CONTRIBUTIONS TO THE THEORY OF BIASED ESTIMATION OF LINEAR MODELS

by

David J. Svendsgarrd

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1162

JUNE 1978

DAVID JOHN SVENDSGAARD. Contributions to the Theory of Biased Estimation of Linear Models. (Under the direction of LAWRENCE L. KUPPER.)

This study investigates estimators that may be used instead of the usual least squares estimator for the parameters of a linear model. These estimators are linear transforms of the least squares estimator and they are usually biased.

The motivation to use a biased estimator is either a desire to obtain an improved estimator or to satisfy a constraint. We use a criterion called General Mean Square Error (GMSE). When there is a constraint we only consider the class of estimators that satisfy it. We then proceed to determine the estimator with the smallest GMSE within this class.

The linear transform of the least squares estimator, which defines this estimator with the smallest GMSE, is usually a function of unknown parameters. Therefore, such an estimator is not useable. However, consideration of the form of this optimal estimator together with the availability of prior information of bounds on the parameter space can result in the development of a useful biased estimator that is superior to many of the other estimators in the class for some values of the parameters.

Rather than base the choice of the transformation on prior knowledge, one can let the data choose the transform. For example, the Stein estimator, the ridge estimator and stepwise regression can be viewed as stochastic linear transforms of the full model least squares estimator.

Some of these estimators are compared and applications are considered.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I wish to express particular gratitude to Dr. Lawrence L. Kupper and Dr. Ronald W. Helms for the guidance that made this study possible. I also wish to thank the other members of my Doctoral Committee — Dr. R.R. Kuebler, Dr. M.A. Shiffman, and Dr. H.A. Tyroler for their various contributions. I also want to thank Susan Stapleton for typing the manuscript.

Finally, a special tribute must be given my wife, Sylvia, who encouraged me and shared the frustrations.

This research was supported by the Department of Biostatistics through the National Institute of Environmental Health Sciences Training Grant 3-T01ES00133 Biostatistics for Reseach and Environmental Health.

LIST OF TABLES

LIST OF FIGURES

Chapter I

GENERAL CONSIDERATIONS

1.1 Introduction

When estimating the parameters of a linear model, one generally has available an unbiased estimator. The least squares estimator is an unbiased estimator when fitted to the true model. Under the assumption of uncorrelated random errors with zero means, the least squares estimator of any linear combination of the parameters has the smallest variance of any unbiased estimator that is a linear combination of the observations. However, smaller mean square error of an estimator for some linear combinations of the parameters is obtainable if a biased estimator is considered. In this dissertation, biased estimators are explored. These estimators are later compared in terms of their mean square error.

The literature of response surface analysis has been motivated by similar considerations. We find the setting of response surface analysis a convenient starting point to discuss this problem.

1.2 Response Surface Methodology

Response surface analysis is concerned with a functional relationship

$$\eta = f(\xi_1, \xi_2, \ldots, \xi_q; \theta_1, \theta_2, \ldots, \theta_r) = f(\underset{\sim}{\xi}, \underset{\sim}{\theta})$$

between a response variable η, and q independent variables $\xi_1, \xi_2, \ldots, \xi_q$. The relationship involves the parameters $\Theta_1, \Theta_2, \ldots, \Theta_r$. These parameters are generally unknown, but can be estimated from experimental data. Typically, such data will consist of N measured values of η corresponding to N specified combinations of levels of the ξ's.

The response relationship is usually interpreted geometrically as representing a surface in a space whose coordinates are the $q + 1$ variables $\xi_1, \ldots, \xi_q$ and η. The N experimental combinations of levels of the ξ_i's are represented by points in the space of the independent variables. The collection of these N points

$$\{\underset{\sim}{\xi}_u\}_{u=1}^{N}, \; \underset{\sim}{\xi}' = (\xi_{1u}, \xi_{2u}, \ldots, \xi_{qu}),$$

is called the experimental design.

In practice, the form of the response function $f(\underset{\sim}{\xi}, \underset{\sim}{\theta})$ is generally unknown and the usual procedure is to assume that the unknown function can be represented at all points $\underset{\sim}{\xi}$ by a polynomial in $\underset{\sim}{\xi}$. Justification for considering the polynomial model comes from the fact that functions with certain smoothness properties can be approximated about a point by partial sum of a Taylor Series. Therefore, we shall consider that the true model can be written as

$$\eta = \underset{1\text{xp}}{\underset{\sim}{x}'} \; \underset{\text{px}1}{\underset{\sim}{\beta}}$$

where $\underset{\sim}{\beta}$ is the vector of coefficients of the polynomial and the elements of $\underset{\sim}{x}$ are standardized functions of the variables in $\underset{\sim}{\xi}$. Specifically, the ith element of the vector x_u corresponding to the point $\underset{\sim}{\xi}_u$ is

$$x_{iu} = \prod_{j=1}^{q} \left(\frac{\xi_{ju}-\bar{\xi}_j}{s_j} \right)^{h_{ij}} ,$$

where $\bar{\xi}_j = \frac{1}{N} \sum_{u=1}^{N} \xi_{ju}$, the s_j are scale factors, and the h_{ij}'s are non-negative integers. For example, if $\underset{\sim}{\xi}'_u = (\xi_{1u}, \xi_{2u})$, $\bar{\xi}_1 = \bar{\xi}_2 = 0$, and $s_j = 1$ then for a second order polynomial (p=6) we might have

$$\underset{\sim}{x}'_u = (1, \xi_{1u}, \xi_{2u}, \xi_{1u}^2, \xi_{2u}^2, \xi_{1u}\xi_{2u}) .$$

In this case, the matrix whose ith jth element is h_{ij} is

$$\begin{bmatrix} 0 & 1 & 0 & 2 & 0 & 1 \\ 0 & 0 & 1 & 0 & 2 & 1 \end{bmatrix} .$$

Let $\underset{\sim}{X}$ (Nxp) be the matrix whose ith, uth element is x_{iu} and $\underset{\sim}{Y}$ (Nx1) be the vector of uncorrelated responses whose uth element is observed when the level of the factors is $\underset{\sim}{\xi}_u$. If $\underset{\sim}{\varepsilon}$ is the vector of deviations of $\underset{\sim}{Y}$ from $\underset{\sim}{X}\underset{\sim}{\beta}$ then

$$\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} .$$

We assume these deviations are random uncorrelated variables with mean zero and variance σ^2. We also assume $\underset{\sim}{X}$ has full column rank, that is, $(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$ exists.

Given this model, it is possible to derive the conclusions of the Gauss-Markov Theorem:

"The variance of $\underset{\sim}{x}'\underset{\sim}{b}$ is smaller when $\underset{\sim}{b} = \hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}$ than for any other unbiased estimator (of $\underset{\sim}{\beta}$) which is a linear combination of the elements of the vector $\underset{\sim}{Y}$."

Thus, the guarantees of $\underset{\sim}{x}'\hat{\underset{\sim}{\beta}}$ provided by the Gauss-Markov Theorem are true regardless of whether or not $\underset{\sim}{x}'$ is in the sampled region and regardless of physical constraints that restrict the range of $\underset{\sim}{x}'\underset{\sim}{\beta}$. Since these specifications are not a necessary prerequisite to using the conclusions of the Gauss-Markov Theorem, there is no consideration that for some $\underset{\sim}{x}$, $\underset{\sim}{x}'\hat{\underset{\sim}{\beta}}$ may be less reliable than for other $\underset{\sim}{x}$. Also there is no guarantee that $\hat{\underset{\sim}{\beta}}$ will not take on nonsensical values.

Although the use of $\hat{\underset{\sim}{\beta}}$ may be justified by means of various criteria, one overriding characteristic $\hat{\underset{\sim}{\beta}}$ possesses is:

"Of the estimators which are linear combinations of the $\underset{\sim}{Y}$, $\hat{\underset{\sim}{\beta}}$ best fits the model to the data." That is, the sum of squared errors of fit

$$\phi(\underset{\sim}{b}) = (\underset{\sim}{Y}-\underset{\approx}{X}\underset{\sim}{b})'(\underset{\sim}{Y}-\underset{\approx}{X}\underset{\sim}{b}) \tag{1.2.1}$$

is minimized when $\underset{\sim}{b} = \hat{\underset{\sim}{\beta}}$. Since $\underset{\sim}{Y} - \underset{\approx}{X}\hat{\underset{\sim}{\beta}}$ is unexplained error, $\phi(\underset{\sim}{b})$ is a measure of precision. When p is large due to fitting a high degree polynomial, experience has taught that unreliable estimates of η outside the region of experimentation can occur. Since high degree polynomials make $\phi(\underset{\sim}{b})$ small, the unwarranted use of high degree polynomials treats the data as being more precise than the data usually are. That is, if $\underset{\sim}{\beta}$ were known, then $\phi(\underset{\sim}{\beta})$ would be larger than $\phi(\hat{\underset{\sim}{\beta}})$ on the average. That is,

$$(N-p)\sigma^2 = E(\phi(\hat{\underset{\sim}{\beta}})) < E(\phi(\underset{\sim}{\beta})) = N\sigma^2 .$$

In summary, the sum of squared deviations of $\underset{\sim}{Y}$ from the model fitted by the least squares estimator are smaller on the average than the actual deviations are. The more parameters that are estimated, the more the discrepancy between these two sums of squares.

There are various methods that can be used to better balance the fitting and extrapolation errors. The judicious choice of a low order polynomial and/or variable selection routines often yield models whose extrapolation characteristics will be better behaved. The choice of the model used to fit the data may be based on considerations of regions of interest of the ξ variables. For example, the assumption of linearity might be plausible over certain regions but violate physical laws when the real line is considered. Regardless of the motivation, we assume that one has specified a region of interest R of the $\underset{\sim}{\xi}$ variables over which it is desired to approximate η by some approximating function η_a. We will consider the case where η_a is related to η as

$$\eta = \underset{\sim}{x}'\underset{\sim}{\beta} = \underset{\sim}{x}_1'\underset{\sim}{\beta}_1 + \underset{\sim}{x}_2'\underset{\sim}{\beta}_2 = \eta_a + \underset{\sim}{x}_2'\underset{\sim}{\beta}_2$$

so that $\eta_a = \underset{\sim}{x}_1'\underset{\sim}{\beta}_1$ is the form of the model to be fit.

A general form of criterion that has been used for variable selection, smoothing of the response observations, selection of the experimental design, and parameter estimation is general mean square error

$$J(\underset{\sim}{M}, \underset{\sim}{b}) = E(\underset{\sim}{b} - \underset{\sim}{\beta})' \underset{\sim}{M}(\underset{\sim}{b} - \underset{\sim}{\beta})$$

$$= \operatorname{tr} \underset{\sim}{M}\, E(\underset{\sim}{b} - \underset{\sim}{\beta})(\underset{\sim}{b} - \underset{\sim}{\beta})' \ .$$

$\underset{\sim}{M}_{pxp}$ is invariably taken to be a positive definite symmetric matrix.

The integrated mean square error (IMSE) is a special case of the form $J(\underset{\sim}{M},\underset{\sim}{b})$. To see this, let R be the region of interest. The IMSE of $\underset{\sim}{b}$ is

$$J = \frac{N\Omega}{\sigma^2} \int_R E(\underset{\sim}{x}'\underset{\sim}{b} - \underset{\sim}{x}'\underset{\sim}{\beta})^2 \, d\underset{\sim}{x}$$

where $\Omega^{-1} = \int_R d\underset{\sim}{x}$. If we denote the moment matrix as

$$\underset{\sim}{W} = \Omega \int_R \underset{\sim}{x}\underset{\sim}{x}' d\underset{\sim}{x}$$

then

$$J = \frac{N\Omega}{\sigma^2} \int_R E(\underset{\sim}{x}'\underset{\sim}{b} - \underset{\sim}{x}'\underset{\sim}{\beta})^2 \, d\underset{\sim}{x}$$

$$= \frac{N}{\sigma^2} \text{ tr } \underset{\sim}{W} \, E(\underset{\sim}{b}-\underset{\sim}{\beta})(\underset{\sim}{b}-\underset{\sim}{\beta})'$$

$$= J\left(\frac{N\underset{\sim}{W}}{\sigma^2}, \underset{\sim}{b} \right).$$

It can be shown that

$$J = V + B$$

where $V = \frac{N\Omega}{\sigma^2} \int_R \text{Var}(\underset{\sim}{x}'\underset{\sim}{b}) d\underset{\sim}{x}$,

and $B = \frac{N\Omega}{\sigma^2} \int_R (E(\underset{\sim}{x}'\underset{\sim}{b}) - \underset{\sim}{x}'\underset{\sim}{\beta})^2 \, d\underset{\sim}{x}$.

V is the integrated variance of $\underset{\sim}{x}'\underset{\sim}{b}$ and B is the integrated bias error of $\underset{\sim}{x}'\underset{\sim}{b}$.

A similar partition exists for $J(\underset{\sim}{M}, \underset{\sim}{b})$. We shall denote these partitions by V and B and call them the variance error and the bias error (with respect to $\underset{\sim}{M}$).

When the model is written in the form:

$$\eta = \underset{\sim}{x}_1'\underset{\sim}{\beta}_1 + \underset{\sim}{x}_2'\underset{\sim}{\beta}_2 \tag{1.2.2}$$

we will denote the submatrices of $\underset{\sim}{W}$ corresponding to this partitioning of $\underset{\sim}{x}$ as

$$\underset{\sim}{W} = \begin{bmatrix} \underset{\sim}{W}_1 & \underset{\sim}{W}_2 \\ \underset{\sim}{W}_2' & \underset{\sim}{W}_3 \end{bmatrix} .$$

In a situation where the model is further partitioned

$$\eta = (\underset{\sim}{x}_1' \,\vdots\, \underset{\sim}{x}_{2a}' \,\vdots\, \underset{\sim}{x}_{2b}') \begin{bmatrix} \underset{\sim}{\beta}_1 \\ \underset{\sim}{\beta}_{2a} \\ \underset{\sim}{\beta}_{2b} \end{bmatrix} , \qquad (1.2.3)$$

we denote the submatrices by

$$\underset{\sim}{W} = \begin{bmatrix} \underset{\sim}{W}_1 & \underset{\sim}{W}_{2a} & \underset{\sim}{W}_{2b} \\ \underset{\sim}{W}_{2a} & \underset{\sim}{W}_{aa} & \underset{\sim}{W}_{ab} \\ \underset{\sim}{W}_{2b} & \underset{\sim}{W}_{ab}' & \underset{\sim}{W}_{bb} \end{bmatrix} .$$

The partitioning of $\underset{\sim}{X}$ corresponding to the partitioning in (1.2.2) will be denoted by

$$\underset{\sim}{X} = (\underset{\sim}{X}_1 \,\vdots\, \underset{\sim}{X}_2),$$

and corresponding to (1.2.3) by

$$\underset{\sim}{X} = (\underset{\sim}{X}_1 \,\vdots\, \underset{\sim}{X}_{2a} \,\vdots\, \underset{\sim}{X}_{2b}) .$$

1.3 Review of the Literature

Many estimators for the parameters of the linear model can be viewed as a linear transform of the least squares estimator of the full model. That is,

$$\underset{\sim}{b} = \underset{\sim}{K}\,\hat{\underset{\sim}{\beta}} \ .$$

The motivation for considering such an estimator comes from one of two sources. One source is the desire to obtain a better estimator than the least squares estimator. The second source is a knowledge of constraints or limitations on the parameter space. There is also the possibility that both sources of motivation are acting.

These sources of motivation closely coincide with the three aims of regression analysis

- Parameter estimation
- Prediction or smoothing
- Identification of explanatory variables.

Under parameter estimation, the full model least squares estimator is generally available. The motivation here is usually to obtain a better estimator. Under prediction or smoothing, there may be a constraint that makes the full model least squares estimator unacceptable. Under identification of explanatory variables, it is often the case that both sources of motivation are involved.

In our discussion, $\underset{\sim}{K}$ will be either stochastic or deterministic. Usually, a stochastic $\underset{\sim}{K}$ is used to obtain an estimator that is better than the least squares estimator. However, when prior information about the parameters is available, a deterministic $\underset{\sim}{K}$ may be considered for purposes of obtaining an improved estimator. Classically, a deterministic $\underset{\sim}{K}$ is used to satisfy a constraint.

First, let us consider a deterministic $\underset{\sim}{K}$ when no prior information is available and one's goal is to obtain an estimator with better general mean square error characteristics than full model. Using this criterion Theil (1971) has shown that the optimal choice of $\underset{\sim}{K}$ involves the unknown coefficients. Thus, it appears that knowledge of the unknown parameters is a prerequisite to obtaining an improved estimator under such conditions. Moreover, Barnard (1963) has shown that when the range of $\underset{\sim}{x}'\underset{\sim}{\beta}$ is unbounded then among estimators having bounded mean square error, the least squares estimator $\underset{\sim}{x}'\hat{\underset{\sim}{\beta}}$ has minimum mean square error. Ignoring other considerations, then it is only in those instances when $\underset{\sim}{x}'\underset{\sim}{\beta}$ is bounded that a biased estimator can compete with the least squares estimator in terms of mean square error.

Let us now consider stochastic choices of $\underset{\sim}{K}$.

In the multivariate normal setting when there are 3 or more variables, Stein (1960) has exhibited a whole class of estimators for the mean vector that have mean square error characteristics smaller than the least squares estimator. Motivated by the results of Stein, Sclove (1968) has considered the univariate regression situation where $p > 3$ and the error is normally distributed. When the design matrix is orthogonal, Sclove has shown there is a class of estimators of $\underset{\sim}{\beta}$ having smaller general mean square error than the least squares estimator. He shows that this result holds for any form of general mean square error as long as $\underset{\sim}{M}$ is a symmetric positive definite matrix.

Now let us return to considerations involved with using a deterministic $\underset{\sim}{K}$.

For example, the estimator obtained by only fitting a subset of the variables of the true model using the least squares estimator is a transform of the full model least squares estimator. The IMSE of such an estimator depends on the design.

Using integrated mean square error (IMSE) criteria, Box and Draper (1959) (BD) considered choosing the design that would minimize the bias error portion of IMSE when a fixed subset of the variables of the true model was fit by least squares. To achieve minimum bias error they found that certain moments of the design had to agree with the corresponding moments of a uniform distribution on R.

Whether or not the Box and Draper scheme achieves smaller IMSE than obtained by the least squares estimator fitting the full model depends on the variance σ^2 and the magnitude of the coefficients associated with the variables that are not included in the selected subset.

Karson, Manson and Hader (KMH) (1969) minimized the bias error of IMSE by choice of estimator rather than choice of design. Any remaining design flexibility could be used to minimize the variance error portion of IMSE. When designs are used that make all the coefficients of the full model estimable, the KMH scheme always achieves smaller IMSE than does the BD scheme. However, the IMSE of the least squares estimator of the full model may be smaller depending on the coefficients associated with variables not included in the selected subset. This is to be expected in view of the work of Barnard which was referred to earlier.

Finally, we shall consider the use of a deterministic $\underset{\sim}{K}$ when prior information is available. This problem has received relatively little attention. A reason that it has not received more attention is the subjectivity involved here. We assume that one has prior information that can be expressed as a bound on the magnitude of some combination of the parameters of the model.

Sometimes a natural bound is apparent. More often, the prior information is not so easy to specify. In such a situation, the easiest thing to do is to ignore prior information. When the purpose of the analysis is to report the data, ignoring prior information has an air of objectivism to it.

When the objective is to use the results of the analysis for prediction or parameter estimation, ignoring prior information can lead to failure. This failure can result from an intolerably large mean square error of the estimator of $\underset{\sim}{x}'\underset{\sim}{\beta}$ over the range of reasonable values that $\underset{\sim}{x}'\underset{\sim}{\beta}$ can assume.

Usually, prior information is incorporated into an estimation strategy using a Bayesian framework. This requires specification of a prior density on the parameters.

It is felt that specification of a bound on certain combinations of the parameters is a more straightforward decision than specification of a prior density.

With this in mind, Kupper and Meydrech (KM)(1975) considered the problem of minimizing IMSE when a fixed subset of the variables of the true model are fit. Instead of using the least squares estimator of the parameter vector as Box and Draper did, they considered shrinking the least square estimator. The amount of shrinkage was based on an a priori bound on parameters of the variables that were excluded from the fitted model.

Design considerations were also incorporated into their scheme in order to reduce IMSE. They found that when their a priori bound was correct, their method resulted in smaller IMSE than was achieved by the Box and Draper method. Smaller IMSE was also achievable even when the parameters exceeded the presumed bound by a factor of two.

In a quite different context, Hoerl and Kennard (HK)(1970a) developed the ridge estimator. The use of the ridge estimator was developed to identify explanatory variables when the variables of the design matrix were highly correlated. They compared their estimator in terms of general mean square error. Specifically, they considered the mean of the squared length of the deviations of the estimator from the true parameter values. In terms of general mean squared error, this is equivalent to using $\underset{\sim}{M} = \underset{\sim}{I}$.

They found that the ridge estimator was better (in terms of the criteria) than the least squares estimator when a certain bound on the parameters held.

The ridge trace is a plot of the elements of the ridge estimator versus a scalar that defines a class of possible ridge estimators. The ridge trace is used to choose the ridge estimator from this class. Since this choice involves the data, their ridge estimator is actually a stochastic transform of the least squares estimator. Since their comparison was based on the transform being deterministic, this comparison is open to question.

1.4 An Outline of the Study

The preceding literature review has briefly described the estimators which we shall study. Certain aspects of these estimators will be dealt with in more detail in Chapter 2. In that chapter, we shall be noting a general similarity of these estimators while keeping in mind the purpose of these estimators. We believe that it is important to distinguish these purposes, since our eventual goal is to compare estimators.

Consider the difficulties that can occur if these purposes are not distinguished when comparing estimators. The way that comparisons are made places certain restrictions on the general validity of results. When one considers the varied purposes of these estimators, it becomes clear that the results of comparisons are valid under only special circumstances. In the extreme case, when there are no circumstances when one would consider either estimator appropriate, the result of comparing two estimators is of no interest.

To avoid states of affairs of this nature, we shall classify estimators by purpose in Chapter two. When we compare estimators, we shall be comparing only estimators that are within the same purpose group.

In Chapter three we shall describe the new results we have obtained about estimators that are transforms of the least squares estimator.

In Chapter four, we compare estimators of the form $\underset{\sim}{K} \hat{\underset{\sim}{\beta}}$. This requires that we select criteria for comparison that take into account the circumstances for which the original author developed the estimator. Also in this chapter, we incorporate a biased estimator into an iterative method for finding the maximum likelihood estimator for a dose response model. We also consider the bias introduced by extropolating.

Lastly, we describe what routes seem profitable for future research.

Chapter II

A REVIEW OF BIASED ESTIMATION

2.1 Introduction

In this chapter, our aim is to review certain estimators which have been proposed in the literature. These estimators have the characterization:

$$\underset{\sim}{b} = \underset{\sim}{K}\hat{\underset{\sim}{\beta}} \tag{2.1.1}$$

where $\hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}$ is the least squares estimator. This characterization facilitates the comparison of these estimators in terms of their general mean square error. Define the class K where $\underset{\sim}{b}$ is in K if there is a matrix $\underset{\sim}{K}$ such that (2.1.1) is satisfied.

As will be seen, this characterization holds for many estimators. Some of these estimators have been developed for a very specific purpose. Comparisons involving such limited-purpose estimators can cause problems. Therefore, before comparing two estimators, we want to insure that the intersection of the two sets describing the purposes of each of the estimators is not empty. To further delineate the generality of comparison results, one must also take into account the restrictions placed on the comparison. For example, we may restrict comparisons to only estimators of the form $\underset{\sim}{b} = \underset{\sim}{K}\hat{\underset{\sim}{\beta}}$, where $\underset{\sim}{K}$ is a matrix of *constants*.

In summary, we have to consider the purposes of the estimators and the restrictions surrounding the comparison. The generality of the comparison will depend on the intersection of three sets:

- The set of restrictions placed on the comparison;

• The two sets describing the purposes of each of the estimators.

In the extreme case when the intersection of these three sets is empty, we would consider the comparison invalid.

Therefore, to insure valid comparisons we will consider three purpose-classification groups. In this chapter, each estimator will be classified into one of these groups, and comparisons will be made only among estimators in the same group.

The three groups are:

1. The form of the true model is assumed to be known and one wishes to estimate the parameters of the model.
2. The form of the true model is not known, and one wishes to estimate certain relevant features of the true model. These features may be expressed in terms of a model which will be used to extrapolate (or predict) or to simply condense the data.
3. The form of the true model is not known and one wants to identify explanatory variables.

In this chapter we will introduce estimators according to the way the original authors introduced them. Also, generally known results about these estimators will be introduced at this stage. First, we will discuss estimators used for parameter estimation. Here, we take the full model least squares estimator as a viable alternative to other estimators mentioned in this section. These other estimators are:

• The minimum general mean square error estimator of the form $\underset{\sim}{K}\hat{\underset{\sim}{\beta}}$ where $\underset{\sim}{K}$ is a matrix of constants.

• Estimators satisfying the Stein restriction as applied to regression theory by Sclove and Branchik.

• Ridge Estimator .

Before the introduction of the ridge estimator, we shall discuss the concept of spurious correlations.

In the second section, we consider estimators that might be used when smoothing and/or prediction is the aim. In that section we only consider the case when the full model least squares estimator is not a viable alternative. Thus, the integrated bias of the estimators, which we will consider, will generally be nonzero. Here, we will compare (in terms of IMSE) those design and estimation requirements that result in minimizing the integrated bias. Also, the Kupper-Meydrech estimator and the reduced model least squares estimator are discussed.

In the third section, we discuss the estimation situation when identification of explanatory variables is the purpose of the analysis. The inconsistency arising from using percentiles of F distribution for determining cutoff points in stepwise regression is mentioned. Other variable selection criteria are mentioned and the use of ridge regression is also mentioned.

2.2 Estimation of Regression Coefficients

When interest centers on estimating the coefficients $\underset{\sim}{\beta}$ of a model, it is usually assumed that the true model is being fit. The usual least squares estimator $\hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}$ is unbiased for $\underset{\sim}{\beta}$ under this assumption. Therefore, motivation for considering a transformed estimator

$$\underset{\sim}{b} = \underset{p\times p}{\underset{\sim}{K}} \; \underset{p\times 1}{\hat{\underset{\sim}{\beta}}}$$

of $\hat{\underset{\sim}{\beta}}$ comes mainly from a desire to minimize mean square error in some form. That is, there is a willingness to accept non-zero bias in an estimator in order that the total error, bias plus variance, be reduced.

Various forms of mean square error criteria have been considered. Many of these forms can be expressed as

$$J(\underset{\sim}{M},\underset{\sim}{b}) = \text{tr } \underset{\sim}{M}E(\underset{\sim}{b}-\underset{\sim}{\beta})(\underset{\sim}{b}-\underset{\sim}{\beta})'.$$

For example, Mallows (1973) takes $\underset{\sim}{M}$ to be $\underset{\sim}{X}'\underset{\sim}{X}/\sigma^2$; Hoerl and Kennard (1970) use the identity matrix for $\underset{\sim}{M}$; and Box and Draper's (1959) integrated mean squared error can be viewed as letting

$$\underset{\sim}{M} = \frac{N\Omega}{\sigma^2}\int_R \underset{\sim}{x}\underset{\sim}{x}'d\underset{\sim}{x} .$$

So $E(\underset{\sim}{b}-\underset{\sim}{\beta})(\underset{\sim}{b}-\underset{\sim}{\beta})'$, called the mean square error matrix of $\underset{\sim}{b}$, is a matrix common to many forms of mean square error criteria. When $\underset{\sim}{M}$ is a symmetric positive definite matrix, the mean square error matrix of b is the only quantity that need be considered for the purpose of comparing estimators with respect to the same mean square error criterion. The following theorem (Finkbeiner, 1960) illustrates the importance of this matrix.

A real symmetric pxp matrix $\underset{\sim}{M}$ represents a positive definite quadratic form if and only if

$$\underset{\sim}{M} = \underset{\sim}{P}\underset{\sim}{P}'$$

for some real non-singular pxp matrix $\underset{\sim}{P}$.

Since $\underset{\sim}{E}_b = E(\underset{\sim}{b}-\underset{\sim}{\beta})(\underset{\sim}{b}-\underset{\sim}{\beta})'$ is non-negative definite and $\underset{\sim}{M}$ is positive definite we have that

$$\text{tr } \underset{\sim}{M}\underset{\sim}{E}_b = \text{tr } \underset{\sim}{P}\underset{\sim}{P}'\underset{\sim}{E}_b = \text{tr } \underset{\sim}{P}'\underset{\sim}{E}_b\underset{\sim}{P} \geq 0,$$

since all the diagonal elements of $\underset{\sim}{P}'\underset{\sim}{E}_b\underset{\sim}{P}$ are non-negative.

Therefore, for the purpose of comparing $J(\underset{\sim}{M},\underset{\sim}{b})$ for arbitrary $\underset{\sim}{b}_1$ and $\underset{\sim}{b}_2$, we have that

$$J(\underset{\sim}{M},\underset{\sim}{b}_1) \geq J(\underset{\sim}{M},\underset{\sim}{b}_2)$$

whenever the matrix $\underset{\sim}{E}_{\underset{\sim}{b}_1} - \underset{\sim}{E}_{\underset{\sim}{b}_2}$ is non-negative definite and $\underset{\sim}{M}$ is positive definite.

When $\underset{\sim}{K}$ is stochastic, the mean square error matrix of $\underset{\sim}{b} = \underset{\sim}{K}\hat{\underset{\sim}{\beta}}$ is:

$$\begin{aligned}
\underset{\sim}{E}_b &= E(\underset{\sim}{b}-\underset{\sim}{\beta})(\underset{\sim}{b}-\underset{\sim}{\beta})' = \text{Var } \underset{\sim}{K}\hat{\underset{\sim}{\beta}} + (E(\underset{\sim}{K}\hat{\underset{\sim}{\beta}})-\underset{\sim}{\beta})(E(\underset{\sim}{K}\hat{\underset{\sim}{\beta}})-\underset{\sim}{\beta})' \\
&= E(E(\underset{\sim}{K})\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta})(E(\underset{\sim}{K})\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta})' \\
&+ \sigma^2 E(\underset{\sim}{K}-E(\underset{\sim}{K}))[(\underset{\sim}{X}'\underset{\sim}{X})^{-1} + \underset{\sim\sim}{\beta\beta'}/\sigma^2](\underset{\sim}{K}-E(\underset{\sim}{K}))' \\
&+ \text{Cov } (\underset{\sim}{K},\hat{\underset{\sim}{\beta}}) \text{ Cov } (\underset{\sim}{K},\hat{\underset{\sim}{\beta}})' + \text{Cov } (\underset{\sim}{K},\hat{\underset{\sim}{\beta}})\underset{\sim}{\beta}'(E(\underset{\sim}{K})-\underset{\sim}{I})' \\
&+ (E(\underset{\sim}{K})-\underset{\sim}{I})\underset{\sim}{\beta}(\text{Cov}(\underset{\sim}{K},\hat{\underset{\sim}{\beta}}))'
\end{aligned} \tag{2.2.1}$$

This expression for the mean square error matrix is obtained by using the identities

$$\text{Var } \underset{\sim}{K}\hat{\underset{\sim}{\beta}} = E[\text{Var}(\underset{\sim}{K}\hat{\underset{\sim}{\beta}}|\underset{\sim}{K})] + \text{Var } [E(\underset{\sim}{K}\hat{\underset{\sim}{\beta}}|\underset{\sim}{K})]$$

and

$$E(\underset{\sim}{K}\hat{\underset{\sim}{\beta}}) = \text{Cov } (\underset{\sim}{K},\hat{\underset{\sim}{\beta}}) + E(\underset{\sim}{K})\underset{\sim}{\beta} .$$

This expression for $\underset{\sim}{E}_b$ allows one to gain insight into the difference between using a deterministic $\underset{\sim}{K}$ and a stochastic $\underset{\sim}{K}$ matrix. When $\underset{\sim}{K}$ is deterministic (a matrix of constants),

$$\underset{\sim}{K} = E(\underset{\sim}{K})$$

and

$$\text{Cov } (\underset{\sim}{K},\hat{\underset{\sim}{\beta}}) = \underset{\sim}{0}.$$

Thus, only the first matrix to the right of the last equal sign in expression (2.2.1) is nonzero in this case. Theil (1971) has considered when

$$\underset{\sim}{b} = \underset{\sim}{K}^*\hat{\underset{\sim}{\beta}} + \underset{\sim}{H}\underset{\sim}{Y} = \underset{\sim}{b}^* + \underset{\sim}{b}_H,$$

where

$$\underset{\sim}{K}^* = \underset{\sim}{\beta}[\sigma^2+\underset{\sim}{\beta}'\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{\beta}]^{-1}\ \underset{\sim}{\beta}'\underset{\sim}{X}'\underset{\sim}{X}$$

and $\underset{\sim}{H}$ is an arbitrary pxN matrix, the general mean square error can be expressed as

$$\underset{\sim}{E}_b = E(\underset{\sim}{b}-\underset{\sim}{\beta})(\underset{\sim}{b}-\underset{\sim}{\beta})' = E(\underset{\sim}{b}^*-\underset{\sim}{\beta})(\underset{\sim}{b}^*-\underset{\sim}{\beta})' + \underset{\sim}{H}_1 ,$$

where $\underset{\sim}{H}_1$ is non-negative definite. It follows that when a criterion of the form $J(\underset{\sim}{M},\underset{\sim}{b})$ is used and when $\underset{\sim}{b}$ is any linear combination of $\underset{\sim}{Y}$, then $\underset{\sim}{b}^*$ will be the linear combination that minimizes $J(\underset{\sim}{M},\underset{\sim}{b})$.

Unfortunately, $\underset{\sim}{K}^*$ involves unknown parameters. Therefore, $\underset{\sim}{b}^*$ is not a useful estimator. A particular choice of $\underset{\sim}{K}$ may achieve small $J(\underset{\sim}{M},\underset{\sim}{b})$ for some σ^2 and $\underset{\sim}{\beta}$, but Barnard (1963) has shown that the least squares estimator of $\underset{\sim}{x}'\underset{\sim}{\beta}$ is that linear combination of $\underset{\sim}{Y}$ having minimum mean square error in the class of bounded mean square error estimators when the range $\underset{\sim}{x}'\underset{\sim}{\beta}$ is unbounded. His result is generalized to the relevant setting in Chapter 3.

Therefore, the justification for using the least squares estimator does not require that one be restricted to the class of unbiased estimators. If the range of $\underset{\sim}{x}'\underset{\sim}{\beta}$ can not be restricted and prudence requires only consideration of bounded mean square error estimates, then the least squares estimator is the only resort. It is when the magnitude of the mean square error of the least squares estimator is unacceptably large that improved estimators in some larger class must be sought.

When $\underset{\sim}{K}$ is stochastic, however, there are certain estimators $\underset{\sim}{b}$ of the form

$$\underset{\sim}{b} = \underset{\sim}{K}\hat{\underset{\sim}{\beta}}$$

that have the property that when $\hat{\underset{\sim}{\beta}} \sim N(\underset{\sim}{\beta}, \sigma^2(\underset{\sim}{X}'\underset{\sim}{X})^{-1})$ and $p \geq 3$ then

$$J(\underset{\sim}{M},\underset{\sim}{b}) < J(\underset{\sim}{M},\hat{\underset{\sim}{\beta}}) \tag{2.2.2}$$

for all $\underset{\sim}{\beta}$.

James and Stein (1961) proved that (2.2.2) holds when $\underset{\sim}{M} = \underset{\sim}{I}$ and $(\underset{\sim}{X}'\underset{\sim}{X})^{-1} = \underset{\sim}{I}$. They found that (2.2.2) holds for the class of estimators characterized by $\underset{\sim}{K}$ being a scalar matrix (a multiple of $\underset{\sim}{I}$). Bhattacharya (1966) found a class of estimators for which (2.2.2) holds when $\underset{\sim}{M}$ is diagonal with diagonal elements d_i where

$$d_1 \geq d_2 \geq \ldots \geq d_p \ . \tag{2.2.3}$$

This class of estimators is characterized by $\underset{\sim}{K}$ being diagonal. He also shows how to obtain estimators satisfying (2.2.2) when $\underset{\sim}{X}'\underset{\sim}{X}$ is nonsingular and $\underset{\sim}{M}$ is positive definite.

Since this is the only case of general interest, we shall show exactly what the $\underset{\sim}{K}$ matrix looks like for this class of estimators.

Bhattacharya (1967) has shown that there is a nonsingular matrix $\underset{\sim}{L}$ so that when we define

$$\underset{\sim}{\alpha} = \underset{\sim}{L}\underset{\sim}{\beta}$$

and

$$\underset{\sim}{Z} = \underset{\sim}{X}\underset{\sim}{L}^{-1}$$

then the variance-covariance matrix of $\hat{\underset{\sim}{\alpha}} = \underset{\sim}{L}\hat{\underset{\sim}{\beta}}$ is

$$\text{Var } \hat{\underset{\sim}{\alpha}} = \text{Var } \underset{\sim}{L}\hat{\underset{\sim}{\beta}} = \sigma^2 \underset{\sim}{L}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{L}' = \sigma^2 \underset{\sim}{I}$$

and $\underset{\sim}{L}$ will diagonalize $\underset{\sim}{M}$. That is, there is a diagonal matrix $\underset{\sim}{D}$ so that

$$\underset{\sim}{L}'\underset{\sim}{D}\underset{\sim}{L} = \underset{\sim}{M}.$$

The columns of $\underset{\sim}{L}$ can be arranged so that (2.2.3) is satisfied. Thus Bhattacharya has found the transform $\underset{\sim}{L}$ so that in the transformed space, we have the conditions that the appropriate $\underset{\sim}{M}$ is diagonal and the elements of $\hat{\underset{\sim}{\alpha}}$ are uncorrelated. By Bhattacharya's extension to the results of James and Stein we know that there is a $\underset{\sim}{K}$ matrix so that

$$J(\underset{\sim}{D},\underset{\sim}{K}\hat{\underset{\sim}{\alpha}}) \le J(\underset{\sim}{D},\hat{\underset{\sim}{\alpha}}) \quad \text{for all } \underset{\sim}{\alpha}. \qquad (2.2.4.)$$

Now, consider an arbitrary estimator $\underset{\sim}{b}$, and subscript the $J(\underset{\sim}{M},\underset{\sim}{b})$ notation to clarify what parameter vector we mean. Then

$$\begin{aligned} J_{\underset{\sim}{\beta}}(\underset{\sim}{M},\underset{\sim}{b}) &= E(\underset{\sim}{b}-\underset{\sim}{\beta})'\underset{\sim}{M}(\underset{\sim}{b}-\underset{\sim}{\beta}) \\ &= E(\underset{\sim}{b}-\underset{\sim}{\beta})'\underset{\sim}{L}'\underset{\sim}{D}\underset{\sim}{L}(\underset{\sim}{b}-\underset{\sim}{\beta}) \\ &= E(\underset{\sim}{L}\underset{\sim}{b}-\underset{\sim}{L}\underset{\sim}{\beta})'\underset{\sim}{D}(\underset{\sim}{L}\underset{\sim}{b}-\underset{\sim}{L}\underset{\sim}{\beta}) \\ &= E(\underset{\sim}{L}\underset{\sim}{b}-\underset{\sim}{\alpha})'\underset{\sim}{D}(\underset{\sim}{L}\underset{\sim}{b}-\underset{\sim}{\alpha}) \\ &= J_{\underset{\sim}{\alpha}}(\underset{\sim}{D},\underset{\sim}{L}\underset{\sim}{b}) . \end{aligned}$$

We can use this last relationship to transform both sides of (2.2.4) to obtain

$$J(\underset{\sim}{M},\underset{\sim}{L}^{-1}\underset{\sim}{K}\underset{\sim}{L}\hat{\underset{\sim}{\beta}}) \le J(\underset{\sim}{M}, \hat{\underset{\sim}{\beta}}) \quad \text{for all } \underset{\sim}{\beta}.$$

Now let us present a matrix $\underset{\sim}{K}$ which has this property.

Let

$$\hat{\underset{\sim}{\alpha}}_{(j)} = (\hat{\alpha}_1,\hat{\alpha}_2,\ldots,\hat{\alpha}_j) \quad \text{and}$$

d_i be the ith diagonal of $\underset{\sim}{D}$ so that (2.2.3) is satisfied.

Now, let

$$f_j = \begin{cases} 1, & j = 1,2; \\ 1-(j-2)\phi(\hat{\beta})/(N-p+2)\hat{\alpha}'_{(j)}\hat{\alpha}_{(j)}, & j = 3,4,\ldots,p. \end{cases}$$

$\underset{\sim}{K}$ is then the diagonal matrix whose ith element is

$$\sum_{j=1}^{p} (d_j - d_{j+1}) f_j / d_i \ ,$$

where $d_{p+1} \equiv 0$.

2.2.1 The IMSE of Some Biased Estimation Procedures for Linear Models Where Spurious Correlations Exist Among the Independent Variables.

We now consider a particular difficulty that is sometimes faced in parameter estimation. This difficulty concerns the effect of spurious correlations. The effect of this problem may be so serious that the aims of the study can not be achieved. In this case, the purpose of the study may be downgraded from parameter estimation to identification of explanatory variables or prediction and smoothing. Thus, this difficulty may be experienced with any of the three purposes. We choose to consider its effect only when the purpose is parameter estimation.

Spurious correlations are apparent associations between unrelated quantities. When the data are observed rather than controlled, spurious correlations may arise. It is enlightening for the understanding of an aspect of this problem to consider a region of interest and the effect of spurious correlations on IMSE when the usual least squares estimator is used. An alternative estimator, the ridge estimator, has been proposed as a better estimator in such a situation. After discussing spurious correlations we shall compare these two estimators.

2.2.2 An Aspect of the Problem of Spurious Correlations in the Independent Variables.

It is believed that an important aspect of the problem of spurious correlations can simply be viewed as the multidimensional analog of the one-dimensional extrapolation problem.

Consider a simple case where one is attempting to estimate the coefficients of a linear model of the form:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 .$$

One might offhandedly believe that fitting such a function would provide a reasonable predictor for values of the independent variables (x_1, x_2) that are in the rectangle:

$$\min x_{1i} \leq x_1 \leq \max x_{1i} \text{ and } \min x_{2i} \leq x_2 \leq \max x_{2i},$$

where these minima and maxima are taken over the N values of the independent variables observed. See Figure 2.1.

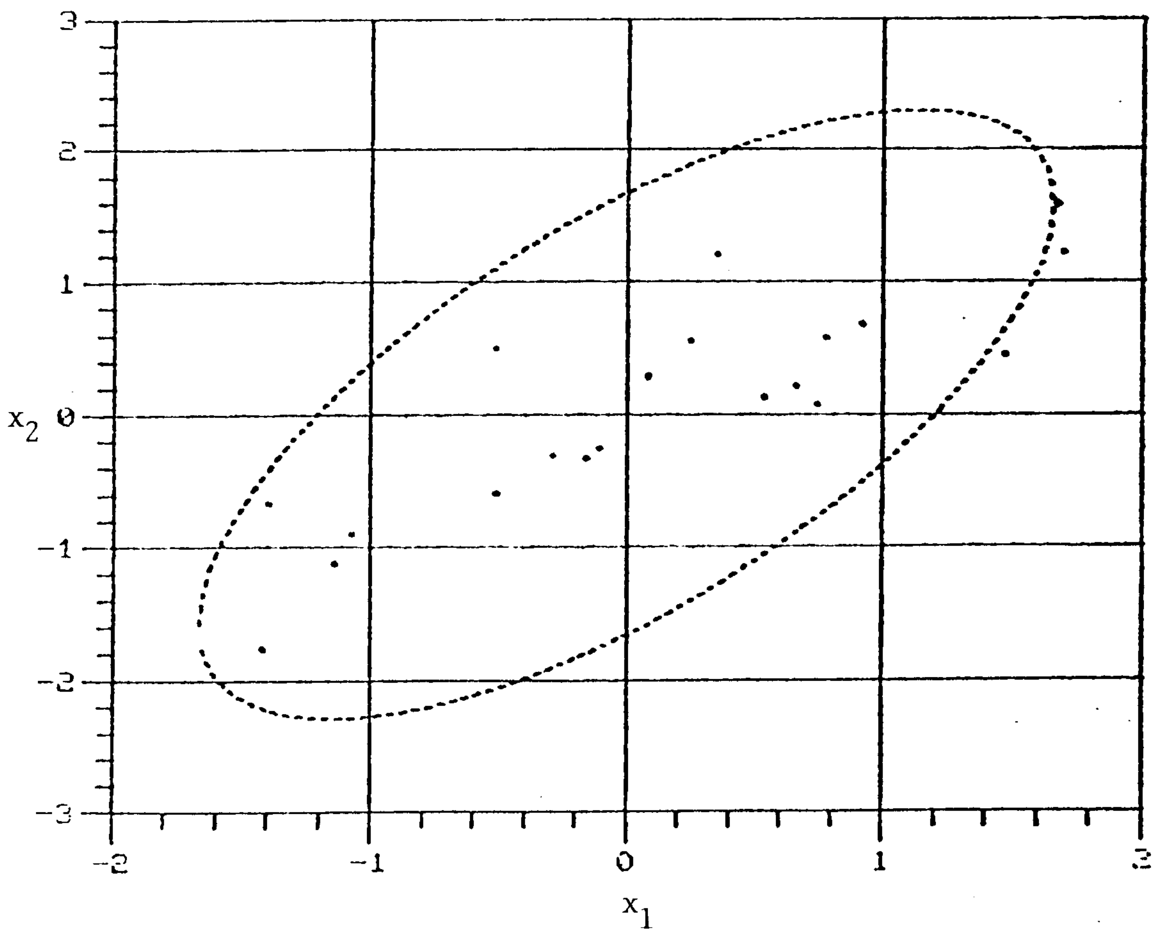
The figure shows a typical scatter plot of x_1 and x_2 for which the Pearson product moment coefficient would be near one. As can be seen the regions near the top left and bottom right corners are sparsely sampled and relatively distant from the region of experimentation. Using simulated data, Marquardt (1975) considers such a situation and has shown that when there is high correlation, the models with coefficients determined by usual least squares predict well in the region of experimentation, but poorly near the corners of the square. One way to circumvent the problem is to let the distribution of points (x_1, x_2) dictate the region of interest. For example, one could choose an ellipse as shown in the figure as the region of interest. If x_1 and x_2 are associated, then it will be rare that a point (x_1,x_2) lie outside the ellipse and so it is reasonable that there would be reduced interest in this region. On the other hand, if x_1 and x_2 are not associated, that is, the correlation is spurious, then there may still be a great deal of interest in the part of the rectangle outside the ellipse. In this situation and at other times when the region of interest must remain fixed, Marquardt proposes the ridge estimator as an alternative to the least squares estimator for the parameters.

Thus, spurious correlations mean that the experimenter has interest in regions of the independent variables that are sparsely sampled. The one-dimensional extrapolation problem occurs because the experimenter has interest in a range of the independent variable that has been sparsely sampled.

The extrapolation problem occurs because the experimenter who uses the least squares estimator must tacitly assume that the fitted function holds in this sparsely sampled region. Similarly, the experimenter who uses the least squares estimator in the multivariate case may be (unknowingly) making a similar assumption.

Figure 2.1

Scatter diagram of two highly correlated variables



The points were obtained from a table of random numbers. The underlying distribution is bivariate normal with zero means, unit variances and a correlation of 0.9. The ellipse is a 75 percentile of this distribution.

2.2.3 Ridge Regression

Hoerl and Kennard (HK) (1970a) have proposed a method of estimating the parameters of a linear model. This technique, called ridge regression, is designed to provide estimates with desirable properties when spurious correlations exist among the independent variables.

Hoerl and Kennard assume that the variables have been standardized so that $\underset{\sim}{X}'\underset{\sim}{X}$ is in correlation form. Since they consider only the situation when $\underset{\sim}{X}$ is of full column rank (even though it may have a "near singularity"), this implies that there is no constant term in the model. In general, we take this to mean that the constant term has been subtracted away. This may involve estimating it by the mean of the $\underset{\sim}{Y}$ vector. When this is the case, we ignore the fact that this constant has been estimated when we compute the general mean square error.

The ridge estimator is

$$\underset{\sim}{b}_k = (\underset{\sim}{X}'\underset{\sim}{X}+k\underset{\sim}{I})^{-1}\underset{\sim}{X}'\underset{\sim}{Y}$$

where k is a small positive quantity. In practice, k is determined from the ridge trace, the graph of the $\underset{\sim}{b}_k$ coefficient vector versus k. In the presence of variables so correlated that $\underset{\sim}{X}'\underset{\sim}{X}$ is ill conditioned, some elements of the $\underset{\sim}{b}_k$ vector may vary greatly with a small change in k. It is recommended that one choose the smallest value of k, say k^*, so that for values of k greater than k^*, there is little change in the elements of $\underset{\sim}{b}_k$. The ridge trace can also be used for variable selection.

The ridge estimator has not been considered in a formal region of interest framework. Therefore, IMSE criterion could not be used as criterion for evaluating the properties of the ridge estimator. Most writers have chosen instead to use total MSE. We shall first compare the least squares estimator to the ridge estimator using total MSE criterion, and then we shall show how this comparison is effected when a region of interest is defined and IMSE criteria is used.

HK have shown that there exists a range of k values such that smaller total MSE is achieved using the ridge estimator $\underset{\sim}{b}_k$ with k chosen from this set of values than when $k = 0$ (the least squares estimator for the full model).

To see this, let $\underset{\sim}{P}$ be such that $\underset{\sim}{X}'\underset{\sim}{X} = \underset{\sim}{P}'\underset{\sim}{\lambda}\underset{\sim}{P}$ and $\underset{\sim}{P}'\underset{\sim}{P} = \underset{\sim}{I}$ where $\underset{\sim}{\lambda} = \text{diag}\ (\lambda_1, \lambda_2, \ldots, \lambda_p)$ and $\lambda_i \geq \lambda_j$ for $i > j$. Define $\hat{\underset{\sim}{\phi}}_k = P\underset{\sim}{b}_k$ and $\underset{\sim}{\phi} = \underset{\sim}{P}\underset{\sim}{\beta}$, then if $L_k = (\underset{\sim}{b}_k - \underset{\sim}{\beta})'(\underset{\sim}{b}_k - \underset{\sim}{\beta})$, it follows that $E(L_k)$ is the total MSE of these coefficients. Now for any k

$$
\begin{aligned}
E(L_k) &= E(\underset{\sim}{b}_k - \underset{\sim}{\beta})'(\underset{\sim}{b}_k - \underset{\sim}{\beta}) \\
&= E(\hat{\underset{\sim}{\phi}}_k - \underset{\sim}{\phi})'(\hat{\underset{\sim}{\phi}}_k - \underset{\sim}{\phi}) \\
&= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + \sum_{i=1}^{p} \frac{\phi_i^2\, k^2}{(\lambda_i + k)^2} .
\end{aligned}
$$

The total MSE of the least squares estimator $(k=0)$ is

$$
E(L_0) = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} .
$$

When $k \leq k^* = \dfrac{\sigma^2}{\phi^2_{max}}$, where ϕ^2_{max} is the maximum of the squared elements of the vector $\underset{\sim}{\phi}$, we have:

$$
E(L_k) \leq \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + \sum_{i=1}^{p} \frac{\phi^2_{max}\, kk^*}{(\lambda_i + k)^2}
$$

$$= \sum_{i=1}^{p} \frac{\sigma^2\lambda_i+\sigma^2 k}{(\lambda_i+k)^2}$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i+k} \le \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} .$$

That is, the total MSE of the coefficient vector is less for the full model least squares estimator when

$$k < \frac{\sigma^2}{\phi^2_{max}} .$$

Thus, if k is chosen sufficiently small one can obtain an improvement over the least squares estimator. However, to know that k has been chosen properly requires some prior knowledge of the true parameters.

Now, let us see how this comparison is affected when IMSE criteria is used. First, a region of interest R has to be specified. Consider an R whose center is at the origin. Let

$$W_s = \Omega \int_R \underset{\sim}{x}\underset{\sim}{x}'d\underset{\sim}{x} .$$

Then, the IMSE of the ridge estimator is:

$$J_R = E(\underset{\sim}{b}_k-\underset{\sim}{\beta})'\underset{\sim}{W}_s(\underset{\sim}{b}_k-\underset{\sim}{\beta}) .$$

Let $\underset{\sim}{H}'\underset{\sim}{H}$ be the square root decomposition of $\underset{\sim}{W}_s$ (i.e. $\underset{\sim}{W}_s = \underset{\sim}{H}'\underset{\sim}{H}$) and define

$$\underset{\sim}{\psi}_k = \underset{\sim}{H}^{-1}\underset{\sim}{P}\underset{\sim}{b}_k \quad \text{and} \quad \underset{\sim}{\psi} = \underset{\sim}{H}^{-1}\underset{\sim}{P}\underset{\sim}{\beta};$$

then,

$$(\underset{\sim}{b}_k-\underset{\sim}{\beta})' \underset{\sim}{W}_s(\underset{\sim}{b}_k-\underset{\sim}{\beta}) = (\underset{\sim}{\psi}_k-\underset{\sim}{\psi})'\underset{\sim}{P}'\underset{\sim}{H}^{-1}\underset{\sim}{W}_s\underset{\sim}{H}^{-1}\underset{\sim}{P}(\underset{\sim}{\psi}_k-\underset{\sim}{\psi})$$

$$= (\underset{\sim}{\psi}_k-\underset{\sim}{\psi})'\underset{\sim}{P}'\underset{\sim}{H}'^{-1}\underset{\sim}{H}'\underset{\sim}{H}\underset{\sim}{H}^{-1}\underset{\sim}{P}(\hat{\underset{\sim}{\psi}}_k-\underset{\sim}{\psi}) = (\underset{\sim}{\psi}_k-\underset{\sim}{\psi})'\underset{\sim}{P}'\underset{\sim}{P}(\underset{\sim}{\psi}_k-\underset{\sim}{\psi})$$

$$= (\underset{\sim}{\psi}_k-\underset{\sim}{\psi})'(\underset{\sim}{\psi}_k-\underset{\sim}{\psi}) \ .$$

The same reasoning applies here that applied when using total MSE criterion. It follows that when

$$0 < k < \sigma^2/\psi^2_{max}$$

where ψ^2_{max} is the maximum element of the squares of the elements of the parameter vector $\underset{\sim}{\psi}$, the ridge estimator will achieve smaller IMSE than the least squares estimator.

2.3 The Case Where the Aim is Smoothing or Prediction

There are numerous instances where it is best to consider a model that is much simpler than what might be considered the true model.

This situation usually occurs when there are a large number of possibly important factors. In the interest of keeping the situation relatively simple, one might choose to work with an approximating function, and forego the complexity of finding the true model.

Usually, an adequately fitting approximation function will leave unexplained a sizeable amount of the deviation in the data. This variation will consist of bias and variance errors. The distinction between these two types of error is important to keep in mind namely bias error will not average out in the long run. Thus confidence intervals based on the assumption that the true model has been fit will tend to paint a rosier situation than exists. Of course, the confidence intervals will be wider than those obtained using the true model, but this bias contribution will not "average out" so that the usual interpretation of the confidence interval must be altered.

In order to more honestly portray the data, one might assume that the true model is only slightly more complicated than the fitted model. For example, when fitting a first order polynomial to the data, one might assume that the true model is a second degree polynomial.

This is the situation we consider. We assume that one is fitting a model that is not the form of the assumed model. The reason one is doing this is because of smoothing or prediction considerations. These considerations rule out the alternative of using $\hat{\beta}$ to estimate β. For example, in a dose-response situation one may be interested in low dose levels. The instrument used to measure response may be insensitive at low dose levels. For this reason the dose is applied at higher levels where the response can be measured and the fitted model is used to extrapolate back to the low dose region of interest. Time and again, experiments have indicated that response is a concave function of dose in this region. One might assume that the true response is a concave function but fit a straight line. A fitted straight line in such a situation will usually overpredict response. Thus, one is able to achieve a safety factor that may not be achievable with a concave function. Also, if one fits a quadratic function there is no guarantee that the fitted function will not show an increase in response in low dose regions or negative values of response. For reasons such as these one might rule out the option of estimating β with no bias.

2.3.1 Achieving Minimum Bias with Approximating Functions

In both the BD and KMH approaches, minimum bias error was achieved. BD achieved minimum bias by choice of design and KMH achieved it by choice of estimator. Later, Karson (1970) showed that minimum bias could be achieved by a combination of design and estimation considerations.

A simple way of viewing minimum bias techniques will be presented along with a comparison of the resultant IMSE achieved by these techniques when the true model is misspecified.

2.3.1.1 Correcting an Estimator so that its Bias Error is Minimized

Suppose that the true model is:

$$\eta(\underset{\sim}{x}) = \underset{\sim}{x}'\underset{\sim}{\beta} = \underset{\sim}{x}_1'\underset{\sim}{\beta}_1 + \underset{\sim}{x}_2'\underset{\sim}{\beta}_2 = \underset{\sim}{x}_1'\underset{\sim}{\beta}_1 + \underset{\sim}{x}_{2a}'\underset{\sim}{\beta}_{2a} + \underset{\sim}{x}_{2b}'\underset{\sim}{\beta}_{2b},$$

and it is desired to fit an approximating function of the form:

$$\eta(\underset{\sim}{x}) = \underset{\sim}{x}_1'\underset{\sim}{b}_1 \ .$$

KMH have shown that any estimator $\underset{\sim}{b}_1$ with expected value $\underset{\sim}{\beta}_1 + \underset{\sim}{W}_1^{-1}\underset{\sim}{W}_2\underset{\sim}{\beta}_2$ will achieve minimum integrated squared bias (B). For example, KMH suggest using the estimator $\hat{\underset{\sim}{\beta}}_1 + \underset{\sim}{W}_1^{-1}\underset{\sim}{W}_2\hat{\underset{\sim}{\beta}}_2$, where $\begin{bmatrix} \hat{\underset{\sim}{\beta}}_1 \\ \hat{\underset{\sim}{\beta}}_2 \end{bmatrix}$ are the least square estimators using the full model. That is,

$$\begin{bmatrix} \hat{\underset{\sim}{\beta}}_1 \\ \hat{\underset{\sim}{\beta}}_2 \end{bmatrix} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\,\underset{\sim}{X}'\underset{\sim}{Y} = \left(\begin{bmatrix} \underset{\sim}{X}_1' \\ \hline \underset{\sim}{X}_2' \end{bmatrix} \begin{bmatrix} \underset{\sim}{X}_1 \,\vdots\, \underset{\sim}{X}_2 \end{bmatrix} \right)^{-1} \begin{bmatrix} \underset{\sim}{X}_1' \\ \underset{\sim}{X}_2' \end{bmatrix} \underset{\sim}{Y} \ .$$

In order to use the KMH estimator, $\underset{\sim}{\beta}_1 + \underset{\sim}{W}_1^{-1}\underset{\sim}{W}_2\underset{\sim}{\beta}_2$ must be estimable. Satisfying this condition may require additional observations. Probably the usual estimator that is commonly employed is

$$\tilde{\underset{\sim}{\beta}}_1 = (\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\,\underset{\sim}{X}_1'\underset{\sim}{Y} \ .$$

Among other desired properties, it minimizes the sum of unexplained squared error $(\underset{\sim}{Y}-\underset{\sim}{X}_1\underset{\sim}{b})'(\underset{\sim}{Y}-\underset{\sim}{X}_1\underset{\sim}{b})$. Since KMH have shown that minimum bias error imposes a constraint only on an estimator's expected value, one might consider correcting $\tilde{\beta}_1$ by adding a vector of correcting terms C_1 to $\tilde{\beta}_1$. What is the C_1 necessary to obtain minimum bias error in this case? That is, what C_1 satisfies

$$E(\tilde{\beta}_1) + C_1 = \beta_1 + W_1^{-1}W_2\beta_2 \ ?$$

Since

$$E(\tilde{\beta}_1) + C_1 = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + C_1 = \beta_1 + A\beta_2 + C_1 ,$$

where

$$A = (X_1'X_1)^{-1}X_1'X_2 ,$$

it follows that

$$C_1 = (W_1^{-1}W_2 - A)\beta_2 .$$

Using this concept, it is easy to understand how estimation and design considerations have been employed to achieve minimum bias error.

The BD approach is to use the estimator $\tilde{\beta}_1$ and set C_1 equal to zero by using the design condition $A = W_1^{-1}W_2$. This then precludes the necessity of having to worry about β_2. The KMH method is to estimate C_1 by $C_1 = (W_1^{-1}W_2 - A)\hat{\beta}_2$. This follows since $\tilde{\beta}_1 = \hat{\beta}_1 - A\hat{\beta}_2$. They use design flexibility to minimize V.

It has been shown by KMH that given complete choice of design their approach will result in smaller IMSE than the BD approach. Whether, in a particular problem, one prefers to use one of these two approaches, may depend on how much control over the independent variables one has, cost of sampling in subregions in R, and other considerations. An alternative approach given by Karson that also results in minimum B is to use both design and estimation restrictions. How this is achieved can be seen by partioning C_1 as follows:

$$\underset{\sim}{C}_1 = (\underset{\sim}{W}_1^{-1}\underset{\sim}{W}_{2a}-\underset{\sim}{A}_a)\underset{\sim}{\beta}_{2a} + (\underset{\sim}{W}_1^{-1}\underset{\sim}{W}_{2b}-\underset{\sim}{A}_b)\underset{\sim}{\beta}_{2b},$$

where $\underset{\sim}{A} = (\underset{\sim}{A}_a \mid \underset{\sim}{A}_b) = ((\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1'\underset{\sim}{X}_{2a} \mid (\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1'\underset{\sim}{X}_{2b})$ and $\underset{\sim}{W}_2 = (\underset{\sim}{W}_{2a} \mid \underset{\sim}{W}_{2b})$.

When this partition is possible, one may choose the design restriction $\underset{\sim}{A}_b = \underset{\sim}{W}_1^{-1}\underset{\sim}{W}_{2b}$ and estimate the non-zero term by $(\underset{\sim}{A}_a-\underset{\sim}{W}_1^{-1}\underset{\sim}{W}_{2a})\hat{\underset{\sim}{\beta}}_{2a}$.

Since all three approaches will achieve the same minimum value of bias error, comparison of their IMSE's requires us to consider only the variance error. The variance error of the KMH estimator is

$$V_{KMH} = N\ \mathrm{tr}(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{W}_1 + N\ \mathrm{tr}(\underset{\sim}{W}_1^{-1}\underset{\sim}{W}_2-\underset{\sim}{A})'\underset{\sim}{W}_1(\underset{\sim}{W}_1^{-1}\underset{\sim}{W}_2-\underset{\sim}{A})\underset{\sim}{S}^{-1}$$

where

$$\underset{\sim}{S} = \underset{\sim}{X}_2'\underset{\sim}{X}_2 - \underset{\sim}{X}_2'\underset{\sim}{X}_1(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1'\underset{\sim}{X}_2 .$$

If the dimensions of $\underset{\sim}{S}^{-1}$ are greater than one, this matrix can be partioned. Therefore denote the partion of $\underset{\sim}{S}^{-1}$ that agrees with the partition of $\underset{\sim}{W}_2 = (\underset{\sim}{W}_{2a} \mid \underset{\sim}{W}_{2b})$ as:

$$\underset{\sim}{S}^{-1} = \left[\begin{array}{c|c} \underset{\sim}{S}_{11} & \underset{\sim}{S}_{12} \\ \hline \underset{\sim}{S}_{12}' & \underset{\sim}{S}_{22} \end{array}\right] .$$

Since the variance error of the BD estimator is $V_{BD} = N\ \mathrm{tr}(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{W}_1$, the variance error of the KMH estimator is:

$$V_{KMH} = V_{BD} + N\ \mathrm{tr}(\underset{\sim}{W}_1^{-1}\underset{\sim}{W}_{2a}-\underset{\sim}{A}_a)'\underset{\sim}{W}_1(\underset{\sim}{W}_1^{-1}\underset{\sim}{W}_{2a}-\underset{\sim}{A}_a)\underset{\sim}{S}_{11}$$

$$+ \{2N\ \mathrm{tr}(\underset{\sim}{W}_1^{-1}\underset{\sim}{W}_{2b}-\underset{\sim}{A}_a)'\underset{\sim}{W}_1(\underset{\sim}{W}_1^{-1}\underset{\sim}{W}_{2b}-\underset{\sim}{A}_b)\underset{\sim}{S}_{12}$$

$$+ N\ \mathrm{tr}(\underset{\sim}{W}_1^{-1}\underset{\sim}{W}_{2b}-\underset{\sim}{A}_b)'\underset{\sim}{W}_1(\underset{\sim}{W}_1^{-1}\underset{\sim}{W}_{2b}-\underset{\sim}{A}_b)\underset{\sim}{S}_{22}\} .$$

In the above expression, denote the second term on the right by ΔV_a and the bracketed term by ΔV_b. The above expression becomes

$$V_{KMH} = V_{BD} + \Delta V_a + \Delta V_b \ .$$

The KMH method uses the design which minimizes V_{KMH}. Since

$$\min V_{KMH} \le \min V_{BD} + \min \Delta V_a + \min \Delta V_b = \min V_{BD} \le V_{BD},$$

it follows that the KMH approach always attains a variance error no larger than that attained by the BD method. Also, the variance error of the Karson estimator is $V_K = V_{BD} + \Delta V_a$.

By the same argument, it follows that the Karson method results in an IMSE that is between those attained by the BD and KMH methods.

In summary, it has been shown that the KMH method is preferable to the BD method, in terms of small IMSE,when the true model is correctly specified. The Karson method is better than the BD method, but not as good as the KMH method, when only IMSE criteria are considered.

2.4 Identification of Explanatory Variables

When the aim of the analysis is identification of explanatory variables, and linear models are employed, it is common to focus attention on tests of significance rather than estimates of the parameters of the model. If the experiment involves a small number of variables, the analysis is rather straightforward. In this case, it is possible to estimate $\underset{\sim}{\beta}$, the parameter vector, using $\hat{\underset{\sim}{\beta}}$.

When the data are from an uncontrolled study and the dependent variable is observed at many different combinations of levels of the independent variables, the statistical power of these significance tests may be small. In this situation, there still remains a desire to identify the explanatory variables. In such a case, linear effects (plus possibly quadratic effects and crossproduct terms) may be considered the true model, while higher order terms are ignored.

In order to identify the explanatory variables at this stage, it is quite common that stepwise regression is employed. Even when the terms of the true model are included in the terms being analyzed, the usual F tests associated with stepwise regression are not valid. Pope and Webster (1972) have considered the forward selection procedure and the F statistic used at the ith step to decide whether or not to include one additional variable in the equation. Superficially, the F statistic has a central F distribution under the normality assumption and under the null hypothesis

$$H_0: \ \beta_i = 0,$$

where β_i is the coefficient associated with a variable not already included in the equation. This test is done for all such variables and if the maximum of these F statistics does not exceed some cutoff value the selection process stops. The cutoff value is usually chosen to match a certain significance level based on the central F distribution. Pope and Webster have listed sufficient conditions for this maximum F statistic to follow a central F distribution, and they note that these conditions must not hold due to the way the procedure works. So, although the procedure is clearcut, there is no probability available to measure the goodness of the procedure.

Having described an inconsistency with using F tests and stepwise regression, we feel that parameter estimation should be given more attention, even in this situation. Consider how the estimator resulting from a stepwise regression procedure relates to $\hat{\beta}$, the least squares estimator when all the terms (linear, quadratic and crossproduct) are included in the model.

Assume that the selection procedure stops after p_s elements have been included in the prediction equation so that b_s is a $p_s \times 1$ vector. Then assume that the first p_s elements of $\hat{\beta}$ are those terms included in the prediction equation. Then $b_s = K_s\hat{\beta}$ where K_s is p_s by p and

$$K_s = \left[I_{p_s} \;\vdots\; \underset{p_s \times (p-p_s)}{A_s} \right] .$$

Here $A_s = (X_s'X_s)^{-1}X_s'X_n$ where X_s is the matrix whose columns are the observed values of the variables in the equation. The columns of X_n are the observed values of variables not included in the equation. Note that K_s must be considered stochastic, since the assumed permutation of the elements of β and value of p_s depends on the Y observations.

Criteria other than the use of cutoff points based on the F distribution have been used for basis of variable selection in stepwise regression procedures.

The R^2 statistic has also been employed to judge whether an additional variable should be added to the equation. The square of the multiple correlation coefficient is defined to be:

$$R^2 = \frac{(\bar{Y}-X\hat{\beta})'(\bar{Y}-X\hat{\beta})}{(Y-\bar{Y})'(Y-\bar{Y})} = \frac{F_0}{\frac{n-p}{p-1} + F_0} ,$$

where F_0 tests $\beta=0$ for the model $\eta=\beta_0+x'\beta$, and where $\bar{Y}$ is the vector whose N elements are the average of the $\underset{\sim}{Y}$. R^2 is usually expressed as a percentage. Including more variables into the equation will increase R^2. If including the ith variable would only result in a "small" increase in R^2, this variable is not included and the selection procedure stops. Although it is difficult to objectively decide how small is a "small" increase in R^2, the tabulation of the R^2 along with the variable selected at each step offers a way to report the essential features of the data.

The C_p statistic proposed by Mallows (1973) is

$$C_p = \frac{(\underset{\sim}{Y}-\underset{\sim}{X}_s\underset{\sim}{b}_s)'(\underset{\sim}{Y}-\underset{\sim}{X}_s\underset{\sim}{b}_s)}{\hat{\sigma}^2} - n + 2p$$

where $\hat{\sigma}^2$ is an estimator for σ^2. This statistic estimates the expected value of:

$$(\underset{\sim}{b}_s-\underset{\sim}{\beta})'\underset{\sim}{X}'\underset{\sim}{X}(\underset{\sim}{b}_s-\underset{\sim}{\beta})/\sigma^2$$

which is the scaled sum of the squared errors. Mallows' idea is that C_p be plotted against p for as many of the regressions as possible. Such a plot gives a intuitive feel of the relative explanatory power of using different subsets of the possible regressions. The use of such plots is very much subjective, but can be seen as controlling for numbers of observations and numbers of variables in the equation. Since the R^2 statistic can usually be made one by including enough variables in the equation, R^2 doesn't control for the number of variables in the equation. The similarity of R^2 to a correlation coefficient is probably the reason for its stronger appeal.

Helms (1974) considers average estimated variance (AEV) as a method of variable selection. This involves the choice of a region of interest and consideration of moments. The AEV automatically incorporates information about bias and variance when one enters or deletes a variable. When the data are collected and the moments are $(\underset{\sim}{X}'\underset{\sim}{X})/N$, the AEV becomes

$$\text{AEV} = s_1^2 p_1/N,$$

where

$$s_1^2 = (\underset{\sim}{Y}-\underset{\sim}{X}_1\tilde{\underset{\sim}{\beta}}_1)'(\underset{\sim}{Y}-\underset{\sim}{X}_1\tilde{\underset{\sim}{\beta}}_1)/(N-p_1).$$

When these moments are used and when only submodels having the same rank of $\underset{\sim}{X}_1$ are considered then the AEV is a monotonic function of C_p. Consequently, in this case both statistics lead to the same conclusion.

The developers of ridge regression felt that omitting variables is an unwise procedure. They felt that the true model should be fit, and if it is desirable to identify important variables, then those variables whose associated estimated coefficients have large magnitudes should be considered most important. Since these variables are scaled by dividing each variable by the square root of its sum of squares, comparisons of these magnitudes are invariant to multiplication of the variable by a constant. In this sense, ridge analysis transforms the problem of variable selection back to a problem of parameter estimation. However, when one considers the objective of variable selection, it is not clear what criterion one would want to use to judge the goodness of different estimation methods.

Except for ridge regression, all of these procedures use the possibly biased least squares estimator $\tilde{\underset{\sim}{\beta}}_1$ that minimizes the sum of squares $(\underset{\sim}{Y}-\underset{\sim}{X}_1\underset{\sim}{b}_1)'(\underset{\sim}{Y}-\underset{\sim}{X}_1\underset{\sim}{b}_1)$. A reason for preferring $\tilde{\underset{\sim}{\beta}}_1$ to other estimators of the form $\underset{\sim}{b}_1 = K\tilde{\underset{\sim}{\beta}}_1$ is that $\tilde{\underset{\sim}{\beta}}_1$ yields the optimal value for any of these statistics. That is, given the matrix $\underset{\sim}{X}_1$ then R^2 is maximized and C_p and AEV are minimized using $\tilde{\underset{\sim}{\beta}}_1$.

Even so, the least squares estimator being used here is not $\hat{\beta}_1$ but the estimator $\tilde{\beta}_1$. If this estimator is used, then the estimator that minimizes the squares of the residuals from this model is the stagewise estimator $\hat{\beta}_S$.

Stagewise regression is a procedure that has been used in economic data analysis. If the model is

$$\eta = x_1'\beta_1 + x_2'\beta_2$$

one might estimate β_1 by

$$\tilde{\beta}_1 = (X_1'X_1)^{-1}X_1'Y$$

and then regress the residuals of this regression onto X_2 to get estimates for β_2. So the stagewise estimator $\hat{\beta}_S$ is:

$$\hat{\beta}_S = \begin{bmatrix} \tilde{\beta}_1 \\ (X_2'X_2)^{-1}X_2'(Y-X_1\tilde{\beta}_1) \end{bmatrix}$$

whereas the usual estimator is:

$$\hat{\beta} = \begin{bmatrix} \tilde{\beta}_1 \\ 0 \end{bmatrix} + \begin{bmatrix} -A \\ I \end{bmatrix} [(X_2-X_1A)'(X_2-X_1A)]^{-1}(X_2-X_1A)'Y ,$$

where $A = (X_1'X_1)^{-1} X_1'X_2$. Wallace (1964) has considered the case where

$$\eta = \beta_1 x_1 + \beta_2 x_2$$

and has shown that the mean square of $\tilde{\beta}_1$ is less than the mean square error of $\hat{\beta}_1$ when

$$\frac{\beta_2^2}{\mathrm{Var}\,(\hat{\beta}_2)} < 1$$

and $\tilde{\beta}_2$ attains smaller MSE than $\hat{\beta}_2$ when

$$\frac{\beta_1^2}{\mathrm{Var}(\hat{\beta}_1)} < \frac{2-r_{12}^2}{r_{12}^2}$$

where r_{12}^2 is the coefficient of correlation between x_1 and x_2.

Chapter III
NEW RESULTS

3.1 Introduction

In Chapter II we have reviewed some of the estimators we will be considering. An attempt was made to present these estimators in one of the three "purpose" classification categories. In this chapter we will not always follow the classification labels presented in the last chapter. We want to assess certain aspects of an estimator's robustness. That is, we want to assess certain characteristics of an estimator (or a slight modification of an estimator) when the estimator is used for a purpose other than the original intended purpose. Sometimes an estimator may perform well but the assumptions involving its implementation may be unwarranted. Problems of this nature will also be discussed.

3.2 Parameter Estimation

Consider the situation in which one desires to estimate the parameter vector $\underset{\sim}{\beta}$ with small IMSE. We shall consider some estimators of $\underset{\sim}{\beta}$, say $\underset{\sim}{b}$, in the class K.

First, consider transforms $\underset{\sim}{K}$ which are deterministic. That is

$$\underset{\sim}{b} = \underset{\sim}{K}\hat{\underset{\sim}{\beta}}$$

where $\underset{\sim}{K}$ is a matrix of constants. We shall only consider those estimators $\underset{\sim}{b}$ whose integrated bias error

$$\underset{\sim}{\beta}'(\underset{\sim}{K}-\underset{\sim}{I})'\underset{\sim}{W}(\underset{\sim}{K}-\underset{\sim}{I})\underset{\sim}{\beta}$$

is less than N_0. That is, the integrated bias error is bounded. Since $\underset{\sim}{W}$ is positive definite, $(\underset{\sim}{K}-\underset{\sim}{I})'\underset{\sim}{W}(\underset{\sim}{K}-\underset{\sim}{I})$ is non-**negative** definite if $K \neq I$. Thus, there is a matrix $\underset{\sim}{P}$ such that

$$\underset{\sim}{P}'\underset{\sim}{P} = \underset{\sim}{I}$$

and

$$\underset{\sim}{P}'\underset{\sim}{\lambda}\underset{\sim}{P} = (\underset{\sim}{K}-\underset{\sim}{I})'\underset{\sim}{W}(\underset{\sim}{K}-\underset{\sim}{I})$$

where $\underset{\sim}{\lambda}$ is the diagonal matrix of eigenvalues of which at least one eigenvalue is strictly positive. Let λ_i be the value of a strictly positive eigenvalue and let $\underset{\sim}{P}_i$ be an associated eigenvector with length 1. Then let $\underset{\sim}{\beta} = 2\sqrt{N_0/\lambda_i}\,\underset{\sim}{P}_i$. So

$$\underset{\sim}{\beta}'(\underset{\sim}{K}-\underset{\sim}{I})'\underset{\sim}{W}(\underset{\sim}{K}-\underset{\sim}{I})\underset{\sim}{\beta} = 4N_0 > N_0 \ .$$

This is inconsistent with our earlier assumption. Thus, the only estimator in K which is a deterministic transform of $\hat{\underset{\sim}{\beta}}$ and has bounded integrated bias error for all $\underset{\sim}{\beta}$ is the identity transform.

Now, let us consider cases where there is prior knowledge of upper bounds on some of the parameters or prior knowledge of upper bounds on functions of the parameters.

First, consider when there is an upper bound on $\eta^2(\underset{\sim}{x})/\sigma^2$ for $\underset{\sim}{x}$ in R. Specifically, say that for $\underset{\sim}{x}$ in R

$$\eta^2(\underset{\sim}{x})/\sigma^2 < U.$$

We shall consider the shrunken(i.e., shrinking towards 0),estimator $\underset{\sim}{b}_{MW}$ studied by Mayer and Willke. The transform $\underset{\sim}{K}$ for this estimator is $\underset{\sim}{K} = \delta_{MW}\underset{\sim}{I}$ where δ_{MW} is a scalar so that

$$\underset{\sim}{b}_{MW} = \delta_{MW}\hat{\underset{\sim}{\beta}}$$

is a scalar multiple of the least squares estimator. The IMSE of $\underset{\sim}{b}_{MW}$ is

$$J(\underset{\sim}{W},\underset{\sim}{b}_{MW}) = \delta_{MW}^2 V_{FM} + (1-\delta_{MW})^2 B_0, \qquad (3.2.1.)$$

where

$$V_{FM} = N\ tr(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{W},$$

and is the variance error of $\hat{\underset{\sim}{\beta}}$; and

$$B_0 = \frac{N}{\sigma^2}\underset{\sim}{\beta}'\underset{\sim}{W}\underset{\sim}{\beta}$$

is the bias error of the zero estimator, $\underset{\sim}{b} \equiv \underset{\sim}{0}$. The choice of δ_{MW} that has smallest IMSE is

$$\delta_{MW}^* = \frac{B_0}{V_{FM}+B_0}.$$

Since δ_{MW}^* involves the unknown parameters, as was the case for a general deterministic $\underset{\sim}{K}$ in chapter II, it is not a useable choice of scalar for an estimator of the form $\underset{\sim}{b}_{MW}$.

It is possible to obtain an estimator with smaller IMSE than $\hat{\underset{\sim}{\beta}}$ has when the parameter space is bounded. To see this, first note that the bound we have assumed on $\eta^2(x)/\sigma^2$ implies

$$B_0 = \frac{N\Omega}{\sigma^2}\int_R \eta^2(x)d\underset{\sim}{x} \le N\Omega\int_R U d\underset{\sim}{x} = NU\ .$$

Now,

$$J(\underset{\sim}{W},\underset{\sim}{b}_{MW}) < J(\underset{\sim}{W},\hat{\underset{\sim}{\beta}})$$

when

$$\delta_{MW}^2 V_{FM} + (1-\delta_{MW})^2 B_0 < V_{FM} .$$

This is true when

$$\frac{B_0 - V_{FM}}{V + B_0} < \delta_{MW} < 1 .$$

Also

$$J(\underset{\sim}{W}, \underset{\sim}{b}_{MW}) < J(\underset{\sim}{W}, \underset{\sim}{0})$$

when

$$0 < \delta_{MW} < \frac{2B_0}{V_{FM} + B_0} .$$

Combining ,we have that

$$J(\underset{\sim}{W}, \underset{\sim}{b}_{MW}) < \min(J(\underset{\sim}{W}, 0), J(\underset{\sim}{W}, \hat{\underset{\sim}{\beta}}))$$

when

$$\min\left(0, \frac{2B_0}{V_{FM} + B_0}\right) < \delta_{MW} < \max\left(\frac{B_0 - V_{FM}}{V_{FM} + B_0}, 1\right) .$$

Based on the above inequalities, an estimator that attains small IMSE for a wide range of B_0 is

$$\underset{\sim}{b}_{MW}^{+} = \max\left(0, \frac{U - V_{FM}}{U + V_{FM}}\right) \hat{\underset{\sim}{\beta}} . \qquad (3.2.2)$$

Second, we consider the case where $\underset{\sim}{K}$ is stochastic. Consider the Stein estimator that was described in Section 2.2. Since IMSE is just a special case of general mean square error, the results in Section 2.2 about this estimator will apply. Therefore, we assume that the errors are normally distributed and we assume that $p \geq 3$.

Now let us compare this stochastic estimator, say $\underset{\sim}{b}_S$, to $\underset{\sim}{b}_{MW}^{+}$. Consider the set of parameter values $\underset{\sim}{\beta}$ for which

$$\underset{\sim}{\beta}'\underset{\sim}{W}\underset{\sim}{\beta}/\sigma^2 < U.$$

When $\underset{\sim}{\beta}'\underset{\sim}{W}\underset{\sim}{\beta}/\sigma^2$ equals U, the IMSE of $\underset{\sim}{b}^+_{MW}$ equals the IMSE of $\hat{\underset{\sim}{\beta}}$. Since the IMSE of $\underset{\sim}{b}_s$ is always less than or equal to the IMSE of $\hat{\underset{\sim}{\beta}}$, the IMSE of $\underset{\sim}{b}_s$ is less than the IMSE of $\underset{\sim}{b}^+_{MW}$ if $\underset{\sim}{\beta}'\underset{\sim}{W}\underset{\sim}{\beta}/\sigma^2 = U$.

Now let us consider the other extreme where $\underset{\sim}{\beta} = 0$. From (3.2.1) and (3.2.2) we have

$$J(\underset{\sim}{W},\underset{\sim}{b}_{MW}) \leq \left(\frac{U-V_{FM}}{U+V_{FM}}\right)^2 V_{FM}. \tag{3.2.3}$$

Bhattacharyya has shown that the IMSE of $\underset{\sim}{b}_s$ is bounded by

$$\sigma^2 \left\{d_1+d_2+\frac{2}{N-p+2}\sum_{i=3}^{p} d_i\right\}.$$

Since

$$\begin{aligned} V &= \sigma^2 tr(\underset{\sim}{X}'\underset{\sim}{X})^{-1}W \\ &= \sigma^2 tr\ \underset{\sim}{L}^{-1}\underset{\sim}{L}'^{-1}\underset{\sim}{L}'\underset{\sim}{D}\underset{\sim}{L} \\ &= \sigma^2 \sum_{i=1}^{p} d_i \end{aligned}$$

we have that the IMSE of $\underset{\sim}{b}_s$ is bounded by $h\,V_{FM}$ where

$$h = \left\{\frac{d_1+d_2+\frac{2}{N-p+2}\sum_{i=3}^{p} d_i}{\sum_{i=1}^{P} d_i}\right\}. \tag{3.2.4}$$

Equating the multiples of V_{FM} in (3.2.3) and (3.2.4), we see that $J(\underset{\sim}{W},\underset{\sim}{b}_S) < J(\underset{\sim}{W},\underset{\sim}{b}^+_{MW})$ when

$$U > \left(\frac{1+h^{\frac{1}{2}}}{1-h^{\frac{1}{2}}}\right) V_{FM}, \tag{3.2.5}$$

and $\underset{\sim}{\beta} = \underset{\sim}{0}$.

If one is unwilling to assume a bound on $\underset{\sim}{\beta}'\underset{\sim}{W}\underset{\sim}{\beta}/\sigma^2$ smaller than the multiple of V_{FM} in (3.2.5), then $\underset{\sim}{b}_s$ would be preferable no matter what the true value of $\underset{\sim}{\beta}'\underset{\sim}{W}\underset{\sim}{\beta}/\sigma$.

We will now consider the effect that spurious correlations have on the bound in (3.2.5). First, consider the orthogonal case (no spurious correlations), that is, $(\underset{\sim}{X}'\underset{\sim}{X})^{-1} = \underset{\sim}{I}$. We assume that $\underset{\sim}{M} = \underset{\sim}{I}$. Then the d_i's are all one and from (3.1.4)

$$h = \frac{2+2(p-2)/N-p+2}{p}$$

$$= \frac{2N}{p(N-p+2)} ,$$

so that the multiple of V_{FM} in (3.1.5) is

$$\frac{1+h^{\frac{1}{2}}}{1-h^{\frac{1}{2}}} = \frac{\sqrt{p(N-p+2)} + \sqrt{2N}}{\sqrt{p(N-p+2)} - \sqrt{2N}} .$$

If $p = 10$ and $N = 36$ as in the 10 factor example of HK (1970b) then

$$\frac{1+h^{\frac{1}{2}}}{1-h^{\frac{1}{2}}} = 3.06 .$$

If $(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$ is any nonsingular matrix and $M = I$, then the d_i's are eigenvalues of $(\underset{\sim}{X}'\underset{\sim}{X})$. This would be the situation when the correlations between the independent variables are viewed as being spurious. When $\underset{\sim}{X}'\underset{\sim}{X}$ is ill conditioned, the eigenvalues are very spread out. In this situation the factor

$$\frac{1+h^{\frac{1}{2}}}{1-h^{\frac{1}{2}}}$$

will be larger then when $\underset{\sim}{X}'\underset{\sim}{X} = \underset{\sim}{I}$. We substituted into (3.2.4) the eigenvalues from the 10 factor example of Hoerl and Kennard (1970) to obtain

$$\frac{1+h^{\frac{1}{2}}}{1-h^{\frac{1}{2}}} = 6.89.$$

Thus, we see that the multiple of V_{FM} in (3.2.5) is larger when spurious correlations are present than in the orthogonal situation.

Apparently, in the ill-conditioned situation, $\underset{\sim}{b}_{MW}^{+}$ requires less exact prior knowledge in order to be competitive with $\underset{\sim}{b}_S$.

3.2.1. Prior Information About Exactly One of the Elements of $\underset{\sim}{\beta}/\sigma^2$.

Now let us consider the case where prior information is expressible in slightly different form. Instead of a bound on η^2/σ^2, we assume an a priori bound exists on one of the elements of $\underset{\sim}{\beta}/\sigma^2$, say β_2. We wish to incorporate this prior knowledge into our estimation procedure for $\underset{\sim}{\beta}$ because we feel that using $\hat{\underset{\sim}{\beta}}$ will result in an intolerably large IMSE. That is, failure is likely to result if our IMSE is too large. We do not define "too large" and ask the reader to bear with us. An application of this estimator follows in 4.1.

Consider the estimator

$$\underset{\sim}{x}'\underset{\sim}{b}_\delta = \underset{\sim}{x}'\underset{\sim}{K}\,\hat{\underset{\sim}{\beta}}$$

$$= \underset{\sim}{x}_1'\,\tilde{\underset{\sim}{\beta}}_1 - \delta(\underset{\sim}{x}_1'\,\underset{\sim}{A}\hat{\beta}_2 - x_2'\,\hat{\beta}_2).$$

where $\tilde{\underset{\sim}{\beta}}_1 = (\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\,\underset{\sim}{X}_1'\underset{\sim}{Y}$,

$$\hat{\beta}_2 = (\underset{\sim}{X}_2'(\underset{\sim}{I}-\underset{\sim}{X}_1(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1')\underset{\sim}{X}_2)^{-1}\underset{\sim}{X}_2'(\underset{\sim}{I}-\underset{\sim}{X}_1(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1')\underset{\sim}{Y},$$

and $\underset{\sim}{A} = (\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1'\underset{\sim}{X}_2$

and δ is a scalar constant. Note that when $\delta = 0$, $\underset{\sim}{b}_\delta$ is the least squares estimator for fitting the reduced model $\eta = \underset{\sim}{x}_1\underset{\sim}{\beta}_1$. When $\delta = 1$, then $\underset{\sim}{b}_\delta = \hat{\underset{\sim}{\beta}}$. Let us determine $\underset{\sim}{K}$. Now if $\hat{\underset{\sim}{\beta}}_1 = \tilde{\underset{\sim}{\beta}}_1 - \underset{\sim}{A}\hat{\beta}_2$ then

$$\underset{\sim}{x}'\underset{\sim}{b}_\delta = [\underset{\sim}{x}_1 \,\vdots\, x_2]'\begin{bmatrix} \underset{\sim}{I} & (1-\delta)\underset{\sim}{A} \\ \underset{\sim}{0} & \delta \end{bmatrix}\begin{bmatrix} \hat{\underset{\sim}{\beta}}_1 \\ \hat{\beta}_2 \end{bmatrix} = \underset{\sim}{x}'\underset{\sim}{K}\hat{\underset{\sim}{\beta}}\,, \qquad (3.2.1.1)$$

whence $$\underset{\sim}{K} = \begin{bmatrix} \underset{\sim}{I} & (1-\delta)\underset{\sim}{A} \\ \underset{\sim}{0} & \delta \end{bmatrix}.$$

The expected value of $\underset{\sim}{b}_\delta$ is

$$\underset{\sim}{K}\underset{\sim}{\beta}$$

and the bias of $\underset{\sim}{x}'\underset{\sim}{b}_\delta$ is

$$\underset{\sim}{x}'(\underset{\sim}{K}-\underset{\sim}{I})\underset{\sim}{\beta} = x_1'(1-\delta)\underset{\sim}{A}\beta_2 + (\delta-1)x_2\beta_2$$

$$= (1-\delta)\{\underset{\sim}{x}_1\underset{\sim}{A}\beta_2 - x_2\beta_2\}$$

$$= (1-\delta)(\underset{\sim}{x}_1'\underset{\sim}{A} - x_2)\,\beta_2$$

The integrated bias is

$$B_\delta = (1-\delta)^2 \frac{N\beta_2^2}{\sigma^2} \operatorname{tr}(\underset{\sim}{A}'\underset{\sim}{W}_1\underset{\sim}{A} - 2\underset{\sim}{A}'\underset{\sim}{W}_2 + W_3)$$

The integrated variance of $\underset{\sim}{x}'\underset{\sim}{b}_\delta$ is

$$V_\delta = N \operatorname{tr} \underset{\sim}{K}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{K}'\underset{\sim}{W} .$$

Now $\underset{\sim}{K}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{K}' =$

$$\begin{bmatrix} \underset{\sim}{I} & (1-\delta)\underset{\sim}{A} \\ \underset{\sim}{0} & \delta \end{bmatrix} \cdot \left(\begin{bmatrix} (\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1} & \underset{\sim}{0} \\ \underset{\sim}{0} & 0 \end{bmatrix} + \begin{bmatrix} -\underset{\sim}{A} \\ 1 \end{bmatrix} G[-\underset{\sim}{A}' \;\; 1] \right)$$

$$\cdot \begin{bmatrix} \underset{\sim}{I} & (1-\delta)\underset{\sim}{A} \\ 0 & \delta \end{bmatrix}$$

$$= \begin{bmatrix} (\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1} & \underset{\sim}{0} \\ \underset{\sim}{0} & 0 \end{bmatrix} + \delta^2 \begin{bmatrix} -\underset{\sim}{A} \\ 1 \end{bmatrix} G^{-1}[-\underset{\sim}{A}' \;\; 1] ,$$

where $$G = \underset{\sim}{X}_2'(\underset{\sim}{I}-\underset{\sim}{X}_1(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1)\underset{\sim}{X}_2 .$$

So $$V_\delta = V_{RM} + \delta^2 V_\Delta ,$$

where V_{RM} is the integrated variance of $\tilde{\underset{\sim}{\beta}}$ and V_Δ is the integrated variance of $\hat{\underset{\sim}{\beta}}$ minus the integrated variance of $\tilde{\underset{\sim}{\beta}}$. The IMSE of $\underset{\sim}{b}_\delta$ is

$$J(\underset{\sim}{W}, \underset{\sim}{b}_\delta) = V_{RM} + \delta^2 V_\Delta + (1-\delta)^2 B_{RM} .$$

Now

$$B_{RM} = \frac{N}{\sigma^2} \beta_2^2 \ \mathrm{tr} \begin{bmatrix} -\underset{\sim}{A} \\ 1 \end{bmatrix} G^{-1} [-\underset{\sim}{A}' \ \vdots \ 1] \underset{\sim}{W}$$

$$= \frac{N}{\sigma^2} \beta_2^2 G \ \mathrm{tr} [-\underset{\sim}{A}' \vdots \ 1] \underset{\sim}{W} \begin{bmatrix} -\underset{\sim}{A} \\ 1 \end{bmatrix}$$

$$= \frac{\beta_2^2 G}{\sigma^2} V_\Delta \ ,$$

so that

$$J(\underset{\sim}{W}, \underset{\sim}{b}_\delta) = V_{RM} + \left(\delta^2 + (1-\delta)^2 \frac{\beta_2^2 G}{\sigma^2} \right) V_\Delta .$$

Note that $J(\underset{\sim}{W}, \underset{\sim}{b}_0)$ and $J(\underset{\sim}{W}, \underset{\sim}{b}_1)$ are the IMSE for $\tilde{\underset{\sim}{\beta}}_1$ and $\hat{\underset{\sim}{\beta}}$, , respectively. It can be shown that $J(\underset{\sim}{W}, \underset{\sim}{b}_\delta)$ is minimized when

$$\delta = \frac{\beta_2^2 G}{\beta_2^2 G + \sigma^2} .$$

Now

$$J(\underset{\sim}{W}, \underset{\sim}{b}_\delta) - J(\underset{\sim}{W}, \underset{\sim}{b}_0) = \left\{ \delta^2 + (\delta^2 - 2\delta) \frac{\beta_2^2 G}{\sigma^2} \right\} V_\Delta .$$

and

$$J(\underset{\sim}{W}, \underset{\sim}{b}_\delta) - J(\underset{\sim}{W}, \underset{\sim}{b}_1) = \left\{ \delta^2 - 1 + (1-\delta)^2 \frac{\beta_2^2 G}{\sigma^2} \right\} V_\Delta .$$

Since

$$J(\underset{\sim}{W},\underset{\sim}{b}_\delta)-J(\underset{\sim}{W},\underset{\sim}{b}_0) < 0 \quad \text{when} \quad 0 < \delta < \frac{2\beta_2^2 G}{\sigma^2+\beta_2^2 G}$$

and

$$J(\underset{\sim}{W},\underset{\sim}{b}_\delta)-J(\underset{\sim}{W},\underset{\sim}{b}_1) < 0 \quad \text{when} \quad \frac{\beta_2^2 G-\sigma^2}{\beta_2^2 G+\sigma^2} < \delta < 1,$$

we would prefer the estimator (3.2.1.1) whenever

$$\max\left(0 \ , \frac{\beta_2^2 \ G - \sigma^2}{\beta_2^2 \ G + \sigma^2}\right) < \delta < \min\left(1, \frac{2\beta_2^2 \ G}{\beta_2^2 \ G+\sigma^2}\right) \ . \qquad (3.2.1.2)$$

To proceed further, suppose that it is possible to specify a bound U so that

$$\frac{\beta_2^2}{\sigma^2} < U.$$

When $\beta_2^2/\sigma^2 < U$, then the interval

$$\max\left(0, \frac{UG-1}{UG+1}\right) < \delta < \min\left(1, \frac{2UG}{UG+1}\right)$$

is contained in the interval (3.2.1.2). If we agree to use

$$\delta^* = \max\left(0, \frac{UG-1}{UG+1}\right)$$

then it follows that

$$J(\underset{\sim}{W}, \underset{\sim}{b}_\delta *) < \min\{J(\underset{\sim}{W}, \underset{\sim}{b}_0), J(\underset{\sim}{W}, \underset{\sim}{b}_1)\}$$

if

$$\frac{U-G^{-1}}{UG+3} < \beta_2/\sigma^2 < U \ .$$

3.3. A Characterization of Some Biased Estimators

In situations where one is not fitting the true model, there is a desire to achieve small IMSE for the model one has chosen to fit. In the last section we have seen one aspect pertaining to this problem. That aspect is the danger that such an estimator will have an unbounded IMSE. Specifically, if any $\underset{\sim}{\beta}$ in p dimensional space is possible, and one has chosen a bound U on the IMSE, then there are $\underset{\sim}{\beta}$ and σ^2 combinations that will yield an IMSE larger than U for any estimator $\underset{\sim}{K}\hat{\underset{\sim}{\beta}}$ where $\underset{\sim}{K}$ is deterministic and not equal to $\underset{\sim}{I}$. In practice, this fact is not always considered. Our argument rests on the notion that whenever one fits a reduced model to data, then he has prior convictions that certain values of $\underset{\sim}{\beta}$ and σ^2 will not occur. We further assume that he is willing to state his prior convictions in the form of a bound on the parameter space. We will use these assumptions somewhat freely in what follows.

Consider the case where the true response is

$$\eta = \underset{\sim}{x}'\underset{\sim}{\beta} = \underset{\sim}{x}_1'\underset{\sim}{\beta}_1 + \underset{\sim}{x}_2'\underset{\sim}{\beta}_2 \ . \qquad (3.3.1)$$

Let $\underset{N\times 1}{\underset{\sim}{Y}}$ be a vector of uncorrelated responses and let

$\underset{N\times p}{\underset{\sim}{X}} = (\underset{N\times p_1}{\underset{\sim}{X}_1} \ \vdots \ \underset{N\times p_2}{\underset{\sim}{X}_2})$, so that

$$E(\underset{\sim}{Y}) = \underset{\sim}{X}\underset{\sim}{\beta} = \underset{\sim}{X}_1\underset{\sim}{\beta}_1 + \underset{\sim}{X}_2\underset{\sim}{\beta}_2 \ \text{ and } \ \operatorname{Var} \underset{\sim}{Y} = \sigma^2 \underset{\sim}{I} \ .$$

The KM, BD, KMH, and ridge estimators can all be characterized as linear transforms of the least squares estimator of the full model. That is,

$$b = \underset{pxp}{\underset{\sim}{K}}\ \hat{\underset{\sim}{\beta}}$$

where $\underset{\sim}{K} = \begin{bmatrix} \underset{\sim}{K}_1 \\ \hdashline \underset{\sim}{0} \end{bmatrix}$ and $\underset{\sim}{K}_1$ is a $p_1 xp$ matrix. Define

$$\underset{p_1 xp_2}{\underset{\sim}{A}} = (\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\ \underset{\sim}{X}_1'\underset{\sim}{X}_2$$

Then for the BD estimator, $\underset{\sim}{b}_{BD} = (\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1'\underset{\sim}{Y}$, so that the matrix $\underset{\sim}{K}_1$ for this estimator takes on the value

$$\underset{\sim}{K}_1 = (\underset{\sim}{I}_{p_1} \vdots \underset{p_1 xp_2}{\underset{\sim}{A}})$$

The KMH estimator has a $\underset{\sim}{K}_1$ matrix of the form

$$\underset{\sim}{K}_1 = (\underset{\sim}{I}_{p_1} \vdots \underset{\sim}{W}_1^{-1}\ \underset{\sim}{W}_2)$$

The ridge estimator has a $\underset{\sim}{K}_1$ matrix of the form

$$\begin{bmatrix} \underset{\sim}{I}_{p_1} & \vdots & \underset{p_1 xp_2}{\underset{\sim}{0}} \end{bmatrix} (\underset{\sim}{X}'\underset{\sim}{X} + k\underset{\sim}{I})^{-1}\underset{\sim}{X}'\underset{\sim}{X}$$

where k is a scalar. The value of k to be used is based on $\underset{\sim}{Y}$, whence k is stochastic. The KM estimator has a $\underset{\sim}{K}_1$ matrix of

$$\underset{\sim}{K}_{KM}\ [\underset{\sim}{I} \vdots \underset{\sim}{A}] = [\underset{\sim}{K}_{KM} \vdots \underset{\sim}{K}_{KM}\ \underset{\sim}{A}],$$

where $\underset{\sim}{K}_{KM}$ is a matrix of constants. These constants are chosen based on prior bounds on the parameter space.

3.3.1 Consideration of the Estimators of the form $\tilde{\underset{\sim}{\beta}}_1 + \underset{\sim}{K}_2\hat{\underset{\sim}{\beta}}_2$

At this point we will consider a particular class. To simplify the discussion, we shall refer to only the first p_1 elements $\underset{\sim}{b}_1$ of the estimator $\underset{\sim}{b}$. That is

$$\underset{1xp}{\underset{\sim}{b}} = \begin{bmatrix} \underset{1xp_1}{\underset{\sim}{b}_1} \\ \underset{1xp_2}{\underset{\sim}{0}} \end{bmatrix} .$$

We shall work in the form

$$\begin{bmatrix} \tilde{\underset{\sim}{\beta}}_1 \\ \hat{\underset{\sim}{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \underset{\sim}{I} & \underset{\sim}{A} \\ \underset{\sim}{0} & \underset{\sim}{I} \end{bmatrix} \begin{bmatrix} \hat{\underset{\sim}{\beta}}_1 \\ \hat{\underset{\sim}{\beta}}_2 \end{bmatrix}$$

or

$$\hat{\underset{\sim}{\beta}} = \begin{bmatrix} \hat{\underset{\sim}{\beta}}_1 \\ \hat{\underset{\sim}{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \underset{\sim}{I} & -\underset{\sim}{A} \\ \underset{\sim}{0} & \underset{\sim}{I} \end{bmatrix} \begin{bmatrix} \tilde{\underset{\sim}{\beta}}_1 \\ \hat{\underset{\sim}{\beta}}_2 \end{bmatrix} .$$

So that for the BD estimator

$$[\underset{\sim}{I}\ \underset{\sim}{A}] \begin{bmatrix} \hat{\underset{\sim}{\beta}}_1 \\ \hat{\underset{\sim}{\beta}}_2 \end{bmatrix} = [\underset{\sim}{I}\ \underset{\sim}{A}] \begin{bmatrix} \underset{\sim}{I} & -\underset{\sim}{A} \\ \underset{\sim}{0} & \underset{\sim}{I} \end{bmatrix} \begin{bmatrix} \tilde{\underset{\sim}{\beta}}_1 \\ \hat{\underset{\sim}{\beta}}_2 \end{bmatrix} = [\underset{\sim}{I}\ \underset{\sim}{0}] \begin{bmatrix} \tilde{\underset{\sim}{\beta}}_1 \\ \hat{\underset{\sim}{\beta}}_2 \end{bmatrix}$$

and for the KMH estimator

$$[\underset{\sim}{I} \vdots \underset{\sim}{W}_1^{-1}\underset{\sim}{W}_2]\begin{bmatrix}\hat{\underset{\sim}{\beta}}_1\\ \hat{\underset{\sim}{\beta}}_2\end{bmatrix} = [\underset{\sim}{I} \vdots \underset{\sim}{W}^{-1}\underset{\sim}{W}_2]\begin{bmatrix}\underset{\sim}{I} & -\underset{\sim}{A}\\ \underset{\sim}{0} & \underset{\sim}{I}\end{bmatrix}\begin{bmatrix}\tilde{\underset{\sim}{\beta}}_1\\ \hat{\underset{\sim}{\beta}}_2\end{bmatrix} =$$

$$= (I \vdots (W_1^{-1}W_2-A))\begin{bmatrix}\tilde{\underset{\sim}{\beta}}\\ \hat{\underset{\sim}{\beta}}_2\end{bmatrix} .$$

We shall consider estimators of the form

$$\underset{p_1 x1}{\underset{\sim}{b}} = (\underset{p_1 x p_1}{\underset{\sim}{I}} \vdots \underset{p_1 x p_2}{\underset{\sim}{K}_2})\begin{bmatrix}\tilde{\underset{\sim}{\beta}}_1\\ \hat{\underset{\sim}{\beta}}_2\end{bmatrix} \tag{3.3.2}$$

In this class of estimators, $\underset{\sim}{b}_{BD}$ attains minimum integrated variance and $\underset{\sim}{b}_{KMH}$ attains minimum integrated bias. Let us proceed to find that $\underset{\sim}{K}_2$ that minimizes the IMSE, the sum of integrated variance and integrated bias. For an arbitrary $\underset{\sim}{K}_2$, the integrated mean square error of the estimator (3.2.2) when (3.2.1) holds is

$$J(\underset{\sim}{W}, \underset{\sim}{b}) = N \text{ tr}\{(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1} + \underset{\sim}{K}_2\underset{\sim}{M}^{-1}\underset{\sim}{K}_2'\}\underset{\sim}{W}_1$$

$$+ \frac{N}{\sigma^2} \underset{\sim}{\beta}_2'\{\underset{\sim}{K}_2'\underset{\sim}{W}_1\underset{\sim}{K}_2 + 2(\underset{\sim}{A}'\underset{\sim}{W}_1 - \underset{\sim}{W}_2')\underset{\sim}{K}_2 + \underset{\sim}{W}_3$$

$$- 2\,\underset{\sim}{W}_2'\underset{\sim}{A} + \underset{\sim}{A}'\underset{\sim}{W}_1\underset{\sim}{A}\}\underset{\sim}{\beta}_2 .$$

Here $\underset{\sim}{M} = \underset{\sim}{X}_2'(\underset{\sim}{I}-\underset{\sim}{X}_1(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1)\underset{\sim}{X}_2$. This is what G was. We will be using the symbol G for another purpose. We want to differentiate J with respect to $\underset{\sim}{K}_2$ and set the expression to zero. Instead of directly differentiating J with respect to $\underset{\sim}{K}_2$, let $\underset{\sim}{K}_2 = \underset{\sim}{H}\underset{\sim}{G}$, where $\underset{\sim}{H}$ is square, and then differentiate J with respect to $\underset{\sim}{H}$. The derivative is

$$\frac{\partial J}{\partial \underset{\sim}{H}} = \frac{2N}{\sigma^2}\{\underset{\sim}{H}\underset{\sim}{W}_1\underset{\sim}{G}\beta_2\beta_2'\underset{\sim}{G}' + (\underset{\sim}{W}_1\underset{\sim}{A}-\underset{\sim}{W}_2)\beta_2\beta_2'\underset{\sim}{G}\} + 2N\underset{\sim}{H}\underset{\sim}{W}_1\underset{\sim}{G}\underset{\sim}{M}^{-1}\underset{\sim}{G}',$$

and $\frac{\partial J}{\partial \underset{\sim}{H}} = 0$ when

$$\underset{\sim}{H}\underset{\sim}{W}_1\underset{\sim}{G}\left(\frac{\beta_2\beta_2'}{\sigma^2} + \underset{\sim}{M}^{-1}\right)\underset{\sim}{G}' = (\underset{\sim}{W}_2-\underset{\sim}{W}_1\underset{\sim}{A})\frac{\beta_2\beta_2'}{\sigma^2}\underset{\sim}{G}'. \tag{3.3.3}$$

In order that (3.2.3) hold for all $\underset{\sim}{G}$, it must be that

$$\underset{\sim}{H}\underset{\sim}{W}_1\underset{\sim}{G}\left(\frac{\beta_2\beta_2'}{\sigma^2} + \underset{\sim}{M}^{-1}\right) = (\underset{\sim}{W}_2-\underset{\sim}{W}_1\underset{\sim}{A})\frac{\beta_2\beta_2'}{\sigma^2},$$

or

$$\underset{\sim}{H}\underset{\sim}{W}_1\underset{\sim}{G} = (\underset{\sim}{W}_2-\underset{\sim}{W}_1\underset{\sim}{A})\frac{\beta_2\beta_2}{\sigma^2}\left(\frac{\beta_2\beta_2'}{\sigma^2} + \underset{\sim}{M}^{-1}\right)^{-1}.$$

A solution that holds for all $\frac{\beta_2\beta_2'}{\sigma^2}$ is

$$\underset{\sim}{G} = (\underset{\sim}{W}_2-\underset{\sim}{W}_1\underset{\sim}{A})\frac{\beta_2\beta_2'}{\sigma^2}\left(\frac{\beta_2\beta_2'}{\sigma^2} + \underset{\sim}{M}^{-1}\right)^{-1}$$

and $\underset{\sim}{H} = \underset{\sim}{W}_1^{-1}$.

It appears that the minimum IMSE estimator of the form (3.3.2) is

$$b^* = \tilde{\beta}_1 + (\underset{\sim}{W}_1^{-1}\underset{\sim}{W}_2-\underset{\sim}{A})\frac{\beta_2\beta_2'}{\sigma^2}\left(\frac{\beta_2\beta_2'}{\sigma^2} + \underset{\sim}{M}^{-1}\right)\hat{\beta}_2 . \tag{3.3.4}$$

In order to prove that $\underset{\sim}{b}^*$ is the minimum IMSE estimator of the form (3. 3.2), it is necessary to take an arbitrary $\underset{\sim}{K}_2$ and show that

$$J(\underset{\sim}{W}, \underset{\sim}{b}^*) < J(\underset{\sim}{W}, \underset{\sim}{b}) .$$

Let K_2^* be defined so that

$$b^* = \tilde{\beta}_1 + K_2^* \hat{\beta}_2 .$$

Then

$$J(\underset{\sim}{W}, \underset{\sim}{b}^*) - J(\underset{\sim}{W}, \underset{\sim}{b}) = \frac{N}{\sigma^2} \beta_2'\{\underset{\sim}{K}_2^{*'} \underset{\sim}{W}_1 \underset{\sim}{K}_2^* - \underset{\sim}{K}_2 \underset{\sim}{W}_1 \underset{\sim}{K}_2 +$$

$$2(\underset{\sim}{A}'\underset{\sim}{W}_1 - \underset{\sim}{W}_2')(\underset{\sim}{K}_2^* - \underset{\sim}{K}_2)\}\beta_2$$

$$+ N \operatorname{tr}(\underset{\sim}{K}_2^* \underset{\sim}{M}^{-1} \underset{\sim}{K}_2^{*'} - \underset{\sim}{K}_2 \underset{\sim}{M}^{-1} \underset{\sim}{K}_2')\underset{\sim}{W}_1$$

$$= N \operatorname{tr} \underset{\sim}{K}_2^* \left(\underset{\sim}{M}^{-1} + \frac{\beta_2 \beta_2'}{\sigma^2} \right) \underset{\sim}{K}_2^{*'} \underset{\sim}{W}_1$$

$$+ 2 \frac{N}{\sigma^2} \beta_2'(\underset{\sim}{A}'\underset{\sim}{W}_1 - \underset{\sim}{W}_2')(\underset{\sim}{K}_2^* - \underset{\sim}{K}_2)\beta_2$$

$$- N \operatorname{tr} \underset{\sim}{K}_2(\underset{\sim}{M}^{-1} + \beta_2\beta_2'/\sigma^2)\underset{\sim}{K}_2'\underset{\sim}{W}_1 . \qquad (3.3.5)$$

From (3.3.4) we have

$$K_2^* = (\underset{\sim}{W}_1^{-1}\underset{\sim}{W}_2 - \underset{\sim}{A})\beta_2\beta_2'/\sigma^2(\underset{\sim}{M}^{-1} + \beta_2\beta_2'/\sigma^2)^{-1}$$

and letting

$$\underset{\sim}{Q} = \underset{\sim}{M}^{-1} + \beta_2\beta_2'/\sigma^2 \quad \text{then we have}$$

$$J(\underset{\sim}{W}, \underset{\sim}{b}^*) - J(\underset{\sim}{W}, \underset{\sim}{b}) =$$

$$N \operatorname{tr} \underset{\sim}{K}_2^* \underset{\sim}{Q} \underset{\sim}{K}_2^{*'} \underset{\sim}{W}_1$$

$$- 2N \operatorname{tr} \beta_2\beta_2'/\sigma^2 (\underset{\sim}{A}'\underset{\sim}{W}_1 - \underset{\sim}{W}_2')\underset{\sim}{K}_2^*$$

$$-2N \operatorname{tr} \beta_2\beta_2'/\sigma^2 (A'W_1-W_2')K_2$$

$$- N \operatorname{tr} K_2 Q K_2' W_1 \; . \tag{3.3.6}$$

Now

$$-2N \operatorname{tr}(\beta_2\beta_2'/\sigma^2)(A'W_1-W_2')K_2^* = -2N \operatorname{tr} K_2^* Q K_2^{*\prime} W_1$$

and

$$-2N \operatorname{tr}(\beta_2\beta_2'/\sigma^2)(A'W_1-W_2')K_2 = 2N \operatorname{tr} K_2' W_1 K_2^* Q \; .$$

Substituting these expressions into (3.3.6) and completing the square we have

$$J(W, b^*) - J(W, b) = - N \operatorname{tr}(K_2^*-K_2)'W_1(K_2^*-K_2)Q \; . \tag{3.3.7}$$

Now W_1 and Q are positive semidefinite so $(K_2^*-K_2)'W_1(K_2^*-K_2)$ is also positive semidefinite. This means that the entire product is positive semidefinite. Since the trace of a positive semidefinite matrix is nonnegative we have from (3.3.7) that the IMSE of b^* is always less than the IMSE of b. So b^* attains minimum IMSE of all estimators of the form (3.3.2).

The form of b^* is of interest for several reasons. First, it can be seen that generally b^* involves unknown parameters, and therefore is not generally useable. However, if the design satisfies the BD condition, then $K_2^* = 0$ and the BD estimator is the minimum IMSE estimator of this form.

Now let's see a way that the KMH estimator is related to b^*. The expected value of b^* is

$$E(b^*) = \beta_1 + A\beta_2 + (W_1^{-1}W_2-A) \frac{\beta_2\beta_2'}{\sigma^2} Q^{-1} \beta_2 \; .$$

Using the fact that M is positive definite and (See Lemma 1, Mayer and Willke (1973)) that

$$Q^{-1} = \left[1 - \frac{1}{1+\frac{\beta_2'\beta_2}{\sigma^2}} \quad \frac{\beta_2\beta_2'}{\sigma^2} \right]$$

it can be shown that

$$\frac{\beta_2\beta_2'}{\sigma^2}\left(\frac{\beta_2\beta_2'}{\sigma^2} + M^{-1}\right) = \frac{1}{\sigma^2+\beta_2' M\beta_2} \quad \beta_2\beta_2' M \quad .$$

It follows that the expected value of b^* is

$$E(b^*) = \beta_1 + A\beta_2 + (W_1^{-1}W_2-A)\left[\frac{\beta_2'M\beta_2}{\sigma^2+\beta_2'M\beta_2}\right]\beta_2 \ . \qquad (3.3.8)$$

Thus, $K_2\hat{\beta}_2$ for K_2 given by (3. 38) is

$$\frac{\beta_2' M \beta_2}{\sigma^2+\beta_2'M\beta_2}(W_1^{-1}W_2-A)\beta_2 = \delta_E(W_1^{-1}W_2-A)\beta_2 \ ,$$

which is just a scalar (δ_E) multiple of the KMH minimum bias error correction term. Clearly, δ_E is a function of unknown parameters and δ_E is a monotonically increasing function of $\beta_2 M\beta_2/\sigma^2$ bounded above by one.

Equation (3. 38) can be used to gain different views of the BD and KMH estimators. The BD estimator can be viewed as minimum variance unbiased estimator of $E(b^*)$ when $\beta_2'M\beta_2$ is zero. The KMH estimator can be viewed as the minimum variance unbiased estimator of

$$\lim_{\beta_2' M \beta_2 \to \infty} E(b^*) \ .$$

If the BD and KMH estimator had been developed in order to attain the same expected value as $\underset{\sim}{b}^*$, then both estimators do so when $\underset{\sim}{\beta}_2' \underset{\sim}{M} \underset{\sim}{\beta}_2/\sigma^2$ takes on an extreme value (when $\underset{\sim}{\beta}_2' \underset{\sim}{M} \underset{\sim}{\beta}_2/\sigma^2 = 0$ for the BD estimator or for very large values of $\underset{\sim}{\beta}_2' \underset{\sim}{M} \underset{\sim}{\beta}_2/\sigma^2$ for the KMH estimator).

Since we considered $E(\underset{\sim}{b}^*)$ to gain a different view of the KMH estimator, we will confine our remarks to the KMH estimator. First, compare the relationship between the KMH estimator and $E(\underset{\sim}{b}^*)$ to the relationship between $\hat{\underset{\sim}{\beta}}$ and the expected value of the minimum IMSE estimator in K. We have not formally considered this last relationship, but certainly this last relationship must be similar to the first relationship. Consider $\delta^*_{MW} \hat{\underset{\sim}{\beta}}$, an estimator discussed on page 43. Its expected value is $\delta^*_{MW} \underset{\sim}{\beta}$, and $\hat{\underset{\sim}{\beta}}$ is unbiased for $\lim\limits_{B_0 \to \infty} E(\delta^*_{MW} \hat{\underset{\sim}{\beta}})$. Thus, it would appear that these two relationships are very similar. Since the relationships are similar, one might feel that the KMH estimator is similar to $\hat{\underset{\sim}{\beta}}$—an often critized estimator but also an often used estimator. However, this comparison ignores an important point.

The KMH estimator is intended to be used in situations where the integrated bias error is nonzero. The full model least squares estimator $\hat{\underset{\sim}{\beta}}$ is intended for use when the integrated bias is zero. The KMH estimator is the minimum bias estimator for estimates of the form

$$\begin{bmatrix} \underset{\sim}{K}_1 \\ \underset{\sim}{0} \end{bmatrix} \hat{\underset{\sim}{\beta}} \, .$$

However, the increase in integrated variance of the KMH estimator may not be warranted in light of the fact that one is going to end up with a biased estimator.

The KMH estimator has been praised by its authors because one can use it without having any prior knowledge of the parameters. We see this as misleading to potential users of this estimator. As we stated earlier, we assume users have prior convictions that certain values of $\underset{\sim}{\beta}$ and σ^2 will not occur. Thus the fact that the KMH estimator doesn't use the prior knowledge is actually a hindrance. It's something one can use when he has prior knowledge, but for some reason doesn't want to use the prior knowledge. In other words, the KMH estimator is a fancy way of camouflaging one's subjectivity in a maze of matrices. It would seem that if one had some prior knowledge, it would be more honest to try to quantify it instead of hiding it. When one considers that such quantification may lead to an improved estimator, it seems worth-while to attempt quantification of prior knowledge.

3.3.2. The SK Estimator.

We now consider the Svendsgaard-Kupper (SK) estimator. Consider the deterministic $\underset{\sim}{K}$ matrix

$$\underset{\sim}{K}' = \begin{bmatrix} \underset{\sim}{K}_1 \\ \hdashline \underset{\sim}{0} \end{bmatrix}$$

where $\underset{\sim}{K}_1 = [I_{p_1} \vdots \underset{\sim}{A} + \delta(\underset{\sim}{W}_1^{-1}\underset{\sim}{W}_2 - \underset{\sim}{A})]$

so that

$$\begin{aligned} \underset{\sim}{b}_{SK} &= \underset{\sim}{K}\,\hat{\underset{\sim}{\beta}} \\ &= \hat{\underset{\sim}{\beta}} + \underset{\sim}{A}\hat{\underset{\sim}{\beta}}_2 + \delta(\underset{\sim}{W}_1^{-1}\underset{\sim}{W}_2 - \underset{\sim}{A})\hat{\underset{\sim}{\beta}}_2 \\ &= \tilde{\underset{\sim}{\beta}}_1 + \delta(\underset{\sim}{W}^{-1}\underset{\sim}{W}_2 - \underset{\sim}{A})\hat{\underset{\sim}{\beta}}_2\,, \end{aligned} \qquad (3.39)$$

where δ is a constant. The IMSE of $\underset{\sim}{b}_{SK}$ is

$$J(\underset{\sim}{W}, \underset{\sim}{b}_{SK}) = (\delta-1)^2 N\gamma + \delta^2 N\tau + \frac{N}{\sigma^2} \underset{\sim}{\beta}_2'(\underset{\sim}{W}_3 - \underset{\sim}{W}_2^1 \underset{\sim}{W}_1^{-1} \underset{\sim}{W}_2)\underset{\sim}{\beta}_2$$

$$+ N \operatorname{tr} \underset{\sim}{W}_1(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}$$

where $\gamma = \dfrac{\underset{\sim}{\beta}_2'\underset{\sim}{W}_0\underset{\sim}{\beta}_2}{\sigma^2}$ and $\tau = \operatorname{tr}[\underset{\sim}{W}_0(\underset{\sim}{X}_2'\underset{\sim}{C}\underset{\sim}{X}_2)^{-1}]$

with

$$\underset{\sim}{W}_0 = (\underset{\sim}{W}_1^{-1}\underset{\sim}{W}_2 - \underset{\sim}{A})' \; \underset{\sim}{W}_1 \; (\underset{\sim}{W}_1^{-1}\underset{\sim}{W}_2 - \underset{\sim}{A})$$

and $\underset{\sim}{C} = \underset{\sim}{I} - \underset{\sim}{X}_1(\underset{\sim}{X}_1'\underset{\sim}{X}_1)^{-1}\underset{\sim}{X}_1'$. Let us change the arguments of $J(\cdot)$ to $J(\delta)$. Note that $J(0)$ and $J(1)$ are, respectively, the corresponding BD and KMH expressions for J.

Now,

$$J(\delta) - J(0) = N(\delta^2-2\delta)\gamma + N\delta^2\tau$$

and

$$J(\delta) - J(1) = N(\delta-1)^2\gamma + N(\delta^2-1)\tau.$$

Since

$$J(\delta) - J(0) < 0 \quad \text{when} \quad 0 < \delta < \frac{2\gamma}{(\gamma+\tau)}$$

and

$$J(\delta) - J(1) < 0 \quad \text{when} \quad \frac{(\gamma-\tau)}{(\gamma+\tau)} < \delta < 1,$$

we would prefer the estimator (3.3.9) whenever

$$\max\left[0, \frac{\gamma-\tau}{\gamma+\tau}\right] < \delta < \min\left[1, \frac{2\gamma}{\gamma+\tau}\right]. \tag{3.3.10}$$

To proceed further, suppose that it is possible to specify an upper bound U for γ. This can be done, for example, in the following way. Because the BD integrated squared bias can be written as the sum of two non-negative terms, namely

$$N\gamma + \frac{N}{\sigma^2} \underset{\sim}{\beta}_2'(\underset{\sim}{W}_3 - \underset{\sim}{W}_2'\underset{\sim}{W}_1^{-1}\underset{\sim}{W}_2)\underset{\sim}{\beta}_2 ,$$

and because this quantity should be less than

$$\frac{N\Omega}{\sigma^2} \int_R \eta^2 d\underset{\sim}{x} ,$$

which is the value of B for that estimator identically equal to zero, it follows that

$$\gamma < U \quad \text{when} \quad \frac{\eta^2}{\sigma^2} < U.$$

Thus, one can construct an upper bound for γ by considering η^2 and σ^2.

If we assume that $\gamma < U$, then the interval

$$\max\left(0, \frac{U-\tau}{U+\tau}\right) < \delta < \min\left(1, \frac{2\gamma}{\gamma+\tau}\right)$$

is contained in the interval (3.3.10). If we agree to use

$$\delta^* = \max\left(0, \frac{U-\tau}{U+\tau}\right) ,$$

it follows that

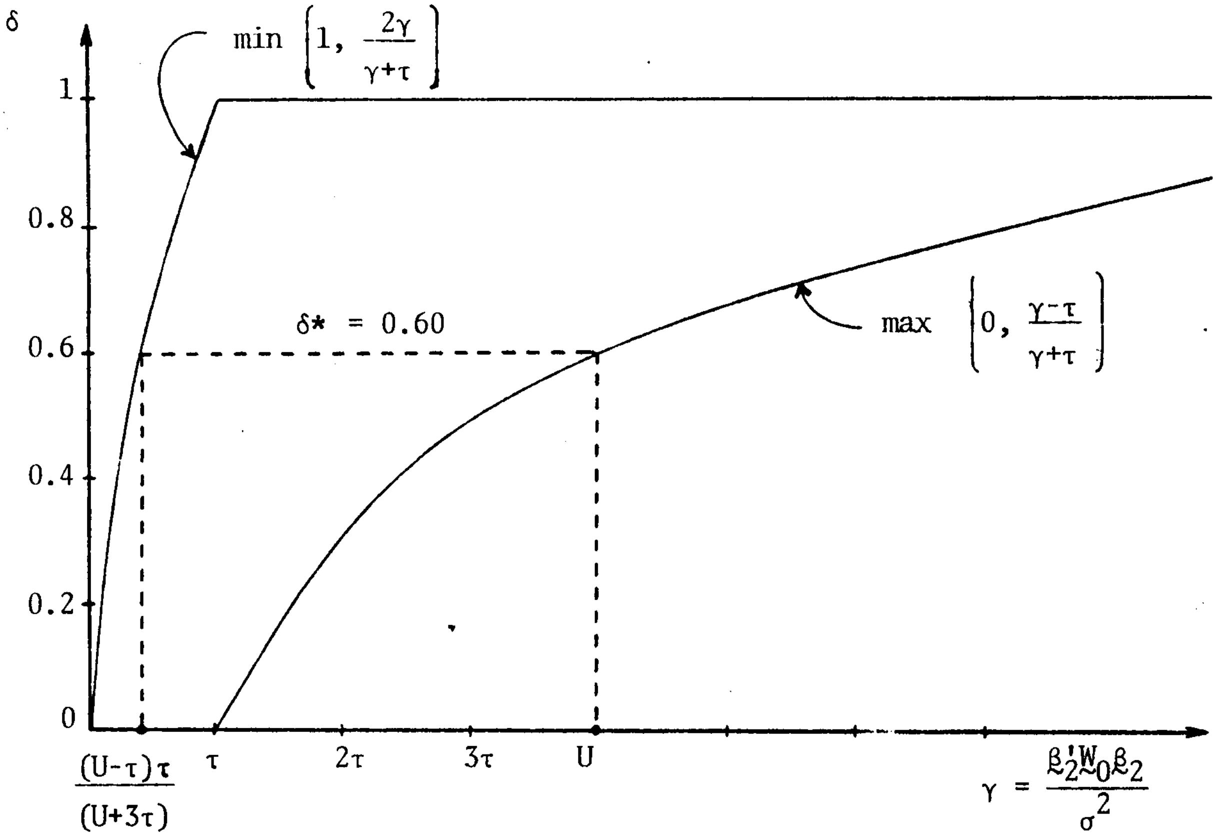
$$J(\delta^*) < \min[J(0), J(1)]$$

if

$$\frac{(U-\tau)\tau}{(U+3\tau)} < \gamma < U.$$

Figure (3.3.2.1) illustrates these concepts.

FIGURE (3.3.2.1). Illustration of Mechanism for Choosing δ^*



3.3.3. Estimation of γ.

It is possible to find an unbiased estimator of γ by using the following easily demonstrated result:

If

$$E(\underset{\sim}{\xi}) = \underset{\sim}{\mu} \quad \text{and} \quad Var(\underset{\sim}{\xi}) = \underset{\sim}{V},$$

then

$$E(\underset{\sim}{\xi}'\underset{\sim}{Q}\underset{\sim}{\xi}) = \underset{\sim}{\mu}'\underset{\sim}{Q}\underset{\sim}{\mu} + tr(\underset{\sim}{Q}\underset{\sim}{V}).$$

In our case,

$$E(\hat{\underset{\sim}{\beta}}_2) = \underset{\sim}{\beta}_2 \quad \text{and} \quad Var(\hat{\underset{\sim}{\beta}}_2) = \sigma^2(\underset{\sim}{X}_2'\underset{\sim}{C}\underset{\sim}{X}_2)^{-1},$$

so that

$$E(\hat{\beta}_2'\underset{\sim}{W}_0\hat{\beta}_2) = \sigma^2(\gamma+\tau).$$

Under the usual normality assumptions,

$$s^2 = \frac{1}{(N-p)}\underset{\sim}{Y}'[\underset{\sim}{I}-\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}']\underset{\sim}{Y}$$

is distributed as $\frac{\sigma^2}{(N-p)}\chi^2_{(N-p)}$, and s^2 and $\hat{\underset{\sim}{\beta}}_2'\underset{\sim}{W}_0\hat{\underset{\sim}{\beta}}_2$ are independent. Since one can readily show that

$$E(1/s^2) = \left(\frac{N-p}{N-p-2}\right)\frac{1}{\sigma^2},$$

it follows that

$$\hat{\gamma} = \frac{(N-p-2)\hat{\underset{\sim}{\beta}}_2'\underset{\sim}{W}_0\underset{\sim}{\beta}_2}{(N-p)s^2} - \tau$$

is an unbiased estimator of γ.

3.3.4. A Confidence Interval for γ.

When one has made a prior assumption, it is prudent to make an a posteriori check of the assumption. Since the use of the SK estimator requires the specification of an upper bound on η^2/σ^2, it is wise to check this. Accordingly, a confidence level on γ has been derived. This derivation is based on the assumption that the deviations from the true model are normally distributed.

Let $\underset{\sim}{S}^{-1} = LL'$ where $\underset{\sim}{L}$ is lower trianglar and $\underset{\sim}{z} = \underset{\sim}{L}^{-1}\underset{\sim}{\beta}_2$. If λ_{max} is the largest eigenvalue of $\underset{\sim}{L}'\underset{\sim}{W}_0\underset{\sim}{L}$ then by a result in Rao [1965],

$$\frac{\underset{\sim}{\beta}_2'\underset{\sim}{W}_0\underset{\sim}{\beta}_2}{\underset{\sim}{\beta}_2'\underset{\sim}{S}\underset{\sim}{\beta}_2} = \frac{\underset{\sim}{z}'\underset{\sim}{L}'\underset{\sim}{W}\ \underset{\sim}{L}\ \underset{\sim}{z}}{\underset{\sim}{z}'\underset{\sim}{L}'\underset{\sim}{L}'^{-1}\underset{\sim}{L}^{-1}\underset{\sim}{L}\ \underset{\sim}{z}} = \frac{\underset{\sim}{z}'\underset{\sim}{L}'\underset{\sim}{W}_0\underset{\sim}{L}\ \underset{\sim}{z}}{\underset{\sim}{z}'\underset{\sim}{z}} \le \lambda_{max} .$$

So

$$\gamma \le \lambda_{max} \frac{\underset{\sim}{\beta}_2'\ \underset{\sim}{S}\ \underset{\sim}{\beta}_2}{\sigma^2} .$$

Now

$$F = \frac{\hat{\underset{\sim}{\beta}}_2'\ \underset{\sim}{M}\ \hat{\underset{\sim}{\beta}}_2}{p_2\ s^2} \sim F'_{p_2,\ N-p}\left(\frac{\underset{\sim}{\beta}_2'\underset{\sim}{S}\ \underset{\sim}{\beta}_2}{\sigma^2}\right)$$

where $F'_{\nu_1,\nu_2}(\lambda)$ denotes the noncentral F distribution with ν_1 and ν_2 degrees of freedom and noncentrality parameter λ. In this case $p_2 = \text{rank}\ \underset{\sim}{M}$ and $\lambda = \underset{\sim}{\beta}_2'\underset{\sim}{S}\ \underset{\sim}{\beta}_2/\sigma^2$. Let $F'_{p_2,N-p,\alpha}(\lambda)$ denote the αth percentile of this distribution. Since $\Pr[F'_{p_2,N-p,\alpha}(\lambda) < F]$ is a decreasing function of λ, the value of λ, say $\hat{\lambda}$, which satisfies

$$F'_{p_2,N-p,\alpha}(\hat{\lambda}) = F$$

will be the upper bound for all values of λ for which

$$F_{p_2,N-p,\alpha}(\lambda) \le F.$$

Thus

$$\Pr\left[\frac{\underset{\sim}{\beta}_2 \; \underset{\sim}{S} \underset{\sim}{\beta}_2}{\sigma^2} < \hat{\lambda}\right] \ge 1 - \alpha. \qquad (3.3.4.1)$$

Since $\gamma \le \lambda_{max} \dfrac{\underset{\sim}{\beta}_2' \underset{\sim}{S} \; \underset{\sim}{\beta}_2}{\sigma^2}$, we have that

$$\frac{\gamma}{\lambda_{max}} \le \frac{\underset{\sim}{\beta}_2' \; \underset{\sim}{S} \underset{\sim}{\beta}_2}{\sigma^2} \quad . \qquad (3.3.4.2)$$

Substituting (3.3.4.2) into (3.3.4.1), it follows that

$$\Pr[\gamma < \lambda_{max} \hat{\lambda}] \ge 1 - \alpha.$$

Thus, one would expect $\gamma < \lambda_{max} \hat{\lambda}$ with confidence $(1-\alpha) \times 100$ percent. As long as U was greater than $\lambda_{max} \hat{\lambda}$, one could be confident that U had not been chosen too small. If U was smaller than $\lambda_{max} \hat{\lambda}$ or $\hat{\gamma}$, then it would appear that the data did not substantiate the small choice of U. Varying U based on $\lambda_{max} \hat{\lambda}$ would make δ stochastic. The IMSE of such a resulting estimator has not been investigated. If U had been carefully chosen, then one should not place a great deal of weight on this confidence interval.

3.3.5 An Application of the Ridge Estimator to Smoothing and Prediction Situations.

Earlier we considered estimators when the least squares estimator

$\hat{\beta}$ was for some reason not an acceptable alternative estimator. In some situations this restriction may be made without thorough consideration. For example, a straight line fit to data may be considered smoother than a quadratic fit to these same data. The integrated variance is always smaller when the straight line is fit, and one may feel that he is achieving smaller IMSE by fitting a straight line. Nothing we have discussed so far contradicts this conception if the true quadratic coefficient is small. However, the ridge estimator offers a way to modify all the elements of $\hat{\beta}$ for the purpose of achieving smaller IMSE. This method may be preferable in those situations when one's goal is to achieve an estimator with smaller IMSE than $\hat{\beta}$. When σ^2 is large and the columns of $\underset{\sim}{X}$ are highly correlated, the IMSE of $\hat{\beta}$ can be very large. In such situations there is not always a great deal of prior knowledge available on any of the elements of β. Therefore, the elimination of variables in order to reduce the integrated variance can be quite subjective and can appear arbitrary.

We shall develop an estimator for use in such a situation. Consider an arbitrary estimator $\underset{\sim}{b}$ with IMSE

$$J(\underset{\sim}{W}, \underset{\sim}{b}) = V_{\underset{\sim}{b}} + B_{\underset{\sim}{b}} \quad .$$

The IMSE of $\underset{\sim}{b}$ can be expressed as

$$J(\underset{\sim}{W}, \underset{\sim}{b}) = V_{\underset{\sim}{b}} + \frac{B_{\underset{\sim}{b}}}{B_{\underset{\sim}{0}}} B_{\underset{\sim}{0}}$$

where $B_{\underset{\sim}{0}}$ is the bias error of the zero estimator. Now, if a priori we know a number $\vartheta_{\underset{\sim}{b}}$ such that

$$\vartheta_{\underset{\sim}{b}} \geq \frac{B_{\underset{\sim}{b}}}{B_{\underset{\sim}{0}}}$$

then

$$J(\underset{\sim}{W}, \underset{\sim}{b}) \leq V_{\underset{\sim}{b}} + \vartheta_{\underset{\sim}{b}} B_{\underset{\sim}{0}} .$$

We may prefer $\underset{\sim}{b}_1$ to $\underset{\sim}{b}_2$ if

$$V_{\underset{\sim}{b}_1} \leq V_{\underset{\sim}{b}_2} \tag{3.3.5.1}$$

when

$$\vartheta_{\underset{\sim}{b}_1} \leq \vartheta_{\underset{\sim}{b}_2} . \tag{3.3.5.2}$$

When (3.3.5.1) and (3.3.5.2) held we would know that the bound on the IMSE of $\underset{\sim}{b}_1$ was less than the bound on the IMSE of $\underset{\sim}{b}_2$. There may be parameter coefficients $\underset{\sim}{\beta}$ such that even when (3.3.5.1) and (3.3.5.2) hold we would have

$$J(\underset{\sim}{W}, \underset{\sim}{b}_2) < J(\underset{\sim}{W}, \underset{\sim}{b}_1) .$$

We shall temporarily ignore this possibility, and choose the estimator $\underset{\sim}{b}$ that attains minimum integrated variance given that the bound on the ratio $\vartheta_{\underset{\sim}{b}}$ is fixed.

The integrated bias of the zero estimator is

$$B_{\underset{\sim}{0}} = \underset{\sim}{\beta}' \underset{\sim}{W} \underset{\sim}{\beta} \frac{N}{\sigma^2} .$$

The ratio of this integrated bias to the bias error of an estimator of the form $\hat{\underset{\sim}{b}} = \underset{\sim}{K}\underset{\sim}{\beta}$ is

$$\Omega \int_R \frac{(\underset{\sim}{x}'(\underset{\sim}{K}-\underset{\sim}{I})\underset{\sim}{\beta})^2}{\underset{\sim}{\beta}' \underset{\sim}{W} \underset{\sim}{\beta}} d\underset{\sim}{x}$$

and by an application of the Cauchy Schwarz inequality this ratio is bounded by

$$\mathrm{tr}(\underset{\sim}{K}-\underset{\sim}{I})\underset{\sim}{W}^{-1}(\underset{\sim}{K}-\underset{\sim}{I})'\underset{\sim}{W} \quad .$$

We select an estimator of the form $\underset{\sim}{b} = \underset{\sim}{K}\hat{\underset{\sim}{\beta}}$ where this bound is fixed (at an a priori value ϑ). As a Lagrangian problem, we want to minimize

$$F = N \ \mathrm{tr} \ \underset{\sim}{K}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{K}'\underset{\sim}{W} + \frac{1}{k}(\mathrm{tr}(\underset{\sim}{K}-\underset{\sim}{I})\underset{\sim}{W}^{-1}(\underset{\sim}{K}-\underset{\sim}{I})\underset{\sim}{W}-\vartheta) \quad .$$

The derivative is

$$\frac{\partial F}{\partial \underset{\sim}{K}} = NW(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{K}' + \frac{1}{k}(\underset{\sim}{K}-\underset{\sim}{I})' \qquad (3.3.5.3)$$

where k is chosen so that

$$\vartheta = \mathrm{tr}(\underset{\sim}{K}-\underset{\sim}{I})\underset{\sim}{W}^{-1}(\underset{\sim}{K}-\underset{\sim}{I})'\underset{\sim}{W}$$

is true. Setting (3.3.5.3) to zero and solving for $\underset{\sim}{K}$, we have that the solution matrix is

$$\underset{\sim}{K}^{+} = (N\underset{\sim}{W}k + (\underset{\sim}{X}'\underset{\sim}{X}))^{-1} \ \underset{\sim}{X}'\underset{\sim}{X} \ .$$

This completes the derivation of this estimator. We shall now make some comments about this estimator.

First, this estimator is of the general form of a ridge estimator. For a general ridge estimator the transform of the least squares coefficients is

$$(\underset{\sim}{K} + \underset{\sim}{X}'\underset{\sim}{X})^{-1} \ \underset{\sim}{X}'\underset{\sim}{X} \ .$$

Thus, $\underset{\sim}{b}^{+} = \underset{\sim}{K}^{+}\hat{\underset{\sim}{\beta}}$ is a generalized ridge estimator.

Second, we shall consider how ϑ may be chosen. Say that a priori we know that $\eta^2/\sigma^2 < U$ on R. Then

$$J(\underset{\sim}{W}, \underset{\sim}{b}^+) < J(\underset{\sim}{W}, \hat{\underset{\sim}{\beta}})$$

when

$$V_{\underset{\sim}{b}^+} + \vartheta_{\underset{\sim}{b}^+} U < V_{\hat{\underset{\sim}{\beta}}} \ .$$

Solving for $\vartheta_{\underset{\sim}{b}^+}$ we have that the IMSE of $\underset{\sim}{b}^+$ is smaller than the IMSE of $\hat{\underset{\sim}{\beta}}$ when

$$\vartheta_{\underset{\sim}{b}^+} \leq \frac{V_{\hat{\underset{\sim}{\beta}}} - V_{\underset{\sim}{b}^+}}{U} \ . \qquad (3.3.5.4)$$

Therefore, when $U > \eta^2/\sigma^2$ on R is true, any choice of k so that (3.3.5.4) holds will result in an estimator with smaller IMSE than the IMSE attained by $\hat{\underset{\sim}{\beta}}$. Obviously, when one is very conservative and will admit to only a large value for U then one might as well use $\hat{\underset{\sim}{\beta}}$. We say this since equation (3.3.5.4) indicates that only small choices of $\vartheta_{\underset{\sim}{b}^+}$ will be available. However, we see that equation (3.3.5.4) gives us a method to assess if U is too large.

3.4 Identification of Explanatory Variables

We shall only discuss the ridge estimator in this section. The problem of how to best select k has received considerable attention. A number of methods for selecting k have been reviewed by R. L. Obenchain (1975). He lists 6 methods of selecting k besides the use of $\hat{\underset{\sim}{\beta}}$.

We would like to list a seventh method. That is, choose k so that the residual sum of squares

$$\phi(\hat{\beta}_k) = (\underset{\sim}{Y} - \underset{\sim}{X}\hat{\beta}_k)'(\underset{\sim}{Y} - \underset{\sim}{X}\hat{\beta}_k)$$

satisfies

$$\frac{\phi(\hat{\beta}_k)}{N} = \frac{\phi(\hat{\beta})}{N-p} = \hat{\sigma}^2 \quad .$$

The IMSE of such an estimator is difficult to evaluate, since it is a stochastically shrunken estimator. Also, such an estimator is apparently not in S, the class of estimators that satisfy the Stein restriction.

However, choosing k in this way does result in an estimator that has some nice properties.

- If there is no error then k is zero.
- It has the same RSS that would be expected to result if β were substituted for $\hat{\beta}_K$.

Using the 10-Factor example from Hoerl-Kennard (Page 71) we find that a k of about .15 satisfies

$$\frac{\phi(\hat{\beta}_k)}{36} = \frac{\phi(\hat{\beta})}{26} \quad .$$

Since the other nonzero k's were .007, .02, .039, .11 (.2, .3) and .54; we find that .15 is near the median.

Applying this technique to Hoerl and Kennard's second example, we obtain a k of about .08. Hoerl and Kennard suggest using a k between .2 and .4.

Many other characteristics of this method of choosing k can be studied. However, as an objective method, it appears reasonable and gives similar results to other methods. We suggest that its simplicity is a main reason to use it.

Chapter IV

APPLICATIONS AND FURTHER COMPARISONS

4.1 An Application of Biased Estimation Theory to an Iterative Maximum Likelihood Search

4.1.1. Background

Parametric dose-response models have great utility in quantifying the relationship between an agent and a health effect of that agent. Knowledge of such relationships is key information for the determination of air quality standards.

The parameters of such models are estimated from observed data. The maximum likelihood estimator (MLE) of these parameters have good statistical attributes, and so the MLE is often calculated. Even for simple mathematical models, this calculation usually involves iterative methods. Such iterative schemes do not always successfully converge to the MLE.

We have attempted to determine the MLE for the cumulative logistic function adjusted for natural responsiveness. The data were obtained in an epidemiological panel study. Generally, there are no control groups for such studies and existing iterative methods can perform poorly under such circumstances. Difficulty in using one such method was experienced in this case.

The success of the iterative method that we used depended heavily on our ability to start the iteration with good starting estimates of the parameters (starting values). If one set of starting values doesn't result in convergence, one tries another set. Complex subjective judgments may be required to obtain good starting values. Since starting with the MLE as a set of starting values will usually result in convergence one can always be accused of using poor starting values when the iterative scheme fails.

There are a couple of ways that an iterative scheme can fail. When very poor starting values are used, convergence to a relative maximum may take place, or the specified number of iterations may be used up before convergence to the MLE has occurred. Another type of failure results when the revised estimates overshoot the MLE and get progressively worse. This type of failure occurs when a certain matrix requiring inversion is ill-conditioned. It is this type of failure that was experienced in our case, and it is this type of problem that our modification is designed to handle.

This type of failure can be viewed as the fault of unbiased estimation. At each iteration the revised estimates are derived from a solution vector. This solution vector can be considered the least squares estimator of the parameters of a particular linear model. Although the least squares estimator attains minimum variance among unbiased estimators, this variance can be large.

The theory of biased estimation is based on the fact that mean squared error (bias plus variance) can be smaller for a slightly biased estimator with much smaller variance, than for an unbiased estimator. This theory is applied to develop a class of estimators appropriate for this problem. The result is an iterative maximum likelihood search strategy in which the user of the method simply specifies a bound on the difference between the MLE and the starting value for the proportion of natural responses. The method using the least squares estimator can be viewed as considering this difference to be unbounded. Thus, the user can specify a very crude bound and expect to do better by incorporating this bound into the modified method.

First, we present the theory involved. This includes a description of the dose-response model, a description of the iterative method, the development of a modification to this iterative method, and a discussion on the subjective input required to use the modification.

Second, we do an empirical evaluation of this modification. The data used in the evaluation are described. Next, we describe the evaluation, list the results, and draw conclusions.

4.1.2. Theoretical Development

4.1.2.1. Description of the Cumulative Logistic Adjusted for Natural Responsiveness Dose-Response Model.

Consider a dose-response experiment where dose levels Z_i at N different levels are applied to n_i subjects and at each dose level r_i subjects respond.

If the administration of a dose Z causes a proportion $P(Z)$ of the test subjects to respond and other independent factors acting on the test subjects during the experiment cause a proportion C to respond, then the total expected proportion responding will be

$$P'(Z) = P(Z) + C - P(Z)C = C + (1-C)\,P(Z). \qquad (4.1.2.1)$$

This equation is called Abbott's Formula. If $P(Z)$ is the cumulative logistic function

$$P(Z) = \frac{1}{1 + \exp[-\,(A+BZ)]},$$

then $P'(Z)$ is known as the cumulative logistic adjusted for natural responsiveness. This model is motivated by the concept of a tolerance.

An individual's tolerance is defined as that level of dose Z_0 such that doses higher than Z_0 always cause a response. The purpose of the model is to make inferences about the distribution of tolerances in a target population. In fitting such models, it is usual to assume that the n_i subjects exposed to a dose level Z_i were randomly selected from the target population. Assume in addition that the proportion of tolerances less than Z is $P(Z)$ for each Z. When these two assumptions are correct the probability of observing r_i responses from the n_i subjects dosed at level Z_i for $i = 1,2,\ldots,N$ is

$$L(A,B,C) = \prod_{i=1}^{N} \binom{n_i}{r_i} P'(Z_i)^{r_i} (1-P'(Z_i))^{n_i-r_i} .$$

In general, $L(A,B,C)$ is called the likelihood function. Those choices of A, B and C (denoted by $\hat{A}$, $\hat{B}$, and $\hat{C}$) that maximize $L(A,B,C)$ are the MLE.

4.1.2.2. <u>Description of an existing iterative method</u>.

Based on an approximation to the first derivative of the log likelihood evaluated at the MLE by the first order terms of the Taylor-Maclaurin expansion, Finney (1971) has derived the expressions used in iteratively solving for the MLE of the parameters of the probit dose-response model adjusted for natural responsiveness. These expressions can be easily applied to the logistic model described in (4.1.2.1). When A_s, B_s, and C_s are the starting values for the parameters A, B, and C respectively, at the iteration number S, the revised estimates $(A_{s+1}, B_{s+1}, C_{s+1})$ are

$$\begin{bmatrix} A_{s+1} \\ B_{s+1} \\ C_{s+1} \end{bmatrix} = \begin{bmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ C_s \end{bmatrix}$$

The formulae for $\hat{\alpha}_0$, $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are algebraically the same as the formulae for the least squares estimator of the parameters of the linear model

$$y_i = \alpha_0 x_{0_i} + \alpha_1 x_{1_i} + \alpha_2 x_{2_i} + \varepsilon_i, \qquad i = 1,2,\ldots,N, \qquad (4.1.2.2.1)$$

where the ε_i are uncorrelated random variables with zero mean and common variance σ^2. The x_{0i}, x_{1i}, and x_{2i} are all determined from A_s, B_s and C_s, and the formula for the y_i also involves r_i. The formulae below define the x_{ji} and y_i for $i = 1,2,\ldots,N$ and $j = 0,1$, and 2.

$$\hat{P}_i = \frac{\frac{r_i}{n_i} - C_s}{1 - C_s};$$

$$P_i = \frac{1}{1 + \exp - (A_s + B_s Z_i)};$$

$$Q_i = 1 - P_i;$$

$$W_i = P_i Q_i;$$

$$w_i = \frac{W_i^2}{Q_i \left(P_i + \frac{C_s}{1-C_s} \right)};$$

$$y_i = (A_s + B_s Z_i + \frac{\hat{P}_i - P_i}{W_i})(n_i w_i)^{\frac{1}{2}} ;$$

$$x_{0i} = (n_i w_i)^{\frac{1}{2}};$$

$$x_{1i} = Z_i (n_i w_i)^{\frac{1}{2}};$$

and

$$x_{2i} = \frac{Q_i}{W_i} (n_i w_i)^{\frac{1}{2}} .$$

If C_1 is obtained from inspection of the data, A_1 and B_1 can be obtained by iteration to maximize the likelihood for this value of C_1. The appropriate expressions involved in this first stage of iteration are algebraically equivalent to the least squares estimators

$$\begin{bmatrix} \tilde{\alpha}_0 \\ \tilde{\alpha}_1 \end{bmatrix} = \begin{bmatrix} A_1 \\ B_1 \end{bmatrix}$$

for the parameters of the model

$$y_i = \alpha_0 x_{0i} + \alpha_1 x_{1i} + \varepsilon_i, \qquad i = 1,2,\ldots,N \qquad (4.1.2.2.2)$$

which is obtained from (4.1.2.2.1) by setting α_2 to zero. Of course, starting values for A and B are required even for this first stage of iteration, and for B_0 one could use zero and set

$$A_0 = \ln \frac{\overline{P} - C_1}{1 - \overline{P}} ,$$

where

$$\overline{P} = \frac{\sum r_i}{\sum n_i} \quad .$$

This choice of starting value for A_0 maximizes the likelihood function for these values of B_0 and C_1. Having obtained revised estimates for A and B after a number of iterations using this first stage, one could proceed to the second stage of iteration using (4.1.2.2.1) and these revised estimates as starting values.

Based on empirical results, it appears that the success of such a two stage scheme depends on how well C_1 is chosen and on what criterion is used to determine when to stop first stage iterations and transition to the second stage. Very stringent criteria would be wasteful of computer time if C_1 is poorly chosen, and its use does not guarantee success when C_1 is chosen close to $\hat{C}$. When a very loose criterion is employed, e.g. omitting the first stage entirely, failure rates are high. The failures are usually the result of high correlations among the x_{ji} resulting in a singularity in the matrix requiring inversion. Invariably, such a singularity is preceeded by an unreasonably large overestimation of the magnitude of α_2.

When we regard (4.1.2.2.1) as the true model whose parameters are to be estimated, the problems mentioned in the last paragraph are a familiar weakness of least squares estimation. That is, even though the least squares estimator achieves minimum variance among unbiased estimators, this variance can be very large when the independent variables are highly correlated. In those cases cited above, we view this variance as being intolerably large. Possibly a slightly biased estimator with much smaller variance can achieve small enough mean squared error (bias plus variance) to usually avoid that type of problems.

4.1.2.3 Development of an Alternative Estimator via the Consideration of Biased Estimation Theory.

Consider the class of biased estimators of the form

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \tilde{\alpha}_0 \\ \tilde{\alpha}_1 \\ 0 \end{bmatrix} + \begin{bmatrix} k_0 \\ k_1 \\ k_2 \end{bmatrix} \hat{\alpha}_2 , \qquad (4.1.2.3.1)$$

where k_0, k_1 and k_2 are constants. We shall consider using the a_i in place of the $\hat{\alpha}_i$ in a single phase iteration scheme. Note that if the k_i are zero then the a_i are the least squares estimators of the model described in (4.1.2.2.2), which are biased when α_2 is nonzero. It is also possible to obtain the $\hat{\alpha}_i$ from (4.1.2.3.1) by letting the k_i take on certain values.

We shall use an error criterion that seems appropriate for this problem and show how the k_i can be chosen so that for a wide range of α_2 values, better performance in terms of this criterion is attained by an estimator using the chosen k_i's than is attained by using either of the estimators mentioned in the above paragraph. The choice of the k_i is made by specifying an upper bound on α_2^2/σ^2.

It turns out that the formula for the k_i that we suggest using involves the specified upper bound and the correlation between the x_{ji}. Generally, this correlation changes from one iteration to the next, so that even though only one bound is used, the k_i vary between iterations. In those cases where successful convergence takes place, the estimators are similar to those that were suggested for the two stage scheme. Initially, the x_{ji} are highly correlated and the k_i are zero. In later iterations, the correlations become smaller and the a_i's approach the $\hat{\alpha}_i$'s. Thus, instead of selecting a criterion for deciding when to jump from the first to second stage of iterations, the use of the proposed estimator eliminates the need to make this decision by automatically incorporating it into the computation process based on the value of the specified upper bound.

Let the true model at each iteration be

$$\eta = a_0x_0 + \alpha_1x_1 + \alpha_2x_2,$$

and consider approximating η over an interval of interest $\xi_0 \le x \le \xi_1$ with

$$\hat{\eta} = a_0x_0 + a_1x_1 + a_2x_2$$

where (a_0, a_1, a_2) is a vector of estimated regression coefficients. A reasonable measure of the closeness of $\hat{\eta}$ to η is integrated mean square error

$$J = \frac{N\Omega}{\sigma^2} \int_{\xi_0}^{\xi_1} E(\hat{\eta}-\eta)^2 dx$$

where

$$\Omega^{-1} = \int_{\xi_0}^{\xi_1} dx.$$

This is the situation where one may consider using the estimator discussed in Section 3.2.1. Using that estimator suggests that we set

$$\underset{\sim}{k} = \begin{bmatrix} k_0 \\ k_1 \\ k_2 \end{bmatrix} = \delta \begin{bmatrix} -\hat{\mu}_0 \\ -\hat{\mu}_1 \\ 1 \end{bmatrix}$$

where $\hat{\mu}_0$ and $\hat{\mu}_1$ are the least square estimators of μ_0 and μ_1 (respectively) for the model

$$E(x_{2i}) = \mu_0 x_{0i} + \mu_1 x_{1i} \quad . \tag{4.1.2.3.2}$$

The value of δ suggested by (3.2.1.3) is in this case

$$\delta^* = \max(0,\ (Us^2-1)/(Us^2+1)),$$

where s^2 is the sum of squared deviations from fitting (4.1.2.3.2). Finally, U is our upper bound on α_2^2/σ^2. Now, let us consider how to choose U.

One way to choose a value U that bounds α_2^2/σ^2 is to find an upper bound for α_2^2 and a lower bound for σ^2. Therefore, first consider α_2 as being $C-C_s$, since $C_{s+1} = \hat{\alpha}_2+C_s$. Partial justification for considering $\alpha_2 = C-C_s$ can be based on a result from large sample theory: "... if first approximations are of nonzero efficiency, one cycle of computation will yield fully efficient estimates" Finney (1971). By (4.1.2.1), C is a proportion, so that if we agree to set C_1 to some value between zero and one then $(C-C_1)^2 = \alpha_2^2$ will be less than one.

Considering only the variability contributed to y_i by $\hat{P}_i$. we have

$$\sigma^2 = \text{Var } y_i = \frac{(1-C)(C+(1-C)P_i(1-P_i))}{(1-C_s)(C_s+(1-C_s)P_i(1-P_i))} \; .$$

When we assume that the starting values are equal to the respective true parameters of the model, we have that σ^2 is one. For these reasons we suggest using $U = 1$.

Smaller values for U such as $\hat{P}_1^2$ might be used, but we have only investigated the case where $\underset{\sim}{k}$ is deterministic. The use of any information gained from the data in fixing $\underset{\sim}{k}$ would require that $\underset{\sim}{k}$ be treated as random.

In practice, the error may be too large even when U is carefully chosen. So if the iteration routine fails one should consider varying U. However, it would seem unnecessary to choose a value of U larger than one.

4.1.3. Empirical Evaluation of Modification

Hammer et al (1974) reported on data from student nurses in Los Angeles who completed daily symptom diaries during the period of their training. The total numbers of yes/no responses of four symptom categories pooled over days having the same range of maximum hourly oxidant levels are listed in Table 4.1.1 along with the midpoint of these oxidant ranges. The symptom categories are Headache, Eye Discomfort, Cough and Chest Discomfort with no accompanying fever, chills or temperature. Due to the restrictions, the positive responses are indicators of only mild personal discomfort.

Assuming independence between days of responses from the same student nurse, the number of positive responses are taken to be binomially distributed with parameters P_i' and n_i where

$$P'_{k_i} = C + (1-C)/(1+\exp[-(A+BZ_i)]).$$

The MLE's for this model are listed in Table 4.1.2 for each symptom.

Iterations were run using three different starting values for C and four different bounds on α_2^2/σ^2. The starting values for C were:

(a) The observed proportion of total responses corresponding to the lowest oxidant level. For all four symptom categories this proportion turned out to differ from the MLE for C by less than 0.01. The minimum observed proportion was used as a starting value for C in the case of Chest Discomfort in order to avoid the computation of the logarithm of a negative number in the process of getting a starting value for A.

(b) One half the proportion used in (a).

(c) Zero.

The starting value for B was zero and the starting value for A was

$$A_s = \ln(\bar{P}-C_s)/(1-\bar{P})$$

where

$$\bar{P} = (\sum r_i)/\sum n_i \ .$$

The four different values for U used were $\hat{P}_1^2$, $\hat{P}_9^2$, 1, and ∞. Infinity corresponds to using the least squares estimator for (4.1.2.2.1).

Iterations were continued until the change in the likelihood was zero (to the accuracy of the computer), or until 30 iterations, or until the program faulted. The maximum number of iterations used on any one run was 25.

Table 4.1.3 lists the occurrences of successful convergence. From this table it can be seen that the percent of successful convergences was higher in all cases when a finite bound on σ_2^2/σ^2 was used. When $\hat{P}_1^2$ was specified as a bound, the failure was due to convergence to the MLE of A and B for the starting value C_1. The highest percent of successful convergences when a starting value for C of either 0 or $\frac{1}{2}\hat{P}_1$ was used occurred when U was set equal to one.

4.1.4. Conclusions

If good starting values are used, convergence can always take place using Finney's formula. However, how good these starting values must be depends on the data.

The modification of Finney's formula developed here yields successful convergence for starting values that are too poor to be used directly in Finney's formula. When very poor starting values for C are used neither the modification nor Finney's formula works all the time. However, for some values of U, a higher success rate can be obtained using the modification. Moreover, when seemingly adequate C starting values are used, successful convergence is obtained using the modification in cases where the use of Finney's formula had failed.

The use of smaller U values seems to yield poorer success rates for poor choices of C starting values, but can achieve improved success rates for seemingly adequate C starting values.

It is recommended that a number of C starting values be used with this modification. When no other information is available, U should be taken to be one. If convergence fails due to a singularity, a tighter bound should be tried. A plot of the likelihood obtained at the final iteration versus the corresponding C value is helpful for deciding if more starting values should be tried and what values to use.

Table 4.1.1

Los Angeles Student Nurse Data

Daily Maximum Hourly Oxidant Level, pphm		Total Exposed Nurses x Days	Headache		Eye Discomfort		Cough		Chest Discomfort	
			Yes	Proportion of Yeses	Yes	Proportion of Yeses	Yes	Proportion of Yeses	Yes	Proportion of Yeses
i	Z_i	n_i	r_i	$\hat{P}_i$	r_i	$\hat{P}_i$	r_i	$\hat{P}_i$	r_i	$\hat{P}_i$
1	3	14656	1539	.105	733	.050	1334	.091	264	.018
2	7	10856	1162	.107	586	.054	1075	.099	195	.018
3	9	2030	215	.106	114	.056	207	.102	39	.019
4	12	10912	1200	.110	644	.059	1026	.094	196	.018
5	17	8352	952	.114	576	.069	810	.097	142	.017
6	22	3780	438	.116	348	.092	344	.091	60	.016
7	27	1500	173	.115	168	.112	144	.096	30	.020
8	35	603	81	.134	107	.177	71	.118	14	.023
9	45	159	24	.151	51	.321	27	.170	9	.057

Table 4.1.2

Maximum Likelihood Estimates for Each Symptom

Symptom	A	B	C
Headache	- 4.626920811	.04070218878	.09535299228
Eye Discomfort	- 5.045747967	.09309640264	.04166289253
Cough	- 9.949532964	.1690304341	.09442183274
Chest Discomfort	-13.26019686	.2243214541	.01772926043

Table 4.1.3

Successful (S) and Failing (F) Iteration Attempts

U	Symptom	Starting Value for C		
		0	$\frac{1}{2}P_1$	P_1
P_1^2	Eye Discomfort	S	S	S
	Headache	F	F	S
	Cough	F	F	S
	Chest Discomfort	F	F	S
	% Success	25	25	100
P_9^2	Eye Discomfort	S	S	F
	Headache	F	F	S
	Cough	F	F	S
	Chest Discomfort	F	F	F*
	% Success	25	25	50
1	Eye Discomfort	S	S	F
	Headache	S	S	S
	Cough	F	F*	S
	Chest Discomfort	F	F*	F*
	% Success	50	50	50
∞	Eye Discomfort	F*	F*	F*
	Headache	F*	F*	F*
	Cough	F*	F*	S
	Chest Discomfort	F*	F*	F*
	% Success	0	0	25

*Indicates failure occurred due to program default.

4.2. A COMPARISON OF SOME ESTIMATORS OF THE FORM $\underset{\sim}{K}\hat{\underset{\sim}{\beta}}$

4.2.1. Introduction

In this section we will compare some estimators that are of the form $\underset{\sim}{K}\hat{\underset{\sim}{\beta}}$. The comparisons will be done in such a way that some light will be shed on answers to such questions as:

•For poor designs (highly correlated independent variables), is the BD estimator preferable to the KMH estimator? Does the SK estimator achieve sizeable reduction in IMSE as compared to the KMH estimator when the scalar δ is chosen conservatively?

•If one's goal is the reduction in IMSE, are there full model estimators (i.e., estimators of the form $\underset{\sim}{K}\hat{\underset{\sim}{\beta}}$ where $\underset{\sim}{K}$ is nonsingular) that are competitive with reduced model estimators in achieving small IMSE?

•How does $\underset{\sim}{b}^+$, the estimator developed in Section 3.2.4, compare to the KMH and BD estimators? Since $\underset{\sim}{b}^+$ is a generalized ridge estimator, the results of such comparisons are indicative of comparisons between the ridge estimator and these two other estimators.

In the general setting of linear models, the answers to these questions are in terms of complex matrix expressions whose interpretations often depend on the values of unknown coefficients and on the coordinates of the design points. In general, these coefficients and these coordinates are free to vary in multidimensional space. For these reasons we shall restrict the scope of our comparisons. We shall consider a particular model, a simple class of designs, and a certain criterion.

The model is discussed in Section 4.2.2. The class of designs is discussed in Section 4.2.3. The criterion and the motivation for considering this criterion is discussed in Section 4.2.4. An explanation of the comparison is in 4.2.5. The results of the comparison are described in Section 4.2.6, and our interpretations of these results are given in Section 4.2.7.

4.2.2. The Model

We consider the situation where the true model is not known. We are interested in estimating the response surface over a particular region of interest. We want to use a model that is general enough to detect important irregularities in the response surface, and yet we don't want a model that is so complex that it has little chance of verification in subsequent experiments. In many situations, the most complex model that can be used is the second degree polynomial. We shall assume that the second degree polynomial in two variables is the true model. Specifically, we assume that the true model is of the form

$$\eta = \beta_0+\beta_1x_1+\beta_2x_2+\beta_{11}x_1^2+\beta_{22}x_2^2+\beta_{12}x_1x_2.$$

We assume that the β's are unknown coefficients. Later, we will assume that on a priori bound on η^2/σ^2 can be specified. Such an assumption implies some knowledge of the β's. However, no additional specific knowledge of the β's will be assumed.

4.2.3. The Class of Designs

We take the region of interest to be the square bounded by $-1 \le x_1 \le 1$ and $-1 \le x_2 \le 1$. The designs we shall consider will be generated by the scalar ζ. These designs are nine point designs. The coordinates of these nine points are

$$
\begin{aligned}
(x_1, x_2) = \; & (-\zeta,\ \zeta) \\
& (\tfrac{1}{2} - \tfrac{\zeta}{2},\ \tfrac{1}{2} + \tfrac{\zeta}{2}), \\
& (1,\ 1) \\
& (-\tfrac{1}{2} - \tfrac{\zeta}{2},\ -\tfrac{1}{2} + \tfrac{\zeta}{2}), \\
& (0,\ 0), \\
& (\tfrac{1}{2} + \tfrac{\zeta}{2},\ \tfrac{1}{2} - \tfrac{\zeta}{2}), \\
& (-1,\ -1), \\
& (-\tfrac{1}{2} + \tfrac{\zeta}{2},\ -\tfrac{1}{2} - \tfrac{\zeta}{2}), \\
& (\zeta,\ -\zeta).
\end{aligned}
$$

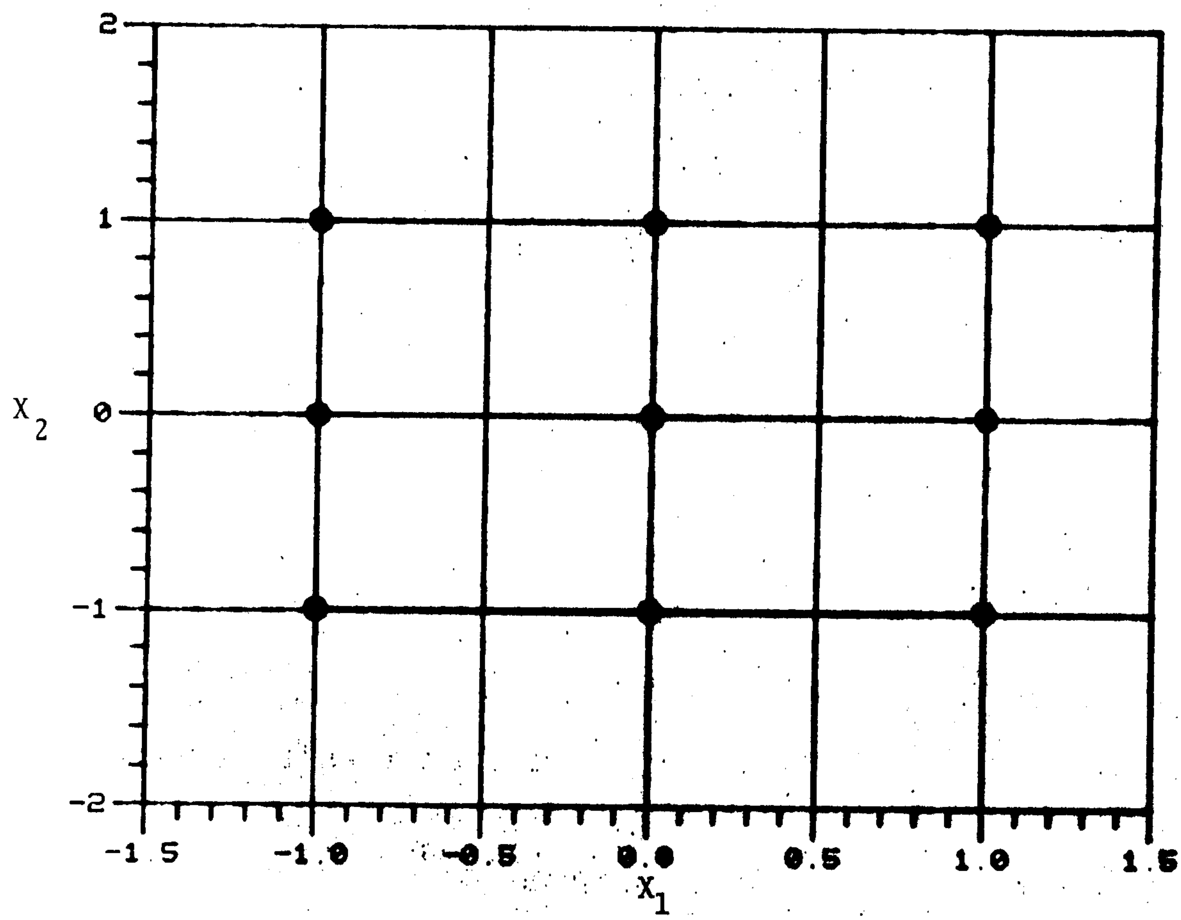
Figures 4.2.1 and 4.2.2 show the plots of these nine points when ζ is 1 and when ζ is 1/2, respectively.

The correlation coefficient of x_1 and x_2 is $(1-\zeta^2)/(1+\zeta^2)$. The mean of the design points for x_1 and x_2 is always zero irrespective of what value ζ assumes. When ζ is one, we have a completely orthogonal design. As ζ approaches 0, x_1 and x_2 become increasingly correlated until all points lie on the line $x_1 = x_2$ when ζ is zero. We will consider five designs in this class specified by $\zeta = 1, .75, .5, .25$ and $.125$.

For this model and this region of interest, the moment matrix is

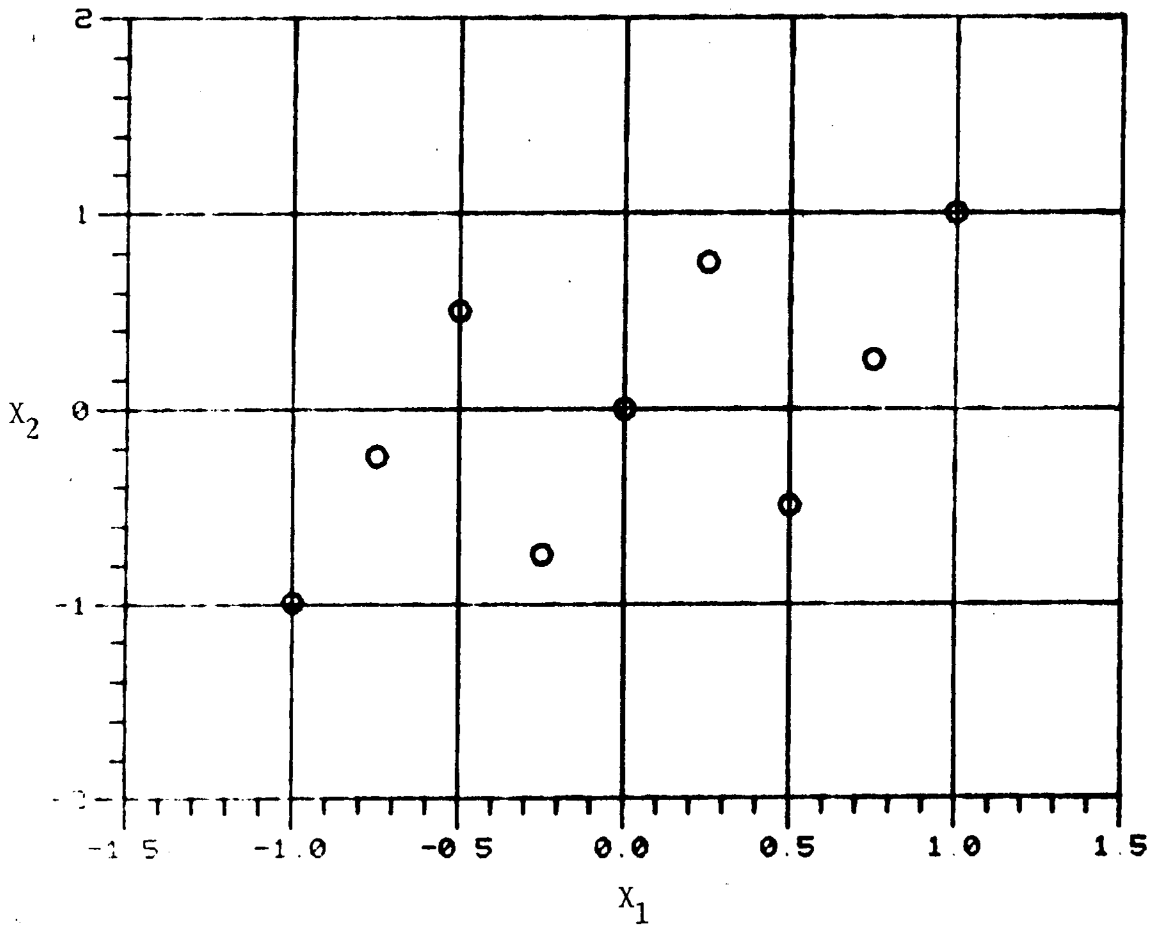
$$\underset{\sim}{W} = \Omega \int_R \underset{\sim}{x}\underset{\sim}{x}'d\underset{\sim}{x} = \begin{bmatrix} 1 & 0 & 0 & 1/3 & 1/3 & 0 \\ 0 & 1/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/5 & 1/9 & 0 \\ 1/3 & 0 & 0 & 1/9 & 1/5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/9 \end{bmatrix}$$

FIGURE 4.2.1



Locations of design points when $\zeta = 1$

FIGURE 4.2.2



Locations of design points when $\zeta = \frac{1}{2}$.

4.2.4. The Criterion Used for Comparison

We believe that IMSE is a reasonable choice for comparison of estimators. However, the computation of the IMSE of biased estimators generally requires the specification of the coefficients of the model. Such specification would limit the generality of the results of the comparisons.

We shall instead compare estimators by equating their integrated variance (when possible) and observing a bound on the ratio between the integrated bias of one estimator and the integrated bias of the zero estimator. The two ratios, for each of the two estimators, are compared. We would prefer that estimator whose ratio is smaller. Let's see what inference can be made about the IMSE of the two estimators after comparing these ratios.

Consider two estimators, $\underset{\sim}{b}_1$ and $\underset{\sim}{b}_2$, whose respective IMSE is $J_i = V_i + B_i$, $i = 1,2,$. If the bias error of the zero estimator is B_0. then let ϕ_i for $i = 1$ and $i = 2$ be the respective bounds on the ratio of B_i to B_0. Then we have

$$J_i = V_i + (B_i/B_0)B_0 \leq V_i + \phi_i B_0.$$

When we have equated the integrated variance so that $V_1 = V_2$ then $\phi_1 < \phi_2$ implies that the bound on the IMSE of $\underset{\sim}{b}_1$ is less than the bound on the IMSE of $\underset{\sim}{b}_2$ for all B_0. For certain values of the coefficients, it may be that the IMSE of $\underset{\sim}{b}_2$ is smaller than the IMSE of $\underset{\sim}{b}_1$. However, with little specific knowledge of the coefficients, this seems an objective method for comparing estimators. We shall call $\underset{\sim}{b}_1$ a more efficient estimator than $\underset{\sim}{b}_2$ when ϕ_1 is less than ϕ_2 given that $V_1 = V_2$.

When the two V_i's can not be equated, then we consider the difference

$$V_1 - V_2 + (\phi_1-\phi_2)B_0,$$

set this difference equal to zero, and then solve for B_0. When a bound M on B_0 satisfies

$$B_0 < M < \frac{V_2-V_1}{\phi_1-\phi_2}$$

then we know that $\underset{\sim}{b}_1$ is more "efficient" than $\underset{\sim}{b}_2$.

We interject a word of caution at this point. For many criteria, it is possible to find optimal estimators. An optimal estimator may perform well for a particular criterion, but it may be poor when the criterion is modified slightly.

The optimal estimator for the criterion used here is the $\underset{\sim}{b}^+$ estimator. Since this estimator is similar to the ridge estimator, it is tempting to infer that the results of these comparisons apply to the ridge estimator. However, the fact that the scalar k of the ridge estimator is not generally chosen in the manner that we will choose it weakens this inference.

4.2.5. Explanation of the Comparison

In this section we are interested in comparing four estimators:

- The full model estimator.
- The $\underset{\sim}{b}^+$ estimator.
- The BD estimator.
- The KMH estimator.

Each of these estimators is a deterministic transform of the least squares estimator $\hat{\underset{\sim}{\beta}}$. A typical estimator is $\underset{\sim}{b} = \underset{\sim}{K}\hat{\underset{\sim}{\beta}}$. The full model estimator, the BD estimator, and the KMH estimator have an integrated variance that depends on the design only. The only one of the four estimators whose integrated variance can be modified once the design is fixed is the $\underset{\sim}{b}^+$ estimator. We shall vary the scalar k of the $\underset{\sim}{b}^+$ estimator so as to equate the integrated variance error of $\underset{\sim}{b}^+$ with that of either the BD or KMH estimator. Only choosing k equal to zero will equate the integrated variance error of $\underset{\sim}{b}^+$ estimator with the full model estimator. Therefore, the estimator $\underset{\sim}{b}^+$ will be compared with the full model estimator when the k is fixed so that the integrated variance of the KMH estimator and the integrated variance of $\underset{\sim}{b}^+$ are equal.

4.2.5.1. Method of Obtaining Bound on the Ratio of Bias Errors

Consider an estimator of the form $b = K\hat{\underset{\sim}{\beta}}$. The bias error of this estimator is

$$\frac{N\Omega}{\sigma^2}\int_R (\underset{\sim}{x}'(\underset{\sim}{K}-\underset{\sim}{I})\underset{\sim}{\beta})^2 d\underset{\sim}{x}.$$

The bias error of the zero estimator is obtained by setting k to zero and is B_0 where

$$B_0 = \frac{N}{\sigma^2}\,\underset{\sim}{\beta}'\underset{\sim}{W}\underset{\sim}{\beta} .$$

The ratio of these two bias errors is

$$\Omega\int \frac{(\underset{\sim}{x}'(\underset{\sim}{K}-\underset{\sim}{I})\underset{\sim}{\beta})^2}{\underset{\sim}{\beta}'\underset{\sim}{W}\underset{\sim}{\beta}}\, d\underset{\sim}{x} . \qquad (4.2.5.1.1)$$

A version of the C-S inequality states that "Let $\underset{\sim}{A}$ be a p.d. m x m matrix and $\underset{\sim}{U} \in E_m$, (i.e., $\underset{\sim}{U}$ is an m-vector); then

$$\sup_{\underset{\sim}{X} \in E_m} \frac{(\underset{\sim}{U}'\underset{\sim}{X})^2}{\underset{\sim}{X}'\underset{\sim}{A}\underset{\sim}{X}} = \underset{\sim}{U}'\underset{\sim}{A}^{-1}\underset{\sim}{U}$$

and the supremum is attained at $\underset{\sim}{X}^* = \underset{\sim}{A}^{-1}\underset{\sim}{U}$." Rao(1965). We shall apply this fact to the integrand of(4.2.1.1.1). In this case letting $\underset{\sim}{U}' = \underset{\sim}{x}'(\underset{\sim}{K}-\underset{\sim}{I})$, $\underset{\sim}{X} = \underset{\sim}{\beta}$, and $\underset{\sim}{A} = \underset{\sim}{W}$, we have

$$\sup_{\underset{\sim}{\beta} \in E_p} \frac{(\underset{\sim}{x}'(\underset{\sim}{K}-\underset{\sim}{I})\underset{\sim}{\beta})^2}{\underset{\sim}{\beta}'\underset{\sim}{W}\underset{\sim}{\beta}} = \underset{\sim}{x}'(\underset{\sim}{K}-\underset{\sim}{I})\underset{\sim}{W}^{-1}(\underset{\sim}{K}-\underset{\sim}{I})'\underset{\sim}{x} \ .$$

By the monotone convergence theorem ,

$$\sup \Omega \int_R \frac{(\underset{\sim}{x}'(\underset{\sim}{K}-\underset{\sim}{I})\underset{\sim}{\beta})^2}{\underset{\sim}{\beta}'\underset{\sim}{W}\underset{\sim}{\beta}} \, d\underset{\sim}{x} \le \Omega \int \underset{\sim}{x}'(\underset{\sim}{K}-\underset{\sim}{I})\underset{\sim}{W}^{-1}(\underset{\sim}{K}-\underset{\sim}{I})'\underset{\sim}{x} \, d\underset{\sim}{x}$$

$$= \mathrm{tr}\ \Omega \int_R \underset{\sim}{x}'(\underset{\sim}{K}-\underset{\sim}{I})\underset{\sim}{W}^{-1}(\underset{\sim}{K}-\underset{\sim}{I})'\underset{\sim}{x} \, d\underset{\sim}{x} = \mathrm{tr}\ \underset{\sim}{W}(\underset{\sim}{K}-\underset{\sim}{I})\underset{\sim}{W}^{-1}(\underset{\sim}{K}-\underset{\sim}{I})' \ .$$

4.2.6. Results

Table 4.2.1 lists the results of our comparisons.

First, let us consider the comparison between the BD and the $\underset{\sim}{b}^+$ estimator. We see that for all choices of ζ, the bound ϕ of the $\underset{\sim}{b}^+$ estimator is smaller than the bound ϕ of the BD estimator. This means that for all values of the integrated bias of the zero estimator, the $\underset{\sim}{b}^+$ estimator has a smaller upper bound on its IMSE. Thus, the $\underset{\sim}{b}^+$ estimator with this choice of k is more efficient than the BD estimator.

Secondly, comparing the KMH and $\underset{\sim}{b}^+$ we see this same relationship between the bounds on the ratio. Thus $\underset{\sim}{b}^+$ with this choice of k is more efficient than the KMH estimator.

Thirdly, we compare the full model estimator with $\underset{\sim}{b}^+$ when k has been chosen to equate the integrated varuance of $\underset{\sim}{b}^+$ to the integrated variance of the KMH estimator. For $\zeta = 1$, we have

$$\frac{V_{FM} - V_{\underset{\sim}{b}^+}}{\phi_{\underset{\sim}{b}^+} - \phi_{FM}} = \frac{4.25 - 3}{.0857-0} < 10 .$$

So if one could specifv a bound smaller than 10 on η^2/σ^2 one would achieve smaller IMSE using $\underset{\sim}{b}^+$ with k fixed as above rather than using the full model estimator. This results holds for all $\underset{\sim}{\beta}$, and the result holds for ζ set at one. As ζ was decreased toward zero, this result held for larger values of the integrated bias of the zero estimator. When ζ is .125,

$$\frac{V_{FM} - V_{\underset{\sim}{b}^+}}{\phi_{\underset{\sim}{b}^+} - \phi_{FM}} = \frac{4020 - 1210}{.212 - 0} \cong 15000.$$

Fourthlv, we compare the KMH and full model estimators. When ζ = 1 then

$$\frac{V_{FM} - V_{KMH}}{\phi_{KMH} - \phi_{FM}} = \frac{4.25 - 3}{3 - 0} \cong .4,$$

and when $\zeta = .125$

$$\frac{V_{FM} - V_{KMH}}{\phi_{KMH} - \phi_{FM}} = \frac{4020 - 1210}{3 - 0} \cong 1000$$

We see that the KMH estimator is preferable to the full model only for choices of bounds that are smaller than the bounds that were true for the MVFR estimator.

Fifthly, we compare the BD and full model estimators. When $\zeta=1$, then

$$\frac{V_{FM} - V_{BD}}{\phi_{BD} - \phi_{FM}} = \frac{4.25 - 2.15}{7.25 - 0} \cong .2,$$

and when $\zeta = .125$

$$\frac{V_{FM} - V_{BD}}{\phi_{BD} - \phi_{FM}} = \frac{4020 - 34.6}{3.99} \simeq 1000 .$$

So the BD and KMH estimators are very similar in their comparison to the full model estimator.

Lastly, we compare the BD and KMH estimators. When $\zeta = 1$, then

$$\frac{V_{KMH} - V_{BD}}{\phi_{BD} - \phi_{KMH}} = \frac{3 - 2.15}{7.25 - 3} \cong .25.$$

When ζ is 3/4, 1/2, 1/4 and 1/8, then this bound is .14, .67, 67.4, and 1187.3, respectively. Thus, the BD estimator seems preferable to the KMH estimator only when the independent variables are highly correlated.

4.2.7. Interpretations of the Results

Let us see what light has been shed on the answers to the questions posed in Section 4.2.1. For the first question it would appear that these comparisons indicate that as the independent variables become more correlated the BD estimator becomes more clearly preferable. We must qualify this statement since the answer depends on the values of the unknown parameters. Here it would seem that the SK estimator could be used. However, the SK estimator can do no better than the BD estimator in achieving small integrated variance. The KM estimator may be preferable when the integrated variance of the BD estimator is too large.

It would seem that we have answered the second question in the affirmative. It appears that the $\underset{\sim}{b}^{+}$ is more efficient than either the BD or KMH estimators. Since this estimator is similar to the ridge estimator, one might expect that the ridge estimator is more efficient than either the BD or KMH estimators. However, it may be that smaller values of k are used in practice. If this is the case, then $\underset{\sim}{b}^{+}$ will have characteristics that are similar to the full model estimator.

Table 4.2.1

Comparison of Biased Estimators

		Variance Error			Bound on Bias Error Ratio		Variance Error				Bound on Bias Error Ratio	
ζ	r	BD	b^+	k used	BD	b^+	Full	KMH	b^+	k used	KMH	b^+
1	0	2.15	2.15	3.4	7.25	.301	4.25	3	3	1.41	3	.0857
3/4	.28	2.55	2.55	3.801	4.77	.476	5.44	2.8	2.8	3.2	3	.378
1/2	.60	3.69	3.69	2.36	3.93	.798	15.5	4.31	4.31	1.88	3	.654
1/4	.882	9.88	9.88	.51	3.83	1.03	240	65.8	65.8	.0442	3	.277
1/8	.969	34.6	34.6	.1095	3.99	1.09	4020	1210	1210	.00194	3	.212

4.3 AN APPLICATION OF BIASED ESTIMATION TO THE LOW DOSE EXTRAPOLATION PROBLEM

4.3.1 Introduction

In screening environmental contaminants for their effects on the health of a large population, often the goal is to obtain information about the effects when the exposure dose of the contaminant is low. The exposure to low levels of a contaminant may result in a serious health effect (death or presence of tumors) in only a small proportion of the population. However, even a proportion that might normally be considered small may be of concern to public health officials when the target population is large (say, over a million).

When the response rate is low, say less than one in a hundred, there is a fair chance that a reasonable size sample (say, 50 individuals) will contain no affected individuals. Thus, there is a real danger that hazardous contaminants can pass undetected through such a screening.

When all outcomes are the same (no detections), the estimated variance is zero. However, the variance must have been nonzero if we have missed detection due to using too small a sample size. Therefore, when we do make a detection we must realize that it is unlikely that the observed proportions will coincide with proportions that would be obtained in repeated experiments. In order that we are not mislead by such random fluctuation we may plan at the outset to limit the amount of detailed information to be extracted from the data. Although using large numbers of experimental units can improve precision, there is always a limit on the number of experimental units available.

When one conducts an experiment with a small number of experimental units, one is taking a risk that the information to be derived from the experimental data lacks sufficient precision. Since we have ruled out increasing the number of experimental units, we shall confine our remarks to two alternatives that can also improve precision.

First, one might decide to allocate the experimental units so that the barest minimum of detailed information can be extracted from the data. Such a sacrifice would be made in order that this minimum amount of information is as precise as possible. For example, when the shape of the response curve is unknown and the number of experimental units is fixed, then a two-point design can more precisely estimate the differential response between the two points than is possible when some units are allocated to other dose levels. However, confining the design to two levels sacrifices information about the gradient.

Second, one might decide to extrapolate. That is, choose experimental doses higher than the interval of interest, fit a curve to these points, and use the points lying on the curve in the interval of interest as estimates of the response. In practice, a few units are usually tested at the background dose as a control. Thus, one might prefer to call this method an interpolation rather than an extrapolation. However, the term extrapolation is usually used here. Perhaps this term persists because the number of control units is usually small, that is,the resulting precision is only able to detect background response levels of, say, five percent.

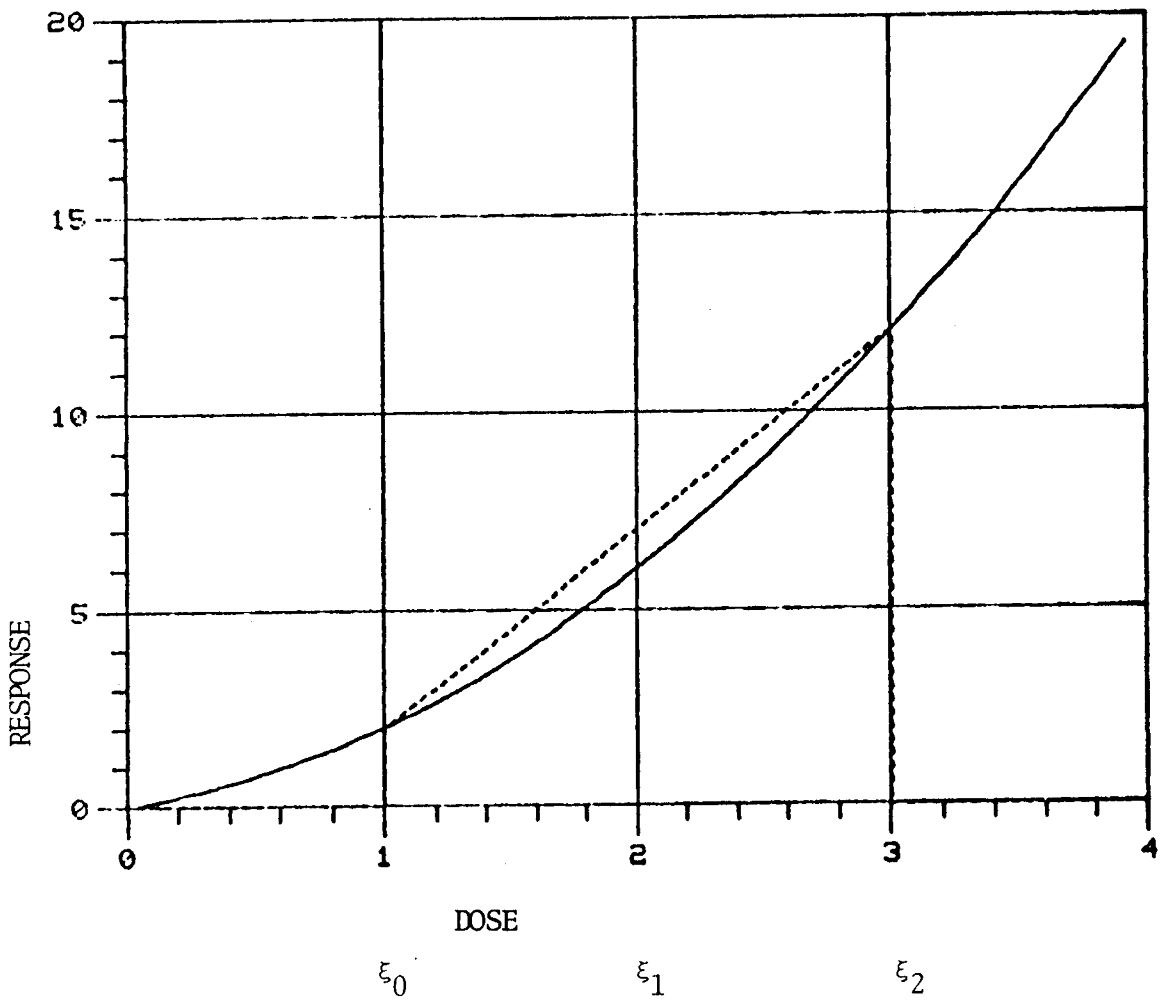
Usually a straight line is chosen as the curve fit to the dose levels when interpolation is used. There are a number of reasons to choose a linear function. First, the parameters of a linear function can be estimated more precisely (less variance) than the parameters of a more complex function. Second, most dose-response curves are sigmoid in shape. Thus, intervening points on a line (connecting points on the lower concave portion of the dose-response curve) will lie above (on the average) the true response curve. See Figure 4.3.1.

Here, we shall discuss some methods that can be employed to increase the precision of estimates of the dose-response curve. We assume that the experimenter is able to dose at any reasonable level that he desires.

First, we assume that the true function is unknown, that goal of the experiment is the estimation of Δ (the differential response between dose levels ξ_0 and ξ_1), that a two-point design at dose levels ξ_0 and $\xi_2 > \xi_1$ is employed, that the estimator for Δ (say $\hat{\Delta}_2$) is based on linear interpolation between ξ_0 and ξ_2, and that only the variance of $\hat{\Delta}_2$ is the criterion used. In this setting we find that the variance of $\hat{\Delta}_2$ is smallest when ξ_2 is as large as possible.

Second, we assume that the true model is a quadratic function. Since $\hat{\Delta}_2$ will be biased, we use mean square error (MSE) criterion rather than variance criterion. We find that the square of the quadratic term divided by the variance of the experimental error must be bounded in order to guarantee that the MSE of $\hat{\Delta}_2$ is smaller using $\xi_2 > \xi_1$ rather than $\xi_2 = \xi_1$.

FIGURE 4.3.1



The overprediction of a concave upward response function at a dose equal to 2. The straight line passes thru the response function at dose levels of 1 and 3.

Third, we consider a particular quadratic, namely $\beta_2 x^2$. Since the true model is unknown we cannot deny that this may be the true form of the function. We find under special circumstances that we would only want to choose ξ_2 only slightly larger than ξ_1 in order to minimize MSE. We can not recommend this as a solution, but only consider it as information to motivate making further constraining assumptions about the form of the dose-response curve.

Fourth, we assume that ξ_2 has been chosen arbitrarily larger than ξ_1. We assume that the true dose-response curve is a quadratic function with positive coefficients. We then find functions that bound this class of functions, and estimate them.

Finally, we show what level lower than ξ_2 to use in a subsequent experiment to best refute the upperbound on this class. We also recommend how sample size should be chosen for this subsequent experiment.

4.3.2. An Example

Consider a population of experimental units (humans, mice, etc.). Conceptually, it is possible to imagine the average response, $\eta(x)$, that would be obtained if the whole population were administered a dose x. We can further imagine that the function η could be defined for dose levels other than x. Although it would be impossible to verify by experimentation, it is conceptually possible to view $\eta(x)$ as a function defined for all dose levels x in some interval. We call $\eta(x)$ the true response.

Let us assume that the experimenter has interest in determining information about the response on an interval $R = \{x \mid \xi_0 < x < \xi_1\}$, which is a subinterval of the interval where $\eta(x)$ is defined. For example, ξ_0 might be the background concentration of an agent, and ξ_1 might represent a very small increase over ξ_0 that could be realized if no environmental control were exercised with regard to this agent. The term "very small" means that the concentration would be much smaller than typical concentrations that laboratories are designed to handle. Let $\underset{\sim}{Y}$ be the vector of observed responses from N experimental units which have been exposed to doses $(x_1, x_2, \ldots, x_N) = \underset{\sim}{x}'$. Some or all of these dose levels may be the same. Let $\underset{\sim}{\eta}(x) = (\eta(x_1), \eta(x_2), \ldots, \eta(x_N))$ be the true response vector. We assume that

$$\underset{\sim}{Y} = \underset{\sim}{\eta}(x) + \underset{\sim}{\varepsilon},$$

where $\underset{\sim}{\varepsilon}$ is a vector of random uncorrelated errors with zero means and common variance σ^2. Biological differences, errors in applying the dose, sampling error and instrument error are some of the sources of error that make the vector $\underset{\sim}{\varepsilon}$ nonzero. So far we have made no assumptions about the functional form of $\eta(x)$.

A minimal amount of useful information about $\eta(x)$ is provided by the difference

$$\Delta = \eta(\xi_1) - \eta(\xi_0).$$

The parameter Δ is most meaningful when η is a monotonic increasing function of dose. This population parameter can be estimated rather precisely as compared to trying to estimate a more complex aspect of $\eta(x)$ on the interval R.

For reasons such as these, the experimenter may be content with an experimental design that can provide information only about Δ. This is especially true when the number of experimental units is small.

We shall discuss only the case where two dose levels are used.

If one obtains n_0 response measurements at ξ_0 and n_1 response measurements at ξ_1, then an unbiased estimator for Δ is the difference of the two means

$$\hat{\Delta}_1 = \bar{Y}_{\xi_1} - \bar{Y}_{\xi_0} .$$

Here, $\bar{Y}_{\xi_1}$ and $\bar{Y}_{\xi_0}$ are the averages of the responses made at dose levels ξ_1 and ξ_0, respectively.

The variance of $\hat{\Delta}_1$ is

$$\text{Var } \hat{\Delta}_1 = \sigma^2 \left(\frac{1}{n_0} + \frac{1}{n_1}\right).$$

It may be judged that this variance is too large. Such a judgment is usually based on the size of type I and type II errors. In order to compute the type II error, one must specify how small a difference one desires to detect.

When these computations indicate that the variance is too large, one could decide to increase n_0 and n_1.

We assume that this alternative is not available. We will consider extrapolating. That is, instead of using ξ_1 as a dose level, we will dose at a level higher than ξ_1, and using linear interpolation, estimate Δ. In this case, our estimator for Δ is

$$\hat{\Delta}_2 = (\xi_1 - \xi_0)\tilde{\beta}_1,$$

where $\tilde{\beta}_1$ is the estimated slope from fitting the straight line by least squares.

If we made n_2 and n_0 response measurements at $\xi_2 (\xi_2 > \xi_1)$ and ξ_0, respectively, then the variance of $\hat{\Delta}_2$ is

$$\text{Var } \hat{\Delta}_2 = \left[\frac{\xi_1 - \xi_0}{\xi_2 - \xi_0}\right]^2 \sigma^2 \left[\frac{1}{n_0} + \frac{1}{n_2}\right] .$$

If $n_2 = n_1$ then the variance of $\hat{\Delta}_2$ will be less than the variance of $\hat{\Delta}_1$ whenever

$$\frac{\xi_1 - \xi_0}{\xi_2 - \xi_0} < 1.$$

Thus, based on only the variance, we would prefer that ξ_2 be greater than ξ_1. This result comes about because we are ignoring the fact that $\eta(x)$ may not be a straight line.

From figure 4.3.1, we see that on the average $\hat{\Delta}_2$ will tend to overestimate Δ. This tendency is not taken into account when only variance is considered.

In order to partially account for this tendency, let us assume that $\eta(x)$ is only slightly more complex than a straight line. We will assume that $\eta(x)$ is a quadratic in x, namely

$$\eta(x) = \beta_0 + \beta_1 x + \beta_2 x^2 .$$

Under this assumption, $\hat{\Delta}_2$ is a biased estimator for Δ, while $\hat{\Delta}_1$ remains unbiased. The expected value of $\hat{\Delta}_2$ is

$$\begin{aligned} E(\hat{\Delta}_2) &= (\xi_1-\xi_0)E(\tilde{\beta}_1) \\ &= (\xi_1-\xi_0)(\beta_1+a\beta_2), \end{aligned}$$

where

$$a = (\xi_2+\xi_0).$$

Since

$$\Delta = (\xi_1-\xi_0)\beta_1 + (\xi_1^2-\xi_0^2)\beta_2 ,$$

the mean square error (MSE) of $\hat{\Delta}_2$ is then

$$\begin{aligned} &\mathrm{Var}\,\hat{\Delta}_2 + (E(\hat{\Delta}_2)-\Delta)^2 \\ &= \sigma^2 \left(\frac{1}{n_0} + \frac{1}{n_2}\right)\left(\frac{\xi_1-\xi_0}{\xi_2-\xi_0}\right)^2 + (\xi_1-\xi_0)^2(\xi_2-\xi_1)^2\beta_2^2 , \quad (4.3.2.1) \end{aligned}$$

and $\hat{\Delta}_2$ will attain smaller MSE than $\hat{\Delta}_1$ when (assuming $n_1 = n_2$)

$$\beta_2^2/\sigma^2 < \frac{\left(\frac{1}{n_0} + \frac{1}{n_2}\right)\left(1 - \left(\frac{\xi_1-\xi_0}{\xi_2-\xi_0}\right)^2\right)}{(\xi_1-\xi_0)^2(\xi_2-\xi_1)^2} \tag{4.3.2.2}$$

$$= \left(\frac{1}{n_0} + \frac{1}{n_2}\right)(\xi_2+\xi_1-2\xi_0)/(\xi_1-\xi_0)^2(\xi_2-\xi_0)^2(\xi_2-\xi_1) = M$$

Thus, $\hat{\Delta}_2$ is preferable to $\hat{\Delta}_1$ only when (4.3.2.2) holds.

4.3.2.1. An Example where Bias Is More Important than Variance.

In the last section we have considered that the true model is a quadratic. From (4.3.2.1) we see that when β_2 is zero then one attains smaller MSE for larger values of ξ_2. We have also shown how β_2^2/σ^2 must be bounded to attain smaller MSE using $\hat{\Delta}_2$ rather than $\hat{\Delta}_1$. We shall now consider a situation where the optimum choice for ξ_2 is only slightly larger than ξ_1.

We assume that the response $\eta(x)$ is an incidence rate. Therefore, $\eta(x)$ is always less than one. We assume that the true response is a second degree polynomial in x, namely

$$\eta(x) = \beta_2 x^2.$$

We have set β_0 and β_1 to zero for our convenience. However, in a sense, this choice of β_0 and β_1 yields a response model whose curvature is somewhat extreme, when, in fact, β_0 and β_1 are positive.

It is often assumed when working with proportions that the variance is due solely to sampling error. We shall also make this assumption. However, it should be kept in mind that this is probably a lower bound on the true variance in many practical applications. If the variance is actually larger, then our assumption will yield a value of ξ_2 which is smaller than the optimum value. However, the realization that this assumption may be wrong does justify using the least squares estimator $\hat{\Delta}_2$ instead of some weighted least squares estimator.

With our assumptions about the model and the form of the variance, we have that

$$\mathrm{MSE}(\hat{\Delta}_2) = \left\{ \frac{\beta_2\xi_0^2(1-\beta_2\xi_0^2)}{n_0} + \frac{\beta_2\xi_2^2(1-\beta_2\xi_2^2)}{n_2} \right\} \left(\frac{\xi_1-\xi_0}{\xi_2-\xi_0} \right)^2$$

$$+ (\xi_2-\xi_1)^2(\xi_1-\xi_0)^2\beta_2^2 . \qquad (4.3.2.1.1)$$

Taking the derivative of $\mathrm{MSE}(\hat{\Delta}_2)$ with respect to ξ_2 and setting it to zero and multiplying by it by $(\xi_2-\xi_0)^3/2\beta_2(\xi_1-\xi_0)^2$, we have

$$(\xi_2-2\beta_2\xi_2^3)(\xi_2-\xi_0)/n_2-\xi_2^2(1-\beta_2\xi_2^2)/n_2 + \beta_2(\xi_2-\xi_1)(\xi_2-\xi_0)^3 = 0. \qquad (4.3.2.1.2)$$

If we multiplied this expression out we would find that the only term that didn't have a β_2 in it is $-\xi_2\xi_0$. Consider the situation where ξ_0 is zero. Then we can divide (4.3.2.1.2) by $\xi_2^3\beta_2$. We are left with

$$-\xi_2/n_2 + \xi_2-\xi_1 = 0.$$

Solving for ξ_2 we have

$$\xi_2 = \xi_1 n_2/(n_2-1).$$

Using this value for ξ_2 we have that

$$\mathrm{MSE}\ \hat{\Delta}_2 = \beta_2\xi_1^2/n_2 - \beta_2^2\xi_1^4/(n_2-1).$$

The ratio of MSE $\hat{\Delta}_2$ to the MSE of $\hat{\Delta}_1$ is

$$\frac{n_2}{n_2-1} - \frac{1}{(n_2-1)(1-\beta_2\xi_1^2)} .$$

We see that for typically small values of $\Delta = \beta_2\xi_1^2$ there will be only slight improvement.

We have derived the optimal choice for ξ_2 using a number of assumptions that seem weighted in favor of not interpolating at all. In practice, larger choices of ξ_2 may be justifiable on other grounds. In view of the fact that $\xi_1^2\beta_2$ will be small in this application, we do not recommend that ξ_2 be selected in this way.

4.3.3 Obtaining Bounds on the Dose-Response Curve.

In the last section we have seen that one would not optimally choose ξ_2 much larger than ξ_1 when a straight line is used for extrapolation and the true model is a very restrictive quadratic. Let's say one does not know the true model. Then he can't rule out that the true model may behave similarly to the quadratic that was used in the previous section. That is, the MSE may be minimized for the true model only when ξ_2 is quite close to ξ_1. Now MSE consists of both bias and variance, and variance can be controlled by choosing ξ_2 large. Therefore, it seems that when ξ_2 is large and one fits a straight line one needs to be most concerned about the bias.

When one fits a straight line to dose-response data on the grounds that it is conservative, one is making some assumptions about the form of the true dose-response curve. He is assuming that the true dose-response curve is both monotonically increasing and concave upward over the interval from ξ_0 to ξ_2. When these assumptions hold then one can infer that the straight line will lie above the true response on the average. This is often considered a rather weak inference. It could be that Δ is zero, and yet the estimate of Δ that is computed from the straight line fit suggests that the environmental contaminant poses a real danger to the health of some large target population. Rather than take action to control the contaminant, it may be decided that it is preferable to obtain a better bound on Δ.

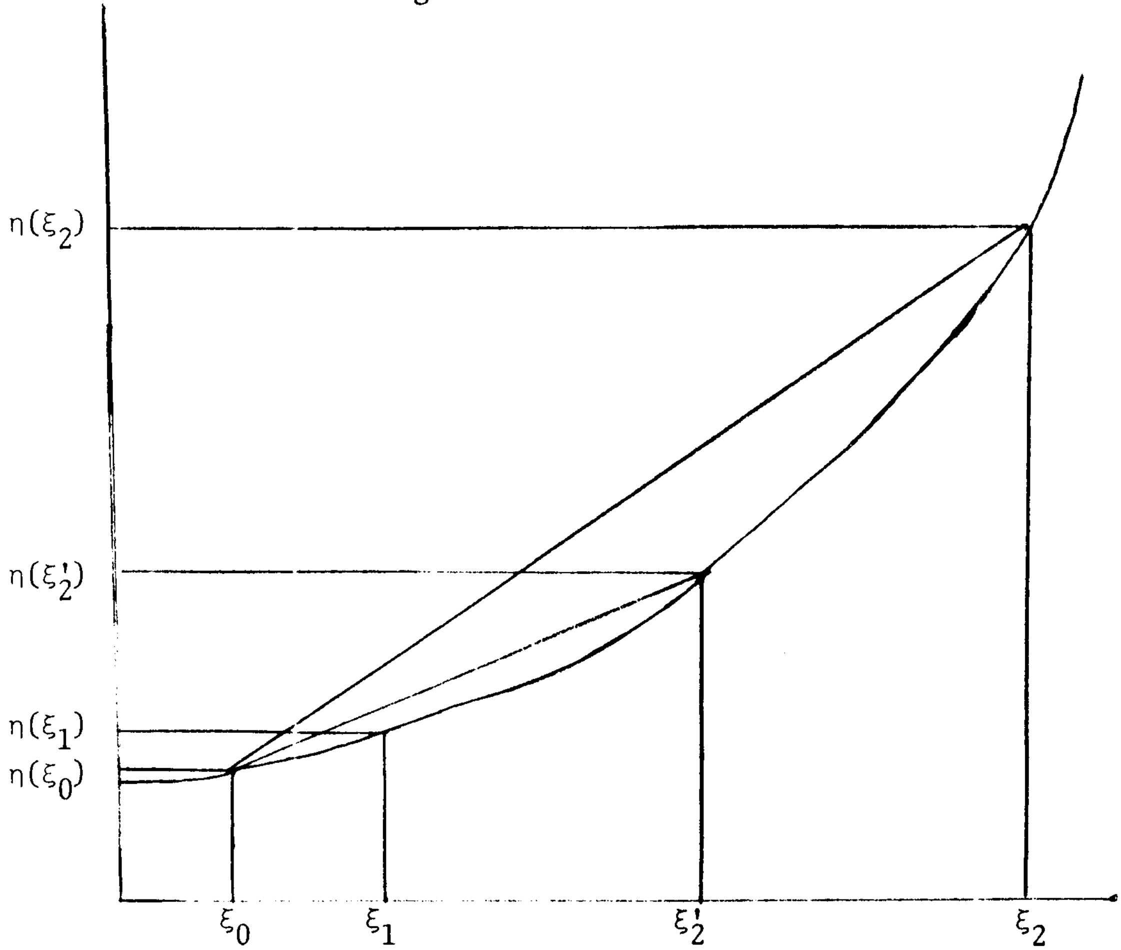
If the assumptions about the shape of the dose response curve are true, then doing a new experiment using a dose level ξ_2', where ξ_2' is less than ξ_2, will result in an estimate for Δ which is still conservative and yet this new estimate will have a smaller expected value than the previous estimate for Δ (See Figure 4.3.3.1). However, many animals might be needed before one could safely reject the old estimate for Δ and accept the new estimate.

Thus, there seems to be a need to decide how many animals should be used for such a subsequent experiment. There is also a need to determine ξ_2'.

It would seem that this problem is solvable by viewing it in the light of classical hypothesis testing. Viewed in this light we might want to test whether $\eta(\xi_2')$ is larger than $\hat{\eta}_L(\xi_2')$, the estimator of $\eta(\xi_2')$ based on linear interpolation between ξ_0 and ξ_2, or if whether $\eta(\xi_2')$ is zero. The problem is not that simple.

Consider choosing the point ξ_2', which is in the interval $[\xi_1,\xi_2]$, that necessitates the use of the least number of animals to reject $\eta(\xi_2') = \hat{\eta}_L(\xi_2')$ (and accept $\eta(\xi_2') = 0$) when $\eta(\xi_2') = \hat{\eta}_L(\xi_2')$ is false. Now, if we select $\xi_2'=\xi_2$ then $\hat{\eta}_L(\xi_2')-0$ is maximized among choices of ξ_2' from the interval $[\xi_1,\xi_2]$. Thus, it would seem that this criterion would indicate that a second experiment be repeated at the two

Figure 4.3.3.1



The figure shows that the bias error at ξ_1 resulting from the linear interpolation between $(\xi_0, \eta(\xi_0))$ and $(\xi_2', \eta(\xi_2'))$ is smaller than the bias error at ξ_1 resulting from the linear interpolation between $(\xi_0, \eta(\xi_0))$ and $(\xi_2, \eta(\xi_2))$. This result holds for any function that is monotone concave upward on (ξ_0, ξ_2) whenever $\xi_1 < \xi_2' < \xi_2$.

doses ξ_0 and ξ_2. Such a second experiment does not really seem to relevantly address the problem of obtaining a better bound on Δ.

On the other hand, if we choose ξ_2' to minimize the bias error of $\eta_L(\xi_1)$ we are lead to choose ξ_2' equal to ξ_1. Presumably, we do not have sufficient animals to test at ξ_1 and obtain sufficient precision.

In order to objectively select a "reasonable" point ξ_2' that lies between ξ_1 and ξ_2 we make an additional assumption. We tentatively assume that

$$\eta(x) = \beta_0 + \beta_1 x + \beta_2 x^2,$$

where β_1 and β_2 are non-negative. This assumption about the signs of β_1 and β_2 is not so constraining that we cannot define β_0, β_1 and β_2 to be those values that make $\eta(x)$ coincide with the true unknown response at ξ_0 and ξ_2 when the true response is monotonically increasing. Before we consider the inferences that can be made when this assumption holds, let us first consider how constraining the assumptions about β_0, β_1, and β_2 are.

First, consider the function

$$\eta_A(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$$

where this function is equal to the true unknown function at the points ξ_0, ξ_1 and ξ_2. In order for $\eta_A(x)$ to satisfy these requirements it need not follow that $\eta_A(x)$ is monotone or concave upward on (ξ_0,ξ_2). However, if the true unknown function is both monotone and concave upward then the parameters of $\eta_A(x)$ must satisfy the conditions

$$\alpha_2 > 0 \text{ and } \alpha_1 > -\alpha_2(\xi_1+\xi_0).$$

The condition $\alpha_2 > 0$ arises from the fact that $\eta_A(\xi_1)$ must be less than the coordinate corresponding to ξ_1 on the straight line passing through $(\xi_0,\eta_A(\xi_0))$ and $(\xi_2, \eta_A(\xi_2))$. The condition $\alpha_1 > -\alpha_2(\xi_1 + \xi_0)$ arises since $\eta_A(\xi_0) < \eta_A(\xi_1)$.

Since α_1 need not be positive, it does not follow that our assumption is general enough to allow $\eta(x)$ to coincide with the true dose-response model at ξ_0, ξ_1 and ξ_2. However, α_2 is a measure of how close the true response at ξ_0, ξ_1 and ξ_2 lie to a straight line. In this sense then, when α_2 is small, the true unknown response will be approximately linear. When α_2 is small, α_1 cannot be too negative. This is because α_1 must be larger than $-\alpha_2(\xi_1+\xi_0)$. Now let us consider the error of $\eta(x)$ at $x = \xi_1$ when α_1 is negative.

Since $\eta(x)$ coincides with the true unknown response (and with $\eta_A(x)$) at ξ_0 and ξ_2, it follows that

$$\beta_0 - \alpha_0 + (\beta_1-\alpha_1)\xi_0 + (\beta_2-\alpha_2)\xi_0^2 = 0 \qquad (4.3.3.1)$$

and

$$\beta_0 - \alpha_0 + (\beta_1-\alpha_1)\xi_2 + (\beta_2-\alpha_2)\xi_2^2 = 0. \qquad (4.3.3.2)$$

The error of $\eta(x)$ at $x = \xi_1$ is

$$\eta(\xi_1) - \eta_A(\xi_1) = (\beta_0-\alpha_0) + (\beta_1-\alpha_1)\xi_1 + (\beta_2-\alpha_2)\xi_1^2. \qquad (4.3.3.3)$$

Substituting the value of $\beta_0 - \alpha_0$ defined by (4.3.3.1) into (4.3.3.3), and substituting the value of $\beta_1 - \alpha_1$ defined by the difference of (4.3.3.1) and (4.3.3.2) into (4.3.3.3), we have that

$$\eta(\xi_1) - \eta_A(\xi_1) = (\beta_1-\alpha_1)\left[(\xi_1-\xi_0) + \frac{\xi_1^2-\xi_0^2}{\xi_2-\xi_0}\right]. \qquad (4.3.3.4)$$

Equation(4.3.3.4) shows that $\eta(\xi_1)-\eta_A(\xi_1)$ is always positive. Also the equation shows that $\eta(\xi_1) - \eta_A(\xi_1)$ is minimized if β_1 is taken to be zero when α_1 is negative. It appears that $\eta(\xi_1) - \eta_A(\xi_1)$ decreases as ξ_2 gets larger. This is true if the true response is a quadratic so that α_1 is fixed. However, α_1 can be expected to change with ξ_2 when the true unknown response is more complex than a quadratic.

Thus, we consider

$$\eta(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

to be the quadratic that coincides with the true response at ξ_0, ξ_1 and ξ_2 if the coefficients are positive. When this is not possible we assume that

$$\eta(x) = \beta_0 + \beta_2 x^2$$

where $\eta(x)$ coincides with the true response at ξ_0 and ξ_2. In this case, our assumption is conservative since $\eta(\xi_1) - \eta_A(\xi_1)$ is positive.

Now, consider the class of functions Q, where $\eta(x)$ is in Q if

$$\eta(x) = \beta_0 + \beta_1 x + \beta_2 x^2 ,$$

and β_1 and β_2 are non-negative. A relevant property of functions in Q is :

For any $\eta(x)$ in Q, there are two corresponding functions $\eta_L(x)$ and $\eta_Q(x)$ such that these functions are in Q anl

$$\eta_Q(x) \leq \eta(x) \leq \eta_L(x),$$

for $\xi_0 < x \leq \xi_2$, and

$$\eta_Q(\xi_i) = \eta(\xi_i) = \eta_L(\xi_i),$$

for $i = 0$ or 2. We can show that this holds by considering

$$\eta_L(x) = \beta_0 + \theta_1 x$$

and $\eta_Q(x) = \beta_0 + \theta_2 x^2$,

where $$\theta_1 = \beta_1 + \beta_2(\xi_2+\xi_0)$$

and $$\theta_2 = (\beta_1+\beta_2(\xi_2+\xi_0))/(\xi_2+\xi_0).$$

For these particular functions,

$$\eta_Q(x) < \eta(x) < \eta_L(x)$$

when $\beta_1 > 0$, $\beta_2 > 0$, and $x < \xi_0 + \xi_2$. Since $0 \leq \xi_0 < \xi_2$, we have that $\{x|\xi_0 < x < \xi_2\}$ is a subset of the set $\{x|x < \xi_0+\xi_2\}$. We require $\beta_2 > 0$ so that $\eta(x) < \eta_L(x)$ and we require $\beta_1 > 0$ so that $\eta_Q(x) < \eta(x)$. Under our assumptions about β_1 and β_2 we know that, since

$$\Delta = \eta(\xi_1) - \eta(\xi_0),$$

then

$$\eta_Q(\xi_1) - \eta(\xi_0) < \Delta < \eta_L(\xi_1) - \eta(\xi_0).$$

Hence when our assumptions are true about $\eta(x)$, we know that obtaining bounds on Δ is straightforward.

From a two-point design and for any $\eta(x)$ in Q, we can find unbiased estimators for the functions $\eta_L(x)$ and $\eta_Q(x)$ corresponding to $\eta(x)$. These estimators are

$$\hat{\eta}_L(x) = \{\xi_2\bar{Y}_{\xi_0} - \xi_0\bar{Y}_{\xi_2} + (\bar{Y}_{\xi_2}-\bar{Y}_{\xi_0})x\}/(\xi_2-\xi_0) \qquad (4.3.3.5)$$

and

$$\hat{\eta}_Q(x) = \{\xi_2^2\bar{Y}_{\xi_0} - \xi_0^2\bar{Y}_{\xi_2} + (\bar{Y}_{\xi_2}-\bar{Y}_{\xi_0})x^2\}/(\xi_2^2 - \xi_0^2) \ . \qquad (4.3.3.6)$$

We have that

$$E(\hat{\eta}_L(x)) = \beta_0 + \theta_1 x,$$

and

$$E(\hat{\eta}_Q(x)) = \beta_0 + \theta_2 x^2.$$

It follows that $\hat{\eta}_L(x)$ and $\hat{\eta}_Q(x)$ are unbiased for $\eta_L(x)$ and $\eta_Q(x)$ respectively.

Further, when $\text{Var}\ \bar{Y}_{\xi_0} = \text{Var}\ \bar{Y}_{\xi_2}$ and $\bar{Y}_{\xi_0}$ and $\bar{Y}_{\xi_2}$ are uncorrelated, then by the Gauss-Markov theorem we have at the point x that $\hat{\eta}_L(x)$ and $\hat{\eta}_Q(x)$ are best linear unbiased estimates of $\eta_L(x)$ and $\eta_Q(x)$, respectively. When $\text{Var}\ \bar{Y}_{\xi_0} \neq \text{Var}\ \bar{Y}_{\xi_2}$ then since we are using a two-point design we know that $\bar{Y}_{\xi_0}$ and $\bar{Y}_{\xi_2}$ are the maximum likelihood estimators for $\eta(\xi_0)$ and $\eta(\xi_2)$. So $\hat{\eta}_L(x)$ and $\hat{\eta}_Q(x)$ are the maximum likelihood estimators for $\eta_L(x)$ and $\eta_Q(x)$, respectively, because of the invariance property of the maximum likelihood estimators.

We see that if the true dose-response function is in Q then

$$E\{\hat{\eta}_Q(x)\} \leq \eta(x) \leq E\{\hat{\eta}_L(x)\}$$

for $\xi_0 < x < \xi_2$. Thus, by making a rather nonrestrictive assumption, we find that we can bound the true response function. However nonrestrictive the assumption, we must admit to a degree of arbitrariness and thus the need to validate this assumption.

In the validation we assume that the response is an incidence rate and that the frequency is binomially distributed. Consider a subsequent two point experiment at $x = \xi_0$ and $x = \xi_2'$. Since our goal is to make inferences about Δ, we clearly only wish to consider those values of ξ_2' such that $\xi_1 \leq \xi_2' < \xi_2$. The estimates of $\eta_L(\xi_2')$ and $\eta_Q(\xi_2')$ can be computed from the data of the first experiment. These are

$$\hat{\eta}_L(\xi_2') = \left\{\xi_2\bar{Y}_{\xi_0} - \xi_0\bar{Y}_{\xi_2} + (\bar{Y}_{\xi_2}-\bar{Y}_{\xi_0})\xi'_2\right\}/(\xi_2-\xi_0)$$

$$\hat{\eta}_Q(\xi_2') = \left\{\xi_2^2\bar{Y}_{\xi_0} - \xi_0^2\bar{Y}_{\xi_2} + (\bar{Y}_{\xi_2}-\bar{Y}_{\xi_0})\xi'^2_2\right\}/(\xi_2^2-\xi_0^2) \ .$$

We want to choose ξ_2' and the number of experimental units, n_2', so that it is possible to "optimally" test whether there is a significant difference between $\eta_L(\xi_2')$ and $\eta_Q(\xi_2')$. Since we don't know these values we do the test on $\hat{\eta}_L(\xi_2')$ and $\hat{\eta}_Q(\xi_2')$. We quantify this test as follows:

Let $\bar{Y}_{\xi_2'} = r_2'/n_2$ be the observed incidence rate at ξ_2'. The likelihood ratio test of the hypothesis

H_0: r_2' is distributed binomially with parameters n_2' and

$E\{\bar{Y}_{\xi'_2}\} = \hat{\eta}_L(\xi'_2)$

versus the alternative hypothesis

H_1: r_2' is distributed binomially with parameters n_2' and

$E\{\bar{Y}_{\xi_2'}\} < \hat{\eta}_L(\xi_2')$

is to reject H_0 when (using the approximate large sample distribution)

$$-2 \log_e L_{n_2'} > C_1$$

where

$$L_{n_2'} = \left(\frac{\hat{\eta}_L(\xi_2')}{\bar{Y}_{\xi_2'}}\right)^{n_2' \bar{Y}_{\xi_2'}} \left(\frac{1 - \hat{\eta}_L(\xi_2')}{1 - \bar{Y}_{\xi_2'}}\right)^{n_2'(1-\bar{Y}_{\xi_2'})} .$$

The constant C_1 is determined by selecting the significance level of the test. Let us now consider how the required sample size could be determined.

One way to select sample size is to determine a proportion so that if the observed proportion is less than this value then H_0 is rejected at the selected level. It would seem that if $\bar{Y}_{\xi_2'} \leq \hat{\eta}_Q(\xi_2')$ then one would expect to reject H_0, since

$$E(\hat{\eta}_Q(\xi_2')) = \eta_Q(\xi_2') < \eta_L(\xi_2').$$

When $\bar{Y}_{\xi_2'} = \hat{\eta}_Q(\xi_2')$, then $L_{n_2'}$ becomes

$$L_{n_2'} = \left(\frac{\hat{\eta}_L(\xi_2')}{\hat{\eta}_Q(\xi_2')}\right)^{n_2'\hat{\eta}_Q(\xi_2')} \left(\frac{1 - \hat{\eta}_L(\xi_2')}{1 - \hat{\eta}_Q(\xi_2')}\right)^{n_2'(1-\hat{\eta}_Q(\xi_2'))} . \tag{4.3.3.7}$$

Since $L_{n_2'} = L_1^{n_2'}$, we would reject H_0 at the 5% significance level when

$$-2\ n_2' \log_e L_1 > 2.71 = \chi^2_{1,0.1} .$$

The required number of observations to reject H_0 when $\bar{Y}_{\xi_2'} \leq \hat{\eta}_Q(\xi_2')$ is

$$n_2' = 2.71/(-2 \log_e L_1). \tag{4.3.3.8}$$

By choosing ξ_2' to maximize $-\log_e L_1$, equation (4.3.3.8) shows that we will minimize our sample size requirements.

4.3.3.1. Results

In this section we present tables listing our prescribed values of n_2' and ξ_2' for various choices of $\bar{Y}_{\xi_0}$, $\bar{Y}_{\xi_2}$, ξ_0 and ξ_2. First,we will describe how these tables were constructed and then we will comment on the implications of this analysis.

For fixed values of $\bar{Y}_{\xi_0}$, $\bar{Y}_{\xi_2}$, ξ_0, ξ_2' and ξ_2, we used (4.3.3.5) and (4.3.3.6) to compute $\eta_L(\xi_2')$ and $\eta_Q(\xi_2')$, respectively. n_2' was set to 1, and expression (4.3.3.7) was used to compute L_1, and from this we used (4.3.3.8) to compute n_2'. We varied ξ_2' until n_2' was minimized. These values of ξ_2' that minimized n_2', say ξ_2^*, are correct to the second decimal place. Table 4.3.3.1.1 shows n_2^* (the minimized value of n_2') and ξ_2^* for various values of $\bar{Y}_{\xi_0}$ and $\bar{Y}_{\xi_2}$ when $\xi_0 = 0$ and $\xi_2 = 1$. Table 4.3.3.1.2 shows n_2' and ξ_2^* when $\xi_0 = 1$ and $\xi_2 = 10$.

Both tables show that n_2^* decreases as either $\bar{Y}_{\xi_2}$ increases or $\bar{Y}_0$ decreases. These tables also show, that except when $\bar{Y}_{\xi_2}$ is zero, ξ_2^* decreases as either $\bar{Y}_{\xi_2}$ increase or $\bar{Y}_{\xi_0}$ decreases. When $\bar{Y}_{\xi_0}$ is zero then ξ_2^* increases with $\bar{Y}_{\xi_2}$.

Now let us consider the implications of these tables. Assume the experimenter has used a two-point design and estimated Δ based on $\hat{\eta}_L(x)$. He considers conducting a subsequent experiment to check on the conservativeness of his estimate for Δ. He is willing to reject $\eta(\xi_2^*) \geq \hat{\eta}_L(\xi_2^*)$ at the 5 percent significance level when $\bar{Y}_{\xi_2^*} \leq \hat{\eta}_Q(\xi_2^*)$. That is, he is willing to re-estimate Δ based only on $\bar{Y}_{\xi_2^*}$ and ignoring $\bar{Y}_{\xi_2}$ when the above hypothesis is rejected.

For example, Table 4.3.3.1.1 shows that if $\bar{Y}_{\xi_2} = 0.5$ and $\bar{Y}_{\xi_0} = 0$, then he should conduct the experiment at $\xi_2' = 0.30$ and he needs 24 animals. If he chooses ξ_2' other than 0.30 he will require more animals. Also if he desires to reject for some other value of $\bar{Y}_{\xi_2^*}$ between $\eta_Q(\xi_2^*)$ and $\eta_L(\xi_2^*)$ he will need more animals. If he finds that he has insufficient resources to meet the sample size requirements, this analysis indicates that he should not conduct a subsequent experiment for this purpose. His conclusion should be that Δ lies somewhere between the estimates based on $\hat{\eta}_Q(\xi_2)$ and $\hat{\eta}_L(\xi_2)$.

Table 4.3.3.1.1

Sample Size (Per Group) n_2^* and value of ξ_2^* to reject $\eta(\xi_2^*) \geq \hat{\eta}_L(\xi_2^*)$ when $\overline{Y}_{\xi_2^*} \leq \hat{\eta}_Q(\xi_2^*)$. ($\xi_2^*$ is the value of ξ_2' that requires the smallest n_2' for this test)

$\xi_0 = 0$, $\xi_2 = 1$

Significance probability = 0.05 (one-sided test)

Cell entries are n_2^* / ξ_2^*

$\overline{Y}_{\xi_0}$ \ $\overline{Y}_{\xi_2}$	0.01	0.015	0.02	0.05	0.10	0.20	0.25	0.30	0.50
0	1329 / 0.29	885 / 0.29	663 / 0.29	264 / 0.29	131 / 0.29	64 / 0.30	51 / 0.30	42 / 0.30	24 / 0.30
0.01		20589 / 0.47	5955 / 0.46	649 / 0.40	212 / 0.37	83 / 0.35	62 / 0.34	50 / 0.34	27 / 0.34
0.015			29041 / 0.48	969 / 0.43	260 / 0.39	92 / 0.36	68 / 0.36	53 / 0.36	28 / 0.36
0.02				1473 / 0.44	317 / 0.41	102 / 0.38	74 / 0.37	57 / 0.37	29 / 0.36
0.05					1125 / 0.46	183 / 0.42	118 / 0.41	85 / 0.41	36 / 0.40
0.10						521 / 0.46	257 / 0.45	158 / 0.45	51 / 0.43
0.2							2984 / 0.49	793 / 0.48	105 / 0.47
0.25								3424 / 0.49	159 / 0.49
0.30									256 / 0.49

Table 4.3.1.2

Sample Size (Per Group) n_2^* and value of ξ_2^* to reject $\eta(\xi_2^*) \geq \hat{\eta}_L(\xi_2^*)$ when $\overline{Y}_{\xi_2^*} \leq \hat{\eta}_Q(\xi_2^*)$. ($\xi_2^*$ is the value of ξ_2' that requires the smallest n_2' for this test)

$\xi_0 = 1 \quad \xi_2 = 10$

Significance probability = 0.05 (1 - tail test)

Cell entries are $\boxed{\begin{matrix} n_2^* & \\ & \xi_2^* \end{matrix}}$

$\overline{Y}_{\xi_0}$ \ $\overline{Y}_{\xi_2}$	0.010	0.015	0.02	0.05	0.10	0.20	0.25	0.30	0.50
0	2152 3.7	1433 3.7	1074 3.7	427 3.7	211 3.8	103 3.8	82 3.9	67 3.9	38 4.1
0.01		30951 5.3	8997 5.1	997 4.6	329 4.4	129 4.4	98 4.2	78 4.2	41 4.2
0.015			43582 5.3	1480 4.8	401 4.5	143 4.3	106 4.3	83 4.3	43 4.3
0.02				2234 5.0	487 4.7	158 4.4	115 4.4	87 4.4	45 4.4
0.05					1699 5.1	280 4.8	181 4.7	130 4.7	56 4.6
0.10						786 5.2	389 5.1	240 5.0	78 4.9
0.2							4465 5.4	1190 5.3	158 5.2
0.25								5124 5.4	238 5.3
0.50									384 5.4

4.3.4 Summary

We have considered some of the errors in extrapolation. Some of these errors are due to the variability of the data. A separate error component arises when we incorrectly specify the true model.

We have shown that even when we specify the wrong model, we can control random variability by dosing at high levels and using linear interpolation to estimate the differential response between doses at lower levels.

We have seen that an important error in linear interpolation is model misspecification. For a particular model, we have seen that the best choice for controlling variability and bias is to interpolate only slightly. When the model is correctly specified we need only control variability. In this case we are lead to dosing at as high a level as possible.

The class of models that contains these two extreme examples must be considered quite flexible in its ability to contain a function that approximates the true function. For any function in this class we can use a two-point design to obtain unbiased estimates of functions that bound the first function.

When these bounds are too wide for decision-making purposes we have shown how to design an experiment to distinguish between these bounds. This design suggests what dose level to use and the sample size requirements.

The second experiment may still involve some interpolation. In view of the fact that our assumption about the forms of the model may be too narrow we suggest computing similar bounds from this second experiment. If one believes that the true response is approximately linear at low dose, then there is reason to believe that at small enough dose the linear model will not be rejected.

Consideration of the fact that many observations are required to distinguish between bounds when $\bar{Y}_{\xi_2} - \bar{Y}_{\xi_0}$ is about 0.01 indicates that attempting to obtain experimental data at lower response levels may not be worthwhile.

CHAPTER V

FUTURE RESEARCH

One of the aims of this dissertation was to develop estimators that could attain smaller general mean square error than the least squares estimator.

Many of the estimators that were studied can attain smaller mean square error than the least squares estimator, but only for a subset of the parameter space. Thus, using these estimators because they have small mean square error requires that one know that the true parameters lie in this subset. Thus, prior knowledge is required in this case.

In many instances, only crude knowledge of certain bounds on the parameters is necessary to insure that the parameters lie in the required subspace. It was felt that there would be many instances where such knowledge would exist for many data sets.

However, in the cases we have investigated, when such bounds exist, they are usually very large and the incorporation of this information into the estimator would probably not be worth the effort of having to explain what one was doing.

In those instances where a smaller bound is known, one often finds that a transformation is used. For example, proportions are bounded by one, but one might employ linear model methods by transforming to the logits of these proportions. Since the logit is not bounded, there is no obvious choice for a bound on the logit in general.

This suggests that the use of transformations, which map a finite interval into an unbounded interval, may be viewed as a way of incorporating prior knowledge into the estimator.

Transformations also have an ability to reduce the terms in a linear model. For example, the log transformation can remove multiplicative interaction terms. Thus, transformations may be able to reduce the number of parameters to be estimated. However, the use of transformations has some disadvantages.

In the first place, using transformations of the dependent variable makes the interpretation of the parameters more complex. To reduce the complexity, one might try to make inferences about the untransformed variable from the estimates obtained in the transformed space. There are theoretical reasons why such inferences are invalid.

For example, if Y is lognormally distributed, that is

$$Y = \exp Z,$$

where $Z \sim N(\mu, \sigma^2)$ then

$$E(Y) = \exp\{\mu + \sigma^2/2\},$$

whereas

$$\exp\ E\ \ell n\ Y = \exp\ EZ = \exp\ \mu\ .$$

So, in general

$$E(Y) \neq f^{-1}\ E(f(Y)).$$

That is, one can not talk about how Y will behave on the average if one determines how $f(Y)$ behaves and then transforms back using f^{-1}.

In the case of the log transformation we see that such a procedure introduces bias. It might be of interest to conduct research to determine the mean square error of such a procedure, and compare it to some biased estimator.

We have investigated linear transformations of the independent variable. For example, if the true model is

$$\eta = \underset{\sim}{x}'\underset{\sim}{\beta}$$

and one fits

$$\eta = \underset{\sim}{z}'\underset{\sim}{a}$$

where $\underset{\sim}{z} = \underset{\sim}{M}\underset{\sim}{x}$, then the integrated bias error is minimized when

$$\underset{\sim}{a} = (\underset{\sim}{M}\underset{\sim}{W}\underset{\sim}{M}')^{-1}\underset{\sim}{M}\underset{\sim}{W}\hat{\underset{\sim}{\beta}} \ .$$

One might consider a $\underset{\sim}{z}$ defined as above when using principal component regression.

We have considered only estimators that are linear transformations of the least squares estimator. There is at least one application involving a fixed nonlinear transformation of the least squares estimator. This is the regression problem when there are errors in the measurement of the independent variables. When the variances of the measurement error in both variables are known, one solution in estimating the coefficients is to rotate the axis a certain amount, obtain the least squares estimator using the transformed variables and obtain the required estimator by transforming back. This required estimator is a least squares estimator using the transformed variables.

For this reason, it may be of interest to consider nonlinear transforms of the least squares estimator. If such a study were done, the nonlinear transform might be approximated by a linear transform.

In this case, it may be possible to apply some of the theory developed here.

LIST OF REFERENCES

Barnard, G.A. 1960. "The Logic of Least Squares." *Journal of the Royal Statistical Society* Series B25: 124-127.

Bhattacharya, P.K. 1966. "Estimating the mean of a multivariate normal population with general quadratic loss function." *Annals of Mathematical Statistics*, 37, 1819-1824.

Box, G.E.P. and N.R. Draper. 1959. "A Basis for the selection of a Response Surface Design. *Journal of the American Statistical Association* 54: 622-654.

Finkbeiner, D.T. 1960. *Introduction to Matrices and Linear Transformations.* W.H. Freeman and Company, San Francisco.

Finney, D.J. 1971. *Probit Analysis.* Cambridge.

Hammer, D.I., V. Hasselblad, B. Portnoy, and P.F. Wehrle. 1974. "Los Angeles Student Nurse Study." *Archives of Environmental Health* 28: 255-260.

Helms, R.W. 1974. "The Average Estimated Variance Criterion for the Selection-of-Variables Problem in General Linear Models." *Technometrics* 16: 261-74.

Hoerl, A.E. and R.W. Kennard. 1970a. "Biased Estimation for Non-orthogonal Problems." *Technometrics* 12: 55-67.

Hoerl, A.E. and R.W. Kennard, 1970b. "Ridge Regression: Applications to Non-orthogonal Problems." *Technometrics* 12: 69-82.

James, W. and C. Stein. 1961. Estimation with Quadratic loss. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1, 361-379. University of California Press, Berkeley and Los Angeles.

Karson, M.J. 1970. "Design Criterion for Minimum Bias Estimation of Response Surfaces." *Journal of the American Statistical Association* 65: 1565-1572.

Karson, M.J., A.R. Manson and R.J. Hader. 1969. "Minimum Bias Estimation and Experimental Design for Response Surfaces." *Technometrics*, 11: 561-475.

Kupper, L.L. and E.F. Meydrech. 1974. "Experimental Design Considerations Based on a New Approach to Mean Square Error Estimation of Response Surfaces." *Journal of the American Statistical Association* 69: 461-463.

Mallows, C.L. 1973. "Some Comments on C_p." *Technometrics* 15: 661-75.

Marquardt, D.W. and R.D. Snee. 1975. "Ridge Regression in Practice." *The American Statistician* 29: 3-20.

Mayer, L.S. and T.A. Willke.1973. "On Biased Estimation in Linear Models. *Technometrics* 15: 497-508.

Pope, P.T. and J.T. Webster. 1972. "The Use of an F-Statistic in Stepwise Regression Procedures." *Technometrics* 14: 327-340.

Rao, C.R. 1965. *Linear Statistical Inference and It's Applications.* John Wiley and Sons, New York.

Sclove, S.L. 1968. "Improved Estimators for Coefficients in Linear Regression." *Journal of the American Statistical Association* 63: 597-606.

Theil, H. 1971. *Principles of Econometrics.* North Holland Publishing Company, Amsterdam.

Wallace, T.D. 1964. "Efficiencies for Stepwise Regressions." *Journal of the American Statistical Association* 59: 1179-82.