Monographs
on Statistics and
Applied Probability 95

Subset Selection in Regression

Second Edition

Alan Miller

CHAPMAN & HALL/CRC

Subset Selection in Regression

Second Edition

MONOGRAPHS ON STATISTICS AND APPLIED PROBABILITY

General Editors

V. Isham, N. Keiding, T. Louis, N. Reid, R. Tibshirani, and H. Tong

1 Stochastic Population Models in Ecology and Epidemiology *M.S. Barlett* (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. Barlett* (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, 2nd edition *J.S. Maritz* (1995)
18 Residuals and Influence in Regression *R.D. Cook and S. Weisberg* (1982)
19 Applications of Queueing Theory, 2nd edition *G.F. Newell* (1982)
20 Risk Theory, 3rd edition *R.E. Beard, T. Pentikäinen 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 Composition Data *J. Aitchison* (1986)
26 Density Estimation for Statistics 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 and K.D. Glazebrook (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. Bandorff-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.W. Ng (1990)

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

38 Cyclic and Computer Generated Designs, 2nd edition
J.A. John and E.R. Williams (1995)

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)

42 Statistical Reasoning with Imprecise Probabilities *P. Walley* (1991)

43 Generalized Additive Models *T.J. Hastie and R.J. Tibshirani* (1990)

44 Inspection Errors for Attributes in Quality Control
N.L. Johnson, S. Kotz and X. Wu (1991)

45 The Analysis of Contingency Tables, 2nd edition *B.S. Everitt* (1992)

46 The Analysis of Quantal Response Data *B.J.T. Morgan* (1992)

47 Longitudinal Data with Serial Correlation—A state-space approach
R.H. Jones (1993)

48 Differential Geometry and Statistics *M.K. Murray and J.W. Rice* (1993)

49 Markov Models and Optimization *M.H.A. Davis* (1993)

50 Networks and Chaos—Statistical and probabilistic aspects
O.E. Barndorff-Nielsen, J.L. Jensen and W.S. Kendall (1993)

51 Number-Theoretic Methods in Statistics *K.-T. Fang and Y. Wang* (1994)

52 Inference and Asymptotics *O.E. Barndorff-Nielsen and D.R. Cox* (1994)

53 Practical Risk Theory for Actuaries
C.D. Daykin, T. Pentikäinen and M. Pesonen (1994)

54 Biplots *J.C. Gower and D.J. Hand* (1996)

55 Predictive Inference—An introduction *S. Geisser* (1993)

56 Model-Free Curve Estimation *M.E. Tarter and M.D. Lock* (1993)

57 An Introduction to the Bootstrap *B. Efron and R.J. Tibshirani* (1993)

58 Nonparametric Regression and Generalized Linear Models
P.J. Green and B.W. Silverman (1994)

59 Multidimensional Scaling *T.F. Cox and M.A.A. Cox* (1994)

60 Kernel Smoothing *M.P. Wand and M.C. Jones* (1995)

61 Statistics for Long Memory Processes *J. Beran* (1995)

62 Nonlinear Models for Repeated Measurement Data
M. Davidian and D.M. Giltinan (1995)

63 Measurement Error in Nonlinear Models
R.J. Carroll, D. Rupert and L.A. Stefanski (1995)

64 Analyzing and Modeling Rank Data *J.J. Marden* (1995)

65 Time Series Models—In econometrics, finance and other fields
D.R. Cox, D.V. Hinkley and O.E. Barndorff-Nielsen (1996)

66 Local Polynomial Modeling and its Applications *J. Fan and I. Gijbels* (1996)

67 Multivariate Dependencies—Models, analysis and interpretation
D.R. Cox and N. Wermuth (1996)

68 Statistical Inference—Based on the likelihood *A. Azzalini* (1996)

69 Bayes and Empirical Bayes Methods for Data Analysis
B.P. Carlin and T.A Louis (1996)

70 Hidden Markov and Other Models for Discrete-Valued Time Series
I.L. Macdonald and W. Zucchini (1997)

71 Statistical Evidence—A likelihood paradigm *R. Royall* (1997)

72 Analysis of Incomplete Multivariate Data *J.L. Schafer* (1997)

73 Multivariate Models and Dependence Concepts *H. Joe* (1997)

74 Theory of Sample Surveys *M.E. Thompson* (1997)

75 Retrial Queues *G. Falin and J.G.C. Templeton* (1997)

76 Theory of Dispersion Models *B. Jørgensen* (1997)

77 Mixed Poisson Processes *J. Grandell* (1997)

78 Variance Components Estimation—Mixed models, methodologies and applications
P.S.R.S. Rao (1997)

79 Bayesian Methods for Finite Population Sampling
G. Meeden and M. Ghosh (1997)

80 Stochastic Geometry—Likelihood and computation
O.E. Barndorff-Nielsen, W.S. Kendall and M.N.M. van Lieshout (1998)

81 Computer-Assisted Analysis of Mixtures and Applications—
Meta-analysis, disease mapping and others *D. Böhning* (1999)

82 Classification, 2nd edition *A.D. Gordon* (1999)

83 Semimartingales and their Statistical Inference *B.L.S. Prakasa Rao* (1999)

84 Statistical Aspects of BSE and vCJD—Models for Epidemics
C.A. Donnelly and N.M. Ferguson (1999)

85 Set-Indexed Martingales *G. Ivanoff and E. Merzbach* (2000)

86 The Theory of the Design of Experiments *D.R. Cox and N. Reid* (2000)

87 Complex Stochastic Systems
O.E. Barndorff-Nielsen, D.R. Cox and C. Klüppelberg (2001)

88 Multidimensional Scaling, 2nd edition *T.F. Cox and M.A.A. Cox* (2001)

89 Algebraic Statistics—Computational Commutative Algebra in Statistics
G. Pistone, E. Riccomagno and H.P. Wynn (2001)

90 Analysis of Time Series Structure—SSA and Related Techniques
N. Golyandina, V. Nekrutkin and A.A. Zhigljavsky (2001)

91 Subjective Probability Models for Lifetimes
Fabio Spizzichino (2001)

92 Empirical Likelihood *Art B. Owen (2001)*

93 Statistics in the 21st Century *Adrian E. Raftery, Martin A. Tanner,
and Martin T. Wells (2001)*

94 Accelerated Life Models: Modeling and Statistical Analysis
Mikhail Nikulin and Vilijandas Bagdonavičius (2001)

95 Subset Selection in Regression, Second Edition
Alan Miller (2002)

Subset Selection in Regression

Second Edition

Alan Miller

CHAPMAN & HALL/CRC

A CRC Press Company
Boca Raton London New York Washington, D.C.

Library of Congress Cataloging-in-Publication Data

Miller, Alan J.
Subset selection in regression / Alan Miller.-- 2nd ed.
p. cm. -- (Monographs on statistics and applied probability ; 95)
Includes bibliographical references and index.
ISBN 1-58488-171-2
1. Regression analysis. 2. Least squares. I. Title. II. Series.

QA278.2 .M56 2002
519.5′36--dc21 2002020214

International Standard Book Number 1-58488-171-2
Library of Congress Card Number 2002020214
Printed in the United States of America 1 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

Contents

Preface to first edition

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 what is known about estimation and presents some new material. 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 disappointed. I hope that it will disturb many readers and awaken them to the dangers of using automatic packages that 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 using 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 that is emphasized in this monograph.

The topic of subset selection in regression is one that 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 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 that are specific to each particular data set and distract attention from the main topic. I am sure that this applies particularly 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 6-month sabbatical period at the University of Waterloo over the northern hemisphere winter of 1979/1980. I am grateful to Jerry Lawless and others at Waterloo for the help and encouragement they gave me. Hari Iyer is to be thanked for organizing a series of lectures 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, Queensland) 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

Preface to second edition

What has happened in this field since the first edition was published in 1990?

The short answer is that there has been very little real progress. The increase in the speed of computers has been used to apply subset selection to an increasing range of models, linear, nonlinear, generalized linear models, to regression methods which are more robust against outliers than least squares, but we still know very little about the properties of the parameters of the best-fitting models chosen by these methods. From time-to-time simulation studies have been published, e.g. Adams (1990), Hurvich and Tsai (1990), and Roecker (1991), which have shown, for instance, that prediction errors using ordinary least squares are far too small, or that nominal 95% confidence regions only include the true parameter values in perhaps 50% of cases.

Perhaps the most active area of development during the 1990s has been into Bayesian methods of model selection. Bayesian methods require the assumption of prior probabilities for either individual models or variables, as well as prior distributions for the parameters, that is, the regression coefficients and the residual variance, in addition to the assumptions required by the frequentist approach; these may be that the model is linear with independent residuals sampled from a normal distribution with constant variance. In return, the Bayesian methods give posterior probabilities for the models and their parameters. Rather than make Chapter 5 (the previous Chapter 6) even larger, a new chapter has been added on Bayesian methods.

There is a major divergence between the Bayesian methods that have been developed and those described in the first edition. Most authors in this field have chosen not to select a single model. The underlying Bayesian structure is that of a mixture of models, perhaps many millions of them, each with an associated probability. The result, the posterior model, is not to take a gamble by selecting just one model involving a subset of variables that by chance has come out best using the data at hand and the procedures which we have chosen to use, but it is a weighted average of all of the models with moderately large weights (posterior probabilities) given to some models and very small weights given to most of them. Most authors have then taken the lesser gamble of using a Bayesian Model Average (BMA) of perhaps the top 5% or 10% of these models.

A disadvantage of using a BMA of, say the top 5% of models, is that if

we started with 50 candidate predictor models, this BMA may still use 40 of them. An objective of subset selection in many applications is to substantially reduce the number of variables that need be measured in the future.

For those Bayesians who are brave enough to grab hold of the single model with the highest posterior, there is very little advice in the literature. The problems are basically the same as with the frequentist approach. Very little work has been done on the properties of the extreme best-fit model. Some new ideas are given in the last section of Chapter 7.

The bootstrap aggregation (bagging) method of Breiman (1996) bears a similarity to Bayesian model averaging. It uses bootstrapping of the original data set. Instead of picking just one model, the models are averaged. It does not use posterior probabilities to weight the different models found. However, if one model is found several times from different replicates, it is given weight according to the number of times that the same model is selected. This method came to the author's attention too late to be covered adequately in this monograph.

There has been little progress on algorithms to search for the globally optimum subsets of regressors. Problems in global optimization are very difficult. There has been substantial progress in recent years in finding global optima of continuous functions. There is an excellent web site maintained by Arnold Neumaier on this subject at the University of Vienna:

`http:\\www.mat.univie.ac.at/~neum/glopt.html`

However, the problem here is one of combinatorial optimization. In this field, the algorithms seem to be for specific problems such as the travelling salesman problem, or packing problems in one, two, or three dimensions, such as the knapsack problem.

In the statistical literature, the new methods have been Monte Carlo methods, particularly those that have been used by people developing software for Bayesian subset selection. So far, these methods have all been local optimization methods, but with a stochastic element, which gives a chance of breaking away and finding a better local optimum. So far, only an exhaustive search incorporating branch-and-bound (also known as leaps-and-bounds) is available, giving a guarantee that the global optimum has been found. This is only feasible for up to about 30 or so predictors. Despite claims to the contrary, the Monte Carlo methods give absolutely no guarantee that the optimum that has been found is any more than a local optimum. The two-at-a-time algorithms from a random start, which are described in Chapter 3, appear to be the best available at the moment.

The flood of theoretical papers on the asymptotic properties of stopping rules continues unabated. Scant recognition is given to this literature in this monograph, even though an issue of *Statistica Sinica* was almost entirely devoted to this subject. See Shao (1997) for a survey of this vast literature.

The problems in this field are essentially small sample problems. In this regard, there have been a number of advances. Several modifications to Mallows'

C_p have been proposed. These are described in Chapter 5. Also a modification to Akaike's AIC by Hurvich and Tsai (1989) seems to give good results. Perhaps the most important development in this area is the Risk Inflation Criterion (RIC) of Foster and George (1994). In those applications in which there are typically large numbers of available predictor variables, such as meteorology and near-infrared spectroscopy, users of subset selection procedures have typically used F-to-enter values of the order of 10 rather than the default value of 4 used in the stepwise regression algorithm of Efroymson or the value of approximately 2 which is implicit in the AIC or Mallows' C_p. The RIC appeals to extreme-value theory to argue that when the number of available predictors, k, increases then the largest value of the F-to-enter statistic from variables that are not in the true model will increase as well. In a way, this is looking at the asymptotics in k rather than those in n.

As far as estimation of regression coefficients is concerned, there has been essentially no progress. Chapter 6 outlines some progress that has been made with a similar problem, that of estimating the mean of the population that yielded the largest sample mean. For instance, suppose we plant 40 different varieties of potatoes, and choose the one which produced the largest yield per hectare in our experiment. It is probably fortuitous that this variety came out best, and it is likely that its true yield is lower than in our experiment, but how much lower? (Note: There is a huge literature in this field on the design of such experiments, and on maximizing the PCS, the probability of correct selection, but very little literature on the estimation problem.) The most important contributions in this field known to the author are those of Venter and Steel (1991), and Cohen and Sackrowitz (1989).

For the moment, the best method of estimation that can be recommended for estimating regression coefficients when the same data are used for both model selection and estimation, is to bootstrap the standardized residuals. The model selection procedure is then applied to the bootstrap replicates. The regression coefficients are estimated for the model that has been chosen with the original data, but the difference is noted between the bootstrap replicates in which the same model is selected, and for all cases. The regression coefficients from the original data set are then adjusted by this difference. This method is described in the bootstrap section in Chapter 5. There is no current theory to back up this method. This is not a satisfactory situation, but nobody seems to have found anything better.

In the mathematical derivations in this edition, there is even more emphasis upon QR-orthogonalization than in the first edition. Most statisticians learn about principal components (PC), which are closely related to the singular value decomposition (SVD) that is much better known to numerical analysts. These methods (PC and SVD) yield orthogonal directions that usually involve all the predictor variables. The orthogonal directions of the QR-orthogonalization involve an increasing number of the predictors. The first column of the $\boldsymbol{Q}$-matrix is in the direction of the first column of the $\boldsymbol{X}$-matrix. The second column is that part of the second column of $\boldsymbol{X}$ that is

orthogonal to the first column, after scaling. The third column is that part of the third column of $\boldsymbol{X}$ that is orthogonal to the first two columns, again after scaling. This orthogonalization makes far more sense in the context of subset selection than the PC/SVD decompositions.

QR-factorizations are not only accurate methods for least-squares calculations, but they also lead to simple statistical interpretations.

An important property of the QR-orthogonalization is that the projections of the dependent variable on these directions all have the same standard deviation, equal to the residual standard deviation for the full model. This means that the projections, which have the same units as the dependent variable whether that be pounds or kilometres per hour or whatever units have a simple interpretation to the layman. In contrast, the analysis of variance, which is often applied to examine the fit of nested models, has units of, say (pounds)^2 or (kph)^2. How often are the units shown in an analysis of variance?

A little time spent understanding these least-squares projections could be very rewarding. Summing squares of projections leads to much simpler mathematical derivations than the grotesque expressions involving the inverses of parts of $\boldsymbol{X'X}$-matrices multiplied by parts of $\boldsymbol{X'y}$ products. On the negative side, the order of the predictor variables is of vital importance, when the predictors are not orthogonal.

The basic planar rotation algorithm is important as the fast way to change from one subset model to another, simply by changing the order of variables and then omitting the later variables. A deficiency of most of the attempts to use Bayesian methods has been that most of them have used very slow methods to change from one model to another. A notable exception in this regard is the paper by Smith and Kohn (1996).

Much of the fear of ill-conditioning or collinearity seems to stem from a lack of knowledge of the methods of least-squares calculation, such as the Householder reduction, or of Gram-Schmidt-Laplace orthogonalization, or of Jacobi-Givens planar rotation algorithms. Those who are tempted to leave out predictors with correlations of, say 0.8 or 0.9, with other predictors, or to average them or use principal components of the moderately highly correlated variables, may want to consider Bayesian Model Averaging as another alternative. Those who are just nervous because of accuracy problems are probably still using methods based upon the normal equations, and should switch to better computational methods. If high correlations worry you, try looking at near-infrared spectroscopy. I had one set of data for which the majority of correlations between predictors exceeded 0.999; many exceeded 0.9999. Correlations less than 0.999 were sometimes a warning of an error in the data.

I have retained my speed timings performed in the 1980s using an old Cromemco computer. Though ordinary personal computers now run these tests perhaps a thousand times faster, it is the relative times that are of importance.

Software in Fortran for some of the selection procedures in this book can be downloaded from my web site at:

`http://www.ozemail.com.au/~milleraj/`

Finally, I would like to acknowledge my gratitude to CSIRO, which has continued to allow me to use its facilities as an Honorary Research Fellow after my retirement in 1994. I would like to thank Ed George for reading the first draft of Chapter 7, even though he was in the process of getting married and moving home at the time. I would also like to thank Mark Steel for his comments on the same chapter; I have used some of them and disagree with others. Rob Tibshirani persuaded me to pay more attention to bootstrapping and the lasso, and I am grateful to him.

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 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 be denoted by X's. It may be necessaryto 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-infrared 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.
4. The fitting of splines or polynomials, often in two or more dimensions, to functions or surfaces.

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 meteorological data, or with 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 case in practice. In many cases, the original set of measured variables will be augmented with other variables constructed from them. Such variables could include the squares of variables, to allow for curvature in the relationship, or simple products of variables, to allow the gradient of Y on one regressor, say X_1, to vary linearly with the value of another variable, say X_2. Usually a quadratic term is only included in the model if the corresponding linear term is also included. Similarly, a product (interaction) of two variables is only included if at least one of the separate variables is included. The computational methods which will be discussed in Chapter 3 for finding best-fitting subsets assume that there are no restrictions such as these for the inclusion or exclusion 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 log-normal 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 focussed 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 and generalized linear models (particularly the fitting of logistic relationships), the complexity is greatly increased. Also, though there are many ways of fitting

regression lines, least squares 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 some work done on multivariate subset selection. The reader is referred to Seber (1984) and Sparks (1985) for an introduction to this subject. For more recent references, see the paper by Brown et al. (2000).

An entirely different form of empirical modelling is that of classification and regression trees (CART). 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 of the data are then split again. 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 overfitting 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, Friedman, Olshen and Stone (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 the predictors can be explored, the techniques known as projection pursuit may be appropriate. 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 , ... , X_k; that is

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

where the residuals, ϵ, have zero mean and are independently sampled from the same distribution which has a finite variance σ^2. The coefficients β_0, β_1, ... , β_k will usually be unknown, so let us estimate them using least squares. The least-squares estimates of the regression coefficients, to be denoted by b's, are given in matrix notation by

$$\boldsymbol{b} \;=\; (\boldsymbol{X'X})^{-1}\boldsymbol{X'y},$$

where

$$\boldsymbol{b'} \;=\; (b_0,\; b_1,\; ...,\; b_k),$$

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

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

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

Then, from standard least-squares theory (see e.g. Seber (1977) page 364), we have that

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

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

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

where the superscript $\boldsymbol{^{-T}}$ denotes the inverse of the transpose, then it follows that

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

Now $\boldsymbol{x'R^{-1}}$ 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 $\boldsymbol{X}$-matrix without the intermediate step of forming the 'sum of squares and products' matrix $\boldsymbol{X'X}$.

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

$$\boldsymbol{X = (X_A,\ X_B)},$$

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

$$\boldsymbol{X_A'X_A = R_A'R_A},$$

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 so that the factorization down to row p and the inverse down to row p are independent of the following rows. The Cholesky factorization of $\boldsymbol{X'X}$ can be shown to exist and to be unique except for signs provided that $\boldsymbol{X'X}$ is a positive-definite matrix.

Then if $\boldsymbol{x_A}$ consists of the first $(p+1)$ elements of $\boldsymbol{x}$ and $\boldsymbol{b_A}$ is the corresponding vector of least-squares regression coefficients for the model with only p variables, we have similarly to (1.2) that

$$\text{var}(\boldsymbol{x_A'b_A}) = \sigma^2 (\boldsymbol{x_A'R_A^{-1}})(\boldsymbol{x_A'R_A^{-1}})' \ ; \tag{1.3}$$

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

$$\text{var}(\boldsymbol{x'b}) \;\geq\; \text{var}(\boldsymbol{x'_A b_A}).$$

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 $\boldsymbol{Y} = 7$ 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

$$\boldsymbol{b_A} \;=\; (\boldsymbol{X'_A X_A})^{-1}\boldsymbol{X'_A y}$$

and hence

$$\begin{aligned} E(\boldsymbol{b_A}) &= (\boldsymbol{X'_A X_A})^{-1}\boldsymbol{X'_A X}\boldsymbol{\beta} \\ &= (\boldsymbol{X'_A X_A})^{-1}\boldsymbol{X'_A}(\boldsymbol{X_A},\ \boldsymbol{X_B})\boldsymbol{\beta} \\ &= (\boldsymbol{X'_A X_A})^{-1}(\boldsymbol{X'_A X_A},\ \boldsymbol{X'_A X_B})\boldsymbol{\beta} \\ &= \boldsymbol{\beta_A} \;+\; (\boldsymbol{X'_A X_A})^{-1}\boldsymbol{X'_A X_B}\boldsymbol{\beta_B}, \end{aligned}$$

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 $\boldsymbol{x}$ is then

$$\begin{aligned} \boldsymbol{x'\beta} - E(\boldsymbol{x'_A b_A}) &= \boldsymbol{x'_A\beta_A} \;+\; \boldsymbol{x'_B\beta_B} \\ &\qquad -\; \boldsymbol{x'_A\beta_A} - \boldsymbol{x'_A}(\boldsymbol{X'_A X_A})^{-1}\boldsymbol{X'_A X_B}\boldsymbol{\beta_B} \\ &= \{\boldsymbol{x'_B} \;-\; \boldsymbol{x'_A}(\boldsymbol{X'_A X_A})^{-1}\boldsymbol{X'_A X_B}\}\boldsymbol{\beta_B} \qquad (1.4) \end{aligned}$$

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 5 because of the very substantial problems of bias when the model has not been selected independently of the data. Note that the addition of extra variables does not generally reduce the bias for every vector $\boldsymbol{x}$. Also, the best subset for prediction is a function of the range of vectors $\boldsymbol{x}$ 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 omission bias, and for this reason, all of Chapter 6 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.

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 neither 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 will be attempted in Chapter 5.

1.3 Alternatives to using subsets

The main reasons for not using all of the available predictor variables are 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, or that we want to reduce the cost of measuring or acquiring the data on many variables in future. Three alternatives that use all of the variables are (i) using 'shrunken' estimators as in ridge regression, (ii) using orthogonal (or nonorthogonal) linear combinations of the predictor variables, and (iii) using Bayesian model averaging.

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

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

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 centered by subtracting the sample average and then scaled so that the diagonal elements of $\boldsymbol{X'X}$ are all equal to one. In this form the $\boldsymbol{X'X}$-matrix is the sample correlation matrix of the original predictor variables. There is a very large literature on