p variables are those

in subset S, then we can write this as

$$\mathbf{Y} = \sum_{j=1}^{p} (r_{jy} + \hat{\delta}_j)Q_j + \sum_{j=p+1}^{k} (r_{jy} + \hat{\delta}_j)Q_j + \boldsymbol{\varepsilon}.$$

The second term on the right-hand side is the omission bias. If the X-variables can be considered as random variables, then the second term can be treated as additional residual variation and the residual variance for subset S can be estimated as

$$\left\{(n-k)\hat{\sigma}^2 + \sum_{j=p+1}^{k} (r_{jy} + \hat{\delta}_j)^2\right\} \Big/ (n-k). \tag{5.15}$$

To obtain the regression coefficients for subset S we must transform back from the orthogonal Q-variables to the original X-variables using

$$\boldsymbol{X}_s = \boldsymbol{Q}_s \boldsymbol{R}.$$

Hence the estimate of the regression coefficients for subset S, $\hat{\beta}_s$, are obtained by solving by back-substitution

$$\boldsymbol{R}_s \hat{\boldsymbol{\beta}}_s = (\boldsymbol{Q}_s \mathbf{Y} + \boldsymbol{\delta}_s) \tag{5.16}$$

and hence the covariance matrix for $\hat{\boldsymbol{\beta}}_s$ is

$$\boldsymbol{R}_s^{-1} \boldsymbol{V}(\boldsymbol{\delta}_s) \boldsymbol{R}_s^{-T} \tag{5.17}$$

where $\boldsymbol{V}(\boldsymbol{\delta}_s)$ is the covariance matrix for the subset of parameters $\boldsymbol{\delta}_s$, extracted from the covariance matrix for δ obtained from the second derivatives (5.14).

5.5.1 *Conditional maximum likelihood – two competing variables*

Let us see how the conditional maximum likelihood method performs when we have two competing variables of which one is to be selected. For simplicity, suppose we are fitting a model without an intercept. Let the orthogonal reductions of X and Y be

$$\boldsymbol{X} = (\mathbf{X}_1, \mathbf{X}_2)$$

$$= \boldsymbol{Q} \begin{pmatrix} \boldsymbol{R} \\ 0 \end{pmatrix}$$

and

$$\mathbf{Y} = \boldsymbol{Q} \begin{pmatrix} \mathbf{T} \\ e \end{pmatrix}$$

where the Cholesky factor of X is

$$\boldsymbol{R} = \begin{pmatrix} r_{11} & r_{12} \\ & r_{22} \end{pmatrix}$$

and the LS projections are

$$\boldsymbol{T} = \begin{pmatrix} t_1 \\ t_2 \end{pmatrix}.$$

The regression sum of squares if only variable X_1 is used is then t_1^2. To find the regression sum of squares if variable X_2 is used instead, we need to reverse the ordering of the columns for X_1 and X_2. To obtain the corresponding orthogonal reduction, we reverse the columns of $\boldsymbol{R}$ and then use a planar rotation to return it to upper-triangular form, i.e. we find c and s such that

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix} \begin{pmatrix} r_{12} & r_{11} \\ r_{22} & \end{pmatrix} = \begin{pmatrix} (r_{12}^2 + r_{22}^2)^{1/2} & cr_{11} \\ & -sr_{11} \end{pmatrix}.$$

Hence $c = r_{12}/(r_{12}^2 + r_{22}^2)^{1/2}$ and $s = r_{22}/(r_{22}^2 + r_{22}^2)^{1/2}$. Applying the rotation to the T-vector, the projections become

$$\begin{pmatrix} ct_1 + st_2 \\ -st_1 + ct_2 \end{pmatrix}.$$

If we select the variable with the larger regression sum of squares, variable X_1 is selected when

$$|t_1| > |ct_1 + st_2|. \tag{5.18}$$

In practice, both variables would normally be selected if both $|t_1|$ and $|t_2|$ are large, while only X_1 would be selected if $|t_1|$ is large while $|t_2|$ is small. For keen competition for selection, we need the two sides of (5.18) to be nearly equal. If $|t_1|$ is large and $|t_2|$ is small, this occurs when c is close to $+1$ or -1, which is when $|r_{12}|$ is large compared with $|r_{22}|$. It can be shown that c is equal to the sample correlation between X_1 and X_2 (redefined by not subtracting out variable means), so that the condition for only one variable to be selected and for keen competition for selection is that X_1 and X_2 are highly correlated.

Figure 5.4 shows the regions of the (t_1, t_2)-space in which each of the two variables is selected.

Let (τ_1, τ_2) denote the expected values of the LS projections. The

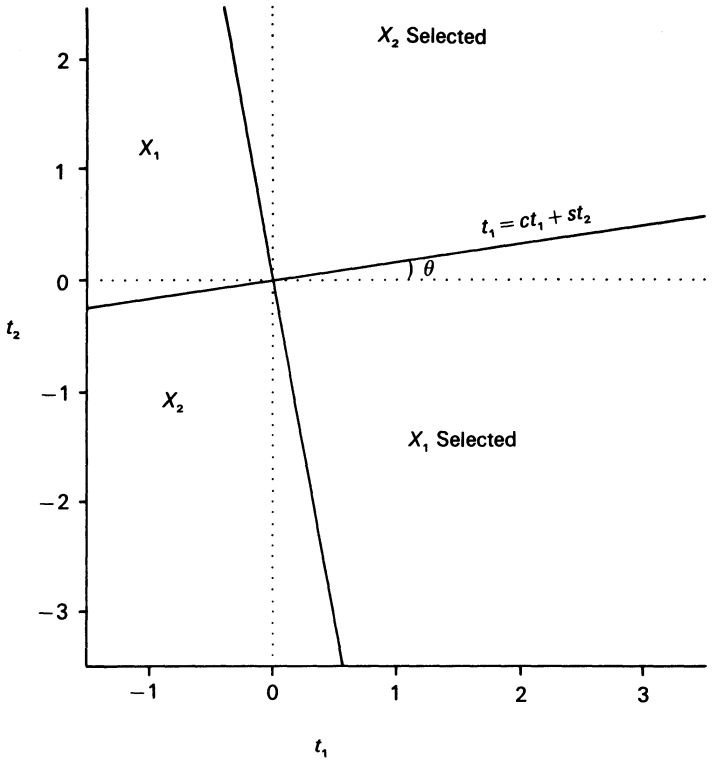


Fig. 5.4 *The regions in the space of the projections* (t_1, t_2), *in which each variable is selected.*

actual sample projections will be from a bivariate normal distribution centred on this point. The probability of selection of variable X_1 can then be found by integrating the normal frequency function over the appropriate region in the (t_1, t_2)-space. The integration is simplified if we rotate the axes to new axes along the boundaries between the selection regions. The boundary lines are

$$t_1 = \frac{s}{1-c} t_2$$

and

$$t_1 = \frac{-s}{1+c} t_2.$$

The angle θ shown in Fig. 5.4 is such that

$$\cos\theta = s/(2-2c)^{1/2}$$
$$\sin\theta = (1-c)/(2-2c)^{1/2}.$$

NB c and $\cos\theta$ are not the same.

Let

$$z_1 = t_1\cos\theta + t_2\sin\theta$$
$$z_2 = t_2\cos\theta - t_1\sin\theta,$$

so that z_1 is measured along the axis closest to t_1 and z_2 is measured along the perpendicular axis. Let (ω_1, ω_2) be the coordinates of the population values of the projections relative to the new axes. The probability of selection of X_1 is then

$$\begin{aligned} P &= \text{prob}(Z_1 < 0 \text{ and } Z_2 > 0) + \text{prob}(Z_1 > 0 \text{ and } Z_2 < 0) \\ &= \Phi(-\omega_1/\sigma)\,[1-\Phi(-\omega_2/\sigma)] + [1-\Phi(-\omega_1/\sigma)]\,\Phi(-\omega_2/\sigma) \end{aligned} \tag{5.19}$$

where Φ is the standard normal distribution function, and σ is the standard deviation of the LS projections. If we set $\sigma = 1$, then the derivatives of P are

$$\frac{\mathrm{d}p}{\mathrm{d}\,\omega_1} = \phi(\omega_1) - 2\phi(\omega_1)\,\Phi(\omega_2)$$

$$\frac{\mathrm{d}p}{\mathrm{d}\,\omega_2} = \phi(\omega_2) - 2\phi(\omega_2)\,\Phi(\omega_1)$$

and then the conditional ML estimates of (ω_1, ω_2) are obtained by solving

$$\hat{\omega}_i = z_i - \frac{1}{P}\frac{\mathrm{d}P}{\mathrm{d}\,\omega_i}, \qquad i = 1, 2. \tag{5.20}$$

Figure 5.5 shows the conditional ML solutions, denoted by squares, related to the LS projections, denoted by crosses, for a set of LS points on a regular grid in one of the regions in which X_1 is selected.

We see that for points more than about two standard deviations away from a boundary, the ML solution is very close to the LS point. Where the LS point is fairly close to only one of the boundaries, the line to the ML solution is roughly perpendicular to the boundary.

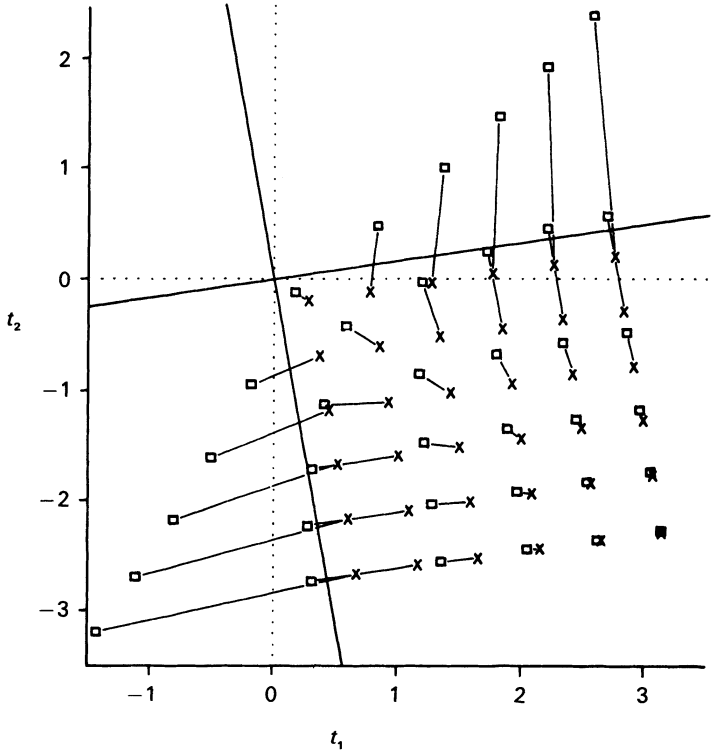


Fig. 5.5 *Showing the ML solutions (denoted by squares) which correspond to certain LS projections (denoted by crosses) for the case in which variable X_1 is selected.*

Notice that for LS points for which t_2 is small, there is very little change in t_1 from the LS point to the ML solution. t_1 changes substantially only when t_2 is large; in such cases in practice, both X_1 and X_2 would be selected.

The regression coefficient for the Y on X_1 regression is t_1/r_{11} for LS, and $\hat{t}_1/r_{11}$ for ML. This means that for most points in the region in which X_1 only would be selected, there is a small reduction in going from the LS to the ML regression coefficient. The difference is not more than about 0·4 × the standard deviation when $|t_1| > |t_2|$.

When ML is the method used for estimation, it is often instructive to examine the shape of the likelihood surface. In the present case,

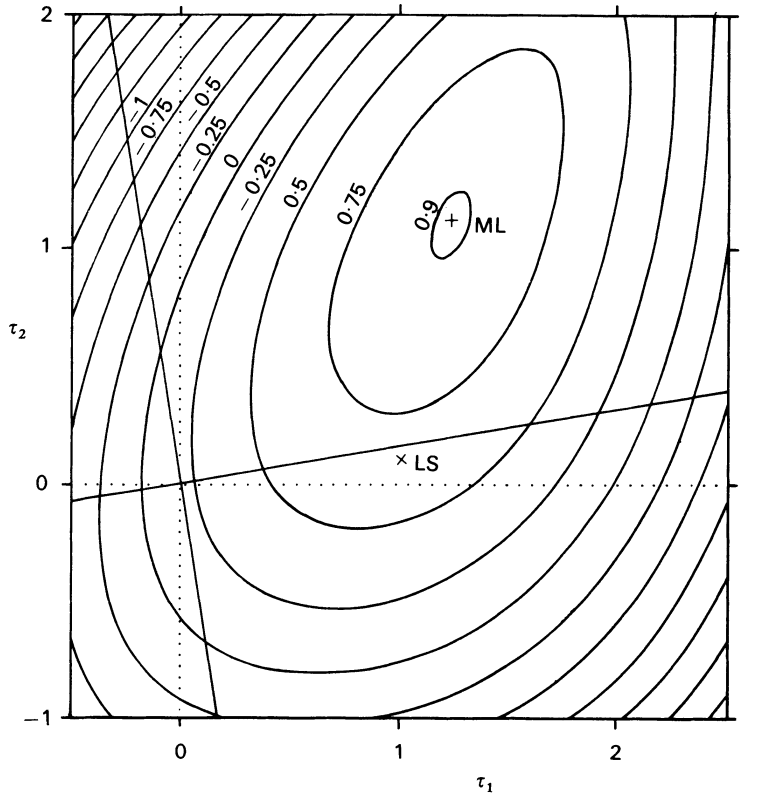


Fig. 5.6 *Showing likelihood contours for the case of two competing variables when the LS projections are close to a decision boundary.*

this is of most interest when the LS projections are close to one of the boundaries. Figure 5.6 shows the conditional likelihood contours for one such point. This point, for which the LS projections are (1·0, 0·1) is much closer to the boundary than any of the points in Fig. 5.5. Except for the innermost contour, the contours are for evenly spaced values of the *LCL*. The values shown alongside the contours are of

$$-\tfrac{1}{2}(t_1-\tau_1)^2-\tfrac{1}{2}(t_2-\tau_2)^2-\log P,$$

that is, the constant term $-\log(2\pi)$ has been omitted. In terms of the rotated projections, the values shown are of

$$-\tfrac{1}{2}(z_1-\omega_1)^2-\tfrac{1}{2}(z_2-\omega_2)^2-\log P.$$

We see that, even in this fairly extreme case, the likelihood surface is smooth and not far from symmetric. In many real-life situations, likelihood surfaces contain singularities, discontinuities, saddle points, or the contours are 'banana-shaped'; in contrast, these surfaces are fairly well behaved for our problem.

We have seen how the ML estimates relate to the LS estimates, but how do they relate to the true values which we are trying to estimate? Still working with our rotated projections, we would like to know how close $E(\hat{\omega}_i)$ is to ω_i.

We have an iterative procedure for finding $\hat{\omega}_i$ given samples LS projections (z_1, z_2), from (5.19). Hence

$$E(\hat{\omega}_i) = \int_{z_1}\int_{z_2} \hat{\omega}_i(z_1, z_2) f(z_1, z_2) \mathrm{d}z_1 \, \mathrm{d}z_2 \Big/ \int_{z_1}\int_{z_2} f(z_1, z_2) z_1 \, \mathrm{d}z_2 \tag{5.21}$$

where $\hat{\omega}_i(z_1, z_2)$ is the solution of (5.19), $f(z_1, z_2)$ is the bivariate normal density

$$f(z_1, z_2) = (2\pi)^{-1} \exp[-\tfrac{1}{2}(z_1 - \omega_1)^2 - \tfrac{1}{2}(z_2 - \omega_2)^2]$$

and the integration is over that region of the (z_1, z_2)-space in which variable X_1 is selected. Note that (ω_1, ω_2) is not necessarily in this region.

The variance of $\hat{\omega}_i$ can be obtained similarly by replacing $\hat{\omega}_i$ in (5.21) with its square.

Before embarking on numerical integration to evaluate (5.21), we should investigate whether the integral is finite. From Fig. 5.5 we see that as the LS point approaches a boundary, the ML point moves very rapidly away from it on the other side. Does it move away too rapidly for the integral to converge?

If the LS point (z_1, z_2) has large, positive z_1, say greater than two standard deviations from zero, while z_2 is small and negative, as for the LS point in Fig. 5.6, then $\hat{\omega}_1$ will be close to z_1 while z_2 will be large and positive. For population projections (ω_1, ω_2), the probability that the sample projections will be such that variable X_1 is selected is approximately

$$P = \text{prob}(Z_2 < 0) + \text{prob}(Z_1 < 0),$$

neglecting the small probability that both Z_1 and Z_2 could both be negative. For large ordinate, the area under the tail of the normal distribution, $[1 - \Phi(x)]$, is approximately equal to $\phi(x)/x$. Hence we

have

$$P \approx \frac{\exp(-\frac{1}{2}\hat{\omega}_1^2)}{\hat{\omega}_1(2\pi)^{1/2}} + \frac{\exp(-\frac{1}{2}\hat{\omega}_2^2)}{\hat{\omega}_2(2\pi)^{1/2}}$$

$$= (1-\alpha)P + \alpha P \text{ say.}$$

Then by differentiating the *LCL* using the approximation for P, we find that the ML equations require the solution of

$$0 = z_1 - \hat{\omega}_1 + (1-\alpha)(\hat{\omega}_1 + 1/\hat{\omega}_1) \tag{5.22}$$

$$0 = z_2 - \hat{\omega}_2 + \alpha(\hat{\omega}_2 + 1/\hat{\omega}_2). \tag{5.23}$$

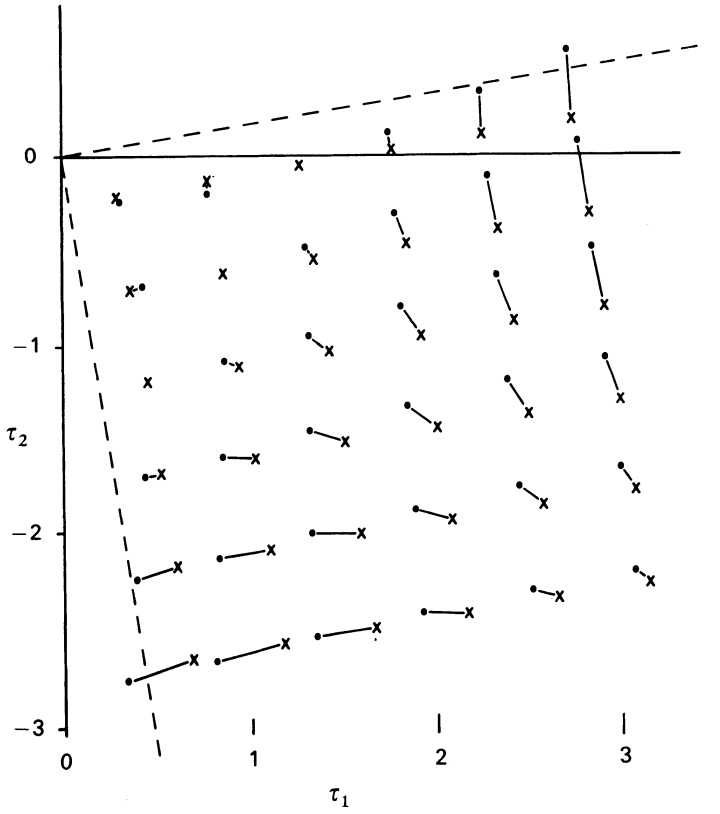


Fig. 5.7 *Showing the locations of the expected ML estimates connected to the true positions for the case of two competing variables. Crosses denote the true values with dots for the expected ML estimates.*

If we forget for the moment that α is a function of $\hat{\omega}_1$ and $\hat{\omega}_2$, then these equations are simple, single-variable, quadratic equations. Taking the appropriate root,

$$\hat{\omega}_1 = \{z_1 + [z_1^2 + 4\alpha(1-\alpha)]^{1/2}\}/(2\alpha)$$

and as $4\alpha(1-\alpha)$ cannot be greater than 1, so that it is dominated by z_1^2, when z_1 is moderately large we have that

$$\hat{\omega}_1 \approx z_1/\alpha.$$

That is, $\hat{\omega}_1$ will be greater than z_1 for sufficiently large z_1. In fact, α will usually be fairly close to 1.

Turning to the other equation (5.23), we find that if we substitute $\alpha = 1$ then we obtain

$$\hat{\omega}_2 \approx 1/z_2,$$

which suggests that $\hat{\omega}_2$ tends to infinity sufficiently rapidly to mean that the integral (5.21) is divergent. Fortunately, numerical evaluation of the solutions of (5.22) and (5.23) shows that α falls away from 1 as z_2 tends to zero, and the values of $\hat{\omega}_1$ and $\hat{\omega}_2$ tend towards being identical. The moments of the ML estimates appear to exist, based upon experience with their numerical evaluation.

The integration of the numerator in (5.21) has been carried out using the routine DQNG from QUADPACK (Piessens *et al.*, 1980).

Figure 5.7 shows the expected values of the ML estimates, denoted by dots, connected to the true values, denoted by crosses. The size of the bias is very small in most cases.

5.5.2 Conditional maximum likelihood for k orthogonal predictors

(a) Largest only selected

Let us suppose that we have k orthogonal predictors, $X_1, X_2, \ldots, X_k$, of which one and only one is to be selected. If the expected values of the LS projections for each of these variables are $\tau_1, \tau_2, \ldots, \tau_k$, then the probability that variable X_1 is the one which is selected is the probability that its sample projection, t_1, is greater in absolute value than the other projections, $t_2, \ldots, t_k$. Assuming the residual standard deviation = 1, then

Table 5.11 *Maximum likelihood (ML) solutions conditional upon the selection of variable X_1 corresponding to given least-squares (LS) projections for 5 and 10 competing orthogonal predictors*

LS projn	*ML soln*	*LS projn*	*ML soln*
2·000	0·446	2·000	0·987
1·600	2·458	1·000	1·501
1·280	1·781	0·500	0·690
1·024	1·342	0·250	0·338
0·819	1·035	0·125	0·168
LS projn	*ML soln*	*LS projn*	*ML soln*
2·000	0·435	2·000	0·727
1·600	2·330	1·000	1·411
1·280	1·697	0·500	0·652
1·024	1·286	0·250	0·320
0·819	0·997	0·125	0·159
0·655	0·782	0·063	0·080
0·524	0·618	0·031	0·040
0·419	0·491	0·016	0·020
0·336	0·391	0·008	0·010
0·268	0·312	0·004	0·005

$$\text{prob.}(X_1 \text{ selected})$$

$$= \int_0^{\infty} \phi(t_1 - \tau_1) \prod_{i=2}^{k} [\Phi(t_1 - \tau_i) - \Phi(-t_1 - \tau_i)]\, dt_1$$

$$+ \int_{-\infty}^{0} \phi(t_1 - \tau_1) \prod_{i=2}^{k} [\Phi(-t_1 - \tau_i) - \Phi(t_1 - \tau_i)]\, dt_1. \qquad (5.24)$$

Using this formula, the *LCL* is easily evaluated and maximized to obtain estimates of τ_1 (and of the other τ_i).

Table 5.11 shows the ML solutions for four cases. In the two cases on the left-hand side, the largest LS projection is not much larger than the second largest. Conditional ML has shrunk the estimate for the first projection by about 1·5 standard deviations. In the two cases on the right-hand side, the LS projections fall away more

rapidly and the ML estimate for the first projection has not been shrunk as much.

If the signs of some of the projections in Table 5.11 are reversed, the signs of the corresponding ML solutions are reversed but their magnitudes remain the same. This was one test which was used to check that the program written to calculate the results shown was correct.

We have no way of knowing from Table 5.11 whether the amount of shrinkage of the first projection is too great or too little, as we do not know the size of the expected value of the first projection. However, if the expected values of all of the projections are equal, then for a normal distribution we can look up tables of order statistics to find out how far we can expect the largest sample projection to be above its expected value. This is found to be 1·16 and 1·54 standard deviations for samples of 5 and 10 respectively. Thus the shrinkages shown in Table 5.11 appear to be of the right order of magnitude for the worst case. (This heuristic argument is crude as it neglects the fact that we are looking at absolute values of projections. If all of the expected values of the projections are say 0·446, using the first case in Table 5.11, the largest sample projection could turn out to be say $-2{\cdot}00$. If we look at the first-order statistic for absolute values of samples from a normal distribution with zero mean, they are found to be at 1·57 and 1·88 standard deviations from zero.)

To see how well, or badly, conditional ML performs at estimating the populations, we need to evaluate the expected value $E(\hat{\tau}_1 | X_1$ selected; $\tau)$, of the ML solution over all projections which lead to the selection of X_1, when the population projections are given by vector $\boldsymbol{\tau}$. This requires multidimensional numerical integration, where the kernel is $\hat{\boldsymbol{\tau}}_1(t)$ multiplied by the probability density of the LS projection vector $\mathbf{t}$. Thus each evaluation of the kernel requires the solution of the conditional ML equation. Such numerical integration is most efficiently performed using Monte Carlo methods.

The cases for which the expected value of the ML solution have been evaluated are

$$k = 5, 10, 20$$
$$\tau_1 = 2{\cdot}0, 1{\cdot}8, 1{\cdot}6, 1{\cdot}4, 1{\cdot}2$$
$$\tau_i = \tau_1 \alpha^{i-1} \qquad \text{for } \alpha = 1{\cdot}0, 0{\cdot}9, 0{\cdot}8, 0{\cdot}5.$$

NB Using these expected values for the projections, the largest LS

projection generated was between 2 and 4 in most cases, which is typically where users of stepwise regression packages choose their cut-off points.

Rather than use only those cases in which the first LS projection was the largest, all cases were used. The ML estimation was then for the variable with the largest sample LS projection, which was often not the first variable. This of course is what usually happens in practice when we are lucky if the chosen variable(s) are also those with the largest expected projections.

Each case was replicated 100 times.

Table 5.12(a) *Means (avge $\hat{\tau}_{[1]}$) and variances (var $\hat{\tau}_{[1]}$) of the maximum likelihood solutions for 5 competing variables when only one is selected. Simulation results for 100 replications. NB Expected values of projections given by $\tau_i = \tau_1 \alpha^{i-1}$; $\tau(t_{max})$ = the average of the τ_i's corresponding to the largest generated sample projection t_i*

k	τ_1	α	$\tau(t_{max})$	*Avge* $\hat{\tau}_{[1]}$	*Var.* $\hat{\tau}_{[1]}$
5	2·00	1·0	2·00	1·50	1·68
		0·9	1·67	1·16	0·91
		0·8	1·52	1·18	0·85
		0·5	1·44	1·09	1·29
5	1·80	1·0	1·80	1·29	1·24
		0·9	1·50	1·09	0·87
		0·8	1·29	1·20	1·05
		0·5	1·11	1·10	2·53
5	1·60	1·0	1·60	1·22	1·01
		0·9	1·31	0·97	0·62
		0·8	1·12	1·12	0·76
		0·5	0·98	0·80	1·23
5	1·40	1·0	1·40	1·12	0·76
		0·9	1·16	1·09	0·84
		0·8	0·99	0·84	0·68
		0·5	0·78	0·69	0·85
5	1·20	1·0	1·20	0·84	0·50
		0·9	0·99	0·99	0·92
		0·8	0·82	0·87	0·78
		0·5	0·72	0·77	1·17

Table 5.12(b) *As Table 5.12(a) but for 10 competing variables*

k	τ_1	α	$\tau(t_{max})$	*Avge* $\hat{\tau}_{[1]}$	*Var.* $\hat{\tau}_{[1]}$
10	2·00	1·0	2·00	0·98	1·16
		0·9	1·49	0·96	0·83
		0·8	1·33	0·89	0·76
		0·5	1·15	0·96	2·92
10	1·80	1·0	1·80	0·83	0·54
		0·9	1·35	0·94	0·67
		0·8	1·12	0·89	1·09
		0·5	0·95	0·97	2·66
10	1·60	1·0	1·60	0·88	0·69
		0·9	1·16	0·84	0·54
		0·8	0·95	0·73	1·34
		0·5	0·77	0·77	1·79
10	1·40	1·0	1·40	0·95	0·81
		0·9	1·00	0·78	0·60
		0·8	0·77	0·78	1·28
		0·5	0·57	0·93	3·14
10	1·20	1·0	1·20	0·81	0·57
		0·9	0·88	0·73	0·50
		0·8	0·68	0·65	0·73
		0·5	0·49	0·78	1·96

Tables 5.12(a)–(c) show the average values of the ML solutions, denoted by $\tau(t_{max})$, the average value of the corresponding τ_i, denoted by avge $\hat{\tau}_{[1]}$, and the variance of the estimates.

We note that when $\alpha = 1{\cdot}0$ or $0{\cdot}9$, the ML solutions are too small (too much shrinkage from the LS estimate), by between $0{\cdot}5$ and 1 standard deviation. For smaller values of α, when fewer of the LS projections are close together, the ML solutions show less bias, and in the case $\alpha = 0{\cdot}5$ there is a bias in the same direction as the LS estimates, i.e. ML has not applied as much correction as necessary. Overall though, the bias in the ML estimates is much smaller than that of the LS estimates.

Notice though that some of the variances of the ML estimates are very large. LS estimates when there is no selection, have variance $= 1{\cdot}0$. We see that for the smaller values of α, there variances are often

Table 5.12(c) *As Table 5.12(a) but for 20 competing variables*

k	τ_1	α	$\tau(t_{max})$	*Avge* $\hat{\tau}_{[1]}$	*Var.* $\hat{\tau}_{[1]}$
20	2·00	1·0	2·00	0·99	1·29
		0·9	1·34	0·83	0·90
		0·8	1·09	0·80	1·68
		0·5	0·89	0·94	3·31
20	1·80	1·0	1·80	0·87	1·04
		0·9	1·23	0·71	0·69
		0·8	0·96	0·88	1·99
		0·5	0·83	1·16	4·48
20	1·60	1·0	1·60	0·77	0·68
		0·9	0·99	0·67	0·47
		0·8	0·79	0·77	1·76
		0·5	0·69	1·09	4·34
20	1·40	1·0	1·40	0·89	0·96
		0·9	0·81	0·67	0·70
		0·8	0·65	0·79	2·05
		0·5	0·45	0·83	3·13
20	1·20	1·0	1·20	0·72	0·65
		0·9	0·75	0·70	1·26
		0·8	0·55	0·81	2·00
		0·5	0·44	1·03	3·27

much greater than 1. The large contributions have come from a small number of cases in which the sample projection has been moderately large and negative when its expected value, τ, has been of the order of say $+0{\cdot}2$.

(b) Selection of more than one variable with a cut-off

The case in which one and only one variable out of k is selected, was investigated to discover something about the behaviour of the max-*LCL* estimator, though it is a case which does sometimes occur in real life. A more common situation though is that in which all variables which satisfy some cut-off rule are selected, so let us look at that as well.

In this case, still assuming orthogonal predictors, the probability of selecting a particular subset is simply the product of the

independent probabilities of selecting each individual variable in the subset. Hence the logarithm of the *a priori* probability of selection is equal to the sum of the individual logarithms of probabilities of selection. That is, we can just look at the problem one variable at a time.

Let τ_j be the expected value of the LS projection for variable X_j, and let t_j be the sample value. Suppose we select variable X_j if $|t_j| > C$ where C is the cut-off value. Assume that the residual standard deviation, $\sigma = 1$, so that if we choose say $C = 2$ then it is equivalent to selecting the variable if its 't-value' in the common terminology, i.e. the regression coefficient divided by its standard error, is greater than 2 or the F-to-enter is greater than $4(= C^2)$. For simplicity of presentation, this is blurring over the distinction between using the population standard deviation and its sample estimate.

The contribution of variable X_j to the *LCL* is

$$-\tfrac{1}{2}(\tau_j - t_j)^2 - \log_e P_j. \tag{5.25}$$

The probability of selection of X_j is then the probability that either $t_j > C$ or that $t_j < -C$, and hence

$$P_j = 1 - \Phi(C - \tau_j) + \Phi(-C - \tau_j)$$

Substituting in (5.25) and differentiating with respect to τ_j we obtain the following equation to solve for the max-*LCL* estimate of τ_j when X_j is selected:

$$0 = -(\tau_j - t_j) - [\phi(C - \tau_j) - \phi(-C - \tau_j)]/P_j,$$

where Φ is the standard normal distribution function, and ϕ is its first derivative, the normal probability density. This equation can be easily be solved numerically; for instance, Newton-Raphson converges rapidly. If $\hat{\tau}(t_j)$ is the max-*LCL* estimate of τ_j corresponding to an LS projection, t_j, then it is straightforward to use numerical integration to calculate the expected value of $\hat{\tau}(t_j)$ and its variance as functions of the cut-off value C.

Figure 5.8 shows the expected value of the max-*LCL* estimate compared with the expected value of the LS estimate when a cut-off of $C = 2{\cdot}0$ is used. It shows that the bias in the max-*LCL* estimator is very much smaller except for large values of τ when the bias is small for both estimators.

Figure 5.9 compares the variances of the two estimators for a

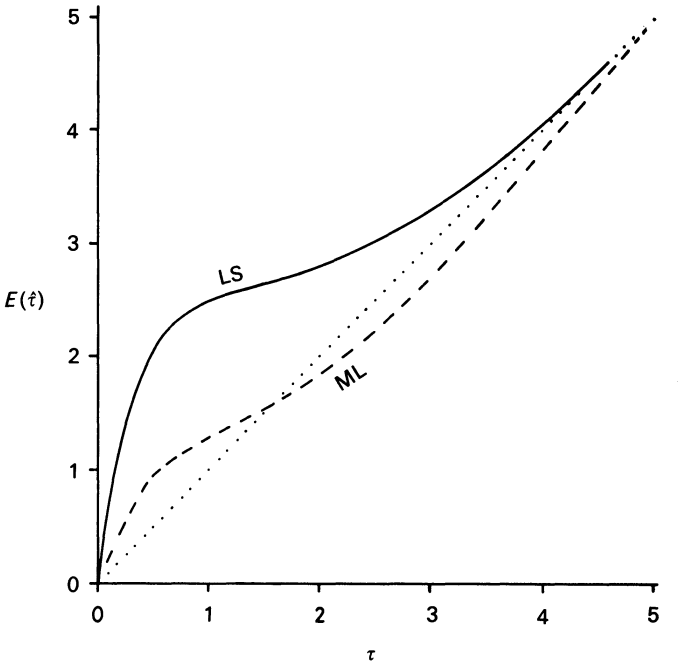


Fig. 5.8 *Expected value of the max-LCL estimate (ML), against* τ, *compared with the expected value of the LS-estimate (LS), for cut-off* $C = 2{\cdot}0$ *(roughly equivalent to an F-to-enter of* 4·0*). The dotted line is* $\hat{t} = \tau$.

cut-off of $C = 2{\cdot}0$. We shall see in the next chapter that in making predictions, the second moments of the estimators are very important. For most values of the expected projection, τ_j, both variances are greater than 1·0, which is the variance when there is no selection, The max-*LCL* estimator has much smaller variance except for large values of τ.

The high variance of the max-*LCL* estimator is due to the substantial shrinkage which it applies when the sample projection, t_j, is close to the cut-off point. For cut-offs between 1 and 3, when t_j is just on the positive cut-off boundary, it is shrunk to a value between 0·445 and 0·485. This is a substantial shrinkage and hence a large variance results. A small quantity added to the probability of selection would greatly reduce the variance and only have a small influence on the bias.

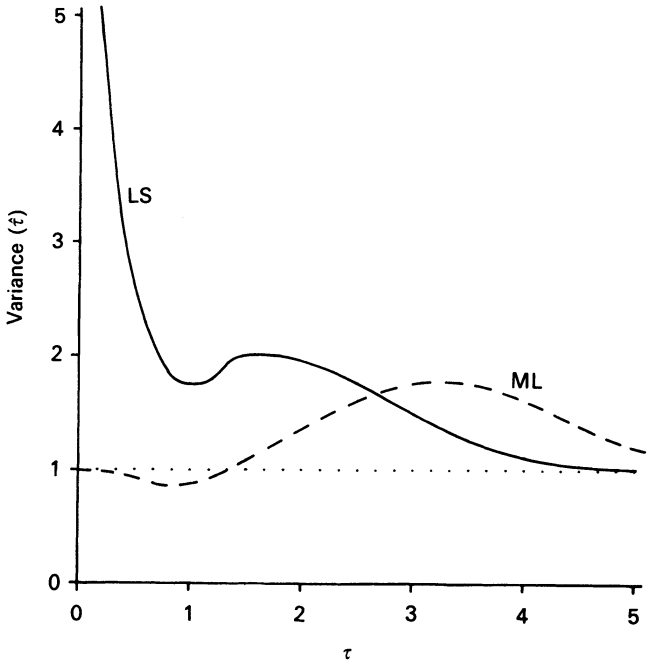


Fig. 5.9 *Variances of the max-LCL(ML) and LS estimators against* τ *for a cut-off of* $C = 2{\cdot}0$.

5.6 Estimation – summary and further work

We have seen that when LS is used to estimate regression coefficients after subset selection, there can be biases of the order of one to two standard errors in the estimates. A number of alternative estimation methods have been examined. Two of these merit further research – the jack-knife estimator (5.9), and the ML estimator (5.11).

Maximizing the logarithm of the conditional likelihood (*LCL*) given by (5.11) for both real and artificial data sets has produced some cases in which it has performed very well in eliminating bias, and others in which it has not done as well. Unfortunately, the maximum *LCL* estimates sometimes have an unacceptably large variance. The cases in which it performs poorly are those in which the estimated probability of selection has become extremely small during the iterations.

It appears that a substantial improvement can be achieved by replacing $\log_e P(\delta)$ in (5.11) with $\log_e [P(\delta) + \varepsilon]$ where ε is small and positive. The quantity should probably be a function of the sample size, the number of predictors, etc. The addition of such a quantity will mean that the *LCL* is approximately quadratic instead of linear in the projections away from the max-*LCL* point.

An argument which can be used to derive an addition to $\log_e P$ is as follows. As we do not know P, we must estimate it from simulations. Though these simulations give unbiased estimates of P, if we take logarithms, we have biased estimates of $\log_e P$. A Taylor series approximation can be used to correct for this bias. Let us suppose that P is estimated simply as

$$\hat{P} = r/n$$

where r is the number of cases in which the subset of interest was selected out of n sets of simulated data. For simplicity of presentation, weighting of the cases has been ignored. Then

$$\begin{aligned} E\{\log_e \hat{P}\} &= E\{\log_e [P + (\hat{P} - P)]\} \\ &= \log_e P + E\left\{\frac{\hat{P} - P}{P} - \frac{1}{2}\left(\frac{\hat{P} - P}{P}\right)^2 + \cdots\right\} \\ &\approx \log_e P - \frac{1 - P}{2nP} \end{aligned}$$

as the variance of $\hat{P}$ from simple binomial sampling is $P(1 - P)/n$. This suggests that we replace $\log_e \hat{P}$ with

$$\log_e \hat{P} + \frac{1 - \hat{P}}{2n\hat{P}}.$$

Using a Taylor series expansion again, we see that this is approximately $\log_e (\hat{P} + \varepsilon)$ where $\varepsilon = (1 - \hat{P})/(2n)$, or, as $\hat{P}$ will usually be very small, $\varepsilon \approx 1/(2n)$. It is then a straightforward exercise to modify this correction term for the weighting derived in Appendix 5A.

These two Taylor series expansions are extremely crude. For the expansion of $\log_e (1 + x)$ as far as the quadratic to be sufficiently accurate for practical purposes, x must be much less than 1·0 in absolute value. In this case, P will often be very small, say of the order of 10^{-4}, while the possible value of $\hat{P}$ are $0, 1/n, 2/n, \ldots$, etc. Thus some of the values of $(\hat{P} - P)$ will be much larger than P in many cases.

The use of Monte Carlo methods to estimate part or all of a likelihood function is becoming more common. A good discussion of methods and problems in this area is contained in Diggle and Gratton (1984). We are fortunate in this case that only a small part of the likelihood function needs to be estimated using simulation.

Appendix 5A Conditional maximum likelihood algorithm

We want to maximize the quantity

$$LCL = \text{constant} - (RSS_k + \sum \delta_j^2)/(2\sigma^2) - \log_e P(\delta) \qquad (5.11)$$

with respect to the changes, δ_j, from the LS projections. The problem is to evaluate the probability of selection, $P(\boldsymbol{\delta})$, and if possible, its first derivatives, as then we have from (5.12) that

$$\delta_j = -\sigma^2 \frac{(\mathrm{d}P/\mathrm{d}\delta_j)}{P}. \qquad (5.13)$$

This can be solved iteratively by summing starting values for the δ_j's (say zeros), calculating the derivatives, obtaining new estimates for the δ_j's, etc., until convergence. Experience has shown that this process has converged in all cases investigated.

Let t_j be the jth LS projection, and let τ_j be its expected value. That is, τ_j is its expected value for all samples of the same size with the same values for the X-predictors, irrespective of whether the sample leads to the selection of our subset of interest. In our earlier notation then we have

$$\mathbf{t} = \boldsymbol{Q}'\mathbf{Y},$$
$$\boldsymbol{\tau} = \boldsymbol{R}\boldsymbol{\beta},$$

and

$$\boldsymbol{\delta} = \boldsymbol{\tau} - \mathbf{t}$$

The probability P that our subset is selected is then

$$P = \text{prob}\left(\sum_{j=1}^{p} t_j^2 > \sum_{j=1}^{p} [t_j^{(s)}]^2, \qquad s = 1, \ldots, N_s \right)$$

where N_s is the number of other subsets compared with our subset using whatever selection procedure was used, and $t_j^{(s)}$ is the jth projection for the sth alternative subset where the projections for that subset have been reordered so that the first p are for the selected

variables. The elements of $t_j^{(s)}$ are linear combinations of the t_j's, i.e.

$$t_j^{(s)} = \sum_{i=1}^{k} c_{ij} t_i$$

where the coefficients c_{ij} can be obtained by multiplying together the planar rotations used to go from one ordering of the projections to the other.

For any trial values for the unknown population projections, $\tau_j, j = 1, \ldots, k$, we could estimate P by generating sets of projections $t_j = \tau_j + \varepsilon_j$ with the ε_j's sampled from a normal distribution with zero mean and variance σ^2. We could generate 100 or 1000 such sets and run our selection procedure on each. The proportion of times on which our subset is selected can then be used as an estimate of P. As P will usually be very small in the region of the ML solution, this requires very large numbers of cases to estimate $\log_e P$ and its first derivatives with reasonable accuracy.

A way to reduce the amount of effort needed to estimate small proportions is to use biased samples, and then to compensate for that bias. This method has been called 'importance sampling' (Kahn, 1956), though it is probably also known by other names.

Let $f_0(\mathbf{t})$ be the density function desired for the projections, $\mathbf{t}$, and let $f(\mathbf{t})$ be the density function from which we sample. The outcome from each sample is then given weight equal to $f_0(\mathbf{t})/f(\mathbf{t})$. Suppose we estimate some statistic, $T(\mathbf{t})$, from the sampled projections. If we had sampled from the true distribution $f_0(\mathbf{t})$, then the expected value of the statistic is

$$E[T(\mathbf{t})] = \int T(\mathbf{t}) f_0(\mathbf{t}) \, \mathrm{d}t.$$

Sampling from distribution $f(t)$ and giving weights as above, the expected value for the weighted statistic is

$$\begin{aligned} E[T(\mathbf{t})] &= \int w T(\mathbf{t}) f(\mathbf{t}) \, \mathrm{d}t \\ &= \int T(\mathbf{t}) f_0(\mathbf{t}) \, \mathrm{d}t. \end{aligned}$$

That is, the estimates are unbiased.

If the LS projections have a multivariate normal distribution with mean, τ_0 and variance σ^2, and we sample from a distribution with

some other mean τ but the same variance, then the two density functions are

$$f_0(\mathbf{t}) = (2\pi\sigma^2)^{-k/2} \exp\left\{ -\sum_{i=1}^{k} (t_i - \tau_{i0})^2/(2\sigma^2) \right\}$$

and

$$f(\mathbf{t}) = (2\pi\sigma^2)^{-k/2} \exp\left\{ -\sum_{i=1}^{k} (t_i - \tau_i)^2/(2\sigma^2) \right\}$$

where t_i, τ_{i0}, τ_i are the ith elements in the vectors $\mathbf{t}, \boldsymbol{\tau}_0$ and $\boldsymbol{\tau}$ respectively. The weight given to a LS projection, $\mathbf{t}$, is then

$$\begin{aligned} w(\mathbf{t}) &= \exp\left\{ \left[-\sum_{i=1}^{k} (t_i - \tau_{i0})^2 + \sum_{i=1}^{k} (t_i - \tau_i)^2 \right] \Big/ (2\sigma^2) \right\} \\ &= \exp\left\{ (2\sigma^2)^{-1} \sum_{i=1}^{k} \left(t_i - \frac{\tau_i + \tau_{i0}}{2} \right)(\tau_{i0} - \tau_i) \right\}. \end{aligned} \tag{5A.1}$$

The statistic we want to estimate is the probability of selection of our chosen subset. Let $S_j = 1$ if the jth set of LS projections leads to the selection of our subset of interest, $= 0$ otherwise. Then our estimate of P is

$$\hat{P} = \sum w_j S_j / \sum w_j. \tag{5A.2}$$

An advantage of this method is that the same sample can be reused as we improve our estimate of the ML solution vector, $\boldsymbol{\tau}_0$, just changing the weights. Furthermore, estimates of the derivatives of $\log_e P$ can be obtained simply by differentiating the expression for $\log_e \hat{P}$, as follows:

$$\frac{\mathrm{d} \log_e \hat{P}}{\mathrm{d}\boldsymbol{\tau}_0} = \frac{\mathrm{d} \log(\sum w_j S_j)}{\mathrm{d}\boldsymbol{\tau}_0} - \frac{\mathrm{d} \log(\sum w_j)}{\mathrm{d}\boldsymbol{\tau}_0}$$

The differentials of the weights are given by

$$\frac{\mathrm{d}w}{\mathrm{d}\boldsymbol{\tau}_0} = w(\mathbf{t} - \boldsymbol{\tau}_0)/\sigma^2$$

so that

$$\frac{\mathrm{d} \log_e \hat{P}}{\mathrm{d}\boldsymbol{\tau}_0} = (1/\sigma^2) \left\{ \frac{\sum w_j S_j (\mathbf{t}_j - \boldsymbol{\tau}_0)}{\sum w_j S_j} - \frac{\sum w_j (\mathbf{t}_j - \boldsymbol{\tau}_0)}{\sum w_j} \right\}.$$

Hence, by substituting in (5.13), we derive

$$\delta = \frac{\sum w_j(\mathbf{t}_j - \tau_0)}{\sum w_j} - \frac{\sum w_j S_j(\mathbf{t}_j - \tau_0)}{\sum w_j S_j}$$
$$= (\sum w_j \mathbf{t}_j / \sum w_j) - (\sum w_j S_j \mathbf{t}_j / \sum w_j S_j). \qquad (5A.3)$$

The first term on the right-hand side of (5A.3) is the weighted average of the LS projections over all the generated sets, while the second term is averaged over only those subsets which lead to the selection of our subset of interest. Thus the resultant method is one which we might have chosen intuitively.

Appendix 5B An application of the ML algorithm

As an illustration, let us take the POLLUTE data set and consider the estimation of the regression coefficients for the best-fitting subset of four predictors (out of the 15). These four are number 1 (average annual rainfall), 2 (January temperature), 9 (% nonwhite population) and 14 (SO_2 concentration); the dependent variable is age-adjusted mortality rate in deaths per 100 000 population per annum. Referring to Table 3.17, we see that other subsets which fit well as (2, 6, 9, 14) and (2, 5, 6, 9), both of which contain variable number 6 (median years of education).

Table 5B.1 shows the experimental design for the Monte Carlo experiment. Apart from the first point, the design is centred around

Table 5B.1 *Design of Monte Carlo experiment for ML estimation*

Number of design points	*The centre points*	
1	$c_i = t_i + s_i\sigma/\sqrt{p}$	$i = 1,\ldots,k$
p	$c_i = t_i + \sigma$	$i = 1,\ldots,p$
	$c_j = t_j - \sigma/p$	$j \neq i$
p	$c_i = t_i - \sigma$	$i = 1,\ldots,p$
	$c_j = t_j + \sigma/p$	$j \neq i$
1	$c_i = t_i - \sigma/p$	$i = 1,\ldots,k$
1	$c_i = t_i + \sigma/p$	$i = 1,\ldots,k$

Note: s_i is the sign of t_i if $i \leqslant p$, and the opposite otherwise.

the LS projections, t_i. The first point was added to make sure that the subset selected using the real data was also selected for a moderate number of the artificial data sets. In the two groups of p sets of points, each projection for one of the selected variables is either increased or decreased by σ in an attempt to estimate the effect of changing that projection.

For each design point, a set of 15 projections, $c_i + \hat{\sigma} \cdot \varepsilon_i, i = 1, \ldots, 15$, was generated, where the ε_i's were sampled from the standard normal distribution, and the standard deviation, $\hat{\sigma} = 34{\cdot}9$, was the usual residual standard deviation estimate using all 15 predictors. For each case, an exhaustive search was carried out to find whether the subset (1, 2, 9, 14) was the best-fitting subset of four predictors. To reduce the amount of computation, forward selection was used first. If the subset of four predictors gave a smaller residual sum of squares than that for subset (1, 2, 9, 14), then it was unnecessary to carry out the exhaustive search.

Each design point in Table 5B.1 was replicated 100 times, thus the LS projections were generated for 1100 sets of artificial data, each consisting of 60 observations with the same values for the predictor variables as in the real data set. To achieve some variance reduction, the replications were carried out in pairs with the signs of each ε_i reversed for the second replicate. Thus the generated projections had averages which were exactly equal to the design point values. This entire experiment was carried out twice.

Table 5B.2 shows the number of times, out of 100, that the subset (1, 2, 9, 14) was selected at each design point. We see from design points 3 and 5 that adding to the projections for variables 2 and 14, and hence reducing their magnitude as they are both negative, has substantially reduced the probability of selection. Similarly, from design points 6 and 8 we see that substracting from the positive projections for variables 1 and 9 has reduced the probability of selection.

Using the original LS projections as the first estimates of the ML solutions, the probabilities of selection, P, and the log-likelihood, L, relative to those for the usual normal regression model with $P = 1$, were calculated for consecutive iterations and are shown in Table 5B.3. As the log-likelihood for the usual normal regresion model is constant $- RSS_k$, the quantity used for L is

$$L = -\sum \delta_j^2/(2\sigma^2) - \log_e P(\boldsymbol{\delta}).$$

Table 5B.2 *Frequency, out of* 100, *of selection of subset* (1, 2, 9, 14) *at each design point*

Design point	*Experiment 1*	*Experiment 2*
1	53	51
2	52	49
3	24	20
4	45	29
5	21	29
6	24	27
7	49	46
8	33	26
9	49	45
10	35	42
11	36	30

Table 5B.3 *Probabilities of selection and log-likelihood for iterations of the ML estimation method*

	Experiment 1		*Experiment 2*	
Iteration	*Sel. prob. P*	*Log-l' hood L*	*Sel. prob. P*	*Log-l' hood L*
0	0·166	1·80	0·251	1·38
1	0·0101	3·53	0·0651	2·23
2	1·87E-4	5·41	0·0124	2·86
3	1·15E-5	5·74	1·73E-3	3·41
4	9·24E-6	6·00	3·29E-5	5·22
5	1·12E-5	6·05	5·51E-7	5·73
6	9·09E-6	6·07	7·91E-7	5·84
7	1·06E-5	6·08	8·92E-7	5·84

The convergence criterion used for stopping was that the change in log-likelihood had to be less than 0·01. In this case, in which only a small number of the other variables are 'competing' for selection (particularly variable 6), the estimated probability of selection has fallen to around 1 part in a million. In other cases, it has sometimes been of the order of 1.E-20 or smaller at the ML solution.

Table 5B.4 shows the final estimated projections, while Table 5B.5 shows the regression coefficients calculated from them. Notice that

Table 5B.4 *ML and LS projections*

Variable no.	*LS projection*	*ML projections*	
		Expt 1	*Expt 2*
1	243·4	221·0	225·6
2	−37·0	13·8	6·8
9	274·8	270·7	274·6
14	−151·6	−123·3	−104·3
3	−59·1	−94·9	−32·3
4	15·0	27·4	26·5
5	−11·2	2·1	43·8
6	−89·0	−101·6	−174·0
7	−16·9	−11·1	−15·0
8	−30·2	−33·5	−40·8
10	−14·1	−36·3	21·8
11	2·1	12·3	−22·1
12	11·3	53·0	53·0
13	−46·5	−96·3	−54·3
15	3·2	54·1	−16·1

Table 5B.5 *LS and ML regression coefficients for the POLLUTE data set*

Variable no.	*LS regn coeff.*	*Experiment 1*		*Experiment 2*	
		ML regn coeff.	*Std. error*	*ML regn coeff.*	*Std. error*
1	2·059	1·637	0·507	1·619	0·487
2	−1·772	−1·190	0·387	−1·372	0·417
9	4·079	4·167	0·495	4·349	0·437
14	0·3305	0·2687	0·0585	0·2273	0·0610

the projections for selected variables 1, 2 and 14 have been reduced in magnitude by amounts of the order of 50–100% of $\hat{\sigma}$, while the projection for variable number 9 is almost unchanged. The projections for most of the unselected variables have been increased, particularly that for variable 6. When we look at the regression coefficients we see that for the dominant variable 9 it has increased very slightly while the other three have decreased by one or more standard errors.

CHAPTER 6

How many variables?

6.1 Introduction

The stopping rule to apply in any situation depends upon

1. the objectives, and
2. the estimation method.

For instance, if the objective is to minimize prediction errors in some sense, e.g. by minimizing the mean squared error of prediction, then a larger subset will often be appropriate than if there is to be a trade off between the future cost of measuring predictors and the accuracy achieved.

The choice of subset, as well as the size of subset, will also depend upon the region of the X-space in which we wish to predict. A fairly common situation is one in which one variable, say X_1, is expected to vary or be varied much more in the future than in the past. If the range of values of X_1 is small in the calibration sample, the influence of this variable may be so small that any of the automatic procedures described in Chapter 3 will not select that variable. In such circumstances, it may be necessary to force X_1 to be selected, provided that a regression coefficient can be estimated for it which has an estimated standard error which is smaller than the coefficient itself.

Typically we would like to minimize with respect to the size of subset, p, a sum of the kind

$$\|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}_p\hat{\boldsymbol{\beta}}_p\|_2^2 = \sum_{i=1}^{N}\left(\sum_{j=1}^{k} x_{ij}\beta_j - \sum_{j=1}^{k} I_j x_{ij}\hat{\beta}_j\right)^2 + \sum_{j=1}^{k} C_j I_j \quad (6.1)$$

where $I_j = 1$ if the jth variable is in the subset, and $=0$ otherwise, C_j is the cost of measuring the jth variable, where the matrix $\boldsymbol{X} = \{x_{ij}\}$ of values of the predictor variables is specified, as is the method

of obtaining the estimated regression coefficients, $\hat{\beta}_j$, and N is the number of future values to be predicted.

As the future values of x_{ij} for which predictions will be required are often not known, the matrix X is often taken to be identical with the observation matrix or to be the multivariate normal distribution with covariance matrix equal to the sample covariance matrix for the calibration data. This sometimes leads to some simple analytic results, provided that unbiased estimates are available for the β_j's.

There are almost no examples in the literature of the use of any other assumed future values for the predictor variables, other than Galpin and Hawkins (1982, 1986). They demonstrate that different subsets should be chosen for different ranges of future X- values, though their derivation neglects the bias due to selection.

For one future prediction for which $\mathbf{x}$ is the vector of values of the k predictors, the variance of the predictor $x'\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the vector of least-squares (LS) regression coefficients, is

$$\sigma^2 \mathbf{x}'(X'X)^{-1}\mathbf{x}$$

where σ^2 is the residual variance with all the predictors in the model. If we take the gamble of ignoring the biases due to selection, and assume the same residual variance for selected subsets, then we could minimize this quadratic form substituting the appropriate subsets of the matrix X and vector $\mathbf{x}$, to find a suitable subset for prediction at the point $\mathbf{x}$. An algorithm for minimizing quadratic forms over subsets of variables has been given by Ridout (1988); this could be used for this purpose. In a practical situation, this could be used to indicate a possible subset to use for future prediction, despite the biases.

The representation (6.1) can also be used when the objective is to estimate the regression coefficients. This can be treated as equivalent to estimating $X\boldsymbol{\beta}$ when each row of X contains zeros in all except one position. Other X-matrices can similarly be constructed if the purpose is to estimate contrasts. For the simplest contrasts, each row of X contains one element equal to $+1$, one element equal to -1, and the remainder equal to zero.

Unfortunately, almost all of the available theory assumes that we have unbiased LS estimates (i.e. no selection bias) of the regression coefficients. This would apply if separate, independent data are used for model selection and for estimation. For instance, Bendel and Afifi (1977) appear at first glance to have solved the common problem

of the choice of stopping rule using LS regression coefficients for the case of one data set for both selection and estimation until one notices the requirement that 'the subset is selected without reference to the regression sample'. In the derivations of mean squared error of prediction (*MSEP*) which follow later in this chapter, we shall see that in many practical cases, the minimum *MSEP* is obtained by using all of the available predictors, i.e. with no selection, if no correction is made for competition bias in selection.

The basic criterion we will use will be that of minimizing the *MSEP*, but it will be shown that in practice this often yields the same results as using either a false F-to-enter criterion, or a likelihood criterion such as the Akaike information criterion.

6.2 Mean squared errors of prediction (*MSEP*)

We will consider two models for the predictor variables which Thompson (1978) described as the fixed and random models.

(a) Fixed model
The values of the X-variables are fixed or controllable, as for instance when the data are from a controlled experiment.

(b) Random model
The X-variables are random variables. This type of model is relevant to observational data when the X-variables cannot be controlled.

In many practical cases using observational data, there will be a mixture of fixed and random variables. For instance, in studies of ozone concentration in the atmosphere, some of the predictors could be season, day of the week, time of day, location, all of which will be known, i.e. fixed variables, while other predictors, such as meteorological variables and concentrations of pollutants such as nitric oxides and hydrocarbons, will be random variables.

6.2.1 MSEP for the fixed model

Let $\boldsymbol{X}_A$ denote the $n \times p$ matrix consisting of the p columns of $\boldsymbol{X}$ for the p selected variables. If the prediction equation is to contain a constant or intercept term, one of these columns will be a column of 1's. For convenience it will be assumed that the columns of $\boldsymbol{X}$

have been ordered so that the first p are those for the selected variables. Let X be partitioned as

$$X = (X_A, X_B)$$

where X_B is an $n \times (k-p)$ matrix, and let $\mathbf{b}_A$ be the vector of LS regression coefficients, where

$$\mathbf{b}_A = (X_A'X_A)^{-1}X_A'\mathbf{Y} \tag{6.2}$$

Let $\mathbf{x}$ be one vector of values of the predictors for which we want to predict Y, and let $\hat{y}(\mathbf{x})$ denote the predicted value using the LS regression coefficients, i.e.

$$\hat{y}(\mathbf{x}) = \mathbf{x}_A'\mathbf{b}_A.$$

If the true relationship between Y and the X-variables is

$$Y = X\boldsymbol{\beta} + \varepsilon$$

where the residuals, ε, have zero mean and variance σ^2, then the prediction error is

$$\hat{y}(\mathbf{x}) - \mathbf{x}'\boldsymbol{\beta} = \mathbf{x}_A'\mathbf{b}_A - \mathbf{x}'\boldsymbol{\beta} \tag{6.3}$$

where

$$\mathbf{x}' = (\mathbf{x}_A', \mathbf{x}_B'),$$

i.e. it includes the values of the other $(k-p)$ predictors which were not selected. The prediction error given by (6.3) can be regarded as having the following components:

1. a sampling error in $\mathbf{b}_A$;
2. a bias in $\mathbf{b}_A$, the selection bias, if the same data were used both to select the model and to estimate the regression coefficients; and
3. a bias due to the omission of the other $(k-p)$ predictors.

If there is no selection bias, which would be the case if independent data had been used to select the model, then

$$\begin{aligned} E(\mathbf{b}_A) &= \gamma_A, \quad \text{say} \\ &= (X_A'X_A)^{-1}X_A'X\boldsymbol{\beta} \\ &= (X_A'X_A)^{-1}X_A'(X_A, X_B)\begin{pmatrix}\boldsymbol{\beta}_A \\ \boldsymbol{\beta}_B\end{pmatrix} \\ &= \boldsymbol{\beta}_A + (X_A'X_A)^{-1}X_A'X_B\boldsymbol{\beta}_B \end{aligned} \tag{6.4}$$

that is, $\boldsymbol{\beta}_A$ is augmented by the regression of $X_B\boldsymbol{\beta}_B$ on the p selected variables.

Now let us rewrite the prediction error (6.3) using (6.4) above as:

$$\begin{aligned}\hat{y}(\mathbf{x}) - \mathbf{x}'\boldsymbol{\beta} &= \mathbf{x}_A'\{[\mathbf{b}_A - E(\mathbf{b}_A)] + [E(\mathbf{b}_A) - \boldsymbol{\gamma}_A] + [\boldsymbol{\gamma}_A - \boldsymbol{\beta}_A]\} - \mathbf{x}_B'\boldsymbol{\beta}_B \\ &= \mathbf{x}_A'\{(1) + (2) + (X_A'X_A)^{-1}X_A'X_B\boldsymbol{\beta}_B\} - \mathbf{x}'\boldsymbol{\beta}_B \\ &= \mathbf{x}_A'\{(1) + (2)\} + [\mathbf{x}_A'(X_A'X_A)^{-1}X_A'X_B - \mathbf{x}_B']\boldsymbol{\beta}_B \\ &= \mathbf{x}_A'\{(1) + (2)\} + (3)\end{aligned}$$

where the messy expression for (3) is for the projection of the Y-variable on that part of the X_B predictors which is orthogonal to X_A.

The expected squared error of prediction is then

$$\begin{aligned}E[\hat{y}(\mathbf{x}) - \mathbf{x}'\boldsymbol{\beta}]^2 &= \mathbf{x}_A'\{V(\mathbf{b}_A) + [E(\mathbf{b}_A) - \gamma_A][E(\mathbf{b}_A) - \gamma_A]'\}\mathbf{x}_A \\ &\quad + (\text{omission bias})^2 \\ &= \mathbf{x}_A'\{V(\mathbf{b}_A) + (\text{sel. bias})(\text{sel. bias})'\}\mathbf{x}_A \\ &\quad + (\text{omission bias})^2 \qquad (6.5)\end{aligned}$$

where 'sel. bias' denotes the selection bias, and $V(\mathbf{b}_A)$ is the covariance matrix of the p elements of $\mathbf{b}_A$ about their (biased) expected values.

Note that equation (6.5) still holds if $\mathbf{b}_A$ has been estimated using a method other than LS.

To derive the *MSEP* from (6.5), we need to supply a set of $\mathbf{x}$-vectors over which to average. A simple, well-known result can be obtained if we make the following choice of $\mathbf{x}$-vectors and the following assumptions about $\mathbf{b}_A$:

1. The future $\mathbf{x}$-vectors will be the same as those in the X-matrix used for the estimation of the regression coefficients.
2. There is no selection bias.
3. $V(\mathbf{b}_A) = \sigma^2(X_A'X_A)^{-1}$.

The second and third conditions apply for LS regression coefficients when the subset has been chosen independently of the data used for the estimation of $\mathbf{b}_A$.

Subject to these conditions, we have that

$$E(RSS_p) = \sum_{i=1}^{n} (\text{omission bias})_i^2 + (n-p)\sigma^2 \qquad (6.6)$$

where $(\text{omission bias})_i$ is the bias in estimating the ith observation caused by omitting the $(k-p)$ predictors. We can obtain an estimate of the sum of squares of thse omission biases if we replace the

left-hand side of (6.6) with the sample RSS_p. Note that when the same data are used for both selection and estimation, the sample value of RSS_p is liable to be artificially low for the selected subset.

Now we need to find the sum of the $\mathbf{x}'_A V(\mathbf{b}_A)\mathbf{x}_A$ terms in (6.5). That is, we need to calculate

$$\sigma^2 \sum_{i=1}^{n} \mathbf{x}'_i (X'_A X_A)^{-1} \mathbf{x}_i$$

where $\mathbf{x}'_i$ is the ith row of X_A. By examining which terms are being multiplied, we see that this is the sum of diagonal elements, i.e. the trace of a matrix product, and that

$$\begin{aligned} \sigma^2 \sum_{i=1}^{n} \mathbf{x}'_i (X'_A X_A)^{-1} \mathbf{x}_i &= \sigma^2 \operatorname{trace} X_A (X'_A X_A)^{-1} X'_A \\ &= \sigma^2 \operatorname{trace} (X'_A X_A)^{-1} X'_A X_A \\ &= \sigma^2 \operatorname{trace} I_{p \times p} \\ &= p\sigma^2. \end{aligned} \tag{6.7}$$

It is assumed that $X'_A X_A$ is of full rank (why select a subset with redundant variables?). Here we have used the property that for any pair of matrices $\boldsymbol{P}, \boldsymbol{Q}$ with appropriate dimensions,

$$\operatorname{trace} \boldsymbol{PQ} = \operatorname{trace} \boldsymbol{QP}.$$

Finally, summing the terms (6.5) over the n future observations and using (6.6) and (6.7), we have that the sum of squared errors

$$\begin{aligned} &\approx p\sigma^2 + RSS_p - (n-p)\sigma^2 \\ &= RSS_p - (n-2p)\sigma^2. \end{aligned} \tag{6.8}$$

If we divide through by σ^2 we obtain the well-known Mallows's C_p statistic (Mallows, 1973):

$$C_p = \frac{RSS_p}{\sigma^2} - (n-2p). \tag{6.9}$$

In practice, σ^2 is replaced with the unbiased estimate

$$\hat{\sigma}^2 = \frac{RSS_k}{(n-k)}$$

that is, the estimate of the residual variance for the full model.

The mean squared error of prediction (*MSEP*) is defined as

$$E(y - \hat{y}(\mathbf{x}))^2,$$

where y is a future value of the Y-variable. As we have so far been looking at differences between $\hat{y}(\mathbf{x})$ and its expected value, we must now add on an extra term for the future difference between the expected value and the actual value. Hence the *MSEP* is obtained from (6.8) by dividing through by n and then adding an extra σ^2. Finally, for the fixed-variables model with unbiased LS estimates of regression coefficients,

$$MSEP \approx (RSS_p + 2p\sigma^2)/n. \tag{6.10}$$

Minimizing Mallows' C_p has been widely used as a criterion in subset selection despite the requirement for unbiased regression coefficients. Most of the applications have been to situations with predictors which are random variables whereas the derivation requires that the X's be fixed or controllable variables.

Mallows himself warns that minimizing C_p can lead to the selection of a subset which gives an *MSEP*, using LS regression coefficients, which is much worse than if there were no selection at all and all predictors are used. His demonstration uses k orthogonal predictors; a similar derivation follows. First, though, we present a shorter, simpler derivation of the *MSEP* and hence of Mallow's C_p based upon an orthogonal projection approach.

Write an orthogonal reduction of $\boldsymbol{X}$ as

$$\begin{aligned}\boldsymbol{X} &= (\boldsymbol{X}_A, \boldsymbol{X}_B) \\ &= (\boldsymbol{Q}_A, \boldsymbol{Q}_B)\begin{Bmatrix} \boldsymbol{R}_A & \boldsymbol{R}_{AB} \\ & \boldsymbol{R}_B \end{Bmatrix}\end{aligned}$$

where $\boldsymbol{Q}_A$, $\boldsymbol{Q}_B$ are orthogonal matrices, and $\boldsymbol{R}_A$, $\boldsymbol{R}_B$ are upper triangular (though $\boldsymbol{R}_{AB}$ is not). Let the vector of orthogonal projection be

$$\begin{Bmatrix} \boldsymbol{Q}_A' \\ \boldsymbol{Q}_B' \end{Bmatrix} Y = \begin{Bmatrix} \mathbf{t}_A \\ \mathbf{t}_B \end{Bmatrix}$$

with expected values τ_A, τ_B for the vectors $\mathbf{t}_A$, $\mathbf{t}_B$. The LS regression coefficients for the selected variables are obtained by solving

$$\boldsymbol{R}_A \mathbf{b}_A = \mathbf{t}_A.$$

Hence the prediction errors, if our future X is exactly the same as the X used for estimation, are given by

$$\begin{aligned} X\boldsymbol{\beta} - X_A\mathbf{b}_A &= QR\boldsymbol{\beta} - Q_AR_A\mathbf{b}_A \\ &= Q\boldsymbol{\tau} - Q_A\mathbf{t}_A. \end{aligned}$$

Hence, because of the orthogonality of the columns of Q, the sum of squared errors

$$\begin{aligned} &= E(X\boldsymbol{\beta} - X_A\mathbf{b}_A)'(X\boldsymbol{\beta} - X_A\mathbf{b}_A) \\ &= \sum_{i=1}^{p} \operatorname{var}\mathbf{t}_{Ai} + \sum_{i=p+1}^{k} \tau_{Bi}^2 \end{aligned} \tag{6.11}$$

where $\mathbf{t}_{Ai}$, τ_{Bi} are the ith elements of $\mathbf{t}_A$ and $\boldsymbol{\tau}_B$ respectively. As

$$E(RSS_p) = \sum_{i=p+1}^{k} \tau_{Bi}^2 + (n-p)\sigma^2,$$

and the projections $\mathbf{t}_{Ai}$ have variance σ^2, we have that the sum of squared errors

$$= p\sigma^2 + E(RSS_p) - (n-p)\sigma^2.$$

Replacing the expected value of RSS_p with its sample value gives formula (6.8).

Now let us look at the case of orthogonal predictors with estimation and selection from the same data. Adding variable X_i to the subset of selected variables reduces the residual sum of squares by $\mathbf{t}_i^2$, hence the approximate $MSEP$, or equivalently Mallows' C_p, is minimized by including all of those variables for which $\mathbf{t}_i^2 > 2\hat{\sigma}^2$. Hence if the $\mathbf{t}_i$'s are normally distributed with expected values τ_i and, for convenience, $\sigma = 1$, then

$$E(t_i|\text{variable } X_i \text{ is selected}) = \frac{\displaystyle\int_{-\infty}^{-\sqrt{2}} t\phi(t-\tau_i)\,\mathrm{d}t + \int_{\sqrt{2}}^{\infty} t\,\phi(t-\tau_i)\,\mathrm{d}t}{\Phi(-\sqrt{2}-\tau_i) + 1 - \Phi(\sqrt{2}-\tau_i)}$$

where ϕ and Φ are the density function and distribution function respectively for the standard normal distribution. Here, for simplicity, it is being assumed that $\hat{\sigma} = 1$; a more rigorous derivation would use the t-distribution instead of the normal.

The true sum of squared errors in this case is then given by

$$\sum_{i=1}^{p} (\text{selection bias})_i^2 + \sum_{i=1}^{p} \operatorname{var}(\mathbf{t}_{Ai}) + \sum_{i=p+1}^{k} \tau_{Bi}^2 \tag{6.12}$$

Table 6.1 *Expected values of contribution to the sum of squared errors against expected projections for orthogonal predictors, using Mallows' C_p as the stopping rule*

τ_i	*Contribution if selected*	*Contribution if rejected*	*Wtd average from (6.13)*
0·2	3·46	0·04	0·61
0·4	3·01	0·16	0·70
0·6	2·48	0·36	0·84
0·8	1·99	0·64	1·02
1·0	1·59	1·00	1·20
1·2	1·28	1·44	1·37
1·4	1·06	1·96	1·51
1·6	0·90	2·56	1·60
1·8	0·79	3·24	1·65
2·0	0·73	4·00	1·64
2·5	0·72	6·25	1·49
3·0	0·81	9·00	1·27
3·5	0·90	12·25	1·11
4·0	0·96	16·00	1·04
4·5	0·989	20·25	1·009
5·0	0·998	25·00	1·002

where the variance of the projections of selected variables is no longer equal to σ^2. Table 6.1 shows the contributions to the sum of squared errors according to whether a variable is selected (the first two terms of (6.12)) or rejected (last term of (6.12)), as a function of the expected projection, τ_i.

In practice, we do not know the expected values of the projections, so where should we apply the cut-off? To answer this question, let us look at the sum of squared errors for a mixture of τ_i's. Let $\tau_1 = 10$, so that the first variable will almost always be selected, and let $\tau_i = 10\alpha^{i-1}$ for some α between 0 and 1. Let the residual standard deviation, $\sigma = 1$. Let C be the cut-off value such that variable X_i is selected if $|t_i| \geqslant C$, and rejected otherwise. For variable X_i, the expected contribution to the sum of squared errors is

$$\begin{aligned}&\text{prob}(|t_i| \geqslant C)\, E[(t_i - \tau_i)^2 \text{ given } |t_i| \geqslant C] + \text{prob}(|t_i| < C)\cdot\tau_i^2 \\ &= \int_{-\infty}^{-C} (t-\tau_i)^2 \phi(t-\tau_i)\,dt + \int_{C}^{\infty} (t-\tau_i)^2 \phi(t-\tau_i)\,dt \\ &\quad + \tau_i^2 \int_{-C}^{C} \phi(t-\tau_i)\,dt. \qquad (6.13)\end{aligned}$$

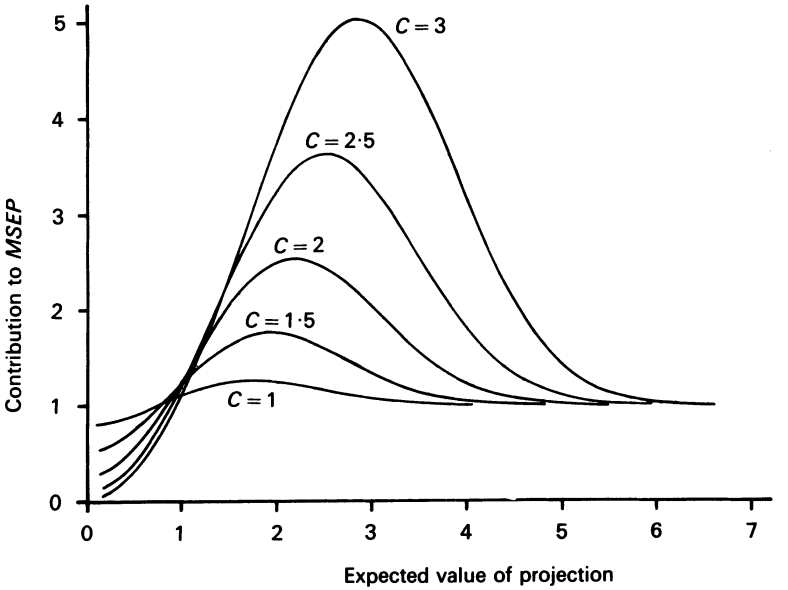


Fig. 6.1 *Expected contributions to the MSEP for the case of orthogonal predictors, against the expected value of a projection, τ, for a range of cut-off values, C.*

This quantity is shown as function $m(\tau)$ in Fig. 4 of Mallows (1973), and is shown here in Fig. 6.1.

We see from Fig. 6.1 that if the true but unknown projections are less than about 0·8 in absolute value then we should reject that variable. In this case, the higher the value of C the better. But if the unknown true projections are greater than about 1·1 in absolute value, then a large value of C is very undesirable.

Figures 6.2 and 6.3 show the error sums of squares for the mixture of τ_i's described above for $\alpha = 0{\cdot}8$ and 0·6 respectively. In Fig. 6.2, for $k = 10$ available predictors, the smallest true projection is $\tau_{10} = 1{\cdot}34$, so that there are no very small projections. In this case, the use of any cut-off is undesirable. Even in the case of $k = 20$, when the smallest true projection is $\tau_{20} = 0{\cdot}144$, there are sufficiently many true projections between 1 and 10 that nothing is gained by using any cut-off. For larger values of k or smaller values of α (see Fig. 6.3), when there are sufficiently many true small projections, moderate gains in reducing the error sum of squares can be obtained

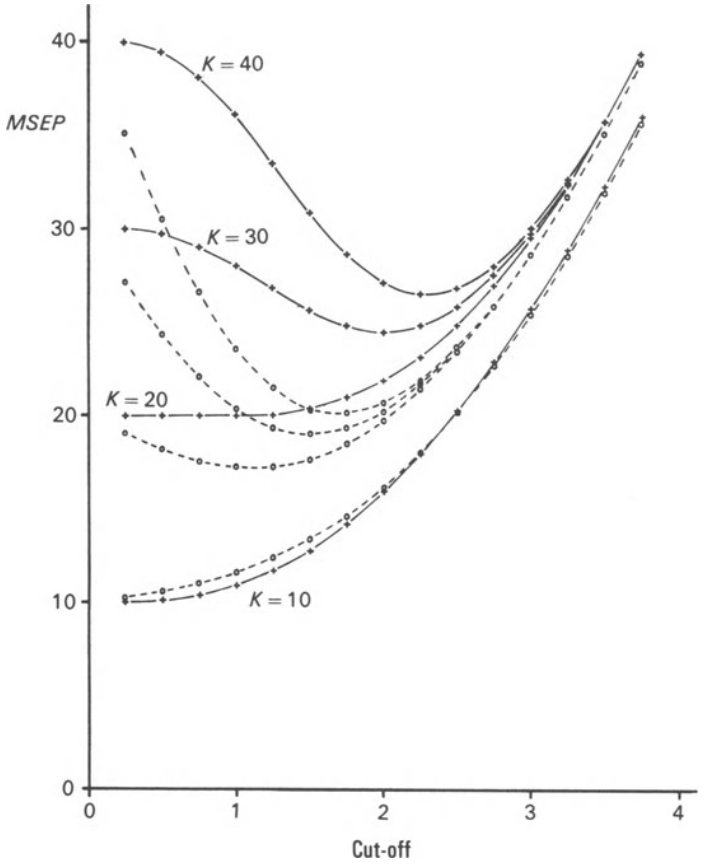


Fig. 6.2 *MSEP against cut-off for* $k = 10, 20, 30$ *or* 40 *predictors, and* $\tau_i = 10\alpha^{i-1}$ *with* $\alpha = 0{\cdot}8$. *The broken lines indicate the MSEP which would be obtained if unbiased estimates of the projections could be obtained when variables are selected.*

by using a cut-off in the vicinity of 2·0 standard deviations, as opposed to the $\sqrt{2}$ for Mallows' C_p.

Figures 6.2 and 6.3 also show, as broken lines, the error sums of squares which are obtained if unbiased LS regression coefficients are used, e.g. by using independent data for model selection and estimation. This illustrates the potential improvement which can be

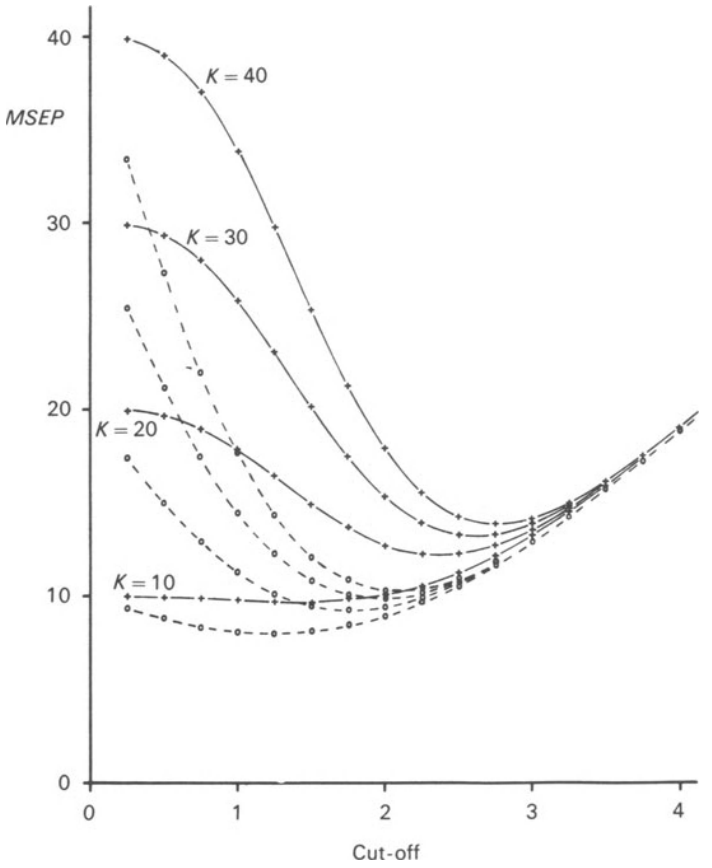


Fig. 6.3 *MSEP against cut-off for* $k = 10, 20, 30$ or 40 *predictors, and* $\tau_i = 10\alpha^{i-1}$ *with* $\alpha = 0{\cdot}6$. *The broken lines indicate the MSEP which would be obtained if unbiased estimates of the projections, with variance* = 1, *could be obtained when variables are selected.*

obtained if the bias in the LS regression coefficients can be removed or reduced.

It appears surprising that for $k = 10$ and $\alpha = 0{\cdot}8$ in Fig. 6.2, the error sum of squares is slightly lower when biased LS regression coefficients are used rather than unbiased ones, for cut-offs, C, less than about 2·5. The explanation is that the variance of the biased regression coefficients is substantially smaller than that of the unbiased coefficients, and this difference more than outweighs the

square of the bias. For instance, for the smallest true projection of 1·34 and $C = 1$, the expected value of t_{10} when variable X_{10} is selected is 1·886, giving a bias of 0·546, but its variance is only 0·597.

Looking back to equation (6.12), we see that the first term in the sum of squared errors is the square of the selection bias. If we can halve this bias, then its contribution to the squared error will be divided by four. We can anticipate that the likelihood estimation method described in section 5.4 will thus give squared errors closer to the broken lines than to the solid ones in Figs 6.2 and 6.3, and hence that a stopping rule such as minimizing Mallows' C_p or an F-to-enter of about 2·0, will produce something close to the optimal linear subset predictor.

The conclusions of this section are that for prediction when future values of the predictor variables will take the same values as those in the calibration sample:

1. Using biased LS regression coefficients estimated from the same data as were used to select the model, a cut-off value of about 2·0 standard deviations for the absolute value of the LS projections (roughly equivalent to an F-to-enter of 4·0) is about optimal when an appreciable fraction of the true (but unknown!) projections are less than about 0·8 standard deviations in absolute value, otherwise no selection should be used.
2. It is desirable to try to reduce or eliminate the bias in the LS regression coefficients. If this can be achieved, then using a cut-off of between 1·5 and 2·0 standard deviations for the sample LS projections may be about optimal.

The results described in this section have been for orthogonal predictors. In this case, the contribution of a variable to the sum of squared errors is independent of the other variables which have been selected, and so there is no selection bias. The results should be applied with caution in cases where the predictors in the calibration data are not orthogonal. It should be emphasized that these results are for an assumed geometric progression of values of the true projections; there is no evidence that this pattern is realistic.

6.2.2 *MSEP for the random model*

In this case, the omission bias in the fixed model becomes additional residual variation. However, in practice, the columns of the X-matrix

will usually be far from orthogonal, and there will often be considerable competition between variables for selection.

Suppose that

$$\mathbf{Y} = X_A\boldsymbol{\beta}_A + \boldsymbol{\varepsilon}$$

where the residuals ε, have zero expected value and $E(\boldsymbol{\varepsilon}^2) = \boldsymbol{\sigma}_A^2$, i.e. the residual variance is a function of the subset. Assume for the moment that we have unbiased estimates $\mathbf{b}_A$ of $\boldsymbol{\beta}_A$ and $\mathbf{s}_A^2$ of $\boldsymbol{\sigma}_A^2$.

For one future prediction, the prediction error is

$$\hat{\mathbf{y}}(\mathbf{x}_A) - \mathbf{x}_A'\boldsymbol{\beta}_A = \mathbf{x}_A'(\mathbf{b}_A - \boldsymbol{\beta}_A).$$

Hence the predictions are unbiased with variance:

$$\mathbf{x}_A' V(\mathbf{b}_A)\mathbf{x}_A,$$

where $V(\mathbf{b}_A)$ is the covariance matrix of $\mathbf{b}_A$.

If we use independent data for selection and estimation, which would give us unbiased LS regression coefficients, then the covariance of the regression coefficients is

$$V(\mathbf{b}_A) = \sigma_A^2(X_A'X_A)^{-1},$$

where X_A is the X-matrix used for estimation (not that used for model selection). If we now average the squared prediction errors over $\mathbf{x}_A$'s comprising the rows of X_A, we derive

$$MSEP = \frac{\sigma_A^2}{n}\sum_{i=1}^{n} x_i'(X_A'X_A)^{-1}x_i + \sigma_A^2$$

$$= \sigma_A^2\frac{n+p}{n} \tag{6.14}$$

from (6.7), where p is the rank of $X_A'X_A$. If we replace σ_A^2 with its sample estimate, then the estimated $MSEP$ is

$$\approx \frac{RSS_p}{(n-p)}\frac{n+p}{n}. \tag{6.15}$$

This expression for the $MSEP$ (see Rothman, 1968) usually gives numerical values which are almost the same as those from the fixed-variables model. If we replace the σ^2 in (6.8) by $RSS_p/(n-p)$ then we obtain (6.15).

At the minimum *MSEP* we have that

$$\frac{RSS_p(n+p)}{(n-p)n} \leqslant \frac{RSS_{p+1}(n+p+1)}{(n-p-1)n}.$$

A little rearrangement shows that at the minimum

$$(n+p)(n-p-1)(RSS_p - RSS_{p+1}) \leqslant 2n\, RSS_{p+1}$$

or

$$\frac{RSS_p - RSS_{p+1}}{RSS_{p+1}/(n-p-1)} \leqslant \frac{2n}{n+p}. \tag{6.16}$$

The left-hand side of (6.16) is the usual '*F*-to-enter' statistic, so that when the *MSEP* is minimized the '*F*-to-enter' for the next larger subset is less than $2n/(n+p)$, or a little less than 2 if $n \gg p$.

For the random model though, it is extremely unreasonable to assume that future x's will take the same values as in the calibration sample. It may be reasonable though in some circumstances to assume that future x's will be sampled from the same distribution as the x's in the calibration sample. The *MSEP* for future x's is

$$MSEP = \sigma_A^2 + E[\mathbf{x}'_A \boldsymbol{M}_2(\mathbf{b}_A - \boldsymbol{\beta}_A)\mathbf{x}_A]$$

where $\boldsymbol{M}_2(\mathbf{b}_A - \boldsymbol{\beta}_A)$ is the matrix of second moments of $\mathbf{b}_A$ about $\boldsymbol{\beta}_A$. The expectation has to be taken over both the future x's and $\boldsymbol{M}_2$.

Again let us assume that we have unbiased estimates $\mathbf{b}_A$ of $\boldsymbol{\beta}_A$ with covariance matrix

$$V(\mathbf{b}_A) = \sigma_A^2 (\boldsymbol{X}'_A \boldsymbol{X}_A)^{-1},$$

as would be the case if independent data had been used for model selection and parameter estimation. We have then that

$$MSEP = \sigma_A^2 + \sigma_A^2 E[\mathbf{x}'_A (\boldsymbol{X}'_A \boldsymbol{X}_A)^{-1} \mathbf{x}_A] \tag{6.17}$$

In circumstances in which the future x's are already known, substitution in (6.17) then gives the *MSEP*. Galpin and Hawkins (1982, 1986) do precisely that, and show that in some cases quite different subsets should be chosen for different x's to give the best predictions in the *MSEP* sense.

A simple general formula for the *MSEP* can be obtained if we assume

1. that a constant is being fitted in the model; and

2. that the calibration sample and the future x's are independently sampled from the same multivariate normal distribution.

We have then from (6.17)

$$MSEP = \sigma_A^2\{1 + (1/n) + E[x_A^{*\prime}(X_A^{*\prime}X_A^*)^{-1}\mathbf{x}_A^*]\}$$

where the asterisks indicate that the sample mean of the calibration data has been removed from each variable. The $(1/n)$ is for the variance of the mean. The vectors are now of length $(p-1)$.

If the X-variables are sampled from a multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}_A$, then the variance of the future $\mathbf{x}_A^*$'s, after allowing for the removal of the sample mean, is

$$[1 + (1/n)]\boldsymbol{\Sigma}_A.$$

An estimate of $\boldsymbol{\Sigma}_A$ is provided by

$$V = (X_A^{*\prime}X_A^*)/(n-1);$$

hence we can write

$$\mathbf{x}_A^{*\prime}(X_A^{*\prime}X_A^*)^{-1}\mathbf{x}_A^* = \frac{n+1}{n(n-1)}\mathbf{t}'V^{-1}\mathbf{t} \tag{6.18}$$

where $\mathbf{t}$ is a vector of statistics with zero mean and covariance matrix $\boldsymbol{\Sigma}_A$. Now $\mathbf{t}'V^{-1}\mathbf{t}$ is Hotelling's T^2-statistic (see any standard text on multivariate analysis, e.g. Morrison, 1967, pp. 117–24, or Press, 1972, pp. 123–6). The quantity

$$(n-p+1)T^2/((p-1)(n-1))$$

is known to have an F-distribution with $(p-1)$ and $(n-p+1)$ degrees of freedom for the numerator and denominator respectively. The expected value of $F(v_1, v_2)$ is $v_2/(v_2-2)$, where v_1, v_2 are the numbers of degrees of freedom, and hence the expected value of (6.18) is

$$\frac{n+1}{n(n-1)}\ \frac{(n-1)(p-1)}{n-p+1}\ \frac{n-p+1}{n-p-1} = \frac{(n+1)(p-1)}{n(n-p-1)}.$$

Finally, we derive

$$MSEP = \sigma_A^2\left\{1 + (1/n) + \frac{(n+1)(p-1)}{n(n-p-1)}\right\}. \tag{6.19}$$

This result is due to Stein (1960), though the derivation given here is that of Bendel (1973). Notice that here the p variables include the constant in the count, in line with the usual derivation of Mallows' C_p, which is the equivalent result for the fixed-variables case. Other authors, including Thompson (1978), Oliker (1978) and Bendel, quote the result without the constant included in the count for p.

For large n and $p \ll n$, the last term in (6.19) is approximately $(p-1)/n$ so that the numerical value of (6.19) is nearly the same as that of (6.14). As p/n increases though, the last term of (6.19) increases rapidly.

If we replace σ_A^2 with the estimate $RSS_p/(n-p)$ then the *MSEP* becomes

$$\frac{RSS_p}{n-p}\,\frac{(n+1)(n-2)}{n(n-p-1)}. \tag{6.20}$$

Minimizing this estimated *MSEP* with respect to p is then equivalent to minimizing

$$\frac{RSS_p}{(n-p)(n-p-1)}.$$

At the minimum we have that

$$\frac{RSS_p}{(n-p)(n-p-1)} < \frac{RSS_{p+1}}{(n-p-1)(n-p-2)}$$

or, after some rearrangement, that

$$\frac{RSS_p - RSS_{p+1}}{RSS_{p+1}/(n-p-1)} < \frac{2(n-p-1)}{n-p-2}.$$

That is, at the minimum the 'F-to-enter' statistic is less than a quantity which is just greater than 2. In practice, minimizing the estimated *MSEP* given by (6.20) often selects the same size of subset as minimizing Mallows' C_p.

A somewhat surprising feature of all the formulae for the *MSEP* which we have derived is that they do not involve the X-matrix. That is, the estimated *MSEP* is the same whether the X-variables are highly correlated or almost independent. The basic reason for this independence of X is that the same pattern of correlations has

been assumed for future x's. These correlations are important though if either

1. the future x's will have a different pattern from those used for calibration; or
2. the same data are used for both model selection and estimation of the regression coefficients.

6.2.3 *A simulation with random predictors*

In most practical cases the same data are used for both model selection and for estimation. To examine the effect of competition bias in inflating the *MSEP*, a simulation experiment has been carried out. The experiment was a 5×2^3 complete factorial with replication. The four factors were as follows:

1. Number of available predictors, $k = 10(10)50$, plus a constant.
2. Sample size, n, either small ($n = 2k$) or moderate ($n = 4k$).
3. Ill-conditioning of $X'X$, either low or moderate.
4. Size of smallest LS projection, either small or moderate.

To obtain the desired amount of ill-conditioning, a method similar to that of Bendel (1973) was employed. A diagonal matrix was constructed with diagonal elements decreasing geometrically except for a small constant, δ, added to prevent them from becoming too small. The initial values of the diagonal elements were

$$\lambda_1 = \frac{k(1-\alpha)(1-\delta)}{(1-\alpha^k)} + \delta,$$

and

$$\lambda_i = \alpha(\lambda_{i-1} - \delta) + \delta \text{ for } i = 2, \ldots, k.$$

δ was arbitrarily set equal to 0·001. The λ's are then of course the eigenvalues of this matrix, and are chosen to sum to k, which is the trace of a $k \times k$ correlation matrix. Similarity transformations of the kind

$$\boldsymbol{A}_{r+1} = \boldsymbol{P}\boldsymbol{A}_r\boldsymbol{P}^{-1}$$

preserve the eigenvalues. Random planar rotations were applied to each pair of rows and columns. A further $(k-1)$ similarity transformations were then used to produce 1's on the diagonal. Thus if the diagonal block for rows i and $i+1$ is

$$\begin{pmatrix} z & x \\ x & m \end{pmatrix}$$

then the required similarity transformation is

$$\begin{pmatrix} c & s \\ -s & c \end{pmatrix}\begin{pmatrix} w & x \\ x & z \end{pmatrix}\begin{pmatrix} c & -s \\ s & c \end{pmatrix}=\begin{pmatrix} 1 & x^* \\ x^* & z^* \end{pmatrix}$$

where the tangent ($t = s/c$) satisfies

$$t^2(1-z)-2tx+(1-w)=0,$$

that is

$$t=\frac{x\pm\{x^2-(1-z)(1-w)\}^{1/2}}{(1-z)}.$$

In the rare event that the roots of this quadratic were not real, rows i and $i+2$, or i and $i+3$ if necessary, were used. As the diagonal elements must average 1·0, it is always possible to find a diagonal element z on the opposite side of 1·0 from w so that the product $(1-z)(1-w)$ is negative and hence the roots are real.

The matrix so generated was used as the covariance matrix $\boldsymbol{\Sigma}$, of a multivariate normal distribution. If we form the Cholesky factorization

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}'$$

where $\boldsymbol{L}$ is a lower-triangular matrix, then a single sample can be generated from this distribution using

$$\boldsymbol{x} = \boldsymbol{L}\boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is a vector of elements ε_i which are sampled independently from the standard normal distribution. This follows as the covariance matrix of the elements of x is

$$\begin{aligned} E(\boldsymbol{x}\boldsymbol{x}') &= E(\boldsymbol{L}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{L}') \\ &= \boldsymbol{L}\boldsymbol{L}' \\ &= \boldsymbol{\Sigma}. \end{aligned}$$

The projections of the Y-variable were chosen so that their expected values, $\tau_i = n^{1/2}\gamma^{i-1}$ for $i = 1,\ldots,k$. The expected values of the regression coefficients for these projections were then obtained by solving

$$\boldsymbol{L}'\boldsymbol{\beta} = \boldsymbol{\tau},$$

and each value y of Y was generated as

$$y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon,$$

where the residuals, ε, were sampled from the standard normal distribution.

To decide upon suitable values to use for α, the eigenvalues (principal components) of some correlation matrices were examined. Table 6.2 shows the eigenvalues for the DETROIT, LONGLEY, POLLUTE and CLOUDS data sets. A crude way of choosing a value for α for each data set is to fit the largest and smallest (nonzero) eigenvalues, ignoring δ. If we call these values λ_{max} and λ_{min}, then the fitted value of α is $(\lambda_{min}/\lambda_{max})^{1/(r-1)}$ where r is the rank. Table 6.3 shows these values, together with similar values calculated for the data sets examined by Bendel (1973, p. 91).

Based upon Table 6.3, the values chosen for α and γ in the simulations were as follows:

No. of predictors (k)	10	20	30	40	50
High α or γ	0.75	0.80	0.85	0.90	0.95
Low α or γ	0.45	0.55	0.70	0.75	0.80

NB The number of predictors k shown above excludes the constant which was fitted.

Table 6.2 *Eigenvalues (principal components) of some correlation matrices*

Data set			
DETROIT	*LONGLEY*	*POLLUTE*	*CLOUDS*
2·83	2·15	2·13	2·85
1·38	1·08	1·66	2·26
0·90	0·45	1·43	1·81
0·32	0·12	1·16	1·37
0·26	0·051	1·11	1·05
0·18	0·019	0·98	0·45
0·14		0·78	0·40
0·12		0·69	0·31
0·075		0·61	0·22
0·050		0·47	0·11
0·020		0·41	0·093
		0·36	0·060
		0·34	0·021
		0·21	
		0·070	

Table 6.3 *Ratios of smallest to largest eigenvalues of correlation matrices. Data for the last eight data sets have been calculated from the table on p. 91 of Bendel (1973)*

Data set	*Rank(r)*	λ_1	λ_r	$(\lambda_1/\lambda_r)^{1/(r-1)}$
DETROIT	11	2·83	0·020	0·37
LONGLEY	6	2·15	0·019	0·15
POLLUTE	15	2·13	0·070	0·61
CLOUDS	13	2·85	0·021	0·44
AFI	13	3·9	0·02	0·64
CMA	16	4·2	0·14	0·80
CRD	11	3·2	0·11	0·71
CWE	19	2·1	0·27	0·89
MCC	15	2·4	0·10	0·80
MEY	23	8·8	0·05	0·79
ROC[a]	12	3·5	0·36	0·81
ROC[b]	12	7·7	0.17	0·71

A high value for α meant that the X-predictors were not highly correlated, while a low value meant that they were. A high value for γ meant that the smallest projection was of the order of the residual standard deviation ($\sigma = 1$) when $n = 2k$, and this led to the selection of large subsets. The smaller values of γ meant that a moderate number of the expected projections were very large compared with the noise in the sample projections, and this led to the selection of much smaller subsets.

The method used to find subsets which fitted well was sequential replacement. This was used as a compromise in speed between the widely used Efroymson algorithm, and the use of the exhaustive search algorithm.

Using 10 replicates, that is, 10 artificial data sets for each case, and minimizing the estimated *MSEP* given by (6.20) as the stopping rule, the sample means and standard deviations of the sizes of selected subsets were as given in Table 6.4. It is interesting to note that the degree of ill-conditioning of the Σ-matrix, as indicated by α, had very little effect upon the size of subset selected.

For comparison, Mallows' C_p was also used as a stopping rule, though its derivation is only for the fixed predictors case. In 69% of

Table 6.4 *Sample means and standard deviations (in brackets) of sizes of selected subsets using minimum estimated MSEP as the stopping rule. The constant has been excluded from the count of variables in the table below*

		Large α		*Small* α	
k	n	*Large* γ	*Small* γ	*Large* γ	*Small* γ
10	20	4·9 (1·0)	4·4 (1·6)	4·8 (1·6)	4·9 (2·2)
	40	6·3 (1·3)	4·7 (0·8)	5·9 (1·1)	4·0 (0·9)
20	40	10·1 (2·4)	6·7 (1·9)	8·3 (2·4)	7.2 (2·9)
	80	10·1 (2·5)	7·2 (1·9)	9·5 (2·4)	6·6 (3·1)
30	60	14·1 (2·7)	10·6 (2·9)	12·8 (2·3)	12·1 (5.7)
	120	15·8 (2·4)	11·6 (2·5)	15·1 (2·8)	10·5 (2·6)
40	80	20·9 (2·8)	14·5 (2·1)	22·0 (4·4)	13·9 (3·0)
	160	23·2 (2·1)	13·4 (3·1)	23·3 (2·2)	13·4 (3.7)
50	100	33·9 (3.3)	18·0 (3·1)	35·0 (3·0)	19·9 (4.2)
	200	40·3 (2·6)	18·3 (4·1)	37·0 (4·1)	17·7 (3·6)

cases, it selected the same size of subset, in 1% of cases it selected a larger subset; in the remaining 30% of cases it selected a smaller subset. When the true *MSEP*'s were compared, those selected using Mallows' C_p were smaller in 21% of cases and larger in 10% than the true values when the estimated *MSEP* was minimized.

Table 6.5 shows the average values of three different *MSEP*'s when the stopping rule used was that of minimizing the estimated *MSEP*. The first of these, labelled (a) in the table, is the estimated *MSEP* given by (6.20). The second is the true *MSEP*. As we know the true population values of the regression coefficients, the error in a future prediction for a known vector, $\mathbf{x}_A$, of values of the predictors in the selected subset A is

$$x'_A(\mathbf{b}_A - \boldsymbol{\beta}_A) + \eta$$

where $\mathbf{b}_A, \boldsymbol{\beta}_A$ are the vectors of estimated and population regression coefficients, and η is a residual with standard deviation σ_A. Future $\mathbf{x}_A$'s can be generated as

$$\mathbf{x}_A = L_A \boldsymbol{\varepsilon}$$

where L_A is the lower-triangular Cholesky factorization of those rows and columns of $\boldsymbol{\Sigma}$ relating to variables in subset A, and $\boldsymbol{\varepsilon}$ is a vector

Table 6.5 *Average MSEP's at the minimum of the estimated MSEP: (a) average estimates MSEP for LS regression coefficients; (b) average true MSEP for LS regression coefficients, and (c) average MSEP for unbiased LS regression coefficients. Each average is based upon 10 replications*

		Large α						*Small* α					
		Large γ			*Small* γ			*Large* γ			*Small* γ		
k	n	*(a)*	*(b)*	*(c)*	*(a)*	*(b)*	*(c)*	*(a)*	*(b)*	*(c)*	*(a)*	*(b)*	*(c)*
10	20	1·32	2·22	1·96	1·05	1·77	1·53	1·20	2·03	1·79	0·98	2·20	1·47
	40	1·13	1·58	1·45	1·05	1·34	1·26	1·16	1·44	1·44	1·01	1·22	1·22
20	40	1·17	2·00	1·82	0·98	1·54	1·36	1·10	1·63	1·56	0·92	1·66	1·37
	80	1·00	1·39	1·29	1·01	1·26	1·13	1·09	1·32	1·29	0·95	1·22	1·16
30	60	1·14	1·83	1·56	1·00	1·67	1·44	1·08	1·68	1·66	0·84	1·65	1·41
	120	1·06	1·29	1·28	1·04	1·24	1·16	1·07	1·29	1·26	1·00	1·24	1·16
40	80	1·08	1·86	1·74	0·87	1·56	1·30	1·08	1·95	1·67	0·96	1·54	1·33
	160	1·14	1·36	1·31	1·04	1·22	1·14	1·06	1·35	1·28	1·05	1·23	1·14
50	100	1·28	2·44	2·32	0·93	1·65	1·34	1·17	2·20	2·00	0·97	1·73	1·40
	200	1·22	1·44	1·45	0·97	1·26	1·16	1·20	1·48	1·43	1·01	1·21	1·15

of elements sampled from the standard normal distribution. Hence the prediction error can be written as

$$\boldsymbol{\varepsilon}' \boldsymbol{L}_A' (\mathbf{b}_A - \boldsymbol{\beta}_A) + \boldsymbol{\eta}$$

and hence the true *MSEP* is

$$(\mathbf{b}_A - \boldsymbol{\beta}_A)' \boldsymbol{L}_A \boldsymbol{L}_A' (\mathbf{b}_A - \boldsymbol{\beta}_A) + \sigma_A^2. \tag{6.21}$$

This quantity is shown as (b) in Table 6.5.

The third *MSEP* shown in Table 6.5, as (c), is that for unbiased LS regression coefficients, such as would be obtained from an independent set of data from the same population with the same sample size. This *MSEP* has been calculated using (6.19) with the known population variance σ_A^2.

A number of important hypotheses are suggested by Table 6.5. First we notice that when the sample size is only double the number of available predictors (excluding the constant), the estimated *MSEP* is an underestimate (column (a) versus column (b)). The true *MSEP* using LS regression coefficients is between 20 and 60% larger than that estimated. However, when the sample size is four times the

number of available predictors, there is no evidence of any underestimation of the *MSEP*. There are two contributory factors to this. First, the larger sample size allows much better discrimination between subsets. This means that with the larger sample sizes there is less competition between close subsets and hence less bias in the regression coefficients. This will be shown more clearly in Table 6.8. The other contributory factor is the increased number of degrees of freedom for the residual sum of squares. The artificially high regression sum of squares for the best-fitting subset has a relatively smaller effect in depressing the residual sum of squares when $n = 4k$ than when $n = 2k$.

The *MSEP*'s in column (c), which are those which would be obtained using an independent set of n observations to estimate the regression coefficients, are usually slightly smaller than the true *MSEP*'s using the biased LS regression coefficients (column (b)). Hence, for situations similar to those simulated here, it is very much better to use all of the data to select the model and to use the same data to obtain biased LS regression coefficients for the selected model, than to split the data into equal parts and use one part for model selection and the other for estimation. The penalty is that if the first path is chosen, the estimate of the *MSEP* is optimistically small.

Figure 6.4 shows the (true) *MSEP*'s using LS regression coefficients from an independent data set, for two cases from the simulations. Both cases are for 40 available predictors and sample size = 80. Both were the first sets generated for $\alpha = 0.90$ (low correlations between predictors), but one is for $\gamma = 0.90$ (all projections moderately large), while the other is for $\gamma = 0.75$ (many small projections).

The A and B in Fig. 6.4 indicate the sizes of subset which minimized the (false) estimated *MSEP* given by (6.20) for $\gamma = 0.75$ and $\gamma = 0.90$ respectively. For $\gamma = 0.90$, minimizing Mallows' C_p led to the same stopping point; for $\gamma = 0.75$, minimizing Mallows' C_p picked a subset of 19 variables instead of 9.

The curves of *MSEP* versus p are usually fairly flat near the minimum, so that the exact choice of stopping point is often not very critical.

The *MSEP*'s were also calculated for the case in which independent data are used to estimate unbiased regression coefficients for the subsets using LS and assuming the same set of values for the

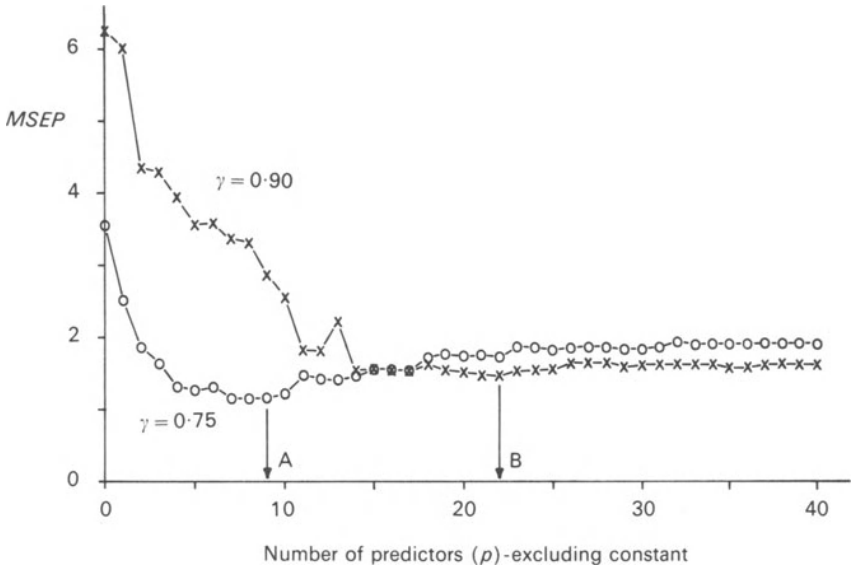


Fig. 6.4 *True MSEP using LS for simulated data for* $K = 40$ *available predictors and a sample size of* 80 *observations. See text for further description.*

X-variables. In most cases, these values were close to those using the biased LS coefficients from the same data. They were sufficiently close that they would not have been clearly distinguishable if they had been added to Fig. 6.4.

Probably the most frequently asked question with respect to subset selection in regression is 'What stopping rule should be used?' Table 6.6 was drawn up to illustrate the answer. If we generalize Mallows' C_p to

$$C_p(M) = \frac{RSS_p}{\hat{\sigma}^2} - (n - Mp) \tag{6.22}$$

where $M = 2$ for Mallows' C_p, then minimizing this quantity gives a range of stopping rules. Large values of M will lead to the selection of small subsets, and vice versa. M can be regarded as a penalty or cost for each additional variable in the selected subset.

Table 6.6 shows the true *MSEP*'s using this stopping rule and LS regression coefficients for $M = 1, 2$ and 3. The conclusion is clear. For large γ we should use a small M (selecting a large subset), while

Table 6.6 *Average true MSEP's using LS regression coefficients and the generalized Mallows' C_p as the stopping rule for $M = 1, 2,$ and 3 ($M = 2$ for Mallows' C_p)*

		Large α						*Small* α					
		Large γ			*Small* γ			*Large* γ			*Small* γ		
k	n	$M=1$	$M=2$	$M=3$	$M=1$	$M=2$	$M=3$	$M=1$	$M=2$	$M=3$	$M=1$	$M=2$	$M=3$
10	20	2·23	2·14	2·19	1·83	1·79	1·83	2·23	2·03	2·22	2·45	2·20	1·87
	40	1·53	1·59	1·61	1·31	1·30	1·29	1·45	1·44	1·51	1·26	1·22	1·22
20	40	2·06	2·05	1·82	1·62	1·45	1·40	1·73	1·64	1·71	1·87	1·52	1·46
	80	1·43	1·40	1·36	1·27	1·26	1·22	1·35	1·32	1·31	1·26	1·23	1·22
30	60	1·84	1·73	1·81	1·87	1·57	1·46	1·89	1·68	1·69	1·73	1·62	1·40
	120	1·30	1·28	1·27	1·27	1·22	1·19	1·33	1·29	1·24	1·34	1·23	1·20
40	80	1·95	1·84	1·79	1·67	1·52	1·43	2·04	1·91	1·92	1·86	1·47	1·39
	160	1·35	1·36	1·37	1·27	1·20	1·18	1·35	1·34	1·35	1·31	1·19	1·17
50	100	2·27	2·44	2·42	1·77	1·61	1·60	2·22	2·15	2·27	1·86	1·69	1·54
	200	1·39	1·43	1·48	1·31	1·24	1·22	1·43	1·49	1·51	1·25	1·21	1·20

for small γ we should use a large M (selecting a small subset). Thus the answer to the question, 'What stopping rule should we use?' does not have one answer. In some circumstances it is preferable to include many predictors, in others we should select a small number.

In practice, nothing equivalent to the γ of the simulations will be available, so the finding from Table 6.6 is of limited value. Suppose the rule of minimizing the estimated *MSEP* is used, with the *MSEP* falsely estimated from (6.20). If this selects a small subset, with $p < k/2$ say, it suggests that there may be many predictors which make little or no real contribution after the more important predictors have been included. This corresponds to the small γ case, and suggests that we should then use the modified Mallows' C_p with $M = 3$ say. On the other hand, if minimizing (6.20) selects a large subset, with $p > k/2$ say, then we should use $M = 1$ say. It appears that it is undesirable to pick a subset which is close to half the number of available predictors. In such cases, the number of alternative subsets of the same size is a maximum, and we can anticipate large competition biases. It is of course possible to construct artificial examples in which the best subset for prediction is of exactly half the number of available predictors, and doubtless such cases will occur sometimes in real-life examples.

If we use all of the available predictors without considering subset selection, then the regression coefficients are unbiased. This is sometimes a better strategy. Table 6.7 shows how often, out of the 10 replications in each case, the *MSEP* using all the available predictors was better than or equal to that for the selected subset using biased LS regression coefficients for that subset. Using all the predictors was nearly always as good as or only slightly worse for large γ, but rarely so for small γ. Note that in a few cases, particularly for $k = 10$, the selected subset was that of all k predictors.

Table 6.8 shows the estimated average bias of the LS regression coefficients after standardization and adjustment for sign. If $\mathbf{b}_{Ai}$ and β_{Ai} are the sample regression coefficient and its expected value respectively for variable X_i in subset A, then the standardized and sign-adjusted difference used was

$$\frac{b_{Ai} - \beta_{Ai}}{s_{\mathrm{LS}}(\mathbf{b}_{Ai})} \operatorname{sign}(\beta_{Ai}),$$

where the estimated standard errors, $s_{\mathrm{LS}}(\mathbf{b}_{Ai})$, are those for LS regression when the model has been chosen *a priori*, that is they are

Table 6.7 *Frequencies, out of* 10, *for which the MSEP using all available predictors (using LS regression coefficients) was smaller or equal to the true MSEP for the subset selected by minimizing the estimated MSEP*

		Large α		*Small* α	
k	n	*Large* γ	*Small* γ	*Large* γ	*Small* γ
10	20	6	2	2	2
	40	6	5	6	3
20	40	5	1	0	1
	80	5	5	4	3
30	60	3	2	1	2
	120	4	3	5	2
40	80	3	0	2	2
	160	5	3	5	2
50	100	6	2	4	2
	200	8	3	7	2

Table 6.8 *Average estimated bias and sample standard deviation (in brackets) of LS regression coefficients after standardization and sign adjustment, for selected subsets*

		Large α		*Small* α	
k	n	*Large* γ	*Small* γ	*Large* γ	*Small* γ
10	20	0·26 (0·69)	0·72 (0·70)	0·36 (0·52)	0·23 (0·22)
	40	0·31 (0·36)	0·72 (0·25)	0·49 (0·26)	0·38 (0·26)
20	40	0·38 (0·72)	0·72 (0·50)	1·03 (0·26)	1·17 (0·89)
	80	0·55 (0·64)	0·37 (0·24)	0·58 (0·16)	1·29 (1·44)
30	60	0·58 (0·52)	0·52 (0·83)	0·92 (0·27)	0·65 (0·93)
	120	0·31 (0·46)	0·57 (0·24)	0·36 (0·31)	0·66 (0·58)
40	80	0·54 (0·55)	0·46 (0·60)	0·88 (0·34)	0·68 (0·45)
	160	0·48 (0·56)	0·58 (0·35)	0·67 (0·41)	0·64 (0·44)
50	100	0·49 (0·81)	0·47 (0·50)	1·03 (0·53)	0·57 (0·77)
	200	0·17 (0·40)	0·24 (0·50)	0·38 (0·31)	0·41 (0·46)

the square roots of the diagonal elements of

$$s^2_{LS}(\mathbf{b}_A) = \sigma^2_A \operatorname{diag}(X'_A X_A)^{-1},$$

and σ^2_A is estimated from the residual sum of squares, RSS_A, for subset A in the usual way, i.e.

$$\sigma^2_A = RSS_A/(n-p),$$

where p is the number of variables, including the constant, in subset A.

The quantities shown in Table 6.8 are the averages of all the standardized differences between sample and population regression coefficients for all the selected variables except the constant in the model. Thus for the first entry, the numbers of regression coefficients used for the 10 replicates were 6, 3, 4, 4, 5, 10, 7, 4, 7 and 4.

We notice firstly that all of the averages in Table 6.8 are positive. The overall average of the bias estimates in the table is 0·57 of a standard error. As each subset selected would have contained a number of 'dominant' variables for which the bias would have been very small, a substantial proportion of the biases for other variables were well in excess of one standard error.

Table 6.9 contains a simple analysis of variance of the average estimated biases, with only main effects fitted and the interactions used to estimate the residual variance. The only effects which are significant (other than the constant in the model) are those related to α and k. The average estimated bias for large α is 0·47 of a standard

Table 6.9 *Analysis of variance of estimated biases of LS regression coefficients*

Factor	*Sum of squares*	*Deg. of freedom*	*Mean square*	*F-ratio*
α	0·388	1	0·388	8·68*
γ	0·041	1	0·041	0·92
k	0·538	4	0·135	3·01†
n	0·156	1	0·156	3·49
Residual	1·431	32	0·045	
Total	2·554	39		

*Significant at the 1% level.
†Significant at the 5% level.

error, while that for small α is 0·67 of a standard error. With small α, some of the X-predictors are highly correlated so that there is considerable competition among the predictors for selection. The average estimated biases are 0·43, 0·76, 0·57, 0·62 and 0·47 standard errors for the five values of k.

Another important feature of the simulation results in Table 6.8 is that most of the standard deviations are substantially less than 1.

The low variance of the biased LS regression coefficients is the basic explanation for the relatively good *MSEP*'s in the columns labelled (b) in Table 6.5 compared with the *MSEP*'s for unbiased LS regression coefficients in the columns labelled (c). If we refer back to formula (6.5), we see that an important part of the *MSEP* is

$$V(\mathbf{b}_A) + (\text{sel. bias})(\text{sel. bias})',$$

which is the second moment matrix of the regression coefficients. In the standardized units used for Table 6.8, for unbiased LS regression coefficients from an independent data set, the diagonal elements of $\mathbf{b}_A$ are all equal to 1·0, while the selection bias is zero. Thus the second moments for unbiassed LS regression coefficients are all equal to 1·0. Using the estimated biases in Table 6.8 and their sample standard deviations, only 6 of the 40 estimated second moments exceeds 1·0, while many are less than 0·5. This means that, in terms of the *MSEP*, using LS estimates from an independent data set will usually yield worse predictions than those obtained by using the biased LS estimates from the model selection data set.

Thus subset selection is very much like ridge regression or James and Stein/Sclove regression in trading bias in the parameter estimates for a reduced variance. However, the regression coefficients from subset selection are biased in the direction of being too large, while those from the shrinkage estimators are too small.

How well do we do if we use a subset selection procedure and then apply some form of shrinkage to the subset of selected variables? Table 6.10 shows the *MSEP*'s for shrinkage using the Sclove (1968) estimator and for ridge regression using the Lawless and Wang (1976) shrinkage parameter. The Sclove shrinkage always reduces the *MSEP* by a few per cent, but ridge regression can be disastrous. This is in broad agreement with the findings of others, e.g. Dempster, Schatzoff and Wermuth (1977) and Lawless (1978), for the case in which no selection of variables is made.

There have been few cases in the literature in which shrinkage has been applied after subset selection. The two known to this author

Table 6.10 *MSEP using LS regression coefficients compared with those for shrunken estimates for the selected subset using the Sclove estimator and the Lawless–Wang ridge estimator*

			Large γ			*Small* γ		
k	n	α	*LS*	*Sclove*	*Ridge*	*LS*	*Sclove*	*Ridge*
10	20	0·75	1·004	0·986	0·956	0·924	0·901	0·878
	40		1·011	1·005	1·030	0·925	0·917	0·903
	20	0·45	0·828	0·816	1·150	0·954	0·941	2·206
	40		0·943	0·938	2·510	0·834	0·831	0·857
20	40	0·80	0·906	0·875	0·947	0·742	0·723	0·679
	80		0·965	0·954	1·026	0·942	0·936	0·941
	40	0·55	0·754	0·729	77·2	0·844	0·826	12·4
	80		0·900	0·892	17·2	0·865	0·854	22·1
30	60	0·85	0·905	0·873	0·869	0·833	0·799	0·879
	120		0·901	0·894	0·932	0·899	0·892	0·901
	60	0·70	0·781	0·759	20·7	0·852	0·833	15·2
	120		0·902	0·901	29·0	0·910	0·903	17·4
40	80	0·90	0·891	0·869	0·885	0·822	0·800	0·775
	160		0·921	0·914	0·917	0·892	0·884	0·881
	80	0·75	0·936	0·915	56·9	0·805	0·782	75·3
	160		0·942	0·938	54.3	0·896	0·889	38·1
50	100	0·95	1·033	1·003	0·983	0·826	0·813	0·811
	200		0·978	0·971	0·980	0·909	0·903	0·901
	100	0·80	0·991	0·967	57·9	0·883	0·856	26·8
	200		0·995	0·986	46·3	0·891	0·885	29·8

at the time of writing are Copas (1983) and Hoerl, Schuenemeyer and Hoerl (1986).

Copas was very much aware of the bias in the regression coefficients and regressed independent Y-values against those predicted from subset selection for the given set of values of the predictor variables. It is difficult to compare his results with those here as he was mainly fitting logistic, not linear regressions; estimates are almost invariably biased for models which are nonlinear in the parameters. In most cases, the slope of the regression of independent Y-values against those predicted was much less than 1, being smallest for the smallest subsets.

The paper by Hoerl, Schuenemeyer and Hoerl (1986) uses the ridge trace of Hoerl and Kennard (1970b) to select the subset of variables, and then uses three different estimators of the regression coefficients. Throughout this paper, the emphasis is upon minimizing the mean squared error of the regression coefficients, rather than the *MSEP*. Only one table reports results for the *MSEP*, and this indicates a good performance for the Lawless–Wang (1976) estimator applied after ridge analysis is used to select the subset. The Efroymson stepwise procedure is also compared, using ordinary (and hence biased) estimators of the regression coefficients; it gave quite large *MSEP*'s in some cases, often performing much worse than ordinary LS applied to the full set of predictors.

As explained in section 3.9, ridge regression largely ignores the correlation between the dependent variable and the eigenvectors associated with the smaller principal components. In the simulations performed here, there was no attempt to associate the dependent variable with the larger principal components. Thus these simulations have generated some of the kind of data on which Fearn (1983) warned that ridge regression performs badly. The cases with small values of α had many small eigenvalues, and the performance of ridge regression in these cases was particularly poor. Such cases are probably more typical of the physical sciences where the predictor variables may be similar quantities measured at different times or locations, or the variables may be constructed by taking polynomials, logarithms, cross-products, etc., of a small set of original variables.

6.3 Cross-validation and the *PRESS* statistic

The *PRESS* (prediction sum of squares) statistic is a cross-validation statistic suggested by Allen (1974) for model selection. Cross-

validation consists of setting aside part of the data, usually just one observation at a time, and predicting that data from the remainder.

In calculating the *PRESS* statistic, for a given set of p predictors, each observation, y_i, is predicted from the LS regression equation obtained from the other $(n-1)$ observations. If $\hat{y}_{ip}$ denotes the predicted value for y_i, then the *PRESS* statistic for a particular subset of p predictors is

$$PRESS_p = \sum_{i=1}^{n} (y_i - \hat{y}_{ip})^2. \tag{6.23}$$

The thought of performing n multiple regressions for every subset in which we are interested can be very daunting. Fortunately there is a mathematical result which can reduce the amount of computation very substantially. We use the well-known formula for a rank one update of the inverse of a nonsingular matrix A:

$$(A + \mathbf{x}\mathbf{x}')^{-1} = A^{-1} - A^{-1}\mathbf{x}(1 + \mathbf{x}'A^{-1}\mathbf{x})^{-1}\mathbf{x}'A^{-1} \tag{6.24}$$

where $\mathbf{x}$ is a vector of new 'data'. In our case, the matrix $X_p'X_p$ corresponds to A, where X_p is that part of X corresponding to the p predictors of interest. However we want to remove $\mathbf{x}\mathbf{x}'$, not add it. It is easily shown that the corresponding formula for downdating is

$$(A - \mathbf{x}\mathbf{x}')^{-1} = A^{-1} + A^{-1}\mathbf{x}(1 - \mathbf{x}'A^{-1}\mathbf{x})^{-1}\mathbf{x}'A^{-1}. \tag{6.25}$$

NB Neither (6.24) nor (6.25) should ever be used for computational purposes, though (6.24) is still often used in the Kalman filter. Updating or downdating the Cholesky factorization is far more accurate. Unless A^{-1} itself is required after every update, rather than quantities calculated from it, updating or downdating the Cholesky factorization is often faster also.

If $\mathbf{b}_{ip}$ is the vector of LS regression coefficients based upon the $(n-1)$ observations excluding the ith, then

$$\begin{aligned}\hat{y}_{ip} &= \mathbf{x}_i'\mathbf{b}_{ip} \\ &= \mathbf{x}_i'(A - \mathbf{x}_i\mathbf{x}_i')^{-1}(X_p'Y - \mathbf{x}_i y_i) \\ &= \mathbf{x}_i'(A^{-1} + A^{-1}\mathbf{x}_i d_i^{-1}\mathbf{x}_i'A^{-1})(X_p'\mathbf{Y} - \mathbf{x}_i y_i)\end{aligned}$$

where the scalar $d_i = 1 - \mathbf{x}_i'A^{-1}\mathbf{x}_i$.

If $\mathbf{b}_p$ is used to denote the vector of LS regression coefficients when all n observations are used, i.e.

$$\mathbf{b}_p = A^{-1}X_p'\mathbf{Y},$$

then we find that

$$\begin{aligned} y_i - \hat{y}_{ip} &= (y_i - \mathbf{x}_i'\mathbf{b}_p)/d_i \\ &= e_i/d_i \end{aligned} \tag{6.26}$$

where e_i is the LS residual using all the observations.

The d_i's can easily be calculated from the Cholesky factorization. Thus if

$$\boldsymbol{A} = \boldsymbol{X}_p'\boldsymbol{X}_p = \boldsymbol{R}_p'\boldsymbol{R}_p$$

where $\boldsymbol{R}_p$ is upper triangular, then

$$\begin{aligned} d_i &= 1 - \mathbf{x}_i'R^{-1}(\mathbf{x}_i'R^{-1})' \\ &= 1 - \mathbf{z}_i'\mathbf{z}_i \end{aligned}$$

where $\mathbf{z}_i' = \mathbf{x}_i'\boldsymbol{R}^{-1}$.

Finally we derive

$$PRESS_p = \sum_{i=1}^{n} (e_i/d_i)^2. \tag{6.27}$$

The d_i's above are the diagonal elements of $(\boldsymbol{I} - \boldsymbol{V})$ where $\boldsymbol{V}$ is the projection operator, sometimes also known as the 'hat' matrix:

$$\boldsymbol{V} = \boldsymbol{X}_p(\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p'.$$

It was shown in (6.7) that the sum of the diagonal elements of this matrix equals p. Hence the average value of the d_i's is $(1 - p/n)$. If $n \gg p$, then

$$\begin{aligned} PRESS_p &\approx \sum_{i=1}^{n} e_i^2/(1 - p/n)^2 \\ &= RSS_p/(1 - p/n)^2. \end{aligned}$$

Hence

$$PRESS_p \approx RSS_p \frac{n^2}{(n-p)^2}.$$

After dividing by the sample size, n, this is very similar to formula (6.20) for the (false) estimated *MSEP*, so that we can expect that minimizing the *PRESS* statistic with respect to p will often pick the same size of subset as minimizing the estimated *MSEP*.

Using the *PRESS* statistic is not using true cross-validation as we do not repeat the same procedure on each of the sets of $(n - 1)$

observations as was applied to the full set of n observations. That is, we do not repeat the subset selection procedure each time that a different observation is left out. If we do this, then the selected subsets will not always be the same as that selected using the full data. To illustrate this, let us look at the STEAM and POLLUTE data sets again.

The STEAM data set contained $k = 9$ predictors and $n = 25$ observations. Using the full data set, and exhaustive search as the selection procedure, the estimated *MSEP*'s for the best-fitting subsets of each size are:

No. of predictors	0	1	2	3	4
Estimated *MSEP*	2·765	0·861	0·462	0·418	0·428
No. of predictors	5	6	7	8	9
Estimated *MSEP*	0·448	0·416	0·448	0·492	0·555

Here the number of predictors shown excludes the constant which was fitted in all models.

Notice that the best-fitting subset of three predictors (variables numbered 4, 5, 7) gave almost the same estimated *MSEP* as the best-fitting subset of six predictors (numbers 1, 3, 5, 7, 8, 9).

Omitting one observation at a time and repeating the exercise of finding the best-fitting subset, followed by applying the stopping rule of minimizing the estimated *MSEP* gave the frequencies of selection shown in Table 6.11.

The value of the *PRESS* statistic was 11·09, whereas the true cross-validation sum of squares of prediction errors was 16·41. Dividing by n to obtain the average squared prediction error gives 0·444 for *PRESS*, which is close to the estimated *MSEP* using all the data of 0·416, but much smaller than the value of 0·656 for true cross-validation.

In the case of the POLLUTE data set ($k = 15$, $n = 60$), the estimated *MSEP*'s for the best-fitting subsets of all sizes, using all the data are:

No. of predictors	0	1	2	3	4	5	6	7
Estimated *MSEP*	3934	2385	1844	1577	1373	1332	1295	1298
No. of predictors	8	9	10	11	12	13	14	15
Estimated *MSEP*	1327	1332	1359	1409	1465	1530	1599	1673

Table 6.11

Selected subset (variable nos)	*Frequency*
1 3 5 7 8 9	5
4 5 7	18
1 2 5 7	1
4 5 7 8	1

Table 6.12

Selected subset (variable nos)	*Frequency*
1 2 3 6 9 14	46
1 2 3 5 6 9 14	8
1 2 6 9 14	1
1 3 8 9 10 14	1
1 2 3 8 9 10 14	1
1 2 3 5 6 9 12 13	1
1 2 3 6 8 9 12 13	1
1 2 3 4 5 6 9 12 13	1

The best-fitting subsets of six and seven variables were those numbered (1, 2, 3, 6, 9, 14) and (1, 2, 3, 5, 6, 9, 14).

Omitting one observation at a time, the selected subsets were as shown in Table 6.12.

The value of the *PRESS* statistic was 79 490 compared with 116 673 for the true cross-validation sum of squares of prediction errors.

If the n observations are independently sampled from the same population, then the ith observation is independent of the other $(n-1)$ used to predict it. Hence $(y_i - \hat{y}_{ip})^2$ is an unbiased estimate of the squared error of prediction for the ith case, for all i. Hence the true cross-validation sum of squares, divided by the sample size, gives an unbiased estimate of the *MSEP* for our procedure based upon $(n-1)$ observations. We can expect that the true *MSEP* for n observations will be slightly smaller than this.

Notice, however, that consecutive values of $(y_i - \hat{y}_{ip})^2$, though unbiased, will be correlated with each other.

A forward validation procedure for predicting each y_i using only the previous y_j's has been proposed by Hjorth (1982). This attempts to allow for the change in sample size, and the span of the X-variables, by using generalized cross-validation. Unfortunately this assumes the covariance matrix $\sigma^2(X'_{(i)}X_{(i)})^{-1}$ for the LS regression coefficients, where $X_{(i)}$ is the matrix of values of the selected X-variables using observations $1, 2, \ldots, i-1$. This does not allow for the effect of selection, and as we have seen earlier, will usually overestimate the sizes of the variances.

In this brief section, only 'one out at a time' cross-validation has been considered. For a review of the statistical literature on this subject, see Stone (1978).

Appendix 6A Approximate equivalence of stopping rules

In section 6.2.3, a generalized form of Mallows' C_p:

$$C_p(M) = \frac{RSS_p}{\hat{\sigma}^2} - (n - Mp) \tag{6.22}$$

was introduced. There are many different stopping rules in use, and most of them can be shown to yield a similar stopping point to minimizing $C_p(M)$ for some value of M, usually a value close to 2·0 which is that for Mallows' C_p.

6A.1 F-to-enter

If $C_p(M)$ is a minimum at $p = m$, then $C_{m+1}(M) \geqslant C_m(M)$, where $C_m(M)$ and $C_{m+1}(M)$ are for the best-fitting subsets of m and $(m+1)$ variables which have been found. Substitution from (6.21) and a little rearrangement shows that

$$\frac{RSS_m - RSS_{m+1}}{\hat{\sigma}^2} \leqslant M. \tag{6A.1}$$

The left-hand side of (6A.1) is approximately the F-to-enter statistic. The difference is that in $C_p(M)$, $\hat{\sigma}^2$ is defined as

$$\frac{RSS_k}{n-k},$$

that is, it is the residual variance estimate with all the available

predictors in the model, whereas the denominator of the F-to-enter statistic is

$$\frac{RSS_{m+1}}{n-m-1}.$$

These quantities will usually be very similar provided that $n \gg k$. Hence at the minimum of $C_p(M)$, the F-to-enter statistic will usually be less than M.

If the F-to-enter statistic is used as the stopping rule for progressively increasing subset sizes, stopping as soon as it falls below M, then we could stop earlier than if we use the rule of minimizing $C_p(M)$. This is because, unless the X-predictors are orthogonal, the quantity $C_p(M)$ may have several local minima, and the F-to-enter stopping rule tends to stop at the first. If k is small, say of the order of 10–20, then stopping when the F-to-enter first drops below M will usually find the minimum of $C_p(M)$, but for larger k it will often pick a smaller subset.

6A.2 Adjusted R^2 or Fisher's A-statistic

The adjusted R^2-statistic (adjusted for degrees of freedom) is usually defined as

$$A_p = 1 - (1 - R_p^2)\frac{n-1}{n-p} \tag{6A.2}$$

where

$$R_p^2 = 1 - (RSS_p/RSS_1).$$

This is an appropriate definition when a constant is being fitted. When no constant is being fitted, R_p^2 is often redefined as

$$R_p^2 = 1 - (RSS_p/RSS_0)$$

where RSS_0 means just the total sum of squares of the values of the Y-variable without subtraction of the mean. In this case, it is appropriate to replace the $(n-1)$ in (6A.2) with n.

Maximizing the adjusted R^2-statistic can be shown to be identical to minimizing the quantity

$$s_p^2 = \frac{RSS_p}{(n-p)},$$

with respect to p, whichever of the two definitions for A_p is used. The quantity, i.e. s_p^2, is the residual mean square for the p-variable subset.

A little rearrangement shows that at the minimum of s_p^2 we have

$$\frac{RSS_m - RSS_{m+1}}{RSS_{m+1}/(n-m-1)} \leqslant 1,$$

this is, the F-to-enter statistic has a value not greater than 1. Hence maximizing the adjusted R^2-statistic is approximately equivalent to minimizing $C_p(M)$ with $M = 1$.

6A.3 Akaike's information criterion (AIC)

The AIC (Akaike, 1969) is to minimize $-2(L_p - p)$ where L_p is the log-likelihood of a model with p parameters. In applying it to selecting subsets of regression variables, L_p is the log-likelihood for the best-fitting subset of p predictors, after maximization with respect to the regression coefficients. The AIC has been widely used as a stopping rule in the field of time-series analysis in much the same way that Mallows' C_p is used as the practical solution to exactly the same mathematical problem in selecting subsets of regression variables.

Several modifications of the AIC have been proposed (e.g. Akaike, 1977; Rissanen, 1978; Schwarz, 1978; Hannan and Quinn, 1979). These all have the form $-2(L_p - pf(n))$ where $f(n)$ is some slowly increasing function of the sample size, n, such as $\log_e n$ or $\log_e(\log_e n)$. These modifications have been suggested for the situation in which the subsets of variables of each size have been predetermined, so there is no competition bias for selection. Often the variables in the model will be the consecutive autoregressive terms. This is the situation treated by Kennedy and Bancroft (1971). Asymptotic comparisons of some of these criteria have been made by Stone (1977, 1979), who also slowed the asymptotic equivalence of cross-validation and the AIC. These asymptotic results give no indication of the behaviour for finite sample sizes, and contain no discussion of the bias in parameter estimates resulting from model selection.

The AIC has often been used as the stopping rule for selecting ARIMA (auto-regressive, integrated, moving-average) models where selection in not only between models with different numbers of

parameters but also between many models of the same size. In this situation, there is competition bias, which does not appear to have been considered in the time-series literature.

The AIC, and the various modifications of it, can be applied in situations in which normality is not assumed, in which case the ML fitting procedure may not be equivalent to LS fitting.

In the linear regression situation, the crude log-likelihood, neglecting the issue of model selection, is

$$L_p = -(n/2)\log_e(2\pi\sigma_p^2) - \frac{1}{2\sigma_p^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

$$= -(n/2)\log_e(2\pi\sigma_p^2) - RSS_p/(2\sigma_p^2).$$

As the ML estimate for σ_p^2 is

$$\hat{\sigma}_p^2 = RSS_p/n,$$

where the division is by n not the more usual $(n-p)$, we have that

$$L_p - pf(n) = \text{constant} - (n/2)\log_e(RSS_p) - pf(n).$$

Changing signs and taking exponential we see that minimizing these modified AIC's is equivalent to minimizing

$$RSS_p \exp[(2p/n)f(n)].$$

At the minimum, with respect to p, we have that

$$RSS_m \exp[(2m/n)f(n)] \leqslant RSS_{m+1} \exp[((2m+2)/n)f(n)]$$

or that

$$RSS_m \leqslant RSS_{m+1} \exp[2f(n)/n].$$

A little rearrangement then gives the F-to-enter statistic at the minimum,

$$\frac{RSS_m - RSS_{m+1}}{RSS_{m+1}/(n-m-1)} \leqslant (n-m-1)(e^{(2/n)f(n)} - 1).$$

Provided that $(2/n)f(n)$ is small, the right-hand side above is approximately

$$2\left(1 - \frac{m+1}{n}\right)f(n).$$

Thus using the AIC in its original form, i.e. with $f(n) = 1$, is equivalent to minimizing $C_p(M)$ with M a little less than 2.

If the estimate used for $\hat{\sigma}_p^2$ is

$$\hat{\sigma}_p^2 = RSS_p/(n-p)$$

then maximizing $(L_p - pf(n))$ is equivalent to minimizing

$$RSS_p \exp\left[(2p/n)f(n) - \log_e(n-p) - p/n\right].$$

Continuing as before, we find that at the minimum the F-to-enter statistic for the AIC is not greater than a quantity approximated by

$$2 - \frac{m+1}{n}.$$

Thus minimizing the AIC tends to select slightly larger subsets than minimizing Mallows' C_p.

CHAPTER 7

Conclusions and some recommendations

Let us conclude by posing a number of questions and examining how far they can be answered.

Question 1 How can we test whether there is any relationship between the predictors and the predictand?

This is a frequent question in the social and biological sciences. Data have been collected on say 20 or 50 predictors, and this may have been augmented with constructed variables such as reciprocals, logarithms, squares and interactions of the original variables. An automatic computer package may have selected 5 or 10 of these predictors, and has probably output an R^2 value for the selected subset. Could we have done as well if the predictors had been replaced with random numbers or columns from the telephone directory?

If the package used the Efroymson stepwise algorithm, sometimes simply called stepwise regression, then Table 4.4 or formula (4.1) can be used to test whether the value of R^2 could reasonably have arisen by chance if there is no real relationship between the Y-variable and any of the X-variables. Clearly, the more exhaustive the search procedure used, the higher the R^2 value which can be achieved. References are given in section 4.1 to tables for other search algorithms, though there is scope for the extension of these tables. Some of these tables allow for nonorthogonality of the predictors, others do not. In fact, the degree of correlation among the predictors does not make much difference to the distribution of R^2.

Alternatively, if the number of observations exceeds the number of available predictors, the Spjøtvoll test described in section 4.2 can be used to test whether the selected subset fits significantly better than just a constant.

Question 2 Does one subset fit significantly better than another?

If the number of observations exceeds the number of available predictors, then the Spjøtvoll test described in section 4.2 provides a satisfactory answer to this question. An attractive feature of the Spjøtvoll test is that it does not require the assumption that either model is the true model. The test though is that one model fits better than another over the range of the X-variables in the available data. It is possible to modify the test though to apply for extrapolated X's, though this has not been described in detail here.

The Spjøtvoll test is fairly conservative, that is it tends to say that subsets do not differ significantly unless one is very strikingly better than the other.

In the case in which the number of available predictors equals or exceeds the number of observations, there is no general test available or possible. The situation is akin to that in the analysis of designed experiments when there is no replication. If the experimenter is prepared to take the gamble that high-order interactions can be used as a measure of the residual variation, then an analysis can proceed. Similarly, if the researcher gambles on some variables having no effect, or he thinks that he has a reasonable estimate of residual variation from other data sources, then some kind of risky analysis can proceed. Of course, if the judgement that certain variables have no effect is taken after a preliminary analysis of the data, the resulting estimate of residual variance is liable to be artificially small.

Question 3 How do we find subsets which fit well?

Many automatic procedures have been described in Chapter 3. Exhaustive search, using a branch-and-bound algorithm, for the best-fitting subsets of all sizes is typically feasible if we have not more than about 25 available predictors. It is often sensible to try one of the cheap methods first though, say sequential replacement. This will usually show up the 'dominant' variables and give an idea of the likely size of the final subset. At this stage it is often wise to use some of the standard regression diagnostic tools (see e.g. Belsley, Kuh and Welsch 1980; Gunst and Mason, 1980; or Cook and Weisberg, 1982). These could show up a nonlinear relationship with one of the dominant variables, or outliers, or very influential observations.

If the cheap method has shown say that there is very little reduction in the *RSS* between fitting say eight variables and fitting all of them, then an exhaustive search can be restricted to subsets of eight or fewer variables. Such a search may be feasible when it is not feasible to search for the best-fitting subsets of all sizes.

As a very rough rule, the feasible number of subsets which can be searched is of the order of 10^7. There are $2^{25} = 3{\cdot}3 \times 10^7$ possible subsets out of 25 predictors, including the empty subset and the complete set of 25, so that this is close to the limit of feasibility for subsets of all sizes. If we have say 50 available predictors then an exhaustive search for best-fitting subsets of all sizes will usually not be feasible, but it will be feasible to search for the best-fitting subset of six or fewer variables.

If an exhaustive search is not feasible, then a sequential procedure which adds or removes two variables at a time will sometimes find much better-fitting subsets than one-at-a-time algorithms.

Question 4 How many variables should be included in the final subset, assuming that it is required for prediction?

Before that question can be answered, we need to know what method is to be used to estimate the regression coefficients. If ordinary LS, or one of the robust alternatives, is to be used with no attempt to correct for selection bias, then using all the available predictors will often yield predictions with a smaller *MSEP* than any subset.

If the conditional ML method of section 5.4, or some other method, is used to adjust partially for selection bias, then minimizing the (falsely) estimated *MSEP* given by (6.20) is often a reasonable stopping rule with random predictors, while minimizing Mallows' C_p is the equivalent for fixed predictors. However, it is always possible to construct examples using fixed orthogonal predictors for which any given stopping rule will perform badly. This follows directly from Mallows (1973). In most practical cases, the stopping rule is not critical, provided that there is a correction for selection bias in the regression coefficients. The use of Mallows' C_p, even when the predictors are random, or of Akaike's information criterion, or an F-to-enter of just under 2.0, or of minimizing the *PRESS* statistic, can all be expected to give about the same result as using the true *MSEP*.

If the cost of measuring the variables is an important consideration

then a stopping rule which selects a smaller subset should be used such as using a higher F-to-enter. At the moment there are no accurate formulae for the true *MSEP* after subset selection, using either LS regression coefficients, or bias-corrected coefficients.

Notice that for the purpose of prediction we are looking at F-to-enter's of the order of 1.5–2.0. If the Spjøtvoll test is being used for hypothesis testing, a test at say the 5% level may be equivalent to using an F-to-enter of say 8–15, depending upon the numbers of predictors and observations, and the structure of the sample correlations among the predictors.

All of the above assumes that future predictions will be for X-variables which span the same space as those used for the model selection and calibration. It is extremely hazardous to extrapolate beyond this region. The emphasis throughout this monograph has been upon finding models which fit and describe relationships within the space of the X-predictors. Unless there is established theory to justify a particular form of model, there is no reason for believing it will fit well outside of the calibration region.

Question 5 How should we estimate regression coefficients?

A conditional ML method was described in section 5.4. It uses simulation, is very slow, neglects the bias due to the stopping rule, and appears to over-correct for the bias. However the cost of the computer time is now usually very small compared with the cost of the data collection, and to the value which can be attached to the predictions in some cases. A simple alternative, which has not been investigated here, is the jack-knife suggested near the end of section 5.3, using the square root of the sample size.

The 'off-the-peg' alternatives, such as James–Stein/Sclove shrinkage and ridge regression, are not designed to reduce selection bias. They are intended primarily to reduce the variance of the regression coefficients at the expense of adding a small amount of bias. The variance of LS regression coefficients of best-fitting subsets is often very much smaller than those for models chosen independently of the data, so that subset selection has already done what these shrinkage estimators are designed to do. The simulation results in Table 6.10 show that the use of Sclove shrinkage of the regression coefficients of the selected variables always gave a small improvement in *MSEP* over the use of LS estimates, while the use of

ridge regression can sometimes give a larger improvement and can sometimes be disastrous.

Question 6 Can the use of subset regression techniques for prediction be justified?

There are many practical situations in which the cost of measuring the X-predictors is a major consideration. If there is no cost associated with obtaining future X-predictors, then the use of ridge regression using the Lawless–Wang ridge parameter, or the Sclove estimator, will often be preferable alternatives to subset selection. They have the important advantage that their properties are known. If cost of measurement of the X-predictors is a consideration then the loss due to poor predictions should be traded off against it in deciding how many predictors to use. True cross-validation, as described in section 6.3, can be used to obtain a realistic estimate of the *MSEP* provided that no extrapolation outside of the space of the X-predictors is required.

Question 7 What alternatives are there to subset selection?

In many cases in say the social or biological sciences, relationships between variables are monotonic and a simple linear regression (perhaps after using a transformation such as taking logarithms) is an adequate empirical approximation. In the physical sciences, the shape of the regression curve must often be approximated with more precision. One way to do this is by augmenting the predictor variables with polynomial or cross-product terms. This often gives rise to situations in which the cheap 'one-at-a-time' selection procedures pick poor subsets, while an exhaustive search procedure is not feasible. One alternative for this situation is to use projection pursuit (see e.g. Huber, 1985; Friedman, 1987). There is often some prior knowledge of the system being modelled in the physical sciences which enables say partial differential equations to be formulated which partially describe the system and leave only part of it to be modelled empirically. This will often be preferable to the use of black-box techniques.

References

Abramowitz, M. and Stegun, I. (eds) (1964) *Handbook of Mathematical Functions.* US Govt Printing Office, Washington, DC.

Aitkin, M.A. (1974) Simultaneous inference and the choice of variable subsets in multiple regression. *Technometrics*, **16**, 221–7.

Akaike, H. (1969) Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.*, **21**, 243–7.

Akaike, H. (1977) On entropy maximisation principle, in *Applications of Statistics* (ed. P.R. Krishnaiah), North Holland, Amsterdam, pp. 27–41.

Allen, D.M. (1974) The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16**, 125–7.

Armstrong, R.D., Beck, P.O. and Kung, M.T. (1984) Algorithm 615: the best subset of parameters in least absolute value regression. *ACM Trans. on Math. Software (TOMS)*, **10**, 202–6.

Armstrong, R.D. and Kung, M.T. (1982) An algorithm to select the best subset for a least absolute value regression problem. *TIMS Studies of Management Sci.*, **33**, 931–6.

Atkinson, A.C. (1985) *Plots, Transformations and Regression.* Oxford Univ. Press.

Banachiewicz, T. (1938) Méthode de résolution numérique des équations linéaires, du calcul des déterminants et des inverses et de réduction des formes quadratiques. *Comptes-rendus mensuels des sciences mathématiques et naturelles, Acad. Polonaise des Sci. et des Lettres*, 393–404.

Bancroft, T.A. and Han, C-P. (1977) Inference based on conditional specification: a note and a bibliography. *Internat. Statist. Rev.*, **45**, 117–27.

Barnett, V.D. and Lewis, T. (1978) *Outliers in Statistical Data.* Wiley, New York.

Bartlett, M.S. (1951) An inverse matrix adjustment arising in discriminant analysis. *Ann. Math. Statist.*, **22**, 107–11.

Baskerville, J.C. and Toogood, J.H. (1982) Guided regression modeling for prediction and exploration of structure with many explanatory variables. *Technometrics*, **24**, 9–17.

Beale, E.M.L. (1970) Note on procedures for variable selection in multiple regression. *Technometrics*, **12**, 909–14.

Beale, E.M.L., Kendall, M.G. and Mann, D.W. (1967) The discarding of variables in multivariate analysis. *Biometrika*, **54**, 357–66.

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York.

Bendel, R.B. (1973) Stopping rules in forward stepwise-regression. PhD thesis, Biostatistics Dept, Univ. of California at Los Angeles.

Bendel, R.B. and Afifi, A.A. (1977) Comparison of stopping rules in forward 'stepwise' regression. *J. Amer. Statist. Assoc.*, **72**, 46–53.

Benedetti, J.K. and Brown, M.B. (1978) Strategies for the selection of log-linear models. *Biometrics*, **34**, 680–6.

Berk, K.N. (1978a) Gauss–Jordan v. Choleski, in *Comput. Science and Statist: 11th Annual Symposium on the Interface*. Inst. of Statist., N. Carolina State Univ., pp. 321–4.

Berk, K.N. (1978b) Comparing subset regression procedures. *Technometrics*, **20**, 1–6.

Biondini, R., Simpson, J. and Woodley, W. (1977) Empirical predictors for natural and seeded rainfalls in the Florida Area Cumulus Experiment (FACE), 1970–1975. *J. Appl. Meteor.*, **16**, 585–94.

Borowiak, D. (1981) A procedure for selecting between two regression models. *Commun. in Statist.*, **A10**, 1197–203.

Boyce, D.E., Farhi, A. and Weischedel, R. (1974) *Optimal Subset Selection: Multiple Regression, Interdependence and Optimal Network Algorithms*, Lecture notes in Economics and Mathematical Systems 103. Springer-Verlag, Berlin.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) *Classification and Regression Trees*. Wadsworth, Belmont, C.A.

Brown, M.B. (1976) Screening effects in multidimensional contingency tables. *Appl. Statist.*, **25**, 37–46.

Brown, R.L., Durbin, J. and Evans, J.M. (1975) Techniques for testing the constancy of regression relationships over time. *J. Roy. Statist. Soc., B*, **37**, 149–63.

Butler, R.W. (1982) Bounds on the significance attained by the best-fitting regressor variable. *Appl. Statist.*, **31**, 290–2.

Butler, R.W. (1984) The significance attained by the best-fitting regressor variable. *J. Amer. Statist. Assoc.*, **79**, 341–8.

Chan, T.F. (1982) Algorithm 581: an improved algorithm for computing the singular value decomposition. *ACM Trans. Math. Software (TOMS)*, **8**, 84–8.

Chan, T.F., Golub, G.H. and LeVeque, R.J. (1983) Algorithms for computing the sample variance: analysis and recommendations. *The Amer. Statistician*, **37**, 242–7.

Clarke, M.R.B. (1980) Choice of algorithm for a model-fitting system, in *Compstat 1980*. Physica-Verlag, Vienna, pp. 530–6.

Clarke, M.R.B. (1981) Algorithm AS163: a Givens algorithm for moving from one linear model to another without going back to the data. *Appl. Statist.*, **30**, 198–203

Cook, R.D. and Weisberg, S. (1982) *Residuals and Influence in Regression*. Chapman and Hall, London.

Copas, J. B. (1983) Regression, prediction and shrinkage. *J. Roy. Statist. Soc., B*, **45**, 311–54, incl. discussion.

Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics*. Chapman and Hall, London.

Cox, D.R. and Snell, E.J. (1974) The choice of variables in observational studies. *Appl. Statist.*, **23**, 53–9.

Cox, D.R. and Snell, E.J. (1981) *Applied Statistics: Principles and examples*. Chapman and Hall, London.

Dempster, A.P., Schatzoff, M. and Wermuth, N. (1977) A simulation study of alternatives to ordinary least squares. *J. Amer. Statist. Assoc.*, **72**, 77–106.

Diehr, G. and Hoflin, D.R. (1974) Approximating the distribution of the sample R^2 in best subset regressions. *Technometrics*, **16**, 317–20.

Diggle, P.J. and Gratton, R.J. (1984) Monte Carlo methods of inference for implicit statistical models. *J. Roy. Statist. Soc.*, Series B, **46**, 193–227, incl. discussion.

Dijkstra, D.A. and Veldkamp, J.H. (1988) Data-driven selection of regressors and the bootstrap, in *On Model Uncertainty and its Statistical Implications* (ed. T.K. Dijkstra). Springer-Verlag, Berlin, pp. 1–16.

Dongarra, J.J., Bunch, J.R., Moler, C.B. and Stewart, G.W. (1979) *LINPACK Users Guide*. Soc. for Industrial and Appl. Math., Philadelphia.

Draper, N.R., Guttman, I. and Kanemasu, H. (1971) The distribution of certain regression statistics. *Biometrika*, **58**, 295–8.

Draper, N.R., Guttman, I. and Lapczak, L. (1979) Actual rejection levels in a certain stepwise test. *Commun. in Statist.*, **A8**, 99–105.

Draper, N.R. and Smith, H. (1981) *Applied Regression Analysis*, 2nd edn. Wiley, New York.

Draper, N.R. and van Nostrand, R.C. (1979) Ridge regression and James–Stein estimation: review and comments. *Technometrics*, **21**, 451–66.

Edwards, D. and Havranek, T. (1987) A fast model selection procedure for large families of models. *J. Amer. Statist. Assoc.*, **82**, 205–13.

Efron, B. and Morris, C. (1973) Combining possibly related estimation problems. *J. Roy. Statist. Soc., B*, **35**, 379–421.

Efroymson, M.A. (1960) Multiple regression analysis, in *Mathematical Methods for Digital Computers* (eds A. Ralston and H.S. Wilf). Wiley, New York, pp. 191–203.

Elden, L. (1972) Stepwise regression analysis with orthogonal transformations. Master's thesis, unpubl. report, Mathematics Dept, Linkoping Univ., Sweden.

Everitt, B. (1974) *Cluster Analysis*. Heinemann, London.

Farebrother, R.W. (1974) Algorithm AS79: Gram–Schmidt regression. *Appl. Statist.*, **23**, 470–6.

Farebrother, R.W. (1978) An historical note on recursive residuals. *J. Roy. Statist. Soc., B*, **40**, 373–75.

Farebrother, R.W. (1988) *Linear Least Squares Computations*. Marcel Dekker, New York.

Fearn, T. (1983) A misuse of ridge regression in the calibration of a near infrared reflectance instrument. *App. Statist.*, **32**, 73–9.

Fisher, J.C. (1976) Homicide in Detroit: the role of firearms. *Criminology*, **14**, 387–400.

Forsythe, A.B., Engelman, L., Jennrich, R. and May, P.R.A. (1973) A stopping rule for variable selection in multiple regression. *J. Amer. Statist. Assoc.*, **68**, 75–7.

Forsythe, G.E. and Golub, G.H. (1965) On the stationary values of a second-degree polynomial on the unit sphere. *SIAM J.*, **13**, 1050–68.

Freedman, D.A., Navidi, W. and Peters, S.C. (1988) On the impact of variable selection in fitting regression equations, in *On Model Uncertainty and its Statistical Implications* (ed. T.K. Dijkstra). Springer-Verlag, Berlin, pp. 1–16.

Friedman, J.H. (1987) Exploratory projection pursuit. *J. Amer. Statist. Assoc.*, **82**, 249–66.

Furnival, G.M. (1971) All possible regressions with less computation. *Technometrics*, **13**, 403–8.

Furnival, G.M. and Wilson, R.W. (1974) Regression by leaps and bounds. *Technometrics*, **16**, 499–511.

Gabriel, K.R. and Pun, F.C. (1979) Binary prediction of weather events with several predictors, in *6th Conference on Prob. and Statist. in Atmos. Sci.* Amer. Meteor. Soc., pp. 248–53.

Galpin, J.S. and Hawkins, D.M. (1982) *Selecting a Subset of Regression Variables so as to Maximize the Prediction Accuracy at a Specified Point.* Technical Report, Nat. Res. Inst. for Math. Sciences, CSIR, P.O. Box 395, Pretoria, South Africa.

Galpin, J.S. and Hawkins, D.M. (1986) Selecting a subset of regression variables so as to maximize the prediction accuracy at a specified point. *J. Appl. Statist.*, **13**, 187–98.

Garside, M.J. (1965) The best subset in multiple regression analysis. *Appl. Statist.*, **14**, 196–200.

Garside, M.J. (1971a) Algorithm AS 37: inversion of a symmetric matrix. *Appl. Statist.*, **20**, 111–12.

Garside, M.J. (1971b) Some computational procedures for the best subset problem. *Appl. Statist.*, **20**, 8–15.

Garside, M.J. (1971c) Algorithm AS 38: best subset search. *Appl. Statist.*, **20**, 112–15.

Gentle, J.E. and Hanson, T.A. (1977) Variable selection under L_1, in *Proc. Statist. Comput. Section, Amer. Statist. Assoc.*, pp. 228–30.

Gentle, J.E. and Kennedy, W.J. (1978) Best subsets regression under the minimax criterion, in *Comput. Science and Statist.: 11th Annual Symposium on the Interface.* Inst. of Statist., N. Carolina State Univ, pp. 215–17.

Gentleman, J.F. (1975) Algorithm AS88: generation of all ${}^{N}C_{R}$ combinations by simulating nested Fortran DO-loops. *Appl. Statost.*, **24**, 374–6.

Gentleman, W.M. (1973) Least squares computations by Givens transformations without square roots. *J. Inst. Maths. Applics*, **12**, 329–36.

Gentleman, W.M. (1975) Error analysis of QR decomposition by Givens transformations. *Linear Algebra and its Applics*, **10**, 189–97.

Golub, G.H. (1969) Matrix decomposition and statistical calculations, in *Statistical Computation* (eds R.C. Milton and J.A. Nelder). Academic Press, New York, pp. 365–97.

Golub, G.H. and Styan, G.P.H. (1973) Numerical computations for univariate linear models. *J. Statist. Comput. Simul.*, **2**, 253–74.

Goodman, L.A. (1971) The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics*, **13**, 33–61.

Gray, H.L. and Schucany, W.R. (1972) *The Generalized Jackknife Statistic.* Marcel Dekker, New York.

Grossman, S.I. and Styan, G.P.H. (1972) Optimality properties of Theil's BLUS residuals. *J. Amer. Statist. Assoc.*, **67**, 672–3.

Gunst, R.F. and Mason, R.L. (1980) *Regression Analysis and its Application.* Marcel Dekker, New York.

Hall, P. (1989) On projection pursuit regression. *Ann. Statist.*, **17**, 573–88.

Hammarling, S. (1974) A note on modifications to the Givens plane rotation. *J. Inst. Maths. Applics*, **13**, 215–18.

Hannan, E.J. and Quinn, B.G. (1979) The determination of the order of an autoregression. *J. Roy. Statist. Soc., B*, **41**, 190–5.

Hartigan, J.A. (1975) *Clustering Algorithms.* Wiley, New York.

Hawkins, D.M. (1980) *Identification of Outliers.* Chapman and Hall, London.

Healy, M.J.R. (1968a) Algorithm AS 6: triangular decomposition of a symmetric matrix. *Appl. Statist.*, **17**, 195–7.

Healy, M.J.R. (1968b) Algorithm AS 7: inversion of a positive semi-definite symmetric matrix. *Appl. Statist.*, **17**, 198–9.

Hemmerle, W.J. (1975) An explicit solution for generalized ridge regression. *Technometrics*, **17**, 309–14.

Hemmerle, W.J. and Brantle, T.F. (1978) Explicit and constrained generalized ridge regression. *Technometrics*, **20**, 109–20.

Hjorth, U. (1982) Model selection and forward validation. *Scand. J. Statist.*, **9**, 95–105.

Hocking, R.R. (1976) The analysis and selection of variables in linear regression. *Biometrics*, **32**, 1–49.

Hocking, R.R. and Leslie, R.N. (1967) Selection of the best subset in regression analysis. *Technometrics*, **9**, 531–40.

Hocking, R.R., Speed, F.M. and Lynn, M.J. (1976) A class of biased estimators in linear regression. *Technometrics*, **18**, 425–37.

Hoerl, A.E. and Kennard, R.W. (1970a) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

Hoerl, A.E. and Kennard, R.W. (1970b) Ridge regression: applications to nonorthogonal problems. *Technometrics*, **12**, 69–82.

Hoerl, A.E., Kennard, R.W. and Baldwin, K.F. (1975) Ridge regression: some simulations. *Commun. in Statist.*, 4, 105–23.

Hoerl, A.E., Kennard, R.W. and Hoerl, R.W. (1985) Practical use of ridge regression: a challenge met. *Appl. Statist.*, **34**, 114–20.

Hoerl, R.W., Schuenemeyer, J.H. and Hoerl, A.E. (1986) A simulation of biased estimation and subset selection regression techniques. *Technometrics*, **28**, 369–80.

Huber, P.J. (1985) Projection pursuit. *Ann. Stat.*, **13**, 435–525, incl. discussion.

James, W. and Stein, C. (1961) Estimation with quadratic loss, in *Proc. 4th Berkeley Symposium on Probability and Statistics*, (eds J. Neyman and L. Le Com), pp. 362–79.

Jeffers, J.N.R. (1967) Two case studies in the application of principal component analysis. *Appl. Statist.*, **16**, 225–36.

Jennings, L.S. and Osborne, M.R. (1974) A direct error analysis for least squares. *Numer. Math.*, **22**, 325–32.

Jennrich, R.I. (1977) Stepwise regression, in *Statistical Methods for Digital Computers* (eds K. Enslein, A. Ralston, and H.S. Wilf). Wiley, New York, pp. 58–75.

Jolliffe, I.T. (1982) A note on the use of principal components in regression. *Appl. Statist.*, **31**, 300–3.

Jones, M.C. and Sibson, R. (1987) What is projection pursuit? *J. Roy. Statist. Soc., A*, **150**, 1–36, incl. discussion.

Judge, G.G. and Bock, M.E. (1978) *The Statistical Implications of Pre-test and Stein-rule Estimators in Econometrics*. North-Holland, Amsterdam.

Kahaner, D., Tietjen, G. and Beckmann, R. (1982) Gaussian quadrature formulas for $\int_0^\infty e^{-x^2} g(x)dx$. *J. Statist. Comput. Simul.*, **15**, 155–60.

Kahn, H. (1956) Use of different Monte Carlo sampling techniques, in *Symposium on Monte Carlo Methods* (ed. H.A. Meyer). Wiley, New York, pp. 146–90.

Kailath, T. (1974) A view of three decades of linear filtering theory. *IEEE Trans. on Inf. Theory*, **IT-20**, 145–81.

Kendall, M.G. and Stuart, A. (1961) *The Advanced Theory of Statistics*, Vol. 2. Griffin, London.

Kennedy, W.J. and Bancroft, T.A. (1971) Model building for

prediction in regression based upon repeated significance tests. *Ann. Math. Statist.*, **42**, 1273–84.

Kudo, A. and Tarumi, T. (1974) An algorithm related to all possible regression and discriminant analysis. *J. Japan. Statist. Soc.*, **4**, 47–56.

LaMotte, L.R. and Hocking, R.R. (1970) Computational efficiency in the selection of regression variables. *Technometrics*, **12**, 83–93.

Lawless, J.F. (1978) Ridge and related estimation procedure: theory and practice. *Commun. in Statist.*, **A7**, 139–64.

Lawless, J.F. (1981) Mean squared error properties of generalized ridge estimators. *J. Amer. Statist. Assoc.*, **76**, 462–6.

Lawless, J.F. and Wang, P. (1976) A simulation study of ridge and other regression estimators. *Commun. in Statist.*, **A5**, 307–23.

Lawrence, M.B., Neumann, C.J. and Caso, E.L. (1975) Monte Carlo significance testing as applied to the development of statistical prediction of tropical cyclone motion, in *4th Conf. on Prob. and Statist. in Atmos. Sci.* Amer. Meteor. Soc., pp. 21–4.

Lawson, C.L. and Hanson, R.J. (1974) *Solving Least Squares Problems.* Prentice-Hall, New Jersey.

Lawson, C.L., Hanson, R.J., Kincaid, D.R. and Krogh, F.T. (1979) Basic linear algebra subprograms for FORTRAN usage. *ACM Trans. on Math. Software (TOMS)*, **5**, 308–23.

Lee. T-S. (1987) Algorithm AS 223: optimum ridge parameter selection. *Appl. Statist.*, **36**, 112–18.

Linhart, H. and Zucchini, W. (1986) *Model Selection.* Wiley, New York.

Longley, J.W. (1967) An appraisal of least squares programs for the electronic computer from the point of view of use. *J. Amer. Statist. Assoc.*, **62**, 819–41.

Longley, J.W. (1981) Modified Gram–Schmidt process vs. classical Gram–Schmidt. *Commun. in Statist.*, **B10**, 517–27.

Lovell, M.C. (1983) Data mining. *The Rev. of Econ. and Statist.*, **65**, 1–12.

McCabe, G.P., Jr (1978) Evaluation of regression coefficient estimates using α-acceptability. *Technometrics*, **20**, 131–9.

McDonald, G.C. and Galarneau, D.I. (1975) A Monte Carlo evaluation of some ridge-type estimators. *J. Amer. Statist. Assoc.*, **70**, 407–16.

McDonald, G.C. and Schwing, R.C. (1973) Instabilities of regression estimates relating air pollution to mortality. *Technometrics*, **15**, 463–82.

McKay, R.J. (1979) The adequacy of variable subsets in multivariate regression. *Technometrics*, **21**, 475–9.

Maindonald, J.H. (1984) *Statistical computation.* Wiley, New York.

Mallows, C.L. (1973) Some comments on C_p. *Technometrics*, **15**, 661–75.

Mantel, N. (1970) Why stepdown procedures in variable selection. *Technometrics*, **12**, 621–5.

Mason, R.L. and Gunst, R.F. (1985) Selecting principal components in regression. *Statist. and Prob. Letters*, **3**, 299–301.

Miller, R.G. (1962) Statistical prediction by discriminant analysis, in *Meteor. Monographs.* Amer. Meteor. Soc.

Miller, R.G. (1974) The jackknife – a review. *Biometrika*, **61**, 1–15.

Miller, A.J. (1984) Selection of subsets and regression variables. *J. Roy. Statist. Soc., A*, **147**, 389–425, incl. discussion.

Miller, A.J. (1989) Updating means and variance. *J. Comput. Phys.*, **85**, 500–1.

Morgan, J.A. and Tatar, J.F. (1972) Calculation of the residual sum of squares for all possible regressions. *Technometrics*, **14**, 317–25.

Morrison, D.F. (1967) *Multivariate Statistical Methods.* McGraw-Hill, New York.

Naes, T., Irgens, C., and Martens, H. (1986) Comparison of linear statistical methods for calibration of NIR instruments. *Appl. Statist.*, **35**, 195–206.

Narendra, P.M. and Fukunaga, K. (1977) A branch and bound algorithm for feature subset selection. *IEEE Trans. Comput., C*, **26**, 917–22.

Narula, S.C. and Wellington, J.F. (1977a) Prediction, linear regression and minimum sum of relative errors. *Technometrics*, **19**, 185–90.

Narula, S.C. and Wellington, J.F. (1977b) An algorithm for the minimum sum of weighted absolute errors regression. *Commun. in Statist.*, **B6**, 341–52.

Narula, S.C. and Wellington, J.F. (1979) Selection of variables in linear regression using the sum of weighted absolute errors criterion. *Technometrics*, **21**, 299–306.

Nash, J.C. (1979) *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation.* Adam Hilger (Inst. of Physics), Bristol.

Nelder, J.A. and Mead, R. (1965) A simplex method for function minimization. *Comput. J.*, **7**, 308–13.

Obenchain, R.L. (1975) Ridge analysis following a preliminary test of the shrunken hypothesis. *Technometrics*, **17**, 431–41.

Oliker, V.I. (1978) On the relationship between the sample size and the number of variables in a linear regression model. *Commun. in Statist.*, **A7**, 509–16.

Osborne, M.R. (1976) On the computation of stepwise regressions. *Aust. Comput. J.*, **8**, 61–8.

Piessens, R., de Doncker-Kapenga, E., Uberhuber, C. and Kahaner, D. (1983) *QUADPACK, A Quadrature Subroutine Package*, Series in Computational mathematics, **1**. Springer-Verlag: Berlin.

Plackett, R.L. (1950) Some theorems in least squares. *Biometrika*, **37**, 149–57.

Platt, C.A. (1982) Bootstrap stepwise regression. *Proc. Bus. and Econ. Sect., Amer. Statist. Assoc.*, 586–9.

Pope, P.T. and Webster, J.T. (1972) The use of an F-statistic in stepwise regression procedures. *Technometrics*, **14**, 327–40.

Press, S.J. (1972) *Applied Multivariate Analysis*. Holt, Rinehart and Winston, New York.

Quenouille, M.H. (1956) Notes on bias in estimation. *Biometrika*, **43**, 353–60.

Rencher, A.C. and Pun, F.C. (1980) Inflation of R^2 in best subset regression. *Technometrics*, **22**, 49–53.

Ridout, M.S. (1988) An improved branch and bound algorithm for feature subset selection. *Appl. Statist.*, **37**, 139–47.

Rissanen, J. (1978) Modeling by shortest data description. *Automatica*, **14**, 465–71.

Roodman, G. (1974) A procedure for optimal stepwise MSAE regression analysis. *Operat. Res.*, **22**, 393–9.

Rothman, D. (1968) Letter to the editor. *Technometrics*, **10**, 432.

Rushton, S. (1951) On least squares fitting of orthonormal polynomials using the Choleski method. *J. Roy. Statist. Soc., B*, **13**, 92–9.

Savin, N.E. and White, K.J. (1978) Testing for autocorrelations with missing observations. *Econometrika*, **46**, 59–67.

Schatzoff, M., Tsao, T., and Fienberg, S. (1968) Efficient calculation of all possible regressions. *Technometrics*, **10**, 769–79.

Scheffe, H. (1959) *The Analysis of Variance*. Wiley, New York.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–4.

Sclove, S.L. (1968) Improved estimators for coefficients in linear regression. *J. Amer. Statist. Assoc.*, **63**, 596–606.

Seber, G.A.F. (1977) *Linear Regression Analysis.* Wiley, New York.

Seber, G.A.F. (1984) *Multivariate Observations.* Wiley, New York.

Silvey, S.D. (1975) *Statistical Inference.* Chapman and Hall, London.

Smith, B.T., Boyle, J.M., Dongarra, J.J., Garbow, B.S., Ikebe, Y., Klema, V.C. and Moler, C.B. (1976) *Matrix Eigensystem Routines – EISPACK Guide.* Springer-Verlag, Berlin.

Smith, G. and Campbell, F. (1980) A critique of some ridge regression methods. *J. Amer. Statist. Assoc.*, **75**, 74–103.

Sparks, R.S., Zucchini, W. and Coutsourides, D. (1985) On variable selection in multivariate regression. *Commun. in Statist.*, **A14**, 1569–87.

Spjøtvoll, E. (1972a) Multiple comparison of regression functions. *Ann. Math. Statist.*, **43**, 1076–88.

Spjøtvoll, E. (1972b) A note on a theorem of Forsythe and Golub. *SIAM. J. Appl. Math.*, **23**, 307–11.

Sprevak, D. (1976) Statistical properties of estimates of linear models. *Technometrics*, **18**, 283–9.

Steen, N.M., Byrne, G.D. and Gelbard, E.M. (1969) Gaussian quadratures for the integrals $\int_0^\infty \exp(-x^2)f(x)dx$ and $\int_0^b \exp(-x^2)f(x)dx$. *Math. of Comput.*, **23**, 661–71.

Stein, C. (1960) Multiple regression, in *Contributions to Probability and Statistics* (eds I. Olkin *et al.*). Stanford Univ. Press, Stanford.

Stein, C.M. (1962) Confidence sets for the mean of a multivariate normal distribution. *J. Roy. Statist. Soc., B*, **24**, 265–96.

Stewart, G.W. (1973) *Introduction to Matrix Computations.* Academic Press, New York.

Stewart, G.W. (1977) Perturbation bounds for the QR factorization of a matrix. *SIAM J. Numer. Anal.*, **14**, 509–18.

Stone, M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc., B*, **39**, 44–7.

Stone, M. (1978) Cross-validation: a review. *Math. Operationsforsch., Ser. Statist.*, **9**, 127–39.

Stone, M. (1979) Comments on model selection criteria of Akaike and Schwarz. *J. Roy. Statist. Soc., B*, **41**, 276–8.

Tarone, R.E. (1976) Simultaneous confidence ellipsoids in the general linear model. *Technometrics*, **18**, 85–7.

Thompson, M.L. (1978) Selection of variables in multiple regression: Part I. A review and evaluation. Part II. Chosen procedures, computations and examples. *Internat. Statist. Rev.*, **46**, 1–19 and 129–46.

Vinod, H.D. and Ullah, A. (1981) *Recent Advances in Regression Methods.* Marcel Dekker, New York.

Wallace, T.D. (1977) Pretest estimation in regression: a survey. *Amer. J. Agric. Econ.*, **59**, 431–43.

Wampler, R.H. (1970) A report on the accuracy of some widely used least squares programs. *J. Amer. Statist. Assoc.*, **65**, 549–65.

Wampler, R.H. (1979a) Solutions to weighted least squares problems by modified Gram–Schmidt with iterative refinement. *ACM Trans. on Math. Software (TOMS)*, **5**, 457–65.

Wampler, R.H. (1979b) Algorithm 544. L2A and L2B, weighted least squares solutions by modified Gram–Schmidt with iterative refinement. *ACM Trans. on Math. Software (TOMS)*, **5**, 494–9.

Ward, L.L. (1973) Is uncorrelating the residuals worth it? Master's thesis, Unpubl. MA thesis, Mathematics Dept, McGill Univ., Montreal, Canada.

Weisberg, S. (1980) *Applied Linear Regression.* Wiley, New York.

Wellington, J. F. and Narula, S.C. (1981) Variable selection in multiple linear regression using the minimum sum of weighted absolute errors criterion. *Commun. in Statist.*, **B10**, 641–8.

Wilkinson, J.H. (1965) *The Algebraic Eigenvalue Problem.* Oxford Univ. Press, Oxford.

Wilkinson, L. and Dallal, G.E. (1981) Tests of significance in forward selection regression with an *F*-to-enter stopping rule. *Technometrics*, **23**, 377–80.

Wilkinson, J.H. and Reinsch, C. (1971) *Handbook for Automatic Computations*, Vol. II: *Linear Algebra.* Springer-Verlag, Berlin.

Zirphile, J. (1975) Letter to the editor. *Technometrics*, **17**, 145.

Zurndorfer, E.A. and Glahn, H.R. (1977) Significance testing of regression equations developed by screening regression, in *5th Conf. on Prob. and Statist. in Atmos. Sci.* Amer. Meteor. Soc., pp. 95–100.

Index