Influential data cases when the C-p criterion is used for variable selection in multiple linear regression

Daniël Wilhelm Uys

Dissertation presented for the Degree of Doctor of Philosophy at the University of Stellenbosch

Promoter: Prof. S.J. Steel
Co-promoter: Dr. J.O. van Vuuren

April 2003

DECLARATION

I, the undersigned, hereby declare that the work contained in this dissertation is my own original work and that I have not previously in its entirety or in part submitted it at any university for a degree.

Signature:

Date:

SUMMARY

Influential data cases when the C-p criterion is used for variable selection in multiple linear regression

by

Daniël Wilhelm Uys

Degree:	**Doctor of Philosophy**
Department:	**Statistics and Actuarial Science**
Faculty:	**Science**
University:	**Stellenbosch**
Promoter:	**Prof. S.J. Steel**

In this dissertation we study the influence of data cases when the C_p criterion of Mallows (1973) is used for variable selection in multiple linear regression. The influence is investigated in terms of the predictive power and the predictor variables included in the resulting model when variable selection is applied. In particular, we focus on the importance of identifying and dealing with these so called selection influential data cases before model selection and fitting are performed. For this purpose we develop two new selection influence measures, both based on the C_p criterion. The first measure is specifically developed to identify individual selection influential data cases, whereas the second identifies subsets of selection influential data cases. The success with which these influence measures identify selection influential data cases, is evaluated in example data sets and in simulation. All results are derived in the coordinate free context, with special application in multiple linear regression.

OPSOMMING

Invloedryke waarnemings as die C-p kriterium vir veranderlike seleksie in meervoudige lineêre regressie gebruik word

deur

Daniël Wilhelm Uys

Graad:	**Doktor in Wysbegeerte**
Departement:	**Statistiek en Aktuariële Wetenskap**
Fakulteit:	**Natuurwetenskappe**
Universiteit:	**Stellenbosch**
Promotor:	**Prof. S.J. Steel**

In hierdie proefskrif ondersoek ons die invloed van waarnemings as die C_p kriterium van Mallows (1973) vir veranderlike seleksie in meervoudige lineêre regressie gebruik word. Die invloed van waarnemings op die voorspellingskrag en die onafhanklike veranderlikes wat ingesluit word in die finale geselekteerde model, word ondersoek. In besonder fokus ons op die belangrikheid van identifisering van en handeling met sogenaamde seleksie invloedryke waarnemings voordat model seleksie en passing gedoen word. Vir hierdie doel word twee nuwe invloedsmaatstawwe, albei gebaseer op die C_p kriterium, ontwikkel. Die eerste maatstaf is spesifiek ontwikkel om die invloed van individuele waarnemings te meet, terwyl die tweede die invloed van deelversamelings van waarnemings op die seleksie proses meet. Die sukses waarmee hierdie invloedsmaatstawwe seleksie invloedryke waarnemings identifiseer word beoordeel in voorbeeld datastelle en in simulasie. Alle resultate word afgelei binne die koördinaatvrye konteks, met spesiale toepassing in meervoudige lineêre regressie.

Opgedra aan my Pa en Ma

ACKNOWLEDGMENTS

I would like to express my sincerest appreciation to the following persons and organizations:

- My promoter, Prof. S.J. Steel, for his unfailing commitment and outstanding guidance throughout the study.
- The Department of Statistics and Actuarial Science, Stellenbosch University, for providing me with adequate computer and software facilities.
- The Harry Crossley Trust for research funding.
- My parents, family and friends for their continuous interest and moral support.

CONTENTS

TERMINOLOGY AND NOTATION

Matrix and vector algebra are applied throughout the dissertation. Many results are also expressed in terms of projections of vectors on linear subspaces. In this regard we introduce here some terminology and notation conventions which frequently arise. We also mention several standard results from linear algebra.

The vector space of all n-dimensional real vectors will be denoted by R^n. Let

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

be column vectors in R^n and let M be a linear subspace of R^n. The dimension of M is denoted by $\dim(M)$. Also, let $\mathbf{A}$ be an $n \times \dim(M)$ basis matrix for M (i.e., M is spanned by the linearly independent column vectors of $\mathbf{A}$). The following standard notation will be used.

- The transpose of $\mathbf{A}$ is denoted by $\mathbf{A}'$, and that of $\mathbf{a}$ by $\mathbf{a}'$.

- The inner or dot product, denoted by $\langle \mathbf{a}, \mathbf{b} \rangle$, of $\mathbf{a}$ and $\mathbf{b}$ is calculated as

$$\mathbf{a}'\mathbf{b} = \mathbf{b}'\mathbf{a} = \sum_{i=1}^{n} a_i b_i.$$

- $\|\mathbf{a}\| = \sqrt{\langle \mathbf{a}, \mathbf{a} \rangle} = \sqrt{\sum_{i=1}^{n} a_i^2}$ denotes the Euclidean length of $\mathbf{a}$ in R^n.

- The linear subspace which is the orthogonal complement of M, written as $M^\perp$, is the set of all vectors in R^n orthogonal to M. Hence, $\langle \mathbf{a}, \mathbf{b} \rangle = 0$ for all $\mathbf{a} \in M$ and $\mathbf{b} \in M^\perp$. The dimension of $M^\perp$ is $\dim(M^\perp) = n - \dim(M)$.

- Let L be a linear subspace contained in M, i.e. $L \subset M$. The linear subspace M mod L, denoted by $M \mid L$, is then the set of all vectors in M that are orthogonal to L. The dimension of $M \mid L$ is $\dim(M \mid L) = \dim(M) - \dim(L)$.

- The orthogonal projection of $\mathbf{a}$ onto M, $M^{\perp}$ and $M \mid L$ is denoted by $P_M\mathbf{a}$, $P_{M^{\perp}}\mathbf{a}$ and $P_{M|L}\mathbf{a}$ respectively. In particular, since $\mathbf{A}$ is a basis matrix for M, the projection of $\mathbf{a}$ onto M is given by $P_M\mathbf{a} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{a}$, where $P_M = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$. The matrix P_M is symmetric and idempotent and is referred to as the projection or hat matrix with respect to M. Note also that

  $$P_M\mathbf{a} = \sum_{i=1}^{\dim(M)} \frac{\langle \mathbf{a}, \mathbf{a}_i \rangle}{\|\mathbf{a}_i\|^2}\mathbf{a}_i,$$

  where $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_{\dim(M)}$ are the column vectors of $\mathbf{A}$. Similar expressions can be found for $P_{M^{\perp}}\mathbf{a}$ and $P_{M|L}\mathbf{a}$.

- The square matrix $\mathbf{I}_n$ denotes the $n \times n$ identity matrix, and $\mathbf{1}$ the vector of any size with all elements equal to 1.

- $\mathbf{0}$ is used for the null matrix or null vector of any size.

- For any scalar c, $c^+ = \max\{0, c\}$ denotes its truncated value at 0.

- $N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the m-variate normal distribution with $m \times 1$ mean vector $\boldsymbol{\mu}$ and $m \times m$ variance-covariance matrix $\boldsymbol{\Sigma}$.

- The chi-squared distribution with n degrees of freedom and non-centrality parameter λ is denoted by $\chi_n'^2(\lambda)$ if $\lambda > 0$ and by χ_n^2 if $\lambda = 0$.

CHAPTER 1

INTRODUCTION

"The presence of outliers can have a profound effect on model selection, parameter estimation and prediction for a wide range of models." Glendinning (2001)

In this dissertation we study the influence of data cases when variable selection is applied in multiple linear regression. The study therefore involves the simultaneous consideration of the following three well established statistical topics: influential data cases, variable selection and multiple linear regression. Thus, to start with in the first chapter, we give a short overview of these topics. Section 1.1 briefly touches on regression analysis, with specific focus on the multiple linear regression model. In Section 1.2 we discuss the selection of variables in multiple linear regression. Section 1.3 considers some aspects of influential data cases. In this section we pay special attention to existing literature on the influence of data cases if variable selection is applied in multiple linear regression . Finally, in Section 1.4 we briefly discuss the intention and restrictions of the dissertation. Section 1.4 also gives an outline of the subsequent chapters of the dissertation.

1.1 Regression analysis

Regression analysis is one of the most widely used statistical techniques for fitting models to data (Atkinson and Riani, 2000, p.1). Its frequent use and many applications in various disciplines have made regression analysis well documented in the statistical literature. Neter et al. (1990), Ryan (1997) and Draper and Smith (1998) provide recent and comprehensive discussions of regression analysis and its applications.

Regression analysis is a statistical technique used to investigate the possible relationship between quantitative variables. This relationship is postulated between a single response variable on the one hand and a set of predictor variables on the other, and is described by means of a regression model. Since the postulated relationship between the response variable and the set of predictor variables in a regression analysis is of a statistical nature, the regression model used to describe this relationship is probabilistic. Such a probabilistic model, and therefore also a regression model, includes two components: a deterministic and a random component. The deterministic component consists of a mathematical function, which describes the way in which the expected response varies as a function of the set of predictor variables. The random component accounts for deviation of the response from its expected value. This deviation is in part attributable to incomplete or incorrect specification of the set of predictor variables, i.e. the response variable is influenced by other factors than those represented in the set of predictor variables.

A regression model that gives a satisfactory description of the relationship between the response variable and the set of predictor variables can be considered, in a certain sense, to be the final product of a regression analysis. Such a regression model can be utilized to evaluate the strength of the relationship between the response variable and the individual predictor variables, and to predict the response variable at given values of the predictor variables.

Of importance, for the purpose of this dissertation, is the case where the expected value of the response variable is modeled as a linear function of more than one numerical (as opposed to categorical) predictor variable. We therefore give a short overview of this so-called multiple linear regression model and its assumptions.

1.1.1 The multiple linear regression model and its assumptions

Consider a random response variable Y (also called the dependent variable), and a set of m predictor variables, denoted by $x_1, ..., x_m$. These predictor variables, also referred to as independent variables, explanatory variables or regressors, are assumed non-random and their values determined beforehand. In cases where one or more of the regressors are in fact random, we view our analysis as being conditional on given values of these variables. One of the simplest ways of describing the dependence of Y on $x_1, ..., x_m$ is by means of a *multiple linear regression model*. This posits the expected value of Y to be a linear function of $x_1, ..., x_m$.

Mathematically the model is written as

$$Y = \beta_0 + \sum_{j=1}^{m} \beta_j x_j + \varepsilon. \tag{1.1}$$

In (1.1), $\beta_1, ..., \beta_m$ are unknown parameters, referred to as regression coefficients. The regression coefficient β_j reflects the importance of the jth predictor variable in determining the value of the response. The unknown parameter β_0 is known as the intercept and by setting its value equal to zero, the regression model to be fit is forced through the origin. Note that Y is linear in terms of the regression coefficients and consequently the model in (1.1) is referred to as the multiple *linear* regression model. The random or error component, ε, accounts for deviation of Y from the deterministic component, $\beta_0 + \sum_{j=1}^{m} \beta_j x_j$. This deviation is caused either by the randomness of Y itself, or by factors influencing Y which are not included in the set of m predictor variables, or both. It is assumed that ε is a continuous random variable with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2 > 0$. Frequently, it is also assumed that ε is normally distributed. This then implies that Y is also continuous, distributed normally with $E(Y) = \beta_0 + \sum_{j=1}^{m} \beta_j X_j$ and $Var(Y) = \sigma^2$. Under the assumption that ε is normally distributed, (1.1) is known as the *normal multiple linear regression model.*

Suppose now that the multiple linear regression model in (1.1) gives a satisfactory description of the relationship between Y and the set of m predictor variables. In order to utilize this regression model to predict Y at a set of given values of $x_1, ..., x_m$, or to identify among the m predictor variables those significantly influencing Y, the unknown regression coefficients in (1.1) need to be estimated. For this purpose a regression sample is required, where each sample element includes an observation of Y and a corresponding set of predetermined values of $x_1, ..., x_m$. Suppose a random regression sample of n such cases, where n exceeds m, is available. Let Y_i denote the ith random sample element of Y, and $x_{i1}, ..., x_{im}$ the corresponding ith set of predetermined values of the m predictor variables. From (1.1) we can write

$$Y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \varepsilon_i, \qquad i = 1, ..., n, \tag{1.2}$$

where ε_i denotes the ith random error variable.

It is convenient to express (1.2) as follows in matrix notation:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{1.3}$$

In (1.3), $\mathbf{Y} = [Y_1, Y_2, ..., Y_n]'$ is an $n \times 1$ random response vector, and $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_m]'$ is an $(m+1) \times 1$ vector containing the intercept parameter β_0 and the m regression coefficients. The matrix $\mathbf{X}$ (also referred to as the design matrix) in (1.3) is an $n \times (m+1)$ matrix of predetermined values of the m predictor variables, i.e.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & & & & \vdots \\ 1 & x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & \vdots & \vdots & & & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nm} \end{bmatrix} = [\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m]. \tag{1.4}$$

The first column of $\mathbf{X}$ contains elements all equal to 1 in order to provide for the intercept parameter β_0. For all $i = 1, ..., n$ and all $j = 1, ..., m$, the element in the ith row and jth column of $\mathbf{X}$, x_{ij}, represents the ith predetermined value of the jth predictor variable, x_j. We assume that the set of m predictor variables in $\mathbf{X}$ does not exclude important ones with a significant influence on the response variable, but that it may include redundant ones which will hopefully be omitted from the final regression model when a variable selection technique is applied. We further assume that the columns of $\mathbf{X}$ are linearly independent, implying that $\mathbf{X}$ is of full rank $m+1$, and that $\mathbf{X}'\mathbf{X}$ is a positive definite matrix.

In (1.3), $\boldsymbol{\varepsilon} = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_n]'$ is an $n \times 1$ vector of random error variables. We assume these random error variables to be independently normally distributed with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2 > 0$. This implies that the random response vector, $\mathbf{Y}$, is a vector of independently normally distributed random variables, i.e. $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$. The assumption that $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ is not valid in all practical applications of multiple linear regression. In such cases, assuming $\boldsymbol{\varepsilon}$ differently distributed will be appropriate. Carroll and Ruppert (1988) provide techniques, such as transformation and weighting, to deal with situations where the normality assumption is violated. Many other non-parametric procedures (see for example Hastie and Tibshirani, 1990), robust procedures (see for example Rousseeuw and Leroy, 1987) and asymptotic procedures (see for example Arnold, 1981, Chapter 10) have also been developed in this regard. These procedures will not be considered here since our main focus will be on identification of selection influential data cases, under the classical assumption that $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$.

In general, we assume that the regression sample is of reasonable size, in particular that $n > m$. We also assume that the sample is acquired with sufficient accuracy, so that it is unnecessary to provide for measurement errors in (1.3).

Finally, note that the multiple linear regression model, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, can be viewed as a special case of the standard normal linear model given by

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} \tag{1.5}$$

where $\boldsymbol{\mu} = [\mu_1, \mu_2, ..., \mu_n]'$, an $n \times 1$ unknown parameter vector is assumed to belong to M, a known $(m+1)$-dimensional linear subspace of $\mathcal{R}^n$. In order to see this, let the linear subspace M be spanned by the column vectors, $\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m$, of the the design matrix $\mathbf{X}$ in (1.4). Therefore, $\mathbf{X}$, an $n \times (m+1)$ matrix of full rank $m+1$, is a basis matrix for M. Since $\mathbf{X}$ is a basis matrix for M, the vector $\boldsymbol{\mu} \in M$ can be expressed in the form $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ in exactly one way, where $\boldsymbol{\beta} \in \mathcal{R}^{m+1}$ is also a unique vector containing the intercept parameter β_0, and regression coefficients $\beta_1, ..., \beta_m$. Since $\mathbf{X}$ is a basis matrix, the linear subspace M is coordinatized, so that (1.3), obtained from (1.5) by selecting $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, is referred to as the coordinatized version of the standard normal linear model. See Arnold (1981, Chapter 4) for more details in this regard.

By not committing oneself to a specific basis for the linear subspace M, but rather working within the coordinate free framework implied by (1.5), all the results which are derived gain in generality. Although the focus in this dissertation is on variable selection in multiple linear regression, some results will therefore be applicable within a wider linear model context than merely the multiple linear regression model.

There are other arguments in favour of using the coordinate free approach (see Arnold, 1981, p. 55). For example, contrary to the case in multiple linear regression , a natural basis matrix for the subspace M does not always exist. If, for example, the analysis of variance problem is considered in terms of the standard linear model, there is often no natural basis matrix for

M. In such cases it is appealing to use the coordinate free version of the linear model. In the remainder of our work we will move between the coordinate free linear model, and the so-called coordinatized version. The latter case will be characterized by the assumption of a specific basis for M, usually summarised in a design matrix $\mathbf{X}$. In any case, even if results are derived by using a coordinate free approach, application of these results in the multiple linear regression model will always be indicated.

Assume now that $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. In order to estimate $\boldsymbol{\mu}$ from the available regression sample, we only need to estimate the vector $\boldsymbol{\beta}$, since the elements of $\mathbf{X}$ are predetermined values of the m predictor variables. In the next section we consider estimation of the vector $\boldsymbol{\beta}$ and the error variance, σ^2.

1.1.2 Estimating the parameters in the multiple linear regression model

Suppose a regression sample of size n is available for estimating the parameters of the multiple linear regression model. The method of least squares is popularly used to estimate $\boldsymbol{\beta}$, entailing minimization of $\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \|^2$ with respect to $\boldsymbol{\beta}$. Using standard calculus, we obtain the $m+1$ so-called normal equations

$$\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y},$$

where $\widehat{\boldsymbol{\beta}}$ is the $\boldsymbol{\beta}$ minimising $\| \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \|^2$. From these normal equations, the least squares estimator, $\widehat{\boldsymbol{\beta}} = \left[\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_m\right]'$, of $\boldsymbol{\beta}$ follows as

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{1.6}$$

and under the assumptions of the normal multiple linear regression model, it is easy to show that $\widehat{\boldsymbol{\beta}}$ is normally distributed with $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, and variance-covariance matrix equal to $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. The least squares estimator of $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu}$ follows from (1.6) as

$$\begin{aligned} \widehat{\boldsymbol{\mu}}(M) &= [\widehat{\mu}_1(M), \widehat{\mu}_2(M), ..., \widehat{\mu}_n(M)]' \\ &= \left[\widehat{Y}_1(M), \widehat{Y}_2(M), ..., \widehat{Y}_n(M)\right]' \\ &= \mathbf{X}\widehat{\boldsymbol{\beta}} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \end{aligned} \tag{1.7}$$

For reasons which will become clear later, the notation in (1.7) emphasises the dependence of the estimators on the subspace M.

Note that, since $\mathbf{X}$ is a basis matrix for the linear subspace M, the orthogonal projection of $\mathbf{Y}$ on M is given by

$$P_M\mathbf{Y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{1.8}$$

which equals the right-hand side of (1.7). The matrix $P_M = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, is referred to as the projection matrix with respect to M. It is an $n \times n$ symmetric, idempotent matrix. Note that $P_M\mathbf{Y}$ belongs, as does $\boldsymbol{\mu}$, to the linear subspace M. Provided, once again, that the assumptions of the normal multiple linear regression model are satisfied, the least squares estimator in (1.7) is normally distributed with $E\left(\widehat{\boldsymbol{\mu}}(M)\right) = \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu}$ and variance-covariance matrix given by $\sigma^2 P_M$. The estimator $P_M\mathbf{Y}$ has the attractive property of being the minimax estimator of $\boldsymbol{\mu}$ with respect to squared error loss, since $\max\limits_{\boldsymbol{\mu}} E\left\|P_M\mathbf{Y} - \boldsymbol{\mu}\right\|^2 \leq \max\limits_{\boldsymbol{\mu}} E\left\|\xi(\mathbf{Y}) - \boldsymbol{\mu}\right\|^2$ for any estimator $\boldsymbol{\xi}(\mathbf{Y})$ of $\boldsymbol{\mu}$. The corresponding minimax risk is given by

$$E\left\|P_M\mathbf{Y} - \boldsymbol{\mu}\right\|^2 = \sigma^2 \dim(M) = \sigma^2(m+1). \tag{1.9}$$

$P_M\mathbf{Y}$ in (1.8) is also, if geometrically interpreted, the closest point in M to $\mathbf{Y}$, and therefore minimizes the squared distance between $\mathbf{Y}$ and M.

Turning to estimation of the error variance, σ^2, we find that it is unbiasedly estimated by

$$\begin{aligned} \widehat{\sigma}^2 &= \frac{\left\|\mathbf{Y} - P_M\mathbf{Y}\right\|^2}{n-(m+1)} \\ &= \frac{\left\|P_{M^\perp}\mathbf{Y}\right\|^2}{n-(m+1)} \\ &= \frac{\sum_{i=1}^n \left(Y_i - \widehat{Y}_i(M)\right)^2}{n-(m+1)} \end{aligned} \tag{1.10}$$

where $M^\perp$ is the orthogonal complement of M. Also note that $\sum_{i=1}^n \left(Y_i - \widehat{Y}_i(M)\right)^2$ is the error sum of squares. The standard result, $\widehat{\sigma}^2 \sim \frac{\sigma^2}{n-(m+1)}\chi^2_{n-(m+1)}$ is easily established.

Maximum likelihood estimators can also be derived for $\boldsymbol{\beta}$, $\boldsymbol{\mu}$ and σ^2. Under the assumption that $\boldsymbol{\varepsilon} \backsim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, the log-likelihood function for the multiple linear regression model is

$$\ln L(\mathbf{Y}, \mathbf{X}, \boldsymbol{\beta},\sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}\left\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right\|^2. \tag{1.11}$$

Maximizing (1.11) partially with respect to each of the β_j, we obtain the same estimator of $\boldsymbol{\beta}$, and consequently of $\boldsymbol{\mu}$, as the least squares estimators in (1.6) and (1.7). If (1.11) is maximized with respect to σ^2, the maximum likelihood estimator for σ^2 is found to be

$$\widetilde{\sigma}^2 = \frac{\| \mathbf{Y} - P_M\mathbf{Y} \|^2}{n}$$

which is a biased estimator. These least squares and maximum likelihood estimators of $\boldsymbol{\beta}$, $\boldsymbol{\mu}$ and σ^2, and their properties, are discussed further in Arnold (1981, Chapter 5 and 6).

1.2 Variable selection in multiple linear regression

In multiple linear regression the value of the jth regression coefficient, β_j, reflects the importance of the jth predictor variable, i.e. it reflects the influence that the jth predictor variable, x_j, has on the response variable. In other words, if β_j is zero (or close to zero) this typically implies that the influence of x_j on $\mathbf{Y}$ is small, or omissible, so that x_j is in a certain sense redundant in the regression analysis. A value of β_j significantly away from zero, on the other hand, indicates that x_j has a substantial influence on $\mathbf{Y}$. Therefore, certain predictor variables are more important than others in determining the value of the response, and this is reflected in the values of the β_j's. The predictor variables with corresponding regression coefficients equal to zero (or close to zero) should be omitted from the regression model, whereas important ones should be retained. Since the regression coefficients are unknown, a decision on the inclusion or exclusion of predictor variables from the regression model is therefore a rather uncertain matter. The process of deciding which predictor variables to include in the regression model, and which to exclude, is known as *variable selection*.

Variable selection is often one of the first steps in a multiple linear regression analysis, and it is not surprising that a vast literature on the subject is available. Many different procedures or techniques for selecting variables have therefore been developed through the years. These techniques are broadly divided into four categories, namely: step-wise techniques (see Miller, 2002), Bayesian techniques (an extensive list of references are given in Burnham and Anderson, 1998, p.127), cross-validation selection techniques (see Liu et al., 1999) and all possible subsets techniques (see Snyman, 1994, Chapters 2 and 3). Murtaugh (1998) evaluates the performance of several of these selection techniques. Our attention in this dissertation is however restricted to a selection technique based on the all possible subsets approach.

Selection of a subset of predictor variables to be included in the regression model affects estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$. *Firstly*, consider the effect of variable selection on estimation of $\boldsymbol{\beta}$. In the full model, i.e. if the full set of m predictor variables is included in the regression model, $\boldsymbol{\beta}$ is unbiasedly estimated by its least squares estimator in (1.6). In a reduced model, i.e. if a subset of predictor variables is selected for inclusion in the regression model, the least squares estimator of $\boldsymbol{\beta}$ is obtained as in (1.6), but now, together with the vector $\mathbf{1}$ to provide for the intercept, the selected predictor variables comprise the column vectors of the design matrix. Hocking (1974) shows that the variances of the estimated regression coefficients in the full model are always larger than the corresponding variances of the estimated regression coefficients in a reduced model. However, the estimators of the regression coefficients in the reduced model are biased, unless the regression coefficients corresponding to the omitted predictor variables are zero, or the subset of retained variables is orthogonal to the subset of omitted variables. Hocking (1974) also shows that, if the regression coefficients of the omitted variables are smaller than the standard deviations of their corresponding estimators, then the variances of the unbiased estimated regression coefficients in the full model are larger than the corresponding mean squared errors of the biased estimated regression coefficients in the reduced model.

Secondly, the effect of variable selection on estimation of $\boldsymbol{\mu}$ is very similar. The unbiased estimator of $\boldsymbol{\mu}$, obtained from the full model, is given in (1.7). Again, in the reduced model the elements of $\boldsymbol{\mu}$ are biasedly estimated, but with smaller variance than when the full model is used. Also, if the regression coefficients of the omitted variables are smaller than the standard deviations of their corresponding estimated regression coefficients, then the variances of the unbiased estimators of the elements in $\boldsymbol{\mu}$, in the full model, are larger than the corresponding mean squared errors of the biased estimators of these elements in the reduced model (see also

Hocking, 1974, in this regard). Failure to omit such predictor variables (i.e., those whose regression coefficients are smaller than the standard deviations of their estimated regression coefficients), therefore leads to a loss of precision in estimation of the elements of $\boldsymbol{\mu}$.

It is clear from the preceding discussion that, although $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ are usually biasedly estimated when only a subset of predictor variables is included in the regression model, their corresponding elements may frequently be estimated with greater precision. More accurate estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$ is therefore an important reason for applying variable selection in multiple linear regression. However, there are other reasons as well. The cost involved, for example, in obtaining values of certain predictor variables may be high. It may therefore be preferable to exclude these costly variables from the regression model. Many other economical and also practical reasons for reducing the initial set of m predictor variables to a smaller, more manageable, subset are given in Linhart and Zucchini (1986, pp. 2,111) and Miller (2002). Our main concern, however, is estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$. Note that once an appropriate estimator of $\boldsymbol{\beta}$ is obtained the estimator of $\boldsymbol{\mu}$ is fully described. Hence, our focus will be on increasing the accuracy of estimation of $\boldsymbol{\mu}$ by applying variable selection in multiple linear regression. We elaborate on this point in the next paragraph.

It was indicated earlier that the least squares estimator, $P_M\mathbf{Y}$, of $\boldsymbol{\mu}$ has certain desirable properties. Among others, it estimates $\boldsymbol{\mu}$ unbiasedly. Now consider a *fixed* subspace L of M with $\dim(L) = l + 1 < m + 1$. In a multiple linear regression context, a subset of the columns of $\mathbf{X}$ forms a basis for L. If $\boldsymbol{\mu}$ actually belongs to L, then $P_L\mathbf{Y}$ will also estimate $\boldsymbol{\mu}$ unbiasedly. Moreover, in terms of mean squared error, $P_L\mathbf{Y}$ will then be a better estimator of $\boldsymbol{\mu}$ than $P_M\mathbf{Y}$ since

$$\begin{aligned} E\left\|P_L\mathbf{Y} - \boldsymbol{\mu}\right\|^2 &= (l+1)\sigma^2 \qquad \text{(provided } \boldsymbol{\mu} \in L\text{)} \\ &< (m+1)\sigma^2 \\ &= E\left\|P_M\mathbf{Y} - \boldsymbol{\mu}\right\|^2 . \end{aligned}$$

Typically $\boldsymbol{\mu}$ will of course not belong to L, and then it becomes more difficult to choose between the two estimators, $P_L\mathbf{Y}$ and $P_M\mathbf{Y}$. In this more frequently occurring case, the mean squared error of $P_L\mathbf{Y}$ becomes

$$\begin{aligned} E\left\|P_L\mathbf{Y}-\boldsymbol{\mu}\right\|^2 &= E\left\|P_L\left(\mathbf{Y}-\boldsymbol{\mu}\right)-P_{L^\perp}\boldsymbol{\mu}\right\|^2 \\ &= E\left\|P_L\left(\mathbf{Y}-\boldsymbol{\mu}\right)\right\|^2+\left\|P_{L^\perp}\boldsymbol{\mu}\right\|^2 \\ &= \dim(L)\sigma^2+\left\|P_{L^\perp}\boldsymbol{\mu}\right\|^2 \\ &= (l+1)\sigma^2+\left\|P_{L^\perp}\boldsymbol{\mu}\right\|^2. \end{aligned}$$

The final term in this expression, $\left\|P_{L^\perp}\boldsymbol{\mu}\right\|^2$, is the squared bias resulting from the fact that $P_L\mathbf{Y}\in L$, while $\boldsymbol{\mu}\notin L$. It is clear that in this case $P_L\mathbf{Y}$ will be a better estimator of $\boldsymbol{\mu}$ than $P_M\mathbf{Y}$, if and only if

$$(l+1)\sigma^2+\left\|P_{L^\perp}\boldsymbol{\mu}\right\|^2<(m+1)\sigma^2 \Longleftrightarrow \left\|P_{L^\perp}\boldsymbol{\mu}\right\|^2<(m-l)\sigma^2. \tag{1.12}$$

This condition has the following interpretation: $P_L\mathbf{Y}$ will be a better estimator of $\boldsymbol{\mu}$ than $P_M\mathbf{Y}$ if and only if the squared bias resulting from the fact that $\boldsymbol{\mu}\notin L$, is less than the increase in variance when we move from $P_L\mathbf{Y}$ to $P_M\mathbf{Y}$. This is a classic illustration of the bias versus variance trade-off.

In a multiple linear regression context, the fixed linear subspace L of M with $\dim(L)=l+1$ is spanned by the vector $\mathbf{1}$ and a subset of the columns $\mathbf{x}_1,...,\mathbf{x}_m$ of $\mathbf{X}$. The subspace L is therefore associated with a particular subset of predictor variables. This subset of predictor variables, together with the vector $\mathbf{1}$, form the column vectors of the design matrix associated with L. Let this $n\times(l+1)$ matrix be denoted by $\mathbf{X}_L$. The condition in (1.12) will be satisfied if the values of the regression coefficients corresponding to those predictor variables not included in $\mathbf{X}_L$, are zero (or close to zero).

The column vectors of $\mathbf{X}_L$ form a basis for the linear subspace L. The least squares estimator of $\boldsymbol{\mu}$ that corresponds with L, i.e.

$$\begin{aligned} P_L\mathbf{Y} &= \left[\widehat{\mu}_1(L),\widehat{\mu}_2(L),...,\widehat{\mu}_n(L)\right]' \\ &= \left[\widehat{Y}_1(L),\widehat{Y}_2(L),...,\widehat{Y}_n(L)\right]' \end{aligned}$$

is obtained from (1.7), with $\mathbf{X}_L$ replacing $\mathbf{X}$, i.e.

$$P_L\mathbf{Y}=\mathbf{X}_L(\mathbf{X}_L'\mathbf{X}_L)^{-1}\mathbf{X}_L'\mathbf{Y}. \tag{1.13}$$

Note that in obtaining the estimator $P_L\mathbf{Y}$, fewer regression coefficients need to be estimated than in obtaining the estimator $P_M\mathbf{Y}$. This implies that the total variance associated with $P_L\mathbf{Y}$ is smaller than the total variance associated with $P_M\mathbf{Y}$.

The all possible subsets approach to variable selection in multiple linear regression considers the family of all linear subspaces L in M. Let $\mathcal{L}$ denote this family of linear subspaces in M, spanned by subsets of the column vectors of $\mathbf{X} = [\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m]$. Using the all possible subsets approach, our aim is to identify a member of $\mathcal{L}$, or equivalently, a linear subspace L in M, that satisfies the condition in (1.12). However, this condition may be satisfied by several subspaces in M. Therefore, the linear subspace L, among those not violating (1.12), whose corresponding least squares estimator, $P_L\mathbf{Y}$, estimates $\boldsymbol{\mu}$ most accurately should be identified. This implies that accurate estimation of $\boldsymbol{\mu}$ depends on whether an appropriate linear subspace $L \in \mathcal{L}$ is selected. Note that the bias versus variance trade-off phenomenon when $\boldsymbol{\mu}$ is estimated by $P_L\mathbf{Y}$, rather than by its traditional estimator, $P_M\mathbf{Y}$, is referred to by Burnham and Anderson (1998, p.23) within the context of the principle of parsimony. Application of this principle, in an all possible subsets approach, leads to parsimonious, and accurate estimation of $\boldsymbol{\mu}$.

In practical applications of variable selection, it is impossible to verify condition (1.12) since $\boldsymbol{\mu}$ and σ^2 are unknown parameters. Consequently, the selection of a linear subspace has to be based on the available regression sample. A frequently implemented approach is therefore to use the sample data to estimate

$$E\left\|P_L\mathbf{Y} - \boldsymbol{\mu}\right\|^2 = (l+1)\sigma^2 + \left\|P_{L^\perp}\boldsymbol{\mu}\right\|^2$$

for every $L \in \mathcal{L}$. Let $\widehat{L} \equiv \widehat{L}(\mathbf{Y})$ denote the subspace having the minimal estimated mean squared error, i.e.

$$E\left\|\widehat{P_{\widehat{L}}\mathbf{Y} - \boldsymbol{\mu}}\right\|^2 = \min_{L\in\mathcal{L}}\left\{E\left\|\widehat{P_L\mathbf{Y} - \boldsymbol{\mu}}\right\|^2\right\}.$$

Then $\widehat{L}$ is the data-dependent subspace of choice, hopefully providing parsimonious and accurate estimators of the unknown quantities in the model. Note that in a multiple linear regression context we can write $\widehat{L} = span\{\mathbf{1}, \mathbf{x}_j : j \in J_{\widehat{L}}\}$, where $J_{\widehat{L}} \subset \{1, 2, ..., m\}$ is the subset of indices corresponding to the column vectors of $\mathbf{X}$ that, together with the vector $\mathbf{1}$, span $\widehat{L}$. The least squares estimator of $\boldsymbol{\mu}$ corresponding to $\widehat{L}$ is given by $P_{\widehat{L}}\mathbf{Y}$. This estimator has the attractive property that $P_{\widehat{L}}\mathbf{Y} \in \widehat{L}$, thereby reflecting our belief that $\boldsymbol{\mu} \in \widehat{L}$. Also, we may write

$$P_{\widehat{L}}\mathbf{Y} = \widehat{\beta}_0\mathbf{1} + \sum_{j \in J_{\widehat{L}}} \widehat{\beta}_j \mathbf{x}_j, \tag{1.14}$$

the implication being that $\widehat{\beta}_j = 0$ for all $j \notin J_{\widehat{L}}$. Selecting $\widehat{L}$ therefore describes the estimator of $\boldsymbol{\mu}$ completely, and simultaneously identifies those predictor variables thought to have a significant influence on the response. Note that, although the condition in (1.12) does not necessarily hold for the data-dependently selected subspace $\widehat{L}$, the motivation to estimate $\boldsymbol{\mu}$ by $P_{\widehat{L}}\mathbf{Y}$, rather than by $P_M\mathbf{Y}$, is still valid in these practical applications of variable selection. Thus, by selecting $\widehat{L}$, a parsimonious and hopefully more accurate estimator of $\boldsymbol{\mu}$ is obtained.

Finally, we show that accurate estimation of $\boldsymbol{\mu}$ amounts to accurate response prediction. Suppose a future observation $\mathbf{Y}^*$, with the same structure as $\mathbf{Y}$, needs to be predicted. Let $\mathbf{Y}^* = \boldsymbol{\mu} + \boldsymbol{\varepsilon}^*$, where $\boldsymbol{\varepsilon}^* \backsim N_n(\mathbf{0}, \sigma^2\mathbf{I})$, independently of $\boldsymbol{\varepsilon}$. The corresponding predictive risk when $\xi(\mathbf{Y})$ is used as a predictor for $\mathbf{Y}^*$ is given by $E\left\|\xi(\mathbf{Y}) - \mathbf{Y}^*\right\|^2$. Snyman (1994, p.1-6) shows that

$$E\left\|\xi(\mathbf{Y}) - \mathbf{Y}^*\right\|^2 = n\sigma^2 + E\left\|\xi(\mathbf{Y}) - \boldsymbol{\mu}\right\|^2. \tag{1.15}$$

Since $n\sigma^2$ in (1.15) does not depend on $\boldsymbol{\mu}$, it is clear that finding a low risk when predicting a new observation implies a low mean squared error when $\xi(\mathbf{Y})$ is used as estimator for $\boldsymbol{\mu}$.

1.3 Influential data cases

Studying the influence of individual or small groups of data cases on the results of an analysis of a data set is an important area in statistics. The influence which data cases have on the results of a statistical analysis is often quantified in terms of an influence measure, calculated from the data set at hand. Various such influence measures, with applications in different statistical fields, have been developed through the years. For example, influence measures for data cases in a discriminant analysis can be found in Fung (1992,1994,1995), and Steel and Louw (2001). Also, in time series analysis, contributions appear in Peña et al. (2001, Chapter 6), and Baragona et al. (2001).

The literature on measures of influence in multiple linear regression is comprehensive, and we only briefly refer to some contributions. Cook's distance (Cook, 1977) is, for example, a well-established measure to evaluate the influence of individual data cases in multiple linear regression. Belsley, Kuh and Welsch (1980) also provide an extensive discussion of influence measures for individual data cases in a regression analysis. Atkinson (1994), on the other hand, consider robust influence measures for subsets of data cases. These influence measures were mainly developed to detect effects such as masking and swamping in multiple linear regression (see also Van Vuuren (1998), Barrett and Gray (1997), Rancel and Sierra (2000), and Wisnowski et al. (2001) in this regard). Another contribution is that of Li, Martin and Morris (2001), who proposed a graphical technique for detecting influential cases in regression analysis.

Studying the influence of individual data cases on the results of a statistical analysis gives us a deeper understanding of the underlying structure of the relationships amongst the variables represented in the data. Frequently, a first step in this study entails using an appropriate influence measure to identify data cases having a significant influence on the results of the analysis. Such identified data cases are considered influential and are often referred to as *outliers* in the statistical literature. Once influential data cases have been identified, they can be dealt with in several different ways. Probably the best-known practice is to omit these cases from the data set where after the statistical analysis is repeated on the remaining data cases. Allocating weights to the data cases, in such a manner that influential cases are down-weighted, thereby decreasing their influence on the results of the analysis, is another frequently used procedure to deal with influential cases. Note that allocation of zero-weights to certain data cases is equivalent to omitting such cases from the data set. Many statistical techniques have also been robustified in an attempt to deal with influential data cases. An important reference in the area of robust regression techniques is the book by Rousseeuw and Leroy (1987).

In this dissertation we restrict our attention to the influence of data cases on a multiple linear regression analysis when an initial variable selection technique is applied to the data. An important question to us is therefore: when is a data case considered *selection influential*? An attractive answer seems to be to regard a data case as selection influential if its omission from the data set leads to a different set of predictor variables being included in the selected model, or, if the same set of predictor variables is still selected, the fitted model differing significantly from the model fitted to the complete data set. In particular, suppose a variable selection technique is applied to the complete data set. Once a regression model has been selected, a single data case is omitted from the data set. The same variable selection technique is now applied to the retained data cases, i.e. the reduced data set. If the subset of predictor variables included in the selected model obtained from the complete data set differs from the subset of predictor variables included in the selected model obtained from the reduced data set, the particular data case is obviously selection influential. However, if the same set of variables is selected in both cases, but subsequent results obtained from the two fitted models, for example response predictions, are substantially different, then the omitted data case will also be deemed selection influential. Note that subsequent results obtained from the two models may vary dramatically even though the subsets of predictor variables included in the two models are identical. This may occur as a result of significant differences between the estimated regression coefficients of the corresponding predictor variables in the two fitted models.

Arguing along the same lines, a subset of data cases will be considered selection influential if the subset meets either of the criteria mentioned above for an individual data case to be deemed selection influential. Selection influence measures are measures that can be used to identify such individual and subsets of selection influential data cases. It is essential that these selection influence measures explicitly take an initial variable selection step into account. If this is not done the influence measures are conditional, i.e. given a specific set of predictors in the model. Non-selection influence measures may then just as well be used to identify influential data cases. It is therefore important that selection influence measures are defined unconditionally. In this regard, Léger and Altman (1993) proposed a selection influence measure, based on an *unconditional* selection version of Cook's distance. Their proposal is discussed in greater detail later in this section.

Generally, the influence which data cases have on the selection process and the subsequent results of the regression analysis could be regarded as either positive or negative. Omitting for example certain data cases from the data set may lead to a model with greater predictive power. These cases have a negative influence and should therefore rather be omitted from the data set before variable selection is performed. On the other hand, the retention of certain data cases in the data set may cause the "appropriate" variables to be selected, and may also improve the fit of the model. The influence of such cases could be considered positive, and these cases should obviously be retained. Finally, the influence of data cases may be negative if only a specific subset of predictor variables is considered, while the influence of the same data cases may change to positive once the full set of predictor variables is considered.

Making matters even more complex is the fact that the identification of selection influential cases clearly depends on the specific variable selection technique being used. This implies that data cases may be considered selection influential if, for example, the C_p criterion (see Mallows, 1973) is used for variable selection, but not necessarily if another selection technique is applied.

What can be done once specific data cases have been identified as selection influential? One possible solution is to use a robustified variable selection technique. Since classical variable selection criteria are typically based on least *squares* estimation, they are bound to be more sensitive to outlying cases in the data than selection techniques based on more robust norms. (Ronchetti (1997) shows the extreme sensitivity of many classical model selection techniques.) Most robustified variable selection techniques are therefore not based on projection type least squares estimators, but rather on more robust estimators which try to minimize the effect of influential cases on the results of the analysis. We refer briefly to some of these robust variable selection techniques in the next paragraph.

Ronchetti and Staudte (1994) use a simple artificial setting to illustrate the sensitivity of the C_p criterion to individual outlying cases. In an attempt to diminish this sensitivity the authors develop a robust version of this criterion. The robust version is obtained by using M-estimation rather than the usual least squares estimation to estimate the parameters in the model under consideration. Weights are defined for each of the cases in the data set, so that the robust version of the C_p criterion is based on a weighted sum of squared residuals. Sommer and Staudte (1995) build on the results presented by Ronchetti and Staudte by again replacing least squares estimation by M-estimation, but now a weight function differing from the one in the 1994-paper is used. I east squares estimation is also replaced by suitable M-estimation in a robust version of variable selection based on a cross-validation argument proposed by Ronchetti, Field and

Blanchard (1997), and in a robust version of variable selection based on the so-called Wald test by Sommer and Huggins (1996). Burman and Nolan (1995) present a general Akaike-type criterion, which is applicable not only to least squares modelling, but also to models estimated by using a wide variety of other loss functions. Lastly, a generalisation of the Kullback-Leibler distance is used in the definition of Akaike's model selection criterion by Shi and Tsai (1998). Based on this generalisation, they propose three new robust versions of the Akaike criterion.

Although satisfactory results seem to be obtained when robust selection techniques are applied in multiple linear regression, these techniques will not be considered here. In this dissertation, selection influence measures will be developed for identifying selection influential data cases when a multiple linear regression analysis is preceded by variable selection. Once these cases have been identified, they will be omitted from the data set and the regression analysis, including the initial selection step, will be repeated. A two-step approach to clean the data before a final variable selection step, is therefore followed. The final regression model obtained by using this approach will hopefully now include those predictor variables with genuinely important influence on the response variable. As a result, it is to be hoped that more accurate response predictions should also be provided by this model.

We conclude this section by a short overview of some of the existing contributions on selection influential data cases in a multiple linear regression context.

Weisberg (1981) shows how Mallows' C_p criterion can be written as a sum of n terms. This is done within a regression context, with each term in the sum corresponding to one of the n data cases. The break-up for a particular model, corresponding to a given linear subspace $L \subset M$, is given by

$$C_p = \sum_{i=1}^{n} C_{pi} = \sum_{i=1}^{n} \left(\widehat{\sigma}^2 v_{ii} + \left(\widehat{Y}_i(M) - \widehat{Y}_i(L) \right)^2 - \widehat{\sigma}^2 \left(u_{ii} - v_{ii} \right) \right),$$

where

- $\widehat{\sigma}^2$ is the least squares estimator in (1.10) of σ^2
- v_{ii} is the ith diagonal element of $P_L = \mathbf{X}_L(\mathbf{X}_L'\mathbf{X}_L)^{-1}\mathbf{X}_L'$ (i.e. the leverage of the ith case in the reduced regression model)
- $\widehat{Y}_i(M)$ is the ith predicted value obtained from $P_M\mathbf{Y}$ in (1.8)

- $\widehat{Y}_i(L)$ is the ith predicted value obtained from $P_L\mathbf{Y}$ in (1.13)
- u_{ii} is the ith diagonal element of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ (i.e. the leverage of the ith case in the full regression model)

Weisberg states that the size of the ith C_p-term, C_{pi}, of a given reduced model should be evaluated in terms of v_{ii}. The application of this proposal is illustrated in an example. In the example specific points which seem to be influential with respect to specific models are identified. Weisberg stops short, however, of recommending what, if anything, should be done about such points. In Chapter 2 of this dissertation we return to this aspect and give a detailed discussion of the expansion of the C_p criterion as the sum of n terms. We also show how this break-up can be done within a coordinate free context, thereby enhancing its applicability.

Ahn and Park (1987) also consider the C_p statistic in the form derived by Weisberg (1981). Different possibilities regarding the manner in which individual terms in this representation of C_p can be weighted, are investigated. The intention is to down-weight influential cases. Such weighting of the individual cases gives rise to a new variable selection criterion, a so-called weighted C_p criterion. Their proposed procedure therefore considers variable selection and detection of influential data cases simultaneously. Following the application of their proposed ideas in an illustrative example, Ahn and Park conclude with interesting reasons why their integrated approach of combining variable selection and identification of influential cases is desirable above the two-step approach to clean the data before variable selection is undertaken. One of these reasons is that the two-step approach identifies data cases as influential, *conditional* on the model containing all predictor variables, i.e. the full model. A data case may therefore seem influential as a result of extreme values in predictors that "do not really matter", i.e. predictors that do not show a significant relationship with the response variable and will typically not be selected.

Chatterjee and Hadi (1988) investigate the effect of simultaneously omitting a variable and an observation from the data set. This effect is measured in terms of changes in the least squares regression coefficients, the residual sum of squares, the fitted values, and the predicted value of the omitted observation. The authors state that a study of the statistic which they propose can enable one to identify situations in which a single data case is responsible for retaining or eliminating a specific variable. The example illustrating the application of their statistic clearly shows how influential a single data case can be with respect to the retention or elimination of a given variable.

Peixoto and Lamotte (1989) propose a procedure where a dummy variable is added for each observation in the data set, and variable selection is then performed. The dummy variables which are selected identify outliers in the data. Simultaneously, important predictor variables are identified, taking the influence of possible outliers into account. The intention with the paper is therefore to investigate the use of variable selection to identify predictors and outliers simultaneously. It emerges from an extensive discussion of an example in the paper that remaining problems with the proposed technique are specification of the number of outliers in the data, and specification of the number of predictor variables that should be included in the model. Answers to these questions are still open problems.

Léger and Altman (1993) propose the following steps to identify selection influential data cases:

- Choose a variable selection technique to be applied (for example C_p selection)
- Apply the chosen selection technique to the full data set, select variables and find the vector of fitted values
- Repeat the previous selection step, but now with the ith case omitted
- Calculate a standardised distance between the two vectors of fitted values to measure the influence of the omitted case.

A so-called unconditional influence measure (referred to as Cook's unconditional distance) is obtained if the variable selection is repeated after a data case was omitted from the data set. This is in contrast to a so-called conditional measure (referred to as Cook's conditional distance) when all calculations use the same model, i.e. the model selected from the full data set. The unconditional influence measure is preferable as stated in the final paragraph of the paper: "Variable selection in regression is one of the most used statistical techniques. Although the estimation aspect of this technique has been studied extensively, assessment of influence has always been done conditionally on the selected model due to the difficulty of incorporating the selection process. As shown here, assessment of influence can be done satisfactorily, and use of unconditional Cook's distance helps in understanding the data and in choosing the model."

In Chapter 4 of this dissertation we also propose an unconditional measure for identifying individual influential data cases.

Kim and Park (1995) propose graphical techniques which can be used to show the joint effect of deleting predictors and removing data cases from a regression model. The following situations are considered:

- where an observation is deleted after a variable has been deleted
- where a variable is deleted after an observation has been omitted
- where multiple observations are removed after deleting multiple variables
- where multiple variables are deleted after multiple observations have been omitted.

The graphical displays for the above situations clarify the interrelationship among variables and data cases, and thus give a better understanding of the roles of variables and observations in a regression model.

Gupta and Hang (1996) initially develop new measures of influence of individual data points when no variable selection is done. These measures focus on the change in the residuals when an observation is omitted from the data set. The authors continue by applying a selection technique which was proposed by them in a 1988 paper. They show how this criterion can be applied while taking the potential effect of an individual data point into account.

Hoeting, Raftery and Madigan (1996) point out that the model which is selected can depend upon the order in which variable selection and outlier identification are carried out. A method is thus proposed where variable selection and identification of outliers are combined. In this regard a Bayesian approach is proposed to simultaneously select variables and identify outliers. They state that this approach can identify multiple outliers, and that is also successful in dealing with masking effects. The authors use equal prior probabilities for all possible models, and a variance-inflation model to provide for outliers. They express a preference for complete Bayesian model averaging, but state that it is also possible to evaluate the adequacy of any given model in terms of its posterior probability. Outliers are identified based on so-called outlier posterior probability. This is defined for a given observation to be the sum of the posterior model probability across models in which the observation was classified as an outlier. The paper also contains a section on Bayesian model averaging via simultaneous variable selection and outlier identification. In this section of the paper, a Markov Chain Monte Carlo technique which can be used to approximate the model averaging process is described.

Kim and Hwang (2000) derive an expression for the C_p criterion when k data cases are omitted, in terms of quantities calculated from the fit of the full model and the model under consideration, using all the data. It is clear from the two examples that they examine that omitting certain data cases sometimes changes the selected model and sometimes not.

1.4 Intention, restrictions and outline of the dissertation

The main concern of this dissertation is identification of selection influential data cases when variable selection is applied in multiple linear regression. Once such cases have been identified, these cases are omitted from the data set, and the selection process is repeated on the reduced data set. By using such an approach, we hopefully identify those predictor variables with a significant influence on the response variable. Reducing the initial set of predictor variables to a smaller, more manageable subset in this way, hopefully also results in accurate, yet parsimonious estimation of $\boldsymbol{\mu}$.

We narrow our field of study by making the following additional assumptions:

- The variable selection technique applied to the normal multiple linear regression model, throughout the dissertation, is the all possible subsets approach based on the C_p criterion of Mallows (1973, 1995).

- We assume that no subjective considerations, such as professional judgement, is involved when applying this selection technique to any data set.

- The design matrix $\mathbf{X}$ in the multiple linear regression model, which is of full column rank $m + 1 < n$, is assumed to contain all important and possibly some redundant predictor variables.

- Also, although we do consider the case where $\mathbf{X}$ is orthogonal, our attention will primarily be focused on the non-orthogonal case.

- Whenever the error variance, σ^2, is assumed to be known, its value equals 1, without loss of generality. If σ^2 is assumed unknown, it is unbiasedly estimated by

$$\widehat{\sigma}^2 = \frac{\|P_{M^\perp}\mathbf{Y}\|^2}{n-(m+1)}$$

provided $m+1<n$.

The layout of the remaining text of the dissertation is as follows:

- **Chapter 2** mainly focuses on Mallows' C_p statistic. Theoretical developments based on this selection technique are presented using the coordinate free approach. These results are also specifically applied to variable selection within a multiple linear regression context.
- In **Chapter 3** we consider the influence of data cases when variable selection is applied in multiple linear regression. The effect of such selection influential data cases in a multiple linear regression analysis is illustrated by means of example- and simulated data sets.
- **Chapter 4 and Chapter 5** are devoted to the identification of selection influential cases. In Chapter 4 we propose an influence measure for detecting single selection influential data cases, whereas in Chapter 5 the identification of a subset of selection influential data cases is considered. An influence measure is also derived for this purpose. The performance of this measure is illustrated by means of example data sets and simulation.
- In **Chapter 6** we conclude with a discussion of the contribution made in this dissertation. Some limitations of the dissertation and some promising future research options are also pointed out.

Finally, the following Appendices appear at the end of the dissertation:

- **Appendix A** provides some results regarding the coordinate free approach to linear model selection. These results are applied in the main text.
- In **Appendix B** we list the three example data sets which are utilised for illustration purposes in the main text.
- **Appendix C** provides some of the FORTRAN programs used to do the numerical computations which are reported on in the dissertation.

CHAPTER 2

LINEAR MODEL SELECTION AND ESTIMATION

2.1 Introduction

In this chapter we deal with the problems of linear model selection and the estimation of the parameters of the selected model. Section 2.2 deals with these problems in the broader framework of the coordinate free approach introduced in Section 1.1. In Section 2.3 we apply the coordinate free approach to the C_p subspace selection criterion introduced by Mallows (1973). Special attention is given to the fact that the C_p criterion is an unbiased estimator of the mean squared error of $P_L\mathbf{Y}$ as an estimator of $\boldsymbol{\mu}$, i.e. $E\left\|P_L\mathbf{Y}-\boldsymbol{\mu}\right\|^2$. This section also deals with expanding this mean squared error of $P_L\mathbf{Y}$ as the sum of n terms, whereafter estimation of the individual terms in this expansion is considered. Lastly, the results of Section 2.3 are specialised within a multiple linear regression context in Section 2.4.

2.2 A coordinate free approach to linear model selection

In this section we utilise a coordinate free approach to investigate problems which arise during linear model selection and subsequent parameter estimation. Discussing these problems in a coordinate free context offers the advantage that the results which are obtained can be applied in a wider sense to many special cases of the linear model. The scope of the results is therefore broadened, and not restricted to variable selection in multiple linear regression only.

Consider therefore again the coordinate free formulation of the standard normal linear model, viz.

$$\mathbf{Y}=\boldsymbol{\mu}+\boldsymbol{\varepsilon}\sim N_n(\boldsymbol{\mu},\sigma^2\mathbf{I}_n)$$

where $\mathbf{Y}$ is the n-component response vector, $\boldsymbol{\mu}=E(\mathbf{Y})$, and $\boldsymbol{\varepsilon}$ represents the normally distributed error term. It is assumed that $\boldsymbol{\mu}$ belongs to a known linear subspace M of $\mathcal{R}^n$, where $\dim(M)=m+1$. We write $\dim(M)=m+1$, rather than $\dim(M)=m$, to make explicit the consistent inclusion of an intercept term in the multiple linear regression models which are discussed later. Our objective is to estimate $\boldsymbol{\mu}$ accurately, or to predict a future observation of the response vector accurately. These objectives are equivalent, as argued for example by Snyman (1994, p.2-1). We focus in our discussion on the problem of accurate estimation of $\boldsymbol{\mu}$. The least squares estimator, $P_M\mathbf{Y}$, of $\boldsymbol{\mu}$ has several desirable properties, including that of estimating $\boldsymbol{\mu}$ unbiasedly. The mean squared error, or the expected squared error of estimation (ESEE), of $P_M\mathbf{Y}$ is given by

$$E\left\|P_M\mathbf{Y}-\boldsymbol{\mu}\right\|^2=\sigma^2(m+1). \tag{2.1}$$

If nothing more could be assumed about $\boldsymbol{\mu}$, it would be hard to propose any other estimator than $P_M\mathbf{Y}$. The principle of parsimony, however, suggests that one should aspire to use the simplest possible model that satisfactory explains the data. In the light of this principle we therefore search for a lower-dimensional subspace L of M, where the projection of the data vector onto L estimates $\boldsymbol{\mu}$ accurately. Let $\mathcal{L}$ denote the family of all possible such subspaces L of M. Also, denote the dimension of such a typical subspace L by $l+1$. The problem of estimating $\boldsymbol{\mu}$ within this extended context, entails first of all using the data to select a member of $\mathcal{L}$ believed to contain $\boldsymbol{\mu}$, and then estimating $\boldsymbol{\mu}$ accordingly. Selecting a specific linear subspace L, which is presumed to contain $\boldsymbol{\mu}$, amounts to selecting a particular linear model. The terms *linear model* and *linear subspace* will therefore be used interchangeably.

Since the data are used to select the subspace thought to contain $\boldsymbol{\mu}$, the selected subspace will be denoted by $\widehat{L}$, or by $\widehat{L}(\mathbf{Y})$. After having identified $\widehat{L}$, we wish to estimate $\boldsymbol{\mu}$ accordingly by using an estimator which belongs to $\widehat{L}$. The projection estimator, $P_{\widehat{L}}\mathbf{Y}$, meets this requirement, since $P_{\widehat{L}}\mathbf{Y}\in\widehat{L}$. Note that $P_{\widehat{L}}\mathbf{Y}$ is the ordinary least squares estimator of $\boldsymbol{\mu}$ with respect to the subspace $\widehat{L}$ in the sense that $P_{\widehat{L}}\mathbf{Y}$ is the vector in $\widehat{L}$ which is closest (in a least squared sense) to the response vector $\mathbf{Y}$. Although other estimators of $\boldsymbol{\mu}$ may be used after $\widehat{L}$ has been identified, for example a Stein estimator shrinking $\mathbf{Y}$ towards $\widehat{L}$, we will only consider projection type estimators in this dissertation. Note also therefore that once a linear subspace $\widehat{L}$ has been selected, the corresponding estimator, $P_{\widehat{L}}\mathbf{Y}$, of $\boldsymbol{\mu}$ is fully determined.

How should one proceed to identify $\widehat{L}$? Consider a data-independent subspace L of M. The corresponding estimator of $\boldsymbol{\mu}$ is $P_L\mathbf{Y}$, with ESEE

$$E\left\|P_L\mathbf{Y}-\boldsymbol{\mu}\right\|^2=(l+1)\sigma^2+\left\|P_{L^\perp}\boldsymbol{\mu}\right\|^2. \tag{2.2}$$

The first term on the right-hand side of (2.2) is a variance term, while the second term reflects the bias which we incur if we estimate $\boldsymbol{\mu}$ by $P_L\mathbf{Y}$ rather than $P_M\mathbf{Y}$. Of course, the bias will be zero if $\boldsymbol{\mu} \in L$. Obviously we would like to estimate $\boldsymbol{\mu}$ by a $P_L\mathbf{Y}$ for which (2.2) is small. The attractive idea of using that L for which (2.2) is a minimum is impractical since the right-hand side of (2.2) contains unknown quantities. However, these quantities can be estimated, and this possibility gives rise to the following frequently used strategy for identifying $\widehat{L}$: estimate the ESEE of every member L of $\mathcal{L}$, and take $\widehat{L}$ to be the subspace with minimum estimated ESEE. Since M is also considered to be a member of $\mathcal{L}$, it is hoped that the ESEE of the estimator $P_{\widehat{L}}\mathbf{Y}$ which is identified in this way will be less than that of the non-selection estimator $P_M\mathbf{Y}$. We refer to this strategy for identifying $\widehat{L}$ as an *all possible subspace approach*.

An all possible subspace approach to model selection can therefore be implemented by estimating the right-hand side of (2.2) for every L, and then selecting that L having the minimum estimated mean squared error. It should be noted that other strategies can also be used in an all possible subspace approach, for example taking $\widehat{L}$ to be the subspace which minimises a different (estimated) measure of error (and not necessarily the estimated ESEE of $P_L\mathbf{Y}$). Also, when applying an all possible subspace approach in terms of estimated ESEE it is strictly speaking not required to consider all $L \in \mathcal{L}$ for possible selection. Snyman (1994, Section 2.3) argues that only those subspaces which, for a given dimension, are nearest to the observation vector $\mathbf{Y}$, need to be considered. Hence, if we let $\widehat{L}(\mathbf{Y}, l)$ denote the $(l+1)$-dimensional subspace characterised by

$$\left\| P_{\widehat{L}(\mathbf{Y},l)}\mathbf{Y} \right\|^2 = \max\left\{ \|P_L\mathbf{Y}\|^2 : L \in \mathcal{L} \text{ and } \dim(L) = l+1 \right\} \; (l = 0, 1, ..., m),$$

then $\widehat{L}(\mathbf{Y})$ will be selected from $\widehat{L}(\mathbf{Y},0), \widehat{L}(\mathbf{Y},1), ..., \widehat{L}(\mathbf{Y},m)$. This argument considerably reduces the number of models which need be considered: from 2^m to $m+1$. The interested reader is referred to Section 2.3 of Snyman (1994) for more details in this regard.

Numerous other methods, besides an all possible subspace approach, have been proposed in the literature for the purpose of model selection. Many of these methods are based on sequences of hypothesis tests (see Miller, 2002), while others follow a Bayesian approach (see Burnham and Anderson, 1998, p.127). In this dissertation our attention will be restricted to a single selection method, based on an all possible subspace approach, known as C_p selection.

2.3 A coordinate free approach to Mallows' C_p criterion

Proposed by C.L. Mallows (Mallows, 1973), C_p selection is well known, very commonly used in practice and thoroughly investigated over the years (see Spjøtvoll (1977), Mallows (1995) and Chiu (2000) in this regard). It was indicated in the previous section that an all possible subspace approach to model selection requires estimation of the ESEE of $P_L\mathbf{Y}$ for each subspace L. The C_p criterion is obtained if these ESEE's are estimated unbiasedly. An unbiased estimator of (2.2) is obtained as follows. Since

$$E\left\|P_{L^\perp}\mathbf{Y}\right\|^2 = (n-\dim(L))\,\sigma^2 + \left\|P_{L^\perp}\boldsymbol{\mu}\right\|^2,$$

an unbiased estimator of $\left\|P_{L^\perp}\boldsymbol{\mu}\right\|^2$ in (2.2) is given by $\left\|P_{L^\perp}\mathbf{Y}\right\|^2 - (n-\dim(L))\,\widehat{\sigma}^2$, where $\widehat{\sigma}^2$ is the unbiased estimator of σ^2 defined in (1.10). For a fixed subspace L, an unbiased estimator of the ESEE in (2.2) is therefore given by

$$\left\|P_{L^\perp}\mathbf{Y}\right\|^2 + (2\dim(L)-n)\,\widehat{\sigma}^2 = \left\|P_{L^\perp}\mathbf{Y}\right\|^2 + (2(l+1)-n)\,\widehat{\sigma}^2. \tag{2.3}$$

The estimator in (2.3) is referred to as Mallows' C_p criterion for the subspace L, and will be denoted by $C_p(\mathbf{Y},L)$. Note that $\left\|P_{L^\perp}\mathbf{Y}\right\|^2 = \left\|\mathbf{Y}-P_L\mathbf{Y}\right\|^2$ in (2.3) is the residual sum of squares associated with the linear subspace L. In the statistical literature Mallows' C_p is frequently expressed relative to $\widehat{\sigma}^2$, viz.

$$\frac{Cp(\mathbf{Y},L)}{\widehat{\sigma}^2} = \frac{\left\|P_{L^\perp}\mathbf{Y}\right\|^2}{\widehat{\sigma}^2} + 2(l+1) - n. \tag{2.4}$$

Since division by $\widehat{\sigma}^2$ only has a scaling effect, and since $\widehat{\sigma}^2$ is the same for all subspaces, choosing an L for which (2.3) is a minimum, will result in the same L for which (2.4) is a minimum. Although the form of C_p in (2.4) is generally more familiar, both (2.3) and (2.4) are important in later discussions.

Since the formulation of the standard linear model assumes that $\boldsymbol{\mu} \in M$, it follows that $P_{L^\perp}\boldsymbol{\mu} = P_{M|L}\boldsymbol{\mu}$. The ESEE in (2.2) can therefore also be written as

$$E\left\|P_L\mathbf{Y}-\boldsymbol{\mu}\right\|^2 = (l+1)\sigma^2 + \left\|P_{M|L}\boldsymbol{\mu}\right\|^2$$

which is estimated unbiasedly by

$$\begin{aligned} & \left\|P_{M|L}\mathbf{Y}\right\|^2 + (2(l+1) - \dim(M))\,\widehat{\sigma}^2 \\ = & \left\|P_{M|L}\mathbf{Y}\right\|^2 + (2(l+1) - (m+1))\,\widehat{\sigma}^2. \end{aligned} \tag{2.5}$$

The estimators in (2.3) and (2.5) are equal, since

$$\begin{aligned} C_p(\mathbf{Y}, L) &= \|P_{L^\perp}\mathbf{Y}\|^2 + (2(l+1) - n)\,\widehat{\sigma}^2 \\ &= \left\|P_{M|L}\mathbf{Y}\right\|^2 + \|P_{M^\perp}\mathbf{Y}\|^2 + (2(l+1) - n)\,\widehat{\sigma}^2 \\ &= \left\|P_{M|L}\mathbf{Y}\right\|^2 + (2(l+1) - (m+1))\,\widehat{\sigma}^2, \end{aligned}$$

where the last equality follows from the definition of $\widehat{\sigma}^2$ in (1.10). The symbol $\widetilde{C}_p(\mathbf{Y}, L)$ will be used to refer to the form (2.5) of the C_p criterion.

Applying C_p in a practical situation entails calculating the criterion in (2.3) or (2.5) for all linear subspaces $L \in \mathcal{L}$, and then selecting the subspace with the minimum criterion value. The mean squared errors of the estimators corresponding to all L in M therefore have to be estimated. The selected subspace clearly depends on the data, and this is reflected in the symbols which are used to denote the selected subspace, viz. $\widehat{L}$ or $\widehat{L}\,(\mathbf{Y})$. The corresponding estimator of $\boldsymbol{\mu}$ is denoted by $P_{\widehat{L}}\mathbf{Y}$ or by $P_{\widehat{L}(Y)}\mathbf{Y}$.

What can be said about the ESEE of $P_{\widehat{L}(Y)}\mathbf{Y}$? Since $\widehat{L}(\mathbf{Y})$ is data-dependent, its ESEE is not estimated unbiasedly by the minimum value of the C_p criterion. In fact, the minimum value of the C_p criterion underestimates the ESEE of $P_{\widehat{L}(Y)}\mathbf{Y}$. This follows quite easily, since $\min\{C_p(\mathbf{Y}, L) : L \in \mathcal{L}\} \leq C_p(\mathbf{Y}, L)$ for all $L \in \mathcal{L}$, and for each $\mathbf{Y}$, implies that

$$\begin{aligned} E\left[\min\{C_p(\mathbf{Y}, L) : L \in \mathcal{L}\}\right] &\leq \min\{E\left[C_p(\mathbf{Y}, L)\right] : L \in \mathcal{L}\} \\ &= \min\{E\left\|P_L\mathbf{Y} - \boldsymbol{\mu}\right\|^2 : L \in \mathcal{L}\} \\ &\leq \sum_{L \in \mathcal{L}} E\left\|P_L\mathbf{Y} - \boldsymbol{\mu}\right\|^2 P\left\{\widehat{L}(\mathbf{Y}) = L\right\} \\ &= E\left\|P_{\widehat{L}(Y)}\mathbf{Y} - \boldsymbol{\mu}\right\|^2. \end{aligned}$$

Breiman (1992, Section 2.1) shows a similar result for a specific example in a regression context. Mallows (1995, p. 362) states in this regard that using the estimator corresponding to the data-dependent minimum criterion value, can result in a much larger ESEE than when $P_M\mathbf{Y}$ is used to estimate $\boldsymbol{\mu}$.

We proceed to derive the mean squared error (MSE) of the C_p criterion as an estimator of (2.2) using its form in (2.5). For a fixed linear subspace L, (2.5) is an unbiased estimator of the ESEE in (2.2). Therefore, using the result in Lemma A.4 in Appendix A, and the fact that $\frac{(n-m-1)\widehat{\sigma}^2}{\sigma^2} \sim \mathcal{X}^2_{n-m-1}$, it follows that

$$\begin{aligned}
MSE\left\{\widetilde{C}_p(\mathbf{Y},L)\right\} &= Var\left\{\widetilde{C}_p(\mathbf{Y},L)\right\} \\
&= Var\left\{\left\|P_{M|L}\mathbf{Y}\right\|^2 + (2(l+1)-(m+1))\,\widehat{\sigma}^2\right\} \\
&= Var\left(\left\|P_{M|L}\mathbf{Y}\right\|^2\right) + (2(l+1)-(m+1))^2\, Var\left(\widehat{\sigma}^2\right) \\
&= 4\sigma^2\left\|P_{M|L}\boldsymbol{\mu}\right\|^2 + 2\sigma^4(m-l) + (2(l+1)-(m+1))^2\,\frac{2\sigma^4}{n-(m+1)} \\
&= 4\sigma^2\left\|P_{M|L}\boldsymbol{\mu}\right\|^2 + 2\sigma^4\left((m-l) + \frac{(2(l+1)-(m+1))^2}{n-(m+1)}\right).
\end{aligned}$$

Note finally that when $L = M$ we use $P_M\mathbf{Y}$ to estimate $\boldsymbol{\mu}$. The ESEE in (2.1) is unbiasedly estimated by $C_p(\mathbf{Y},M) = \widehat{\sigma}^2(m+1)$ with

$$MSE\left\{C_p(\mathbf{Y},M)\right\} = 2\sigma^4\frac{(m+1)^2}{n-(m+1)}.$$

2.3.1 Expansion and estimation of $E\left\|P_L\mathbf{Y}-\boldsymbol{\mu}\right\|^2$ as the sum of n terms

One of our objectives is to identify, in the sample of size n, data cases that are influential when the C_p criterion is used for subspace selection. Since the C_p criterion for a given subspace L estimates the ESEE corresponding to this subspace, a first step in quantifying the selection influence of the individual cases in the data set is to express $E\left\|P_L\mathbf{Y}-\boldsymbol{\mu}\right\|^2$ as the sum of n terms. Once this has been achieved, the individual terms can be estimated, thereby obtaining the contribution of each data case to the total estimated ESEE. To accomplish such an expansion, consider the random vector $\mathbf{Z}$ defined by $\sigma\mathbf{Z} = \mathbf{Y}-\boldsymbol{\mu}$. Since $\mathbf{Y} \sim N_n\left(\boldsymbol{\mu},\sigma^2\mathbf{I}_n\right)$, it readily follows that $\mathbf{Z} \sim N_n\left(\mathbf{0},\mathbf{I}_n\right)$. We therefore obtain the following expansion for the ESEE corresponding to a given subspace L:

$$\begin{aligned} E\left\|P_L\mathbf{Y}-\boldsymbol{\mu}\right\|^2 &= E\left\|\sigma P_L\mathbf{Z}+P_L\boldsymbol{\mu}-\boldsymbol{\mu}\right\|^2 \\ &= E\left\|\sigma P_L\mathbf{Z}+P_L\boldsymbol{\mu}-(P_L\boldsymbol{\mu}+P_{L^\perp}\boldsymbol{\mu})\right\|^2 \\ &= \sigma^2 E\left\|P_L\mathbf{Z}\right\|^2+\left\|P_{L^\perp}\boldsymbol{\mu}\right\|^2. \end{aligned} \tag{2.6}$$

The expression in (2.6) is now used to expand the ESEE as a sum of n terms. Let $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_n$ be the standard orthonormal basis fcr $\mathcal{R}^n$. The first term in (2.6) may be written as:

$$\begin{aligned} \sigma^2 E\left\|P_L\mathbf{Z}\right\|^2 &= \sigma^2\sum_{i=1}^n E\left\langle \mathbf{u}_i, P_L\mathbf{Z}\right\rangle^2 \\ &= \sigma^2\sum_{i=1}^n E\left\langle P_L\mathbf{u}_i, \mathbf{Z}\right\rangle^2 \\ &= \sigma^2\left(\sum_{i=1}^n \left\|P_L\mathbf{u}_i\right\|^2+\sum_{i=1}^n \left\langle P_L\mathbf{u}_i, E(\mathbf{Z})\right\rangle^2\right) \\ &= \sigma^2\sum_{i=1}^n \left\|P_L\mathbf{u}_i\right\|^2. \end{aligned} \tag{2.7}$$

Similarly, the second term in (2.6) becomes

$$\left\|P_{L^\perp}\boldsymbol{\mu}\right\|^2=\sum_{i=1}^n \left\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i\right\rangle^2. \tag{2.8}$$

Combining (2.7) and (2.8) we arrive at the following expansion of the ESEE as a sum of n terms:

$$\begin{aligned} E\left\|P_L\mathbf{Y}-\boldsymbol{\mu}\right\|^2 &= \sigma^2\sum_{i=1}^n \left\|P_L\mathbf{u}_i\right\|^2+\sum_{i=1}^n \left\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i\right\rangle^2 \\ &= \sum_{i=1}^n \left\{\sigma^2\left\|P_L\mathbf{u}_i\right\|^2+\left\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i\right\rangle^2\right\} \end{aligned} \tag{2.9}$$

If we write $P_L\mathbf{Y}=[\widehat{\mu}_1(L),\widehat{\mu}_2(L),....\widehat{\mu}_n(L)]'=\left[\widehat{Y}_1(L),\widehat{Y}_2(L),...,\widehat{Y}_n(L)\right]'$, then

$$E\left\|P_L\mathbf{Y}-\boldsymbol{\mu}\right\|^2=\sum_{i=1}^n E\left[\widehat{\mu}_i(L)-\mu_i\right]^2.$$

Now, $E\left[\widehat{\mu}_i(L) - \mu_i\right]^2$ is the mean squared error of $\widehat{\mu}_i(L)$ as an estimator of μ_i, and we see that

$$\begin{aligned}
E\left[\widehat{\mu}_i(L) - \mu_i\right]^2 &= E\left[\langle P_L\mathbf{Y}, \mathbf{u}_i\rangle - \langle \boldsymbol{\mu}, \mathbf{u}_i\rangle\right]^2 \\
&= E\left[\langle P_L\mathbf{Y}, \mathbf{u}_i\rangle - E\left\langle P_L\mathbf{Y}, \mathbf{u}_i\right\rangle + E\left\langle P_L\mathbf{Y}, \mathbf{u}_i\right\rangle - \langle \boldsymbol{\mu}, \mathbf{u}_i\rangle\right]^2 \\
&= Var\left[\langle P_L\mathbf{Y}, \mathbf{u}_i\rangle\right] + \left[E\left\langle P_L\mathbf{Y}, \mathbf{u}_i\right\rangle - \langle \boldsymbol{\mu}, \mathbf{u}_i\rangle\right]^2 \\
&= \sigma^2 \left\|P_L\mathbf{u}_i\right\|^2 + \left\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i\right\rangle^2.
\end{aligned}$$

This is exactly the ith term in (2.9). We can therefore view (2.9) as an expansion of the ESEE of $P_L\mathbf{Y} = [\widehat{\mu}_1(L), \widehat{\mu}_2(L), ..., \widehat{\mu}_n(L)]'$ as the sum of the mean squared errors of $\widehat{\mu}_1(L), \widehat{\mu}_2(L), ..., \widehat{\mu}_n(L)$ as estimators of $\mu_1, \mu_2, ..., \mu_n$ respectively.

We now consider estimation of the individual terms of the ESEE. Consider the ith term in (2.9), given by

$$E(\widehat{\mu}_i(L) - \mu_i)^2 = \sigma^2 \left\|P_L\mathbf{u}_i\right\|^2 + \left\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i\right\rangle^2. \tag{2.10}$$

We start by estimating the first term on the right-hand side of (2.10). Since $\left\|P_L\mathbf{u}_i\right\|^2$ is known, the first term, $\sigma^2 \left\|P_L\mathbf{u}_i\right\|^2$, can be estimated unbiasedly by $\widehat{\sigma}^2 \left\|P_L\mathbf{u}_i\right\|^2$, where $\widehat{\sigma}^2$ is the unbiased estimator of σ^2 defined in (1.10). Since $E\left\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\right\rangle^2 = \left\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i\right\rangle^2 + \sigma^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2$, an unbiased estimator of the second term, $\left\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i\right\rangle^2$, on the right-hand side of (2.10) is given by $\left\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\right\rangle^2 - \widehat{\sigma}^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2$. The ith term in (2.10) is therefore unbiasedly estimated by

$$C_p(\mathbf{Y}, L, i) = \widehat{\sigma}^2 \left\|P_L\mathbf{u}_i\right\|^2 + \left\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\right\rangle^2 - \widehat{\sigma}^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2. \tag{2.11}$$

The symbol $C_p(\mathbf{Y}, L, i)$ is used to denote the right-hand side of (2.11), since the sum over i of the right-hand side of (2.11) can be shown to equal the C_p criterion in (2.3). To see this, the following result is required:

$$\sum_{i=1}^{n} \left\langle P_L\mathbf{u}_i, \mathbf{u}_i\right\rangle = \dim(L). \tag{2.12}$$

To establish (2.12), let $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_{l+1}$ be an orthonormal basis for the subspace L. Then,

$$\begin{aligned}
\sum_{i=1}^{n} \langle P_L \mathbf{u}_i, \mathbf{u}_i \rangle &= \sum_{i=1}^{n} \left\langle \sum_{j=1}^{\dim(L)} \langle \mathbf{v}_j, \mathbf{u}_i \rangle \mathbf{v}_j, \mathbf{u}_i \right\rangle \\
&= \sum_{i=1}^{n} \sum_{j=1}^{\dim(L)} \langle \mathbf{v}_j, \mathbf{u}_i \rangle^2 \\
&= \sum_{j=1}^{\dim(L)} \sum_{i=1}^{n} \langle \mathbf{v}_j, \mathbf{u}_i \rangle^2 \\
&= \sum_{j=1}^{\dim(L)} \|\mathbf{v}_j\|^2 \\
&= \dim(L).
\end{aligned}$$

It now follows that

$$\begin{aligned}
\sum_{i=1}^{n} C_p(\mathbf{Y}, L, i) &= \sum_{i=1}^{n} \left\{ \widehat{\sigma}^2 \|P_L \mathbf{u}_i\|^2 + \langle \mathbf{Y}, P_{L^\perp} \mathbf{u}_i \rangle^2 - \widehat{\sigma}^2 \|P_{L^\perp} \mathbf{u}_i\|^2 \right\} \qquad (2.13) \\
&= \widehat{\sigma}^2 \sum_{i=1}^{n} \left[\langle P_L \mathbf{u}_i, \mathbf{u}_i \rangle - \langle P_{L^\perp} \mathbf{u}_i, \mathbf{u}_i \rangle \right] + \|P_{L^\perp} \mathbf{Y}\|^2 \\
&= \widehat{\sigma}^2 \left[(l+1) - (n - (l+1)) \right] + \|P_{L^\perp} \mathbf{Y}\|^2 \\
&= \widehat{\sigma}^2 (2(l+1) - n) + \|P_{L^\perp} \mathbf{Y}\|^2 \\
&= C_p(\mathbf{Y}, L).
\end{aligned}$$

Summarising: For a given subspace L we can write the ESEE of the corresponding estimator, $P_L \mathbf{Y}$, as in (2.9):

$$E \|P_L \mathbf{Y} - \boldsymbol{\mu}\|^2 = \sum_{i=1}^{n} \left\{ \sigma^2 \|P_L \mathbf{u}_i\|^2 + \langle \boldsymbol{\mu}, P_{L^\perp} \mathbf{u}_i \rangle^2 \right\}.$$

The ith term is this sum is estimated unbiasedly by $C_p(\mathbf{Y}, L, i)$ in (2.11), and the sum over i of the $C_p(\mathbf{Y}, L, i)$-values is simply $C_p(\mathbf{Y}, L)$ of (2.3).

We now turn to the alternative form of the C_p criterion given in (2.5), viz. $\widetilde{C}_p(\mathbf{Y}, L)$. We will show that results similar to those established above, are also valid for $\widetilde{C}_p(\mathbf{Y}, L)$. Note firstly that since $\boldsymbol{\mu} \in M$ it follows that $P_{L^\perp} \boldsymbol{\mu} = P_{M|L} \boldsymbol{\mu}$. The ESEE in (2.6) can therefore also be expressed as

$$\begin{aligned} E\left\|P_L\mathbf{Y}-\boldsymbol{\mu}\right\|^2 &= E\left\|\sigma P_L\mathbf{Z}-P_{L^\perp}\boldsymbol{\mu}\right\|^2 \\ &= E\left\|\sigma P_L\mathbf{Z}-P_{M|L}\boldsymbol{\mu}\right\|^2 \\ &= \sigma^2 E\left\|P_L\mathbf{Z}\right\|^2+\left\|P_{M|L}\boldsymbol{\mu}\right\|^2 \end{aligned} \tag{2.14}$$

so that an equivalent form of (2.9) is

$$\begin{aligned} E\left\|P_L\mathbf{Y}-\boldsymbol{\mu}\right\|^2 &= \sigma^2\sum_{i=1}^n\left\|P_L\mathbf{u}_i\right\|^2+\sum_{i=1}^n\left\langle\boldsymbol{\mu},P_{M|L}\mathbf{u}_i\right\rangle^2 \\ &= \sum_{i=1}^n\left\{\sigma^2\left\|P_L\mathbf{u}_i\right\|^2+\left\langle\boldsymbol{\mu},P_{M|L}\mathbf{u}_i\right\rangle^2\right\}. \end{aligned} \tag{2.15}$$

The ith term of the ESEE in (2.15) is now unbiasedly estimated by

$$\widetilde{C}_p(\mathbf{Y},L,i)=\widehat{\sigma}^2\left\|P_L\mathbf{u}_i\right\|^2+\left\langle\mathbf{Y},P_{M|L}\mathbf{u}_i\right\rangle^2-\widehat{\sigma}^2\left\|P_{M|L}\mathbf{u}_i\right\|^2. \tag{2.16}$$

It should be noted that $\widetilde{C}_p(\mathbf{Y},L,i)$ can be written in two other equivalent forms. Firstly, since

$$\begin{aligned} \left\langle\mathbf{Y},P_{M|L}\mathbf{u}_i\right\rangle^2 &= \left\langle P_M\mathbf{Y}+P_{M^\perp}\mathbf{Y},P_{M|L}\mathbf{u}_i\right\rangle^2 \\ &= \left\langle P_M\mathbf{Y},P_{M|L}\mathbf{u}_i\right\rangle^2 \end{aligned}$$

it follows that

$$\widetilde{C}_p(\mathbf{Y},L,i)=\widehat{\sigma}^2\left\|P_L\mathbf{u}_i\right\|^2+\left\langle P_M\mathbf{Y},P_{M|L}\mathbf{u}_i\right\rangle^2-\widehat{\sigma}^2\left\|P_{M|L}\mathbf{u}_i\right\|^2. \tag{2.17}$$

Similarly, since

$$\begin{aligned} \left\langle\mathbf{Y},P_{M|L}\mathbf{u}_i\right\rangle^2 &= \left\langle P_{M|L}\mathbf{Y},P_{M|L}\mathbf{u}_i\right\rangle^2 \\ &= \left\langle P_{M|L}\mathbf{Y},P_{M|L}\mathbf{u}_i+P_{M^\perp}\mathbf{u}_i\right\rangle^2 \\ &= \left\langle P_{M|L}\mathbf{Y},P_{L^\perp}\mathbf{u}_i\right\rangle^2 \\ &= \left\langle P_L\mathbf{Y}+P_{M|L}\mathbf{Y},P_{L^\perp}\mathbf{u}_i\right\rangle^2 \\ &= \langle P_M\mathbf{Y},P_{L^\perp}\mathbf{u}_i\rangle^2 \end{aligned}$$

it also follows that

$$\widetilde{C}_p(\mathbf{Y},L,i)=\widehat{\sigma}^2\left\|P_L\mathbf{u}_i\right\|^2+\langle P_M\mathbf{Y},P_{L^\perp}\mathbf{u}_i\rangle^2-\widehat{\sigma}^2\left\|P_{M|L}\mathbf{u}_i\right\|^2. \tag{2.18}$$

Taking the sum over i of the proposed estimator in (2.16) or, equivalently, (2.17) or (2.18) again yields the C_p criterion given in (2.5):

$$\begin{aligned}
\sum_{i=1}^{n} \widetilde{C}_p(\mathbf{Y},L,i) &= \sum_{i=1}^{n} \left\{ \widehat{\sigma}^2 \left\|P_L\mathbf{u}_i\right\|^2 + \left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i \right\rangle^2 - \widehat{\sigma}^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2 \right\} \qquad (2.19)\\
&= \widehat{\sigma}^2 \sum_{i=1}^{n} \left(\left\langle P_L\mathbf{u}_i, \mathbf{u}_i \right\rangle - \left\langle P_M\mathbf{u}_i - P_L\mathbf{u}_i, \mathbf{u}_i \right\rangle \right) + \left\|P_{M|L}\mathbf{Y}\right\|^2 \\
&= \widehat{\sigma}^2 \sum_{i=1}^{n} \left(2\left\langle P_L\mathbf{u}_i, \mathbf{u}_i \right\rangle - \left\langle P_M\mathbf{u}_i, \mathbf{u}_i \right\rangle \right) + \left\|P_{M|L}\mathbf{Y}\right\|^2 \\
&= \widehat{\sigma}^2 (2(l+1) - (m+1)) + \left\|P_{M|L}\mathbf{Y}\right\|^2 \qquad \text{(using (2.12))} \\
&= \widetilde{C}_p\left(\mathbf{Y},L\right).
\end{aligned}$$

Note that (2.13) and (2.19) are expansions of Mallows' C_p criterion as the sum of n terms, given within the coordinate free framework. The respective ith cases of these expansions, i.e. $C_p(\mathbf{Y},L,i)$ in (2.11) and $\widetilde{C}_p(\mathbf{Y},L,i)$ in (2.16) or (2.17) or (2.18), are unbiased estimators of the ith term of the expansion of the ESEE in (2.9). Note that this ith term of the expansion of the ESEE in (2.9) is identical to the ith term of the expansion of the ESEE in (2.15), since

$$\begin{aligned}
E(\widehat{\mu}_i(L) - \mu_i)^2 &= \sigma^2 \left\|P_L\mathbf{u}_i\right\|^2 + \left\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i \right\rangle^2 \\
&= \sigma^2 \left\|P_L\mathbf{u}_i\right\|^2 + \left\langle P_{L^\perp}\boldsymbol{\mu}, \mathbf{u}_i \right\rangle^2 \\
&= \sigma^2 \left\|P_L\mathbf{u}_i\right\|^2 + \left\langle P_{M|L}\boldsymbol{\mu}, \mathbf{u}_i \right\rangle^2 \\
&= \sigma^2 \left\|P_L\mathbf{u}_i\right\|^2 + \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i \right\rangle^2. \qquad (2.20)
\end{aligned}$$

2.3.2 Different estimators of $\left\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i \right\rangle^2 = \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i \right\rangle^2$

In the previous section, two unbiased estimators of the ith term in the expansion (2.9) or (2.15) were introduced, viz. $C_p(\mathbf{Y},L,i)$ in (2.11) and $\widetilde{C}_p(\mathbf{Y},L,i)$ in (2.16). We now compare these two unbiased estimators in terms of their respective variances in order to determine which estimator is relatively more efficient.

Consider first the case where it can be assumed *that the value of* σ^2 *is known*. Then, $C_p(\mathbf{Y}, L, i) = \sigma^2 \|P_L \mathbf{u}_i\|^2 + \langle \mathbf{Y}, P_{L^\perp} \mathbf{u}_i \rangle^2 - \sigma^2 \|P_{L^\perp} \mathbf{u}_i\|^2$, and hence

$$\begin{aligned} Var\{C_p(\mathbf{Y}, L, i)\} &= Var\left\{\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2 + \sigma^2 \|P_L \mathbf{u}_i\|^2 - \sigma^2 \|P_{L^\perp}\mathbf{u}_i\|^2\right\} \\ &= Var\left\{\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2\right\} \\ &= E\left\{\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2 - \langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i\rangle^2 - \sigma^2 \|P_{L^\perp}\mathbf{u}_i\|^2\right\}^2 \\ &= E\{\widehat{\gamma}_1(\mathbf{Y}, L, i) - \gamma(\boldsymbol{\mu}, L, i)\}^2 \\ &= Var\{\widehat{\gamma}_1(\mathbf{Y}, L, i)\} \end{aligned}$$

where

$$\widehat{\gamma}_1(\mathbf{Y}, L, i) = \langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2 - \sigma^2 \|P_{L^\perp}\mathbf{u}_i\|^2 \tag{2.21}$$

and

$$\gamma(\boldsymbol{\mu}, L, i) = \langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i\rangle^2 = \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\right\rangle^2. \tag{2.22}$$

Similarly,

$$\begin{aligned} Var\left\{\widetilde{C}_p(\mathbf{Y}, L, i)\right\} &= Var\left\{\left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^2\right\} \\ &= Var\{\widehat{\gamma}_2(\mathbf{Y}, L, i)\} \end{aligned}$$

where

$$\widehat{\gamma}_2(\mathbf{Y}, L, i) = \left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^2 - \sigma^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2. \tag{2.23}$$

Hence, the objective of comparing the variances of $C_p(\mathbf{Y}, L, i)$ and $\widetilde{C}_p(\mathbf{Y}, L, i)$ can be accomplished by comparing the variances of $\widehat{\gamma}_1(\mathbf{Y}, L, i)$ and $\widehat{\gamma}_2(\mathbf{Y}, L, i)$. For notational convenience, let $\gamma(\boldsymbol{\mu}, L, i)$ in (2.22) be denoted by γ. Also, let $\widehat{\gamma}_1(\mathbf{Y}, L, i)$ in (2.21) and $\widehat{\gamma}_2(\mathbf{Y}, L, i)$ in (2.23) respectively be denoted by $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$. The difference between the variances of $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ can be written as

$$\begin{aligned} Var(\widehat{\gamma}_1) - Var(\widehat{\gamma}_2) &= E(\widehat{\gamma}_1 - \gamma)^2 - E(\widehat{\gamma}_2 - \gamma)^2 \\ &= E\left(\widehat{\gamma}_1^2\right) - E\left(\widehat{\gamma}_2^2\right). \end{aligned} \tag{2.24}$$

Since $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$, it follows that $\left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle \sim N\left(\left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\right\rangle, \left\|P_{M|L}\mathbf{u}_i\right\|^2 \sigma^2\right)$. Hence, $\left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle = \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\right\rangle + \left\|P_{M|L}\mathbf{u}_i\right\| \sigma Z$, where Z is a standard normal random variable. It is well known that

$$E\left(Z^{2r}\right) = (2r-1)(2r-3)...(3)(1) \text{ for } r = 1, 2, 3, ...$$

Hence, taking $r = 2$ we find $E\left(Z^4\right) = 3$. Combining this with the well known results $E\left(Z\right) = E\left(Z^3\right) = 0$, $E\left(Z^2\right) = 1$, and the expression above for $\left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i \right\rangle$, we find that

$$E\left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i \right\rangle^4 = \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i \right\rangle^4 + 6\sigma^2 \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i \right\rangle^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2 + 3\sigma^4 \left\|P_{M|L}\mathbf{u}_i\right\|^4. \quad (2.25)$$

Using (2.25), $E\left(\widehat{\gamma}_2^2\right)$ can be expressed as

$$\begin{aligned} E\left(\widehat{\gamma}_2^2\right) &= E\left\{\left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i \right\rangle^2 - \sigma^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2\right\}^2 \\ &= E\left\{\left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i \right\rangle^4 + \sigma^4 \left\|P_{M|L}\mathbf{u}_i\right\|^4 - 2\sigma^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2 \left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i \right\rangle^2\right\} \\ &= \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i \right\rangle^4 + 6\sigma^2 \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i \right\rangle^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2 + 3\sigma^4 \left\|P_{M|L}\mathbf{u}_i\right\|^4 + \sigma^4 \left\|P_{M|L}\mathbf{u}_i\right\|^4 \\ &\quad -2\sigma^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2 \left\{\left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i \right\rangle^2 + \sigma^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2\right\} \\ &= \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i \right\rangle^4 + 4\sigma^2 \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i \right\rangle^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2 + 2\sigma^4 \left\|P_{M|L}\mathbf{u}_i\right\|^4. \end{aligned} \quad (2.26)$$

Similarly if we write $E\left\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i \right\rangle^4$ as in (2.25) and use the fact that $\left\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i \right\rangle^2 = \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i \right\rangle^2$, we find that

$$\begin{aligned} E\left(\widehat{\gamma}_1^2\right) &= E\left\{\left\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i \right\rangle^2 - \sigma^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2\right\}^2 \\ &= \left\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i \right\rangle^4 + 4\sigma^2 \left\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i \right\rangle^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2 + 2\sigma^4 \left\|P_{L^\perp}\mathbf{u}_i\right\|^4 \\ &= \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i \right\rangle^4 + 4\sigma^2 \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i \right\rangle^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2 + 2\sigma^4 \left\|P_{L^\perp}\mathbf{u}_i\right\|^4. \end{aligned}$$

Therefore, it follows that

$$\begin{aligned} E\left(\widehat{\gamma}_1^2\right) - E\left(\widehat{\gamma}_2^2\right) &= 4\sigma^2 \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i \right\rangle^2 \left\{\left\|P_{L^\perp}\mathbf{u}_i\right\|^2 - \left\|P_{M|L}\mathbf{u}_i\right\|^2\right\} \\ &\quad +2\sigma^4 \left\{\left\|P_{L^\perp}\mathbf{u}_i\right\|^4 - \left\|P_{M|L}\mathbf{u}_i\right\|^4\right\}. \end{aligned} \quad (2.27)$$

Note that, since $P_{M|L}\mathbf{u}_i = P_M\mathbf{u}_i - P_L\mathbf{u}_i$, it follows that $\mathbf{u}_i = P_L\mathbf{u}_i + P_{L^\perp}\mathbf{u}_i = P_M\mathbf{u}_i - P_{M|L}\mathbf{u}_i + P_{L^\perp}\mathbf{u}_i$. Therefore, it is clear that

$$\begin{aligned} P_{L^\perp}\mathbf{u}_i &= \mathbf{u}_i - P_M\mathbf{u}_i + P_{M|L}\mathbf{u}_i \\ &= P_{M^\perp}\mathbf{u}_i + P_{M|L}\mathbf{u}_i \end{aligned}$$

and consequently

$$\begin{aligned}\left\|P_{L^\perp}\mathbf{u}_i\right\|^2 &= \left\|P_{M^\perp}\mathbf{u}_i\right\|^2+\left\|P_{M|L}\mathbf{u}_i\right\|^2\\ &\geqq \left\|P_{M|L}\mathbf{u}_i\right\|^2.\end{aligned}$$

Combined with (2.27) this implies that $E\left(\widehat{\gamma}_1^2\right)-E\left(\widehat{\gamma}_2^2\right)\geqq 0$. Assuming therefore that σ^2 is known, it follows from (2.24) that $Var(\widehat{\gamma}_1)\geqq Var(\widehat{\gamma}_2)$, and therefore

$$Var\left\{C_p(\mathbf{Y},L,i)\right\}\geqq Var\left\{\widetilde{C}_p(\mathbf{Y},L,i)\right\}. \tag{2.28}$$

We now briefly discuss a special case, illustrating the magnitude of the difference between $Var\left(\widehat{\gamma}_1\right)$ and $Var\left(\widehat{\gamma}_2\right)$. We consider the case where σ^2 is known, and we use simulation to approximate $Var\left(\widehat{\gamma}_1\right)$ and $Var\left(\widehat{\gamma}_2\right)$. In this numerical evaluation the vectors

$$\mathbf{m}_1=[\overbrace{1,...,1}^{10},\overbrace{0,...,0}^{10}]' \text{ and } \mathbf{m}_2=[\overbrace{0,...,0}^{10},\overbrace{1...,1}^{10}]' \tag{2.29}$$

in $\mathcal{R}^{20}$ are chosen as basis vectors for the two-dimensional linear subspace $M=span\left\{\mathbf{m}_1,\mathbf{m}_2\right\}$. Note that $\mathbf{m}_1$ and $\mathbf{m}_2$ are linearly independent. Also, for any $\alpha,\beta\in\mathcal{R}$ the vector

$$\boldsymbol{\mu}=\alpha\mathbf{m}_1+\beta\mathbf{m}_2\in M.$$

Consider now the linear subspace $L=span\left\{\mathbf{m}_2\right\}$, which is contained in M. The following results will be helpful in later calculations. For a vector $\mathbf{x}=[x_1,...,x_{20}]'$ it follows that

$$P_L\mathbf{x}=\frac{\langle\mathbf{x},\mathbf{m}_2\rangle}{\left\|\mathbf{m}_2\right\|^2}\mathbf{m}_2=\left(\frac{1}{10}\sum_{i=11}^{20}x_i\right)\mathbf{m}_2 \tag{2.30}$$

and therefore

$$\left\|P_L\mathbf{x}\right\|^2=10\left(\frac{1}{10}\sum_{i=11}^{20}x_i\right)^2. \tag{2.31}$$

Also,

$$P_M\mathbf{x}=\frac{\langle\mathbf{x},\mathbf{m}_1\rangle}{\left\|\mathbf{m}_1\right\|^2}\mathbf{m}_1+\frac{\langle\mathbf{x},\mathbf{m}_2\rangle}{\left\|\mathbf{m}_2\right\|^2}\mathbf{m}_2=\left(\frac{1}{10}\sum_{i=1}^{10}x_i\right)\mathbf{m}_1+\left(\frac{1}{10}\sum_{i=11}^{20}x_i\right)\mathbf{m}_2 \tag{2.32}$$

and therefore

$$P_{M|L}\mathbf{x}=P_M\mathbf{x}-P_L\mathbf{x}=\left(\frac{1}{10}\sum_{i=1}^{10}x_i\right)\mathbf{m}_1 \tag{2.33}$$

so that

$$\left\| P_{M|L}\mathbf{x} \right\|^2 = 10 \left(\frac{1}{10} \sum_{i=1}^{10} x_i \right)^2 . \tag{2.34}$$

For given values of α and β, the vector $\boldsymbol{\mu}$ is known and the parameter $\gamma(\boldsymbol{\mu}, L, i)$ in (2.22) can easily be obtained. This is done most simply by calculating the squared difference between the ith element of $\boldsymbol{\mu}$ and the ith element of $P_L\boldsymbol{\mu}$, i.e. $\gamma(\boldsymbol{\mu}, L, i) = [\langle \boldsymbol{\mu}, \mathbf{u}_i \rangle - \langle P_L\boldsymbol{\mu}, \mathbf{u}_i \rangle]^2$. Note that $P_L\boldsymbol{\mu}$ is calculated using the result in (2.30). With the value of σ^2 known, a random vector $\boldsymbol{\varepsilon}$ is simulated from an $N_{20}\left(\mathbf{0},\sigma^2\mathbf{I}_{20}\right)$-distribution. Consequently, the response vector, $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, can be obtained. The estimator $\widehat{\gamma}_1(\mathbf{Y}, L, i)$ in (2.21) can now also be calculated, where its first term, $\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i \rangle^2$, is obtained using again the result in (2.30). The second term, $\sigma^2 \left\| P_{L^\perp}\mathbf{u}_i \right\|^2 = \sigma^2 \left(1 - \left\| P_L\mathbf{u}_i \right\|^2\right)$, of this estimator is acquired from the result in (2.31). In a similar way, $\widehat{\gamma}_2(\mathbf{Y}, L, i)$ in (2.23) is obtained using the results in (2.33) and (2.34).

In the simulation study we choose $\beta = 1$, while the value of α is chosen to vary from -30 to 30 in steps of 1. For each of these α-values a different $\boldsymbol{\mu}$ vector is obtained. By repeatedly simulating error vectors, $\boldsymbol{\varepsilon} \sim N_{20}\left(\mathbf{0},\sigma^2\mathbf{I}_{20}\right)$, ten-thousand response vectors, $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, are obtained for a given $\boldsymbol{\mu}$ vector. Using these response vectors, the corresponding estimators, $\widehat{\gamma}_1(\mathbf{Y}, L, i)$ and $\widehat{\gamma}_2(\mathbf{Y}, L, i)$, are calculated for each of the $i = 1, ..., 20$ cases of the parameter $\gamma(\boldsymbol{\mu}, L, i)$. Since very similar results are calculated for all $i = 1, ..., 10$, we only present the simulation results for $i = 1$. We do not comment on simulation results for values of $i = 11, ..., 20$, since the parameter of interest for these cases equals zero. For a particular $\boldsymbol{\mu}$ vector, the respective variances of $\widehat{\gamma}_1(\mathbf{Y}, L, 1)$ and $\widehat{\gamma}_2(\mathbf{Y}, L, 1)$ are approximated by

$$Var\left\{\widehat{\gamma}_1(\mathbf{Y}, L, 1)\right\} \cong \frac{\sum_{j=1}^{10000} \left(\widehat{\gamma}_1(\mathbf{Y}, L, 1)_j - \gamma(\boldsymbol{\mu}, L, 1)\right)^2}{10000} \tag{2.35}$$

and

$$Var\left\{\widehat{\gamma}_2(\mathbf{Y}, L, 1)\right\} \cong \frac{\sum_{j=1}^{10000} \left(\widehat{\gamma}_2(\mathbf{Y}, L, 1)_j - \gamma(\boldsymbol{\mu}, L, 1)\right)^2}{10000}. \tag{2.36}$$

Program C1 in Appendix C was used to obtain the approximated variances in (2.35) and (2.36). In Figure 2.1, the ratio of the approximate variances in (2.35) and (2.36) is plotted for different values of $\boldsymbol{\mu} = \alpha\mathbf{m}_1 + \beta\mathbf{m}_2$, where $\beta = 1$ and $\alpha = -30(1)30$. These ratios are shown for values of σ which equal: $20, 10, 5, 2, 1.5, 1, 0.5$ and 0.2. In Figure 2.1 the "outside" plotted curve, which is symmetrical around $\alpha = 0$, corresponds to $\sigma = 20$, while the most "inner" symmetrical curve corresponds to $\sigma = 0.2$.

Since the plotted ratios in Figure 2.1 are consistently greater than 1, the simulation results confirm the general analytical result in (2.28). Specifically , for α-values far from zero (i.e. large negative and positive values), $Var\left\{\widehat{\gamma}_1(\mathbf{Y}, L, 1)\right\}$ is approximately ten times larger than $Var\left\{\widehat{\gamma}_2(\mathbf{Y}, L, 1)\right\}$, irrespective of the value of σ. Depending on the value of σ, the difference between the two variances grows dramatically as α moves nearer to zero.

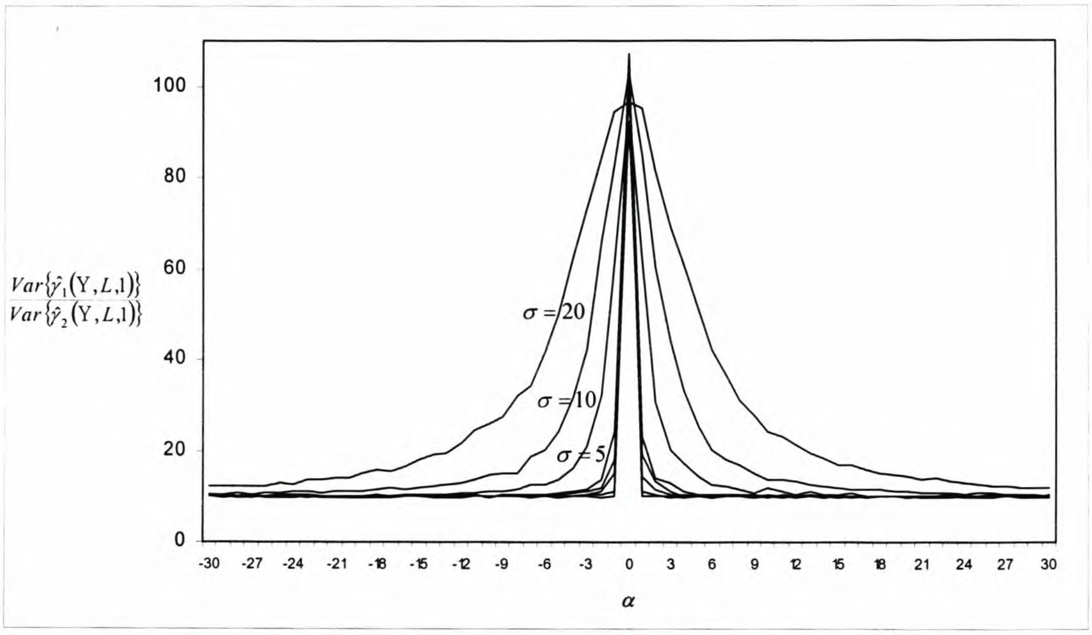


Figure 2.1: Plot of $\frac{Var\{\widehat{\gamma}_1(\mathbf{Y},L,1)\}}{Var\{\widehat{\gamma}_2(\mathbf{Y},L,1)\}}$ for different values of σ^2 if α is incremented from -30 to 30 in steps of 1

Finally, for a given linear subspace L, $Var\left(\widehat{\gamma}_1\right)$ can be expressed as $E\left(\widehat{\gamma}_1^2\right) - \gamma^2$. From (2.27) it thus follows that

$$\begin{aligned} Var\left(\widehat{\gamma}_1\right) &= E\left(\widehat{\gamma}_2^2\right) - \gamma^2 + 4\sigma^2 \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i \right\rangle^2 \left\{ \left\|P_{L^\perp}\mathbf{u}_i\right\|^2 - \left\|P_{M|L}\mathbf{u}_i\right\|^2 \right\} \\ &\quad + 2\sigma^4 \left\{ \left\|P_{L^\perp}\mathbf{u}_i\right\|^4 - \left\|P_{M|L}\mathbf{u}_i\right\|^4 \right\} \\ &= Var\left(\widehat{\gamma}_2\right) + 4\sigma^2 \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i \right\rangle^2 \left\{ \left\|P_{L^\perp}\mathbf{u}_i\right\|^2 - \left\|P_{M|L}\mathbf{u}_i\right\|^2 \right\} \\ &\quad + 2\sigma^4 \left\{ \left\|P_{L^\perp}\mathbf{u}_i\right\|^4 - \left\|P_{M|L}\mathbf{u}_i\right\|^4 \right\}. \end{aligned}$$

Analytically, the ratio of the two variances can therefore be expressed as

$$\begin{aligned}
&\frac{Var\left(\widehat{\gamma}_1\right)}{Var\left(\widehat{\gamma}_2\right)} \\
= \quad & 1+\frac{4\sigma^2\left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\right\rangle^2\left\{\left\|P_{L^\perp}\mathbf{u}_i\right\|^2-\left\|P_{M|L}\mathbf{u}_i\right\|^2\right\}+2\sigma^4\left\{\left\|P_{L^\perp}\mathbf{u}_i\right\|^4-\left\|P_{M|L}\mathbf{u}_i\right\|^4\right\}}{Var\left(\widehat{\gamma}_2\right)} \\
= \quad & 1+\frac{4\sigma^2\left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\right\rangle^2\left\{\left\|P_{L^\perp}\mathbf{u}_i\right\|^2-\left\|P_{M|L}\mathbf{u}_i\right\|^2\right\}+2\sigma^4\left\{\left\|P_{L^\perp}\mathbf{u}_i\right\|^4-\left\|P_{M|L}\mathbf{u}_i\right\|^4\right\}}{4\sigma^2\left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\right\rangle^2\left\|P_{M|L}\mathbf{u}_i\right\|^2+2\sigma^4\left\|P_{M|L}\mathbf{u}_i\right\|^4}
\end{aligned} \tag{2.37}$$

where the last step follows from (2.26). Consider now the particular case where $\boldsymbol{\mu} = 10\mathbf{m}_1 + 1\mathbf{m}_2 = [\overbrace{10,...,10}^{10},\overbrace{1,...,1}^{10}]'$, and $\sigma = 1$. For $i = 1$ the ratio in (2.37) becomes

$$\begin{aligned}
\frac{Var\left(\widehat{\gamma}_1\right)}{Var\left(\widehat{\gamma}_2\right)} &= 1+\frac{4(100)\left\{1-0.1\right\}+2\left\{1-0.01\right\}}{4(100)\left\{0.1\right\}+2\left\{0.01\right\}} \\
&= 10.045.
\end{aligned}$$

The simulated value for this particular case is equal 10.292. The difference, $|10.292 - 10.045| = 0.047$, illustrates the magnitude of the simulation error. In a similar way, $\frac{Var(\widehat{\gamma}_1)}{Var(\widehat{\gamma}_2)} = 100$ for the case where $\boldsymbol{\mu} = 0\mathbf{m}_1 + 1\mathbf{m}_2$ and $\sigma = 1$, with a corresponding simulated value of 100.698.

We now move to the case *where the value of σ^2 is unknown*. The estimators in (2.21) and (2.23) are no longer useful, and are replaced by

$$\widehat{\gamma}_1(\mathbf{Y}, L, i) = \left\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\right\rangle^2 - \widehat{\sigma}^2\left\|P_{L^\perp}\mathbf{u}_i\right\|^2$$

and

$$\widehat{\gamma}_2(\mathbf{Y}, L, i) = \left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^2 - \widehat{\sigma}^2\left\|P_{M|L}\mathbf{u}_i\right\|^2$$

respectively. In these expressions, σ^2 is estimated unbiasedly by $\widehat{\sigma}^2 = \frac{\left\|P_{M^\perp}\mathbf{Y}\right\|^2}{n-(m+1)}$. Steps similar to those required to derive (2.26), give

$$\begin{aligned}
E\left(\widehat{\gamma}_2^2\right) &= E\left\{\left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^2 - \widehat{\sigma}^2\left\|P_{M|L}\mathbf{u}_i\right\|^2\right\}^2 \\
&= E\left\{\left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^4 + \widehat{\sigma}^4\left\|P_{M|L}\mathbf{u}_i\right\|^4 - 2\left\|P_{M|L}\mathbf{u}_i\right\|^2\widehat{\sigma}^2\left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^2\right\} \\
&= \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\right\rangle^4 + 6\sigma^2\left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\right\rangle^2\left\|P_{M|L}\mathbf{u}_i\right\|^2 + 3\sigma^4\left\|P_{M|L}\mathbf{u}_i\right\|^4 \\
&\quad + \left\|P_{M|L}\mathbf{u}_i\right\|^4\left(\frac{2+n-m}{n-m}\right)\sigma^4 - 2\left\|P_{M|L}\mathbf{u}_i\right\|^2 E\left\{\widehat{\sigma}^2\left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^2\right\}
\end{aligned}$$

since $E\left(\widehat{\sigma}^4\right) = \sigma^4\left(\frac{2+n-m}{n-m}\right)$. We can now write $\left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^2 = \left\langle P_{M|L}\mathbf{Y}, \mathbf{u}_i\right\rangle^2$, and since $P_{M|L}\mathbf{Y}$ and $P_{M^\perp}\mathbf{Y}$ are independent random vectors, we conclude that $\widehat{\sigma}^2$ and $\left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^2$ are independently distributed. The final term in the expression above therefore becomes

$$\begin{aligned} & 2\left\|P_{M|L}\mathbf{u}_i\right\|^2 E\left(\widehat{\sigma}^2\right) E\left(\left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^2\right) \\ = \; & 2\left\|P_{M|L}\mathbf{u}_i\right\|^2 \sigma^2\left(\left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\right\rangle^2 + \sigma^2\left\|P_{M|L}\mathbf{u}_i\right\|^2\right), \end{aligned}$$

and we finally see that

$$E\left(\widehat{\gamma}_2^2\right) = \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\right\rangle^4 + 4\sigma^2\left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\right\rangle^2\left\|P_{M|L}\mathbf{u}_i\right\|^2 + 2\sigma^4\left\|P_{M|L}\mathbf{u}_i\right\|^4\left(\frac{1+n-m}{n-m}\right).$$

However, if we try to find $E\left(\widehat{\gamma}_1^2\right)$ we run into difficulties at the final step. It is found that

$$\begin{aligned} E\left(\widehat{\gamma}_1^2\right) &= E\left\{\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^4 + \widehat{\sigma}^4\left\|P_{L^\perp}\mathbf{u}_i\right\|^4 - 2\widehat{\sigma}^2\left\|P_{L^\perp}\mathbf{u}_i\right\|^2\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2\right\} \\ &= E\left(\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^4 + \widehat{\sigma}^4\left\|P_{L^\perp}\mathbf{u}_i\right\|^4\right) - 2\left\|P_{L^\perp}\mathbf{u}_i\right\|^2 E\left(\widehat{\sigma}^2\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2\right). \end{aligned}$$

The argument used in the derivation of $E\left(\widehat{\gamma}_2^2\right)$ is now no longer valid, since $\widehat{\sigma}^2$ and $\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2$ are not independent. The term $E\left(\widehat{\sigma}^2\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2\right)$ can therefore not be simplified. The implication is that no analytical progress can be made when we try to prove that $E\left(\widehat{\gamma}_1^2\right) - E\left(\widehat{\gamma}_2^2\right) \geqq 0$. Numerical evaluation of $Var\left(\widehat{\gamma}_1\right)$ and $Var\left(\widehat{\gamma}_2\right)$ by simulation was therefore considered. The results were very similar to the results obtained from the simulation study for the case where σ^2 is known. Without going into a detailed discussion, we would like to emphasize that in all cases considered in the simulation study it was found that $Var\left(\widehat{\gamma}_1^2\right) \geqq Var\left(\widehat{\gamma}_2^2\right)$.

2.3.3 Truncating the estimators of $\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i\rangle^2 = \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\right\rangle^2$

In the previous section, two estimators of $\gamma(\boldsymbol{\mu}, L, i) = \langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i\rangle^2 = \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\right\rangle^2$ were considered: $\widehat{\gamma}_1(\mathbf{Y}, L, i)$ in (2.21)and $\widehat{\gamma}_2(\mathbf{Y}, L, i)$ in (2.23). It is clear that $\gamma(\boldsymbol{\mu}, L, i)$ can never be negative. The same is obviously not true for $\widehat{\gamma}_1(\mathbf{Y}, L, i)$ or $\widehat{\gamma}_2(\mathbf{Y}, L, i)$. A natural idea that now suggests itself is to consider truncated versions of $\widehat{\gamma}_1(\mathbf{Y}, L, i)$ and $\widehat{\gamma}_2(\mathbf{Y}, L, i)$, thereby eliminating the unsatisfactory situation which may otherwise arise whereby a quantity which we know to be non-negative is estimated to be negative. This leads to two further estimators of

$\gamma(\boldsymbol{\mu}, L, i)$, viz.

$$\begin{aligned} \widehat{\gamma}_1^+(\mathbf{Y}, L, i) &= \max\left\{0, \langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2 - \widehat{\sigma}^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2\right\} \\ &= \max\left\{0, \widehat{\gamma}_1(\mathbf{Y}, L, i)\right\} \end{aligned} \tag{2.38}$$

and

$$\begin{aligned} \widehat{\gamma}_2^+(\mathbf{Y}, L, i) &= \max\left\{0, \left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^2 - \widehat{\sigma}^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2\right\} \\ &= \max\left\{0, \widehat{\gamma}_2(\mathbf{Y}, L, i)\right\} \end{aligned} \tag{2.39}$$

For notational brevity we once again denote the truncated estimators in (2.38) and (2.39) by $\widehat{\gamma}_1^+$ and $\widehat{\gamma}_2^+$ respectively. It is clear that $Var\left(\widehat{\gamma}_1^+\right) \leqq Var\left(\widehat{\gamma}_1\right)$, with a similar result holding for $\widehat{\gamma}_2^+$ and $\widehat{\gamma}_2$. An interesting question concerns the magnitude of the reduction in variance obtainable by replacing $\widehat{\gamma}_1$ or $\widehat{\gamma}_2$ by their respective truncated versions. It does not seem possible to answer this question analytically, and we therefore investigate a special case by means of simulation.

We consider the same situation as in the previous section, i.e. a two-dimensional linear space M spanned by the vectors $\mathbf{m}_1$ and $\mathbf{m}_2$ in (2.29). The subspace L of M is spanned by $\mathbf{m}_2$, and the problem is to estimate $\gamma(\boldsymbol{\mu}, L, i)$. We wish to investigate the following questions:

(i) How does $Var\left(\widehat{\gamma}_1^+\right)$ compare to $Var\left(\widehat{\gamma}_1\right)$?

(ii) How does $Var\left(\widehat{\gamma}_2^+\right)$ compare tu $Var\left(\widehat{\gamma}_2\right)$?

(iii) How does $Var\left(\widehat{\gamma}_1^+\right)$ compare to $Var\left(\widehat{\gamma}_2^+\right)$?

Using simulation to answer these questions, entails the following. Consider a fixed set of parameter values, i.e. values of α, β (and therefore $\boldsymbol{\mu}$), and σ. Calculate γ from these parameter values. Now generate a large number of observation vectors $\mathbf{Y}$ from the appropriate normal distribution, and calculate the four estimators, viz. $\widehat{\gamma}_1, \widehat{\gamma}_1^+, \widehat{\gamma}_2$ and $\widehat{\gamma}_2^+$ for each observation. (Note that in calculating these estimators, the value of σ^2 is unbiasedly estimated by $\widehat{\sigma}^2 = \frac{\|\mathbf{Y}-P_M\mathbf{Y}\|^2}{20-2}$). This gives a large number of realisations of each of the four estimators, and we can therefore approximate the variances of these estimators as in (2.35) or (2.36).

The results which were obtained in this way, for the same σ-values as in Figure 2.1, are summarised in Figures 2.2, 2.3 and 2.4. In Figure 2.2, we graph a simulation approximation of $\frac{Var\{\widehat{\gamma}_1(\mathbf{Y},L,1)\}}{Var\left\{\widehat{\gamma}_1^+(\mathbf{Y},L,1)\right\}}$. Firstly, consider small σ-values presented by the most inner symmetrical curves around $\alpha = 0$. It is clear that the variance of $\widehat{\gamma}_1(\mathbf{Y}, L, 1)$ and the variance of its truncated version are identical for α-values far from zero. However, the truncated version shows up to a 30% reduction in variance for α-values in the vicinity of zero. If larger σ-values are considered, which are presented by the outside symmetrical curves, we see that the performance of $\widehat{\gamma}_1^+(\mathbf{Y}, L, 1)$ is in general better than that of $\widehat{\gamma}_1(\mathbf{Y}, L, 1)$ for a wider range of α-values.

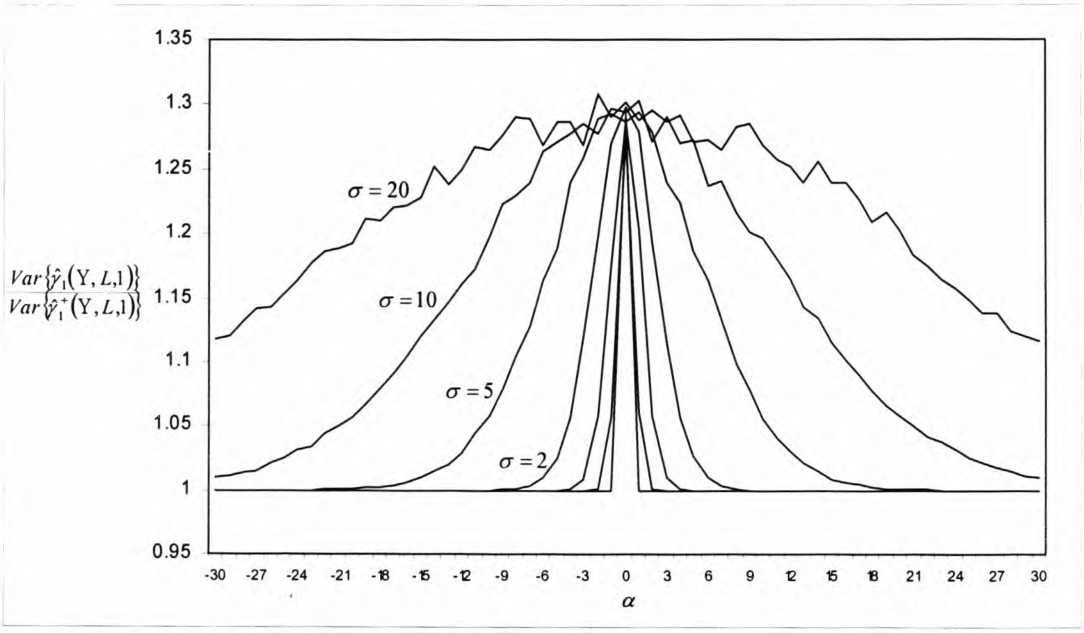


Figure 2.2: Plot of $\frac{Var\{\widehat{\gamma}_1(\mathbf{Y},L,1)\}}{Var\left\{\widehat{\gamma}_1^+(\mathbf{Y},L,1)\right\}}$ for different values of σ^2 if α is incremented from -30 to 30 in steps of 1

Next, in Figure 2.3 we depict our simulation approximation of $\frac{Var\{\widehat{\gamma}_2(\mathbf{Y},L,1)\}}{Var\left\{\widehat{\gamma}_2^+(\mathbf{Y},L,1)\right\}}$. Once again it is seen that $Var\left\{\widehat{\gamma}_2^+(\mathbf{Y}, L, 1)\right\}$ is smaller than $Var\left\{\widehat{\gamma}_2(\mathbf{Y}, L, 1)\right\}$ only for α-values near zero. The range of α-values for which $Var\left\{\widehat{\gamma}_2^+(\mathbf{Y}, L, 1)\right\} \leqq Var\left\{\widehat{\gamma}_2(\mathbf{Y}, L, 1)\right\}$ becomes larger as σ increases, but this range is smaller than in Figure 2.2.

Finally, Figure 2.4 shows $\frac{Var\left\{\widehat{\gamma}_1^+(\mathbf{Y},L,1)\right\}}{Var\left\{\widehat{\gamma}_2^+(\mathbf{Y},L,1)\right\}}$. We see that the results which are obtained are very similar to those in Figure 2.1, where the ratios of the variances of the non-truncated estimators are considered. Once again the variance of $\widehat{\gamma}_1^+(\mathbf{Y}, L, 1)$ is approximately 10 times larger than that of $\widehat{\gamma}_2^+(\mathbf{Y}, L, 1)$ for α-values far from zero. These ratios become much larger for α-values near zero.

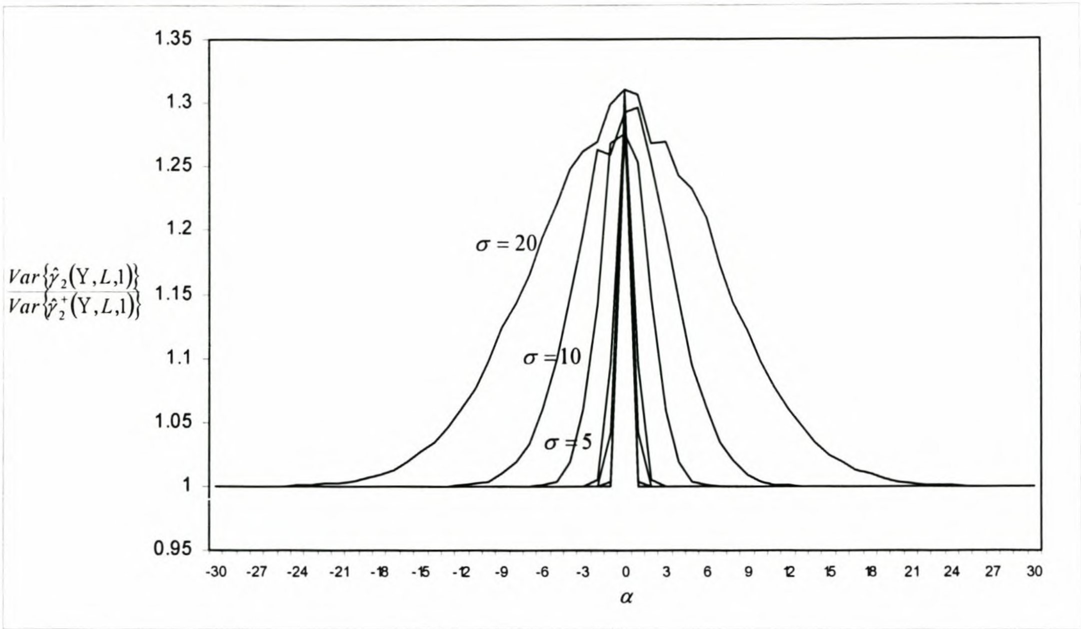


Figure 2.3: Plot of $\frac{Var\{\widehat{\gamma}_2(\mathbf{Y},L,1)\}}{Var\left\{\widehat{\gamma}_2^+(\mathbf{Y},L,1)\right\}}$ for different values of σ^2 if α is incremented from -30 to 30 in steps of 1

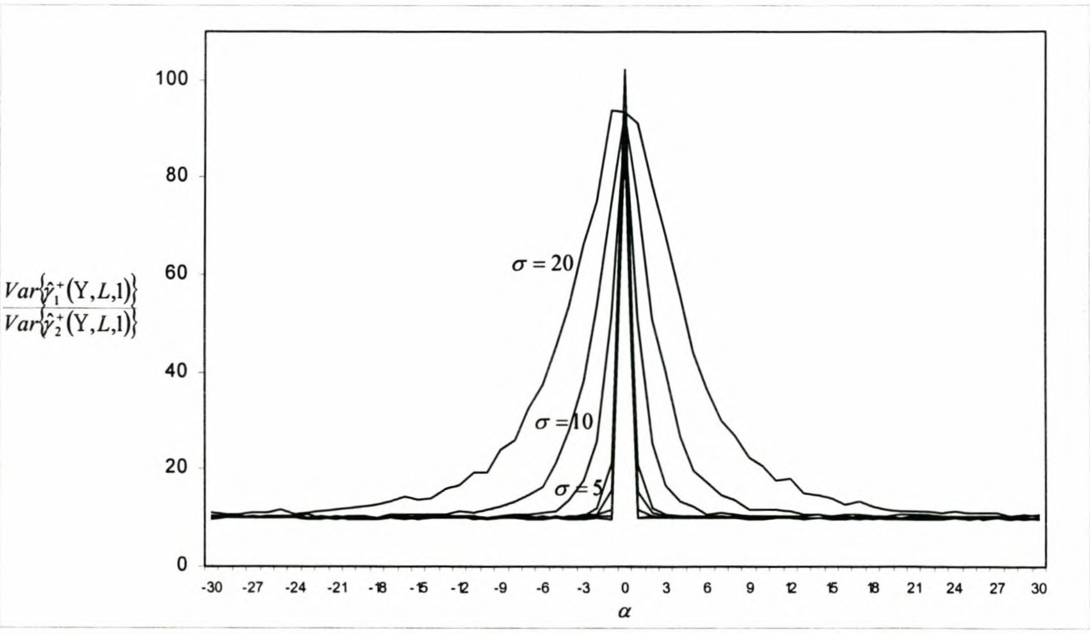


Figure 2.4: Plot of $\frac{Var\left\{\widehat{\gamma}_1^+(\mathbf{Y},L,1)\right\}}{Var\left\{\widehat{\gamma}_2^+(\mathbf{Y},L,1)\right\}}$ for different values of σ^2 if α is incremented from -30 to 30 in steps of 1

Recall that $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ are estimators of $\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i\rangle^2 = \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\right\rangle^2$, and that $\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i\rangle^2 = \left\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\right\rangle^2$ appears in the ith term of the expansion of the ESEE corresponding to the subspace L (see (2.9)). This ith term is estimated unbiasedly by $C_p(\mathbf{Y}, L, i)$ in (2.11) and $\widetilde{C}_p(\mathbf{Y}, L, i)$ in (2.16). Truncating $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ therefore leads to truncated versions of $C_p(\mathbf{Y}, L, i)$ and $\widetilde{C}_p(\mathbf{Y}, L, i)$. In particular, the truncated version of $C_p(\mathbf{Y}, L, i)$ is given by

$$C_p^+(\mathbf{Y}, L, i) = \widehat{\sigma}^2 \left\|P_L\mathbf{u}_i\right\|^2 + \max\left\{0, \langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2 - \widehat{\sigma}^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2\right\}$$

while the truncated version of $\widetilde{C}_p(\mathbf{Y}, L, i)$ is given by

$$\widetilde{C}_p^+(\mathbf{Y}, L, i) = \widehat{\sigma}^2 \left\|P_L\mathbf{u}_i\right\|^2 + \max\left\{0, \left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^2 - \widehat{\sigma}^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2\right\}.$$

These truncated estimators are improvements of $C_p(\mathbf{Y}, L, i)$ and $\widetilde{C}_p(\mathbf{Y}, L, i)$ respectively when our objective is to estimate the ith term in (2.9). Note also that this term is simply estimated by $\widehat{\sigma}^2 \left\|P_L\mathbf{u}_i\right\|^2$ if truncation is applied. Naturally, the truncated versions of $C_p(\mathbf{Y}, L, i)$ and $\widetilde{C}_p(\mathbf{Y}, L, i)$ also lead to truncated versions of the C_p criterion. In particular, (2.13) becomes

$$C_p^+(\mathbf{Y}, L) = \sum_{i=1}^{n} \left[\widehat{\sigma}^2 \left\|P_L\mathbf{u}_i\right\|^2 + \max\left\{0, \langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2 - \widehat{\sigma}^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2\right\}\right]. \tag{2.40}$$

while the modified version (2.19) of the C_p criterion becomes

$$\widetilde{C}_p^+(\mathbf{Y}, L) = \sum_{i=1}^{n} \left[\widehat{\sigma}^2 \left\|P_L\mathbf{u}_i\right\|^2 + \max\left\{0, \left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^2 - \widehat{\sigma}^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2\right\}\right]. \tag{2.41}$$

Snyman (1994, p.2-5) also argues along these lines by proposing that the C_p criterion in (2.3) and (2.5) be truncated at zero. The motivation for this follows from the fact that the C_p criterion gives an unbiased estimate of the non-negative ESEE for a fixed linear subspace L. According to Snyman these truncated estimators also produce smaller mean squared errors than the estimators in (2.3) and (2.5). Note that the proposed truncated C_p criteria in (2.40) and (2.41) differ from those proposed by Snyman, since in (2.40) and (2.41) the estimators of the ith cases in the expansion of C_p are truncated if they are less than zero. Taking therefore the sum over all these estimators (where some are now truncated) provides C_p criteria different from those proposed by Snyman.

2.4 Variable selection and estimation in multiple linear regression

In this section we apply the coordinate free concepts, developed previously, to variable selection in multiple linear regression. In multiple linear regression the columns of the design matrix, $\mathbf{X} = [\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m]$, form a natural basis for the $(m+1)$ dimensional linear subspace M. In a regression context it is therefore more convenient to work within the coordinatized framework. The normal multiple linear regression model, as a special case of the coordinatized form of the standard normal linear model, is given by

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2.42}$$

In (2.42), estimation of $\boldsymbol{\mu} \in M$ entails estimation of the parameter vector $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_m]'$. For this purpose the least squares estimator of $\boldsymbol{\beta}$, as defined in (1.6), is frequently used. Since the column vectors of $\mathbf{X}$ are predetermined, the estimator of $\boldsymbol{\mu}$, $\mathbf{X}\widehat{\boldsymbol{\beta}}$ in (1.7), is fully described once $\boldsymbol{\beta}$ has been estimated.

When provision is made for variable selection in multiple linear regression we additionally assume that $\boldsymbol{\mu}$ possibly belongs to an unknown member of the family of linear subspaces contained in M. Denote this family by $\mathcal{L}$ and let L denote such a typical linear subspace with $\dim(L) = l + 1$. The subspace L is spanned by the vector $\mathbf{1}$ and a subset of l predictor variables, i.e. $L = span\{\mathbf{1}, \mathbf{x}_j : j \in J_L\} \subset M$, where $J_L \subset \{1, 2, ..., m\}$. Let $\mathbf{X}_L$ be the $n \times (l+1)$ matrix with column vectors $\mathbf{1}$ and $\mathbf{x}_j$, $j \in J_L$. Then the least squares estimators of β_0 and the regression coefficients corresponding to the variables spanning L are given by $\widehat{\boldsymbol{\beta}}_L = (\mathbf{X}_L'\mathbf{X}_L)^{-1}\mathbf{X}_L'\mathbf{Y}$. The corresponding estimator of $\boldsymbol{\mu}$ follows as $\mathbf{X}_L\widehat{\boldsymbol{\beta}}_L$.

In this dissertation we apply the C_p criterion, which for a given L (i.e., a given subset of predictor variables) unbiasedly estimates the ESEE, as variable selection technique. For a particular data set we therefore determine the estimated ESEE for every possible subset of predictor variables and then select the subset having the minimum estimated mean squared error.

As was previously done in the coordinate free formulation we now apply our results on the expansion of the ESEE and the C_p criterion to multiple linear regression.

2.4.1 Expansion and estimation of the ESEE as the sum of n terms within a multiple linear regression context

In this section we express the ESEE as the sum of n terms when variables are selected in a multiple linear regression context. As was done in the coordinate free setup we also estimate the individual terms in the expansion of the ESEE. Consider therefore again the coordinate free expansion of the ESEE in (2.9), viz.

$$E\left\|P_L\mathbf{Y}-\boldsymbol{\mu}\right\|^2=\sum_{i=1}^{n}\left\{\sigma^2\left\|P_L\mathbf{u}_i\right\|^2+\langle\boldsymbol{\mu},P_{L^\perp}\mathbf{u}_i\rangle^2\right\}, \tag{2.43}$$

where $\mathbf{u}_i, i=1,...,n$, form the standard orthonormal basis for $\mathcal{R}^n$. Recall that the vectors $\mathbf{1}$ and $\mathbf{x}_j$, $j\in J_L$, form a basis for the linear subspace L, and that the matrix $\mathbf{X}_L$ has these vectors as columns. The projection matrix, P_L, can therefore be expressed as $P_L=\mathbf{X}_L(\mathbf{X}_L'\mathbf{X}_L)^{-1}\mathbf{X}_L'$. This is an $n\times n$ idempotent and symmetric matrix. The left-hand side of (2.43) becomes

$$\begin{aligned}
&E\left\|P_L\mathbf{Y}-\boldsymbol{\mu}\right\|^2\\
=\;&E\left\|\mathbf{X}_L(\mathbf{X}_L'\mathbf{X}_L)^{-1}\mathbf{X}_L'\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right\|^2\\
=\;&E\left\{\left[\mathbf{X}_L(\mathbf{X}_L'\mathbf{X}_L)^{-1}\mathbf{X}_L'\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right]'\left[\mathbf{X}_L(\mathbf{X}_L'\mathbf{X}_L)^{-1}\mathbf{X}_L'\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right]\right\}.
\end{aligned}$$

Regarding the right-hand side of (2.43), note first that

$$\begin{aligned}
\left\|P_L\mathbf{u}_i\right\|^2 &= \langle P_L\mathbf{u}_i,P_L\mathbf{u}_i\rangle\\
&= \mathbf{u}_i'P_L'P_L\mathbf{u}_i\\
&= \mathbf{u}_i'P_L\mathbf{u}_i,
\end{aligned} \tag{2.44}$$

and this is the ith diagonal element of P_L, i.e. the ith diagonal element of $\mathbf{X}_L(\mathbf{X}_L'\mathbf{X}_L)^{-1}\mathbf{X}_L'$. We will denote this scalar by v_{ii}, i.e.

$$\left\|P_L\mathbf{u}_i\right\|^2=\mathbf{u}_i'\mathbf{X}_L(\mathbf{X}_L'\mathbf{X}_L)^{-1}\mathbf{X}_L'\mathbf{u}_i=v_{ii}. \tag{2.45}$$

Let J_L^C denote the complement of J_L, i.e. J_L^C contains the indices of the predictor variables not represented in $\mathbf{X}_L$. The second term in the summation on the right-hand side of (2.43) now becomes

$$\begin{aligned}
\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i\rangle^2 &= \langle \mathbf{X}\boldsymbol{\beta}, P_{L^\perp}\mathbf{u}_i\rangle^2 \\
&= \left\langle \beta_0\mathbf{1} + \sum_{j\in J_L}\beta_j\mathbf{x}_j + \sum_{j\in J_L^C}\beta_j\mathbf{x}_j, P_{L^\perp}\mathbf{u}_i\right\rangle^2 \\
&= \left\langle \beta_0\mathbf{1} + \sum_{j\in J_L^C}\beta_j\mathbf{x}_j, P_{L^\perp}\mathbf{u}_i\right\rangle^2 \\
&= \left[\beta_0\left\langle \mathbf{1}, P_{L^\perp}\mathbf{u}_i\right\rangle + \sum_{j\in J_L^C}\beta_j\left\langle \mathbf{x}_j, P_{L^\perp}\mathbf{u}_i\right\rangle\right]^2,
\end{aligned}$$

so that the ESEE can be written as

$$\begin{aligned}
E\left\|P_L\mathbf{Y} - \boldsymbol{\mu}\right\|^2 &= \sum_{i=1}^{n}\left\{\sigma^2 v_{ii} + \langle \mathbf{X}\boldsymbol{\beta}, P_{L^\perp}\mathbf{u}_i\rangle^2\right\} \\
&= \sum_{i=1}^{n}\left\{\sigma^2 v_{ii} + \left[\beta_0\left\langle \mathbf{1}, P_{L^\perp}\mathbf{u}_i\right\rangle + \sum_{j\in J_L^C}\beta_j\left\langle \mathbf{x}_j, P_{L^\perp}\mathbf{u}_i\right\rangle\right]^2\right\}. \qquad (2.46)
\end{aligned}$$

Since $\langle \mathbf{X}\boldsymbol{\beta}, P_{L^\perp}\mathbf{u}_i\rangle^2 = \left\langle \mathbf{X}\boldsymbol{\beta}, P_{M|L}\mathbf{u}_i\right\rangle^2$, an equivalent form of (2.46) is given by

$$E\left\|P_L\mathbf{Y} - \boldsymbol{\mu}\right\|^2 = \sum_{i=1}^{n}\left\{\sigma^2 v_{ii} + \left[\beta_0\left\langle \mathbf{1}, P_{M|L}\mathbf{u}_i\right\rangle + \sum_{j\in J_L^C}\beta_j\left\langle \mathbf{x}_j, P_{M|L}\mathbf{u}_i\right\rangle\right]^2\right\}. \qquad (2.47)$$

An unbiased estimator of

$$\begin{aligned}
&\left[\beta_0\left\langle \mathbf{1}, P_{L^\perp}\mathbf{u}_i\right\rangle + \sum_{j\in J_L^C}\beta_j\left\langle \mathbf{x}_j, P_{L^\perp}\mathbf{u}_i\right\rangle\right]^2 \\
= &\left[\beta_0\left\langle \mathbf{1}, P_{M|L}\mathbf{u}_i\right\rangle + \sum_{j\in J_L^C}\beta_j\left\langle \mathbf{x}_j, P_{M|L}\mathbf{u}_i\right\rangle\right]^2 \qquad (2.48)
\end{aligned}$$

in (2.46) and (2.47) is given by

$$\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2 - \widehat{\sigma}^2\left\|P_{L^\perp}\mathbf{u}_i\right\|^2 = \left[Y_i - \left(\widehat{\beta}_0 + \sum_{j\in J_L}\widehat{\beta}_j\mathbf{x}_j\right)\right]^2 - \widehat{\sigma}^2\left\|P_{L^\perp}\mathbf{u}_i\right\|^2$$

where $\widehat{\sigma}^2$ is an unbiased estimator of σ^2. The ith term in (2.46) and (2.47) is therefore unbiasedly estimated by

$$\begin{aligned} C_p(\mathbf{Y}, L, i) &= \widehat{\sigma}^2 v_{ii} + \langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2 - \widehat{\sigma}^2 \left\| P_{L^\perp}\mathbf{u}_i \right\|^2 \\ &= \widehat{\sigma}^2 v_{ii} + (\langle \mathbf{Y}, \mathbf{u}_i\rangle - \langle P_L\mathbf{Y}, \mathbf{u}_i\rangle)^2 - \widehat{\sigma}^2 \left(\left\|\mathbf{u}_i\right\|^2 - \left\|P_L\mathbf{u}_i\right\|^2 \right) \\ &= \widehat{\sigma}^2 v_{ii} + \left(Y_i - \widehat{Y}_i(L)\right)^2 - \widehat{\sigma}^2(1 - v_{ii}) \end{aligned} \tag{2.49}$$

where Y_i is the ith component of $\mathbf{Y}$, and $\widehat{Y}_i(L)$ is the ith predicted value, i.e. the ith component of $P_L\mathbf{Y}$ as given in (1.13). Taking the sum over i in (2.49) and applying the result in (2.12) we obtain

$$\begin{aligned} \sum_{i=1}^{n} C_p(\mathbf{Y}, L, i) &= \sum_{i=1}^{n} \left\{ \widehat{\sigma}^2 v_{ii} + \left(Y_i - \widehat{Y}_i(L)\right)^2 - \widehat{\sigma}^2(1 - v_{ii}) \right\} \\ &= \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i(L)\right)^2 + \sum_{i=1}^{n} \widehat{\sigma}^2 \left(2v_{ii} - 1\right) \\ &= RSS_L + \widehat{\sigma}^2 \left(2(l+1) - n\right) \end{aligned} \tag{2.50}$$

where RSS_L is the residual sum of squares associated with the subset of predictor variables which spans the linear subspace L. Note that (2.50) is an expression for the C_p criterion for which the coordinate free counterpart is given in (2.3).

By simply replacing β_0 and β_j, $j \in J_L^C$, in (2.46) or (2.47) by their corresponding least squares estimators obtained from $\widehat{\boldsymbol{\beta}} = \left[\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_m\right]'$ in (1.6), it follows that

$$\begin{aligned} & E\left[\widehat{\beta}_0 \langle \mathbf{1}, P_{L^\perp}\mathbf{u}_i\rangle + \sum_{j\in J_L^C} \widehat{\beta}_j \langle \mathbf{x}_j, P_{L^\perp}\mathbf{u}_i\rangle\right]^2 \\ = & \left[E\left(\widehat{\beta}_0 \langle \mathbf{1}, P_{L^\perp}\mathbf{u}_i\rangle\right) + \sum_{j\in J_L^C} E\left(\widehat{\beta}_j \langle \mathbf{x}_j, P_{L^\perp}\mathbf{u}_i\rangle\right)\right]^2 \\ & + Var\left[\widehat{\beta}_0 \langle \mathbf{1}, P_{L^\perp}\mathbf{u}_i\rangle + \sum_{j\in J_L^C} \widehat{\beta}_j \langle \mathbf{x}_j, P_{L^\perp}\mathbf{u}_i\rangle\right] \\ = & \langle \mathbf{a}_i, \boldsymbol{\beta}\rangle^2 + Var\left\langle \mathbf{a}_i, \widehat{\boldsymbol{\beta}}\right\rangle \\ = & \langle \mathbf{a}_i, \boldsymbol{\beta}\rangle^2 + \sigma^2 \mathbf{a}_i' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{a}_i \end{aligned} \tag{2.51}$$

where $\mathbf{a}_i = [1, a_{i1}, a_{i2}, ..., a_{im}]'$ with

$$a_{ij} = \begin{cases} 0, \text{if } j \in J_L \\ \langle \mathbf{x}_j, P_{L^\perp}\mathbf{u}_i \rangle, \text{if } j \in J_L^C \end{cases}.$$

Note that $Var\left\langle \mathbf{a}_i, \widehat{\boldsymbol{\beta}} \right\rangle = \sigma^2 \mathbf{a}_i' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{a}_i$ in (2.51) can be written as

$$\begin{aligned} Var\left[\widehat{\beta}_0 \langle \mathbf{1}, P_{L^\perp}\mathbf{u}_i \rangle + \sum_{j \in J_L^C} \widehat{\beta}_j \langle \mathbf{x}_j, P_{L^\perp}\mathbf{u}_i \rangle\right] &= Var\left\langle \widehat{\beta}_0 \mathbf{1} + \sum_{j \in J_L^C} \widehat{\beta}_j \mathbf{x}_j, P_{L^\perp}\mathbf{u}_i \right\rangle \\ &= Var\left\langle P_{M|L}\mathbf{Y}, P_{L^\perp}\mathbf{u}_i \right\rangle \\ &= Var\left\langle P_{M|L}\mathbf{Y}, P_{M|L}\mathbf{u}_i \right\rangle \\ &= \sigma^2 \left\| P_{M|L}\mathbf{u}_i \right\|^2 \\ &= \sigma^2 \left(\|P_M\mathbf{u}_i\|^2 - \|P_L\mathbf{u}_i\|^2 \right) \\ &= \sigma^2 (u_{ii} - v_{ii}) \qquad (2.52) \end{aligned}$$

where u_{ii} is the ith diagonal element of $P_M = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Also note that $\left\langle \mathbf{a}_i, \widehat{\boldsymbol{\beta}} \right\rangle^2$ in (2.51) can be written as

$$\begin{aligned} \left[\widehat{\beta}_0 \langle \mathbf{1}, P_{L^\perp}\mathbf{u}_i \rangle + \sum_{j \in J_L^C} \widehat{\beta}_j \langle \mathbf{x}_j, P_{L^\perp}\mathbf{u}_i \rangle\right]^2 &= \left\langle \widehat{\beta}_0 \mathbf{1} + \sum_{j \in J_L} \widehat{\beta}_j \mathbf{x}_j + \sum_{j \in J_L^C} \widehat{\beta}_j \mathbf{x}_j, P_{L^\perp}\mathbf{u}_i \right\rangle^2 \\ &= \langle P_M\mathbf{Y}, P_{L^\perp}\mathbf{u}_i \rangle^2 \\ &= \left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i \right\rangle^2. \end{aligned}$$

If $\widehat{\sigma}^2$ estimates σ^2 unbiasedly, it follows that (2.48) is unbiasedly estimated by

$$\left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i \right\rangle^2 - \widehat{\sigma}^2 (u_{ii} - v_{ii}),$$

which implies that the ith term in (2.46) and (2.47) is unbiasedly estimated by

$$\begin{aligned} \widetilde{C}_p(\mathbf{Y}, L, i) &= \widehat{\sigma}^2 v_{ii} + \left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i \right\rangle^2 - \widehat{\sigma}^2 (u_{ii} - v_{ii}) \\ &= \widehat{\sigma}^2 v_{ii} + \left(\widehat{Y}_i(M) - \widehat{Y}_i(L) \right)^2 - \widehat{\sigma}^2 (u_{ii} - v_{ii}), \qquad (2.53) \end{aligned}$$

where $\widehat{Y}_i(M)$ is the ith predicted value obtained from $P_M\mathbf{Y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Taking the sum over i in (2.53) and applying the result in (2.12) we obtain the form of the C_p criterion for which the coordinate free counterpart is given in (2.5), viz.

$$\begin{aligned}\sum_{i=1}^{n}\widetilde{C}_p(\mathbf{Y},L,i) &= \sum_{i=1}^{n}\left\{\widehat{\sigma}^2 v_{ii}+\left(\widehat{Y}_i(M)-\widehat{Y}_i(L)\right)^2-\widehat{\sigma}^2\left(u_{ii}-v_{ii}\right)\right\} \qquad (2.54)\\ &= \sum_{i=1}^{n}\left(\widehat{Y}_i(M)-\widehat{Y}_i(L)\right)^2+\sum_{i=1}^{n}\widehat{\sigma}^2\left(2v_{ii}-u_{ii}\right)\\ &= \sum_{i=1}^{n}\left(\widehat{Y}_i(M)-\widehat{Y}_i(L)\right)^2+\widehat{\sigma}^2\left(2(l+1)-(m+1)\right).\end{aligned}$$

Weisberg (1981) also gives the expansion in (2.54) for the C_p criterion.

Finally in this section we consider the expansion of the ESEE and estimation of its individual terms for the special case when the vector $\mathbf{1}$ and the set of predictor variables, $\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_m$, are mutually orthogonal. If the linear transformation, $\sigma\mathbf{Z}=\mathbf{Y}-\boldsymbol{\mu}$, where $\sigma\mathbf{Z}\sim N_n(\mathbf{0},\sigma^2\mathbf{I}_n)$, is applied to $\mathbf{Y}$, it follows from (2.14) that the ESEE can once again be expressed as

$$E\left\|P_L\mathbf{Y}-\boldsymbol{\mu}\right\|^2=\sigma^2E\left\|P_L\mathbf{Z}\right\|^2+\left\|P_{M|L}\boldsymbol{\mu}\right\|^2 \qquad (2.55)$$

for the subset of mutually orthogonal predictor variables which spans the linear subspace L. Let $\mathbf{u}_1,\mathbf{u}_2,...,\mathbf{u}_n$ again denote the standard orthonormal basis for $\mathcal{R}^n$. Then

$$P_L\mathbf{Z}=\begin{bmatrix}\langle P_L\mathbf{Z},\mathbf{u}_1\rangle\\ \langle P_L\mathbf{Z},\mathbf{u}_2\rangle\\ \vdots\\ \langle P_L\mathbf{Z},\mathbf{u}_n\rangle\end{bmatrix}=\begin{bmatrix}\langle\mathbf{Z},P_L\mathbf{u}_1\rangle\\ \langle\mathbf{Z},P_L\mathbf{u}_2\rangle\\ \vdots\\ \langle\mathbf{Z},P_L\mathbf{u}_n\rangle\end{bmatrix}. \qquad (2.56)$$

Since the predictor variables $\mathbf{x}_j$, $j\in J_L$, form an orthogonal basis for L, the projection of $\mathbf{u}_i$ on L can be written as

$$P_L\mathbf{u}_i=\sum_{j\in J_L}\frac{\langle\mathbf{u}_i,\mathbf{x}_j\rangle}{\left\|\mathbf{x}_j\right\|^2}\mathbf{x}_j=\sum_{j\in J_L}\frac{x_{ij}}{\left\|\mathbf{x}_j\right\|^2}\mathbf{x}_j.$$

The inner product of $P_L\mathbf{u}_i$ and $\mathbf{Z}$ is therefore given by

$$\langle P_L\mathbf{u}_i,\mathbf{Z}\rangle=\sum_{j\in J_L}\frac{x_{ij}}{\left\|\mathbf{x}_j\right\|^2}\langle\mathbf{x}_j,\mathbf{Z}\rangle. \qquad (2.57)$$

From (2.56) and (2.57) it thus follows that

$$\begin{aligned} \|P_L\mathbf{Z}\|^2 &= \sum_{i=1}^{n} \langle P_L\mathbf{u}_i, \mathbf{Z}\rangle^2 \\ &= \sum_{i=1}^{n} \left\{ \sum_{j\in J_L} \frac{x_{ij}}{\|\mathbf{x}_j\|^2} \langle \mathbf{x}_j, \mathbf{Z}\rangle \right\}^2 . \end{aligned} \tag{2.58}$$

Taking the expected value of (2.58) it follows that

$$\begin{aligned} E\|P_L\mathbf{Z}\|^2 &= \sum_{i=1}^{n} \left\{ \sum_{j\in J_L} \frac{x_{ij}^2}{\|\mathbf{x}_j\|^4} E\langle \mathbf{x}_j, \mathbf{Z}\rangle^2 \right\} + \sum_{i=1}^{n} \left\{ \sum_{j\in J_L} \sum_{\substack{k\in J_L \\ j\neq k}} \frac{x_{ij}x_{ik}}{\|\mathbf{x}_j\|^2 \|\mathbf{x}_k\|^2} E\left(\langle \mathbf{x}_j, \mathbf{Z}\rangle \langle \mathbf{x}_k, \mathbf{Z}\rangle\right) \right\} \\ &= \sum_{i=1}^{n} \left\{ \sum_{j\in J_L} \frac{x_{ij}^2}{\|\mathbf{x}_j\|^4} \|\mathbf{x}_j\|^2 \right\} + \sum_{i=1}^{n} \left\{ \sum_{j\in J_L} \sum_{\substack{k\in J_L \\ j\neq k}} \frac{x_{ij}x_{ik}}{\|\mathbf{x}_j\|^2 \|\mathbf{x}_k\|^2} \langle \mathbf{x}_j, \mathbf{x}_k\rangle \right\} \\ &= \sum_{i=1}^{n} \left\{ \sum_{j\in J_L} \frac{x_{ij}^2}{\|\mathbf{x}_j\|^2} \right\} . \end{aligned} \tag{2.59}$$

Since the subset of predictor vectors in J_L^C is orthogonal to the subset of predictor variables which are associated with L, and $\langle \boldsymbol{\mu}, \mathbf{x}_j\rangle = \langle \beta_0\mathbf{1} + \sum_{k=1}^{m} \beta_k\mathbf{x}_k, \mathbf{x}_j\rangle = \beta_j \|\mathbf{x}_j\|^2$, it follows in a similar way that

$$\begin{aligned} \|P_{M|L}\boldsymbol{\mu}\|^2 &= \sum_{i=1}^{n} \langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\rangle^2 \\ &= \sum_{i=1}^{n} \left\{ \sum_{j\in J_L^C} \frac{x_{ij}}{\|\mathbf{x}_j\|^2} \langle \boldsymbol{\mu}, \mathbf{x}_j\rangle \right\}^2 \\ &= \sum_{i=1}^{n} \left\{ \sum_{j\in J_L^C} \beta_j x_{ij} \right\}^2 \\ &= \sum_{i=1}^{n} \sum_{j\in J_L^C} \beta_j^2 x_{ij}^2 + \sum_{j\in J_L} \sum_{\substack{k\in J_L \\ j\neq k}} \beta_j\beta_k \left(\sum_{i=1}^{n} x_{ij}x_{ik} \right) \\ &= \sum_{i=1}^{n} \sum_{j\in J_L^C} \beta_j^2 x_{ij}^2 . \end{aligned} \tag{2.60}$$

If (2.59) and (2.60) are substituted in (2.55), the ESEE is therefore expressed as the sum of n terms as follows:

$$E\left\|P_L\mathbf{Y}-\boldsymbol{\mu}\right\|^2=\sum_{i=1}^{n}\left\{\sigma^2\sum_{j\in J_L}\frac{x_{ij}^2}{\left\|\mathbf{x}_j\right\|^2}+\sum_{j\in J_L^C}\beta_j^2x_{ij}^2\right\}. \tag{2.61}$$

The ith term of (2.61) is unbiasedly estimated by

$$\widehat{\sigma}^2\sum_{j\in J_L}\frac{x_{ij}^2}{\left\|\mathbf{x}_j\right\|^2}+\sum_{j\in J_L^C}x_{ij}^2\left(\widehat{\beta}_j^2-\widehat{\sigma}^2\frac{1}{\left\|\mathbf{x}_j\right\|^2}\right).$$

This follows since $\sigma^2\sum_{j\in J_L}\frac{x_{ij}^2}{\|\mathbf{x}_j\|^2}$ is unbiasedly estimated by $\widehat{\sigma}^2\sum_{j\in J_L}\frac{x_{ij}^2}{\|\mathbf{x}_j\|^2}$ and

$$\begin{aligned}E\left(\sum_{j\in J_L^C}\widehat{\beta}_j^2x_{ij}^2\right) &= \sum_{j\in J_L^C}\beta_j^2x_{ij}^2+\sum_{j\in J_L^C}x_{ij}^2Var\left(\widehat{\beta}_j\right)\\ &= \sum_{j\in J_L^C}\beta_j^2x_{ij}^2+\sum_{j\in J_L^C}x_{ij}^2Var\left\langle\widehat{\boldsymbol{\beta}},\mathbf{u}_j\right\rangle\\ &= \sum_{j\in J_L^C}x_{ij}^2\left(\beta_j^2+\sigma^2\mathbf{u}_j'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{u}_j\right).\end{aligned} \tag{2.62}$$

Since $\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_m$ are mutually orthogonal it follows that

$$\mathbf{X}'\mathbf{X}=\begin{bmatrix}\mathbf{x}_1'\\ \mathbf{x}_2'\\ \vdots\\ \mathbf{x}_m'\end{bmatrix}\begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_m\end{bmatrix}=diag\left(\left\|\mathbf{x}_1\right\|^2,\left\|\mathbf{x}_2\right\|^2,...,\left\|\mathbf{x}_m\right\|^2\right).$$

The jth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ is therefore $\frac{1}{\|\mathbf{x}_j\|^2}$. Consequently, the final expression in (2.62) equals

$$\sum_{j\in J_L^C}x_{ij}^2\left(\beta_j^2+\sigma^2\frac{1}{\left\|\mathbf{x}_j\right\|^2}\right).$$

CHAPTER 3

SELECTION INFLUENTIAL DATA CASES

3.1 Introduction

Consider a multiple linear regression setting, and suppose that variable selection has to be done. What do we mean in this context by a selection influential data case? In Section 1.3 of Chapter 1, a data case is defined as selection influential if the set of variables that is selected from the full data set (i.e., the data set containing all the data cases) differs from the set of variables selected from the reduced data set (i.e., the data set without the case under consideration), or alternatively, given that the set of selected variables remains unchanged, if the fitted model changes significantly upon omission of the data case being considered. From this definition it is clear that variable selection has to be repeated on the reduced data set if we wish to determine whether a given data case is selection influential. In fact, identifying data cases whose omission leads to a change in the set of selected variables is quite simple. Once variable selection has been performed on the full data set, the following simple steps can be performed to establish which of the data cases is selection influential in the sense of changing the set of selected variables if they are omitted:

- omit a data case to obtain a reduced data set
- apply the same variable selection technique that was used to obtain the set of predictor variables from the full data set, to the reduced data set
- compare the set of selected variables obtained from the reduced data set with the set of selected variables obtained from the full data set
- whenever these two sets of selected variables differ, the data case in question is deemed selection influential.

In Section 3.2 we apply these steps to three example data sets, thereby illustrating the effect that omitting an individual data case may have on the variable selection process. Attention is restricted to identifying selection influential cases whose omission causes the set of selected predictor variables to change. Data cases whose omission does not change the set of selected variables, but does lead to significant changes in the fit of the identified model, are not considered in this section. In Section 3.3, however, we study the effect of both types of selection influential data cases in a limited simulation study.

An interesting question is whether one can use a traditional influence measure, proposed in a non-selection context, to identify data cases whose omission changes the set of selected variables. Possibly the best known such traditional measure of influence was proposed by Cook (1977). Cook's influence measure (or Cook's distance) was proposed to identify data cases having a significant influence on the set of predicted values obtained from a multiple linear regression model. Calculating Cook's distance for a given data case is based on the well known idea in multiple linear regression of studying the influence of individual cases by separately omitting them from the data, and repeating the analysis on the reduced data set. This is then followed by a comparison of the results obtained from the two analyses, i.e. the analysis based on the full data set, and the analysis based on the reduced data set. In the case of Cook's distance, the *prediction vector* is recalculated after a data case has been omitted. More specifically, if $P_M\mathbf{Y}_{(-i)}$ denotes the prediction vector calculated on the data set without case i, Cook's distance for case i is defined by

$$D_i = \frac{\left\|P_M\mathbf{Y} - P_M\mathbf{Y}_{(-i)}\right\|^2}{(m+1)\widehat{\sigma}^2}, \qquad i = 1, ..., n. \tag{3.1}$$

In (3.1), $\widehat{\sigma}^2$ is the estimate of the error variance obtained from the full data set. It is important to note that $P_M\mathbf{Y}_{(-i)}$ in (3.1) contains a prediction for case i, although this particular case is not used in fitting a model to $\mathbf{Y}_{(-i)}$.

In order to use Cook's distance to decide whether a given data case should be deemed influential, consider the following $(1-\alpha) \times 100\%$ confidence region for the unknown vector $\boldsymbol{\mu}$ (equal to $\mathbf{X}\boldsymbol{\beta}$ in the multiple linear regression case), viz. the set of all points $\boldsymbol{\mu}$ such that

$$\frac{\left\|P_M\mathbf{Y} - \boldsymbol{\mu}\right\|^2}{(m+1)\widehat{\sigma}^2} \leq F_{m+1,n-m-1,\alpha},$$

where $F_{m+1,n-m-1,\alpha}$ is the $(1-\alpha)th$-percentile of the central F-distribution with $(m+1)$ and $(n-m-1)$ degrees of freedom. This is of course a valid $(1-\alpha) \times 100\%$ confidence region for $\boldsymbol{\mu}$ only if the random error variable $\boldsymbol{\varepsilon}$ in (1.3) follows an $N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ distribution. Although

D_i in (3.1) does not follow an F-distribution, an observed value of D_i close to $F_{m+1,n-m-1,\alpha}$ implies that the removal of the ith case moves the least squares estimate, $P_M\mathbf{Y}$, of $\boldsymbol{\mu}$ to the edge of a $(1-\alpha)\times 100\%$ confidence region for $\boldsymbol{\mu}$. In this regard, Weisberg (1985) argues that observed values of D_i larger than 1, typically correspond to movement to the edge of a 50% confidence region and beyond. In view of these considerations, it seems to have become common practice to identify data cases with observed values of D_i larger than 1 as influential according to Cook's distance.

In each of the following examples, Cook's distance is calculated for all the data cases in order to investigate whether this measure succeeds in identifying selection influential cases. We will see that although this does indeed happen in some cases, i.e. data cases deemed to be influential when judged in terms of Cook's distance are also found to be *selection* influential, there are other instances where cases which are strongly selection influential do not give particularly large values for Cook's distance. It therefore seems that selection influential data points can not always be identified by making use of traditional measures of influence. This is not unexpected, especially in high dimensional cases. A data case which is selection influential because of its nature in lower dimensional space, may not always be deemed influential when viewed in terms of all its coordinates. Since a measure such as Cook's distance evaluates data cases in terms of all their coordinates, it is understandable that not all selection influential cases will be identified by such a measure.

3.2 Illustrative examples

3.2.1 The Hald data

In our first example, we consider the much analysed cement hardening data of Hald given by Draper and Smith (1998). The data set has 4 predictor variables with 13 observations. Each predictor variable represents the amount of a specific chemical substance used in cement to manufacture clinkers. A complete description of these variables, together with the data set, is given in Table B.1 of Appendix B.

The full regression model includes an intercept term and is given by

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

where we assume that $\varepsilon \backsim N(0, \sigma^2)$. The full model error variance estimate, defined in (1.10), is $\widehat{\sigma}^2 = 5.983$. Least squares estimation of β_0 and the four regression coefficients results in the following estimated regression model:

$$\widehat{Y} = 62.405 + 1.551x_1 + 0.510x_2 + 0.102x_3 - 0.144x_4.$$

Since an intercept term is included in the regression model, there are $\sum_{l=0}^{4} \binom{4}{l} = 16$ possible regression models considered for selection. The corresponding 16 linear subspaces contained in the linear space M are spanned by the vector $\mathbf{1}$, together with an appropriate subset of the vectors formed by the observations on the predictor variables. Note that M itself is spanned by the vector $\mathbf{1}$ and the vectors corresponding to all four the predictor variables. Applying C_p selection to the full data set yields a minimum C_p criterion of 2.678, with predictor variables 1 and 2 being selected. In order to demonstrate the effect that individual data cases have on the variable selection process, each of the 13 data cases is omitted one at a time. The selection process is then repeated on the reduced data sets with only 12 data cases. The minimum value of the C_p criterion, and the corresponding predictor variables which are selected on each of the 13 reduced data sets, are given in Table 3.1.

Program C2 in Appendix C was utilised for calculating the respective minimum C_p criteria and selected variables from the full and reduced data sets. Also shown in Table 3.1 is Cook's distance for each of the 13 data cases. Program C3 in Appendix C was utilised for this purpose.

As stated previously, predictor variables 1 and 2 are selected on the full data set. From Table 3.1 it is clear that, except for cases 3, 6 and 8, the separate omission of all other data cases also results in predictor variables 1 and 2 being selected. However, omitting case 3 results in variables 1, 3 and 4 being selected, whereas the separate omission of both cases 6 and 8 results in variables 1, 2 and 3 being selected. Since omitting any one of data cases 3, 6 and 8 leads to a different model being selected than the model selected from the full data set, these cases are identified as selection influential. Note that although these cases can at this stage be judged to be selection influential, it is still uncertain whether their influence on subsequent analysis, such as response prediction, is beneficial or detrimental.

Data case omitted	Cp criterion	Selected variables	Cook's distance
case 1	3.125	1,2	0.000002
case 2	3.429	1,2	0.057225
case 3	2.964	1,3,4	0.300863
case 4	3.072	1,2	0.059270
case 5	3.325	1,2	0.001821
case 6	1.061	1,2,3	0.083369
case 7	3.234	1,2	0.064285
case 8	1.172	1,2,3	0.393533
case 9	3.000	1,2	0.037532
case 10	2.989	1,2	0.020677
case 11	1.496	1,2	0.170840
case 12	3.523	1,2	0.015322
case 13	1.898	1,2	0.110239

Table 3.1: The C_p criterion, selected variables and Cook's distance calculated for each of the data cases of the Hald data

It is important to bear in mind that a restricted definition of the term "*selection influential*" is applied when only those data cases whose omission leads to a change in the model which is selected, are deemed selection influential. It may easily happen that omission of a given data case leaves the set of predictor variables unchanged, but that the two fitted models in question differ to such an extent that subsequent analyses from the two models, for example, prediction of future cases, give significantly different results. Clearly, such data cases should also be classified as selection influential. It is obviously not as easy to identify selection influential cases of the latter type as it is to identify cases whose omission causes the set of selected variables to change. At this point, however, our objective is not to identify selection influential cases of this type. A selection influence measure capable of identifying such cases is developed in Section 4.2. Our objective for the moment is merely to illustrate the effect that separate omission of data cases such as cases 3, 6 and 8 of the Hald data, can have on the subset of predictor variables which is selected.

How does Cook's distance fare in identifying the selection influential data cases in the Hald data? From Table 3.1 we see that the largest Cook's distance of nearly 0.4 is obtained if case 8 is omitted. Since $F_{5,8,0.836} \approx 0.4$, it implies that omitting case 8 move $P_M\mathbf{Y}$ to the edge of a $(1 - 0.836) \times 100\% = 16.4\%$ confidence region, which is a relatively small movement. Since the data case corresponding to the largest Cook's distance is therefore not deemed to be influential, the same judgement will be applicable to all the other observations in the Hald data. It is therefore clear that for this particular data set, Cook's distance does not identify selection influential data cases, such as cases 3, 6 and 8, as influential. This is not always the situation, as will be seen in the next example.

3.2.2 The fuel data

We further illustrate the effect that individual observations has on the variable selection process by using the fuel data of Weisberg (1985, p.35-36, and 126). This data set contains observations for each of the 50 states in the USA. There are 4 predictor variables. The response variable is the 1972 fuel consumption (in gallons per capita). The complete data set, with a description of the predictor variables, is given in Table B.2 of Appendix B.

The full model is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

where we assume that $\varepsilon \backsim N(0, \sigma^2)$. The estimate of the error variance obtained from the full model is $\widehat{\sigma}^2 = 7452.009$. When an intercept term is included in the regression model, there are 16 possible models to be considered for selection. The full estimated regression model is given by

$$\widehat{Y} = 235.534 - 8.171x_1 + 11.886x_2 - 68.788x_3 + 2.868x_4.$$

Predictor variables 2 and 3 are selected on the full data set with minimum $C_p = 2.517$. Table 3.2 shows the value of the minimum C_p criterion and the corresponding selected variables obtained from each of the reduced data sets resulting when a single data case is omitted from the full data set. Cook's distance in (3.1) is also calculated for each of the data cases and presented in Table 3.2. From Table 3.2 it is clear that, except for the reduced data sets obtained if cases 40, 49 and 50 are respectively omitted, the selection process on all the other reduced data sets also results in predictor variables 2 and 3 being selected. Cases 40, 49 and 50 are therefore definitely identified as selection influential. Also of interest in this data set is the fact that the value of the C_p criterion corresponding to these three reduced data sets is negative. This of course contradicts the fact that the C_p criterion is an estimate of the ESEE in (2.2), which is always non-negative.

For the fuel data, Cook's distance exceeds the value 1 for case 50. As argued earlier, this is a widely used, albeit somewhat arbitrary criterion for deciding whether a given data case is influential in terms of Cook's distance. Evidently, case 50 should be regarded as influential. It is also clear that the Cook distances obtained for cases 40 and 49 are relatively larger than the rest, suggesting that cases 40 and 49 may potentially be influential. Note that cases 40, 49 and 50 are exactly those identified as being selection influential. We therefore conclude that in the case of the fuel data, the data cases identified by Cook's distance as influential, are exactly the cases identified as selection influential by the leave-one-out strategy and repeated application of C_p selection.

Data case omitted	Cp criterion	Selected variables	Cook's distance
case 1	3.431	2,3	0.0000758
case 2	2.968	2,3	0.0036505
case 3	3.147	2,3	0.0022111
case 4	2.876	2,3	0.0104932
case 5	1.471	2,3	0.0253789
case 6	3.243	2,3	0.0013565
case 7	3.497	2,3	0.0112335
case 8	3.420	2,3	0.0005405
case 9	3.220	2,3	0.0047663
case 10	3.339	2,3	0.0033309
case 11	2.937	2,3	0.0039010
case 12	3.495	2,3	0.0022333
case 13	3.486	2,3	0.0007867
case 14	3.315	2,3	0.0024369
case 15	3.187	2,3	0.0046100
case 16	3.225	2,3	0.0014734
case 17	3.441	2,3	0.0001039
case 18	0.763	2,3	0.0377199
case 19	0.592	2,3	0.1248168
case 20	3.009	2,3	0.0159463
case 21	3.483	2,3	0.0016878
case 22	3.443	2,3	0.0000626
case 23	3.510	2,3	0.0019533
case 24	3.311	2,3	0.0049361
case 25	2.870	2,3	0.0040405
case 26	3.495	2,3	0.0000001
case 27	3.427	2,3	0.0010035
case 28	3.470	2,3	0.0000911
case 29	3.514	2,3	0.0000230
case 30	3.503	2,3	0.0013168
case 31	3.467	2,3	0.0001371
case 32	3.491	2,3	0.0020462
case 33	2.341	2,3	0.0327127
case 34	3.454	2,3	0.0013042
case 35	3.337	2,3	0.0028581
case 36	3.362	2,3	0.0050388
case 37	3.073	2,3	0.0000004
case 38	2.487	2,3	0.0067815
case 39	2.506	2,3	0.0244878
case 40	-7.188	2,3,4	0.2161615
case 41	3.272	2,3	0.0023620
case 42	2.309	2,3	0.0172968
case 43	3.455	2,3	0.0006214
case 44	3.035	2,3	0.0152122
case 45	0.090	2,3	0.1696183
case 46	3.179	2,3	0.0024320
case 47	3.425	2,3	0.0008320
case 48	3.418	2,3	0.0051837
case 49	-1.863	2,3,4	0.2774447
case 50	-10.443	1,2,3	1.4696553

Table 3.2: The C_p criterion, selected variables and Cook's distance calculated for each of the data cases of the fuel data

3.2.3 The evaporation data

In our final example, we show the effect of individual selection influential observations by applying the C_p criterion to the evaporation data given by Freund (1979). This data set is also analysed by Becker et al. (1988) and Snyman (1994). Ten independent predictor variables are measured on 46 consecutive days, together with the amount of evaporation from the soil, which represents the response.

The complete data set, with a description of the predictor variables, is given in Table B.3 of Appendix B. Nine of the 10 predictor variables fall into three groups of three highly correlated variables each so that the full model can be written as

$$Y = \beta_0 + \sum_{j=1}^{3} \beta_j x_j + \sum_{j=4}^{6} \beta_j x_j + \sum_{j=7}^{9} \beta_j x_j + \beta_{10} x_{10} + \varepsilon$$

where we once again assume that $\varepsilon \backsim N(0, \sigma^2)$. The estimated error variance, calculated from the full model, is $\widehat{\sigma}^2 = 42.351$. If we assume that it may happen that no predictor variables are selected (i.e., the model contains only the intercept term), there are 2^{10} possible models to be considered for selection. Calculating the C_p criterion for the complete data set yields a minimum value of 3.759, with variables 1,3,6,8 and 9 being selected. The least squares estimate of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}} = \left[\begin{array}{ccccccccccc} -54.075 & 2.232 & 0.205 & -0.743 & 0.501 & 0.304 & 0.092 & 1.110 & 0.751 & -0.556 & 0.009 \end{array} \right]'.$$

As in the previous two illustrative examples, we once again apply C_p selection to the reduced data sets obtained by omitting case i, $i = 1, ..., 46$, from the full data set. The resulting minimum C_p values and the corresponding selected predictor variables for these reduced sets, are presented in Table 3.3. The highlighted cases in the first column of Table 3.3 are the cases which, if separately omitted, results in other predictor variables being selected on the corresponding reduced data set than predictor variables 1,3,6,8 and 9, the variables that are selected if the full data set is used. These highlighted cases are therefore considered selection influential, since omission of any one of these cases causes a change in the subset of selected predictor variables.

Cook's distance calculated for each of the 46 observations of the evaporation data indicates that none of the observations should be classified as influential (since none of these Cook's distances exceed the value 1). Nevertheless, we highlight distances in Table 3.3 which are relatively larger (i.e., distances larger than 0.2) than the other distances, thereby identifying potentially influential cases. The cases corresponding to these relatively larger Cook's distances are however not always selection influential in the sense that a different subset of predictor variables is selected. Consider foı example data case 2. If this case is omitted, the subset of predictor variables selected on the corresponding reduced data set is identical to the subset selected on the complete data set. Data case 2 is therefore not selection influential in the sense that a different subset of predictor variables is selected when this case is omitted from the full data set. However, its Cook's distance of 0.294 is relatively large. On the other hand, some data cases, for example case 18, which is definitely selection influential in the sense that a different subset of predictor variables is selected, has a very small Cook's distance. It is therefore clear that for the evaporation data, Cook's distance is definitely not always successful in identifying selection influential data cases.

Data case omitted	Cp criterion	Selected variables	Cook's distance
case 1	4.656	1,3,6,8,9	0.004795
case 2	-1.131	1,3,6,8,9	0.293887
case 3	3.539	1,3,6,8,9	0.026810
case 4	4.759	1,3,6,8,9	0.000022
case 5	3.926	1,3,6,8,9	0.047007
case 6	3.878	1,3,6,9	0.029748
case 7	4.631	1,3,6,8,9	0.001364
case 8	-0.408	1,3,6,9,10	0.126389
case 9	4.738	1,3,6,8,9	0.002053
case 10	4.585	1,3,6,9,10	0.002574
case 11	4.756	1,3,6,8,9	0.004128
case 12	4.481	1,3,6,8,9	0.000561
case 13	4.625	1,3,6,8,9	0.006815
case 14	4.743	1,3,6,8,9	0.000605
case 15	4.757	1,3,6,8,9	0.000585
case 16	4.549	1,3,6,8,9	0.000769
case 17	4.756	1,3,6,8,9	0.000441
case 18	4.573	1,3,6,9	0.000238
case 19	4.742	1,3,6,8,9	0.004860
case 20	4.744	1,3,6,8,9	0.002742
case 21	2.654	1,3,6,8,9	0.063002
case 22	2.974	1,3,6,9,10	0.049839
case 23	4.629	1,3,6,8,9	0.000421
case 24	1.722	1,3,6,9	0.059969
case 25	4.756	1,3,6,8,9	0.000003
case 26	3.714	1,3,5,7,8,9	0.063056
case 27	4.266	1,3,6,9,10	0.006986
case 28	4.756	1,3,6,8,9	0.000186
case 29	4.667	1,3,6,9,10	0.003374
case 30	4.750	1,3,6,8,9	0.000919
case 31	3.067	1,3,4,8,9	0.427610
case 32	3.148	1,3,6,9,10	0.268503
case 33	-3.262	1,3,6,9,10	0.180341
case 34	4.759	1,3,6,8,9	0.006716
case 35	4.759	1,3,6,8,9	0.000001
case 36	4.520	1,3,6,8,9	0.004725
case 37	1.624	1,3,6,8,9	0.061130
case 38	2.640	1,3,6,8,9	0.087699
case 39	4.002	1,3,6,8,9	0.012991
case 40	2.210	6,9,10	0.038762
case 41	-5.461	1,3,6,9	0.287897
case 42	4.018	1,3,6,8,9	0.011932
case 43	4.759	1,3,6,8,9	0.000053
case 44	4.755	1,3,6,8,9	0.000916
case 45	4.756	1,3,6,8,9	0.002249
case 46	4.252	1,3,6,9,10	0.006726

Table 3.3: The C_p criterion, selected variables and Cook's distance calculated for each of the data cases of the evaporation data

3.3 Simulation study

The three illustrative examples in the previous section confirm the possibility of selection influential data cases being present in a data set. From these examples we also learn that traditional influence measures, such as Cook's distance, are not always helpful in identifying selection influential data cases. An explanation for this must be sought in the fact that traditional measures of influence are calculated using the full set of predictor variables. Selection influential data cases are frequently influential only with respect to a subset of the predictor variables, and will typically not be identified as influential by the traditional influence measures. When investigating the influence of data cases if variable selection is applied in multiple linear regression, it is therefore important to consider the influence of such cases in lower dimensional spaces as well, and not only their influence on the regression model fitted to the complete set of predictor variables.

We earlier defined a data case to be selection influential if the set of variables that is selected from the full data set (i.e., the data set containing all the data cases) differs from the set of variables selected from the reduced data set (i.e., the data set without the case under consideration), or alternatively, given that the set of selected variables remains unchanged, if the fitted model changes significantly upon omission of the data case being considered. In our analysis of the three example data sets we focused on the first aspect of this definition, and we did not attempt to investigate the extent to which selection influential data cases may influence subsequent post-selection analyses of the data. In this section we therefore report on a simulation study that was undertaken to conduct such a deeper investigation into the effects of selection influential data cases.

The primary aim of the simulation study of this section was to study in detail the effect that inclusion of a selection influential data case has on the set of variables that is selected, as well as the effect on the predicted values obtained from the selected model. Note that in this process we focus on selection influential data cases that are *included* in the data set before variable selection is done, rather than on the effect that *exclusion* of selection influential data cases may have on the results of our analysis. Also, we do not try to investigate procedures that may be used to *identify* selection influential points; we merely investigate the results if such a case is deliberately inserted into a data set.

How would one go about deliberately inserting a selection influential data case into a data set? Consider the following illustrative example. A multivariate data set of 20 observations is first of all simulated from a $N_5(\mathbf{0}, \mathbf{I}_5)$ distribution. This multivariate data set will constitute the observations of 5 uncorrelated predictor variables, x_1, x_2, x_3, x_4 and x_5. Twenty additional values are also simulated from a $N(0,1)$ distribution. These values, denoted by $\widehat{\varepsilon}_1, \widehat{\varepsilon}_2, ..., \widehat{\varepsilon}_{20}$, constitute realisations of the random component of the multiple linear regression model. We now let $\beta_0 = 1$ and $\beta_1 = \beta_2 = 3$, $\beta_3 = \beta_4 = \beta_5 = 0$. Using (1.1), we can now calculate the ith value of the response variable, Y_i, $i = 1, 2, ..., 20$, viz.

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \widehat{\varepsilon}_i \\ &= 1 + 3x_{i1} + 3x_{i2} + \widehat{\varepsilon}_i. \end{aligned}$$

The resulting regression sample includes observations on the response variable and each of the 5 predictor variables for each of the 20 data cases. These data will be referred to as the ordinary data set. If C_p selection is applied to this data set it results in predictor variables x_1 and x_2 being selected. This is not surprising, since the corresponding regression coefficients of these predictors are large.

Consider now the linear relationship between Y and the predictor variable x_1. The scatter plot and least squares regression line for (Y_i, x_{i1}), where $i = 1, ..., 20$ are shown in Figure 3.1. The coefficient of determination, r^2, for this fitted regression line implies that 65% of the variation in Y is explained by variation in x_1.

We now weaken the linear relationship between Y and x_1 by replacing the largest value of x_1 by its smallest (i.e., for our example we replace 2.18 by -2.87). All other data cases in the regression sample are kept unchanged. For the case where the value of x_1 has been changed, the corresponding Y-value, which equals 6.47, and the values of x_2, x_3, x_4 and x_5 are left unchanged. The change in the regression sample therefore only affects the largest value of x_1.

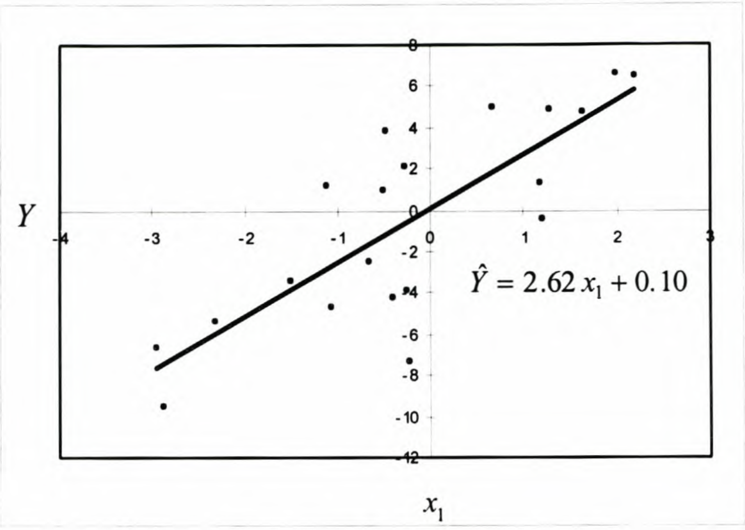


Figure 3.1: Scatter plot of Y and x_1. A least squares regression line is fitted to the data.

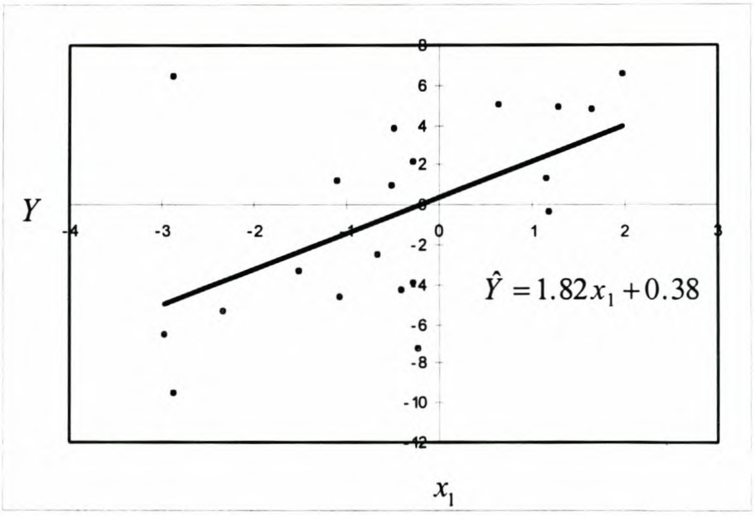


Figure 3.2: Scatter plot of Y and x_1, where the largest value of x_1 has been replaced by its smallest. A least squares regression line is also fiited to this data.

If a least squares regression line is fitted to Y and the "new" x_1, only 31% of the variation in Y is explained by variation in x_1. The change in x_1-values therefore definitely causes the linear relationship between Y and x_1 to deteriorate. The scatter plot and least squares regression line of Y and the "new" x_1, are shown in Figure 3.2. Clearly, if we study this scatter plot, the "new" data point, with coordinates $(-2.87; 6.47)$ will arouse suspicion and certainly be looked upon as an outlier. However, since its influence is evident only if the linear relationship between x_1

and the response is considered, it will probably not be regarded stray if seen in the context of the complete multivariate regression sample.

The complete regression sample that includes the data case with altered x_1-value will be referred to as the modified data set. If C_p selection is applied to the modified data set, predictors x_1, x_2 and x_5 are selected. The data case that had its x_1-value changed can therefore definitely be identified as selection influential, since we now select a set of predictor variables that differs from the set that was selected before. Note that application of variable selection to a modified data set will not always lead to a change in the subset of variables that is selected. This will depend on the magnitude of the change, and on the configuration of the other data points in the multivariate data set. We will therefore refer to a data point that has had one of its coordinates adjusted in the manner described above as *possibly* selection influential.

There are of course many different ways to insert a potentially selection influential data point into a data set. The method that we described above replaces the largest value of an important predictor (i.e., a predictor with a large absolute regression coefficient value) by its corresponding smallest value. Clearly, replacing the smallest value of an important predictor by its largest value will also cause the linear relationship between this predictor and the response to weaken. We could also alter a value of a predictor in such a way that the linear relationship between the response and that predictor is strengthened rather that weakened. It would also be possible to weaken or strengthen the linear relationship between the response and a *subset* of the predictor variables by altering values of the predictors in the subset under consideration. Another possibility would be to change a value of the response variable. For example, we could replace the largest response observation with the smallest, or the other way round. This will typically tend to weaken the linear relationship between the response and each of the individual important predictors. In our limited simulation study we restricted attention to the method that replaces the largest observed value of an important predictor by its corresponding smallest value to weaken the linear relationship between the response variable and the predictor concerned. We also looked only at cases where a single possibly selection influential data point is inserted into the data. It would of course be possible to extend this to situations where more than one such point are inserted.

Our objective in the simulation study is to investigate the effect of a single possibly selection influential data case that is inserted into a data set. We measure this effect in terms of the subset of predictors that is selected, and in terms of the predicted response values that are obtained from the fitted model. Possibly selection influential data points are inserted into data sets by using the method described above, i.e. by replacing the largest observed value of an important predictor by its corresponding smallest value, thereby obtaining the so-called modified data set. Throughout the discussion we focus on variable selection based on the C_p criterion. The effect referred to above is measured by comparing two post-selection fitted models: the model obtained by doing variable selection and model fitting on the ordinary data set, and the model obtained similarly from the modified data set.

Consider firstly a comparison of the two selected and fitted models in terms of the selected predictors. In as simulation study the values of the regression coefficients are of course known. It is therefore possible to identify a so-called "correct" model, i.e. the model that contains all the predictors with significantly non-zero regression coefficients, and none of the predictors with regression coefficients (close to) zero. A desirable property of a variable selection technique is that it should select this "correct" model with high probability. The term "probability of correct selection" (PCS) will be used in this regard. The PCS of a selection technique is therefore the probability, under repeated sampling from the underlying distribution, of selecting all the predictors with significantly non-zero regression coefficients, but none of the predictors with regression coefficients (close to) zero. In our simulation we approximate the PCS by the empirical proportion of times that the "correct" model is selected. This is done for selection on the ordinary data set, as well as for selection on the modified data. A comparison of the two empirical proportions reveals the extent to which the PCS is influenced by the insertion of possibly selection influential data points.

Consider secondly a comparison of the two selected and fitted models in terms of their respective *average prediction errors* (APEs). Let $\mathbf{Y}^*$ denote a new (future) n-component response vector that has to be predicted. It is assumed that $\mathbf{Y}^*$ is obtained from the same mechanism (distribution) that yielded the ordinary regression sample. Note in particular that none of the components of $\mathbf{Y}^*$ has therefore been altered in any way. Let $\widehat{\mathbf{Y}}^*$ denote a prediction of $\mathbf{Y}^*$ obtained by utilising a fitted multiple linear regression model. In our context, $\widehat{\mathbf{Y}}^*$ will either be the prediction obtained from the model fitted to the ordinary data set, or the prediction obtained from the model fitted to the modified data set. The APE of $\widehat{\mathbf{Y}}^*$ is simply $E\left\|\widehat{\mathbf{Y}}^* - \mathbf{Y}^*\right\|^2$. This unknown quantity can be approximated in our simulation by replacing the expectation by an empirical average over a sufficiently large number of simulation repetitions. We do this for $\widehat{\mathbf{Y}}^*$ obtained from the ordinary data set, and for $\widehat{\mathbf{Y}}^*$ obtained from the modified data set. A comparison of the two empirical averages provides information regarding the influence on APE of the possibly selection influential case inserted into the modified data set.

3.3.1 Design of the simulation study

As stated in the previous section, the objective of the simulation study is to investigate the effect on the PCS and the APE of a post-selection fitted multiple linear regression model if a possibly selection influential data point is inserted into the data set. It is assumed throughout that variable selection is done using the C_p criterion. The stated objective is achieved by comparing two simulation estimates of the PCS, and two simulation estimates of the APE. The first member of the set of two estimated PCS-values is obtained from ordinary, unchanged data sets, while the second member is obtained from the corresponding modified data sets. A similar procedure is followed to obtain the two estimated APE-values. Different combinations of the following factors are employed in the simulation study:

- The sample size of the simulated data set. The following sample sizes were used in the study: $n = 20, 50$, and 100.

- The number of predictor variables in the simulated data set. For $n = 20$, we used $m = 5$, and for $n = 50$ and $n = 100$, we used $m = 5$ and $m = 10$.

- The correlation amongst the predictor variables (in this regard we study equi-correlated cases, i.e. cases where the same correlation is assumed to hold for any pair of predictor variables). The common value of the correlation between any two predictors was varied over 0 (the orthogonal regressor case), 0.5 and 0.9.

- The sample data sets are simulated at fixed, predetermined values of the regression coefficients. Two different configurations were used in this regard. In the first case, we set $\beta_1 = \beta_2 = ... = \beta_m = s$, and then increment the common value s from 0 in steps of 0.1 up to 1.5, and thereafter in steps of 0.25 up to 3. In the second regression coefficient configuration we also start by setting $\beta_1 = \beta_2 = ... = \beta_m = 0$, but thereafter only a subset of the β-values are incremented. In particular, for simulated data sets containing $m = 5$ predictors we increment the common value of β_1 and β_2 from 0 in steps of 0.1 up to 1.5, and thereafter in steps of 0.25 up to 3. For the cases where we had $m = 10$ predictors, the values of $\beta_1, \beta_2, \beta_3, \beta_4$ and β_5 are incremented in the same manner.

- It should finally be noted that we consistently used $\beta_0 = 1$ as intercept parameter in the regression models, and that the error variance was, without loss of generality, kept constant at $\sigma^2 = 1$.

Although there are of course many other combinations of the factors listed above that could be investigated in a simulation study, it is hoped that the cases actually covered in our study do provide an indication of the variation in the effect that a possibly selection influential data case may have on the PCS and the APE of a fitted multiple linear regression model.

Consider now a given combination of the factors listed above, i.e. given values of the sample size n, the number of predictor variables m, the correlation ρ amongst the predictor variables, and a given configuration of β-values. The steps followed in the actual simulation study may be summarised as followed.

Step 1: Generate n observations from an m-variate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$, where the diagonal elements of $\mathbf{\Sigma}$ all equal 1, and the off-diagonal elements all equal ρ. These values constitute observations on the m predictor variables $x_1, x_2, ..., x_m$, and they make up the so-called ordinary data set. Create an exact replicate of this data set, and then replace the largest value of predictor variable x_1 by its corresponding smallest value, thereby obtaining the so-called modified data set.

Step 2: Generate n observations from an $N_n(\mathbf{0}, \mathbf{I}_n)$ distribution. These represent observations of the error term in the regression model.

Step 3: Use expression (1.1) to calculate n values of the response variable Y. In this calculation the simulated values of the predictor variables, the simulated values of the error variable, and the relevant values of the regression coefficients are used. The n-component vector that is thereby obtained represents the response in both the ordinary and the modified data sets.

Step 4: Apply C_p variable selection to the ordinary and the modified data sets, thereby obtaining two selected, fitted models.

Step 5: Determine whether the "correct" model was selected on the ordinary data set. If so, put PCS(ordinary, i) $= 1$, otherwise put PCS(ordinary, i) $= 0$. Here, i is used to index the simulation repetitions. Proceed similarly for the modified data set, putting PCS(modified, i) $= 1$ or 0 depending on whether the "correct" model was selected or not on the modified data set.

Step 6: Generate new values of the predictor variables and of the error term as described above, and use these to calculate a new response vector $\mathbf{Y}^*$. Utilise the estimated regression coefficients from the model fitted to the ordinary data set, together with the newly generated values of the predictor variables, to calculate a prediction $\widehat{\mathbf{Y}}^*$(ordinary) for $\mathbf{Y}^*$.
Calculate APE(ordinary, i) $= \left\|\widehat{\mathbf{Y}}^* - \mathbf{Y}^*\right\|^2$. Repeat these steps for the model obtained from the modified data set, thereby obtaining APE(modified, i). Note that the unchanged values of the newly generated predictor variables are used to calculate APE(modified, i).

Step 7: Repeat steps two to six, 200 times, and calculate the average, over i, of the PCS(ordinary, i), PCS(modified, i), APE(ordinary, i) and APE(modified, i) values.

Step 8: Repeat steps one to seven, 50 times, and average the quantities obtained in step 7 over these 50 repetitions.

We include Program C4 in Appendix C as an example of the programs used for obtaining the simulated results. In particular, Program C4 was utilised to obtain the simulated APE and PCS for selected models from data sets of size $n = 20$ with $m = 5$ uncorrelated predictors. The β-configuration in this specific case is $\beta_1 = \beta_2 = ... = \beta_5 = 0(0.1)1.5$ and thereafter $\beta_1 = \beta_2 = ... = \beta_5 = 1.5(0.25)3$.

The simulation results are plotted in Figures 3.3 to 3.10. Figures 3.3 to 3.6 deal with the APE, while Figures 3.7 to 3.10 represent the PCS-results. The cases where we use $m = 5$ predictor variables are dealt with in Figures 3.3, 3.4, 3.7 and 3.8 (these are the plots presented in landscape format), whereas the results for the cases where we use $m = 10$ predictors are represented in Figures 3.5, 3.6, 3.9 and 3.10 (these are the plots presented in portrait format). Consider for example the results depicted in Figure 3.3. The β-configuration in this case is $\beta_1 = \beta_2 = ... = \beta_5 = 0(0.1)1.5$ and thereafter $\beta_1 = \beta_2 = ... = \beta_5 = 1.5(0.25)3$. The solid lines represent the APE of the models selected with C_p from the ordinary simulated data sets, whereas the dotted lines indicate the APE of the models selected with C_p from the corresponding modified data sets. Moving horizontally across the landscape formatted page containing Figure 3.3, the first row of plots represents results obtained for simulated data sets of sizes 20, 50 and 100 respectively, with the equicorrelation parameter $\rho = 0$. The second and third rows of plots are organised similarly, but with $\rho = 0.5$ and $\rho = 0.9$ respectively. The layout in Figure 3.4 is similar to that of Figure 3.3, but now we deal with the β-configuration having $\beta_1 = \beta_2 = 0(0.1)5$, followed by $\beta_1 = \beta_2 = 1.5(0.25)3$, with $\beta_3 = \beta_4 = \beta_5 = 0$ throughout.

Figures 3.5 to 3.10 are organised similarly. In each case a clear indication is provided of the sample size n, the number of predictor variables m, the correlation ρ amongst the predictor variables, and the configuration of β-values.

The results represented in Figures 3.3 to 3.10 are interpreted in the next section.

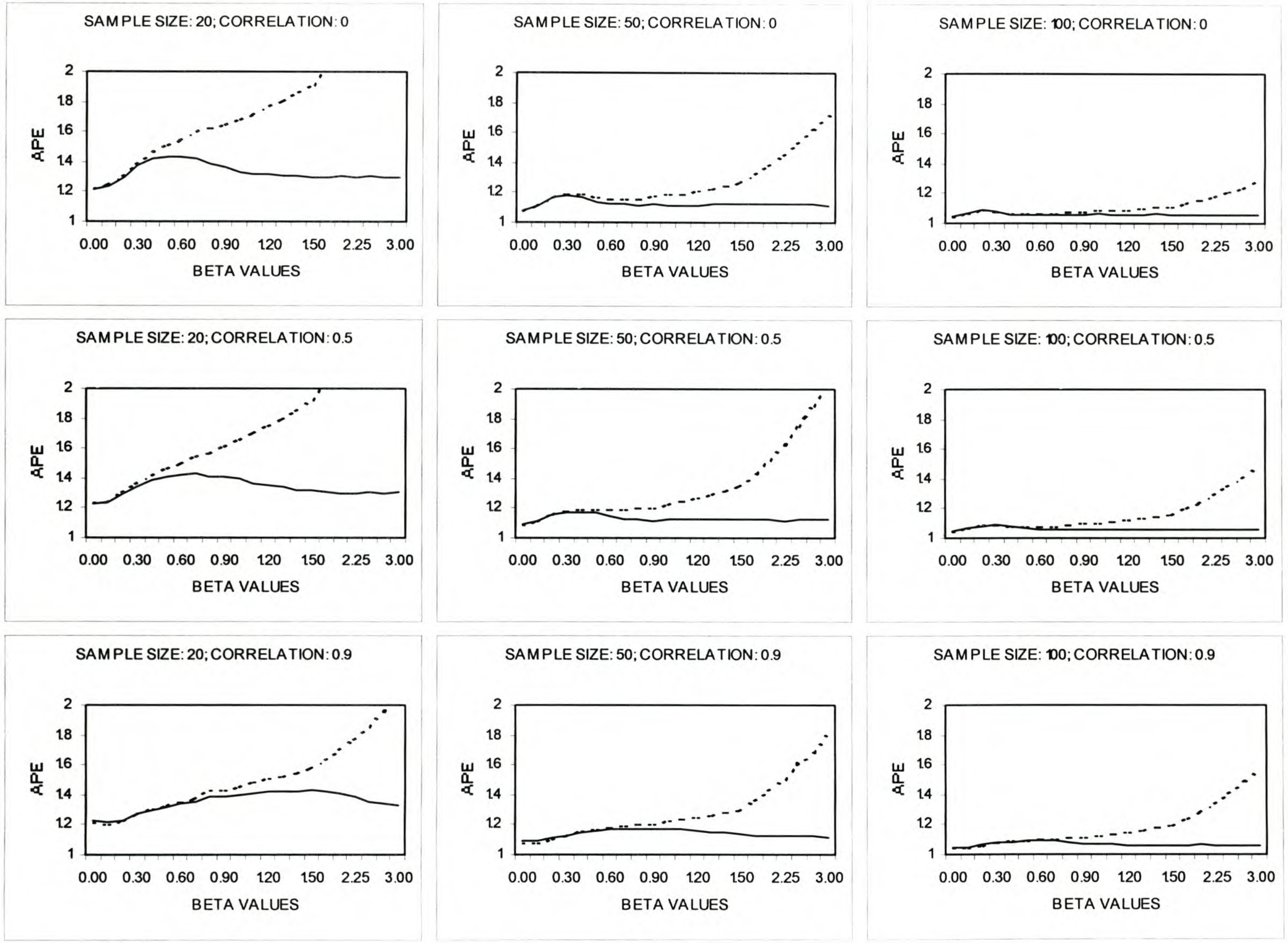


Figure 3.3: The APE of models selected from simulated data sets with $m = 5$ predictors;

$\beta_0 = 1$, $\beta_1 = \beta_2 = \ldots = \beta_5 = 0(0.1)1.5$ and $\beta_1 = \beta_2 = \ldots = \beta_5 = 1.5(0.25)3$

LEGEND: ——: ordinary data set and - - -: modified data set

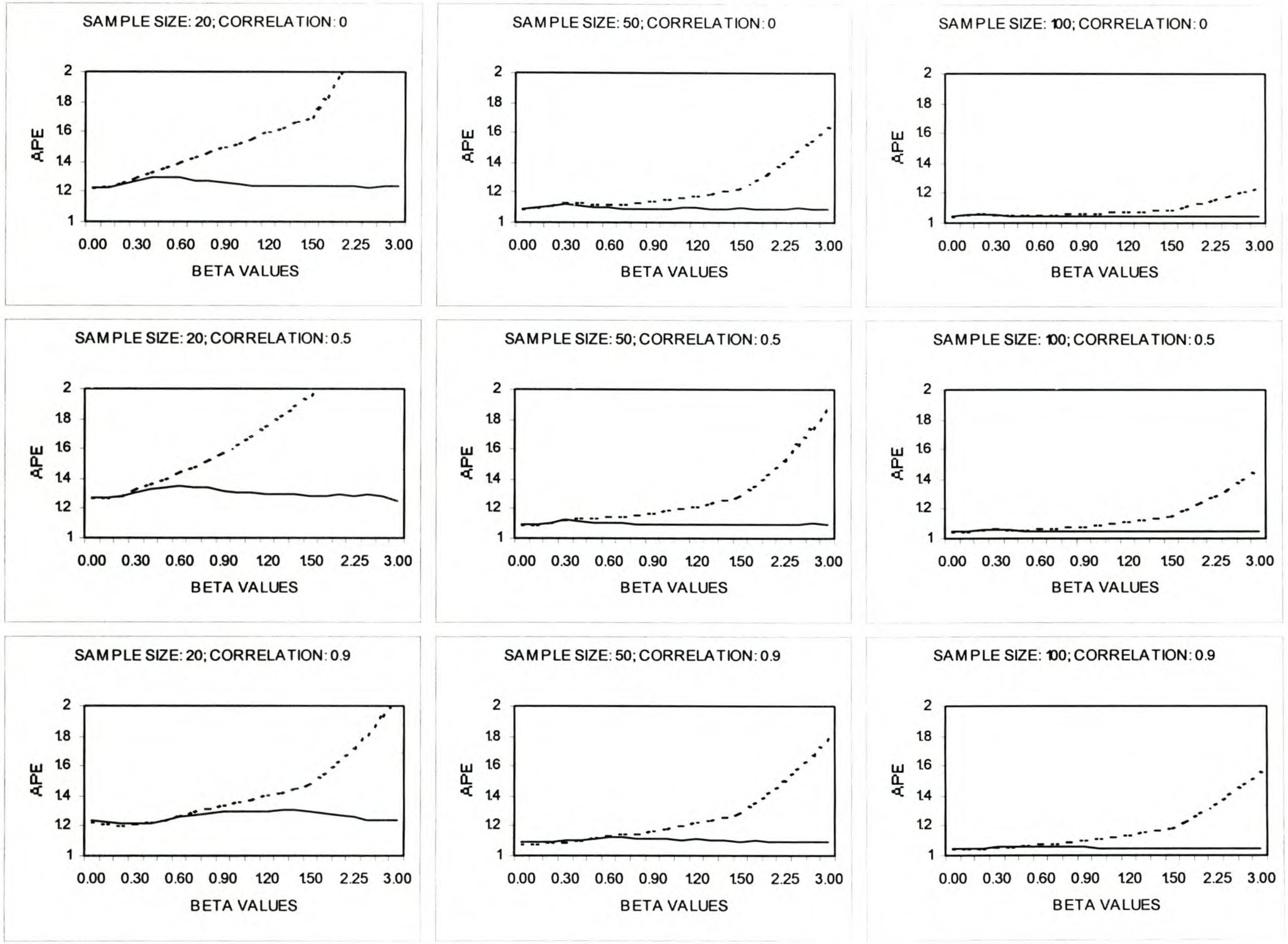
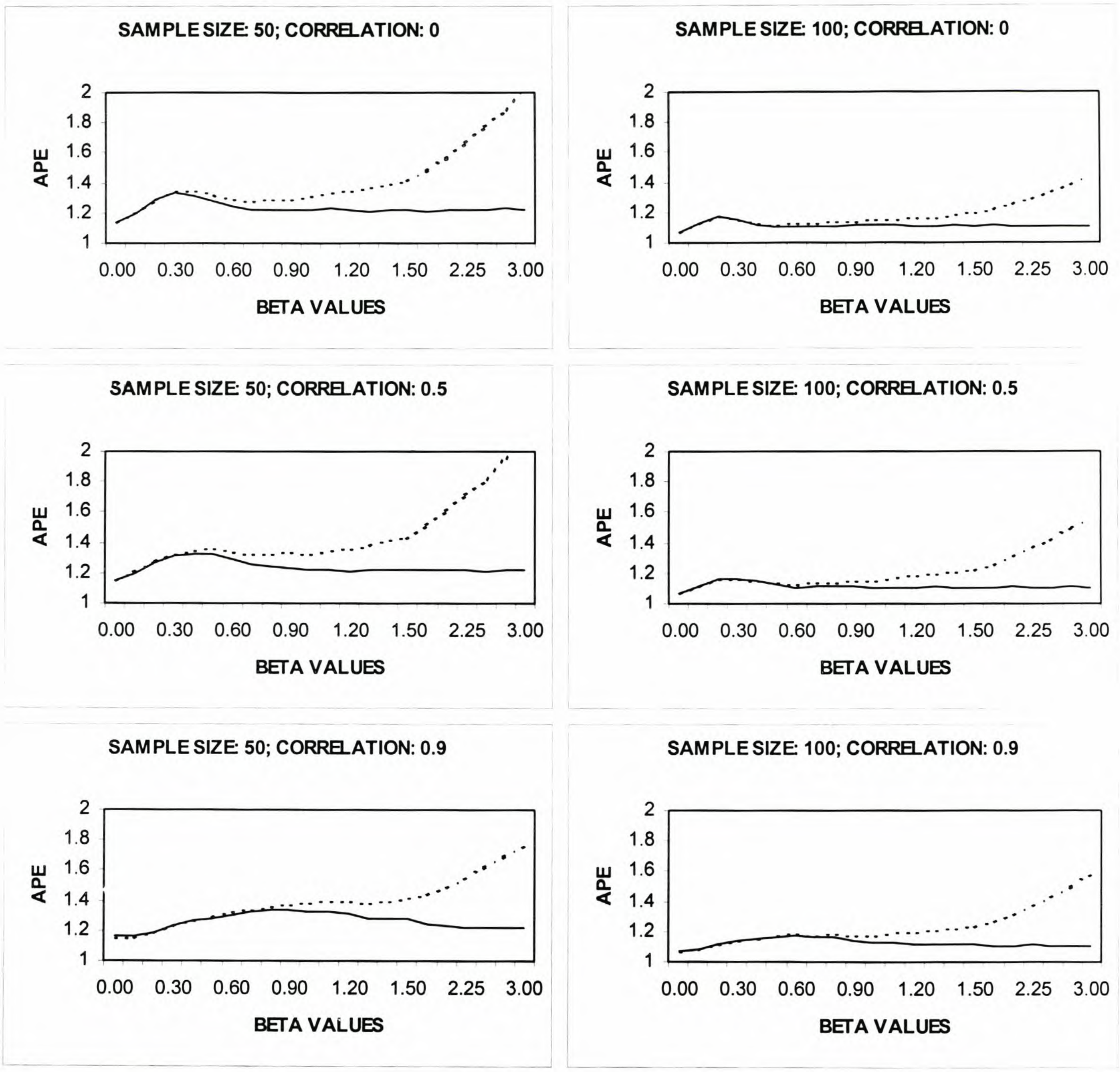


Figure 3.4: The APE of models selected from simulated data sets with $m = 5$ predictors;

$\beta_0 = 1$, $\beta_1 = \beta_2 = 0(0.1)1.5$ and $\beta_1 = \beta_2 = 1.5(0.25)3$ and $\beta_3 = \beta_4 = \beta_5 = 0$

LEGEND: ——: ordinary data set and - - -: modified data set



SAMPLE SIZE: 50; CORRELATION: 0

SAMPLE SIZE: 100; CORRELATION: 0

SAMPLE SIZE: 50; CORRELATION: 0.5

SAMPLE SIZE: 100; CORRELATION: 0.5

SAMPLE SIZE: 50; CORRELATION: 0.9

SAMPLE SIZE: 100; CORRELATION: 0.9

Figure 3.5: The APE of models selected from simulated data sets with $m = 10$ predictors; $\beta_0 = 1$, $\beta_1 = \beta_2 = ... = \beta_{10} = 0(0.1)1.5$ and $\beta_1 = \beta_2 = ... = \beta_{10} = 1.5(0.25)3$

LEGEND: ——: ordinary data set and - - -: modified data set

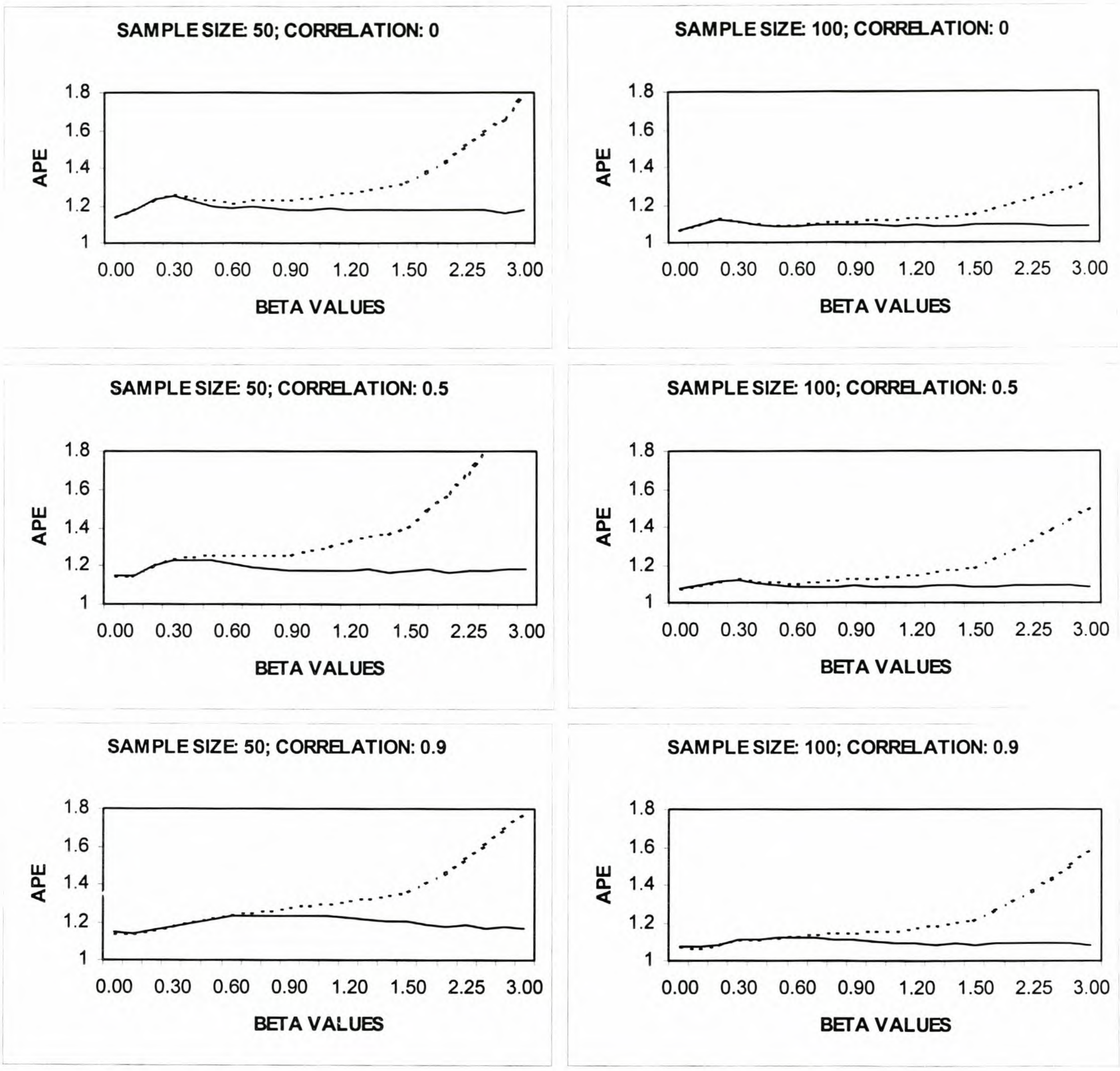


Figure 3.6: The APE of models selected from simulated data sets with $m = 10$ predictors; $\beta_0 = 1$, $\beta_1 = \beta_2 = ... = \beta_5 = 0(0.1)1.5$ and $\beta_1 = \beta_2 = ... = \beta_5 = 1.5(0.25)3$ and $\beta_6 = \beta_7 = ... = \beta_{10} = 0$

LEGEND: ——: ordinary data set and - - -: modified data set

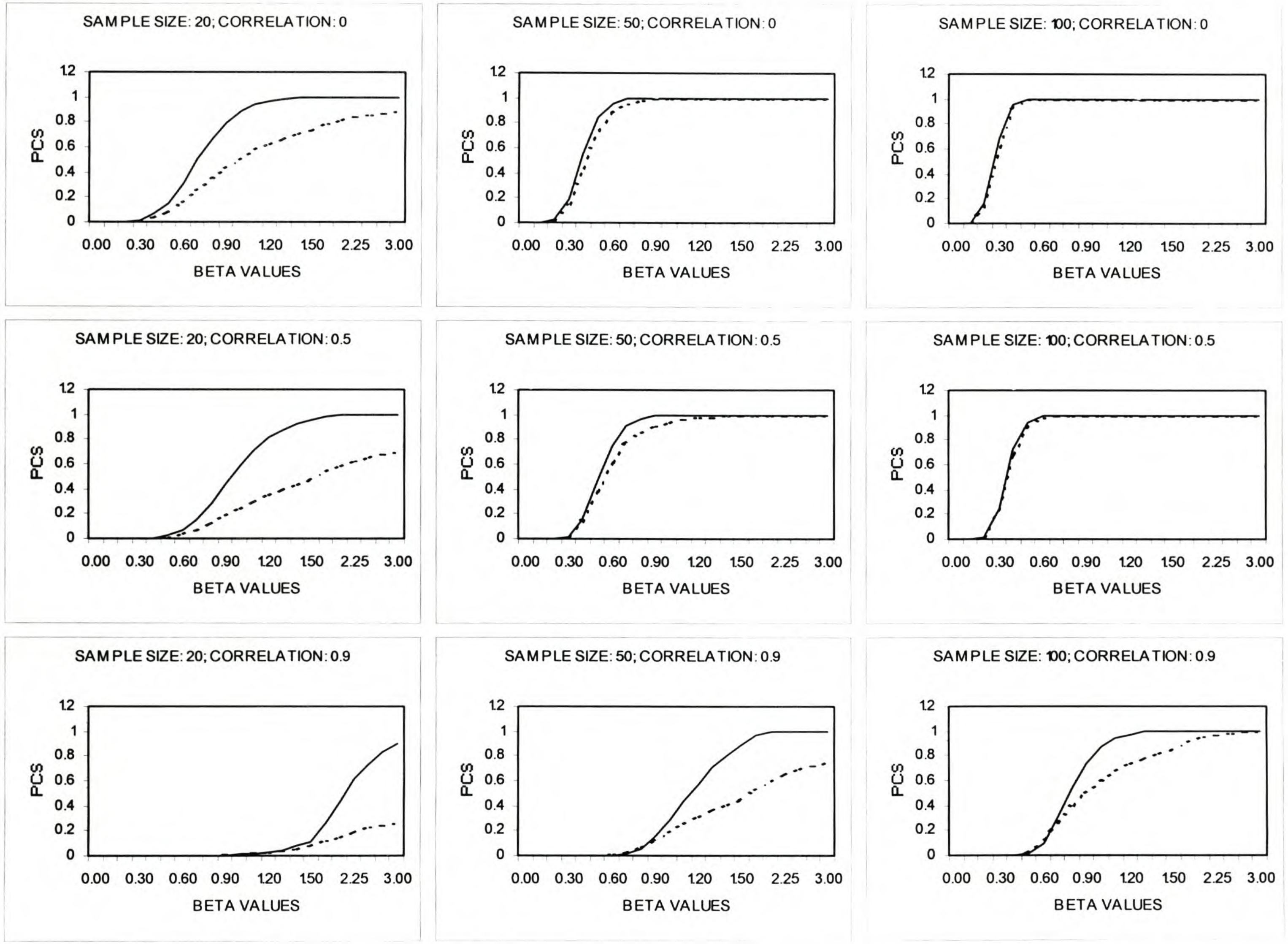


Figure 3.7: The PCS for models selected from simulated data sets with $m = 5$ predictors

$\beta_0 = 1$, $\beta_1 = \beta_2 = ... = \beta_5 = 0(0.1)1.5$ and $\beta_1 = \beta_2 = ... = \beta_5 = 1.5(0.25)3$

LEGEND: ——: ordinary data set and - - -: modified data set

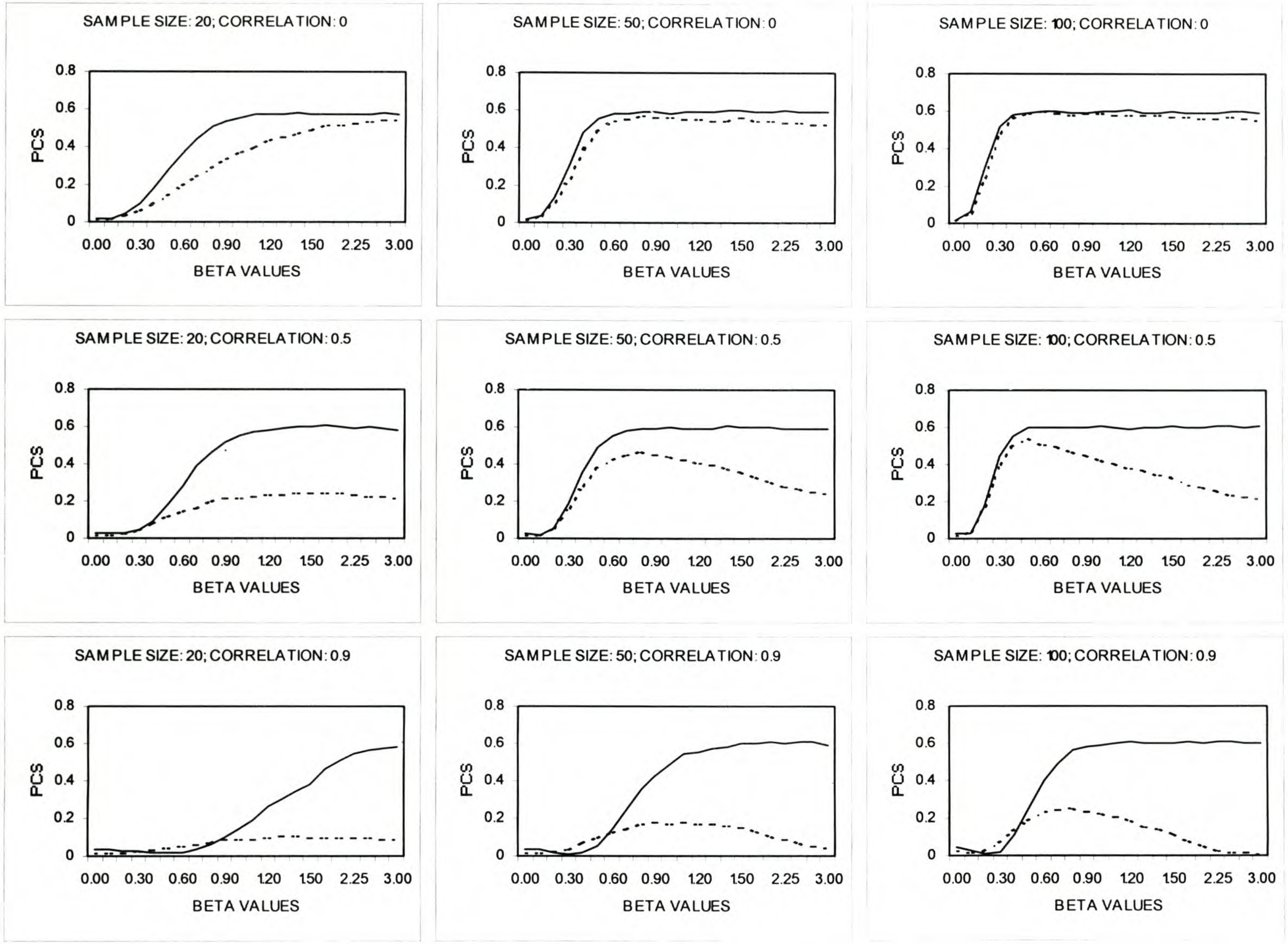


Figure 3.8: The PCS for models selected from simulated data sets with $m = 5$ predictors; $\beta_0 = 1$, $\beta_1 = \beta_2 = 0(0.1)1.5$ and $\beta_1 = \beta_2 = 1.5(0.25)3$ and $\beta_3 = \beta_4 = \beta_5 = 0$

LEGEND: ——: ordinary data set and - - -: modified data set

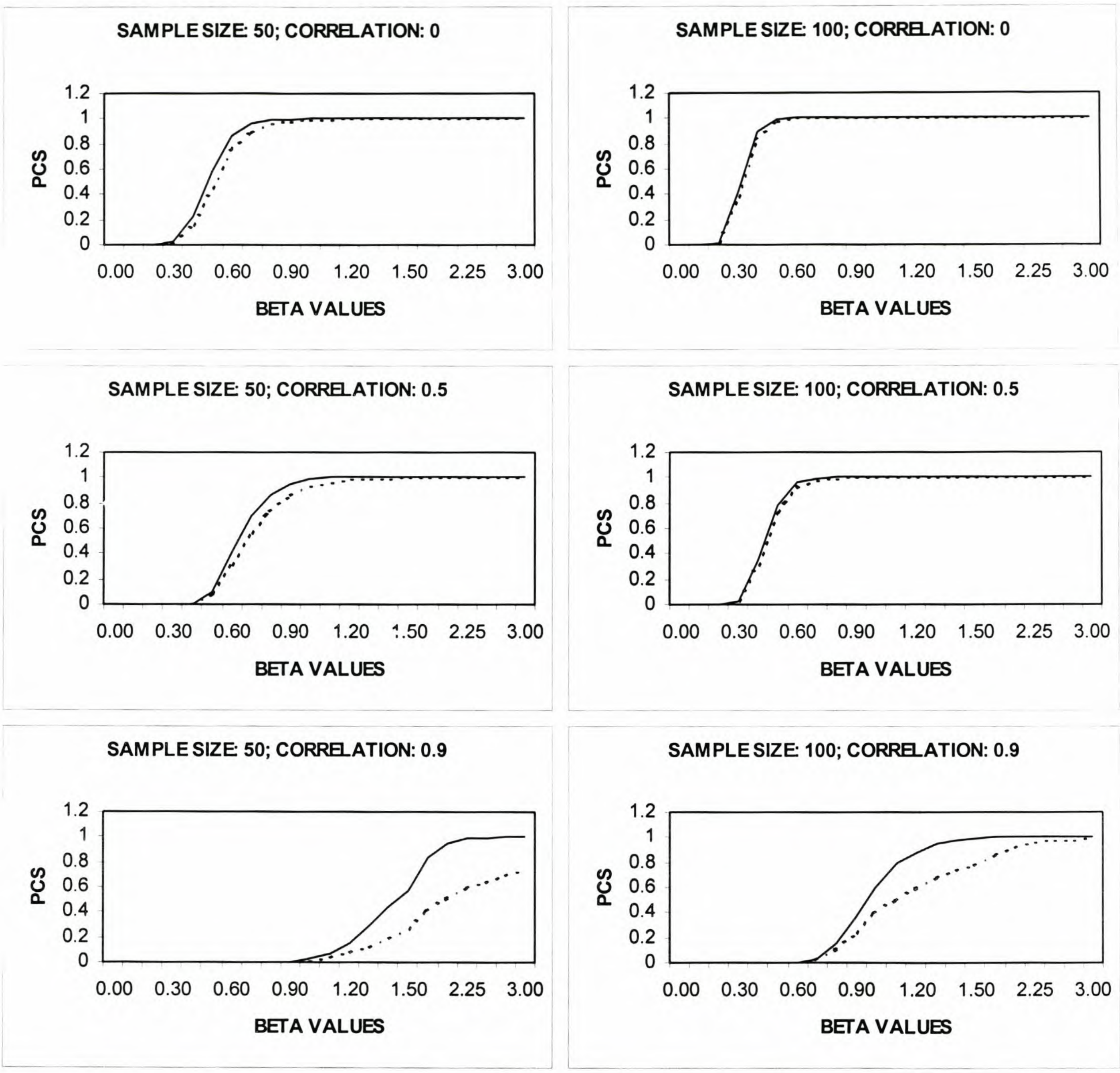


Figure 3.9: The PCS for models selected from simulated data sets with $m = 10$ predictors; $\beta_0 = 1$, $\beta_1 = \beta_2 = ... = \beta_{10} = 0(0.1)1.5$ and $\beta_1 = \beta_2 = ... = \beta_{10} = 1.5(0.25)3$

LEGEND: ——: ordinary data set and - - -: modified data set

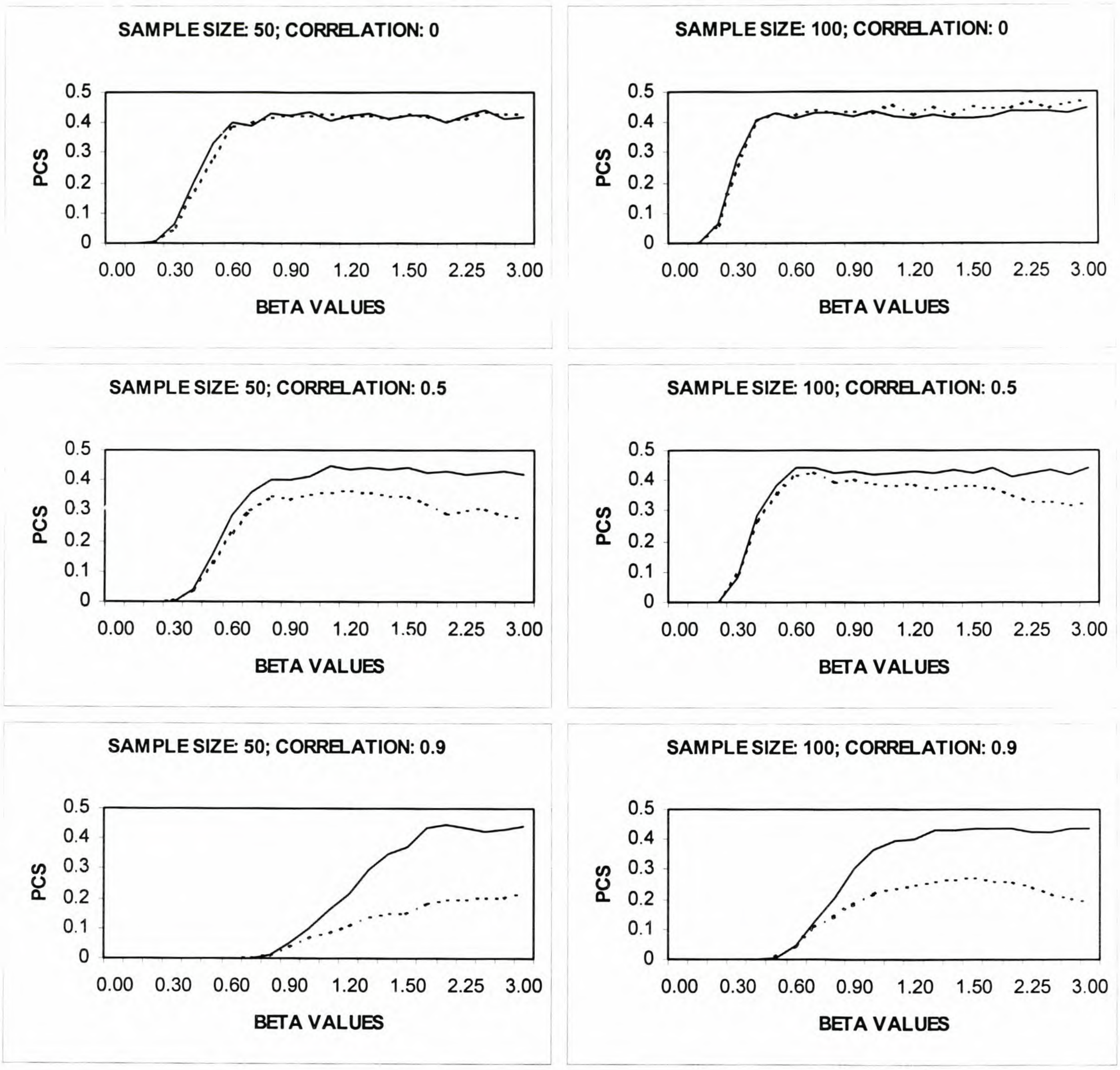


Figure 3.10: The PCS for models selected from simulated data sets with $m = 10$ predictors; $\beta_0 = 1$, $\beta_1 = \beta_2 = ... = \beta_5 = 0(0.1)1.5$ and $\beta_1 = \beta_2 = ... = \beta_5 = 1.5(0.25)3$ and $\beta_6 = \beta_7 = ... = \beta_{10} = 0$

LEGEND: ——: ordinary data set and - - -: modified data set

3.3.2 Discussion of the results of the simulation study

Consider first the simulation results for the APE of the models obtained by applying selection to the ordinary and to the modified data sets. These results are plotted in Figures 3.3 to 3.6. Recall that the solid lines show the APE of the models selected from the ordinary data sets, whereas the dotted lines represent the APE of the models selected from the modified data sets.

The first noticeable feature of the APEs in Figures 3.3 to 3.6 is that the APEs of the models selected from the modified data sets increase as the values of the non-zero regression coefficients increase, whereas the APEs of the models selected from the ordinary data sets reach a maximum before stabilizing at a constant that varies with factors such as the sample size n, the number of predictor variables m, the correlation ρ amongst the predictor variables, and the particular configuration of β-values. It is evident that the presence of a possibly selection influential data point causes the APE of a selected model to deteriorate quite severely. How can this be explained? Recall that the modified data set is obtained from the ordinary data set by replacing the largest value of predictor by its smallest value. The general effect of this is to weaken the linear relationship between the response variable Y and x_1. This may cause x_1 to be wrongly excluded from a selected model in which it should be included, or, in cases where x_1 is in fact selected, it may worsen the fit of the resulting model. In both cases the APE of the selected model will deteriorate. Recall also that x_1 is one of the predictor variables with a non-zero regression coefficient. Therefore, as we move towards the right on any of the graphs in Figures 3.3 to 3.6, the regression coefficient of x_1 increases, i.e. x_1 becomes more important as an explanatory variable for the response. This implies that omission of x_1 from the selected model, or a poor fit of the model with respect to x_1, will tend to have more serious consequences, and this is clearly reflected in Figures 3.3 to 3.6.

Consider now an unimportant predictor, i.e. a predictor with zero regression coefficient. What would happen if the largest value of such a predictor was replaced by its corresponding smallest observation? The effect would be much less severe than in cases involving important predictors. Clearly, as the non-zero regression coefficients increase, it becomes less likely for an unimportant variable to be selected, or, even in cases where such a variable is selected, its contribution to the fitted model will be small. Under these circumstances one therefore expects the behaviour of the APEs of the models selected from the modified data sets to remain very similar to that of the APEs of the models selected from the ordinary data sets.

What can be said about the influence of the sample size n, and the correlation ρ amongst the predictor variables, on the APEs? It is clear from Figures 3.3 to 3.6 that in all cases the APE decreases as the sample size increases. This is true for the models selected from the ordinary data sets, as well as the models selected from the modified data sets. It is also clear that the APEs of the models selected from the modified data sets tend to increase more rapidly, with increasing non-zero regression coefficients, as the correlation amongst the predictors increases.

We now turn to a consideration of the simulation results for the PCS, shown in Figures 3.7 to 3.10. Once again it is clear that the PCS-values for the models selected from the ordinary data sets (the solid lines in these figures) differ from those for the models selected from the modified data sets (the dotted lines in these figures), the difference in some cases being quite large. Consider first in this regard Figures 3.7 and 3.9. These figures show the PCS for the cases where all the predictors are important, i.e. the cases where all m the β-values are increased from 0 up to 3. It is clear that in all cases the "correct" model (i.e. the model containing all the predictors) is eventually selected with probability 1. However, the PCS for the models selected from the ordinary data sets approaches 1 more quickly than does the PCS for the models selected from the modified data sets, especially for small sample sizes. It is also clear that at a fixed sample size both sets of PCS-values approach the limit of 1 more slowly if the correlation amongst the predictors increases.

Figures 3.8 and 3.10 deal with the cases where only some of the regression coefficients are increased: two out of five in Figure 3.8, and five out of ten in Figure 3.10. Now the PCS of the models selected from the ordinary data sets no longer increases to 1 as the non-zero regression coefficients become larger. In Figure 3.8 the limiting PCS-value seems to be approximately 0.6, and in Figure 3.10 it is approximately 0.4. An explanation for this must be sought in the fact that at least one of the unimportant predictors is frequently selected together with the important ones, leading to an "incorrect" model being selected. For the models selected from the modified data sets, the situation is still worse. As the values of the non-zero regression coefficients become larger, the PCS-values of these models decrease to 0.

This has two causes: firstly, as in the case of the models selected from the ordinary data sets, unimportant variables are frequently selected; secondly, because of the presence of a possibly selection influential data case, one of the important predictors, viz. x_1, is frequently wrongly omitted from the selected model. We see that once again a larger sample size has a positive influence on the PCS-values, but that a larger correlation amongst the predictors has a detrimental effect.

CHAPTER 4

IDENTIFICATION OF SINGLE SELECTION INFLUENTIAL DATA CASES

4.1 Introduction

In Section 3.2 of Chapter 3 we showed, by means of illustrative examples, that a traditional influence measure, such as Cook's distance, is not always successful in identifying selection influential cases if variable selection is applied in multiple linear regression. This is not surprising since traditional influence measures evaluate data cases in terms of all their coordinates. No variable selection is therefore involved when such a measure is utilised for identifying influential data cases. Our concern, however, is to measure the influence of data cases if variable selection is applied in multiple linear regression. In the statistical literature such measures, referred to as *selection influence measures* of data cases, have been developed. Contributions include that of Weisberg (1981), Ahn and Park (1987), Chatterjee and Hadi (1988), Peixoto and Lamotte (1989), Léger and Altman (1993), Kim and Park (1995), Gupta and Huang (1996), Hoeting, Raftery and Madigan (1996), and Kim and Hwang (2000). These contributions were discussed in Section 1.3 of Chapter 1. Of importance in all these selection influence measures is to ensure that the measure explicitly takes an initial variable selection step into account. If this approach is not followed the influence measure is effectively calculated conditionally on a predetermined subset of predictors. Léger and Altman (1993) illustrate this aspect very elegantly when they distinguish between a *conditional* and *unconditional* selection version of Cook's distance. We briefly explain the difference between the two versions.

In (3.1), Cook's distance for the ith case in a data set is defined to be

$$D_i = \frac{\left\| P_M \mathbf{Y} - P_M \mathbf{Y}_{(-i)} \right\|^2}{(m+1)\widehat{\sigma}^2}, \qquad i = 1, ..., n. \tag{4.1}$$

Since both $P_M\mathbf{Y}$ and $P_M\mathbf{Y}_{(-i)}$ depend on M it is clear that no variable selection is involved in (4.1). We now indicate how a conditional selection version of (4.1) may be defined. Let

$\widehat{L} \equiv \widehat{L}(\mathbf{Y})$ denote the data-dependent selected linear subspace obtained by applying a variable selection technique to an available regression sample. The subspace $\widehat{L}$, spanned by the vector $\mathbf{1}$ together with the $\widehat{l}$ vectors of observations corresponding to the selected variables, is of dimension $\widehat{l}+1$. The least squares estimator of $\boldsymbol{\mu}$, corresponding to $\widehat{L}$, is given by $P_{\widehat{L}}\mathbf{Y}$, defined in (1.14). This estimator is calculated from the full data set, i.e. the data set containing all the observations. Let $P_{\widehat{L}}\mathbf{Y}_{(-i)}$ denote the least squares estimator of $\boldsymbol{\mu}$ calculated from the complete data set excluding data case i. Note that the same subspace, viz. $\widehat{L}$ identified from the complete data set, is used to calculate both $P_{\widehat{L}}\mathbf{Y}$ and $P_{\widehat{L}}\mathbf{Y}_{(-i)}$. Also, similar to $P_M\mathbf{Y}_{(-i)}$ in (4.1), $P_{\widehat{L}}\mathbf{Y}_{(-i)}$ contains a prediction for case i, although this case is not used in calculating the estimated parameters (β-coefficients) implicit in $P_{\widehat{L}}\mathbf{Y}_{(-i)}$. The conditional Cook's distance for the ith case, denoted by D_i^c, is now defined as

$$D_i^c = \frac{\left\|P_{\widehat{L}}\mathbf{Y} - P_{\widehat{L}}\mathbf{Y}_{(-i)}\right\|^2}{\left(\widehat{l}+1\right)\widehat{\sigma}^2_{\widehat{L}}}, \tag{4.2}$$

where $\widehat{\sigma}^2_{\widehat{L}}$ denotes the least squares estimator of the error variance, based on the corresponding predictors spanning $\widehat{L}$.

Léger and Altman (1993) obtain a so-called unconditional selection version of Cook's distance for data case i by arguing that it is necessary to repeat variable selection using the data set without case i. This yields a linear subspace $\widehat{L}^*$, with $\widehat{L}^*$ possibly different from $\widehat{L}$. The least squares estimator of $\boldsymbol{\mu}$ based on $\widehat{L}^*$ is denoted by $P_{\widehat{L}^*}\mathbf{Y}_{(-i)}$. The unconditional Cook's distance for case i, denoted by D_i^u, is now defined by

$$D_i^u = \frac{\left\|P_{\widehat{L}}\mathbf{Y} - P_{\widehat{L}^*}\mathbf{Y}_{(-i)}\right\|^2}{\left(\widehat{l}+1\right)\widehat{\sigma}^2}, \tag{4.3}$$

where $\widehat{\sigma}^2$, instead of $\widehat{\sigma}^2_{\widehat{L}}$ as in (4.2), is used to estimate the error variance, σ^2. The differences between these two selection versions of Cook's distance are discussed by Léger and Altman (1993). They conclude that the unconditional version is preferable since it explicitly takes the selection effect into account.

We now argue along the same lines to propose a simple and easy to implement influence measure for identifying individual selection influential data cases in Section 4.2. In Section 4.3 this new influence measure is applied to the three example data sets introduced in Chapter 3.

4.2 A new selection influence measure

It is standard statistical practice in multiple linear regression to measure the influence of a single data case in an analysis as follows: analyse the complete data set, and calculate a (summary) measure, say M; repeat the analysis after omitting the case under consideration, and calculate $M_{(-i)}$; quantify the influence of case i in terms of a function, $f(M, M_{(-i)})$, of M and $M_{(-i)}$.

The measure of selection influence that we propose is also based on a leave-one-out strategy. The following questions need to be resolved in this regard:

- What is meant by an analysis of the complete or the reduced data set?
- What measure, M, should be used?
- How should we define the function $f(.)$?

Regarding the *first* question, in a variable selection context an analysis of the data set entails applying a given variable selection technique, and fitting the model corresponding to the predictor variables that form a basis for the selected linear subspace $\widehat{L}$, to the data. Consequently, if we wish to study the influence of a single data case in such an analysis, it is necessary to apply the selection technique under consideration to the full data set and again to the reduced data set. This is of course in line with the unconditional approach recommended by Léger and Altman (1993). Turning to the *second* question, different choices of M can be made, depending on the aspect of the fitted model which is of interest. In variable selection the number of selected variables and the lack of fit of the corresponding model are typically of interest. These quantities are combined in selection criterion such as Akaike's information criteria (see Akaike, 1973 and 1974) and in model selection using the adjusted coefficient of multiple determination (see Draper and Smith, 1998). The same is found in Mallows' C_p criterion, where for a fixed subspace L, the lack of fit component (i.e. the scaled residual sum of squares, $\|\mathbf{Y} - P_L\mathbf{Y}\|^2 / \widehat{\sigma}^2$ in (2.4)) is combined with the number of variables (i.e. the value of l given in the last term, $(2(l+1) - n)$, of (2.4)). It therefore seems sensible to take M simply equal to the criterion employed in the selection method. This implies that $f(.)$ has to based on the difference in the value of the selection criterion before and after omitting case i. This difference, $M - M_{(-i)}$, can then divided by M in order to calculate the relative change in the selection criterion. The proposed selection influence measure for the ith case is therefore given by

$$f\left(M, M_{(-i)}\right) = \frac{M - M_{(-i)}}{M}, \tag{4.4}$$

where M denoted the selection criterion under consideration. Note that (4.4) can be calculated for all selection criteria where the particular criterion is a combination of some sort of goodness-of-fit measure and a penalty function (such a penalty function usually includes the number of predictors of the particular selected model as one of its components, (see Kundu and Murali, 1996, in this regard)). However, since we restrict our attention in this dissertation solely to Mallows' C_p procedure as variable selection technique, the selection influence measure in (4.4) becomes

$$\begin{aligned} & f\left(C_p\left(\mathbf{Y}, \hat{L}\right)/\widehat{\sigma}^2, C_p\left(\mathbf{Y}_{(-i)}, \hat{L}^*\right)/\widehat{\sigma}^2\right) \\ = & \frac{C_p\left(\mathbf{Y}, \hat{L}\right)/\widehat{\sigma}^2 - C_p\left(\mathbf{Y}_{(-i)}, \hat{L}^*\right)/\widehat{\sigma}^2}{C_p\left(\mathbf{Y}, \hat{L}\right)/\widehat{\sigma}^2} \\ = & \frac{C_p\left(\mathbf{Y}, \hat{L}\right) - C_p\left(\mathbf{Y}_{(-i)}, \hat{L}^*\right)}{C_p\left(\mathbf{Y}, \hat{L}\right)}. \end{aligned} \tag{4.5}$$

The value of $C_p\left(\mathbf{Y}, \hat{L}\right)/\widehat{\sigma}^2$ in (4.5) is acquired from the full data set, viz.

$$\frac{C_p\left(\mathbf{Y}, \hat{L}\right)}{\widehat{\sigma}^2} = \frac{\left\|\mathbf{Y} - P_{\hat{L}}\mathbf{Y}\right\|^2}{\widehat{\sigma}^2} + 2(\hat{l}+1) - n, \tag{4.6}$$

where $\hat{L}$, with $\dim(\hat{L}) = \hat{l}+1$, denotes the data dependent linear subspace for which (4.6) is a minimum. In a similar way is $C_p\left(\mathbf{Y}_{(-i)}, \hat{L}^*\right)/\widehat{\sigma}^2$ in (4.5) calculated on the full data set with case i omitted, i.e.

$$\frac{C_p\left(\mathbf{Y}_{(-i)}, \hat{L}^*\right)}{\widehat{\sigma}^2} = \frac{\left\|\mathbf{Y}_{(-i)} - P_{\hat{L}^*}\mathbf{Y}_{(-i)}\right\|^2}{\widehat{\sigma}^2} + 2(\hat{l}^*+1) - (n-1). \tag{4.7}$$

Note that calculation of $C_p\left(\mathbf{Y}_{(-i)}, \hat{L}^*\right)/\widehat{\sigma}^2$ requires the omission of data case i before finding $\hat{L}^*$, which is the linear subspace for which (4.7) is a minimum. It is clear that the subspace $\hat{L}^*$ will for some data cases be equal to $\hat{L}$, and for other cases be different from $\hat{L}$. As already illustrated by the example data sets in Chapter 3, $\hat{L}^*$ typically equals $\hat{L}$ for the majority of data cases. Also note that in the calculation of $C_p\left(\mathbf{Y}_{(-i)}, \hat{L}^*\right)/\widehat{\sigma}^2$, the estimator for the error variance is obtained from the full data set. The use of this error variance estimator is supported by considerations given by Léger and Altman (1993) for using $\widehat{\sigma}^2$ in the denominator of the unconditional Cook's distance in (4.3).

The proposed influence measure in (4.5) is large if the relative difference between $C_p\left(\mathbf{Y}, \hat{L}\right)/\widehat{\sigma}^2$ and $C_p\left(\mathbf{Y}_{(-i)}, \hat{L}^*\right)/\widehat{\sigma}^2$ is large. If this is true for an omitted data case i, the particular case is considered selection influential. Note that negative values of $C_p\left(\mathbf{Y}, \hat{L}\right)/\widehat{\sigma}^2$ in (4.5) may occur. These negative values may cause misrepresentation of the relative difference between $C_p\left(\mathbf{Y}, \hat{L}\right)/\widehat{\sigma}^2$ and $C_p\left(\mathbf{Y}_{(-i)}, \hat{L}^*\right)/\widehat{\sigma}^2$, i.e. the relative differences for certain data cases may now be incorrectly larger than others if, for example $C_p\left(\mathbf{Y}, \hat{L}\right) - C_p\left(\mathbf{Y}_{(-i)}, \hat{L}^*\right)$ in the numerator, and $C_p\left(\mathbf{Y}, \hat{L}\right)$ in the denominator of (4.5) are negative. We overcome this difficulty and ensure that (4.5) is always positive by omitting the subtraction of n in (4.6) and $(n-1)$ in (4.7). Calculating the influence measure in this way ensures that large values of (4.5) corresponds with significant relative changes between $C_p\left(\mathbf{Y}, \hat{L}\right)/\widehat{\sigma}^2$ and $C_p\left(\mathbf{Y}_{(-i)}, \hat{L}^*\right)/\widehat{\sigma}^2$.

In the next section we illustrate the proposed influence measure by applying it to the three example data sets that were introduced in Chapter 3.

4.3 Illustrative examples

4.3.1 The Hald data

Reconsider the Hald data from Chapter 3, which consisting of four predictor variables and 13 observations. Applying C_p to the full data set, without subtracting the term n in (4.6), results in a criterion value of 15.678, with predictor variables 1 and 2 being selected. The C_p criterion in (4.7), once again without subtracting the term $(n - 1)$, is also calculated for each of the reduced data sets obtained by omitting a single observation. For completeness we again list the corresponding predictor variables, as in Table 3.1, selected on each of these reduced data sets in Table 4.1. Also reported in Table 4.1 for each of the reduced data sets are the values of the proposed influence measure, Cook's conditional distance in (4.2), and Cook's unconditional distance in (4.3), and an estimated average prediction error. Program C5 in Appendix C was used to calculate Cook's unconditional distance for the Hald data. The manner in which the entries in the last column of Table 4.1 were obtained, is explained below.

From Table 4.1 it is clear that the three influence measures attain maximum values when different data cases are omitted. Our proposed influence measure is a maximum if case 6 is omitted, whereas Cook's conditional and unconditional distances are maxima if cases 10 and 8 are respectively omitted. In our discussion in Section 3.2.1 data cases 6 and 8 were identified as selection influential, since their separate omission lead to a change in the selected model. In order to obtain an indication of whether it would be beneficial in the sense of improved response prediction to leave out cases 6 or 8, we calculate estimates of the expected squared prediction error for the full data set and for each of the 13 reduced data sets. An explanation of how these values, shown in the last column of Table 4.1, are obtained, follows in the next paragraph.

Data case omitted	Selected variables	Influence measure	Cook's conditional distance	Cook's unconditional distance	Average prediction error
case 1	1,2	0.0353	0.05	0.06	12.2
case 2	1,2	0.0159	0.02	0.03	12.2
case 3	1,3,4	0.0456	0.02	0.84	12.3
case 4	1,2	0.0387	0.05	0.06	14.2
case 5	1,2	0.0226	0.01	0.01	13.0
case 6	1,2,3	0.1670	0.11	0.64	9.5
case 7	1,2	0.0283	0.07	0.08	12.4
case 8	1,2,3	0.1599	0.08	1.83	7.3
case 9	1,2	0.0432	0.04	0.05	11.5
case 10	1,2	0.0440	0.23	0.28	12.8
case 11	1,2	0.1392	0.14	0.16	11.5
case 12	1,2	0.0099	0.01	0.01	12.6
case 13	1,2	0.1136	0.13	0.16	10.2

Table 4.1: The selected variables, proposed selection influence measure, Cook's conditional and unconditional distance, and the estimated average prediction error for each of the reduced data sets of the Hald data

Firstly, consider the full data set. Randomly select 10 of the 13 cases to form a training data set of 10 cases and a test data set of 3 cases. Apply the C_p criterion to the training data set and use the selected model to calculate predictions for the 3 cases of the test data set. From these, calculate the average squared prediction error for the test data set. Random selection of a training data set and calculation of the average squared prediction error for the test data set, based on the selected model from the training data set, is repeated 20000 times. Record is kept of the 20000 average prediction errors. Their average is calculated in order to obtain an estimate of the expected squared prediction error. For the full data set this value equals 10.53.

The process explained above is then repeated for each reduced data set, now selecting only 9 cases for the training data set and again 3 cases for the test data set. The resulting values are shown in the last column of Table 4.1.

We observe the following: The values of the estimated expected squared prediction error for the reduced data sets, when cases 6 and 8 are respectively omitted, are significantly smaller than 10.53. In fact, the smallest value of this measure is obtained if data case 8 is omitted. Note that case 8 is identified by Cook's unconditional selection influence measure as potentially the most selection influential point in the data set. The values obtained for the estimated expected squared prediction error therefore seem to substantiate this pronouncement. Note that in this example the new selection influence proposed in (4.5), and Cook's unconditional selection influence distance, do not agree as to the most prominent selection influential case. We see, however, that the value of measure (4.5) is quite large for data case 8 (0.1599 compared to 0.1670 for data case 6). In our next illustrative example we do find these two measures agreeing as to the most prominent selection influential data case.

4.3.2 The fuel data

Recall that the fuel data, introduced in Chapter 3 consist of 50 observations on a response and each of four predictor variables. Applying C_p selection in this instance to the full data set results in variables 2 and 3 being selected. The corresponding minimum value of the C_p criterion equals 52.517. We calculate the influence measure in (4.5) as before: for each i, omit data case i, and repeat C_p selection to obtain a sequence of selected subspaces. In each case, calculate the quantity in (4.7), and combine this with the value 52.517 to find the measure in (4.5) for each value of $i, i = 1, 2, ..., 50$. The resulting values of (4.5) are shown in Table 4.2, together with the variables that are selected in each case, the values of Cook's conditional and unconditional selection distances, and the estimated expected squared prediction errors.

It is clear from Table 4.2 that all three influence measures (the measure proposed in (4.5), Cook's conditional selection distance in (4.2), and Cook's unconditional distance in (4.3)) are relatively large for cases 40, 49 and 50. In fact, measure (4.5) and Cook's unconditional selection distance both reach a maximum at case 50. We repeated the calculations described in the previous example to obtain estimated expected squared prediction errors. For the full data set this estimate equals 9850.24. This value was obtained by performing 20000 repetitions, at each repetition randomly dividing the 50 observations into 40 training observations and 10 test observations. The estimated expected squared prediction errors for each of the reduced data sets were calculated similarly, now using 39 observations in the training data sets and 10 observations in the test data sets. The resulting values are shown in the last column of Table 4.2. We observe a sharp reduction in the estimated expected squared prediction error from 9850.24 for the full data set to only 6012.5 for the reduced data set missing observation 50. The estimated expected squared prediction errors for the reduced data sets without observations 40 and 49 are also significantly smaller than that of the full data set, but not as low as that of the reduced data set missing observation 50.

Closer inspection of the third and sixth columns of Table 4.2 reveals a strong correspondence between the values of the selection influence measure (4.5) and the estimated expected squared prediction errors. This is reflected in the correlation coefficient of -0.9716 between these two sets of numbers. Note that the sign of the correlation coefficient suggests that omission of a data case that is deemed to be selection influential according to measure (4.5) generally leads to a reduction in the estimated expected squared prediction error. The correlation between the unconditional Cook's selection distance and the estimated expected squared prediction error is -0.9657, almost as strong as that between measure (4.5) and the estimated expected squared prediction error. Finally, the correlation between the values of measure (4.5) and the unconditional Cook's selection distances is 0.9276, confirming a strong positive relationship between these two measures for this example.

The proposed influence measure shows that the maximum relative difference between $C_p\left(\mathbf{Y}, \hat{L}\right)$ and $C_p\left(\mathbf{Y}_{(-i)}, \hat{L}^*\right)$ is obtained if data case 50 is omitted. Cook's unconditional selection distance confirms the implication that case 50 is selection influential. The low value obtained for the average prediction error when case 50 is omitted supplies strong evidence that omitting this case before doing variable selection improves the predictive power of the resulting model. The following question arises: How should one judge the significance of a value of the proposed influence measure? In other words, should one recommend to a practitioner analysing this data set, to omit case 50 before doing variable selection? We attempt to answer this question by comparing the influence measure of data case 50 with the largest influence measures obtained in a large number of bootstrap samples.

The bootstrap method, proposed by Efron and Tibshirani (1993), is applied within a regression context in either one of two ways. The one technique involves the sampling of "pairs". A bootstrap data set is obtained by selecting response observations, together with their corresponding m-dimensional independent observation vectors, randomly from the original data set. The selection is done with replacement, and repeated until n "pairs", each of the form $(Y_i; x_{i1}, x_{i2}, ..., x_{im})$, for $i = 1, ..., n$, have been obtained. The other technique, and the one we apply to the fuel data set in order to judge the significance of the corresponding influence measure if data case 50 is omitted, involves random selection of residuals. Appropriately, this technique is known as residual bootstrapping and is applied for a regression sample of size n, in the following way: Determine the vector of residuals, of the form $\mathbf{r} = \mathbf{Y} - \mathbf{P}_M\mathbf{Y}$, once a linear regression model has been fitted to the complete data set. Random selection of n of these residuals with replacement yields a bootstrap vector of residuals, denoted by $\mathbf{r}_b$. If we calculate $\mathbf{r}_b$+ $\mathbf{P}_M\mathbf{Y}$ a new bootstrap response vector, denoted by $\mathbf{Y}_b$, is obtained. This newly formed bootstrap vector and the original set of unchanged predictor variable values constitute the bootstrap sample.

Consider now 10000 of these bootstrap samples, each of size 50 for the fuel data. The proposed influence measure in (4.5) is calculated for every single omitted data case in each of the 10000 bootstrap samples. By keeping record of the largest value of (4.5) in every bootstrap sample, we are provided with a set of values to which we can compare the largest influence measure (i.e. if data case 50 is omitted) of the fuel data set. The distribution of these 10000 largest bootstrap influence measures is shown in Figure 4.1.

Data case omitted	Selected variables	Influence measure	Cook's conditional distance	Cook's unconditional distance	Average prediction error
case 1	2,3	0.0016	0.00169	0.00175	10138.1
case 2	2,3	0.0105	0.00400	0.00414	10012.7
case 3	2,3	0.0070	0.00386	0.00399	10028.1
case 4	2,3	0.0122	0.01194	0.01234	10394.8
case 5	2,3	0.0389	0.01716	0.01774	10152.9
case 6	2,3	0.0052	0.00856	0.00884	10104.9
case 7	2,3	0.0004	0.00146	0.00151	10272.8
case 8	2,3	0.0018	0.00225	0.00233	10240.8
case 9	2,3	0.0056	0.00324	0.00335	9966.7
case 10	2,3	0.0034	0.00152	0.00157	10106.5
case 11	2,3	0.0110	0.00599	0.00619	10086.7
case 12	2,3	0.0004	0.00063	0.00065	10288.4
case 13	2,3	0.0006	0.00039	0.00040	10230.5
case 14	2,3	0.0038	0.00158	0.00164	10237.8
case 15	2,3	0.0063	0.00317	0.00328	10098.8
case 16	2,3	0.0056	0.00209	0.00216	10010.3
case 17	2,3	0.0015	0.00051	0.00053	10066.1
case 18	2,3	0.0524	0.03876	0.04006	9537.0
case 19	2,3	0.0557	0.19215	0.19863	9625.1
case 20	2,3	0.0097	0.01626	0.01681	10051.1
case 21	2,3	0.0006	0.00089	0.00092	10250.7
case 22	2,3	0.0014	0.00134	0.00138	10276.3
case 23	2,3	0.0001	0.00016	0.00016	10215.2
case 24	2,3	0.0039	0.00255	0.00263	10228.6
case 25	2,3	0.0123	0.00606	0.00626	10037.9
case 26	2,3	0.0004	0.00030	0.00031	10219.8
case 27	2,3	0.0017	0.00191	0.00197	10088.2
case 28	2,3	0.0009	0.00050	0.00051	10157.2
case 29	2,3	0.0001	0.00002	0.00002	10097.2
case 30	2,3	0.0003	0.00037	0.00038	10296.7
case 31	2,3	0.0010	0.00096	0.00100	10086.6
case 32	2,3	0.0005	0.00077	0.00080	10260.0
case 33	2,3	0.0224	0.04700	0.04859	9911.4
case 34	2,3	0.0012	0.00155	0.00160	10104.9
case 35	2,3	0.0034	0.00553	0.00571	10246.1
case 36	2,3	0.0029	0.00313	0.00324	10106.4
case 37	2,3	0.0085	0.00340	0.00351	10145.4
case 38	2,3	0.0196	0.01067	0.01103	9957.6
case 39	2,3	0.0192	0.03887	0.04019	9951.4
case 40	2,3,4	0.2038	0.32901	0.87563	7977.2
case 41	2,3	0.0047	0.00329	0.00340	10198.4
case 42	2,3	0.0230	0.01741	0.01800	9858.6
case 43	2,3	0.0012	0.00055	0.00056	10365.2
case 44	2,3	0.0092	0.00918	0.00949	9997.8
case 45	2,3	0.0652	0.15736	0.16267	9429.6
case 46	2,3	0.0064	0.00250	0.00258	10048.5
case 47	2,3	0.0017	0.00115	0.00119	10290.6
case 48	2,3	0.0019	0.00178	0.00184	10140.8
case 49	2,3,4	0.1024	0.35151	1.06643	8630.7
case 50	1,2,3	0.2658	0.31159	2.40291	6012.5

Table 4.2: The selected variables, proposed selection influence measure, Cook's conditional and unconditional distance, and the estimated average prediction error for each of the reduced data sets of the fuel data

The proposed influence measure, if data case 50 is omitted, equals 0.2658. We evaluate this value in terms of the clearly positively skewed distribution of largest bootstrap influence measures in Figure 4.1. The vertical line drawn in the class interval $(0.26^{-}; 0.29)$ on the histogram in Figure 4.1 shows the position of 0.2658 in the distribution. The proportion of bootstrap influence measures that are smaller than 0.2658, equals 0.8824. This implies that the value 0.2658 lies close to the 90th percentile of the bootstrap distribution. Providing this information to any practitioner who analyses the fuel data, will surely be helpful in the decision of whether case 50 has to be omitted before subsequent analysis is performed.

It is important to bear in mind that the proposed influence measure only identifies individual possibly selection influential data cases. If it is, for example, decided to reject case 50 from the fuel data set, the influence measure should be recalculated on the $n - 1$ remaining observations to identify other possibly selection influential data cases. In Chapter 5 we propose a method for identifying more than one selection influential case simultaneously.

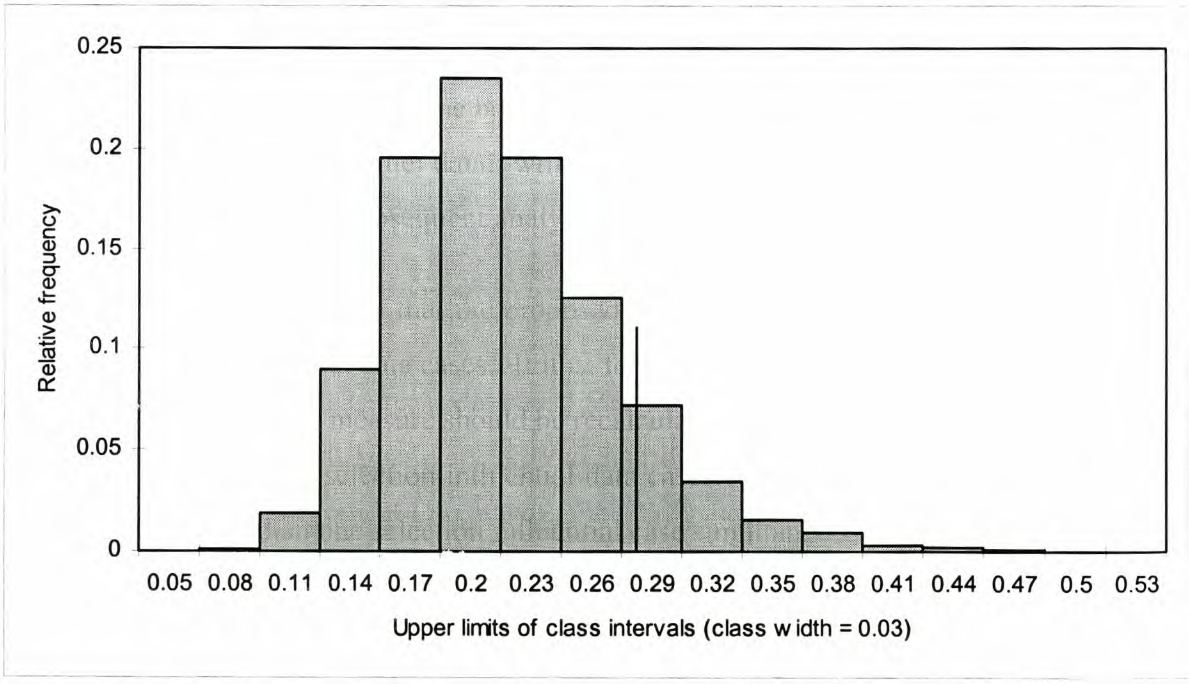


Figure 4.1: Histogram of largest proposed influence measures obtained in 10000 bootstrap samples of the fuel data

4.3.3 The evaporation data

The influence measure proposed in (4.5) was also applied to the evaporation data introduced in Chapter 3. Recall that ten independent predictor variables are measured on 46 consecutive days. Variable selection on the full data set yields the C_p criterion in (4.6) which equals $C_p\left(\mathbf{Y},\hat{L}\right)/\widehat{\sigma}^2+46=49.579$, with variables 1,3,6,8 and 9 being selected. Table 4.3 shows the following for each of the reduced data sets: the selected variables; values of the proposed influence measure; Cook's conditional and unconditional distances, and the estimated expected squared prediction errors. Random selection of 37 cases from the full data set in order to constitute a training data set, and using the model selected from this set to determine the average prediction error for the remaining 9 cases, resulted in an estimated expected squared prediction error of 71.78, in 20000 repetitions. The estimated values for the reduced data sets are based on 20000 repetitions, each time using 36 cases in the training data set and 9 cases in the test data set. As for the previous two examples, these values are also listed in Table 4.3.

The largest influence measures are obtained when case 41 is omitted, followed by case 33. The reduced data sets when cases 41 and 33 are respectively omitted also give the smallest values for the estimated average squared prediction error. The strong correspondence between the proposed influence measure and the estimated average squared prediction error for all 46 reduced data sets is reflected in a correlation coefficient of -0.9712. Note that Cook's unconditional distance is a maximum if a different case (i.e. case 40) is omitted. A correlation coefficient of only -0.4532 is obtained when the relationship between Cook's unconditional selection distance and the estimated expected squared prediction error is considered. Finally, as expected, the weak relationship between the proposed influence measure and Cook's unconditional distance is reflected in a correlation coefficient of 0.5022.

Omission of case 41 yields the largest value of measure (4.5) for the evaporation data. The low value obtained for the estimated average prediction error when case 41 is omitted provides strong evidence that omitting this case before doing variable selection improves the predictive power of the resulting model. Even stronger evidence is acquired when the value of the influence measure, if data case 41 is omitted, is compared with the largest influence measures obtained from 10000 bootstrap samples of the evaporation data.

Data case omitted	Selected variables	Influence measure	Cook's conditional distance	Cook's unconditional distance	Average prediction error
case 1	1,3,6,8,9	0.002079	0.004413	0.004761	77.8
case 2	1,3,6,8,9	0.118377	0.464169	0.500764	63.2
case 3	1,3,6,8,9	0.024529	0.044216	0.047702	74.7
case 4	1,3,6,8,9	0.000001	0.000001	0.000001	75.9
case 5	1,3,6,8,9	0.016745	0.017850	0.019257	75.7
case 6	1,3,6,9	0.017722	0.035181	0.435609	75.8
case 7	1,3,6,8,9	0.002581	0.004219	0.004551	76.4
case 8	1,3,6,9,10	0.103856	0.122133	0.711800	63.0
case 9	1,3,6,8,9	0.000423	0.000425	0.000458	76.0
case 10	1,3,6,9,10	0.003505	0.000656	0.569935	76.1
case 11	1,3,6,8,9	0.000075	0.000072	0.000077	75.5
case 12	1,3,6,8,9	0.005596	0.003556	0.003836	74.9
case 13	1,3,6,8,9	0.002701	0.003464	0.003737	76.0
case 14	1,3,6,8,9	0.000332	0.000259	0.000280	76.1
case 15	1,3,6,8,9	0.000051	0.000049	0.000053	77.5
case 16	1,3,6,8,9	0.004226	0.002995	0.003231	74.9
case 17	1,3,6,8,9	0.000060	0.000049	0.000053	75.9
case 18	1,3,6,9	0.003754	0.002727	0.388582	75.7
case 19	1,3,6,8,9	0.000343	0.000346	0.000374	77.1
case 20	1,3,6,8,9	0.000313	0.000427	0.000461	77.5
case 21	1,3,6,8,9	0.042305	0.036039	0.038880	72.9
case 22	1,3,6,9,10	0.035879	0.015650	0.614829	74.8
case 23	1,3,6,8,9	0.002618	0.001122	0.001210	75.3
case 24	1,3,6,9	0.061039	0.049505	0.421704	72.6
case 25	1,3,6,8,9	0.000060	0.000034	0.000037	76.1
case 26	1,3,5,7,8,9	0.021005	0.018032	0.522548	73.6
case 27	1,3,6,9,10	0.009917	0.004282	0.548482	76.4
case 28	1,3,6,8,9	0.000070	0.000033	0.000035	75.2
case 29	1,3,6,9,10	0.001863	0.000178	0.543130	75.5
case 30	1,3,6,8,9	0.000192	0.000292	0.000315	76.1
case 31	1,3,4,8,9	0.034015	0.001401	1.346172	72.6
case 32	1,3,6,9,10	0.032391	0.070408	0.861339	74.1
case 33	1,3,6,9,10	0.161212	0.046721	0.834903	61.9
case 34	1,3,6,8,9	0.000009	0.000010	0.000011	76.6
case 35	1,3,6,8,9	0.000001	0.000001	0.000001	76.6
case 36	1,3,6,8,9	0.004813	0.006332	0.006831	75.6
case 37	1,3,6,8,9	0.063002	0.081253	0.087659	70.4
case 38	1,3,6,8,9	0.042600	0.130336	0.140611	70.9
case 39	1,3,6,8,9	0.015217	0.012991	0.014015	75.1
case 40	6,9,10	0.051232	0.022216	1.933175	73.1
case 41	1,3,6,9	0.205396	0.379901	0.776770	55.6
case 42	1,3,6,8,9	0.014908	0.015983	0.017243	75.0
case 43	1,3,6,8,9	0.000002	0.000001	0.000001	76.3
case 44	1,3,6,8,9	0.000089	0.000033	0.000036	75.5
case 45	1,3,6,8,9	0.000076	0.000037	0.000040	75.1
case 46	1,3,6,9,10	0.010194	0.004262	0.546985	74.7

Table 4.3: The selected variables, proposed selection influence measure, Cook's conditional and unconditional distance, and the estimated average prediction error for each of the reduced data sets of the evaporation data

The histogram in Figure 4.2 shows the distribution of these values. The vertical line shows the value of the influence measure if data case 41 is omitted, which equals 0.2054. This value lies above the 97th percentile of the distribution, since 97.7% of the bootstrap influence measures are smaller than 0.2054. Evaluating therefore the influence measure for case 41 in terms of the distribution in Figure 4.2 will clearly suggest to any practitioner that omitting case 41, and thereby using only variables 1, 3, 6 and 9 in the final regression model, is appropriate.

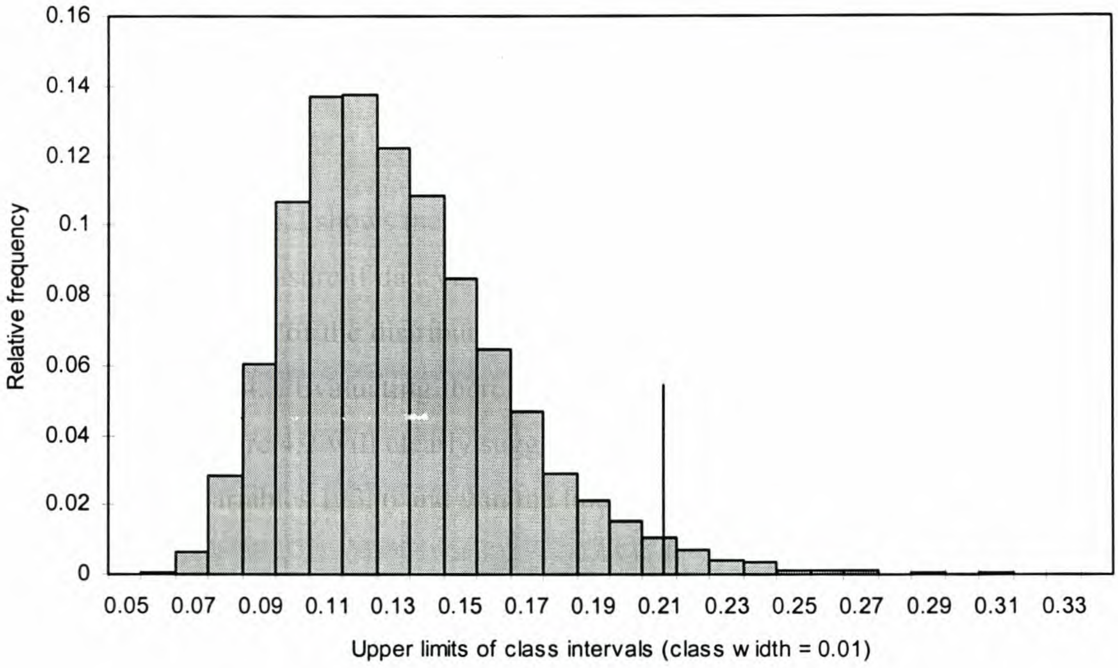


Figure 4.2: Histogram of largest proposed influence measures obtained in 10000 bootstrap samples of the evaporation data

Remark: After analysing the three example data sets, we come to the following conclusion: The proposed influence measure can easily be calculated for any multivariate regression sample that needs to be analysed. Once the influence measure has been obtained, a decision has to be taken on whether to omit the data case with the largest influence measure, before selection is repeated on the reduced data set. The magnitude by which the estimated average prediction error decreases, if calculated for the complete and reduced data set, provides us with a good indication of whether the data case should be omitted. The illustrated bootstrap approach can also be utilised to judge the significance of the largest proposed influence measure. We strongly recommend that both these aspects (i.e., the estimated average prediction error and the bootstrap distributicn) should be taken into consideration before the data case with the largest influence measure is merely excluded from the regression sample. ■

CHAPTER 5

IDENTIFICATION OF MULTIPLE SELECTION INFLUENTIAL DATA CASES

5.1 Introduction

In Chapter 4 we proposed an influence measure which can be used in multiple linear regression to identify single selection influential data cases. We also illustrated that in certain situations it will be beneficial in terms of the average prediction error of the resulting model to omit such a data case before variable selection is applied. In this chapter we develop a procedure which can be used for simultaneous identification of more than one selection influential data case. In order to explain the general argument underlying this proposed influence measure, consider once again the expansion of the mean squared error of $P_L\mathbf{Y}$ as an estimator of $\boldsymbol{\mu}$, as the sum of n terms, viz.

$$\begin{aligned} E\left\|P_L\mathbf{Y}-\boldsymbol{\mu}\right\|^2 &= \sum_{i=1}^{n}\left\{\sigma^2\left\|P_L\mathbf{u}_i\right\|^2+\langle\boldsymbol{\mu},P_{L^\perp}\mathbf{u}_i\rangle^2\right\} \\ &= \sum_{i=1}^{n}\left\{\sigma^2\left\|P_L\mathbf{u}_i\right\|^2+\left\langle\boldsymbol{\mu},P_{M|L}\mathbf{u}_i\right\rangle^2\right\}. \end{aligned} \tag{5.1}$$

Consider an unbiased estimator of this quantity, viz.

$$\sum_{i=1}^{n} D_{j,i}, \tag{5.2}$$

where the index j refers to a given subspace L. If t denotes the total number of possible subspaces (models) that can be considered for selection, we have $t = 2^m - 1$, with m denoting the total number of predictors in a multiple linear regression setup. If we use C_p variable selection, we effectively identify the subspace $\widehat{L}$ for which (5.2) is a minimum. We can therefore think of the tn values $D_{j,i}$, $i = 1, 2, ..., n$; $L \in \mathcal{L}$, as the basic data that have to be used in the variable selection process. Now consider the $t \times n$ matrix $\mathbf{D}$ with (j,i)th element equal to $D_{j,i}$. Note that the row index, j, in this matrix refers to the different subspaces or models, and that the column index, i, refers to the different points in the data set. For a given value of j we have a given subspace L, and the sum over i, i.e. the sum over the different columns, of the entries in the jth row gives us the quantity in (5.2).

How can we use the information summarised in the matrix $\mathbf{D}$ to identify potentially selection influential data cases? In general one's feeling is that if $d_{j,i}$, the (j,i)th observed value of $D_{j,i}$ in matrix $\mathbf{D}$, is very large or very small, it would signify that the observation i may be selection influential with respect to the subspace (or model) corresponding to j. If $d_{j,i}$ is very large compared to the other entries in the ith column, it would mean that observation i plays a significant role in preventing the subspace corresponding to j from being selected (remember that we select the subspace having the smallest row total). In such a case omitting observation i may easily cause the selected subspace to change. Similarly, if $d_{j,i}$ is very small compared to the other entries in the ith column, it would mean that observation i plays a significant role in promoting the selection of the subspace corresponding to j. Once again we may find that omitting observation i under such circumstances may cause the selected subspace to change. Viewed in this light our problem is therefore to decide whether any observed value of $D_{j,i}$ in the matrix $\mathbf{D}$ can be regarded as being extreme (i.e. very large or very small). The crucial question now is: how can such a decision be made?

It has be admitted that the question posed in the final sentence of the previous paragraph has many possible answers. Our answer to this question is to make use of the underlying distribution of $D_{j,i}$, the random variable yielding the observed value $d_{j,i}$. Using our knowledge of this distribution we intend to calculate or estimate the p-value defined by

$$P\left\{D_{j,i} > d_{j,i}\right\}. \tag{5.3}$$

If the calculated or estimated p-value is very small, it would signify that $d_{j,i}$ is probably a significantly large observation. Similarly, if the calculated or estimated p-value is very large, it would signify that $d_{j,i}$ is probably a significantly small observation. In both cases our conclusion would be that observation i is probably selection influential with respect to subspace L. In Sections 5.2 and 5.3 we proceed to a more detailed exposition of this approach. Section 5.2 deals with the case where σ^2 is known, and we consider p-values based on the underlying distributions of $\widehat{\gamma}_1(\mathbf{Y}, L, i)$ in (2.21), and $\widehat{\gamma}_2(\mathbf{Y}, L, i)$ in (2.23). In Section 5.3 we apply the above argument in the situation where σ^2 is unknown, and we consider p-values based on the underlying distributions of $C_p(\mathbf{Y}, L, i)$ in (2.11) and $\widetilde{C}_p(\mathbf{Y}, L, i)$ in (2.16). In this section, we also consider the p-values based on the underlying distribution of the random variables obtained if $C_p(\mathbf{Y}, L, i)$ and $\widetilde{C}_p(\mathbf{Y}, L, i)$ are expressed relative to the unbiased estimator of σ^2, i.e. the random variables $\frac{C_p(\mathbf{Y},L,i)}{\widehat{\sigma}^2}$ and $\frac{\widetilde{C}_p(\mathbf{Y},L,i)}{\widehat{\sigma}^2}$. We shall see that it is not possible to derive the underlying distribution of all these random variables. However, the distribution of the random variables considered in Section 5.2 and the distributions of those that can be derived in Section 5.3 turn out to be non-central chi-squared distributions. Of importance in the non-central chi-squared distribution, before the actual p-values can be determined, is estimation of the non-centrality parameter. This aspect is considered in Section 5.4. We finally apply the proposed influence measure to example data sets in Section 5.5, and evaluate its effectivity by means of simulation in Section 5.6.

5.2 The case where σ^2 is known

We consider first the case where the value of σ^2 is assumed known. This simplifies the development of our procedure, and is not an unrealistic assumption when the error degrees of freedom in a multiple linear regression is large. Since the value of σ^2 is known, the unknown quantity in the ESEE in (5.1), for a given subspace L, is simply

$$\sum_{i=1}^{n} \langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i\rangle^2 = \sum_{i=1}^{n} \langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\rangle^2 = \sum_{i=1}^{n} \gamma(\boldsymbol{\mu}, L, i).$$

This quantity is estimated unbiasedly by

$$\sum_{i=1}^{n} \widehat{\gamma}_1(\mathbf{Y}, L, i) = \sum_{i=1}^{n} \widehat{\gamma}_2(\mathbf{Y}, L, i)$$

where

$$\widehat{\gamma}_1(\mathbf{Y}, L, i) = \langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2 - \sigma^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2$$

and

$$\widehat{\gamma}_2(\mathbf{Y}, L, i) = \langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\rangle^2 - \sigma^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2.$$

are the two unbiased estimators of $\gamma(\boldsymbol{\mu}, L, i)$, respectively introduced in (2.21) and (2.23) of Section 2.3.2 of Chapter 2. Firstly, consider the estimator in (2.21). If $\widehat{\gamma}_1(\mathbf{y}, L, i)$ denotes the observed value of $\widehat{\gamma}_1(\mathbf{Y}, L, i)$, the p-value in (5.3) becomes

$$\begin{aligned} & P\left\{D_{j,i} > d_{j,i}\right\} \\ = & P\left\{\sigma^2 \left\|P_L \mathbf{u}_i\right\|^2 + \widehat{\gamma}_1(\mathbf{Y}, L, i) > \sigma^2 \left\|P_L \mathbf{u}_i\right\|^2 + \widehat{\gamma}_1(\mathbf{y}, L, i)\right\} \\ = & P\left\{\widehat{\gamma}_1(\mathbf{Y}, L, i) > \widehat{\gamma}_1(\mathbf{y}, L, i)\right\}. \end{aligned} \tag{5.4}$$

How can this p-value in (5.4) be calculated? For notational simplicity, let L be the jth linear subspace of all $\sum_{l=1}^{m} \binom{m}{l} - 1 = 2^m - 1$ possible linear subspaces containing at least one predictor as basis vector. Since $\mathbf{Y} \sim N_n\left(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n\right)$ it follows that

$$\left\langle \mathbf{Y}, P_{L^\perp} \mathbf{u}_i \right\rangle \sim N\left(\left\langle \boldsymbol{\mu}, P_{L^\perp} \mathbf{u}_i \right\rangle, \sigma^2 \left\|P_{L^\perp} \mathbf{u}_i\right\|^2\right). \tag{5.5}$$

It now follows that for the ith case and the jth linear subspace L, the random variable

$$V_{ij} = \frac{\left\langle \mathbf{Y}, P_{L^\perp} \mathbf{u}_i \right\rangle^2}{\sigma^2 \left\|P_{L^\perp} \mathbf{u}_i\right\|^2} \tag{5.6}$$

follows a non-central chi-squared distribution with 1 degree of freedom, non-centrality parameter

$$\lambda_{1ij}^2 = \frac{\left\langle \boldsymbol{\mu}, P_{L^\perp} \mathbf{u}_i \right\rangle^2}{\sigma^2 \left\|P_{L^\perp} \mathbf{u}_i\right\|^2} \tag{5.7}$$

and probability density function (pdf)

$$f_{V_{ij}}(v) = e^{-\frac{1}{2}\lambda_{1ij}^2} \sum_{s=0}^{\infty} \frac{\left(\frac{1}{2}\lambda_{1ij}^2\right)^s}{s!} \frac{v^{\frac{1}{2}(1+2s)-1} e^{-\frac{1}{2}v}}{\Gamma\left(\frac{1}{2}(1+2s)\right) 2^{\frac{1}{2}(1+2s)}}, v > 0. \tag{5.8}$$

In short we write $V_{ij} \sim \chi_1'^2(\lambda_{1ij}^2)$. The p-value that corresponds with the ith data case and the jth linear subspace L is now calculated as follows:

$$\begin{aligned} & P\left\{\widehat{\gamma}_1(\mathbf{Y}, L, i) > \widehat{\gamma}_1(\mathbf{y}, L, i)\right\} \\ = & P\left(\left\langle \mathbf{Y}, P_{L^\perp} \mathbf{u}_i \right\rangle^2 - \sigma^2 \left\|P_{L^\perp} \mathbf{u}_i\right\|^2 > \left\langle \mathbf{y}, P_{L^\perp} \mathbf{u}_i \right\rangle^2 - \sigma^2 \left\|P_{L^\perp} \mathbf{u}_i\right\|^2\right) \\ = & P\left(\left\langle \mathbf{Y}, P_{L^\perp} \mathbf{u}_i \right\rangle^2 > \left\langle \mathbf{y}, P_{L^\perp} \mathbf{u}_i \right\rangle^2\right) \\ = & P\left(\frac{\left\langle \mathbf{Y}, P_{L^\perp} \mathbf{u}_i \right\rangle^2}{\sigma^2 \left\|P_{L^\perp} \mathbf{u}_i\right\|^2} > \frac{\left\langle \mathbf{y}, P_{L^\perp} \mathbf{u}_i \right\rangle^2}{\sigma^2 \left\|P_{L^\perp} \mathbf{u}_i\right\|^2}\right) \\ = & P\left(V_{ij} > v_{ij}\right), \text{ where } v_{ij} = \frac{\left\langle \mathbf{y}, P_{L^\perp} \mathbf{u}_i \right\rangle^2}{\sigma^2 \left\|P_{L^\perp} \mathbf{u}_i\right\|^2} \\ = & \int_{v_{ij}}^{\infty} f_{V_{ij}}(v) dv. \end{aligned} \tag{5.9}$$

Expression (5.9) cannot be used directly to calculate the required p-value, since the pdf $f_{V_{ij}}(v)$ depends on the unknown vector $\boldsymbol{\mu}$ through the non-centrality parameter λ_{1ij}^2. Later on we will discuss estimation of the non-centrality parameter, our intention then being to calculate the required p-value from (5.9) with λ_{1ij}^2 replaced by its estimated value.

Remark: An interesting question that arises is whether an alternative form of the p-value may not be more informative. Consider in this regard the sum of the n estimators of the n parameters, $\gamma(\boldsymbol{\mu}, L, i)$, $i = 1, ..., n$, in the jth linear subspace L, viz.

$$\widehat{\gamma}_1(\mathbf{Y}, L) = \sum_{i=1}^{n} \widehat{\gamma}_1(\mathbf{Y}, L, i).$$

For all $i = 1, ..., n$, the following inequality holds:

$$\widehat{\gamma}_1(\mathbf{Y}, L, i) \leq \widehat{\gamma}_1(\mathbf{Y}, L).$$

Is it possible to use this additional information when we attempt to identify selection influential data cases in the linear subspace L? Consider in this regard the following conditional p-value:

$$P\left[\widehat{\gamma}_1(\mathbf{Y}, L, i) > \widehat{\gamma}_1(\mathbf{y}, L, i) \mid \widehat{\gamma}_1\left(\mathbf{Y}, L\right) = \widehat{\gamma}_1(\mathbf{y}, L)\right], \tag{5.10}$$

where $\widehat{\gamma}_1(\mathbf{y}, L) = \sum_{i=1}^{n} \widehat{\gamma}_1(\mathbf{y}, L, i)$ is the observed value of $\widehat{\gamma}_1\left(\mathbf{Y}, L\right)$. In order to calculate (5.10), we have to consider the distribution of $\widehat{\gamma}_1(\mathbf{Y}, L, i)$ conditional on $\widehat{\gamma}_1\left(\mathbf{Y}, L\right)$. This conditional distribution involves the joint pdf of

$$\widehat{\gamma}_1\left(\mathbf{Y}, L, 1\right), \widehat{\gamma}_1\left(\mathbf{Y}, L, 2\right), ..., \widehat{\gamma}_1\left(\mathbf{Y}, L, n\right).$$

These random variables are not in general independent. Consider in this regard the random variables $\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i \rangle$ and $\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_k \rangle$, for $i \neq k$. Recall that $\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i \rangle$ follows a $N\left(\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i \rangle, \sigma^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2\right)$ distribution. Hence, using the result in Lemma A.4 in Appendix A, the covariance between $\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i \rangle$ and $\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_k \rangle$ can be written as

$$\begin{aligned}
Cov\left(\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i \rangle, \langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_k \rangle\right) &= \sigma^2 \left\langle P_{L^\perp}\mathbf{u}_i, P_{L^\perp}\mathbf{u}_k \right\rangle \\
&= \sigma^2 \left\langle P_{L^\perp}\mathbf{u}_i, \mathbf{u}_k \right\rangle \\
&= \sigma^2 \left\langle \mathbf{u}_i - P_L\mathbf{u}_i, \mathbf{u}_k \right\rangle \\
&= \sigma^2 \left\{\langle \mathbf{u}_i, \mathbf{u}_k \rangle - \langle P_L\mathbf{u}_i, \mathbf{u}_k \rangle\right\} \\
&= \sigma^2 \left\langle P_L\mathbf{u}_i, \mathbf{u}_k \right\rangle.
\end{aligned}$$

In general, $\sigma^2 \langle P_L \mathbf{u}_i, \mathbf{u}_k \rangle$ does not equal zero, showing that $\langle \mathbf{Y}, P_{L^\perp} \mathbf{u}_i \rangle$ and $\langle \mathbf{Y}, P_{L^\perp} \mathbf{u}_k \rangle$, for $i \neq k$, are not in general independent. This makes it very difficult to determine the joint pdf of $\widehat{\gamma}_1 (\mathbf{Y}, L, 1), \widehat{\gamma}_1 (\mathbf{Y}, L, 2), ..., \widehat{\gamma}_1 (\mathbf{Y}, L, n)$. We conclude that calculation of the conditional probability in (5.10) is infeasible. ■

The p-value in (5.4) is based on the unbiased estimator $\widehat{\gamma}_1 (\mathbf{Y}, L, i)$ of $\gamma(\boldsymbol{\mu}, L, i)$. Similar p-values can also be obtained from the unbiased estimator $\widehat{\gamma}_2 (\mathbf{Y}, L, i)$ of $\gamma(\boldsymbol{\mu}, L, i)$. Again applying the general argument in Section 5.1, using now $\widehat{\gamma}_2 (\mathbf{Y}, L, i)$, the random variable yielding the observed value $\widehat{\gamma}_2(\mathbf{y}, L, i)$, as estimator of $\gamma(\boldsymbol{\mu}, L, i)$, the p-value in (5.3) becomes

$$\begin{aligned}
& P \{D_{j,i} > d_{j,i}\} \\
= \;& P \left\{\sigma^2 \|P_L \mathbf{u}_i\|^2 + \widehat{\gamma}_2(\mathbf{Y}, L, i) > \sigma^2 \|P_L \mathbf{u}_i\|^2 + \widehat{\gamma}_2(\mathbf{y}, L, i)\right\} \\
= \;& P \{\widehat{\gamma}_2(\mathbf{Y}, L, i) > \widehat{\gamma}_2(\mathbf{y}, L, i)\} \\
= \;& P \left(\langle \mathbf{Y}, P_{M|L} \mathbf{u}_i \rangle^2 > \langle \mathbf{y}, P_{M|L} \mathbf{u}_i \rangle^2\right) \\
= \;& P \left(\widetilde{V}_{ij} > \widetilde{v}_{ij}\right), \text{ where } \widetilde{v}_{ij} = \frac{\langle \mathbf{y}, P_{M|L} \mathbf{u}_i \rangle^2}{\sigma^2 \left\|P_{M|L} \mathbf{u}_i\right\|^2} \\
= \;& \int_{\widetilde{v}_{ij}}^{\infty} f_{\widetilde{V}_{ij}}(v) dv \qquad (5.11)
\end{aligned}$$

In (5.11) $f_{\widetilde{V}_{ij}}(v)$ is the pdf of the random variable

$$\widetilde{V}_{ij} = \frac{\langle \mathbf{Y}, P_{M|L} \mathbf{u}_i \rangle^2}{\sigma^2 \left\|P_{M|L} \mathbf{u}_i\right\|^2}. \qquad (5.12)$$

that follows a non-central chi-squared distribution with 1 degree of freedom and non-centrality parameter

$$\lambda_{2ij}^2 = \frac{\langle \boldsymbol{\mu}, P_{M|L} \mathbf{u}_i \rangle^2}{\sigma^2 \left\|P_{M|L} \mathbf{u}_i\right\|^2}. \qquad (5.13)$$

The probability density function of $\grave{V}_{ij}$ is given by

$$f_{\widetilde{V}_{ij}}(v) = e^{-\frac{1}{2}\lambda_{2ij}^2} \sum_{s=0}^{\infty} \frac{\left(\frac{1}{2}\lambda_{2ij}^2\right)^s}{s!} \frac{v^{\frac{1}{2}(1+2s)-1} e^{-\frac{1}{2}v}}{\Gamma\left(\frac{1}{2}(1+2s)\right) 2^{\frac{1}{2}(1+2s)}}, v > 0. \qquad (5.14)$$

Should the p-value in (5.9) or the p-value in (5.11) be used in order to decide if the ith data case is selection influential in the linear subspace L? In order to answer this question we should bear in mind that in determining either of these p-values, the corresponding non-centrality parameter of V_{ij} in (5.7) and of $\widetilde{V}_{ij}$ in (5.13) have to be estimated. Once the non-centrality parameters have been estimated we shall be in a better position to decide which one of the two p-values is more appropriate for the purpose of identifying selection influential data cases in the subspace L. Estimation of these non-centrality parameters receives thorough attention in Section 5.4 of this chapter. In the next section we first consider the case where σ^2 is unknown.

5.3 The case where σ^2 is unknown

5.3.1 P-values based on the distribution of $C_p(\mathbf{Y}, L, i)$

Identification of selection influential data cases when σ^2 is unknown, can also be based on the p-value approach. As in the previous section we first consider $\widehat{\gamma}_1(\mathbf{Y}, L, i)$ as an estimator of $\gamma(\boldsymbol{\mu}, L, \imath)$. Since the value of σ^2 is unknown, $\widehat{\gamma}_1(\mathbf{Y}, L, i)$ is now defined by

$$\widehat{\gamma}_1(\mathbf{Y}, L, i) = \langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i \rangle^2 - \widehat{\sigma}^2 \left\| P_{L^\perp}\mathbf{u}_i \right\|^2 \tag{5.15}$$

where σ^2 is estimated by its least squares estimator,

$$\widehat{\sigma}^2 = \frac{\left\|\mathbf{Y} - P_M\mathbf{Y}\right\|^2}{n-(m+1)} = \frac{\left\|P_{M^\perp}\mathbf{Y}\right\|^2}{n-(m+1)}$$

given in (1.10). If the term $\widehat{\sigma}^2 \left\|P_L\mathbf{u}_i\right\|^2$ is added to $\widehat{\gamma}_1(\mathbf{Y}, L, i)$, we obtain the expression in (2.11), which is the ith term in the expansion of the C_p criterion as the sum of n terms, i.e.

$$\begin{aligned} C_p(\mathbf{Y}, L, i) &= \widehat{\sigma}^2 \left\|P_L\mathbf{u}_i\right\|^2 + \widehat{\gamma}_1(\mathbf{Y}, L, i) \\ &= \widehat{\sigma}^2 \left\|P_L\mathbf{u}_i\right\|^2 + \langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i \rangle^2 - \widehat{\sigma}^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2 . \end{aligned}$$

Consider now the distribution of $C_p(\mathbf{Y}, L, i)$. Let

$$C_p(\mathbf{y}, L, i) = s^2 \left\|P_L\mathbf{u}_i\right\|^2 + \langle \mathbf{y}, P_{L^\perp}\mathbf{u}_i \rangle^2 - s^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2$$

denote the ith observed value of the random variable $C_p(\mathbf{Y}, L, i)$, where s^2 is the observed value of $\widehat{\sigma}^2$.

We can again argue that for the jth linear subspace L, the corresponding ith data case is selection influential if the value of $C_p(\mathbf{Y}, L, i)$ is very large or very small. Therefore, the corresponding ith data case is considered selection influential with respect to the linear subspace L if a significantly large or small value is obtained for a p-value in (5.3) which is now of the form

$$\begin{aligned}
& P\left\{D_{j,i} > d_{j,i}\right\} \\
= \; & P\left\{C_p(\mathbf{Y}, L, i) > C_p(\mathbf{y}, L, i)\right\} \\
= \; & P\{\widehat{\sigma}^2 \left\|P_L \mathbf{u}_i\right\|^2 + \langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2 - \widehat{\sigma}^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2 > s^2 \left\|P_L \mathbf{u}_i\right\|^2 + \langle \mathbf{y}, P_{L^\perp}\mathbf{u}_i\rangle^2 \\
& -s^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2 \} \\
= \; & P\{ \langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2 + \widehat{\sigma}^2 \left(2 \left\|P_L \mathbf{u}_i\right\|^2 - 1\right) > \langle \mathbf{y}, P_{L^\perp}\mathbf{u}_i\rangle^2 \\
& + s^2 (2 \left\|P_L \mathbf{u}_i\right\|^2 - 1)\}.
\end{aligned} \tag{5.16}$$

How can we calculate the p-value in (5.16)? Note that the random variable $C_p(\mathbf{Y}, L, i)$ is the sum of two functions of two different random variables. The first, $\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2 = \langle P_{L^\perp}\mathbf{Y}, \mathbf{u}_i\rangle^2$, is a function of $P_{L^\perp}\mathbf{Y}$, and the second, $\widehat{\sigma}^2 \left(2 \left\|P_L \mathbf{u}_i\right\|^2 - 1\right) = \frac{\left\|P_{M^\perp}\mathbf{Y}\right\|^2 \left(2\left\|P_L \mathbf{u}_i\right\|^2 - 1\right)}{n-(m+1)}$, is a function of $P_{M^\perp}\mathbf{Y}$. Since L is a linear subspace of M, the random variables $P_{L^\perp}\mathbf{Y} = \mathbf{Y} - P_L\mathbf{Y}$ and $P_{M^\perp}\mathbf{Y} = \mathbf{Y} - P_M\mathbf{Y}$ are not independent. This has the implication that the two functions of these random variables, $\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2$ and $\widehat{\sigma}^2 \left(2 \left\|P_L \mathbf{u}_i\right\|^2 - 1\right)$, are also not independent. Due to this dependence of $\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2$ and $\widehat{\sigma}^2 \left(2 \left\|P_L \mathbf{u}_i\right\|^2 - 1\right)$ it is difficult to write (5.16) in a form that is suitable for calculations. Better progress is possible with p-values based on the distribution of $\widehat{\sigma}^2 \left\|P_L \mathbf{u}_i\right\|^2 + \left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^2 - \widehat{\sigma}^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2$. Such p-values are discussed in the next section.

5.3.2 P-values based on the distribution of $\widetilde{C}_p(\mathbf{Y}, L, i)$

The same difficulties mentioned in the previous paragraph are not experienced if

$$\widehat{\gamma}_2\left(\mathbf{Y}, L, i\right) = \left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^2 - \widehat{\sigma}^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2$$

is considered as an estimator of $\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i\rangle^2$. If the term $\widehat{\sigma}^2 \left\|P_L \mathbf{u}_i\right\|^2$ is added to $\widehat{\gamma}_2\left(\mathbf{Y}, L, i\right)$, we obtain the expression in (2.16), which is the ith term in the expansion of the C_p criterion as the sum of n terms, i.e.

$$\widetilde{C}_p(\mathbf{Y}, L, i) = \widehat{\sigma}^2 \left\|P_L \mathbf{u}_i\right\|^2 + \left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^2 - \widehat{\sigma}^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2.$$

The corresponding p-value based on the sampling distribution of $\widetilde{C}_p(\mathbf{Y}, L, i)$ is of the form

$$\begin{aligned} & P\left\{\widetilde{C}_p(\mathbf{Y}, L, i) > \widetilde{C}_p(\mathbf{y}, L, i)\right\} \qquad (5.17) \\ = \; & P\{\widehat{\sigma}^2 \left\|P_L \mathbf{u}_i\right\|^2 + \left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i \right\rangle^2 - \widehat{\sigma}^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2 > s^2 \left\|P_L \mathbf{u}_i\right\|^2 + \left\langle \mathbf{y}, P_{M|L}\mathbf{u}_i \right\rangle^2 \\ & -s^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2 \} \\ = \; & P\{ \left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i \right\rangle^2 + \widehat{\sigma}^2 \left(2\left\|P_L \mathbf{u}_i\right\|^2 - \left\|P_M \mathbf{u}_i\right\|^2\right) > \left\langle \mathbf{y}, P_{M|L}\mathbf{u}_i \right\rangle^2 \\ & + s^2 \left(2\left\|P_L \mathbf{u}_i\right\|^2 - \left\|P_M \mathbf{u}_i\right\|^2\right) \} \end{aligned}$$

where $\widetilde{C}_p(\mathbf{y}, L, i)$ in (5.17) is the observed value of the random variable $\widetilde{C}_p(\mathbf{Y}, L, i)$.

Remark: Note that (5.17) is again a specialisation of the general argument proposed for the influence measure in Section 5.1, here $\widetilde{C}_p(\mathbf{Y}, L, i) = D_{j,i}$ and $\widetilde{C}_p(\mathbf{y}, L, i) = d_{j,i}$. ■

For notational simplicity we let

$$\begin{aligned} \widetilde{C}_p(\mathbf{Y}, L, i) &= \left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i \right\rangle^2 + \widehat{\sigma}^2 \left(2\left\|P_L \mathbf{u}_i\right\|^2 - \left\|P_M \mathbf{u}_i\right\|^2\right) \\ &= X_{ij} + W_{ij} \\ &= R_{ij}, \qquad (5.18) \end{aligned}$$

where $X_{ij} = \left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i \right\rangle^2$, and $W_{ij} = \widehat{\sigma}^2 \left(2\left\|P_L \mathbf{u}_i\right\|^2 - \left\|P_M \mathbf{u}_i\right\|^2\right)$, and where the index j again refers to a particular linear subspace L contained in M. In this notation the p-value in (5.17) becomes

$$P(R_{ij} > r_{ij}) \qquad (5.19)$$

where r_{ij} is the observed value of R_{ij}. The distribution of R_{ij} is required to determine the p-value in (5.17) or (5.19). In order to obtain this distribution, recall that in (5.12) it is stated that

$$\widetilde{V}_{ij} = \frac{\left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i \right\rangle^2}{\sigma^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2} \sim \chi_1'^2(\lambda_{2ij}^2),$$

with $\lambda_{2ij}^2 = \frac{\left\langle \mu, P_{M|L}\mathbf{u}_i \right\rangle^2}{\sigma^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2}$, and probability density function $f_{\widetilde{V}_{ij}}(v)$. For the ith data case and the jth linear subspace L, we therefore have

$$X_{ij} = \widetilde{V}_{ij}\sigma^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2 . \qquad (5.20)$$

This transformation from $\{v : f_{\widetilde{V}_{ij}}(v) > 0\}$ onto $\{x : f_{X_{ij}}(x) > 0\}$, yields the probability density function of X_{ij} as

$$f_{X_{ij}}(x) = \frac{1}{\sigma^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2} f_{\widetilde{V}_{ij}}\left(\frac{x}{\sigma^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2}\right). \tag{5.21}$$

The result that

$$T = \frac{(n-(m+1))\,\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-(m+1)} \tag{5.22}$$

where χ^2 is the central chi-squared distribution and $n-(m+1)$ its degrees of freedom, can easily be established. The pdf of T is given by

$$f_T(t) = \frac{t^{\frac{1}{2}(n-(m+1))-1} e^{-\frac{1}{2}t}}{\Gamma\left(\frac{1}{2}(n-(m+1))\right) 2^{\frac{1}{2}(n-(m+1))}}, t > 0. \tag{5.23}$$

A transformation from $\{t : f_T(t) > 0\}$ onto $\{w : f_{W_{ij}}(w) > 0\}$, where

$$W_{ij} = \frac{\sigma^2\left(2\left\|P_L\mathbf{u}_i\right\|^2 - \left\|P_M\mathbf{u}_i\right\|^2\right)}{(n-(m+1))} T, \tag{5.24}$$

yields the probability density function of W_{ij} as

$$f_{W_{ij}}(w) = \frac{(n-(m+1))}{\sigma^2\left\|P_{M|L}\mathbf{u}_i\right\|^2} f_T\left(\frac{w(n-(m+1))}{\sigma^2\left\|P_{M|L}\mathbf{u}_i\right\|^2}\right). \tag{5.25}$$

Note that the transformations in (5.20) and (5.24) were chosen such that their sum yields the random variable in (5.18), i.e.

$$\begin{aligned} & X_{ij} + W_{ij} \\ = \; & \widetilde{V}_{ij}\sigma^2\left\|P_{M|L}\mathbf{u}_i\right\|^2 + \frac{\sigma^2\left(2\left\|P_L\mathbf{u}_i\right\|^2 - \left\|P_M\mathbf{u}_i\right\|^2\right)}{(n-(m+1))} T \\ = \; & \left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i \right\rangle^2 + \widehat{\sigma}^2\left(2\left\|P_L\mathbf{u}_i\right\|^2 - \left\|P_M\mathbf{u}_i\right\|^2\right) \\ = \; & R_{ij}. \end{aligned}$$

Consider also a transformation from $\{w : f_{W_{ij}}(w) > 0\}$ onto $\{s : f_S(s) > 0\}$ by letting $S = W_{ij}$. Then $R_{ij} = X_{ij} + S$, and the joint pdf of R_{ij} and S is given by

$$f_{R_{ij},S}(r,s) = f_{X_{ij},W_{ij}}(r+s, s) \tag{5.26}$$

for $-\infty < r < \infty$ and $s > 0$. If we now assume that X_{ij} and W_{ij} are independently distributed (the validity of this assumption is established below), the joint probability density function of

X_{ij} and W_{ij} in (5.26) can be written as

$$f_{X_{ij},W_{ij}}(r+s,s) = f_{X_{ij}}(r+s)f_{W_{ij}}(s). \tag{5.27}$$

The marginal probability density function of R_{ij} follows as

$$f_{R_{ij}}(r) = \int_0^\infty f_{X_{ij}}(r+s)f_{W_{ij}}(s)ds.$$

By using the marginal probability density function of R_{ij}, the p-value in (5.17) and (5.19) can be written as

$$\begin{aligned} P(R_{ij} &> r_{ij}) = \int_{r_{ij}}^\infty f_{R_{ij}}(r)dr \\ &= \int_{r_{ij}}^\infty \int_0^\infty f_{X_{ij}}(r+s)f_{W_{ij}}(s)dsdr \\ &= \int_0^\infty f_{W_{ij}}(s)\left(\int_{r_{ij}}^\infty f_{X_{ij}}(r+s)dr\right)ds. \end{aligned}$$

Setting $r+s=u$ it follows that

$$\begin{aligned} P\left(R_{ij} > r_{ij}\right) &= \int_0^\infty f_{W_{ij}}(s)\left(\int_{r_{ij}+s}^\infty f_{X_{ij}}(u)du\right)ds \\ &= \int_0^\infty f_{W_{ij}}(s)\left(\int_{r_{ij}+s}^\infty \frac{1}{\sigma^2\left\|P_{M|L}\mathbf{u}_i\right\|^2} f_{\tilde{V}_{ij}}\left(\frac{u}{\sigma^2\left\|P_{M|L}\mathbf{u}_i\right\|^2}\right)du\right)ds. \end{aligned}$$

Finally, by setting $\frac{u}{\sigma^2\left\|P_{M|L}\mathbf{u}_i\right\|^2} = v$, the p-value in (5.17) and (5.19) becomes

$$P\left(R_{ij} > r_{ij}\right) = \int_0^\infty f_{W_{ij}}(s)\left(\int_{\frac{r_{ij}+s}{\sigma^2\left\|P_{M|L}\mathbf{u}_i\right\|^2}}^\infty f_{\tilde{V}_{ij}}\left(v\right)dv\right)ds. \tag{5.28}$$

It should be noted that the value of σ^2 in (5.28) is unknown. We therefore have to replace σ^2 by an estimate, $\widehat{\sigma}^2$, before (5.28) can be used to calculate an estimated p-value.

Recall that the expression for $P\left(R_{ij} > r_{ij}\right)$ above is based on the assumption that X_{ij} and W_{ij} are independent random variables. We now verify that this assumption indeed holds. Note that $X_{ij} = \left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^2 = \left\langle P_{M|L}\mathbf{Y}, \mathbf{u}_i\right\rangle^2$ is a function of $P_{M|L}\mathbf{Y}$, and that $W_{ij} = \widehat{\sigma}^2\left(2\left\|P_L\mathbf{u}_i\right\|^2 - \left\|P_M\mathbf{u}_i\right\|^2\right) = \frac{\left\|P_{M^\perp}\mathbf{Y}\right\|^2\left(2\|P_L\mathbf{u}_i\|^2 - \|P_M\mathbf{u}_i\|^2\right)}{n-(m+1)}$ is a function of $P_{M^\perp}\mathbf{Y}$. It is sufficient therefore to show that $P_{M|L}\mathbf{Y}$ is independent of $P_{M^\perp}\mathbf{Y}$, since this would imply that any function of $P_{M|L}\mathbf{Y}$ is independent of any function of $P_{M^\perp}\mathbf{Y}$. Thus, in our situation, the independence of $P_{M|L}\mathbf{Y}$ and $P_{M^\perp}\mathbf{Y}$ will yield independence of X_{ij} and W_{ij}. Since $P_{M|L}\mathbf{Y}$ and $P_{M^\perp}\mathbf{Y}$ are both normal random vectors, independence will follow if we can show that $E\left(P_{M|L}\mathbf{Y}\right)\left(P_{M^\perp}\mathbf{Y}\right)' = E\left(P_{M|L}\mathbf{Y}\right)\left[E\left(P_{M^\perp}\mathbf{Y}\right)\right]'$, since then the variance-covariance matrix of these two random vectors will be the null matrix. For this purpose, let

- $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n$ be an orthonormal basis for $\mathcal{R}^n$
- $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_l$ be an orthonormal basis for L, and
- $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_l,.\mathbf{v}_{l+1}, ..., \mathbf{v}_m$ be an orthonormal basis for M, with $m < n$.

It now follows that

$$\begin{aligned}
& E\left(P_{M|L}\mathbf{Y}\right)\left(P_{M^\perp}\mathbf{Y}\right)' \\
= \; & E\left(\sum_{i=l+1}^{m} \langle \mathbf{v}_i, \mathbf{Y}\rangle \mathbf{v}_i\right)\left(\sum_{j=m+1}^{n} \langle \mathbf{v}_j, \mathbf{Y}\rangle \mathbf{v}_j'\right) \\
= \; & \sum_{i=l+1}^{m} \sum_{j=m+1}^{n} \mathbf{v}_i\mathbf{v}_j' E\left[\langle \mathbf{v}_i, \mathbf{Y}\rangle \langle \mathbf{v}_j, \mathbf{Y}\rangle\right] \\
= \; & \sum_{i=l+1}^{m} \sum_{j=m+1}^{n} \mathbf{v}_i\mathbf{v}_j' E\left[\left(\sum_{k=1}^{n} v_{ik}Y_k\right)\left(\sum_{q=1}^{n} v_{jq}Y_q\right)\right] \\
= \; & \sum_{i=l+1}^{n} \sum_{j=m+1}^{n} \mathbf{v}_i\mathbf{v}_j' \left[\sum_{k=1}^{n} v_{ik}v_{jk}E\left(Y_k^2\right) + \sum_{k\neq q} v_{ik}v_{jq}E\left(Y_kY_q\right)\right] \\
= \; & \sum_{i=l+1}^{m} \sum_{j=m+1}^{n} \mathbf{v}_i\mathbf{v}_j' \left[\sigma^2 \sum_{k=1}^{n} v_{ik}v_{jk} + \left\{\sum_{k=1}^{n} v_{ik}v_{jk}\mu_k^2 + \sum_{k\neq q} v_{ik}v_{jq}\mu_k\mu_q\right\}\right] \\
= \; & \sum_{i=l+1}^{m} \sum_{j=m+1}^{n} \mathbf{v}_i\mathbf{v}_j' \left[\sigma^2 \langle \mathbf{v}_i, \mathbf{v}_j\rangle + \langle \mathbf{v}_i, \boldsymbol{\mu}\rangle \langle \mathbf{v}_j, \boldsymbol{\mu}\rangle\right] \\
= \; & \sum_{i=l+1}^{m} \sum_{j=m+1}^{n} \mathbf{v}_i\mathbf{v}_j' \left[\langle \mathbf{v}_i, \boldsymbol{\mu}\rangle \langle \mathbf{v}_j, \boldsymbol{\mu}\rangle\right] \\
= \; & \left(\sum_{i=l+1}^{m} \mathbf{v}_i \langle \mathbf{v}_i, \boldsymbol{\mu}\rangle\right)\left(\sum_{j=m+1}^{n} \mathbf{v}_j' \langle \mathbf{v}_j, \boldsymbol{\mu}\rangle\right) \\
= \; & E\left(P_{M|L}\mathbf{Y}\right) E\left(P_{M^\perp}\mathbf{Y}\right)'.
\end{aligned}$$

This shows the independence of $P_{M|L}\mathbf{Y}$ and $P_{M^\perp}\mathbf{Y}$, implying therefore also independence of X_{ij} and W_{ij}.

The independence of $P_{M|L}\mathbf{Y}$ and $P_{M^\perp}\mathbf{Y}$ also implies that $\widetilde{V}_{ij}$ and T are independently distributed, since $\widetilde{V}_{ij}$ is a function of $P_{M|L}\mathbf{Y}$, and T is a function of $P_{M^{\perp}}\mathbf{Y}$. The distribution of

$$F = \frac{\widetilde{V}_{ij}/1}{T/(n-(m+1))} = \frac{\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i \rangle^2}{\widehat{\sigma}^2 \left\| P_{M|L}\mathbf{u}_i \right\|^2}$$

will thus be a non-central F distribution with degrees of freedom 1 and $(n-(m+1))$ and non-centrality parameter λ^2_{2ij} as given in (5.13). In short we write

$$F = \frac{\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i \rangle^2}{\widehat{\sigma}^2 \left\| P_{M|L}\mathbf{u}_i \right\|^2} \sim F'_{1,(n-(m+1))}\left(\lambda^2_{2ij}\right).$$

5.3.3 P-values based on the distribution of $\frac{C_p(\mathbf{Y},L,i)}{\widehat{\sigma}^2}$

In Section 5.3.1 we investigated the possibility of calculating p-values using the distribution of $C_p(\mathbf{Y}, L, i)$. In this section we briefly discuss p-values based on the distribution of $C_p(\mathbf{Y}, L, i)$, expressed relative to $\widehat{\sigma}^2$. Consider therefore the random variable

$$\begin{aligned} \frac{C_p(\mathbf{Y},L,i)}{\widehat{\sigma}^2} &= \left\|P_L\mathbf{u}_i\right\|^2 + \frac{\langle \mathbf{Y}, P_{L^{\perp}}\mathbf{u}_i\rangle^2}{\widehat{\sigma}^2} - \left\|P_{L^{\perp}}\mathbf{u}_i\right\|^2 \\ &= (2\left\|P_L\mathbf{u}_i\right\|^2 - 1) + \frac{\langle \mathbf{Y}, P_{L^{\perp}}\mathbf{u}_i\rangle^2}{\widehat{\sigma}^2} \end{aligned} \tag{5.29}$$

This random variable is an estimator of the ith term of the ESEE expressed relative to σ^2, viz.

$$\frac{E\left[\widehat{\mu}_i(L) - \mu_i\right]^2}{\sigma^2} = \frac{E\left[\widehat{Y}_i(L) - \mu_i\right]^2}{\sigma^2} = \left\|P_L\mathbf{u}_i\right\|^2 + \frac{\langle \boldsymbol{\mu}, P_{L^{\perp}}\mathbf{u}_i\rangle^2}{\sigma^2}. \tag{5.30}$$

Let s^2 denote the observed value of $\widehat{\sigma}^2$. Since $(2\left\|P_L\mathbf{u}_i\right\|^2 - 1)$ in (5.29) is independent of the data, the general p-value in (5.3), where $D_{j,i}$ now estimates (5.30), becomes

$$P\left\{\frac{\langle \mathbf{Y}, P_{L^{\perp}}\mathbf{u}_i\rangle^2}{\widehat{\sigma}^2} > \frac{\langle \mathbf{y}, P_{L^{\perp}}\mathbf{u}_i\rangle^2}{s^2}\right\} \tag{5.31}$$

where $\frac{\langle \mathbf{y}, P_{L^{\perp}}\mathbf{u}_i\rangle^2}{s^2}$ is an observation of the random variable $\frac{\langle \mathbf{Y}, P_{L^{\perp}}\mathbf{u}_i\rangle^2}{\widehat{\sigma}^2}$. In Section 5.3.1, where p-values based on the distribution of $C_p(\mathbf{Y}, L, i)$ were discussed, it was pointed out that $\langle \mathbf{Y}, P_{L^{\perp}}\mathbf{u}_i\rangle^2$ and $\widehat{\sigma}^2$ are not independently distributed. Because of this dependence progress with the p-value in (5.31) is very difficult. We therefore rather consider p-values based on the distribution of $\widetilde{C}_p(\mathbf{Y}, L, i)/\widehat{\sigma}^2$ in the next section.

5.3.4 P-values based on the distribution of $\frac{\widetilde{C}_p(\mathbf{Y},L,i)}{\widehat{\sigma}^2}$

Consider again the ith term of the ESEE expressed relative to σ^2, as given in (5.30). We now estimate $\frac{\langle \mu, P_{M|L}\mathbf{u}_i\rangle^2}{\sigma^2}$ in (5.30) unbiasedly by

$$\frac{\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\rangle^2}{\widehat{\sigma}^2} - \left\|P_{M|L}\mathbf{u}_i\right\|^2.$$

This leads to the ith term in the expansion of the $\widetilde{C}_p$ criterion, expressed relative to $\widehat{\sigma}^2$, being

$$\begin{aligned}\frac{\widetilde{C}_p(\mathbf{Y},L,i)}{\widehat{\sigma}^2} &= \left\|P_L\mathbf{u}_i\right\|^2 + \frac{\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\rangle^2}{\widehat{\sigma}^2} - \left\|P_{M|L}\mathbf{u}_i\right\|^2 \\ &= \left(2\left\|P_L\mathbf{u}_i\right\|^2 - \left\|P_M\mathbf{u}_i\right\|^2\right) + \frac{\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\rangle^2}{\widehat{\sigma}^2}. \qquad (5.32)\end{aligned}$$

Note that $\left(2\left\|P_L\mathbf{u}_i\right\|^2 - \left\|P_M\mathbf{u}_i\right\|^2\right)$ is independent of the data. We therefore consider the distribution of $\frac{\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\rangle^2}{\widehat{\sigma}^2}$, which is denoted by S_{ij}. Let $s_{ij} = \frac{\langle \mathbf{y}, P_{M|L}\mathbf{u}_i\rangle^2}{s^2}$ be an observation of S_{ij}. The p-value in (5.3), where (5.30) is estimated by $\left(2\left\|P_L\mathbf{u}_i\right\|^2 - \left\|P_M\mathbf{u}_i\right\|^2\right) + S_{ij}$, now becomes

$$\begin{aligned} & P\left\{S_{ij} > s_{ij}\right\} \\ = \; & P\left\{\frac{\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\rangle^2}{\widehat{\sigma}^2} > \frac{\langle \mathbf{y}, P_{M|L}\mathbf{u}_i\rangle^2}{s^2}\right\} \\ = \; & P\left\{\frac{\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\rangle^2}{\widehat{\sigma}^2\left\|P_{M|L}\mathbf{u}_i\right\|^2}\left(\frac{\sigma^2}{\sigma^2}\right) > \frac{\langle \mathbf{y}, P_{M|L}\mathbf{u}_i\rangle^2}{s^2\left\|P_{M|L}\mathbf{u}_i\right\|^2}\right\} \\ = \; & P\left\{\frac{\frac{\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\rangle^2}{\sigma^2\left\|P_{M|L}\mathbf{u}_i\right\|^2}}{\frac{\widehat{\sigma}^2}{\sigma^2}} > \frac{\langle \mathbf{y}, P_{M|L}\mathbf{u}_i\rangle^2}{s^2\left\|P_{M|L}\mathbf{u}_i\right\|^2}\right\} \\ = \; & P\left\{\frac{\widetilde{V}_{ij}}{U} > \frac{\langle \mathbf{y}, P_{M|L}\mathbf{u}_i\rangle^2}{s^2\left\|P_{M|L}\mathbf{u}_i\right\|^2}\right\} \qquad (5.33)\end{aligned}$$

where $\widetilde{V}_{ij} = \frac{\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\rangle^2}{\sigma^2\left\|P_{M|L}\mathbf{u}_i\right\|^2} \sim \chi_1'^2(\lambda_{2ij}^2)$ (cf. (5.12)) and $U = \frac{\widehat{\sigma}^2}{\sigma^2}$. Note that $T = U(n-(m+1))$ follows a chi-squared distribution with $(n-(m+1))$ degrees of freedom (cf. (5.22)) and probability density function $f_T(t)$ given in (5.23). This transformation from $\{t : f_T(t) > 0\}$ onto $\{u : f_U(u) > 0\}$, yields the following probability density function for U:

$$f_U(u) = (n-(m+1))f_T(u(n-(m+1)).$$

Since $\widetilde{V}_{ij}$ and U are independent we can easily condition on U. Consequently the p-value in

(5.33) becomes

$$\begin{aligned}
P(S_{ij} &> s_{ij}) = P\left(\widetilde{V}_{ij} > s_{ij}U\right) \\
&= E_U\left\{P\left(\widetilde{V}_{ij} > Us_{ij}|U=u\right)\right\} \\
&= E_U\left\{P\left(\widetilde{V}_{ij} > us_{ij}\right)\right\} \\
&= E_U\left\{\int_{us_{ij}}^{\infty} f_{\widetilde{V}_{ij}}(v)dv\right\} \\
&= \int_0^{\infty}\left\{\int_{us_{ij}}^{\infty} f_{\widetilde{V}_{ij}}(v)dv\right\} f_U(u)du \\
&= \int_0^{\infty}\left\{\int_{us_{ij}}^{\infty} f_{\widetilde{V}_{ij}}(v)dv\right\} (n-(m+1))f_T(u(n-(m+1))du. \qquad (5.34)
\end{aligned}$$

The integral in (5.34) can be calculated using numerical integration. A significantly large or small answer will once again indicate that the corresponding ith case is selection influential with respect to the linear subspace L. The quality of the conclusions drawn in this regard from (5.34) will again strongly depend on estimation of the non-centrality parameter λ_{2ij}^2 of the non-central chi-squared random variable $\widetilde{V}_{ij}$. This aspect should be investigated before (5.34) is used for the identification of selection influential data cases.

Remark: We conclude this section by showing how the p-value in (5.34) can be refined by using an unbiased estimator of $\frac{1}{\sigma^2}$ in the estimation of $\frac{E[\widehat{\mu}_i(L)-\mu_i]^2}{\sigma^2}$. In order to obtain such an unbiased estimator of $\frac{1}{\sigma^2}$, reconsider the random variable T given in (5.22) which follows a chi-squared distribution with $E(T) = n-(m+1)$ and

$$\begin{aligned}
E\left(\frac{1}{T}\right) &= E\left(\frac{\sigma^2}{(n-(m+1))\widehat{\sigma}^2}\right) \\
&= \frac{1}{\Gamma\left(\frac{1}{2}(n-(m+1))\right)2^{\frac{1}{2}(n-(m+1))}}\int_0^{\infty} t^{\left(\frac{1}{2}(n-(m+1))-1\right)-1}e^{-\frac{1}{2}t}dt \\
&= \frac{\Gamma\left(\frac{1}{2}(n-(m+1))-1\right)2^{\frac{1}{2}(n-(m+1))-1}}{\Gamma\left(\frac{1}{2}(n-(m+1))\right)2^{\frac{1}{2}(n-(m+1))}} \\
&= \frac{\Gamma\left(\frac{1}{2}(n-(m+1))-1\right)}{2\Gamma\left(\frac{1}{2}(n-(m+1))\right)} \\
&= \frac{1}{n-(m+1)-2}.
\end{aligned}$$

If we now consider

$$E\left(\frac{1}{\widehat{\sigma}^2}\right) = \frac{n-(m+1)}{\sigma^2} E\left(\frac{1}{T}\right) = \left(\frac{n-(m+1)}{n-(m+1)-2}\right)\frac{1}{\sigma^2},$$

it is clear that an unbiased estimator of $\frac{1}{\sigma^2}$ is given by

$$\left(\frac{n-(m+1)-2}{n-(m+1)}\right)\frac{1}{\widehat{\sigma}^2}. \tag{5.35}$$

To determine the refined p-value we now consider the distribution of the random variable $\widetilde{S}_{ij} = \frac{(n-(m+1)-2)\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\rangle^2}{(n-(m+1))\widehat{\sigma}^2}$ with an observed value of $\widetilde{s}_{ij} = \frac{(n-(m+1)-2)\langle \mathbf{y}, P_{M|L}\mathbf{u}_i\rangle^2}{(n-(m+1))s^2}$. Arguing along the same lines as above, the p-value can be determined from

$$\begin{aligned} P(\widetilde{S}_{ij} &> \widetilde{s}_{ij}) \\ &= \int_0^\infty \left\{ \int_{u\widetilde{s}_{ij}}^\infty f_{\widetilde{V}_{ij}}(v)dv \right\} (n-(m+1)) f_T(u(n-(m+1))du. \end{aligned} \tag{5.36}$$

■

5.4 Estimation of the non-centrality parameter

5.4.1 The case where σ^2 is known

In the previous section we developed a method to identify multiple selection influential data points by using a p-value approach. In particular, for the case where σ^2 is known, we can either use the p-value in (5.9) based on the non-central chi-squared distribution of V_{ij} in (5.6) or the p-value in (5.11) based on the non-central chi-squared distribution of $\widetilde{V}_{ij}$ in (5.12). Determining either of these p-values requires estimation of the corresponding non-centrality parameter of the relevant non-central chi-squared distribution. These non-centrality parameters of V_{ij} and $\widetilde{V}_{ij}$ are respectively given in (5.7) and (5.13) as

$$\lambda_{1ij}^2 = \frac{\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i\rangle^2}{\sigma^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2} = \frac{\gamma(\boldsymbol{\mu}, L, i)}{\sigma^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2}$$

and

$$\lambda_{2ij}^2 = \frac{\langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\rangle^2}{\sigma^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2} = \frac{\gamma(\boldsymbol{\mu}, L, i)}{\sigma^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2}.$$

Should we now use the p-value based on the distribution of V_{ij} or the p-value based on the distribution of $\widetilde{V}_{ij}$ in an attempt to identify selection influential data cases? To answer this question we should consider estimation of the non-centrality parameters of the non-central chi-squared distributions of V_{ij} and $\widetilde{V}_{ij}$. The accuracy with which these non-centrality parameters are estimated will naturally affect the accuracy of estimation of the p-values. Here we will measure the accuracy in terms of the variances of unbiased estimators of λ_{1ij}^2 and λ_{2ij}^2. The estimated p-value corresponding to the smallest variance non-centrality parameter estimator will be preferred in the identification of selection influential data cases.

For notational convenience let the non-centrality parameter of the ith case and the jth linear subspace L, λ_{1ij}^2, be denoted by λ_1^2. In a similar way, let λ_{2ij}^2 be denoted by λ_2^2. Estimation of either λ_1^2 or λ_2^2 only involves the estimation of $\langle \boldsymbol{\mu}, P_{L^\perp}\mathbf{u}_i\rangle^2 = \langle \boldsymbol{\mu}, P_{M|L}\mathbf{u}_i\rangle^2 = \gamma(\boldsymbol{\mu}, L, i)$, since the respective terms in the denominators of the two non-centrality parameters are known. In (2.21) and (2.23) of Chapter 2, we presented $\widehat{\gamma}_1(\mathbf{Y}, L, i)$ and $\widehat{\gamma}_2(\mathbf{Y}, L, i)$ as two unbiased estimators of $\gamma(\boldsymbol{\mu}, L, i)$. As previously, for notational convenience, let $\gamma(\boldsymbol{\mu}, L, i)$ be denoted by γ, $\widehat{\gamma}_1(\mathbf{Y}, L, i)$ by $\hat{\gamma}_1$ and $\widehat{\gamma}_2(\mathbf{Y}, L, i)$ by $\widehat{\gamma}_2$. Also in Section 2.3.2 of Chapter 2 we showed analytically, and by means of a limited simulation study, for the case where σ^2 is known, that $\widehat{\gamma}_2$ is a relatively more efficient estimator of γ, than $\widehat{\gamma}_1$, i.e.

$$Var\left(\widehat{\gamma}_2\right) \leqq Var\left(\widehat{\gamma}_1\right).$$

For this reason we prefer to use $\widehat{\gamma}_2$ as an unbiased estimator of γ when estimating either λ_1 or λ_2 unbiasedly. The unbiased estimators of λ_1 and λ_2 are therefore given by

$$\widehat{\lambda}_1^2 = \frac{\widehat{\gamma}_2}{\sigma^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2} \tag{5.37}$$

and

$$\widehat{\lambda}_2^2 = \frac{\widehat{\gamma}_2}{\sigma^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2}. \tag{5.38}$$

Consider now the variances of these two estimators:

$$Var\left(\widehat{\lambda}_1^2\right) = \frac{Var\left(\widehat{\gamma}_2\right)}{\left(\sigma^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2\right)^2} = \frac{Var\left(\widehat{\gamma}_2\right)}{\sigma^4 \left\|P_{L^\perp}\mathbf{u}_i\right\|^4}$$

and

$$Var\left(\widehat{\lambda_2^2}\right) = \frac{Var\left(\widehat{\gamma}_2\right)}{\left(\sigma^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2\right)^2} = \frac{Var\left(\widehat{\gamma}_2\right)}{\sigma^4 \left\|P_{M|L}\mathbf{u}_i\right\|^4}.$$

Belsley, Kuh and Welsch (1980, p. 66) show that

$$\left\|P_M\mathbf{u}_i\right\|^2 \leqq 1.$$

This implies that

$$\left\|P_M\mathbf{u}_i\right\|^2 - \left\|P_L\mathbf{u}_i\right\|^2 \leqq 1 - \left\|P_L\mathbf{u}_i\right\|^2 \Longleftrightarrow \left\|P_{M|L}\mathbf{u}_i\right\|^2 \leqq \left\|P_{L^\perp}\mathbf{u}_i\right\|^2,$$

and consequently that

$$Var\left(\widehat{\lambda_1^2}\right) \leqq Var\left(\widehat{\lambda_2^2}\right). \tag{5.39}$$

Since the non-centrality parameter λ_1^2 of the non-central chi-squared random variable V_{ij} in (5.6) is estimated with smaller variance than the non-centrality parameter λ_2^2 of non-central chi-squared random variable $\widetilde{V}_{ij}$ in (5.12), we utilise the p-value in (5.9), based on the distribution of V_{ij}, rather than the p-value in (5.11) based on the distribution of $\widetilde{V}_{ij}$, to identify selection influential data cases.

Remark: As a matter of interest, note that if γ in λ_1^2 is unbiasedly estimated by $\widehat{\gamma}_1$ rather than $\widehat{\gamma}_2$, the inequality in (5.39) does not always hold. To see this let

$$\widehat{\lambda_3^2} = \frac{\widehat{\gamma}_1}{\sigma^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2} = \frac{\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2}{\sigma^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2} - 1$$

be the corresponding unbiased estimator of λ_1^2 with

$$Var\left(\widehat{\lambda_3^2}\right) = \frac{Var\left(\widehat{\gamma}_1\right)}{\left(\sigma^2 \left\|P_{L^\perp}\mathbf{u}_i\right\|^2\right)^2} = \frac{Var\left(\widehat{\gamma}_1\right)}{\sigma^4 \left\|P_{L^\perp}\mathbf{u}_i\right\|^4}.$$

Since $Var\left(\widehat{\gamma}_2\right) \leqq Var\left(\widehat{\gamma}_1\right)$ and $\left\|P_{M|L}\mathbf{u}_i\right\|^2 \leqq \left\|P_{L^\perp}\mathbf{u}_i\right\|^2$, it follows that

$$\frac{Var\left(\widehat{\gamma}_2\right)}{Var\left(\widehat{\gamma}_1\right)} \leqq 1 \tag{5.40}$$

and

$$\frac{\left\|P_{M|L}\mathbf{u}_i\right\|^2}{\left\|P_{L^\perp}\mathbf{u}_i\right\|^2} \leqq 1.$$

If we assume now that

$$\frac{\left\|P_{M|L}\mathbf{u}_i\right\|^4}{\left\|P_{L^\perp}\mathbf{u}_i\right\|^4} \leqq \frac{Var\left(\widehat{\gamma}_2\right)}{Var\left(\widehat{\gamma}_1\right)} \leqq 1$$

it follows that

$$\frac{Var\left(\widehat{\gamma}_1\right)}{\sigma^4\left\|P_{L^\perp}\mathbf{u}_i\right\|^4} \leqq \frac{Var\left(\widehat{\gamma}_2\right)}{\sigma^4\left\|P_{M|L}\mathbf{u}_i\right\|^4}$$

and hence

$$Var\left(\widehat{\lambda}_3^2\right) \leqq Var\left(\widehat{\lambda}_2^2\right).$$

However, on the other hand, if we assume that

$$\frac{Var\left(\widehat{\gamma}_2\right)}{Var\left(\widehat{\gamma}_1\right)} \leqq \frac{\left\|P_{M|L}\mathbf{u}_i\right\|^4}{\left\|P_{L^\perp}\mathbf{u}_i\right\|^4} \leqq 1,$$

it follows in a similar way that

$$Var\left(\widehat{\lambda}_3^2\right) \geqq Var\left(\widehat{\lambda}_2^2\right).$$

The relative sizes of the variances of $\widehat{\lambda}_3^2$ and $\widehat{\lambda}_2^2$ therefore depends on which of $\frac{Var(\widehat{\gamma}_2)}{Var(\widehat{\gamma}_1)} \leqq 1$ or $\frac{\left\|P_{M|L}\mathbf{u}_i\right\|^4}{\left\|P_{L^\perp}\mathbf{u}_i\right\|^4} \leqq 1$ is closer to 1.

We also used simulation to investigate the relative sizes of $Var\left(\widehat{\lambda}_2^2\right)$ and $Var\left(\widehat{\lambda}_3^2\right)$, and found in almost every case that $Var\left(\widehat{\lambda}_3^2\right)$ is smaller than $Var\left(\widehat{\lambda}_2^2\right)$. For example, consider again the numerical evaluation of $Var\left(\widehat{\gamma}_1\right)$ and $Var\left(\widehat{\gamma}_2\right)$ in Section 2.3.2 of Chapter 2. The vectors

$$\mathbf{m}_1 = [\overbrace{1,...,1}^{10},\overbrace{0,...,0}^{10}]' \text{ and } \mathbf{m}_2 = [\overbrace{0,...,0}^{10},\overbrace{1...,1}^{10}]'$$

were chosen as basis vectors for the two-dimensional linear subspace $M = span\left\{\mathbf{m}_1, \mathbf{m}_2\right\}$. We also defined the linear subspace $L = span\left\{\mathbf{m}_2\right\}$. Vectors of the form

$$\boldsymbol{\mu} = \alpha\mathbf{m}_1 + \beta\mathbf{m}_2 \in M$$

were obtained for $\beta = 1$ and α-values which varied from -30 to 30 in steps of size 1. The variances of $\widehat{\gamma}_1(\mathbf{Y}, L, 1)$ and $\widehat{\gamma}_2(\mathbf{Y}, L, 1)$, as estimators $\gamma(\mathbf{Y}, L, 1)$, were then approximated by repeatedly simulating error vectors $\boldsymbol{\varepsilon} \sim N_{20}(\mathbf{0}, \sigma^2\mathbf{I}_{20})$ and computing $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$. Remember that $Var\left(\widehat{\lambda}_3^2\right) \leqq Var\left(\widehat{\lambda}_2^2\right)$ if and only if $\frac{\sigma^4\left\|P_{M|L}\mathbf{u}_i\right\|^4}{\sigma^4\left\|P_{L^\perp}\mathbf{u}_i\right\|^4} \leqq \frac{Var(\widehat{\gamma}_2)}{Var(\widehat{\gamma}_1)}$. From Figure 2.1 in Chapter 2 it is apparent that the ratio of the approximated variances, $\frac{Var\left(\widehat{\gamma}_1^2(\mathbf{Y},L,1)\right)}{Var\left(\widehat{\gamma}_2^2(\mathbf{Y},L,1)\right)}$, regardless of the value of σ, is more or less equal to 10 for large positive and negative α-values. For α-values close to zero the ratio increases to a value which approximately equals 100 for $\sigma = 1$. Using (2.31) and (2.34) it follows that

$$\|P_{L^\perp}\mathbf{u}_1\|^2 = 1 \text{ and } \left\|P_{M|L}\mathbf{u}_1\right\|^2 = 0.1.$$

Consequently $\frac{\left\|P_{L^\perp}\mathbf{u}_i\right\|^4}{\left\|P_{M|L}\mathbf{u}_i\right\|^4} = 100$ which, except when $\alpha = 0$ for some of the σ-values, is larger than $\frac{Var\left(\widehat{\gamma}_1^2(\mathbf{Y},L,1)\right)}{Var\left(\widehat{\gamma}_2^2(\mathbf{Y},L,1)\right)}$. It therefore seems reasonable to conclude here that $\frac{\left\|P_{M|L}\mathbf{u}_i\right\|^4}{\left\|P_{L^\perp}\mathbf{u}_i\right\|^4} < \frac{Var(\widehat{\gamma}_2)}{Var(\widehat{\gamma}_1)}$, which implies that $Var\left(\widehat{\lambda}_3^2\right) < Var\left(\widehat{\lambda}_2^2\right)$.

Also in a multiple linear regression context, it becomes very clear why $Var\left(\widehat{\lambda}_3^2\right)$ is usually smaller than $Var\left(\widehat{\lambda}_2^2\right)$. From (2.45) and (2.52) in Section 2.4.1, we know that $\|P_{L^\perp}\mathbf{u}_i\|^2 = 1 - v_{ii}$ and $\left\|P_{M|L}\mathbf{u}_i\right\|^2 = u_{ii} - v_{ii}$, where v_{ii} is the ith diagonal element of $P_L = \mathbf{X}_L(\mathbf{X}_L'\mathbf{X}_L)^{-1}\mathbf{X}_L'$, and u_{ii} is the ith diagonal element of $P_M = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. We found in practical applications of multiple linear regression analysis that values obtained for u_{ii} were significantly smaller than 1. Since $v_{ii} \leqq u_{ii}$, this implies that v_{ii} is even smaller, which consequently results in a very small value for $\frac{u_{ii}-v_{ii}}{1-v_{ii}}$. Also for subsets which contain a large number of predictor variables, values of u_{ii} and v_{ii} are close to each other. Especially for these cases, values for $\frac{u_{ii}-v_{ii}}{1-v_{ii}}$ is extremely small. We found in most of the regression applications that these values of $\frac{u_{ii}-v_{ii}}{1-v_{ii}}$ were usually smaller than $\frac{Var(\widehat{\gamma}_2)}{Var(\widehat{\gamma}_1)}$, which implies that $Var\left(\widehat{\lambda}_3^2\right)$ is usually smaller than $Var\left(\widehat{\lambda}_2^2\right)$. ■

Our estimator of choice for λ_1^2 is, however its unbiased estimator $\widehat{\lambda}_1^2$ in (5.37). Note however that the non-centrality parameter λ_1^2 is always positive. However, λ_1^2 may be negatively estimated by its unbiased estimator,

$$\begin{aligned}\widehat{\lambda}_1^2 &= \frac{\left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^2 - \sigma^2\left\|P_{M|L}\mathbf{u}_i\right\|^2}{\sigma^2\left\|P_{L^\perp}\mathbf{u}_i\right\|^2} \\ &= \frac{\left\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\right\rangle^2}{\sigma^2\left\|P_{L^\perp}\mathbf{u}_i\right\|^2} - \frac{\left\|P_{M|L}\mathbf{u}_i\right\|^2}{\left\|P_{L^\perp}\mathbf{u}_i\right\|^2},\end{aligned}$$

given in (5.37). In order to prevent this, the estimator is truncated at 0, i.e. we rather use

$$\left(\widehat{\lambda_1^2}\right)^+ = \max\left\{0; \widehat{\lambda_1^2}\right\}. \tag{5.41}$$

Although $\left(\widehat{\lambda_1^2}\right)^+$ is a biased estimator of λ_1^2, its mean squared error will be smaller than the mean squared error of the unbiased estimator $\widehat{\lambda_1^2}$ of λ_1^2. Saxena and Alam (1982) also suggest this type of truncated estimator when estimating the non-centrality parameter of the non-central chi-squared distribution. Shao and Strawderman (1995) propose further minor modifications to this type of estimator. Venter and Steel (1990) show that in estimating the non-centrality parameter, its truncated form and also various empirical Bayes estimators perform much better than the unbiased estimator.

In a simulation study reported in Section 5.6 of this chapter, we illustrate how the p-value in (5.9), can be utilised for identifying selection influential data cases. In this simulation study we estimate the corresponding non-centrality parameter of the chi-squared distribution of V_{ij} in (5.6), by the truncated estimator in (5.41).

5.4.2 The case where σ^2 is unknown

We now turn to estimation of the non-centrality parameter for the case were σ^2 is unknown. For this case, as pointed out earlier, no progress is made in deriving p-values based on the distribution of $C_p(\mathbf{Y}, L, i)$ or the distribution of $\frac{C_p(\mathbf{Y},L,i)}{\widehat{\sigma}^2}$, due to the dependence of $\langle \mathbf{Y}, P_{L^\perp}\mathbf{u}_i\rangle^2$ and $\widehat{\sigma}^2$. A p-value based on the distribution of $\widetilde{C}_p(\mathbf{Y}, L, i)$ was also presented, but its usefulness for identifying selection influential data cases is hampered by the fact that σ^2 need to be estimated in the lower bound of one of the integrals in (5.28). This leaves us with the p-value based on the distribution of $\frac{\widetilde{C}_p(\mathbf{Y},L,i)}{\widehat{\sigma}^2}$, which was derived in (5.34), for the identification of selection influential data cases. A modified version, where $\frac{1}{\sigma^2}$ is unbiasedly estimated, appears in (5.36). Computation of either (5.34) or (5.36) entails integration of the probability density function of the random variable $\widetilde{V}_{ij}$ which follows a non-central chi-squared distribution with 1 degree of freedom and non-centrality parameter

$$\lambda_2^2 = \frac{\gamma}{\sigma^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2}.$$

If γ and $\frac{1}{\sigma^2}$ in the above non-centrality parameter are unbiasedly estimated by respectively $\widehat{\gamma}_2$

in (2.23) of Chapter 2 and $\left(\frac{n-(m+1)-2}{n-(m+1)}\right)\frac{1}{\widehat{\sigma}^2}$ in (5.35), we obtain an unbiased estimator of λ_2^2. Note that this estimator can also be truncated at zero, thereby eliminating the unsatisfactory situation which may arise whereby a quantity which we know to be non-negative is estimated to be negative. In Section 5.5 that follows we identify selection influential data cases by applying the p-value approach to the same example data sets introduced earlier.

5.5 Illustrative examples

In practical regression samples the value of the error variance, σ^2, is unknown. If our objective is to utilise a p-value approach to identify selection influential data cases, we have to consider the p-values in (5.34) and (5.36) for this purpose. Making, for example, use of the p-value in (5.34), we have to determine a probability of the form

$$P(S_{ij} > s_{ij}) = P\left(\frac{\langle \mathbf{Y}, P_{M|L}\mathbf{u}_i\rangle^2}{\widehat{\sigma}^2} > \frac{\langle \mathbf{y}, P_{M|L}\mathbf{u}_i\rangle^2}{s^2}\right)$$

for each of the data cases and for every linear subspace L of M. In obtaining these p-values we have to estimate, for each data case and every linear subspace, the corresponding non-centrality parameter

$$\lambda_{2ij}^2 = \frac{\gamma(\boldsymbol{\mu}, L, i)}{\sigma^2 \left\|P_{M|L}\mathbf{u}_i\right\|^2}$$

of the non-central chi-squared random variable $\widetilde{V}_{ij}$ in (5.12). In Section 5.4.2 we showed how λ_{2ij}^2 can be estimated unbiasedly. Note however that the unbiased estimator of λ_{2ij}^2 is not computable when the linear subspace L equals M, since then

$$\left\|P_{M|L}\mathbf{u}_i\right\|^2 = \left\|P_M\mathbf{u}_i\right\|^2 - \left\|P_L\mathbf{u}_i\right\|^2 = 0.$$

Consequently, we cannot use this approach to determine whether a data case is selection influential if the regression model is fitted to the complete set of predictor variables. A more serious problem also arose when we tried to use (5.34) or (5.36) to identify selection influential data points in example data sets. We found that the majority of p-values obtained for other subspaces than M, equalled either 0 or 1, thereby implying that the majority of cases in the corresponding data set are selection influential. A possible reason for these inconclusive results is poor estimation of the corresponding λ_{2ij}^2 required to calculate the p-values for these cases in the particular subspaces of M. Recall from Section 5.4.1 that, for the case where σ^2 is known, the variance of the estimator $\widehat{\lambda}_{2ij}^2$ of λ_{2ij}^2 tends to be large due to the presence of $\left\|P_{M|L}\mathbf{u}_i\right\|^2$ in the denominator of the estimator. This also applies if we consider estimation of λ_{2ij}^2 for the case where σ^2 is unknown. The p-value in (5.34) is therefore often unsuccessful in identifying selection influential data cases. The same applies for the p-value in (5.36).

How should we then proceed to identify selection influential cases in a data set where σ^2 in unknown? A fairly naive approach, that seems to give good results in the examples that were investigated, is to assume that the estimate of σ^2, obtained from the regression sample, is in fact the known value of the parameter. In particular, this procedure entails the following:

- For a data set under consideration, σ^2 is merely estimated by its unbiased estimator in (1.10). The obtained estimated value is now assumed to be the known value of σ^2.
- The p-value in (5.9) is utilised for identifying selection influential cases in the data set.

Note that the non-centrality parameter associated with the p-value in (5.9) is $\widehat{\lambda}_1^2$, which is estimated by the truncated estimator in (5.41), as if σ^2 is known. Estimation of a non-centrality parameter that includes the term $\left\|P_{M|L}\mathbf{u}_i\right\|^2$ in the denominator, is thus specifically avoided.

We found that the method described above produces satisfactory results if applied to practical data sets. In the next section we apply this method to the three example data sets already introduced in Sections 3.2 and 4.3.

5.5.1 The Hald data

Consider again the Hald data which include 4 predictors and 13 observations. The full model variance estimate is $\widehat{\sigma}^2 = 5.983$. This estimate is now assumed to be the known value of σ^2. We now utilise the p-value in (5.9) in an attempt to identify selection influential data cases. For each of the 13 observations, we calculate (5.9) for all subspaces L in $M = span\{\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$. There are $2^4 - 1 = 15$ such linear subspaces containing at least 1 predictor variable. Each of these subspaces L, with $\dim(L) = l + 1$, is spanned by the vector $\mathbf{1}$ and a corresponding subset of l predictors, i.e. $L = span\{\mathbf{1}, \mathbf{x}_j : j \in J_L\} \subset M$, where $J_L \subset \{1, 2, 3, 4\}$. The corresponding non-centrality parameter needed to obtain these p-values, is estimated by (5.41).

These p-values for all linear subspaces L of M (thus 15 for each of the 13 data cases) are shown in Table 5.1. Program C6 in Appendix C was utilised for calculating the p-values and average p-values of the Hald data.

How should we now proceed in order to identify selection influential data cases? Consider the p-value calculated for the ith data case and the linear subspace L, viz.

$$P(\widehat{\gamma}_1(\mathbf{Y}, L, i) > \widehat{\gamma}_1(\mathbf{y}, L, i)).$$

If a significantly large or small value is obtained for the p-value, the corresponding ith data case is identified selection influential within the linear subspace L. In order to determine if a case is selection influential in all linear subspaces containing a given jth predictor variable as a basis vector, we calculate the average of all p-values for those subspaces that include predictor $\mathbf{x}_j$ as a basis vector. The ith data case is considered selection influential if any of its m average p-values is significantly large or small.

The 4 average p-values, for each of the 13 data cases of the Hald data, are also shown in Table 5.1. We observe that the average p-values for data case 6 are significantly smaller than the average p-values obtained for other data cases. At a 10% significance level, data case 6 will be regarded as selection influential. No other data cases are deemed selection influential at this significance level. Note, as a matter of interest, that data case 6 was also identified as selection influential by the new influence measure proposed in Chapter 4.

	Observation						
Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	0.486	0.746	0.718	0.222	0.569	0.044	0.731
2	0.792	0.748	0.718	0.776	0.505	0.045	0.241
3	0.492	0.737	0.731	0.221	0.528	0.045	0.731
4	0.475	0.750	0.732	0.772	0.417	0.045	0.241
1,2	0.491	0.618	0.509	0.436	0.589	0.079	0.505
1,3	0.490	0.752	0.720	0.221	0.445	0.044	0.732
1,4	0.461	0.418	0.303	0.257	0.414	0.101	0.604
2,3	0.482	0.747	0.773	0.830	0.531	0.044	0.229
2,4	0.490	0.751	0.723	0.776	0.745	0.045	0.241
3,4	0.475	0.775	0.204	0.992	0.519	0.053	0.228
1,2,3	0.936	0.434	0.560	0.449	0.930	0.084	0.450
1,2,4	0.971	0.491	0.406	0.397	0.896	0.090	0.464
1,3,4	0.786	0.581	0.216	0.313	0.793	0.100	0.502
2,3,4	0.613	0.961	0.176	0.240	0.860	0.115	0.465
1,2,3,4	0.998	0.449	0.294	0.400	0.898	0.086	0.457
Variable							
1	0.702	0.561	0.466	0.337	0.692	0.079	0.555
2	0.722	0.650	0.520	0.538	0.744	0.074	0.381
3	0.659	0.680	0.459	0.458	0.688	0.072	0.474
4	0.659	0.647	0.382	0.518	0.693	0.079	0.400

	Observation					
Model	**8**	**9**	**10**	**11**	**12**	**13**
1	0.071	0.261	0.541	0.852	0.330	0.832
2	0.069	0.712	0.437	0.788	0.318	0.438
3	0.059	0.255	0.423	0.214	0.332	0.834
4	0.076	0.714	0.434	0.790	0.319	0.890
1,2	0.341	0.410	0.429	0.147	0.694	0.182
1,3	0.058	0.259	0.569	0.842	0.330	0.834
1,4	0.054	0.774	0.532	0.960	0.438	0.782
2,3	0.993	0.909	0.428	0.150	0.389	0.133
2,4	0.062	0.716	0.434	0.812	0.318	0.890
3,4	0.913	0.837	0.415	0.147	0.493	0.138
1,2,3	0.090	0.485	0.874	0.329	0.700	0.257
1,2,4	0.118	0.556	0.806	0.298	0.629	0.291
1,3,4	0.125	0.704	0.726	0.222	0.511	0.401
2,3,4	0.318	0.898	0.393	0.147	0.358	0.615
1,2,3,4	0.091	0.502	0.833	0.283	0.643	0.261
Variable						
1	0.119	0.494	0.664	0.491	0.534	0.480
2	0.260	0.649	0.579	0.369	0.506	0.383
3	0.331	0.606	0.583	0.292	0.469	0.434
4	0.220	0.713	0.572	0.457	0.464	0.533

Table 5.1: Individual p-values and average p-values calculated for the 13 data cases of the Hald data

5.5.2 The fuel data

The p-value in (5.9) was also calculated for the fuel data with 4 predictors and 50 observations. We assume that the full model error variance estimate $\widehat{\sigma}^2 = 7452.009$ is the known value of σ^2. The different subspaces L for which (5.9) needs to be calculated, for each of the 50 data cases, are the same as for the Hald data. For each of the 50 observations the average of the corresponding p-values of the subspaces including x_j, where $j = 1, 2, 3, 4$ are also calculated. Note that 8 of these subspaces include predictor variable x_j. The calculated p-values in (5.9) and the 4 average p-values for each of the 50 data cases are shown in Table 5.2.

The average p-values obtained for cases 40 and 50 are extremely small. These cases will therefore be considered as selection influential. Other average p-values which are also relatively small or large are highlighted. These are for cases 18, 19, 29 (large average p-values for x_2, x_3 and x_4) and 45. Note that case 50 (which is here one of the subset of selection influential data cases) was also identified as selection influential by the influence measure proposed in (4.5).

Recall from a previous analysis of the fuel data in Section 4.3.2 that predictors 2 and 3 were selected if C_p selection is applied to the full data set. Omitting cases 40 and 50 simultaneously before C_p selection is applied to thc reduced data, which now only consist of 48 cases, results in predictors 1,2 and 3 being selected. In order to obtain an indication of whether it would be beneficial in the sense of improved response prediction to leave out cases 40 and 50 simultaneously, we calculate an estimate of the expected squared prediction error for the reduced data set of 48 cases. This is done in a similar way as in Section 4.3.2, where estimates of the expected squared prediction error were calculated for the full data set and for each of the reduced data sets, where only one observation was omitted at a time. For the reduced data set with cases 40 and 50 omitted, we randomly select 38 cases to form the training data set. The remaining 10 observations constitute the test data set. Recall from Section 4.3.2 that the selected model obtained by applying C_p to the training data set is used to calculate the average squared prediction error for the test data set. In 20000 such repetitions (i.e. each time randomly select a training data set and calculate the average squared prediction error for the test data set), we obtain an estimate of 4457.881 for the expected squared prediction error. This estimate is significantly smaller than the estimate obtained for the full data set (i.e. 9850.24). This estimate is also smaller than the respective estimates of the expected squared prediction error of the reduced data sets when case 40 and case 50 are individually omitted. As presented in the last column of Table 4.2, for case 40 the estimate was 7977.2, and for case 50 it was 6012.5. Based on these estimates of the respective average squared prediction error, we therefore recommend that cases 40 and 50 be omitted from the fuel data set, and that the selected model, obtained by applying C_p to the corresponding reduced data set, be utilised for prediction of future data cases.

	Observation									
Model	1	2	3	4	5	6	7	8	9	10
1	0.8553	0.9867	0.7490	0.2306	0.1126	0.6621	0.3405	0.4461	0.3126	0.3249
2	0.8125	0.5548	0.7910	0.2538	0.1299	0.4173	0.3512	0.4549	0.4719	0.5306
3	0.5235	0.4989	0.6956	0.2400	0.1095	0.5279	0.3363	0.5413	0.3049	0.4778
4	0.7956	0.6597	0.9916	0.2314	0.1126	0.4228	0.3433	0.4428	0.2778	0.3662
1,2	0.7072	0.6938	0.9588	0.2480	0.1346	0.4475	0.3461	0.4615	0.4848	0.4623
1,3	0.7829	0.9096	0.8837	0.2383	0.1188	0.7327	0.3341	0.6295	0.3456	0.3652
1,4	0.8766	0.9554	0.7642	0.2308	0.1089	0.6511	0.3407	0.4428	0.3299	0.3345
2,3	0.7718	0.4637	0.5470	0.4236	0.1527	0.6007	0.8891	0.7556	0.5861	0.6749
2,4	0.7071	0.6836	0.9261	0.2786	0.1772	0.4246	0.3382	0.4851	0.3981	0.4464
3,4	0.5515	0.5721	0.7765	0.2469	0.1176	0.5829	0.3400	0.5825	0.2905	0.4294
1,2,3	0.9023	0.5877	0.6901	0.3922	0.1627	0.8746	0.8496	0.7902	0.6013	0.5931
1,2,4	0.6755	0.7435	0.9964	0.2661	0.1673	0.4478	0.3396	0.4823	0.4196	0.4304
1,3,4	0.7572	0.8693	0.9131	0.2360	0.1122	0.7402	0.3385	0.5982	0.3752	0.3831
2,3,4	0.8997	0.5750	0.6595	0.5111	0.2201	0.7358	0.6509	0.8748	0.4928	0.5736
1,2,3,4	0.9514	0.6307	0.7254	0.4612	0.2048	0.8765	0.6840	0.8656	0.5229	0.5499
Variable										
1	0.8135	0.7971	0.8351	0.2879	0.1403	0.6791	0.4466	0.5895	0.4240	0.4304
2	0.8034	0.6166	0.7868	0.3543	0.1687	0.6031	0.5561	0.6463	0.4972	0.5327
3	0.7675	0.6384	0.7364	0.3437	0.1498	0.7089	0.5528	0.7047	0.4399	0.5059
4	0.7768	0.7112	0.8441	0.3078	0.1526	0.6102	0.4219	0.5968	0.3883	0.4392

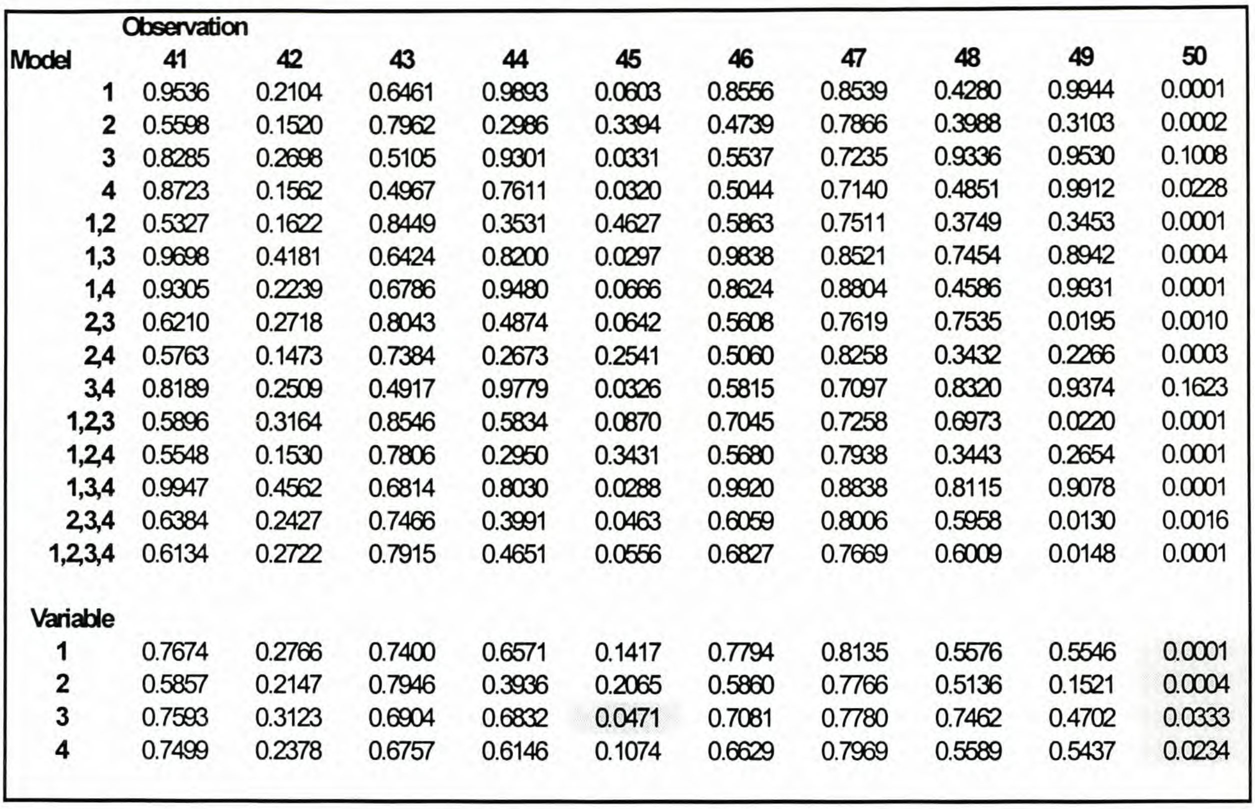
	Observation									
Model	11	12	13	14	15	16	17	18	19	20
1	0.8282	0.4046	0.4813	0.3631	0.7728	0.6167	0.9009	0.1531	0.0529	0.7083
2	0.5308	0.5594	0.5602	0.6781	0.5590	0.6024	0.7422	0.0565	0.1472	0.5054
3	0.8836	0.5301	0.8434	0.4387	0.9813	0.4762	0.7502	0.1992	0.0540	0.7132
4	0.9472	0.4140	0.5684	0.4434	0.8991	0.5576	0.7836	0.0915	0.0540	0.7173
1,2	0.5156	0.5261	0.5183	0.5940	0.5191	0.6565	0.8069	0.0608	0.1319	0.6542
1,3	0.7611	0.5199	0.6646	0.3478	0.7947	0.6044	0.9471	0.3632	0.0531	0.7082
1,4	0.8132	0.4289	0.4855	0.3647	0.8005	0.5704	0.8680	0.1616	0.0529	0.7094
2,3	0.4464	0.8817	0.8607	0.6534	0.5667	0.5898	0.7834	0.0970	0.0872	0.4761
2,4	0.5422	0.4382	0.5264	0.6421	0.4700	0.7638	0.8459	0.0556	0.1692	0.4739
3,4	0.8946	0.4564	0.8084	0.4235	0.9276	0.5404	0.8145	0.1896	0.0539	0.7144
1,2,3	0.4328	0.9275	0.7895	0.5692	0.5251	0.6439	0.8536	0.1215	0.0785	0.6244
1,2,4	0.5300	0.4488	0.5105	0.5995	0.4677	0.7577	0.8619	0.0577	0.1529	0.5568
1,3,4	0.7422	0.6254	0.6756	0.3493	0.8278	0.5521	0.9078	0.3904	0.0531	0.7094
2,3,4	0.4560	0.7862	0.8054	0.6189	0.4770	0.7481	0.8912	0.0895	0.0978	0.4458
1,2,3,4	0.4446	0.8410	0.7751	0.5743	0.4745	0.7417	0.9085	0.1029	0.0881	0.5306
Variable										
1	0.6335	0.5903	0.6126	0.4702	0.6478	0.6429	0.8818	0.1764	0.0829	0.6502
2	0.4873	0.6761	0.6683	0.6162	0.5074	0.6880	0.8367	0.0802	0.1191	0.5334
3	0.6327	0.6960	0.7778	0.4969	0.6968	0.6121	0.8570	0.1942	0.0707	0.6153
4	0.6713	0.5548	0.6444	0.5019	0.6680	0.6540	0.8602	0.1424	0.0902	0.6072

Table 5.2: Individual p-values and average p-values calculated for data cases 1 to 20 of the fuel data

	Observation									
Model	21	22	23	24	25	26	27	28	29	30
1	0.6545	0.8241	0.6478	0.8083	0.3557	0.6642	0.8474	0.5723	0.8608	0.9195
2	0.7200	0.5251	0.7119	0.6744	0.3268	0.7841	0.7664	0.6062	0.9082	0.5664
3	0.6075	0.9296	0.6270	0.8354	0.3134	0.6899	0.6372	0.6871	0.9773	0.6098
4	0.6264	0.8165	0.6092	0.8493	0.2936	0.9634	0.9509	0.5982	0.9909	0.7260
1,2	0.7115	0.5482	0.7824	0.5567	0.3623	0.6722	0.7482	0.6108	0.8688	0.5049
1,3	0.6234	0.7949	0.8088	0.7979	0.4532	0.9003	0.7268	0.7276	0.9016	0.7706
1,4	0.6491	0.7622	0.6449	0.7904	0.3463	0.6527	0.8415	0.5510	0.8456	0.9022
2,3	0.8545	0.7873	0.9361	0.6497	0.4236	0.8828	0.7639	0.8295	0.9593	0.9068
2,4	0.6280	0.5963	0.7786	0.6424	0.3686	0.7597	0.7638	0.6979	0.9272	0.5417
3,4	0.6117	0.8237	0.6392	0.8442	0.3337	0.7016	0.6384	0.7552	0.9648	0.6124
1,2,3	0.8457	0.8551	0.8052	0.5309	0.4895	0.9676	0.7925	0.8381	0.9182	0.7917
1,2,4	0.6441	0.5961	0.8109	0.5765	0.3834	0.6963	0.7533	0.6768	0.8981	0.5111
1,3,4	0.6209	0.8623	0.7972	0.7764	0.4352	0.8799	0.7330	0.6803	0.8830	0.7834
2,3,4	0.7514	0.9548	0.8163	0.6183	0.4977	0.9112	0.7681	0.9615	0.9783	0.8661
1,2,3,4	0.7717	0.9551	0.7652	0.5495	0.5228	0.9979	0.7850	0.9320	0.9474	0.8036
Variable										
1	0.6901	0.7747	0.7578	0.6733	0.4186	0.8039	0.7785	0.6986	0.8904	0.7484
2	0.7409	0.7272	0.8008	0.5998	0.4218	0.8340	0.7677	0.7691	0.9257	0.6865
3	0.7109	0.8703	0.7744	0.7003	0.4336	0.8664	0.7306	0.8014	0.9412	0.7680
4	0.6629	0.7959	0.7327	0.7059	0.3976	0.8203	0.7792	0.7316	0.9294	0.7183

	Observation									
Model	31	32	33	34	35	36	37	38	39	40
1	0.8078	0.6519	0.9180	0.5522	0.4379	0.7800	0.8078	0.2198	0.7714	0.0005
2	0.5552	0.6817	0.9736	0.4315	0.9131	0.9663	0.4886	0.2157	0.7229	0.0007
3	0.7060	0.4620	0.4710	0.9462	0.3596	0.7535	0.5407	0.2200	0.8115	0.0005
4	0.9591	0.8094	0.9855	0.5119	0.4121	0.7194	0.6479	0.1917	0.7716	0.0005
1,2	0.6129	0.7359	0.9940	0.4395	0.9239	0.9612	0.6324	0.2313	0.8962	0.0007
1,3	0.5989	0.4054	0.5642	0.9905	0.3636	0.9335	0.6486	0.2979	0.7810	0.0005
1,4	0.8157	0.6559	0.9100	0.5665	0.4300	0.7752	0.9383	0.2194	0.7717	0.0005
2,3	0.8227	0.8732	0.2782	0.8021	0.6717	0.6953	0.5358	0.3104	0.3147	0.0008
2,4	0.5771	0.6958	0.9441	0.4205	0.8458	0.9472	0.7464	0.2247	0.7901	0.0006
3,4	0.6851	0.4526	0.4486	0.9146	0.3614	0.7708	0.7377	0.2254	0.8025	0.0005
1,2,3	0.9353	0.7531	0.3073	0.8340	0.6588	0.6137	0.7723	0.3488	0.4230	0.0009
1,2,4	0.6083	0.7262	0.9735	0.4273	0.8673	0.9218	0.8297	0.2326	0.8796	0.0006
1,3,4	0.6020	0.4058	0.5908	0.9627	0.3627	0.9218	0.7855	0.2976	0.7823	0.0005
2,3,4	0.8681	0.8389	0.2545	0.7517	0.7502	0.6032	0.9014	0.3328	0.3550	0.0007
1,2,3,4	0.9273	0.7712	0.2760	0.7839	0.7202	0.5770	0.9983	0.3514	0.4101	0.0007
Variable										
1	0.7385	0.6382	0.6917	0.6946	0.5956	0.8105	0.8016	0.2748	0.7144	0.0006
2	0.7384	0.7595	0.6251	0.6113	0.7939	0.7857	0.7381	0.2809	0.5990	0.0007
3	0.7682	0.6203	0.3988	0.8732	0.5310	0.7336	0.7400	0.2980	0.5850	0.0006
4	0.7553	0.6695	0.6729	0.6674	0.5937	0.7795	0.8231	0.2594	0.6954	0.0006

Table 5.2 *(continued)*: Individual p-values and average p-values calculated for data cases 21 to 40 of the fuel data



	Observation									
Model	41	42	43	44	45	46	47	48	49	50
1	0.9536	0.2104	0.6461	0.9893	0.0603	0.8556	0.8539	0.4280	0.9944	0.0001
2	0.5598	0.1520	0.7962	0.2986	0.3394	0.4739	0.7866	0.3988	0.3103	0.0002
3	0.8285	0.2698	0.5105	0.9301	0.0331	0.5537	0.7235	0.9336	0.9530	0.1008
4	0.8723	0.1562	0.4967	0.7611	0.0320	0.5044	0.7140	0.4851	0.9912	0.0228
1,2	0.5327	0.1622	0.8449	0.3531	0.4627	0.5863	0.7511	0.3749	0.3453	0.0001
1,3	0.9698	0.4181	0.6424	0.8200	0.0297	0.9838	0.8521	0.7454	0.8942	0.0004
1,4	0.9305	0.2239	0.6786	0.9480	0.0666	0.8624	0.8804	0.4586	0.9931	0.0001
2,3	0.6210	0.2718	0.8043	0.4874	0.0642	0.5608	0.7619	0.7535	0.0195	0.0010
2,4	0.5763	0.1473	0.7384	0.2673	0.2541	0.5060	0.8258	0.3432	0.2266	0.0003
3,4	0.8189	0.2509	0.4917	0.9779	0.0326	0.5815	0.7097	0.8320	0.9374	0.1623
1,2,3	0.5896	0.3164	0.8546	0.5834	0.0870	0.7045	0.7258	0.6973	0.0220	0.0001
1,2,4	0.5548	0.1530	0.7806	0.2950	0.3431	0.5680	0.7938	0.3443	0.2654	0.0001
1,3,4	0.9947	0.4562	0.6814	0.8030	0.0288	0.9920	0.8838	0.8115	0.9078	0.0001
2,3,4	0.6384	0.2427	0.7466	0.3991	0.0463	0.6059	0.8006	0.5958	0.0130	0.0016
1,2,3,4	0.6134	0.2722	0.7915	0.4651	0.0556	0.6827	0.7669	0.6009	0.0148	0.0001
Variable										
1	0.7674	0.2766	0.7400	0.6571	0.1417	0.7794	0.8135	0.5576	0.5546	0.0001
2	0.5857	0.2147	0.7946	0.3936	0.2065	0.5860	0.7766	0.5136	0.1521	0.0004
3	0.7593	0.3123	0.6904	0.6832	0.0471	0.7081	0.7780	0.7462	0.4702	0.0333
4	0.7499	0.2378	0.6757	0.6146	0.1074	0.6629	0.7969	0.5589	0.5437	0.0234

Table 5.2 *(continued)*: Individual p-values and average p-values calculated for data cases 41 to 50 of the fuel data

5.5.3 The evaporation data

Finally, consider the evaporation data with 10 predictors and 46 observations. The full model error variance estimate $\widehat{\sigma}^2 = 42.351$ is assumed to be the known value of σ^2 if the p-value in (5.9) is calculated for each of the 46 cases and $2^{10} - 1 = 1023$ linear subspaces spanned by the vector $\mathbf{1}$ and at least one predictor variable. As for the Hald and fuel data, the non-centrality parameter, associated with the p-value in (5.9), is estimated by the truncated estimator in (5.41). For each observation the averages of all those p-values for those subspaces that include $\mathbf{x}_j$, where $j = 1, 2, ..., 10$, as basis vector, are also calculated. The 10 average p-values for each of the 46 cases are shown in Table 5.3.

We observe that the average p-values calculated for cases 33 and 41 are extremely small. According to the proposed p-value approach, these two cases will certainly be deemed selection influential. Other average p-values which are relatively small are also highlighted in Table 5.3. Such cases are 2, 8 and 24 which may also be regarded as selection influential if a higher significance level is considered.

We also consider estimation of the expected squared prediction error for the reduced data set where observations 33 and 41 are omitted simultaneously. The estimation is done in a similar way as in Section 4.3.3, but now dividing the 44 observations into 35 training observations and 9 test observations in each of the 20000 repetitions. The resulting estimated expected squared prediction error for this reduced data set is 46.189, which is relatively smaller than the estimated expected squared prediction error for the full data set (i.e. 71.78), and also for the respective estimated values of the respective reduced data sets when observations 33 and 41 are omitted individually. From Table 4.3 these estimates equal 61.9 if case 33 is omitted and 55.6 if case 41 is omitted. We suggest that the selected model (which includes predictors 1, 3, 6, 9 and 10) obtained if C_p is applied to the reduced data set (i.e. where both cases 33 and 41 are omitted) be utilised for prediction of future observations.

	Observation									
Variable	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0.6791	0.3887	0.3766	0.7534	0.4017	0.5609	0.8187	0.0709	0.6776	0.7542
2	0.6629	0.3472	0.4398	0.7302	0.5546	0.5301	0.7899	0.0963	0.6828	0.7501
3	0.6483	0.3595	0.4595	0.7408	0.4484	0.5281	0.7816	0.1078	0.6497	0.7294
4	0.6470	0.3839	0.4344	0.6967	0.4794	0.4867	0.7948	0.1499	0.6558	0.7536
5	0.6236	0.3796	0.4597	0.7278	0.5020	0.5237	0.8006	0.1096	0.6289	0.7633
6	0.6280	0.3777	0.3756	0.6803	0.4594	0.4760	0.8054	0.1129	0.7064	0.7374
7	0.6320	0.3739	0.4523	0.7332	0.4984	0.5410	0.7821	0.1306	0.6402	0.7536
8	0.6465	0.2500	0.4936	0.7432	0.5237	0.4747	0.8049	0.0384	0.6498	0.7564
9	0.6147	0.0464	0.5809	0.8218	0.6851	0.3964	0.8024	0.0256	0.6160	0.7617
10	0.6929	0.3756	0.5146	0.7387	0.5206	0.5937	0.7982	0.1218	0.6841	0.7489

	Observation									
Variable	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	0.7479	0.6113	0.7519	0.6398	0.7429	0.5637	0.8016	0.7238	0.6357	0.7405
2	0.7497	0.6375	0.7447	0.6153	0.7440	0.5582	0.7939	0.6985	0.6071	0.7532
3	0.7549	0.6400	0.7648	0.6588	0.7467	0.5622	0.7759	0.6850	0.6285	0.7397
4	0.7525	0.6476	0.7654	0.6330	0.7427	0.5755	0.7912	0.7126	0.6105	0.7578
5	0.7545	0.6579	0.7470	0.6392	0.7537	0.5523	0.8077	0.6965	0.5737	0.7333
6	0.7704	0.6391	0.7692	0.6649	0.7680	0.5791	0.7844	0.7034	0.7236	0.7782
7	0.7509	0.6461	0.7422	0.6286	0.7382	0.5487	0.7825	0.7071	0.5947	0.7615
8	0.7521	0.6633	0.7530	0.6334	0.7641	0.5605	0.8045	0.7079	0.5963	0.7575
9	0.7976	0.7700	0.7558	0.6718	0.8809	0.5951	0.8430	0.7162	0.5645	0.7837
10	0.7576	0.7249	0.8186	0.6652	0.7612	0.6119	0.7981	0.7647	0.6686	0.7559

	Observation									
Variable	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
1	0.4725	0.2064	0.7307	0.1672	0.8003	0.6048	0.5813	0.8118	0.6995	0.7123
2	0.3700	0.2110	0.6817	0.1540	0.7120	0.6431	0.5597	0.8094	0.6000	0.7144
3	0.4504	0.1801	0.6945	0.1361	0.7378	0.6498	0.5407	0.8138	0.6248	0.7136
4	0.4542	0.1702	0.6596	0.1055	0.6970	0.6762	0.5470	0.8076	0.5898	0.7168
5	0.4765	0.1668	0.6918	0.1249	0.7241	0.6340	0.5366	0.7753	0.6058	0.7170
6	0.3395	0.1809	0.6198	0.0985	0.7430	0.6671	0.6178	0.8204	0.6186	0.7202
7	0.4956	0.1703	0.7021	0.1251	0.7187	0.6190	0.5415	0.8131	0.6302	0.7313
8	0.4896	0.1502	0.7035	0.1097	0.7104	0.6158	0.4616	0.8101	0.6002	0.7122
9	0.4227	0.1342	0.7467	0.1058	0.7079	0.4208	0.3914	0.8345	0.6520	0.7812
10	0.5811	0.1359	0.7648	0.1400	0.7444	0.6675	0.5058	0.8174	0.4979	0.7378

	Observation									
Variable	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
1	0.4985	0.1968	0.0074	0.7808	0.7893	0.7874	0.4241	0.5108	0.4379	0.1760
2	0.4856	0.2872	0.0072	0.7625	0.7865	0.7730	0.3287	0.4364	0.4409	0.1581
3	0.4690	0.2724	0.0078	0.7469	0.7879	0.7791	0.3904	0.4812	0.4487	0.1765
4	0.3071	0.2788	0.0074	0.7789	0.8093	0.7810	0.4323	0.4723	0.4488	0.1598
5	0.4648	0.3163	0.0073	0.7589	0.7842	0.7860	0.3868	0.4770	0.5041	0.1590
6	0.5741	0.4072	0.0069	0.7465	0.8068	0.7912	0.3283	0.4280	0.4534	0.1603
7	0.4387	0.3176	0.0069	0.7921	0.7803	0.7877	0.3872	0.4731	0.4856	0.1603
8	0.4800	0.3184	0.0073	0.7717	0.7755	0.7597	0.4286	0.5600	0.5215	0.1673
9	0.4070	0.4528	0.0075	0.7957	0.8144	0.8127	0.1976	0.4348	0.6471	0.1727
10	0.4710	0.2268	0.0062	0.7496	0.7831	0.8060	0.4096	0.5239	0.5611	0.1502

Table 5.3: Average p-values calculated for data cases 1 to 40 of the evaporation data

	Observation					
Variable	41	42	43	44	45	46
1	0.0063	0.7173	0.8355	0.8385	0.7358	0.5730
2	0.0078	0.7469	0.8244	0.8380	0.6555	0.4747
3	0.0077	0.6524	0.8492	0.8369	0.7059	0.5087
4	0.0082	0.7528	0.8436	0.8584	0.6894	0.5039
5	0.0069	0.7246	0.8337	0.8400	0.6810	0.5052
6	0.0060	0.6576	0.8571	0.8560	0.7580	0.4918
7	0.0089	0.7266	0.8481	0.8103	0.7084	0.5332
8	0.0091	0.7247	0.8507	0.8469	0.6858	0.5072
9	0.0098	0.7655	0.8764	0.8667	0.7313	0.5350
10	0.0099	0.7521	0.8318	0.8287	0.6099	0.4453

Table 5.3 *(continued)*: Average p-values calculated for data cases 41 to 46 of the evaporation data

5.6 Simulation study

The simulation study described in Section 3.3 of Chapter 3 illustrates the effect that inclusion of a possibly selection influential data case into a data set may have on the properties of the resulting multiple linear regression model. This effect was studied in terms of the APE (average prediction error) and the PCS (probability of correct selection) of the model selected by means of the C_p criterion. The APE and the PCS of the model selected and fitted on an ordinary data set (i.e., a data set not containing a deliberately inserted possibly selection influential data case) were compared to the respective same quantities for a model selected and fitted on a modified data set (i.e., a data set into which a possibly selection influential data case had been inserted). It became clear that the presence of possibly selection influential cases in a data set has negative consequences: the APE tends to increase, and the PCS tends to decrease. The impact that factors such as the sample size, the number of predictor variables, and the correlation amongst the predictor variables have on the effect of a possibly selection influential data case, was also investigated.

A *possibly* selection influential data case included in a data set is deemed to be *definitely* selection influential if this data case complies with the definition of a selection influential data case. This would be the case if its omission from the modified data set leads to a different set of variables being selected, or, given that the same set of predictors is selected, if there is a significant change in the fit of the selected model. Note that the deliberately inserted possibly selection influential data case is obviously not the only case in the modified data set that may turn out to be selection influential. Any other individual "ordinary" data case may of course also turn out to be selection influential. Furthermore, any subset of data cases that may or may not include the possibly selection influential case, may turn out to be selection influential. This is of course also true for the ordinary data set. The fact that the ordinary data set does not include a deliberately inserted possibly selection influential case, does not guarantee that no individual or groups of selection influential data cases will be found in this set. It should also be noted that a procedure designed to identify selection influential data cases is subject to two kinds of error: it may wrongly identify a non-selection influential case as being influential, or it may wrongly classify a case that is in fact selection influential, as being not so.

The simulation study described in this section was undertaken to investigate the performance of a procedure for identifying selection influential data cases based on the p-value defined in (5.9). More specifically, four sets of APE values and PCS values were generated and compared. The first two sets are simply the values obtained in the simulation study described in Chapter 3. The third and fourth sets were generated as follows. Consider a typical ordinary data set, and its modified version containing a possibly selection influential data case, as simulated in Chapter 3. The p-value approach, based on the p-value defined in (5.9), is applied to these data sets in order to identify selection influential cases. We discuss below the cut-off point for the calculated p-values that was used to decide whether a given case is selection influential or not. Note that this p-value is based on the assumption that the value of the error variance, σ^2, is known. We comment in more detail on this assumption below. The cases in the modified data set identified as selection influential are omitted, and similarly for the ordinary data set. Variable selection using the C_p criterion is now applied to the modified and to the ordinary data sets after omission of the cases that were identified as selection influential. The APE and the PCS of the resulting models are approximated by simulation as already described in Chapter 3.

The four sets of APE and PCS values compared in the simulation study can therefore be summarised as follows.

(a) A set obtained for ordinary data sets, i.e. data sets not containing deliberately inserted possibly selection influential data cases, and without any attempt being made to identify and omit such cases.

(b) A set obtained for modified data sets, i.e. data sets that do contain deliberately inserted possibly selection influential data cases. Once again, no attempt is made to identify and omit such cases.

(c) A set obtained for ordinary data sets, but now we do apply the p-value approach in an attempt to identify and omit selection influential data cases.

(d) A set obtained for modified data sets, but once again we do apply the p-value approach in an attempt to identify and omit selection influential data cases.

The first two sets of APE and PCS values were already compared in Chapter 3, illustrating the effect that the presence of selection influential data cases may have. The third and fourth sets of APE and PCS values enable us to investigate the effect that application of our proposed procedure for dealing with selection influential data cases may have. The third set of values shows the effect if this procedure is applied to a data set that actually does not contain any selection influential cases, and the fourth set shows the effect if the data set does indeed contain such cases.

In order to identify selection influential cases in a data set by applying the p-value in (5.9), we proceed in a similar manner as for the example data sets in Section 5.5. This entails the following:

- Consider the ith data case in the simulated data set of size n. There are m predictor variables, and the vectors $\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m$ form a basis for the $(m+1)$-dimensional linear space M. There are $2^m - 1$ possible linear subspaces of that includes at least 1 predictor variable. Each of these subspaces L of M, with $\dim(L) = l + 1$, is spanned by the vector $\mathbf{1}$ and a subset of l predictor variables, i.e. $L = span\left\{\mathbf{1}, \mathbf{x}_j : j \in J_L\right\} \subset M$, where $J_L \subset \{1, 2, ..., m\}$. The p-value in (5.9) is calculated for each combination of a subspace L of M and a data case i.
- In order to decide if a given data case is selection influential, we calculate m average

p-values. The jth of these average p-values is obtained by calculating the average of all the p-values (for the ith data case) for those subspaces that include predictor $\mathbf{x}_j$ as a basis vector. The ith data case is deemed selection influential if at least one of its m average p-values is significantly large or significantly small.

We reproduce the results of the simulation study conducted in Chapter 3 in Figures 5.1 to 5.8. Recall that these results show the APE and the PCS of the selected models if C_p variable selection is applied to a large number of simulated ordinary and their corresponding modified data sets. As in the simulation study of Chapter 3, several factors that were felt could have an influence on the APE and PCS were varied over several levels. For ease of reference these factors are once again listed below.

- The sample size of the simulated data set. The following sample sizes were used in the study: $n = 20, 50$, and 100.
- The number of predictor variables in the simulated data set. For $n = 20$, we used $m = 5$, and for $n = 50$ and $n = 100$, we used $m = 5$ and $m = 10$.
- The correlation amongst the predictor variables (in this regard we study equi-correlated cases, i.e. cases where the same correlation is assumed to hold for any pair of predictor variables). The common value of the correlation between any two predictors was varied over 0 (the orthogonal regressor case), 0.5 and 0.9.
- The sample data sets are simulated at fixed, predetermined values of the regression coefficients. Two different configurations were used in this regard. In the first case, we set $\beta_1 = \beta_2 = ... = \beta_m = s$, and then increment the common value s from 0 in steps of 0.1 up to 1.5, and thereafter in steps of 0.25 up to 3. In the second regression coefficient configuration we also start by setting $\beta_1 = \beta_2 = ... = \beta_m = 0$, but thereafter only a subset of the β-values are incremented. In particular, for simulated data sets containing $m = 5$ predictors we increment the common value of β_1 and β_2 from 0 in steps of 0.1 up to 1.5, and thereafter in steps of 0.25 up to 3. For the cases where we had $m = 10$ predictors, the values of $\beta_1, \beta_2, \beta_3, \beta_4$ and β_5 are incremented in the same manner.
- It should finally be noted that we consistently used $\beta_0 = 1$ as intercept parameter in the regression models.

Figures 5.1 to 5.4 show the APE of the selected models if C_p variable selection is applied to the ordinary data sets (represented by a solid black line), and their corresponding modified data sets (represented by a dotted black line). The PCS if C_p variable selection is applied to the ordinary data sets (solid black line) and their corresponding modified data sets (dotted black line) is shown in Figures 5.5 to 5.8. Also shown in these figures are the APE and the PCS of the models selected from the ordinary (the solid red lines) and the corresponding modified (the dotted red lines) data sets if the p-value approach based on (5.9) is first used to screen these data sets for selection influential pcints, and any points identified as such are removed before selecting and fitting a model. In this regard it must be pointed out that a data case was deemed selection influential, and omitted from the data set before a model was selected and fitted, if at least one of its m average p-values (described above) was less than 0.01 or larger than 0.99. Other cut-off points (0.025 and 0.975, as well as 0.05 and 0.95) were also used, but the resulting procedures were found to perform worse, and the results are therefore not included here.

Finally, before moving on to a discussion of the results, we comment on the fact that the value of the error variance, σ^2, is assumed known throughout the simulation study. This is a reasonable assumption in multiple linear regression if the degrees of freedom available for estimation of σ^2 is large. In small sample cases this assumption may not always be reasonable. However, as we saw earlier, the procedure that we investigate in our simulation study performs quite well in example data sets (where the value of σ^2 is of course unknown) if we treat the estimated value of σ^2 as if it were the known value of this parameter. Since we do not currently have a fully satisfactory way of otherwise dealing with the case where the value of σ^2 is unknown, we restricted our simulation study to the limited situation of a known value for the error variance.

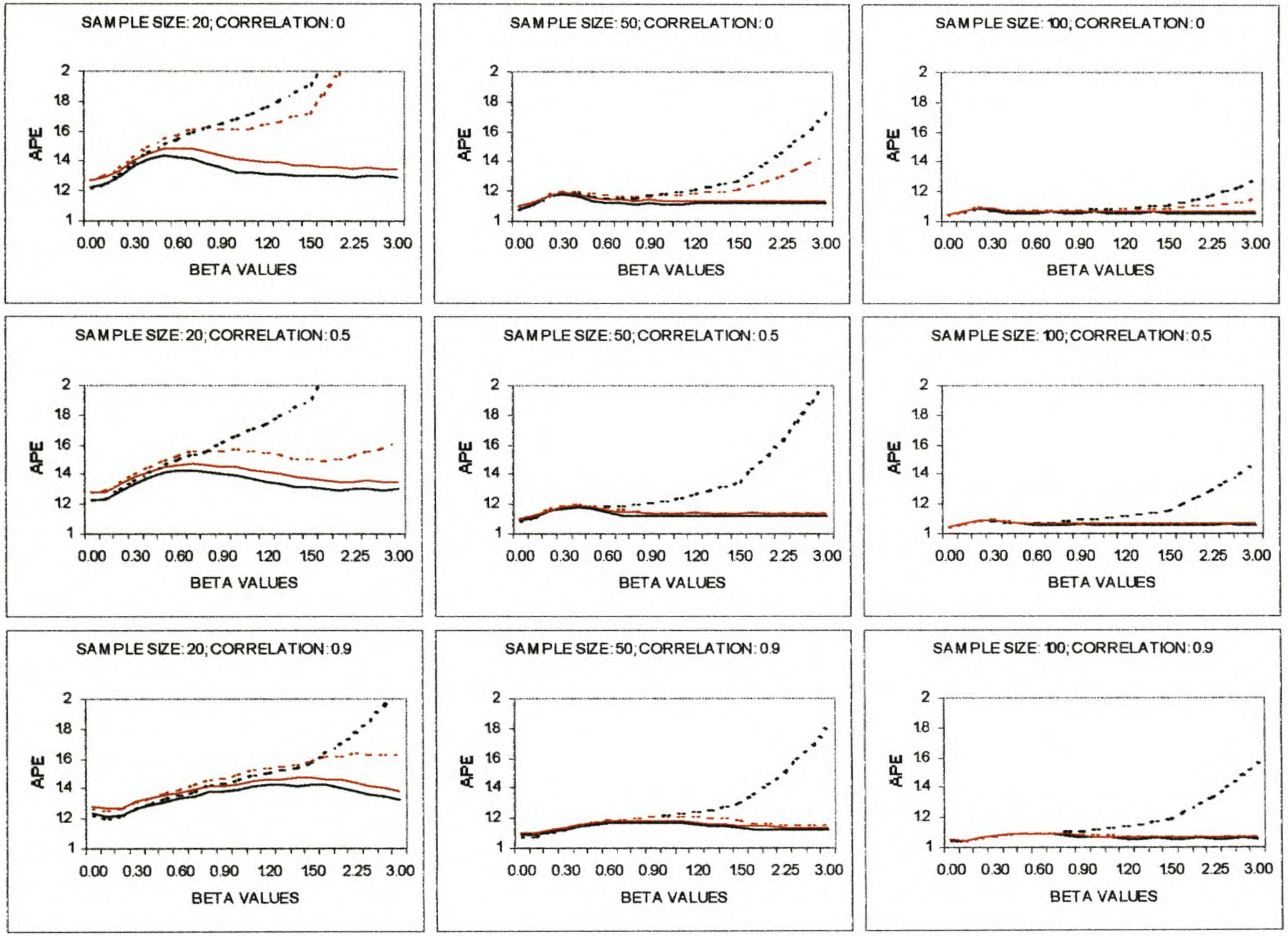


Figure 5.1: The APE of models selected from simulated data sets with $m = 5$ predictors; $\beta_0 = 1, \beta_1 = \beta_2 = ... = \beta_5 = 0(0.1)1.5$ and $\beta_1 = \beta_2 = ... = \beta_5 = 1.5(0.25)3$

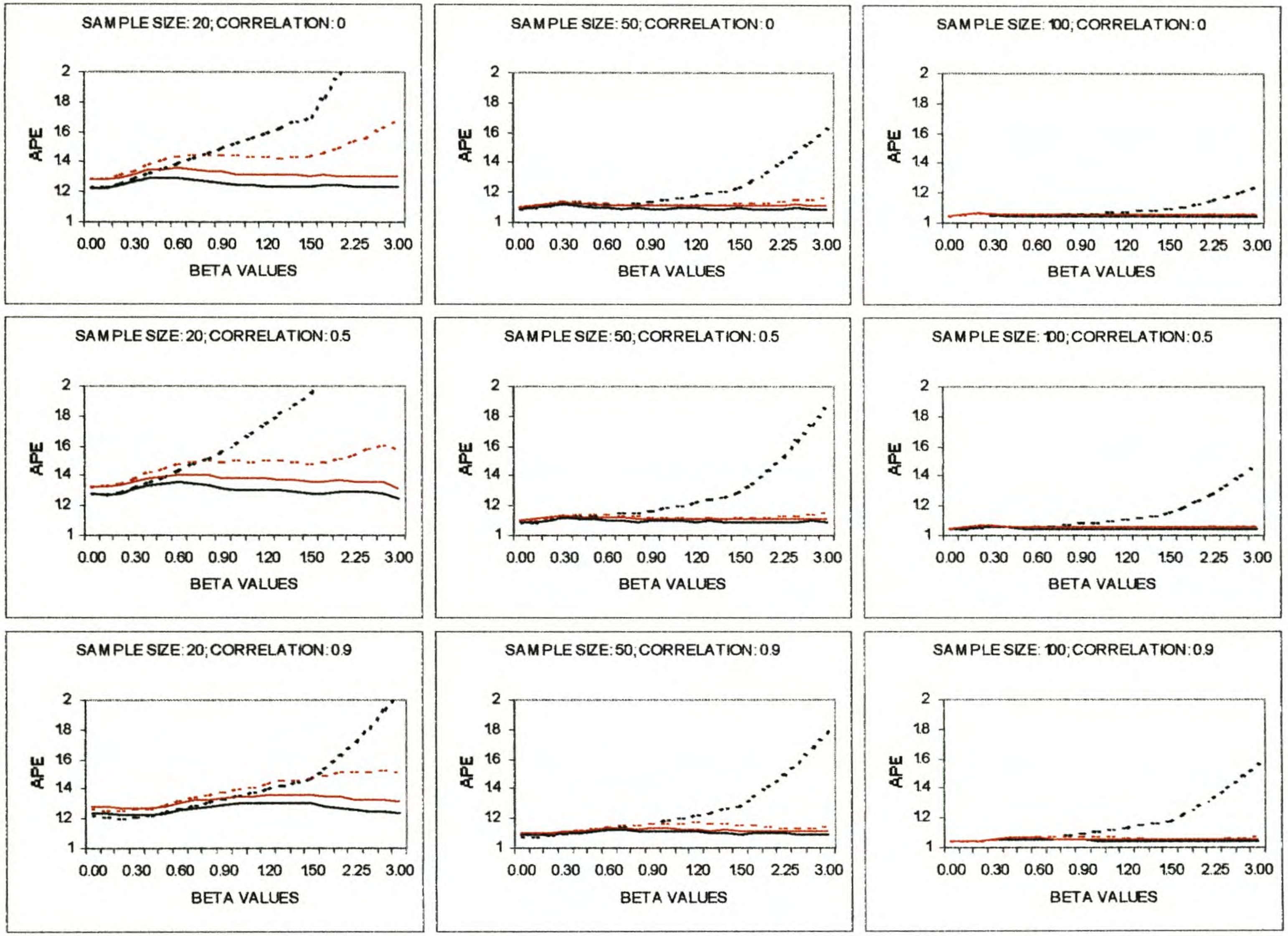


Figure 5.2: The APE of models selected from simulated data sets with $m = 5$ predictors; $\beta_0 = 1$, $\beta_1 = \beta_2 = 0(0.1)1.5$ and $\beta_1 = \beta_2 = 1.5(0.25)3$ and $\beta_3 = \beta_4 = \beta_5 = 0$

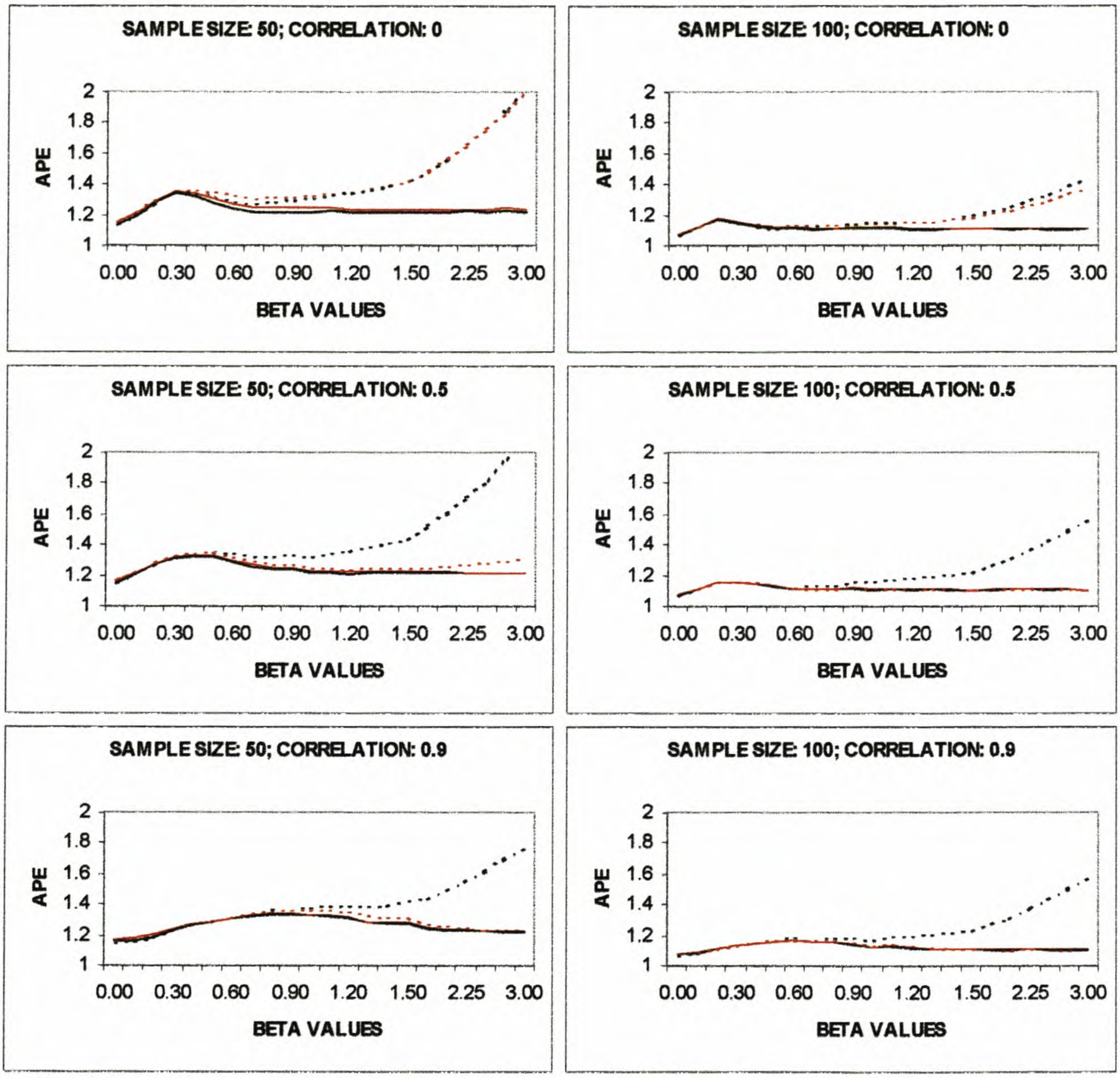


Figure 5.3: The APE of models selected from simulated data sets with $m = 10$ predictors; $\beta_0 = 1$, $\beta_1 = \beta_2 = ... = \beta_{10} = 0(0.1)1.5$ and $\beta_1 = \beta_2 = ... = \beta_{10} = 1.5(0.25)3$

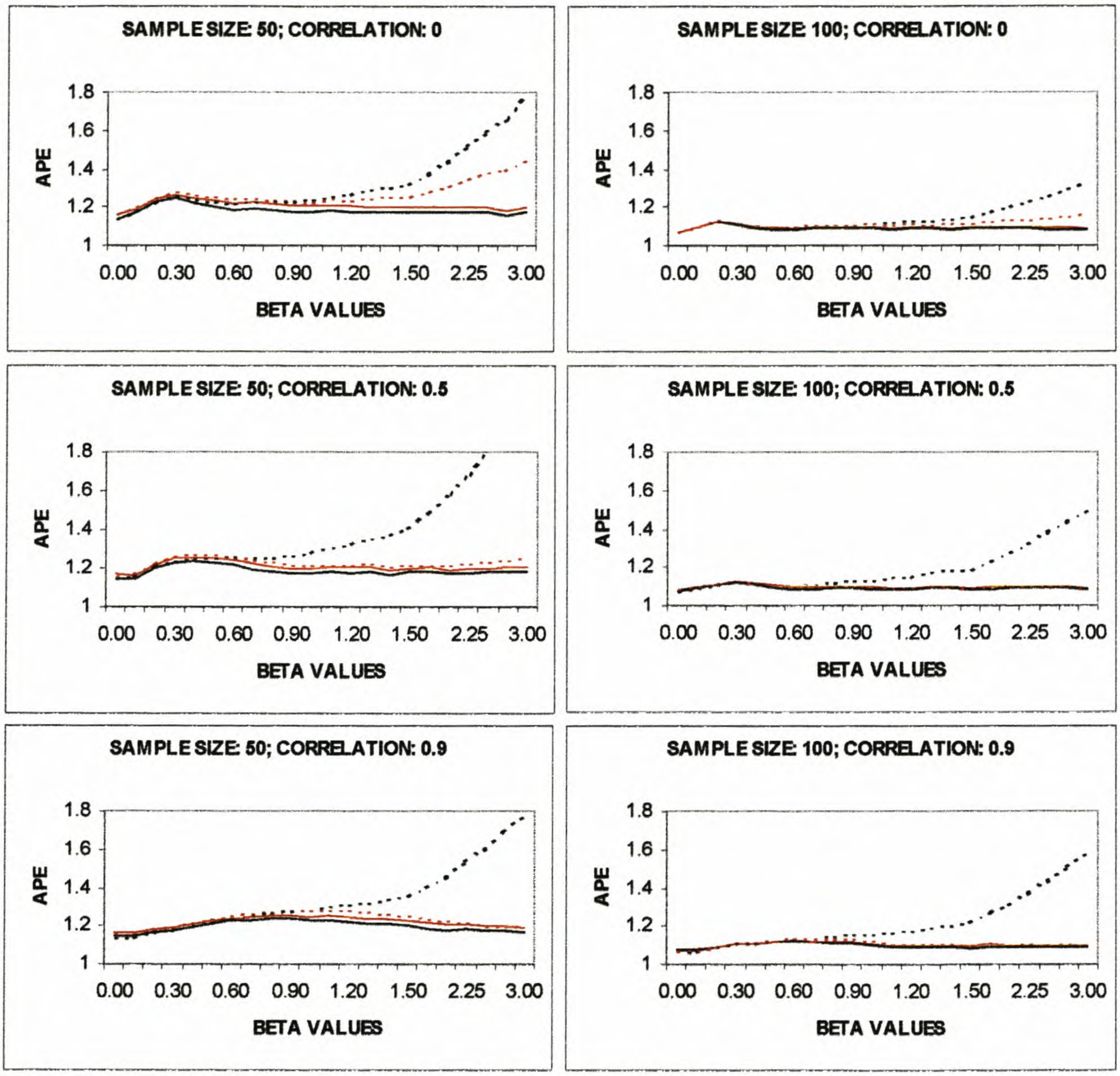


Figure 5.4: The APE of models selected from simulated data sets with $m = 10$ predictors; $\beta_0 = 1$, $\beta_1 = \beta_2 = ... = \beta_5 = 0(0.1)1.5$ and $\beta_1 = \beta_2 = ... = \beta_5 = 1.5(0.25)3$ and $\beta_6 = \beta_7 = ... = \beta_{10} = 0$

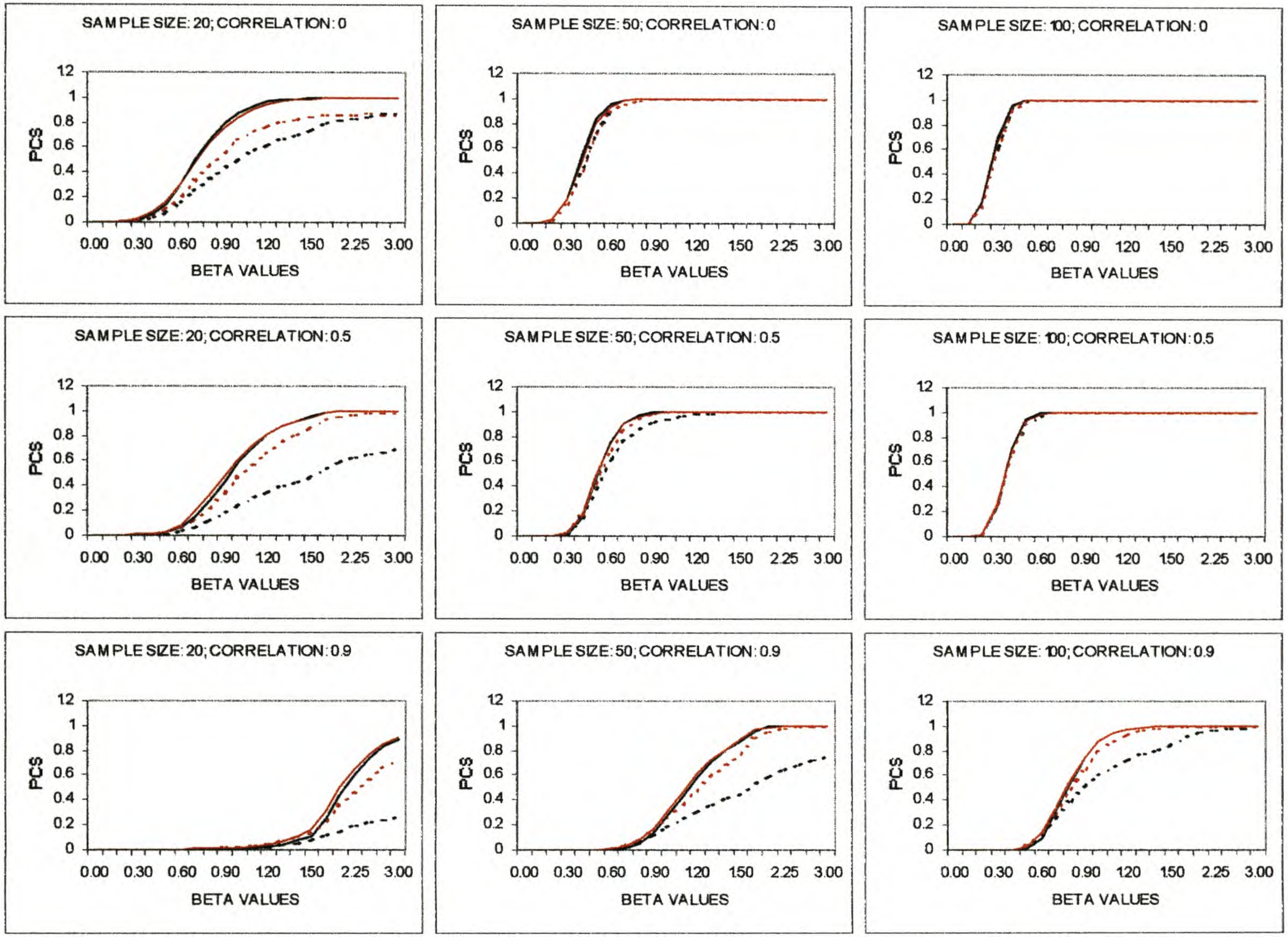


Figure 5.5: The PCS for models selected from simulated data sets with $m = 5$ predictors; $\beta_0 = 1$, $\beta_1 = \beta_2 = ... = \beta_5 = 0(0.1)1.5$ and $\beta_1 = \beta_2 = ... = \beta_5 = 1.5(0.25)3$

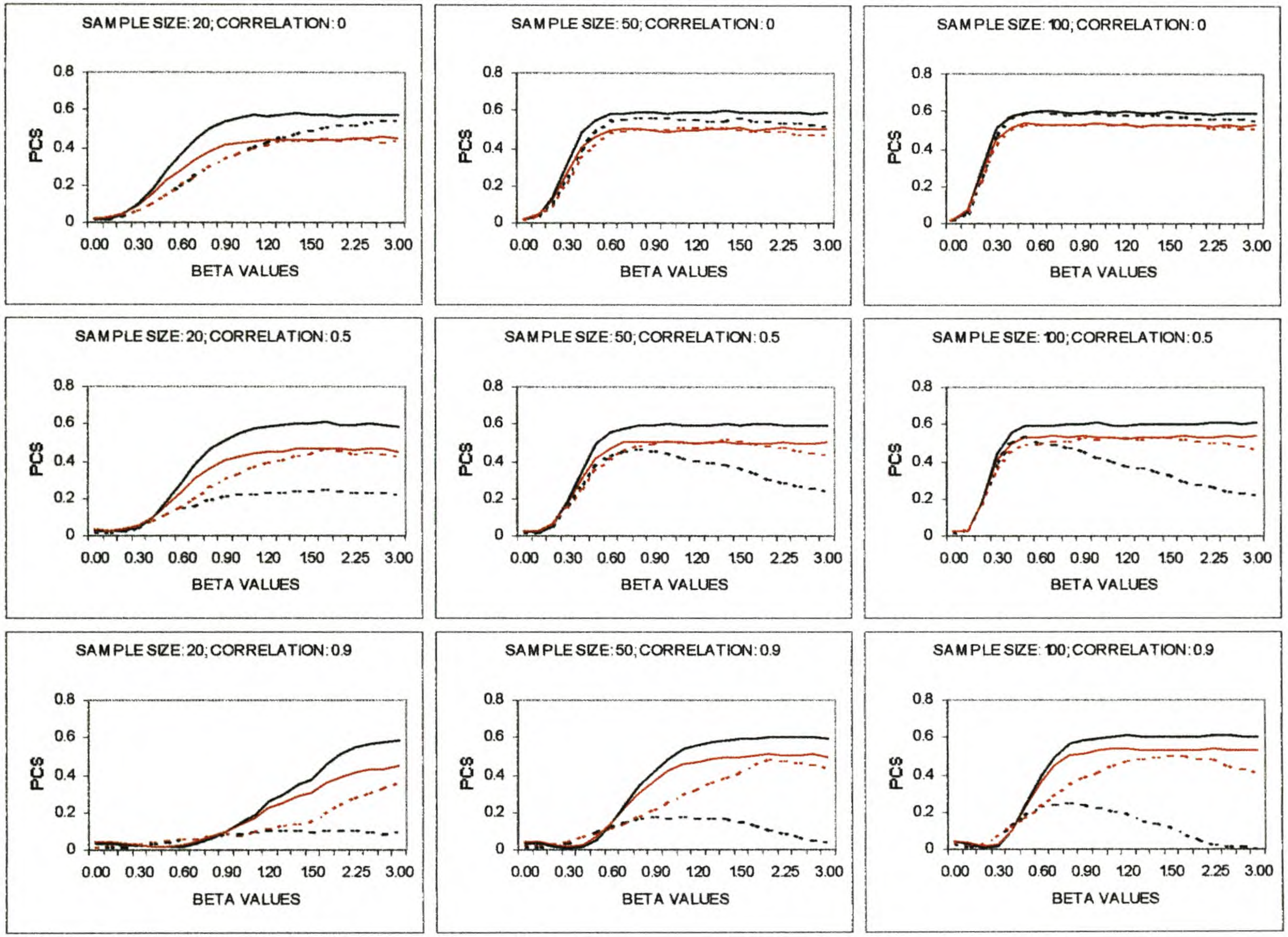


Figure 5.6: The PCS for models selected from simulated data sets with $m = 5$ predictors; $\beta_0 = 1$, $\beta_1 = \beta_2 = 0(0.1)1.5$ and $\beta_1 = \beta_2 = 1.5(0.25)3$ and $\beta_3 = \beta_4 = \beta_5 = 0$

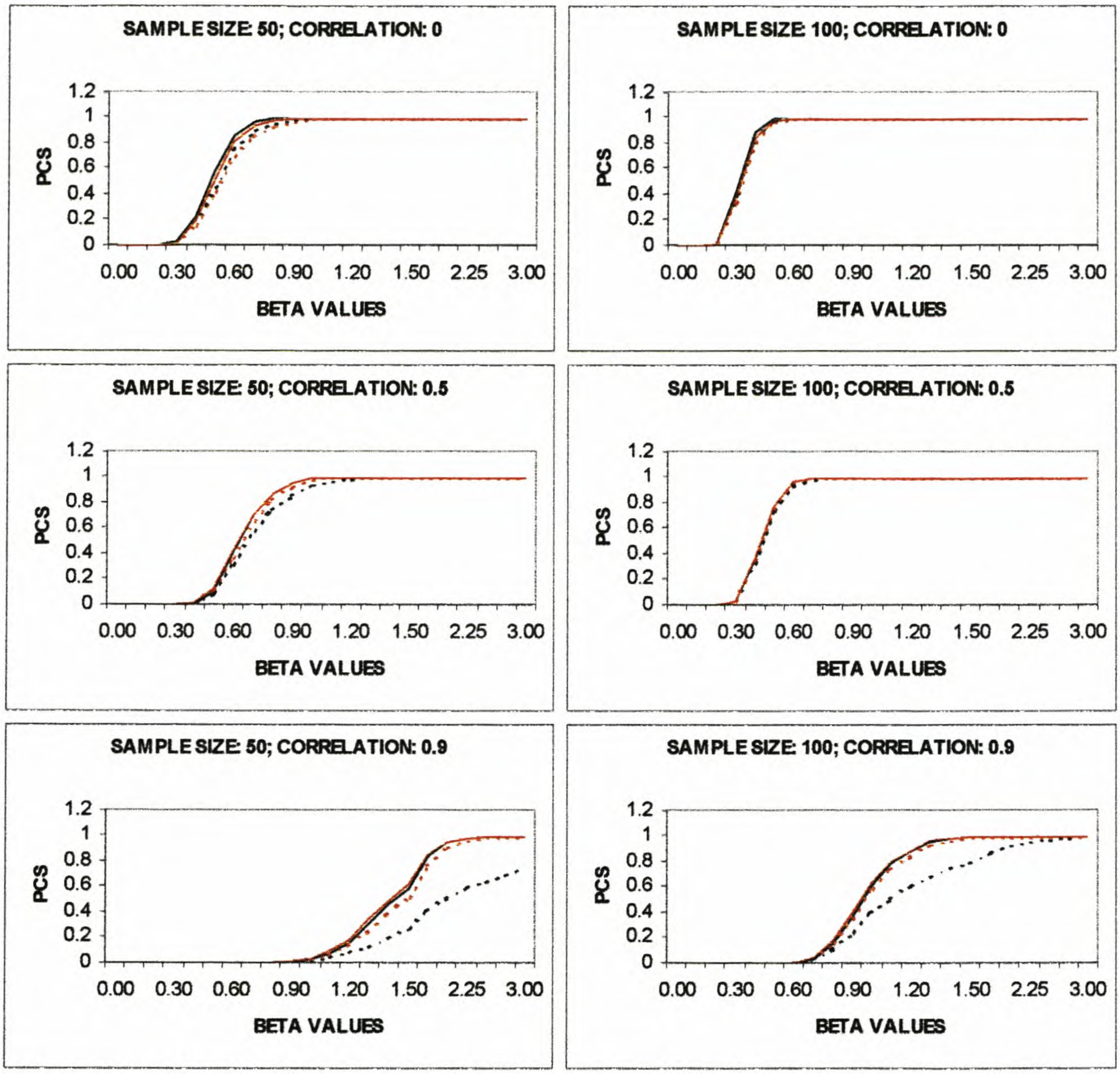


Figure 5.7: The PCS for models selected from simulated data sets with $m = 10$ predictors; $\beta_0 = 1, \beta_1 = \beta_2 = ... = \beta_{10} = 0(0.1)1.5$ and $\beta_1 = \beta_2 = ... = \beta_{10} = 1.5(0.25)3$

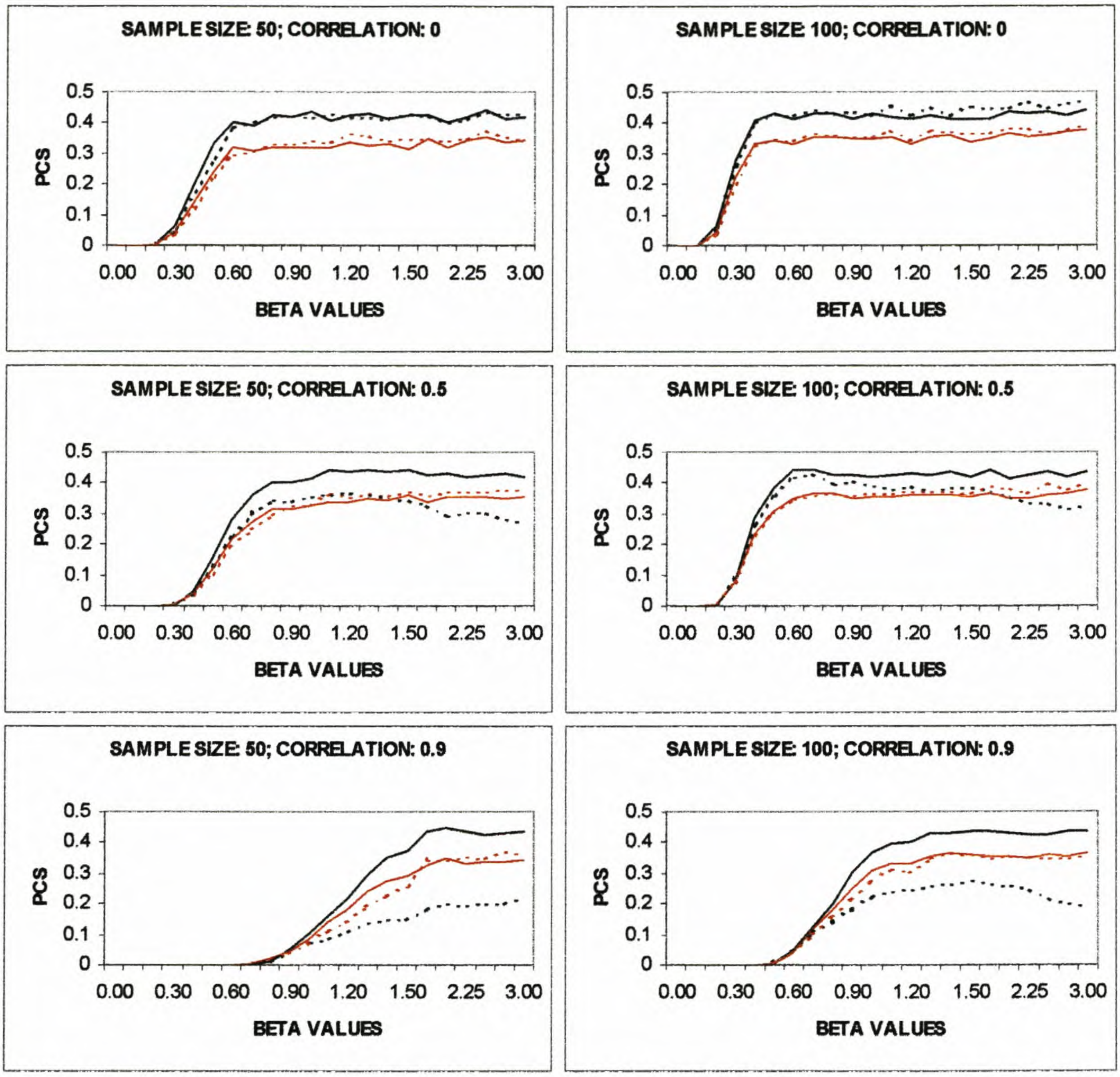


Figure 5.8: The PCS for models selected from simulated data sets with $m = 10$ predictors; $\beta_0 = 1$, $\beta_1 = \beta_2 = ... = \beta_5 = 0(0.1)1.5$ and $\beta_1 = \beta_2 = ... = \beta_5 = 1.5(0.25)3$ and $\beta_6 = \beta_7 = ... = \beta_{10} = 0$

5.6.1 Discussion of the simulation results

Firstly, consider the APEs plotted in Figures 5.1 to 5.4. Recall that the four APEs shown in each of these figures correspond to models selected and fitted from the following data sets:

(a) Ordinary data sets, i.e. data sets not containing deliberately inserted possibly selection influential data cases, and without any attempt being made to identify and omit such cases. The corresponding APE is shown as a solid black line.

(b) Modified data sets, i.e. data sets that do contain deliberately inserted possibly selection influential data cases. Once again, no attempt is made to identify and omit such cases. The corresponding APE is shown as a dotted black line.

(c) Ordinary data sets, but now we do apply the p-value approach in an attempt to identify and omit selection influential data cases. The corresponding APE is shown as a solid red line.

(d) Modified data sets, but once again we do apply the p-value approach in an attempt to identify and omit selection influential data cases. The corresponding APE is shown as a dotted red line.

We note first of all that the solid black line and the solid red line virtually coincide in all the cases. The only slight exceptions are the small sample cases ($n = 20$). This implies that application of the p-value approach in cases where it is not really required, i.e. in situations where the ordinary data set contains no deliberately inserted possibly selection influential data cases, does not lead to a dramatic increase in the APE of the resulting model. In general, therefore, it seems that the price paid in terms of increased APE if the p-value approach is applied to a data set that typically does not contain selection influential data cases, is fairly small.

Comparing the two dotted lines, we observe that the dotted red line is almost always significantly below the dotted black line. This is especially the case when the sample size is fairly large, or even in small sample cases when the correlation amongst the predictor variables is large. The only exceptions occur when the sample size is small ($n = 20$), especially when the predictor variables are (close to) being uncorrelated, or for the uncorrelated cases shown in Figure 5.3. The general conclusion to be drawn from these results is that application of the p-value approach to identify selection influential data cases, followed by omission of these cases before model selection and fitting are done, pays off in terms of reduced APE if the data set does indeed contain selection influential cases. This payoff is especially significant in large sample cases when there is a fairy high correlation amongst the predictor variables.

It is interesting to observe that in several cases the dotted red line comes very close to the two solid lines. This is true for large sample sizes, and high correlation amongst the predictor variables. The implication is that in such situations the p-value approach applied to data sets containing selection influential data cases fares almost as well as the p-value approach applied to ordinary data sets.

Consider *secondly* the PCSs plotted in Figures 5.5 to 5.8. In the cases where all the regression coefficients are increased from 0 to 3 (shown in Figure 5.5 for $m = 5$ and in Figure 5.7 for $m = 10$). the two solid lines are virtually indistinguishable, irrespective of sample size and the correlation amongst the predictor variables. In these cases we also have the dotted red line almost always above the dotted black line, confirming the value in terms of increased PCS of first applying the p-value approach before selecting and fitting a regression model to data. The situation becomes much more complicated if only some of the regression coefficients are moved away from zero (shown in Figure 5.6, where two out of $m = 5$ regression coefficients are increased, and in Figure 5.8, where five out of $m = 10$ coefficients are increased). Although it is difficult to discern general patterns in these figures, it does seem as if the p-value approach is especially worthwhile in the cases where the correlation amongst the predictor variables is large. In these cases the dotted red line is generally above the dotted black line, in some instances even exceeding the solid red line. The behaviour in the uncorrelated cases in Figure 5.8 is somewhat difficult to explain. It should, however, be noted that all the PCS values in this case are fairly small.

All in all the APE and PCS graphs suggest that the proposed p-value approach for identifying and eliminating selection influential cases in a data set is worthwhile.

CHAPTER 6

CONCLUDING REMARKS

6.1 Looking back

In this dissertation we first of all studied the influence of data cases if variable selection is done in multiple linear regression by using the C_p criterion. This influence was investigated in terms of the average prediction error (APE) of the resulting model, and in terms of the probability of correct selection (PCS) when variable selection is applied. A definition was developed of the concept of a *selection influential* data case. A simulation study revealed that the presence of potentially selection influential data cases in a data set leads to an increase in the APE of the resulting model, while the PCS is reduced. These findings emphasize the importance of identifying and dealing with (possibly by omitting them) selection influential data cases before model selection and fitting are performed. Attention was therefore devoted to the problem of identifying selection influential data cases. Two new measures were developed for this purpose. The first new measure is based on a leave-one-out strategy. This measure simply quantifies the relative change in the selection criterion (the C_p criterion in our study) if a data case is omitted from the full data set. Important in this regard is the fact that variable selection is repeated on the reduced data set before the measure is calculated. This leads to a so-called *unconditional* measure of selection influence (see Leger and Altman, 1993, for arguments in favour of such an approach). A bootstrap procedure was proposed and applied in example data sets as an aid for deciding whether a given value of the maximum relative change in the C_p criterion (over all data cases omitted one at a time) should be interpreted as a significant indication of the corresponding data case being selection influential. The new measure was applied in several example data sets, and it seems that it succeeds in identifying individual points that should be considered for omission because of their undue influence on the selection criterion.

The second new measure of selection influence that was developed and investigated in the dissertation is based on a p-value approach. The general idea behind this measure is as follows.

The basic problem in variable selection may be viewed as one of using the available data to identify a single linear subspace L from a family of subspaces $\mathcal{L}$. In the process a corresponding subset of the predictor variables in the multiple linear regression setup is identified. A commonly used strategy to achieve this end is to define a measure of (estimation) error, and to calculate an estimate of this error for each of the available subspaces. The selected subspace is then simply the subspace with smallest estimated (estimation) error. In our study we used the expected squared error of estimation (ESEE) as a measure of error. The ESEE corresponding to a given linear subspace is defined by

$$E\left\|P_L\mathbf{Y}-\boldsymbol{\mu}\right\|^2=\sum_{i=1}^{n}\sigma^2\left\|P_L\mathbf{u}_i\right\|^2+\left\langle\boldsymbol{\mu},P_{L^\perp}\mathbf{u}_i\right\rangle^2=\sum_{i=1}^{n}E\left(\widehat{\mu}_i(L)-\mu_i\right)^2.$$

The C_p criterion for the subspace L is an unbiased estimate of the ESEE above. We showed in the dissertation how this criterion may be written as a sum of n terms in a coordinate free context, where the ith term in this sum represents the contribution of the ith data case to the criterion value. It can now be argued that the ith data case is selection influential with respect to a given linear subspace if the contribution of this case to the value of the selection criterion for the particular subspace is very large or very small. To decide whether this contribution is indeed very large or very small, we proposed using a p-value based on the underlying distribution of the calculated contribution. This approach was developed in some detail, applied to example data sets, and investigated in a simulation study. It seems from the results that the measure has merit in terms of identifying selection influential data cases.

6.2 Looking forward

There are several interesting and challenging problems for further research. *Firstly*, utilizing the p-value approach described briefly in the previous section to identify selection influential data cases, involves estimation of a non-centrality parameter of a non-central chi-squared distribution. More specifically, in order to calculate the p-value for the ith data case and the jth linear subspace L, we need to estimate the non-centrality parameter. This was achieved satisfactorily if we could assume the value of σ^2 to be known. Incorporating the estimated non-centrality parameter into the p-value approach for identifying selection influential data cases could therefore be done successfully in these cases. However, estimating the non-centrality parameter if we could not assume the value of σ^2 to be known, proved more difficult. In fact, incorporating an estimate of the non-centrality parameter for these cases into the p-value approach, proved to be a problem that we could not resolve satisfactorily. This is therefore an area that is open to further research.

Secondly, another more basic problem that deserves further attention can be described as follows. Let t denote the total number of possible subspaces (models) that can be considered for selection in a particular application. In a multiple linear regression setup we have $t = 2^m - 1$, with m denoting the total number of predictor variables. If we use C_p variable selection, we effectively identify the subspace $\widehat{L}$ for which the estimated ESEE defined above is a minimum. Let $D_{j,i}$ be the contribution of the ith data case to the value of the C_p criterion for the jth subspace, $i = 1, 2, ..., n$; $j = 1, 2, ..., t$. We can think of the tn values $D_{j,i}$, $i = 1, 2, ..., n$; $j = 1, 2, ..., t$, as the basic data that have to be used in the variable selection process. Now consider the $t \times n$ matrix $\mathbf{D}$ with (j, i)th element equal to $D_{j,i}$. Note that the row index, j, in this matrix refers to the different subspaces or models, and that the column index, i, refers to the different points in the data set. For a given value of j we have a given subspace L, and the sum over i, i.e. the sum over the different columns, of the entries in the jth row gives the value of the selection criterion for this subspace. A basic question that may now be considered is: how can we use the information summarised in the matrix $\mathbf{D}$ to identify potentially selection influential data cases? In general one would feel that if $d_{j,i}$, the (j, i)th observed value of $D_{j,i}$ in the matrix $\mathbf{D}$, is very large or very small, it would signify that observation i may be selection influential with respect to the subspace (or model) corresponding to j. If $d_{j,i}$ is very large compared to the other entries in the ith column, it would mean that observation i plays a significant role in preventing the subspace corresponding to j from being selected (remember that we select the subspace having the smallest row total). In such a case omitting observation i may easily cause the selected subspace to change. Similarly, if $d_{j,i}$ is very small compared to the other entries in the ith column, it would mean that observation i plays a significant role in promoting the selection of the subspace corresponding to j. Once again we may find that omitting observation i under such circumstances may cause the selected subspace to change. Viewed in this light our problem is therefore to decide whether any observed value, $d_{j,i}$, of $D_{j,i}$ in the matrix $\mathbf{D}$ can be regarded as being extreme (i.e. very large or very small). The crucial question now is: how can such a decision be made? In our discussion we utilised a p-value approach to answer this question. Other approaches are certainly possible, and should be investigated.

Finally, there are several other more minor questions that deserve further research.

(a) Should data cases that have been identified as selection influential, be omitted from the data set, or should these cases rather be down-weighted before a model is selected and fitted?

(b) What can be said about identification of selection influential data points when the error term in a multiple linear regression model is not normally distributed?

(c) In the dıssertation attention was restricted to variable selection making use of the C_p criterion. This choice can be justified from the fact that in practical applications the C_p criterion is frequently used for variable selection. There are of course many other selection criteria that can be used, and investigating selection influence for these other criteria will be a worthwhile exercise.

(d) Using the p-value approach to decide whether data points are selection influential requires comparing calculated (or estimated) p-values to a given significance level. Specifying the latter is a problem that deserves further attention. For example, should we take into account the fact that nm p-values are compared to the significance level, and adjust the p-value for this multiplicity of comparisons? One possibility would be to use a Bonferroni type of adjustment, but this requires further investigation.

APPENDIX A
THE COORDINATE FREE APPROACH

In the multiple linear regression model

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\mu}$ is assumed to belong to the known $(m+1)$-dimensional linear subspace M of R^n, the design matrix $\mathbf{X}$, of full rank $(m+1)$, is a basis matrix for M. The linear subspace M is therefore coordinatized by $\mathbf{X}$. If we do not commit ourselves to a specific basis for M, results derived in the coordinate free linear subspace M, gain in generality. Other reasons why the coordinate free approach, also referred to as the vector space approach, is more appealing are listed in Arnold (1981, p.55). Also in this regard Snyman (1994, p.A-1) stated the following: "By using the vector space approach we win more freedom for the representation of linear models and related statistics. This approach promotes elegant, concise and simple arguments with the added benefit of direct geometric interpretability (important statistics can often be expressed as projections on a particular subspace or as lengths of such projections)."

In this Appendix we present some of the definitions and lemmas which are used in order to obtain the coordinate free results in the main text. Other standard results which are often encountered in coordinate free theory are also given. The proofs of these lemmas are omitted, and can be found in Arnold (1981, Chapter 2) or linear algebra textbooks.

i) **Definition A.1**

(a) Let R^n be the *vector space*, containing all n-dimensional vectors. Also let $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_p \in R^n$. Then $\mathbf{u}$ is a *linear combination* of the $\mathbf{v}_i$ if there exist $a_1, a_2, ..., a_p \in R^1$ such that $\mathbf{u} = a_1\mathbf{v}_1 + ... + a_p\mathbf{v}_p$.

(b) Let V be a set, $V \subset R^n$. Then V is a *linear subspace* if V is closed under the operation of taking linear combinations, that is, for all $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_p \in V$, and for all $a_1, a_2, ..., a_p \in R^1$, $\mathbf{u} = a_1\mathbf{v}_1 + ... + a_p\mathbf{v}_p \in V$.

(c) Let $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_p \in V$, where V is a linear subspace. The following results hold:

I. The $\mathbf{v}_i$ *span* V if every $\mathbf{v} \in V$ can be written as a linear combination of the $\mathbf{v}_i$.

II. The $\mathbf{v}_i$ are *linearly independent* if $a_1\mathbf{v}_1 + \ldots + a_p\mathbf{v}_p = \mathbf{0}$ implies that $a_1 = \ldots = a_p = 0$.

III. The $\mathbf{v}_i$ are a *basis* for V if they are linearly independent and they span V.

IV. If, in addition to (c), $\|\mathbf{v}_i\| = 1$ and $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0$ for $i \neq j$, then the $\mathbf{v}_i$ are an *orthonormal basis* for V.

V. The $n \times p$ matrix $\mathbf{X}$ is a *basis matrix* for V if the columns of $\mathbf{X}$ form a basis for V.

VI. The $n \times p$ matrix $\mathbf{X}$ is an *orthonormal basis matrix* for V if the columns of $\mathbf{X}$ form a orthonormal basis for V.

ii) **Lemma A.1**

(a) Let V be a linear subspace of R^n. Then the following are true:

I. V has an orthonormal basis.

II. Any two bases for V have the same number of vectors. This number is called the dimension of V, denoted by $\dim(V)$.

III. If $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_p \in V$ form a basis of V, then every vector in V can be written in exactly one way as a linear combination of the $\mathbf{v}_i$.

IV. If V has dimension p and $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_p \in V$ and the $\mathbf{v}_i$ are linearly independent, then the $\mathbf{v}_i$ form a basis for V.

(b) Let $\mathbf{X}$ be a basis matrix for the p-dimensional linear subspace $V \subset R^n$. Then:

I. $\mathbf{X}$ is an $n \times p$ matrix of rank p and $\mathbf{X}'\mathbf{X}$ is invertible.

II. The vector $\mathbf{v} \in V$ if and only if $\mathbf{v} = \mathbf{Xb}$ for some vector unique vector $\mathbf{b} \in R^p$.

iii) **Definition A.2**

Let $W \subset V$ be linear subspaces of R^n.

(a) The orthogonal complement of V, written as $V^{\perp}$, is the set of all vectors orthogonal to V.

(b) The set of all vectors in V that are orthogonal to W is denoted by $V \cap W^{\perp} = V \mid W$.

iv) **Lemma A.2**

Consider the linear subspaces $W \subset V \subset R^n$. Then:

(a) $\dim(V^{\perp}) = n - \dim(V)$ and $\dim(V \mid W) = \dim(V) - \dim(W)$.

(b) If the dimension of W equals the dimension of V, then $V = W$.

v) **Definition A.3**

Let V be a linear subspace of R^n, with orthonormal basis $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_p \in V$. For any $\mathbf{y} \in R^n$:

(a) The orthogonal projection of $\mathbf{y}$ onto V is defined as $P_V\mathbf{y} = \sum_{i=1}^{p} \langle \mathbf{y}, \mathbf{v}_i \rangle \mathbf{v}_i$.

(b) The orthogonal projection of $\mathbf{y}$ onto V is a vector $\mathbf{v}$ such that $\mathbf{v} \in V$ and $\mathbf{y} - \mathbf{v} \in V^{\perp}$.

vi) **Lemma A.3**

(a) Let $\mathbf{X}$ be a basis matrix for the p-dimensional linear subspace $V \in R^n$. The projection of $\mathbf{y}$ onto V is given by $P_V\mathbf{y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ with squared norm $\|P_V\mathbf{y}\|^2 = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

(b) Let $\mathbf{X} = \left[\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p\right]$ be an orthonormal basis matrix for the linear subspace $V \subset R^n$. Then $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$, $P_V\mathbf{y} = \mathbf{X}\mathbf{X}'\mathbf{y} = \sum_{i=1}^{p} \langle \mathbf{y}, \mathbf{x}_i \rangle \mathbf{x}_i$ and $\|P_V\mathbf{y}\|^2 = \|\mathbf{X}'\mathbf{y}\|^2$.

(c) Consider the linear subspaces $W \subset V \subset R^n$. Also, let $\mathbf{x}, \mathbf{y}$ be any vectors in R^n and a, b any real constants. Then:

I. $P_V = P_V'$, $P_V P_V = P_V$, $\dim(V) = rank(P_V) = tr(P_V)$.

II. $\mathbf{x} \in V$ if and only if $P_V\mathbf{x} = \mathbf{x}$, and $\mathbf{x}$ is orthogonal to V, written as $\mathbf{x} \perp V$, if and only if $P_V\mathbf{x} = \mathbf{0}$.

III. $\mathbf{x} = P_V\mathbf{x} + P_{V^{\perp}}\mathbf{x}$ and $P_{V|W}\mathbf{x} = P_V\mathbf{x} - P_W\mathbf{x} = P_{V^{\perp}}\mathbf{x} + P_{W^{\perp}}\mathbf{x}$.

IV. $\langle P_V\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, P_V\mathbf{y} \rangle$.

V. From the Pythagoras identity $\|a\mathbf{x} + b\mathbf{y}\|^2 = a^2\|\mathbf{x}\|^2 + b^2\|\mathbf{y}\|^2 + 2ab\langle \mathbf{x}, \mathbf{y} \rangle$ follows identities such as

(i) $\|\mathbf{x}\|^2 = \|P_V\mathbf{x}\|^2 + \|P_{V^{\perp}}\mathbf{x}\|^2$

(ii) $\left\|P_{V|W}\mathbf{x}\right\|^2 = \|P_V\mathbf{x}\|^2 - \|P_W\mathbf{x}\|^2 = \|P_{W^{\perp}}\mathbf{x}\|^2 - \|P_{V^{\perp}}\mathbf{x}\|^2$.

vii) **Lemma A.4**

(a) Let $\mathbf{Y} : n \times 1$ be a random vector from a multivariate normal distribution $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

I. If $\mathbf{a}$ is a fixed vector in R^n, then the linear function $\mathbf{a}'\mathbf{Y} = \langle \mathbf{a}, \mathbf{Y} \rangle \sim N(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a})$.

II. If $\mathbf{A}$ is a constant $q \times n$ matrix of rank q, where $q \leqq n$, then $\mathbf{AY} \sim N_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$.

III. If $\boldsymbol{\Sigma}$ is nonsingular, then $(\mathbf{Y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \boldsymbol{\mu}) \sim \chi^2_n$.

(b) Let $\mathbf{a}, \mathbf{b}$ be any fixed vectors in R^n and let $\mathbf{Y} : n \times 1 \sim N_n(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$. Then:

I. $\mathbf{a}'\mathbf{Y} = \langle \mathbf{a}, \mathbf{Y} \rangle \sim N(\mathbf{a}'\boldsymbol{\mu}, \|\mathbf{a}\|^2 \sigma^2)$.

II. $Cov\left(\langle \mathbf{a}, \mathbf{Y} \rangle, \langle \mathbf{b}, \mathbf{Y} \rangle\right) = \langle \mathbf{a}, \mathbf{b} \rangle \sigma^2$.

(c) Let V be some linear subspace of R^n and $\mathbf{Y} : n \times 1 \sim N_n(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$. Then:

I. $P_V \mathbf{Y} \sim N_n(P_V \boldsymbol{\mu}, \sigma^2 P_V)$.

II. $\frac{\|P_V \mathbf{Y}\|^2}{\sigma^2} \sim \chi'^2_{\dim(V)}\left(\frac{\|P_V \boldsymbol{\mu}\|^2}{\sigma^2}\right)$.

III. $E\|P_V \mathbf{Y}\|^2 = \sigma^2 \dim(V) + \|P_V \boldsymbol{\mu}\|^2$.

IV. $Var\|P_V \mathbf{Y}\|^2 = 2\sigma^2 \dim(V) + 4\sigma^2 \|P_V \boldsymbol{\mu}\|^2$.

APPENDIX B

EXAMPLE DATA SETS

i) **The Hald data** (Draper and Smith, 1998)

The Hald data are given in Table B.1. The description of the columns are as follows:

Column 1: Observation number

The response variable:

Y: The heat evolved in calories per gram of cement

The four predictor variables are:

x1: The amount of tricalcium aliminate

x2: The amount of tricalcium silicate

x3: The amount of tetracalcium alumino ferrite

x4: The amount of dicalcium silicate

Observation	Y	x1	x2	x3	x4
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

Table B.1: The Hald data; 13 observations; Y response; 4 predictor variables

ii) **The fuel data** (Weisberg, 1985)

The fuel data are given in Table B.2. The description of the columns are as follows:

Column 1: Observation number

The response variable:

Y: The 1972 fuel consumption (in gallons per capita)

The four predictor variables are:

x1: The amount of tax on a gallon of fuel (in cents)

x2: The percentage of the population with a driver's license

x3: The average income (in thousands of dollars)

x4: The total length of roads (in thousands of miles)

Observation	Y	x1	x2	x3	x4
1	541	9	52.5	3.571	1.976
2	524	9	57.2	4.092	1.25
3	561	9	58	3.865	1.586
4	414	7.5	52.9	4.87	2.351
5	410	8	54.4	4.399	0.431
6	457	10	57.1	5.342	1.333
7	344	8	45.1	5.319	11.868
8	467	8	55.3	5.126	2.138
9	464	8	52.9	4.447	8.577
10	498	7	55.2	4.512	8.507
11	580	8	53	4.391	5.939
12	471	7.5	52.5	5.126	14.186
13	525	7	57.4	4.817	6.93
14	508	7	54.5	4.207	6.58
15	566	7	60.8	4.332	8.159
16	635	7	58.6	4.318	10.34
17	603	7	57.2	4.206	8.508
18	714	7	54	3.718	4.725
19	865	7	72.4	4.716	5.915
20	640	8.5	67.7	4.341	6.01
21	649	7	66.3	4.593	7.834
22	540	8	60.2	4.983	0.602
23	464	9	51.1	4.897	2.449
24	547	9	51.7	4.258	4.686
25	460	8.5	55.1	4.574	2.619
26	566	9	54.4	3.721	4.746
27	577	8	54.8	3.448	5.399
28	631	7.5	57.9	3.846	9.061
29	574	8	56.3	4.188	5.975
30	534	9	49.3	3.601	4.65
31	571	7	51.8	3.64	6.905
32	554	7	51.3	3.333	6.594
33	577	8	57.8	3.063	6.524
34	628	7.5	54.7	3.357	4.121
35	487	8	48.7	3.528	3.495
36	644	6.58	62.9	3.802	7.834
37	640	5	56.6	4.045	17.782
38	704	7	58.6	3.897	6.385
39	648	8.5	66.3	3.635	3.274
40	968	7	67.2	4.345	3.905
41	587	7	62.6	4.449	4.639
42	699	7	56.3	3.656	3.985
43	632	7	60.3	4.3	3.635
44	591	7	50.8	3.745	2.611
45	782	6	67.2	5.215	2.302
46	510	9	57.1	4.476	3.942
47	610	7	62.3	4.296	4.083
48	524	7	59.3	5.002	9.794
49	551	8	45.2	5.162	3.246
50	345	5	64.8	4.995	0.602

Table B.2: The Fuel data; Observations 1 to 50; Y response; 4 predictor variables

iii) **The evaporation data** (Freund, 1979)

The evaporation data is given in Table B.3. The description of the columns are as follows:

Column 1: Observation number

Column 2: Month

Column 3: Day

The response variable:

Y: The amount of evaporation from the soil

The ten predictor variables are:

x1: The maximum daily soil temperature

x2: The minimum daily soil temperature

x3: The integrated area under the soil temperature curve

x4: The maximum daily air temperature

x5: The minimum daily air temperature

x6: The integrated area under the daily temperature curve

x7: The maximum daily relative humidity

x8: The minimum daily relative humidity

x9: The integrated area under the daily humidity curve

x10: The total wind measured in miles per day

Obsvervation	Month	Day	Y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1	6	6	30	84	65	147	85	59	151	95	40	398	273
2	6	7	34	84	65	149	86	61	159	94	28	345	140
3	6	8	33	79	66	142	83	64	152	94	41	388	318
4	6	9	26	81	67	147	83	65	158	94	50	406	282
5	6	10	41	84	68	167	88	69	180	93	46	379	311
6	6	11	4	74	66	131	77	67	147	96	73	478	446
7	6	12	5	73	66	131	78	69	159	96	72	462	294
8	6	13	20	75	67	134	84	68	159	95	70	464	313
9	6	14	31	84	68	161	89	71	195	95	63	430	455
10	6	15	38	86	72	169	91	76	206	93	56	406	604
11	6	16	43	88	73	178	91	76	208	94	55	393	610
12	6	17	47	90	74	187	94	76	211	94	51	385	520
13	6	18	45	88	72	171	94	75	211	96	54	405	663
14	6	19	45	88	72	171	92	70	201	95	51	392	467
15	6	20	11	81	69	154	87	68	167	95	61	448	184
16	6	21	10	79	68	149	83	68	162	95	59	436	177
17	6	22	30	84	69	160	87	66	173	95	42	392	173
18	6	23	29	84	70	160	87	68	177	94	44	392	76
19	6	24	23	84	70	168	88	70	169	95	48	398	72
20	6	25	16	77	67	147	83	66	170	97	60	431	183
21	6	26	37	87	67	166	92	67	196	96	44	379	76
22	6	27	50	89	69	171	92	72	199	94	48	393	230
23	6	28	36	89	72	180	94	72	204	95	48	394	193
24	6	29	54	93	72	186	92	73	201	94	47	386	400
25	6	30	44	93	74	188	93	72	206	95	47	389	339
26	7	1	41	94	75	199	94	72	208	96	45	370	172
27	7	2	45	93	74	193	95	73	214	95	50	396	238
28	7	3	42	93	74	196	95	70	210	96	45	380	118
29	7	4	50	96	75	198	95	71	207	93	40	365	93
30	7	5	48	95	76	202	95	69	202	93	39	357	269
31	7	6	17	84	73	173	96	69	173	94	58	418	128
32	7	7	20	91	71	170	91	69	168	94	44	420	423
33	7	8	15	88	72	179	89	70	189	93	50	399	415
34	7	9	42	89	72	179	95	71	210	98	46	389	300
35	7	10	44	91	72	182	96	73	208	95	43	384	193
36	7	11	41	92	74	196	97	75	215	96	46	389	195
37	7	12	49	94	75	192	96	69	198	95	36	380	215
38	7	13	53	96	75	195	95	67	196	97	24	354	185
39	7	14	53	93	76	198	94	75	211	93	43	364	466
40	7	15	21	88	74	188	92	73	198	95	52	405	399
41	7	16	1	88	74	178	90	74	197	95	61	447	232
42	7	17	44	91	72	175	94	70	205	94	42	380	275
43	7	18	44	92	72	190	95	71	209	96	44	379	166
44	7	19	46	92	73	189	96	72	208	93	42	372	189
45	7	20	47	94	75	194	95	71	208	93	43	373	164
46	7	21	50	96	76	202	96	71	208	94	40	368	139

Table B.3: The Evaporation data; Observations 1 to 46; Month; Day; Y response; 10 predictor variables

APPENDIX C

FORTRAN PROGRAMS

```
C     PROGRAM C1
C     THIS PROGRAM GIVES AS OUTPUT THE VALUES OF THE APPROXIMATED
C     VARIANCES IN (2.36) AND (2.37).
C     THE VALUE OF SIGMA IS ASSUMED TO BE KNOWN

      USE MSIMSL
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      PARAMETER (NN=20,NMC=10000,MUTEL=61)
      DIMENSION Z(NN),Y(NN),AMU(NN),UM(NN,NN),PARAM(NN),PMY(NN),UV(NN)
      DIMENSION BERAMER1(NN),BERAMER2(NN),PROJ(NN),VERH1(MUTEL)
      DIMENSION SOMVAR1(NN),SOMVAR2(NN),PROJMU(NN),VERH2(MUTEL)

      CHARACTER*70 FILEOUT

C     CREATE AN OUTPUT FILE
      FILEOUT='C:\OUTPUT.TXT

C     SET UP THE VALUES OF MU
      NN2=NN/2
C     SIGMA EQUALS ONE
      SIGMA=1.0D0
      DO 600 II=1,MUTEL
      DO 15 I=1,NN
      IF (I.LE.NN2) AMU(I)=1.0D0*II-31.0D0
      IF (I.GT.NN2) AMU(I)=1.0D0
15    CONTINUE

C     SET UP THE VALUES OF THE U-VECTORS
      DO 20 I=1,NN
      DO 19 J=1,NN
      UM(I,J)=0.0D0
19    CONTINUE
      UM(I,I)=1.0D0
20    CONTINUE

C     DETERMINE THE PARAMETER WHICH HAS TO BE ESTIMATED
      CALL PL(AMU,PROJMU)
      DO 30 I=1,NN
      DO 28 J=1,NN
      UV(J)=UM(I,J)
28    CONTINUE
      CALL PL(UV,PROJ)
      ANTW=0.0D0
      DO 29 L=1,NN
      ANTW=ANTW+PROJ(L)*PROJ(L)
29    CONTINUE
      PARAM(I)=((AMU(I)-PROJMU(I))**2.0D0)
30    CONTINUE
```

```
      DO 31 I=1,NN
      SOMVAR1(I)=0.0D0
      SOMVAR2(I)=0.0D0
31    CONTINUE

C     START THE LOOP FOR GENERATING NEW ERRORS
      DO 500 III=1,NMC

C     DETERMINE THE Y-VALUES
      CALL DRNNOR(NN,Z)
      DO 90 I=1,NN
      Y(I)=AMU(I)+SIGMA*Z(I)
90    CONTINUE

C     DETERMINE THE VALUES OF THE TWO ESTIMATORS
      DO 100 I=1,1
      DO 96 J=1,NN
      UV(J)=UM(I,J)
96    CONTINUE
      CALL PL(UV,PROJ)
      ANTW=0.0D0
      DO 97 L=1,NN
      ANTW=ANTW+PROJ(L)*PROJ(L)
97    CONTINUE
      SOM=0.0D0
      DO 98 K=1,NN
      SOM=SOM+Y(K)*PROJ(K)
98    CONTINUE
      BERAMER1(I)=((Y(I)-SOM)**2.0D0)-(SIGMA*SIGMA)*(1.0D0-ANTW)

      CALL PML(UV,PROJ)
      SOM1=0.0D0
      SOM2=0.0D0
      DO 99 K=1,NN
      SOM1=SOM1+PROJ(K)*PROJ(K)
      SOM2=SOM2+PROJ(K)*Y(K)
99    CONTINUE
      BERAMER2(I)=(SOM2**2.0D0)-SOM1*(SIGMA*SIGMA)

100   CONTINUE

      DO 110 I=1,NN
      SOMVAR1(I)=SOMVAR1(I)+(BERAMER1(I)-PARAM(I))**2.0D0
      SOMVAR2(I)=SOMVAR2(I)+(BERAMER2(I)-PARAM(I))**2.0D0
110   CONTINUE

500   CONTINUE

      DO 510 I=1,NN
      SOMVAR1(I)=SOMVAR1(I)/NMC
      SOMVAR2(I)=SOMVAR2(I)/NMC
510   CONTINUE

      VERH1(II)=SOMVAR1(1)/SOMVAR2(1)
600   CONTINUE

      OPEN(1,FILE=FILEOUT)
      DO 900 I=1,MUTEL
      WRITE(1,*) I,VERH1(I)
900   CONTINUE
      CLOSE(1)

1000  STOP
      END
```

```
C     SUBROUTINE FOR CALCULATING THE PROJECTION OF A VECTOR (X) ON M
      SUBROUTINE PM(X,PMX)
      USE MSIMSL
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      PARAMETER (NN=20,NMC=1000)
      DIMENSION X(NN),PMX(NN)
      NN2=NN/2
      SOM=0.0D0
      DO 10 I=1,NN2
      SOM=SOM+X(I)
10    CONTINUE
      GEM1=SOM/NN2
      SOM=0.0D0
      DO 20 I=NN2+1,NN
      SOM=SOM+X(I)
20    CONTINUE
      GEM2=SOM/NN2
      DO 30 I=1,NN
      IF (I.LE.NN2) PMX(I)=GEM1
      IF (I.GT.NN2) PMX(I)=GEM2
30    CONTINUE
      RETURN
      END

C     SUBROUTINE FOR CALCULATING THE PROJECTION OF A VECTOR (X) ON L
      SUBROUTINE PL(X,PLX)
      USE MSIMSL
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      PARAMETER (NN=20,NMC=1000)
      DIMENSION X(NN),PLX(NN)
      NN2=NN/2
      SOM=0.0D0
      DO 10 I=NN2+1,NN
      SOM=SOM+X(I)
10    CONTINUE
      GEM1=SOM/NN2
      DO 20 I=1,NN
      IF (I.LE.NN2) PLX(I)=0.0D0
      IF (I.GT.NN2) PLX(I)=GEM1
20    CONTINUE
      RETURN
      END

C     SUBROUTINE FOR CALCULATING THE PROJECTION OF A VECTOR (X) ON M|L
      SUBROUTINE PML(X,PMLX)
      USE MSIMSL
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      PARAMETER (NN=20,NMC=1000)
      DIMENSION X(NN),PP1(NN),PP2(NN),PMLX(NN)
      CALL PM(X,PP1)
      CALL PL(X,PP2)
      DO 10 I=1,NN
      PMLX(I)=PP1(I)-PP2(I)
10    CONTINUE
      RETURN
      END
```

```
C       PROGRAM C2
C       THIS PROGRAM GIVES AS OUTPUT THE Cp CRITERION AND SELECTED VARIABLES IN TABLE 3.1 FOR
        THE HALD DATA.
C       THE Cp SELECTION CRITERION IS APPLIED TO THE COMPLETE DATA SET.
C       THE ith DATA CASE IS THEN OMITTED FROM THE DATA SET AND Cp IS CALCULATED FOR EACH OF
C       THE REDUCED DATA SET.
C       NOTE THAT IN CALCULATING Cp FOR THE REDUCED DATA SETS, THE ERROR VARIANCE IS OBTAINED
C       FROM THE COMPLETE DATA SET

        USE MSIMSL
        IMPLICIT DOUBLE PRECISION (A-H,O-Z)
        PARAMETER(IP=4,NN=13,IPP1=IP+1,N=NN-1)

        PARAMETER(NVAR1=IPP1,LDCOV1=NVAR1,NOBS1=NN,ICRIT1=3,NBEST1=1,NGOOD1=10,IPRINT1=0,
        &LDCOEF1=NBEST1*IP,NSIZE1=IP,LINDVAR1=NGOOD1*NSIZE1*(NSIZE1+1)/2,NTBEST1=NBEST1)

        DIMENSION CRIT1(NGOOD1*NSIZE1),COEF1(LDCOEF1,5)
        DIMENSION ICRITX1(NSIZE1+1),IVARX1(NSIZE1+1),INDVAR1(LINDVAR1)
        DIMENSION ICOEFX1(NTBEST1+1)
        DIMENSION COV(IPP1,IPP1)
        DIMENSION XMEAN(IPP1)
        DIMENSION INCD(1,1)

        DIMENSION XY(NN,IPP1),XY1(N,IPP1),X(NN,IP),XX(N,P),Y(NN)
        DIMENSION B(0:IP),IRYW(NN),CPWEG1(NN),AKRIT(NN),KIESB(NN,IP)

        CHARACTER*70 FILEIN
        CHARACTER*70 FILEOUT

C       THE HALD DATA ARE USED AS INPUT FILE
        FILEIN='C:\HALD.TXT'
        FILEOUT='C:\OUTPUT.TXT'

C       READ THE DATA INTO XY
        OPEN(1,FILE=FILEIN)
        DO 1 I=1,NN
        READ(1,*) XY(I,IPP1),(XY(I,J),J=1,IP)
1       CONTINUE
        CLOSE(1)

C       THE VALUES OF THE PREDICTORS ARE PLACED IN X
        DO 3 J=1,IP
        DO 2 I=1,NN
        X(I,J)=XY(I,J)
      2 CONTINUE
      3 CONTINUE

C       THE VALUES OF THE RESPONSE ARE PLACED IN Y
        DO 5 I=1,NN
        Y(I)=XY(I,IPP1)
5       CONTINUE

        NOBS=NN
        LDX=NN
        NIND=IP
        INTCEP=1
        CALL DRLSE(NOBS,Y,NIND,X,LDX,INTCEP,B,SST,SSE)
        SIG=SSE/(NN-IP-1)
```

```
C     THE Cp CRITERION IS APPLIED TO THE FULL DATA SET
      IDO=0
      NROW=NN
      NVAR=IPP1
      LDX=NN
      IFRQ=0
      IWT=0
      MOPT=0
      ICOPT=1
      LDCOV=IPP1
      LDINCD=1

      CALL DCORVC(IDO,NROW,NVAR,XY,LDX,IFRQ,IWT,MOPT,ICOPT,XMEAN,COV,
      &LDCOV,INCD,LDINCD,NOBS,NMISS,SUMWT)

      CALL DRBEST(NVAR1,COV,LDCOV1,NOBS1,ICRIT1,NBEST1,NGOOD1,
      &IPRINT1,ICRITX1,CRIT1,IVARX1,INDVAR1,ICOEFX1,COEF1,LDCOEF1)

      IMIN=1
      AMIN=CRIT1(1)
      DO 10 I=2,IP
      IF (CRIT1(ICRITX1(I)).LT.AMIN) THEN
      AMIN=CRIT1(ICRITX1(I))
      IMIN=I
      ENDIF
10    CONTINUE
      IB=IVARX1(IMIN)

C     THE MATRIX XX, THAT CONTAINS THE SELECTED COLUMNS OF XY, IS SET UP
C     NOTE THAT IMIN = THE NUMBER OF PREDICTORS SELECTED
      DO 20 J=1,IMIN
      ITT=INDVAR1(IB+J-1)
      DO 15 I=1,NN
      XX(I,J)=X(I,ITT)
15    CONTINUE
20    CONTINUE

      NOBS=NN
      NIND=IMIN
      LDX=NN
      INTCEP=1
      CALL DRLSE(NOBS,Y,NIND,XX,LDX,INTCEP,B,SST,SSE)
      CPVOL=SSE/SIG+(2*(IMIN+1)-NN)
C     WRITE THE Cp CRITERION CALCULATED ON THE COMPLETE DATA SET ON THE SCREEN
      WRITE(6,*) CPVOL

C     THE LOOP THAT OMITS THE ith DATA CASE IS STARTED
      DO 50 II=1,NN
      NWEG=1
      IRYW(1)=II
C     SUBROUTINE THAT OMITS THE ith DATA CASE IS CALLED
      CALL WEGLAAT(NWEG,IRYW,XY,XY1)

C     SUBROUTINE THAT CALCULATES THE Cp CRITERION FOR THE REDUCED DATA SET IS CALLED
      CALL WAARWEG(II,XY1,SIG,CPWEG,KIESB)
      CPWEG1(II)=CPWEG
50    CONTINUE
```

```
      OPEN(1,FILE=FILEOUT)
      DO 950 I=1,NN
      WRITE(1,96C) I,CPWEG1(I),(KIESB(I,J),J=1,IP)
950   CONTINUE
      CLOSE(1)
960   FORMAT(I2,2X,F20.6,2X,10(I2,1X))

1000  STOP
      END

C     SUBROUTINE THAT OMITS THE ith DATA CASE
      SUBROUTINE WEGLAAT(NWEG,IRYW,XY,XY1)
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      PARAMETER (IP=4,NN=13,IPP1=IP+1,N=NN-1)
      DIMENSION XY(NN,IPP1),XY1(N,IPP1),IRYW(NN)
      ITEL=0
      DO 100 I=1,NN
      II=0
      DO 5 J=1,NWEG
      IF (IRYW(J).EQ.I) II=1
5     CONTINUE
      IF (II.EQ.1) GOTO 100
      ITEL=ITEL+1
      DO 10 J=1,IPP1
      XY1(ITEL,J)=XY(I,J)
10    CONTINUE
100   CONTINUE
      RETURN
      END

C     SUBROUTINE THAT CALCULATES THE Cp CRITERION FOR THE REDUCED DATA SET
      SUBROUTINE WAARWEG(II,XY1,SIG,ANT,KIESB)
      USE MSIMSL
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      PARAMETER(IP=4,NN=13,IPP1=IP+1,N=NN-1)
      PARAMETER(NVAR1=IPP1,LDCOV1=NVAR1,NOBS1=NN,ICRIT1=3,NBEST1=1,NGOOD1=10,IPRINT1=0,
      &LDCOEF1=NBEST1*IP,NSIZE1=IP,LINDVAR1=NGOOD1*NSIZE1*(NSIZE1+1)/2,NTBEST1=NBEST1)

      DIMENSION CRIT1(NGOOD1*NSIZE1),COEF1(LDCOEF1,5)
      DIMENSION ICRITX1(NSIZE1+1),IVARX1(NSIZE1+1),INDVAR1(LINDVAR1)
      DIMENSION ICOEFX1(NTBEST1+1)
      DIMENSION COV(IPP1,IPP1)
      DIMENSION XMEAN(IPP1)
      DIMENSION INCD(1,1)

      DIMENSION Y1(N),X1(N,IP),X2(N,IP),XY1(N,IPP1)
      DIMENSION B(0:IP),KIESB(NN,IP)

      DO 6 I=1,N
      DO 4 J=1,IP
      X1(I,J)=XY1(I,J)
4     CONTINUE
      Y1(I)=XY1(I,IPP1)
6     CONTINUE
```

```
C       THE Cp CRITERION IS APPLIED TO THE REDUCED DATA SET
        IDO=0
        NROW=N
        NVAR=IPP1
        LDX=N
        IFRQ=0
        IWT=0
        MOPT=0
        ICOPT=1
        LDCOV=IPP1
        LDINCD=1

        CALL DCORVC(IDO,NROW,NVAR,XY1,LDX,IFRQ,IWT,MOPT,ICOPT,XMEAN,COV,
        &LDCOV,INCD,LDINCD,NOBS,NMISS,SUMWT)

        CALL DRBEST(NVAR1,COV,LDCOV1,NOBS1,ICRIT1,NBEST1,NGOOD1,
        &IPRINT1,ICRITX1,CRIT1,IVARX1,INDVAR1,ICOEFX1,COEF1,LDCOEF1)

        IMIN=1
        AMIN=CRIT1(1)
        DO 10 I=2,IP
        IF (CRIT1(ICRITX1(I)).LT.AMIN) THEN
        AMIN=CRIT1(ICRITX1(I))
        IMIN=I
        ENDIF
10      CONTINUE
        IB=IVARX1(IMIN)

        DO 20 J=1,IMIN
        ITT=INDVAR1(IB+J-1)
        KIESB(II,J)=ITT
        DO 15 I=1,N
        X2(I,J)=X1(I,ITT)
15      CONTINUE
20      CONTINUE

        NOBS=N
        NIND=IMIN
        LDX=N
        INTCEP=1
        CALL DRLSE(NOBS,Y1,NIND,X2,LDX,INTCEP,B,SST,SSE)
        ANT=SSE/SIG+(2*(IMIN+1)-N)

        RETURN
        END
```

```
C       PROGRAM C3
C       THIS PROGRAM GIVES AS OUTPUT THE COOK DISTANCES IN TABLE 3.1 FOR THE HALD DATA

        USE MSIMSL
        IMPLICIT DOUBLE PRECISION (A-H,O-Z)
        PARAMETER(IP=4,NN=13,IPP1=IP+1)
        DIMENSION XY(NN,IPP1),Y(NN),X(NN,IP),X1(NN,IPP1),X1Y(NN,IPP1),XY1(NN-1,IPP1)
        DIMENSION YKAP(NN),YKAP1(NN)
        DIMENSION B(IPP1),IRYW(NN),D(NN),DU(NN)

        CHARACTER*70 FILEIN
        CHARACTER*70 FILEOUT

        FILEIN='C:\HALD.TXT'
        FILEOUT='C:\OUTPUT.TXT'

        OPEN(1,FILE=FILEIN)
        DO 5 I=1,NN
        READ(1,*) XY(I,IPP1),(XY(I,J),J=1,IP)
5       CONTINUE
        CLOSE(1)

        DO 6 I=1,NN
        Y(I)=XY(I,IPP1)
6       CONTINUE

        DO 8 I=1,NN
        DO 7 J=1,IP
        X(I,J)=XY(I,J)
7       CONTINUE
8       CONTINUE

C       CALCULATE THE MSE OF THE MODEL FITTED TO THE FULL DATA SET
        NOBS=NN
        NIND=IP
        LDX=NN
        INTCEP=1
        CALL DRLSE(NOBS,Y,NIND,X,LDX,INTCEP,B,SST,SSE)
        GSKF=SSE/(NN-IP-1)

C       CALCULATE THE PREDICTED VALUES FROM THE MODEL FITTED TO THE FULL DATA SET
        DO 35 I=1,NN
        YKAP(I)=B(1)
35      CONTINUE

        DO 45 I=1,NN
        DO 40 J=1,IP
        YKAP(I)=YKAP(I)+(B(J+1)*X(I,J))
40      CONTINUE
45      CONTINUE

C       THE LOOP THAT OMITS THE ith DATA CASE IS STARTED
        DO 50 II=1,NN
        D(II)=0.0D0
        NWEG=1
        IRYW(1)=II

C       SUBROUTINE THAT OMITS THE ith DATA CASE IS CALLED
        CALL WEGLAAT(NWEG,NN,IRYW,XY,XY1)
```

```
C       SUBROUTINE THAT CALCULATES THE PREDICTED VALUES OF THE MODEL FITTED TO THE
C       REDUCED DATA SET IS CALLED
        CALL WAARWEG(XY1,XY,II,YKAP1)

C       CALCULATE THE SQUARED DIFFERENCES BETWEEN THE TWO SETS OF PREDICTED VALUES
        DO 46 I=1,NN
        D(II)=D(II)+((YKAP(I)-YKAP1(I))*(YKAP(I)-YKAP1(I)))
46      CONTINUE

C       CALCULATE COOK'S DISTANCE
        DU(II)=D(II)/(GSKF*(IP+1))
50      CONTINUE

        OPEN(1,FILE=FILEOUT)
        DO 950 I=1,NN
        WRITE(1,960) I,DU(I)
950     CONTINUE
        CLOSE(1)
960     FORMAT(I2,2X,F12.8)
1000    STOP
        END

C       SUBROUTINE THAT OMITS THE ith DATA CASE
        SUBROUTINE WEGLAAT(NWEG,N,IRYW,XX,X1)
        IMPLICIT DOUBLE PRECISION (A-H,O-Z)
        PARAMETER (IP=4,NN=13,IPP1=IP+1)
        DIMENSION XX(NN,IPP1),X1(NN-1,IPP1),IRYW(NN)

        ITEL=0
        DO 100 I=1,N
        II=0
        DO 5 J=1,NWEG
        IF (IRYW(J).EQ.I) II=1
5       CONTINUE
        IF (II.EQ.1) GOTO 100
        ITEL=ITEL+1
        DO 10 J=1,IPP1
        X1(ITEL,J)=XX(I,J)
10      CONTINUE
100     CONTINUE

        RETURN
        END

C       SUBROUTINE THAT CALCULATES THE PREDICTED VALUES OF THE MODEL FITTED TO
C       THE REDUCED DATA SET
        SUBROUTINE WAARWEG(XY1,XY,II,YKAP1)
        USE MSIMSL
        IMPLICIT DOUBLE PRECISION (A-H,O-Z)
        PARAMETER(IP=4,NN=13,IPP1=IP+1,N=NN-1)
        DIMENSION Y1(N),X3(1,IP),YKAP1(NN),XY(NN,IPP1),XY1(N,IPP1),B(IPP1)

        DO 6 I=1,N
        Y1(I)=XY1(I,IPP1)
6       CONTINUE

        DO 21 J=1,IP
        X3(1,J)=XY(II,J)
21      CONTINUE
```

```
C     CALCULATE THE ESTIMATED COEFFICIENTS OF THE MODEL FITTED TO THE REDUCED DATA SET
      NOBS=N
      NIND=IP
      LDX=N
      INTCEP=1
      CALL DRLSE(NOBS,Y1,NIND,XY1,LDX,INTCEP,B,SST,SSE)

C     CALCULATE THE PREDICTED VALUES
      DO 25 I=1,NN
      YKAP1(I)=B(1)
25    CONTINUE

      IF (II.EQ.1) THEN
              DO 28 J=1,IP
              YKAP1(1)=YKAP1(1)+X3(1,J)*B(J+1)
28            CONTINUE
              DO 32 I=1,N
              DO 30 J=1,IP
              YKAP1(I+1)=YKAP1(I+1)+XY1(I,J)*B(J+1)
30            CONTINUE
32            CONTINUE
      ENDIF

      IF (II.EQ.NN) THEN
              DO 56 I=1,N
              DO 54 J=1,IP
              YKAP1(I)=YKAP1(I)+XY1(I,J)*B(J+1)
54            CONTINUE
56            CONTINUE
              DO 58 J=1,IP
              YKAP1(NN)=YKAP1(NN)+X3(1,J)*B(J+1)
58            CONTINUE
      ENDIF

      IF ((II.NE.1).AND.(II.NE.NN)) THEN
              DO 62 I=1,II-1
              DO 60 J=1,IP
              YKAP1(I)=YKAP1(I)+XY(I,J)*B(J+1)
60            CONTINUE
62            CONTINUE

              DO 64 J=1,IP
              YKAP1(II)=YKAP1(II)+X3(1,J)*B(J+1)
64            CONTINUE

              DO 70 I=II+1,NN
              DO 65 J=1,IP
              YKAP1(I)=YKAP1(I)+XY(I,J)*B(J+1)
65            CONTINUE
70            CONTINUE
      ENDIF

      RETURN
      END
```

```
C     PROGRAM C4
C     THIS PROGRAM GIVES AS OUTPUT THE APEs AND PCS-VALUES OF THE SELECTED MODELS IF Cp IS
C     APPLIED TO THE ORDINARY AND THEIR CORRESPONDING MODIFIED DATA SETS.  THE APEs OF
C     THE SELECTED MODELS FOR THIS SPECIFIC SIMULATION PROGRAM ARE PLOTTED IN TOP LEFT
C     CORNER (LANDSCAPE FORMAT) OF FIGURE 3.3. THE CORRESPONDING PCS-RESULTS ARE PLOTTED
C     IN THE TOP LEFT CORNER (LANDSCAPE FORMAT) OF FIGURE 3.7

      USE MSIMSL
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      PARAMETER (IP=5,NN=20,IPP1=IP+1,NMOD=2**IP,NMC1=50,NMC2=200)
      PARAMETER (XVARIANSIE=1.0D0,FOUTVARIANSIE=1.0D0,NUMBETAS=21)
      PARAMETER (XRHO=0.0D0,IXINVLOED=1,NBNIENUL=5)

      DIMENSION XM(NN,IP),Y(NN),XM1(NN,IP),YN(NN),XMI(NN,IP),XMI1(NN,IP)
      DIMENSION SIGMAM(IP,IP),SIGINV(IP,IP),RSIG(IP,IP),Z(NN)
      DIMENSION APEG(0:NUMBETAS),APEI(0:NUMBETAS)
      DIMENSION PCSG(0:NUMBETAS),PCSI(0:NUMBETAS)
      DIMENSION B(0:IP),BKAP(0:IP),KIESB(IP)

      CHARACTER*70 FILEIN
      CHARACTER*70 FILEOUT

      FILEOUT='C:\OUTPUT.TXT'

C     THE VALUE OF BETA 1 IS SET EQUAL TO 1
      B(0)=1.0D0
      ZSIG=DSQRT(FOUTVARIANSIE)

C     PREPARE THE SETUP FOR GENERATING THE DESIGN MATRICES
      DO 3 I=1,IP
      DO 2 J=1,IP
      SIGMAM(I,J)=XRHO
2     CONTINUE
      SIGMAM(I,I)=XVARIANSIE
3     CONTINUE
      TOL=1.0D2*DMACH(4)
      CALL DCHFAC(IP,SIGMAM,IP,TOL,IRANK,RSIG,IP)

C     CERTAIN SIMULATION COUNTERS ARE INITIALISED
      DO 5 I=0,NUMBETAS
      APEG(I)=0.0D0
      APEI(I)=0.0D0
      PCSG(I)=0.0D0
      PCSI(I)=0.0D0
5     CONTINUE

C     THE FOLLOWING LOOP IS REPEATED FOR DIFFERENT DESIGN MATRICES
      DO 800 IX=1,NMC1

C     GENERATE A NEW DESIGN MATRIX AND DETERMINE THE MAXIMUM AND
C     MINIMUM VALUES OF THE FIRST PREDICTOR
      CALL DRNMVN(NN,IP,RSIG,IP,XM,NN)
      AMIN=XM(1,IXINVLOED)
      AMAX=XM(1,IXINVLOED)
      NMIN=1
      NMAX=1
      DO 8 I=2,NN
      IF (XM(I,IXINVLOED).LE.AMIN) THEN
                  AMIN=XM(I,IXINVLOED)
                  NMIN=I
      ENDIF
```

```
        IF (XM(I,IXINVLOED).GE.AMAX) THEN
                  AMAX=XM(I,IXINVLOED)
                  NMAX=I
        ENDIF
8       CONTINUE

C       INCLUDE A POSSIBLY SELECTION INFLUENTIAL DATA CASE IN THE DESIGN MATRIX
        DO 12 I=1,NN
        DO 11 J=1,IP
        XMI(I,J)=XM(I,J)
11      CONTINUE
12      CONTINUE
        XMI(NMAX,IXINVLOED)=AMIN

C       THE LOOP THAT INCREMENTS THE BETA VALUES IS STARTED.
C       NOTE THAT THE BETA COEFFICIENT CONFIGURATION WHERE ALL THE BETA-VALUES
C       ARE INCREMENTED IS USED
        DO 500 IB=0,NUMBETAS

        IF (IB.LE.15) THEN
                  DO 13 J=1,NBNIENUL
                  B(J)=0.1D0*IB
13                CONTINUE
        ENDIF
        IF (IB.GT.15) THEN
                  DO 14 J=1,NBNIENUL
                  B(J)=1.5D0+0.25D0*(IB-15)
14                CONTINUE
                  ENDIF
        IF (NBNIENUL.LT.IP) THEN
                  DO 16 J=NBNIENUL+1,IP
                  B(J)=0.0D0
16                CONTINUE
        ENDIF

C       DETERMINE VALUES OF THE RESPONSE VARIABLE
        DO 400 IFOUT=1,NMC2
        CALL DRNNOR(NN,Z)
        DO 20 I=1,NN
        Y(I)=B(0)+ZSIG*Z(I)
        DO 19 J=1,IP
        Y(I)=Y(I)+B(J)*XM(I,J)
19      CONTINUE
20      CONTINUE

C       THE SUBROUTINE THAT APPLIES Cp TO THE ORDINARY DATA SET IS CALLED
        CALL ALLEMODELLE(XM,Y,NVAR,KIESB)

        DO 30 I=1,NN
        DO 29 J=1,NVAR
        XM1(I,J)=XM(I,KIESB(J))
29      CONTINUE
30      CONTINUE

        NOBS=NN
        NIND=NVAR
        LDX=NN
        INTCEP=1
        CALL DRLSE(NOBS,Y,NIND,XM1,LDX,INTCEP,BKAP,SST,SSE)
```

```
C     GENERATE A NEW VECTOR OF Y-VALUES
      CALL DRNNOR(NN,Z)
      DO 33 I=1,NN
      YN(I)=B(0)+ZSIG*Z(I)
      DO 32 J=1,IP
      YN(I)=YN(I)+B(J)*XM(I,J)
32    CONTINUE
33    CONTINUE

C     DETERMINE THE SUM OF THE APEs OF THE SELECTED MODELS
C     FROM THE ORDINARY DATA SETS
      S=0.0D0
      DO 40 I=1,NN
      YKAP=BKAP(0)
      DO 35 J=1,NVAR
      YKAP=YKAP+XM1(I,J)*BKAP(J)
35    CONTINUE
      S=S+(YKAP-YN(I))**2.0D0
40    CONTINUE
      APEG(IB)=APEG(IB)+S

C     DETERMINE THE SUM OF THE PCS-VALUES IF Cp IS APPLIED TO
C     THE ORDINARY DATA SETS
      IKIES=1
      IF (NVAR.NE.NBNIENUL) IKIES=0
      DO 45 J=1,NVAR
      IF (KIESB(J).NE.J) IKIES=0
45    CONTINUE
      IF (IKIES.EQ.1) PCSG(IB)=PCSG(IB)+1.0D0

C     THE SUBROUTINE THAT APPLIES Cp TO THE MODIFIED
C     DATA SETS IS CALLED
      CALL ALLEMODELLE(XMI,Y,NVAR,KIESB)

      DO 50 I=1,NN
      DO 49 J=1,NVAR
      XMI1(I,J)=XMI(I,KIESB(J))
      XM1(I,J)=XM(I,KIESB(J))
49    CONTINUE
50    CONTINUE

      NOBS=NN
      NIND=NVAR
      LDX=NN
      INTCEP=1
      CALL DRLSE(NOBS,Y,NIND,XMI1,LDX,INTCEP,BKAP,SST,SSE)

C     DETERMINE THE SUM OF THE APEs OF THE SELECTED MODELS FROM THE
C     MODIFIED DATA SETS
      S=0.0D0
      DO 60 I=1,NN
      YKAP=BKAP(0)
      DO 55 J=1,NVAR
      YKAP=YKAP+XM1(I,J)*BKAP(J)
55    CONTINUE
      S=S+(YKAP-YN(I))**2.0D0
60    CONTINUE
      APEI(IB)=APEI(IB)+S
```

```
C       DETERMINE THE SUM OF PCS-VALUES IF Cp IS APPLIED TO THE
C       MODIFIED DATA SETS
        IKIES=1
        IF (NVAR.NE.NBNIENUL) IKIES=0
        DO 65 J=1,NVAR
        IF (KIESB(J).NE.J) IKIES=0
65      CONTINUE
        IF (IKIES.EQ.1) PCSI(IB)=PCSI(IB)+1.0D0

400     CONTINUE
500     CONTINUE
800     CONTINUE

C       CALCULATE THE APE AND PCS OF THE SELELCTED MODELS FROM THE
C       ORDINARY AND MODIFIED DATA SETS
        DO 850 I=0,NUMBETAS
        APEG(I)=APEG(I)/(NMC1*NMC2*NN)
        APEI(I)=APEI(I)/(NMC1*NMC2*NN)
        PCSG(I)=PCSG(I)/(NMC1*NMC2)
        PCSI(I)=PCSI(I)/(NMC1*NMC2)
850     CONTINUE

        OPEN(1,FILE=FILEOUT)
        DO 920 I=0,NUMBETAS
        WRITE(1,975) APEG(I),APEI(I),PCSG(I),PCSI(I)
920     CONTINUE
620     CONTINUE
        CLOSE(1)

975     FORMAT(F9.6,2X,F9.6,2X,F9.6,2X,F9.6)

1000    STOP
        END


C       SUBROUTINE THAT APPLIES THE Cp CRITERION
        SUBROUTINE ALLEMODELLE(XM,Y,NVAR,KIESB)
        USE MSIMSL
        IMPLICIT DOUBLE PRECISION (A-H,O-Z)
        PARAMETER (IP=5,NN=20,IPP1=IP+1,NMOD=2**IP,NMC1=100,NMC2=100)

        PARAMETER(NVAR1=IPP1,LDCOV1=NVAR1,NOBS1=NN,
        &ICRIT1=3,NBEST1=1,NGOOD1=IP,IPRINT1=0,LDCOEF1=NBEST1*IP,
        &NSIZE1=IP,LINDVAR1=NGOOD1*NSIZE1*(NSIZE1+1)/2,NTBEST1=NBEST1)

        DIMENSION XM(NN,IP),Y(NN),XY(NN,IPP1)
        DIMENSION KIESB(IP)

        DIMENSION CRIT1(NGOOD1*NSIZE1),COEF1(LDCOEF1,5)
        DIMENSION ICRITX1(NSIZE1+1),IVARX1(NSIZE1+1),INDVAR1(LINDVAR1)
        DIMENSION ICOEFX1(NTBEST1+1)
        DIMENSION COV(IPP1,IPP1),XMEAN(IPP1)
        DIMENSION INCD(1,1)

        DO 5 I=1,NN
        XY(I,IPP1)=Y(I)
        DO 4 J=1,IP
        XY(I,J)=XM(I,J)
4       CONTINUE
5       CONTINUE
```

```
      IDO=0
      NROW=NN
      NVAR=IPP1
      LDX=NN
      IFRQ=0
      IWT=0
      MOPT=0
      ICOPT=1
      LDCOV=IPP1
      LDINCD=1

      CALL DCORVC(IDO,NROW,NVAR,XY,LDX,IFRQ,IWT,MOPT,ICOPT,XMEAN,
      &COV,LDCOV,INCD,LDINCD,NOBS,NMISS,SUMWT)

      CALL DRBEST(NVAR1,COV,LDCOV1,NOBS1,ICRIT1,NBEST1,NGOOD1,
      &IPRINT1,ICRITX1,CRIT1,IVARX1,INDVAR1,ICOEFX1,COEF1,LDCOEF1)

      IMIN=1
      AMIN=CRIT1(1)
      DO 12 I=2,IP
      IF (CRIT1(ICRITX1(I)).LT.AMIN) THEN
                  AMIN=CRIT1(ICRITX1(I))
                  IMIN=I
      ENDIF
12    CONTINUE
      NVAR=IMIN
      IBES=IVARX1(IMIN)

      DO 15 J=1,IP
      KIESB(J)=0
15    CONTINUE
      DO 20 J=1,IMIN
      KIESB(J)=INDVAR1(IBES+J-1)
20    CONTINUE

      RETURN
      END
```

```
C     PROGRAM C5
C     THIS PROGRAM GIVES AS OUTOUT COOK'S UNCONDITIONAL DISTANCE IN TABLE 4.1 FOR THE HALD DATA

      USE MSIMSL
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      PARAMETER(IP=4,NN=13,IPP1=IP+1)

      PARAMETER(NVAR1=IPP1,LDCOV1=NVAR1,NOBS1=NN,ICRIT1=3,NBEST1=1,NGOOD1=10,IPRINT1=0,+
      LDCOEF1=NBEST1*IP,NSIZE1=IP,LINDVAR1=NGOOD1*NSIZE1*(NSIZE1+1)/2,NTBEST1=NBEST1)

      DIMENSION CRIT1(NGOOD1*NSIZE1),COEF1(LDCOEF1,5)
      DIMENSION ICRITX1(NSIZE1+1),IVARX1(NSIZE1+1),INDVAR1(LINDVAR1)
      DIMENSION ICOEFX1(NTBEST1+1)
      DIMENSION COV(IPP1,IPP1)
      DIMENSION XMEAN(IPP1)
      DIMENSION INCD(1,1)

      DIMENSION XY(NN,IPP1),Y(NN),X(NN,IP),X1(NN,IPP1),XY1(NN-1,IPP1)
      DIMENSION YKAP(NN),YKAP1(NN)
      DIMENSION B(IPP1),IRYW(NN),D(NN),DU(NN)

      CHARACTER*70 FILEIN
      CHARACTER*70 FILEOUT

      FILEIN='C:\HALD.TXT'
      FILEOUT='C:\OUTPUT.TXT'

      OPEN(1,FILE=FILEIN)
      DO 5 I=1,NN
      READ(1,*) XY(I,IPP1),(XY(I,J),J=1,IP)
5     CONTINUE
      CLOSE(1)

      DO 6 I=1,NN
      Y(I)=XY(I,IPP1)
6     CONTINUE

      DO 8 I=1,NN
      DO 7 J=1,IP
      X(I,J)=XY(I,J)
7     CONTINUE
8     CONTINUE

C     CALCULATE THE MSE OF THE MODEL FITTED TO THE FULL DATA SET
      NOBS=NN
      NIND=IP
      LDX=NN
      INTCEP=1
      CALL DRLSE(NOBS,Y,NIND,X,LDX,INTCEP,B,SST,SSE)
      GSKF=SSE/(NN-IP-1)

C     Cp IS APPLIED TO THE FULL DATA SET
      IDO=0
      NROW=NN
      NVAR=IPP1
      LDX=NN
      IFRQ=0
      IWT=0
      MOPT=0
      ICOPT=1
      LDCOV=IPP1
      LDINCD=1

      CALL DCORVC(IDO,NROW,NVAR,XY,LDX,IFRQ,IWT,MOPT,ICOPT,XMEAN,COV,+
      LDCOV,INCD,LDINCD,NOBS,NMISS,SUMWT)

      CALL DRBEST(NVAR1,COV,LDCOV1,NOBS1,ICRIT1,NBEST1,NGOOD1,+
      IPRINT1,ICRITX1,CRIT1,IVARX1,INDVAR1,ICOEFX1,COEF1,LDCOEF1)
      IMIN=1
      AMIN=CRIT1(1)
      DO 10 I=2,IP
      IF (CRIT1(ICRITX1(I)).LT.AMIN) THEN
              AMIN=CRIT1(ICRITX1(I))
      IMIN=I
      ENDIF
10    CONTINUE
      IB=IVARX1(IMIN)

C     THE COLUMNS OF X1 CONTAIN THE PREDICTORS SELECTED FROM THE FULL DATA SET
      DO 20 J=1,IMIN
      ITT=INDVAR1(IB+J-1)
      DO 15 I=1,NN
      X1(I,J)=XY(I,ITT)
15    CONTINUE
20    CONTINUE

C     CALCULATE THE ESTIMATED REGRESSION COEFFICIENTS
      NOBS=NN
      NIND=IMIN
      LDX=NN
      INTCEP=1
      CALL DRLSE(NOBS,Y,NIND,X1,LDX,INTCEP,B,SST,SSE)
      IVERB=IMIN

C     CALCULATE THE PREDICTED VALUES
      DO 35 I=1,NN
      YKAP(I)=B(1)
35    CONTINUE

      DO 45 I=1,NN
      DO 40 J=1,IMIN
      YKAP(I)=YKAP(I)+(B(J+1)*X1(I,J))
40    CONTINUE
45    CONTINUE

C     THE LOOP THAT OMITS THE ith DATA CASE IS STARTED
      DO 50 II=1,NN
      D(II)=0.0D0
      NWEG=1
      IRYW(1)=II

C     SUBROUTINE THAT OMITS THE ith DATA CASE IS CALLED
      CALL WEGLAAT(NWEG,NN,IRYW,XY,XY1)

C     SUBROUTINE THAT APPLIES Cp TO THE REDUCED DATA SET IS CALLED.  THIS SUBROUTINE ALSO
C     CALCULATES THE PREDICTED VALUES
      CALL WAARWEG(XY1,XY,II,YKAP1)

C     CALCULATES THE SUM OF THE SQUARED DIFFERENCES BETWEEN THE TWO SETS OF PREDICTED VALUES
      DO 46 I=1,NN
      D(II)=D(II)+((YKAP(I)-YKAP1(I))*(YKAP(I)-YKAP1(I)))
46    CONTINUE

C     CALCULATE COOK'S UNCONDITIONAL DISTANCE
      DU(II)=D(II)/(GSKF*(IVERB+1))
50    CONTINUE

      OPEN(1,FILE=FILEOUT)
      DO 950 I=1,NN
      WRITE(1,960) I,DU(I)
950   CONTINUE
      CLOSE(1)
960   FORMAT(I2,2X,F12.8)
1000  STOP
      END
```

```
C       SUBROUTINE THAT OMITS THE ith DATA CASE
        SUBROUTINE WEGLAAT(NWEG,N,IRYW,XX,X1)
        IMPLICIT DOUBLE PRECISION (A-H,O-Z)
        PARAMETER (IP=4,NN=13,IPP1=IP+1)
        DIMENSION XX(NN,IPP1),X1(NN-1,IPP1),IRYW(NN)
        ITEL=0
        DO 100 I=1,N
        II=0
        DO 5 J=1,NWEG
        IF (IRYW(J).EQ.I) II=1
5       CONTINUE
        IF (II.EQ.1) GOTO 100
        ITEL=ITEL+1
        DO 10 J=1,IPP1
        X1(ITEL,J)=XX(I,J)
10      CONTINUE
100     CONTINUE

        RETURN
        END

C       SUBROUTINE THAT APPLIES Cp TO THE REDUCED DATA SET
        SUBROUTINE WAARWEG(XY1,XY,II,YKAP1)
        USE MSIMSL
        IMPLICIT DOUBLE PRECISION (A-H,O-Z)
        PARAMETER(IP=4,NN=13,IPP1=IP+1,N=NN-1)


        PARAMETER(NVAR1=IPP1,LDCOV1=NVAR1,NOBS1=N,ICRIT1=3,NBEST1=1,NGOOD1=10,IPRINT1=0,+
        LDCOEF1=NBEST1*IP,NSIZE1=IP,LINDVAR1=NGOOD1*NSIZE1*(NSIZE1+1)/2,NTBEST1=NBEST1)

        DIMENSION CRIT1(NGOOD1*NSIZE1),COEF1(LDCOEF1,5)
        DIMENSION ICRITX1(NSIZE1+1),IVARX1(NSIZE1+1),INDVAR1(LINDVAR1)
        DIMENSION ICOEFX1(NTBEST1+1)
        DIMENSION COV(IPP1,IPP1)
        DIMENSION XMEAN(IPP1)
        DIMENSION INCD(1,1)

        DIMENSION Y1(N),X2(N,IP),X3(1,IP),YKAP1(NN),YKAP(NN)
        DIMENSION XY(NN,IPP1),XY1(N,IPP1),B(IPP1)

        DO 6 I=1,N
        Y1(I)=XY1(I,IPP1)
6       CONTINUE

C       Cp IS APPLIED TO THE REDUCED DATA SET
        IDO=0
        NROW=N
        NVAR=IPP1
        LDX=N
        IFRQ=0
        IWT=0
        MOPT=0
        ICOPT=1
        LDCOV=IPP1
        LDINCD=1

        CALL DCORVC(IDO,NROW,NVAR,XY1,LDX,IFRQ,IWT,MOPT,ICOPT,XMEAN,COV,+
        LDCOV,INCD,LDINCD,NOBS,NMISS,SUMWT)

        CALL DRBEST(NVAR1,COV,LDCOV1,NOBS1,ICRIT1,NBEST1,NGOOD1,+
        IPRINT1,ICRITX1,CRIT1,IVARX1,INDVAR1,ICOEFX1,COEF1,LDCOEF1)

        IMIN=1
        AMIN=CRIT1(1)
        DO 10 I=2,IP
        IF (CRIT1(ICRITX1(I)).LT.AMIN) THEN
                AMIN=CRIT1(ICRITX1(I))
                IMIN=I
        ENDIF
10      CONTINUE
        IB=IVARX1(IMIN)
```

```
      DO 20 J=1,IMIN
      ITT=INDVAR1(IB+J-1)
      DO 15 I=1,N
      X2(I,J)=XY1(I,ITT)
15    CONTINUE
20    CONTINUE

      DO 21 J=1,IMIN
      ITT=INDVAR1(IB+J-1)
      X3(1,J)=XY(II,ITT)
21    CONTINUE

      NOBS=N
      NIND=IMIN
      LDX=N
      INTCEP=1
      CALL DRLSE(NOBS,Y1,NIND,X2,LDX,INTCEP,B,SST,SSE)

C     CALCULATE THE PREDICTED VALUES FORM THE MODEL FITTED TO THE REDUCED DATA SET
      DO 25 I=1,NN
      YKAP1(I)=B(1)
25    CONTINUE

      IF (II.EQ.1) THEN
              DO 28 J=1,IMIN
              YKAP1(1)=YKAP1(1)+X3(1,J)*B(J+1)
28            CONTINUE
              DO 32 I=1,N
              DO 30 J=1,IMIN
              YKAP1(I+1)=YKAP1(I+1)+X2(I,J)*B(J+1)
30            CONTINUE
32            CONTINUE
      ENDIF

      IF (II.EQ.NN) THEN
              DO 56 I=1,N
              DO 54 J=1,IMIN
              YKAP1(I)=YKAP1(I)+X2(I,J)*B(J+1)
54            CONTINUE
56            CONTINUE
              DO 58 J=1,IMIN
              YKAP1(NN)=YKAP1(NN)+X3(1,J)*B(J+1)
58            CONTINUE
      ENDIF

      IF ((II.NE.1).AND.(II.NE.NN)) THEN
              DO 62 I=1,II-1
              DO 60 J=1,IMIN
              YKAP1(I)=YKAP1(I)+X2(I,J)*B(J+1)
60            CONTINUE
62            CONTINUE

              DO 64 J=1,IMIN
              YKAP1(II)=YKAP1(II)+X3(1,J)*B(J+1)
64            CONTINUE

              DO 70 I=II+1,NN
              DO 65 J=1,IMIN
              YKAP1(I)=YKAP1(I)+X2(I-1,J)*B(J+1)
65            CONTINUE
70            CONTINUE
      ENDIF

      RETURN
      END
```

```
C       PROGRAM C6
C       THIS PROGRAM GIVES AS OUTPUT THE P-VALUES AND AVERAGE P-VALUES IN TABLE 5.1 FOR
C       THE HALD DATA.
C       THE ERROR VARIANCE IS ESTIMATED FORM THE MODEL FITTED TO THE FULL DATA SET.
C       THIS ESTIMATE IS ASSUMED TO BE THE KNOWN VALUE OF THE ERROR VARIANCE.

        USE MSIMSL
        IMPLICIT DOUBLE PRECISION (A-H,O-Z)
        PARAMETER (IP=4,NN=13,IPP1=IP+1,IPP2=IP+2,KOMB=10,NMOD=(2**IP)-1)
        DIMENSION XY(NN,IPP1),XX(NN,IP),XX1(NN,IPP1),Y(NN),XTX(IPP1,IPP1),XTXI(IPP1,IPP1),XXTXI(NN,IPP1)
        DIMENSION B(0:IP),XXTXIX(NN,NN),UW(NN),YKAP(NN),CP(NMOD,NN),ALAMDA(NMOD,NN)
        DIMENSION HAKIE(NMOD,NN),PWAARDES(NMOD,NN),PSOM(IP,NN),MODEL(NMOD,IPP2)

        CHARACTER*70 FILEIN
        CHARACTER*70 FILEOUT

        FILEIN='C:\HALD.TXT'
        FILEOUT='C:\OUTPUT.TXT'

C       READ THE DATA INTO XY
        OPEN(1,FILE=FILEIN)
        DO 1 I=1,NN
        READ(1,*) XY(I,IPP1),(XY(I,J),J=1,IP)
1       CONTINUE
        CLOSE(1)

C       THE VALUES OF THE PREDICTORS ARE PLACED IN XX
        DO 3 J=1,IP
         DO 2 I=1,NN
        XX(I,J)=XY(I,J)
2       CONTINUE
3       CONTINUE

C       THE RESPONSE VALUES ARE PLACED IN Y
        DO 5 I=1,NN
        Y(I)=XY(I,IPP1)
5       CONTINUE
C       CALCULATE THE MEAN SQUARED ERROR OF THE MODEL FITTED TO THE FULL DATA SET
        NOBS=NN
        NIND=IP
        LDX=NN
        INTCEP=1
        CALL DRLSE(NOBS,Y,NIND,XX,LDX,INTCEP,B,SST,SSE)
        SIGMA=SSE/(NN-IPP1)

C       CALCULATE THE DIAGONAL ELEMENTS OF THE PROJECTION MATRIX THAT CORRESPONDS TO
        THE COMPLETE SET OF PREDICTORS
        DO 10 I=1,NN
        XX1(I,1)=1.0D0
        DO 8 J=1,IP
        XX1(I,J+1)=XX(I,J)
8       CONTINUE
10      CONTINUE

        CALL DMXTXF(NN,NIND+1,XX1,NN,NIND+1,XTX,IPP1)
        CALL DLINDS(NIND+1,XTX,IPP1,XTXI,IPP1)
        DO 15 I=1,NN
        DO 14 J=1,NIND+1
        S=0.0D0
        DO 13 K=1,NIND+1
        S=S+XX1(I,K)*XTXI(K,J)
13      CONTINUE
        XXTXI(I,J)=S
14      CONTINUE
```

```
15    CONTINUE
      DO 18 I=1,NN
      DO 17 J=1,NN
      S=0.0D0
      DO 16 K=1,NIND+1
      S=S+XXTXI(I,K)*XX1(J,K)
16    CONTINUE
      XXTXIX(I,J)=S
17    CONTINUE
18    CONTINUE

      DO 25 I=1,NN
      UW(I)=XXTXIX(I,I)
      YKAP(I)=B(0)
      DO 20 J=1,NIND
      YKAP(I)=YKAP(I)+XX(I,J)*B(J)
20    CONTINUE
25    CONTINUE

C     THE SUBROUTINE THAT CALCULATES THE OBSERVED VALUE AND ESTIMATED NON-CENTRALITY
C     PARAMETER FOR THE ith CASE AND EVERY LINEAR SUBSPACE IS CALLED
      CALL ALLEMODELLE(NN,XX,Y,UW,YKAP,SIGMA,CP,ALAMDA,HAKIE)

      DO 50 I=1,NMOD
      DO 40 J=1,NN
C     THE NON-CENTRALITY PARAMETER IS TRUNCATED
      ALAMDA(I,J)=ALAMDA(I,J)
      IF (ALAMDA(I,J).LT.0.0D0) THEN
      ALAMDA(I,J)=0.0D0
      ENDIF

C     CALCULATE THE P-VALUES
      DF=1.0D0
      PWAARDES(I,J)=1.0D0-DCSNDF(HAKIE(I,J),DF,ALAMDA(I,J))
40    CONTINUE
50    CONTINUE

C     THE SUBROUTINE WHICH CALCULATES ALL POSSIBLE SUBSETS OF PREDICTORS IS CALLED
      CALL MODELINDEKS(MODEL)

C     CALCULATE THE AVERAGE P-VALUES
      DO 342 I=1,NN
      DO 341 J=1,IP
      PSOM(J,I)=0.0D0
      DO 340 K=1,NMOD
      DO 339 L=3,IPP2
      IF (MODEL(K,L).EQ.J) THEN
      PSOM(J,I)=PSOM(J,I)+PWAARDES(K,I)
      ENDIF
339   CONTINUE
340   CONTINUE
341   CONTINUE
342   CONTINUE

      DO 360 I=1,NN
      DO 350 J=1,IP
      PSOM(J,I)=PSOM(J,I)/((2**IP)/2)
350   CONTINUE
360   CONTINUE

      OPEN(1,FILE=FILEOUT)
      DO 600 NTELMOD=1,NMOD
      WRITE(1,605) NTELMOD,(PWAARDES(NTELMOD,I),I=1,NN)
600   CONTINUE
```

```
      WRITE(1,*)
      DO 601 J=1,IP
      WRITE(1,605) J,(PSOM(J,I),I=1,NN)
601   CONTINUE

      CLOSE(1)
605   FORMAT(I2,2X,50(F12.6,2X))

1000  STOP
      END

C     THE SUBROUTINE WHICH CALCULATES THE OBSERVED VALUE AND ESTIMATED NON-CENTRALITY
C     PARAMETER FOR THE ith CASE AND EVERY LINEAR SUBSPACE
C     RANDOM VARIABLE
      SUBROUTINE ALLEMODELLE(N,X,Y,UW,YKAP,SIGMA,CP,ALAMDA,HAKIE)
      USE MSIMSL
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      PARAMETER (IP=4,NN=13,IPP1=IP+1,KOMB=10,NMOD=(2**IP)-1)
      DIMENSION X(NN,IP),XX(NN,IP),Y(NN),B(0:IP),XX1(NN,IPP1)
      DIMENSION XTX(IPP1,IPP1),XTXI(IPP1,IPP1),XXTXI(NN,IPP1)
      DIMENSION XXTXIX(NN,NN),VW(NN),UW(NN),YKAP(NN),RW(NN),CP(NMOD,NN)
      DIMENSION ALAMDA(NMOD,NN),HAKIE(NMOD,NN),YKAP1(NN)

C     CONSIDER THE MODELS THAT INCLUDE 1 PREDICTOR VARIABLE
      NTELMOD=0

      DO 30 J1=1,IP
      NTELMOD=NTELMOD+1
      DO 10 I=1,N
      XX1(I,1)=1.0D0
      XX1(I,2)=X(I,J1)
      XX(I,1)=X(I,J1)
10    CONTINUE

      NIND=1
      NOBS=N
      LDX=NN
      INTCEP=1
      CALL DRLSE(NOBS,Y,NIND,XX,LDX,INTCEP,B,SST,SSE)

      CALL DMXTXF(N,NIND+1,XX1,NN,NIND+1,XTX,IPP1)
      CALL DLINDS(NIND+1,XTX,IPP1,XTXI,IPP1)
      DO 15 I=1,N
      DO 14 J=1,NIND+1
      S=0.0D0
      DO 13 K=1,NIND+1
      S=S+XX1(I,K)*XTXI(K,J)
13    CONTINUE
      XXTXI(I,J)=S
14    CONTINUE
15    CONTINUE
      DO 18 I=1,N
      DO 17 J=1,N
      S=0.0D0
      DO 16 K=1,NIND+1
      S=S+XXTXI(I,K)*XX1(J,K)
16    CONTINUE
      XXTXIX(I,J)=S
17    CONTINUE
18    CONTINUE

      DO 25 I=1,N
      VW(I)=XXTXIX(I,I)
      YKAP1(I)=B(0)
      DO 24 J=1,NIND
      YKAP1(I)=YKAP1(I)+XX(I,J)*B(J)
24    CONTINUE
      CP(NTELMOD,I)=(((YKAP1(I)-Y(I))**2)/SIGMA)+VW(I)-(1.0D0-VW(I))
      ALAMDA(NTELMOD,I)=(YKAP1(I)-YKAP(I))**2-(SIGMA*(UW(I)-VW(I)))
      ALAMDA(NTELMOD,I)=ALAMDA(NTELMOD,I)/(SIGMA*(1.0D0-VW(I)))
      HAKIE(NTELMOD,I)=((YKAP1(I)-Y(I))**2)/(SIGMA*(1.0D0-VW(I)))
25    CONTINUE
30    CONTINUE

C     CONSIDER THE MODELS THAT INCLUDE 2 PREDICTOR VARIABLES
      NTELMOD=NTELMOD+1

      DO 60 J1=1,IP-1
      DO 59 J2=J1+1,IP

      DO 35 I=1,N
      XX1(I,2)=X(I,J1)
      XX1(I,3)=X(I,J2)
      XX(I,1)=X(I,J1)
      XX(I,2)=X(I,J2)
35    CONTINUE

      NIND=2
      NOBS=N
      LDX=NN
      INTCEP=1
      CALL DRLSE(NOBS,Y,NIND,XX,LDX,INTCEP,B,SST,SSE)

      CALL DMXTXF(N,NIND+1,XX1,NN,NIND+1,XTX.IPP1)
      CALL DLINDS(NIND+1,XTX,IPP1,XTXI,IPP1)
      DO 40 I=1,N
      DO 39 J=1,NIND+1
      S=0.0D0
      DO 38 K=1,NIND+1
      S=S+XX1(I,K)*XTXI(K,J)
38    CONTINUE
      XXTXI(I,J)=S
39    CONTINUE
40    CONTINUE
      DO 44 I=1,N
      DO 43 J=1,N
      S=0.0D0
      DO 42 K=1,NIND+1
      S=S+XXTXI(I,K)*XX1(J,K)
42    CONTINUE
      XXTXIX(I,J)=S
43    CONTINUE
44    CONTINUE

      DO 50 I=1,N
      VW(I)=XXTXIX(I,I)
      YKAP1(I)=B(0)
      DO 49 J=1,NIND
      YKAP1(I)=YKAP1(I)+XX(I,J)*B(J)
49    CONTINUE
      CP(NTELMOD,I)=(((YKAP1(I)-Y(I))**2)/SIGMA)+VW(I)-(1.0D0-VW(I))
      ALAMDA(NTELMOD,I)=(YKAP1(I)-YKAP(I))**2-(SIGMA*(UW(I)-VW(I)))
      ALAMDA(NTELMOD,I)=ALAMDA(NTELMOD,I)/(SIGMA*(1.0D0-VW(I)))
      HAKIE(NTELMOD,I)=((YKAP1(I)-Y(I))**2)/(SIGMA*(1.0D0-VW(I)))
50    CONTINUE
59    CONTINUE
60    CONTINUE
```

```
C     CONSIDER THE MODELS THAT INCLUDE 3 PREDICTOR VARIABLES
      NTELMOD=NTELMOD+1
      DO 90 J1=1,IP-2
      DO 89 J2=J1+1,IP-1
      DO 88 J3=J2+1,IP

      DO 65 I=1,N
      XX1(I,2)=X(I,J1)
      XX1(I,3)=X(I,J2)
      XX1(I,4)=X(I,J3)
      XX(I,1)=X(I,J1)
      XX(I,2)=X(I,J2)
      XX(I,3)=X(I,J3)
65    CONTINUE

      NIND=3
      NOBS=N
      LDX=NN
      INTCEP=1
      CALL DRLSE(NOBS,Y,NIND,XX,LDX,INTCEP,B,SST,SSE)

      CALL DMXTXF(N,NIND+1,XX1,NN,NIND+1,XTX,IPP1)
      CALL DLINDS(NIND+1,XTX,IPP1,XTXI,IPP1)

      DO 70 I=1,N
      DO 69 J=1,NIND+1
      S=0.0D0
      DO 68 K=1,NIND+1
      S=S+XX1(I,K)*XTXI(K,J)
68    CONTINUE
      XXTXI(I,J)=S
69    CONTINUE
70    CONTINUE
      DO 78 I=1,N
      DO 77 J=1,N
      S=0.0D0
      DO 76 K=1,NIND+1
      S=S+XXTXI(I,K)*XX1(J,K)
76    CONTINUE
      XXTXIX(I,J)=S
77    CONTINUE
78    CONTINUE

      DO 86 I=1,N
      VW(I)=XXTXIX(I,I)
      YKAP1(I)=B(0)
      DO 84 J=1,NIND
      YKAP1(I)=YKAP1(I)+XX(I,J)*B(J)
84    CONTINUE
      CP(NTELMOD,I)=(((YKAP1(I)-Y(I))**2)/SIGMA)+VW(I)-(1.0D0-VW(I))
      ALAMDA(NTELMOD,I)=(YKAP1(I)-YKAP(I))**2-(SIGMA*(UW(I)-VW(I)))
      ALAMDA(NTELMOD,I)=ALAMDA(NTELMOD,I)/(SIGMA*(1.0D0-VW(I)))
      HAKIE(NTELMOD,I)=((YKAP1(I)-Y(I))**2)/(SIGMA*(1.0D0-VW(I)))
86    CONTINUE
88    CONTINUE
89    CONTINUE
90    CONTINUE

C     CONSIDER THE MODEL THAT INCLUDES ALL 4 PREDICTOR VARIABLES
      NTELMOD=NTELMOD+1

      DO 92 I=1,N
      XX1(I,2)=X(I,1)
      XX1(I,3)=X(I,2)
      XX1(I,4)=X(I,3)
      XX1(I,5)=X(I,4)
      XX(I,1)=X(I,1)
      XX(I,2)=X(I,2}
      XX(I,3)=X(I,3)
      XX(I,4)=X(I,4)
92    CONTINUE

      NIND=4
      NOBS=N
      LDX=NN
      INTCEP=1
      CALL DRLSE(NOBS,Y,NIND,XX,LDX,INTCEP,B,SST,SSE)

      CALL DMXTXF(N,NIND+1,XX1,NN,NIND+1,XTX,IPP1)
      CALL DLINDS(NIND+1,XTX,IPP1,XTXI,IPP1)

      DO 100 I=1,N
      DO 99 J=1,NIND+1
      S=0.0D0
      DO 98 K=1,NIND+1
      S=S+XX1(I,K)*XTXI(K,J)
98    CONTINUE
      XXTXI(I,J)=S
99    CONTINUE
100   CONTINUE
      DO 108 I=1,N
      DO 107 J=1,N
      S=0.0D0
      DO 106 K=1,NIND+1
      S=S+XXTXI(I,K)*XX1(J,K)
106   CONTINUE
      XXTXIX(I,J)=S
107   CONTINUE
108   CONTINUE

      DO 116 I=1,N
      VW(I)=XXTXIX(I,I)
      YKAP1(I)=B(0)
      DO 112 J=1,NIND
      YKAP1(I)=YKAP1(I)+XX(I,J)*B(J)
112   CONTINUE
      CP(NTELMOD,I)=(((YKAP1(I)-Y(I))**2)/SIGMA)+VW(I)-(1.0D0-VW(I))
      ALAMDA(NTELMOD,I)=(YKAP1(I)-YKAP(I))**2-(SIGMA*(UW(I)-VW(I)))
      ALAMDA(NTELMOD,I)=ALAMDA(NTELMOD,I)/(SIGMA*(1.0D0-VW(I)))
      HAKIE(NTELMOD,I)=((YKAP1(I)-Y(I))**2)/(SIGMA*(1.0D0-VW(I)))
116   CONTINUE

117   CONTINUE
118   CONTINUE
119   CONTINUE
120   CONTINUE

      RETURN
      END

C     THE SUBROUTINE THAT CALCULATES ALL POSSIBLE SUBSETS OF PREDICTORS
      SUBROUTINE MODELINDEKS(MODEL)
      USE MSIMSL
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      PARAMETER (IP=4,NN=13,IPP1=IP+1,NMOD=(2**IP)-1,IPP2=IP+2)
      DIMENSION MODEL(NMOD,IPP2)
```

```
C     CONSIDER THE MODELS THAT INCLUDE 1 PREDICTOR VARIABLE
      NTELMOD=0
      NIND=1
      DO 30 J1=1,IP
      NTELMOD=NTELMOD+1
      MODEL(NTELMOD,1)=NTELMOD
      MODEL(NTELMOD,2)=NIND
      MODEL(NTELMOD,3)=J1
30    CONTINUE

C     CONSIDER THE MODELS THAT INCLUDE 2 PREDICTOR VARIABLES
      NIND=2
      DO 60 J1=1,IP-1
      DO 59 J2=J1+1,IP
      NTELMOD=NTELMOD+1
      MODEL(NTELMOD,1)=NTELMOD
      MODEL(NTELMOD,2)=NIND
      MODEL(NTELMOD,3)=J1
      MODEL(NTELMOD,4)=J2
59    CONTINUE
60    CONTINUE

C     CONSIDER THE MODELS THAT INCLUDE 3 PREDICTOR VARIABLES
      NIND=3
      DO 90 J1=1,IP-2
      DO 89 J2=J1+1,IP-1
      DO 88 J3=J2+1,IP
      NTELMOD=NTELMOD+1
      MODEL(NTELMOD,1)=NTELMOD
      MODEL(NTELMOD,2)=NIND
      MODEL(NTELMOD,3)=J1
      MODEL(NTELMOD,4)=J2
      MODEL(NTELMOD,5)=J3
88    CONTINUE
89    CONTINUE
90    CONTINUE

C     CONSIDER THE MODEL THAT INCLUDES ALL 4 PREDICTOR VARIABLES
      NTELMOD=NTELMOD+1
      NIND=4
      MODEL(NTELMOD,1)=NTELMOD
      MODEL(NTELMOD,2)=NIND
      MODEL(NTELMOD,3)=1
      MODEL(NTELMOD,4)=2
      MODEL(NTELMOD,5)=3
      MODEL(NTELMOD,6)=4

      RETURN
      END
```

BIBLIOGRAPHY

Ahn, B.J. and Park, S.H. (1987). On simultaneous consideration of variable selection and detection of influential cases. *Journal of the Korean Statistical Society*, **16**, 10-20.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (eds. B.N. Petrov and F. Csàki). Akademia Kiadó, Budapest, 267-281.

Akaike, H. (1974). A new look at the statistical model identification. *I.E.E.E. Trans. Auto. Control*, **19**, 716-723.

Arnold, S.F. (1981). *The theory of linear models and multivariate analysis.* Wiley, New York.

Atkinson, A.C. (1994). Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, **89**, 1329-1339.

Atkinson, A.C. and Riani, M. (2000). *Robust diagnostic regression analysis*. Springer-Verlag, New York.

Baragona, R., Battaglia, F. and Calzini, C. (2001). Genetic algorithms for the identification of additive and innovation outliers in time series. *Computational Statistics & Data Analysis, **37**, 1-12.*

Barrett, B.E. and Gray, J.B. (1997). Leverage, residual, and interaction diagnostics for subsets of cases in least squares regression. *Computational Statistics & Data Analysis*, **26**, 39-52.

Becker, R.A., Chambers, J.M. and Wilks, A.R. (1988). *The new S language*. Wadsworth and Brooks Cole, California.

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics*. Wiley, New York.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, **87**, 738-754.

Burnham, P. and Nolan, D. (1995). A general Akaike-type criterion for model selection in robust regression. *Biometrika*, **82**, 877-886.

Burnham, K.P. and Anderson, D.R. (1998). *Model selection and inference. A practical information-theoretic approach*. Springer-Verlag, New York.

Carroll, R.J. and Ruppert, D. (1988). *Transformation and weighting in regression.* Chapman and Hall, New York.

Chatterjee, S. and Hadi, A.S. (1988). Impact of simultaneous omission of a variable and an observation on a linear regression equation. *Computational Statistics & Data Analysis,* **6**, 129-144.

Chiu, S-T. (2000). Test of significance in order selection. *Journal of Statistical Computation and Simulation*, **65**, 23-42.

Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**, 15-18.

Draper, N.R. and Smith, H. (1998). *Applied regression analysis (3rd edition)*. Wiley, New York.

Efron, B. and Tibshirani, R.J. (1993). *An introduction to the bootstrap*. Chapman and Hall, New York.

Freund, R.J. (1979). Multicollinearity etc.: Some "new" examples. Proceedings of Statistical Computing Section. *American Statistical Association*, 111-112.

Fung, W.K. (1992). Some diagnostic measures in discriminant analysis. *Statistical and Probability Letters*, **13**, 279-285.

Fung, W.K. (1995). Diagnostics in linear discriminant analysis. *Journal of the American Statistical Association*, **90**, 952-956.

Fung, W.K. (1996). Diagnosing influential observations in quadratic discriminant analysis. *Biometrics*, **52**, 1235-1241.

Glendinning, H.G. (2001). Selecting sub-set autoregressions from outlier contaminated data. *Computational Statistics & Data Analysis,* **36**, 179-207.

Gupta, S.S. and Huang, D-Y. (1988). Selecting important independent variables in linear regression models. *Journal of Statistical Planning and Inference,* **20**, 155-167. See also Erratum (1990), *Journal of Statistical Planning and Inference,* **24**, 269.

Gupta, S.S. and Huang, D-Y. (1996). On detecting influential data and selecting regression variables. *Journal of Statistical Planning and Inference*, **53**, 421-435.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized additive models.* Chapman and Hall, London.

Hoeting, J., Raftery, A.E. and Madigan, D. (1996). A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics & Data Analysis,* **22**, 251-270.

Hocking, R.R. (1974). Misspecification in regression. *The American Statistician*, **28**, 39-40.

Kim, C. and Hwang, S. (2000). Influential subsets on the variable selection. *Communications in Statistics: Theory and Methods*, **29**, 335-347.

Kim, S.S. and Park, S.H. (1995). Dynamic plots for displaying the roles of variables and observations in regression model. *Computational Statistics & Data Analysis,* **19**, 401-418.

Kundu, D. and Murali, G. (1996). Model selection in linear regression. *Computational Statistics & Data Analysis,* **22**, 461-469.

Léger, C. and Altman, N. (1993). Assessing influence in variable selection problems. *Journal of the American Statistical Association*, **88**, 547-556.

Li, B, Martin, E.B. and Morris, A.J. (2001). A graphical technique for detecting influential cases in regression analysis. *Communications in Statistics: Theory and Methods*, **30**, 463-483.

Linhart, H. and Zucchini, W. (1986). *Model selection.* Wiley, New York.

Liu, H., Weiss, R.E., Jennrich, R.I. and Wegner, N.S. (1999). PRESS model selection in repeated measures data. *Computational Statistics & Data Analysis,* **30**, 169-184.

Mallows, C.L. (1973). Some comments on Cp. *Technometrics*, **15**, 661-675.

Mallows, C.L. (1995). More Comments on Cp. *Technometrics*, **37**, 362-372.

Miller, A.J. (2002). *Subset selection in regression (2nd edition)*. Chapman and Hall, London.

Murtaugh, P.A. (1998). Methods of variable selection in regression modelling. *Communications in Statistics: Simulation and Computation*, **27**, 711-734.

Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1990). *Applied linear statistical models.* Irwin, Chicago.

Peixoto, J.L. and Lamotte, L.R. (1989). Simultaneous identification of outliers and predictors using variable selection techniques. *Journal of Statistical Planning and Inference*, **23**, 327-343.

Peña, D., Tia, G.C. and Tsay, R.S. (2001). *A course in time series analysis*. Wiley, New York.

Rancel, M.M.S. and Sierra, M.A.G. (2000). Procedures for the identification of multiple influential observations using local influence. *The Indian Journal of Statistics*, **62**, 135-143.

Ronchetti, E. (1997). Robustness aspects of model choice. *Statistica Sinica*, **7**, 327-338.

Ronchetti, E., Field, C. and Blanchard, W. (1997). Robust linear model selection by cross-validation. *Journal of the American Statistical Association*, **92**, 1017-1023.

Ronchetti, E. and Staudte, R.G. (1994). A robust version of Mallows's C_p. *Journal of the American Statistical Association*, **89**, 550-559.

Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust regression and outlier detection.* Wiley, New York.

Ryan, T.P. (1997). *Modern regression analysis.* Wiley, New York.

Saxena, K.M.L. and Alam, K. (1982). Estimation of the non-centrality parameter of a chi squared distribution. *The Annals of Statistics*, **10**, 1012-1016.

Shao, P.Y.S. and Strawderman, W.E. (1995). Improving on the positive part of the UMVUE of a noncentrality parameter of a noncentral chi-square distribution. *Journal of Multivariate Analysis*, **53**, 52-66.

Shi, P. and Tsai, C. (1998). A note on the unification of the Akaike information criterion. *Journal of the Royal Statistical Society B*, **60**, 551-558.

Snyman, J.L.J. (1994). *Model selection and estimation in multiple linear regression.* PhD thesis completed in the Department of Statistics and Operational Research at the Potchefstroom University for Christian Higher Education.

Sommer, S. and Huggins, R.M. (1996). Variable selection using the Wald test and robust C_p. *Journal of the Royal Statistical Society*, **45**, 15-29.

Sommer, S. and Staudte, R.G. (1995). Robust variable selection in regression in the presence of outliers and leverage points. *Australian Journal of Statistics*, **37**, 323-336.

Spjϕtvoll, E. (1977). Alternatives to plotting C_p in multiple regression. *Biometrika*, **64**, 1-8.

Steel, S.J. and Louw, N. (2001). Variable selection in discriminant analysis: measuring the influence of individual cases. *Computational Statistics & Data Analysis,* **37**, 249-260.

Van Vuuren, J.O. (1998). *Collinearity-influential observations and outliers in the multiple* *linear regression model.* PhD thesis completed in the Department of Statistics and Actuarial Science at Stellenbosch University.

Venter, J.H. and Snyman, J.L.J. (1994). Computationally intensive variable selection criteria. *COMPSTAT*, Vienna, 494-499.

Venter, J.H and Steel, S.J. (1990). Estimating risk reduction in Stein estimation. *The Canadian Journal of Statistics*, **18**, 221-232.

Weisberg, S. (1981). A statistic for allocating C_p to individual cases. *Technometrics*, **23**, 27-31.

Weisberg, S. (1985). *Applied Linear Regression (2nd edition).* Wiley, New York.

Wisnowski, J.W., Montgomery, D.C. and Simpson, J.R. (2001). A comparative analysis of multiple outlier detection procedures in the linear regression model. *Computational Statistics & Data Analysis,* **36**, 351-382.

Wu, Y. (2001). An M-estimation-based model selection criterion with data-oriented penalty. *Journal of Statistical Computation and Simulation*, **70**, 71-87.