THE SELECTION OF COVARIATES FOR THE RELATIONSHIP BETWEEN BLOOD-LEAD AND ABILITY

Gillian M Raab

Doctor of Philosophy

University of Edinburgh

1989

The field work of the Edinburgh Lead Study, described in chapter 5, was conducted by a research team of which I was a member. The data described in chapter 5 were compiled by the staff of that team, under my direction. All the analyses of these data which are reported in this thesis, however, have been performed by me. Part of the text of chapter 5 is taken from a paper of which I was first author (Raab et al 1989).

I declare that, with the exceptions mentioned above, this thesis has been composed by me, and the work is my own.

Gillian M Raab, August 1989.

ACKNOWLEDGEMENTS

I wish to thank all of my colleagues in the Medical Statistics Unit for their support and encouragement in the writing of this thesis. Edith Stewart, our department secretary, has been a particular support.

I am indebted to everyone who worked on the Edinburgh Lead Study, without whom this work would have been impossible. My particular thanks must go to Mary Fulton, the study director, and to Ruth Hunter and Linda Boyd for their organisation and analysis of the large body of data which was collected.

I am grateful to Lindsay Paterson and Bill Adams who gave generously of their time to check some of the matrix algebra, and to my supervisor Robin Prescott for his advice and encouragement.

My thanks are due to my family. My daughter Anna printed out much of the final version of the text and my son Jonathan typed the contents pages and prodded me into meeting deadlines. Finally, to my husband Charlie, for doing more to help than I could possibly enumerate.

ABSTRACT

This thesis arose from a problem in the analysis of data from the Edinburgh Lead Study. The data were to be used to estimate the influence of children's blood lead levels on their mental abilities, controlling for other factors which might confound this relationship. The other factors were summarised as a set of covariate scores, and the question arose as to which of these scores should be included in a multiple regression whose purpose was to estimate the coefficient of blood-lead. This problem has arisen in other studies of the influence of lead on ability, and a variety of solutions have been implemented. The statistical and epidemiological literature offers little guidance.

The problem is formalised by proposing regression models with various assumptions. Expressions are derived for the mean-square-error of the parameter of special interest (here the blood-lead coefficient) in terms of quantities which can be calculated from the data. Various stepwise procedures are proposed for selecting a sub-set of covariates to include in the regression equation. These include the usual stepwise procedures, as well as new ones based on the various mean-square-error criteria and on changes in the coefficient of interest. These procedures are studied for the data from the Edinburgh Lead Study and evaluated by simulation in different ways.

The potential for variance reduction from sub-models, compared to including all covariates, is a function of the multiple correlation between the variable of special interest and the variables which could be omitted from the model. The results suggest that, unless this correlation exceeds 0.2, inferences should be based on a regression with the full set of covariates. The greatest benefit is obtained from sub-set selection procedures when the multiple correlation is increased as a result of a decrease in the residual degrees of freedom. In these circumstances the multiple correlation will be high, but its value will fall when the usual adjustment for degrees of freedom is applied. The simulation results suggest that sub-set selection will be beneficial when the residual degrees of freedom for the full model are less than three time the number of covariates.

The method which performed best was to select, at each step, the variable which made the largest change in the coefficient of interest. Stopping rules for this criterion are proposed. This method was less prone than the other methods to underestimate the variance of the coefficient of interest, when this is evaluated in the usual way for the final model. But it performed badly and underestimated this variance, for artificial data where the population multiple correlation between the variable of special interest and the covariates was high. This suggests that sub-set selection should not be used when the estimated multiple correlation adjusted for degrees of freedom is high.

These criteria applied to the Lead Study data would suggest that the effect of lead on ability should be assessed by adjusting for all the covariate scores.

Contents

Chapter 11 Simulations for multivariate normal data

Chapter 1

Introduction and overview

1.1 Introduction

It always comes as a surprise to me that, despite the large number of books and papers on applied statistics published every year, practical statistical analysis gives rise to problems for which no solution appears in the literature. The Edinburgh Lead Study has been a good example of this. I have been involved with this study, as a member of the project team, from its first planning through to data collection, validation and final analyses. As well as the applied papers which have presented results of the effects of lead exposure on children, and of the contribution of environmental lead to children's exposures, particular features of the study have contributed to developments in statistical methodology (Raab & Zhou 1987, Paterson & Raab (in preparation)).

The main aim of the Edinburgh Lead Study was to estimate the influence of children's blood-lead on their scores in ability tests. The study was an observational cross-sectional study of children with a restricted age range. Where observational data are used to attempt to draw conclusions about causal mechanisms, the influence of other concomitant variables must be taken into account (see for example Blalock 1964 and Cochran1984). This

gives some protection against drawing false conclusions that are the result of confounding by variables which are not under experimental control. This was reflected in the Lead Study field work, where most of the data collection effort was focussed on obtaining information on potential confounding variables. This emphasis continued during the analysis phase, when the data collected on potential confounders were reduced to a series of 33 scores which could be measured for each child.

A question then arose for which I could find no useful guidance in the statistical literature. It was the following:

> " Which of these scores should be included as covariates when we estimate the influence of blood-lead levels on children's ability and attainment ?".

This was a bigger question than could be tackled within the time available for analysis. All that emerged from my initial literature review was that the variance of the coefficient of interest (here, that of blood-lead) is always reduced when other covariates are excluded, and thus there may well be sub-models for which the coefficient has a lower mean-square-error (MSE). However, some related literature on prediction contained suggestions that improvements from variable exclusion may be less valuable than they appear, when the covariates to exclude are selected with reference to the data. Also, one could not be sure that the confidence intervals which one might calculate after a variable selection procedure would be valid.

In the circumstances, the safest course was to adjust for all the available covariates. The substantive papers on the

effects of lead on children's abilities (Fulton et al 1987, Thomson et al 1989 and Raab et al 1989) all estimate the influence of blood-lead, adjusted for all available covariates. However, the exclusion of certain covariates by stepwise procedures has become so much the norm in the medical and epidemiological literature, that the results for models which excluded certain variables were also presented. It was fortunate that the conclusions drawn from the data about the influence of lead on children's abilities, did not change when the analysis was carried out with a sub-set of the covariates. However, we cannot guarantee that this would be the case for every set of data.

The work I present in this thesis has allowed me to return to this problem. My aim is to provide guidelines for selecting covariates in studies which share the design characteristics of the Edinburgh Lead Study.

The selection of a sub-set of the covariates is not the only method which can be used to obtain improved estimators by reducing the dimensions of the problem. Principal components analysis of the covariates, and ridge regression methods might also be considered. However, sub-set selection methods are by far the most frequently used in practice, and this is the justification for restricting attention to them.

1.2 Computing methods

The Edinburgh Lead Study data were analysed by BMDP (Dixon et al 1985). However for the purposes of testing out procedures, a more flexible package which also provided matrix manipulation was necessary. All the analyses which I present in this thesis, including the simulations in chapters 9, 10 and 11 have been performed using GENSTAT IV (Alvey et al 1980). For graphical presentation the data, some of the GENSTAT results were read into S (Becker & Chambers 1984), and also on some occasions into MINITAB (MINITAB Inc 1986) to allow a quick interactive assessment of the results.

1.3 Overview

Following a review of the statistical and epidemiological literature, my first method of tackling this problem was to devise criteria which could be calculated from the data, and which would assess the mean-square-error (MSE) of estimation for the coefficient of special interest (denoted by β^*) for a sub-model. There were several such criteria which I have denoted by the general term G_p, with various additional sub-scripts, where p refers to the number of covariates in the model. This theory is developed along similar lines to the equivalent theory which has derived quantities to use when selecting sub-sets for prediction (eg C_p, S_p and A_p).

Although the performance of these G_p criteria, as evaluated in subsequent chapters, has not lived up to the expectations which I initially held out for them, they have been invaluable in helping to understand the structure of the problem. In particular, they forced me to draw the important distinction between a random-effects model and a fixed-effects model. This distinction is not important when one is concerned with prediction, where random-effects criteria and fixed-effects criteria perform in a very similar way. However for estimating a single coefficient the distinction is very important. In its simplest terms it determines whether one estimates the residual variance from the full model or from the reduced model. In practical epidemiology a random-effects model will almost always be required.

In chapters 5, 6 and 7 I look closely at the Lead Study data, and at the way in which various sub-set selection procedures operate on them. The selection procedures examined include those which are derived from the G_p procedures, selection by the significance of the covariates in relation to the outcome, and selection by choosing the variables which will have the greatest influence on the estimate of β^*.

Without performing any simulation procedures, a detailed study of how some of these selection procedures operated on the real data pointed to potential problems. In fact, similar problems could be seen to occur for the C_p and S_p criteria, in relation to prediction, which were examined in the context of selecting variables related to outcome.

These problems were confirmed in the simulations presented in chapter 9, where data were generated with a similar structure to the Lead Study data. Selection by some of the G_p criteria gave results which were so close to those for the full model as to be a complete waste of computer time, while others performed worse than the full model. After this chapter only three selection procedures remained worthy of further consideration. They were :

(1) selection by the value of the residual-mean-square (RMS), the method most commonly used in practical epidemiology;

(2) selection by one of the G_p criteria (G'_{rp}) ;

(3) selecting the variable into the model which gives the greatest absolute change in the estimate of β^* ($\Delta(b^*)$).

None of these performed any better than the full model for the complete Lead Study data. However, when they were further evaluated in chapter 10 on similar data with smaller sample sizes, all three gave improvements relative to the full model. The RMS procedure gave greater improvements when used with a nominal significance level Of 0.05, than were obtained when the stepwise procedure was carried further to give models in the region where the minimum of C_p would be found. A stopping criterion (C) for ($\Delta(b^*)$) was considered in terms of the squared change in the estimate b* as a fraction of the variance of b*. A value of 0.1 for C seemed to be suitable. Only the ($\Delta(b^*)$) procedure seemed to be free of problems of under-estimating the variance of b*.

Finally, the three procedures were given the more severe test of selecting variables from multivariate normal data, which

were selected to have awkward properties. The ($\Delta(b^*)$) procedure came out best from this evaluation, but it has problems in some circumstances. In particular it can give poor estimates when the other covariates are strongly related to the variable of special interest (blood-lead in this case), and under-estimates of the variance of b^* can also occur in this case.

My final recommendation is that sub-set selection should not be performed at all, but the full model should be used, unless the residual degrees of freedom in the analysis are less than three times the number of covariates. It is reassuring that the real Lead Study data have more residual degrees of freedom than this, and so our original decision to include all the covariates in the analysis would be vindicated. If sub-set selection is to be done, the best procedure would seem to be ($\Delta(b^*)$), although one cannot always be sure that it will perform well, especially when the true dependence between the variable of special interest and the other covariates is strong. The value of the adjusted multiple correlation between the variable of special interest and all the other covariates can be used to judge when this is the case.

Chapter 2

Selection of covariates : review of epidemiological practice

2.1 Introduction

In recent years there have been many papers in the statistical journals on the criteria for selecting a subset of variables to use in a multiple regression. These will be reviewed in chapter 3. However, none of these have addressed the problem of interest here, namely :- "Which variables should be included in the analysis of observational studies, when one independent variable is the focus of interest?"

As long ago as 1965 Cochrane commented on the statistical contribution to the analysis of observational studies

> "This type of research, dealing with the acquisition of knowledge that may help us to lead happier and more harmonious lives is potentially important, yet I have the impression that it has been somewhat neglected by the statistical profession."

The recent statistical literature does not appear to have done much to remedy this situation.

In contrast, there have been a considerable number of papers in the epidemiological literature on the analysis of observational studies in general, and in particular on the choice of confounding variables. In this chapter I will review the recommendations in

books and articles which discuss the methodology of epidemiological studies, and exemplify how these have influenced the analysis of lead/ability studies.

2.2 Covariates and confounders : definitions and concepts.

The primary goal of epidemiology can be considered to be "to discover manipulable causes" (Weed 1986) and we often try to attain this goal with studies which are observational rather than experimental in design. Thus the adjustment of estimates for other explanatory variables is one of the most important statistical tools for use in this field.

What type of extra variables should be included in an observational study whose primary purpose is to investigate the relationship between a risk factor and an outcome measure ? This question is addressed by many epidemiological texts, and is also discussed by Smith et al (1983) in the context of lead/ability studies. The consensus is that we should be controlling for variables which

(1) are associated with the outcome, and may be a cause of the outcome;

(2) may be associated with the risk in the study population;

(3) should not be a cause of the risk (as this would result in overcontrol).

In what follows I will assume that the variables being discussed are acceptable in terms of this definition. It will be assumed, in this and the following chapters, that we have obtained

error-free measures of all variables which could modify the relationship between the risk factor and the outcome. Thus we are not in the situation of measuring an apparent association between a risk factor and an outcome variable which is the result of an unmeasured confounding variable. In reality, we will never know whether this assumption holds. However, the success of observational studies in identifying risk factors for disease suggests that this assumption may not be unreasonable.

As more is learnt about the determinants of the outcomes of interest (eg children's abilities) the number of such variables which are measured can become large. Also, the availability of computer programs to perform multiple regression analysis and logistic regression analysis relatively easily and cheaply has now made it possible for studies to collect and analyse data with many covariates.

Most reports of such studies use some data-dependent method of selecting a sub-set of the covariates which are included in the final reported analysis. The terminology which I will adopt here is to use the terms "covariate" and "concomitant variable" for variables which are candidates for inclusion in the regression equation. The term "confounder" applies to such a variable whose inclusion in the equation would appreciably alter the estimate of the coefficient of interest. Although this definition is somewhat unsatisfactory, because we have no criterion for what is an appreciable alteration, it is in line with current usage in the epidemiological literature (eg Miettenin & Cook 1981). An

additional uncertainty about this definition is the fact that it does not distinguish between the alteration of the coefficient for the data which have been collected, and the difference between the marginal and partial coefficients in the population from which the subjects in the study can be considered a sample. I will use the latter as a definition of a genuine confounder. Thus all covariates are potential confounders to a greater or lesser extent. The variable selection process can be considered as the process of selecting the most important confounders from the covariates.

How do most studies select the confounders from the covariates? The procedures which are available for selecting regression variables in the most commonly used computer programs dominate the published results. Draper and Smith (1981, chapter 6) give a critical review of these, and emphasise that these procedures can "easily be abused by the amateur statistician" ; it is not clear to me why they do not include the professional in this caution. The main methods which are used to select variables are forwards or backwards selection procedures or stepwise procedures which are a combination of these two (Efroymson 1960). These methods have the advantage of specifying an analysis strategy, although most text books and computer manuals (eg Draper & Smith 1981, Daniel & Wood 1971, Minitab 1986) stress that these methods should not be followed in a totally automatic way, but that attempts should be made at every stage to interpret the coefficients and assess their plausibility. It is unlikely that this suggestion will be of much help in epidemiological studies. The covariates are usually chosen because they are thought likely to influence the outcome, and so it

is a-priori plausible to control for any one of them. Epidemiologists have no difficulty in suggesting mechanisms for observed associations, and explanations of coefficients with a sign opposite from what is expected can often sound convincing. It is unlikely that one would be able to distinguish a covariate where the association was due to the play of chance from one with a genuine association of similar strength.

The evaluation of epidemiological studies is particularly difficult when no analysis strategy has been described. When there are a total of k potential confounders, in addition to the variable of special interest, the results could be reported after controlling for any one member of the 2^k sets of covariates. Thus for 33 covariates we have a choice of more than eight thousand million possible regression models from which the influence of the risk factor on the outcome could be estimated. The possibility of selecting, from among all of those, the one which is in best agreement with the investigators' previous beliefs cannot be discounted unless the analysis policy is fully and objectively described. This point does not seem to have been considered in the epidemiological papers reviewed below.

2.3 Review of recommendations in the epidemiological literature

I will review the papers in this field in chronological order, although they do not represent the development of a single theme. Rather, each one seems to make a set of rules or prescriptions by which to carry out analysis, often justified by a successful

application to one example. Some of these papers use the analysis of categorical data as their examples, but the arguments are sufficiently general for them to apply to continuous data.

In 1974 Fisher and Patil discussed the choice of confounding variables in cross-classified data. They argue that to select confounders one must examine the relationship of each covariate with the outcome, taking into account the influence of all the other covariates. In reply to this Miettenin (1974) argued that their examination of every possible conditional relationship would have "too low a productivity". He suggested that examination of the data should start with an analysis of the simple relationships, and no variable be considered further unless controlling for this variable alone would "indicate confounding". Essentially, Fisher and Patil are arguing for some form of backwards elimination method, whereas Miettenin suggests a forward selection approach, with screening at the first step.

In their short note on significance levels in stepwise regression Kupper, Stewart & Williams (1976) point out that the selection process can invalidate the usual F statistics used in stepwise regression. They propose that significance levels derived from the Bonferroni inequality will provide useful upper bounds for the p-value to attach to the "most significant" regression coefficient. Their discussion applies mostly to exploratory data analysis where there is potential interest in any of the explanatory factors, and hence the control of type I errors is of the greatest importance, if the literature is not to be swamped by false leads.

Their recommendations imply that the nominal p-values to be used in a multiple regression should be much smaller than the conventional 0.05 level when many regressors are being considered.

Quite the opposite advice is given by Dales & Ury (1979) who deal with the case of controlling for covariates. They quote Bancroft (1964) in suggesting that p-values larger than the standard ones of perhaps 0.25-0.5 or even 0.7-0.8 should be used. They argue that the rationale of evaluating a covariate for its confounding potential is quite different from that underlying the usual significance test, and that the question of whether or not the relationship between the outcome and the covariate could have occurred by chance is not directly relevant. Significance tests lay the burden of proof on rejecting the null hypothesis, whereas in assessing confounding the onus should be on showing that the covariate could not possibly distort the relationships being investigated. They suggest a policy of comparing the estimates of the disease/risk factor association with and without control for the covariates. This seems sound advice, although it ignores the influence of variable selection on p-values and, as they comment, "there are no established guidelines or cut-off points for such a selection".

Bancroft (1964) is interesting in its own right. It deals with the general problem of inference procedures which use preliminary tests of significance, and presents simulation results for the comparison of two means with a preliminary test of equality of the variances in the two samples. He presents guidelines for

the circumstances when the use of the preliminary test may give increased power without affecting the nominal significance level of the final test. Bancroft anticipated that the availability of high speed computers would prove helpful in giving guidance for many other such problems, and refers to the selection of covariates as an example. The reality has turned out somewhat differently. High speed computers have increased the number of preliminary significance tests which are carried out, largely ignoring their influence on the significance level of the final test.

Quite another set of criteria for carrying out stepwise regression procedures in epidemiology are put forward by Kleinbaum Kupper & Morgenstern in their text book *Epidemiological Research* (1982); see also Kupper & Hogan (1978). They call their procedure Hierarchical Model Simplification (chapter 21). They suggest various strategies all of which start with a model which consists of all the covariates, the risk factor of special interest and various interactions. They suggest in particular that one should start by testing the significance of the interaction of each covariate with the risk factor of special interest. The analysis then proceeds in a forward and backward stepwise manner, testing higher order interactions with extreme significance levels and removing insignificant terms from the model. There is a suggestion that the main effects of all *important confounders* should never be removed from the model, as this might result in sacrificing validity for precision. They propose that such variables might be removed from the model if "deletion of the main effect does not materially alter the exposure-related coefficients" This is in line with their

recommendation in chapter 13 that "the use of a statistical test is not appropriate to assess confounding". However this point does not seem to have been appreciated by some epidemiologists who cite this book as the source of their modelling strategy (see Schroeder et al 1984 below). Also no guidance is given as to what is an *important confounder* or how one judges when a coefficient has been *materially altered*.

To conclude, there are a wide variety of strategies which have been suggested for choosing a sub-set of covariates in epidemiological studies. The lack of consensus on the best procedure, and the lack of any criterion for deciding when an appreciable bias is being introduced by a confounder, have helped to fill the correspondence columns of epidemiological journals. To give but one example, Mantel (1986) criticised the fact that Rona et al (1985) had not included social class as a covariate in their analysis of the effects of passive smoking on children's growth. The authors of the original study replied as follows:

> "We are well aware that a variable not <u>significantly</u> related to the dependent variable should not be automatically deleted from the model as it may nevertheless affect the relationship between the dependent variable and the independent variable of special interest. However a relationship between a factor, and the independent variable of special interest , in this case smoking, is not in itself a reason for its inclusion in the model, as the factor must also have an association with the outcome." (Rona et al 1986).

They go on to show that the relationship which they have estimated is little influenced by the inclusion of social class.

2.4 The choice of covariates in studies of the effect of lead on childrens' abilities

The literature on the effects of lead on children's mental abilities is a large one and several reviews are available (Rutter (1980), Pocock and Ashby (1985), Lansdown and Yule (1986), Smith (1985), Lester Grant (1985)). I will not attempt to be comprehensive here but will concentrate on a few large studies which have been influential and have used different strategies in the selection of covariates.

Needleman and his colleagues in Boston (1979) presented the first study of the effect of lead on children where data were collected for a substantial number of potential confounding variables. They present an analysis of covariance which compares 58 children with high-tooth lead levels with 100 with low tooth-lead levels. A total of 39 covariates were identified, but only four variables were controlled for in the analysis of covariance. The criterion used to select covariates was a difference at $p<0.1$ between the high and low lead groups, but one variable (parental IQ) was controlled in the analysis which did not differ at this level between groups. No information is presented about the relationship between the outcome measure (child's IQ) and the covariates. This study has been the subject of much criticism, particularly with respect to the manner in which the 158 children were selected for analysis from a much larger original group (EPA 1985). The answers which have been given to such criticisms (Needleman 1983) have not always reassured us that the investigators were aware of the biases

which can arise in the conduct of epidemiological studies, or that they had any coherent analysis policy to guide their analysis.

Largely as a result of these criticisms Smith et al (1983) replicated the Boston study on a UK population with a larger sample size. A total of 403 children were selected from a much larger group who donated teeth, in three groups, high lead, low lead and a sample from the centre of the tooth lead distribution. A detailed parental interview collected a large amount of data on family background which was condensed into a set of scores for concomitant variables. The scores were derived by selecting the items in the interview which showed a significant relationship to IQ, and then by grouping similar items together. The analysis of covariance was then carried out with the "application of stepwise procedures both with lead level and with outcome variables", although no details are given of the strategy employed. In the final analysis there was adjustment for between five and seven covariates, depending on the outcome which was being studied.

In a further analysis of the same data (Pocock et al 1986) the covariate scores and a different analysis policy are described in detail. Covariates were selected from the 17 available by a forward stepwise procedure to identify an 'optimal' regression model defined in terms of Mallows C_p criterion (see chapter 3). This examines the relationships between the covariates and the outcome variables to minimise the expected prediction error, and resulted in ten covariates being chosen. The risk factor (tooth lead) was entered into the regression after this stepwise procedure. Although

this method has the advantage of being well specified, and of including certain terms which are related to the outcome at less than conventional significance levels, it takes no account of the relationships between the covariates and the lead values, which are crucial in assessing the degree of confounding. The substantial conclusions of the original paper were not changed by the reanalysis.

The American groups who have reported results of lead studies in recent years have been much influenced by Kleinbaum, Kupper and Morgenstern's text book (1981). Two papers, in particular, give details of the analysis strategy employed. Schroeder et al (1984) analysed data from 104 children and eight concomitant variables. Following the text book's rules they start with a model containing all the covariates and various interaction terms and proceed to delete all but one covariate by backward's elimination. In another study Bellinger et al (1984) analysed the results from 216 infants who were selected in three groups of high, medium and low cord-blood lead. They collected information on 120 potential confounding variables. Their analysis policy is complex, and is described in great detail. Briefly, they started with forward stepwise procedures to identify the best predictors of outcome (mental development) without including the exposure (lead levels). Extreme p-values were used to allow for variable selection. The lead variable was then included in the equation and other covariates removed from the model if their exclusion "did not substantially alter the magnitude or precision of the blood lead coefficient". This procedure resulted in the selection of only two covariates from

the original 120. The authors point out that the extent to which a variable is a confounder cannot be judged by the significance of the association between the confounder and the lead level, illustrating that one of their two final choices resulted in an alteration in the lead coefficient although its relationship with lead levels had a p-value of only 0.13. The results of this study were unusual in that the significance of the lead coefficient was enhanced, rather than diminished, by the inclusion of the covariates, and in doing so it passed through the conventional "5% level".

2.5 Summary of variable selection for lead studies.

To summarise, a variety of different strategies have been used in the selection of covariates in lead exposure studies. These include those which examine the relationships only between the covariates and lead, or only between the covariates and outcome. Where more complex strategies which look at both these relationships have been carried out, they seem to result in control for only a very few covariates from a much larger set. The studies discussed above are summarised in table 2.1.

Table 2.1

STUDY	No of subjects N	Total covariates k-2[π]	Covariates used p-2[π]	Method of selection
Needleman et al 1979	158	39	4	Relationship of covariates to lead.
Smith et al 1983	403	17	5	No adequate details
Reanalysis of Smith et al 1986	377	17	10	Forward stepwise with C_p
Schroeder et a 19841	104	8	1	Hierarchical model simplification
Bellinger et al 1984	216	120	2	Hierarchical model simplification

[π] As k and p stand for the total number of variables in the full and reduced models and the lead exposure variable and a constant term are always included, so the additional covariates in the two models are k-2 and p-2.

2.6 Questions to be answered

Any approach which examines covariates to see if they may be confounders must surely attempt to look at both the relationship between the covariates and the outcome, and the relationship between the covariates and the variable of special interest. However none of the proposed strategies seems to have any rationale, nor has there been satisfactory consideration of the problem of selection bias from the large number of sets of potential confounders.

The crucial questions in relation to the selection of covariates in observational studies are

(1) How does the selection process affect the bias and precision of the coefficient for the influence of the risk factor on outcome?

(2) Given an answer to (1) what selection process, if any, should we chose ?

(3) Can we find a method of estimating the bias and precision of the risk/outcome coefficient after a variable selection process, and thus derive a valid confidence interval for the coefficient of special interest?

These questions are of a statistical nature, and their answer presupposes a formal approach to defining models and pocedures. The chapters which follow will develop this approach.

Chapter 3

Statistical approaches to variable selection

3.1 Introduction

Most of the statistical literature on the topic of variable selection in regression is concerned with the use of a regression relationship for prediction. Although this is not the problem here, yet the results are relevant and will be discussed later in the chapter. The only papers which consider the effect of variable selection on the estimate of a single regression parameter are those which deal with clinical trials. I will discuss them first, but initially I will outline the various statistical models which have been proposed, the assumptions made and the notation to be used.

3.2 Types of model considered; fixed or random covariates

In all cases it will be assumed that the outcome variable is a random variable y which has been observed for n individuals, giving an n-vector Y of independent observations. Lower case will be used, throughout this thesis, for random variables and upper case will be used for their realisations and also for fixed quantities. Greek letters will be used for unknown parameters and the corresponding Roman letters for their estimates. The expectation of y for fixed values of a k-vector of covariates X is given by

$$E(y) = X\beta,$$

where β is a k-vector of parameters. The first elements (β^* and X^*) of β and X correspond to the coefficient and value of the variable of special interest. We observe the random variable y for the fixed values of the covariates in the n rows of the n x k matrix X. The residuals $y - X\beta$ are assumed to have a distribution which is independent of X with variance σ^2. Notice that we are assuming that all relevant covariates have been measured, and that all are free from measurement error. It is possible that some elements of β may be zero, but this cannot be known a priori.

The model will be termed a "fixed-effects" model when all the results are conditional on the observed values of X. This is the model which is appropriate to industrial experiments when the Xs are chosen as fixed design points. When the model is extended to consider some or all of the X's as the realisations of random variables x, we will be dealing with a random-effects model. Notice that results from the fixed-effect model will apply to the random-effects model conditionally on the particular X's observed. The random-effects model often makes the assumption that the x variables follow a multivariate normal distribution. If the distribution of the residuals is normal then the joint distribution of x and y is also multivariate normal.

Which model is more appropriate for the analysis of covariance in epidemiological studies? Clearly the covariates are not fixed in the same sense as they are in experimental studies, and thus a random-effects model would seem more appropriate. An exception might be X^* in the case when two groups are compared, one containing

the risk factor and the other free from it. Even when this is not the case, studies are often designed to include a spread of values of the dependent variable of special interest and one might argue for always treating it as a fixed-effect. A model with X* treated as fixed and the other covariates x as random variables will be termed a random model with fixed X*. This is the model which has been used when considering the selection of covariates in clinical trials. In this situation we have the additional assumption that the expected value of x is the same across treatment groups, or more generally, that the distribution of the random covariates is not dependent on the value of the fixed covariate X*. Without this assumption it would appear to be a suitable model for epidemiological studies. Multivariate normality for the other covariates is unlikely to be found in practice either for observational studies or for clinical trials.

For most of the results which follow it will not be necessary to assume that the residuals of y for fixed Xs follow a normal distribution. However, this assumption will be necessary for such things as the calculation of confidence intervals and significance tests. When the residual degrees of freedom are large we would expect these tests to be robust to modest departures from normality.

Obviously, the condition of normally distributed residuals will be fulfilled for multivariate normal data. When we are dealing with a random-effects model for which some of the xs are not normally distributed (eg if they are categorical variables) and if the condition of normality of the residuals holds for the full model

which contains all the covariates, it is unlikely that it would hold for sub-models with one or more of the non-normal x variables omitted. An exception to this would be the case when the non-normal covariates have distributions which are independent of y.

3.3 Selection for a random-effects model with fixed X*.

The first discussion of selection of covariates for this model appears in a paper by Cochran (1965) which deals specifically with the problems of observational studies. He makes the assumption of equal population means of x in exposed and unexposed populations, which is really only appropriate for randomised clinical trials, but which one might hope to achieve by a suitable design in an observational study. He calculates the coverage of the usual confidence interval for the unadjusted estimate of β^*, conditional on the t-statistic for estimating the difference between exposure groups on the value of a single covariate x. On the basis of these results he suggests one should consider using the adjusted estimate when the t statistic is greater than about 1.5, and that this will be especially beneficial when the correlation between y and x is high. However he does not consider the implications for such a policy on the significance tests and confidence intervals for β^*.

This problem has been considered more recently, in the context of clinical trials, by papers which have investigated various strategies by simulations. Forsythe (1977) considered the case of a single covariate, using a simulation of a clinical trial with two groups of 16 patients each. Values of x and y were generated from

the bivariate normal distribution with a range of correlations such that ρ^2 ranged from 0 to 0.75. Six strategies for variable selection were considered.

(1) Always adjust for the covariate;
(2) Never adjust for the covariate;
(3) Only adjust if x is correlated with $p<0.05$ with y;
(4) Adjust if the estimate of β^* is more significant after adjustment;
(5) Adjust if the means of x are significantly different ($p<0.05$) in the two treatment groups (ie X is correlated with X*);
(6) Both of conditions (3) and (5) are met.

Forsythe investigated the size of tests , with nominal values of 0.05, for assessing treatment-effects after using these strategies. Strategies (1) and (2) , as one would predict, did not influence the size. Strategy (3) resulted in a slightly increased size the middle range of ρ^2, whereas strategies (5) and (6) had sizes which were less than the nominal values , especially at large ρ^2 (0.03 for a nominal p-value of 0.05 when ρ^2 was 0.75). As expected, strategy (4) resulted in the nominal p-values being too extreme, sometimes considerably so. Forsythe also estimates the power of the various strategies, showing that all of the policies of adjustment have advantages when ρ^2 is large. However, the power comparisons made between methods are difficult to evaluate because of the different sizes of the tests.

Shirley and Newnham (1984) report a somewhat similar simulation study in the context of a toxicological experiment. They do not appear to have been aware of Forsythe's work. They consider strategies of type (3) with p-values ranging from 0.05 to 0.25. The simulated data were based on a real toxicological experiment where

the outcomes were organ weights, and the possibility of controlling for the x variable, body weight, was considered. Data were simulated for two treatment groups with 6 animals per group. The covariate was simulated with the correlation with outcome found in the data, and also with no correlation with the outcome.

They concluded that, for the case when x and y are correlated, there will be no gain in power for the adaptive strategies over always adjusting for the covariate, if the nominal significance levels for the test of differences between groups (which can be quite misleading) are adjusted to correspond to the true size of the test. When there was no correlation between x and y the adaptive procedures gave a modest improvement in power of about 6-9% compared to always adjusting.

Forsythe's work was extended to multivariate covariates by Schluchter and Forsythe (1985). They considered the case k=5 with two treatment groups, and evaluated various strategies including always adjusting for the covariates, never adjusting, and a total of 16 strategies for selection which can be considered in three groups:

(1) Select covariates correlated with y;
(2) Select covariates whose means differ across groups;
(3) Both of conditions (1) and (2).

All these adaptive methods are considered with the preliminary tests carried out at the 0.05 and 0.25 level. Within group (1) there are four methods: testing the significance of the joint relationship between all the 5 covariates and y; testing the marginal relationship between each column of x and y; testing the partial

relationship after controlling for all the other members of x; and a forward stepwise regression of y on x. In all cases the X to Y relationship is examined within the treatment groups, ie after adjusting for X*.

Simulated data were generated with (y, x) having the same multivariate normal distribution in each treatment group. The design was a factorial one with respect to the following parameters:

Sample size (8, 15 or 32 per group);
Common correlation among the x (0 or 0.9);
Magnitude of the multiple correlation coefficient R^2 between x and y (.1, .4 or .7);
Pattern of distribution of R^2 between the covariates (one only or all x's equally).

The correct type I error rates for the adaptive strategies were different from their nominal 5% level.

Type (1) strategies

For those strategies which selected on the basis of the relationship between X and Y, the correct type I error rates were greater than their nominal value of 5%, the-effect being greatest for the stepwise and partial correlation methods. The size of this-effect depends strongly on the sample size (greatest in small samples) and on the significance level of the intermediate tests (greatest for 0.25). These results can be explained by the underestimation of the residual variance in variable selection procedures (Berk 1978, Rencher and Pun 1980 and Pinault 1988) which is at its most severe when several variables are competing for selection and the ratio k/(n-k-1) is large. The-effect can be a large one giving true significance levels of 10% and greater, but the authors suggest it can be safely ignored when the ratio of k/(n-k-1) is less than 0.1.

Type (2) strategies

The strategies which selected on the relationship between X and X* had true significance levels which were less than 5%. This-effect was independent of sample size but dependent on the value of R^2, and the level of the significance test. The average type I errors (nominal significance level 5%) for a preliminary test at $p<0.25$ were 4.4%, 3.7% , 2.3% and for $p<0.05$ 4.6%, 4.1%, 3.1% at R^2 values of .1,.4 and .7 respectively. This-effect can be understood as follows. When x and y are correlated, the occasions when y differs by treatment group will also tend to be those where x will differ by treatment group. Adjustment for the covariates will reduce the number of such occurrences which appear to show a significant-effect of y on treatment.

Type (3) strategies

The true significance levels of these strategies were intermediate between those for (1) and (2). However they were more often biassed to small significance levels (as in type (2)), especially as the sample size increased.

The power of the procedures are compared by calculating the confidence intervals for the resultant estimates of β^*. The exact variances of never adjusting for covariates or always adjusting for covariates are calculated, and the authors show that adjustment is beneficial when $R^2 > k/(n-3)$. This was the case for the simulation data when R^2 was .4 and .7, with the benefit being greatest for the largest sample size. The only conditions for which the adaptive methods performed better than the best of the other two was when the R^2 was concentrated on a single covariate and methods were of type (1). The stepwise method seemed to perform best among methods of type (1) in this case. When R^2 was diffuse methods of type (1) performed similarly or somewhat worse than always adjusting. Methods of types (2) and (3) had variances which lay closer to the

unadjusted estimates than to the adjusted ones, and so could result in large losses when R^2 was large.

The authors conclude that the safest strategy is to adjust for all covariates when the sample size is large in relation to the number of covariates, since significance tests are valid and the efficiency is not much impaired. They also suggest that a stepwise procedure may be of benefit, but that it should only be used if some method such as the bootstrap (Efron 1979) or the jacknife (Miller 1974) is used to obtain a valid test for the treatment-effect after this procedure.

The most important result to follow from this study is that, even when no bias is introduced by failing to adjust for a covariate (a condition which will not usually be met in epidemiology), the significance test for the variable of special interest may be invalidated by the selection process. It seems likely that the test may be conservative when the distribution of y is highly dependent on x, but x and x* are independent, and the selection procedure requires that x be related to x* before an adjustment is made. Conversely, when the selection process is on the basis of the relationship between x and y the significance levels may be too extreme, especially when $k/(n-k-2)$ is large, say, greater than .1.

In epidemiological studies the omission of any of the covariates may introduce a bias, which will be in addition to the problems of the invalid significance tests discussed above. The sample sizes for which these results have been obtained are

generally smaller than those used in epidemiological studies. But to ensure the avoidance of bias the only safe strategy would seem to be to adjust for all of the covariates. However, it is easy to prove (see next chapter) that if we were in the favourable position of having perfect knowledge of the parameters the mean-square-error of an estimate of β^* based on a reduced model can often be less than that for the full model, even when the estimate based on the reduced model is biassed. This is the motivation for seeking a mean-square-error criterion which has more justification than the somewhat ad-hoc rules suggested by Schluchter and Forsythe (1985). Similar criteria have been derived for the prediction of future values of y from a regression equation, and I will review these below.

3.4 Mean-square-error criteria for prediction

The various criteria which have been suggested for minimising prediction mean-square-error are conveniently reviewed by Thompson (1978a and b). The two which she recommends are each derived from the same principle of finding an expression for the mean-square-error of prediction averaged over a set of x variables with the same dispersion as the predictor set, and then replacing the parameters in this expression with terms derived from the data which have the same expectations.

For the fixed-effects model the appropriate criterion is C_p (Mallows 1973). For a sub-model which contains p covariates, the total mean-square-error of estimation calculated over the current sample can be shown to be

$$\Sigma\ MSE_{pred} = \Sigma_Y(bias)^2 + p\ \sigma^2 + n\sigma^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.1)$$

where $\Sigma_Y(bias)^2$ is the sum of the square of the estimated bias in prediction, over the whole sample, caused by omitting covariates other than the p. Now the expected value for the residual sum of squares from this model with the p covariates is just

$$E(RSS_p) = \Sigma_Y(bias)^2 + (n-p)\ \sigma^2.$$

Hence the quantity

$$RSS_p + (2p - n)\sigma^2 + n\sigma^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.2)$$

will have expectation equal to 3.1. By replacing σ^2 by the estimate (s^2) from the full model we get a quantity which can be estimated from the data, and which has the same expectation as (3.1) and (3.2). If we ignore the last term (which is common to all models) and standardise by dividing by s^2, we get the criterion

$$C_p = RSS_p\ /\ s^2 - n + 2p.$$

For the random-effects model with x and y following a jointly normal distribution the equivalent criterion is S_p. A procedure equivalent to minimising S_p was suggested by Narula (1974) but is introduced by Hocking (1976) and Thompson (1978) without reference to Narula's paper. For this model the x variables are assumed to follow a k-1 dimensional normal distribution (one of the k covariates being a vector of 1s corresponding to the grand mean).

Conditional on the observed values (X_p) of p of the covariates, y will have a normal distribution with its mean at the predicted value of y for a given X_p (expressed in terms of the partial regression coefficients β_p) and with variance σ_p^2 (the variance of the partial conditional distribution).

Predicting a new value of y from X_p will have a mean-square-error of prediction

$$\sigma_p^2/n \{1 + n + T^2/(n-1)) \quad \ldots\ldots\ldots\ldots\ldots (3.3)$$

where T^2 has a non-central Hotelling's T^2 distribution, over the population of all possible regression samples, with p-1 and n-1 degrees of freedom, and non-centrality parameter

$$\lambda = n\ (X_p - \mu_p)' \Sigma_p^{-1} (X_p - \mu_p)$$

where μ_p and Σ_p are the mean and covariance matrix of the p-1 variables included in the model. The expected value of 3.3 over all regression samples is thus

$$\sigma_p^2/n \{1 + n + (p-1 + \lambda)/(n-p-1)).$$

Now, considering the expectation of this expression over future values of X_p, since λ is just n times a quantity with a χ^2 distribution with p-1 degrees of freedom, we can obtain the expected value of the mean-square-error of prediction from the reduced model as

$$\sigma_p^2 \quad \{ (n+1)(n-2) \} / \{ n(n-p-1) \} \ldots\ldots\ldots (3.4),$$

a result reported by Kerridge (1967).

Ignoring the terms which contain only n, and estimating σ_p^2 from the residual sum of squares from the model containing the p terms we obtain the criterion

$$S_p = RSS_p / \{(n-p-1)(n-p)\}.$$

NOTE; The derivations of this criterion by Thompson and Hocking each contain algebraic errors, which are confusing, although in neither case is the final value for S_p incorrect.

Both of these criteria are quantities which can be calculated from the data for any specified subset of the covariates. The various subsets can then be compared, and those with low values of the criterion indicate a low prediction mean-square-error for that predictor set. In addition subsets with values of C_p close to p are considered as giving evidence that the remaining variables contribute only noise to the prediction. For the full model the value of C_p is exactly p.

Various authors have considered the asymptotic properties of these criteria. In order for the asymptotics to make sense, the number of parameters (k) must tend to infinity as the sample size tends to infinity. If this were not the case, the model which includes all the covariates would always be preferred because the variance part of the criteria tends to zero as n goes to infinity,

while the sum of the biases remains finite. However, if the number of potential covariates increases with sample size this no longer holds.

Brieman & Freedman (1983) show that the optimal number of regressors to minimise the mean-square-error of prediction is a small fraction of the number of data points (ie for optimal prediction $p/n \to 0$ when n and p each tend to infinity). They show that for multivariate normal data the S_p criterion provides an asymptotically optimal rule for selecting regressors. It is easy to see that under these asymptotics the criteria S_p and C_p are equivalent, and are also equivalent to the final prediction error criterion of Akaike (1970) and his information criterion (Akaike 1974). Shibata (1981) derives yet another criterion which he also calls S_p but which in our notation is

$$RSS_p \ (n+2p) \ / \ n,$$

and shows that the selection of covariates which minimises this criterion will attain the lower bound for the mean-square-error of prediction. His derivation does not require the assumption of multivariate normality, and his criterion is asymptotically equivalent to the other four mentioned above.

What is the relevance of these results to epidemiological studies? Both n and p can be large in epidemiological studies, and the concept of p and n increasing together is not unreasonable if studies for outcomes with a large number of potential predictors are designed to be correspondingly large. For such large studies, we would expect to obtain optimal prediction from a fraction of the

covariates, and we would expect either C_p or S_p to select a similar set of covariates. That this will be approximately true when $p<<n$ is clear from the definition of the two criteria. Obtaining the minimum value for either criterion is equivalent to

$$\frac{d}{dp}(RSS_p) = -2p\sigma^2,$$

since σ^2 and σ_p^2 are equivalent to the first order in p/n. In my experience of other studies I have found that the number of predictors which give a minimum of the C_p criterion is the number included in a forward stepwise regression which stops when the F statistic to include a further variable in the model no longer exceeds 2 (F-to-enter set to 2). This is clearly related to the results above, although the two criteria (minimum C_p, and an F-to-enter of 2) are only strictly equivalent when the model with the larger number of covariates is the full model. This will be illustrated for the lead study data in the chapter 6.

3.5 Criticism of regression procedures which select x variables

Multiple regression analysis has become one of the most widely used of statistical techniques and is available as part of almost all statistical packages. Most computer packages also include stepwise regression algorithms, and the additional feature of an "optimum regression" routine which will select the subset of a given size with the smallest residual sums of squares is considered to be

worth advertising. A large body of work (eg Hocking & Leslie 1967, Furnival & Wilson 1974) has been devoted to devising search procedures which will select such subsets with the minimum amount of computation. It is now possible, with the computing facilities generally available, to obtain the regression which will minimise a criterion such as C_p or S_p for up to about 25 predictors ie selecting from 2^{25} (over 30 million) possible subsets. It has been estimated (Copas in discussion of Miller 1984) that over 10^6 regressions are carried out per day worldwide, many of which involve subset selection. However, statisticians have now come to realise that the apparent benefits which may be obtained from subset selection may be illusory.

The arguments are well expressed in the paper by Miller (1984) and in the subsequent discussion. The properties of least-squares regression, which are used to derive results for hypothesis tests and for the properties of criterion functions, are only valid when the subsets being compared are specified in advance without reference to the data. For example, when the same data are used to select a subset with a MSE criterion for prediction, as are used to make the prediction, the estimate s^2 will be an underestimate of σ^2 and the apparent benefit which one can obtain from a reduced set of predictors will be greater than is really the case. Miller (1984) illustrates this for a small simulation based on real data with n=31 and k=14. Copas (1983) considers the case of orthogonal predictors with an n of 50 and 5 covariates. He shows that using a variable selection procedure is often worse than using all the covariates, and can even be worse than using no covariates when there is

competition for selection between the regressors. These results are related to the problem of the inflated significance levels for selection procedures discussed by Schluchter and Forsythe (1985).

Miller distinguishes three possible sources of bias in estimating a least-squares coefficient by sub-set selection:

(1) omission bias;

(2) competition bias, in choosing between subsets of the same size;

(3) stopping-rule bias, in choosing the number of predictors to use.

In practice, the predictions and estimated coefficients are likely to be derived from the same data as are used for subset selection. One could seldom justify collecting one set of data to determine which sub-set to use, and then a completely new one to estimate the coefficients. Thus, Mallow's defence of his criterion (discussion of Miller (1984), p418) that he had not made any claims for it in these circumstances, has a rather hollow ring. . His C_p criterion has become enormously popular because of its inclusion in so many regression packages and the associated graphical methods (C_p against p plots) encourage its use.

Any methods which are suggested in this thesis for selecting subsets of variables in epidemiological studies will have to be evaluated in terms of their real application. Thus, I will hope to discover any biases of the types (1) to (3) above before suggesting

that a subset selection method may be useful in estimating the coefficient of the variable of special interest.

Chapter 4

Mean square error criteria for estimation of β^*

4.1 Introduction

The mean-square-error (MSE) criteria, G_p, which are derived here, relate to the MSE of the estimate of the coefficient, β^*, of the variable of special interest. There are two possible criteria. The first (G_{FP}) is derived from the fixed-effects model, and the second (G_{RP}) from the random-effects model with fixed X*. These correspond to the two criteria C_p and S_p for prediction. Each of the two criteria (G_{RP} and G_{FP}) consist of a sum of two terms, a variance term, and a term for the squared bias which can be evaluated separately.

For the random-effects model it is necessary to condition on the observed values of the covariates included in the model. If we make the assumption of multivariate normality for y and the x variables in the model (except X*), it is possible to divide the variance part of G_{RP} into two parts. The expectation of one of these two parts over the distribution of the xs can be evaluated directly. Unfortunately, the need to evaluate the second term prevents us from using this result to derive a further MSE criterion. However, it allows us to break down the variance part of G_{RP} into two parts, each of which has a clear interpretation.

The derivations are similar to those for C_p and S_p outlined in the previous chapter. The quantities G_p have the property that, for any selected subset of the data, their expectation (under the appropriate model) will be equal to the MSE of the estimate b* of the coefficient β^* which is of special interest. Again the subscript p refers to a model which contains p covariates. As the covariates must always include a constant term and X*, the minimum possible value of p is 2.

4.2 Fixed effects model

The first regression model to be considered is that for which all the independent variables, including the variable of special interest, are considered as fixed effects. It is the model for which the C_p criterion was derived. It is not the ideal model for the consideration of epidemiological studies (see chapter 3), but it has the advantage that it does not require any distributional assumptions for the independent variables.

Let the matrix of observations X be reordered and partitioned into two matrices [P:Q] where P includes X* and the p-1 other covariates included in the regression model, and where Q contains the k-p covariates which are omitted. The vector of coefficients is partitioned conformably into β_P and β_Q. It is also convenient for the special variable X* to occupy the first column of P. Now we can estimate β^* from the reduced model and obtain an estimate (b^*_P) from the first element of $(P'P)^{-1}P'Y$ with variance

from the corresponding element of $\sigma^2(P'P)^{-1}$. Notice that it is the residual variance from the full model which enters into the variance estimate here, which is a consequence of modelling Xs as fixed effects. The estimate of the coefficient of interest will always have a smaller variance than the estimate of β^* from the full model. This result is well known and discussed in various papers (Walls & Weeks 1969, Rao 1971, Narula & Ramburg 1972, Rosenberg & Levy 1972 and Hocking 1974). However, the following derivation, using results from least-squares theory, helps to clarify when a reduction in variance will be expected.

The sum-of-squares matrix from which the variance of b^*_P is derived can be written as

$$\begin{bmatrix} X^{*\prime}X^* & X^{*\prime}P'' \\ P''^{\prime}X^* & P''^{\prime}P'' \end{bmatrix},$$

where P" is the matrix P without the first column which contains X*. The first element of the inverse of this matrix is just

$$[X^{*\prime}[1-P''[P''^{\prime}P'']^{-1}P''^{\prime}]X^*]^{-1}$$

This is the inverse of the residual sum of squares of X* from the least-squares fit of X* on P". Thus the inclusion of extra variables (the matrix Q in our example) can only decrease this sum of squares, and hence increase the variance of b^*_P. This shows that the variables which will have the worst effect on the variance of b^*_P are those which are the best predictors of X*.

Of course, the estimate based on the reduced model will be biassed and the magnitude of this bias will be the first element of the vector $(P'P)^{-1}P'Q\beta_Q$. Various conditions can lead to a selection of the matrix Q which will introduce zero bias. If the matrices P and Q are orthogonal, then there will be no bias even for finite β_Q. However, this situation would confer no benefit in terms of reduced variance, because X* and Q would be orthogonal. If all the coefficients β_Q are zero then there will be no bias, even when P and Q are not orthogonal. It is this situation which would appear to confer the greatest advantage for improved estimation of β^*. So the variables which we might seek to exclude from the regression are those which are related to the exposure, but which do not act as predictors of the outcome.

Now for any partition of the X matrix we can compute a MSE matrix for the estimated coefficients as

$$\sigma^2(P'P)^{-1} + (P'P)^{-1}P'Q\,\beta_Q\beta_Q'Q'P(P'P)^{-1} \quad \ldots\ldots\ldots\ldots (4.1),$$

and we are interested in the first element of this matrix which corresponds to the MSE for our variable of special interest. To obtain an estimate of this quantity from the data, we need an estimate of σ^2 and of the $(k-p)x(k-p)$ matrix $\beta_Q\beta_Q'$ for the true regression coefficients of Q in the full model. We can obtain an unbiassed estimate (s^2) of σ^2 from the residual sum of squares after fitting the full model, and we can find an estimate of $\beta_Q\beta_Q'$ from the estimate of β_Q for the full model.

The estimate of β_Q from the full model is obtained as

$$\begin{bmatrix} (P'P) & (P'Q) \\ (Q'P) & (Q'Q) \end{bmatrix}^{-1} \; [\; P : Q \;]' \; Y$$

and the inverse matrix becomes

$$\begin{bmatrix} [P'P-P'Q(Q'Q)^{-1}Q'P]^{-1} & -(P'P)^{-1}P'Q[Q'Q-Q'P(P'P)^{-1}P'Q]^{-1} \\ -[Q'Q-Q'P(P'P)^{-1}P'Q]^{-1}Q'P(P'P)^{-1} & [Q'Q-Q'P(P'P)^{-1}P'Q]^{-1} \end{bmatrix}$$

Thus the estimate becomes

$$b_Q = [Q'(1-P(P'P)^{-1}P')Q]^{-1}[-Q'P(P'P)^{-1}P' + Q']Y$$

$$= [Q'(1-P(P'P)^{-1}P')Q]^{-1}Q'(1-P(P'P)^{-1}P')Y$$

with variance matrix $[Q'(1-P(P'P)^{-1}P')Q]^{-1}\sigma^2$. Because this is an unbiassed estimate of β_Q the expectation of $b_Q b_Q'$ is given by

$$\beta_Q\beta_Q' + [Q'(1-P(P'P)^{-1}P')Q]^{-1}\sigma^2, \text{ and}$$

an unbiassed estimator of $\beta_Q\beta_Q'$ is

$$b_Q b_Q' - [Q'(1-P(P'P)^{-1}P')Q]^{-1}s^2 \ldots\ldots\ldots\ldots (4.2)$$

When this is substituted into 4.1 and σ^2 is estimated by s^2 we obtain an unbiassed estimator of the MSE matrix for the reduced model which is

$$(P'P)^{-1}s^2$$

$$+ (P'P)^{-1}P'Q[b_Q b_Q']Q'P(P'P)^{-1}$$

$$- (P'P)^{-1}P'Q[Q'(1-P(P'P)^{-1}P')Q]^{-1}Q'P(P'P)^{-1}s^2 \quad \ldots\ldots\ldots (4.3)$$

and we can pick out the first element of this which corresponds to β^*, which will become the MSE criterion G_{FP}.

This is not as bad as it looks. For G_{FP} the first term is the estimate of variance of b^*_P from the reduced model, which uses the estimate of σ^2 from the full model.

For the full model we can write

$$\begin{bmatrix} P'P & P'Q \\ Q'P & Q'Q \end{bmatrix} \begin{bmatrix} b_P \\ b_Q \end{bmatrix} = [\, P : Q \,]' \; Y$$

which gives $(P'P)\, b_P + P'Q\, b_Q = P'Y$

and thus $\quad b_P - (P'P)^{-1} P'Y = -(P'P)^{-1} P'Q\, b_Q.$

The left hand side of this equation is just the difference between the estimates of β_P from the full model and from the reduced model. Thus the second term of the first element of 4.3 is just the square of the difference between the estimates of β^* from the full model and from the reduced model.

A matrix equality which is easily derived from the expression for the inverse matrix given above permits us to write the third term of 4.3 as

$$\{ (P'P)^{-1} - [P'P - P'Q(Q'Q)^{-1}Q'P]^{-1} \} s^2$$

and so we see that the first element of the third term of 4.3 is the difference between the estimate of variance of β^* from the reduced model (our first term) and its estimated variance from the full model. This term is negative semi-definite because the estimated variance from the reduced model cannot be greater than that from the full model.

If we carry out regression calculations on the reduced model and on the full model, and obtain the following statistics for the parameter of special interest β^* calculated in the usual way from each regression equation, as if it were the correct one

	Reduced model	Full model
estimate of β^*	b^*_P	b^*_{full}
estimated variance of b*	v^*_P	v^*_{full}
residual sum of squares	RSS_P	RSS_{full}

then the estimate of MSE for the parameter of special interest becomes

$$G_{FP} = (b^*_{full} - b^*_P)^2 + 2v^*_P \{RSS_{full}/(n-k))/\{RSS_P/(n-p)) - v^*_{full}$$

The multiplying factor in the second term is required because the regression output for the reduced model will calculate the estimated variance of b^*_P from RSS_P, rather than from RSS_{full} as we require. This quantity has the property that it will give an unbiassed estimate of the MSE of b* for any specified submodel defined by the matrices P and Q. Note that for the full model this expression reduces to the estimated variance of b*.

We can only be assured that the properties above hold if the sub-set P has been selected without reference to the data (see chapter 3 for a discussion of this with respect to prediction MSE criteria). However, keeping this caution in mind, the value of G_{FP} can be calculated for any model being considered, and the model with the smallest value chosen. Forward or backward stepwise procedures could be designed with this criterion used to include or exclude variables. These strategies will be described for the lead study data in chapter 7, and evaluated by simulation in chapter 9.

4.3 Relationship of G_{FP} to other criteria

The G_{FP} criterion can be considered as a special case of the prediction MSE criterion of Allen (1971). This criterion, A_P, refers to the prediction of a future value of y at one particular point in the X space. This criterion has been discussed recently by Galpin and Hawkins (1986), who suggest various search procedures based on it and on related criteria. The MSE of b* is equivalent to the prediction of the fitted value of the outcome at a point where

all the covariates (including the grand mean) are zero except for x^* which has the value 1.

The G_{FP} criterion is also equivalent to a criterion discussed by Schluchter (1985) for the choice of covariates in a clinical trial, and Schluchter's criterion is a special case of a criterion for the selection of covariates in the analysis of covariance proposed by Linhart & Zucchini (1982). Schluchter's derivation refers to the case when β^* is a treatment effect which is a one/zero variable defined by the treatment allocation. Although his expression for the criterion is expressed in terms of the within-treatment-group variance-covariance matrices, it is equivalent to G_{FP} for this case. He suggests various stepwise procedures for choosing between models. These involve considering the model with one additional covariate as though it were the full model.

Related work, taking a hypothesis-testing approach, has been carried out by Toro-Vizcarrando and Wallace (1968) and Wallace and Toro-Vizcarrando (1969). They consider the more general issue of using a constraint on the X variables in a multiple regression, and comparing the MSE matrix for the restricted model with the variance covariance matrix for the full model. If the difference between these two matrices is positive definite (MSE matrix larger) then the full model should be used, because it will give a lower variance for any linear combination of the βs. The condition that this matrix is positive definite can be shown to be equivalent to the F-statistic for testing the full model relative to the restricted model

exceeding a certain value. Under the null hypothesis that the differences between the matrices is not positive definite, this F-ratio will have the non-central F distribution, with non-centrality parameter 1 (or 1/2 in the non-standard parameterisation used by Toro-Vizcarrando and Wallace). This distribution (which they tabulate) can be used to test whether there is evidence in favour of using the more complex (full) model for estimating the βs.

A similar approach can be developed for the estimation of only one coefficient. The difference in MSE matrices between b_P and the full model becomes, from 4.1

$$\{[P'(1-Q(Q'Q)^{-1})Q')P]^{-1}-(P'P)^{-1})\sigma^2 - (P'P)^{-1}P'Q\beta_Q\beta_Q'P(P'P)^{-1}\ldots (4.4)$$

The element corresponding to β* of the final term is the square of the expected bias in β* when estimating from the reduced model. The first two terms become, from the matrix equality used in the previous section

$$\{(P'P)^{-1}P'Q[Q'Q-Q'P(P'P)^{-1}P'Q]^{-1}Q'P(P'P)^{-1})\sigma^2 \ldots\ldots\ldots\ldots\ldots\ldots (4.5)$$

which is the variance of $(P'P)^{-1}P'Qb_Q$, whose first element is estimated bias of b_P*. The ratio, F, of the square of the estimated bias to its variance can be written as

$$F = \frac{\text{1st element } [\ (P'P)^{-1}P'Qb_Qb_Q'Q'P(P'P)^{-1}\ ]}{\text{1st element } [\ \{(P'P)^{-1}P'Q[Q'Q-Q'P(P'P)^{-1}P'Q]^{-1}Q'P(P'P)^{-1})s^2\ ]}.$$

Under the null hypothesis that 4.4 is zero F will have the non-central F distribution with 1 and n-k degrees of freedom and non-centrality parameter 1, which has been tabulated by Toro-Vizcarrando and Wallace(1969). The distribution of this statistic can be used to test departures from the null hypothesis which suggest that the more complex model is better.

The G_{FP} criterion can be written as

$$G_{FP} = v^*_{full} + (\text{est. bias})^2 - 2\ \text{var}(\text{est. bias}).$$

Thus a reduced model is to be preferred over the full model, in terms of the G_{FP} criterion, whenever $F > 2$. The 5% , 10% and 25% points of the non-central F distribution are never lower than 6.97, 5.20 and 2.79. Thus using the F statistic as suggested by Toro-Vizcarrando and Wallace will select simpler models than using G_{FP}. This is because it requires that the more complex model must not only give an improvement in the MSE criterion, but also that we must have evidence that this improvement is more than a chance effect.

Yet another procedure might be possible. One might require that the statistic F showed evidence that the omission of the covariates Q introduced a bias into the estimate of β^*. This would imply testing F against the hypothesis that the bias is zero, ie referring it to the central F distribution. This would favour more complex models than the above use of the F statistic. The limiting values for the percentage points of F'_m for large m and p-values of 5%, 10% and 25% are just 3.84, 2.69 and 1.32. The G_{FP} criterion is

equivalent, for large residual degrees of freedom, to a test for significant bias in β^*_Q at a p-value of 0.1584.

In the special case when the two models being considered differ by only one covariate, the vector b_Q becomes a scalar, and the ratio F becomes identical to the F test for the introduction of the additional covariate. In this case the criterion for the minimum G_{FP}, used to compare each model with the one with a single additional covariate (as if this were the correct model), becomes identical to the choice of the model which gives the minimum C_p (derived from the assumption that the full model is correct) from all the models with just one additional covariate.

4.4 Random-effects model with fixed X*

In order to develop this model, it is convenient to modify the notation introduced in section 3.2. The regression model, conditional on fixed values of X has the same form as before, with a total of k covariates which include the constant term and X*. If we write the regression model, conditional on fixed X as

$$E(y \mid X) = \beta_o + X^* \beta^* + X \beta \quad \ldots\ldots\ldots (4.6)$$

then the matrix X becomes the (k-2) dimensioned matrix of regressors other than the constant and X*. Now, when we consider the Xs as realisations of a random variable, x, we can derive (4.6) from the following assumptions:

Let $y = \gamma_o + X^* \gamma^* + \varepsilon_y$

and $x = \delta_o + X^* \delta^* + \varepsilon_x$ ..... (4.7)

where $(\varepsilon_y, \varepsilon_x)$ follow a (k-1) dimensional joint distribution with mean zero and variance covariance matrix which is independent of X*, and this distribution is such that $E(\varepsilon_y \mid \varepsilon_x = e_x) = e_x \beta$. Substituting this into equation 4.7 we obtain

$$E(y \mid X) = \gamma_o + X^* \gamma^* + (X - \delta_o - \delta^* X^*) \beta$$

$$= (\gamma_o - \beta\delta_o) + X^* (\gamma^* - \beta\delta^*) + X \beta \text{ ..... } (4.8),$$

which is of the same form as we require for 4.6 to be satisfied. We also require that the joint distribution of ε_x and ε_y is such that the conditional distribution of ε_y given e_x is the same for all e_x and hence for all X, and in particular, has a variance (σ^2) which is independent of X. When considering sub-models we also require that this condition is fulfilled for y and for the subset of x included in the model. In this case the appropriate variance of the conditional distribution will depend on which xs are included, and will be denoted by σ^2_p. These conditions are fulfilled when ε_x and ε_y follow a multivariate normal distribution.

Now if we estimate β^* from a sub-model of the x matrix, then our estimate will be biassed by the contribution from the omitted regressors to the second term in equation 4.8. Conditional on the p-2 columns of X included in the model, the variance of the estimate b^*_p of β^* will be the first element of $(P'P)^{-1} \sigma_p^2$, where the matrix

P is exactly as referred to in the sections above, and where σ_p^2 is the variance of y conditional on only those covariates which are included in the model. The MSE of b^*_p, conditional on the covariates in the model becomes

$$(\text{bias})^2 + \text{1st element of } \{ (P'P)^{-1} \} \sigma^2_p \quad \ldots\ldots\ldots (4.9);$$

and, estimating σ^2_p by the residual variance from the submodel, s^2_p, and estimating the squared bias from the expression 4.2 , we obtain a quantity whose expectation is 4.9 from the first element of

$$(P'P)^{-1}s_p^2$$
$$+ (P'P)^{-1}P'Q[b_Q b_Q']Q'P(P'P)^{-1}$$
$$- (P'P)^{-1}P'Q[Q'(1-P(P'P)^{-1}P')Q]^{-1}Q'P(P'P)^{-1}s^2.$$

We can write this as

$$G_{Rp} =$$
$$(b^*_{full}-b^*_p)^2 + v^*_p [1 + (RSS_{full}/(n-k))/(RSS_p/(n-p))] - v^*_{full},$$

using the same notation as in section 4.2. This criterion is very similar to G_{Fp}, differing only in that the estimate of variance from the sub-model uses the residual-mean-square from that model rather than from the full model.

4.5 Multi-variate normal x variables

Where x and y follow a multivariate normal distribution, we can avoid conditioning on the Xs included in the model for the variance part of the mean-square-error criterion. The squared bias still requires us to condition on all the observed Xs, and thus to estimate the squared bias from 4.2.

If the submatrix of X to be included in the model is X_p (dimension p-2) we can calculate the sample sums of squares and products matrix of X^* and X_p, about their means, as

$$\begin{bmatrix} S_{**} & S_{*p} \\ S_{p*} & S_{pp} \end{bmatrix} \quad \ldots\ldots\ldots\ldots\ldots\ldots (4.10)$$

where $S_{**} = \Sigma\ (X^* - \bar{X}^*)^2$ is a scalar, and S_{pp}, S_{*p} are calculated similarly as the vector and matrix of sums of squares and cross products for X_p and X^*. All summations are over the n observations.

Writing the inverse of this matrix as

$$\begin{bmatrix} A_{**} & A_{*p} \\ A_{p*} & A_{pp} \end{bmatrix} \quad \ldots\ldots\ldots\ldots\ldots\ldots (4.11),$$

the estimated variance of b^*_p, conditional on X_p is $A_{**}\ \sigma_p^2$. Using the fact that 4.10 and 4.11 are inverses, and that the first element of each is a scalar, we can write

$$A_{**} = S_{**}^{-1}\ (\ 1 + S_{**}^{-1}\ (S_{*p}\ A_{pp}\ S_{p*})\ ) \ldots\ldots\ldots (4.12)$$

This quantity is identical to the first element of $(P'P)^{-1}$ in the notation used above.

Now we can write $S_{p*} = (X_p - \bar{X}_p)\ (X^* - \bar{X}^*)$

or equivalently $S_{p*} = (X_p)\ (X^* - \bar{X}^*)$.

This can be considered as a realisation of a vector of fixed linear combinations of the random variables x_p. The mean of S_{p*} will be $(\delta^*_p X^* - \delta^*_p \bar{X}^*)\ (X^* - \bar{X}^*)$, which is just $S_{**}\delta^*_p$, where the vector δ^*_p is the rearrangement of the vector δ^* to correspond to x_p. Now the variance-covariance matrix Σ_{pp} of ε_p and hence of x_p, conditional on the fixed values of X^*, is estimated by $A_{pp}^{-1}/(n-2)$. The variance-covariance matrix of $\Sigma\ (x_p)\ (X^* - \bar{X}^*)$ is $S_{**}\ \Sigma_{pp}$, estimated by $\{S_{**}\ A_{pp}^{-1}/(n-2)\}$. Now we can form a quantity which has Hotelling's non-central T^2 distribution from the quadratic form of the vector $\Sigma\ (x_p)\ (X^* - \bar{X}^*)$ with the inverse of its estimated variance covariance matrix

$$T^2 = \{S_{*p}\ [S_{**}\ A_{pp}^{-1}/(n-2)]^{-1}\ S_{p*}\}$$

$$= \{(n-2)\ S_{**}^{-1}\ (S_{*p}\ A_{pp}\ S_{p*})\}\ .$$

The second term in the brackets of expression 4.12 is thus $T^2/(n-2)$ where T^2 has Hotelling's non-central T^2 distribution with $(p-2)$ and $(n-2)$ degrees of freedom, and with non-centrality parameter

$$\lambda = S_{**}\ \delta^*_p\ \Sigma_{pp}^{-1}\ \delta^*_p.$$

Now the expected value of T^2 is $[(n-2)(p-2)/(n-p-1) + \lambda]$. Thus taking expectations over the distribution of x, we get

$$E(A_{**}) = S_{**}^{-1} [1 + (p-2) / (n-p-1) + \lambda/(n-2)]$$
$$= S_{**}^{-1} [(n-3) / (n-p-1) + \lambda/(n-2)] \quad \ldots\ldots\ldots (4.13)$$

Thus we can express the variance part of G_{Rp} as the sum of two parts. The first part is

$$S_{**}^{-1} s_p^2 (n-3)/ (n-p-1) \quad \ldots\ldots\ldots\ldots (4.14)$$

which is the expected variance of β_* for the case when all the elements of δ^*_p are zero, ie the variable of special interest is uncorrelated with the x variables in the model. The second term is

$$S_{**}^{-1} \quad s_p^2 \lambda / (n-2) \quad \ldots\ldots\ldots\ldots\ldots\ldots (4.15)$$

and represents the increase in the variance of the estimate of β^* which arises from the the correlation of X^* with the p-2 x variables included in the model.

We cannot use the sum of these two terms to derive a further criterion, because λ contains the unknown parameters δ^*. In order to estimate λ we must condition on X_p and estimate δ^*. To estimate λ we are led to consider the expectation of the term $(S_{*p} A_{pp} S_{*p})$ from 4.12 which depends on λ. When we estimate λ from this and substitute back into 4.13, we arrive back at the expression 4.12 which is conditional on the x variables in the model.

However, we can use the expression 4.13 to divide the variance part of G_{RP} into two parts which correspond to 4.14 (estimated directly) and 4.15 (estimated as the difference between G_{RP} and 4.15. This second part is the estimate of a quantity which is positive definite, although its estimate may be negative.

We can write G_{RP} as the sum of three parts as follows :

$$G_{RP} = v^*_{x*} \{RSS_P(n-2)(n-3)\} / \{RSS_{x*}(n-p)(n-p-1)\} \qquad \text{(variance1)}$$

$$+ v^*p - v^*_{x*} \{RSS_P(n-2)(n-3)\} / \{RSS_{x*}(n-p)(n-p-1) \qquad \text{(variance2)}$$

$$+ (b^*_{full}-b^*_P)^2 + v^*_{P(RSSfull}/(n-k-2)\}/\{RSS_P/(n-p-2)\}-v^*_{full}$$

$$\text{(bias}^2\text{)}$$

.... (4.16)

where the notation is the same as that used in section 4.2 with the extension that v_{x*} and RSS_{x*} are the expressions, for the model which contains only X^* and a constant, of the variance of the estimate of β^* (calculated as if this was the correct model) and the residual sum-of-squares for Y.

4.6 The relation of G_{RP} to another criterion

In their paper about mean square errors for prediction Breiman & Freedman (1983) give an incomplete reference to a technical report from Stanford University by Freedman and Moses which they say

derives a mean-square-error criterion for estimating the main effect in clinical trials. My correspondence with Moses, suggests that this Technical report was never completed. However, Moses kindly supplied me with some draft lecture notes (Moses 1983/87) on this work. These mention briefly such a criterion which can be shown to be equivalent for the situation of two treatment group to the expression 4.14. Because, one can assume for randomised trials that no bias is introduced by omitting covariates, the bias term is not required and all the δ^*, and hence term λ is exactly zero. Thus the expression G_{RP} reduces to 4.14 in this case. This expression is a constant multiple of S_P (chapter 3) for any given set of data. Thus, selecting a model which gives the minimum on this criterion will give identical results to the selection of the model which minimises S_P.

Moses, however, does not appear to have done any evaluation of this criterion. The remainder of the draft notes report the results of a small simulation, similar to that done by Schluchter and Forsythe, which evaluates the effect of the usual stepwise regression procedures on the estimation of the treatment effect. The results of this are broadly in agreement with Schluchter and Forsythe's.

4.7 Summary

Thus it can be seen that the two MSE criteria G_{FP} and G_{RP} can be used to select subsets of regressors by searching for models which give small values of the criteria. The use of the criteria

could be combined with stepwise procedures to produce a great many possible strategies. For example, a stepwise search could be used which selects the next covariate to enter as the one which will give the smallest value of the criterion being used. Alternatively each more complex model to be considered, with say, one or more additional covariates, could be treated as though it were the full model in the calculation of the MSE criteria, and compared with the current model on this basis.

The criteria could also be used as stopping rules in search procedures which are not directed by them. For example some conventional stepwise procedures could be used, or else special search procedures which might, for example, choose the covariates which produce the largest change in the estimated coefficient for the variable of special interest.

The two criteria G_{FP} and G_{RP} can each be split into a component corresponding to the squared bias, and one corresponding to the variance. The estimate of the squared bias can give a negative quantity, although we know that the term which is being estimated is positive definite. The same is true for one of the two terms in the variance part of G_{RP}. Each of the two criteria could be modified by replacing negative estimates of positive definite quantities by zero.

The various possibilities will be illustrated for the lead study data in the chapters which follow.

Chapter 5

The Edinburgh Lead Study: description of the data

5.1 Study design

The Edinburgh Lead Study was set up in 1983 at the instigation of the Medical Research Council with support from the Scottish Home and Health Department. Its main aim was to investigate the association between blood-lead levels and mental abilities in a population of Edinburgh school children, taking into account a wide range of other influences.

Most of the centre of Edinburgh was built in the nineteenth century or earlier. Many homes still retain some of their original lead plumbing, and the water is plumbosolvent. Thus water lead makes a substantial contribution to some children's lead intake (Raab, Laxen & Fulton 1987). Unlike many other inner city areas, the population of central Edinburgh is affluent and includes a high proportion of owner-occupiers and of people in professional and managerial occupations (SASPAK, 1983). The advantage of this situation for a study of lead exposure is that we may be expected to find higher lead levels occurring in children who were not subject to other influences which might result in poor results in the ability scores. Thus Edinburgh was selected precisely because we hoped that the extent of confounding between lead exposure and other variables would be small.

The basic design was a cross-sectional study of children in their third and fourth years of primary schooling (6 to 9 year olds) at local authority schools in a defined area of central Edinburgh. The schools were approached in random order, with all stages of the study in each school being completed within two to three months. The field work lasted from August 1983 to June 1985. The main outcome measure was an ability score which was standardised on a reference population to have a mean of 100 and a standard deviation of 15. Previous studies (Needleman et al 1971, and Smith et al 1983) had found differences of the order of 5 points on similarly standardised scores, between a high-lead group and a low-lead group. To detect a 5 point difference between two groups with 95% power and using a 5% significance level would require 234 pupils per group. Although our study was cross-sectional, rather than a two group study, this gave us some indication of the target numbers we should aim for. It was decided to aim for a sample size of 500 children.

The first step in each school was the compilation of a list of eligible children, and requests to the parents for their children's participation. Further details of the eligibility criteria and other aspects of the design have been published (Raab et al 1985, Fulton et al 1987). A medical team then visited the school to obtain venous blood samples which were assayed for lead. A main study sample was then selected from the blood lead levels in each school, which included all children in the top quartile of the blood-lead distribution and a one-in-three (approximately) sample of the remainder. The study was continued until the numbers in the main study sample reached our target. This required us to include 18

schools (excluding 2 small schools which were used in a pilot project). From the total of 1210 eligible children, parental consent was obtained for 948 (78%) to take part. A satisfactory blood sample and a successful lead assay were obtained for 855 children (90% of those whose parents agreed), and 501 of these children were selected into the main study sample.

The selected children were tested by a psychologist who visited the school. The test battery consisted of measures of inspection and reaction time, and ability and attainment tests. The latter were all taken from the British Ability Scales (BAS) (Elliott, Murray and Pearson 1978, Elliot 1983) which have been recently validated and standardised on a United Kingdom population. The attainment tests were of reading and number skills. Five ability tests were used which together give a combined score (BASC score), standardised on the same scale as the WISC-R IQ score (Wechsler 1978). Subsequently, an extensive home interview with one parent (usually the mother) collected data on the child's home and family background. This included tests of the parent's vocabulary and spatial ability (Raven, Court and Raven 1978). Behaviour ratings (Rutter 1967) for each child were completed by parents and teachers.

The full set of outcome variables were ;

(1) five ability tests from the BAS which were combined to give a combined score (BASC), standardised to a mean of 100 and with a standard deviation of 15;

(2) two attainment tests (reading and number skills);

(3) two tests of mental speed (inspection time and reaction time);

(4) parents' and teachers ratings of the children's behaviour.

In this thesis I will consider only the BASC, which was the main outcome measure, and also the basis on which the size of the study was planned.

5.2 Univariate statistics for blood-lead, BASC and covariates

The geometric mean blood-lead for the 855 children was 104 μg/l (mean of natural logs 4.64 s.d. 0.37) and the distribution appeared normal after log transformation. For the selected sample of 501 children the geometric mean was 115 μg/l (mean of natural logs 4.75 s.d. 0.38), and the distribution of the log values also appeared reasonably close to a normal distribution (fig 5.1). The log values were used as x* in the regression of outcomes on blood-lead. The justification for this choice was not that the blood-lead values were normally distributed, but that when Pocock et al (1987) examined the dose-response relationship in the data of Smith et al (1983) they found that it appeared to be more linear when the lead levels were measured on a log scale.

Figure 5.1: Log blood lead

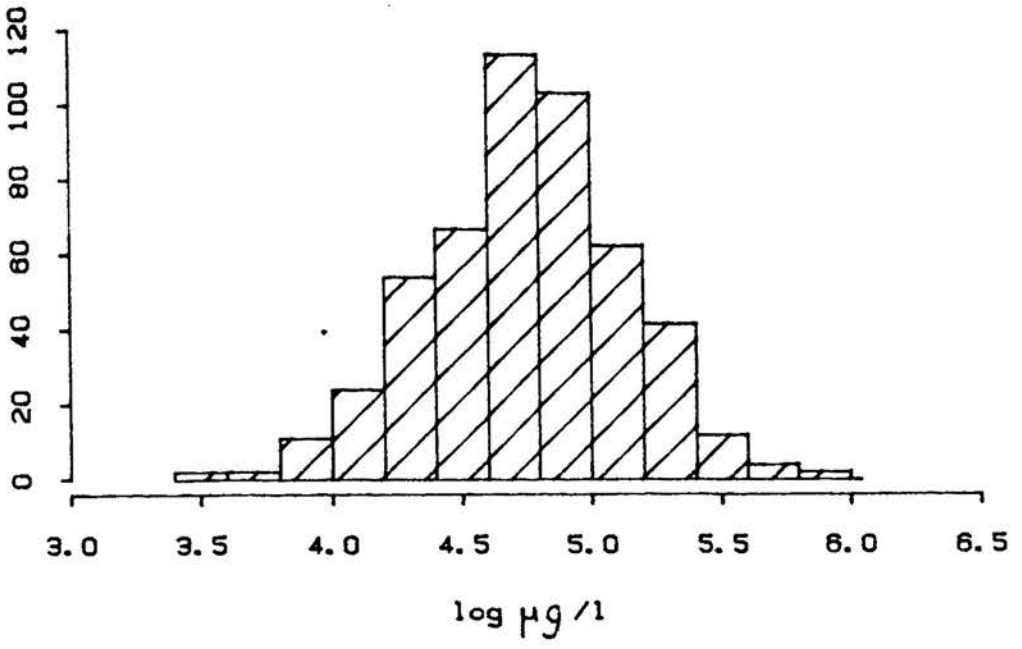
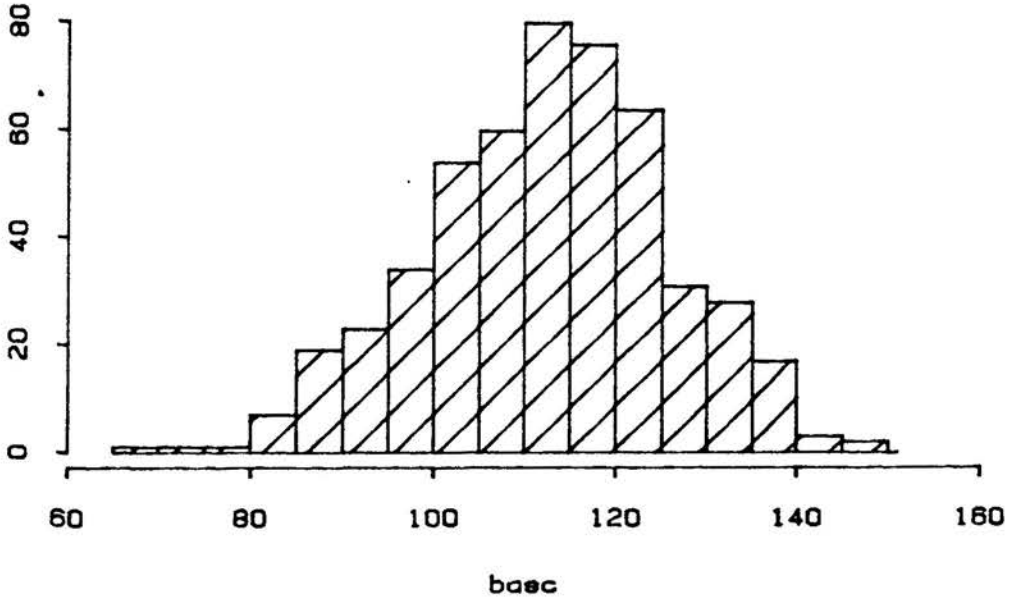


Figure 5.2: British abilility scale (combined) (BASC)



The mean score for the BASC test, for the 501 children for whom complete data were available, was 112, s.d. 13.4. Thus the children performed considerably better than the random sample of the UK population on which the tests had been standardised. The distribution of the BASC is shown in figure 5.2.

A large part of the parent's interview was devoted to questions which probed areas which might result in confounding of the blood-lead/ability relationship. The interview data were analysed, without reference either to the outcome scores or to the blood-lead results, to produce scores for 33 confounding variables. These variables were chosen because of their potential relationship with children's performance and their choice and scoring (where this was relevant) was based on the published results of long term child development studies (Douglas et al 1964, Douglas 1977, Davie et al 1972, Kellmer-Pringle, Butler and Davis 1966), reports in the psychological literature (Rutter & Madge 1976, Fogelman et al 1978) and experience in other lead studies, particularly the Institute of Child Health/Southampton study (Smith et al, 1983).

Some of the covariates are simple factual items e.g. age, sex of child, family size and birth order, some are based on well-established classifications e.g. social class and educational qualifications, and some have been measured by standardised tests, e.g. parent's ability. We also constructed more complex covariates by combining a number of related items in the data collected at interview and scoring them. Examples are:- parent's general and mental health scores, family structure score and child's interest

score. The items contributing to these scores are listed in the appendix to this chapter. Table 5.1 gives details and mnemonics for all the covariates. Scores for each variable were obtained for all children. For one-parent families, values were imputed for a second parent.

Table 5.1: Description of the covariates.

Mnemonic	Type of variable	Description
AGEINT	continuous	age of child in months
SEX	binary	male=1 female=2
MOVESCH	binary	change of school in past year (1=yes, 2=no)
CLASSYR	binary	year of schooling (3 or 4)
TIMEDAY	binary	a.m.=1 p.m.=2
HANDED	binary	right=1 left=2
FAMHIST	score	score for problems in family history
FSOC	score	father's and mother's Social class
MSOC	"	4=I&II, 3=IIInm, 2=IIIm, 1=IV&V
MQUALIF	score	father's and mother's qualifications
FQUALIF	"	from 0=none to 6=degree (see table 5.2)
UNEMPLO	binary	unemployed father or single mother
WORKMUM	score	working mother, 1=part-time, 2=full-time
PARHLTH	score	score for parents' health problems
PARMENT	score	score for parents' mental health problems
TOTCIGS	continuous	total cigarettes smoked by both parents
CARPHON	score	car/phone ownership 0=none, 1=one, 2=both.
CONSUME	score	total consumer goods owned from 4 items.
OCCUPRA	continuous	persons per room
FAMSIZE	score	family size 1=1 , 2=2, 3=3 or more children
BIRTHOR	score	1= first, 2= second, 3= third or more
GESTAT	score	0=38+weeks, 1=34-37 weeks, 2=<34 weeks
BRTHWT	binary	0=>2500g, 1=<2500g
BRTHSCO	score	score for problems at birth
MEDHIST	binary	history of child's medical probs 0=no, 1=yes
STHEIGH	continuous	age-standardised height
OFFSCHL	continuous	number of days off school in past year
CHILDIN	score	score for child's activities outside school
PARCHCO	score	score for parent/child communication
PARPART	score	score for parents' participation with child
PARSCHL	score	score for parents' involvement with school
PVOC	continuous	parent's vocabulary test
PMAT	continuous	parent's matrices test

The distributions of the covariates for the 501 children are shown in figure 5.3. It is obvious that many are very far from being normally distributed.

Figure 5.3: Distribution of covariates

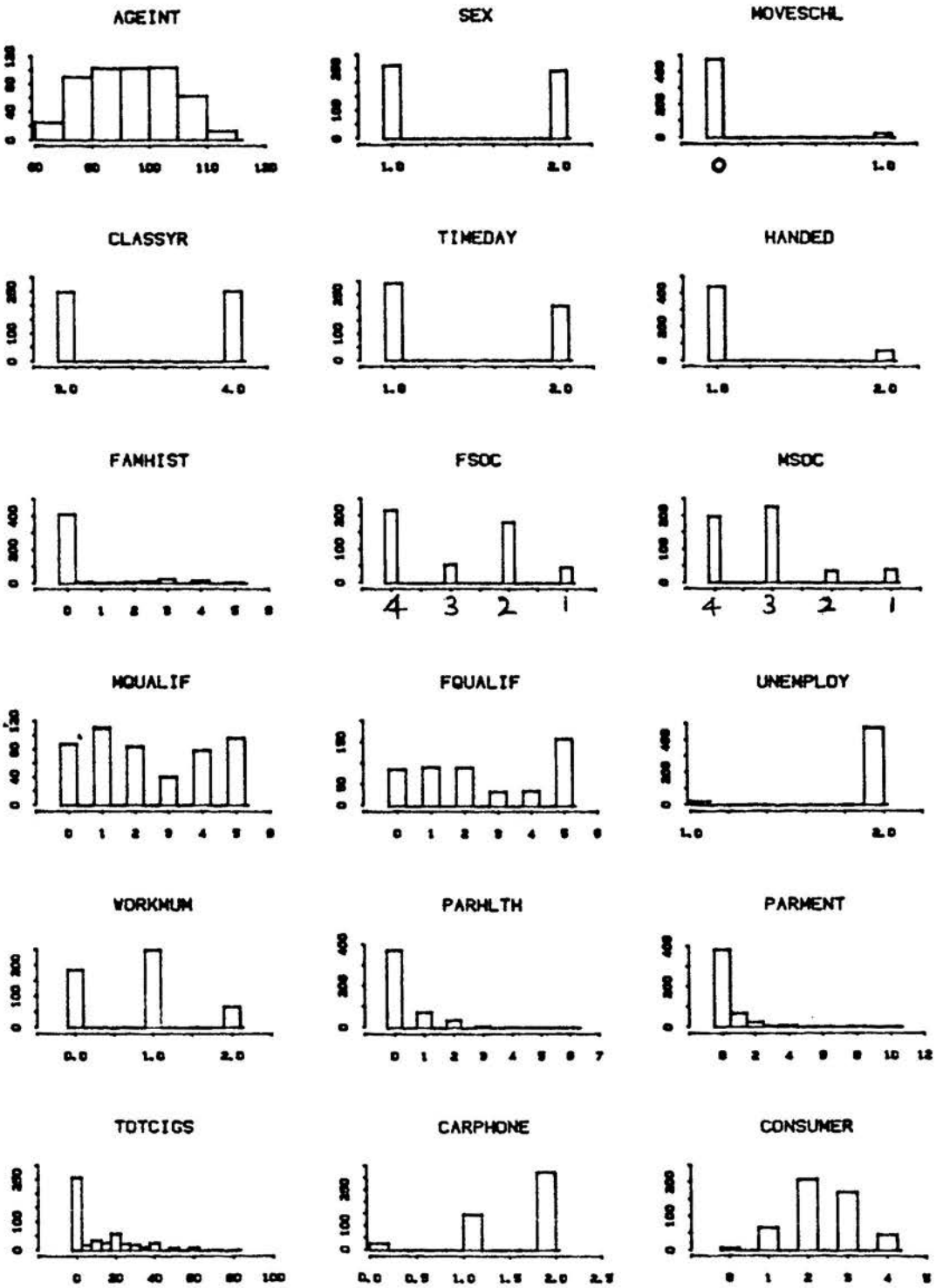
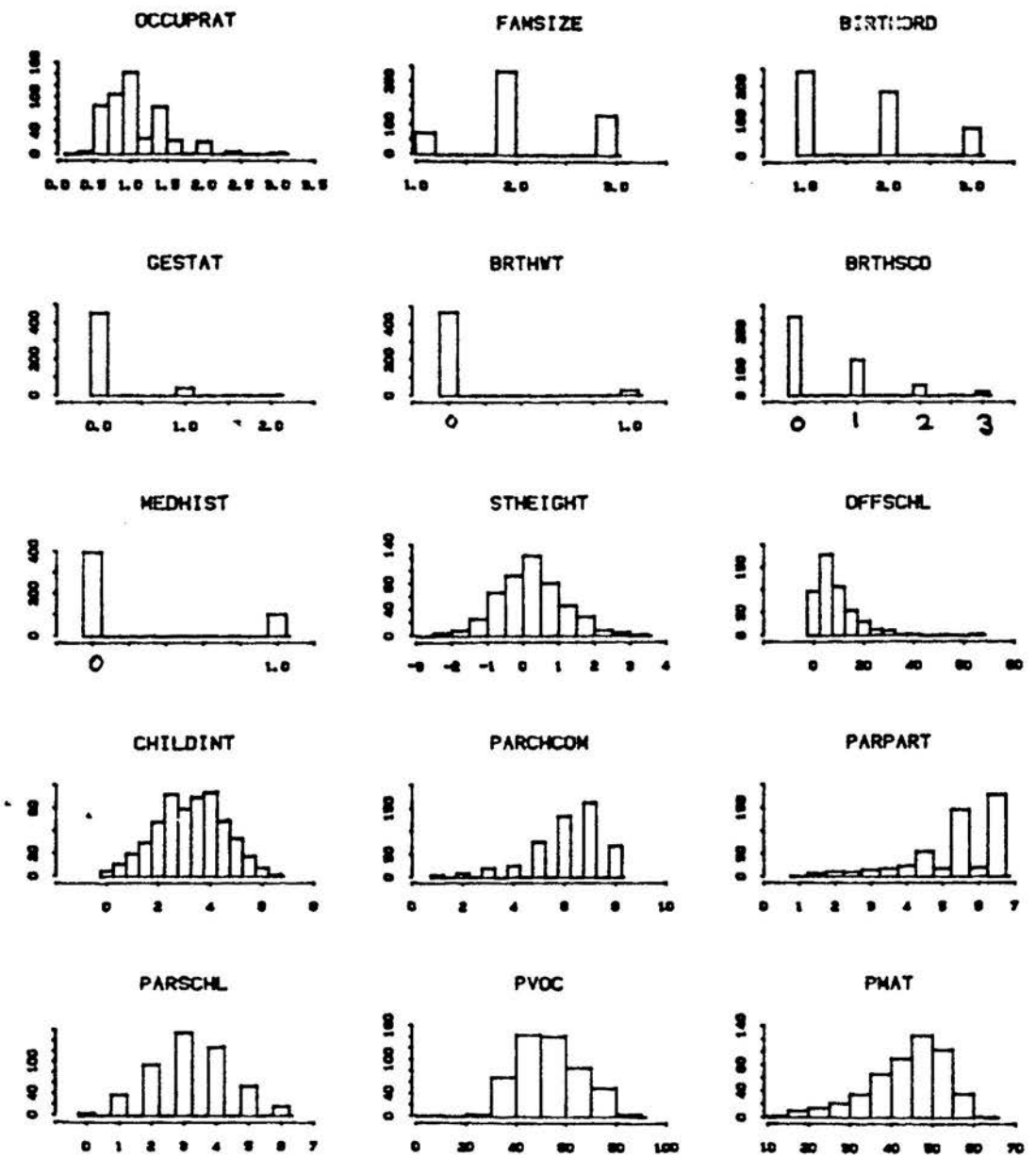


Figure 5.3 (contd.)



For those covariates which were continuous variables, or scores with more than two categories the relationships with the BASC were examined. The results are shown in table 5.2. The variables are ordered by the absolute value of their correlation with BASC.

Table 5.2: BASC and scored covariates.

Confounder	Categories	No. of children	BASC score mean	correlation with BASC
Parent's vocabulary score PVOC	<30	5	89.4	.52
	30-39	69	102.4	
	40-49	144	106.5	
	50-59	142	114.7	
	60-69	86	120.0	
	70+	55	121.1	
Mother's qualifications MQUALIF	None	88	101.6	.52
	Commercial/apprent.	111	108.1	
	Ordinary school cert.	85	109.2	
	Higher school cert.	41	116.2	
	Further education	79	118.3	
	Degree	97	121.5	
Father's qualifications FQUALIF	None	87	102.2	.49
	Commercial/apprent.	92	107.2	
	Ordinary school cert.	91	110.9	
	Higher school cert.	35	109.9	
	Further education	36	116.8	
	Degree	160	120.0	
Parent's matrices score PMAT	<20	12	96.3	.46
	20-29	35	98.3	
	30-39	100	107.2	
	40-49	214	113.0	
	50+	140	118.7	
Child's interest score CHILDIN	0 -<1	16	95.7	.40
	1 -<2	53	102.3	
	2 -<3	119	111.4	
	3 -<4	138	112.3	
	4 -<5	116	114.0	
	5 -<6	51	121.0	
	6 -<7	8	126.9	

Table 5.2 (contd)

Confounder	Categories	No. of children	BASC score mean	correlation with BASC
Mother's social class	I, II	198	118.4	.39
	III non-manual	227	109.0	
	III manual	36	106.7	
MSOC	IV, V	40	101.9	
Father's social class	I, II	217	117.9	.37
	III non-manual	56	109.0	
	III manual	181	107.5	
FSOC	IV, V	47	105.7	
Parental participation with child	1 -<2	14	98.6	.33
	2 -<3	22	102.9	
	3 -<4	33	107.6	
	4 -<5	77	109.2	
PARPART	5 -<6	162	111.9	
	6 -<7	193	116.0	
Occupancy ratio	<0.5 persons/room	5	122.4	-.32
	0.5 - 0.65	56	117.2	
	0.66 - 0.99	134	115.8	
OCCUPY	1	143	112.3	
	>1 - 1.5	111	108.1	
	>1.5	52	103.1	
Standardised height	<-2.00	6	94.8	.26
	-2.00 to -1.01	35	105.3	
	-1.00 to -0.01	160	110.4	
STHEIGHT	0.00 to 0.99	206	113.0	
	1.00 to 1.99	77	115.2	
	> 2.00	17	119.6	
Parent/child communication	bad 1 - 4	59	103.9	.24
	5	77	108.1	
	6	133	113.8	
PARCHCO	7	163	114.5	
	good 8	69	113.8	
Cigarettes smoked per day (both parents)	None	246	114.6	-.22
	1 - 10	58	113.9	
	11 - 20	88	107.2	
	21 - 40	87	110.0	
TOTCIGS	41 - 80	22	105.0	
Age	< 7:0	16	121.0	-.18
	7:0 - 7:5	99	114.4	
AGEINT	7:6 - 7:11	120	112.4	
	8:0 - 8:5	130	110.7	
	8:6 - 8:11	105	112.1	
	> 9:0	31	103.0	

Table 5.2 (contd)

Confounder	Categories	No. of children	BASC score mean	correlation with BASC
Parental involvement with school PARSCHL	bad 0	5	101.6	.16
	1	39	107.1	
	2	96	108.6	
	3	157	113.1	
	4	130	114.7	
	5	56	112.0	
	good 6	18	114.2	
Car/telephone ownership CARPHON	neither	28	101.6	.15
	either	148	111.9	
	both	325	112.9	
Absence from school in last year OFFSCHL	0 days	39	113.8	-.15
	1-10	306	113.0	
	11-20	119	111.1	
	21-30	25	103.5	
	> 30	12	107.7	
Gestation GESTAT	38+ weeks	455	112.5	-.12
	34-37	43	107.7	
	<34	3	101.7	
Family size FAMSIZE	1 child	74	108.9	.08
	2	289	112.5	
	3+	138	112.6	
Birth problem score BIRTHSCO	good 0	307	112.5	-.07
	1	138	111.7	
	2	41	110.4	
	bad 3	15	107.7	
Parent's health score PARHLTH	good 0	376	112.6	-.06
	1	76	109.9	
	2	37	109.9	
	bad 3+	12	112.9	
Birth order BIRTHORD	1st	241	112.6	-.04
	2nd	182	111.7	
	3rd+	78	111.0	
Working mother (or single father) WORKMUM	no paid employment	185	112.1	.03
	part-time	249	111.4	
	full-time	67	114.1	
Family history FAMHIST	0	411	112.0	-.03
	0.5 - 2.5	40	113.1	
	3.0 - 5.0	50	110.7	

Table 5.2 (contd)

Confounder	Categories	No. of children	BASC score mean	correlation with BASC
Parent's mental health score PARMENT	0 (good)	384	11.9	.02
	1	68	112.8	
	2	26	110.8	
	3+ (bad)	23	112.3	
Consumer goods	none	8	111.0	.01
	1 item	67	112.9	
	2	209	111.8	
	3	171	111.4	
CONSUME	all 4 items	46	114.3	

This table shows that the results from previous studies have, in the main, been confirmed in the present data. The relationships between many of our derived scores and the BASC scores are in the directions anticipated, and show a consistent dose-response relationship in most cases. Possible exceptions are the social class measures, where children from social classes I and II seem to score higher than would be expected from a simple linear scale. However, this pattern was not maintained in further investigation of this dose-response relationship, after control for PVOC and parental education. Also AGEINT seems to be more strongly influenced by the extreme groups than by those in the middle of the distribution. However, this variable has an interesting relationship with CLASSYR (stage of schooling), and their joint influence on ability explains this univariate relationship with AGEINT (see next chapter).

5.3 Relationship between covariates, blood lead (LNPBBL) and BASC

The correlation matrix between the covariates is shown in table 5.3 which also gives the correlation of the covariates with log blood lead and with the BASC score. This correlation matrix was used in McQuitty's elementary linkage analysis (McQuitty, 1957), which gives a visual representation of the relationships between variables, shown in Figure 5.4. This procedure links each variable with the variable to which it is most highly correlated (shown by an arrow), thus forming clusters of related variables. The dotted arrows show the highest correlation from any variable in a cluster with a variable in another cluster.

The covariates relate much more strongly to the BASC score than they do to the blood lead levels, as might be expected from the way in which they were selected. The highest correlations between covariates and BASC are those for the parents' test scores, educational qualifications and social class. The specially constructed scores also relate strongly to BASC, especially CHILDIN and PARPART, as does STHEIGHT and OCCUPRAT. The covariate with the strongest relationship with LNBLPB is standardised height, with the parent's vocabulary test and the social class and education measures also showing relationships with blood-lead although none of these are very strong.

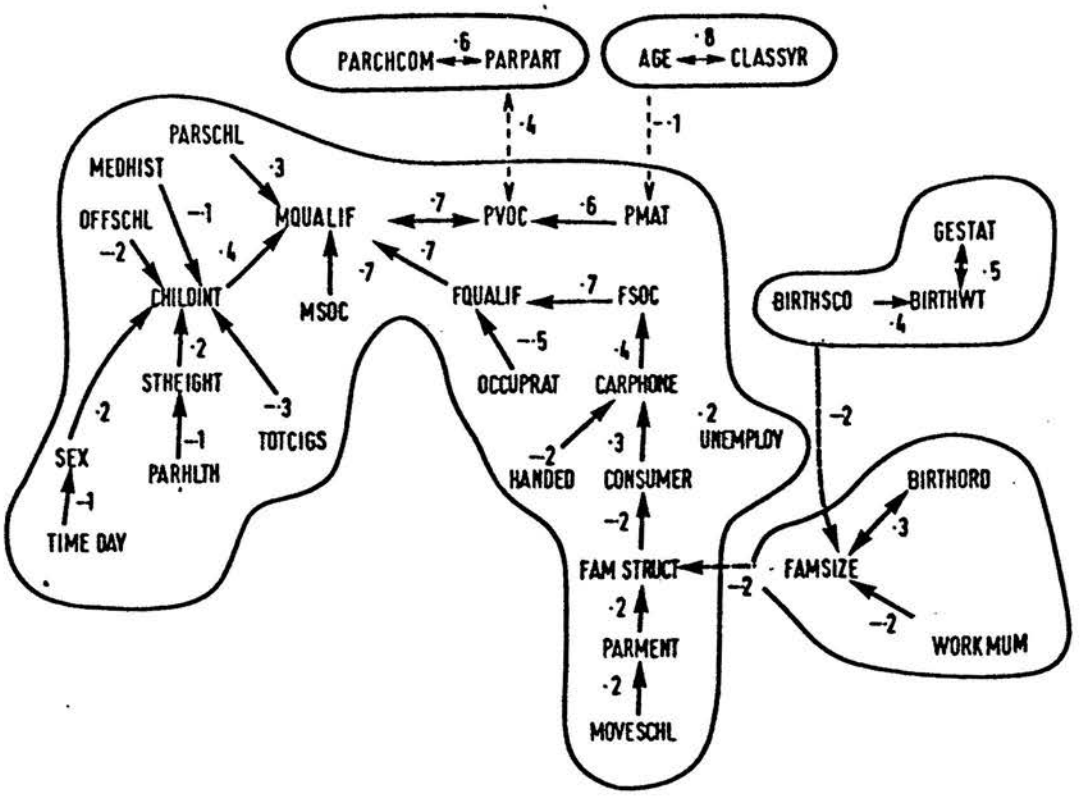
Table 5.3: Correlation matrix of confounders (unadjusted for school)

	AGEINT	SEX	MOVSCHL	CLASSYR	TIMEDAY	HANDED	FAMHIST	FSOC	MSOC	MQUALIF	FQUALIF	UNEMPLO	WORKMUM	PARHLTH	PARMENT	TOTCIGS	CARPHON	CONSUME	OCCUPRA	FAMSIZE	BIRTHOR	GESTAT	BRTHWT	BRTHSCO	MEDHIST	STHEIGH	OFFSCHL	CHILDIN	PARCHCO	PARPART	PARSCHL	PVOC	PMAT
AGEINT	■■	.	9	81	.	8	.	.	-9	-9	.	.	.	13	.	.	.	.	.	.	.	.	.	7	.	.	.	.	.	.	-8	-14
SEX	.	■■	.	.	-8	.	.	.	7	.	.	8	.	.	.	-7	9	10	.	.	.	.	.	-11	9	.	16	14	.	.	.	7
MOVESCH	9	.	■■	.	.	.	.	.	.	.	.	.	.	.	15	.	.	.	11	.	.	.	.	.	-7	.	.	.	.	.	.	.
CLASSYR	81	.	.	■■	.	10	.	.	.	.	.	7	.	8	.	.	.	.	.	7	.	.	.	7	.	.	.	.	.	.	.	-7
TIMEDAY	.	-8	.	.	■■	.	.	.	.	.	.	.	.	.	.	.	.	-7	.	.	.	.	.	.	.	.	.	.	.	.	.	.
HANDED	8	.	.	10	.	■■	.	.	.	-7	-11	.	-8	.	.	8	-16	.	12	9	.	.	.	.	-10	.	.	-12	.	.	-16	-8
FAMHIST	.	.	.	.	.	.	■■	-21	.	.	.	-11	.	11	19	9	-21	-24	-8	-20	.	.	.	.	.	.	-15	-10	.	.	.	.
FSOC	.	.	.	.	.	.	-21	■■	47	55	67	12	.	.	.	-19	38	14	-39	20	.	.	.	-12	-10	15	.	27	22	26	22	48	35
MSOC	-9	7	.	.	.	.	.	47	■■	66	54	10	11	.	.	-22	33	.	-38	10	.	.	.	-11	.	20	.	33	18	28	25	54	50
MQUALIF	-9	.	.	.	.	-7	.	55	66	■■	73	8	16	.	.	-30	30	.	-45	14	.	-9	-11	-8	-9	22	-11	39	29	39	34	74	53
FQUALIF	.	.	.	.	.	-11	.	67	54	73	■■	11	8	.	.	-32	33	.	-49	16	.	.	-8	-9	-7	19	-11	39	29	40	33	63	48
UNEMPLO	.	8	.	7	.	.	-11	12	10	8	11	■■	12	.	.	-14	17	.	.	.	9	.	.	.	.	10	.	13	9	.	.	9	.
WORKMUM	.	.	.	.	.	-8	.	.	11	16	8	12	■■	.	.	.	11	.	.	-22	10	.	.	.	.	.	.	.	.	.	.	17	9
PARHLTH	13	.	.	8	.	.	11	.	.	.	.	.	.	■■	.	.	.	.	.	-8	.	.	.	.	9	-14	.	-10	-10	-10	.	.	-12
PARMENT	.	.	15	.	.	.	19	.	.	.	.	.	.	.	■■	.	-8	.	.	.	.	.	.	-7	.	.	.	.	.	.	7	.	.
TOTCIGS	.	-7	.	.	.	8	9	-19	-22	-30	-32	-14	.	.	.	■■	-23	-10	27	-16	.	.	8	.	.	-13	.	-34	-14	-19	-19	-25	-23
CARPHON	.	9	.	.	.	-16	-21	38	33	30	33	17	11	.	.	-8	-23	■■	27	-33	8	9	.	.	.	15	-7	23	13	12	12	25	21
CONSUME	.	10	.	.	-7	.	-24	14	.	.	.	.	.	.	.	-10	27	■■	.	11	15	7	8	.	.	.	.	7	.	.	.	.	.
OCCUPRA	.	.	11	.	.	12	-8	-39	-38	-45	-49	.	.	.	.	27	-33	.	■■	11	9	.	7	.	.	-19	10	-31	-23	-28	-24	-44	-31
FAMSIZE	.	.	.	7	.	9	-20	20	10	14	16	.	-22	-8	.	-16	8	11	11	■■	29	.	.	-19	.	.	-14	17	.	.	9	7	9
BIRTHOR	.	.	.	.	.	.	.	.	.	.	.	9	10	.	.	.	9	15	9	29	■■	.	.	-11	.	.	.	7	.	-12	.	.	.
GESTAT	.	.	.	.	.	.	.	.	.	-9	.	.	.	.	.	.	.	7	.	.	.	■■	50	34	.	.	.	7	.	.	-9	.	.
BRTHWTD	.	.	.	.	.	.	.	.	.	-11	-8	.	.	.	.	8	.	8	7	.	.	50	■■	40	.	.	.	.	.	.	-7	-8	-11	-7
BRTHSCO	7	.	.	.	.	.	.	-12	-11	-8	-9	.	.	.	-7	.	.	.	.	-19	-11	34	40	■■	.	.	.	.	.	.	-10	-8	.
MEDHIST	.	-11	.	7	.	.	.	-10	.	-9	-7	.	.	9	.	.	.	.	.	.	.	.	.	.	■■	.	.	-12	-9	.	.	.	.
STHEIGH	.	9	-7	.	.	-10	.	15	20	22	19	10	.	-14	.	-13	15	.	-19	.	.	.	.	.	.	■■	.	23	11	13	12	20	18
OFFSCHL	.	.	.	.	.	.	.	.	.	-11	-11	.	.	.	.	.	-7	.	10	-14	.	.	.	.	.	.	■■	-17	.	-9	.	-12	-9
CHILDIN	.	16	.	.	.	-15	27	33	39	39	13	.	-10	.	-34	23	7	-31	17	7	7	.	.	-12	23	-17	■■	32	37	26	36	32
PARCHCO	.	14	.	.	.	-12	.	22	18	29	29	9	.	-10	.	-14	13	.	-23	.	.	.	.	.	-9	11	.	32	■■	58	20	35	23
PARPART	.	.	.	.	.	.	-10	26	28	39	40	.	.	-10	.	-19	12	.	-28	.	-12	.	-7	.	.	13	-9	37	58	■■	30	43	31
PARSCHL	.	.	.	.	.	.	.	.	22	25	34	33	.	.	.	7	-19	12	.	-24	9	.	-9	-8	-10	.	12	.	26	20	30	■■	32	25
PVOC	-8	.	.	.	.	-16	.	48	54	74	63	9	17	.	.	-25	25	.	-44	7	.	.	-11	-8	.	20	-12	36	35	43	32	■■	56
PMAT	-14	7	.	-7	.	-8	.	35	50	53	48	.	9	-12	.	-23	21	.	-31	9	.	.	-7	.	.	18	-9	32	23	31	25	56	■■
BASC	-18	.	.	-10	.	-14	.	37	39	51	49	.	.	.	.	-22	15	.	-32	7	.	-12	-14	-7	-7	26	-15	40	24	33	16	52	46
LNBLPB	-9	-8	-8	.	.	.	.	-15	-8	-15	-12	.	.	.	.	8	.	.	.	.	.	.	.	.	.	8	-20	.	.	.	.	-13	.

AGEINT SEX MOVSCHL CLASSYR TIMEDAY HANDED FAMHIST FSOC MSOC MQUALIF FQUALIF UNEMPLO WORKMUM PARHLTH PARMENT TOTCIGS CARPHON CONSUME OCCUPRA FAMSIZE BIRTHOR GESTAT BRTHWT BRTHSCO MEDHIST STHEIGH OFFSCHL CHILDIN PARCHCO PARPART PARSCHL PVOC PMAT

NOTES : Correlations times 100 are rounded to nearest whole number, and . represents a correlation which is less than 0.065 in absolute value.

Figure 5.4 McQuitty's elementary linkage anlysis for the 33 covariates



The covariates show strong interrelationships, as can be seen in Table 5.3. The linkage analysis divides them into five groups. Four small groups contain items on the age of the child, variables relating to birth, family size and parent-child communication. The fifth and largest group includes those variables which relate to the social and educational background of the family. There appear to be two sub-branches within this group, one of which relates to the social situation of the family, centred on the social class measures and the other relating to the quality of the child's home life centering on the child's interest score.

5.4 Blood lead, BASC and covariates by school

The analyses presented so far have ignored the blocking factor, school. Both the BASC and LNBLPB vary between schools. The analyses of variance for BASC, LNBLPB and selected covariates are presented in table 5.4. The variables which have been ommitted from this table are those with very skew distributions, because they would be unlikely to meet the assumptions of the analysis of variance. However, certain binary variables and those with short ordinal scales have been included, where the distribution is reasonably symmetrical, because the fairly large numbers which we are dealing with (the smallest school contributed 14 pupils, and the largest 54) will ensure that the distribution of school means will approach normality.

Table 5.4: Analyses of variance within and between schools.

VARIATE	SUMS OF SQUARES between schools	SUMS OF SQUARES within schools	MEAN SQUARES between schools	MEAN SQUARES within schools	F Ratio* (17/483 df)
BASC	19964	69442	1174.4	143.7	**8.17**
LNBLPB	11.72	61.53	0.689	0.127	**5.41**
AGEINT	3646	23808	214.5	49.3	**4.35**
SEX	3.48	121.55	0.201	0.251	0.81
CLASSYR	5.13	120.55	0.302	0.249	1.22
TIMEDAY	3.79	117.68	0.223	0.244	0.92
FSOC	138.3	441.0	8.135	0.914	**8.91**
MSOC	95.8	284.8	5.632	0.590	**9.55**
MQUALIF	541.1	1069.7	31.8	2.2	**14.41**
FQUALIF	621.7	1231.6	36.6	2.5	**14.34**
WORKMUM	9.27	214.9	0.54	0.45	1.23
CONSUMER	30.1	359.3	1.77	0.74	**2.37**
OCCUPRAT	20.6	61.1	1.21	0.13	**9.56**
FAMSIZE	20.7	183.2	1.21	0.37	**3.20**
BIRTHORD	8.1	257.9	0.47	0.53	0.89
STHEIGHT	37.5	395.3	2.208	0.818	**2.70**
OFFSCHL	2123	33180	124.9	68.7	1.81
CHILDINT	213.7	626.9	12.57	1.30	**9.68**
PARCHCOM	145.1	914.3	8.53	1.89	**4.51**
PARPART	156.6	690.7	9.21	1.43	**6.44**
PARSCHL	93.0	697.2	5.47	1.44	**3.79**
PVOC	23910	52917	1406.5	108.7	**13.05**
PMAT	7704	36612	453.2	75.8	**5.98**

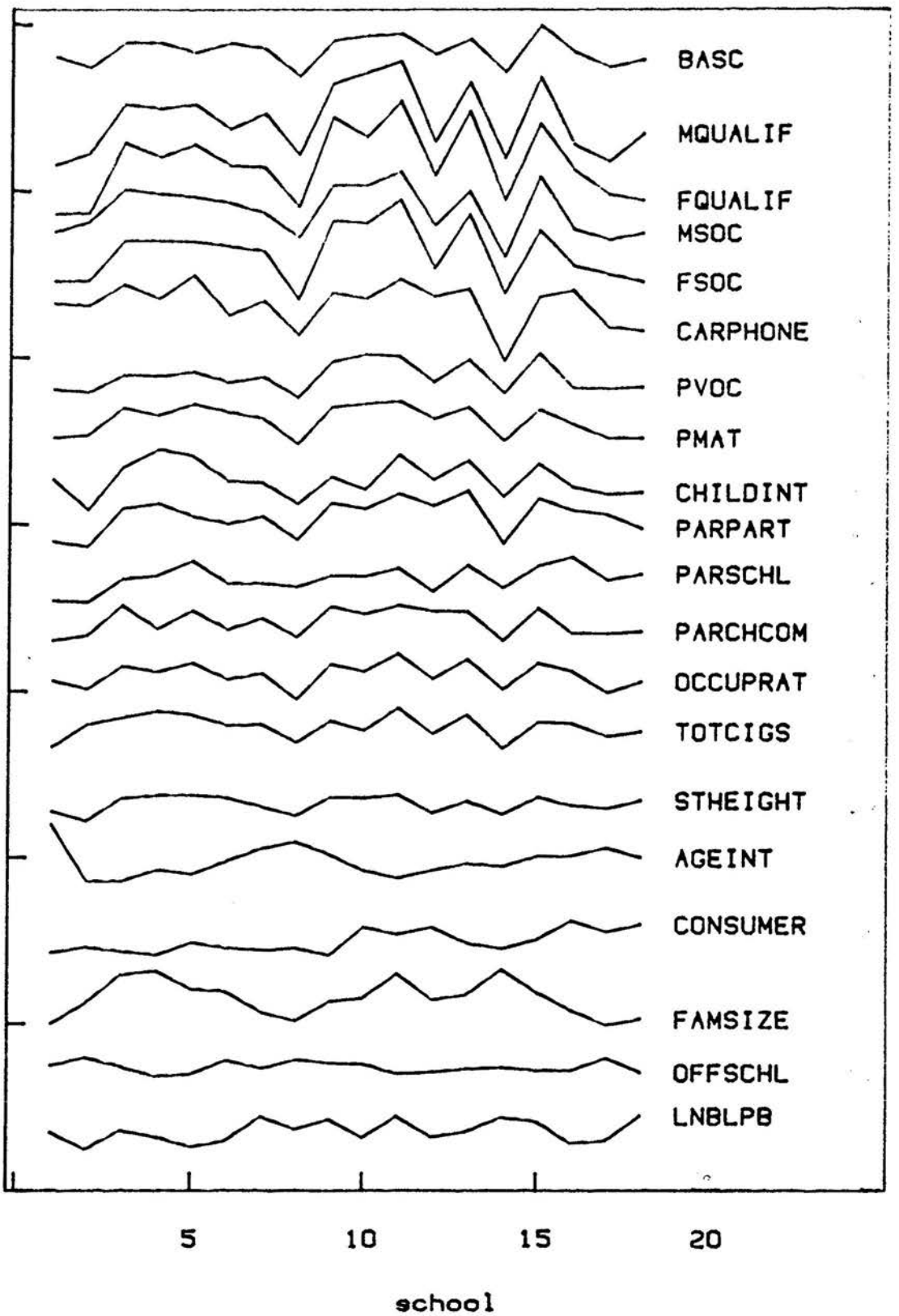
*Percentage points for the F ratio - 1.66(p=0.05) and 2.07(p=0.01)

The distributions of values within and between schools, for those variables with skew distributions, were compared using Kruskal-Wallis tests. No differences between schools were found for the variables MOVESCHL, HANDED, FAMHIST, UNEMPLOY, PARHLTH, PARMENT, GESTAT, BIRTHWT, and MEDHIST. There were very pronounced differences for TOTCIGS and CARPHONE (χ^2 values of 73 and 57 with 17 degrees of freedom, both $p<0.001$) and somewhat less so for OFFSCH ($\chi^2 = 33$, $p=0.01$).

The profiles of the school means for all the variables which vary by by school are shown in Figure 5.5. The schools are numbered in the order in which they were approached. Each variable is scaled by the range of the values in the whole sample, and measured on a scale from 0 (corresponding to the lowest value) and 100 (for the highest value). The variables are ordered so that those with similar patterns of between-school variation are placed together. The variables OCCUPRAT and TOTCIGS have been reversed on this scale, because they show the reverse pattern to the other variables with which they are correlated. The scale has also been reversed for LNBLPB, although the pattern is less clear in this case.

A common pattern of profiles of school means can be seen for the BASC and the 14 variables (MQUAL to STHEIGHT) which are shown immediately below it on Figure 5.5. These variables form the core of the main cluster in Figure 5.4, with the exception of PARSCHL and PARCHCOM which form a separate cluster which nevertheless has a high correlation with the main cluster. Variables which show similar between-school patterns have been grouped together. AGEINT has a different pattern between schools, which corresponds to the time of the school year when the children were seen. CONSUMER has an increasing trend over the period of the study. The variables FAMSIZE, OFFSCHL and LNBLPB show somewhat different patterns.

Figure 5.5 School mean profiles for selected variables. Each variable is measured on a scale determined by the range of values in the data on a scale from 0 to 100.



These differences between schools do not, however, account for all the inter-relationships between the covariates seen in Table 5.3 and Figure 5.4 above. The correlations adjusted for school means are given in Table 5.6 and the corresponding cluster analysis is shown in Figure 5.6. Substantial correlations remain between the variables, in a pattern which is not dissimilar to that for the unadjusted data. Those correlations forming the core of the cluster analysis are reduced in absolute value by between 0.06 and 0.18. Most other correlations are either reduced by a small amount, or relatively unaffected.

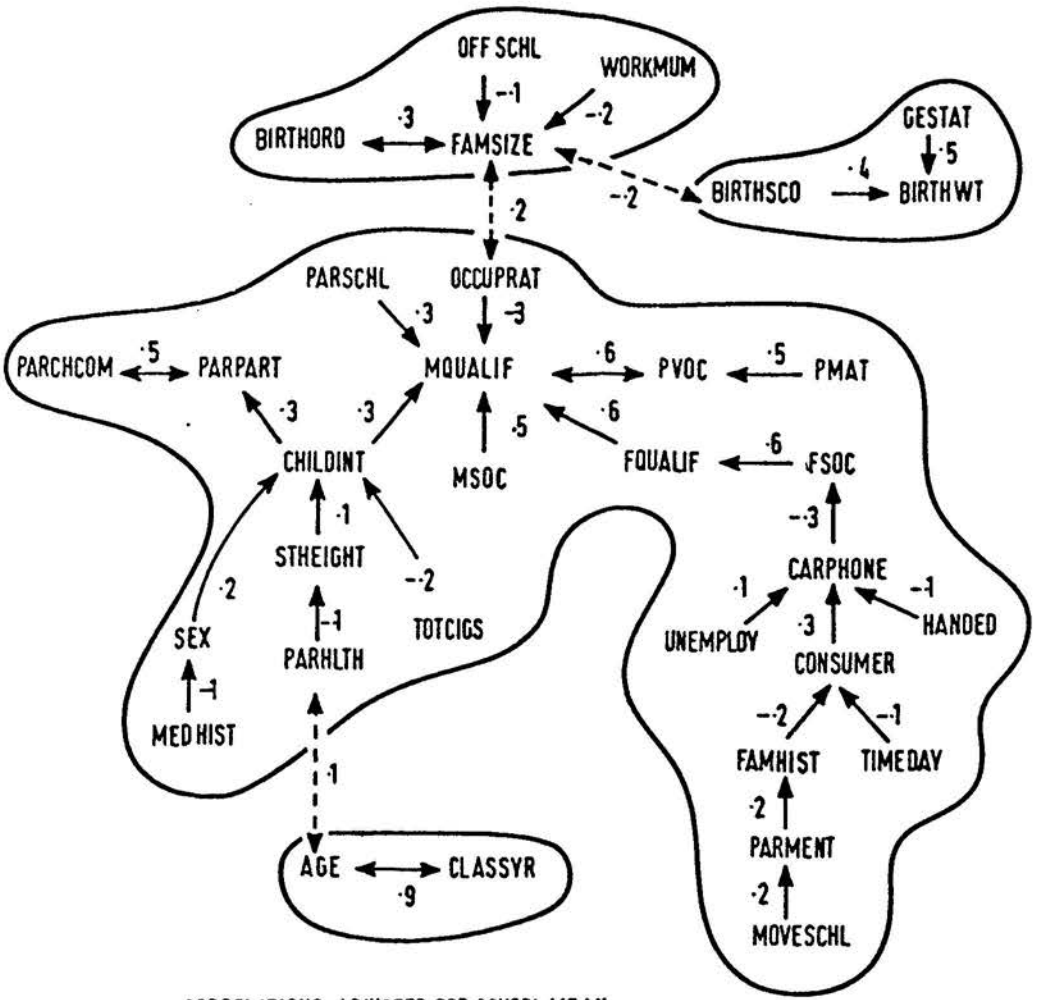
So far I have not discussed the correlation between BASC and LNBLPB, which is the item of major interest in this study. It will be presented along with the results of the regression analyses in the next section.

Table 5.6: Correlation matrix of confounders (adjusted for school).

	AGEINT	SEX	MOVSCHL	CLASSYR	TIMEDAY	HANDED	FAMHIST	FSOC	MSOC	MQUALIF	FQUALIF	UNEMPLO	WORKMUM	PARHLTH	PARMENT	TOTCIGS	CARPHON	CONSUME	OCCUPRA	FAMSIZE	BIRTHOR	GESTAT	BRTHWT	BRTHSCO	MEDHIST	STHEIGH	OFFSCHL	CHILDIN	PARCHCO	PARPART	PARSCHL	PVOC	PMAT
AGEINT	■■	.	8	85	.	10	.	.	.	.	.	.	.	14	.	.	.	9	.	.	8	.	.	.	.	.	.	10	.	.	.	.	-7
SEX	.	■■	.	.	-7	.	.	.	.	.	.	8	.	.	.	-7	9	13	.	.	.	.	.	.	-11	9	.	17	15	.	.	.	.
MOVESCH	8	.	■■	.	.	.	.	.	.	.	.	.	.	.	15	.	.	.	8	.	.	.	.	.	.	.	.	.	.	7	.	.	.
CLASSYR	85	.	.	■■	.	11	.	.	.	.	.	7	.	9	.	.	.	.	.	9	.	.	.	.	.	.	.	.	.	.	.	.	.
TIMEDAY	.	-7	.	.	■■	.	.	.	.	.	.	.	.	.	.	.	.	-8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
HANDED	10	.	.	11	.	■■	.	.	.	.	.	.	-7	.	.	7	-12	.	7	10	8	.	.	.	.	-9	.	.	-9	.	-12	.	.
FAMHIST	.	.	.	.	.	.	■■	-23	.	.	.	-9	7	10	20	7	-19	-23	-11	-22	.	.	.	.	.	.	-14	.	-9	.	.	.	.
FSOC	.	.	.	.	.	.	-23	■■	31	39	55	.	.	.	.	.	30	17	-21	14	.	.	.	-7	-7	.	.	11	9	10	14	30	20
MSOC	.	.	.	.	.	.	.	31	■■	55	39	.	12	.	.	-9	22	.	-22	.	.	.	.	.	.	9	.	18	.	14	19	39	39
MQUALIF	.	.	.	.	.	.	.	39	55	■■	62	.	19	.	.	-18	19	.	-27	.	-14	-10	-8	.	.	11	.	26	14	25	27	63	42
FQUALIF	.	.	.	.	.	.	.	55	39	62	■■	.	10	.	.	-20	22	.	-31	7	-8	.	.	.	.	7	.	24	13	23	24	47	34
UNEMPLO	.	8	.	7	.	.	-9	.	.	.	.	■■	12	.	.	-9	13	.	.	.	7	.	.	.	.	7	.	.	.	.	.	.	.
WORKMUM	.	.	.	.	.	-7	7	.	12	19	10	12	■■	.	.	.	9	.	.	-21	10	.	.	.	.	.	.	.	.	.	.	19	9
PARHLTH	14	.	.	9	.	.	10	.	.	.	.	.	.	■■	.	.	.	.	.	-7	.	.	.	.	9	-12	.	.	-9	-7	.	.	-9
PARMENT	.	.	15	.	.	.	20	.	.	.	.	.	.	.	■■	.	-10	.	-8	.	.	.	.	-9	.	.	.	.	.	.	11	7	.
TOTCIGS	.	-7	.	.	.	7	7	.	9	-18	-20	-9	.	.	.	■■	-14	-11	15	-10	.	.	8	.	.	.	-25	.	-9	-12	-15	-12	
CARPHON	.	9	.	.	.	-12	-19	30	22	19	22	13	9	.	-10	-14	■■	29	-21	.	7	8	.	.	9	.	10	.	.	.	13	9	
CONSUME	9	13	.	.	-8	.	-23	17	.	.	.	.	.	.	.	-11	29	■■	.	13	14	9	8	.	.	.	10	.	.	.	.	.	.
OCCUPRA	.	.	8	.	.	7	-11	-21	-22	-27	-31	.	.	.	-8	15	-21	.	■■	24	15	.	.	.	-8	.	-13	-10	-12	-14	-27	-15	
FAMSIZE	.	.	.	9	.	10	-22	14	.	.	7	.	-21	-7	.	-10	.	13	24	■■	29	.	.	-17	.	.	-10	9	.	-8	.	.	.
BIRTHOR	8	.	.	.	.	8	.	.	.	-14	-8	7	10	.	.	.	7	14	15	29	■■	.	.	-9	.	-8	.	.	.	-17	-7	.	-10
GESTAT	.	.	.	.	.	.	.	.	.	-10	.	.	.	.	.	.	8	9	.	.	.	■■	52	34	.	.	.	7	.	.	-9	.	.
BRTHWTD	.	.	.	.	.	.	.	.	.	-8	.	.	.	.	.	8	.	8	.	.	.	52	■■	40	.	.	.	.	.	.	.	-7	.
BRTHSCO	.	.	.	.	.	.	.	-7	.	.	.	.	.	.	-9	.	.	.	-17	-9	34	40	■■	.	.	.	9	.	9	-8	.	.	
MEDHIST	.	-11	.	.	.	.	.	-7	.	.	.	.	.	9	.	.	.	.	.	.	.	.	.	.	■■	.	.	-8	.	.	.	.	.
STHEIGH	.	9	.	.	.	-9	.	.	9	11	7	7	.	-12	.	.	9	.	-8	.	-8	.	.	.	.	■■	.	14	.	.	.	10	18
OFFSCHL	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-10	.	.	.	.	.	.	■■	-10	.	.	.	-8	.
CHILDIN	10	17	.	.	.	.	-14	11	18	26	24	.	.	.	.	-25	10	10	-13	9	.	7	.	9	-8	14	-10	■■	24	26	19	22	20
PARCHCO	.	15	.	.	.	-9	.	9	.	14	13	.	.	-9	.	.	.	.	-10	.	.	.	.	.	.	.	.	24	■■	54	15	22	11
PARPART	.	.	7	.	.	.	-9	10	14	25	23	.	.	-7	.	-9	.	.	-12	-8	-17	.	.	9	.	.	.	26	54	■■	22	30	20
PARSCHL	.	.	.	.	.	.	.	14	19	27	24	.	.	.	11	-12	.	.	-14	.	-7	-9	.	-8	.	.	.	19	15	22	■■	26	18
PVOC	.	.	.	.	.	-12	.	30	39	63	47	.	19	.	7	-15	13	.	-27	.	.	.	-7	.	.	10	-8	22	22	30	26	■■	47
PMAT	-7	.	.	.	.	.	.	20	39	42	34	.	9	-9	.	-12	9	.	-15	.	-10	.	.	.	.	9	.	20	11	20	18	47	■■
BASC	-16	.	.	-10	.	-10	.	22	23	38	35	.	.	.	.	-12	.	.	-15	.	-10	-11	-11	.	.	18	-12	29	12	21	8	39	36
LNBLPB	-9	-7	-10	.	.	.	.	-11	.	.	.	-8	.	.	.	8	-7	.	.	.	.	.	.	.	-10	-20	-9	.	.	.	.	.	.
	AGEINT	SEX	MOVSCHL	CLASSYR	TIMEDAY	HANDED	FAMHIST	FSOC	MSOC	MQUALIF	FQUALIF	UNEMPLO	WORKMUM	PARHLTH	PARMENT	TOTCIGS	CARPHON	CONSUME	OCCUPRA	FAMSIZE	BIRTHOR	GESTAT	BRTHWT	BRTHSCO	MEDHIST	STHEIGH	OFFSCHL	CHILDIN	PARCHCO	PARPART	PARSCHL	PVOC	PMAT

NOTES : Correlations times 100 are rounded to the nearest whole number, and . represents a correlation which is less than 0.065 in absolute value.

Figure 5.6. McQuitty's elementary linkage analysis for the 33 covariates, adjusted for school.



5.5 Regression analyses with BASC as the dependent variable.

Taking BASC as the dependent variable, the estimated regression coefficient for LNBLPB (b*) is modified by the addition of the blocking factor school and/or the 33 covariates. The results are given in Table 5.7. The standard errors and t-values quoted are derived in the usual way, by assuming that the model being fitted is the correct one. The variance of the estimated coefficient is calculated as the product of the residual mean square from the model multiplied by a constant (see chapter 4). These two factors of the variance b* are also given in table 5.7.

Table 5.7 Regression coefficients (b*) for various models. ▼

Model	b*	s.e	t-value	FACTORS OF VARIANCE OF b* RMS	Multiplier
Lead only	-5.45	1.54	-3.53	174.8	13.65 $\times 10^{-3}$
Lead & schools	-3.89	1.52	-2.56	142.1	16.26 $\times 10^{-3}$
Lead & covariates	-3.18	1.30	-2.45	104.9	16.05 $\times 10^{-3}$
Lead, schools & covariates	-3.81	1.37	-2.79	98.8	18.90 $\times 10^{-3}$

▼ These results differ very slightly from the results quoted in Fulton et al (1987), because in that paper the variable WORKMUM was considered as a factor with 3 levels.

The results show a statistically significant relationship between blood lead and ability, the negative coefficient implying that high lead levels are associated with poor scores on the BASC tests. The relationship, although significant, is not strong in terms of the correlation between LNBLPB and ABILITY. Their

univariate correlation is 0.156, reducing to a partial correlation of 0.102 after controlling for schools, and to 0.084 after controlling for schools and the 33 covariates together.

The univariate regression coefficient for LNBLPB is modified to a rather similar extent by either the schools or the covariates, or by both taken together. The lowest absolute value for the coefficient is achieved by adjustment for covariates only. The residual mean square is more sharply reduced by the covariates than by the blocking factor, school. However, fitting school after the covariates still gives a significant improvement to the fit of the model (F ratio 2.69 df 17/449 $p<0.01$). The multiplier of the variance of the LNBLPB coefficient, which relates to the multiple correlation of the other covariates with blood lead, is similar for either the covariates or the schools taken separately, and is increased even further when both schools and covariates are included in the model.

The regression which controlled for schools and all the covariates was the one which was reported in the account of this study in the medical literature (Fulton et al 1987), as giving the most secure estimate of the effects of lead on the BASC score. This decision was taken, before the data were analysed, because of the uncertainty of the validity of the standard errors based on any data-dependent selection procedure. However, some results of such procedures were also quoted to highlight covariates which were important.

5.6 Regression coefficients for the covariates

These will be discussed in much more detail in the next chapter. The strength of the univariate associations between the BASC score and the covariates can be assessed from the correlations in table 5.3, and after controlling for the blocking factor, school, in table 5.6. The relationships with BASC score in the models which include all the covariates are given in table 5.8.

Table 5.8: Regression coefficients(b) and t values for covariates

Model	Covariates		Covariates +LNBLPB		Covariates +schools		Covariates +school+LNBLPB	
	b	t	b	t	b	t	b	t
AGEINT	-.30	-2.7	-.34	-3.1	-.54	-4.2	-.57	-4.4
SEX	-1.7	-1.7	-1.9	-1.9	-1.8	-1.9	-2.0	-2.1
MOVESCHL	.07	.	.25	.	1.6	.	1.0	.
CLASSYR	1.4	.	1.8	.	4.1	2.2	4.4	2.4
TIMEDAY	-.02	.	-.46	.	-.01	.	-.23	.
HANDED	-.09	-1.8	-2.8	-1.8	-2.2	-1.5	-2.3	.
FAMHIST	-.01	.	-.15	.	-.11	.	-.10	.
FSOC	.73	.	.55	.	.64	.	.44	.
MSOC	.02	.	.08	.	-.46	.	.41	.
MQUALIF	.75	1.5	.63	.	.82	1.6	.75	1.5
FQUALIF	.78	1.8	.80	1.9	.87	2.1	.	2.2
UNEMPLO	-1.9	.	-2.1	.	-2.3	.	-2.6	.
WORKMUM	-1.3	-1.6	-1.2	-1.6	-1.2	-1.6	-1.1	-1.5
PARHLTH	.35	.	.39	.	.61	.	.66	.
PARMENT	.31	.	.33	.	.33	.	.31.	.
TOTCIGS	.00	.	.01	.	-.01	.	.00	.
CARPHON	-1.7	-1.8	-1.6	-1.7	-1.7	-1.8	-1.7	-1.9
CONSUMER	.59	.	.67	.	.65	.	.76	.
OCCUPRA	.22	.	.08	.	.34	.	.01	.
FAMSIZE	-.89	.	.75	.	-.67	.	-.35	.
BIRTHOR	.00	.	-.14	.	-.01	.	-.22	.
GESTAT	-.45	-2.6	-4.5	-2.6	-4.6	-2.6	-4.4	-2.5
BRTHWT	-1.3	.	-1.4	.	-.64	.	-1.0	.
BIRTHSCO	.13	.	.23	.	.29	.	.41	.
MEDHIST	-.31	.	-.67	.	-.66	.	-1.1	.
STHEIGHT	1.5	2.9	1.3	2.4	1.4	2.7	1.1	2.1
OFFSCHL	-0.9	-1.7	-.10	-1.8	-.11	-1.9	-.12	-2.2
CHILDINT	2.1	4.6	2.1	4.7	2.2	4.8	2.2	4.8
PARCHCOM	-.22	.	-.20	.	-.06	.	-.02	.
PARPART	.52	.	.47	.	.32	.	.28	.
PARSCHL	-1.2	-2.8	-1.1	-2.6	-1.1	-2.6	-1.0	-2.4
PVOC	.21	3.3	.20	3.2	.18	2.8	.17	2.7
PMAT	.23	3.6	.24	3.8	.24	3.7	.25	3.9

NOTE: t values lower than 1.5 in absolute value are shown by "."

Comparing the multivariate regressions with the univariate associations, we can see that the variables AGEINT, STHEIGHT, OFFSCHL, CHILDIN, GESTAT, PVOC and PMAT maintain their associations with BASC in the multivariate regression. The variables MSOC and FSOC, PARPART, PARSCHL, BRTHWT and BIRTHSCO and TOTCIGS no longer show significant relationships with BASC after controlling for the other variables. The effect of FQUALIF and MQUALIF is still apparent in the multiple regression, but is considerably reduced. The direction of the association between BASC and each of the three variables CLASSYR, CARPHON and PARSCHL is reversed in the multiple regressions.

The results of such multiple regressions must be interpreted with caution when, as here, there are high correlations between the x variables. Two correlated factors may each appear unimportant in a multiple regression, where either together or separately they make a considerable contribution to the regression. The way in which the inclusion of one variable affects the coefficients of the others will be further described in the context of the stepwise analyses.

Appendix to chapter 5.

Items contributing to constructed scores.

BIRTHSCO Type of delivery, admission to a Special Care Baby Unit, duration of hospital stay.

MEDHIST Hospital admission for head injury, history of fits, presence of chronic or recurrent illness.

PARHLTH History of chronic illness or accident, general practitioner, outpatient or inpatient care, scored for both parents and combined. Single parent score is doubled.

PARMENT History of depression, anxiety or other psychiatric problem, general practitioner outpatient or inpatient hospital care, prescription of psychotropic drugs, chronic mental illness in last 10 years. Scored for both parents and combined. Single parent score is doubled.

FAMHIST Measures departure from nuclear family using loss of natural parent(s), most recent disruption in carers, age of child when these events occurred, current carers, father working away from home.

CHILDIN Assesses child's regular activities based on number of books in the home, use of library, attendance at organised activities including recreational/sporting, artistic/musical and instructional/educational activities, frequency of these activities. A higher score indicates a wider range or higher frequency of activity.

PARCHCO Talking with parent(s) about school or study tests and games, supervision of homework, read stories by parents. A higher score indicates a higher degree of communication.

PARPART Joint activities with a parent, in sports, outings, indoor games, reading stories, annual holiday, meals together. A higher score indicates a higher degree of parental participation.

PARSCHL Self-initiated parental visits to school, attendance at school and parents' meetings, child's visit to school with parent before starting, discussing child's progress at school. A higher score indicates more parental involvement.

Chapter 6

Stepwise procedures based on the residual sum of squares for the lead-study data

6.1 Adjusted and unadjusted data

The lead study data include the blocking factor, "school", as well as the other covariates. We saw in the last chapter that entering the factor "school" into the regression equation produced a large shift in the lead coefficient, to a value similar to that achieved with all the covariates. However, most of the covariates still retained some predictive power for the BASC scores after the factor "school" was fitted. This allows us to treat the data in two different ways, in looking at the results of the variable selection procedures:

(1) Ignore the factor "school", and examine the influence of the selection of covariates, taking the corresponding "full model" as that which contains lead and covariates only. For this case the unadjusted lead coefficient is -5.45 and the full model the lead coefficient is -3.18 (see Table 5.7)

(2) Consider the variable selection process as starting after the factor "school" has been fitted. This is equivalent to considering the residuals of BASC from the factor "school", with the partial residuals of the covariates from the factor "school" as the set of covariates. The unadjusted lead coefficient is then -3.89 and the full model lead coefficient is -3.81 (see Table 5.7).

These two cases will be referred to as the "unadjusted" and "school adjusted".

6.2 The forward stepwise procedure

In chapters 3 and 4 criteria were introduced, both for prediction (C_p and S_p) and for the MSE of the lead coefficient (the two G_p criteria). If these were combined with the two methods of treating the data (section 6.1) and the large number of possible search strategies which they might provide, it is clear that this part of this thesis could be extended almost indefinitely. However, I will describe the results of only a few of the analyses which I performed, and will use these to illustrate the apparent properties of the statistics discussed.

To gain an overall impression of how the various criteria behave, this chapter will report their values for the usual forward stepwise regression procedure, selecting the next variable at each step which minimises the residual sum of squares. The constant term and the blood-lead variable will be included first in all cases.

Tables 6.1 and 6.2 give the results for the prediction criteria for the unadjusted and school-adjusted data. The column RMS_p is the residual-mean-square, or $RSS_p/(n-p)$. The value of the F-ratio is the F-to-enter value for the variable which is entering the equation at each step. The estimate of β^* for each model is also given, along with the corresponding t statistic, calculated from that model as if it were the correct one.

Table 6.1: Prediction criteria for forward stepwise procedure, unadjusted data

Variable	p	F ratio	RMS_p	C_p	S_p	b_p*	t-value
X* only	2		174.80	334.51	0.35100	-5.45	-3.53
PVOC	3	177.0	129.23	118.47	0.26001	-3.18	-2.38
CHILDINT	4	37.11	120.49	77.85	0.24292	-3.04	-2.35
PMAT	5	25.33	114.87	52.11	0.23205	-3.51	-2.77
AGEINT	6	15.32	111.64	37.80	0.22600	-3.98	-3.17
FQUALIF	7	12.66	109.07	26.64	0.22124	-3.71	-2.99
GESTAT	8	10.75	106.96	17.67	0.21740	-3.49	-2.83
PARSCHL	9	5.70	105.95	13.92	0.21579	-3.26	-2.65
STHEIGHT	10	6.09	104.87	9.83	0.21401	-2.69	-2.16
SEX	11	4.37	104.15	7.48	0.21298	-2.86	-2.30
OFFSCHL	12	3.00	103.73	6.52	0.21256	-3.01	-2.42
WORKMUM	13	2.33	103.45	6.23	0.21242	-2.98	-2.40
CARPHONE	14	1.97	103.24	6.29	0.21243	-2.98	-2.40
HANDED	15	2.35	102.96	5.99	0.21228	-3.02	-2.43
MQUALIF	16	2.37	102.67	5.67	0.21212	-2.89	-2.33
CONSUMER	17	1.45	102.57	6.25	0.21237	-2.97	-2.39
PARPART	18	1.15	102.54	7.13	0.21274	-2.94	-2.37
FSOC	19	0.92	102.56	8.22	0.21322	-2.84	-2.28
FAMSIZE	20	1.01	102.55	9.24	0.21366	-2.80	-2.25
CLASSYR	21	0.92	102.57	10.34	0.21414	-2.93	-2.33
UNEMPLOY	22	1.03	102.56	11.33	0.21457	-2.97	-2.37
PARMENT	23	0.53	102.66	12.80	0.21523	-2.97	-2.36
PARHLTH	24	0.39	102.80	14.42	0.21596	-2.98	-2.37
MEDHIST	25	0.29	102.95	16.14	0.21674	-3.06	-2.42
PARCHCOM	26	0.32	103.10	17.82	0.21750	-3.06	-2.41
BRTHWT	27	0.22	103.27	19.60	0.21832	-3.06	-2.41
TIMEDAY	28	0.22	103.44	21.39	0.21914	-3.11	-2.44
FAMHIST	29	0.12	103.63	23.27	0.22002	-3.13	-2.45
BRTHSCO	30	0.11	103.82	25.15	0.22090	-3.15	-2.46
TOTCIGS	31	0.09	104.02	27.06	0.22180	-3.17	-2.47
BIRTHOR	32	0.03	104.24	29.03	0.22273	-3.19	-2.48
MOVESCHL	33	0.01	104.46	31.01	0.22368	-3.18	-2.46
MSOC	34	0.01	104.68	33.00	0.22463	-3.18	-2.46
OCCUPRAT	35	0.00	104.90	35.00	0.22560	-3.18	-2.45

Table 6.2: Prediction criteria for forward stepwise procedure, school-adjusted data

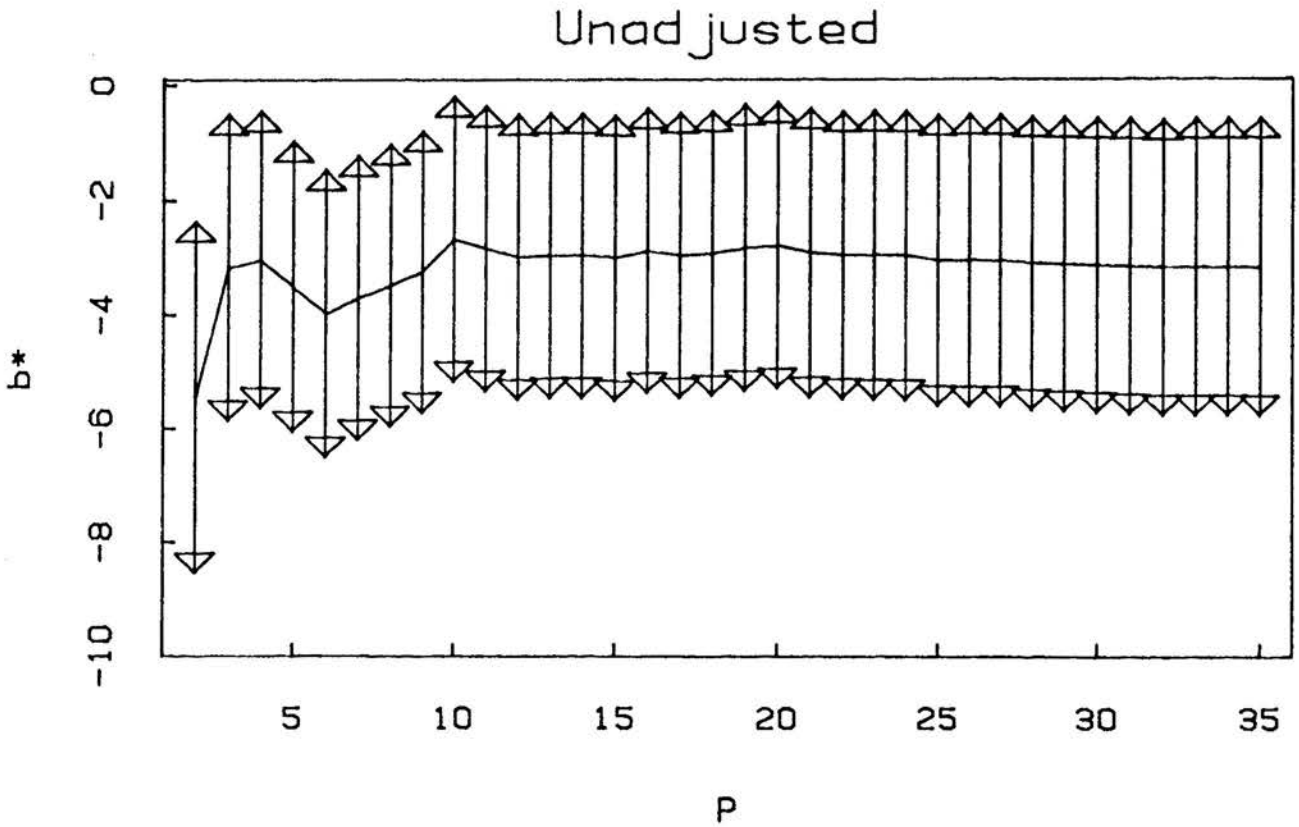
Variable	p	F ratio	RMS_p	C_p	S_p	b_p*	t-value
X* only	19		142.10	154.17	0.29543	-3.89	-2.56
PVOC	20	82.57	121.57	130.81	0.25327	-3.26	-2.32
PMAT	21	26.86	115.37	101.45	0.24085	-3.59	-2.61
CHILDINT	22	21.42	110.66	79.46	0.23150	-3.44	-2.56
AGEINT	23	19.61	106.52	60.32	0.22331	-3.93	-2.97
FQUALIF	24	12.82	103.95	48.83	0.21838	-3.83	-2.93
GESTAT	25	9.68	102.09	40.84	0.21493	-3.60	-2.77
PARSCHL	26	5.06	101.23	37.65	0.21357	-3.43	-2.65
CLASSYR	27	4.75	100.44	34.83	0.21234	-3.53	-2.74
OFFSCHL	28	5.14	99.57	31.65	0.21095	-3.84	-2.97
SEX	29	4.48	98.84	29.17	0.20986	-4.03	-3.12
STHEIGHT	30	4.61	98.09	26.59	0.20871	-3.48	-2.66
CARPHONE	31	3.04	97.67	25.58	0.20825	-3.59	-2.74
WORKMUM	32	2.35	97.39	25.26	0.20809	-3.59	-2.75
MQUALIF	33	2.29	97.12	25.01	0.20797	-3.54	-2.71
HANDED	34	2.44	96.82	24.63	0.20778	-3.59	-2.75
CONSUMER	35	2.01	96.61	24.66	0.20777	-3.70	-2.83
UNEMPLOY	36	1.47	96.52	25.23	0.20801	-3.78	-2.89
PARHLTH	37	1.00	96.52	26.25	0.20846	-3.80	-2.91
MEDHIST	38	0.99	96.52	27.28	0.20891	-3.96	-3.01
PARPART	39	0.77	96.57	28.53	0.20947	-3.94	-2.99
FSOC	40	0.45	96.68	30.09	0.21018	-3.85	-2.91
PARMENT	41	0.46	96.80	31.64	0.21089	-3.82	-2.88
BRTHSCO	42	0.43	96.92	33.22	0.21161	-3.85	-2.90
MSOC	43	0.37	97.05	34.85	0.21236	-3.84	-2.89
FAMSIZE	44	0.27	97.20	36.59	0.21317	-3.81	-2.86
MOVESCHL	45	0.21	97.37	38.38	0.21401	-3.74	-2.79
BRTHWT	46	0.15	97.55	40.23	0.21488	-3.76	-2.80
BIRTHOR	47	0.10	97.75	42.14	0.21578	-3.79	-2.81
TIMEDAY	48	0.07	97.95	44.07	0.21670	-3.81	-2.82
FAMHIST	49	0.06	98.15	46.01	0.21763	-3.81	-2.82
TOTCIGS	50	0.01	98.37	48.00	0.21860	-3.82	-2.81
PARCHCOM	51	0.00	98.59	50.00	0.21957	-3.82	-2.80
OCCUPRAT	52	0.00	98.81	52.00	0.22055	-3.81	-2.79

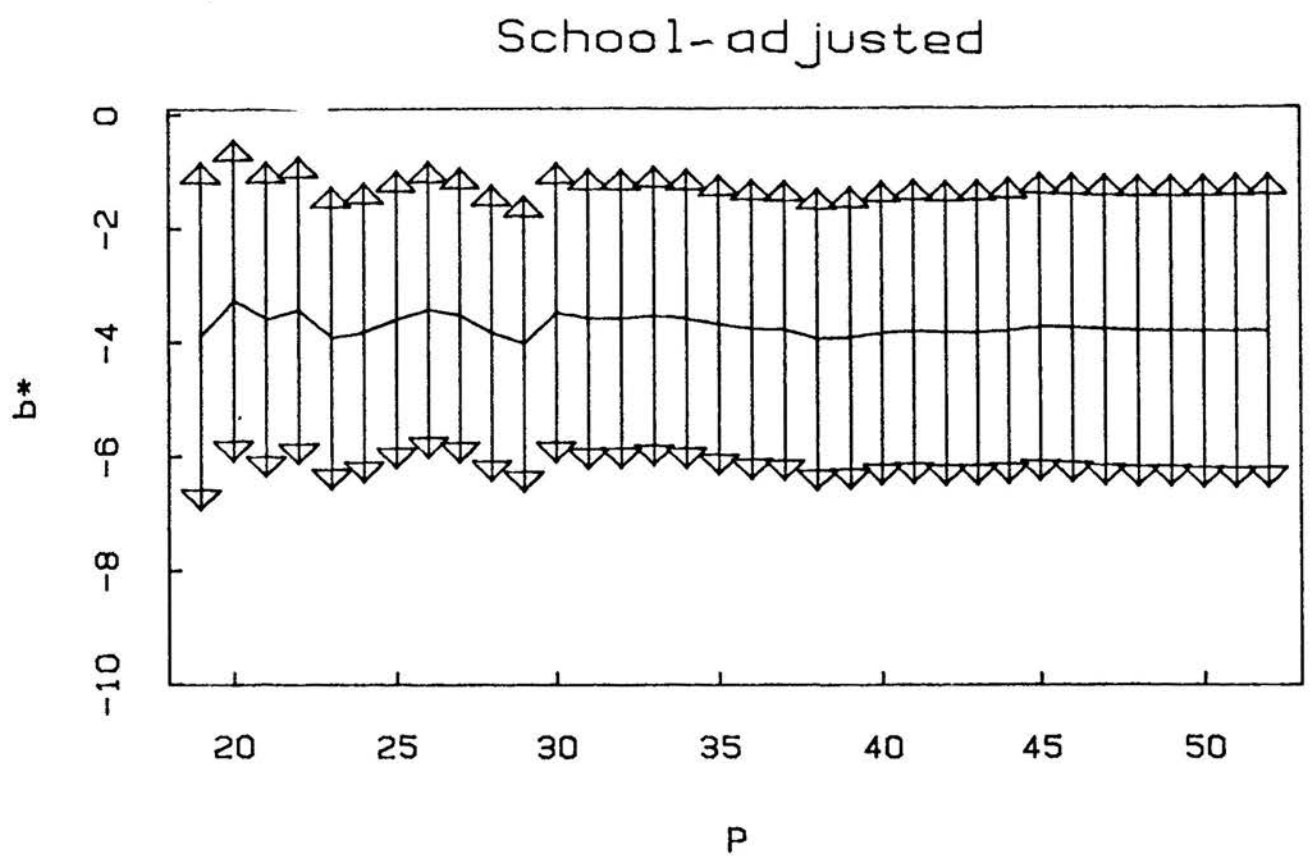
The order in which the variables enter the equation for the unadjusted and adjusted data is remarkably similar. The two tests of parent's ability and the score for child's interest enter first, followed by AGE, FQUALIF, GESTAT and PARSCHL. These seven variables are the first to enter in both cases, and except for the reversal of the second and third, they enter in the same order. For these first seven variables the F-to-enter statistics are considerably greater for the unadjusted data. After the first seven variables there is much more variation in the order in which variables enter, and the F-to-enter statistics are not always larger for the unadjusted data.

For the unadjusted data the estimates coefficient b* is reduced in absolute value from -5.45 to -3.18 at the first step, which is almost identical to the value for the full model with 33 covariates. After the first step the estimated coefficient fluctuates until around the 10th step, when it settles down around -3.0 and then gradually rises to -3.18 over the last 10 steps.

The school-adjusted coefficient of -3.89 is modified to -3.26 at the first step, and at subsequent steps tends to drift back towards the value for the unadjusted data. Its fluctuations are somewhat less than for the unadjusted data, and it settles down to a value of about -3.8 after the 16th step.

Figure 6.1 : Estimates of β* from a forward stepwise procedure, along with their 95% confidence limits





These coefficients and their confidence intervals, calculated in the usual manner are shown in figure 6.1. At every step the coefficient of blood lead would be judged statistically significant by a conventional test at the 5% significance level.

6.3 The criteria C_p and S_p

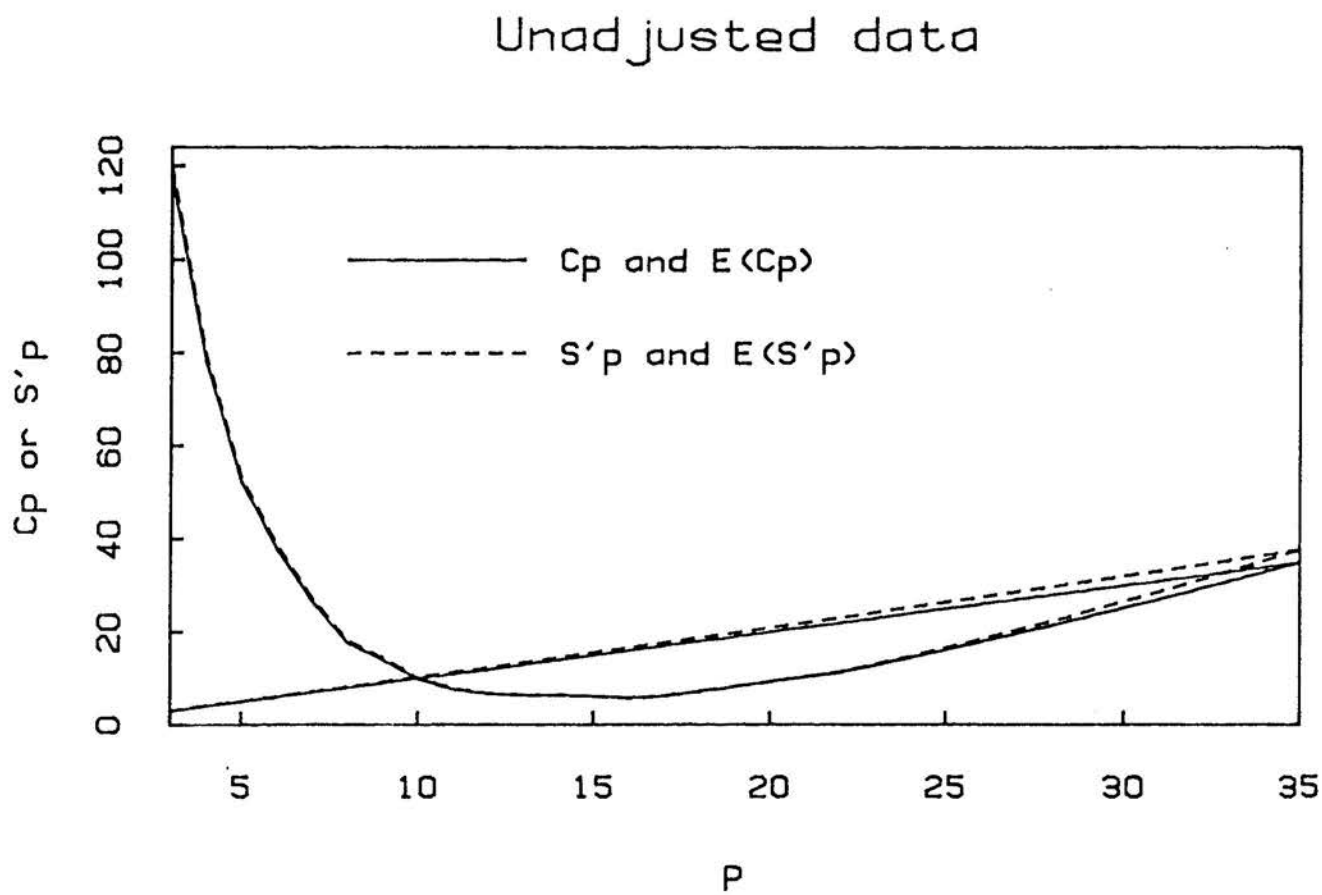
The similar behaviour of the two criteria C_p and S_p is apparent for both the adjusted and unadjusted data. For the unadjusted data the minima of both criteria are achieved at the 14th step when p=16. For the adjusted data the minima are at the 15th and 16th step for C_p and S_p respectively. Note also that the minima for these criteria, and even the local minimum for the unadjusted data, occur at the last step for which the F-to-enter statistic exceeds 2. Thus the asymptotics discussed in chapter 3 seem to make some sense.

The comparison of C_p and S_p is facilitated by rescaling S_p in the same manner as C_p is scaled. The expression which has expectation equal to the MSE of prediction in terms of S_p is $S_p(n+1)(n-3)/n$. If this is adjusted in the same manner as C_p is obtained from the expression (3.2) we obtain

$$S'_p = \frac{(n+1)(n-3)}{n\, s^2} S_p - n ,$$

which is in equivalent units to C_p. The equivalence of the two criteria is seen in the plots of C_p and S'_p against p in figure 6.2, and the way in which they have identical small fluctuations

Figure 6.2: Plots of C_p and S'_p against p



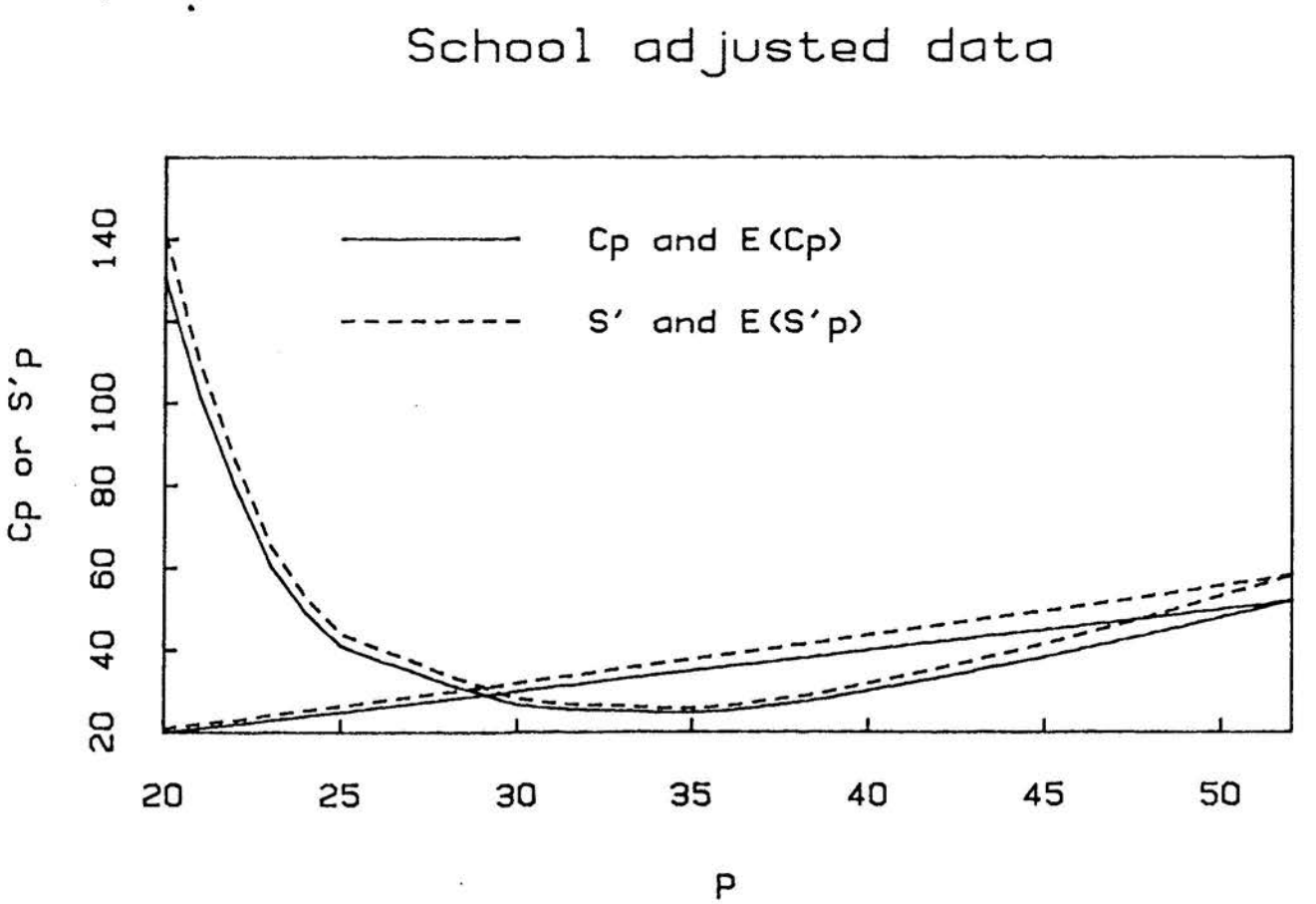
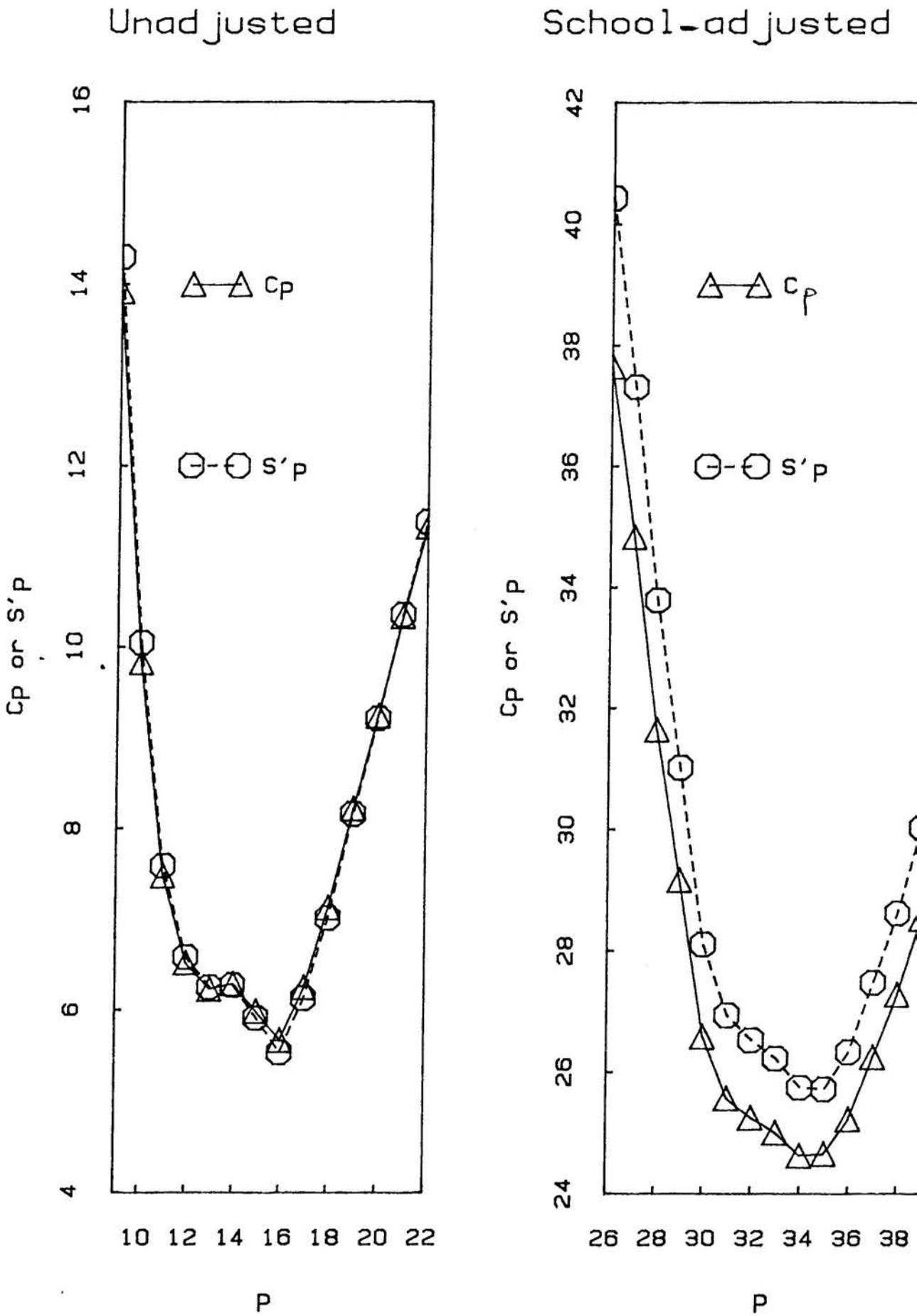


Figure 6.3: Enlargement of part of Figure 6.2



is seen clearly in figure 6.3, which is a section of the previous figure plotted with an exaggerated y scale.

The S'_p criterion is larger than C_p for the larger values of p associated with the adjusted data. When the difference between the two criteria ($S'_p - C_p$) is expanded in terms of order 1/n, the leading term becomes $((p+1)^2-(p+3))/n$ times s_p^2/s^2, which explains the larger values of S'_p for the unadjusted data, for which $(p+1)^2$ is of the same order as n.

For the school adjusted data, considering the dummy variables for "school" as if they are normally distributed variables seems particularly inappropriate. It would be better to take a mixed model for which the factor "school" is considered fixed, and in which the subsequent variables are assumed to have a joint distribution. This would produce a criterion which would be a combination of C_p (for the fixed effects) and S_p for the random effects. This would be straightforward, but has not been developed here, because prediction is not the main focus of interest.

If the covariates omitted from the model have no value in predicting y (ie the βs are all identically zero) the expected value of C_p would be p. This is the justification for the C_p against p plot. By a similar argument, the value of S'p when no further variables have any predictive power becomes $(n+1)(n-2)/(n-p-1)-n$. The lines for these expected values of the two criteria are also shown in figure 6.2.

Both the C_p and S'_p plots start above their expected values, and then dip well below them, and finally approach their expected values (as they must do) as all the covariates are entered. This seems to be a feature of the C_p plots which have been published for other data eg Hocking 1976 (p24, 25 &27), Thompson (1984b p143), Mallows in the discussion of Mitchell & Beauchamp (1988 p1035) and the C_p values quoted by Pocock et al (1987) for Smith et al's data which attain a minimum value of 8.5 when p has the value 12. These plots alone should make us realise that what we are observing, in searching for low values of the criteria, are the minima from a distribution of all possible combinations of covariates which contribute nothing further to the prediction of y. Of course, we cannot be assured that the forward stepwise procedure used here has attained the lowest possible value of either of these criteria. However, since we have already obtained values which are doing better than they should, any further search would be likely to be investigating irrelevant noise.

6.4 The G_p criteria

Table 6.3 gives the values of the criteria G_{Fp}, G_{Rp} and their constituent parts for the forward stepwise procedure for the unadjusted data. The random-effects model for the school-adjusted data treats the "school" dummy variables as fixed effects, and the other covariates as random effects. The results for the MSE criteria for this model are given in table 6.4.

Table 6.3: MSE criteria for the unadjusted data, selection by forward stepwise procedure

	p	Est(Bias^2)	v_{FP}	G_{FP}	$V1_{RP}$	$V2_{RP}$	G_{RP}
none	2	4,9070	1,4323	6,3394	2,3868	0,0	7,2939
PVOC	3	-0,2288	1,4559	1,2271	1,7680	0,0255	1,5647
CHILDINT	4	-0,2076	1,4564	1,2488	1,6518	0,0211	1,4652
PMAT	5	-0,1134	1,4644	1,3510	1,5779	0,0257	1,4901
AGEINT	6	0,4282	1,4780	1,9062	1,5367	0,0362	2,0012
FQUALIF	7	0,0781	1,4834	1,5616	1,5043	0,0380	1,6205
GESTAT	8	-0,1030	1,4879	1,3849	1,4782	0,0389	1,4141
PARSCHL	9	-0,1812	1,4965	1,3153	1,4673	0,0442	1,3303
STHEIGHT	10	0,1106	1,5514	1,6620	1,4552	0,0957	1,6614
SEX	11	-0,0243	1,5584	1,5341	1,4482	0,0990	1,5229
OFFSCHL	12	-0,0892	1,5656	1,4764	1,4453	0,1028	1,4589
WORKMUM	13	-0,0789	1,5659	1,4870	1,4443	0,0999	1,4653
CARPHONE	14	-0,0772	1,5660	1,4887	1,4444	0,0967	1,4640
HANDED	15	-0,0916	1,5667	1,4751	1,4434	0,0942	1,4461
MQUALIF	16	-0,0283	1,5735	1,5452	1,4424	0,0976	1,5117
CONSUMER	17	-0,0620	1,5775	1,5154	1,4440	0,0985	1,4804
PARPART	18	-0,0504	1,5780	1,5276	1,4465	0,0960	1,4921
FSOC	19	0,0197	1,5895	1,6092	1,4498	0,1042	1,5737
FAMSIZE	20	0,0462	1,5908	1,6370	1,4528	0,1024	1,6014
CLASSYR	21	-0,0137	1,6072	1,5935	1,4560	0,1155	1,5578
UNEMPLOY	22	-0,0309	1,6088	1,5779	1,4590	0,1140	1,5421
PARMENT	23	-0,0312	1,6088	1,5776	1,4635	0,1111	1,5433
PARHLTH	24	-0,0355	1,6092	1,5737	1,4684	0,1084	1,5414
MEDHIST	25	-0,0375	1,6334	1,5959	1,4737	0,1293	1,5655
PARCHCOM	26	-0,0359	1,6336	1,5976	1,4789	0,1265	1,5695
BRTHWT	27	-0,0362	1,6336	1,5973	1,4845	0,1236	1,5718
TIMEDAY	28	-0,0344	1,6451	1,6108	1,4901	0,1321	1,5878
FAMHIST	29	-0,0335	1,6486	1,6151	1,4960	0,1325	1,5950
BRTHSCO	30	-0,0304	1,6536	1,6232	1,5020	0,1345	1,6062
TOTCIGS	31	-0,0261	1,6587	1,6326	1,5081	0,1366	1,6187
BIRTHOR	32	-0,0164	1,6683	1,6519	1,5145	0,1432	1,6413
MOVESCHL	33	-0,0042	1,6805	1,6763	1,5209	0,1525	1,6692
MSOC	34	-0,0024	1,6824	1,6800	1,5274	0,1514	1,6764
OCCUPRAT	35	0,0	1,6848	1,6848	1,5340	0,1508	1,6848

Table 6.4: MSE criteria for the school-adjusted data, selection by forward stepwise procedure.

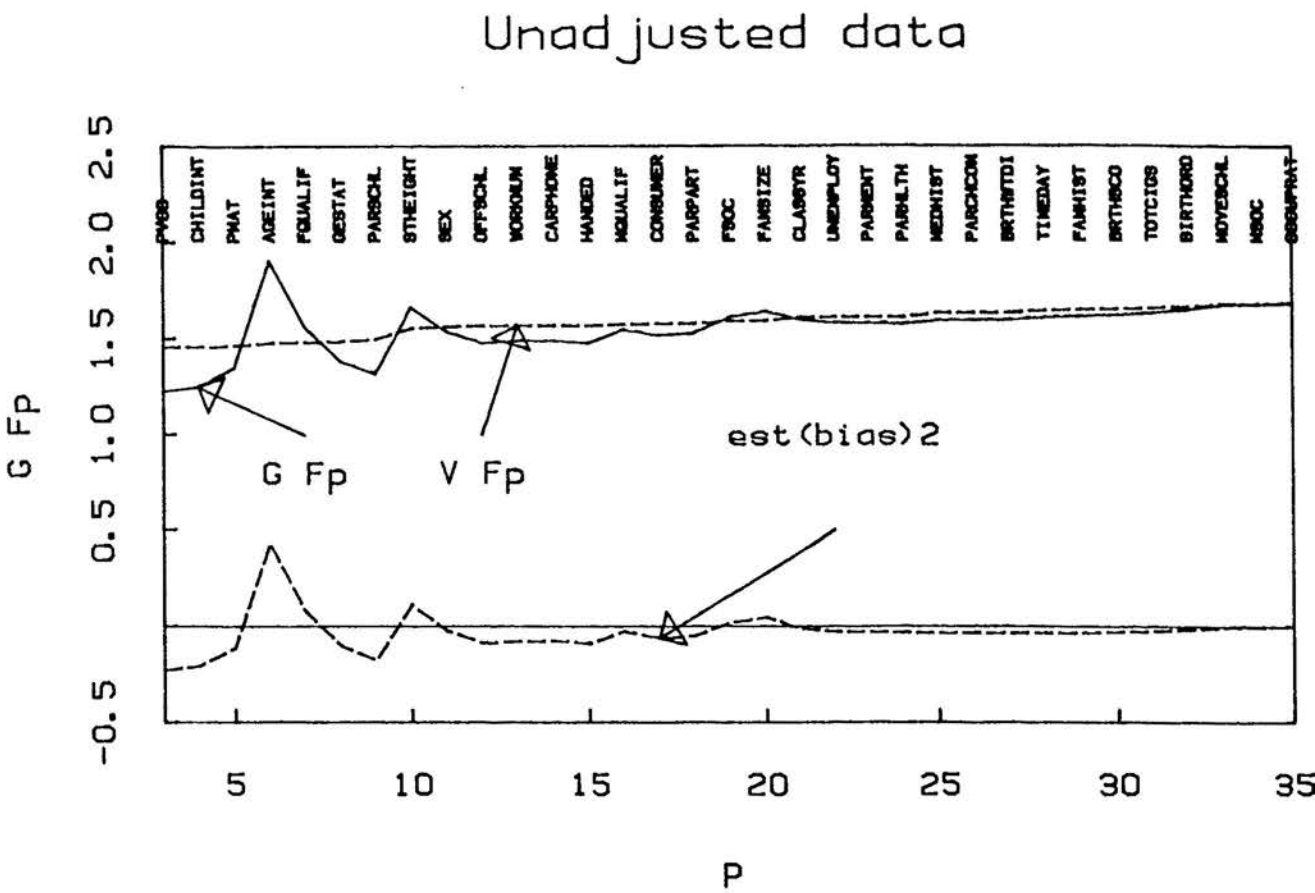
	p	Est(Bias^2)	V_{FP}	G_{FP}	$V1_{RP}$	$V2_{RP}$	G_{RP}
none	19	-0,2560	1,6061	1,3501	2,3105	0,0	2,0545
PVOC	20	0,0487	1,6100	1,6587	1,9802	0,0007	2,0296
PMAT	21	-0,2033	1,6134	1,4101	1,8831	0,0007	1,6804
CHILDINT	22	-0,1121	1,6143	1,5022	1,8100	-0,0020	1,6958
AGEINT	23	-0,2290	1,6257	1,3967	1,7460	0,0066	1,5236
FQUALIF	24	-0,2410	1,6263	1,3853	1,7074	0,0036	1,4700
GESTAT	25	-0,1918	1,6317	1,4399	1,6805	0,0055	1,4942
PARSCHL	26	-0,0809	1,6376	1,5567	1,6698	0,0081	1,5969
CLASSYR	27	-0,1493	1,6400	1,4906	1,6602	0,0069	1,5177
OFFSCHL	28	-0,2094	1,6578	1,4484	1,6493	0,0213	1,4612
SEX	29	-0,1563	1,6659	1,5096	1,6407	0,0257	1,5102
STHEIGHT	30	-0,0301	1,7300	1,7000	1,6318	0,0857	1,6874
CARPHONE	31	-0,0834	1,7335	1,6501	1,6282	0,0854	1,6301
WORKMUM	32	-0,0842	1,7335	1,6493	1,6270	0,0817	1,6244
MQUALIF	33	-0,0591	1,7346	1,6754	1,6260	0,0790	1,6458
HANDED	34	-0,0829	1,7356	1,6527	1,6245	0,0763	1,6179
CONSUMER	35	-0,1128	1,7412	1,6284	1,6245	0,0781	1,5897
UNEMPLOY	36	-0,1206	1,7460	1,6253	1,6263	0,0792	1,5849
PARHLTH	37	-0,1212	1,7465	1,6253	1,6298	0,0762	1,5849
MEDHIST	38	-0,0728	1,7729	1,7002	1,6334	0,0985	1,6591
PARPART	39	-0,0780	1,7735	1,6955	1,6378	0,0956	1,6553
FSOC	40	-0,0736	1,7929	1,7192	1,6433	0,1111	1,6807
PARMENT	41	-0,0735	1,7942	1,7207	1,6488	0,1089	1,6842
BRTHSCO	42	-0,0698	1,7963	1,7264	1,6544	0,1075	1,6921
MSOC	43	-0,0702	1,7968	1,7266	1,6603	0,1045	1,6946
FAMSIZE	44	-0,0668	1,8009	1,7341	1,6666	0,1051	1,7049
MOVESCHL	45	-0,0410	1,8217	1,7807	1,6732	0,1221	1,7543
BRTHWT	46	-0,0407	1,8245	1,7838	1,6800	0,1214	1,7606
BIRTHOR	47	-0,0367	1,8304	1,7937	1,6871	0,1237	1,7741
TIMEDAY	48	-0,0247	1,8431	1,8184	1,6943	0,1328	1,8024
FAMHIST	49	-0,0242	1,8435	1,8193	1,7016	0,1298	1,8071
TOTCIGS	50	-0,0120	1,8557	1,8437	1,7091	0,1384	1,8355
PARCHCOM	51	-0,0090	1,8587	1,8497	1,7167	0,1379	1,8456
OCCUPRAT	52	0,0	1,8678	1,8678	1,7244	0,1434	1,8678

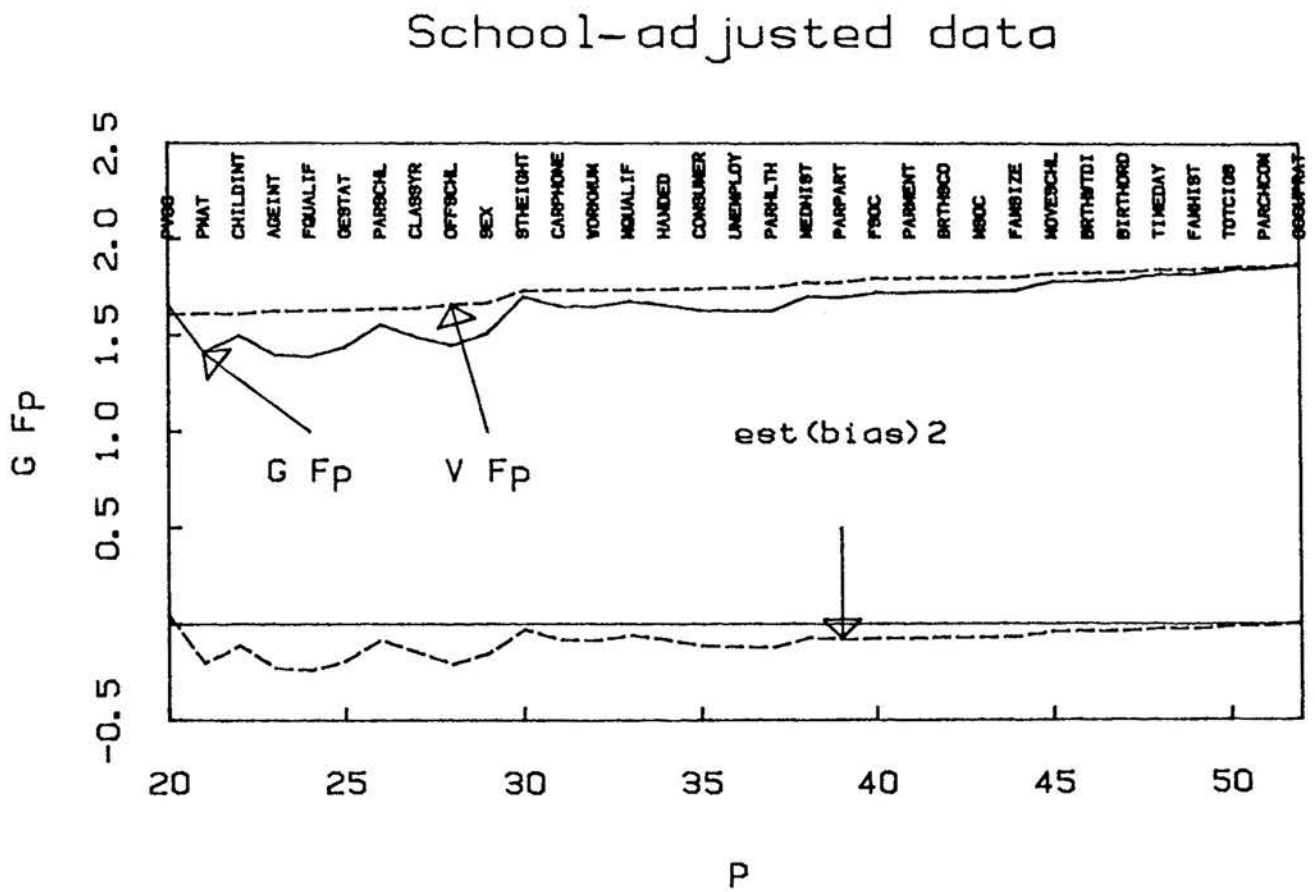
The term Est(bias)^2 is part of both criteria and is evaluated from the final line of expression 4.16. It is the square of the estimated bias minus its estimated variance, and hence has the square of the true bias as its expected value. There is a single variance term V_{FP} for the first criterion, and two such terms $V1_{RP}$ and $V2_{RP}$ for the second criterion which are defined in expression 4.16. The term $V2_{RP}$ depends on the correlations between the x variables included in the model and X*.

The only one of these quantities which requires modification for the model with fixed dummy variables for schools is $V1_{RP}$. The factor (n-2)(n-3) in the middle term of expression 4.16 is replaced by (n-d-1)(n-d-2), where d is the number of levels of the factor school (18 here). The terms with suffices x* in expression 4.16 now refer to the model which contains X* and the factor "school".

The first point to notice is that the Est(bias)^2 term obtains substantial negative values for a large number of the models considered. In particular, this is very marked for the first step of the unadjusted data, and occurs for every step except the first for the school-adjusted data. This comes about because the values of b* for these models happen to be almost identical to the value for the full model, so that the value of the squared bias is much smaller than would be expected for the case when the true value of the bias is zero. Where the Est(bias)^2 term is negative it would suggest that for this model the G_P criteria will be underestimating the MSE of β^*.

Figure 6.4: Values of the G_p criteria and their constituent parts for forward stepwise procedure





The values of the G_p and their constituent parts are plotted on figure 6.4. The criteria always attain their lowest values, for the models examined, when this Est(bias^2) term is substantially negative. For the unadjusted data G_{FP} attains its lowest value after the first covariate is entered, and for G_{RP} it is attained after 7 covariates are entered, since by this step the term Vi_{RP} has fallen substantially. For the adjusted data the lowest value of G_{FP} is obtained with no additional covariates in the model, and for G_{RP} after 9 covariates. These are not minima of the criteria in relation to the values for the adjacent steps, because the values of the G_p tend to fluctuate as the estimate of b^* varies about b^*_{full}.

The variance parts of the G_p criteria tend to behave more regularly. The variance part of G_{FP} is strictly increasing with p, as was proved in chapter 4. It shows the most pronounced increase when the variable STHEIGHT is added into the models, which is the term that had the strongest univariate relationships with blood lead (see Tables 5.3 and 5.6). The first variance term for G_{RP} attains a minimum at the same point as the criterion S_p, since it differs from it only by a constant factor. It is thus highly likely that it, too, will be affected by selection bias, especially when, as here, the models with the minimum residual sums of squares are selected. The second variance term of G_{RP} makes a much smaller contribution for these data. The variables which cause it to increase tend to be those which are most strongly related to X* (see tables 5.3 and 5.6). The small values

of this component here relate to the weak associations in these data between the covariates and X*.

These results suggest that any method which simply searches for low values of the G_p criteria will be picking out sub-models which happen, by chance, to give exactly the same value of b* as the full model for the particular data being examined, and that such models will not have values for the MSE which are as low as the criteria suggest. An alternative would be to count the variance term as equal to zero whenever it is estimated with a negative quantity. For the unadjusted data G_{FP} would still have a minimum at the first step, but G_{RP} now attains a minimum at the 13th step instead of the 6th. The minima achieved would be considerably greater in each case. The results for the school-adjusted data are similar. The same strategy could be used if the second variance term in G_{RP} became negative, although this only occurred once in the data which we have here. Although such procedures would give estimates of the MSE of b* which will be biassed upwards for a model selected without reference to the data, they should have a reduced variance and may prove to be better criteria to use in search procedures or as stopping rules.

6.5 Significance test approaches to MSE

We can apply a significance testing approach, following Toro-Vizcarrando & Wallace as discussed in section 4.2, to the models examined in the stepwise procedure. The F ratio for the bias of

the coefficient of special interest will always be less than 1 when the Est(bias^2) term in tables 6.3 and 6.4 is negative.

For the unadjusted data the value of F for the model with no covariates is 20.4, which is well beyond the 1% point for both the central and non-central F distribution. Thus we have evidence of significant bias compared to the full model and evidence that the estimate of MSE for the blood lead coefficient is significantly greater for the reduced model compared to the full model. Subsequently the F statistic has its greatest value of 3.07 at the 4th step, after AGEINT is entered. This only exceeds the 25% point for the non-central F distribution and the 10% point for central F, so the evidence for a biassed blood lead coefficient is weak, and for an increased MSE for these reduced models is even weaker. At the 4th step the value of G_{FP} is greater than for the full model. This agrees with the comment in section 4.2 that the F statistic approach will treat the reduced model more favourably than G_{FP}. At no other step does the F statistic exceed 2.

For the school-adjusted data the F statistic is less than 1 for the model with no covariates, and at every subsequent step except the first when its value is 1.07. Thus there is no evidence of significant bias for any of the models considered, and this obviously implies no evidence of an increased MSE for b*. Thus all the models examined would be preferred to the full model. This agrees with the fact that value of G_{FP} is less than that for the full model at every step for the school-adjusted data.

6.6 Changes in the coefficients of the other covariates

The influences of the other covariates on the BASC score are modified as each additional term is added to the model. Tables 6.5 and 6.6 give the t-values of variables for the first 10 steps and the full model, for the adjusted and unadjusted data. When a variable has not been entered into the equation the t-value corresponds to the coefficient which that variable would have if it were to be the next to enter the equation. When a variable is already in the equation the t-value corresponds to the coefficient, after the variable which is entered at this step has been included. In each case the t-values are calculated as the ratio of the coefficient to its standard error, as if the model for which it has been calculated were the correct one. The variables have been ordered in the sequence in which they enter the regression.

Those variables for which no t-statistic exceeds $\sqrt{2}$ in absolute value at any of the 33 steps have been excluded from these tables. They are

unadjusted data : CONSUMER, PARMENT, PARHLTH, TIMEDAY, FAMHIST, MOVESCH

school-adjusted data : CONSUMER, PARMENT, PARHLTH, TIMEDAY, FAMHIST, MOVESCH, BRTHSCO, FAMSIZE.

Table 6.5 : Changes in the t-values for the covariates during forward stepwise regression, blood lead entered first. The rows correspond to variables and the columns to steps. Bold type indicates that the variable is included in the model

ENTERS STEP	COEFFICIENT FOR	Variable being entered at this step									
		PVOC 1	CHINT 2	PMAT 3	AGE 4	FQUAL 5	GESTA 6	PARSC 7	STHGH 8	SEX 9	OFFSC 10	FULL
1	PVOC	**13.31**	**10.69**	**6.88**	**6.85**	**4.61**	**4.72**	**4.99**	**4.92**	**4.89**	**4.81**	**3.25**
2	CHINT	*9.58*	**6.09**	**5.37**	**5.70**	**4.93**	**5.27**	**5.57**	**5.15**	**5.42**	**5.12**	**4.68**
3	PMAT	*11.79*	*5.79*	**5.03**	**4.61**	**4.00**	**3.90**	**3.99**	**3.87**	**3.98**	**3.96**	**3.80**
4	AGE	*-4.45*	*-3.93*	*-4.39*	**-3.91**	**-3.90**	**-3.96**	**-3.99**	**-3.85**	**-3.84**	**-3.89**	**-3.03**
5	FQUAL	*12.15*	*5.47*	*4.32*	*3.58*	**3.56**	**3.44**	**3.75**	**3.72**	**3.67**	**3.66**	**1.89**
6	GESTAT	*-2.44*	*-2.86*	*-3.48*	*-3.32*	*-3.40*	**-3.28**	**-3.53**	**-3.47**	**-3.41**	**-3.44**	**-2.56**
7	PARSCHL	*3.70*	*-0.08*	*-1.11*	*-1.42*	*-1.46*	*-2.00*	**-2.39**	**-2.49**	**-2.59**	**-2.54**	**-2.60**
8	STHGH	*5.48*	*3.90*	*2.99*	*2.72*	*2.52*	*2.46*	*2.37*	**2.47**	**2.57**	**2.59**	**2.36**
9	SEX	*-0.39*	*-0.79*	*-1.81*	*-2.06*	*-2.04*	*-1.94*	*-1.84*	*-1.96*	**-2.09**	**-1.95**	**-1.94**
10	OFFSCH	*-3.64*	*-2.52*	*-1.75*	*-1.76*	*-1.88*	*-1.86*	*-1.92*	*-1.85*	*-1.89*	**-1.73**	**-1.83**
11	WMUM	*0.57*	*-1.56*	*-1.32*	*-1.34*	*-1.27*	*-1.21*	*-1.32*	*-1.54*	*-1.57*	*-1.55*	**-1.58**
12	CARPHO	*3.22*	*0.50*	*-0.41*	*-0.76*	*-0.80*	*-1.53*	*-1.36*	*-1.38*	*-1.58*	*-1.47*	**-1.74**
13	HANDED	*-3.02*	*-1.44*	*-1.73*	*-1.80*	*-1.58*	*-1.49*	*-1.58*	*-1.36*	*-1.17*	*-1.29*	**-1.85**
14	MQUAL	*13.00*	*4.98*	*3.96*	*3.15*	*2.98*	*1.48*	*1.05*	*1.18*	*1.09*	*1.18*	**1.25**
16	PPART	*7.68*	*3.13*	*1.75*	*1.50*	*1.38*	*0.96*	*0.87*	*1.20*	*1.23*	*1.11*	**0.97**
17	FSOC	*-8.51*	*-3.51*	*-2.90*	*-2.48*	*-2.49*	*-0.76*	*-0.70*	*-0.69*	*-0.68*	*-0.76*	**-0.86**
18	FAMSZ	*1.77*	*1.11*	*0.19*	*-0.01*	*-0.14*	*-0.54*	*-0.54*	*-0.47*	*-0.38*	*-0.40*	**0.40**
19	CLSSYR	*-2.35*	*-2.56*	*-2.93*	*-2.59*	*0.92*	*0.74*	*0.86*	*0.90*	*0.74*	*0.84*	**1.09**
20	UNEMPL	*0.34*	*-0.76*	*-1.39*	*-1.15*	*-1.14*	*-1.34*	*-1.25*	*-1.27*	*-1.44*	*-1.33*	**-0.89**
23	MEDHST	*-1.76*	*-1.76*	*-1.07*	*-1.05*	*-0.84*	*-0.67*	*-0.69*	*-0.49*	*-0.37*	*-0.57*	**-0.57**
24	PCHCOM	*5.29*	*1.44*	*0.16*	*0.15*	*0.03*	*-0.15*	*-0.21*	*-0.08*	*-0.09*	*0.15*	**-0.50**
25	BRTHWT	*-3.08*	*-2.16*	*-2.17*	*-2.16*	*-2.21*	*-2.19*	*-0.63*	*-0.62*	*-0.62*	*-0.58*	**-0.58**
28	BRTHSC	*-1.40*	*-0.62*	*-1.16*	*-1.01*	*-0.81*	*-0.60*	*0.54*	*0.40*	*0.41*	*0.23*	**0.33**
29	TOTCIG	*-4.78*	*-2.27*	*-0.71*	*-0.36*	*-0.24*	*0.28*	*0.44*	*0.31*	*0.35*	*0.31*	**0.29**
30	BRTORD	*-1.20*	*-1.29*	*-1.82*	*-1.46*	*-1.35*	*-1.26*	*-1.24*	*-1.38*	*-1.20*	*-1.13*	**-0.19**
32	MSOC	*-9.26*	*-3.46*	*-2.57*	*-1.33*	*-1.22*	*-0.35*	*-0.15*	*-0.19*	*-0.06*	*-0.14*	**0.91**
33	OCCUPR	*-7.41*	*-2.65*	*-1.65*	*-1.42*	*-1.31*	*-0.37*	*-0.29*	*-0.40*	*-0.22*	*-0.25*	**0.06**

Table 6.6: Changes in the t-values for the covariates during forward stepwise regression, blood lead and school entered first. Rows correspond to variables and columns to steps

ENTERS STEP	COEFFICIENT FOR	Variable being entered at this step PVOC 1	CHINT 2	PMAT 3	AGE 4	FQUAL 5	GESTA 6	PARSC 7	STHGH 8	SEX 9	OFFSC 10	FULL
1	PVOC	9.09	5.75	5.13	5.22	3.66	3.81	4.09	4.09	3.98	3.93	3.88
2	PMAT	8.70	5.18	4.73	4.45	3.94	3.83	3.92	3.95	3.94	4.05	2.73
3	CHINT	6.63	5.09	4.63	5.19	4.74	5.03	5.30	5.43	5.27	5.57	4.80
4	AGE	-3.86	-4.01	-3.76	-4.43	-4.62	-4.71	-4.74	-4.34	-4.51	-4.61	-4.43
5	FQUAL	8.10	4.55	3.87	3.33	3.58	3.48	3.74	3.74	3.73	3.69	2.18
6	GESTA	-2.40	-2.76	-2.63	-3.08	-3.22	-3.11	-3.35	-3.45	-3.48	-3.43	-2.53
7	PRSCHL	1.86	-0.38	-0.75	-1.40	-1.41	-1.88	-2.25	-2.39	-2.29	-2.36	-2.40
8	CLYR	-2.47	-2.59	-2.34	-2.75	1.91	1.90	2.02	2.18	2.31	2.41	2.39
9	OFFSCL	-2.80	-2.26	-2.25	-1.91	-2.19	-2.19	-2.24	-2.14	-2.27	-2.14	-2.19
10	SEX	-0.85	-0.99	-1.32	-2.13	-2.21	-2.14	-2.06	-2.13	-2.25	-2.12	-2.11
11	STHGHT	3.52	2.99	2.78	2.31	2.15	2.18	2.09	2.15	2.00	2.03	2.11
12	CPHON	0.32	-0.81	-1.00	-1.36	-1.29	-1.94	-1.70	-1.71	-1.71	-1.75	-1.86
13	WMUM	0.50	-1.19	-1.25	-1.29	-1.33	-1.39	-1.47	-1.66	-1.63	-1.63	-1.53
14	MQUAL	8.94	4.24	3.46	2.92	2.98	1.57	1.14	1.29	1.20	1.18	1.50
15	HANDD	-2.20	-1.29	-1.38	-1.73	-1.33	-1.27	-1.39	-1.20	-1.28	-1.24	-1.59
17	UNEMP	-1.27	-1.60	-1.34	-1.66	-1.52	-1.67	-1.53	-1.54	-1.67	-1.73	-1.11
19	MDHIST	-1.16	-1.71	-1.62	-1.19	-1.01	-0.87	-0.89	-0.73	-0.83	-0.85	-0.93
20	PPART	4.64	2.32	2.01	1.15	1.08	0.84	0.75	1.03	1.00	0.97	0.56
21	FSOC	-4.63	-2.31	-2.02	-1.90	-2.15	-0.49	-0.45	-0.47	-0.38	-0.43	-0.70
24	MSOC	-5.06	-2.00	-0.75	-0.41	-0.45	0.34	0.49	0.39	0.63	0.42	0.55
27	BTHWT	-2.51	-2.04	-2.00	-2.04	-2.01	-2.00	-0.45	-0.41	-0.31	-0.26	-0.41
28	BTORD	-2.31	-1.98	-1.60	-2.01	-1.76	-1.55	-1.48	-1.62	-1.37	-1.42	-0.31
31	TOTCG	-2.52	-1.43	-1.13	-0.15	-0.26	0.13	0.24	0.14	0.08	0.04	0.08
32	PCHCOM	2.76	0.95	0.98	0.06	0.06	0.04	0.00	0.16	0.18	0.25	-0.04
33	OCCRT	-3.39	-1.22	-1.13	-0.82	-0.91	-0.17	-0.13	-0.23	-0.16	-0.15	0.04

Certain groups of variables can be taken together. Consider first the 7 variables (PVOC, PMAT, CHILDINT, FQUALIF, MQUALIF, FSOC & MSOC) which are at the core of the main cluster in figures 5.4 and 5.5. The first four of these enter during the first five steps for both adjusted and unadjusted data, and the t-values for the others are reduced as they enter. The variables FSOC & MSOC have little association with outcome, once these first four are entered. MQUALIF still retains some association with outcome and when it enters the model at a later step, the coefficients of the first four are modified to become close to the values they achieve for the full model. The regression coefficients for this group of variables are not influenced by any variables outwith the group. At early steps the t-values for the unadjusted data are much greater, but the values are comparable for the adjusted and unadjusted data, once they have settled down to their final values.

Certain other variables have large t-values at early steps but their association with outcome is much diminished after PVOC, PMAT, CHILDINT & FQUALIF. These are OCCUPRAT, TOTCIGS, FAMSIZE, PARPART, PARCHCOM & MEDHIST. The t-values for STHEIGH and OFFSCHL are reduced in a similar manner, but they retain some association with outcome after control for the first five variables, which is not diminished by the entry of other variables into the model.

The variable PARSCHL, which measures parents' involvement with the school starts by being positively associated with outcome, but its influence is reversed, and remains in this direction, after control for the first four variables. One possible explanation of this is as follows. When this variable was being constructed we were aware of the possibility of over-control, when a parent's visits to the school were prompted by their child's learning problems. Thus we excluded from the score those visits made by the parent that were initiated by teachers who were concerned with the child's progress. However, the reversal of the PARSCHL coefficient when, after control for parental characteristics, more parental involvement with the school is associated with poorer outcome, suggests that we may not have been entirely successful in removing this aspect of PARCHCOM. The influence of the variables WORKMUM and CARPHONE on outcome are also reversed at early steps. Again, various social explanations for these relationships could be suggested. I will not pursue these further here, as they are peripheral to my main topic. However, they do illustrate the way in which data which have unexpected patterns can be interpreted in a plausible manner. Even if the t-values quoted here had a valid probabilistic interpretation (which they have not) we cannot discount the possibility that some of the patterns which we are observing might be noise.

The three related variables GESTAT, BRTHWT and BRTHSCO influence each other's coefficients. GESTAT has the greatest effect on outcome; there being no evidence of the other two

variables having an influence on outcome which is independent of it.

The two variables AGE and CLASSYR must be considered together, and it must remembered that the outcome measure (BASC) was itself standardised for age on a random sample of British children. The scores used in this analysis have been age-adjusted in accordance with the procedures laid down in the test manual. Because the Edinburgh Lead Study population was recruited not by age, but by year of schooling, those children who are younger will have had more schooling relative to their chronological age. This explains the negative coefficient for age, with younger children performing better, and also the reversal of the CLASSYR coefficient once AGE is included in the model.

Before adjustment for the other covariates the coefficient of SEX is such that girls obtain slightly lower scores. This effect is enhanced after adjustment for the other covariates, especially the inclusion of the variable CHILDINT on which girls obtained higher scores than boys. Thus this apparent sex difference in the full model may be an artefact of the different scoring which we may have used for the activities in which boys participate, compared with those which are more common for girls.

This description of the changing coefficients for the covariates is not quite complete, but it illustrates that there is a considerable degree of structure in the multiple correlation of these other x variables with the outcome, BASC. I hope to be able

to model this in later chapters when simulating a model which requires a joint distribution of y and the x variables.

6.7 Selection by backwards elimination

A backward elimination method of variable selection was also tried for the unadjusted and adjusted data. Starting with the full model, covariates were removed from the model at each step by choosing the covariate which produced the smallest increase in the residual sum of squares. The order in which the variables were selected was very nearly the same as the reverse order for the forward stepwise procedure. The two orders for the unadjusted data differed only by moving three covariates by at most three positions. For the adjusted data only one variable needed to be moved three places to make the orders identical. In both cases the differences in order were around the middle of the stepwise procedure.

The details of the backwards elimination method are not shown, because the models chosen are largely the same as those for forward selection. The same models are selected as those giving the minima of all the criteria discussed above.

If there is exact agreement between the ordering of the subsets chosen by the forward selection and backwards elimination, then they also agree with the all-subsets minima for every subset size; Berk (1978) gives a simple proof of this. While this condition, also referred to as "nesting", is not quite fulfilled

here, it is very close to being true. Thus we would expect the subsets selected by the forward and backwards procedures to be those with the minimum sum of squares for their subset size in the majority of cases.

Chapter 7

Other stepwise procedures

7.1 Introduction

Various other stepwise procedures are described here. The MSE criteria introduced in chapter 4 can have three different functions in connection with model selection, each of which could be used with or without the others. Firstly, the value of the criterion can be used to drive the selection procedure, by selecting variables to enter or exclude which will give low values of the criterion. Secondly, the criteria can be used to provide stopping rules, which determine what size of model should be used. Thirdly they can be used, at the end of selection, to estimate the MSE of the coefficient b^*.

The results described here use the criteria in the first sense, starting with forward and backward procedures which select the model at each step which gives the lowest value of the various G_p criteria. The results are different from those for the MSE of prediction, in that the fixed and random effects models select different sets of covariates. Also, unlike the case for prediction MSEs, different sets of covariates are selected by forward and backward selection procedures. The stepwise procedures are carried forward until all variables are entered or removed, and the values of all of the criteria are given for each stepwise procedure, not just that of the criterion being used for selection. This allows

the influences on selection to be understood, and potential biases identified. Later in this chapter modifications are introduced which estimate the G_p criteria in such a way that negative estimates of positive quantities are set to zero.

Examination of these results suggest that either the modified criteria, or else procedures which select models which include terms which will influence the estimate of β^*, may be more sensible than using the G_p for model selection. Finally procedures which select covariates which are associated with X^* are considered.

7.2 Selection to minimise G_{FP}

The results of a forward stepwise procedure to minimise G_{FP} for the adjusted and unadjusted data are in tables 7.1 and 7.2. The minimum of the criterion for the unadjusted data is achieved at the first step, and for the school-adjusted data G_{FP} is minimised at the third step. For the unadjusted data this is the same model as is chosen by the first step of the usual forward stepwise procedure, but for the school-adjusted data the minimum achieved by G_{FP} is lower than that for the model with no covariates which was the lowest found in section 6.2.

Several features are obvious from these tables. The models selected are those with estimates for b_p^* which are very close to those for the full model, so that the Est($bias^2$) term becomes as negative as possible. The stepwise procedure manages to find such a model immediately for the unadjusted data and keeps on finding such

Table 7.1 : Results of forward stepwise procedure based on choosing the minimum value of G_{FP}, unadjusted data. (contd on next page).

Variable entered	p	Est(bias²)	V_{FP}	G_{FP}	V_{1RP}	V_{2RP}	G_{RP}
PVOC	3	-0,22884	1,45593	1,22709	1,76797	0,02554	1,53912
WMUM	4	-0,22883	1,45594	1,22710	1,76643	0,02192	1,53760
HNDED	5	-0,22881	1,45595	1,22714	1,76524	0,01832	1,53643
PMENT	6	-0,22871	1,45606	1,22735	1,77190	0,01491	1,54319
OCCRT	7	-0,22830	1,45611	1,22781	1,75639	0,01126	1,52808
PCHCM	8	-0,22849	1,45625	1,22776	1,75992	0,00785	1,53143
PPART	9	-0,22824	1,45630	1,22806	1,74031	0,00428	1,51208
CPHNE	10	-0,22761	1,45706	1,22945	1,74730	0,00164	1,51969
PRHLT	11	-0,22595	1,45765	1,23170	1,75141	-0,00123	1,52546
CSMR	12	-0,22542	1,45935	1,23393	1,75618	-0,00278	1,53076
BTHWT	13	-0,22399	1,45961	1,23562	1,74729	-0,00603	1,52330
FHIST	14	-0,22299	1,46179	1,23880	1,75274	-0,00703	1,52975
MSOC	15	-0,22015	1,46202	1,24187	1,72402	-0,01018	1,50387
FSIZE	16	-0,22119	1,46347	1,24228	1,72778	-0,01205	1,50659
UNMPL	17	-0,21758	1,46678	1,24920	1,73403	-0,01175	1,51645
BTSCO	18	-0,21207	1,46989	1,25783	1,74017	-0,01171	1,52810
TCIGS	19	-0,21033	1,47404	1,26371	1,74344	-0,01045	1,53311
TMDAY	20	-0,20523	1,47948	1,27426	1,75026	-0,00770	1,54503
CHINT	21	-0,19621	1,47990	1,28369	1,68214	-0,01042	1,48592
CLSYR	22	-0,19690	1,48300	1,28610	1,66256	-0,01029	1,46565
GESTA	23	-0,19417	1,48582	1,29165	1,64802	-0,01053	1,45385
MDHIS	24	-0,18855	1,49531	1,30677	1,65167	-0,00352	1,46312
MVSCL	25	-0,18039	1,50427	1,32388	1,65849	0,00288	1,47810
BORD	26	-0,15356	1,51316	1,35960	1,66001	0,00919	1,50645
FQUAL	27	-0,15878	1,51762	1,35884	1,63065	0,01039	1,47187
OFFSC	28	-0,15569	1,52409	1,36840	1,62458	0,01385	1,46889
MQUAL	29	-0,13981	1,53891	1,39910	1,62671	0,02631	1,48690
SEX	30	-0,13241	1,54825	1,41584	1,62667	0,03281	1,49426
FSOC	31	-0,11048	1,56845	1,45797	1,63044	0,05101	1,51996
PSCHL	32	0,00804	1,57933	1,58737	1,61713	0,05857	1,62517
PMAT	33	-0,07755	1,59256	1,51501	1,57235	0,06709	1,49479
STHGT	34	0,24231	1,65514	1,89745	1,55747	0,12665	1,79978
AGE	35	-0,00000	1,68477	1,68477	1,53396	0,15082	1,53396

Table 7.1 (contd) : Forward selection on G_{Fp}, unadjusted data.

Variable entered	p	F ratio	RMS_p	C_p	S_p	b_p*	t-ratio
PVOC	3	177.03	129.23	118.47	0.26001	-3.18	-2.38
WMUM	4	2.44	128.85	117.47	0.25979	-3.18	-2.38
HNDED	5	2.34	128.51	116.61	0.25961	-3.18	-2.38
PMENT	6	0.14	128.73	118.44	0.26059	-3.18	-2.38
OCCRT	7	6.38	127.35	112.69	0.25831	-3.20	-2.41
PCHCM	8	1.01	127.34	113.46	0.25883	-3.19	-2.40
PPART	9	7.57	125.67	106.40	0.25594	-3.16	-2.40
CPHNE	10	0.03	125.92	108.36	0.25697	-3.17	-2.40
PRHLT	11	0.85	125.96	109.34	0.25758	-3.14	-2.38
CSMR	12	0.67	126.04	110.53	0.25828	-3.18	-2.40
BTHWT	13	4.49	125.14	107.17	0.25697	-3.14	-2.38
FHIST	14	0.49	125.28	108.59	0.25777	-3.18	-2.41
MSOC	15	10.13	122.97	98.71	0.25355	-3.13	-2.39
FSIZE	16	0.94	122.98	99.60	0.25410	-3.17	-2.42
UNMPL	17	0.25	123.17	101.31	0.25502	-3.20	-2.44
BTSCO	18	0.29	123.35	102.96	0.25592	-3.23	-2.46
TCIGS	19	1.10	123.33	103.67	0.25640	-3.16	-2.40
TMDAY	20	0.12	123.56	105.53	0.25741	-3.19	-2.41
CHINT	21	21.52	118.50	83.21	0.24739	-3.09	-2.39
CLSYR	22	7.67	116.88	76.67	0.24451	-3.25	-2.53
GESTA	23	6.24	115.61	71.79	0.24237	-3.11	-2.43
MDHIS	24	0.95	115.62	72.75	0.24291	-3.21	-2.50
MVSCL	25	0.04	115.86	74.71	0.24391	-3.19	-2.47
BORD	26	1.57	115.72	74.98	0.24413	-3.31	-2.56
FQUAL	27	10.57	113.43	65.55	0.23982	-3.09	-2.41
OFFSC	28	3.78	112.77	63.48	0.23892	-3.25	-2.54
MQUAL	29	1.39	112.68	64.00	0.23924	-3.10	-2.41
SEX	30	2.02	112.44	63.84	0.23923	-3.24	-2.52
FSOC	31	0.91	112.46	64.86	0.23979	-3.10	-2.39
PSCHL	32	5.88	111.30	60.62	0.23783	-2.84	-2.20
PMAT	33	15.39	107.99	46.77	0.23124	-3.30	-2.58
STHGT	34	6.48	106.74	42.18	0.22905	-2.66	-2.05
AGE	35	9.18	104.90	35.00	0.22560	-3.18	-2.45

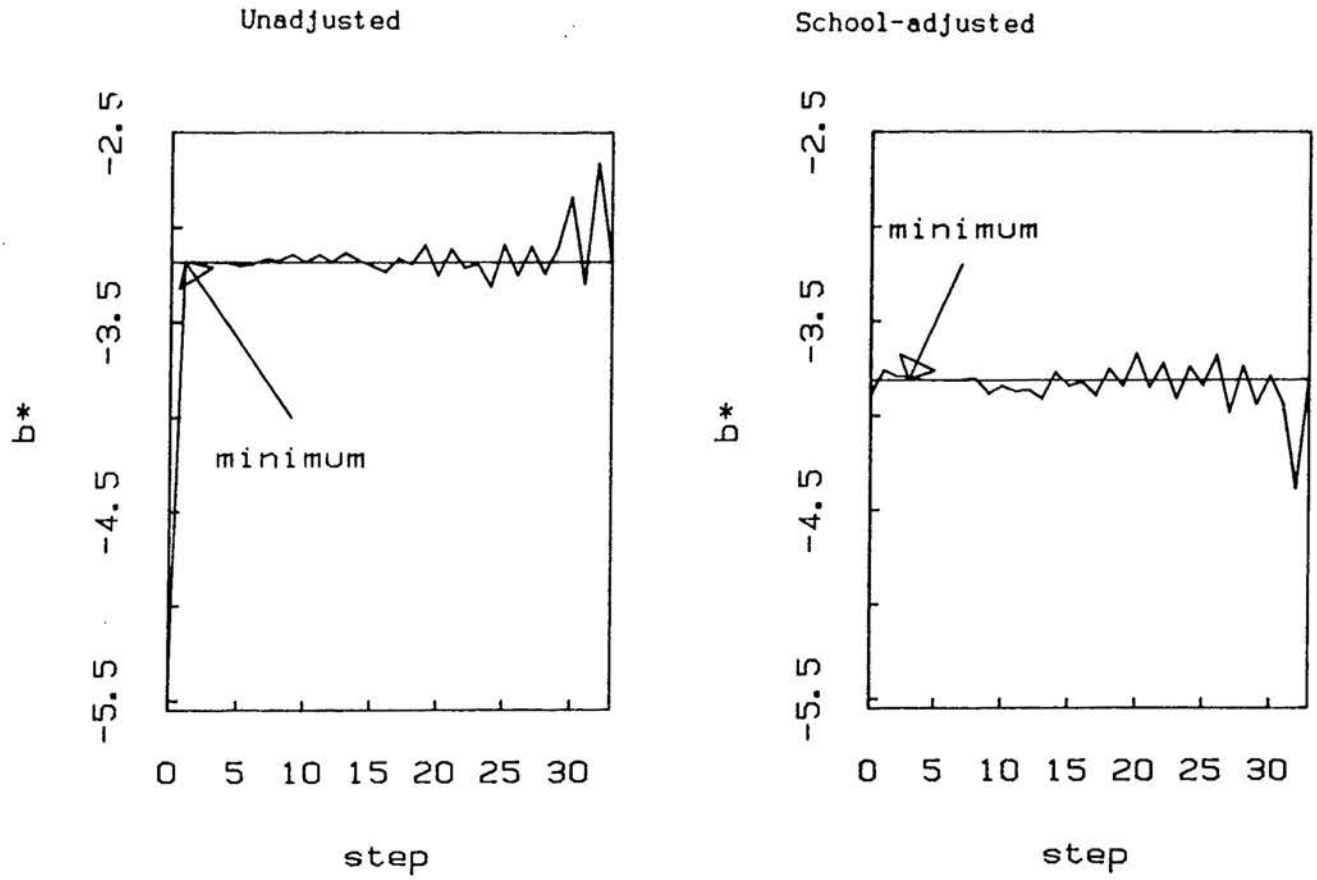
Table 7.2 : Results of forward stepwise procedure based on choosing the minimum value of G_{FP}, school-adjusted data. (contd next page).

Variable entered	p	Est(bias^2)	V_{FP}	G_{FP}	V_{1RP}	V_{2RP}	G_{RP}
PPART	20	-0,25861	1,60659	1,34798	2,22049	-0,00387	1,96188
BTHWT	21	-0,26042	1,60667	1,34625	2,20385	-0,00832	1,94343
HNDED	22	-0,26054	1,60667	1,34613	2,18867	-0,01281	1,92813
CSMR	23	-0,26063	1,60710	1,34646	2,19566	-0,01685	1,93502
PCHCM	24	-0,26040	1,60736	1,34696	2,20477	-0,02114	1,94438
PRHLT	25	-0,25988	1,60786	1,34798	2,21403	-0,02516	1,95415
FHIST	26	-0,25943	1,60830	1,34888	2,22337	-0,02929	1,96394
WMUM	27	-0,25831	1,60932	1,35101	2,23268	-0,03267	1,97437
OCCRT	28	-0,25314	1,60999	1,35685	2,20871	-0,03602	1,95557
PMENT	29	-0,25464	1,61261	1,35797	2,21540	-0,03721	1,96076
PSCHL	30	-0,24953	1,61547	1,36593	2,22376	-0,03812	1,97423
TMDAY	31	-0,24403	1,62174	1,37770	2,23321	-0,03445	1,98917
BTSCO	32	-0,23238	1,62624	1,39386	2,24073	-0,03316	2,00835
GESTA	33	-0,23401	1,63208	1,39806	2,23899	-0,02994	2,00498
BORD	34	-0,23244	1,63473	1,40229	2,24280	-0,03115	2,01035
MVSCL	35	-0,22040	1,64724	1,42683	2,25240	-0,01909	2,03199
CPHNE	36	-0,20318	1,65883	1,45566	2,26083	-0,00823	2,05766
MSOC	37	-0,20350	1,65972	1,45622	2,18708	-0,01150	1,98358
UNMPL	38	-0,19949	1,66766	1,46817	2,19263	-0,00583	1,99315
TCIGS	39	-0,16900	1,67800	1,50900	2,19207	0,00296	2,02308
CLSYR	40	-0,18474	1,68171	1,49698	2,16962	0,00302	1,98488
CHINT	41	-0,17677	1,68220	1,50544	2,05170	-0,00101	1,87494
SEX	42	-0,16648	1,69101	1,52453	2,04475	0,00522	1,87827
FQUAL	43	-0,17073	1,69198	1,52126	1,94040	0,00182	1,76967
FSIZE	44	-0,15566	1,71153	1,55587	1,94718	0,02003	1,79152
FSOC	45	-0,11955	1,73128	1,61172	1,95124	0,03844	1,83169
PMAT	46	-0,10576	1,73432	1,62856	1,85169	0,03564	1,74593
MQUAL	47	-0,12169	1,74079	1,61910	1,82662	0,03799	1,70493
OFFSC	48	-0,09903	1,75178	1,65275	1,82055	0,04547	1,72151
PVOC	49	-0,11224	1,75499	1,64275	1,80075	0,04427	1,68851
MDHIS	50	-0,06908	1,78201	1,71292	1,80547	0,06869	1,73638
AGE	51	0,26239	1,79182	2,05421	1,73372	0,07185	1,99612
STHGT	52	-0,00000	1,86778	1,86778	1,72436	0,14342	1,72436

Table 7.2 (cntd) : Forward selection on G_{F_p}, school-adjusted data.

Variable entered	p	F ratio	RMS_p	C_p	S_p	b_p*	t-ratio
PPART	20	21.57	136.32	202.63	0.28401	-3.76	-2.53
BTHWT	21	5.64	135.02	196.92	0.28188	-3.79	-2.56
HNDED	22	5.34	133.81	191.69	0.27994	-3.79	-2.57
CSMR	23	0.48	133.96	193.04	0.28083	-3.81	-2.58
PCHCM	24	0.02	134.23	195.01	0.28200	-3.81	-2.58
PRHLT	25	0.01	134.51	197.01	0.28318	-3.81	-2.57
FHIST	26	0.00	134.79	199.00	0.28438	-3.81	-2.57
WMUM	27	0.02	135.07	200.98	0.28557	-3.80	-2.56
OCCRT	28	7.16	133.34	193.32	0.28250	-3.88	-2.63
PMENT	29	0.57	133.46	194.54	0.28336	-3.84	-2.60
PSCHL	30	0.23	133.68	196.24	0.28442	-3.87	-2.61
TMDAY	31	0.01	133.96	198.23	0.28563	-3.86	-2.60
BTSCO	32	0.42	134.13	199.65	0.28660	-3.91	-2.63
GESTA	33	2.37	133.74	198.45	0.28637	-3.77	-2.54
BORD	34	1.21	133.68	198.81	0.28686	-3.84	-2.58
MVSCL	35	0.01	133.96	200.80	0.28809	-3.82	-2.56
CPHNE	36	0.26	134.17	202.44	0.28917	-3.89	-2.59
MSOC	37	17.72	129.52	181.22	0.27973	-3.75	-2.54
UNMPL	38	0.83	129.56	182.13	0.28044	-3.84	-2.60
TCIGS	39	2.12	129.25	181.36	0.28037	-3.67	-2.48
CLSYR	40	6.80	127.65	174.57	0.27750	-3.85	-2.61
CHINT	41	28.56	120.45	141.76	0.26242	-3.72	-2.60
SEX	42	3.57	119.78	139.43	0.26153	-3.91	-2.73
FQUAL	43	26.74	113.42	110.74	0.24818	-3.74	-2.68
FSIZE	44	0.41	113.57	112.27	0.24905	-3.84	-2.74
FSOC	45	1.05	113.55	113.06	0.24957	-3.68	-2.61
PMAT	46	26.57	107.52	86.14	0.23684	-3.98	-2.90
MQUAL	47	8.26	105.83	79.29	0.23363	-3.74	-2.74
OFFSC	48	3.52	105.25	77.54	0.23285	-3.94	-2.89
PVOC	49	7.00	103.87	72.18	0.23032	-3.79	-2.79
MDHIS	50	0.82	103.92	73.32	0.23092	-3.94	-2.88
AGE	51	20.71	99.56	54.45	0.22175	-4.39	-3.27
STHGT	52	4.45	98.81	52.00	0.22055	-3.81	-2.79

Figure 7.1 : Values of b*, forward selection by G_{FP}.



models' until it is obliged to enter the variables which swing the estimate of b_P* away from its full model value, during the last few steps. There is a similar pattern for the school-adjusted data. These changes in the estimate of b_P* are illustrated in Figure 7.1.

The models selected are very different from those which minimise the residual sum-of-squares, and the values of C_P and S_P are nowhere near their minima. This is because the fixed-effects model criterion uses the variance term from the full model in its estimates. Another consequence of this is the erratic values for G_{RP} obtained for this sequence of models.

We know that the variance part of the G_{FP} criterion will be strictly increasing as further terms are added to the model. However, we can see that some selection is operating to limit this increase. There is evidence of selection on the second variance term in the random-effects model, a multiple of which is an implicit contributor to the variance of the fixed-effects model. Variables are selected which give small values (many negative) of this term. This term increases when variables correlated with X* are added into the model, and we can see that the variable which has the strongest relationship with X* (STHEIGHT) enters at the very end.

These results, with the large number of negative estimates for quantities which should be positive, must cast some doubt on these methods. Even if they can be shown to select reasonable models the value of the G_{FP} criterion at the end of the procedure will be an underestimate of the MSE, because of the selection effects.

This problem might be ameliorated by replacing a negative estimate of the squared bias by zero in the expression G_{FP}, thus obtaining a new criterion G'_{FP}. The estimates of the MSE criteria for a forward stepwise procedure based on this criterion are given in tables 7.3 and 7.4. The criterion G'_{FP} is not tabulated explicitly since it is equal to G_{FP} when the Est(bias^2) term is positive, and equal to V_{FP} when Est(bias^2) is negative.

The minimum of the new criteria occurs at the first step for both the adjusted and the unadjusted data. This is bound to happen if the models chosen at these two steps are such as to give zero for

Est(bias^2), since the variance part of G_{Fp} is strictly increasing with additional covariates. Examination of later steps shows that, for both the unadjusted data and the adjusted data, models with zero bias terms were selected up to the last three steps. Also the term which was being minimised as further covariates were added to the model was V_{2Rp}, and this was even more marked than for straightforward minimisation of G_{Fp}. These procedures were even less likely than the previous ones to select a models with low RMS_p. For neither the unadjusted nor the adjusted data did the value of C_p fall below p, and it was generally high above it (detailed results not shown).

For selection based on G'_{Fp} the value of b_p^* is not so tightly constrained to be close to the value for the full model. This is illustrated in Figure 7.2. However, it is constrained to be within a certain range of the full model estimate by the requirement that the squared bias be less than its expected value for zero bias. One interpretation of this criterion is that it is selecting models with low variance estimates, subject to a condition on the bias.

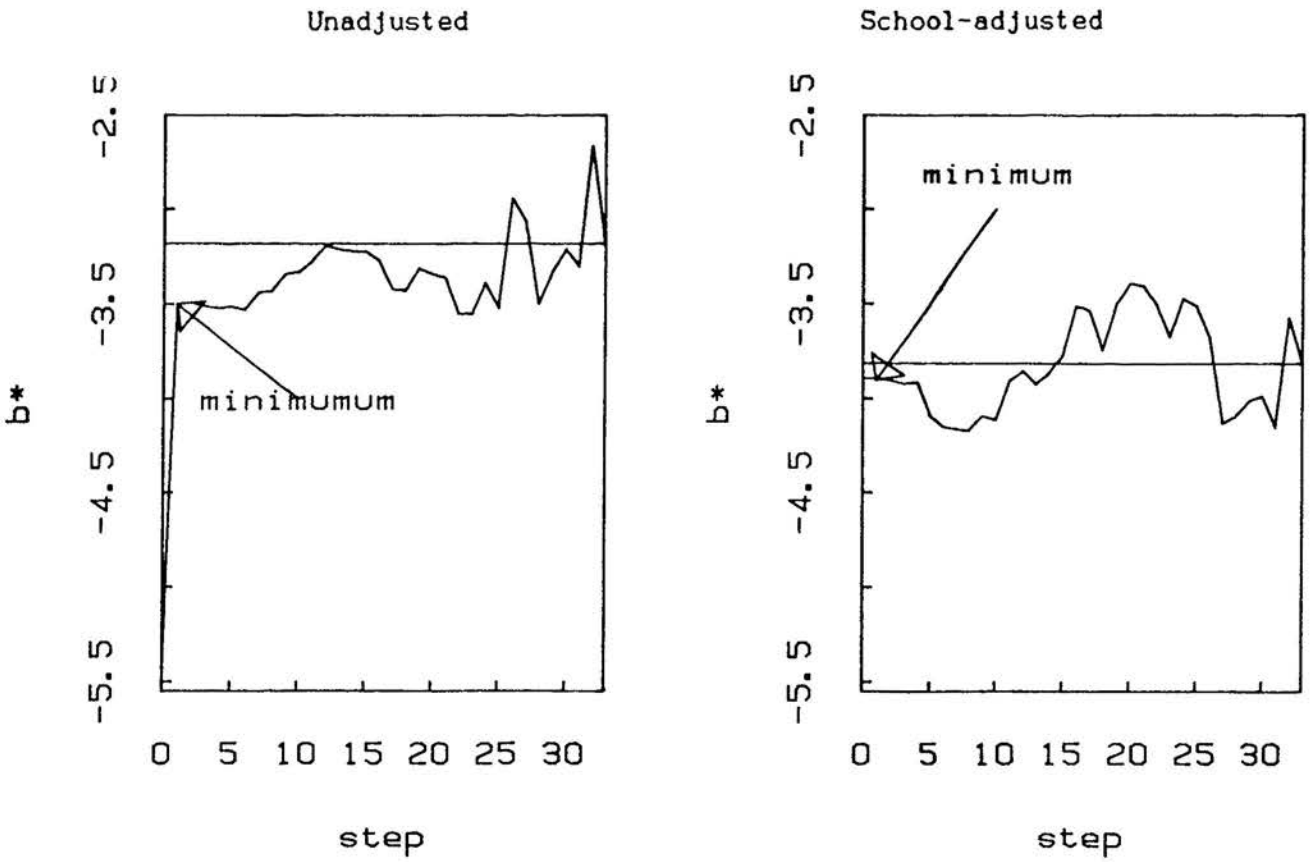
Table 7.3 : Results of forward stepwise procedure based on choosing the minimum value of G'_{FP}, unadjusted data.

Variable entered	p	Est(bias²)	V_{FP}	G_{FP}	V_{1RP}	V_{2RP}	G_{RP}
FQUAL	3	-0,13019	1,45235	1,32216	1,84870	0,02211	1,71850
HNDED	4	-0,13647	1,45237	1,31590	1,84029	0,01828	1,70382
OCCRT	5	-0,12120	1,45246	1,33126	1,82886	0,01455	1,70766
WMUM	6	-0,11681	1,45259	1,33578	1,83537	0,01104	1,71857
PMENT	7	-0,12037	1,45276	1,33239	1,84230	0,00754	1,72193
CPHNE	8	-0,10665	1,45304	1,34640	1,84581	0,00416	1,73916
PPART	9	-0,16417	1,45359	1,28941	1,80460	0,00107	1,64043
PCHCM	10	-0,16569	1,45383	1,28814	1,81182	-0,00232	1,64614
CHINT	11	-0,20464	1,45411	1,24947	1,72515	-0,00540	1,52051
PRHLT	12	-0,20757	1,45458	1,24701	1,73146	-0,00839	1,52389
BTHWT	13	-0,22127	1,45511	1,23385	1,71907	-0,01121	1,49780
MSOC	14	-0,22901	1,45567	1,22666	1,68986	-0,01382	1,46085
CSMR	15	-0,22621	1,45787	1,23166	1,69645	-0,01480	1,47024
FHIST	16	-0,22305	1,46016	1,23712	1,70323	-0,01569	1,48018
FSIZE	17	-0,22092	1,46249	1,24157	1,71027	-0,01656	1,48935
UNMPL	18	-0,21054	1,46531	1,25477	1,71378	-0,01684	1,50324
CLSYR	19	-0,15726	1,46782	1,31056	1,69339	-0,01725	1,53613
TCIGS	20	-0,15172	1,47113	1,31941	1,70043	-0,01704	1,54871
GESTA	21	-0,19246	1,47449	1,28203	1,69526	-0,01665	1,50280
BTSCO	22	-0,18183	1,47752	1,29569	1,70166	-0,01679	1,51983
TMDAY	23	-0,16927	1,48270	1,31343	1,70851	-0,01448	1,53924
OFFSC	24	-0,06109	1,48879	1,42770	1,69782	-0,01102	1,63673
MVSCL	25	-0,05318	1,49613	1,44294	1,70497	-0,00629	1,65179
PSCHL	26	-0,13571	1,50403	1,36832	1,70243	-0,00091	1,56671
SEX	27	-0,05678	1,51082	1,45404	1,70169	0,00317	1,64492
PVOC	28	-0,10765	1,52117	1,41353	1,60551	0,01059	1,49787
BORD	29	-0,14006	1,53003	1,38997	1,60751	0,01657	1,46745
PMAT	30	-0,04102	1,54226	1,50124	1,56338	0,02536	1,52236
MQUAL	31	-0,10646	1,55516	1,44870	1,56310	0,03524	1,45664
FSOC	32	-0,11222	1,57158	1,45936	1,56689	0,04879	1,45467
MDHIS	33	-0,07755	1,59256	1,51501	1,57235	0,06709	1,49479
STHGT	34	0,24231	1,65514	1,89745	1,55747	0,12665	1,79978
AGE	35	-0,00000	1,68477	1,68477	1,53396	0,15082	1,53396

Table 7.4 : Results of forward stepwise procedure based on choosing the minimum value of G'_{Fp}, school-adjusted data.

Variable entered	p	Est(bias²)	V_{Fp}	G_{Fp}	V_{1Rp}	V_{2Rp}	G_{Rp}
HNDED	20	-0,25576	1,60606	1,35030	2,29688	-0,00477	2,04112
PCHCM	21	-0,25493	1,60606	1,35113	2,27497	-0,00945	2,02004
BTHWT	22	-0,25027	1,60612	1,35586	2,25289	-0,01396	2,00263
FHIST	23	-0,25210	1,60652	1,35442	2,26130	-0,01815	2,00921
PMAT	24	-0,18128	1,60693	1,42564	1,98009	-0,01952	1,79881
OCCRT	25	-0,14927	1,60739	1,45812	1,96897	-0,02295	1,81970
PRHLT	26	-0,13865	1,60791	1,46927	1,97585	-0,02650	1,83721
CSMR	27	-0,13233	1,60853	1,47620	1,98384	-0,02999	1,85151
PPART	28	-0,18273	1,60951	1,42678	1,96891	-0,03268	1,78617
WMUM	29	-0,17009	1,61073	1,44063	1,97611	-0,03545	1,80602
CHINT	30	-0,24721	1,61246	1,36525	1,90009	-0,03605	1,65288
PSCHL	31	-0,25128	1,61487	1,36359	1,90462	-0,03732	1,65334
BORD	32	-0,23854	1,61724	1,37870	1,90566	-0,03859	1,66712
PMENT	33	-0,24581	1,62007	1,37426	1,90839	-0,03937	1,66258
MSOC	34	-0,24200	1,62364	1,38164	1,90857	-0,03927	1,66657
FQUAL	35	-0,15143	1,62703	1,47560	1,84970	-0,03817	1,69827
BTSCO	36	-0,15882	1,63075	1,47193	1,85732	-0,03809	1,69850
CLSYR	37	-0,22626	1,63586	1,40961	1,83596	-0,03591	1,60970
PVOC	38	-0,12727	1,64024	1,51297	1,79587	-0,03422	1,66860
MQUAL	39	-0,04573	1,64327	1,59754	1,79016	-0,03468	1,74443
TMDAY	40	-0,04729	1,64961	1,60232	1,79789	-0,03186	1,75060
UNMPL	41	-0,11426	1,65686	1,54260	1,80042	-0,02800	1,68616
CPHNE	42	-0,18378	1,66399	1,48021	1,79349	-0,02416	1,60971
GESTA	43	-0,07564	1,67203	1,59639	1,78202	-0,01937	1,70638
TCIGS	44	-0,09516	1,68025	1,58509	1,78902	-0,01463	1,69386
SEX	45	-0,16187	1,68883	1,52696	1,78386	-0,00945	1,62199
AGE	46	-0,06732	1,69922	1,63190	1,71924	-0,00237	1,65191
MVSCL	47	-0,07550	1,71321	1,63771	1,72646	0,00798	1,65096
FSOC	48	-0,09778	1,72979	1,63201	1,73262	0,02098	1,63484
FSIZE	49	-0,08999	1,75115	1,66116	1,74005	0,03887	1,65007
MDHIS	50	0,02678	1,77910	1,80588	1,74354	0,06338	1,77032
STHGT	51	0,04491	1,85554	1,90045	1,73455	0,13613	1,77946
OFFSC	52	-0,00000	1,86778	1,86778	1,72436	0,14342	1,72436

Figure 7.2 : Values of b*, forward selection by G'_{Fp}.



Backward elimination procedures, based on the same two criteria were also explored. Unlike the case for selection on the residual sum-of-squares, quite a different set of models were obtained by the backwards and forwards procedures. Results for the unadjusted data for the criteria G_{Fp} and G'_{Fp} are given in tables 7.5 and 7.6. The results for the school-adjusted data were similar. Those variables which move b* a long way from its full model value are retained in the model until the last few steps. The minimum for the G_{Fp} procedure is obtained for the same model with just the one covariate (PVOC) as was found for the forward selection. For G'_{Fp} the minimum is obtained a few steps from the end, and is higher than that found for forward selection.

Table 7.5 : Results of backward elimination procedure based on choosing the minimum value of G_{FP}, unadjusted data.

Variable dropped	p	est(bias²)	V_{FP}	G_{FP}	V_{1RP}	V_{2RP}	G_{RP}
MDHIS	35	-0,01806	1,65820	1,64014	1,52844	0,12735	1,51038
FSOC	34	-0,04505	1,63842	1,59338	1,52463	0,11083	1,47959
BORD	33	-0,05689	1,62743	1,57054	1,51819	0,10290	1,46130
MVSCL	32	-0,06620	1,61742	1,55123	1,51178	0,09597	1,44559
TMDAY	31	-0,07721	1,60720	1,53000	1,50626	0,08889	1,42905
BTSCO	30	-0,08058	1,60297	1,52239	1,50006	0,08771	1,41948
MQUAL	29	-0,08162	1,59456	1,51293	1,49992	0,08272	1,41830
CLSYR	28	-0,10215	1,58083	1,47867	1,49782	0,07232	1,39566
TCIGS	27	-0,10402	1,57702	1,47301	1,49181	0,07156	1,38779
WMUM	26	-0,10836	1,57616	1,46780	1,49295	0,07406	1,38459
OCCRT	25	-0,10987	1,57419	1,46432	1,48687	0,07509	1,37700
PPART	24	-0,11143	1,57334	1,46191	1,48373	0,07736	1,37230
CSMR	23	-0,11136	1,57065	1,45929	1,48102	0,07782	1,36966
FSIZE	22	-0,11241	1,57040	1,45799	1,47572	0,08055	1,36331
PCHCM	21	-0,11236	1,57040	1,45803	1,46974	0,08345	1,35738
MSOC	20	-0,11226	1,57039	1,45813	1,46542	0,08643	1,35315
PMENT	19	-0,11202	1,57037	1,45835	1,46044	0,08933	1,34842
FHIST	18	-0,10983	1,56820	1,45838	1,45597	0,09012	1,34615
GESTA	17	-0,11108	1,56331	1,45223	1,47101	0,08940	1,35993
OFFSC	16	-0,12693	1,55587	1,42894	1,47367	0,08534	1,34675
BTHWT	15	-0,12932	1,55545	1,42613	1,48065	0,08855	1,35133
PRHLT	14	-0,12968	1,55483	1,42515	1,47554	0,09084	1,34586
CPHNE	13	-0,13005	1,55472	1,42467	1,47771	0,09408	1,34766
HNDED	12	-0,12959	1,55378	1,42420	1,47647	0,09627	1,34688
UNMPL	11	-0,12621	1,55315	1,42694	1,47658	0,09886	1,35036
PSCHL	10	-0,12346	1,54513	1,42167	1,48541	0,09448	1,36195
SEX	9	-0,14432	1,53771	1,39339	1,49217	0,09051	1,34785
CHINT	8	-0,12129	1,53713	1,41585	1,54745	0,09658	1,42616
FQUAL	7	-0,13503	1,53300	1,39797	1,60125	0,09880	1,46622
AGE	6	-0,06483	1,51617	1,45134	1,62998	0,08504	1,56515
STHGT	5	0,05922	1,46313	1,52235	1,66303	0,02896	1,72225
PMAT	4	-0,22884	1,45593	1,22709	1,76797	0,02554	1,53912
PVOC	3	4,90704	1,43231	6,33935	2,38684	0,00000	7,29388

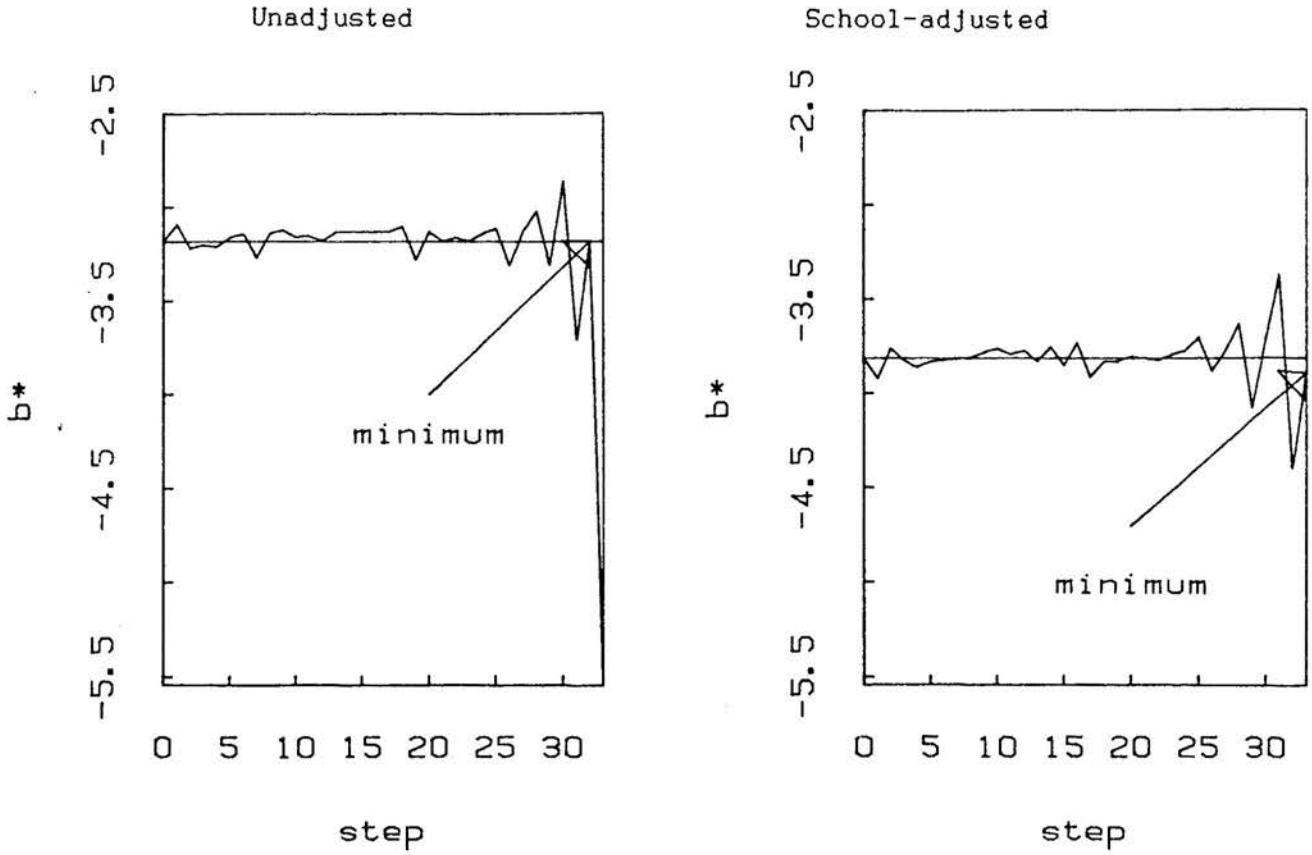
Table 7.6 : Results of backward elimination procedure based on choosing the minimum value of G'_{FP}, unadjusted data.

Variable dropped	p	est(bias²)	V_{FP}	G_{FP}	V_{1RP}	V_{2RP}	G_{RP}
MDHIS	35	-0,01806	1,65820	1,64014	1,52844	0,12735	1,51038
FSOC	34	-0,04505	1,63842	1,59338	1,52463	0,11083	1,47959
MQUAL	33	-0,02369	1,62537	1,60168	1,52395	0,10124	1,50025
AGE	32	0,00747	1,59986	1,60733	1,54828	0,08041	1,55574
STHGT	31	-0,04102	1,54226	1,50124	1,56338	0,02536	1,52236
PMAT	30	-0,14006	1,53003	1,38997	1,60751	0,01657	1,46745
BORD	29	-0,10765	1,52117	1,41353	1,60551	0,01059	1,49787
PVOC	28	-0,05678	1,51082	1,45404	1,70169	0,00317	1,64492
MVSCL	27	-0,07260	1,50348	1,43088	1,69460	-0,00152	1,62200
SEX	26	-0,14513	1,49729	1,35216	1,69529	-0,00494	1,55017
PSCHL	25	-0,06109	1,48879	1,42770	1,69782	-0,01102	1,63673
OFFSC	24	-0,16927	1,48270	1,31343	1,70851	-0,01448	1,53924
TMDAY	23	-0,18183	1,47752	1,29569	1,70166	-0,01679	1,51983
FSIZE	22	-0,19029	1,47299	1,28270	1,69468	-0,01836	1,50439
FHIST	21	-0,19682	1,47025	1,27342	1,68784	-0,01792	1,49102
CLSYR	20	-0,21669	1,46741	1,25072	1,70827	-0,01788	1,49157
FQUAL	19	-0,04329	1,46137	1,41808	1,80406	-0,02253	1,76077
TCIGS	18	-0,01912	1,45829	1,43917	1,79756	-0,02252	1,77843
UNMPL	17	-0,05797	1,45547	1,39750	1,79199	-0,02222	1,73402
CSMR	16	-0,07654	1,45325	1,37670	1,78518	-0,02119	1,70863
BTSCO	15	-0,07838	1,45150	1,37313	1,77784	-0,01959	1,69947
CPHNE	14	-0,10320	1,45090	1,34770	1,77571	-0,01668	1,67251
GESTA	13	-0,00661	1,44765	1,44104	1,78137	-0,01708	1,77476
PRHLT	12	-0,00086	1,44731	1,44645	1,77445	-0,01382	1,77358
PCHCM	11	-0,00268	1,44699	1,44431	1,76724	-0,01057	1,76456
WMUM	10	-0,00973	1,44688	1,43714	1,76146	-0,00710	1,75173
PMENT	9	-0,00575	1,44674	1,44099	1,75470	-0,00368	1,74895
OCCRT	8	0,01530	1,44667	1,46196	1,76742	-0,00020	1,78272
HNDED	7	0,06583	1,44645	1,51228	1,79120	0,00315	1,85703
BTHWT	6	0,15476	1,44579	1,60055	1,80936	0,00603	1,96412
PPART	5	0,37937	1,44441	1,82378	1,85545	0,00816	2,23481
CHINT	4	1,12210	1,44228	2,56438	2,04447	0,01009	3,16657
MSOC	3	4,90704	1,43231	6,33935	2,38684	0,00000	7,29388

The constraint which holds b* close to its full model value can also be seen to operate for the G_{Fp} criterion. It does not have such a marked effect at the initial steps, because the maximum negative value of the Est(bias^2) term is minus the variance of the estimated bias, and this can be shown to decrease systematically towards zero as the number of terms in the model increases. Thus the maximum negative contribution of this term towards G_{Fp} is much smaller at the first step of the backwards elimination procedure, than at the first step of forwards selection. The values of b* are plotted in Figure 7.3.

Negative values of V_{2Rp} are not found for backwards selection by G_{Fp}, and this procedure produces models with much lower RMS_p than forward selection on either G_{Fp} criterion. C_p values less than p were found for the school-adjusted data, although never as low as

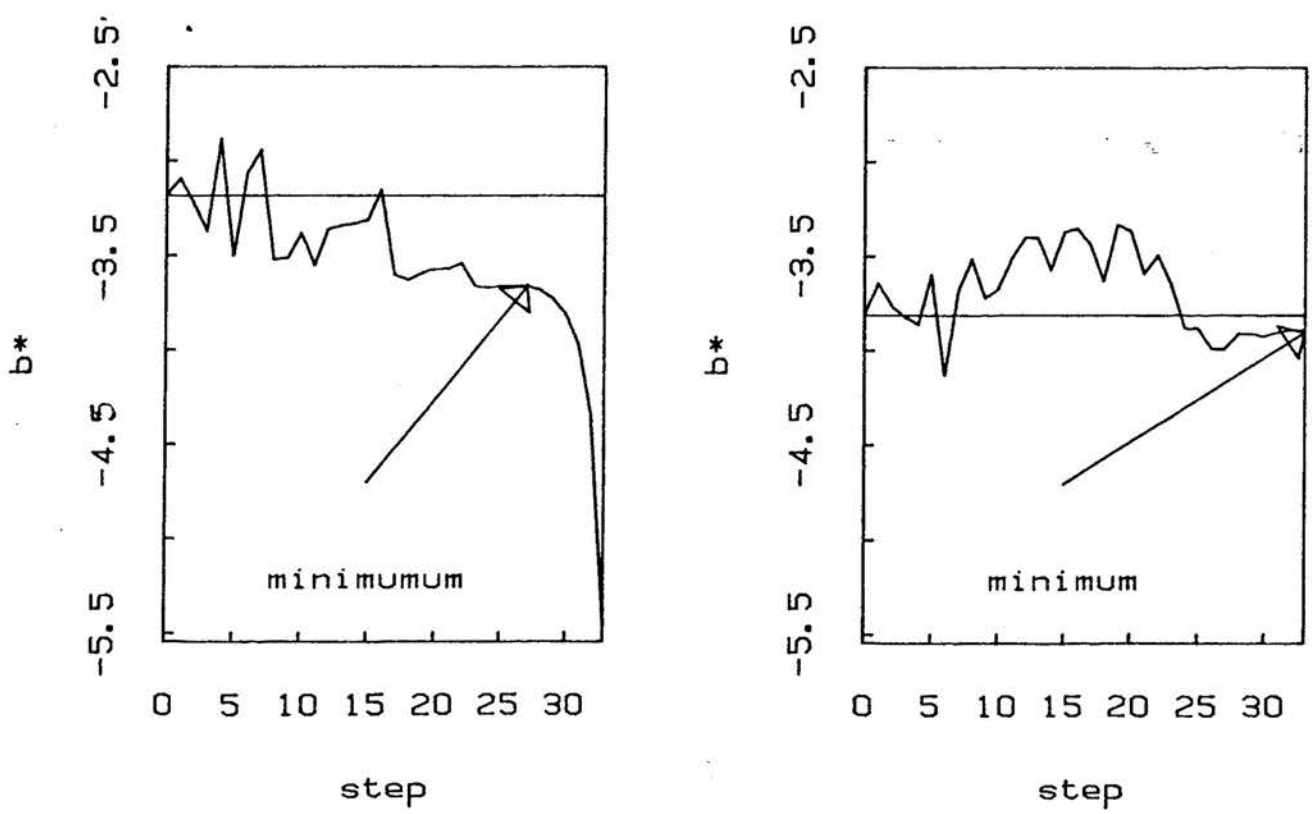
Figure 7.3 : Values of b*, backward elimination with G_{Fp}.



for forward selection. The lowest C_p for table 7.5 was 10.2 when p=8. This presumably occurs because the selection of models which introduce a very small bias in b* will also pick out terms which are unrelated to y. These features were not found for backward selection by G'_{FP} which still seeks negative values of V_{2RP}, and does not find models with low RMS_p.

The minimum values of the criteria tend to occur either at the very end of the procedure, or close to the end. As the variance part of the criteria is strictly decreasing, the minimum value seems to depend on there being a model with a low value of Est(bias^2) (or a negative value in the case of G'_p) among the last few considered.

Figure 7.4 : Values of b*, backward elimination with G'_{FP}.



7.3 Selection by minimising G_{RP}.

The results for a stepwise procedure based on minimising G_{RP} are given in tables 7.7 and 7.8. They share many of the same features as the procedure based on selecting low values of G_{FP}, while differing in being more likely to select variables which predict outcome.

For the unadjusted data the minimum of G_{RP} is achieved at the 9th step, although a value which is almost as low occurs at the 17th step. The value of G_{RP} which is achieved (1.38) is higher than the lowest value (1.33) which was found in the forward selection procedure. For the school-adjusted data the minimum is found at the 11th step (1.50) and this is also higher than the lowest value (1.46) found for the forward selection based on minimising the residual sum-of-squares.

The tendency to select models with exactly the same blood-lead coefficient as the full model is found here, as it was for G_{FP}. This is illustrated in Figure 7.5, where it can be seen to operate less strongly than for selection by G_{FP}.

Negative values of V_{2RP} are also selected preferentially, and this is particularly obvious for the school-adjusted data. In neither case are models selected which give low values of the prediction MSE criteria C_P and S_P.

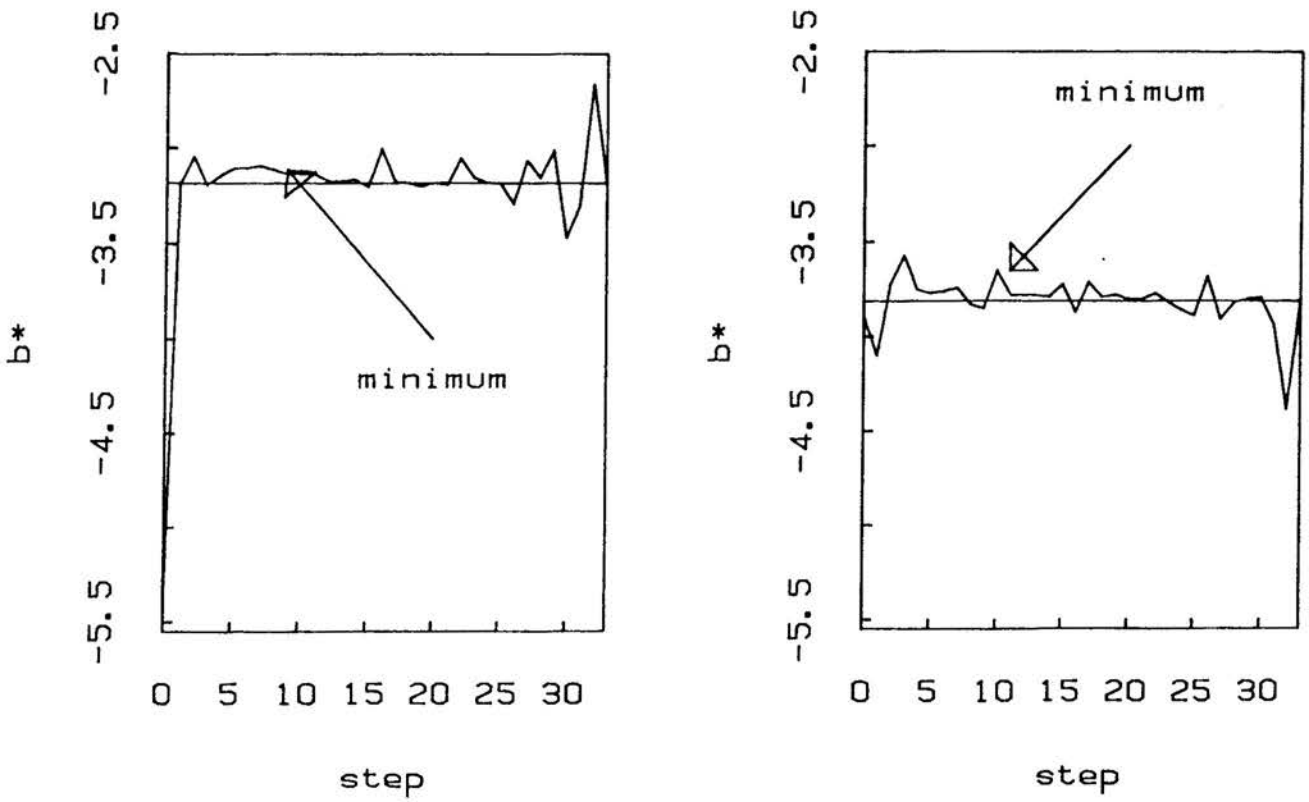
Table 7.7: Results of forward stepwise procedure based on choosing the minimum value of G_{RP}, unadjusted data.

Variable entered	p	estimated bias²	V_{FP}	G_{FP}	V_{1RP}	V_{2RP}	G_{RP}
PVOC	3	-0,22884	1,45593	1,22709	1,76797	0,02554	1,53912
CHINT	4	-0,20761	1,45643	1,24882	1,65177	0,02107	1,44416
CLSYR	5	-0,22576	1,45890	1,23314	1,63023	0,02026	1,40448
MSOC	6	-0,22427	1,45918	1,23490	1,61348	0,01706	1,38920
BTHWT	7	-0,21933	1,45948	1,24014	1,60365	0,01401	1,38431
HNDED	8	-0,21852	1,45948	1,24096	1,60120	0,01072	1,38268
PPART	9	-0,21690	1,45951	1,24261	1,59886	0,00746	1,38196
CPHNE	10	-0,22000	1,45981	1,23981	1,60105	0,00453	1,38105
OCCRT	11	-0,22244	1,46016	1,23772	1,60271	0,00163	1,38027
WMUM	12	-0,22191	1,46017	1,23826	1,60278	-0,00164	1,38088
PCHCM	13	-0,22236	1,46020	1,23784	1,60704	-0,00490	1,38469
CSMR	14	-0,22253	1,46216	1,23963	1,61146	-0,00606	1,38893
PMENT	15	-0,22233	1,46232	1,23999	1,61803	-0,00922	1,39570
PRHLT	16	-0,22144	1,46295	1,24151	1,62440	-0,01189	1,40296
UNMPL	17	-0,21915	1,46534	1,24619	1,62944	-0,01264	1,41029
FQUAL	18	-0,18408	1,46834	1,28426	1,59663	-0,01242	1,41256
OFFSC	19	-0,20824	1,47650	1,26826	1,59188	-0,00691	1,38363
TCIGS	20	-0,20608	1,47869	1,27261	1,59846	-0,00788	1,39238
FHIST	21	-0,20298	1,48148	1,27850	1,60477	-0,00824	1,40179
FSIZE	22	-0,20095	1,48382	1,28287	1,61090	-0,00909	1,40995
BTSCO	23	-0,19644	1,48829	1,29185	1,61758	-0,00766	1,42115
GESTA	24	-0,17653	1,49115	1,31462	1,60359	-0,00787	1,42706
SEX	25	-0,18667	1,49737	1,31070	1,60489	-0,00459	1,41821
TMDAY	26	-0,18109	1,50367	1,32259	1,61118	-0,00124	1,43009
MVSCL	27	-0,17274	1,51203	1,33929	1,61798	0,00431	1,44523
BORD	28	-0,15149	1,52182	1,37034	1,62099	0,01138	1,46951
PSCHL	29	-0,14006	1,53003	1,38997	1,60751	0,01657	1,46745
MDHIS	30	-0,13952	1,54424	1,40472	1,61254	0,02826	1,47302
FSOC	31	-0,09115	1,56468	1,47353	1,61639	0,04655	1,52524
PMAT	32	-0,02433	1,57787	1,55354	1,57176	0,05543	1,54742
MQUAL	33	-0,07755	1,59256	1,51501	1,57235	0,06709	1,49479
STHGT	34	0,24231	1,65514	1,89745	1,55747	0,12665	1,79978
AGE	35	-0,00000	1,68477	1,68477	1,53396	0,15082	1,53396

Table 7.8 : Results of forward stepwise procedure based on choosing the minimum value of G_{RP}, school-adjusted data.

Variable entered	p	estimated bias²	V_{FP}	G_{FP}	V_{1RP}	V_{2RP}	G_{RP}
PMAT	20	-0,17812	1,60653	1,42841	2,00454	-0,00357	1,82642
FQUAL	21	-0,24864	1,61040	1,36176	1,88630	-0,00276	1,63766
CHINT	22	-0,19679	1,61141	1,41462	1,81857	-0,00531	1,62178
CLSYR	23	-0,24919	1,61510	1,36591	1,79475	-0,00490	1,54556
BTHWT	24	-0,25107	1,61518	1,36411	1,78439	-0,00852	1,53331
HNDED	25	-0,25014	1,61521	1,36506	1,77917	-0,01218	1,52903
PPART	26	-0,24782	1,61535	1,36753	1,77854	-0,01573	1,53072
BORD	27	-0,24908	1,61843	1,36935	1,77803	-0,01610	1,52895
WMUM	28	-0,24760	1,61894	1,37133	1,78324	-0,01933	1,53564
PVOC	29	-0,21929	1,62210	1,40282	1,74790	-0,01923	1,52862
CPHNE	30	-0,23973	1,62709	1,38736	1,74327	-0,01756	1,50354
FHIST	31	-0,23979	1,62709	1,38730	1,74951	-0,02130	1,50972
MSOC	32	-0,23924	1,62730	1,38806	1,75651	-0,02487	1,51727
PRHLT	33	-0,23854	1,62892	1,39038	1,76338	-0,02695	1,52483
MQUAL	34	-0,22746	1,63080	1,40334	1,75896	-0,02859	1,53150
SEX	35	-0,22617	1,63781	1,41164	1,75364	-0,02481	1,52746
PSCHL	36	-0,21473	1,64310	1,42837	1,74356	-0,02283	1,52882
CSMR	37	-0,21939	1,64790	1,42852	1,74658	-0,02155	1,52719
PCHCM	38	-0,21700	1,64984	1,43284	1,75400	-0,02336	1,53701
OCCRT	39	-0,21456	1,65315	1,43859	1,76106	-0,02373	1,54650
BTSCO	40	-0,21058	1,65713	1,44655	1,76872	-0,02342	1,55814
PMENT	41	-0,20560	1,66017	1,45457	1,77483	-0,02410	1,56923
UNMPL	42	-0,20300	1,66465	1,46166	1,77997	-0,02328	1,57697
TCIGS	43	-0,19336	1,67232	1,47896	1,78719	-0,01911	1,59383
TMDAY	44	-0,18184	1,68053	1,49869	1,79468	-0,01438	1,61284
GESTA	45	-0,16187	1,68883	1,52696	1,78386	-0,00945	1,62199
OFFSC	46	-0,15303	1,70484	1,55181	1,77933	0,00342	1,62630
FSOC	47	-0,14723	1,72053	1,57330	1,78503	0,01592	1,63780
MVSCL	48	-0,13311	1,73439	1,60128	1,79280	0,02653	1,65969
FSIZE	49	-0,11224	1,75499	1,64275	1,80075	0,04427	1,68851
MDHIS	50	-0,06908	1,78201	1,71292	1,80547	0,06869	1,73638
AGE	51	0,26239	1,79182	2,05421	1,73372	0,07185	1,99612
STHGT	52	-0,00000	1,86778	1,86778	1,72436	0,14342	1,72436

Figure 7.5 : Values of b*, forward selection by G_{RP}.



The random effects criterion can be modified to avoid negative estimates by forcing the values of Est(bias^2) and of V_{2RP} to zero whenever they become negative. This gives a new criterion which we will call G'_{RP}. The results for forward selection by this criterion are given in table 7.9 for the unadjusted data.

Selection to minimise G'_{RP} gives an order of inclusion of the variables closer to the order for the minimisation of the residual sum-of-squares. This was particularly true at the initial steps where the C_p fell consistently. After about step 10 the value of C_p oscillated about the line $C_p=p$, but did not fall far below it. However, the term STHEIGHT was still excluded until near the end

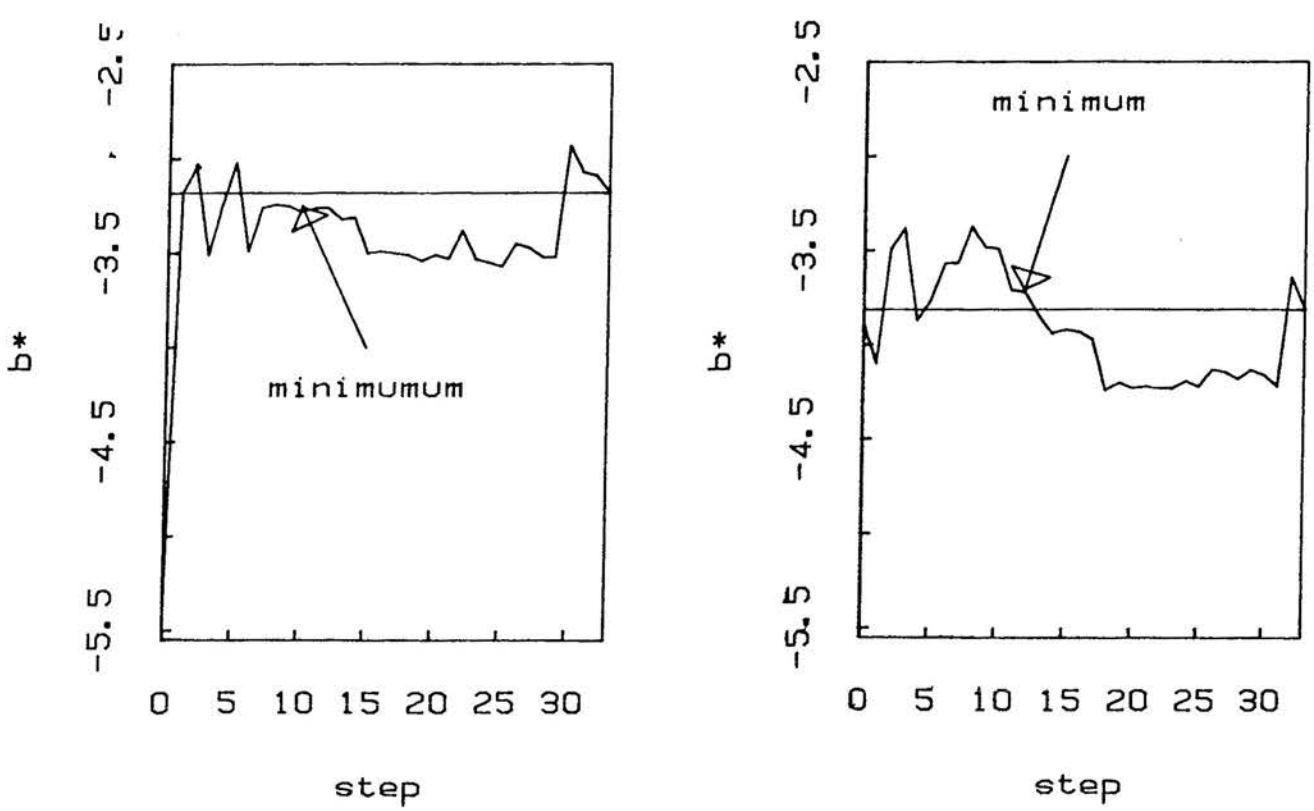
Table 7.9 : Results of forward stepwise procedure based on choosing the minimum value of G'Rp, unadjusted data.

Variable entered	p	estimated bias²	V_{FP}	G_{FP}	V_{1RP}	V_{2RP}	G_{RP}
PVOC	3	-0,22884	1,45593	1,22709	1,76797	0,02554	1,53912
CHINT	4	-0,20761	1,45643	1,24882	1,65177	0,02107	1,44416
PMAT	5	-0,11342	1,46443	1,35101	1,57786	0,02566	1,46445
FQUAL	6	-0,21144	1,46960	1,25815	1,54437	0,02748	1,33293
GESTA	7	-0,18422	1,47424	1,29002	1,51914	0,02877	1,33493
AGE	8	-0,10304	1,48794	1,38490	1,47823	0,03891	1,37518
PSCHL	9	-0,18117	1,49650	1,31534	1,46725	0,04421	1,28609
WMUM	10	-0,18400	1,49668	1,31267	1,46614	0,04128	1,28214
HNDED	11	-0,18297	1,49670	1,31373	1,46591	0,03822	1,28294
CPHNE	12	-0,17824	1,49702	1,31879	1,46555	0,03546	1,28732
PPART	13	-0,18136	1,49731	1,31595	1,46766	0,03272	1,28630
PCHCM	14	-0,18115	1,49731	1,31616	1,47076	0,02970	1,28961
CSMR	15	-0,16633	1,49968	1,33335	1,47258	0,02901	1,30625
PMENT	16	-0,16771	1,49974	1,33203	1,47723	0,02606	1,30952
SEX	17	-0,07235	1,50989	1,43753	1,47266	0,03300	1,40030
MSOC	18	-0,08000	1,51030	1,43030	1,47754	0,03040	1,39754
BTHWT	19	-0,07162	1,51073	1,43911	1,48255	0,02779	1,41092
OCCRT	20	-0,06063	1,51156	1,45093	1,48783	0,02557	1,42720
UNMPL	21	-0,04230	1,51304	1,47074	1,49281	0,02398	1,45051
FSIZE	22	-0,05753	1,51505	1,45752	1,49810	0,02290	1,44057
FHIST	23	-0,04574	1,51617	1,47043	1,50371	0,02092	1,45796
MQUAL	24	-0,11889	1,52532	1,40643	1,50239	0,02688	1,38350
OFFSC	25	-0,02728	1,53294	1,50566	1,49922	0,03123	1,47194
TCIGS	26	-0,00984	1,53585	1,52601	1,50511	0,03102	1,49527
PRHLT	27	0,00519	1,53889	1,54408	1,51121	0,03094	1,51641
FSOC	28	-0,05591	1,55559	1,49968	1,51495	0,04449	1,45905
BTSCO	29	-0,04043	1,55984	1,51941	1,52109	0,04562	1,48066
TMDAY	30	0,00037	1,56755	1,56792	1,52642	0,05020	1,52679
MVSCL	31	0,00972	1,57557	1,58529	1,53293	0,05513	1,54265
STHGT	32	0,00911	1,63229	1,64140	1,51933	0,10782	1,52844
CLSYR	33	-0,02404	1,64857	1,62453	1,52202	0,12076	1,49797
BORD	34	-0,01806	1,65820	1,64014	1,52844	0,12735	1,51038
MDHIS	35	-0,00000	1,68477	1,68477	1,53396	0,15082	1,53396

because of the increase it gives in V_{2Rp}. The selection of models with low, negative values of V_{2Rp} is avoided, as might be expected, because of the modification which sets this component to zero.

The minima of G'_{Rp} occurred at the 10th and 12th step for the unadjusted and school-adjusted data. Selection on this criterion appears to select on the basis of minimum residual sum-of-squares, from among the possible models which have zero estimates for the squared bias contribution and a low value of V_{2Rp}. The values of b* are shown in figure 7.6.

Figure 7.6 : Values of b*, forward selection by G'_{Rp}.



Backward elimination based on G_{Rp} behaved differently from forward selection. The same features were found as for G_{Fp}, described above. Selection towards negative values of V_{2Rp} no longer operated. The order of the variables was closer to those which minimised the residual sum-of-squares than was the case for forward selection on this criterion, and the lowest value of C_p was only very slightly larger than the lowest value found in chapter 6 (6.10 compared to 5.99 at p=15 for unadjusted and 25.08 compared to 24.63 at p=34 for school-adjusted data). A lower value of the G_{Rp} is found for the backwards than the forwards procedure. The results for the unadjusted data are given in Table 7.10, and values of b* are plotted in figure 7.7.

Figure 7.7 : Values of b*, backward selection by G_{Rp}.

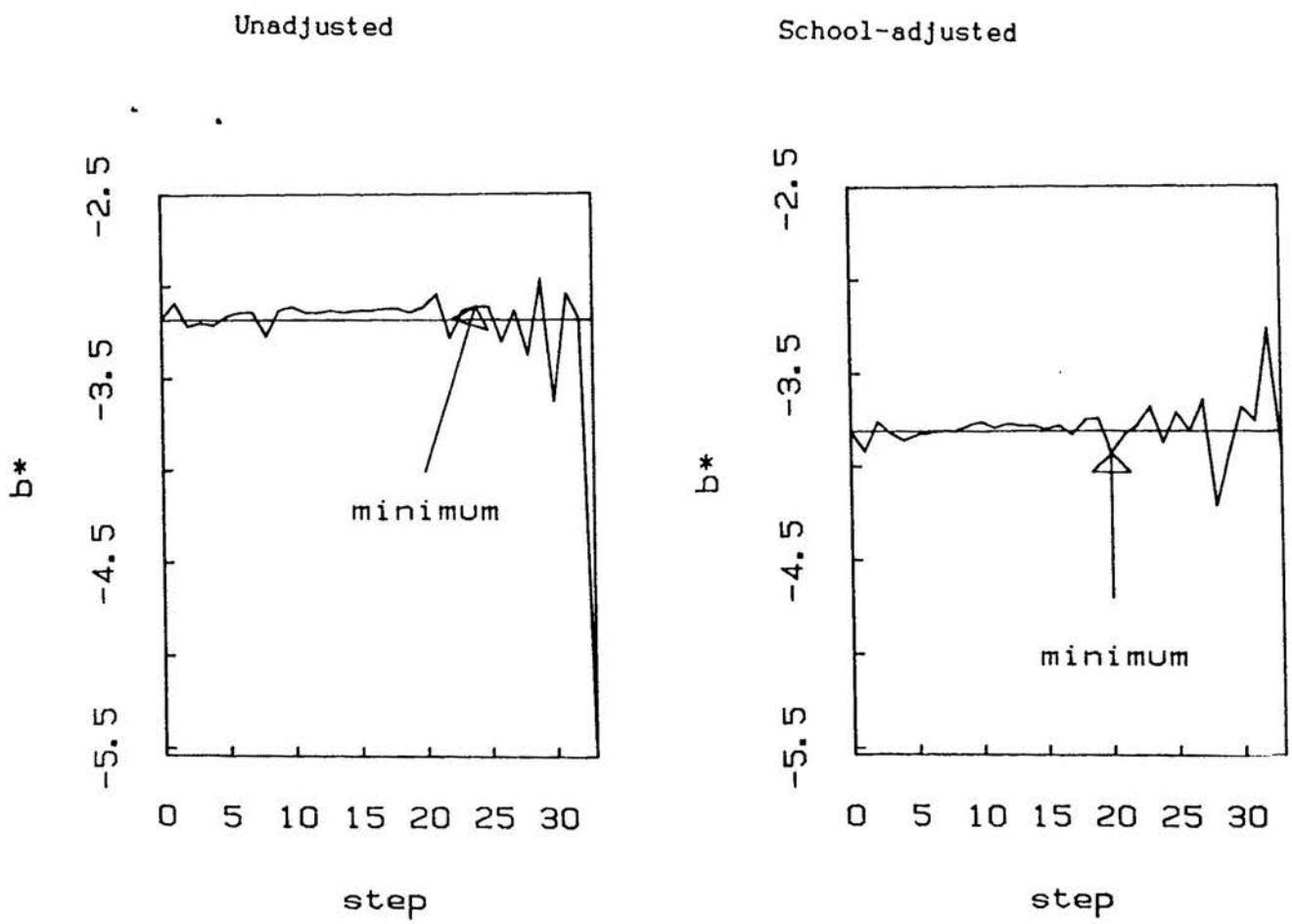


Table 7.10 : Results of backward stepwise procedure based on choosing the minimum value of GRp, unadjusted data.

Variable entered	p	estimated bias²	V_{FP}	G_{FP}	V_{1RP}	V_{2RP}	G_{RP}
MDHIS	35	-0,01806	1,65820	1,64014	1,52844	0,12735	1,51038
FSOC	34	-0,04505	1,63842	1,59338	1,52463	0,11083	1,47959
BORD	33	-0,05689	1,62743	1,57054	1,51819	0,10290	1,46130
MVSCL	32	-0,06620	1,61742	1,55123	1,51178	0,09597	1,44559
TMDAY	31	-0,07721	1,60720	1,53000	1,50626	0,08889	1,42905
BTSCO	30	-0,08058	1,60297	1,52239	1,50006	0,08771	1,41948
OCCRT	29	-0,08271	1,60073	1,51802	1,49372	0,08848	1,41102
MQUAL	28	-0,08421	1,59254	1,50832	1,49361	0,08372	1,40939
CLSYR	27	-0,10426	1,57795	1,47369	1,49159	0,07247	1,38732
TCIGS	26	-0,10548	1,57486	1,46938	1,48558	0,07241	1,38010
PPART	25	-0,10954	1,57396	1,46442	1,48272	0,07466	1,37318
PCHCM	24	-0,10956	1,57392	1,46436	1,47653	0,07756	1,36697
PRHLT	23	-0,10915	1,57315	1,46399	1,47112	0,07976	1,36196
FSIZE	22	-0,11072	1,57289	1,46216	1,46788	0,08257	1,35715
FHIST	21	-0,11058	1,57168	1,46111	1,46243	0,08430	1,35185
PMENT	20	-0,11057	1,57168	1,46112	1,45726	0,08722	1,34669
BTHWT	19	-0,10985	1,57132	1,46147	1,45204	0,08975	1,34219
MSOC	18	-0,10997	1,57132	1,46135	1,44793	0,09268	1,33796
WMUM	17	-0,11254	1,57101	1,45848	1,44771	0,09556	1,33517
UNMPL	16	-0,11016	1,57026	1,46010	1,44561	0,09786	1,33545
CSMR	15	-0,09898	1,56643	1,46745	1,44388	0,09715	1,34491
PSCHL	14	-0,11806	1,55606	1,43800	1,45498	0,09079	1,33691
SEX	13	-0,13341	1,54924	1,41583	1,45879	0,08740	1,32539
HNDED	12	-0,13107	1,54904	1,41798	1,46047	0,09048	1,32941
CPHNE	11	-0,13087	1,54904	1,41818	1,46208	0,09376	1,33121
GESTA	10	-0,12487	1,54565	1,42078	1,48728	0,09513	1,36241
OFFSC	9	-0,14432	1,53771	1,39339	1,49217	0,09051	1,34785
FQUAL	8	-0,11597	1,53312	1,41716	1,52329	0,09085	1,40732
PMAT	7	-0,11035	1,52512	1,41477	1,57869	0,08879	1,46835
STHGT	6	-0,02018	1,47230	1,45212	1,59633	0,03468	1,57615
AGE	5	-0,20761	1,45643	1,24882	1,65177	0,02107	1,44416
CHINT	4	-0,22884	1,45593	1,22709	1,76797	0,02554	1,53912
PVOC	3	4,90704	1,43231	6,33935	2,38684	0,00000	7,29388

By contrast, the G'_{Rp} procedure selects rather similar models whether forward or backwards stepwise procedures are used, especially for the models with few covariates. The results for backwards elimination are given in table 7.11 for the unadjusted data, which can be compared to table 7.9 for forward selection. The values of b* are plotted in figure 7.8, these can be seen to be similar to the mirror images of figure 7.6, as the same terms are retained in the backwards elimination models as were entered into the forwards procedure. The backwards procedures based on G'_{Rp} tends to select models with lower RMS_p than the equivalent forward procedures, with C_p now falling somewhat below p, but not as low as for backwards elimination with G_{Rp}.

Figure 7.8 : Values of b*, backward selection by G'_{Rp}.

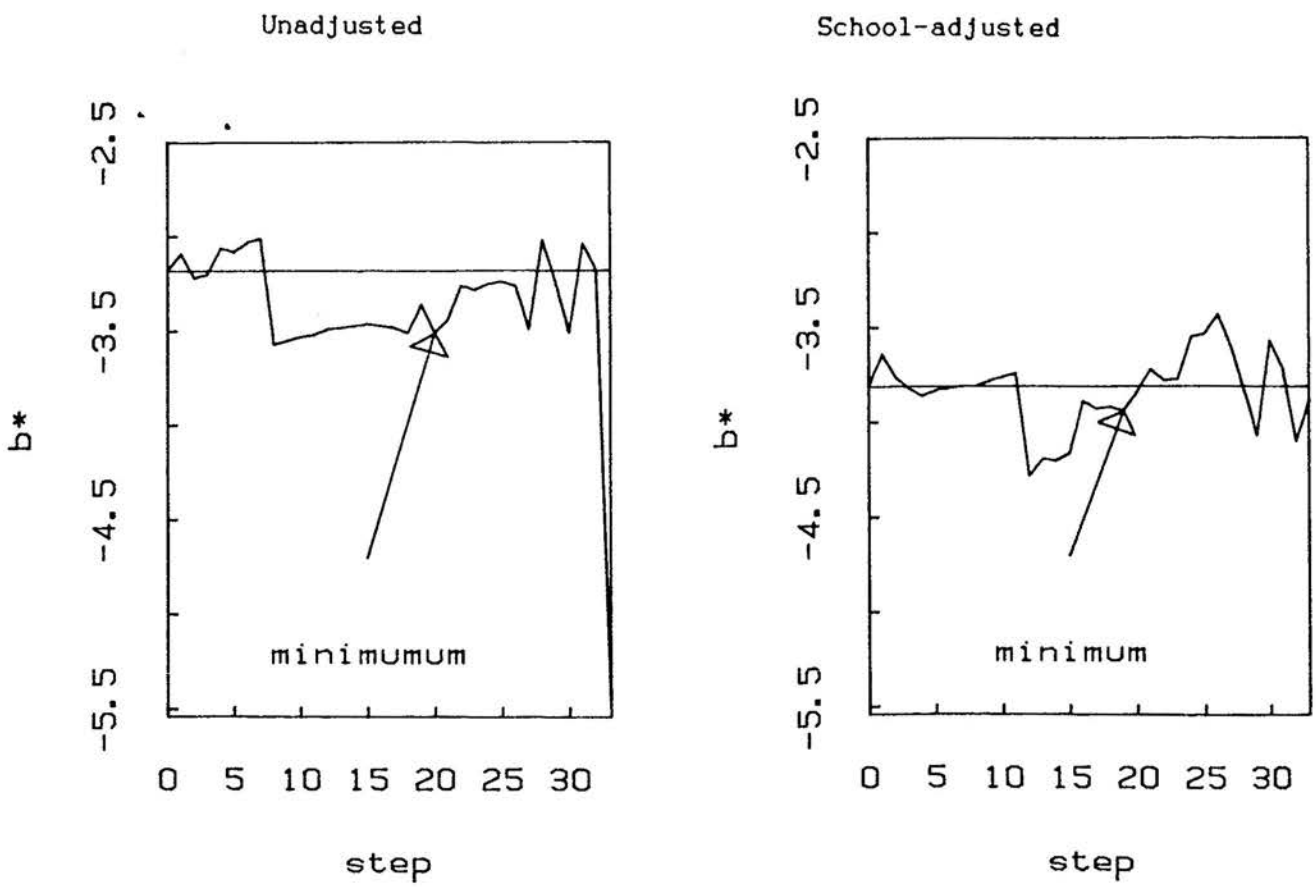


Table 7.11 : Results of backward stepwise procedure based on choosing the minimum value of G'_{Rp}, unadjusted data.

Variable entered	p	estimated bias²	V_{Fp}	G_{Fp}	V_{1Rp}	V_{2Rp}	G_{Rp}
MDHIS	35	-0,01806	1,65820	1,64014	1,52844	0,12735	1,51038
FSOC	34	-0,04505	1,63842	1,59338	1,52463	0,11083	1,47959
BORD	33	-0,05689	1,62743	1,57054	1,51819	0,10290	1,46130
CLSYR	32	-0,05869	1,61303	1,55434	1,51587	0,09185	1,45718
MVSCL	31	-0,07145	1,60266	1,53121	1,50947	0,08457	1,43802
TMDAY	30	-0,06911	1,59346	1,52435	1,50381	0,07849	1,43470
BTSCO	29	-0,06771	1,58869	1,52098	1,49770	0,07679	1,42999
STHGT	28	0,00519	1,53889	1,54408	1,51121	0,03094	1,51641
PRHLT	27	-0,00984	1,53585	1,52601	1,50511	0,03102	1,49527
TCIGS	26	-0,02728	1,53294	1,50566	1,49922	0,03123	1,47194
OCCRT	25	-0,03368	1,53156	1,49788	1,49305	0,03293	1,45937
UNMPL	24	-0,05914	1,52893	1,46978	1,48824	0,03340	1,42910
MSOC	23	-0,06454	1,52834	1,46380	1,48236	0,03586	1,41781
FHIST	22	-0,07249	1,52767	1,45518	1,47688	0,03823	1,40438
BTHWT	21	-0,07830	1,52739	1,44909	1,47178	0,04098	1,39348
PCHCM	20	-0,07310	1,52715	1,45405	1,46683	0,04374	1,39373
PMENT	19	-0,06921	1,52708	1,45786	1,46252	0,04667	1,39331
FSIZE	18	-0,04968	1,52597	1,47629	1,45945	0,04860	1,40977
OFFSC	17	-0,13263	1,51800	1,38537	1,46152	0,04389	1,32888
MQUAL	16	-0,06339	1,50967	1,44627	1,46361	0,03878	1,40022
CSMR	15	-0,11100	1,50647	1,39546	1,46288	0,03867	1,35188
SEX	14	-0,18136	1,49731	1,31595	1,46766	0,03272	1,28630
PPART	13	-0,17824	1,49702	1,31879	1,46555	0,03546	1,28732
CPHNE	12	-0,18297	1,49670	1,31373	1,46591	0,03822	1,28294
HNDED	11	-0,18400	1,49668	1,31267	1,46614	0,04128	1,28214
WMUM	10	-0,18117	1,49650	1,31534	1,46725	0,04421	1,28609
PSCHL	9	-0,10304	1,48794	1,38490	1,47823	0,03891	1,37518
AGE	8	-0,18422	1,47424	1,29002	1,51914	0,02877	1,33493
GESTA	7	-0,21144	1,46960	1,25815	1,54437	0,02748	1,33293
FQUAL	6	-0,11342	1,46443	1,35101	1,57786	0,02566	1,46445
PMAT	5	-0,20761	1,45643	1,24882	1,65177	0,02107	1,44416
CHINT	4	-0,22884	1,45593	1,22709	1,76797	0,02554	1,53912
PVOC	3	4,90704	1,43231	6,33935	2,38684	0,00000	7,29388

7.4 Summary of the results of using the G_p criteria for selection.

Several features of these procedures can be identified, which suggest that they may be unsuitable either to drive a selection procedure, or as the estimator of the MSE after such a procedure.

The first such feature is the tendency for the criteria to select models which give exactly the same b* as the full model. This is at its worst for forwards selection by G_{Fp} or G_{Rp}, but less marked when these criteria are used for backward elimination. This is likely to be undesirable because it may introduce noise by selecting variables which by chance give low values of the bias^2 term; it will also give a systematic downward bias to the values of G_p after selection; and it makes the whole procedure seem irrelevant if all one ends up with is the full model estimate. The modified criteria G'_{Fp} and G'_{Rp} are less affected.

Secondly, we find evidence of selection towards negative values of V_{2Rp}. This is most marked for G'_{Fp} where both forward and backward selection have this feature. It is also seen for forward selection on G_{Fp} and G_{Rp}. The absolute effect of this term is not large for these data, compared with the other terms in the criteria. However, other data sets where X* is more strongly related to the other covariates might show this feature to a larger extent. Again, this may lead to problems because of the noise introduced by irrelevant selection procedures, and the resultant under-estimates of the G_p.

Finally, in common with the procedures examined in chapter 6, we may have under-estimation of the residual variance. The models which were most affected by this here were backwards elimination by G_{FP} or G_{RP} (especially the former). Backward selection by G'_{RP} also produced low values of the residual variance, but for forward selection by G'_{RP} the value of RMS_P did not fall far below its value for the full model.

The minima of the G_{FP} criteria were always achieved for models with very few covariates. That for the G_{RP} criteria occurred with about nine covariates included when the criteria themselves were used for selection, but for larger models when G_{FP} was used for selection. This is because the procedures which are driven by G_{FP} do not necessarily select models with low residual sums of squares.

7.5 Selection for the maximum change in b_P^*.

In observing what happens to the various G_P criteria during the process of variable selection, the argument has strayed from the original derivation of the criteria as estimators of the MSE of b^*. It is clear from the results above that this property is most unlikely to hold when the criteria themselves are used to select variables.

The natural statistic to observe when we are concerned with estimating β^* is its estimate as plotted in figures 7.1 to 7.8, above. If variables are selected which ensure that we include in the model every term which is likely to alter this estimate

substantially, then we may obtain reduced models which are of the type we are seeking, and we may also be able to use the G_p criteria to estimate the MSE of b* from such models. The advantage of forward selection by this criteria, rather than the G_p, is that the procedure driven by changes in b* does not make use of the value of the b* estimate for the full model. Thus it will not be forced to select terms for which est(bias^2) is negative. The same would not be true for backward selection by minimising the changes in b*, so this is not reported here.

The values of the various criteria for such selection procedures are given in tables 7.12 and 7.13, and the changes in b* are plotted in figure 7.9. The values of the squared absolute change in b* are also given as Δ^2b*. This criterion could be used to form a stopping rule if the procedure was stopped once Δ^2b* changed by less than a small fraction of V_{Fp} or V_{Gp}.

Figure 7.9 : Values of b*, selection by maximum change in b*.

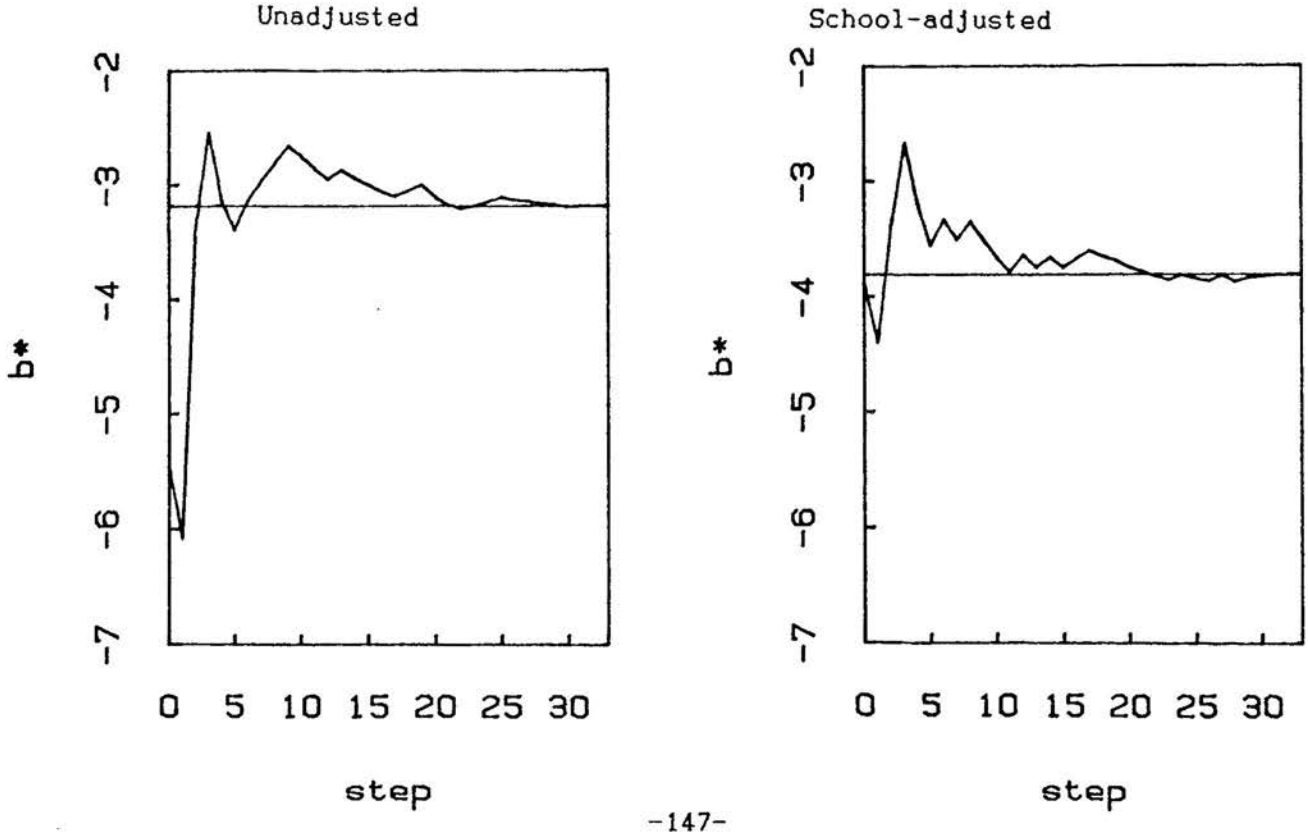


Table 7.12 : Results of selection by maximum change in b*, unadjusted data.

Variable entered	p	est(bias)^2	$\Delta(b^*)^2$	V_{FP}	G_{FP}	V_{1RP}	V_{2RP}	G_{RP}
AGE	3	8,1968	71,74928	1,44491	9,64179	2,30474	0,01561	10,50162
MQUAL	4	-0,1595	7,27307	1,48211	1,32259	1,74827	0,05352	1,58876
STHGT	5	0,2598	0,72392	1,53108	1,79096	1,71239	0,10706	1,97227
PMAT	6	-0,1425	0,36814	1,54089	1,39832	1,60980	0,10812	1,46723
OFFSC	7	-0,0911	0,06122	1,54909	1,45796	1,59447	0,11268	1,50334
FSOC	8	-0,1236	0,06108	1,55987	1,43626	1,58429	0,12030	1,46068
PVOC	9	-0,0726	0,03507	1,56229	1,48969	1,54807	0,11676	1,47547
PSCHL	10	0,0267	0,02196	1,57347	1,60017	1,54857	0,12529	1,57527
GESTA	11	0,1648	0,02277	1,57674	1,74160	1,53419	0,12418	1,69904
SEX	12	0,0795	0,00902	1,58160	1,66111	1,53492	0,12595	1,61444
MDHIS	13	0,0145	0,01282	1,60055	1,61506	1,53920	0,14280	1,55370
CLSYR	14	-0,0116	0,00645	1,61840	1,60673	1,54444	0,15861	1,53277
PPART	15	0,0300	0,00562	1,61942	1,64946	1,53408	0,15513	1,56412
CSMR	16	0,0012	0,00456	1,62780	1,62906	1,53877	0,16085	1,54003
FQUAL	17	-0,0254	0,00434	1,62854	1,60309	1,52737	0,15695	1,50191
HNDED	18	-0,0395	0,00275	1,63013	1,59060	1,52839	0,15520	1,48886
TMDAY	19	-0,0361	0,00151	1,64158	1,60545	1,53433	0,16415	1,49820
CPHNE	20	-0,0263	0,00189	1,64220	1,61589	1,53135	0,16094	1,50504
TCIGS	21	-0,0074	0,00224	1,64679	1,63938	1,53623	0,16265	1,52882
CHINT	22	-0,0321	0,01029	1,64724	1,61509	1,47151	0,15285	1,43936
BORD	23	-0,0294	0,00570	1,65528	1,62579	1,47546	0,15778	1,44597
UNMPL	24	-0,0275	0,00101	1,65610	1,62857	1,47775	0,15540	1,45023
FSIZE	25	-0,0258	0,00081	1,65891	1,63307	1,48306	0,15529	1,45722
WMUM	26	-0,0232	0,00222	1,65984	1,63664	1,48180	0,15263	1,45859
MVSCL	27	-0,0084	0,00065	1,67182	1,66338	1,48789	0,16162	1,47945
PRHLT	28	-0,0097	0,00029	1,67250	1,66275	1,49284	0,15934	1,48309
PMENT	29	-0,0105	0,00022	1,67297	1,66243	1,49770	0,15681	1,48716
FHIST	30	-0,0086	0,00035	1,67583	1,66716	1,50367	0,15673	1,49500
BTSCO	31	-0,0050	0,00012	1,67968	1,67461	1,50998	0,15766	1,50491
BTHWT	32	-0,0045	0,00018	1,68018	1,67564	1,51527	0,15514	1,51073
PCHCM	33	-0,0042	0,00009	1,68053	1,67629	1,52091	0,15249	1,51667
MSOC	34	-0,0023	0,00002	1,68238	1,67999	1,52740	0,15139	1,52501
OCCRT	35	-0,0000	0,00001	1,68477	1,68477	1,53396	0,15082	1,53396

Table 7.13 : Results of selection by maximum change in b*, school-adjusted data.

Variable entered	p	est(bias)^2	$\Delta(b^*)^2$	V_{FP}	G_{FP}	V_{1RP}	V_{2RP}	G_{RP}
AGE	20	0,09468	45,01846	1,61860	1,71327	2,25050	0,01286	2,34517
STHGT	21	0,01175	1,05651	1,68464	1,69639	2,20615	0,09833	2,21790
FSOC	22	1,13789	0,49127	1,70066	2,83855	2,11190	0,11046	3,24979
PMAT	23	0,22027	0,27382	1,70477	1,92505	1,89540	0,09978	2,11567
OFFSC	24	-0,08582	0,13559	1,71903	1,63321	1,87100	0,11080	1,78518
PVOC	25	0,08594	0,05360	1,72102	1,80697	1,78715	0,10405	1,87309
MDHIS	26	-0,03075	0,03133	1,74378	1,71303	1,78996	0,12521	1,75921
GESTA	27	0,09720	0,02604	1,74714	1,84435	1,77087	0,12353	1,86807
CLSYR	28	-0,02112	0,02731	1,75576	1,73464	1,76741	0,12860	1,74629
SEX	29	-0,07999	0,02222	1,76456	1,68457	1,76614	0,13396	1,68615
FQUAL	30	-0,10131	0,01698	1,76598	1,66467	1,73195	0,12891	1,63064
PSCHL	31	-0,06312	0,02289	1,77459	1,71148	1,73002	0,13385	1,66691
UNMPL	32	-0,08381	0,01172	1,77973	1,69591	1,72938	0,13522	1,64557
TCIGS	33	-0,05648	0,00851	1,78652	1,73005	1,73236	0,13858	1,67588
CSMR	34	-0,06863	0,00851	1,79490	1,72626	1,73619	0,14364	1,66756
MVSCL	35	-0,03079	0,00627	1,81615	1,78536	1,74261	0,16241	1,71182
MQUAL	36	-0,00222	0,00519	1,81872	1,81650	1,74287	0,16103	1,74065
BTSCO	37	-0,01948	0,00254	1,82072	1,80123	1,74584	0,15928	1,72636
CPHNE	38	-0,03174	0,00197	1,82123	1,78949	1,73944	0,15513	1,70770
HNDED	39	-0,04003	0,00305	1,82333	1,78330	1,74173	0,15341	1,70170
BTHWT	40	-0,04123	0,00200	1,82608	1,78485	1,74665	0,15258	1,70543
PRHLT	41	-0,03988	0,00154	1,82760	1,78773	1,75058	0,15037	1,71070
PCHCM	42	-0,03692	0,00116	1,82821	1,79129	1,75124	0,14692	1,71432
PPART	43	-0,03789	0,00167	1,82978	1,79189	1,75498	0,14471	1,71709
TMDAY	44	-0,02541	0,00047	1,84133	1,81592	1,76252	0,15317	1,73711
FSIZE	45	-0,01506	0,00056	1,84959	1,83453	1,77002	0,15821	1,75496
CHINT	46	-0,01799	0,00206	1,84968	1,83169	1,69243	0,14732	1,67444
BORD	47	-0,00668	0,00310	1,85672	1,85004	1,69825	0,15077	1,69157
WMUM	48	-0,00960	0,00131	1,85728	1,84767	1,69693	0,14713	1,68733
MSOC	49	-0,00926	0,00030	1,85836	1,84910	1,70342	0,14467	1,69415
PMENT	50	-0,00931	0,00002	1,85841	1,84910	1,70932	0,14110	1,70001
OCCRT	51	-0,00006	0,00004	1,86772	1,86766	1,71691	0,14690	1,71685
FHIST	52	-0,00000	0,00000	1,86778	1,86778	1,72436	0,14342	1,72436

Very different variables are selected from those seen by any of the previous procedures, with terms being entered which increase the term V_{2RP} (the opposite of what happened in many of the procedures described above). The models considered do not have low values of RMS_P until after the variable CHLDINT is included. This occurs at steps 20 and 23 for the unadjusted and school-adjusted data. The results for the G_P criteria suggest that the procedure may give a modest improvement over the full model for the fixed effects criterion if it is stopped after the first few steps, but an improvement for the random effects criterion is not found until after CHILDINT enters, and thus is only very small. For the random effects model some modification of this procedure might be required to make sure that all the terms which predict outcome strongly come into the model early.

7.6 Including variables which are related to blood lead.

Some studies (eg Needleman et al 1979) have used a strategy of controlling for only those variables which are related to the exposure being studied (here blood lead). The discussion and review in chapters 3 and 4 suggests that this will tend to underestimate the effect of the exposure.

One possible strategy is to include those covariates whose univariate associations with blood lead reach a certain level of significance. The covariates which are included by using $p<0.05$ and $p<0.01$ for the adjusted and unadjusted data are given in table 7.14,

along with the resulting value for the lead coefficient and its t-value.

Table 7.14: Selecting covariates related to blood lead.

Unadjusted data

p-value	Covariates	b*	t-ratio
0.05	STHEIGHT MQUALIF FSOC PVOC FQUALIF AGEINT	-2.33	-1.79
0.01	STHEIGHT MQUALIF FSOC PVOC FQUALIF	-1.78	-1.36

School-adjusted data

p-value	Covariates	b*	t-ratio
0.05	STHEIGHT FSOC MEDHIST MOVESCHL PVOC AGEINT	-2.67	-1.88
0.01	STHEIGHT FSOC	-2.13	-1.41

These values are much lower in absolute value than any achieved in any of the stepwise procedures described above (data not shown for most of these). None of the other models examined gave a t-ratio which was less extreme than -2.0, while all of these four do. To understand why this is so we can examine the values for the various MSE criteria and their components for these models.

Table 7.15: Components of G_p for models in table 7.14..

Unadjusted data

p-value	RMS_p	Est(bias)^2	V_{FP}	G_{FP}	V1Rp	V_{2Rp}	G_{Rp}
0.05	115.5	0.5801	1.5424	2.1226	1.5963	0.1021	2.2785
0.01	118.4	1.7949	1.5215	3.3164	1.6326	0.0847	3.5122

School-adjusted data

p-value	RMS_p	Est(bias)^2	V_{FP}	G_{FP}	V1Rp	V_{2Rp}	G_{Rp}
0.05	114.7	1.1730	1.7427	2.1957	1.8872	0.1360	3.1962
0.01	133.2	2.6441	1.6873	4.3315	2.1747	0.1002	4.9190

Apart from the bias term the other main difference from the other models examined so far is the larger values of RMS_p and hence of V_{1Rp}. This is because all the other methods discussed have been influenced by the relationship between y and the other covariates, whereas this method ignores y in the variable selection. Also, as for the selection of variables which will change b*, these models have relatively large values of V_{2Rp}.

Chapter 8

A review of models examined, and a new model

8.1 Introduction

This chapter is really a continuation of chapter 4, in that it extends the range of models considered by looking at models where X* is considered as a realisation of a random variable. The need to consider such models was a result of studying the empirical patterns of results which appeared in the last chapter. The consideration of how the strategies described in the previous chapters should be evaluated by simulation helps to clarify which models are really appropriate.

8.2 Components of the G_p criteria under fixed and random models.

In chapters 6 & 7 the value of the G_p criteria were investigated for a selection of sub-models. Although certain patterns emerged, they were complex, and suggested a need to look at the component parts of the criteria.

The Gp criteria can be built up in terms of the following :

n	- number of observations
k	- parameters in full model, including constant and X*
s_k^2	- residual mean square of Y from regression with X* and X
S_{**}	- sum of squares of X*

and for the particular sub-model considered

$\mathbf{p}$ - number of parameters

$\mathbf{s_p^2}$ - residual mean square of Y

$\mathbf{S_{*\ast p}}$ - residual sum of squares of X* from the regression with the other covariates in the model

$\mathbf{\Delta b^{*2}}$ - squared difference of estimate of β^* from full model value

Bold type in this section will indicate a quantity which is different for different sub-models. The first three such quantities are the number of parameters, two quantities which relate respectively to the relationships of Y to all the covariates in the model, and of X* to the other covariates. The fourth quantity $\mathbf{\Delta b^{*2}}$ depends on the preceding ones, and also on the relationship between X* and Y. $\mathbf{\Delta b^{*2}}$ will be zero if either

$$\mathbf{S_{*\ast p}} = S_{*\ast k} \qquad \text{or} \qquad \mathbf{s_p^2}(n-\mathbf{p}) = s_k^2(n-k),$$

although it can also be zero when these equalities do not hold.

We can write

$$G_{F\,p} = \mathbf{\Delta b^{*2}} - (S_{*\ast k}^{-1} - \mathbf{S_{*\ast p}^{-1}})s_k^2 + \mathbf{S_{*\ast p}^{-1}}s_k^2 \ldots\ldots (8.1)$$

and

$$G_{R\,p} = \mathbf{\Delta b^{*2}} - (S_{*\ast k}^{-1} - \mathbf{S_{*\ast p}^{-1}})s_k^2 + \mathbf{S_{*\ast p}^{-1}}\mathbf{s_p^2} \ldots\ldots (8.2)$$

The random-effects model allows us to divide the term $S_{*\ast p}^{-1}\mathbf{s_p^2}$ into two parts. This interpretation is not available for $G_{F\,p}$, but in chapters 6 & 7 we saw that one of these two parts appeared to be an influence on selection, when $G_{F\,p}$ was driving the

selection procedure. Also, for neither the fixed nor random effects model was it possible to consider the distributional properties of the term $(S_{*\ast k}^{-1} - S_{\ast\ast p}^{-1})s_k^2$, which is the expected value of the squared bias when the true bias is zero. But this quantity also seemed to be an important element in the selection procedures. To achieve both these aims, we need to consider X* as the realisation of a random variable, and introduce yet another regression model, which reduces to the model described in section 4.2 conditional on the observed values of X*.

8.2 A model with random y and x*

Consider the X variables (except for x*) as fixed effects, and suppose that y and x* are each defined in terms of the random quantities ε_y and ε_* as follows :

$$y = X\delta + \varepsilon_y$$
$$x^* = X\delta_* + \varepsilon_*. \quad \ldots\ldots\ldots\ldots\ldots \quad (8.3)$$

Let the joint distribution of ε_y and ε_* be such that

$$E(\varepsilon_y \mid \varepsilon_* = e_*) = e_* \beta^*,$$

and the distribution of ε_y conditional on $\varepsilon_* = e_*$ is independent of e_* with variance σ^2. Then we can write

$$E(y \mid \varepsilon_* = e_*) = X\delta + e_* \beta^*$$
$$= X\delta + (X^* - X\delta_*)\beta^*$$
$$= X(\delta - \delta_* \beta^*) + X^* \beta^*$$

which has the desired form with the vector $(\delta - \delta_* \beta^*)$ being equivalent to the vector β introduced in chapter 3 and used in the fixed-effects model in section 4.2.

Now estimation of β^* is exactly equivalent to estimation from the fixed effects model. However we can derive it from first estimating δ and δ_*, by d_* and d, from 8.3, as follows

$$d = (X'X)^{-1} X'Y$$
$$d_* = (X'X)^{-1} X'X^*.$$

The residuals e_y and e_* become

$$e_y = (1 - X(X'X)^{-1}X')\, Y \text{ and}$$
$$e_* = (1 - X(X'X)^{-1}X')\, X^*$$

and the estimate of β^* can be obtained from the regression of e_y on e_*.

This gives

$$b^* = [X^{*\prime}(1-X(X'X)^{-1}X')(1-X(X'X)^{-1}X')X^*]^{-1}$$
$$X^{*\prime}(1-X(X'X)^{-1}X')(1-(X'X)^{-1}X')Y$$

$$= [X^{*\prime}\ (1-X(X'X)^{-1}X')\ X^*]^{-1}\ X^{*\prime}\ (1-X(X'X)^{-1}X')\ Y$$

with variance $[X^{*\prime}\ (1-X(X'X)^{-1}X')\ X^*]^{-1}\ \sigma^2$..... (8.4)

This result can be compared with the same expression derived from the usual matrix inversion procedures in section 4.2. Now for the model considered here (8.4) is a realisation of a function of x^* which is

$[x^{*\prime}\ (1-X(X'X)^{-1}X')\ x^*]^{-1}\ \sigma^2$..... (8.5)

The term $[X^{*\prime}\ (1-X(X'X)^{-1}X')\ X^*]$ is just S_{***} and its expectation will be $(n - (k-1))\ \sigma_*^2 = (n-k+1)\ \sigma_*^2$, where σ_*^2 is the variance of ε_* in 8.3. When ε_* and ε_y are assumed to be bivariate normal, then S_{***} will be distributed as σ_*^2 times a χ^2 variable with (n-k+1) degrees of freedom. Now the expectation of the inverse of a quantity with a χ^2 distribution with ν degrees of freedom is $1/(\nu-2)$. Thus

$$E(\ S_{***}^{-1}\ \sigma^2) = \sigma^2\ /\ [\ \sigma_*^2\ (n-k-1)\ ]$$..... (8.6)

For the case when all the δ^* are zero the estimate of (8.6) can be obtained by estimating σ_*^2 from the marginal distribution of X^*. In this case S_{**}^2 is distributed as σ_*^2 times χ^2 with (n-1) degrees of freedom. Thus the expected value of $1/S_{**}^2$ is $1/[\sigma_*^2(n-3)]$ and $(n-3)/S_{**}^2$ is an unbiassed estimate of $1/\sigma_*^2$. Substituting this into 8.6 gives

$$\sigma^2(n-3)\ /[S_{**}^2(n-k-1)]$$ ..... (8.7)

which is of the same form as 4.12. The same argument follows through for a sub-matrix of X, when all the δ_* corresponding to the sub-matrix of X included in the model are zero, except that k is replaced by the number of covariates p in the model.

The case when the δ^* are non-zero is different, because σ_*^2 has to be estimated from the partial residuals, which takes us back to 8.4. If we attempt to get an expression for the expectation of 8.5 in terms of the parameters δ^*, σ_*^2 and the fixed quantities X the distribution theory gets complicated, because we have a quantity with a non-central χ^2 distribution in the denominator. Unlike the case for the previous random effects model, we do not get a simple expression which falls into two parts one of which is 8.7. However, the variance for non-zero δ^* will be strictly greater than 8.6 and expression 8.7, with σ^2 replaced by s_k^2 becomes V_{1FP}, the first part of the variance element of G_{FP}, with the second part (V_{2FP}) being obtained by subtraction.

It is possible to extend this model to allow the covariates in the regression model to be realisations of random variables, with the conditions necessary for the regression model to be valid. This leads to another justification of the terms V_{1RP} and V_{2RP}, which does not require multivariate normality for y and the other covariates, but does require bivariate normality for y and x^*.

The fact that we can derive the same sub-division of the variance term in the MSE criteria from two quite different regression models is encouraging. It suggests that it may have

general validity which may go beyond precise model assumptions, so we can hope that it may be a robust procedure.

The model which assumes bivariate normality for y and x* also enables us to compute an expectation for the variance of Δb^{*2}, for the case when all the δ_* for the terms omitted from the model are zero. This applies to both the fixed-effects model and the random-effects model, because conditioning on all the covariates is necessary to estimate the bias.

The term for the variance of Δb^{*2} is $(S_{***k}^{-1} - S_{***p}^{-1})\sigma^2$, the estimate of which is subtracted from Δb^{*2} in 8.1 and 8.2. If the δ^* for the terms omitted from the model are all zero, then S_{***k} and $(S_{***k} - S_{***p})$ will be independently distributed as σ_*^2 times χ^2 distributions with degrees of freedom (n-k) and (k-p). Thus

$$(S_{***k}^{-1} - S_{***p}^{-1}) = B / [A (A + B) \sigma_*^2] \quad \ldots\ldots\ldots (8.8)$$

where A and B are independent χ^2s with (n-k) and (k-p) degrees of freedom, respectively. We can obtain an approximation to the expectation of this by a Taylor's theorem expansion as

$$E(B/A \cdot 1/(A+B)) \approx E(B/A)\, E[1/(A+B)] + \mathrm{cov}[B/A \cdot 1/(A+B)]$$

the last term can be evaluated similarly as

$$\mathrm{cov}[B/A.\ 1/(A+B)] \approx \frac{E(B)\ \ \mathrm{var}(A)}{E(A)^2\ [E(A)+E(A)]^2} - \frac{\mathrm{var}(B)}{E(A)\ [E(A)+E(B)]^2} \quad \ldots (8.9)$$

Because B/A is the ratio of independent χ^2s and (A+B) is χ^2 with n-p degrees of freedom, we know the values of everything in these expressions and find that 8.9 reduces to zero, giving

$$E(S_{***k}^{-1} - S_{***p}^{-1}) \approx \frac{(k-p)}{(n-k-2)(n-p-2)\ \sigma_{**}^2}.$$

Similarly, we would expect the stochastic properties of 8.8 to be similar to those of χ^2_{k-p} divided by $(n-k-2)(n-p-2)\ \sigma_{**}^2$. We can approximate the variance by Taylor's expansion replacing A and B by their expected values in the following expressions

$$\begin{aligned}
\mathrm{var}\{B/[A(A+B)]\} &\approx \mathrm{var}(A)\ \{B(2A+B)/[A^2(A+B)^2]\}^2 + \mathrm{var}(B)\ \{1/(A+B)^2\}^2 \\
&\approx 2A\ \{B(2A+B)/[A^2(A+B)^2]\}^2 + 2B\ \{1/(A+B)^2\}^2 \\
&\approx 2B\ \{\ (A^3+4A^2B+2AB^2+B^3)/A^3\ )\ /\ (A+B)^4.
\end{aligned}$$

This is not quite as simple as we might have hoped. But when n >> k and p the term in {} ≈ 1 and (A+B) ≈ A, giving

$$\begin{aligned}
\mathrm{var}\{B/[A(A+B)]\} &\approx 2B\ /[\ A^2\ (A+B)^2] \\
&\approx 2(k-p)\ /\ [\ (n-k)(n-p)\ ]^2 \text{ and thus}
\end{aligned}$$

$$\mathrm{var}(S_{***k}^{-1} - S_{***p}^{-1}) \approx 2(k-p)\ \{(n-k-2)(n-p-2)\ \sigma_{**}^2\}^2,$$

which is the same as the variance of χ^2_{k-p} divided by $(n-k-2)(n-p-2)\ \sigma_{**}^2$, justifying the approximation suggested above.

For the case when the δ^* for the omitted covariates are not all zero, the quantity B will be a non-central χ^2, and so the

expression for the variance of the estimated bias will be stochastically greater than in the case when the δ^* are all zero.

This result explains the behaviour of the expected $(\text{bias})^2$ term in chapters 6 & 7. This achieved much larger negative values for small values of p, because of the larger values which were achieved by $(S_{***k}^{-1} - S_{***p}^{-1})$ when p was much less than k.

8.3 Comparison of models

Which of the models considered so far is appropriate for epidemiological studies, in general, and for the lead study in particular? We have seen above that the two formulations which we have proposed for a model where the covariates, other than X*, are random lead to exactly the same expressions for the MSE criterion and for its constituent parts. The same is true for the two fixed-effect models. Thus the major decision is whether these other covariates should be treated as fixed or random. Clearly, once the sample has been selected, the inferences which we are interested in are those which are conditional on the observed Xs. However, if we attempt to model the sampling situation (say by simulation) the use of this model would suggest that we could repeat the study many times over with exactly the same choice of X. This is unrealistic. If we could do this, the Xs would be under experimental control and we could ensure that no confounding took place.

I have no doubt that a proper evaluation of the MSE criteria and their uses should have a random X model as its basis. The

disadvantage such a model is that the lack of normality for some of real covariates makes it difficult to produce simulated data which will have similar properties to experimental data. By contrast, it is easy to simulate the fixed effects model by taking the values of X as fixed and generating fitted values of Y from a regression equation. The difficulty with the fixed-effects model simulations is that we cannot be sure that their properties are generalisable to other sets of Xs, rather than being due to the particular pattern of X variables which happens to have been generated for this study. On the other hand, a procedure which appeared to perform very badly for a fixed-effects model would seem unlikely to do well when judged against the more difficult test of a random-effects model.

In the chapter which follows I will start by evaluating various procedures by simulations on the fixed-effects model. The more promising of these will then be evaluated for a random-effects model in the following chapter.

A secondary decision, which has to be taken in setting up these simulations, is whether to model X* as a random or as a fixed-effect. Following the argument above, a random model would seem more sensible because we could not select children on the basis of their blood lead. Thus the random-effect simulations will also consider X* as a random variable. For the model with fixed Xs, however, the X* will be taken as fixed also.

Chapter 9

Simulations for a fixed-effects model

9.1 Properties of the simulated data

To simulate a fixed-effects model, the values of log blood lead and the 33 covariates were taken as fixed. The blocking factor, "school", was ignored in these simulations. The estimated values (b* and b) of the coefficients for the full model fitted to the real data, ignoring the blocking factor "school", were taken to correspond to the population values in the simulation. Thus, in this chapter, the parameter values β and β^* have known values. For each simulation, values for the outcome variable (BASC) were generated by adding a random normal residual, with variance equal to the estimated residual variance from the full model, to each of the 501 fitted values of the vector $X\beta$, The value of σ^2 is thus also a known quantity.

Because we know the parameter values for this model we can calculate the true value of the MSE of estimation of b*, as it is estimated for any sub-model. I will use the expression "true MSE" to refer to this quantity. From expression 4.1 the true MSE is just the first element of

$$\sigma^2(P'P)^{-1} + (P'P)^{-1}P'Q\,\beta_Q\beta_Q'Q'P(P'P)^{-1}$$

For the full model, the last term drops out and this expression simply reduces to the estimated variance of β^* from the regression, using the original data. For any sub-model, we showed in chapter 4 that the difference between the full model estimate of β^* and its sub-model estimate b_P^* is just the first element of $(P'P)^{-1}P'Qb_Q$, where b_Q is the estimate of the omitted βs from the full model. Thus the second term in the true MSE is just the square of the difference between the full model and sub-model estimates of β^*, again using the original data. The first term is also easily evaluated for sub models, and has as its known value the expression which we denoted by V_{FP} in chapter 4. This is the estimate of the variance of b^*_P from the sub-model, with the residual variance from the sub-model replaced by the full model residual variance.

The values of the true MSE are given in table 8.1 for the sub-models chosen in a forward stepwise procedure; see chapter 6. Only 5 of the 31 models which are considered between the model with lead only, and the full model have values of the true MSE which are greater than the value of 1.6848 for the full model. Because the variance term is strictly increasing with p, the models with the lowest true MSE occur for small p; here the best one is for only one additional covariate. However, bad values can also occur for small p, because of large values of the bias term.

Table 9.1: True MSE values for simulated data

Covariates	Bias term	Variance term	Total
None	5.1589	1.4323	6.5912
+PVOC	0.0000	1.4559	1.4559
+CHILDINT	0.0208	1.4564	1.4772
+PMAT	0.1070	1.4644	1.5714
+AGEINT	0.6316	1.4780	2.1096
+FQUALIF	0.2795	1.4834	1.7629
+GESTAT	0.0939	1.4879	1.5818
+PARSCHL	0.0071	1.4965	1.5036
+STHEIGHT	0.2440	1.5514	1.7954
+SEX	0.1021	1.5584	1.6605
+OFFSCHL	0.0300	1.5656	1.5956
+WORKMUM	0.0400	1.5659	1.6059
+CARPHONE	0.0416	1.5660	1.6076
+HANDED	0.0265	1.5667	1.5932
+MQUALIF	0.0830	1.5735	1.6565
+CONSUMER	0.0453	1.5775	1.6228
+PARPART	0.0564	1.5780	1.6344
+FSOC	0.1150	1.5895	1.7045
+FAMSIZE	0.1402	1.5908	1.7310
+CLASSYR	0.0639	1.6072	1.6711
+UNEMPLOY	0.0451	1.6088	1.6539
+PARMENT	0.0450	1.6088	1.6583
+PARHLTH	0.0401	1.6092	1.6493
+MEDHIST	0.0139	1.6334	1.6473
+PARCHCOM	0.0153	1.6336	1.6489
+BRTHWT	0.0150	1.6336	1.6486
+TIMEDAY	0.0053	1.6451	1.6504
+FAMHIST	0.0027	1.6486	1.6513
+BRTHSCO	0.0008	1.6536	1.6544
+TOTCIGSD	0.0000	1.6587	1.6587
+BIRTHORD	0.0001	1.6683	1.6684
+MOVESCHL	0.0001	1.6805	1.6806
+MSOC	0.0000	1.6824	1.6824
full model	0.0	1.6848	1.6848

If inferences about β^* were made from the best model found here, that including PVOC only, the improvement in true MSE would be 1.6848/1.4559, a 16% improvement over inferences based on the full model. To be sure of going beyond the area where a very poor model would be selected it would be necessary to chose a model with a larger p, and modest improvements in efficiency of the order of 10% or even of 5% might be a more realistic goal.

Of course, only one subset of each size has been examined here. To go beyond this, I have looked at the values for the true MSE for all subsets of sizes 1, 2, 3, 4, and 32, 31, 30, 29 from the 33 possible covariates. The results are summarised in table 9.2. They give the percentage of all subsets which have a lower MSE than the full model, as well as the percentages which give improvements of at least 5% and at least 10% over the full model (true MSE less than MSE_{full} divided by 1.05 or 1.10).

Table 9.2: True value of MSE for all subsets.

p-2	No of subsets	No(%) with MSE < 1.6848	No(%) with MSE < 1.6046	No(%) with MSE < 1.5316	MSE best subset	MSE worst subset
1	33	4 (12%)	3 (9%)	1 (3%)	1.4559	9.8819
2	528	108 (20%)	88 (17%)	44 (8%)	1.4559	11.8254
3	5456	1448 (27%)	1156 (21%)	680 (12%)	1.4561	13.1453
4	40920	12633 (31%)	10021 (24%)	6167 (15%)	1.4546	14.3181
29	40920	17058 (42%)	364 (1%)	0 (0%)	1.5623	3.6344
30	5456	2387 (44%)	22 (½%)	0 (0%)	1.5726	3.1460
31	528	246 (47%)	0 (0%)	0 (0%)	1.6072	2.6519
32	33	18 (55%)	0 (0%)	0 (0%)	1.6667	1.9959

The model selected by the stepwise procedure in table 9.1 happened (fortuitously, I think) to, be the model with one covariate which gives the lowest MSE. Enumeration of all subsets, however,

gives a subset of size 4 which has a lower MSE. The larger models have a higher percentage of the subsets showing some improvement, but to a much lesser extent than the smaller models. The best subset with one covariate omitted gives an improvement of only 1.1% in the ratio of MSEs. I have not proceeded to look at subsets of intermediate size. A full enumeration of these would take vast amounts of computing, and a sampling procedure would be complicated.

9.2 Simulation results: initial test of 50 data sets

The same simulated data were used to test all the procedures described in chapters 6 and 7. Initially 50 sets of data were generated, as described above, and the various stepwise procedures were applied to each of them. The estimate of the variance of β^* for the full model for these simulated data will, of course, be different from the known value of this variance. The object of the simulations is to compare the results obtained after the selection procedures with one another and with those from the full model. To do this correctly the paired nature of the data from the different simulations must be taken into account. An initial test of the simulated data was to compare the results from the full model with those from the best model with only one covariate (PVOC only). For this case we know what the correct answer should be. The results of comparing the full model to that with PVOC only are described here, and exemplify the way in which calculations will be done in the next section.

In order to estimate the MSE we need to estimate the variance and the squared bias of the b* which are obtained from the regression/selection procedure. I will consider the squared bias term first. Although the true value of the coefficient is known, we can obtain a better estimate of the bias by comparing the values of b* from the sub-model, with the estimates of b* for the full model from the same simulated data. Using run 2 of the simulated data (table 9.3) as an illustration, we get the following results. The mean b* for the model with PVOC only is -3.1661, compared to the known true value of β^* which is -3.18. This gives an estimated bias of +0.014 with s.e. 0.166. Now, using the differences between the estimates for the full and reduced model, we obtain an estimated bias of +0.047 with a standard error of 0.077, which is less than half the standard error of the estimate which uses only the data from the reduced model. This reduction in variance comes about because of the high correlation between the estimates from the full and reduced models calculated for the same data; in this case the value of the correlation (r) is 0.9105. Thus we will always use the estimate of the bias from the paired data.

Table 9.3: Simulation results (3 sets each with n = 50)

model	mean(b*)	variance(b*)	r	est. MSE (95% conf int)	
full					
run 1	-3.3464	1.8891	*	1.6848	* *
run 2	-3.2134	1.7420	*	1.6848	* *
run 3	-3.0981	1.6556	*	1.6848	* *
PVOC only					
run 1	-3.2920	1.5874	.9468	1.41	(1.24, 1.61)
run 2	-3.1661	1.3812	.9105	1.34	(1.04, 1.78)
run 3	-3.0850	1.6534	.9335	1.64	(1.45, 1.95)

To estimate of the squared bias, we must square the estimated bias and subtract the square of its standard error. Here this procedure would lead to a negative quantity, so we estimate the squared bias at zero, but for the run 1 data it was just positive and has been added into the estimate of MSE. For this particular model we know that the true value of the bias is almost zero.

A similar procedure is used to estimate the ratio of the variances of two estimation procedures, and hence to obtain an estimate of the variance for the reduced model. The ratio of the two variances for run 2 is 0.7929. Now if we assume, as appears reasonable from plots of the results, that the joint distribution of the estimates of β^* is bivariate normal we can use a result for the ratio of multivariate normal variances due to Pitman (1939).

If the ratio of the two variances is denoted by L, the number of observations by m, the correlation of the two sets of estimates by r, and the quantity K is

$$K = 1 + \{2(1-r^2)/(m-2)\}\ t^2_{\alpha,\, m-2},$$

where $t_{\alpha,\, m-2}$ is the two-sided α point of the t distribution with m-2 degrees of freedom, then a $(1-\alpha)$ confidence interval for the variance ratio is

$$\left(1\{ K - (K^2-1)^{1/2}\} ,\ 1\{ K + (K^2-1)^{1/2}\} \right).$$

For the run 2 data the value of K, for $\alpha = 0.05$, becomes 1.0287 and thus a 95% confidence interval for the variance ratio becomes (0.62, 1.06).

A wider confidence interval is obtained by using the simulation results for the reduced model only. The ratio of the estimate of the variance from the sub-model to its known value for the full model is 1.34/1.6848 = 0.795 with a confidence interval (0.57, 1.11), obtained in the usual way from the percentage points of χ^2_{49}. In this case, the true value of 0.86 lies well within both confidence intervals. Because of its narrower interval we will always use the estimate of the variance ratio from the paired data. For full optimality it must be possible to combine the information from both sources by some method such as maximum likelihood. However, in the estimators which follow the correlations are larger than in this example, so we will be ignoring very little information by using the paired-data estimate of the mean and the variance ratio.

An estimate of the variance for the reduced model is obtained by multiplying the ratio and its confidence interval, from the paired data, by the known variance of 1.6848 for the full model. This gives 1.34 (1.04, 1.78) for a MSE whose true value we know to be 1.46. In the general case one would obtain the estimated MSE from this by adding in the estimated squared bias. However, this is estimated as zero here. Also, the variance of the estimate of the squared bias is so much less than the variance of the estimated

variance that it can be treated as a known quantity. This can be seen from a Taylor's approximation to the variance of the squared bias from the figures above, which gives $(2 \times 0.047)^2 \times (0.077)^2 = 0.000052$, which will make no noticeable contribution to the confidence interval for the MSE. This was also true for all other models considered in this chapter. Thus the confidence interval for the estimated MSE is simply obtained by adding the estimated squared bias to the confidence intervals for the estimated variance.

The results for three runs of the test data are given in table 9.3. If we were to use these results to evaluate the variance of the reduced model, a confidence interval based on only 50 observations would be unsatisfactorily wide. However, the same number of simulations gave satisfactory results for some of the other estimators below. The width of the confidence interval for the MSE depends crucially on the correlation r. Values of r close to 1 give narrow confidence intervals. For each estimation procedure an initial run of 50 simulations was done to determine r and hence to calculate the number of simulations required to give a 95% confidence interval with a ratio of 1.10 from one end to the other. A value of 1.001 for k is required for this, and so the number of simulations required is obtained as

$$m \simeq 2\, t^2\, (1 - r^2) \,/\, (0.001).$$

For PVOC only from the run 2 data this would be 1368. We will see below that this was larger than was required for any of the other estimation procedures evaluated here. These further simulations

were not carried out for PVOC only, because this model was only investigated as a test of the simulated data for a case where the answer is already known.

9.3 Simulation results for a forward stepwise procedure

The simulated data were first run through a stepwise regression procedure which selected, at each step, the variable which gave the greatest improvement in the residual mean square of Y. Two stopping rules were evaluated. The first was to stop the selection procedure when the F-to-enter statistic for the next variable no longer exceeds 4. This is equivalent to stopping when the next variable is no longer formally significant at the 5% level. The second rule continued the stepwise procedure until the F-to-enter statistic no longer exceeded 2, a procedure which will give a model close to that with the minimum value of C_p. Although it cannot be guaranteed, without doing a full search of the subsets, that the model chosen will be the very best in terms of C_p, it is very likely that the model from the stepwise procedure will be close to it; see chapter 6 for a discussion of this. For the first 50 simulations, stopping at an F of 4 included between 8 and 17 covariates, with a median of 12, whereas stopping at an F of 2 included from 12 to 21 covariates with a median of 16. The results for the MSE of b* are given in table 9.4.

Table 9.4: Simulation results for forward stepwise procedure minimising the residual sum of squares of Y (m = 50).

stopping rule	mean(b*)	variance(b*)	r	est (bias)^2	est. (95% conf int) MSE
full	-3.3464	1.8891	*	*	1.6848 * *
F=4	-3.1930	1.7203	.9683	.0212	1.555 1.406 1.721
			%of full model		92.3% (83.4%102.2%)
F=2	-3.2086	1.8734	.9849	.0179	1.6886 1.574 1.811
			%of full model		100.2% (93.4%107.5%)

The wide confidence intervals for the MSE prevent any firm conclusion being drawn. The required sample sizes, calculated as suggested above, are of the order of 500 and 250. Thus a further 450 simulations were run and the combined results of these, along with the initial 50, are given in table 9.5.

Table 9.5: Simulation results for forward stepwise procedure minimising the residual sum of squares of Y (m =500).

stopping rule	mean(b*)	variance(b*)	r	est (bias)^2	est. (95% conf int) MSE
full	-3.2023	1.6631	*	*	1.6848 * *
F=4	-3.0617	1.6709	.9652	.0196	1.7124 1.6348 1.7936
			%of full model		101.6% (97.0%106.5%)
F=2	-3.0941	1.6734	.9841	.0120	1.707(1.654, 1.762)
			%of full model		101.3% (98.2%104.6%)

These further results show, that for these data, the estimates based on the reduced models from minimising the residual sum-of-squares confer little advantage in terms of the MSE of b*. In

fact, it is possible that all the computations involved in stepwise or search procedures may even increase the MSE of b*, compared to its full model estimate.

A further problem is that, for both of these stopping rules, the estimated variance from the final model, calculated as though it were the correct one, is an underestimate of its true value. The mean value of the estimated final model variances from the two stopping rules are 92.1% and 92.5% of the full model value (estimated here with a standard error of <0.5%). The value of G_{FP}, calculated for the final model gives similar results

9.4 Simulations for forward stepwise procedures based on minimising the G_P criteria

The same set of 50 sets of simulated data were used and were analysed by stepwise procedures which selected, at each step, the variable which gave the lowest value of the criterion being tested. For these initial 50 simulations the stepwise procedure was continued until all variables were included, and then the estimate of β^* was computed from the model chosen at the step which gave the lowest value of the G_P criterion. Results for the MSE of b* are in table 9.6.

There were two fixed effects model criteria. One was G_{FP}, as defined in chapter 4, and the other, G_{FP}' was derived from it by setting the term for the expected bias to zero whenever its estimate was negative. Similarly, the two random effects models were G_{RP} and

G_{RP}' where the latter was derived from the former by setting the estimated bias term and the term V_{2R}, to zero if they had negative estimates.

The G_{FP} criterion selected models with anything between 1 and 7 covariates, but there was no pattern as to which ones were chosen. It can be seen that this procedure selected a model which gave a value almost identical to the value of b* for the full model calculated from the simulated data - but not, of course, the true value of β^*. Because of the high value of r, we have a very tight confidence interval for the MSE of estimates derived after this procedure and we can see that it offers no advantage over the full model. Thus, the suspicions mentioned in chapter 7 were justified.

Table 9.5: Simulation results for forward stepwise procedures based on minimising the various G_P criteria (m = 50)

criterion	mean(b*)	variance(b*)	r	est (bias)^2	est. (95% conf int) MSE
full	-3.3464	1.8891	*	*	1.6848 * *
G_{FP}	-3.3432	1.8816	.9999	.0000	1.678(1.672, 1.684)
			%of full model		99.6% (99.2%100.0%)
G_{FP}'	-3.5912	1.9818	.9880	.0590	1.8266(1.718, 1.942)
			%of full model		108.4%(102.0%115.3%)
G_{RP}	-3.3367	1.8784	.9996	.0001	1.675(1.656, 1.695)
			%of full model		99.4% (98.3%100.6%)
G_{RP}'	-3.527	1.9243	.9865	.0317	1.748(1.637, 1.867)
			%of full model		103.7% (97.1%110.8%)

The G_{FP}' is performing significantly worse than the full model. It selected models with very few covariates. All but 4 of the 50 simulations gave a model with just one additional covariate. However it selects models which we know have a small value of the true MSE of b*. The most commonly selected was FQUALIF (28 times) which has gives a model with MSE of 1.554, followed by PVOC (11 times) which has the lowest MSE of 1.456.

How can a procedure which selects good models give an estimate of b* worse than the full model? It is wrong to think that a set of models selected from a stepwise algorithm will behave like a mixture of the corresponding models, selected at random. There will be a correlation between the model which happens to be selected and the value obtained for the estimate of b* from the same data. It must be some mechanism such as this which accounts for these results. I will call this increase in variance, over what would be expected from a mixture of the corresponding models, "selection variance" since it is the additional variation introduced by the variable selection process. The mechanism by which it had operated here became clear when the detailed results of the simulations were examined. The models which selected FQUALIF alone had a mean bias which was considerably greater than the true value of the bias for this model when it is selected at random (which is a known quantity here). The results in chapter 6 showed that this procedure selected values which had very low values of their correlation with blood lead. Thus the variable FQUALIF will be selected when its correlation with blood lead is low, and on these occasions it will

move the estimate of β^* a shorter distance from the null model to the full model than it would on the other occasions.

Like G_{FP}, G_{RP} selected models which gave values almost identical to the estimate for the full model, although these were different models with a larger number of covariates included (ranging from 8 to 16 additional covariates). The results, although with a slightly wider confidence interval, are equally useless.

The G_{RP}' criterion produces results which are less strongly correlated with the full model, and hence the MSE has a slightly wider confidence interval. It selected models with between 9 and 20 covariates. Although the results did not appear promising, a further 200 simulations were run to give a better estimate of the MSE. On this occasion the stepwise procedure was stopped at the first local minimum of the criteria. For the first 50 simulations this procedure would give the minimum value on every occasion. The results for the total of 250 simulations are given in table 9.6. They confirm the suggestion in the first 50 simulations, that the sub-model estimates are just a little worse than the full model estimates. Again we must suspect that some kind of selection variance is being introduced, although on a lesser scale than for G_{FP}'.

Table 9.6: Simulation results for forward stepwise procedures based on minimising G_{RP}' criteria (m = 250)

criterion	mean(b*)	variance(b*)	r	est (bias)^2	est. (95% conf int) MSE
full	-3.1854	891	*	*	1.6848 * *
G_{RP}'	-3.3889	43	.9885	.0413	1.737(1.692, 1.784)
			%of full model		103.1%(100.4%105.9%)

Thus we can conclude that forward stepwise selection using any of the G_P criteria will give estimates, for this set of simulated data, which are no better, and can even be worse, than using the full model, and selection variance may be an explanation for this.

9.5 Simulations for backward stepwise procedures based on minimising the G_P criteria.

The same 50 sets of simulated data were used to test the backwards stepwise procedures driven by G_P. Results are in table 9.7.

Table 9.7: Simulation results for backwards stepwise procedures based on minimising the various G_P criteria (m = 50)

criterion	mean(b*)	variance(b*)	r	est (bias)^2	est. (95% conf int) MSE
full	-3.3464	1.8891	*	*	1.6848 * *
G_{FP}	-3.3965	1.8971	.9936	.0020	1.694(1.618, 1.774)
			%of full model		100.5% (96.0%105.3%)
G_{FP}'	-3.7599	1.9157	.9964	.1707	1.879(1.821, 1.340)
			%of full model		111.5%(108.1%115.1%)

(continued on next page)

Table 9.7 continued

criterion	mean(b*)	variance(b*)	r	est (bias)^2	est. (95% conf int) MSE
G_{RP}	-3.3276	1.8716	.9994	.0003	1.669(1.645, 1.693)
			%of full model		99.1% (97.6%100.5%)
G_{RP}'	-3.3464	1.8655	.9839	.0069	1.680 (1.562, 1.807)
			%of full model		99.7.0% (92.7%107.3%)

The results for G_{FP}', G_{RP} and G_{RP}' are very similar to those for forward selection. The models selected by G_{FP} give estimates of β^* which are less highly correlated with the full model than was the case for the forward selection procedure. When the backwards algorithm for G_{FP} was studied in chapter 6, the criterion was reduced as covariates were dropped from the model, but at the last few steps it tended to oscillate because those variables which caused it to increase had to be included. Similar patterns were found for the simulated data, and it could be seen that the final model selected was the one from the last few steps which came closest to the full model. Sometimes none of the last few steps gave a model which was very close. In contrast, the forward procedure was able to find a model with the same value as the full model more easily. This explains the difference between the forward and backward results for this criterion. The same difference was not evident for the other criteria because they selected models with more covariates.

None of these procedures look very promising as methods of selecting sub-models. As for the forward procedures, the only one which appears possible is G_{RP}', and a further 150 simulations were run for this model. Results for the complete set of 200 are given in table 9.8. They are very similar to those for the forward procedure with the same criterion, and are no more encouraging.

Table 9.8: Simulation results for backward stepwise procedures based on minimising G_{RP}' criteria (m = 200)

criterion	mean(b*)	variance(b*)	r	est (bias)^2	est. (95% conf int) MSE
full	-3.1841	1.7647	*	*	1.6848 * *
G_{RP}'	-3.2994	1.8277	.9844	0.0130	1.7580(1.697, 1.8209)
			%of full	model	104.3%(100.7%, 108.1%)

9.6 Selecting models which change the estimate of β*

The same set of simulated data was used to evaluate the procedure of selecting covariates to enter or remove from the model by their influence on the estimate of β*. A forward and backward procedure were each investigated. In the forward procedure, starting with the model which contained lead only, the covariate which entered the model at each step was the one which produced the greatest absolute change in b*. The backwards procedure started with the full model and, at each step, dropped the covariate which produced the smallest change in the estimate b*. Stopping criteria were defined for the squared change in b* in terms of the variance of b*. The forward procedure was stopped when the largest squared change in b* for the next step was less than C times the variance of b* for the full model. Similarly, the backward procedure was

stopped when no covariate could be excluded which would produce a change of less than this amount. Six stopping criteria were evaluated - C = 0.4, 0.2, 0.1, 0.05, 0.01 and 0.001. Results are given in table 9.9.

Table 9.9: Simulation results for stepwise procedures driven by the absolute change in b* (m = 50)

Stopping rule (C)	mean(b*)	variance(b*)	r	est (bias)^2	est. (95% conf int) MSE
full	-3.3464	1.8891	*	*	1.6848 * *
forward selection					
0.4	-2.5341	1.6870	.9522	.6563	2.161(1.983, 2.363)
			%of full model		128.3%(117.7%140.3%)
0.2	-2.8868	1.7042	.9651	.2086	1.729(1.573, 1.902)
					102.6% (93.4%112.9%)
0.1	-3.1186	1.6845	.9709	.0497	1.5521(1.411, 1.707)
					92.1%(83.8%, 101.3%)
0.05	-3.2111	1.5639	.9726	.0161	1.4109(1.284, 1.551)
					83.7%(76.2%, 92.0%)
0.01	-3.1164	1.7047	.9858	.0518	1.577(1.471, 1.680)
					93.3% (87.3% 99.7%)
0.001	-3.2924	1.8584	.9984	.0028	1.660(1.623, 1.699)
					98.5% (96.3%100.8%)
backward elimination					
0.001	-3.3117	1.8551	.9984	.0011	1.656(1.618, 1.694)
					98.3% (96.0%100.6%)
0.01	-3.1671	1.6671	.9885	.0312	1.518(1.429, 1.613)
					90.1%(84.8%, 95.7%)
0.05	-3.177	1.5651	.9705	.0265	1.422(1.292, 1.567)
					84.4%(76.7%, 93.0%)
0.1	-3.095	1.6876	.9703	.0611	1.724
					93.0% (84.5%102.3%)
0.2	-2.825	1.6588	.9636	.2691	1.749(1.594, 1.921)
					103.8%(94.6%, 114.0%)
0.4	-2.552	1.5753	.9350	.6259	2.031(1.839, 2.253)
					120.5%(109.1%, 133.7%)

This begins to look a bit more hopeful. Backward and forward procedures give similar results. The larger values of C (0.4, 0.2) give biassed results with a larger MSE, but stopping at values of C of 0.1, 0.05 and 0.01 seems to be an improvement over the full model. However, these results would estimate that at a value of C=0.05 the efficiency of this procedure has become better than would be possible for any sub-model. This suggests that this set of simulations may be over-estimating the benefits of this procedure, and the post-hoc choice of a value for C may have helped to exaggerate its advantages. A further set of 100 simulations were run, and their results (not including the first 50) are given in table 9.10. Only the forward stepwise results are shown, as forward and backwards procedures gave very similar results.

Table 9.10: Simulation results for stepwise procedures driven by the absolute change in b* - second set of simulations - (m = 100).

Stopping rule (C)	mean(b*)	variance(b*)	r	est (bias)^2	est. (95% conf int) MSE
full	-3.0460	1.9331	*	*	1.6848 * *
forward selection					
0.4	-2.2799	1.7898	.9058	.5808	2.141(1.894, 2.430)
			%of full model		128.3%(112.4%144.2%)
0.2	-2.6932	1.8202	.9428	.1215	1.708(1.484, 1.907)
					102.6% (88.1%113.2%)
0.1	-2.8840	1.9034	.9378	.0221	1.681(1.463, 1.932)
					99.7%(86.8%, 114.6%)
0.05	-3.0223	1.8500	.9565	.0000	1.612(1.432, 1.815)
					95.7%(85.0%, 107.7%)
0.01	-2.9009	1.9250	.9804	.0208	1.698(1.552, 1.818)
					100.8% (92.0%107.9%)
0.001	-3.0372	1.9464	.9980	.0000	1.696(1.653, 1.740)
					100.6% (98.1%103.3%)

This second set of results is less favourable to the selection procedure. There are similarities in that the first two values of C give biassed estimators which are worse than the full model, and the final value of C is too close to the full model to give much improvement. However the apparent benefit of the three middle values is much less evident. To resolve this, a final series of 100 simulations were run, this time evaluating only the values 0.1, 0.05 and 0.01 for C. Results are in table 9.11.

Table 9.11: Simulation results for stepwise procedures driven by the absolute change in b* - third set of simulations - (m = 100)

Stopping rule (C)	mean(b*)	variance(b*)	r	est (bias)^2	est. (95% conf int) MSE
full	-2.9283	1.6712	*	*	1.6848 * *
forward selection					
0.1	-2.7031	1.7116	.9526	.0491	1.774(1.629, 1.933) 105.3%(96.5%, 114.8%)
0.05	-2.7962	1.6367	.9625	.0162	1.666(1.542, 1.800) 98.9%(91.5%, 106.9%)
0.01	-2.7445	1.6712	.9822	.0332	1.718(1.630, 1.811) 102.0% (96.7%107.5%)

These results are more in line with the second set of 100 than with the first set of 50. Table 9.12 gives the combined estimate of MSE relative to the full model for all 250 simulations.

Table 9.12: Simulation results for forward stepwise procedures driven by the absolute change in b* - combined rusults - (m = 250)

Stopping rule (C)	est. MSE relative to full model (95% conf int)		
0.1	100.4%	(93.5%,	107.0%)
0.05	94.6%	(88.9%,	100.3%)
0.01	99.8%	(95.8%,	103.8%)

The combined results suggest that a stopping rule of 0.05 may be giving a slight advantage, but it is not great. This may have come about because of the particular covariates which these data select at this value of the stopping criterion, rather than any particular virtue of the value 0.05 for C in the more general case.

9.7 Conclusions

The results in this chapter have shown that none of the methods proposed for selecting sub-models could be recommended to reduce the MSE of estimation for β^*, for data with structure like that of the Edinburgh Lead Study. Some methods produce increased MSEs compared with the full model ($G_{F:F}$'), whereas others (t statistics, C_p) may tend to underestimate the variance of the regression coefficient. Several of the methods give results which are so highly correlated with the full model values as to be indistinguishable from them.

These results may be true for all data sets, or else they may reflect the fact that the potential for improved estimation from sub-models in the Lead Study data was not very great. There can be

three possible reasons why sub-models will not produce good estimates in terms of the MSE of β^* :

(1) The sub-model estimates all have a large bias.

(2) The covariates are not strongly dependent on X*

(3) The sample size is too large.

The first reason does not apply to the Lead Study data, because we have seen that there are many sub-models with negligible bias. However, both the other two conditions apply, as we saw in chapters 5, 6 and 7.

The squared multiple correlation between X* and the other covariates can be used to assess the potential for reducing the variance of b* in sub-models. For a fixed-effect model, the maximum possible reduction in variance for a sub-model, compared to the full model, is the ratio S_{***}^2/S_{**}^2. This is just $(1-R^{*2})$ where R^{*2} is the multiple correlation of X* with all the other covariates. Here this multiple correlation is only 0.15, so the variance can reduce by a factor of, at most, 0.85.

We saw in chapters 4 and 8 that the variance of b* for the reduced model can be expressed as a sum of two parts, V_1 which depends only on the increase in the residual degrees of freedom for the reduced model, and V_2 whose expected value increases with the population parameter corresponding to the multiple correlation R^{*2}. For the Lead Study data the minimum value of V_1, for the model with no covariates except blood lead, is just (501-35-1)/(501-3) = 0.93

times the equivalent quantity for the full model. The difference between this factor and the overall factor of 0.85 is a measure of the decrease in variance in sub-models which arises because of the genuine correlations between X* and the other covariates.

This result is closely related to the quantity "adjusted R^{*2}" which is often computed for multiple regressions (eg in GENSTAT output). The value for adjusted R^{*2} for the relationship between X* and the other covariates is

$$R^{*2}_{adj} = 1 - (1 - R^{*2})(n-1)/(n-p+1).$$

Its advantage over the unadjusted R^{*2} is that it gives a consistent estimate of the corresponding population parameter; however, like the quantity V_2, it can take negative values.

The value of R^{*2} can be used to assess the potential for improved estimation from sub-models, with a high value indicating that sub-models may be better than the full model. If a low value of R^{*2}_{adj} is obtained from a higher value of the unadjusted R^{*2}, then the improved estimation from sub-models is a consequence of the increase in residual degrees of freedom. If both the adjusted and unadjusted R^{*2} are high, a strong association between X* and the covariates would appear to be the reason for the improved estimation from sub-models. For the Lead Study data we have $R^{*2} = 0.15$ and $R^{*2}_{adj} = 0.09$, so there is very little overall potential for variance reduction because both of the conditions (2) and (3) apply.

However, if the sample size were reduced the performance of sub-models relative to the full model would be enhanced. If the Lead Study had included 200, 100, 75, or 50 children then the minimum value of V_1 would have been .83,. 66, .54 or .30 times its value for the full model.

These results are all derived for the fixed-effects model, which assumes that the variances of all the estimates of b* are expressions which contain the residual variance from the full model. For a random-effects model similar results apply, but for sub-models the expression for the variance contains the residual variance from the sub-models.

The effects of reducing the sample size are investigated in the next chapter, by selecting sub-samples from the set of 501 points. Only three of the procedures discussed in this chapter are evaluated on these reduced samples. These are the methods which gave values different from the full model, and which did not show any undesirable features in this chapter. The methods are selection by the residual sum-of-squares, by G_{RP}' and by the absolute change in the b* estimate. By selecting sub-models from the 501 cases we can also go at least some way towards having a full random effects model for the other covariates, and we can evaluate some of the theory developed in chapter 4.

Chapter 10

Random sub-samples from the X variables; simulation results

10.1 Introduction

The Lead Study data were also used as the basis for the simulation experiments reported in this chapter. Random sub-samples, with replacement, were taken from the 501 rows of the X matrix. Sub-samples of 200, 100, 75 and 50 were investigated. For each sub-sample, values of BASC were generated in the same manner as in the last chapter, assuming that the true conditional relationship between BASC and the covariates (including X*) was the same as that estimated from the real data for the full model. Because we are generating a different set of covariates for each simulation, we now have a random-effects model. The expected value of the estimate of β^* from a sub-model, averaged over all selections from X, will be the same as in the model in the previous chapter. However, for any given selection of X and X*, the expected value of the conditional estimate will differ from this value. It is the variance of these differences, across different selections of X variables, that accounts for the difference in the MSE criteria GFp and GRp between fixed- and random-effect models. Because sampling from the X matrix is with replacement, we can derive the properties of estimates from reduced models (assuming model selection without reference to the data) in a straightforward way.

This formulation is more likely to be applicable to real epidemiological data than is the fixed-effects model in the last chapter. However, the properties of the estimators from the two models are linked, as we will see below.

10.2 Properties of the simulated data

As in the last chapter, I will start by evaluating the estimates of β^* for the simulated data, with the model which contains only log blood lead and the parent's vocabulary score (PVOC).

If the sample contains n_{sub} members, the the expected value of the sums of squares and products of X and X* about their means will be $(n_{sub}-1)/(500)$ times the corresponding full model values. Similarly, the term S_{**p} (chapter 8), which is the residual sum-of-squares from the regression of X* with the (p-2) covariates, will have an expected value for the sub-sample of $(n_{sub}-(p-2)-1)/(501-(p-2)-1) = (n_{sub}-p+1)/(502-p)$ times the equivalent quantity calculated for the same sub-model from all the data. We must remember that p is the total number of covariates including a constant and lead. Thus to first order in $(n_{sub}-p+1)$ the term S^{-1}_{**p} will have an expected value for a sub-sample of size n_{sub} which is $(502-p)/(n_{sub}-p+1)$ times the fixed quantity S^{-1}_{***} for the full model.

For a particular choice of X and X* in a sub-sample we can calculate the bias in the estimate (b^*_p) of β^* for the sub-model from the expression $(P'P)^{-1}\ P'\beta_Q$. The matrices P and Q are the

matrices of selected and unselected covariates with n_{sub} rows, and the coefficients β_Q are fixed for the simulation. This allows us to calculate the true conditional regression coefficient for any sub-model at each simulation (b^*_{true}). The variance of the value of b* estimated from the sub-model about b^*_{true} will be the variance term from G_{Fp}, (V_{Fp} = $SS_{**p}{}^2$ σ^2_{full}). We have fixed σ^2_{full} in the simulations, and we can calculate $SS_{**p}{}^2$ in terms of the X values for the covariates in the sub-model, so we can calculate V_{Fp}, at each simulation, for any sub-model.

Now the value of b^*_{true} will vary from one simulation to the next in such a way that the total variance of the residuals from the true regression line will, on average, be σ^2_p. The quantity σ^2_p is a mixture of the random normal deviate added to the regression, and the variance introduced by the sub-sampling of the X matrix. It's value is just the residual-mean-square from the reduced model for the real data. Thus the total variance of b^*_p will be the variance term from G_{Rp}, (V_{Rp} = $SS_{**p}{}^2$ σ^2p). Again we can compute this for any sub-model, and sub-samples of any size.

Thus we can calculate the terms V_{Rp} and V_{Fp} for any sub-model and choice of sub-samples of the 501 observations, in terms of the quantity $S_{**p}{}^2$ calculated for the sub-sample. Since we have an expression for the sub-sample expectation of $SS_{**p}{}^2$ in terms of the same quantity for the full data, we can calculate expectations for these quantities (to order $1/(n_{sub} - p + 1)$) for any sub-model The expectations of these variances for different values of n_{sub}, for the model which contains PVOC only, are given in table 10.1. We

know that the bias of this model with respect to the true value of β^* is practically zero.

Table 10.1 Theoretical values for the variances of b^*_p (values from 500 simulations given in brackets).

	n_{sub}	$n_{sub}-p+1$	V_{Fp}	V_{Rp}
FULL MODEL				
(b^*_k)	501	467	1.685	1.685
	200	166	4.777 (4.795)	4.777
	100	66	12.240 (13.182)	12.240
	75	41	20.028 (23.430)	20.028
	50	16	55.945 (61.088)	55.945
PVOC ONLY				
(b^*_p)	501	499	1.456	1.794
	200	198	3.654 (3.378)	4.534 (4.640)
	100	98	7.475 (7.748)	9.208 (9.315)
	75	73	10.070 (10.553)	12.405 (12.943)
	50	48	15.747 (16.92)	20.725 (21.120)

For the complete data set, b^*_p for the model with PVOC only no longer has a lower variance than b^*_k, the estimate from full model, when judged as a random-effects model. For a sample size of 200 the variances of b^*_p and b^*_k are comparable, but for smaller sample sizes the variance of b^*_p is smaller than b^*_k. The benefit of the reduced model is less for the random-effects model than it

would be for the corresponding fixed-effects model. However, we noticed in chapters 6 and 7 that lower values of the random-effects criteria were obtained for models with more covariates, for which the residual-mean-square was considerably reduced compared with the single variable model considered here. For example, if we consider some of the models selected in chapter 7 (table 7.9) in procedures which selected by low values of the G_{RP} procedure, we can find a model with p=10 which gives a derived value of V_{RP} of 3.88 and a very small bias for samples of size 200, which is a MSE of 81% of the full model variance.

As a test of the simulated data, and the above approximations, 500 sets of simulated data were constructed at each of these sample sizes and the estimates of β^* for the full and reduced model were calculated for each one. For each set of simulated data the value of b^*_{true} was calculated from the selected rows of the X matrix. This allowed estimates of the term V_{FP}, as the variance of the differences between b^*_P and b^*_{true}. The results are also given in table 10.1. Each variance estimate has a standard error of +/- 6% due to sampling error from the set of 500. It can be seen that the results are in very good agreement with the derived values. There is a possible exception for small values of $(n_{sub}-p+1)$ when the derived values tend to give a under-estimates of the variance of the estimators.

By calculating the value b*true we can consider the variation of the estimator about its mean value in two parts

$\Delta_r = b^*_p - b^*_{true}$ and $\Delta_x = b^*_{true} - \beta^*$

which correspond to the variation conditional on fixed X, and the variation due to different choices of X. For the simulations considered here these two components appeared to be independent, with the largest correlation between them being -0.067. However, we might not expect this to hold for models where the covariates are selected with reference to the data.

A practical problem which occurred in these simulations was that some sub-samples had one or two covariates which had the same value for all members of the sample. The covariate UNEMPLOY was the one for which this happened most frequently. There were 125/500 such sets for $n_{sub} = 50$, 33/500 for $n_{sub} = 75$ and 6/500 for $n_{sub} = 100$. It was, of course, impossible to fit all the covariates to such data. The covariates which had no variation were simply excluded from the model. Less commonly, two covariates from the sub-sample had exactly the same values (or values which differed by a constant) for all members of the sub-sample. In this case only one of the two could be included in the regression. These procedures are exactly what would be done if such data were encountered in practice, and do not present any theoretical problems. The main disadvantage of this feature of the data was the extra computing required to exclude the appropriate variables automatically from each simulation.

10.3 Comparison of the three selection procedures.

Further simulated data were used to investigate the properties of the three variable selection procedures which are still worthy of consideration. As in the last chapter runs with m=50 were tried initially, but the sub-model estimates were found to correlate less strongly with the full sample estimates for the smaller sub-sample sizes, and the simulations were thus continued to give a total number m=500. An exception was the set of simulations driven by G'_{RP}. These were the most time-consuming, because it was necessary to extract the variance estimate and calculate the rather complicated expression for G'_{RP} for every covariate being considered for inclusion in the model. Problems arose with computer runs which exceeded the maximum time permitted. Since the properties of the procedures became clear after 250 simulations, no further runs were done at this stage. Each of the three procedures used a different set of random sub-samples and of randomly generated residuals.

Results for the full model and sub-model estimates of β^* are in tables 10.2 to 10.4. The mean value, over all the simulations, of the estimated variance of b^*_p calculated for the selected sub-model (as if it were the correct one) is also tabulated. No backward selection procedures were considered. Selection by RMS and by G'_{RP} gave similar models for either the backwards or forwards procedures, conditional on the fixed Xs. It would be expected to perform similarly here too, so the considerable extra computation which would be required to simulate the backwards procedures was not considered worthwhile. A backwards procedure was not considered for

the selection by the maximum change in b^*, as it would have very different properties. Forward selection to minimise the RMS with stopping rules at F=4 and F=2 is considered first.

Table 10.2 : Simulation results for forward stepwise procedures by RMS, and F-to-enter of 2 or 4 (m=500).

n_{sub}	$var(b^*_k)$	$var(b^*_p)$	r	mean(se) $b^*_p - b^*_k$	MSE b^*_p	(MSE/$var(b^*_k)$) % (95% C.I)	estim. $var(b^*_p)$
F = 4							
200	5.12	5.30	.87	.02(.05)	5.30	103%(94%-112%)	3.94
100	13.40	10.52	.78	.10(.10)	10.52	79% (71%-88%)	7.92
75	23.65	15.32	.75	.02(.14)	15.32	65% (57%-73%)	10.23
50	69.31	28.27	.56	.50(.31)	28.42	41% (35%-48%)	15.03
F = 2							
200	5.12	5.14	.93	.04(.04)	5.14	100%(94%-107%)	3.93
100	13.40	10.94	.87	.19(.08)	10.97	82% (75%-89%)	7.67
75	23.65	17.85	.83	.23(.12)	17.88	76% (68%-84%)	9.82
50	69.31	35.53	.69	.47(.73)	35.53	51% (45%-58%)	13.65

For n_{sub} = 200, the results are like those for the full model, in that the sub-model estimates are no better than the estimates from models which include all the covariates. However, for n_{sub} = 100, 75 and 50 the selection procedures give estimates of β^*_p which have lower MSE than the full model estimates. The most extreme differences are seen for the lowest sample sizes. The stopping rule of F<4 (equivalent to stopping at a nominal p-value of 0.05) gave better results than F<2 (equivalent to C_p). The estimate of the variance of the estimator at the end of the selection procedure is

much lower than it should be. This is true even when the sub-model gives no improvement in estimation, but is even more marked for the small sample sizes which give improved estimates of b*.

Table 10.3 : Simulation results for forward stepwise procedures for a minimum value of G'_{RP} (m=250).

n_{sub}	$var(b^*_k)$	$var(b^*_p)$	r	mean(se) $b^*_p - b^*_k$	MSE b^*_p	(MSE/$var(b^*_k)$) % (95% C.I)	estim. $var(b^*_p)$
200	5.32	5.36	.95	-.10(.05)	5.37	101%(93%-109%)	3.84
100	13.73	10.88	.91	.06(.09)	10.88	79% (71%-88%)	7.62
75	20.51	14.22	.87	.01(.14)	14.22	69% (61%-78%)	10.04
50	70.38	36.89	.82	.09(.31)	36.89	52% (45%-60%)	14.60

The results for selection by G'_{RP} are very similar to those for the forward stepwise procedure driven by the residual-mean-square of the ability scores. They also share the same feature of under-estimating the variance of b^*_p, with a very similar pattern to the results in table 10.2. This is perhaps not too surprising when we look back at the detailed study in chapter 7. It was seen there that G'_{RP} was selecting variables on the basis of a low value of the residual-mean-square, from among the variables which did not introduce appreciable bias and did not have a strong relationship with blood lead. As these two conditions would not exclude many variables from the lead study data, the patterns are fairly similar. However, this might not be the case for other data sets where there are more sub-models which would give biassed estimates of β^*, and where the correlations between X* and the other covariates were stronger.

Table 10.4 : Simulation results for forward stepwise procedures by maximum absolute change in b^*_p (m=500).

n_{sub}	var(b^*_k)	var(b^*_p)	r	mean(se) $b^*_p - b^*_k$	MSE b^*_p	(MSE/var(b^*_k)) % (95% C.I)	estim. var(b^*_p)
C = 0.1							
200	4.67	4.60	.86	.15(.05)	4.61	99%(91%-108%)	4.78
100	13.96	11.09	.80	.14(.10)	11.10	79% (71%-88%)	10.00
75	22.11	15.38	.71	.19(.15)	15.39	70% (62%-80%)	14.12
50	69.75	25.43	.43	.45(.34)	25.51	36% (31%-42%)	22.32
C = 0.05							
200	4.67	4.74	.90	.14(.04)	4.76	102%(94%-110%)	4.71
100	13.96	11.41	.84	.12(.09)	11.42	82% (74%-90%)	9.86
75	22.11	17.24	.79	.23(.13)	17.28	78% (70%-87%)	14.01
50	69.75	29.93	.57	.44(.30)	30.03	43% (37%-50%)	22.17
C = 0.01							
200	4.67	4.63	.96	.08(.03)	4.64	99% (94%-104%)	4.68
100	13.96	12.74	.94	.09(.06)	12.74	91% (86%-97%)	10.79
75	22.11	20.53	.93	.03(.08)	20.53	93% (85%-97%)	15.22
50	69.75	46.99	.81	.47(.22)	47.16	68% (61%-75%)	26.86

Again there is benefit in estimating from the sub-model except when n_{sub} = 200. The best results are obtained when the stopping criterion C = 0.1 and 0.05. The value of 0.01 for C give worse results, more highly correlated with the full model value. The under-estimate of the variance of b^*_p is much less marked than for the previous two procedures. This is particularly the case for C=0.1 where the variance estimates are only about 10% lower than their estimates from the simulation.

The variance of the quantities Δ_r (the difference between the b^*_p and the true conditional estimate of β for the sub-model) and the correlation between Δ_r and Δ_x were calculated for all these simulations. Also, for each sub-model selected the true value of the conditional variance of b^*_p was computed. These quantities are tabulated in table 10.6, which also gives the range of values of p for the sub-models selected by each stepwise procedure. The column "true" contains the mean of the true values of the conditional variance for the sub-model for each set of simulations.

Table 10.5 : Components of the variance of b^*_p.

Selection procedure	n_{sub}	var(b^*_p)	var(Δ_r) (true)	var(Δ_x)	cor($\Delta_r\Delta_x$)	median and range of p	
RMS F= 4	200	5.30	4.28 (3.94)	0.80	.06	6	2-11
F= 2		5.14	4.45 (4.14)	0.60	.03	11	7-16
G'_{Rp}		5.36	4.47 (3.96)	0.83	.02	9	5-14
change in b							
C = 0.1		4.60	4.04 (4.11)	0.69	-.04	4	1-8
C = 0.05		4.74	4.10 (4.23)	0.53	.04	6	2-10
C = 0.01		4.63	4.40 (4.61)	0.26	-.01	11	7-17
RMS F=4	100	10.52	8.37 (8.12)	2.02	.02	4	1-13
F=2		10.94	9.11 (8.71)	1.63	.03	9	4-17
G'_{Rp}		10.88	8.50 (8.52)	1.92	.06		
change in b							
C = 0.1		11.09	9.77 (8.66)	1.38	-.01	3	1-7
C = 0.05		11.41	9.99 (9.11)	1.13	.04	5	2-10
C = 0.01		12.74	12.07(10.80)	0.64	.00	9	6-20
RMS F=4	75	15.32	11.65(10.82)	3.08	.05	4	1-13
F=2		17.86	14.12(11.85)	2.57	.10	7	2-15
G'_{Rp}		14.22	10.92(11.67)	2.99	.03		
change in b							
C = 0.1		15.38	14.02(12.18)	2.00	-.06	3	1-7
C = 0.05		17.24	15.09(13.11)	1.70	.04	5	1-12
C = 0.01		20.53	19.18(16.79)	1.03	.04	11	6-23
RMS F=4	50	28.27	20.11(16.91)	5.34	.13	3	1-10
F=2		35.54	27.63(19.70)	4.50	.15	7	3-18
G'Rp		36.89	24.99(19.25)	7.41	.17		
change in b							
C = 0.1		25.43	22.51(19.07)	4.45	-.08	3	0-7
C = 0.05		29.93	26.28(21.08)	3.87	-.01	4	0-12
C = 0.01		46.99	44.81(34.87)	2.11	.00	11	4-29

The component of variance Δ_x is always considerably smaller than Δ_r. As we would expect it is lowest for the models which contain the largest number of covariates, and relatively larger for models with few covariates. Selection by the maximum change in b* tends to give lower values of Δ_x, at corresponding values for p. This at first seems counter-intuitive. Since this procedure does not make explicit use of information about the residual sums-of squares, it might be expected that it would omit from the model some variables which are good predictors of the outcome (BASC), and hence this "random-effects" component of the variance would be inflated. However this does not seem to be the case, and the forward stepwise procedure driven by the maximum absolute change in b* seems to hold promise for the random-effects model, as well as for the fixed-effects model.

It is also interesting to note that there is little, if any, evidence of dependence between the two components Δ_r and Δ_x. Thus the variance of the estimate of β^* from the fixed effects model will be an independent contribution to the total variance, and for these data the dominant contribution. Thus we would not expect that any procedures which performed badly for the fixed effect simulations to do better when evaluated for a random-effects model.

The extent to which the column "true" is lower than the column var(Δ_r) is a measure of the extent to which selection variance is operating for the conditional estimates. There is considerable sampling error in the comparison of these quantities from the simulated data, and it must be remembered that the different

stopping criteria within a selection procedure do not provide independent estimates. Taken together, however, they suggest that selection variance may increase the variance of the estimators by from 20% (at n_{sub} = 50) to somewhere between 5% and 10% (at larger sample sizes).

10.4 Further simulations for sub-samples of 75.

To obtain a more precise comparison of the three selection procedures the results for n_{sub} = 75 were repeated on a common set of simulated data for all three procedures. This was a set of 404 simulations which were the first 404 of the 500 given above for selection by the change in b*. This curious choice of number corresponded to the time when the simulations driven by G'_{RP} ran out of computer time. Results are given in table 10.7.

The simulations were extended to cover a wider range of stopping criteria (F and C) for the procedures which minimise the RMS and the absolute change in b*, respectively. Also, as a further check, two of the models which gave poor results on the fixed effects simulations were evaluated for the same data.

Testing the estimators on the same data also allows us to evaluate their correlations. Apart from correlations within procedures, between different levels of the stopping rule, the highest correlation was 0.90 between the G'_{RP} procedure and the RMS with F=2. The other correlations ranged between .56 and this value.

Table 10.6 : Comparison of estimators of b* for n_{sub} = 75 (m = 404).

Model	var(b*)	est(var(b*))	r	bias(se)	MSE (95% C.I.)
Full	23.24	22.66	*	*	100%
RMS					
F=30	16.82	14.17	.60	-.81(.20)	75% (64% - 88%)
F=20	14.85	12.74	.63	-.13(.19)	64% (55% - 75%)
F=10	14.88	11.92	.74	.02(.16)	64% (56% - 73%)
F=8	15.14	11.57	.73	.06(.16)	65% (56% -73%)
F=6	15.36	11.08	.70	.16(.17)	66% (57% - 79%)
F=5	15.29	10.78	.72	.22(.17)	66% (58% - 78%)
F=4	15.81	10.44	.74	.26(.16)	68% (60% - 78%)
F=2	18.12	9.97	.83	.24(.14)	78% (70% - 87%)
G'_{RP}	17.03	10.02	.89	.05(.10)	73% (67% - 80%)
change in b*					
C=1.0	19.56	15.94	.40	-.23(.23)	67% (56% - 80%)
C=0.8	18.63	15.75	.43	-.50(.27)	81% (68% - 96%)
C=0.6	17.32	15.27	.47	-.23(.23)	71% (56% - 80%)
C=0.4	14.47	14.87	.55	.14(.20)	62% (53% - 74%)
C=0.2	13.90	14.33	.65	.21(.18)	59% (51% - 69%)
C=0.1	15.50	14.35	.72	.25(.17)	67% (58% - 77%)
C=0.05	17.35	14.03	.80	.30(.14)	75% (67% - 85%)
C=0.01	21.47	15.16	.94	.13(.08)	92% (86% - 98%)
G_{FP}	20.61	13.39	.94	-.17(.05)	89% (83% - 95%)
G'_{FP}	15.98	13.57	.82	-1.09(.14)	74% (66% - 83%)

The stopping rule for the RMS procedure could be increased considerably, effectively requiring rather extreme significant levels before a covariate is entered, without detriment to the MSE. Examination of the details of the simulations showed that the values of 6, 8 and 10 for F gave a majority of models with just a single covariate, though by no means always the same one. It was not until F= 20, and to an even greater extent at F=30, that some simulations had no covariates, which gave estimates with a negative bias and a worse variance.

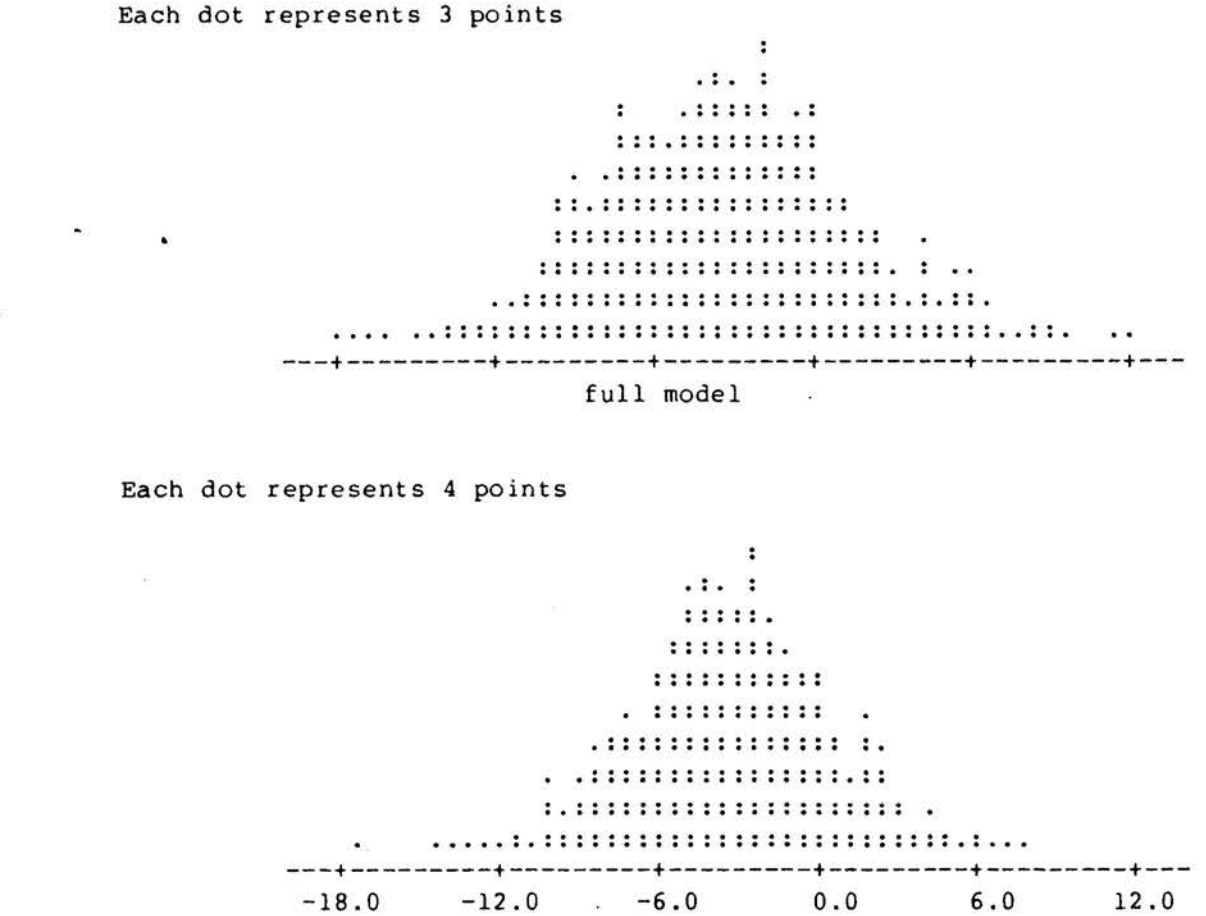
The simulations by the maximum absolute change in b* began to get worse and have a negative bias when C had values of 0.6 and larger. The values 0.2 and 0.4 for C performed rather better than the previous best value of 0.1.

The procedures based on the two G_{FP} procedures did badly. The results for G_{FP} were very much as expected, giving a high correlation with the true value, and much less improvement in variance than were found for the other procedures. The G'_{FP} results were rather different from those for the fixed-effect simulations. For the fixed-effect simulations (for the full sample) this procedure gave a small negative bias and an increase in variance compared to the full model. Here it gives a large bias, and a relatively small variance which together make it's MSE quite modest. However, despite the reasonable MSE value, other less biassed procedures would always be preferred.

To confirm the results which compare the RMS procedure with the change in b* procedure, each procedure was evaluated for the 500 sets of simulated data presented for the other procedure in tables 10.2 and 10.4. This gave a total of 1000 simulations for comparing the two procedures. The results were very similar to those presented above. On average, the procedure of selecting by changes in b* gave slightly smaller variances, but this could have been the result of sampling errors from the simulations. The second set of data also gave estimated variances for b* which were severely biassed for the RMS procedure, but apparently unbiassed for selection by changes in b*, when C has the values 0.1, 0.2 and 0.4.

The joint distribution of these 1000 estimates was examined, and no irregular features were apparent. The histograms of the estimates of b* for the full model and for the best of the covariate selection procedures (changes in b* with C=0.2) are illustrated in figure 10.1. They also serve to remind us that with this sample size, a study of the effects of lead on children's ability would contribute very little, on its own, to increase our knowledge. Thus its most likely use would be as a contribution to a meta-analysis. This would put a greater weight on obtaining unbiassed estimates than a simple MSE criterion would imply.

Figure 10.1 Histograms of b* estimates for the full model, and for selection by maximum changes in b* with C=0.2, n_{sub}=75, m=1000.



10.4 Provisional recommendations

What strategy is suggested by these results? If I were to be presented with another data set tomorrow, with structure similar to that of the Lead Study data, how would I proceed?

The first step would be to calculate the proportion of the variance of X* which could be explained by all the potential confounders. If this quantity is less than 20% then it is unlikely that any sub-model will give improved estimates, and one should proceed to estimate from the full model. This rule does not make direct reference to sample size. However, by using the simple ratio of the residual sum-of-squares due to the regression to the total sum-of-squares (not corrected for degrees of freedom) the proportion of variance explained will be larger as the residual degrees of freedom are reduced. If the covariates and X* are independent the expected value of this percentage would be $100x(p-2)/(n-p+1)$.

If this percentage exceeds 20%, then I would provisionally suggest a selection procedure based on the maximum change in b* , with a stopping criteria of C=0.1. The variance estimate which one obtained after such a procedure might not be too bad. The procedure of selecting by the lowest RMS might perform just as well, but one would not want to rely on the variance estimates obtained after such a procedure. The variance estimates could be corrected a re-sampling procedure (eg Efrom 1979), but this would be complicated, and would involve re-computing the stepwise procedure for each bootstrap sample (Efron & Gong 1983, Snappinn & Knoke 1989).

To what extent is the success of this procedure, in this chapter, a consequence of the special structure of the lead study data ? What features might cause it to go wrong ? There are several possibilities which were not true for the Lead Study data. Firstly, there may be no real models with small biases. We would hope that the selection procedure would not exclude any covariates for such data, but this needs to be tested. Secondly, we might be in a situation where the benefit of the sub-model is due to there being covariates which are correlated with X^* but not with Y, rather than to a reduction in the residual degrees of freedom. Would the procedure work so well here, and still give unbiassed variance estimates? The final substantive chapter attempts to cover a little of this ground. However, the number of possible parameter combinations is so great that it cannot possibly be comprehensive. Detailed examination of the structure of the covariates from other epidemiological studies might provide useful insight into the features one should be looking for, but this is beyond the scope of this thesis.

Chapter 11

Simulations for multivariate normal data

11.1 Generation of the simulated data

To get a wider view of the variable selection procedure, data for y, x* and x were generated with a multivariate normal distribution. The model used is a special case of both random-effects models introduced in chapter 4 and in chapter 8. It is convenient to use the notation introduced for this model in section 4.5. In particular we can write the sample sums of squares and products matrix of X* and X, about their means, as

$$\begin{Bmatrix} S_{kk} & S_{k*} \\ S_{*k} & S_{**} \end{Bmatrix} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (11.1)$$

where $S_{**} = \Sigma\ (X^* - \bar{X}^*)^2$ is a scalar, and $S_{kk'}$, S_{*k} are calculated similarly as the vector and matrix of sums-of-squares and cross products for X and X*. All summations are over the n observations. For convenience, in the algebra which follows, I have reorganised the layout of this matrix to put S_{kk} in the top left hand corner.

Where x and x* are multivariate normal, this sample sums of squares and products matrix will have a Wishart distribution with a variance-covariance matrix, which will be denoted by

$$\begin{bmatrix} \Sigma_{kk} & \Sigma_{k*} \\ \Sigma_{*k} & \sigma_{**} \end{bmatrix} \quad \ldots\ldots\ldots\ldots\ldots\ldots (11.2)$$

where Σ_{k*} is a vector. From this distribution for the independent variables, the vector Y is predicted as a function of X and X* by the equation

$$Y = \beta_0 + X^* \beta^* + X \beta + \varepsilon_y,$$

where ε_y is also normally distributed with mean zero. The joint distribution of Y, X* and X is then also multivariate normal with a variance-covariance matrix

$$\left[\begin{array}{cc|c} \Sigma_{kk} & \Sigma_{k*} & \Sigma_{ky} \\ \Sigma_{*k} & \sigma_{**} & \sigma_{*y} \\ \hline \Sigma_{yk} & \sigma_{y*} & \sigma_{yy} \end{array}\right] \quad \ldots\ldots\ldots\ldots (11.3).$$

The vector of quantities [Σ_{yk} , σ_{y*}] is readily calculated by multiplying the inverse of the matrix 11.2 into the vector [β , β^*], and the quantity σ_{yy} is determined from the variance of the ε_y and the other parameters.

In simulating data from this distribution one can, without loss of generality, take all the means to be zero and all the diagonal elements of 11.3 to be 1. Any variance-covariance matrix can be reduced to this form by a scale and location transformation. Also, since one can reverse the scoring of any of the Xs, the convention

will be adopted that the regression coefficients β and β^* will always be positive or zero.

The first step in setting up the simulated data was to compute the matrix 11.3 from the following quantities which could be changed to alter the parameters of the problem :-

(1) the off-diagonal elements of Σ_{kk}, which are the correlations between the covariates;

(2) the correlations Σ_{k*} between x^* and the other covariates;

(3) the regression coefficients β and β^*.

This completely defines the matrix 11.3, which is easily calculated as described above. I found this approach to the simulation more convenient than defining the matrix 11.3 as the starting point. It helped to keep the value of β and β^* at the same values for different sets of simulations and alter the degree of confounding by adjusting the correlations.

Once the matrix 11.3 has been computed it is easy to generate multivariate normal data with this covariance structure (Morgan 1984). Several options are available, some of which involve generating the sums of squares and products directly from the Wishart distribution (eg Smith & Hocking 1972) . This choice might have used less computer time, but since the time for the generation of the random variables was only a small fraction of the time required to compute the stepwise procedures, the option of generating X, X^* and Y as vectors was chosen. These vectors were first filled with independent normal random variables with mean zero

and variance 1. A Choleski factorisation of the matrix 11.3 was then used to obtain a lower triangular matrix (Z) which was multiplied into the matrix [X, X*, Y] to give a set of vectors with the desired covariance structure.

Since we know the true values of all the parameters of this problem we can calculate the true regression coefficients for the full sample, and also the bias and variance of the estimate of β^* for any specified sub-model. These quantities are easily calculated from the expressions in chapters 4 & 8. In the notation introduced above with the quantities Σ_{pp}, β_p and Σ_{*p} being the sub-matrix and sub-vectors corresponding to the covariates retained in the model, we obtain the following expressions for the mean and variance of the estimate of β^* for the sub-set

Mean(b*) = last element of

$$\begin{bmatrix} \Sigma_{pp} & \Sigma_{p*} \\ \Sigma p* & \sigma_{**} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{p*} \\ \sigma_{**} \end{bmatrix} \quad \ldots\ldots\ldots\ldots\ldots (11.4)$$

and from the expression 8.6, the expected value of the variance of b* from the sub-set is

$$\sigma_{resid}^2 \; / \; [\; \sigma_*^2 \; (n - p - 1) \;]$$

where σ_{resid}^2 is the residual variance of y from the sub-model and σ_*^2 is the residual variance of x* from its regression with the p covariates in the model. These two quantities can be written as

$$\sigma_*^2 = 1 - \Sigma_{*p}\,\Sigma_{pp}^{-1}\,\Sigma_{p*}$$

and

$$\sigma_{resid}^2 = 1 - \begin{bmatrix} \Sigma_{pp} \\ \sigma^{**} \end{bmatrix}' \begin{bmatrix} \Sigma_{pp} & \Sigma_{p*} \\ \Sigma_{p*} & \sigma_{**} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{py} \\ \sigma_{*y} \end{bmatrix}$$

The changes in the expected mean and variance of β^* for subsets, were explored for various choices of the parameters. A selection was made which gave features which were similar to the patterns that can occur in real multiple regressions and which might be good tests of the stepwise procedures. In particular, when some selection procedures seemed to be performing reasonably well, I tried them out on data which might be likely to make them go wrong. Although there was some arbitrariness in this procedure, the initial sets of parameters were selected before any of the simulations of the stepwise procedures were run. Also, I report here on all the tests on simulated data which I have performed, and have not selected runs from among a larger set which favour one particular procedure.

11.2 Data where sub-sets give almost no advantage.

As a severe test of the variable selection procedures, some data were generated for which there were few sub-models which gave an improved MSE compared to the full model. In addition to X*, 10 covariates were chosen, each of which had a correlation of 0.5 with all the others and a correlation of 0.4 with X*. The prediction equation was chosen to give β^* and all the elements of β a value of 0.1. Two sample sizes of 30 and 100 were considered. Because all

the x's have the same distribution, models with any fixed number of the covariates included will all have the same properties. These are tabulated in table 11.1.

Table 11.1 Properties of estimation of β* from sub-models

No covariates	Mean(b*)	Variance(b*)	MSE(b*)	Residual variance of y
n=30				
0	0.5000	0.0278	0.1875	0.7500
1	0.3143	0.0260	0.0720	0.5689
2	0.2356	0.0250	0.0434	0.4922
3	0.1921	0.0247	0.0331	0.4498
4	0.1645	0.0247	0.0289	0.4229
5	0.1455	0.0251	0.0271	0.4043
6	0.1315	0.0256	0.0266	0.3907
7	0.1208	0.0264	0.0268	0.3803
8	0.1124	0.0274	0.0275	0.3721
9	0.1056	0.0285	0.0285	0.3655
10	0.1000	0.0299	0.0299	0.3600
n=100				
0	0.5000	0.0077	0.1677	0.7500
1	0.3143	0.0071	0.0530	0.5689
2	0.2356	0.0066	0.0250	0.4922
3	0.1921	0.0063	0.0148	0.4498
5	0.1455	0.0061	0.0103	0.4043
6	0.1315	0.0060	0.0081	0.3907
7	0.1208	0.0059	0.0069	0.3803
8	0.1124	0.0059	0.0063	0.3721
9	0.1056	0.0058	0.0060	0.3655
10	0.1000	0.0058	0.0058	0.3600

When n is 100 there are no sub-sets with a lower MSE than the full model, and for n=30 only subsets with between 4 and 9 covariates have an improved MSE, and none of these is much better than the full model. In both cases all subsets with a few covariates are severely biassed. Thus, one would not expect any sub-set selection method to give much improvement over the full model - and there is the potential for them to do much worse. Notice also that the relationship between x* and the covariates is quite strong with an

expected value for R^{*2} of .55 when n=30 and .36 when n=100, so my provisional recommendations would suggest that there was potential for sub-set selection. The expected value of the adjusted R^{*2} in both cases is 0.59. Thus the benefit from subsets is largely due to the correlation between X* and X when n=100, and is not large enough to overcome the bias in the sub-models. However when it is augmented by the advantage from increased residual degrees of freedom when n=30, there are modest advantages for some sub-models.

The results for 200 sets of simulated data analysed by the three subset selection procedures are in table 11.2. A stopping rule of F=4 was used for selection by the residual mean square, and of C=0.1 for selection by the maximum absolute change in b* (from now on referred to as $\Delta(b^*)$). In this and subsequent tables approximate values for the s.e.s of the mean and variance estimates are included with every table. Usually these were rather similar for the different methods, and the largest value for any method has been included in the tables.

Table 11.2 Estimates of β^* after variable selection (m=200)
$\beta^* = 0.1$, $\beta = .1$, $\Sigma_{**} = .5$, $\Sigma_{*k} = .4$.

Method	mean(b*)	var(b*)	MSE(b*)	est. var.	no covariates median & range
n=30					
Full	0.1049	0.0286	0.0286		10
RMS	0.1676	0.0279	0.0325	0.0165	2 (1 -4)
G'_{RP}	0.1370	0.0287	0.0301	0.0167	3 (1 -8)
$\Delta(b^*)$	0.1034	0.0252	0.0252	0.0237	3 (1 -6)
s.e.	.01	.003			
n=100					
Full	0.0927	.00590	.00590		10
RMS	0.1390	.00801	.01024	0.0052	4 (1 - 6)
G'_{RP}	0.1050	.00613	.00628	0.0054	5.5(1 - 8)
$\Delta(b^*)$	0.1091	.00625	.00651	0.0062	3 (1 - 6)
s.e.	.005	.0007			

These results show $\Delta(b^*)$ as the clear winner. The two other procedures give biassed estimates when n=30 with MSEs larger than the full model. The ratio of MSEs for $\Delta(b^*)$ relative to the full model for n=30 is 0.881 with a 95% confidence interval of (0.804, 0.965) calculated by the methods of the previous chapter.

When n=100 RMS still performs badly, although selection by G'_{RP} would give acceptable estimates. Both RMS and G'_{RP} under-estimate the variance of the estimates after selection, whereas the variance estimates from $\Delta(b^*)$ seem acceptable. The ratio of MSEs for $\Delta(b^*)$ relative to the full model is 1.10 with a 95% confidence interval of (1.00, 1.20). This suggests that the $\Delta(b^*)$ procedure may be

performing slightly worse than the full model, which is not surprising for these data where all sub-sets are worse than the full model. However, the price in terms of increased MSE is not great.

11.3 Data with a diagonal covariance matrix.

If all the correlation and regression coefficients are zero, then the variance matrix (11.3) becomes the unit matrix. This is the case most favourable to variable selection, because all the sub-models are better than the full model, and no bias is introduced. Simulated data were generated for this case. Again 10 covariates were considered, at the two sample sizes of 30 and 100. Results are in the first two sections of table 11.3.

Table 11.3 Estimates of β* after variable selection (m=200), all covariances and regression coefficients zero

Method	mean(b*)	var(b*)	MSE(b*)	est. var.	no covariates median & range
n=30	10 covariates				
			(true value)		
Full	0.0068	0.0562	(0.0588)		10
RMS	0.0101	0.0348	0.0348	0.0335	0 (0 -3)
G'_{RP}	0.0036	0.0451	0.0451	0.0326	2 (0 -7)
Δ(b*)	0.0041	0.0382	0.0382	0.0368	0 (0 -5)
s.e.	.015	.004			
n=100	10 covariates				
			(true value)		
Full	-0.0173	.0116	(.0115)		10
RMS	-0.0091	.0106	.0107	0.0101	0 (0 - 4)
G'_{RP}	-0.0102	.0110	.0111	0.0100	1 (0 - 5)
Δ(b*)	-0.0093	.0107	.0107	0.0103	0 (0 - 3)
s.e.	.007	.001			

Table 11.3 continued

Method	mean(b*)	var(b*)	MSE(b*)	est. var.	no covariates median & range
n=90	30 covariates				
			(true value)		
Full	-0.0015	.0204	(0.0175)		10
RMS	-0.0015	.0144	.0144	0.0106	1 (0 - 8)
G'_{RP}	-0.0007	.0160	.0160	0.0102	4 (0 - 12)
$\Delta(b^*)$	-0.0015	.0142	.0142	0.0113	0 (0 - 2)
s.e.	.008	.001			

The variable selection procedures perform better than the full model, although the potential for improvement is not great when n=100, because the true variance for the model with no covariates is only 0.0099, compared with 0.0110 for the full model. The G'_{RP} criterion includes a few more covariates, and so does a little worse than the other two. The variance estimates from the reduced models showed little evidence of a downwards bias, with the possible exception of the G'_{RP} when n=30.

To test for problems with selection variance for data with this covariance structure, a further set of simulated data were generated for 30 covariates and a sample size of 90. Results are also in table 11.3. Estimation of b* from the reduced models is, once again, a great improvement on the full model. At first sight there appears to be some under-estimation of the residual variance. However, we know that the true full-model variance of b* for these data should be 0.0175 and for sub-models with 0, 1 and 2 covariates the true variances are 0.0115, 0.0116 and 0.0118. This suggests that the estimates of b* from this particular set of simulated data have, by chance, a larger sample variance than expected. Taking

this into account there is little evidence of under-estimation of the variance of b^*.

11.4 No bias in b^*, but dependence between x and x^*

A set of parameters were sought which would result in sub-models with low MSEs because of dependence between the x variables and x^*, but no dependence between x and y. The parameters evaluated in section 11.3 gave better estimates from the reduced models because of the additional residual degrees of freedom. These parameters will give improved estimation from sub-models because of the larger conditional variance of x^*.

The true value of β^* was 0.2, and the βs for a set of ten covariates were zero. The covariates correlated at 0.3 (first simulation) or at 0.6 (second simulation) with x^*, and at 0.4 with each other. The sample size was fixed at 200, large enough for the effect of reduced residual degrees of freedom to be unimportant. All the estimates of β^* for sub-models are unbiassed. The variances of the estimates of β^* from sub-models (selected without reference to the data) are given in table 11.4.

Table 11.4 Variance of sub-model estimates of β*, no bias but dependence between x and x*

Number of covariates	variance(b*), when correlations(x, x*) are 0.3	0.6
0	.0049	.0049
1	.0054	.0077
2	.0056	.0101
3	.0058	.0124
4	.0059	.0144
5	.0060	.0162
6	.0061	.0180
7	.0062	.0195
8	.0063	.0210
9	.0063	.0223
10	.0064	.0236

The correlation of 0.6 between x and x* gives much greater advantage to the sub- model estimators. The results of the three stepwise procedures for the simulated data are given in table 11.5.

Table 11.3 Estimates of β* after variable selection (m=200) no bias but dependence between x and x*

Method	mean(b*)	var(b*)	est. var.	no covariates median & range
n=200	10 covariates correlation(x, x*)= 0.3			
Full	.1930	0.0069	(0.0064)	10
RMS	.1952	0.0059	0.0050	0 (0 -3)
G'_{RP}	.1948	0.0059	0.0050	0 (0 -4)
Δ(b*)	.1954	0.0068	0.0055	1 (0 -3)
s.e.	.005	.0006		
n=200	10 covariates correlation(x, x*)=0.6			
Full	.1866	0.0255	(0.0236)	10
RMS	.1935	0.0121	0.0059	0 (0 -3)
G'_{RP}	.1967	0.0111	0.0058	0 (0 -5)
Δ(b*)	.1912	0.0221	0.0119	3 (0 -7)
s.e.	.01	.002		

For these parameters the RMS and G'_{RP} methods give lower variances for b* than does $\Delta(b^*)$. The latter gives a variance comparable to that of the full model. For the second set of simulated data, for which the correlation between the x's and x* are 0.6, all the methods tend to underestimate the variance of b*. These results are what one might expect. Since the x's are unrelated to y, RMS and G'_{RP} select few covariates. The $\Delta(b^*)$ method selects more variables because the high correlations between the xs and x* give a range of values of b*, some of which will exceed the value for the stopping rule.

The under-estimated variances of b* for these data must be a consequence of some mechanism other than the under-estimation of the residual variance, since they occur for all three procedures and the scope for under-estimating the residual variance is much less for data with this larger sample size.

The population values of the multiple correlations between x* and the ten covariates are .19 and .78 for the two sets of simulated data. These are the quantities which would be estimated by R^{*2}_{adj}. The corresponding values for the expected values of R^{*2} are 0.23 and 0.79. The population multiple correlation between x* and the ten covariates for the data in section 11.2 was only .31, although the expected values of the proportions of variance explained were larger at .55 and .36 because of the smaller sample numbers.

11.5 More data which give biassed estimates

To test the $\Delta(b^*)$ procedure, a covariance structure was generated was devised which I felt might give the greatest problems to this procedure. The parameters were identical to those in the previous section (11.4), which already gave some problems for this method, except for the values for the regression coefficients for the other x's which were all taken as 0.05. These small values were selected to make it difficult for individual x's to meet the inclusion criterion for changes in b^*. The same two values of the correlations between x^* and the other x's were chosen for a sample size of 200 and 10 covariates. In addition, a third set of data was generated with 20 covariates and a correlation of 0.6 between each of them and x^*.

The true value of β^* was 0.2, and the mean values of b^* for omitting all the covariates from the three data sets were 0.35, 0.50 and 0.80. The results for the three selection procedures are in table 11.4. No subsets, for any of the three data sets, had a MSE for b^* which was less than 97% of that for the full model, and many had values much larger than that for the full model. The third set of data was the most extreme.

Table 11.4 Estimates of β* after variable selection (m=200) more biassed data with dependence between x and x*

Method	mean(b*)	var(b*)	MSE(b*)	est. var.	no covariates median & range
n=200	10 covariates correlation(x, x*)= 0.3				
Full	.1937	0.0056			10
RMS	.2243	0.0057	0.0063	0.0046	2 (1 -4)
G'_{RP}	.2090	0.0059	0.0060	0.0047	3 (1 -7)
Δ(b*)	.2135	0.0053	0.0055	0.0048	2 (1 -3)
s.e.	.005	.0006			
n=200	10 covariates correlation(x, x*)=0.6				
Full	.1884	0.0192			10
RMS	.3499	0.0158	0.0383	0.0059	1 (0 -5)
G'_{RP}	.2746	0.0210	0.0266	0.0058	3 (0 -8)
Δ(b*)	.2401	0.0177	0.0193	0.0102	4 (1 -7)
s.e.	.01	.002			
n=200	20 covariates correlation(x, x*)=0.6				
Full	.2097	0.0089			10
RMS	.4219	0.0158	0.0650	0.0047	1 (0 -5)
G'_{RP}	.3894	0.0196	0.0555	0.0065	3 (0 -8)
Δ(b*)	.2810	0.0141	0.0205	0.0062	4 (1 -7)
s.e.	.01	.002			

The results resemble those of section 11.2, in that Δ(b*) gives better results than the other procedures. However, it no longer gives unbiassed estimates of the variance of b*, and for the third set of data it gives values which are much worse than using the full model.

11.6 Simulations designed to resemble the Lead Study data

The final set of simulations were designed to have a data structure which was similar to the Lead Study data. The regression coefficient for X* was 0.2, and 20 covariates were included which were divided into three groups. The first group were 10 covariates which had no correlations with either X* or Y. The second group of 6 had regression coefficients of 0.1, and correlations of 0.1 with X* and of 0.5 with each other. The third group of 4 covariates had regression coefficients of 0.1, and correlations of -.1 with X* and of .3 with each other. Every variable in the second group had the same correlation of -.2 with every variable in the third group. The model with no covariates here would give a mean estimate of β^* of 0.22. Sample sizes of 200, 100, 50 and 30 were investigated.

The results were very similar to those for the samples from the Lead Study data investigated in chapter 10. For n=200, there was no advantage in any of the sub-set selection procedures. For the smaller sample sizes the selection procedures all gave better results than the full model. For the smallest sample size of 30 this was a reduction by a factor of 0.35 in the MSE of b*. The three methods gave comparable results except for n=30 when the G_{Rp} did less well than the other two, but still much better than the full model. The RMS and G_{Rp} methods gave underestimates of the variance of b^* at n=100, 50 and 30. There was some evidence of underestimation of the variance for the $\Delta_{(b^*)}$ method at n=30 and n=50, but to a much lesser extent than for the other two methods. For example, for n=30 $\Delta_{(b^*)}$ underestimated the variance by a factor

of 0.74, while the factors for the other two methods were 0.33 and 0.44.

Detailed tables of results are not included because the purpose of the simulations was to confirm the results of the previous chapter. The same conclusions applied as were reached in chapter 10. The main benefit for sub-set selection is at reduced sample sizes, and the preferred method is $\Delta_{(b}$*) because it is less prone to underestimation of the variance of the estimates.

11.6 Summary and conclusions

The Δ(b*) procedure, which appeared the most useful in the last chapter, still seems to be the best of the stepwise procedures which I have tried. However, the results in this chapter suggest that its benefits may only apply when the reduced variance of the sub-models is a consequence of an increase in the residual degrees of freedom. The other two stepwise procedures gave better results for the case when no bias was introduced by the covariates, but there was a reduction in the variance for the sub-models as a result of the association between x* and the other covariates. However, this was at the expense of over-optimistic variance estimates. Also, both RMS and G'_{RP} gave disastrous results for models which were rather similar to the zero-bias model, but where an association between the x's and y gave biassed estimates of b*. In the real world, one would not know when this was occurring. The Δ(b*) method performed rather better in this situation, but it could

still give biassed estimates when there was a large degree of confounding, and it also gave under-estimates of the variance of b*.

For some data with this structure $\Delta(b^*)$ can give results which are worse than the full model. The particular covariance structure for which this arose was rather strange, with 20 covariates all equally and weakly related to y, and a sample size of 200. Smaller sample sizes, with the same structure gave satisfactory results. Also, the $\Delta(b^*)$ procedure is not free from problems of underestimated variances, although these are less severe than for the other two procedures considered. It is difficult to generalise from the formal structure of these artificially generated data sets to the real world. A more fruitful approach might be to investigate other real data sets - but I must leave this to others. However, the results are sufficiently worrying to suggest an amendment to the provisional recommendations from the last chapter.

11.7 Final recommendations

If a study is large enough for the number of covariates to be small compared with the residual degrees of freedom, then no sub-set selection should be attempted. Compared with my previous suggestion, this will exclude those cases where improved sub-models might come about because of strong relationships between x* and x. Covariates which are strongly related to x* may be suspect in other respects. They might be possible sources of over-control, and thus should not be included in the regression because they are causally related to x*. I suggest that such variables should be identified

and examined, but they should not be excluded via a stepwise procedure unless the condition on sample size is fulfilled.

This recommendation is very similar to the one made for clinical trials by Schluchter & Forsythe (1985) and discussed in chapter 3. They suggest that no stepwise procedures should be used unless the sample size is small relative to the number of covariates, although their work applies to much smaller sample sizes and numbers of covariates than have been considered here.

It is difficult to make a definite ruling about when the number of covariates is large enough for subset selection to be worthwhile. This may depend on the absolute number of covariates being considered. For the range of 10 - 30 covariates considered in this thesis, a suitable rule might be to attempt no selection unless the residual degrees of freedom are less than three times the number of covariates. This is probably somewhat cautious, because there were examples of benefit from sub-set selection with more residual degrees of freedom. However, it will give some protection against things going as badly wrong as they did for the last data set in section 11.5.

A well-designed study should have more residual degrees of freedom than are required by this rule. It is important to have additional degrees of freedom to check for such features as interactions and linearity of effects. If we refer back to the list of lead studies in table 2.1, however, we see that two of the five

studies listed there could be candidates for variable selection on the basis of such a rule.

If variable selection is to be attempted, then the $\Delta(b^*)$ rule is the best of the those I have evaluated here. It has a considerable number of advantages. It is simple to compute, and is intuitively reasonable. Another advantage is that is immediately generalisable to other regression techniques such as logistic regression and regression methods in survival analysis. Stepwise methods are used extensively in this area, usually based on the deviance statistic. The consequences for statistical inference, and estimation of the use of stepwise procedures for these techniques, remains to be explored.

Another possible approach to increasing the residual degrees of freedom would be to reduce the dimensions of the X variables by a method such as cluster analysis or principal component analysis. Something similar to this was done by Gardner (1973) for a problem with 61 observations and over 100 covariates. A potential advantage of such an approach is that it does not use information on the relationship between the covariates and X^* and Y to reduce the degrees of freedom, so inferences from such reduced models may be more valid. The method used by Gardner, however, did a preliminary selection of the covariates on their correlations with Y, which might lead to an underestimate of the residual-mean-square.

Such methods might run into difficulty where a set of variables are highly correlated, and yet have different

relationships with X* and/or Y. Examination of the correlation matrices for the Lead Study data in tables 5.5 and 5.6 suggest that this is relatively uncommon, although one can identify possible problems. A method such as principal components may have additional benefits over sub-set selection when covariates are subject to errors of measurement. Further exploration of these topics lies beyond the scope of this thesis. However, the results I have obtained here would suggest that no such procedures are likely to be of benefit, compared to the full model, unless the sample size is small relative to the number of covariates.

Notation and abbreviations

This list excludes certain notation which was used only for intermediate quantities within derivations. The matrix X has different meanings for the random-effects and fixed-effects models, and these are listed separately below along with their implications for related quantities.

Fixed and random effects models

n	number of observations
k	number of covariates in full model, including a constant and X* (ie no of additional covariates+2)
p	number of covariates in a sub-model, including a constant and X* (ie no of additional covariates+2)
q	number of omitted covariates (q=k-p)
y	random variable for the outcome variable
Y	n-vector of observations of y
σ^2	variance of y conditional on fixed values of all k covariates
s^2	estimate of σ^2 from the full model
X*	n-vector of observed values of the variable of special interest
β^*	coefficient of X* in the linear equation for y as a function of all k covariates
b^*	estimate of β^* from the model including all covariates
b^*_p	estimate of β^* from the model including only p covariates
RSS_{full}, RSS_p	residual sums of squares of Y from the full and sub-model regressions
RMS_{full}, RMS_p	corresponding mean squares

Fixed Effects Model only

X	nxk matrix of observed values of the k covariates, including a constant and X*
β	regression coefficient corresponding to X in the linear prediction equation $y=X\beta$, for y

β_p	p-vector of regression coefficients for the covariates included in a sub-model (a sub-vector of β)
b_p	estimate of β_p
S_{**}	sample sum of squares of X* about its sample mean
G_{Fp}, G'_{Fp}	Mean-square-error criteria for β^*
V_{Fp}	variance part of G_{Fp}
V_{1Fp}, V_{2Fp}	components of the above
C_p	Mallows mean-square-error criterion for prediction

Random effects model only

x	random vector of p-2 covariates
X	nx(k-2) matrix of observed values of x
β_o, β	regression coefficients corresponding to a constant and X in the linear prediction equation $y= \beta_o + X^*\beta^* + X\beta$, for y
β_p	(p-2)-vector of regression coefficients for the covariates included in a sub-model (a sub-vector of β)
b_p	estimate of β_p
X_p	nx(p-2) matrix of observed values of the xs corresponding to β_p
x^*	random variable corresponding to X*
Σ_{**}, Σ_{*p}, Σ_{pp}	components of the variance co-variance matrix of x* and x_p
S_{**}, S_{*p}, S_{pp}	sample sum-of-squares-and-products matrix for x* and x_p
A_{**}, A_{*p}, A_{pp}	inverse of the matrix formed from S_{**}, S_{*p}, S_{pp}
σ^2_p	variance of y conditional on the p covariates in the model
GRp, G'Rp	Mean-square-error criteria for β^*
V_{Rp}	variance part of G_{Rp}
V_{1Rp}, V_{2Rp}	components of the above
S_p	Mean-square-error criterion for prediction

Random and fixed effects models

P	nxp matrix of observed values of covariates included in the model, including a constant and X*
Q	nxq matrix of excluded covariates
β_q	sub-vector of β corresponding to Q
b_q	estimate of β_q
s_p^2	residual-mean-square of Y for the model with p covariates (same as RMS_p)
MSE_p	mean-square-error of the estimate of β^* from a sub-model containing p covariates
S_{***p}	residual sum of squares of X* for the regression with the other p-1 covariates included in the model
R^{*2}	multiple correlation between X* and all the other k-1 covariates which could be included in the model
Δb^*	difference between the estimates b* for the full and reduced models
m	number of simulations
r	observed correlation of two estimators evaluated for the same set of simulated data

Stepwise procedures

RMS	selection by the minimum value of the residual sum-of-squares
G_{Fp} etc	selection for the minimum value of the various G_p criteria
$\Delta(b^*)$	selection of the model which gives the greatest (for forward procedures) or smallest (backward procedures) change in the estimate of β^*.

References

Akaike H (1970)
Statistical predictor identification *Ann. Inst. Statist. Math.* 22, 203-17.

Akaike H (1974)
A new look at statistical model identification *IEEE Transactions on Automatic Control* 19, 716-23.

Alvey NG et al (1980)
Genstat - a general statistical program, Lawes Agricultural Trust, Rothamsted.

Allen DM (1971)
Mean square error of prediction as a criterion for selecting variables *Technometrics* 13, 469-75

Anderson TW (1957)
Maximum likelihood estimators for a multivariate normal distribution when some observations are missing. *JASA* 52 200-203.

Bancroft TA (1964)
Analysis and inference for incompletely specified models involving the use of preliminary tests of significance. *Biometrics 20, 427-39.*

Becker RA & Chambers JM (1984)
S - an interactive environment for data analysis, Wadsworth, Belmont California.

Bellinger D, Leviton A, Waternaux C & Alldred E (1984)
Methodological issues in modelling the relationship between low level lead exposure and infant development: examples from the Boston lead study. *Env Res* 38, 119-29.

Berk, KN (1978)
Comparing subset regression procedures *Technometrics 20, 1-6.*

Blalock HM (1964)
Causal inferences in non-experimental research, University of North Carolina Press, Chapel Hill.

Brieman L & Freedman D (1983)
How many variables should be entered in a regression equation. *JASA* 78, 131-6.

Cochran WG (1965)
The planning of observational studies of human populations. *JRSS A* 128, 234-266.

Cochran WG (1983)
The design of observational studies Wiley, New York

Copas JB (1983)
Regression, prediction and shrinkage *JRSS B* 45, 311-354

Cox DR (1984)
Present position and potential developments: some personal views on design of experiments and regression. *JRSS B* 47, 306-15.

Dales LG & Ury HK (1979)
An improper use of significance testing in studying covariables *Int J Epidem* 7, 373-5.

Daniel C & Wood FS (1971)
Fitting Equations to data Wiley, New York

Davie R, Butler NR, Goldstein H (1972)
From birth to seven. A report of the national child development study. Logmans, London.

Demets D & Halperin M (1977)
Estimation of a simple regression coefficient in samples arising from a subpopulation procedure. *Biometrics* 33, 47-56.

Dempster AP, Schatzoff M & Wermuth N (1970)
A simulation study of alternatives to least squares. *J Am Stat Soc* 72, 77-106.

Dixon WJ (ed) (1985)
BMDP Statistical Software University of California Press, Berkeley.

Douglas JWB (1964)
The home and the school MacGibbon & Kee, London.

Douglas JWB, Ross JM, Simpson HR (1971)
All our future Panther, London.

Draper N & Smith H (1981)
Applied Regression Analysis - 2nd edn Wiley, New York

Efron B (1979)
Bootstrap methods: another look at the jack-knife. *Ann Statist* 7, 1-26.

Efron B & Gong G (1983)
A leisurely look at the bootstrap, the jacknife and cross-validation *The American Statistician* 37, 36-48.

Efroymson MA (1960)
Multiple regression analysis. In Ralston A & Wilf HS (eds) *Mathematical methods for digital computers*, Wiley, New York.

Elliot CD, Murray DJ, Pearson LS, (1983)
British ability scales, NFER/Nelson, Windsor.

Environmental Protection Agency (1985)
Air quality criteria for lead Chapter 13 Evaluation of human health risk associated with exposure to lead and its compounds EPA, Research Triangle Park USA

Fisher L & Patil K (1974)
Matching and unrelatedness *Am J Epidem* 100, 347-9.

Flack VF & Chang PC (1987)
Frequency of selecting noise variables in subset regression analysis: a simulation study *Am. Statistician 41, 84-6.*

Fogelman KR, Goldstein H, Essen J, Ghodsian M (1978)
Patterns of attainment *Educat Studies* 2, 121-30.

Forsythe AB (1977)
Post-hoc decision to use a covariate *J Chron Dis* 30, 61-64.

Fulton M, Raab GM, Laxen DPH, Thomson GOB, Hunter R, Hepburn W (1987)
Influence of blod lead on the ability and attainment of children in Edinburgh *Lancet 1, 1221-6.*

Furnival GM & Wilson RW (1974)
Regression by leaps and bounds *Technometrics* 16, 499-511

Galpin JS & Hawkins DM (1986)
Variable subset selection for optimal regression prediction at a specified point. *Journal of Applied Statistics* 13, 187-99.

Gardner MJ (1973)
Using the environment to explain and predict mortality *J R Statist Soc A* 136, 421-440.

Grant LD & Davis JM (1989)
Effects of low-level lead exposure on paediatric neurobehavioral development: current findongs and future directions. In *Lead exposure and child development an international assessment* MA Smith, LD Grant & AI Sors eds, Dordrecht, Kluwer Academic Publishers.

Harvey PG, Hamlin MW, Kumar R & Delves T (1984)
Blood lead, behaviour and intelligence test performance in pre-school children. Sc Tot Env 40, 45-60.

Hausman JA & Wise (1981)
Stratification on endogenous variables and estimation : the Gary income management experiment. In *Structural Analysis of discrete data with econometric applications.* CF Manski & D McFadden eds Cambridge Mass, MIT Press.

Hocking RR (1974)
Misspecification in regression *Am Statist* 28, 39-40.

Hocking RR (1976)
The analysis and selection of variables in regression *Biometrics* 32, 1-51.

Hocking & Leslie (1967)
Selection of the best subset in regression analysis *Technometrics* 9, 531-40.

Holt D, Smith TMF & Winter PD (1980)
Regression analysis of data from complex surveys. *JRSS* A 143, 474-87.

Jewell NP (1985)
Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika* 72 11-21.

Kellmer-Pringle ML, Butler NR, Davie R (1966)
11,000 seven year olds. Nat Childrens Bureau, London.

Kendall MG & Stuart A (1967)
The advanced theory of statistics, Vol 2. Griffin, London.

Kleinbaum DG, Kupper LL & Morgenstern H (1982)
Epidemiological Research, Principles and Quantitative Methods Van Nostrand Reinhold, New York

Kupper LL, Stewart JR and Williams KA (1976)
A note on controlling significance levels in stepwise regression. *Amer J Epidem* 103, 13-15.

Kupper LL & Hogan MD (1978)
Interaction in epidemiological studies. *Am J Epidem* 108, 447-53

Linhart H & Zucchini W (1982)
A method for selecting the covariates in analysis of covariance *S Afric Statist J* 16, 97-112.

Lansdown R & Yule W (1987)
The lead debate: the environment, toxicology and health Croom-Helm, London.

Mallows CL (1973)
Some comments on C_p *Technometrics 15, 661-75.*

Mantel N (1986)
Does passive smoking stunt the growth of children? *Am J Epidem* 15, 427-8.

Miettenin O (1974)
Confounding and effect modification *Am J Epidem* 100, 350.

Miettenin O & Cook EF (1981)
Confounding : essence and detection. *Am J Epidem* 114, 593-7.

Miller AJ (1984)
Selection of subsets of regression variables *JRSS A* 147, 389-425.

Miller RG (1974)
The jacknife - a review *Biometrika* 61, 1-15.

MINITAB Inc (1986)
Minitab reference manual. Release 5 State College, Pennsylvania.

Mitchell TJ & Beauchamp JJ (1988)
Bayesian variable selection in linear regression *JASA* 83, 1023-1037

Morgan BJT (1984)
Elements of simulation London, Chapman & Hall

Moses LE (dated 1983, received 1987)
How much help from covariates in a randomised experiment. *Personal communication of draft lecture notes.*

Narula SC (1974)
Predictive mean square error and stochastic regressors *Appl Stat. 23, 11-18.*

Narula S & Ramburg JS (1972)
Letter to the editor. *The American Statistician* 26, 42.

Needleman HL, Gunnoe C, Leviton A, et al (1979)
Deficits in psychologic and classroom performance ofchildren with elevated dentine lead levels *New Engl J Med* 300, 689-95.

Needleman H (1983)
The neuropsychological consequences of low level exposure to lead in childhood. in Rutter MA & Russell Jones R *Lead versus health*, Wiley, Chichester.

Paterson LJP & Raab GM (in preparation)
Sample selection bias in linear regression

Pitman EJG (1939)
A note on normal correlation. *Biometrika* 31, 9.

Pearson K (1902)
On the influence of natural selection on the variability and correlation of organs. *Phil Trans Roy Soc A-i 200 1 - 66.*

Pinault SC (1988)
An analysis of subset regressions for orthogonal designs *Amer Statistician* 42, 275-7.

Pferrmann D, & Holmes D (1985)
Robustness considerations in the choice of a method of inference for regression analysis of survey data. *JRSS A* 148, 268-278.

Pocock SJ & Ashby D (1985)
Environmental lead and children's intelligence: a review of recent epidemiological studies. *Statistician* 34, 31-44.

Pocock SJ, Ashby D, Smith MA (1987)
Lead exposure and children's intellectual performance *Int J Epidem* 16, 57-67.

Raab GM, Fulton M, Laxen DPH, Thomson GOB (1985)
The Edinburgh lead study aspects of deign and progress *The Statistician* 34, 45-57

Raab GM, Fulton M, Thomson GOB, Laxen DPH, Hunter R, Hepburn W (1989)
Blood lead and other influences on mental abilities - results from the Edinburgh lead study. In *Lead exposure and child development an international assessment* MA Smith, LD Grant & AI Sors eds, Dordrecht, Kluwer Academic Publishers.

Raab GM, Hunter R, Fulton M & Laxen DPH (1987)
Lead from dust and water as exposure sources for children. *Environmental geochemistry and health.* 9, 80-85.

Raab GM & Zhou YJ (1987)
The effect of changes of location on least squares estimators for stratified samples. *Biometrika* 74, 216-9.

Raab GM, Thomson GOB, Boyd L, Fulton M & Laxen DPH
Blood levels, reactio time, inspection time and ability in Edinburgh children. *British Journal of Developmental Psychology (to appear)*

Rao P (1971)
Some notes on misspecification in muliple regression. *Am Stat* 25, 37-9.

Raven JC, Court JH, Raven J (1978)
Maual for Raven's progressive matrices and vocabulary scales. HK Lewis, London.

Rencher AC & Pun FC (1980)
Inflation of R^2 in best subset regression procedures *Technometrics* 22, 49-53.

Rona RJ, Chinn S, & Florey C du V (1985)
Exposure to cigarette smoking and children's growth *Int J Epidem*, 14, 402-9.

Rona RJ, Chinn S & Florey C du V (1986)
Reply to N Mantel *Int J Epidem, 15, 428.*

Rosenberg, SH & Levy PS (1972)
A characterization on misspecification in the general linear regression model. *Biometrics* 28, 1129-32.

Rutter MA (1967)
A children's behaviour questionnaire for completion by teachers. *J Child Psych & Psychol* 8, 1-11.

Rutter M (1980)
Raised lead levels and impaired cognitive/behavioural functioning: a review of the evidence. *Developmental Medicine & Child Neurology* Supplement 42.

Rutter M, Madge M (1976)
Cycles of disadvantage Heinemann, London.

SASPAK user manual Release 3 (1983)
Local authority management services and computer comittee.

Schroeder SR, Hawk B, Otto DA, Mushak P & Hicks RE (1985)
Separating the effects of lead and social factors on IQ *Env Res* 38, 144-54.

Schluchter M (1984)
Selection of covariates to minimise mean-squared error ofestimated treatment effects. *ASA Proceedings of the Survey Research Section* 49-50. (to check)

Schluchter M & Forsyth AB (1985)
Post-hoc selection of covariates in randomized experiments. *Commun Stat Theor Meth 14, 679-99.*

Shirley EAC & Newnham P (1984)
The choice between the analysis of variance and covariance. *Statistics in Medicine* 3, 84-92.

Shibata R (1981)
An optimum selection of regression variables *Biometrika* 68, 45-54

Smith M, Delves T, Lansdown R, Clayton B, Graham P (1983)
The effects of lead exposure on urban children: the Institute of Child Health/Southampton study. *Developmental Medicine & Child Neurology* Supplement 47.

Smith M (1985)
Intellectual and Behavioral Consequences of Low Level LeadExposure : A Review of the Literature, *Clinics in Endocrinology & Metabolism* 14, 657-80.

Smith WB & Hocking RR (1972)
Algorithm AS53, Wishart Variance Generator Applied Statistics 21, 341-3.

Snapinn S & Knoke JD (1989)
Estimation of Error Rates in Discriminant Analysis with Selection of Variables *Biometrics 45, 289-99.*

Toro-Vizcarrando C & Wallace TD (1968)
A test for the mean-square-error criterion for restrictions in linear regression. *J Amer Statist Soc* 63, 558-71.

Thompson ML (1978a)
Selection of variables in multiple regression: Part I A review and evaluation. *Int Statist Rev* 46, 1-20.

Thompson ML (1978b)
Selection of variables in multiple regression: Part II. Chosen procedures, computatiomns and examples. *Int Statist Rev* 46, 129-146.

Thomson GOB, Raab GM, Hepburn WS, Hunter R, Fulton M, Laxen DPH (1989)
Blood-lead levels and children's behaviour - results from the Edinburgh Lead Study *J Child Psychol Psychiat* 30, 15-28.

Wallace TD & Toro-Vizarrando C (1969)
Tables for the mean square error test for exact linear restriction in regression *J Amer Staist Soc* 64, 1649-63.

Walls RC & Weeks DL (1969)
A note on the variance of a predicted response in regression *The American Statistician* 23, 24-6.

Weed D (1986)
On the logic of causal inference *Am J Epidem* 123, 965-79.

Winneke G, Beginn U, Ewart T, Havestadt C, Kramer U, Krause C, Thron HL, Wagner HM (1985)
Comparing the effects of perinatal and later childhood lead exposure on neuropsychological outcome. *Env Res* 38, 155-67.