LEAST SQUARES REGRESSION WITH THE MEAN SQUARE ERROR CRITERION

by

Subhash Chander Narula

A thesis submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Department of Industrial and Management
Engineering in the Graduate College of
The University of Iowa

August, 1971

Thesis supervisor: Associate Professor John S. Ramberg

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Subhash Chander Narula

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy in the Department of Industrial and Management Engineering at the August, 1971 graduation.

Thesis committee: John S Ramberg
Thesis supervisor

Fred C. Leone
Member

Henri L. Beenhakker.
Member

Member

James D Broffitt
Member

ACKNOWLEDGEMENTS

I wish to express my appreciation for the valuable guidance and assistance provided by Associate Professor John S. Ramberg during this research and my stay at the University of Iowa. I also extend special thanks to Assistant Professors James W. L. Cole and James D. Broffitt for their suggestions and encouragement, and to Professor Fred C. Leone and Associate Professor Henri L. Beenhakker for serving on my thesis committee.

Last, but by no means least, I wish to thank all those who have made this day possible--especially God.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND SUMMARY

In this thesis, we study modifications of the least squares prediction equation to improve the predictive mean square error, (variance + squared bias), of the equation.

Numerous planned and unplanned studies and experiments are concerned with finding an equation to predict a response, the value of which is usually not available at the time the prediction takes place. Often the investigator has at his disposal a limited number of independent observations on the predictor variables and the corresponding response variable (or the predictand). The problem of the investigator, then, is to derive on the basis of this limited set of data, an equation relating the response variable to the predictor variables. Least squares is a statistical technique which provides a method for estimating the unknown parameters in the prediction equation. This is referred to as regression analysis by many practitioners and although the precise definition of regression is more limited, we use this broader definition here. In the majority of applications of regression analysis, a linear regression equation is employed, i.e., an equation that involves random variables, fixed variables, and parameters and is linear in the parameters and the random variables.

A statement of the problem is given in the next section followed by a literature review and an outline of the present research.

1.1 Statement of the Problem

We denote the response variable by y, the set of k predictor variables by $z_1, z_2, \ldots, z_k$, and assume that the correct model is

$$(1.1) \qquad y = \alpha + \beta_1 z_1 + \beta_2 z_2 + \ldots + \beta_k z_k + \varepsilon,$$

where $\alpha, \beta_1, \beta_2, \ldots, \beta_k$ are the unknown parameters and ε, the random error. If we denote the least squares estimators[1] of $\beta_1, \beta_2, \ldots, \beta_k$ by $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$, the prediction equation can be written as

$$(1.2) \qquad \hat{y} = \bar{y} + \hat{\beta}_1(z_1 - \bar{z}_1) + \hat{\beta}_2(z_2 - \bar{z}_2) + \ldots + \hat{\beta}_k(z_k - \bar{z}_k),$$

where $\bar{y}, \bar{z}_1, \bar{z}_2, \ldots, \bar{z}_k$ represent the sample means of the response variable and the predictor variables respectively and y is the predicted response.

In many planned experiments, it is possible to fix or control the levels of the predictor variables, whereas in observational studies the values of the predictor variables often cannot be fixed or controlled, but are random. We treat these two problems separately.

First we consider the problem where all of the predictor variables are fixed or controllable (or non-stochastic). For this case

[1] The least squares estimator of α is devoted by $\hat{\alpha}$ and is given by $\hat{\alpha} = \bar{y} - \hat{\beta}_1\bar{z}_1 - \hat{\beta}_2\bar{z}_2 - \ldots - \hat{\beta}_k\bar{z}_k$.

we assume that $E(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$ (unknown) and that the observations are independent. Under the above conditions, the least squares estimators are the best (minimum-variance), linear (linear function of the y), unbiased estimators of the unknown parameters.

Next we study the problem where all of the predictor variables are stochastic (or random). In particular, we assume that the response variable and the predictor variables follow a (k + 1)-variate normal distribution with unknown mean vector and unknown covariance matrix so that $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \mathrm{Var}(y|\underline{z})$. This assumption seems reasonable in many practical problems and the results are mathematically tractable. Here we also assume that the observation vectors are independently identically distributed. Under the above assumptions, the least squares estimators (which are also the maximum likelihood estimators for this problem) are the best linear unbiased estimators (BLUE) of the unknown parameters of the model.

For both of these problems, our objective is to improve the predictive mean square error, $E(y - \hat{y})^2$, where y is the unknown value of the response variable for a future observation. We study various modifications of the prediction equation (1.2) for both of these problems, but do not consider problems where both controllable and random predictor variables are present.

The first modification we consider is to use a subset of the predictor variables to predict the unknown value of the response variable. If $z_1, z_2, \ldots, z_p$, the first p ($p \leq k$) predictor variables, are used in the prediction equation (No loss of generality is involved since the numbering of the variables is arbitrary.), the subset prediction equation is given by

$$(1.3) \qquad \tilde{y} = \bar{y} + \tilde{\beta}_1(z_1 - \bar{z}_1) + \tilde{\beta}_2(z_2 - \bar{z}_2) + \ldots + \tilde{\beta}_p(z_p - \bar{z}_p),$$

where $\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_p$ are the least squares estimators of $\beta_1, \beta_2, \ldots, \beta_p$ for the reduced equation. We show that the predictive mean square error (p.m.s.e.) of the prediction equation (1.3) can be smaller than that of (1.2) and obtain a decision rule to select the best subset of the predictor variables in the p.m.s.e. sense. We will refer to this as the subset approach.

In the second modification, which we term the lambda approach, we use all of the predictor variables and a multiplicative constant λ, such that the prediction equation is given by

$$(1.4) \qquad \hat{y}^{\dagger} = \bar{y} + \lambda\{\hat{\beta}_1(z_1 - \bar{z}_1) + \hat{\beta}_2(z_2 - \bar{z}_2) + \ldots + \hat{\beta}_k(z_k - \bar{z}_k)\},$$

where $0 \leq \lambda \leq 1$ and $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ are same as in (1.2). We show that for an appropriate choice of λ, (1.4) has smaller p.m.s.e. than (1.2). We also give an expression to calculate the value of λ and obtain its distribution.

In the third modification, (for the fixed predictor variables only), which we term the ridge approach, we use all of the predictor variables and the prediction equation is given by

$$(1.5) \qquad y^* = \bar{y} + \beta_1^*(z_1 - \bar{z}_1) + \beta_2^*(z_2 - \bar{z}_2) + \ldots + \beta_k^*(z_k - \bar{z}_k),$$

where β_1^*, β_2^*, ..., β_k^* represent the ridge estimators of β_1, β_2, ... β_k. (See Hoerl and Kennard [24, 25].) We suggest a method for finding the ridge estimates of the unknown parameters such that the p.m.s.e. of (1.5) is smaller than that of (1.2).

We also consider a number of combinations of the above modifications for the aforementioned problems.

1.2 Literature Review

Considerable literature on regression analysis already exists. This discussion is intended to provide the reader with some idea of the general nature of the previous work relevant to this thesis.

Before the advent of high speed electronic computers, much of the literature relating to multiple regression was concerned with finding methods for calculating the least squares estimators of the unknown parameters on desk calculators. Dwyer [13] lists 37 references for the years 1932-1941 which dealt essentially with the desk calculator computational techniques. Because of the costs involved in obtaining information on a large number of predictor variables and subsequently monitoring them, as well as to obtain a simple equation, the interest in the problem of selecting a suitable subset of predictor variables began. Cochran [10] was among the earliest researchers who discussed the problem of the deletion or the addition of a predictor variable to a multiple regression equation.

With the advent of the digital computers, numerous techniques were proposed to select the "best" subset regression equation. Some of the important ones are: (1) all possible regressions, (2) forward selection, (3) backward elimination, (4) stepwise regression, and (5) stagewise regression. Draper and Smith [12] give a thorough discussion including advantages and disadvantages of these methods. Bancroft [4] and Wallace [43] studied the bias introduced by the stepwise regression procedures, whereas Ashar [3], Freund, Vail and Clunies-Ross [16], Goldberger [18], and Goldberger and Jochem [19] studied it for the stagewise procedure.

Some of the more recent literature has been concerned with the development of computationally efficient algorithms which, unlike the procedures mentioned earlier, find the "best" subset regression equation in an economically feasible computer time. Among these is Garside [17] who gave an efficient method of generating all possible regression equations. His technique involves the idea that the subsequent subsets differ by one variable only. Schatzoff, Tsao and Fienberg [39] further improved this technique by operating on a minimal submatrix of the cross-product matrix at each step, and taking advantage of the inherent symmetry in the problem. Beale [6, 7] and Beale, Kendall and Mann [9] used a branch and bound method to find the best subset within a given size of a subset. Gorman and Toman [20] used fractional factorial designs for selecting a subset from the $2^k - 1$ regressions. Hocking and Leslie [23] developed a computationally efficient algorithm for calculating the "best" subset equation of all possible sizes. Their technique was further improved by LaMotte and Hocking [30], who also developed an algorithm to compute the "best" subset equation of each size and to give some other subset equations of each size which may be almost equally good. The efficiency of the technique lies in the fact that to find the best subset of the given size, all subsets of that size often need not be evaluated. More important, the unique feature of this algorithm is that it guarantees the "best" subset within a given size.

Recently Beale [8] and Mantel [35] have given a discussion of various methods with differing opinions as to which is the best technique. Longley [33] and Wampler [45] have discussed the accuracy of some often used computer programs for least squares regression.

Draper and Smith [12] discuss certain criteria to select the "best" subset equation, notable being: (1) minimum s^2 (the residual mean square error), and (2) maximum R^2, which is the ratio of the variability explained by the regression equation, namely, the regression sum of squares, to the total variability, namely, the total sum of squares. It is well known that R^2 is a non-decreasing function of the number of variables. To compare the subsets of various sizes, $\bar{R}^2$ was introduced which corrects R^2 for the number of variables in the equation (See Haitovsky [22]). This is given by

$$1 - \bar{R}^2 = n(1 - R^2) \ / \ (n - p + 1),$$

where n is the number of independent observations and p is the number of predictor variables in the equation. More recently Wiorkowski [49] showed that if the number of variables in the equation is a function of the sample size, then R^2 is not a consistent estimator of the population correlation coefficient ρ^2 (say). He suggested an alternative estimator P^2 where

$$P^2 = R^2 - p(1 - R^2) / (n - p + 1).$$

When selecting a subset within a given size, maximum R^2, $\bar{R}^2$, P^2 and regression sum of squares are all equivalent criteria, as also are minimum s^2, and the residual sum of squares. Mallows [34] introduced the C_p statistic, which is an estimator of the squared error (variance + squared bias) summed over all the n data points and is given by

$$C_p = RSS / \hat{\sigma}^2 - (n - 2p - 2)$$

where RSS is the residual sum of squares and $\hat{\sigma}^2$ is an unbiased estimator of the residual variance. Gorman and Toman [20] and Hocking and Leslie [23] used the C_p statistic in their search for the best subset. The statistic C_p can be used to find the "best" subset equation, though minimum C_p is equivalent to the criteria mentioned earlier when selecting the "best" subset within a given size. Recently, Walls and Weeks [44] have shown that the addition of a variable to the regression equation cannot decrease (and usually increases) the variance of the predicted response, although it can decrease the bias.

Webster [46] proved that the subset equation, though biased (see Larson and Bancroft [31] and Wallace [43]), can be a "better" predictor of the response than the full equation when one of the following criteria is used:

(a) $\tilde{y}$ will be said to be a "closer" predictor, (see Pitman [38]), of $\hat{y}$ than y if

$$P[|y - \tilde{y}| < |y - \hat{y}|] > \tfrac{1}{2}.$$

(b) $\tilde{y}$ will be said to be a better "K-neighborhood" predictor of y than $\hat{y}$ if

$$P[|y - \tilde{y}| < K] > P[|y - \hat{y}| < K].$$

Webster [46] noted that the major disadvantage in the use of the above criteria is that except for the case of large K in the K-neighborhood criterion, it is difficult to detect whether the condition is satisfied. This difficulty arises from protecting against the use of biased estimators when the bias may be excessive. Davies [11] discussed the choice of variables in the design of experiments for a linear fitted model. He used the mean square error, m.s.e., of the fitted linear model averaged over the spherical region $z_1^2 + z_2^2 + \ldots + z_k^2 \leq 1$, as the criterion for the inclusion of or the deletion of one variable from the design. Allen [1][2] used the mean square error of prediction as a criterion for selecting variables in subset regression.

[2] During the period in which this thesis was being written, this paper was delivered at the 130th Annual Meeting of the American Statistical Association, the Biometric Society, ENAR and WNAR in Detroit, Michigan. The problem that we consider in section 2.1 is similar to the one described in the paper.

As described in Draper and Smith [12] the partial F-test, which measures the contribution of the variable to the regression sum of squares as though it were added to the model last, is the criterion often used for the exclusion of or the inclusion of a variable from a multiple regression equation. Since the stepwise estimators of the unknown parameters are biased (see Bancroft [4] and Wallace [43]), and the partial F-test depends only on the variance of the estimator, Toro-Vizcarrondo and Wallace [42] used the m.s.e. of the estimator as the criterion and developed a uniformaly most powerful testing procedure for the criterion.

The above literature concerns mainly the problem where all the predictor variables are fixed or controllable. When the predictor variables are assumed to follow a multivariate normal distribution, the literature has mainly been concerned with finding the distribution and proving admissibility of the estimators of the unknown parameters in the model, although Lindley [32] discussed the selection of variables for this problem. His approach has been Bayesian.

Kerridge [28] gave an expression for the unconditional p.m.s.e. for the problem of random predictor variables and [29] suggested the use of a constant between zero and one to improve it.

Stein [41], for the problem of random predictor variables and Sclove [40], for the problem of fixed predictor variables, suggested "better" estimators of the unknown parameters than the usual least squares estimators in the m.s.e. sense, whenever there are three or more unknown parameters in the model. For the problem of fixed variables, Hoerl and Kennard [24, 25] developed ridge estimators which have smaller m.s.e. than the least squares estimators. Marquardt [36], in a recent paper, discussed the class of biased estimators of the unknown parameters employing generalized inverses and established the similarities among the generalized inverses, ridge estimates and the corresponding non-linear estimation procedures.

For the problem when all the predictor variables are fixed, the distribution theory of the least squares estimators of the unknown parameters is well known (see Graybill [21]). But for the problem when all the predictor variables are random the distribution theory is not well known. Kabe [27] derived the distribution of the least squares estimators of the unknown parameters in the model and Fisher [14], Moran [37], and Wilks [47] discussed the distribution of the sample multiple correlation coefficient R^2. Banerjee [5] derived an expression for $E(R^{2m})$, where m is a positive integer.

1.3 Summary

In Chapter 2, we study the problem with all of the predictor variables fixed. For this problem, we discuss the subset approach in Section 2.1 and obtain decision rules to select the "best" subset of the predictor variables. In Section 2.2, we discuss the lambda approach and give expressions to calculate the values of λ. Under the added assumption that the random errors are independently normally distributed, we derive the density function of $\hat{\lambda}$, the estimator of λ.

The ridge approach is studied in Section 2.3 and a method is suggested to calculate the ridge estimates. Various combinations of the above three approaches are studied in Section 2.4.

In Chapter 3, we study the problem where all of the predictor variables are random and follow a multivariate normal distribution. For this problem, we discuss the subset approach in Section 3.1 and obtain decision rules to select the "best" subset. In Section 3.2, we discuss the lambda approach and give expressions to calculate the values of λ. We also derive the density function of $\hat{\lambda}$, the estimator of λ. The subset-lambda approach is discussed in Section 3.3

Numerical examples are given in Chapter 4 followed by directions for future research in Chapter 5.

CHAPTER 2

NON-STOCHASTIC PREDICTOR VARIABLES

We let

$$\underline{z}_i = \begin{bmatrix} z_{i1} \\ z_{i2} \\ \cdot \\ \cdot \\ z_{ik} \end{bmatrix} = [z_{i1}, z_{i2}, \ldots, z_{ik}]' \ , \ (i = 1, 2, \ldots, n)$$

denote n independent vector observations on the k predictor variables and $y_1, y_2, \ldots, y_n$ denote the corresponding observations on the response variable. We define the vector of sample means by $\bar{\underline{z}} = \Sigma\underline{z}_i \ / \ n$ and $\bar{y} = \Sigma y_i \ / \ n$. For brevity, we let $\underline{x}_i$ denote the values of the predictor variables corrected for the sample means, i.e., $\underline{x}_i = \underline{z}_i - \bar{\underline{z}}$ for $(i = 1, 2, \ldots, n)$. The model (1.1) for each observation is then given by

$$(2.1) \qquad y_i = \alpha + \underline{x}_i'\underline{\beta} + \varepsilon_i,$$

where α and $\underline{\beta}$ (a k-component vector) are unknown parameters and ε_i represents the random error for an observation such that $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$ (unknown) for all i, and $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.

The least squares estimators (l.s.e.) of α and $\underline{\beta}$ minimize $\sum(y_i - \hat{y}_i)^2$, where $\hat{y}_i$ is the predicted value of the response variable, and are given by $\hat{\alpha} = \bar{y}$ and $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$, where

$$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \text{ and } X = \begin{bmatrix} \underline{x}_1' \\ \underline{x}_2' \\ \cdot \\ \cdot \\ \cdot \\ \underline{x}_n' \end{bmatrix} = \begin{bmatrix} x_{11}, & x_{12}, & \ldots, & x_{1k} \\ x_{21}, & x_{22}, & \ldots, & x_{2k} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{n1}, & x_{n2}, & \ldots, & x_{nk} \end{bmatrix}$$

Hence the least squares prediction equation is

(2.2) $$\hat{y} = \bar{y} + \underline{x}'\hat{\underline{\beta}}.$$

In many problems, interest centers on the prediction at $\underline{x}_0$ rather than the estimation of the unknown parameters. Hence using (2.2), the predicted response at $\underline{z}_0$ is given by $\hat{y}_0 = \bar{y} + \underline{x}_0'\hat{\underline{\beta}}$. We follow the convention of Johnston [26] and call a random variable $\hat{y}_0$ an unbiased predictor of another random variable y_0 whenever $E(y_0 - \hat{y}_0) = 0$. Since y_0 and $\hat{y}_0$ are independent, the p.m.s.e. is given by

(2.3) $$\begin{aligned} E(y_0 - \hat{y}_0)^2 &= Var(y_0) + Var(\hat{y}_0) \\ &= \{1 + 1/n + \underline{x}_0'(X'X)^{-1}\underline{x}_0\}\sigma^2. \end{aligned}$$

If we are interested in using (2.2) to predict at m points X_0 (where X_0 is a m x k matrix and each element of X_0 represents the value of a predictor variable corrected for its mean), the p.m.s.e. summed over each of the m points is given by

$$(2.4) \qquad \sum_{i=1}^{m} E(y_i - \hat{y}_i)^2 = \sum_{i=1}^{m} \{1 + 1/n + \underline{x}_i'(X'X)^{-1}\underline{x}_i\}\sigma^2$$

$$= [m + m/n + tr\{X_0(X'X)^{-1}X_0'\}]\sigma^2.$$

The p.m.s.e. summed over each of the n original data points X, as a special case of (2.4), is given by

$$(2.5) \qquad \sum_{i=1}^{n} E(y_i - \hat{y}_i)^2 = (1 + n + k)\sigma^2,$$

since m = n, X_0 is X and $tr\{X(X'X)^{-1}X'\} = tr\{(X'X)^{-1}X'X\} = tr(I_k) = k$.

We will now consider some modifications of the prediction equation (2.2) to improve the p.m.s.e. given by (2.3), (2.4) and (2.5).

2.1 The Subset Approach

We partition the k-component vector of predictor variables into two parts, $\underline{x}_i' = [\underline{x}_{i1}', \underline{x}_{i2}']$, where $\underline{x}_{i1}$ (a p-component vector) represents the set of p predictor variables included in the prediction equation and $\underline{x}_{i2}$ (a (p - k) - component vector), those not included. Accordingly we also partition $X = [X_1, X_2]$ and $\underline{\beta}' = [\underline{\beta}_1', \underline{\beta}_2']$ so that the subset prediction equation is given by

$$(2.6) \qquad \tilde{y}_i = \bar{y} + \underline{x}_{i1}'\tilde{\underline{\beta}}_1,$$

where $\tilde{\underline{\beta}}_1 = (X_1'X_1)^{-1}X_1'\underline{y}$ is the l.s.e. of $\underline{\beta}_1$ for the subset equation.

Using (2.6), the predicted response at $\underline{z}_0' = [\underline{z}_{01}', \underline{z}_{02}']$ is given by $\tilde{y}_0 = \bar{y} + \underline{x}_{01}'\tilde{\underline{\beta}}_1$. Although y_0 and $\tilde{y}_0$ are independent, $\tilde{y}_0$ is not an unbiased predictor of y_0 and hence the p.m.s.e. is given by

$$(2.7) \qquad E(y_0 - \tilde{y}_0)^2 = \mathrm{Var}(y_0) + \mathrm{Var}(\tilde{y}_0) + \{E(y_0 - \tilde{y}_0)\}^2$$

$$= \{1 + 1/n + \underline{x}_{01}'(X_1'X_1)^{-1}\underline{x}_{01}\}\sigma^2 + (\underline{x}_0'\underline{\beta} - \underline{x}_{01}'\underline{\Theta}_1)^2$$

$$= \{1 + 1/n + \underline{x}_{01}'(X_1'X_1)^{-1}\underline{x}_{01}\}\sigma^2$$

$$+ \{\underline{x}_{02}'\underline{\beta}_2 - \underline{x}_{01}'(X_1'X_1)^{-1}X_1'X_2\underline{\beta}_2\}^2,$$

where $\underline{\Theta}_1 = \underline{\beta}_1 + (X_1'X_1)^{-1}X_1'X_2\underline{\beta}_2$.

If we use (2.6) to predict at m points $X_0 = [X_{01}, X_{02}]$, the p.m.s.e. summed over each of these m points is given by

$$(2.8) \qquad \sum_{i=1}^{m} E(y_i - \tilde{y}_i)^2 = \sum_{i=1}^{m}[\{1 + 1/n + \underline{x}_{i1}'(X_1'X_1)^{-1}\underline{x}_{i1}\}\sigma^2$$

$$+ (\underline{x}_i'\underline{\beta} - \underline{x}_{i1}'\underline{\Theta}_1)^2]$$

$$= [m + m/n + \mathrm{tr}\{X_{01}(X_1'X_1)^{-1}X_{01}'\}]\sigma^2$$

$$+ \underline{\beta}_2'\{X_{02}' - X_2'X_1(X_1'X_1)^{-1}X_{01}'\} \cdot$$

$$\{X_{02} - X_{01}(X_1'X_1)^{-1}X_1'X_2\}\underline{\beta}_2.$$

The p.m.s.e. summed over each of the n original data points, which is a special case of (2.9), is given by

$$(2.9) \qquad \sum_{i=1}^{n} E(y_i - \tilde{y}_i)^2 = (1 + n + p)\sigma^2$$

$$+ \underline{\beta}_2' X_2' \{I_p - X_1(X_1'X_1)^{-1}X_1'\}X_2\underline{\beta}_2,$$

since $m = n$, X_0 is X and $\mathrm{tr}\{X_1(X_1'X_1)^{-1}X_1'\} = \mathrm{tr}\{(X_1'X_1)^{-1}X_1'X_1\} = \mathrm{tr}(I_p) = p$.

Since our criterion is the p.m.s.e. it would be desirable to use a subset rather than a full set whenever the p.m.s.e. of $\tilde{y}_0$ is less than or equal to the p.m.s.e. of $\hat{y}_0$, i.e., $E(y_0 - \tilde{y}_0)^2 \leq E(y_0 - \hat{y}_0)^2$, i.e., whenever

$$(2.10) \qquad \{1 + 1/n + \underline{x}_{01}'(X_1'X_1)^{-1}\underline{x}_{01}\}\sigma^2$$

$$+ \{\underline{x}_{02}'\underline{\beta}_2 - \underline{x}_{01}'(X_1'X_1)^{-1}X_1'X_2\underline{\beta}_2\}^2$$

$$\leq \{1 + 1/n + \underline{x}_0'(X'X)^{-1}\underline{x}_0\}\sigma^2,$$

or

$$\{\underline{x}_{02}'\underline{\beta}_2 - \underline{x}_{01}'(X_1'X_1)^{-1}X_1'X_2\underline{\beta}_2\}^2$$

$$\leq \{\underline{x}_0'(X'X)^{-1}\underline{x}_0 - \underline{x}_{01}'(X_1'X_1)^{-1}\underline{x}_{01}\}\sigma^2.$$

To gain some insight into the above result and to show that the inequality can actually be satisfied, we consider the case of two predictor variables. The model (2.1) simplifies to

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

Letting $s_{00} = \sum_i (y_i - \bar{y})^2 / (n - 1)$, $s_{ij} = \sum_\ell x_{i\ell} x_{j\ell} / (n - 1)$ $(i , j = 1, 2)$, $s_{0i} = \sum_j (y_j - \bar{y})x_{ij} / (n - 1)$ $(i = 1, 2)$ and $r_{ij} = s_{ij} / \sqrt{s_{ii}s_{jj}}$

(i = 0, 1, 2; j = 1, 2; i ≠ j), (2.10) reduces to $\beta_2^2 \le \sigma^2 / \{(n - 1)s_{22}(1 - r_{12}^2)\}$, which obviously can be true, even when $\beta_2^2 \ne 0$. Whenever this inequality holds the predicted response $\tilde{y}_0 = \bar{y} + \tilde{\beta}_1 x_{01}$ at $\underline{z}_0' = [z_{01}, z_{02}]$ will be better than $\hat{y}_0 = \bar{y} + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02}$ in the p.m.s.e. sense. Similarly, we would prefer to predict the response y_0 at $\underline{z}_0$ using the equation with x_2 alone whenever $\lambda_1^2 \le \sigma^2 / \{(n - 1)s_{11}(1 - r_{12}^2)\}$. If both of the above inequalities are satisfied, we would use the equation with smaller p.m.s.e. The result can be graphically presented as in Figure 1. This result demonstrates that a subset prediction equation can be better than the one with a full set of variables even though the sample R^2 is not as large.

Of course, the value of $\underline{\beta}_2$ and σ^2 is (2.10) are not known in a real application. An obvious decision rule for selecting the best subset of variables would be to select the subset that minimizes $\underline{x}_{01}'(X_1'X_1)^{-1}\underline{x}_{01}\sigma^2 + \{\underline{x}_{02}'\underline{\beta}_2 - \underline{x}_{01}'(X_1'X_1)^{-1}X_1'X_2\underline{\beta}_2\}^2$ for all possible subsets. In a practical situation we substitute the l.s.e. of $\underline{\beta}$ and σ^2 in the above expression and select the subset that minimizes $\underline{x}_{01}'(X_1'X_1)^{-1}\underline{x}_{01}\hat{\sigma}^2 + \{\underline{x}_{02}'\hat{\underline{\beta}}_2 - \underline{x}_{01}'(X_1'X_1)^{-1}X_1'X_2\hat{\underline{\beta}}_2\}^2$. This involves inversion of $2^k - 1$ matrices of the form $(X_1'X_1)$—one for each subset. Garside [17] and Schatzoff, Tsao and Fienberg [39] have given an efficient algorithm for these inversions.

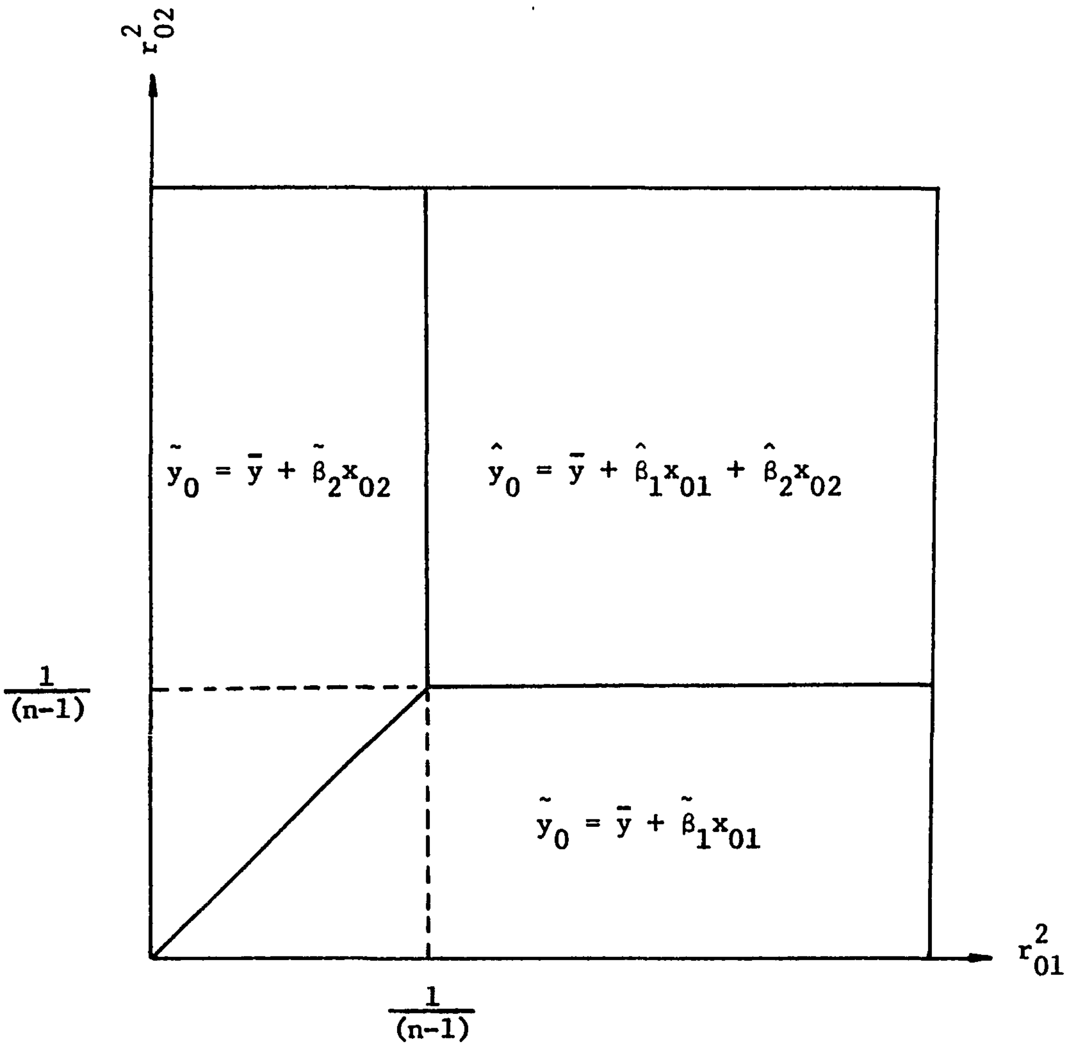
r_{02}^2
$\tilde{y}_0 = \bar{y} + \tilde{\beta}_2 x_{02}$
$\hat{y}_0 = \bar{y} + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02}$
$\frac{1}{(n-1)}$
$\tilde{y}_0 = \bar{y} + \tilde{\beta}_1 x_{01}$
r_{01}^2
$\frac{1}{(n-1)}$

FIGURE 1

TWO VARIABLE SUBSET SELECTION CRITERION

If we use (2.6) to predict at m points X_0, the condition equivalent to (2.10) is given by

$$(2.11) \qquad \underline{\beta}_2'\{X_{02}' - X_2'X_1(X_1'X_1)^{-1}X_{01}'\}\{X_{02} - X_{01}(X_1'X_1)^{-1}X_1'X_2\}\underline{\beta}_2$$
$$\leq \mathrm{tr}\{X_0(X'X)^{-1}X_0' - X_{01}(X_1'X_1)^{-1}X_{01}'\}\sigma^2.$$

Since it is the sum of m inequalities of the form (2.10), it can be satisfied. Here also, as in the previous case, we select the subset which minimizes the expression $\mathrm{tr}\{X_{01}(X_1'X_1)^{-1}X_{01}'\}\hat{\sigma}^2$ $- \hat{\underline{\beta}}_2'\{X_{02}' - X_2'X_1(X_1'X_1)^{-1}X_{01}'\}\{X_{02} - X_{01}(X_1'X_1)^{-1}X_1'X_2\}\hat{\underline{\beta}}_2$ for all $2^k - 1$ subsets.

When we use (2.6) to predict at n original data points X, (2.11) reduces to

$$(2.12) \qquad \underline{\beta}_2'X_2'\{I_p - X_1(X_1'X_1)^{-1}X_1'\}X_2\underline{\beta}_2 \leq (k - p)\sigma^2.$$

since $\mathrm{tr}\{X(X'X)^{-1}X'\} = k$, $\mathrm{tr}\{X_1(X_1'X_1)^{-1}X_1'\} = p$ and
$\underline{\beta}_2'\{X_{02}' - X_2'X_1(X_1'X_1)^{-1}X_{01}'\}\{X_{02} - X_{01}(X_1'X_1)^{-1}X_1'X_2\}\underline{\beta}_2$
$\underline{\beta}_2'X_2'\{I_p - X_1'(X_1'X_1)^{-1}X_1'\}X_2\underline{\beta}_2$.

For this case, the best subset minimizes $p\sigma^2 + \underline{\beta}_2'X_2'\{I_p - X_1(X_1'X_1)^{-1}X_1'\}X_2\underline{\beta}_2$. The C_p statistic developed by Mallows [34] is an estimate of $p + \beta_2'X_2'\{I_p - X_1(X_1'X_1)^{-1}X_1'\}X_2\underline{\beta}_2 \,/\, \sigma^2$. Hocking and Leslie [23] (also LaMotte and Hocking [30]) used the C_p statistic to find the best subset of each size. An obvious choice, therefore, is to use their algorithm and select the subset with minimum C_p.

2.2 The Lambda Approach

In this section, we study how by multiplying the prediction equation (2.2) by a constant $\lambda(0 \leq \lambda \leq 1)$, we can improve the p.m.s.e. The prediction equation is given by

$$(2.13) \qquad \hat{y}^{\dagger} = \bar{y} + \lambda \underline{x}'\underline{\hat{\beta}}.$$

Using (2.13), the predicted response at $\underline{z}_0$ is given by $\hat{y}_0^{\dagger} = \bar{y} + \lambda_{\underline{z}_0} \underline{x}_0'\underline{\hat{\beta}}$ (the subscript $\underline{z}_0$ of λ is to emphasize that λ depends upon $\underline{z}_0$). Although y_0 and $\hat{y}_0^{\dagger}$ are independent, $\hat{y}_0^{\dagger}$ is obviously not an unbiased predictor of y_0 and hence the p.m.s.e. is given by

$$\begin{aligned}(2.14) \qquad E(y_0 - \hat{y}_0^{\dagger})^2 &= \{1 + 1/n + \lambda_{\underline{z}_0}^2 \underline{x}_0'(X'X)^{-1}\underline{x}_0\}\sigma^2 \\ &\quad + (1 - \lambda_{\underline{z}_0})^2 \underline{x}_0'\underline{\beta}\underline{\beta}'\underline{x}_0 \\ &= \lambda_{\underline{z}_0}^2 \underline{x}_0'\{\underline{\beta}\underline{\beta}' + (X'X)^{-1}\sigma^2\}\underline{x}_0 \\ &\quad - 2\lambda_{\underline{z}_0} \underline{x}_0'\underline{\beta}\underline{\beta}'\underline{x}_0 + (1 + 1/n)\sigma^2 + \underline{x}_0'\underline{\beta}\underline{\beta}'\underline{x}_0,\end{aligned}$$

which is a quadratic in $\lambda_{\underline{z}_0}$ and reduces to (2.3) when $\lambda_{\underline{z}_0} = 1$. Since the coefficient of $\lambda_{\underline{z}_0}^2$ in (2.14) is always non-negative, the value of $\lambda_{\underline{z}_0}$ which minimizes $E(y_0 - \hat{y}_0^{\dagger})^2$ is given by

$$(2.15) \qquad \lambda_{\underline{z}_0} = \underline{x}_0'\underline{\beta}\underline{\beta}'\underline{x}_0 \,/\, \underline{x}_0'\{\underline{\beta}\underline{\beta}' + (X'X)^{-1}\sigma^2\}\underline{x}_0.$$

where $\underline{x}_0$ is a non-zero vector. When $\underline{x}_0$ is orthogonal to $\underline{\beta}$, $\lambda_{\underline{z}_0} = 0$. Since in a practical situation the values of $\underline{\beta}$ and σ^2 are not known, the value of $\lambda_{\underline{z}_0}$ cannot be calculated from (2.14). However, an estimate of $\lambda_{\underline{z}_0}$, $\hat{\lambda}_{\underline{z}_0}$ (say), can be obtained by using the l.s.e. of $\underline{\beta}$ and σ^2 in (2.15) so that

$$(2.16) \qquad \hat{\lambda}_{\underline{z}_0} = \underline{x}_0'\hat{\underline{\beta}}\hat{\underline{\beta}}'\underline{x}_0 \; / \; \underline{x}_0'\{\hat{\underline{\beta}}\hat{\underline{\beta}}' + (X'X)^{-1}\hat{\sigma}^2\}\underline{x}_0.$$

Since $\hat{\underline{\beta}}$ and $\hat{\sigma}^2$ are random variables, $\hat{\lambda}_{\underline{z}_0}$ is a random variable and the prediction equation (2.13) for any $\underline{z}_0$ is

$$(2.17) \qquad \hat{y}_0^{\dagger} = \bar{y} + \hat{\lambda}_{\underline{z}_0}\underline{x}_0'\hat{\underline{\beta}}.$$

If we assume that the random error ε is normally distributed, we obtain the distribution of $\hat{\lambda}_{\underline{z}_0}$ as

$$\hat{\lambda}_{\underline{z}_0} = \underline{x}_0'\hat{\underline{\beta}}\hat{\underline{\beta}}'\underline{x}_0 \; / \; \underline{x}_0'\{\hat{\underline{\beta}}\hat{\underline{\beta}}' + (X'X)^{-1}\hat{\sigma}^2\}\underline{x}_0,$$

and hence,

$$\hat{\lambda}_{\underline{z}_0} \; / \; (1 - \hat{\lambda}_{\underline{z}_0}) = \underline{x}_0'\hat{\underline{\beta}}\hat{\underline{\beta}}'\underline{x}_0 \; / \; \{\underline{x}_0'(X'X)^{-1}\underline{x}_0\hat{\sigma}^2\}.$$

Let $u = \underline{x}_0'\hat{\underline{\beta}}\hat{\underline{\beta}}'\underline{x}_0 \; / \; \{\underline{x}_0'(X'X)^{-1}\underline{x}_0\sigma^2\}$ and $v = (n - k - 1)\hat{\sigma}^2 / \sigma^2$, then $\hat{\lambda}_{\underline{z}_0} \; / \; (1 - \hat{\lambda}_{\underline{z}_0}) = (n - k - 1)u \; / \; v$.

Since $u \sim \chi_1^2(\delta)$, where $\delta = \underline{x_0'\beta\beta'}\underline{x}_0 \;/\; \{\underline{x}_0'(X'X)^{-1}\underline{x}_0\sigma^2\}$, and $v \sim \chi^2_{n-k-1}$, and u and v are stochastically independent, $\hat{\lambda}_{\underline{z}_0} \;/\; (1 - \hat{\lambda}_{\underline{z}_0}) \sim F_{1,n-k-1}(\delta)$.

Let $w = \hat{\lambda}_{\underline{z}_0} \;/\; (1 - \hat{\lambda}_{\underline{z}_0})$, then $\hat{\lambda}_{\underline{z}_0} = w \;/\; (1 + w)$.

$$0 < \omega < \infty \rightarrow \quad 0 < \hat{\lambda}_{\underline{z}_0} < 1, \text{ and}$$

$$|J| = |dw| / d\lambda_{\underline{z}_0}| = 1 \;/\; (1 - \hat{\lambda}_{\underline{z}_0})^2.$$

Since

$$f(\omega) = \begin{cases} \sum_{i=0}^{\infty} C_i \, \omega^{(2i - 1) / 2} \;/\; \{1 + \omega / (n - k - 1)\}^{(2i + n - k) / 2} & 0 < \omega < \infty \\ 0 & , \text{otherwise} \end{cases},$$

$$\text{where } C_i = \frac{\Gamma(\frac{2i + n - k}{2})}{\Gamma(\frac{n - k - 1}{2})\,\Gamma(\frac{2i + 1}{2})} \left(\frac{1}{n - k - 1}\right)^{(1 + 2i) / 2} \frac{\delta^i e^{-\delta}}{i!}$$

the density function g of $\hat{\lambda}_{\underline{z}_0}$ is given by,

$$g(\lambda) = \begin{cases} \sum_{i=0}^{\infty} C_i \dfrac{(\lambda)^{(2i - 1) / 2}\,(1 - \lambda)^{-(2i + 3) / 2}}{(1 + \frac{1}{n - k - 1}\frac{\lambda}{1 - \lambda})^{(2i + n - k) / 2}}, & 0 < \lambda < 1 \\ 0 & , \text{otherwise} \end{cases}.$$

Now, if we use the same λ, λ_m(say), to predict each of the m points X_0, the p.m.s.e. summed at each of the m points is given by

$$(2.18) \qquad \sum_{i=1}^{m} E(y_i - \hat{y}_i^{\dagger})^2 = [m + m/n + \lambda_m^2 \mathrm{tr}\{X_0(X'X)^{-1}X_0'\}]\sigma^2 + (1 - \lambda_m)^2\underline{\beta}'X_0'X_0\underline{\beta}$$

$$= \lambda_m^2[\underline{\beta}'X_0'X_0\underline{\beta} + \mathrm{tr}\{X_0(X'X)^{-1}X_0'\}\sigma^2] - 2\lambda_m\underline{\beta}'X_0'X_0\underline{\beta} + (m + m/n)\sigma^2 + \underline{\beta}'X_0'X_0\underline{\beta},$$

which is a quadratic in λ_m and reduces to (2.4) when $\lambda_m = 1$. Since the coefficient of λ_m^2 in (2.18) is always non-negative, the value of λ_m, which minimizes $\sum_{i=1}^{m} E(y_i - \hat{y}_i^{\dagger})^2$, is given by

$$(2.19) \qquad \lambda_m = \underline{\beta}'X_0'X_0\underline{\beta} \;/\; [\underline{\beta}'X_0'X_0\underline{\beta} + \mathrm{tr}\{X_0(X'X)^{-1}X_0'\}\sigma^2].$$

The above expression reduces to

$$(2.20) \qquad \lambda_n = \underline{\beta}'X'X\underline{\beta} \;/\; (\underline{\beta}'X'X\underline{\beta} + k\sigma^2),$$

when λ_n is used to predict each of the n original date points X, since X_0 is X and $\mathrm{tr}\{X(X'X)^{-1}X'\} = k$. In a real problem, the values of λ_m and λ_n cannot be calculated from (2.19) and (2.20) respectively. However, estimates $\hat{\lambda}_m$ and $\hat{\lambda}_n$(say) of λ_m and λ_n, can be obtained by using the l.s.e. of $\underline{\beta}$ and σ^2 in (2.19) and (2.20) respectively, so that

$$(2.21) \qquad \hat{\lambda}_m = \hat{\underline{\beta}}'X_0'X_0\hat{\underline{\beta}} \;/\; [\hat{\underline{\beta}}'X_0'X_0\hat{\underline{\beta}} + \mathrm{tr}\{X_0(X'X)^{-1}X_0'\}\hat{\sigma}^2]$$

and

$$(2.22) \qquad \hat{\lambda}_n = \underline{\hat{\beta}}'X'X\underline{\hat{\beta}} \;/\; (\underline{\hat{\beta}}'X'X\underline{\hat{\beta}} + k\hat{\sigma}^2).$$

If we assume that the random error ε is normally distributed, we obtain the distribution of $\hat{\lambda}_m$ as

$$\hat{\lambda}_m = \underline{\hat{\beta}}'X_0'X_0\underline{\hat{\beta}} \;/\; [\underline{\hat{\beta}}'X_0'X_0\underline{\hat{\beta}} + \mathrm{tr}\{X_0(X'X)^{-1}X_0'\}\hat{\sigma}^2],$$

and hence ,

$$\hat{\lambda}_m \;/\; (1 - \hat{\lambda}_m) = \underline{\hat{\beta}}'X_0'X_0\underline{\hat{\beta}} \;/\; [\mathrm{tr}\{X_0(X'X)^{-1}X_0'\}\hat{\sigma}^2].$$

Let $u = m\underline{\hat{\beta}}'X_0'X_0\underline{\beta} \;/\; [\mathrm{tr}\{X_0(X'X)^{-1}X_0\}\sigma^2]$ and $v = (n - k - 1)\hat{\sigma}^2 \;/\; \sigma^2$, then $\hat{\lambda}_m \;/\; (1 - \hat{\lambda}_m) = (n - k - 1)u \;/\; mv$.

Since $u \sim \chi_m^2(^2)$, where $\delta = m\underline{\beta}'X_0'X_0\underline{\beta} \;/\; [\mathrm{tr}\{X_0(X'X)^{-1}X_0\}\sigma^2]$, and $v \sim \chi^2_{n-k-1}$ and u and v are stochastically independent, $\hat{\lambda}_m \;/\; (1 - \hat{\lambda}_m) \sim F_{m,\; n-k-1}(\delta)$.

Let $w = \hat{\lambda}_m \;/\; (1 - \hat{\lambda}_m)$. Then $\hat{\lambda}_m = w \;/\; (1 + w)$.

$0 < w < \infty; < 0 < \hat{\lambda}_m < 1$, and

$|J| = |dw \;/\; d\lambda_m| = 1 \;/\; (1 - \lambda_m)^2.$

Since

$$f(\omega) = \begin{cases} \sum_{i=0}^{\infty} D_i \dfrac{w^{(2i + m - 2)}}{(1 + \frac{m}{n - k - 1}w)^{(2i + n + m - k - 1) / 2}}, & 0 < w < \infty \\ 0 & , \text{otherwise} \end{cases},$$

where

$$D_i = \frac{\Gamma(\frac{2i + n + m - k - 1}{2})}{\Gamma(\frac{n - k - 1}{2})\Gamma(\frac{2i + m}{2})} (\frac{m}{n - k - 1})^{(2i + m) / 2} \frac{\delta^i e^{-\delta}}{i!},$$

the density function h of λ_m is given by

$$(2.23) \qquad h(\lambda) = \begin{cases} \sum_{i=0}^{\infty} D_i \dfrac{\lambda^{(2i + m - 2) / 2} (1 - \lambda)^{-(2i + m + 2) / 2}}{\{1 + \frac{m}{(n - k - 1)} \frac{\lambda}{(1 - \lambda)}\}^{(2i + n + m - k - 1) / 2}}, & 0 < \lambda < 1 \\ 0 & , \text{otherwise.} \end{cases}$$

The density function of $\hat{\lambda}_n$ also is given by (2.23), with $m = k$ and $\delta = \underline{\beta}'X'X\underline{\beta} / \sigma^2$.

2.3 The Ridge Approach

We denote the sample correlation matrix of the predictor variables by V and the ridge estimators of $\underline{\beta}$ by $\underline{\beta}^*$, then the prediction equation can be written as

$$(2.24) \qquad y^* = \bar{y} + \underline{x\beta}^*.$$

where $\underline{\beta}^* = [V + hI_k]^{-1}X'y = WX'\underline{y}$, $(W = [V + hI_k]^{-1})$, $h \geq 0$.

Now the predicted response at $\underline{x}_0$ is given by $y_0^* = \bar{y} + \underline{x}_0'\underline{\beta}^*$. Although y_0 and y_0^* are independent, y_0^* is not an unbiased predictor of y_0 and hence the p.m.s.e. is given by

$$(2.25)\qquad E(y_0 - y_0^*)^2 = \{1 + 1/n + \underline{x}_0'M_{\underline{z}_0}(X'X)^{-1}M_{\underline{z}_0}'\underline{x}_0\}\sigma^2 + (\underline{x}_0'\underline{\beta} - \underline{x}_0'M_{\underline{z}_0}\underline{\beta})^2$$

$$= \{1 + 1/n + \underline{x}_0'(X'X)^{-1}\underline{x}_0\}\sigma^2$$

$$+ h_{\underline{z}_0}^2\underline{x}_0'W_{\underline{z}_0}\{\underline{\beta}\underline{\beta}' + (X'X)^{-1}\sigma^2\}W_{\underline{z}_0}'\underline{x}_0$$

$$- 2h_{\underline{z}_0}\underline{x}_0'(X'X)^{-1}W_{\underline{z}_0}'\underline{x}_0\sigma^2,$$

where $M_{\underline{z}_0} = [I_k + h_{\underline{z}_0}V^{-1}]^{-1} = I_k - h_{\underline{z}_0}W_{\underline{z}_0}$.

If $h_{\underline{z}_0}$ (the subscript $\underline{z}_0$ of h, W and M is to emphasize that h, W and M depend upon $\underline{z}_0$) is zero, (2.25) reduces to (2.3) and (2.24) to (2.2).

It would be preferable to use the ridge prediction equation rather than the least squares prediction equation whenever the p.m.s.e. of $\hat{y}_0$, i.e., whenever $E(y_0 - y_0^*)^2 \leq E(y_0 - \hat{y}_0)^2$, i.e., whenever

$$(2.26)\qquad h_{\underline{z}_0}\underline{x}_0'W_{\underline{z}_0}\{\underline{\beta}\underline{\beta}' + (X'X)^{-1}\sigma^2\}W_{\underline{z}_0}'\underline{x}_0 \leq 2\underline{x}_0'(X'X)^{-1}W_{\underline{z}_0}'\underline{x}_0\sigma^2,$$

which can trivally be satisfied for $h_{\underline{z}_0} = 0$.

As noted before, the value of $\underline{\beta}$ and σ^2 are not known in a real application. Also the value of $W_{\underline{z}_0}$ depends upon $h_{\underline{z}_0}$. An obvious decision rule to calculate the value of $h_{\underline{z}_0}$ is to choose that value of $h_{\underline{z}_0} \geq 0$ which maximizes

$2\underline{x}_0'(X'X)^{-1}W_{\underline{z}_0}'\underline{x}_0\sigma^2 - h_{\underline{z}_0}\underline{x}_0'W_{\underline{z}_0}\{\underline{\beta\beta}' + (X'X)^{-1}\sigma^2\}W_{\underline{z}_0}'\underline{x}_0$ subject to (2.26) to attain the maximum reduction in the p.m.s.e.

For each value of $h_{\underline{z}_0}$ this involves the inversion of the matrix $[V + h_{\underline{z}_0}I_k]$. In a practical situation we substitute the l.s.e. of $\underline{\beta}$ and σ^2, and select that value of $h_{\underline{z}_0} \geq 0$ which maximizes

$2\underline{x}_0'(X'X)^{-1}W_{\underline{z}_0}'\underline{x}_0\hat{\sigma}^2 - h_{\underline{z}_0}\underline{x}_0'W_{\underline{z}_0}\{\underline{\hat{\beta}\hat{\beta}}' + (X'X)^{-1}\hat{\sigma}^2\}W_{\underline{z}_0}'\underline{x}_0$ subject to

$h_{\underline{z}_0}\underline{x}_0'W_{\underline{z}_0}\{\underline{\hat{\beta}\hat{\beta}}' + (X'X)^{-1}\hat{\sigma}^2\}W_{\underline{z}_0}'\underline{x}_0 \leq 2\underline{x}_0(X'X)^{-1}W_{\underline{z}_0}'\underline{x}_0\hat{\sigma}^2$.

If we use the same value of h, h_m(say), to predict at m points X_0, the p.m.s.e. summed at each of the m points is given by

$$(2.27)\qquad \begin{aligned} \sum_{i=1}^{m}E(y_i - y_i^*)^2 &= [m + m/n + tr\{X_0M_m(X'X)^{-1}M_m'X_0'\}]\sigma^2 \\ &\quad + (X_0\underline{\beta} - X_0M_m\underline{\beta})^2 \\ &= [m + m/n + tr\{X_0(X'X)^{-1}X_0'\}]\sigma^2 \\ &\quad + h_m^2\underline{\beta}'W_m'X_0'X_0W_m\underline{\beta} \\ &\quad + h_m^2 tr\{X_0W_m(X'X)^{-1}W_m'X_0'\}\sigma^2 \\ &\quad - 2h_m tr\{X_0(X'X)^{-1}W_m'X_0'\}\sigma^2. \end{aligned}$$

The inequality equivalent to (2.26) is given by

$$(2.28)\qquad h_m[tr\{X_0W_m(X'X)^{-1}W_m'X_0'\}\sigma^2 + \underline{\beta}'W_m'X_0'X_0W_m\underline{\beta}] \leq 2tr\{X_0(X'X)^{-1}W_m'X_0'\}\sigma^2,$$

which is obviously satisfied for $h_m = 0$. Here also, as in the previous case, we select that value of $h_m \geq 0$ which maximizes

$2\text{tr}\{X_0(X'X)^{-1}W'_mX'_0\}\hat{\sigma}^2 - h_m[\text{tr}\{X_0W_m(X'X)^{-1}W'_mX'_0\}\hat{\sigma}^2 + \underline{\hat{\beta}}'W'_mX'_0X_0W_m\underline{\hat{\beta}}]$ subject to $h_m[\text{tr}\{X_0W_m(X'X)^{-1}W'_mX'_0\}\hat{\sigma}^2 + \underline{\hat{\beta}}'W'_mX'_0X_0W_m\underline{\hat{\beta}}] \leq 2\text{tr}\{X_0(X'X)^{-1}W'_mX'_0\}\hat{\sigma}^2$.

As a special case, the value of h, h_n(say), used to predict each of the n original data points, will maximize $2\text{tr}\{X(X'X)^{-1}W'_nX'\}\hat{\sigma}^2 - h_n[\text{tr}\{XW_n(X'X)^{-1}W'_nX'\}\hat{\sigma}^2 + \underline{\hat{\beta}}'W'_nX'XW_n\underline{\hat{\beta}}]$ subject to

$$(2.29) \qquad h_n[\text{tr}\{XW_n(X'X)^{-1}W'_nX'\}\hat{\sigma}^2 + \underline{\hat{\beta}}'W'_nX'XW_n\underline{\hat{\beta}}] \leq 2\text{tr}\{X(X'X)^{-1}W'_nX'\}\hat{\sigma}^2$$

since X_0 is X and W_m is W_n.

2.4 Combination Approaches

In this section, we study various combinations of the subset, the lambda, and the ridge approaches. When studying these combinations, we restrict ourselves to subset sizes which result in improved p.m.s.e. when the subset approach alone is used.

2.4.1 The Subset-Lambda Approach

When the subset and the lambda approaches are used together, the prediction equation is given by

$$(2.30) \qquad \tilde{y}_i^{\dagger} = \bar{y} + \lambda_1\underline{x}'_{i1}\tilde{\underline{\beta}}_1.$$

Using (2.30), the predicted response at $\underline{z}_0$ is given by $\tilde{y}_0^{\dagger} = \bar{y} + \lambda_{1\underline{z}_0}\underline{x}'_{01}\tilde{\underline{\beta}}_1$ (the subscript $\underline{z}_0$ of λ_1 is to emphasize that λ_1 depends upon $\underline{z}_0$) and the p.m.s.e. by

$$(2.31)\qquad \begin{aligned} E(y_0 - \tilde{y}_0^{\dagger})^2 &= (1 + 1/n + \lambda^2_{1\underline{z}_0}\underline{x}'_{01}(X_1'X_1)^{-1}\underline{x}_{01})\sigma^2 \\ &\quad + (\underline{x}_0'\underline{\beta} - \lambda_{1\underline{z}_0}\underline{x}'_{01}\underline{\Theta}_1)^2 \\ &= \lambda^2_{1\underline{z}_0}\underline{x}'_{01}\{\underline{\Theta}_1\underline{\Theta}_1' + (X_1'X_1)^{-1}\sigma^2\}\underline{x}_{01} \\ &\quad - 2\lambda_{1\underline{z}_0}\underline{x}_0'\underline{\beta\Theta}_1'\underline{x}_{01} + (1 + 1/n)\sigma^2 \\ &\quad + \underline{x}_0'\underline{\beta\beta}'\underline{x}_0. \end{aligned}$$

The p.m.s.e. is a quadratic in $\lambda_{1\underline{z}_0}$ and reduces to (2.7) when $\lambda_{1\underline{z}_0} = 1$. Since the coefficient of $\lambda^2_{1\underline{z}_0}$ in (2.31) is always non-negative, the value of $\lambda_{1\underline{z}_0}$ which minimizes $E(y_0 - \tilde{y}_0^{\dagger})^2$ is given by

$$\lambda_{1\underline{z}_0} = \underline{x}_0'\underline{\beta\Theta}_1'\underline{x}_{01} \;/\; \underline{x}'_{01}\{\underline{\Theta}_1\underline{\Theta}_1' + (X_1'X_1)^{-1}\sigma^2\}\underline{x}_{01},$$

where $\underline{x}_{01}$ is a non-zero vector. When $\underline{x}_0$ is orthogonal to $\underline{\beta}$, $\lambda_{1\underline{z}_0} = 0$. Nothing more can be said about the magnitude or sign of $\lambda_{1\underline{z}_0}$. The value of $\lambda_{1\underline{z}_0}$ cannot be calculated from (2.32) in a real application, but an estimate of $\lambda_{1\underline{z}_0}$, $\tilde{\lambda}_{1\underline{z}_0}$ (say), can be obtained by substituting the l.s.e. of $\underline{\beta}$ and σ^2 in (2.32) so that

$$(2.33)\qquad \tilde{\lambda}_{1\underline{z}_0} = \underline{x}_0'\hat{\underline{\beta}}\hat{\underline{\Theta}}_1'\underline{x}_{01} \;/\; x'_{01}\{\hat{\underline{\Theta}}_1\hat{\underline{\Theta}}_1' + (X'X)^{-1}\hat{\sigma}^2\}\underline{x}_{01},$$

where $\hat{\underline{\Theta}}_1 = \hat{\underline{\beta}}_1 + (X_1'X_1)^{-1}X_1'X_2\hat{\underline{\beta}}_2$. The prediction equation becomes

$$(2.34) \qquad \tilde{y}_i^{\dagger} = \bar{y} + \tilde{\lambda}_{1\underline{z}_1}\underline{x}_{i1}'\tilde{\underline{\beta}}_1.$$

We select that combination of $\tilde{\lambda}_{1\underline{z}_0}$ and the subset of predictor variables which minimizes $E(y_0 - \tilde{y}_0^{\dagger})^2$ when the l.s.e. of $\underline{\beta}$ and σ^2 are substituted in (2.31).

Now, if we use the same λ_1, λ_{1m}(say), to predict at m points X_0, the p.m.s.e. summed at each of the m points is given by

$$(2.35) \qquad \sum_{i=1}^{m}E(y_i - \tilde{y}_i^{\dagger})^2 = [m + m/n + \lambda_{1m}^2 tr\{X_{01}(X_1'X_1)^{-1}X_{01}'\}]\sigma^2$$

$$+ (X_0\underline{\beta} - \lambda_{1m}X_{01}\underline{\Theta}_1)^2$$

$$= \lambda_{1m}^2[\underline{\Theta}_1'X_{01}'X_{01}\underline{\Theta}_1 + tr\{X_{01}(X_1'X_1)^{-1}X_{01}'\}\sigma^2]$$

$$- 2\lambda_{1m}\underline{\beta}'X_0'X_{01}\underline{\Theta}_1 + (m + m/n)\sigma^2 + \underline{\beta}'X_0'X_0\underline{\beta}.$$

Since $\sum_{i=1}^{m}E(y_i - \tilde{y}_i^{\dagger})^2$ is a quadratic in λ_{1m} and the coefficient of λ_{1m}^2 is always non-negative, the value of λ_{1m} that minimizes (2.35), is given by

$$(2.35) \qquad \lambda_{1m} = \underline{\beta}'X_0'X_{01}\underline{\Theta}_1 \;/\; [\underline{\Theta}_1'X_{01}'X_{01}\underline{\Theta}_1 + tr\{X_{01}(X_1'X_1)^{-1}X_{01}'\}\sigma^2].$$

As in previous cases, an estimate of λ_{1m}, $\tilde{\lambda}_{1m}$(say), can be obtained by substituting the l.s.e. of $\underline{\beta}$ and σ^2 in (2.36), so that

$$(2.37) \qquad \tilde{\lambda}_{1m} = \hat{\underline{\beta}}'X_0'X_{01}\hat{\underline{\Theta}}_1 \;/\; [\hat{\underline{\Theta}}_1X_{01}'X_{01}\hat{\underline{\Theta}}_1 + tr\{X_{01}(X_1'X_1)^{-1}X_{01}'\}\sigma^2].$$

Once again, we select that particular combination of λ_{1m} and the subset of variables which minimizes

$$-(\hat{\underline{\beta}}'X_0'X_{01}\hat{\underline{\Theta}}_1)^2 \;/\; [\hat{\underline{\Theta}}_1'X_{01}'X_{01}\hat{\underline{\Theta}}_1 + \mathrm{tr}\{X_{01}(X_1'X_1)^{-1}X_{01}'\}\hat{\sigma}^2].$$

As a special case λ_{1n}, the value of λ_1 used to predict each of the n original data points is given by

$$(2.38) \qquad \lambda_{1n} = \underline{\beta}'X'X_1\underline{\Theta}_1 \;/\; (\underline{\Theta}_1'X_1'X_1\underline{\Theta}_1 + p\sigma^2),$$

since X_0 is X and $\mathrm{tr}\{X_1(X_1'X_1)^{-1}X_1'\} = p$. An estimate of λ_{1n} can be obtained by using the l.s.e. of $\underline{\beta}$ and σ^2 in (2.38) as

$$(2.38a) \qquad \tilde{\lambda}_{1n} = \hat{\underline{\beta}}'X'X_1\underline{\Theta}_1 \;/\; (\hat{\underline{\Theta}}_1'X_1'X_1\hat{\underline{\Theta}}_1 + p\sigma^2).$$

As in previous cases, we select that particular value of λ_{1n} and the subset of variables that minimizes $-(\hat{\underline{\beta}}'X'X_1\hat{\underline{\Theta}}_1)^2 \;/\; (\hat{\underline{\Theta}}_1'X_1'X_1\hat{\underline{\Theta}}_1 + p\hat{\sigma}^2)$.

2.4.2 The Subset-Ridge Approach

We partition the correlation matrix V of the predictor variables as follows:

$$V = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix},$$

where V_{11} is the correlation matrix of the variables included in the subset equation, V_{22} of those not included, and V_{12} (= V_{21}') between the variables included and those not included. The prediction equation is given by

$$(2.39) \qquad \tilde{y}^* = \bar{y} + \underline{x}_{01}'\tilde{\underline{\beta}}_1^*,$$

where $\underset{\sim}{\tilde{\beta}}{}_1^* = [V_{11} + h_1 I_p]^{-1} X_1'\underline{y} = W_1 X_1'\underline{y}$ are the ridge estimators of the reduced model ($W_1 = [V_{11} + h_1 I_p]^{-1}$), $h_1 \geq 0$.

The predicted response at $\underline{z}_0$ is given by $\tilde{y}_0^* = \bar{y} + \underline{x}_{01}'\underline{\tilde{\beta}}_1^*$, and the p.m.s.e. by

$$(2.40) \qquad E(y_0 - \tilde{y}_0^*)^2 = \{1 + 1/n + \underline{x}_{01}' M_{1\underline{z}_0} (X_1'X_1)^{-1} M_{1\underline{z}_0}' \underline{x}_{01}\}\sigma^2 + (\underline{x}_0'\underline{\beta} - \underline{x}_{01}' M_{1\underline{z}_0} \underline{\Theta}_1)^2,$$

where $M_{1\underline{z}_0} = I_p - h_{1\underline{z}_0} W_{1\underline{z}_0}$.

If $h_{1\underline{z}_0}$ (the subscript $\underline{z}_0$ of h_1, W_1, and M_1 is to emphasize that h_1, W_1, and M_1 depend upon $\underline{z}_0$) is zero, (2.39) reduces to (2.6) and (2.40) to (2.7).

Since our criterion is the p.m.s.e., it would be desirable to use the subset-ridge prediction equation rather than the least squares prediction equation whenever the p.m.s.e. of $\tilde{y}_0^*$ is less than or equal to the p.m.s.e. of $\hat{y}_0$, i.e., whenever $E(y_0 - \tilde{y}_0^*)^2 \leq E(y_0 - \hat{y}_0)^2$, i.e., whenever

$$(2.41) \qquad (\underline{x}_0'\underline{\beta} - \underline{x}_{01}' M_{1\underline{z}_0} \underline{\Theta}_1)^2 \leq \{\underline{x}_0'(X'X)^{-1}\underline{x}_0 - \underline{x}_{01}' M_{1\underline{z}_0} (X_1'X_1)^{-1} M_{1\underline{z}_0}' \underline{x}_{01}\}\sigma^2,$$

which can be trivially satisfied for $h_{1\underline{z}_0} = 0$ since it reduces to (2.10).

As stated before, the values of $\underline{\beta}$ and σ^2 are unknown in a real application. Also the value of $M_{1\underline{z}_0}$ depends upon $h_{1\underline{z}_0}$. An obvious decision rule would be to select that value of $h_{1\underline{z}_0}$ and the subset that minimizes $\underline{x}_{01}'M_{1\underline{z}_0}(X_1'X_1)^{-1}M_{1\underline{z}_0}'\underline{x}_{01}\sigma^2 + (\underline{x}_0'\underline{\beta} - \underline{x}_{01}'M_{1\underline{z}_0}\underline{\theta}_1)^2$ subject to (2.41) for all values of $h_{1\underline{z}_0}$ and all possible subsets. In a practical situation we substitute the l.s.e. of $\underline{\beta}$ and σ^2 in the above expressions and select the subset that minimizes $\underline{x}_{01}'M_{1\underline{z}_0}(X_1'X_1)^{-1}M_{1\underline{z}_0}'\underline{x}_{10}\hat{\sigma}^2 + (\underline{x}_0'\hat{\underline{\beta}} - \underline{x}_{01}'M_{1\underline{z}_0}\hat{\underline{\theta}}_1)^2$ subject to $(\underline{x}_0'\hat{\underline{\beta}} - \underline{x}_{01}'M_{1\underline{z}_0}\hat{\underline{\theta}}_1)^2 \leq \{\underline{x}_0'(X'X)^{-1}\underline{x}_0 - \underline{x}_{01}'M_{1\underline{z}_0}(X_1'X_1)^{-1}M_{1\underline{z}_0}'\underline{x}_{01}\}\hat{\sigma}^2$. This involves inversion of matrices of the form $(X_1'X_1)$--one for each subset and a few of the form $[V_{11} + h_{1\underline{z}_0}I_p]$ for each subset. Garside [17] and Schatzoff, Tsao and Fienberg [39] have given an efficient algorithm for these inversions.

If we use the same value of h_1, h_{1m}(say), and the same subset to predict at m points X_0, the p.m.s.e. summed at each of the m points is given by

$$(2.42) \qquad \sum_{i=1}^{m} E(y_i - \tilde{y}_i^*)^2 = [m + m/n + tr\{X_{01}M_{1m}(X_1'X_1)^{-1}M_{1m}'X_{01}'\}]\sigma^2 + (X_0\underline{\beta} - X_{01}M_{1m}\underline{\theta}_1)^2.$$

The inequality equivalent to (2.41) is given by

$$(2.43) \qquad (X_0\underline{\beta} - X_{01}M_{1m}\underline{\theta}_1)^2 \leq tr\{X_0(X'X)^{-1}X_0' - X_{01}M_{1m}(X_1'X_1)^{-1}M_{1m}'X_{01}'\}\sigma^2.$$

Since (2.42) reduces to (2.11) when $h_{1m} = 0$, it can be satisfied. Here also, as in the previous case, we select that value of

$h_{1m} \geq 0$, and the subset which minimizes $\mathrm{tr}\{X_{01}M_{1m}(X_1'X_1)^{-1}M_{1m}'X_{01}'\}\hat{\sigma}^2 + (X_0\hat{\underline{\beta}} - X_{01}M_{1m}\hat{\underline{\theta}}_1)^2$ subject to $(X_0\hat{\underline{\beta}} - X_{01}M_{1m}\hat{\underline{\theta}}_1)^2 \leq \mathrm{tr}\{X_0(X'X)^{-1}X_0' - X_{01}M_{1m}(X_1'X_1)^{-1}M_{1m}'X_{01}'\}\hat{\sigma}^2$.

As a special case, the value of h_1, h_{1n}(say), used to predict each of the *n* original data points and the subset minimize $\mathrm{tr}\{X_1M_{1n}(X_1'X_1)^{-1}M_{1n}'X_1'\}\hat{\sigma}^2 + (X\hat{\underline{\beta}} - X_1M_{1n}\hat{\underline{\theta}}_1)^2$ subject to $(X\hat{\underline{\beta}} - X_1M_{1n}\hat{\underline{\theta}}_1)^2 \leq [k - \mathrm{tr}\{X_1M_{1n}(X_1'X_1)^{-1}M_{1n}'X_1'\}]\hat{\sigma}^2$.

2.4.3 The Lambda-Ridge Approach

When the lambda and the ridge approaches are used together, the prediction equation is given by

$$(2.44) \qquad y^{*\dagger} = \bar{y} + \lambda\underline{x}'\underline{\beta}^*.$$

Using (2.44), the predicted response at $\underline{z}_0$ is given by $y_0^{*\dagger} = \bar{y} + \lambda_{\underline{z}_0}\underline{x}_0'\underline{\beta}^*$ (the subscript $\underline{z}_0$ of λ is to emphasize that λ depends upon $\underline{z}_0$), and the p.m.s.e. by

$$\begin{aligned}(2.45) \qquad E(y_0 - y_0^{*\dagger})^2 &= \{1 + 1/n + \lambda_{\underline{z}_0}^2\underline{x}_0'M_{\underline{z}_0}(X'X)^{-1}M_{\underline{z}_0}'\underline{x}_0\}\sigma^2 \\ &\quad + (\underline{x}_0'\underline{\beta} - \lambda_{\underline{z}_0}\underline{x}_0'M_{\underline{z}_0}\underline{\beta})^2 \\ &= \lambda_{\underline{z}_0}^2\underline{x}_0'M_{\underline{z}_0}\{\underline{\beta}\underline{\beta}' + (X'X)^{-1}\sigma^2\}M_{\underline{z}_0}'\underline{x}_0 - 2\lambda_{\underline{z}_0}\underline{x}_0'\underline{\beta}\underline{\beta}'M_{\underline{z}_0}'\underline{x}_0 \\ &\quad + (1 + 1/n)\sigma^2 + \underline{x}_0'\underline{\beta}\underline{\beta}'\underline{x}_0,\end{aligned}$$

which reduces to (2.14) when $h_{\underline{z}_0} = 0$, to (2.25) when $\lambda_{\underline{z}_0} = 1$, and to (2.3) when $h_{\underline{z}_0} = 0$ and $\lambda_{\underline{z}_0} = 1$. Since $E(y_0 - y_0^{*\dagger})^2$ is a quadratic in $\lambda_{\underline{z}_0}$ and the coefficient of $\lambda_{\underline{z}_0}^2$ is always non-negative, the value of $\lambda_{\underline{z}_0}$ which minimizes it, is given by

$$(2.46) \qquad \lambda_{\underline{z}_0} = \underline{x}_0'M_{\underline{z}_0}\underline{\beta}\underline{\beta}'\underline{x}_0 \; / \; [\underline{x}_0'M_{\underline{z}_0}\{\underline{\beta}\underline{\beta}' + (X'X)^{-1}\sigma^2\}M_{\underline{z}_0}'\underline{x}_0].$$

where $\underline{x}_0$ is non-zero vector. Since in a practical situation that values of $\lambda_{\underline{z}_0}$, $\underline{\beta}$ and σ^2 are not known, the value of $\lambda_{\underline{z}_0}$ cannot be calculated from (2.46). However an estimate of $\lambda_{\underline{z}_0}$, $\lambda_{\underline{z}_0}^*$ (say), can be obtained for each value of $h_{\underline{z}_0}$ by using the l.s.e. of $\underline{\beta}$ and σ^2 in (2.46) so that

$$(2.47) \qquad \lambda_{\underline{z}_0}^* = \underline{x}_0'M_{\underline{z}_0}\hat{\underline{\beta}}\hat{\underline{\beta}}'\underline{x}_0 \; / \; [\underline{x}_0'M_{\underline{z}_0}\{\hat{\underline{\beta}}\hat{\underline{\beta}}' + (X'X)^{-1}\hat{\sigma}^2\}M_{\underline{z}_0}'\underline{x}_0].$$

Then we select the value of $h_{\underline{z}_0}$ for which

$-(\underline{x}_0'M_{\underline{z}_0}\hat{\underline{\beta}}\hat{\underline{\beta}}'\underline{x}_0)^2 \; / \; \underline{x}_0M_{\underline{z}_0}\{\hat{\underline{\beta}}\hat{\underline{\beta}}' + (X'X)^{-1}\hat{\sigma}^2\}M_{\underline{z}_0}'\underline{x}_0$ is minimum.

The prediction equation becomes

$$(2.48) \qquad y^{*\dagger} = \bar{y} + \lambda_{\underline{z}_0}^*\underline{x}_0'\underline{\beta}^*.$$

Now if we use the same λ, λ_m(say) and the same h, h_m(say) to predict at m points X_0, the p.m.s.e. summed at each of the m points is given by

$$(2.49)\qquad \sum_{i=1}^{m} E(y_i - y_i^{*\dagger})^2 = [m + m/n + \lambda_m^2 \mathrm{tr}\{X_0 M_m (X'X)^{-1} M_m' X_0'\}]\sigma^2$$

$$+ (X_0\underline{\beta} - \lambda_m X_0 M_m \underline{\beta})^2$$

$$= \lambda_m^2[\underline{\beta}' M_m' X_0' X_0 M_m \underline{\beta} + \mathrm{tr}\{X_0 M_m (X'X)^{-1} M_m' X_0'\}\sigma^2]$$

$$- 2\lambda_m \underline{\beta}' X_0' X_0 M_m \underline{\beta} + (m + m/n)\sigma^2 + \underline{\beta}' X_0' X_0 \underline{\beta},$$

which reduces to (2.18) when $h_m = 0$, to (2.27) when $\lambda_m = 1$, and to (2.4) when $h_m = 0$ and $\lambda_m = 1$. Since $\sum_{i=1}^{m} E(y_i - y_i^{*\dagger})^2$ is a quadratic in λ_m and the coefficient of λ_m^2 is always non-negative, the value of λ_m which minimizes it is given by

$$(2.50)\qquad \lambda_m = \underline{\beta}' X_0' X_0 M_m \underline{\beta} \;/\; [\underline{\beta}' M_m' X_0' X_0 M_m \underline{\beta} + \mathrm{tr}\{X_0 M_m (X'X)^{-1} M_m' X_0'\}\sigma^2].$$

Here also, as in the previous case, we obtain an estimate of λ_m, λ_m^*(say), by substituting the l.s.e. of $\underline{\beta}$ and σ^2 in (2.50) and select that value of h_m for which $-(\hat{\underline{\beta}}' X_0' X_0 M_m \hat{\underline{\beta}})^2 \;/\; [\hat{\underline{\beta}}' M_m' X_0' X_0 M_m \hat{\underline{\beta}} + \mathrm{tr}\{X_0 M_m (X'X)^{-1} M_m' X_0'\}\hat{\sigma}^2]$ is minimum.

As a special case λ_n, the value of λ used to predict each of the n original data points is given by

$$(2.51)\qquad \lambda_n = \underline{\beta}' X' X M_n \underline{\beta} \;/\; [\underline{\beta}' M_n' X' X M_n \underline{\beta} + \mathrm{tr}\{X M_n (X'X)^{-1} M_n' X'\}\sigma^2],$$

since X_0 is X. We obtain an estimate of λ_n, λ_n^*(say), by substituting the l.s.e. of $\underline{\beta}$ and σ^2 in (2.51) and select that value of h_n for which $-(\hat{\underline{\beta}}' X' X M_n \hat{\underline{\beta}})^2 \;/\; [\hat{\underline{\beta}}' M_n' X' X M_n \hat{\underline{\beta}} + \mathrm{tr}\{X M_n (X'X)^{-1} M_n' X'\}\hat{\sigma}^2]$ is minimum.

2.4.4 The Subset-Lambda-Ridge Approach

When the subset, the lambda, and the ridge approaches are used together, the prediction equation is given by

$$(2.52)\qquad \tilde{y}_i^{*\dagger} = \bar{y} + \lambda_1 \underline{x}'_{i1}\underline{\tilde{\beta}}_1^{*}.$$

Using (2.52), the predicted response at $\underline{z}_0$ is given by $\tilde{y}_0^{*\dagger} = \bar{y} + \lambda_{1\underline{z}_0}\underline{x}'_{01}\underline{\tilde{\beta}}_1^{*}$, and the p.m.s.e. by

$$(2.53)\qquad \begin{aligned} E(y_0 - \tilde{y}_0^{*\dagger})^2 &= \{1 + 1/n + \lambda^2_{1\underline{z}_0}\underline{x}'_{01}M_{1\underline{z}_0}(X_1'X_1)^{-1}M'_{1\underline{z}_0}\underline{x}_{01}\}\sigma^2 \\ &\quad + (\underline{x}'_0\underline{\beta} - \lambda_{1\underline{z}_0}\underline{x}'_{01}M_{1\underline{z}_0}\underline{\Theta}_1)^2 \\ &= \lambda^2_{1\underline{z}_0}\underline{x}'_{01}M_{1\underline{z}_0}\{\underline{\Theta}_1\underline{\Theta}'_1 + (X_1'X_1)^{-1}\sigma^2\}M'_{1\underline{z}_0}\underline{x}_{01} \\ &\quad - 2\lambda_{1\underline{z}_0}\underline{x}'_0\underline{\beta}\underline{\Theta}'_1M'_{1\underline{z}_0}\underline{x}_{01} + (1 + 1/n)\sigma^2 + \underline{x}'_0\underline{\beta}\underline{\beta}'\underline{x}_0. \end{aligned}$$

Since $E(y_0 - \tilde{y}_0^{*\dagger})^2$ is a quadratic in $\lambda_{1\underline{z}_0}$ and the coefficient of $\lambda^2_{1\underline{z}_0}$ is always non-negative, the value of $\lambda_{1\underline{z}_0}$ which minimizes it is given by

$$(2.54)\qquad \lambda_{1\underline{z}_0} = \underline{x}'_0\underline{\beta}\underline{\Theta}'_1M_{1\underline{z}_0}\underline{x}_{01} \,/\, [\underline{x}'_{01}M_{1\underline{z}_0}\{\underline{\Theta}_1\underline{\Theta}'_1 + (X_1'X_1)^{-1}\sigma^2\}M'_{1\underline{z}_0}\underline{x}_{01}],$$

where $\underline{x}_{01}$ is a non-zero vector. In a practical situation, for any value of $h_{1\underline{z}_0}$ and a given subset, we estimate $\lambda_{1\underline{z}_0}$ by substituting the l.s.e. of $\underline{\beta}$ and σ^2 in (2.54). As a decision rule, we select $h_{1\underline{z}_0}$, $\tilde{\lambda}^*_{1\underline{z}_0}$ and

the subset of predictor variables for which
$-(\underline{x}_0'\hat{\underline{\beta}}\hat{\underline{\Theta}}_1'M_{1\underline{z}_0}\underline{x}_{01})^2 \,/\, \underline{x}_{01}'M_{1\underline{z}_0}\{\hat{\underline{\Theta}}_1\hat{\underline{\Theta}}_1' + (X_1'X_1)^{-1}\hat{\sigma}^2\}M_{1\underline{z}_0}'\underline{x}_{01}$ is minimum.

When the same value of h, h_{1m}(say), the same value of λ, λ_{1m}(say), and the same subset is used to predict at m points X_0, the p.m.s.e. summed at each of the m points is given by

$$(2.55)\qquad \Sigma_{i=1}^{m}E(y_i - \tilde{y}_i^{*\dagger})^2 = [m + m/n + \lambda_{1m}^2\mathrm{tr}\{X_{01}M_{1m}(X_1'X_1)^{-1}M_{1m}'X_{01}'\}]^{\ 2}$$

$$+ (X_0\underline{\beta} - \lambda_{1m}X_{01}M_{1m}\underline{\Theta}_1)^2$$

$$= \lambda_{1m}^2[\underline{\Theta}_1'M_{1m}'X_{01}'X_{01}M_{1m}\underline{\Theta}_1$$

$$+ \mathrm{tr}\{X_{01}M_{1m}(X'X\ )^{-1}M_{1m}'X_{01}'\}]\sigma^2$$

$$- 2\lambda_{1m}\underline{\beta}'X_0'X_{01}M_{1m}\underline{\Theta}_1 + (m + m/n)\sigma^2 + \underline{\beta}'X_0'X_0\underline{\beta},$$

which is a quadratic in λ_{1m} and the coefficient of λ_{1m}^2 is always non-negative. Hence the value of λ_{1m} which minimizes it is given by

$$(2.56)\qquad \lambda_{1m} = \underline{\beta}'X_0'X_{01}M_{1m}\underline{\Theta}_1 \,/$$

$$[\underline{\Theta}_1'M_{1m}'X_{01}'X_{01}M_{1m}\underline{\Theta}_1 + \mathrm{tr}\{X_{01}M_{1m}(X'X\ )^{-1}M_{1m}'X_{01}'\}\sigma^2].$$

As in the previous case, we estimate λ_{1m} by substituting the l.s.e. of $\underline{\beta}$ and σ^2 in (2.56). We select h_{1m}, $\tilde{\lambda}_{1m}^{*}$ and the subset of predictor variables which minimizes
$-(\hat{\underline{\beta}}'X_0'X_0M_{1m}\hat{\underline{\Theta}}_1)^2 \,/\, [\hat{\underline{\Theta}}_1'M_{1m}'X_{01}'X_{01}M_{1m}\hat{\underline{\Theta}}_1 + \mathrm{tr}\{X_{01}M_{1m}(X_1'X_1)^{-1}M_{1m}'X_{01}'\}\hat{\sigma}^2].$

As a special case the value of λ_{1n}, used to predict at each of the n original data points, is given by

$$(2.57) \qquad \lambda_{1n} = \underline{\beta}'X'X_1M_{1n}\underline{\Theta}_1 \,/\, [\underline{\Theta}_1'M_{1n}'X_1'X_1M_{1n}\underline{\Theta}_1 + \mathrm{tr}\{X_1M_{1n}(X_1'X_1)^{-1}M_{1n}'X_1'\}\sigma^2].$$

As before we estimate λ_{1n} by substituting the l.s.e. of $\underline{\beta}$ and σ^2 in (2.57) and select h_{1n}, $\tilde{\lambda}^*_{1n}$ and the subset of predictor variables which minimizes

$$-(\hat{\underline{\beta}}'X'X_1'M_{1n}\hat{\underline{\Theta}}_1)^2 \,/\, [\hat{\underline{\Theta}}_1'M_{1n}'X_1'X_1M_{1n}\hat{\underline{\Theta}}_1 + \mathrm{tr}\{X_1M_{1n}(X_1'X_1)^{-1}M_{1n}'X_1'\}\hat{\sigma}^2].$$

CHAPTER 3

STOCHASTIC PREDICTOR VARIABLES

In this Chapter, we consider the problem where the response variable and the predictor variables have a joint (k + 1)-variate normal distribution with unknown mean $\underline{\mu}^* = [\mu_0, \mu_1, \ldots, \mu_k]' = [\mu_0, \underline{\mu}']'$ and unknown covariance matrix

$\Sigma^* = \begin{bmatrix} \sigma_{00} & \underline{\sigma}' \\ & \\ \underline{\sigma} & \Sigma \end{bmatrix}$. For this problem, if the analysis is carried out conditioned on the sample used to estimate the unknown parameters of the model, the results of Chapter 2 apply. In this Chapter, we develop the results when we take expectation over the sample used to estimate the parameters. As in Chapter 2, we let $\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_n$ by n independent (k - component vector) observations on the predictor variables and $y_1, y_2, \ldots, y_n$ by the corresponding observations on the response variable. We define the sample means $\bar{\underline{z}} = \ddagger\underline{z}_i / n$ and $\bar{y} = \ddagger y_i / n$ and $\underline{x}_i = \underline{z}_i - \bar{\underline{z}}$, i.e., the value of the predictor variables corrected for sample means. In addition, we denote the matrix of observations on the predictor variables by Z and the matrix of observations on the predictor variables corrected for the sample means by X so that

$$Z = \begin{bmatrix} \underline{z}_1' \\ \underline{z}_2' \\ \cdot \\ \cdot \\ \cdot \\ \underline{z}_n' \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12}, & \ldots, & z_{1k} \\ z_{21} & z_{22}, & \ldots, & z_{2k} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ z_{n1} & z_{n2}, & \ldots, & z_{nk} \end{bmatrix} \text{ and } X = \begin{bmatrix} \underline{x}_1' \\ \underline{x}_2' \\ \cdot \\ \cdot \\ \cdot \\ \underline{x}_n' \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12}, & \ldots, & x_{1k} \\ x_{21} & x_{22}, & \ldots, & x_{2k} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{n1} & x_{n2}, & \ldots, & x_{nk} \end{bmatrix}$$

We define the sample covariance matrix $S^* = \begin{bmatrix} s_{00} & \underline{s}' \\ \underline{s} & S \end{bmatrix}$, where

$s_{00} = \sum(y_i - \bar{y})^2 / (n - 1)$, $\underline{s} = \sum(y_i - \bar{y})\underline{x}_i / (n - 1)$ and S = $S = \sum\underline{x}_i\underline{x}_i' / (n - 1)$. Since for any given vector $\underline{z}_i$ of the predictor variable, $E(y_i|\underline{z}_i) = \mu_0 + \underline{\sigma}'\Sigma^{-1}(\underline{z}_i - \underline{\mu})$, the model (1.1) becomes

$$(3.1) \qquad y_i = \alpha + (\underline{z}_i - \underline{\mu})'\underline{\beta} + \varepsilon_i, \qquad (i = 1, 2, \ldots, n)$$

where $\alpha = \mu_0 - \underline{\sigma}'\Sigma^{-1}\underline{\mu}$, $\underline{\beta} = \Sigma^{-1}\underline{\sigma}$ and ε_i is random error such that $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = Var(y_i|\underline{z}_i) = \sigma_{00} - \underline{\sigma}'\Sigma^{-1}\underline{\sigma} = \sigma_k^2$ and for each observation is independently normally distributed. Under the above assumptions, the least squares prediction equation is given by

$$(3.2) \qquad \hat{y}_i = \bar{y} + \underline{x}_i'\hat{\underline{\beta}},$$

where $\hat{\underline{\beta}} = S^{-1}\underline{s}$. As noted in Chapter 2, we are interested in the problem of prediction at $\underline{z}_0$ rather than the estimation of the unknown parameters in the model. Using (3.2), the predicted response at $\underline{z}_0$ is given by $\hat{y}_0 = \bar{y} + \underline{x}_0'\hat{\underline{\beta}}$ (where $\underline{x}_0 = \underline{z}_0 - \bar{\underline{z}}$) and the conditional predictive mean square error by

$$(3.3)\qquad \begin{aligned} E\{(y_0 - \hat{y}_0)^2 | \underline{z}_0\} &= E\{(\underline{x}_0'\underline{\beta} + \varepsilon_0 - \bar{\varepsilon} - \underline{x}_0'\hat{\underline{\beta}})^2 | \underline{z}_0\} \\ &= \underline{\beta}' E(\underline{x}_0\underline{x}_0' | \underline{z}_0)\underline{\beta} + E\{(\varepsilon_0 - \bar{\varepsilon})^2 | \underline{z}_0\} \\ &\quad + E\{(\underline{x}_0'\hat{\underline{\beta}})^2 | \underline{z}_0\} - 2\underline{\beta}' E(\underline{x}_0\underline{x}_0'\hat{\underline{\beta}} | \underline{z}_0). \end{aligned}$$

Since $E(\varepsilon_0) = E(\bar{\varepsilon}) = 0$ and ε's are independent for each observation, by Lemmas A1, A3 and A7, (3.3) becomes

$$(3.3a)\qquad \begin{aligned} E\{(y_0 - \hat{y}_0)^2 | \underline{z}_0\} &= \sigma_k^2(1 + 1 / n) \\ &\quad + \sigma_k^2\{(\underline{z}_0 - \underline{\mu})'\Sigma^{-1}(\underline{z}_0 - \underline{\mu}) + k / n\} / (n - k - 2). \end{aligned}$$

The unconditional p.m.s.e. can be obtained by taking the expectation of the conditional p.m.s.e. over $\underline{z}_0$ and is given by

$$(3.4)\qquad E(y_0 - \hat{y}_0)^2 = \sigma_k^2(1 + 1 / n)(n - 2) / (n - k -2)$$

We now consider some modification of (3.2) to improve the conditional and the unconditional p.m.s.e.

3.1 The Subset Approach

As in Chapter 2, we partition the k-component vector of predictor variables into two parts, $\underline{z}_i' = [\underline{z}_{i1}', \underline{z}_{i2}']$, where $\underline{z}_{i1}$ (a p-component vector) represents the set of p predictor variables included in the prediction equation and $\underline{z}_{i2}$(a (k - p) - component vector), those not included. Accordingly we also partition

$\underline{x}_i' = [\underline{x}_{i1}', \underline{x}_{i2}']$, $Z = [Z_1, Z_2]$, $X = [X_1, X_2]$, $\underline{\mu}' = [\underline{\mu}_1', \underline{\mu}_2']$,

$\bar{\underline{z}}' = [\bar{\underline{z}}_1', \bar{\underline{z}}_2']$, $\underline{\sigma}' = [\underline{\sigma}_1', \underline{\sigma}_2']$, $\underline{s}' = [\underline{s}_1', \underline{s}_2']$,

$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ and $S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$ so that the subset prediction equation is given by

$$(3.5) \qquad \tilde{y}_i = \bar{y} + \underline{x}_{i1}'\tilde{\underline{\beta}}_1,$$

where $\tilde{\underline{\beta}}_1 = S_{11}^{-1}\underline{s}_1$.

Using the subset prediction equation, the predicted response at $\underline{z}_0' = [\underline{z}_{01}', \underline{z}_{02}']$ is given by $\tilde{y}_0 = \bar{y} + \underline{x}_{01}'\tilde{\underline{\beta}}_1$ (where $\underline{x}_{01} = \underline{z}_{01} - \bar{\underline{z}}_1$), and the conditional p.m.s.e. by

$$(3.6) \qquad E\{(y_0 - \tilde{y}_0)^2|\underline{z}_0\} = E\{(\underline{x}_0'\underline{\beta} + \varepsilon_0 - \bar{\varepsilon} - \underline{x}_{01}'\tilde{\underline{\beta}}_1)^2|\underline{z}_0\}.$$

By Lemmas A1, A3, A7 and A9, (3.6) becomes

$$\begin{aligned}(3.6a) \qquad E\{(y_0 - \tilde{y}_0)^2|\underline{z}_0\} &= \sigma_k^2 + \sigma_p^2 / n \\ &+ \sigma_p^2\{(\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1) + p / n\} / (n - p - 2) \\ &+ \{(\underline{z}_0 - \underline{\mu})'\underline{\beta} - (\underline{z}_{01} - \underline{\mu}_1)'\underline{\Phi}_1\}^2 \\ &= \sigma_k^2 + \sigma_p^2 / n \\ &+ \sigma_p^2\{(\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1) + p / n\} / (n - p - 2) \\ &+ \{(\underline{z}_{02} - \underline{\mu}_2)'\underline{\beta}_2 - (\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2\}^2,\end{aligned}$$

where $\sigma_p^2 = \sigma_{00} - \underline{\sigma}_1'\Sigma_{11}^{-1}\underline{\sigma}_1$ and $\underline{\Phi}_1 = \underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2$.

By taking the expectation of the conditional p.m.s.e. over $\underline{z}_0$, we obtain the unconditional p.m.s.e. as

$$(3.7) \qquad E(y_0 - \tilde{y}_0)^2 = \sigma_p^2(1 + 1/n)(n - 2) \;/\; (n - p - 2).$$

As before, our objective is to improve the p.m.s.e. We would prefer to predict the response at $\underline{z}_0$ using the subset prediction equation rather than the full equation, whenever the conditional p.m.s.e. of the subset equation is less than or equal to the conditional p.m.s.e. of the full equation, i.e., whenever $E\{(y_0 - \tilde{y}_0)^2|\underline{z}_0\} \leq E\{(y_0 - \hat{y}_0)^2|\underline{z}_0\}$, i.e., whenever

$$\begin{aligned}(3.8) \qquad & \sigma_k^2 + \sigma_p^2 \;/\; n + \sigma_p^2\{(\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1) + p \;/\; n\} \;/\; (n - p - 2) \\ & + \{(\underline{z}_{02} - \underline{\mu}_2)'\underline{\beta}_2 - (\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2\}^2 \\ & \leq \sigma_k^2(1 + 1 \;/\; n) + \\ & + \sigma_k^2\{(\underline{z}_0 - \underline{\mu})'\Sigma^{-1}(\underline{z}_0 - \underline{\mu}) + k \;/\; n\} \;/\; (n - k - 2),\end{aligned}$$

or

$$\begin{aligned} & \{(\underline{z}_{02} - \underline{\mu}_2)'\underline{\beta}_2 - (\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2\}^2 \\ & \leq \sigma_k^2 \;/\; n + \sigma_k^2\{(\underline{z}_0 - \underline{\mu})'\Sigma^{-1}(\underline{z}_0 - \underline{\mu}) + k \;/\; n\} \;/\; (n - k - 2) \\ & - \sigma_p^2 \;/\; n - \sigma_p^2\{(\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1) + p \;/\; n\} \;/\; (n - p - 2).\end{aligned}$$

For better understanding of the above result, we consider the case of two predictor variables. The notation simplifies to

$\underline{\sigma} = [\sigma_{01}', \sigma_{02}']'$, $\underline{\mu} = [\mu_1', \mu_2']'$ and $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$. Also $\underline{z}_{i1} = z_{i1}$, $\underline{x}_{i1} = x_{i1}$, $\underline{\sigma}_i = \sigma_{0i}$ and $\underline{\mu}_i = \mu_i$ for $i = 1, 2$ and $\Sigma_{ij} = \sigma_{ij}$ $(i, j = 1, 2)$. We define $\rho_{ij} = \sigma_{ij} / \sqrt{\sigma_{ii}\sigma_{jj}}$ $(i = 0, 1, 2; j = 1, 2)$, then (3.8) reduces to $\rho_{02}^2 / (1 - \rho_{01}^2) \leq 1 / (n - 3)$, where $n > 4$ and $\rho_{12} = 0$. Which can be true. Hence we use the equation with x_1 alone whenever the above inequality is satisfied. Similarly, we use the equation with x_2 alone whenever $\rho_{01}^2 / (1 - \rho_{02}^2) \leq 1 / (n - 3)$, where $n > 4$ and $\rho_{12} = 0$. If both the inequalities are satisfied, we use the equation with smaller p.m.s.e. For $\rho_{12} = 0$, the above result may be graphically presented as in Figure 2.

Now an obvious decision rule would be to select the subset that minimizes $E\{(y_0 - \tilde{y}_0)^2 | \underline{z}_0\}$ for all possible subsets. Since in a practical situation the values of $\underline{\mu}$ and Σ^* are unknown, we substitute the least squares estimators of $\underline{\mu}$ and Σ^* in (3.6a) and select the subset that minimizes

$$\hat{\sigma}_p^2 / n + \hat{\sigma}_p^2(\underline{x}_{01}'S_{11}^{-1}\underline{x}_{01} + p / n) / (n - p - 2) + (\underline{x}_0'\hat{\underline{\beta}} - \underline{x}_{01}'\underline{\Phi}_1)^2,$$

where $\hat{\sigma}_p^2 = s_{00} - \underline{s}_1'S_{11}^{-1}\underline{s}_1$ and $\hat{\underline{\Phi}}_1 = \hat{\underline{\beta}}_1 + S_{11}^{-1}S_{12}\hat{\underline{\beta}}_2$. This involves $2^k - 1$ matrix inversions--one for each subset. As noted in Chapter 2, the algorithm given by Garside [17] and Shatzoff, Tsao and Fienberg [39] may be used.

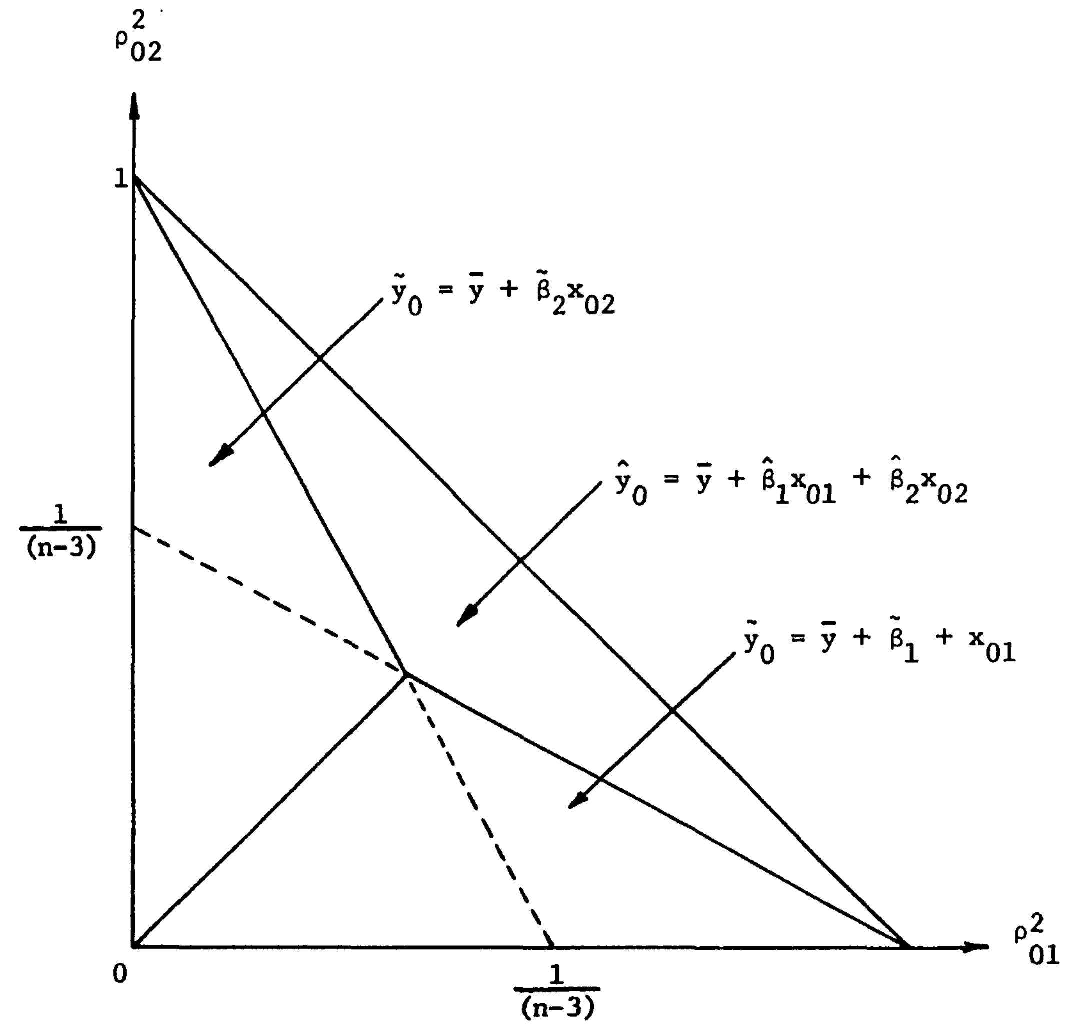
ρ^2_{02}
1
$\frac{1}{(n-3)}$
$\tilde{y}_0 = \bar{y} + \tilde{\beta}_2 x_{02}$
$\hat{y}_0 = \bar{y} + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02}$
$\tilde{y}_0 = \bar{y} + \tilde{\beta}_1 + x_{01}$
0
$\frac{1}{(n-3)}$
ρ^2_{01}

FIGURE 2

TWO VARIABLE SUBSET SELECTION CRITERION

If we want to improve the unconditional p.m.s.e., the condition equivalent to (3.8) is given by

$$(3.9)\qquad \sigma_p^2 \;/\; \sigma_k^2 \leq (n - p - 2) \;/\; (n - k - 2), \qquad n > k + 2.$$

Here also as in the previous case, we select the subset that minimizes $\hat{\sigma}_p^2 \;/\; (n - p - 2)$ subject to $\hat{\sigma}_p^2 \;/\; \hat{\sigma}_k^2 \leq (n - p - 2) \;/\; (n - k - 2)$, where $\hat{\sigma}_k^2 = s_{00} - \underline{s}'S^{-1}\underline{s}$.

3.2 The Lambda Approach

In this section, we study how by multiplying the prediction equation (3.2) by a constant $\lambda(0 \leq \lambda \leq 1)$, we can improve the p.m.s.e. The prediction equation is given by

$$(3.10)\qquad \hat{y}_i^{\dagger} = \bar{y} + \lambda\underline{x}_i'\hat{\underline{\beta}}.$$

Now the predicted response at $\underline{z}_0$ is given by $\hat{y}_0^{\dagger} = \bar{y} + \lambda_{\underline{z}_0}\underline{x}_0'\hat{\underline{\beta}}$ (the subscript $\underline{z}_0$ of λ is to emphasize that λ depends upon $\underline{z}_0$). Although y_0 and $\hat{y}_0^{\dagger}$ are independent, $\hat{y}_0^{\dagger}$ is obviously not an unbiased predictor of y_0 and hence the conditional p.m.s.e. is given by

$$(3.11)\qquad E\{(y_0 - \hat{y}_0^{\dagger})^2 | \underline{z}_0\} = E\{(\underline{x}_0'\underline{\beta} + \varepsilon_0 - \bar{\varepsilon} - \lambda_{\underline{z}_0}\underline{x}_0'\hat{\underline{\beta}})^2 | \underline{z}_0\}$$

$$= (A + B)\lambda_{\underline{z}_0}^2 - 2A\lambda_{\underline{z}_0} + A + \sigma_k^2(1 + 1 \;/\; n),$$

where $A = \underline{\beta}'(\underline{z}_0 - \underline{\mu})(\underline{z}_0 - \underline{\mu})'\underline{\beta} + \underline{\beta}'\Sigma\underline{\beta} \;/\; n$, and

$$B = \sigma_k^2\{(\underline{z}_0 - \underline{\mu})'\Sigma^{-1}(\underline{z}_0 - \underline{\mu}) + k \;/\; n\} \;/\; (n - k - 2).$$

Obviously $E\{(y_0 - \hat{y}_0^{\dagger})^2 | \underline{z}_0\}$ is a quadratic in $\lambda_{\underline{z}_0}$ and reduces to (3.3a) when $\lambda_{\underline{z}_0} = 1$. Since the coefficient of $\lambda_{\underline{z}_0}^2$ in (3.11) is always non-negative, the value of $\lambda_{\underline{z}_0}$ which minimizes the conditional p.m.s.e. is given by

$$(3.12) \qquad \lambda_{\underline{z}_0} = A \;/\; (A + B).$$

In a practical situation the value of $\lambda_{\underline{z}_0}$ cannot be calculated from (3.12) since $\underline{\mu}$ and Σ^* are unknown. However, an estimate of $\lambda_{\underline{z}_0}$, $\hat{\lambda}_{\underline{z}_0}$ (say), can be obtained by using the least squares estimators of $\underline{\mu}$ and Σ^* in (3.12) such that

$$(3.13) \qquad \hat{\lambda}_{\underline{z}_0} = \hat{A} \;/\; (\hat{A} + \hat{B}),$$

where $\hat{A} = \underline{x}_0'\hat{\underline{\beta}}\hat{\underline{\beta}}'\underline{x}_0 + \hat{\underline{\beta}}'S\hat{\underline{\beta}} \;/\; n$, and $\hat{B} = \hat{\sigma}_k(\underline{x}_0'S^{-1}\underline{x}_0 + k \;/\; n) \;/\; (n - k - 2)$.

Hence the prediction equation (3.10) becomes

$$(3.14) \qquad \hat{y}_i^{\dagger} = \bar{y} + \hat{\lambda}_{\underline{z}_i}\underline{x}_i'\hat{\underline{\beta}}.$$

We obtain the unconditional p.m.s.e. by taking the expectation of the conditional p.m.s.e. over $\underline{z}_0$ as

(3.15)

$$E(y_0 - \hat{y}_0^{\dagger})^2 = (1 + 1 \;/\; n)[\{\underline{\beta}'\Sigma\underline{\beta} + k\sigma_k^2 \;/\; (n - k - 2)\}\lambda^2 - 2\underline{\beta}'\Sigma\underline{\beta} + \sigma_{00}].$$

The expression for the unconditional p.m.s.e. is also a quadratic in λ and the coefficient of λ^2 is always non-negative. Hence the value of λ that minimizes $E(y_0 - \hat{y}_0^{\dagger})^2$ is

$$(3.16) \qquad \lambda = \underline{\beta}'\Sigma\underline{\beta} \ / \ \{\underline{\beta}'\Sigma\underline{\beta} + k\sigma_k^2 \ / \ (n - k - 2)\}.$$

In a real problem, an estimate $\hat{\lambda}$(say) of λ is obtained by using the least squares estimator of Σ^* as

$$(3.17) \qquad \hat{\lambda} = \hat{\underline{\beta}}'S\hat{\underline{\beta}} \ / \ \{\hat{\underline{\beta}}'S\hat{\underline{\beta}} + k\hat{\sigma}_k^2 \ / \ (n - k - 2)\}.$$

Since S^* is random so is $\hat{\lambda}$. We proceed as follows to find the distribution of $\hat{\lambda}$.

Define $F = \frac{\hat{\underline{\beta}}'S\hat{\underline{\beta}}}{\hat{\sigma}_k^2} \frac{n - k}{k - 1}$, then the conditional density function of F given Z is (see Anderson [2]) $\frac{(k - 1)}{(n - k)} \cdot \frac{\exp(-\underline{\beta}'S\underline{\beta} \ / \ 2\sigma_k^2)}{\Gamma((n - k) \ / \ 2)}$.

$$\sum_{i=0}^{\infty} \frac{(\underline{\beta}'S\underline{\beta} \ / \ 2\sigma_k^2)^i}{i!} \cdot \frac{(\frac{k - 1}{n - k} f)^{(k + 2i - 3) \ / \ 2}}{(1 + \frac{k - 1}{n - k} f)^{(n + 2i - 1) \ / \ 2}}$$

$$\cdot \frac{\Gamma((n + 2i - 1) \ / \ 2)}{\Gamma((k + 2i - 1) \ / \ 2)},$$

and hence the conditional density function of $F' = \frac{(k - 1)}{(n - k)} \cdot F$ given Z is

(3.18)

$$\frac{\exp(-\underline{\beta}'S\underline{\beta} \;/\; 2\sigma_k^2)}{\Gamma((n - k) \;/\; 2)} \;\cdot$$

$$\sum_{i=0}^{\infty} \frac{(\underline{\beta}'S\underline{\beta} \;/\; 2\sigma^2)^i}{i!} \; \frac{(f')^{(k + 2i - 3) \;/\; 2}}{(1 + f')^{(n + 2i - 1) \;/\; 2}} \; \frac{\Gamma((n + 2i - 1) \;/\; 2}{\Gamma((k + 2i - 1) \;/\; 2}$$

Now let $\underline{\beta}'\Sigma\underline{\beta} \;/\; \sigma_k^2 = \Phi$ and $\underline{\beta}'S\underline{\beta} \;/\; \sigma_k^2 = \Phi\chi_{n-1}^2$ then

$$E\{(\Phi\chi_{n-1}^2 \;/\; 2)^i \exp(-\Phi\chi_{n-1}^2 \;/\; 2)\} = \frac{\Phi^i}{(1 + \Phi)^{(n + 2i - 1) \;/\; 2}} \; \frac{\Gamma((n + 2i - 1) \;/\; 2)}{\Gamma((n - 1) \;/\; 2)}.$$

Applying the above result to (3.18), we get the unconditional density function of F' as

$$\frac{1}{\Gamma((n - k) \;/\; 2)\Gamma((n - 1) \;/\; 2)} \sum_{i=0}^{\infty} \frac{\Phi^i}{(1 + \Phi)^{(n + 2i - 1) \;/\; 2}}$$

$$\cdot \frac{(f')^{(k + 2i - 3) \;/\; 2}}{(1 + f')^{(n + 2i - 1) \;/\; 2}}$$

$$\cdot \frac{\{\Gamma((n + 2i - 1) \;/\; 2)\}^2}{\Gamma((k + 2i - 1) \;/\; 2)} \; \frac{1}{i!},$$

or

$$\frac{(f')^{(k - 3) \;/\; 2}}{(1 + f')^{(n - 1) \;/\; 2}} \cdot \frac{1}{(1 + \Phi)^{(n - 1) \;/\; 2}} \cdot \frac{1}{\Gamma((n - k) \;/\; 2)\Gamma((n - 1) \;/\; 2)}.$$

$$\sum_{i=0}^{\infty} \left\{\frac{\Phi}{1 + \Phi}\right\}^i \left\{\frac{f'}{1 + f'}\right\}^i \frac{\{\Gamma((n + 2i - 1) \;/\; 2)\}^2}{\Gamma((k + 2i - 1) \;/\; 2)} \cdot \frac{1}{i!} \;.$$

Since $\hat{\lambda} = \hat{\underline{\beta}}'S\hat{\underline{\beta}} \;/\; \{\hat{\underline{\beta}}'S\hat{\underline{\beta}} + k\hat{\sigma}_k^2 \;/\; (n - k - 2)\}$, we can write

$$F' = \frac{\hat{\lambda}}{1 - \hat{\lambda}} \cdot \frac{k}{n - k - 2} \text{ and } |J| = |d\hat{\lambda} \mid df'| = \frac{n - k - 2}{k}(1 - \hat{\lambda})^2.$$

Hence g, the density function of $\hat{\lambda}$ is given by

$$g(\lambda) = \begin{cases} C \cdot \sum_{i=0}^{\infty} (\frac{\underline{\beta}'\Sigma\underline{\beta}}{\sigma_{00}})^i \{\frac{k\lambda}{k + (n - 2k - 2)(1 - \lambda)}\}^i \frac{\{\Gamma((n + 2i - 1) / 2)\}^2}{\Gamma((k + 2i - 1) / 2)} \frac{1}{i!} & , 0 < \lambda < 1 \\ 0 & , \text{elsewhere}, \end{cases}$$

$$\text{where } C = \frac{k^{(k - 3) / 2}(n - k - 2)^{(n - k - 2) / 2}}{\Gamma((n - k) / 2) \cdot \Gamma((n - 1) / 2)}$$

$$\cdot \frac{\lambda^{(k - 3) / 2}(1 - \lambda)^{(n - k + 2) / 2}}{\{k + (n - 2k - 2)(1 - \lambda)\}^{(n - 1) / 2}}$$

$$\cdot (\sigma_k^2 / \sigma_{00})^{(n - 1) / 2}.$$

3.3 The Subset-Lambda Approach

When the subset and lambda approaches are used together, the prediction equation is given by

$$(3.19) \qquad \tilde{y}_i^{\dagger} = \bar{y} + \lambda_1 \underline{x}_{i1}' \tilde{\underline{\beta}}_1$$

The predicted response at $\underline{z}_0$ is given by $\tilde{y}_0^{\dagger} = \bar{y} + \lambda_{1\underline{z}_0} \underline{x}_{01}' \tilde{\underline{\beta}}_1$ and the conditional p.m.s.e. by

$$(3.20) \qquad E\{(y_0 - \tilde{y}_0^{\dagger})^2 | \underline{z}_0\} = E\{(\underline{x}_0'\underline{\beta} + \varepsilon_0 - \bar{\varepsilon} - \lambda_{1\underline{z}_0} \underline{x}_{01}' \tilde{\underline{\beta}}_1)^2 | \underline{z}_0\}$$

$$= G\lambda_{1\underline{z}_0}^2 - 2H\lambda_{1\underline{z}_0} + \sigma_k^2 + \sigma_p^2 / n$$

$$+ \underline{\beta}'(\underline{z}_0 - \underline{\mu})(\underline{z}_0 - \underline{\mu})'\underline{\beta} + \underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 / n,$$

where $G = \sigma_p^2\{(\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1) + p / n\} / (n - p - 2)$

$$+ \underline{\Phi}_1'(\underline{z}_{01} - \underline{\mu}_1)(\underline{z}_{01} - \underline{\mu}_1)'\underline{\Phi}_1 + \underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 / n,$$

and $H = \underline{\beta}'(\underline{z}_0 - \underline{\mu})(\underline{z}_{01} - \underline{\mu}_1)'\underline{\Phi}_1 + \underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 / n$. Since $E\{(y_0 - \tilde{y}_o^{\dagger})^2|\underline{z}_0\}$ is a quadratic in $\lambda_{1\underline{z}_0}$ and the coefficient of $\lambda_{1\underline{z}_0}^2$ is always non-negative, the conditional p.m.s.e. is minimized by

(3.21) $\lambda_{1\underline{z}_0} = H / G$

An estimate $\tilde{\lambda}_{1\underline{z}_0}$ of $\lambda_{1\underline{z}_0}$ is obtained by using the least squares estimators of the unknown parameters in the above equation. That is

(3.22) $\tilde{\lambda}_{1\underline{z}_0} = \hat{H} / \hat{G},$

where $\hat{G} = \hat{\sigma}_p^2(\underline{x}_{01}'S_{11}^{-1}\underline{x}_{01} + p / n) / (n - p - 2) + \hat{\Phi}_1'\underline{x}_{01}\underline{x}_{01}'\hat{\Phi}_1$

$$+ \hat{\underline{\Phi}}_1'S_{11}\hat{\underline{\Phi}}_1 / n, \text{ and}$$

$$\hat{H} = \hat{\underline{\beta}}'\underline{x}_0\underline{x}_{01}'\hat{\underline{\Phi}}_1 + \hat{\underline{\Phi}}_1'S_{11}\hat{\underline{\Phi}}_1 / n.$$

We select that particular value of $\hat{\lambda}_{1\underline{z}_0}$ and the subset of the predictor variables that minimizes $E\{(y_0 - \tilde{y}_0^{\dagger})^2|\underline{z}_0\}$ when the least squares estimators of the unknown parameters are used in (3.20).

By taking the expectation of the conditional p.m.s.e. over $\underline{z}_0$, we obtain the unconditional p.m.s.e. as

(3.23) $$E(y_0 - \tilde{y}_0^{\dagger})^2 = (1 + 1 / n)[\{\underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 + p\sigma_p^2 / (n - p - 2)\}\lambda_1^2 - 2\underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1\lambda_1 + \sigma_{00}].$$

Since the coefficient of λ_1^2 in the above expression is always non-negative,

(3.24) $$\lambda_1 = \underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 / \{\underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 + p\sigma_p^2 / (n - p - 2)\},$$

minimizes the unconditional p.m.s.e. As before, we obtain an estimate $\tilde{\lambda}_1$ of λ_1 by using the least squares estimators of the unknown parameters in (3.24). Hence

(3.25) $$\tilde{\lambda}_1 = \hat{\underline{\Phi}}_1'S_{11}\hat{\underline{\Phi}}_1 / \{\hat{\underline{\Phi}}_1'S_{11}\hat{\underline{\Phi}}_1 + p\hat{\sigma}_p^2 / (n - p - 2)\}.$$

Here also, we select the value of λ_1 and the subset of predictor variables which minimizes $-(\hat{\underline{\Phi}}_1'S_{11}\hat{\underline{\Phi}}_1)^2 / \{\hat{\underline{\Phi}}_1'S_{11}\hat{\underline{\Phi}}_1 + p\hat{\sigma}_p^2 / (n - p - 2)\}$.

CHAPTER 4

NUMERICAL EXAMPLES

In this Chapter, we use the results derived in the previous two chapters to analyze a few sets of data.

4.1 The Gorman-Toman Problem

We first present the analysis of the data presented in the paper by Gorman and Toman [20]. Since the data were generated from the equation,

$$y_i = 1 + x_{1i} + x_{2i} + \varepsilon_i,$$

where ε_i = random standard normal deviate N(0, 1), the parameters of the model are known. The data are given in Appendix B. Although the parameters of the model are known, the subset criterion (2.10) using the l.s.e. of the parameters leads us to use the complete equation. In lambda approach, the value of lambda at each point is calculated using the equation (2.16). Although the data were generated at four values of (x_1, x_2), we predict the value of the response at all the lattice points in the square (-1, -1), (-1, 1), (1, 1), (1, -1) every .5 units. The estimated value of lambda for each of the points is given in Figure 3. Since the true value of the response is known for all values of the predictor variables, we calculate the "true" value minus the

x_2 \ x_1	-1.0	-.5	0.0	.5	1.0
1.0	.168	.456	.755	.925	.984
.5	.005	.167	.754	.985	.965
0.0	.599	.592	1.0	.592	.599
-.5	.965	.985	.754	.167	.005
-1.0	.984	.925	.755	.456	.168

FIGURE 3

ESTIMATED LAMBDA VALUE AT EACH POINT

predicted value and call it the "error". The sum of the squared errors is calculated and is given below:

	Sum of the Squared Errors
Full Equation	3.8348
The Lambda Approach	1.9076

In this problem the subset criterion using the l.s.e. of the parameters leads us to use the complete equation and the lambda approach results in 50.25 per cent improvement in the sum of the squared errors.

4.2 Hald's Data

In this section we give an analysis of the often quoted Hald's data (Appendix B) which was first reported by H. Woods, H. H. Steinour and H. R. Starke [50]. The heat evolved during setting and hardening (the response variable) was studied for 13 different compositions of the Portland Cement. The various compositions of Portland Cement were obtained by controlling the amounts of four compounds (here referred to as predictor variables), namely, tricalcium aluminate ($3CaO{\cdot}Al_2O_3$), tricalcium silicate ($3Cal{\cdot}SiO_2$), tetracalcium aluminoferrite ($4CaO{\cdot}Al_2O_3{\cdot}Fe_2O_3$) and β-diacalcium silicate ($3Cal{\cdot}SiO_2$).

We first analyze the data using all the 13 observations to estimate the unknown parameters of the model. In the subset approach, we calculate an estimate of the p.m.s.e. of subset prediction equation for each observation from equation (2.7). For a given observation, the subset with the minimum p.m.s.e. is selected. In this sense the "best" subset of predictor variables is found for each observation. In the lambda approach all of the predictor variables are used in the equation and the value of lambda for each observation is calculated using (2.16). In the ridge approach also all the predictor variables are used. The "ridge estimates" for h values of 0(.005).01(.01).02(.02).1(.1).5 are calculated. Then for each observation, the value of h, which results in the smallest p.m.s.e. as calculated by (2.25), is selected. In this manner, the "best" value of h is found for each observation.

It should be pointed out that in the subset approach, subsets of two and less variables did not result in improved p.m.s.e. Hence when the subset approach is used in combination with the other approaches, we use only the three variable subsets and the full equation.

In the subset-lambda approach, for a given subset of variables, the lambda value (lambda is a function of the subset of variables and the observation point) is calculated by (2.33) and an estimate of the p.m.s.e. by (2.31) for each observation. We then select that particular subset of the predictor variables and the lambda value which results in the minimum p.m.s.e. at the given observation. This process is repeated for all the observations. For the subset-ridge approach, the p.m.s.e. for an observation is estimated by (2.40) for various h values and the subset prediction equations. For each observation, that particular subset of variables and h value are selected which give the smallest p.m.s.e.

In this manner, the "best" subset of predictor variables and the h value are selected for all the observations. For a given point, in the lambda-ridge approach, we calculate the lambda value by (2.47) and an estimate of the p.m.s.e. by (2.45) for various values of h. For each observation, the h and the lambda (lambda is a function of h) value corresponding to the minimum p.m.s.e. is selected. Finally, in the subset-lambda-ridge approach, we calculate, for various values of h and the subset of predictor variables, the value of lambda using (2.54) and an estimate of the p.m.s.e. using (2.53) for each observation. Then for a given observation, we select the h, the lambda value and the subset of predictor variables which gives the smallest p.m.s.e. This procedure is repeated for all the observations.

In the calculations of lambda and the p.m.s.e. the l.s.e. of the unknown parameters are used. Table 1 gives the sum of the residual squares at all the points.

The results tend to look better when we predict the same points which are used to estimate the unknown parameters. Hence we next analyze the data by the jack-knife technique. That is, while predicting an observation, all observations except this particular observation are used to estimate the unknown parameters of the model. This procedure is repeated for all the observations. Because of the large number of calculations involved, we use the subset, the lambda and the subset-lambda approaches only.

In the subset approach, all the observations except the one we want to predict are used to calculate all the possible subset equations.

TABLE 1

HALD'S DATA ANALYSIS ALL APPROACHES

	Residual sum of squares	Percentage Improvement
Usual least squares equation	47.8636	---
The subset approach	47.3451	1.1
The lambda approach	46.6364	2.6
The ridge approach	43.9337	8.2
The subset-lambda approach	47.6617	.4
The subset-ridge approach	46.3232	3.2
The lambda-ridge approach	44.9887	6.2
The subset-lambda-ridge approach	47.6624	.4

TABLE 2

THE JACK-KNIFE ANALYSIS OF HALD'S DATA

	Residual sum of squares	Percentage Improvement
Usual least squares equation	110.3246	---
The subset approach	96.0741	12.9
The lambda approach	106.6395	3.3
The subset-lambda approach	110.2600	.06

Then for the given point, the subset which results in the minimum p.m.s.e. at the point as given by (2.7) is selected. This procedure is repeated for each of the observations. In the lambda approach, all the variables are used in the equation. To estimate the unknown parameters of the model, all observations except the one we want to predict are used. The lambda value for each observation is calculated by (2.16) and the response predicted.

In the subset-lambda approach also, the unknown parameters (for a given observation) are estimated by excluding the particular observation from the sample. Then the lambda value (which is a function of the subset and the observation we want to predict) is calculated using (2.33) and an estimate of the p.m.s.e. using (2.31). Then that lambda value and the subset of predictor variables which results in the smallest p.m.s.e. for the observation is selected. This procedure is followed for all the observations.

In Table 2, we give the residual sum of squares for the various techniques.

4.3 The ACT[3] Data

On the basis of a student's performance on English, Mathematics, Social Sciences and Natural History (here referred to as the predictor variables) tests of the American College Testing Service, we want to predict the first year college grade point average, GPA, of the student (the response variable).

[3] The American College Testing (ACT) Service, Iowa City, Iowa, provided us with these data.

The data are assumed to follow a multivariate normal distribution. Out of a total of 83 observations (on male students only), 25 are selected at random and are used to estimate the parameters of the full equation as well as the subset equations. Based on these estimates the remaining 58 observations are predicted.

It is felt that the suggested techniques will result in greater improvement, if the sample size used to estimate the unknown parameters is small. To study the behavior of these procedures as a "function of sample size", small samples are created artifically in the following manner. From the 25 observations selected earlier 20 are selected at random. Then from these 20, 15 are selected at random and from these 15, 10 are selected at random. Now each of these sets of 25, 20, 15 and 10 observations are used for estimating the unknown parameters of the model and the same 58 observations are predicted each time. Note the nested nature of these sets of observations.

We first give the conditional (conditioned on the sample used to estimate the parameters) analysis of the data using the expressions given in Chapter 2 for calculating the p.m.s.e. and the lambda value. For each set of observations (25, 20, 15 and 10), the following procedure is used to predict the 58 observations.

In the subset approach, for each observation an estimate of the p.m.s.e. is calculated from equation (2.7) and the subset equation with the smallest p.m.s.e. is used to predict the GPA. The difference of the observed GPA and the predicted GPA (the residual) is calculated.

This procedure is repeated for the 58 observations (not used in estimating the parameters) and the residual sum of squares calculated. In the lambda approach all the variables are used in the equation and the lambda value for each observation is calculated by (2.16). As in the subset approach, the residual sum of squares for 58 observations is calculated.

In the subset-lambda approach, for a given subset of variables and the observation the lambda value (lambda is a function of the subset of variables and the values of the predictor variables) is calculated by (2.33) and the p.m.s.e. estimated by (2.31). Then that particular subset of variables and the lambda value are selected which result in the smallest p.m.s.e. for a given observation. The procedure is repeated for the 58 observations and the residual sum of squares calculated. The residual sum of squares and the percentage improvement are given in Table 3.

We next give the unconditional analysis of the data using the expressions given in Chapter 3 for calculating the p.m.s.e. and the lambda value. As before, the data are analyzed using the same set of 25, 20, 15 and 10 observations to estimate the unknown parameters and the same 58 observations are predicted. The following procedure is used for each set.

In the subset approach, the p.m.s.e. is estimated by (3.6) for each observation and the subset with the smallest p.m.s.e. is selected to predict the GPA. For each observation, the residual is calculated

FIRST SET OF RUNS

TABLE 3

ACT DATA: CONDITIONAL ANALYSIS

Residual Sum of Squares (Percentage Improvement)

# of Obs. Technique	10	15	20	25
Full Equation	48.0752	57.5734	56.2861	55.1380
Subset	42.9146 (10.734)	47.7331 (17.091)	49.5122 (12.034)	49.9124 (9.477)
Lambda	48.2408 (-.344)	56.6397 (1.622)	54.0972 (3.152)	51.8885 (5.893)
Subset-Lambda	43.7229 (9.053)	53.4775 (7.114)	53.1197 (5.625)	50.9308 (7.630)

as the observed GPA minus the predicted GPA. For the 58 observations, the sum of the residual squares is calculated. In the lambda approach all the variables are used in the equation and the lambda value for each observation is calculated by (3.13). The residual is calculated at each observation and the residual sum of squares is calculated for the 58 observations predicted.

In the subset-lambda approach, for a given subset of variables and an observation, the value of lambda is calculated using (3.22) and the p.m.s.e. is estimated using (3.20). The subset of variables and the lambda value which result in the smallest p.m.s.e. at a given observation are selected to predict the GPA. The procedure is repeated for each of the 58 observations and the sum of the residual squares calculated. The residual sum of squares and the percentage improvement are given in Table 4.

We would expect the residual sum of squares to increase when a smaller sample is used to estimate the parameters of the model. Due to sample variation, the residual sum of squares when 10 observations are used to estimate the parameters is less than the residual sum of squares when 25 observations are used. Hence another set of 10 observations (all observations in this set being different from the previous 10 observations) is selected at random from the 25 observations. Similarly, another set of 15 (with 10 observations different from the previous set of 15) and 20 (with 5 observations different from the previous set of 20) are selected at random. Once again using these sets of 10, 15, 20 and 25 observations the conditional and the unconditional

FIRST SET OF RUNS

TABLE 4

ACT DATA: UNCONDITIONAL ANALYSIS

Residual Sum of Squares
(Percentage Improvement)

# of Obs. Technique	10	15	20	25
Full Equation	48.0752	57.5734	56.2861	55.1380
Subset	39.0375 (18.814)	47.7967 (16.981)	49.9130 (11.323)	53.9102 (2.227)
Lambda	47.9727 (.213)	56.4499 (1.951)	53.9597 (4.133)	51.8935 (5.854)
Subset-Lambda	40.6033 (15.542)	50.9773 (11.456)	51.1597 (9.107)	49.1435 (10.872)

analysis of the data are carried out and the 58 observations predicted. For this analysis, the residual sum of squares and the percentage improvement are given in Table 5 and 6 for the conditional and the unconditional analysis respectively.

4.4 Discussion of Results

From the Gorman-Toman problem, it can be observed that the lambda approach results in a substantial improvement (50.25 per cent) in the sum of the squared errors. (The subset criterion leads us to use the full equation.) It is also worth observing how the lambda value varies from point to point which suggests that it is worthwhile to calculate the lambda value at each point.

From the analysis of Hald's data, we observe from Table 2 that the subset approach results in maximum improvement (12.92 per cent). The lambda and the subset-lambda approaches also result in improvement. (Note that the ridge approach was not used in this analysis.) We see from Table 1 that although all the approaches result in improved residual sum of squares, the ridge approach (8.21 per cent) results in the maximum improvement.

From the analyses of the ACT data, we observe from Tables 5 and 6 that the subset approach results in maximum improvement followed by the lambda and the subset-lambda approaches. Although further investigation is needed, Tables 5 and 6 suggest that the techniques suggested are specially useful when only small samples are available to estimate the parameters. Also the results of the Hald's data and the ACT data

SECOND SET OF RUNS

TABLE 5

ACT DATA: CONDITIONAL ANALYSIS

# of Obs. Technique	Residual Sum of Squares (Percentage Improvement) 10	 15	 20	 25
Full Equation	125.4816	58.9392	55.5081	55.1380
Subset	50.4203 (59.818)	39.1601 (33.558)	44.8877 (19.133)	49.9126 (9.477)
Lambda	103.0995 (17.836)	51.3753 (12.833)	51.6588 (6.935)	51.8885 (5.893)
Subset-Lambda	47.8036 (61.903)	43.3267 (26.489)	51.3634 (7.467)	50.9308 (7.630)

SECOND SET OF RUNS

TABLE 6

ACT DATA: UNCONDITIONAL ANALYSIS

Residual Sum of Squares
(Percentage Improvement)

# of Obs. Technique	10	15	20	25
Full Equation	125.4816	58.9392	55.5081	55.1380
Subset	54.0352 (56.938)	41.6206 (29.383)	45.3090 (18.376)	53.9102 (2.226)
Lambda	91.1696 (27.344)	50.8653 (13.698)	51.5066 (7.209)	51.8935 (5.884)
Subset-Lambda	44.2705 (64.719)	41.1650 (30.156)	47.4511 (14.515)	49.1435 (10.871)

suggest that the combination approaches do not improve the results over the individual approaches. This point also needs further investigation.

CHAPTER 5

DIRECTIONS FOR FUTURE RESEARCH

We now pose some problems which require further work and suggest some possible extensions of the present research.

Although we have given the p.m.s.e. for the subset equations and the fixed values of λ and h, we have not been able to obtain the p.m.s.e. for the procedures. For example, we derived the p.m.s.e. as a function of λ but have not obtained the p.m.s.e. using $\hat{\lambda}$.

If possible, it would be interesting to derive the conditions under which each of the approaches discussed is the best since we do not know how the various approaches compare. Also the results of Chapter 4 seem to suggest that the improvement increases as the sample size used to estimate the parameter decreases. Hence it will be useful to know the sample size for which the improvement is maximum.

At present, we have to calculate all possible subset regression equations to find the subset with the minimum p.m.s.e. at $\underset{\sim}{z}_0$, the vector of predictor variables. To take full advantage of the subset approach it should be possible to develop an algorithm by which all possible subset regression equations need not be evaluated.

In the lambda approach, at present, neither the expected value nor the variance of $\hat{\lambda}$, the estimator of the unknown constant λ, are known. The properties of the estimator are of interest in that they may lead us to a better estimator of λ.

In the ridge approach, the optimum value of h is calculated by evaluating the p.m.s.e. for various values of h and selecting the h value for which the p.m.s.e. is the minimum. It should be possible to find an expression for h which guarantees the best value of h in the sense that the p.m.s.e. is the minimum for this value. Also it is worthwhile to derive the results of the ridge approach for the unconditional analysis when the random predictor variables follow a multivariate normal distribution.

For the problem in which both the fixed (or controllable) and the random predictor variables are present, the results of Chapter 2 are applicable for the conditional analysis. But we do not know the results for the unconditional analysis for this problem. It is an important problem.

Since two population discriminant problem is analogous to regression analysis, our techniques should be applicable. It will be interesting to find out how these approaches improve the results for this problem.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Allen, David M., "Mean square error of prediction as a criterion for selecting variables," Technical Report Number 16, University of Kentucky, Lexington, Kentucky, December, 1970.

2. Anderson, T. W., An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, New York, 1958.

3. Ashar, Vijay G., "On the use of preliminary tests in regression analysis," A. S. A. 1968 Proc. Buss. and Econom. Statist. Sect. (Pittsburg, Pa.), 337-344.

4. Bancroft, T. A., "On biases in estimation due to the use of preliminary tests of significance," Annals of Mathematical Statistics, 1944, 15, 190-204.

5. Banerjee, D. P., "On the moments of the multiple correlation coefficient in samples from normal population," Journal of the Indian Society of Agricultural Statistics, 1952, 4, 88-90.

6. Beale, E. M. L., "Selecting an optimum subset," Integer and Nonlinear Programming. Ed. J. Abadie, North Holland Publishing Co., Amsterdam, 1970.

7. Beale, E. M. L., "Computational Methods for least squares," (Unpublished).

8. Beale, E. M. L., "A note on procedures for variable selection in multiple regression," Technometrics, 1970, 4, 909-914.

9. Beale, E. M. L., Kendall, M. G., and Mann, D. W., "The discarding of variables in multivariate analysis," Biometrika, 1967, 54, 357-366.

10. Cochran, W. G., "The ommission or addition of an independent variate in multiple linear regression," J. Royal Statistical Soc. Suppl., 1938, B-5, 171-176.

11. Davies, P., "The choice of variables in the design of experiments for linear regression," Biometrika, 1969, 56, 55-63.

12. Draper, N. R. and Smith, H., Applied Regression Analysis, John Wiley and Sons, New York, 1966.

13. Dwyer, P. S., "Recent developments in correlation technique," Journal of the American Statistical Association, 1942, 37, 441-460.

14. Fisher, R. A., "The general sampling distribution of the multiple correlation coefficient," Proc. Roy. Soc., A., 1928, 121, 654-673.

15. Fisk, P. R., "A note on characterization of the multivariate normal distribution," Annal of Mathematical Statistics, 1970, 41, 486-494.

16. Freund, R. J., Vail, R. W., and Clunies-Ross, C. S., "Residual analysis," Journal of the American Statistical Association, 1961, 56, 98-104.

17. Garside, M. J., "The best subset in multiple regression analysis," Applied Statistics, 1965, 14, 196-201.

18. Goldberger, A. S., "Stepwise least squares: Residual analysis and specification error," Journal of the American Statistical Association, 1961, 56, 998-1000.

19. Goldberger, A. S. and Jochems, D. B., "Note on stepwise least squares," Journal of the American Statistical Association, 1961, 56, 105-110.

20. Gorman, J. W. and Toman, R. J., "Selection of variables for fitting equations to data," Technometrics, 1966, 9, 531-540.

21. Graybill, Franklin A., An Introduction to Linear Statistical Models, Vol. 1, McGraw-Hill Book Company, New York, 1961.

22. Haitovsky, Yoel., "A note on the maximization of $\bar{R}^2$," The American Statistician, 1969, 23, 20-21.

23. Hocking, R. R. and Leslie, R. N., "Selection of best subset in regression analysis," Technometrics, 1967, 9, 531-540.

24. Hoerl, A. E. and Kennard, R. W., "Ridge regression: Biased estimation for non-orthogonal problems," Technometrics, 1970, 12, 55-68.

25. Hoerl, A. E. and Kennard, R. W., "Ridge regression: Application to non-orthogonal problems," Technometrics, 1970, 12, 69-82.

26. Johnston, J., Econometric Methods, McGraw-Hill Book Company, New York, 1963.

27. Kabe, D. G., "On the distribution of the regression coefficient matrix of a normal distribution," The Australian Journal of Statistics, 1968, 10, 21-23.

28. Kerridge, D., "Errors of prediction in multiple regression with stochastic regressor variables," Technometrics, 1967, 9, 309-311.

29. Kerridge, D., Private Communication, University of Aberdeen, August 27, 1970.

30. LaMotte, L. R. and Hocking, R. R., "Computational efficiency in the selection of regression coefficients," Technometrics, 1970, 12, 83-94.

31. Larson, Harold J. and Bancroft, T. A., "Biases in prediction by regression for certain incompletely specified models," Biometrika, 1963, 4, 391-402.

32. Lindley, D. V., "The choice of variables in multiple regression," J. Royal Statist. Soc., Ser. B., 1968, 30, 31-66.

33. Longley, James W., "An appraisal of least squares programs for the electronic computer from the point of view of the user," Journal of the American Statistical Association, 1967, 62, 819-841.

34. Mallows, C., "Choosing a subset regression," Presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Kansas, May 7-9, 1964.

35. Mantel, N., "Why step down procedures in variable selection," Technometrics, 1970, 3, 621-626.

36. Marquardt. D. W., "Generalized inverse, ridge regression, biased linear estimation and non-linear estimation," Technometrics, 1970, 12, 591-612.

37. Moran, P. A. P., "The distribution of the multiple correlation coefficient," Proc. Camb. Phil. Soc., 1950, 46, 521-522.

38. Pitman, E. J. G., "The 'closest' estimate of statistical parameters," Proc. Camb. Phil. Soc., 1937, 33, 212-222.

39. Schatzoff, M., Tsao, R., and Fienberg, S., "Efficient calculations of all possible regressions," Technometrics, 1968, 10, 769-780.

40. Sclove, S. L., "Improved estimators for coefficients in linear regression," Journal of the American Statistical Association, 1968, 63, 596-606.

41. Stein, Charles, "Multiple regression," Contributions to Probability and Statistics, "Essays in Honor of Harold Hotelling," 1960, 424-443.

42. Toro-Vizcarrondo, Carlos and Wallace, T. D., "A test of the m.s.e. criterion for restriction in linear regression," Journal of the American Statistical Association, 1968, 63, 558-573.

43. Wallace, T. D., "Efficiencies for stepwise regression," Journal of the American Statistical Association, 1964, 59, 1179-1181.

44. Walls, Robert C. and Weeks, David L., "A note on the variance of a predicted response in regression," The American Statistician, 1969, 23, 24-26.

45. Wampler, Roy H., "A report on the accuracy of some widely used least squares computer programs," Journal of the American Statistical Association, 1970, 65, 549-565.

46. Webster, J. T., "On the use of a biased estimator in linear regression," J. Indian Statistical Association, 1965, 3, 82-90.

47. Wilks, S. S., "On the sampling distribution of the multiple correlation coefficient," Annals of Mathematical Statistics, 1932, 3, 196-202.

48. Williams, J. S., "Some statistical properties of a genetic selection index," Biometrika, 1962, 49, 325-337.

49. Wiorkowski, J. J., "Estimation of the proportion of the variance explained by regression when the number of parameters in the model may depend on sample size," Technometrics, 1970, 12, 915-919.

50. Woods, H., Steinour, H. H., and Starke, H. R., "Effects of composition of Portland Cement on heat evolved during hardening," Industrial and Engineering Chemistry, 1932, 24, 1207-1214.

APPENDICES

APPENDIX A

LEMMAS ON EXPECTATIONS

APPENDIX A

LEMMAS ON EXPECTATIONS

In this appendix, we give some results which are useful in finding the conditional and the unconditional p.m.s.e. for the subset prediction equation. Corresponding results for the full equation can be obtained by letting p equal to k, $\underline{z}_{i1}$ equal $\underline{z}_i$, etc.

<u>Lemma A1</u> $E(\tilde{\underline{\beta}}_1|X_1) = \underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2.$

<u>Proof</u>:

$$E(\tilde{\underline{\beta}}_1'|X_1) = E(\underline{s}_1'S_{11}^{-1}|X_1)$$

$$= E(\underline{s}_1'|X_1)S_{11}^{-1},$$

since S_{11}^{-1} is a function of X_1 only. To find $E(\underline{s}_1'|X_1)$, we find the expected value of a typical element $s_{1i} = \Sigma_m(y_m - \bar{y})x_{im} / (n - 1)$ of $\underline{s}_1$.

$$E(s_{1i}|X_i) = E[\Sigma_m(y_m - \bar{y})x_{im} / (n - 1)|X_1]$$

$$= E[\Sigma_m y_m x_{im} / (n - 1)|X_1]$$

$$= \Sigma_m E(y_m|X_1)x_{im} / (n - 1)$$

$$= \Sigma_m \underline{\sigma}_1'\Sigma_{11}^{-1}\underline{x}_m x_{im} / (n - 1)$$

$$= \underline{\sigma}_1'\Sigma_{11}^{-1}\underline{s}_{1i},$$

where $\underline{s}_{1i}$ is the i-th column vector of S_{11}.

Therefore $E(\underline{s}_1'|X_1) = \underline{\sigma}_1'\Sigma_{11}^{-1}[\underline{s}_{11}, \underline{s}_{12}, \ldots, \underline{s}_{1p}]$

$$= \underline{\sigma}_1'\Sigma_{11}^{-1}S_{11}$$

Hence $E(\underline{s}_1'S_{11}^{-1}|X_1) = \underline{\sigma}_1'\Sigma_{11}^{-1}S_{11}S_{11}^{-1}$

$$= \underline{\sigma}_1'\Sigma_{11}^{-1}.$$

Since by definition $\underline{\beta}' = \underline{\sigma}'\Sigma_{11}^{-1}$, therefore

$\underline{\beta}_1' = \underline{\sigma}_1'(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} - \underline{\sigma}_2'(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{21}\Sigma_{11}^{-1}$ and

$\underline{\beta}_2' = \underline{\sigma}_2'(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1} - \underline{\sigma}_1'(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}$ and

hence $\Sigma_{11}^{-1}\underline{\sigma}_1'$ can be written as $\underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2$.

Therefore $E(\tilde{\underline{\beta}}_1|X_1) = \underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2$.

<u>Corollary A1.1</u> $\quad E(\tilde{\underline{\beta}}_1) = \underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2$.

<u>Proof</u>: $\quad E(\tilde{\underline{\beta}}_1) = E\{E(\tilde{\underline{\beta}}_1|X_1)\}$

$$= E(\underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2), \qquad \text{Lemma A1.}$$

$$= \underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2.$$

<u>Lemma A2</u> $E(\underline{x}'_{01}\tilde{\underline{\beta}}_1|\underline{z}_0) = (\underline{z}_{01} - \underline{\mu}_1)'(\underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2)$.

<u>Proof</u>:

$$\begin{aligned} E(\underline{x}'_{01}\tilde{\underline{\beta}}_1|\underline{z}_0) &= E\{(\underline{z}_{01} - \bar{\underline{z}}_1)'\tilde{\underline{\beta}}_1|\underline{z}_0\} \\ &= E[E\{(\underline{z}_{01} - \bar{\underline{z}}_1)'\tilde{\underline{\beta}}_1|\underline{z}_0, X_1\}] \\ &= E[E\{(\underline{z}_{01} - \bar{\underline{z}}_1)'|\underline{z}_0\}E(\tilde{\underline{\beta}}_1|X_1)] \end{aligned}$$

since, given $\underline{z}_0$ and X_1, $\tilde{\underline{\beta}}_1$ and $\underline{z}_{01} - \bar{\underline{z}}_1$ are independent.

Therefore, $E(\underline{x}'_{01}\tilde{\underline{\beta}}_1|\underline{z}_0) = E\{(\underline{z}_{01} - \underline{\mu}_1)'(\underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2)|\underline{z}_0\}$, Lemma A1.

$$= (\underline{z}_{01} - \underline{\mu}_1)'(\underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2).$$

<u>Corollary A2.1</u> $E(\underline{x}'_{01}\tilde{\underline{\beta}}_1) = 0$

<u>Proof</u>:

$$\begin{aligned} E(\underline{x}'_{01}\tilde{\underline{\beta}}_1) &= E\{E(\underline{x}'_{01}\tilde{\underline{\beta}}_1|\underline{z}_0)\} \\ &= E\{(\underline{z}_{01} - \underline{\mu}_1)'(\underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2)\}, \quad \text{Lemma A2.} \\ &= 0 \qquad\qquad, \text{ since } E(\underline{z}_{01}) = \underline{\mu}_1. \end{aligned}$$

<u>Lemma A3</u> $E(\underline{x}_{01}\underline{x}'_{01}|\underline{z}_0) = (\underline{z}_{01} - \underline{\mu}_1)(\underline{z}_{01} - \underline{\mu}_1)' + \Sigma_{11} / n$.

<u>Proof</u>:

$$\begin{aligned} E(\underline{x}_{01}\underline{x}'_{01}|\underline{z}_0) &= E[(\underline{z}_{01} - \bar{\underline{z}}_1)(\underline{z}_{01} - \bar{\underline{z}}_1)'|\underline{z}_0] \\ &= E(\underline{z}_{01}\underline{z}'_{01} - \underline{z}_{01}\bar{\underline{z}}'_1 - \bar{\underline{z}}_1\underline{z}'_{01} + \bar{\underline{z}}_1\bar{\underline{z}}'_1|\underline{z}_0) \\ &= \underline{z}_{01}\underline{z}'_{01} - \underline{z}_{01}\underline{\mu}'_1 - \underline{\mu}_1\underline{z}'_{01} + E(\bar{\underline{z}}_1\bar{\underline{z}}'_1). \end{aligned}$$

Since $\text{Var}(\bar{\underline{z}}_1) = E(\bar{\underline{z}}_1\bar{\underline{z}}'_1) - E(\bar{\underline{z}}_1)E(\bar{\underline{z}}'_1)$, or $E(\bar{\underline{z}}_1\bar{\underline{z}}'_1) = \Sigma_{11} / n + \underline{\mu}_1\underline{\mu}'_1$,

$$E(\underline{x}_{01}\underline{x}'_{01}|\underline{z}_0) = (\underline{z}_{01} - \underline{\mu}_1)(\underline{z}_{01} - \underline{\mu}_1)' + \Sigma_{11} / n.$$

<u>Corollary A3.1</u> $E(\underline{x}'_{01}S_{11}^{-1}\underline{x}_{01}|\underline{z}_0) =$

$$(n - 1)\{(\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1) + p / n\} / (n - p - 2).$$

<u>Proof</u>:

$$\begin{aligned} E(\underline{x}'_{01}S_{11}^{-1}\underline{x}_{01}|\underline{z}_0) &= E\{tr(\underline{x}'_{01}S_{11}^{-1}\underline{x}_{01}|\underline{z}_0)\} \\ &= E\{tr(S_{11}^{-1}\underline{x}_{01}\underline{x}'_{01})|\underline{z}_0\} \\ &= tr\{E(S_{11}^{-1}\underline{x}_{01}\underline{x}'_{01}|\underline{z}_0)\} \\ &= tr\{E(S_{11}^{-1})E(\underline{x}_{01}\underline{x}'_{01}|\underline{z}_0)\}, \end{aligned}$$

since given $\underline{z}_0$, S_{11}^{-1} and $\underline{x}_{01}$ are independent. Therefore

$$E(\underline{x}'_{01}S_{11}^{-1}\underline{x}_{01}|\underline{z}_0) = tr[(n - 1)\Sigma_{11}^{-1} / (n - p - 2)$$

$$\cdot \{(\underline{z}_{01} - \underline{\mu}_1)(\underline{z}_{01} - \underline{\mu}_1)' + \Sigma_{11} / n\}],$$

since $E(S_{11}^{-1}) = (n - 1)\Sigma_{11}^{-1} / (n - p - 2)$, (see Williams [47]), and lemma 3

$$= (n - 1)tr\{\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1)(\underline{z}_{01} - \underline{\mu}_1)' + I_p / n\} / (n - p - 2)$$

$$= (n - 1)\{(\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1) + p / n\} / (n - p - 2),$$

since $tr\{\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1)(\underline{z}_{01} - \underline{\mu}_1)'\} = (\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1)$ and $tr(I_p) = p$.

<u>Corollary A3.2</u> $E[(\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1)] = p$.

Since $(\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1) \sim \chi_p^2$,

therefore $E\{(\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1)\} = p$.

<u>Corollary A3.3</u> $\quad E\{(\underline{z}_0 - \underline{\mu})(\underline{z}_{01} - \underline{\mu}_1)'\} = \begin{bmatrix} \Sigma_{11} \\ \\ \Sigma_{21} \end{bmatrix}.$

Follows immediately from the definition of variance-covariance matrix, i.e.,

$$E\{(\underline{z}_0 - \underline{\mu})(\underline{z}_0 - \underline{\mu})'\} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ & \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

and hence

$$E\{(\underline{z}_0 - \underline{\mu})(\underline{z}_{01} - \underline{\mu}_1)'\} = \begin{bmatrix} \Sigma_{11} \\ \\ \Sigma_{21} \end{bmatrix}.$$

<u>Lemma A4</u> $\quad \mathrm{Var}(\tilde{\underline{\beta}}_1 | X_1) = \sigma_p^2 S_{11}^{-1} / (n - 1).$

<u>Proof</u>: $\quad \mathrm{Var}(\tilde{\underline{\beta}}_1' | X_1) = \mathrm{Var}(\underline{s}_1' S_{11}^{-1} | X_1)$

$$= S_{11}^{-1} \mathrm{Var}(\underline{s}_1' | X_1) S_{11}^{-1}$$

since S_{11}^{-1} is a function of X_1 alone. To calculate $\mathrm{Var}(\underline{s}_1' | X_1)$, we calculate the covariance between two typical elements

$s_{1i} = \Sigma_m (y_m - \bar{y}) x_{im} / (n - 1)$ and $s_{1j} = \Sigma_\ell (y_\ell - \bar{y}) x_{j\ell} / (n - 1)$ of $\underline{s}_1$.

$$\begin{aligned}\mathrm{Cov}(s_{1i},\ s_{1j}|X_1) &= \mathrm{Cov}[\Sigma_m(y_m - \bar{y})x_{im}\ /\ (n-1),\ \Sigma_\ell(y_\ell - \bar{y})x_{j\ell}\ /\ (n-1)|X_1] \\ &= \mathrm{Cov}[\Sigma_m\Sigma_m y_m x_{im}\ /\ (n-1),\ \Sigma_\ell y_\ell x_{j\ell}\ /\ (n-1)|X_1] \\ &= \Sigma_{m,\ell}\mathrm{Cov}(y_m,\ y_\ell|X_1)x_{im}x_{j\ell}\ /\ (n-1)^2,\end{aligned}$$

since $\mathrm{Cov}(y_m,\ y_\ell|X_1) = \begin{cases}\sigma_p^2 & m = \ell \\ 0 & m \neq \ell\end{cases}$

$$\begin{aligned}\mathrm{Cov}(s_{1i},\ s_{1j}|X_1) &= \Sigma_m\sigma_p^2 x_{\ell m}x_{jm}\ /\ (n-1)^2 \\ &= \sigma_p^2 s_{ij}\ /\ (n-1).\end{aligned}$$

Therefore $\mathrm{Var}(\underline{s}_1'|X_1) = \sigma_p^2 S_{11}\ /\ (n-1)$,

and $$\begin{aligned}\mathrm{Var}(\underline{\tilde{\beta}}_1|X_1) &= S_{11}^{-1}\mathrm{Var}(s_1|X_1)S_{11}^{-1} \\ &= S_{11}^{-1}\sigma_p^2 S_{11}S_{11}^{-1}\ /\ (n-1) \\ &= \sigma_p^2 S_{11}^{-1}\ /\ (n-1).\end{aligned}$$

<u>Corollary A4.1</u> $\mathrm{Var}(\underline{x}_{01}'\underline{\tilde{\beta}}_1|\underline{z}_0,\ X_1) = \sigma_p^2\underline{x}_{01}'S_{11}^{-1}\underline{x}_{01}\ /\ (n-1)$.

<u>Proof</u>: $$\begin{aligned}\mathrm{Var}(\underline{x}_{01}'\underline{\tilde{\beta}}_1|\underline{z}_0,\ X_1) &= \underline{x}_{01}'\mathrm{Var}(\underline{\tilde{\beta}}_1|X_1)\underline{x}_{01} \\ &= \sigma_p^2\underline{x}_{01}'S_{11}^{-1}\underline{x}_{01}\ /\ (n-1).\end{aligned}$$

Lemma A4.

Lemma A5 $\mathrm{Var}(\tilde{\underline{\beta}}_1) = \sigma_p^2\Sigma_{11}^{-1} \,/\, (n - p - 2).$

Proof: $\mathrm{Var}(\tilde{\underline{\beta}}_1) = \mathrm{Var}\{E(\tilde{\underline{\beta}}_1|X_1)\} + E\{\mathrm{Var}(\tilde{\underline{\beta}}_1|X_1)\},$

Since $E(\tilde{\underline{\beta}}_1|X_1) = \underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2$ Lemma A1

$$\mathrm{Var}\{E(\tilde{\underline{\beta}}_1|X_1)\} = \mathrm{Var}(\underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2)$$

$$= 0$$

Also $\mathrm{Var}(\tilde{\underline{\beta}}_1|X_1) = \sigma_p^2 S_{11}^{-1} \,/\, (n - 1).$ Lemma A4

and $E(S_{11}^{-1}) = (n - 1)\Sigma_{11}^{-1} \,/\, (n - p - 2).$

Therefore $E\{\mathrm{Var}(\tilde{\underline{\beta}}_1|X_1)\} = \sigma_p^2\Sigma_{11}^{-1} \,/\, (n - p - 2),$

and hence $\mathrm{Var}(\tilde{\underline{\beta}}_1) = \sigma_p^2\Sigma_{11}^{-1} \,/\, (n - p - 2).$

Lemma A6 $\mathrm{Var}(\underline{x}_{01}'\tilde{\underline{\beta}}_1|\underline{z}_0) =$

$$\sigma_p^2\{(\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1) + p \,/\, n\} \,/\, (n - p - 2)$$

$$+ \underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 \,/\, n.$$

Proof: $\mathrm{Var}(\underline{x}_{01}'\tilde{\underline{\beta}}_1|\underline{z}_0) = E\{\mathrm{Var}(\underline{x}_{01}'\tilde{\underline{\beta}}_1|\underline{z}_0, X_1)\}$

$$+ \mathrm{Var}\{E(\underline{x}_{01}'\tilde{\underline{\beta}}_1|\underline{z}_0, X_1)\}.$$

Since $E(\underline{x}_{01}'\tilde{\underline{\beta}}_1|\underline{z}_0, X_1) = (\underline{z}_{01} - \bar{\underline{z}}_1)'(\underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2),$ Lemma A2.

$$= (\underline{z}_{01} - \bar{\underline{z}}_1)'\underline{\Phi}_1,$$

since $\underline{\Phi}_1 = \underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2$.

$$\text{Therefore } \mathrm{Var}\{E(\underline{x}_{01}'\tilde{\underline{\beta}}_1 | \underline{z}_0, X_1)\} = \mathrm{Var}\{(\underline{z}_{01} - \bar{\underline{z}}_1)'\underline{\Phi}_1 | \underline{z}_0\}$$

$$= \underline{\Phi}_1'\mathrm{Var}\{(\underline{z}_{01} - \bar{\underline{z}}_1) | \underline{z}_0\}\underline{\Phi}_1$$

$$= \underline{\Phi}_1'(\Sigma_{11} / n)\underline{\Phi}_1$$

$$= \underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 / n.$$

Also $\mathrm{Var}(\underline{x}_{01}'\tilde{\underline{\beta}}_1 | \underline{z}_0, X_1) = \sigma_p^2\underline{x}_{01}'S_{11}^{-1}\underline{x}_{01} / (n - 1)$. Corollary A4.1

$$\text{Therefore } E\{\mathrm{Var}(\underline{x}_{01}'\tilde{\underline{\beta}}_1 | \underline{z}_0, X_1)\} = \sigma_p^2 E(\underline{x}_{01}'S_{11}^{-1}\underline{x}_{01} | \underline{z}_0) / (n - 1)$$

$$= \sigma_p^2\{(\underline{x}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1) + p / n\} / (n - p - 2)$$

Corollary A3.1

$$\text{and hence } \mathrm{Var}(\underline{x}_{01}'\tilde{\underline{\beta}}_1 | \underline{z}_0) = \sigma_p^2\{(\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1) + p / n\} / (n - p - 2)$$

$$+ \underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 / n.$$

<u>Corollary A6.1</u> $\mathrm{Var}(\underline{x}_{01}'\tilde{\underline{\beta}}_1) = (1 + 1/n)\{\underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 + p\sigma_p^2 / (n - p - 2)\}$.

<u>Proof:</u> $\mathrm{Var}(\underline{x}_{01}'\tilde{\underline{\beta}}_1) = E\{\mathrm{Var}(\underline{x}_{01}'\tilde{\underline{\beta}}_1 | \underline{z}_0)\} + \mathrm{Var}\{E(\underline{x}_{01}'\tilde{\underline{\beta}}_1 | \underline{z}_0)\}$.

Since $E(\underline{x}_{01}'\tilde{\underline{\beta}}_1 | \underline{z}_0) = (\underline{z}_{01} - \underline{\mu}_1)'(\underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2)$ Lemma A2

$$= \underline{\Phi}_1'(\underline{z}_{01} - \underline{\mu}_1),$$

since $\underline{\Phi}_1 = \underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2$.

Therefore $\mathrm{Var}\{E(\underline{x}_{01}'\tilde{\underline{\beta}}_1 \,|\, \underline{z}_0)\} = \mathrm{Var}\{(\underline{z}_{01} - \underline{\mu}_1)'\underline{\Phi}_1\}$

$$= \underline{\Phi}_1'\mathrm{Var}(\underline{z}_{01} - \underline{\mu}_1)\underline{\Phi}_1$$

$$= \underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1,$$

and $\mathrm{Var}(\underline{x}_{01}'\tilde{\underline{\beta}}_1 \,|\, \underline{z}_0) = \sigma_p^2\{(\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1) + p / n\} / (n - p - 2)$

$$+ \underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1, \qquad \text{Lemma A6.}$$

$$E\{\mathrm{Var}(\underline{x}_{01}'\tilde{\underline{\beta}}_1 \,|\, \underline{z}_0)\} = \sigma_p^2[E\{(z_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1)\} + p / n] / (n - p - 2)$$

$$+ \underline{\Phi}_1'\Sigma_{11}^{-1}\underline{\Phi}_1$$

$$= \sigma_p^2(p + p / n) / (n - p - 2) + \underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 / n,$$

Corollary A3.3

$$= p(1 + 1/n)\sigma_p^2 / (n - p - 2) + \underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 / n$$

Hence $\mathrm{Var}(\underline{x}_{01}'\tilde{\underline{\beta}}_1) = p(1 + 1/n)\sigma_p^2 / (n - p - 2) + \underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 / n$

$$+ \underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1$$

$$= (1 + 1/n)\{\underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 + p\sigma_p^2 / (n - p - 2)\}.$$

Lemma A7 $E\{(\underline{x}_{01}'\tilde{\underline{\beta}}_1)^2 | \underline{z}_0\}$

$$= \sigma_p^2\{(\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1) + p / n\} \; / \; (n - p - 2)$$
$$+ \underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 / n \; + \underline{\Phi}_1'(\underline{z}_{01} - \underline{\mu}_1)(\underline{z}_{01} - \underline{\mu}_1)'\underline{\Phi}_1.$$

Proof: $E\{(\underline{x}_{01}'\tilde{\underline{\beta}}_1)^2 | \underline{z}_0\} = \mathrm{Var}(\underline{x}_{01}'\tilde{\underline{\beta}}_1 | \underline{z}_0) + \{E(\underline{x}_{01}'\tilde{\underline{\beta}}_1 | \underline{z}_0)\}^2$

$$= \sigma_p^2\{(\underline{z}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1) + p / n\} \; / \; (n - p - 2)$$

$$+ \underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 / n + \underline{\Phi}_1'(\underline{z}_{01} - \underline{\mu}_1)(\underline{z}_{01} - \underline{\mu}_1)'\underline{\Phi}_1,$$

since $\mathrm{Var}(\underline{x}_{01}'\tilde{\underline{\beta}}_1 | \underline{z}_0) = \sigma_p^2\{(\underline{x}_{01} - \underline{\mu}_1)'\Sigma_{11}^{-1}(\underline{z}_{01} - \underline{\mu}_1) + p / n\} / (n - p - 2)$

$+ \underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 / n,$ Lemma A6,

and $E(\underline{x}_{01}'\tilde{\underline{\beta}}_1 | \underline{z}_0) = (\underline{z}_{01} - \underline{\mu}_1)'(\underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2),$ Lemma A2,

$= (\underline{z}_{01} - \underline{\mu}_1)'\underline{\Phi}_1,$

since $\underline{\Phi}_1 = \underline{\beta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\underline{\beta}_2.$

Corollary A7.1 $E(\underline{x}_{01}'\tilde{\underline{\beta}}_1)^2 = (1 + 1/n)\{\underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 + p\sigma_p^2 / (n - p - 2)\}$

Proof: $E(\underline{x}_{01}'\tilde{\underline{\beta}}_1)^2 = \mathrm{Var}(\underline{x}_{01}'\tilde{\underline{\beta}}_1) + \{E(\underline{x}_{01}'\tilde{\underline{\beta}}_1)\}^2$

$$= (1 + 1/n)\{\underline{\Phi}_1'\Sigma_{11}\underline{\Phi}_1 + p\sigma_p^2 / (n - p - 2)\},$$

since $\mathrm{Var}(\underline{x}_{01}'\tilde{\underline{\beta}}_1) = (1 + 1/n)\{\underline{\Phi}_1'\Sigma_{11}\underline{\Sigma}_1 + p\sigma_p^2 / (n - p - 2)\}$

, Corollary A6.1,

and $E(\underline{x}_{01}'\tilde{\underline{\beta}}_1) = 0$, Corollary A2.1.

<u>Lemma A8</u> $\mathrm{Cov}(\bar{y}, \underline{x}_{01}'\tilde{\underline{\beta}}_1 | \underline{z}_0, X_1) = 0.$

<u>Proof</u>: $\mathrm{Cov}(\bar{y}, \underline{x}_{01}'\tilde{\underline{\beta}}_1 | \underline{z}_0, X_1)$

$$= \mathrm{Cov}(\bar{y}, \underline{s}_1' S_{11}^{-1} \underline{x}_{01} | \underline{z}_0, X_1)$$

$$= \mathrm{Cov}(\Sigma_m y_m \;/\; n, \Sigma_\ell (y_\ell - \bar{y}) \underline{x}_\ell S_{11}^{-1} \underline{x}_{01} \;/\; (n - 1) | \underline{z}_0, X_1)$$

$$= \mathrm{Cov}(\Sigma_m y_m \;/\; n, \Sigma_\ell y_\ell \underline{x}_\ell S_{11}^{-1} \underline{x}_{01} \;/\; (n - 1) | \underline{z}_0, X_1)$$

$$= \Sigma_{m,\ell} \mathrm{Cov}(y_m, y_\ell | \underline{z}_0, X_1) \underline{x}_\ell S_{11}^{-1} \underline{x}_{01} \;/\; n(n - 1),$$

$$\text{since } \mathrm{Cov}(y_m, y_\ell) = \begin{cases} \sigma_p^2 & m = \ell \\ 0 & m \neq \ell \end{cases},$$

$$\mathrm{Cov}(\bar{y}, \underline{x}_{01}'\tilde{\underline{\beta}}_1 | \underline{z}_0, X_1) = \Sigma_m \sigma_p^2 \underline{x}_m S_{11}^{-1} \underline{x}_{01} \;/\; n(n - 1)$$

$$= \sigma_p^2 (\Sigma_m \underline{x}_m) S_{11}^{-1} \underline{x}_{01} \;/\; n(n - 1)$$

$$= 0 \qquad , \qquad \text{since } \Sigma_m \underline{x}_m = 0.$$

<u>Corollary A8.1</u> $E\{\mathrm{Cov}(\bar{y}, \underline{x}_{01}'\tilde{\underline{\beta}}_1 | \underline{z}_0, X_1)\} = 0$

<u>Proof</u>: Follows immediately from lemma A8, since

$$\mathrm{Cov}(\bar{y}, \underline{x}_{01}'\tilde{\underline{\beta}}_1 | \underline{z}_0, X_1) = 0.$$

<u>Lemma A9</u> $\underline{\sigma}'\cdot\Sigma^{-1}\begin{bmatrix}\Sigma_{11}\\ \\ \Sigma_{21}\end{bmatrix} = \underline{\sigma}_1'.$

<u>Proof:</u> Since $\Sigma^{-1}\Sigma = I$

$$\Sigma^{-1}\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix} = \begin{bmatrix}I_p & 0\\ \\ 0 & I_{k-p}\end{bmatrix}.$$

Hence $$\Sigma^{-1}\begin{bmatrix}\Sigma_{11}\\ \\ \Sigma_{21}\end{bmatrix} = \begin{bmatrix}I_p\\ \\ 0\end{bmatrix}$$

Therefore $$\underline{\sigma}'\Sigma^{-1}\begin{bmatrix}\Sigma_{11}\\ \\ \Sigma_{21}\end{bmatrix} = [\underline{\sigma}_1', \underline{\sigma}_2']\begin{bmatrix}I_p\\ \\ 0\end{bmatrix}$$

$$= \underline{\sigma}_1'.$$

APPENDIX B

THE DATA

APPENDIX B

THE DATA

In this appendix, we give the data used for the examples discussed in Chapter 4.

The data for the Gorman-Toman problem is given in Figure 4, In Table 7, Hald's data is given, followed by the ACT data in Table 8. For the analyses of the ACT data, we give the observations used to estimate the parameters for the first run in Table 9 and for the second run in Table 10.

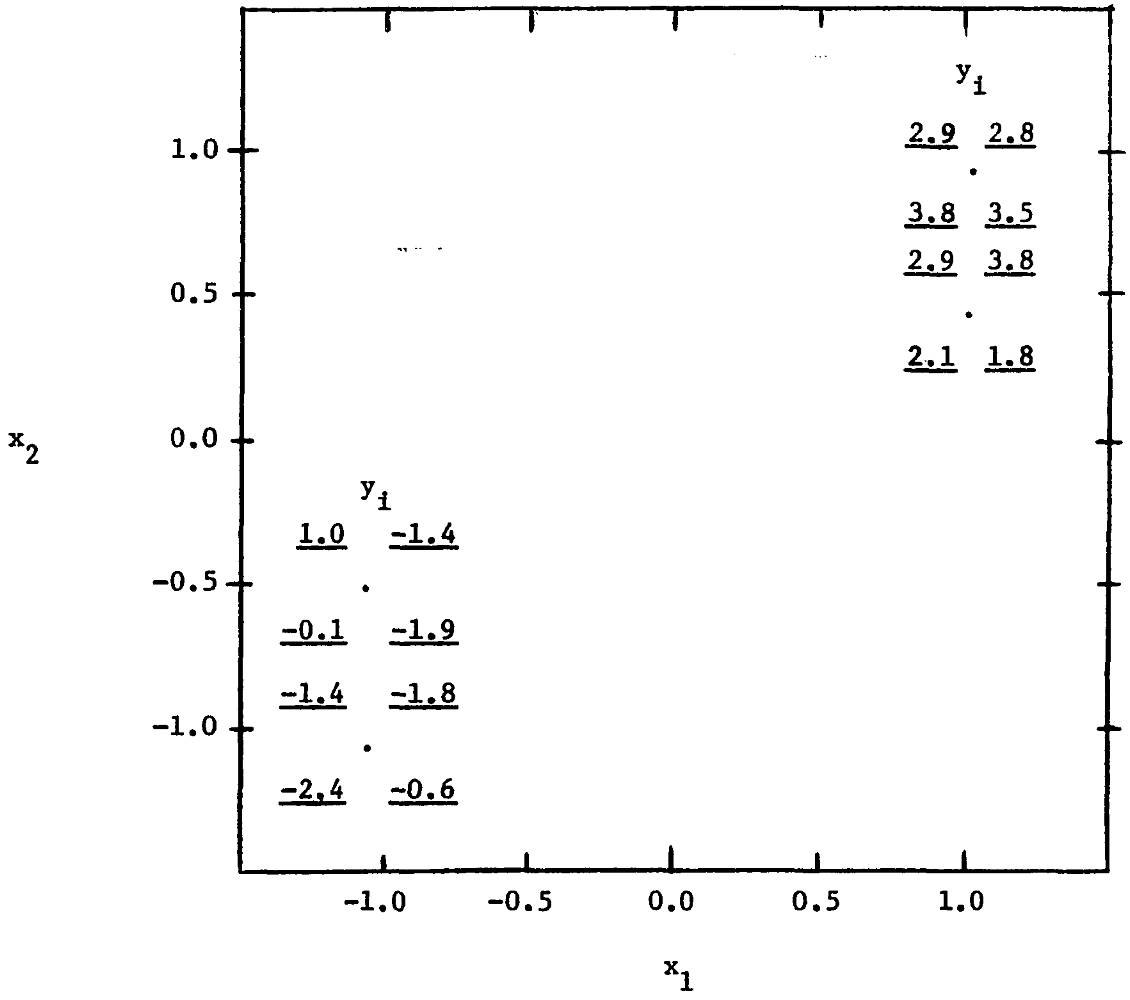

FIGURE 4

GORMAN-TOMAN PROBLEM DATA

TABLE 7

HALD'S DATA

Obs. No.	y[a]	x_1[b]	x_2[c]	x_3[d]	x_4[e]
1	78.5	7.0	26.0	6.0	60.0
2	74.3	1.0	29.0	15.0	52.0
3	104.3	11.0	56.0	8.0	20.0
4	87.6	11.0	31.0	8.0	47.0
5	95.9	7.0	52.0	6.0	33.0
6	109.2	11.0	55.0	9.0	22.0
7	102.7	3.0	71.0	17.0	6.0
8	72.5	1.0	31.0	22.0	44.0
9	93.1	2.0	54.0	18.0	22.0
10	115.9	21.0	47.0	4.0	26.0
11	83.8	1.0	40.0	23.0	34.0
12	113.3	11.0	66.0	9.0	12.0
13	109.4	10.0	68.0	8.0	12.0

[a] y = heat evolved in calories per gram of cement. X_1, X_2, X_3, and X_4 are measured as per cent of the weight of the clinkers from which the cement was made.

[b] X_1 = amount of tricalcium aluminate, $3CaO \cdot Al_2O_3$.

[c] X_2 = amount of tricalcium silicate, $3CaO \cdot SiO_2$.

[d] X_3 = amount of calcium aluminum ferrate, $4CaO \cdot Al_2O_3 \cdot Fe_2O_3$.

[e] X_4 = amount of dicalcium silicate, $2CaO \cdot SiO_2$.

TABLE 8

THE ACT DATA

Obs. No.	y^a	x_1^b	x_2^c	x_3^d	x_4^e
1	0.1	3.0	1.0	9.0	14.0
2	1.7	9.0	25.0	18.0	12.0
3	2.3	17.0	13.0	19.0	19.0
4	0.5	11.0	12.0	15.0	18.0
5	0.8	14.0	12.0	18.0	22.0
6	3.1	19.0	25.0	24.0	27.0
7	1.7	15.0	20.0	21.0	23.0
8	3.1	18.0	23.0	22.0	23.0
9	1.7	9.0	13.0	9.0	15.0
10	2.1	19.0	20.0	20.0	20.0
11	1.1	16.0	12.0	20.0	17.0
12	3.5	19.0	28.0	25.0	27.0
13	1.9	20.0	2.0	20.0	18.0
14	1.4	11.0	11.0	14.0	4.0
15	0.2	8.0	17.0	12.0	13.0
16	2.5	18.0	23.0	20.0	24.0
17	3.0	20.0	18.0	24.0	22.0
18	1.0	18.0	20.0	26.0	27.0
19	0.5	14.0	14.0	7.0	11.0
20	0.1	10.0	18.0	8.0	14.0
21	3.0	20.0	28.0	24.0	25.0
22	1.5	16.0	27.0	8.0	23.0
23	0.7	13.0	14.0	7.0	5.0
24	2.4	17.0	13.0	17.0	23.0
25	2.0	16.0	16.0	19.0	20.0
26	1.8	17.0	20.0	30.0	27.0
27	2.9	13.0	21.0	21.0	15.0
28	2.0	20.0	15.0	22.0	18.0
29	1.9	5.0	9.0	8.0	13.0
30	2.5	1.0	19.0	1.0	14.0
31	0.7	17.0	8.0	16.0	14.0
32	2.6	16.0	19.0	24.0	29.0
33	1.7	13.0	11.0	14.0	12.0
24	1.7	13.0	17.0	14.0	13.0
35	2.4	16.0	27.0	24.0	22.0
36	3.3	19.0	28.0	26.0	30.0
37	1.2	11.0	1.0	5.0	11.0
38	1.3	20.0	17.0	23.0	24.0
39	2.3	14.0	17.0	11.0	16.0
40	1.7	13.0	17.0	21.0	16.0
41	1.1	14.0	21.0	16.0	23.0
42	3.3	22.0	23.0	25.0	27.0

TABLE 8 (Cont'd.)

Obs. No.	y^a	x_1^b	x_2^c	x_3^d	x_4^e
43	1.0	13.0	10.0	16.0	18.0
44	2.1	8.0	19.0	14.0	14.0
45	1.7	8.0	21.0	16.0	15.0
46	3.2	25.0	28.0	27.0	31.0
47	3.8	28.0	28.0	27.0	29.0
48	1.4	17.0	15.0	14.0	19.0
49	1.8	18.0	22.0	17.0	18.0
50	2.1	13.0	23.0	15.0	14.0
51	3.3	15.0	19.0	18.0	23.0
52	0.2	10.0	5.0	2.0	8.0
53	0.7	1.0	2.0	3.0	1.0
54	2.8	15.0	14.0	15.0	19.0
55	2.8	17.0	16.0	17.0	25.0
56	2.8	20.0	23.0	16.0	21.0
57	1.6	11.0	17.0	20.0	20.0
58	1.6	15.0	13.0	22.0	10.0
59	3.2	22.0	29.0	23.0	27.0
60	2.7	17.0	21.0	23.0	22.0
61	2.3	15.0	12.0	11.0	18.0
62	3.4	21.0	22.0	22.0	13.0
63	2.0	13.0	13.0	16.0	11.0
64	2.6	19.0	25.0	12.0	21.0
65	2.1	16.0	15.0	21.0	18.0
66	1.3	22.0	22.0	22.0	22.0
67	2.4	17.0	20.0	21.0	27.0
68	2.0	14.0	11.0	17.0	14.0
69	3.1	20.0	21.0	23.0	27.0
70	2.4	14.0	18.0	18.0	16.0
71	2.4	19.0	17.0	24.0	20.0
72	1.4	15.0	13.0	20.0	25.0
73	2.0	13.0	15.0	14.0	20.0
74	1.6	10.0	23.0	6.0	10.0
75	2.1	10.0	14.0	9.0	11.0
76	2.8	20.0	31.0	22.0	26.0
77	4.0	22.0	27.0	29.0	32.0
78	1.3	12.0	14.0	9.0	14.0
79	1.4	10.0	13.0	20.0	19.0
80	2.0	16.0	21.0	11.0	14.0
81	3.3	20.0	28.0	23.0	24.0
82	1.8	11.0	19.0	18.0	13.0
83	2.5	14.0	20.0	20.0	14.0

TABLE 8 (Cont'd.)

[a] y = first year college grade point average.

[b] x_1 = score on English test.

[c] x_2 = score on Mathematics test.

[d] x_2 = score on Social Sciences test.

[e] x_4 = score on Natural History Test.

TABLE 9

ACT DATA: OBSERVATIONS USED IN THE FIRST RUN

Sample Size			
10	15	20	25
Obs. No.			
12	12	12	12
16	16	16	16
21	21	21	21
23	23	23	23
40	40	40	40
42	42	42	42
47	47	47	47
68	68	68	68
76	76	76	76
77	77	77	77
	46	46	46
	51	51	51
	53	53	53
	82	82	82
	83	83	83
		29	29
		33	33
		35	35
		54	54
		57	57
			17
			20
			37
			65
			73

TABLE 10

ACT DATA: OBSERVATIONS USED IN THE SECOND RUN

Sample Size			
10	15	20	25
Obs. No.			
17	16	12	12
20	17	16	16
29	20	17	17
33	29	20	20
35	33	21	21
46	35	35	23
51	37	37	29
53	46	40	33
65	54	42	35
83	57	47	37
	65	51	40
	68	53	42
	73	54	46
	82	57	47
	83	65	51
		73	53
		76	54
		77	57
		82	65
		83	68
			73
			76
			77
			82
			83

The following 58 observations were predicted each time:

1 - 11, 13 - 15, 18, 19, 22, 24 - 28, 30 - 32, 34, 36, 38, 39, 41, 43 - 45, 48 - 50, 52, 55, 56, 58 - 64, 66, 67, 69 - 72, 74, 75, 78 - 81.