A PERTURBATION/CONDITIONAL RISK APPROACH TO MEASURING
THE RELIABILITY OF THE AVERAGE ESTIMATED VARIANCE CRITERION

By

CHARLES WILLIAM ASH

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 941

JULY 1974
CHARLES WILLIAM ASH. A Perturbation/Conditional Risk Approach to Measuring the Reliability of the Average Estimated Variance Criterion (Under the direction of RONALD W. HELMS.)

In this dissertation an attempt is made to provide tools for evaluating the reliability of decisions based on the AEV criterion for model selection in the general linear model setting. The traditional distribution theory approach to such an evaluation is shown to be intractable due to the complex nature of the joint distribution of the AEV's. A perturbation/conditional risk approach is developed which utilizes the idea of perturbing the observed data and determining the proportion of decisions based on perturbed data which differ from the decision based on the original data. The proportion of changed decisions is shown to be a conditional risk function for an appropriately defined loss function. The approach is described and illustrated with an application to a simple problem of deciding which of two normal distributions has a larger mean. The approach is then applied to the more complicated problem of selecting the "best" of all possible regressions, where "best" is defined in terms of the AEV criterion. Computation of the conditional risk in the AEV situation is shown to be equivalent to evaluating the distribution function of an indeterminate, noncentral quadratic form in normal random variables, which can be evaluated by numerical application of the inversion theorem, a technique originated by Imhof. An algorithm is developed which can be used to perform the calculations with reasonable accuracy. Finally, several examples are given in the use of the conditional risk in conjunction with the AEV.
A PERTURBATION/CONDITIONAL RISK APPROACH
TO MEASURING THE RELIABILITY OF THE
AVERAGE ESTIMATED VARIANCE CRITERION

by

Charles W. Ash

A Dissertation submitted to the faculty of
the University of North Carolina in partial
fulfillment of the requirements for the
degree of Doctor of Philosophy in the
Department of Biostatistics.

Chapel Hill
1974

Approved by:

[Signatures]
Advisor
Reader
Reader
ACKNOWLEDGEMENTS

The author appreciates the assistance of his advisor, Dr. Ronald W. Helms, and of the other members of his advisory committee, Drs. L. L. Kupper, P. A. Lachenbruch, D. Quade, C. M. Suchindran and D. Willhoit. In particular, the author wishes to express his thanks to Dr. Helms for the invaluable aid in all phases of the production of this dissertation.

In addition the author would like to express appreciation for support under the NIEHS Grant 5-T01-ES00133 which enabled him to complete this research.

The author also wishes to thank his typist, Ms. Bea Parker for the typing of the final form of this dissertation.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF APPENDIX TABLES</td>
<td>viii</td>
</tr>
<tr>
<td>I INTRODUCTION AND REVIEW OF THE LITERATURE</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Notation and Assumptions</td>
<td>1</td>
</tr>
<tr>
<td>1.3 Review of the Literature</td>
<td>5</td>
</tr>
<tr>
<td>1.3.1 Criteria for Model Selection</td>
<td>5</td>
</tr>
<tr>
<td>1.3.1.1 The $R^2$ Statistic</td>
<td>6</td>
</tr>
<tr>
<td>1.3.1.2 The Partial F Statistic</td>
<td>6</td>
</tr>
<tr>
<td>1.3.1.3 Mallows' $C_p$ Statistic</td>
<td>8</td>
</tr>
<tr>
<td>1.3.1.4 Mean Square Error of Prediction</td>
<td>9</td>
</tr>
<tr>
<td>1.3.1.5 The Average Estimated Variance</td>
<td>9</td>
</tr>
<tr>
<td>1.3.2 Computational Algorithms for Model Selection</td>
<td>12</td>
</tr>
<tr>
<td>1.3.2.1 The General Sweep Algorithm</td>
<td>13</td>
</tr>
<tr>
<td>1.3.2.2 All Possible Regressions</td>
<td>14</td>
</tr>
<tr>
<td>1.3.2.3 Forward Selection and Backward Elimination</td>
<td>15</td>
</tr>
<tr>
<td>1.3.2.4 Stepwise Regression</td>
<td>16</td>
</tr>
<tr>
<td>1.3.3 Related Problems</td>
<td></td>
</tr>
<tr>
<td>CHAPTER</td>
<td>TITLE</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
</tr>
<tr>
<td>II</td>
<td>PROPERTIES OF THE JOINT DISTRIBUTION OF THE AEV'S FROM ALL POSSIBLE REGRESSIONS</td>
</tr>
<tr>
<td></td>
<td>2.1 Marginal Moments of Unordered AEV's</td>
</tr>
<tr>
<td></td>
<td>2.2 Characteristic Function of the Joint Distribution of the Unordered AEV's in the Case Where All Model Sizes Are Equal</td>
</tr>
<tr>
<td>III</td>
<td>PERTURBATION/CONDITIONAL RISK INFERENCE</td>
</tr>
<tr>
<td></td>
<td>3.1 Decision Theory Example and Notation</td>
</tr>
<tr>
<td></td>
<td>3.2 Perturbations and Conditional Risk</td>
</tr>
<tr>
<td></td>
<td>3.2.1 Perturbations</td>
</tr>
<tr>
<td></td>
<td>3.2.2 Conditional Risk</td>
</tr>
<tr>
<td></td>
<td>3.2.3 A Numerical Example</td>
</tr>
<tr>
<td></td>
<td>3.3 An Example Using the Uniform Distribution</td>
</tr>
<tr>
<td></td>
<td>3.4 Summary</td>
</tr>
<tr>
<td>IV</td>
<td>APPLICATION OF THE PERTURBATION/CONDITIONAL RISK APPROACH TO THE AEV CRITERION</td>
</tr>
<tr>
<td></td>
<td>4.1 Definition of the Problem; Notation</td>
</tr>
<tr>
<td></td>
<td>4.2 Evaluation of the Probabilities Arising from Questions Q1 and Q2</td>
</tr>
<tr>
<td></td>
<td>4.3 Evaluation of the Conditional Risk</td>
</tr>
<tr>
<td></td>
<td>4.4 Applications to Other Criteria</td>
</tr>
<tr>
<td>V</td>
<td>THE ALGORITHM USED FOR THE CALCULATION OF THE CONDITIONAL RISK</td>
</tr>
<tr>
<td></td>
<td>5.1 The Algorithm</td>
</tr>
<tr>
<td></td>
<td>5.2 Implementation of the Algorithm</td>
</tr>
<tr>
<td></td>
<td>5.3 Accuracy of the Algorithm</td>
</tr>
<tr>
<td>CHAPTER</td>
<td>PAGE</td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
</tr>
<tr>
<td>VI</td>
<td></td>
</tr>
<tr>
<td>SOME EXAMPLES ILLUSTRATING THE USE OF THE CONDITIONAL RISK</td>
<td>65</td>
</tr>
<tr>
<td>6.1 Application to the Hald Data</td>
<td>65</td>
</tr>
<tr>
<td>6.2 A Six Variable Example</td>
<td>71</td>
</tr>
<tr>
<td>VII</td>
<td></td>
</tr>
<tr>
<td>SUMMARY AND SUGGESTIONS FOR FUTURE RESEARCH</td>
<td>82</td>
</tr>
<tr>
<td>7.1 Summary and Discussion</td>
<td>82</td>
</tr>
<tr>
<td>7.2 Suggestions for Future Research</td>
<td>83</td>
</tr>
<tr>
<td>LIST OF REFERENCES</td>
<td>84</td>
</tr>
<tr>
<td>APPENDIX I</td>
<td>87</td>
</tr>
<tr>
<td>APPENDIX II</td>
<td>88</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Models for the Hald Data Ranked by AEV with the M Matrix Equal to ((X'X)/n)</td>
</tr>
<tr>
<td>6.2</td>
<td>Conditional Risks for the Hald Data and (M = (X'X)/n)</td>
</tr>
<tr>
<td>6.3</td>
<td>M Matrix for the Six Variable Example</td>
</tr>
<tr>
<td>6.4</td>
<td>Models for the Six Variable Example Ranked by AEV for (M = (X'X)/n)</td>
</tr>
<tr>
<td>6.5</td>
<td>Models for the Six Variable Example Ranked by AEV for (M) as in Table 6.3</td>
</tr>
<tr>
<td>6.6</td>
<td>Conditional Risks for the Six Variable Example and (M = (X'X)/n)</td>
</tr>
<tr>
<td>6.7</td>
<td>Conditional Risks for the Six Variable Example and (M) as in Table 6.3</td>
</tr>
<tr>
<td>6.8</td>
<td>(Pr[AEV_{1}(y) \geq AEV_{j}(y+D)]) for (M) as in Table 6.3</td>
</tr>
<tr>
<td>6.9</td>
<td>(Pr[AEV_{1}(y+D) \geq AEV_{j}(y)]) for (M) as in Table 6.3</td>
</tr>
<tr>
<td>FIGURE</td>
<td>PAGE</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>6.1 &quot;Contour lines&quot; representing sets of points where the joint density of $AEV_1(y+D)$ and $AEV_j(y+D)$ takes on equal values</td>
<td>69</td>
</tr>
</tbody>
</table>
# LIST OF APPENDIX TABLES

<table>
<thead>
<tr>
<th>APPENDIX TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I The Hald Data</td>
<td>87</td>
</tr>
<tr>
<td>II Data for the Six Variable Example</td>
<td>88</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION AND REVIEW OF THE LITERATURE

1.1 Introduction

Since the advent of multiple regression one of the problems faced by statisticians has been that of selecting a subset of independent variables which "best" explains the response. Many criteria for accomplishing this goal have been suggested, among them the Average Estimated Variance. In using this and other criteria, the underlying random variation makes it difficult to determine when two models are really different with respect to the values of the criterion. In this paper we derive an index of the reliability of the decision that two AEV's are different. This index is known as the conditional risk. Examples with data are also given to further clarify the use of the conditional risk.

Since this dissertation is set within the General Linear Models framework let us begin with a short review of some of the General Linear Hypothesis notation and assumptions.

1.2 Notation and Assumptions

This section contains most of the general notation and assumptions which will be used throughout the rest of this dissertation.
The general linear model of full rank is defined as

\[ \mathbf{Y} = X \mathbf{\beta} + \mathbf{E} \]

where

1. \( \mathbf{Y} \) is a \( n \times 1 \) random observed vector.
2. \( \mathbf{E} \) is a \( n \times 1 \) random unobservable vector such that

\[ E(\mathbf{E}) = 0 \text{ and } E(\mathbf{EE}') = \sigma^2 \mathbf{I}_n. \]
3. \( X \) is a known fixed \( n \times p \) matrix of rank \( p \).
4. \( \mathbf{\beta} \) is a \( p \times 1 \) vector of unknown fixed parameters.

No other assumptions about the distribution of \( \mathbf{E} \) will be made unless specifically stated.

Under these assumptions the least squares and best (minimum variance) linear unbiased estimator of \( \mathbf{\beta} \) and the corresponding unbiased estimate of \( \sigma^2 \) are known to be:

\[ \mathbf{b} = (XX')^{-1}X'\mathbf{Y} \quad \text{and} \quad s^2 = (\mathbf{Y}-X\mathbf{b})'(\mathbf{Y}-X\mathbf{b})/n-p. \]

Also, if \( \mathbf{x} \) is any row vector in the space spanned by the rows of \( X \), i.e. \( \mathbf{x} \) is \( 1 \times p \), then the best linear unbiased estimator of \( E[Y(x)] \) is \( \hat{Y} = \mathbf{x} \mathbf{b} \) where

\[ Y(x) = \mathbf{x} \mathbf{\beta} . \]

Consider fitting an incomplete model. Let \( Z_{i} \) be the \( n \times p_{i} \) matrix containing the \( i^{th} \) set of columns of \( \mathbf{x} \) where \( i = 1, \ldots, 2^{p}-1 \) in some ordering. Thus the \( i \)'s correspond to the \( 2^{p}-1 \) possible submodels for a given matrix \( X \) of rank \( p \), \( p_{i} \) is the number of columns in the \( i^{th} \) set. Let \( Z_{ic} \) be the \( n \times (p-p_{i}) \) matrix containing the columns of \( X \) not in \( Z_{i} \). Define \( A_{i} \) as:

\[ A_{i} = \mathbf{I}_n - Z_{i}(Z_{i}'Z_{i})^{-1}Z_{i}' . \]
One can partition the general linear model as

\[ Y = (Z_i^1, Z_i^c) \begin{bmatrix} \beta_i^1 \\ \beta_i^c \end{bmatrix} + \epsilon \]

or

\[ Y = (Z_i^1\beta_i^1 + Z_i^c\beta_i^c) + \epsilon. \]

Now assume also that one fits only the model \( Y = Z_i^1\beta_i^1 + \epsilon. \)

Then the least squares estimator of \( \beta_i^1 \), based on this submodel, is

\[ b_i^1 = (Z_i^1Z_i^1)^{-1}Z_i^1Y; \quad \hat{\beta}_i^1 = b_i^1 + (Z_i^1Z_i^1)^{-1}Z_i^1Z_{ic}\hat{\beta}_{ic}. \]

The corresponding estimator of \( \sigma^2 \) is

\[ s_i^2 = (Y-Z_i^1b_i^1)'(Y-Z_i^1b_i^1)/(n-p_i), \]

with

\[ E(s_i^2) = \sigma^2 + (Z_{ic}\beta_{ic})'A_i(Z_{ic}\beta_{ic})/(n-p_i). \]

The residual sum of squares for the \( i \)th model is defined as:

\[ \text{RSS}_i = (Y-Z_i^1b_i^1)'(Y-Z_i^1b_i^1) \]

and the least squares estimator for \( E[Y(x)] \) for the incomplete model is \( \hat{Y}(z_i^1) = z_i^1b_i^1 \) where \( z_i^1 \) is any column vector in the space spanned by \( Z_i^1 \).

One can define \( R^2 \) as the proportion of the total variance "explained by the model", i.e.,

\[ R^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \text{RSS}}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}. \]

The Integrated Mean Square Error (IMSE) is a parametric index of the average closeness of fit of a fitted model to a "true" or
assumed model (Helms, 1969). The IMSE can be defined as follows.

Let $R$ be a region of interest (for values of $x$) over which one wishes to predict the response $Y(x)$ and let $w(x)$ be a weight function over $R$ which has the properties of a multivariate probability density function (i.e. $w(x) \geq 0$, $\int_R w(x) \, dx = 1$). The the Integrated Mean Square Error of the estimator $\hat{Y}(x)$ for $Y(x)$ is defined as

$$\text{IMSE}(\hat{Y}(x)) = \int_R E[(\hat{Y}(x) - E(Y(x))]^2 w(x) \, dx.$$

The IMSE can easily be partitioned into integrated variance and integrated squared bias components:

$$\text{IMSE}(\hat{Y}(x)) = \int_R \text{Var}(\hat{Y}(x)) w(x) \, dx + \int_R (E(Y(x)) - (\hat{Y}(x)))^2 w(x) \, dx.$$

Let $I$ be a subset of subscripts from the set $\{1, 2, \ldots, p\}$, let $J = I \cup \{j\}$ where $j \notin I$, and define $R_{J|I}$ as the reduction in the residual sum of squares obtained by adding variable $j$ to the model given that the variables included in $I$ are those in the model. Then the "partial $F$ for addition" of the variable number $j$ given the variables specified by $I$ is

$$F_{\text{addition}} = (R_{J|I}(n-p_I-1))/(RSS-R_{J|I})$$

where $p_I$ is the number of variables in the model. The "partial $F$ for deletion" of variable number from the model specified by $J$ is

$$F_{\text{deletion}} = (b_j)^2/(RSS\cdot c_{jj})$$

where $c_{jj}$ is the $jj^{th}$ element of $(Z'_jZ_j)^{-1}$. 

1.3 Review of the Literature

The variable selection problem ("choice of model" problem) in regression is one of selecting that subset of independent variables which is in some sense the "best". We define the "best" subset as the one which has the optimal value for a specified criterion. In actual practice one chooses a small number of models for which the value of the criterion is nearly optimal and assumes that these models are equivalent.

The choice of the model that one will use is then made on the basis of further study of these models and things such as cost, physical meaning, ease in explanation and interpretation, etc. This section contains a discussion of previous work on algorithms and criteria used in model selection and some related problems.

1.3.1 Criteria for Model Selection

In this section we will be discussing criteria for model selection. These criteria should be differentiated from the algorithms which are used in model selection. For instance, stepwise regression is an algorithm which may be used with several of the criteria mentioned in this section (i.e., the partial F statistic, the A&E, etc.). The various algorithms will be discussed in the next section.

Some criteria for model selection which have been widely used by statisticians for selection of variables are

1. $R^2$
2. The Partial F Statistic
3. Mallows Cp Statistic
4. The Mean Square Error of Prediction
5. The Average Estimated Variance.
Expositions on the use of these and other criteria for model selection are given in a number of papers and texts, including Draper and Smith (1966), Hocking (1972), Helms (1971), and Daniel and Wood (1971). The use of these criteria are briefly summarized in the following sections.

1.3.1.1 The $R^2$ Statistic

The $R^2$ statistic can be interpreted as the proportion of the total variance "explained by the model" ($R^2$ is defined in section 1.2). In the past the $R^2$ statistic has been used primarily as a means to select a subset of models for more intensive study (those having large $R^2$).

The use of the $R^2$ statistic is summarized in Chapter 6 of Draper and Smith (1966). One examines $R^2$ for all models of interest and chooses the model containing the fewest variables which has a "satisfactorily high" value of $R^2$. For instance, if by adding a variable to the model, the $R^2$ is increased from .995 to .997, one would select the model with fewer variables. What one means by "satisfactorily" high is very subjective and varies from problem to problem and statistician to statistician.

1.3.1.2 The Partial $F$ Statistic

The partial $F$ statistic (defined in section 2) is used in the following manner. Partial $F$ values are calculated and compared with preselected "critical" $F$ values to determine if the step of adding or deleting a variable reduces the error sum of squares by a "significant" amount. In the forward selection algorithm the variable under consideration for possible addition to the model is
the one which has the highest partial F. This is also the variable which maximizes the reduction in the error sum of squares or has the highest partial correlation with the dependent variable. In the backward elimination algorithm that variable already in the model which has the minimum partial F is deleted if the computed partial F is less than the preselected "F to delete" value. Stepwise regression is a combination of the forward selection and backward elimination algorithms (all three are described in Chapter 6 of Draper and Smith (1966)). The stepwise regression algorithm proceeds as follows. At each step of the algorithm one first checks to see if a variable may be deleted according to the backward elimination procedure. If no variable may be deleted, then the process described above for forward selection is used. One stops when no variables are deleted or added. Note that one must specify both an "F to enter" and an "F to remove" in stepwise regression. One drawback to the use of preselected critical F values is that as the number of variables changes so does the error degrees of freedom and therefore the significance level at which one is operating. This problem has been eliminated in the REGR procedure of the Statistical Analysis System of North Carolina State University. In that regression procedure the user specifies a formal significance level for entry and removal and the "critical values" of the F's to enter and remove are recalculated at each step. This is a problem only if there are few error degrees of freedom.

Stepwise regression using the partial F statistic is intuitively appealing as it includes in the model only those variables which "significantly" reduce the error sum of squares, given the other
variables in the model. However, the meaning of "significance" is not clear here because at best one is working with doubly non-central $F$ distributions (Pope and Webster, 1972) but computing "significance" from central $F$ distributions, and in part because the probability structure of the problem is so complex. Also, this process may include in the model variables with coefficients not significantly different from zero and it may not choose models which are better with respect to some other criteria (e.g., the integrated mean square error of $\hat{y}(x)$).

1.3.1.3 Mallows' $C_p$ Statistic

Within the past few years several other criteria for model selection have been proposed. One of these was suggested by Mallows and presented in a paper by Gorman and Toman (1966). Define

$$\Gamma_p = \frac{1}{s^2} \left[ \sum_i (v_i - \eta_i)^2 + \sum_i \text{var}(\hat{Y}(x)_i) \right]$$

where $v_i = E(Y_i)$ and $\eta_i = E(\hat{Y}(x)_i)$. One can see that $\Gamma_p$ is the Integrated Mean Square Error for a uniform weight function defined on the experimental points. It can be shown that

$$\Gamma_p = E(\text{RSS})/s^2 - (N-2p).$$

The estimator of $\Gamma_p$ proposed by Mallows is

$$C_p = \text{RSS}/s^2 - (N-2p).$$

Gorman and Toman suggested that the $s^2$ computed from the fit of the full model be used as an unbiased estimate of $\sigma^2$ in computing $C_p$. The article explains the use of both graphical and numerical procedures for using the $C_p$ as a criterion for model selection.
Although the proposed procedure is somewhat subjective, it is often equivalent to choosing the model with the minimum value of \( C_p \) where \( p \) is the number of variables in the model. This criterion includes components for both the variance and bias of \( \hat{Y}(x) \) at the experimental points (the \( x_i \)) in selecting a model.

1.3.1.4 Mean Square Error of Prediction

Allen (1971) introduced a criterion called the Mean Square Error of Prediction (MSEP) defined as

\[
\text{MSEP}(\hat{Y}(x)) = E[(Y - \hat{Y}(x))^2] = \sigma^2 + \text{Var}(\hat{Y}(x)) + [E(\hat{Y}(x)) - x_\beta]^2
\]

for a particular value of \( x \).

He proposes that the model with the minimum MSEP at a given point \( x \) is the "best" model for predicting \( Y \) at that point and derives a procedure for selecting the model with the minimum estimated MSEP. One difficulty with this approach is that at different points (\( x \)-values) for which one wishes to predict the value of \( Y \) one might have to use different models.

1.3.1.5 The Average Estimated Variance

Helms (1972) has generalized Allen's method to a criterion named the Average Estimated Variance (AEV). The AEV is defined in the following paragraphs but the basic idea is that for a given model one computes the estimated variance at each point \( x \) in the region of interest and then averages the variance over \( R \) with respect to the weight function, \( w(x) \), discussed earlier in the section on the Integrated Mean Square Error. Since the estimated
variance contains components from both bias and variance, the AEV will be inflated by either a large variance, a large bias or a combination of both. The definition of the AEV follows.

Given the usual general linear model of full rank, the variance of the usual full model estimator of $Y$ at the point $x$ is

$$\text{Var}(\hat{Y}(x)) = \sigma^2 x(x'x)^{-1}x'$$

Assume the weight function $w(x)$ and the region of interest, $R$, are as defined in section 1.2. The weighted average of the variance of $\hat{Y}(x)$ over $R$ is:

$$\text{Average variance of } \hat{Y}(x) = \sigma^2 \int_R x(x'x)^{-1}x'w(x)dx$$

which is shown to be (Graybill (1969), Chapter 9, Theorem 9.1.20)

$$\sigma^2 \text{tr}[(X'X)^{-1}M]$$

where

$$M = \int_R x'x w(x)dx .$$

In order to evaluate this variance one must know $\sigma^2$ or estimate it. Helms has defined the Average Estimated Variance as

$$\text{AEV} = s^2 \text{tr}[(X'X)^{-1}M]$$

where $s^2$ is as defined in section 1.2. If one is dealing with an incomplete model as described in section 1.2 then the AEV is

$$\text{AEV}_{1} = s_1^2 \text{tr}[(Z_1'Z_1)^{-1}M_{11}]$$

and

$$E(\text{AEV}_{1}) = \text{tr}[(Z_1'Z_1)^{-1}M_{11}]*[\sigma^2 + (Z_2'\theta_i \text{ic} A_1(Z_1'\theta_i \text{ic} /\mu - p_i])$$

where $M_{11}$ is the submatrix of $M$ which corresponds to the variables in the model. In the following pages we will use the notation
$T_i = \text{tr}[(Z_i'Z_i)^{-1}M_{ii}]$ for the sake of convenience.

The A\(\text{E}\)V will be inflated by any one or a combination of the following

(a) a high inherent variability ($\sigma^2$)
(b) the omission of important terms ($\beta_i \neq 0$)
(c) the trace of $(Z_i'Z_i)^{-1}M_{ii}$ being large (large variance with respect to the weight function and region of interest).

One uses the A\(\text{E}\)V by selecting for consideration a subset of models which have small A\(\text{E}\)V's. (The actual model chosen for use may be the one with the smallest A\(\text{E}\)V or it may be selected from the subset on the basis of other criterion.) Nothing can be done about $\sigma^2$ so the only possibilities are to reduce the effects of (b) and (c). If one has a choice of $X$ (i.e. one can design the experiment) and $M$ is known, one could design the experiment to minimize $\text{tr}(X'X)^{-1}M$. The only other possibility is to choose the submodel which minimizes the effects of both (b) and (c). One should note that the model which produces the minimum $E(A\text{E}V)$ may not yield an unbiased estimator of $\beta_i$, i.e. the model which includes all variables with $\beta_i \neq 0$. Consider the specific example $M = (X'X)/n$. In this

---

1The choice $M = (X'X)/n$ is a reasonable choice under several circumstances as Helms (1972) has noted. First, if the experiment has been designed, then the experimenter has reflected his region of interest and his weight function in the design of the experiment. Using a "discrete" weight function which takes the value $1/n$ at the experimental points and zero elsewhere (yielding $M = (X'X)/n$) may be quite reasonable. If, on the other hand, the experiment consists of random observations of $x$ and $y$ then $(X'X)/n$ again may be a good choice for $M$, as the observed points should reflect the distribution of $x$. 

---
special case one can look at the effect of adding a single variable
to the model $y = Z_1^\beta + e$. For this model the expected value of the
AEV is

$$E(\text{AEV}_1) = \frac{p_1}{n(\sigma^2 + Q_1)}$$

where $Q_1 = (Z_{i1c}^\beta Z_{i1c})'A_1(Z_{i1c}^\beta Z_{i1c})/(n-p_1)$.

The expected value of the
AEV for the model containing one additional variable, say model $j$, is

$$E(\text{AEV}_j) = \frac{p_1+1}{n} \left[ \sigma^2 + (Q_1 - \delta) \right]$$

where $\delta$ is the change in $Q_1$ due to fitting the additional variable.

The difference in the two expectations is

$$E(\text{AEV}_j) - E(\text{AEV}_1) = \sigma^2/n + Q_1/n - \delta(p_1+1)/n.$$

This of course is the expected value of the change (positive or
negative) in the AEV due to adding the extra variable to the sub-
model. If $\delta > (\sigma^2 + Q_1)/(p_1+1)$ the expected value of the AEV is re-
duced and we add the new variable to our model. This will not
always be the case and therefore the "best" model will not always
yield unbiased estimates of $\beta$.

1.3.2 Computational Algorithms for Model Selection

In this section we will consider algorithms which have been
used in model selection procedures. It is worth noting, however,
that the general algorithm defined in 1.3.2.1 has application to a
wide class of problems, including model selection. Four common
algorithms widely used for attacking the selection-of-variables
problem and based on the general algorithm are described in Chapter 6
of Draper and Smith (1966) and by E. M. L. Beale (1970):
1. All Possible Regressions
2. Forward Selection
3. Backward Elimination
4. Stepwise Regression.

These algorithms are described in sections 1.3.2.2 - 1.3.2.4.

1.3.2.1 The General Sweep Algorithm

Efroymson (1960) proposed the stepwise regression procedure, a modified form of Gaussian elimination as a way of attacking the selection-of-variables problem. In essence the procedure transforms the matrix

\[
\begin{pmatrix}
X'X & X'Y \\
- & - & - & - & 1 & - & - \\
Y'X & Y'Y & 0
\end{pmatrix}
\]

into the matrix

\[
\begin{pmatrix}
I & b & (X'X)^{-1} \\
- & - & - & - & - & - & - \\
0 & RSS & -b'
\end{pmatrix}
\]

by elementary row operations.

The algorithm proceeds in steps, operating on one column of \((X'X)\) at a time. At the point at which the algorithm is operating on the \(i\)th column (corresponding to the \(i\)th independent variable) the \((1,1)\)-element is known as the pivot. Also, in order to save computer memory the algorithm is programmed so the elements of the \(i\)th column of \((X'X)^{-1}\) and \(-b'\) are inserted in place of the column of \((X'X)\) which has been reduced to the identity matrix.

Execution of one step of the algorithm is equivalent to executing a sweep operator similar to that attributed to Beaton as
described by Dempster (1969). A p x p matrix A is said to have been swept on the (k,k) element if A is replaced by a matrix B whose elements are defined as follows:

\[ b_{ik} = \frac{a_{ik}}{a_{kk}} \]
\[ b_{kj} = \frac{a_{kj}}{a_{kk}} \]
\[ b_{ij} = a_{ij} - \frac{a_{kj}a_{ik}}{a_{kk}} \]
\[ b_{kk} = 1/a_{kk} \]

B is denoted by \( \text{SWP}(k)A \). Successive sweeps on the elements \((k_1, k_2, \ldots, k_p)\) is denoted by \( \text{SWP}(k_1, k_2, \ldots, k_p)A \). One should note that the sweep is reversible and that sweeping on the \( i^{th} \) and then the \( j^{th} \) element is equivalent to sweeping in the reverse order.

Dempster (1969) derives a number of properties of the properties of \( \text{SWP} \). Given the matrix A,

\[
A = \begin{pmatrix}
X'X & X'Y \\
- & - & - \\
Y'X & Y'Y
\end{pmatrix}
\]

then

\[
\text{SWP}(k_1, k_2, \ldots, k_p)A = \begin{pmatrix}
(X'X)^{-1} & b \\
- & - \\
-b' & RSS
\end{pmatrix}.
\]

1.3.2.2 All Possible Regressions

In examining all possible regressions of \( p \) independent variables one fits each of the \( 2^p - 1 \) possible models and evaluates the chosen criterion for each model. Garside (1965) described an efficient algorithm for fitting all possible models. The algorithm operates in steps and has the advantage of adding or deleting only
one variable at each step, as follows. The $i^{th}$ variable is used as the pivot in the sweep procedure for the first time on the $2^{i-1}$ step and at each $2^i$ subsequent steps. For example, if $p = 3$ variables, the pivots and the variables in the model are as follows.

<table>
<thead>
<tr>
<th>Step</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pivot</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Variables in the Model</td>
<td>1</td>
<td>1,2</td>
<td>2</td>
<td>2,3</td>
<td>1,2,3</td>
<td>1,3</td>
<td>3</td>
</tr>
</tbody>
</table>

The major advantage of fitting all possible regressions is that one is always able to obtain the "best" fit, as all cases are enumerated. One can also use this technique to select a subset of models to study in more detail. The major disadvantage is that unless the number of independent variables is relatively small the computational problem is huge. Gorman and Toman (1966) suggested using fractional factorial designs in order to reduce the amount of computation.

1.3.2.3 **Forward Selection and Backward Elimination**

The forward selection algorithm starts with no variables in the model and proceeds to add variables in a step by step manner. These steps consist of three elements.

1. Evaluate the selected criterion for each candidate variables for entry into the model.

2. Check to see if a stopping point has been reached.

3. Perform the sweep operation on the variable whose addition to the model produces the most nearly optimum value of the criterion.

These elements are repeated until a suitable stopping point, which depends on the criterion, has been reached.
Backward elimination is similar except that one starts with all the variables in the model and proceeds to delete variables in a step by step manner. The elements of the backward elimination algorithm are:

1. Evaluate the selected criterion for each of the candidate variables for deletion from the model.
2. Check to see if a stopping point has been reached.
3. Perform the sweep operation on the variable whose deletion from the model produces the most nearly optimal value of the criterion.

The major disadvantage of these two procedures is that they do not necessarily lead to the "best" fit, whatever the criterion, and frequently do not lead to as good a fit as stepwise regression. Also, they may not result in the same model. This is partially due to the fact that there may be functional relationships between the independent variables. Suppose, for example, that a variable is added to the model which has a functional relationship with a variable which has previously entered the model.

In this case the "best" model may be one which does not contain the variable which first entered the model. However, in forward selection a variable cannot be removed once it is in the model. A similar situation can occur in backward elimination when a discarded variable becomes important due to the elimination of a variable to which it is functionally related, but once a variable is eliminated it cannot re-enter the model.

1.3.2.4 Stepwise Regression

Stepwise regression, suggested by Efroymson (1960), partially eliminates the problems of forward selection and backward elimination.
One starts with no variables in the model and proceeds in a step by step manner. Each step consists of three elements:

1. Evaluate the selected criterion for every variable both in and out of the model.

2. Check to see if a stopping point has been reached. If so, stop.

3. Perform the sweep operation on the variable which, by addition to or deletion from the model, produces the most nearly optimal value of the criterion.

These three elements constitute a step; steps are repeated until a suitable stopping point has been reached.

1.3.3 Related Problems

There have been some other related but not directly applicable proposals. Garside (1965) suggested that, given a fixed number of variables in the model, the "best" model is the one which minimizes the residual mean square. However, this is equivalent to maximizing $R^2$ or minimizing $C_p$ and really is a somewhat different problem from the general selection-of-variables problem.

Lindley (1968) gives a Bayesian approach to a similar problem. He wishes to choose variables in a way which minimizes the expected loss described below. Let $I$ denote a subset of subscripts for potential variables (for example one might have $I = \{1,4,5,7,19\}$). Let $L(I)$ denote the loss associated with selecting the model with the variables specified by $I$. Assume

$L(I) = (Y - \hat{y}(x_I))^2 + C_I$ where the vector $x_I$ contains the variables in the model and $C_I$ is the cost of observing the $x_I$. Let $J$ denote a set of subscripts not in $I$ (i.e., $I \cap J = \emptyset$). In the usual
regression set up and assuming vague priors Lindley shows that

\[
\bar{E}(L(I)) = \sigma^2 (1 + p/n) + \min_{I}(R(J|I)n^{-1} + CI)
\]

where \( R(J|I) \) is the reduction in the error sum of squares due to adding the \( x_j \) to the model given that the \( x_I \) are already in the model. Thus, if \( C_I = \sum_{I} C_I \) one adds an \( x_j \) to the model if the reduction in the residual sum of squares is greater than \( nC_j \). While this is an interesting problem and solution it is also somewhat different than the general selection of variables problem.
CHAPTER 11

PROPERTIES OF THE JOINT DISTRIBUTION OF THE AEV'S FROM ALL POSSIBLE REGRESSIONS

One of the basic purposes of this dissertation is to attack the problems of selecting the "best" submodel from all $2^p-1$ possible submodels by the AEV criterion and to determine whether one can reliably claim that the model with the smallest observed AEV is a better model, i.e., has a smaller expected AEV, than other models. A traditional approach to this type of problem would be to determine the joint distribution of the ordered AEV's from all possible submodels, say $AEV_1, AEV_2, \ldots, AEV_k$ where $k = 2^p-1$, and test the hypothesis $H_0: \bar{E}(AEV_1) = \min_{1 \leq j \leq k} \bar{E}(AEV_j)$ versus the alternative $H_A: \bar{E}(AEV_2) > \min_{1 \leq j \leq k} \bar{E}(AEV_j)$. In other words we wish to know if the model producing the minimum observed AEV is the model which has the minimum expected AEV.

In order to test such a hypothesis one must have available the joint distribution of $AEV_1, AEV_2, \ldots, AEV_k$, $k = 2^p-1$ and then the joint distribution of the ordered AEV's: $AEV_1, \ldots, AEV_k$.

Although these distribution theory problems have proven to be intractable, the following potentially useful properties of the joint distribution of all possible AEV's have been derived and are included in this chapter: the (marginal) expectation and variance of $AEV_i$ (where the subscript $i$ denotes the AEV for the $i^{th}$ submodel in a list
of the $2^p-1$ submodels). The (marginal) covariance, $\text{cov}(AEV_i, AEV_j)$, is displayed, and the characteristic function of the joint distribution of the unordered $AEV$'s for models of the same size is derived.

2.1 Marginal Moments of Unordered $AEV$'s

The marginal expected value and variance of unordered $AEV$'s are displayed in this section and the marginal covariance between two unordered $AEV$'s arising from different submodels is derived using a result in Searle (1971).

The marginal expected value of a given $AEV$, say $AEV_i$ from the model defined by $Z_i$, is given in Chapter I as

$$E(AEV_i) = T_i(\sigma^2+1/(n-p_i))A_i(\bar{X}_i)$$

where

$$T_i = \text{tr}[(Z_i'Z_i)^{-1}M_{ii}] \quad \text{and}$$

$$A_i = Z_i(Z_i'Z_i)^{-1}Z_i'.$$

Note that $(\bar{X}_i)'A_i(\bar{X}_i) = (Z_i\beta_{ic})'A_i(Z_i\beta_{ic}).$

Under the assumption that $\gamma^N(\mu, \Sigma)$ Searle (1971) on page 66 shows that

$$\text{Var}(Y'BY) = 2\text{tr}(BV)^2 + 4\mu'BVB\mu.$$

Now the $AEV$ for the $i^{th}$ model is

$$AEV_i = [T_i/(n-p_i)]Y'A_iY \quad \text{where} \quad \gamma^N(\bar{X}_i, \sigma^2I).$$

Thus the marginal variance of the $AEV$ from the $i^{th}$ model is

$$\text{Var}(AEV_i) = [T_i/(n-p_i)]^2[2\sigma^2\text{tr}(A_i^2) + 4\sigma^2(\bar{X}_i)'A_i(\bar{X}_i)]$$

or

$$\text{Var}(AEV_i) = \frac{2T_i^2\sigma^2}{n-p_i} \left[\sigma^2 + \frac{2}{n-p_i} (\bar{X}_i)'A_i(\bar{X}_i)\right].$$
Searle (1971, p. 66) also gives a form for the covariance of two quadratic forms:

\[ \text{Cov}(Y'BY, Y'CY) = 2 \text{Tr}(BCV) + 4\mu'(BCV)\mu. \]

Applying this to the AEV's computed from submodels defined by \( Z_i \) and \( Z_j \) yields the marginal covariance

\[ \text{Cov}(\text{AEV}_i, \text{AEV}_j) = \frac{2T_i T_j \sigma^2}{(n-p_i)(n-p_j)} \left[ \sigma^2 \text{Tr}(A_i A_j) + (X)'A_i A_j (X) \right]. \]

2.2 Characteristic Function of the Joint Distribution of the Unordered AEV's in the Case Where All Model Sizes Are Equal

This section contains a derivation of the marginal distribution of a given AEV and a derivation of the characteristic function of the joint distribution of the unordered AEV's for the case where all of the models are of a given size.

The marginal distribution of a single AEV, say \( \text{AEV}_i \), for the model defined by \( Z_i \) under the assumption that \( E \sim N(0, \sigma^2 I) \) may be obtained by using corollary 2.4 on page 58 of Searle (1971):

\[ W \sim N(\mu, I) \text{ implies } W'BW \sim \chi^2(r, \mu', B\mu/2) \]

if \( B \) is idempotent of rank \( r \). With the notation for \( T_i \) and \( A_i \) given in the previous section,

\[ \text{AEV}_i = S_i^2 T_i = T_i Y'A_i Y/(n-p_i), \]
so that
\[ \text{AEV}_i (n-p_i)/T_i \sigma^2 = (\mathbf{Y}/\sigma)'A_i (\mathbf{Y}/\sigma). \]

Since \( A_i \) is idempotent (this is well known) and \( \mathbf{Y}/\sigma \sim N(\mathbf{X}\beta_i/\sigma, I) \)
by the Searle (1971) corollary 2.4
\[ \text{AEV}_i (n-p_i)/T_i \sigma^2 \sim \chi^2 (n-p_i, (\mathbf{X}\beta_i)'A_i (\mathbf{X}\beta_i)/2\sigma^2). \]

Thus the marginal distribution of any one AEV is proportional to a
noncentral \( \chi^2 \) distribution. As noted elsewhere, the noncentrality
parameter is a function of the coefficients (\( \beta_i \)) associated with
variables not in the model, i.e. with variables in \( Z_{ic} \), because
\[ (\mathbf{X}\beta_i)'A_i (\mathbf{X}\beta_i) = (Z_{ic}\beta_{ic})'(I-Z_{ic} (Z_{ic}Z_{ic})^{-1}Z_{ic})(Z_{ic}\beta_{ic}). \]

Although the marginal distribution of a pre-specified \( \text{AEV}_i \)
is interesting, the real goal would be the joint distribution of all
possible \( \text{AEV}'s \). This problem is very similar to the problem of
finding the joint distribution of all possible \( \sigma_i^2 \) or the residual
sum of squares about the \( i^{th} \) model, \( \text{RSS}_i \). A search of the statistical
literature has not revealed a solution to this problem. However,
Arveson and McCabe (1972) have derived the characteristic function of
the joint distribution of the unordered residual sum of squares for
the situation in which the number of variables in the model is pre-
specified. Their derivation may be summarized as follows. Assume
\( \bar{\mathbf{Y}} = X \bar{\beta} + \mathbf{E} \) and \( \bar{\mathbf{Y}} \sim N(\mathbf{X}\bar{\beta}, \sigma^2 I) \). Then the residual sum of
squares for the \( i^{th} \) model is \( \text{RSS}_i = \bar{Y}'A_i \bar{Y} \). Then they assume without
loss of generality that \( \sigma^2 = 1 \). Now one may write \( \bar{Y}'A_i \bar{Y} \) as \( U_i'U_i \)
where \( U_i = B_i \bar{Y}, B_i'B_i = A_i \) and \( B_i'B_i = \mathbf{I} \) where \( B_i \) is a \( (n-p_i) \times n \)
matrix, and \( p_i \) is the prespecified number of variables in the model.
The joint distribution of \( \mathbf{U}' = (U_1', U_2', \ldots, U_r') \) where \( \mathbf{r} = \binom{p}{p_1} \) is multivariate normal with mean vector \( (\eta_1', \ldots, \eta_r') \) with \( \eta_i = B_i' \mathbf{x} \) and covariance matrix \( \Sigma = (\Sigma_{ij}) \) where \( \Sigma_{ij} = B_i' B_j' \). Note that the \( r(n-p_1) \times r(n-p_1) \) matrix \( \Sigma \) may be singular. Let \( \phi(t_1, \ldots, t_r) = \phi(t) \) denote the characteristic function of the joint distribution of \( (RSS_1, RSS_2, \ldots, RSS_r) \) and let \( T = \text{diag}(t_1, t_2, \ldots, t_r) \otimes I_{n-p_1} \).

Arveson and McCabe show
\[
\phi(t) = |I - i t T|^{-1/2} \exp\{i/2[n'T(I - i t T)^{-1} \eta]\}.
\]

Now since \( \text{AEV}_j = K_j RSS_j \) where \( K_j = T_j'(n-p_j) \), the characteristic function of the joint distribution of the AEV's in the case where the model size is constant and equal to \( p_j \) is given by
\[
\phi(t) = |I - i K_j t T|^{-1/2} \exp\{i/2[n'K_j T(I - i K_j t T)^{-1} \eta]\}.
\]

Arveson and McCabe (1973) state that this form of the characteristic function is not useful in practical situations and resort to recommending a simulation in order to determine the properties of the joint distribution.

Our problem is even more difficult than the one that Arveson and McCabe considered as we do not restrict ourselves to a given number of variables and we want the distribution of the ordered AEV's. At the time it appears that the traditional approach to the problem through distribution theory is intractable.
CHAPTER III

PERTURBATION/CONDITIONAL RISK INFERENCE

In practice when one computes the AEV's for a number of possible models the usual decision rule is to select for use the model with the smallest observed AEV. The question immediately arises: How reliable is the choice? That is, is the difference between two AEV's large enough that one can confidently claim that the expected value of the smaller AEV is smaller than the expected value of the larger AEV? One would like to have an index or indicator of the reliability of the formal decision rule: select for use the model with the smallest observed AEV.

The purpose of this chapter is to illustrate, with a simple example, a technique for constructing an index of reliability of a decision rule when the stochastic framework is complicated, as with the AEV. A secondary purpose of the chapter is to establish notation and terms which will be used in the next chapter in which the technique is applied to derive an index of reliability for the AEV decision rule.

3.1 Decision Theory Example and Notation

In this section we will establish standard decision theory notation which will be used throughout the rest of the chapter. It should be noted that the example presented has been solved by standard
techniques and is included here solely for purposes of illustration.

Consider the following decision problem formulated in standard decision theory notation (see e.g., Lindgren, 1965).

Let
\[
U \sim N_{n_1 + n_2}^{(n_1 + n_2) \times 1} \left[ \begin{array}{c} \mu_1 \frac{1}{n_1} \\ \mu_2 \frac{1}{n_2} \\ \frac{1}{n_1 + n_2} \end{array} \right];
\]

let \( U_i \) denote the \( i \)th element of \( U \);

\( \mu_1, \mu_2 \) be real, unknown parameters;

and consider two states of nature which are of interest:

State 1: \( \mu_1 < \mu_2 \)

State 2: \( \mu_1 \geq \mu_2 \).

One will usually also be interested in the magnitude of the difference \( \mu_1 - \mu_2 \), but the problem here is to decide which \( \mu \) is smaller.

Consider two possible actions or decisions:

\( a_1 \): conclude state 1 holds, i.e., \( \mu_1 < \mu_2 \)

\( a_2 \): conclude state 2 holds, i.e., \( \mu_1 \geq \mu_2 \).

We define the decision function (or "action" function)

\[
a(\bar{U}_1 - \bar{U}_2) = \begin{cases} 
  a_1 & \text{if } \bar{U}_1 < \bar{U}_2 \\
  a_2 & \text{if } \bar{U}_1 \geq \bar{U}_2
\end{cases}
\]

where

\[
\bar{U}_1 = \sum_{i=1}^{n_1} U_{i}/n_1 \sim N(\mu_1, 1/n_1)
\]

and

\[
\bar{U}_2 = \sum_{i=n_1+1}^{n_2} U_i/n_2 \sim N(\mu_2, 1/n_2).
\]
We shall assume that there is a loss function which assigns a "loss" to each action and for each value of the parameter $\mu_1 - \mu_2$. For the purposes of this illustration assume the following loss function:

$$L(\mu_1 - \mu_2, a)$$

<table>
<thead>
<tr>
<th>Action</th>
<th>$\mu_1 - \mu_2 &lt; 0$</th>
<th>$\mu_1 - \mu_2 \geq 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$: assume $\mu_1 &lt; \mu_2$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$a_2$: assume $\mu_1 \geq \mu_2$</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

This loss function may be unrealistic because it assumes the same loss no matter how far $\mu_1$ is from $\mu_2$, just so long as the action is erroneous. However, for the purposes of illustration this loss function will be assumed.

Some additional notation is needed for what follows. $\mathbf{U}$ is a random vector; let $\mathbf{u}$ denote the realized value of $\mathbf{U}$. Similarly, $\mathbf{\bar{U}}_1$ and $\mathbf{\bar{U}}_2$ are random variables with realizations

$$\mathbf{\bar{u}}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} u_i/n_1$$
$$\mathbf{\bar{u}}_2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} u_i/n_2.$$ 

Now $a(\mathbf{\bar{U}}_1 - \mathbf{\bar{U}}_2)$ is a random variable with realized value $a(\mathbf{\bar{u}}_1 - \mathbf{\bar{u}}_2)$.

The decision rule imposes a two set partition on the sample space $E^{n_1+n_2}$ (of values of $U_i, i=1, 2, \ldots, n_1+n_2$ or $u_i, i=1, 2, \ldots, n_1+n_2$), namely:

$$T_1 = \{ \mathbf{U}: \mathbf{\bar{U}}_1 < \mathbf{\bar{U}}_2 \} = \{ \mathbf{U}: a(\mathbf{\bar{u}}_1 - \mathbf{\bar{u}}_2) = a_1 \}$$
$$T_2 = \{ \mathbf{U}: \mathbf{\bar{U}}_1 \geq \mathbf{\bar{U}}_2 \} = \{ \mathbf{U}: a(\mathbf{\bar{u}}_1 - \mathbf{\bar{u}}_2) = a_2 \}.$$ 

Clearly $T_1 \cup T_2 = E^{n_1+n_2}$ and $T_1 \cap T_2 = \phi$, i.e., $T_1$ and $T_2$ constitute a partition of the sample space.
Just as \( a(\bar{u}_1-\bar{u}_2) \) is a random variable the loss function, evaluated for a random \( a(\bar{u}_1-\bar{u}_2) \), \( L(\mu_1-\mu_2, a(\bar{u}_1-\bar{u}_2)) \), is a random function. The expected value of the loss function is called the risk function:

\[
R(\mu_1-\mu_2) = E[L(\mu_1-\mu_2, a(\bar{u}_1-\bar{u}_2))]
\]

\[
= \begin{cases} 
0 \cdot \Pr(T_1) + 1 \cdot \Pr(T_2) & \text{if } \mu_1-\mu_2 < 0 \\
1 \cdot \Pr(T_1) + 0 \cdot \Pr(T_2) & \text{if } \mu_1-\mu_2 \geq 0
\end{cases}
\]

\[
= \begin{cases} 
\Pr(T_2) & \text{if } \mu_1-\mu_2 < 0 \\
\Pr(T_1) & \text{if } \mu_1-\mu_2 \geq 0
\end{cases}
\]

With the normality assumptions given above for the distribution of \( \bar{u} \),

\[
\Pr(T_1) = \Pr[(\bar{u}_1-\bar{u}_2) < 0] = 1 - \Phi[(\mu_1-\mu_2)/\sqrt{1/n_1+1/n_2}]
\]

\[
\Pr(T_2) = \Pr[(\bar{u}_1-\bar{u}_2) \geq 0] = \Phi[(\mu_1-\mu_2)/\sqrt{1/n_1+1/n_2}]
\]

where \( \Phi \) denotes the cumulative distribution function of the standard normal \((0,1)\) distribution. Hence with the normality assumptions

\[
R(\mu_1-\mu_2) = \begin{cases} 
1 - \Phi[(\mu_1-\mu_2)/\sqrt{1/n_1+1/n_2}] & \text{if } \mu_1 \geq \mu_2 \\
\Phi[(\mu_1-\mu_2)/\sqrt{1/n_1+1/n_2}] & \text{if } \mu_1 \leq \mu_2
\end{cases}
\]

The material above is well known, of course and is repeated here for the purpose of establishing notation.

3.2 Perturbations and Conditional Risk

In practice if one were using the decision rule for the problem presented in the previous section one would want some measure of reliability of the decision as, for example an estimate of the risk function, i.e. the probability of making the wrong decision.
The distribution theory for the problem above is such that measures of reliability have been worked out. The purpose of this section is to present a technique which can be used to construct a measure of the reliability of the decision procedure when the direct distribution theory approach is intractable.

3.2.1 Perturbations

In the previous section a standard decision theory problem was set up such that the risk function is easily evaluated, given values for the unknown parameters. If the probability structure were more complicated, as is the case in the AEV problem, the technique described in the following paragraphs might be helpful.

The \( U \) vector of the previous section can be written as

\[
U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + E
\]

where \( E \sim N_{n_1+n_2}(0,I) \). One can inquire: Suppose that the data had been different; would the conclusion (decision) have been different?. Or, slightly rephrased: How different would the data have had to have been to change the conclusion? One could also ask what is the least change in the data required to change the conclusion?

One can quantify these questions by first adding (conceptually) a vector of "perturbations" to the data and then asking the questions in terms of the perturbed data. Let \( D \) be an \((n_1+n_2) \times 1\) vector of random perturbations independent of and with the same distribution as the original "errors" \( E: D \sim N_{n_1+n_2}(0,I) \). Now by adding \( D \) to the data, \( U \), we obtain a new random variable, say, \( W \):
\[ \mathbf{w} = \begin{pmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{pmatrix} = \mathbf{u} + \mathbf{d} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \mathbf{e} + \mathbf{d} \]

and

\[ \mathbf{w} \sim N_{n_1+n_2} \left[ \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, 2I \right]. \]

If \( \mathbf{d} \) is added after \( \mathbf{u} \) is realized, we have the conditional variable:

\( (\mathbf{w} | \mathbf{u} = \mathbf{u}) \) or \( (\mathbf{w} | \mathbf{u}) = \mathbf{u} + \mathbf{d} \sim N_{n_1+n_2} \left[ \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \mathbf{I} \right]. \)

We compute the means of \( \mathbf{w}_1 \) and \( \mathbf{w}_2 \) just as we computed the means of \( \mathbf{u}_1 \) and \( \mathbf{u}_2 \):

\[ \bar{\mathbf{w}}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbf{w}_i/n_1, \quad \bar{\mathbf{w}}_2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} \mathbf{w}_i/n_2. \]

The unconditional and conditional distributions are:

\[ \bar{\mathbf{w}}_i \sim N(\mu_i, 2/n_i) \]

\( (\bar{\mathbf{w}}_1 | \mathbf{u}) \sim N(\bar{\mathbf{u}}_i, 1/n_i) \).

Because the conditional variable notation is clumsy, in subsequent paragraphs we will not distinguish between conditional and unconditional variables, but will distinguish between the conditional and unconditional distributions of the variables.

We apply the same decision rule to the \( \bar{\mathbf{w}}_i \) as to the \( \bar{\mathbf{u}}_i \):

\[ a(\bar{\mathbf{w}}_1 - \bar{\mathbf{w}}_2) = \begin{cases} 
\alpha_1: \text{conclude } \mu_1 < \mu_2 \text{ if } \bar{\mathbf{w}}_1 < \bar{\mathbf{w}}_2 \\
\alpha_2: \text{conclude } \mu_1 \geq \mu_2 \text{ if } \bar{\mathbf{w}}_1 \geq \bar{\mathbf{w}}_2 
\end{cases}. \]

Now questions similar to the ones stated at the beginning of the section can be formulated rigorously.
Suppose, for the sake of illustration, we have observed
\( \bar{u}_1 < \bar{u}_2 \) (realized values). One could ask:

Q1: How different would the \( u_1 \) data have had to
have been to have changed the conclusion?
(i.e., make \( \bar{u}_1 \geq \bar{u}_2 \)); or rigorously, what is
the probability measure of the set of
perturbations of \( u_1 \) which results in a
different decision; i.e., what is
\[
\Pr[D_1: \bar{\tilde{u}}_1 \geq \bar{\tilde{u}}_2 | \bar{u}] ?
\]

Q2: How different would the \( u_2 \) data have had to
have been to have changed the conclusion
(i.e., made \( \bar{u}_1 \geq \bar{u}_2 \)); or, rigorously, what
is the probability measure of the set of
perturbations of \( u_2 \) which results in a
different decision; i.e., what is
\[
\Pr[D_2: \bar{\tilde{u}}_1 \geq \bar{\tilde{u}}_2 | \bar{u}] ?
\]

A more sensible question, particularly in this example is

Q3: How different would the data (\( u_1 \) and \( u_2 \)) have
had to have been to have changed the conclu-
sion? or, rigorously, what is the probability
measure of the set of perturbations of
\[
\begin{bmatrix}
\bar{u}_1 \\
\bar{u}_2 
\end{bmatrix}
\]
which would have changed the conclusion; i.e.,
what is
\[
\Pr[D: \bar{\tilde{u}}_1 \geq \bar{\tilde{u}}_2 | \bar{u}] ?
\]

Q1 and Q2 have a different flavor than Q3 in that Q1 and Q2
each fix one mean at its observed value and ask how much the data
going into the other mean would have to change in order to change the conclusion. Q3 treats the two means simultaneously. All three questions have analogous questions in the AEV situation and each of those questions has a meaningful interpretation.

One could also ask:

Q4: What is the smallest change in the data which would change the conclusion? i.e., what is

$$\inf \{ \mathbf{D}: \bar{u}_1 + \mathbf{D}_1 \geq \bar{u}_2 + \mathbf{D}_2 \} \quad ||\mathbf{D}||$$

There are several problems with Q4. One is not sure which norm to use, although the Euclidean norm, $||\mathbf{D}||^2 = \mathbf{D}'\mathbf{D}$, is probably the most appropriate considering that $\mathbf{D}$ is a surrogate for $\mathbf{E}$ and $\mathbf{E} \sim \mathcal{N}(\mathbf{0}, 1)$. Aside from that problem, had the data been different, there is no particular reason to believe the data would have been more likely to occur in the neighborhood of the closest point which changes the conclusion than in the neighborhood of some other point. Finally, knowing the distance to the nearest point which would change the conclusion does not seem to have as satisfying an interpretation as knowing the probability measure of the set of all points (perturbations) which would change the conclusion when the perturbations have the same distribution as the assumed distribution of the original data errors.

The sets for which probabilities are desired in Q1, Q2, and Q3, together with their complements, form two-set partitions of the sample spaces of $D_1$, $D_2$, and $D$, respectively. The point is that although the probabilities can be computed in terms of the distribution of the statistics $\bar{W}_1$ and $\bar{W}_2$, if these distributions were
complicated one could compute the probabilities of sets in the sample space of the perturbations $D_1, D_2$ or $D$, which have simple distributions.

Just to complete the example, the probabilities in Q1, Q2, and Q3 are respectively:

$$Q1: \Pr\{D_1: \tilde{w}_1 \geq \tilde{w}_2 | U\} = \Phi[\sqrt{n_1} (\tilde{u}_1 - \tilde{u}_2)]$$

$$Q2: \Pr\{D_2: \tilde{w}_1 \leq \tilde{u}_1 | U\} = \Phi[\sqrt{n_2} (\tilde{u}_1 - \tilde{u}_2)]$$

$$Q3: \Pr\{D: \tilde{w}_1 - \tilde{w}_2 < 0 | U\} = \Phi[(u_1 - u_2)/\sqrt{1/n_1 + 1/n_2}] .$$

These probabilities are not the risks of the previous section.

One minus the risk can be interpreted as a "distance" between $\mu_1$ and $\mu_2$ in "probability units" in the sense that if the risk (probability of wrong conclusion) is small, the "distance" is large and vice versa. The probabilities of Q1, Q2, and Q3, instead of measuring the distance between the parameters $\mu_1$ and $\mu_2$, actually lead to measures of the "distance" between $\tilde{u}_1$ and $\tilde{u}_2$ in the probability sense described above. Thus the perturbation analysis is a conditional analysis, conditional upon the observed values of $\tilde{u}_1$ and $\tilde{u}_2$.

The interpretation is still reasonable, however, for if the probabilities above are small, the data would have had to have been considerably different to change the conclusion and one can conclude either: (a) $\mu_1$ and $\mu_2$ are far apart in the direction indicated by $\tilde{u}_1 - \tilde{u}_2$; or (b) $\mu_1$ and $\mu_2$ are close together (or far apart in the direction opposite to that indicated by $\tilde{u}_1 - \tilde{u}_2$) and a "rare" or "unusual" or "improbable" vector of errors has been observed. In
case (a) one has drawn the correct conclusion and the small probability of the set of perturbations which would change the conclusion correctly indicates considerable confidence in the conclusion. Case (b), by definition, occurs only rarely; in stochastic situations one must accept the fact that there is always a non-zero probability of reaching wrong conclusions.

In this discussion the terms "small", "far apart", "close together", "rare", etc. have not been defined or quantified because these are interpretive phrases and a risk or probability of error which is quite small in one situation may be unacceptably large in another. However, the point of the discussion is that the smaller the probability (of the set of perturbations which would change the conclusion), the greater the confidence one can place in the conclusion and vice versa. However, rare events can and do occur and in real problems one can never be absolutely certain of a conclusion (otherwise, by definition, the problem is "unreal").

It should be clear from the discussion above that the perturbations need not necessarily have a normal distribution, nor does one necessarily have to add the perturbations. Since the perturbations are surrogate errors, in different applications of this approach one would feel that the perturbations should have the same distribution as the errors (or random component of the model) and should be incorporated into the model in the same manner as the errors. In the present example and in the AEV case (the general linear model), additive, normally distributed perturbations seem most appropriate.
3.2.2 **Conditional Risk**

In this section we continue the theme of the perturbation approach; to augment the intuitive interpretation we develop a risk function for the perturbed data, conditional on the original data, and show this has the same form as some of the probabilities in the previous section.

We continue to use the notation of the previous section. For the decision function \( a(\tilde{W}_1, \tilde{W}_2) \) defined previously, consider the loss function (conditional on the value of \( \bar{u}_1 - \bar{u}_2 \)):

\[
L[\bar{u}_1 - \bar{u}_2, a(\tilde{W}_1, \tilde{W}_2)]
\]

<table>
<thead>
<tr>
<th>Action</th>
<th>States of Nature</th>
<th>( \bar{u}_1 - \bar{u}_2 )</th>
<th>0</th>
<th>( \bar{u}_1 - \bar{u}_2 ) = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_1 ) (conclude ( \mu_1 &lt; \mu_2 ))</td>
<td></td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>( a_2 ) (conclude ( \mu_1 \geq \mu_2 ))</td>
<td></td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Note this is a conditional loss function in the sense that the parameters \( \mu_1 = E(\bar{u}_1) \) and \( \mu_2 = E(\bar{u}_2) \) have been replaced by the conditional expected values of the perturbed data, \( E(\tilde{W}_1 | U) = \bar{u}_1 \) and \( E(\tilde{W}_1 | U) = \bar{u}_2 \), respectively.

The conditional risk, i.e. the expected loss conditional on the observed \( U \), is a function of \( \bar{u}_1 - \bar{u}_2 \):

\[
R(\bar{u}_1 - \bar{u}_2) = E[L[\bar{u}_1 - \bar{u}_2, a(\tilde{W}_1, \tilde{W}_2)|U]]
\]

\[
= \begin{cases} 
\Pr(\tilde{W}_1 \geq \tilde{W}_2 | U) & \text{if } \bar{u}_1 < \bar{u}_2 \\
\Pr(\tilde{W}_1 < \tilde{W}_2 | U) & \text{if } \bar{u}_1 \geq \bar{u}_2 
\end{cases}
\]

The normality assumptions of the previous section imply
\[ R(\bar{u}_1 - \bar{u}_2) = \begin{cases} 
\Phi[(\bar{u}_1 - \bar{u}_2)/\sqrt{1/n_1+1/n_2}] & \text{if } \bar{u}_1 < \bar{u}_2 \\
1 - \Phi[(\bar{u}_1 - \bar{u}_2)/\sqrt{1/n_1+1/n_2}] & \text{if } \bar{u}_1 \geq \bar{u}_2 
\end{cases} \]

These are of course, the same probabilities which arose in Q3 of the previous section. This conditional risk function is the same function as the unconditional risk function derived in the first section of this chapter; here it is evaluated at \( \bar{u}_1 - \bar{u}_2 \); there it was evaluated at \( \mu_1 - \mu_2 \). Thus if \( \bar{u}_1 - \bar{u}_2 \) is close to \( \mu_1 - \mu_2 \), the conditional risk is close to the unconditional risk of the decision rule based on \( \mu_1 - \mu_2 : a(\bar{U}_1 - \bar{U}_2) \). When \( \bar{u}_1 - \bar{u}_2 \) is unusually far from \( \mu_1 - \mu_2 \) (by definition a rare event) then the conditional risk will be quite different from the unconditional risk. Thus one is justified in interpreting a "suitably small" conditional risk as follows: either the conclusion is correct or a rare event has occurred; it is impossible to determine which situation exists but in either case one can place a "large" confidence in the conclusion.

One can also consider the unconditional risk of the decision function \( a(\bar{W}_1 - \bar{W}_2) \). The loss function has the form

\[
L[\mu_1 - \mu_2, a(\bar{W}_1 - \bar{W}_2)]
\]

<table>
<thead>
<tr>
<th>Action</th>
<th>States of Nature</th>
<th>( \mu_1 &lt; \mu_2 )</th>
<th>( \mu_1 \geq \mu_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a_1(\text{conclude } \mu_1 &lt; \mu_2) )</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>( a_2(\text{conclude } \mu_1 \geq \mu_2) )</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

The unconditional risk function is the expected value of the unconditional loss function:
\[ R(\mu_1 - \mu_2) = \begin{cases} 
\text{Pr}(\bar{w}_1 \geq \bar{w}_2) & \text{if } \mu_1 < \mu_2 \\
\text{Pr}(\bar{w}_1 < \bar{w}_2) & \text{if } \mu_1 \geq \mu_2 
\end{cases} \]

Write the normality assumptions

\[ R(\mu_1 - \mu_2) = \begin{cases} 
\Phi\left(\frac{(\mu_1 - \mu_2)/\sqrt{\nu_1^2 + \nu_2^2}}{\nu_1 + \nu_2}\right) & \text{if } \mu_1 < \mu_2 \\
1 - \Phi\left(\frac{(\mu_1 - \mu_2)/\sqrt{\nu_1^2 + \nu_2^2}}{\nu_1 + \nu_2}\right) & \text{if } \mu_1 \geq \mu_2 
\end{cases} \]

3.2.3 A Numerical Example

In this section we present a numerical example in order to further clarify the use of the conditional risk.

Suppose that one observed \( \bar{u}_1 = 4.5 \) and \( \bar{u}_2 = 5.5 \) based on sample sizes of \( n_1 = n_2 = 8 \). The conditional risk would then be

\[ \Phi\left(\frac{(4.5 - 5.5)/\sqrt{1/8 + 1/8}}{1}\right) \]

or

\[ \Phi\left(2(4.5 - 5.5)\right) = 0.0227. \]

Thus if one perturbed the data by adding random perturbations having the distribution \( D \sim N(0, I_{16}) \), 2.27% of the conclusions based on the perturbed data would be different from the conclusion based on the original data. This probability is conditional on the observed values \( \bar{u}_1 = 4.5 \) and \( \bar{u}_2 = 5.5 \). For different values of \( \bar{u}_1 \) and \( \bar{u}_2 \), one would get different conditional risks. Obviously the larger the difference between \( \bar{u}_1 \) and \( \bar{u}_2 \), the smaller the conditional risk.

3.3 An Example Using the Uniform Distribution

Suppose that one has a random variable \( U \sim \text{uniformly on } [\mu-1/2, \mu+1/2] \) and that on the basis of a single observation, \( u \), one
wishes to decide if $\mu$ is greater or less than zero. If $u > 1/2$, then $\mu > 0$ with probability 1 and if $\mu < -1/2$, then $\mu < 0$ with probability 1. Thus with respect to a decision rule and the perturbation approach we need only consider $[-1/2, 1/2]$. Define a decision rule as follows:

$$a(u) = \begin{cases} 
  a_1: \text{conclude } \mu < 0 \text{ if } U < 0 \\
  a_2: \text{conclude } \mu \geq 0 \text{ if } U \geq 0 
\end{cases}$$

If the problem is simply to choose between $\mu < 0$ and $\mu \geq 0$, then an appropriate loss function is:

<table>
<thead>
<tr>
<th>Action</th>
<th>State of Nature</th>
<th>$\mu &lt; 0$</th>
<th>$\mu \geq 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$: conclude $\mu &lt; 0$</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$a_2$: conclude $\mu \geq 0$</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

We could evaluate the expected loss but this would depend explicitly on $\mu$. We will instead proceed directly to the perturbation approach.

The random variable $U$ can be decomposed into constant and random components:

$$U = \mu + E$$

where $\mu$ is the unknown parameter and $E$ is an "error" random variable distributed uniformly on $[-1/2, 1/2]$. Since the perturbation is a post hoc surrogate for the error term, we take the perturbation $D \sim \text{Uniform } (-1/2, 1/2)$ independently of $E$ and add $D$ to $U$ or the realized value, $u$. 
The decision rule based on $u + D$ is:

$$a(u + D) = \begin{cases} 
    a_1: & \text{conclude } \mu < 0 \text{ if } u + D < 0 \\
    a_2: & \text{conclude } \mu \geq 0 \text{ if } u + D \geq 0 .
\end{cases}$$

Define a conditional loss function as:

$$L[u, a(u + D)]$$

<table>
<thead>
<tr>
<th>Action</th>
<th>State of Nature</th>
<th>$u &lt; 0$</th>
<th>$u \geq 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$: conclude $\mu &lt; 0$</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$a_2$: conclude $\mu \geq 0$</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

The expected value of this conditional loss function is the conditional risk:

$$R(u) = \begin{cases} 
    \Pr[(u + D) \geq 0] & \text{if } u < 0 \\
    \Pr[(u + D) < 0] & \text{if } u \geq 0
\end{cases}$$

$$R(u) = \frac{1}{2} - |u| .$$

As before, if the conditional risk is "small", the interpretation is that one can place great confidence in the conclusion. The point of this section is, however, that the perturbation approach need not be confined to normal, additive perturbations.

3.4 **Summary**

The purpose of this chapter has been to present the perturbation/conditional risk approach and interpretation in a familiar setting. The idea is to separate the explanation of this essentially new technique from the description of the application of the technique to the AEV problem, which is given in the next chapter.
CHAPTER IV

APPLICATION OF THE PERTURBATION/CONDITIONAL RISK APPROACH TO THE AEV CRITERION

In this chapter we will apply the perturbation/conditional risk approach to the AEV in order to obtain a measure of reliability of the decision rule: conclude $E(AEV_i) < E(AEV_j)$ if $AEV_i < AEV_j$. We will derive the conditional risk in this situation and show that it may be evaluated numerically. In addition, we will show that the perturbation approach is applicable to other criteria, specifically the $C_p$ statistic.

4.1 Definition of the Problem; Notation

As previously noted the "best" model in the sense of minimizing the average variance with respect to $w(x)$ and $R$ is the one which has the minimum expected AEV. One hopes that the ordered AEV's reflect the order of the expected AEV's. However, since the AEV's are random variables one knows that this will not always be the case. One needs an index which indicates how confident one can be that, given two or more AEV's, the model with the minimum observed AEV is also the model with the minimum expected AEV.

Let us now establish the AEV notation in more general terms. Given a random variable $V \sim N(\mu, \sigma I)$ where $\mu = \mu_0$ we define the AEV for the $k^{th}$ model as:

$$AEV_k(V) = k \mu^T A_k \mu$$
where \( k_\lambda = \text{Tr}[(Z_\lambda \cdot Z_\lambda')^{-1}M_\lambda]/(n-p_\lambda) \) and

\[ A_\lambda = I - Z_\lambda(Z_\lambda \cdot Z_\lambda')^{-1}Z_\lambda'. \]

Note that the AEV is expressed as a function of the model, \( \lambda \), and the random variable \( V \).

Our decision rule, given two AEV's, say \( \text{AEV}_j(V) \) and \( \text{AEV}_j(V) \), is as follows:

\[
a_{\text{AEV}_i(V) - \text{AEV}_j(V)} = \begin{cases} 
  a_1: \text{conclude } E(\text{AEV}_i(V)) < E(\text{AEV}_j(V)) & \text{if } \text{AEV}_i(V) < \text{AEV}_j(V) \\
  a_2: \text{conclude } E(\text{AEV}_i(V)) \geq E(\text{AEV}_j(V)) & \text{if } \text{AEV}_i(V) \geq \text{AEV}_j(V) 
\end{cases}
\]

Define the loss function for this decision rule as:

\[
L[E(\text{AEV}_i(V)) - E(\text{AEV}_j(V)); a]
\]

<table>
<thead>
<tr>
<th>States of Nature</th>
<th>(( \text{AEV}_i(V) &lt; E(\text{AEV}_j(V)) ))</th>
<th>(( \text{AEV}_i(V) \geq E(\text{AEV}_j(V)) ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actions</td>
<td>( \text{AEV}_i(V) &lt; E(\text{AEV}_j(V)) )</td>
<td>( \text{AEV}_i(V) \geq E(\text{AEV}_j(V)) )</td>
</tr>
<tr>
<td>( a_1 ): Assume</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( a_2 ): Assume</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Note that the loss and risk functions have more parameters or arguments than \( E(\text{AEV}_i(V)) - E(\text{AEV}_j(V)) \); specifically the parameters include \( H, c, Z_i, Z_j, M_{i1}, \) and \( M_{jj} \) (submatrices of the matrix of moments, \( M \), of the weight function and region of interest). However for simplicity we shall use the notation shown.
The expected loss or risk is:

\[ R[\mathbb{E}(\text{AEV}_1(Y)) - \mathbb{E}(\text{AEV}_j(Y))] = \]

\[
\begin{cases} 
\Pr\{Y: \text{AEV}_1(Y) \geq \text{AEV}_j(Y)\} & \text{if } \mathbb{E}(\text{AEV}_1(Y)) < \mathbb{E}(\text{AEV}_j(Y)) \\
\Pr\{Y: \text{AEV}_1(Y) < \text{AEV}_j(Y)\} & \text{if } \mathbb{E}(\text{AEV}_1(Y)) \geq \mathbb{E}(\text{AEV}_j(Y))
\end{cases}
\]

\[ = \begin{cases} 
\Pr\{Y: k_1 V' A_1 Y \geq k_j V' A_j Y\} & \text{if } \mathbb{E}(\text{AEV}_1(Y)) < \mathbb{E}(\text{AEV}_j(Y)) \\
\Pr\{Y: k_1 V' A_1 Y < k_j V' A_j Y\} & \text{if } \mathbb{E}(\text{AEV}_1(Y)) \geq \mathbb{E}(\text{AEV}_j(Y)),
\end{cases}
\]

or since \( k_1 V' A_1 Y - k_j V' A_j Y = V'(k_1 A_1 - k_j A_j) Y = V' A Y \)

\[ R[\mathbb{E}(\text{AEV}_1(Y)) - \mathbb{E}(\text{AEV}_j(Y))] = \]

\[
\begin{cases} 
\Pr\{Y' A Y \geq 0\} & \text{if } \mathbb{E}(\text{AEV}_1(Y)) < \mathbb{E}(\text{AEV}_j(Y)) \\
\Pr\{Y' A Y < 0\} & \text{if } \mathbb{E}(\text{AEV}_1(Y)) \geq \mathbb{E}(\text{AEV}_j(Y)).
\end{cases}
\]

Note that \( A \) is a function of \( i \) and \( j \) and should probably be written as \( A_{ij} \) but we have suppressed the subscripts for notational simplicity.

Let us now return to the general linear model situation. In this case we have assumed that the random variable \( Y \) has the form

\( Y = X \beta + \varepsilon \) and that \( \mathbb{E}(\varepsilon) = 0 \) and \( \mathbb{E}(\varepsilon \varepsilon') = \sigma^2 I \). Thus for the \( \text{AEV} \) as a function of \( Y \) the risk is:

\[ R[\mathbb{E}(\text{AEV}_1(Y)) - \mathbb{E}(\text{AEV}_j(Y))] = \]

\[
\begin{cases} 
\Pr\{Y' A Y \geq 0\} & \text{if } \mathbb{E}(\text{AEV}_1(Y)) < \mathbb{E}(\text{AEV}_j(Y)) \\
\Pr\{Y' A Y < 0\} & \text{if } \mathbb{E}(\text{AEV}_1(Y)) \geq \mathbb{E}(\text{AEV}_j(Y))
\end{cases}
\]

\[ = \begin{cases} 
\Pr\{(X \beta + \varepsilon)' A (X \beta + \varepsilon) \geq 0\} & \text{if } \mathbb{E}(\text{AEV}_1(Y)) < \mathbb{E}(\text{AEV}_j(Y)) \\
\Pr\{(X \beta + \varepsilon)' A (X \beta + \varepsilon) < 0\} & \text{if } \mathbb{E}(\text{AEV}_1(Y)) \geq \mathbb{E}(\text{AEV}_j(Y)).
\end{cases}
\]
These probabilities are functions of the unknown parameters $\beta$ and $\sigma^2$. Thus we find ourselves in the same situation as was encountered in the examples of the previous chapter. As a result we can again turn to the perturbation approach.

Let us consider a vector of perturbations $D \sim N(0, s^2I)$, where $s^2$ is the residual mean square for the full model. The perturbations are additive in nature (i.e. the perturbed data are of the form $y + D$). The assumptions with respect to the distribution and additively were made because the perturbations are surrogate errors and thus should have the form of the errors in the general linear model. The replacement of $\sigma^2$ by $s^2$ is done so that we may evaluate the probabilities which arise in the following sections. We use $s^2$ because, without prior knowledge of the form of $\beta$, it is the only estimate of $\sigma^2$ which we know to be unbiased. Note also at the time we are considering perturbations of the data, $s^2$ is not a random variable as the inferences are conditional upon the data.

Suppose, for the sake of illustration, that we have observed $AEV_i(y) < AEV_j(y)$ so that we conclude $E(AEV_i(y)) < E(AEV_j(y))$. In this situation one can again inquire: Is it likely that other data would have led to a different conclusion?

First one can consider the sensitivity of the larger AEV to change in the data. If the larger AEV is sensitive to perturbations (i.e., has a large variance) then a substantial proportion of $AEV_j(y+D)$ values may be smaller than the value being used for comparison, namely $AEV_i(y)$. This consideration leads to the following question which is made rigorous in terms of probability statements:
Q1: How different would the vector \( y \) have had to have been in order for the AEV for the \( j^{th} \) model to have been less than the observed AEV for the \( i^{th} \) model? More rigorously, what is the probability measure of the set of perturbations of \( y \) such that 
\[ \text{AEV}_j(y+D) < \text{AEV}_i(y), \text{ i.e., what is} \]
\[ \Pr[D: \text{AEV}_j(y+D) < \text{AEV}_i(y)]? \]

In a similar manner one can consider the sensitivity of the smaller AEV to change in the data. If \( \text{AEV}_i \) is sensitive to data perturbations (i.e., has large variance), then a substantial proportion of \( \text{AEV}_i(y+D) \) will exceed the comparison value \( \text{AEV}_j(y) \). These considerations lead to the question:

Q2: How different would the vector \( y \) have had to have been in order for the AEV computed for the \( i^{th} \) model to have been greater than the observed AEV for the \( j^{th} \) model? More rigorously, what is the probability measure of the set of perturbations of \( y \) such that \( \text{AEV}_i(y+D) \geq \text{AEV}_j(y), \text{ i.e., what is} \]
\[ \Pr[D: \text{AEV}_i(y+D) \geq \text{AEV}_j(y)]? \]

The two questions above, and the resulting probabilities, are interesting in themselves and also because the probabilities can be evaluated relatively easily (see Section 4.2). However, even though the answers to these questions may lend insight, they are side issues to the main question: How sensitive is the decision itself to perturbations of the data? If a substantial proportion of the perturbations added to the data would result in a different conclusion
one cannot have much confidence in the conclusion based on the data.

These considerations lead to question Q3.

Q3: How different would the vector \( y \) have had to have been in order to have changed the conclusion? More rigorously: What is the probability measure of the set of perturbations of \( y \) which results in a different decision; i.e., what is \( \Pr[D: \text{AEV}_1(y+D) \geq \text{AEV}_j(y+D)] \)?

In Q1 and Q2 we fix an AEV and see how the data would have had to have changed in order to have reversed the order of the original AEV's. Q3 is the same idea as Q3 in the previous chapter. Both AEV's are allowed to vary simultaneously as a function of the perturbed data. Thus the probability which results from Q3 is again the probability of changing the initial conclusion. The sets for which the probabilities are desired in Q1, Q2, and Q3, together with their complements, form two-set partitions on the sample space of \( D \) as was the case in the previous chapter. The method of evaluation of these probabilities will appear in subsequent sections.

It is also possible to express the probability of Q3 in the form of a conditional risk. Earlier in this section we assumed a general decision rule and loss function and derived a general risk function in terms of \( \text{AEV}(V) \). If we make \( V = y + D \) then the risk is a conditional risk. As was shown before

\[
R[\text{IEV}_1(V) - \text{IEV}_j(V)] =
\begin{cases}
\Pr(V'AV \geq 0) & \text{if } \text{IEV}_1(V) < \text{IEV}_j(V) \\
\Pr(V'AV < 0) & \text{if } \text{IEV}_1(V) \geq \text{IEV}_j(V)
\end{cases}
\]
substituting $y + D$ for $y$, we obtain

$$R[AEV_1(y) - AEV_j(y)] =$$

$$\begin{cases} \Pr[(y+D)'A(y+D) \geq 0] & \text{if } AEV_1(y) < AEV_j(y) \\ \Pr[(y+D)'A(y+D) < 0] & \text{if } AEV_1(y) \geq AEV_j(y) \end{cases}.$$

It is interesting to note the similarity between the conditional risk and the unconditional risk derived earlier in this section. In the conditional risk, $E(y) = X\beta$ has been replaced by $y$ and $E$ has been replaced by the surrogate error $D$. Otherwise, the two risk functions are identical. Thus if $y$ is close to $E(y)$ the conditional risk will be close to the unconditional risk.

The interpretation is as before. If the conditional risk is small then we can be confident in the decision that we have made. Conversely, if the conditional risk is large we place little confidence in our conclusion.

4.2 **Evaluation of the Probabilities Arising from Questions Q1 and Q2**

In this section the method of evaluating the probabilities arising from questions Q1 and Q2 of the previous section is described. We show that these probabilities are non-central $\chi^2$ probabilities and indicate how they may be evaluated.

In the previous section it was shown that the probability arising from question Q1 for $AEV_1(y) < AEV_j(y)$ is:

$$\Pr(D; \ AEV_1(y) < AEV_j(y+D) | y) \quad \text{where } D \sim N(0, S^2) \text{ independently of } y.$$ 

This is equivalent to

$$\Pr[(y+D)'k_jA_j(y+D) \leq AEV_1(y)|y] =$$

$$\Pr[(y+D)'A_j(y+D)/s^2 \leq AEV_1(y)/(k_j s^2)|y].$$
Now since \((y+D)/s \sim N(y/s, 1)\) and \(A_j \) is idempotent then by Theorem 4.7 on page 83 of Graybill (1961)

\[(y+D)'A_j(y+D)/s^2 \sim \chi^2(n-p_j, 1/2 y'A_jy/s^2)\]

but

\[y'A_jy/s^2 = AEV_j/k_j s^2\]

thus

\[(y+D)'A_j(y+D)/s^2 \sim \chi^2(n-p_j, 1/2 AEV_j/k_j s^2)\]

and

\[Pr(D|Y) = AEV_j(y)|Y\] \[= Pr[\chi^2(n-p_j, 1/2 AEV_j(y)/k_j s^2) \leq AEV_j(y)/k_j s^2] \]

The probability resulting from Q2, i.e.,

\[Pr(D|Y) \geq AEV_j(y)|Y\], is quite similar in structure to the probability arising from Q1. It can be shown that

\[Pr(D|Y) = AEV_j(y)|Y\] \[= Pr[(y+D)'A_1(y+D)/s^2 \geq AEV_j/k_j s^2] \]

where \((y+D)/s \sim N(y/s, 1)\). Then by Theorem 4.7 on page 83 of Graybill

\[(y+D)'A_1(y+D)/s^2 \sim \chi^2(n-p_1, 1/2 y'A_1y/s^2)\]

but

\[y'A_1y/s^2 = AEV_1(y)/k_1 s^2\]

thus

\[(y+D)'A_1(y+D)/s^2 \sim \chi^2(n-p_1, 1/2 AEV_1(y)/s^2)\].

Therefore, \[Pr(D|Y) \geq AEV_j(y)|Y\] =

\[Pr[\chi^2(n-p_1, 1/2 AEV_1(y)/s^2) \geq AEV_1/k_1 s^2] \]

It should also be noted that the first of these methods tends to be anti-conservative and the second conservative. In other words the first method yields probabilities smaller than the conditional risk and the second probabilities larger than the conditional risk. This is due to the fact that in the first case one is
calculating the probability that a noncentral $\chi^2$ is less than a value which is smaller than its mean (the mean is $n-p_j + AEV_j/k_j s^2$ which, since $AEV_i < AEV_j$, is always larger than $AEV_i/k_i s^2$). Unless the two AEV's are close together the probability will be small. In the second case one is evaluating the probability that a noncentral $\chi^2$ is greater than a value which is usually less than its mean (in this case the mean is $n-p_i - AEV_i/k_i s^2$ which in most cases of interest will be greater than $AEV_j/k_j s^2$). Unless the AEV's are quite far apart the probability will be large.
Thus, these two probabilities are open to clear interpretation only in special cases. If the probability arising from Q1 is large then we can place little confidence in our conclusion as this method is basically anti-conservative. Conversely, if the probability arising from Q2 is small then we can place confidence in our conclusion as this method is basically conservative.

It turns out that in practice probabilities are so large or so small that they are not as good as the conditional risk in measuring the quality of the decisions.

4.3 Evaluation of the Conditional Risk

As was shown in section 4.1, the conditional risk is the probability that an indefinite noncentral quadratic form in normal random variables is greater or less than zero (i.e., \( P(W'BW \geq 0) \)) where \( w \sim N(\mu, cI) \) with \( \mu \neq 0 \) and the eigenvalues of \( B \) may be positive negative or zero). If the conditional risk is to prove useful, then we must be able to evaluate such a probability.

Johnson and Kotz (1970) have noted that comparatively little work has been done on the distribution of indefinite quadratic forms probably due to the difficulty of the area. In addition, most of the work has been done on the central case (e.g. Gurland (1955)).

Johnson and Kotz mention two authors who have done extensive work on the distribution of noncentral, indeterminate quadratic forms. These authors are Imhof (1961) and Press (1966). Imhof's technique requires the use of numerical integration and is the technique used in the evaluation of the conditional risk for the AEV. Imhof has shown that we may obtain \( Pr(Q = W'BW \geq 0) \), where Q is an indefinite quadratic
form, by a numerical inversion of the characteristic function of \( Q \).

The conditional risk is of the form \( \Pr[A_{E_1}(V) \geq A_{E_j}(V)] \). The following theory shows that the conditional risk may be put in the form \( \Pr(V'AV > 0) \) and then shows how it may be evaluated using Imhof's techniques.

**Lemma 1.** The following equality holds:

\[
A_{E_1}(V) - A_{E_j}(V) = V'AV
\]

where

\[
A = k_i A_i - k_j A_j
\]

\[
\Lambda = I - Z_{\xi} (Z_{\xi}'Z_{\xi})^{-1} Z_{\xi}'
\]

\[
k_{\xi} = \text{Tr}[(Z_{\xi}'Z_{\xi})^{-1} M_{\xi}] / (n-p_{\xi})
\]

**Proof:**

\[
A_{E_2}(V) = k_{\xi} V' A_{\xi} V
\]

thus

\[
A_{E_1}(V) - A_{E_j}(V) = k_i V' A_i V - k_j V' A_j V
\]

\[
= V'(k_i A_i - k_j A_j) V
\]

\[
= V'AV.
\]

The matrix \( A \) of Lemma 1 is a real symmetric matrix (thus its eigenvalues are real).

**Lemma 2.** Let \( V \sim N(\mu, \Sigma) \) and \( Q = V'AV \), where \( A \) is real and symmetric. Then one can express \( Q \) in the form

\[
Q = c \sum_{r=1}^{R} \lambda_r \chi^2(1, \delta^2_r) \quad \text{or} \quad Q/c = \sum_{r=1}^{R} \lambda_r \chi^2(1, \delta^2_r)
\]

where:

- \( c \) is a positive constant (such as \( s^2, \sigma^2 \), etc.);
A = P Λ P' is the spectral decomposition of A, i.e., A = P Λ P'
where
Λ = diag (λ₁, λ₂, ..., λₙ); we take the convention
λ₁ ≥ λ₂ ≥ ... ≥ λₙ and
P is a matrix of eigenvectors of A such that P'^TP = PP' = I
The χ²(1, δᵣ²) are independent χ² variables with one d.f. and
noncentrality δᵣ². The variable χ²(h, δ²) is defined by the
relation
χ²(h, δ²) = (X + δ)² + Σ Xᵢ² where the Xᵢ ∼ N(0,1);
and δ = P'[μ/√c] where P is the matrix of eigenvectors of A.

Proof:
V'AV = cVPP'APP'V/c
P Λ P' = Λ = P'AP
Thus
V'AV = cV'P'V/c
let W = P'[U/√c]
then W ∼ N(P'[μ/√c], P'1P) or W ∼ N(P'[μ/√c], I)
V'AV = cW'ΛW = c Σ r λᵣ Wᵣ² or Q/c = Σ λᵣ χ²(1, δᵣ²).
Since the Wᵣ² are independent normal random variables the Wᵣ² are dis-
tributed independent as χ²(1, δᵣ²) where δ = P'[μ/√c] and δᵣ² = (P'r μ)²/c.
(The 1/2 in the noncentrality parameter is omitted here in order to
match Imhof's notation.) Since P(Q > X) = P(Q/c > X/c) we will use the
form Q/c so our setup matches Imhof's.

Although not stated explicitly in theorem proof form, Imhof
(1959) proves the results stated in the following theorems. Where
applicable Imhof's notation has been changed to match the notation of
this paper.
Theorem 1 (Imhof). The characteristic function of \( Q = \sum_{r=1}^{m} \lambda_r \chi_r^2 (h_r, \delta_r^2) \) (where the notation is as given in Lemma 2, \( h_r \) is the multiplicity of the \( r \)th eigenvalue, and \( m \) the number of distinct eigenvalues) is given by

\[
\phi(t) = \sum_{r=1}^{m} \frac{1}{(1-2i\lambda_r t)^{-h_r/2}} \exp \left( \frac{i\delta_r^2 \lambda_r t}{(1-2i\lambda_r t)} \right);
\]

in the particular case where \( h_r = 1 \) (all \( r \)) and \( m = n \)

\[
\phi(t) = \prod_{r=1}^{n} \left(1-2i\lambda_r t\right)^{-1/2} \exp \left( \frac{i\delta_r^2 \lambda_r t}{(1-2i\lambda_r t)} \right).
\]

Theorem 2 (Gil-Pelaz, 1951). Inversion Theorem. Let \( \phi(t) \) denote the characteristic function of a distribution and let \( F(x) \) denote the (cumulative) distribution function of the same distribution. Then

\[
F(x) = 1/2 - 1/\pi \int_{0}^{\infty} t^{-1} I\{e^{itx} \phi(t)\} \, dt
\]

where \( I(z) \) denotes the imaginary part of \( z \).

Remark. The usual inversion formula gives \( F(x) - F(0) \) rather than \( F(x) \) directly. Since \( Q \) takes both positive and negative values, Imhof states that a formula is needed which gives \( F(x) \) directly.

Theorem 3 (Imhof). Let \( Q \) be defined as in Theorem 1 and \( \phi \) be the characteristic function of \( Q \), also given in Theorem 1.

Then

\[
\Pr(Q > x) = 1/2 + 1/\pi \int_{0}^{\infty} \frac{\sin \theta(u)}{u \rho(u)} \, du
\]

where

\[
\theta(u) = 1/2 \sum_{r=1}^{n} \left[ \tan^{-1}(\lambda_r u) + \delta_r^2 \lambda_r \frac{u}{u(1+\lambda_r^2 u^2)^{-1}} \right] - 1/2xu
\]

\[
\rho(u) = \prod_{r=1}^{n} (1+\lambda_r^2 u^2)^{1/4} \exp\left[1/2 \delta_r^2 \lambda_r^2 u^2 (1+\lambda_r^2 u^2)^{-1}\right].
\]
The application of the above theory to the distribution of the difference of two AEV's is summarized in the following corollary.

**Corollary 1.** Let $V \sim N(\mu, \Sigma)$ and let $A = K_i A_i - K_j A_j$

where

$$A_{\ell} = I - Z_{\ell}(Z_{\ell}^T Z_{\ell})^{-1} Z_{\ell}, \quad \ell = i, j$$

$$K_{\ell} = \text{Tr}[(Z_{\ell}^T Z_{\ell})^{-1} M_{\ell\ell}]/(n-p_{\ell})$$

Then

$$\Pr\{V: \text{AEV}_i(V) - \text{AEV}_j(V) > 0\} = 1/2 + 1/n \int_0^{\infty} \frac{\sin \theta(u)}{u \rho(u)} \, du$$

where

$$\theta(u) = 1/2 \sum_{r=1}^{n} \{\tan^{-1}(\lambda_r u) + \delta_r^2 \lambda_r u (1+\lambda_r^2 u^2)^{-1}\}$$

$$\rho(u) = \prod_{r=1}^{n} (1+\lambda_r^2 u^2)^{1/4} \exp[1/2(\lambda_r \delta_r u)^2 (1+\lambda_r^2 u^2)^{-1}]$$

$\Lambda$ is the diagonal matrix of eigenvalues of $A$

$A = P \Lambda P'$ where $P'P = PP' = I$

and $\delta = P'\mu/\sqrt{c}$ where $P$ is the matrix of eigenvectors of $A$.

**Proof:**

By lemma 1 $\text{AEV}_i(V) - \text{AEV}_j(V) = V^T A V$ where $A = K_i A_i - K_j A_j$

taking the spectral decomposition of $A$ and making the transformation $W = P' V/\sqrt{c}$ we obtain

$$\Pr\{V: \text{AEV}_i(V) \geq \text{AEV}_j(V)\} = \Pr(W' \Lambda W > 0)$$

where $\Lambda$ is the matrix of eigenvalues of $A$ and $W \sim N(P'\mu/\sqrt{c}, I)$.

Now by theorems 1, 2, and 3

$$\Pr[W' \Lambda W = \sum_{r=1}^{n} \lambda_r w_r^2 \geq 0] =$$

$$1/2 + 1/n \int_0^{\infty} \frac{\sin \theta(u)}{u \rho(u)}$$
where
\[ \theta(u) = \frac{1}{2} \sum_{r=1}^{n} \left[ \tan^{-1}(\lambda_r u) + \delta_r^2 \lambda_r u (1 + \lambda_r^2 u^2)^{-1} \right] \]

\[ \rho(u) = \prod_{r=1}^{n} (1 + \lambda_r^2 u^2)^{1/4} \exp \left[ \frac{1}{2} (\lambda_r \delta_r u)^2 (1 + \lambda_r^2 u^2)^{-1} \right] \]

and \( \delta = P' \mu / \sqrt{c} \) where \( P \) is the matrix of eigenvectors of \( \Lambda \).

In our case \( \mu = \mathbf{y} \) and \( c = s^2 \).

The integration mentioned above must be accomplished numerically. Imhoff concluded the integrand is stable enough for numerical integration. He tested the stability by computing the integral using two different numerical integration techniques until the integrals agreed to within .001. In all cases considered, the agreement was achieved without difficulty. The actual algorithm used for this dissertation will be discussed in Chapter V.

In summary, we have shown that it is practical to compute the probabilities derived in section 4.1 and thus the conditional risk is useful when attempting to discriminate between two AEV's.

4.4 Applications to Other Criteria

The perturbation technique is also applicable to other criteria for model selection. This is true because many of the criteria are essentially based on the residual sum of squares for the \( i^{th} \) model and complexity of the joint distribution of these residual sums of squares is what has hampered efforts to evaluate these criteria in a reasonable way. We are going to deal in this section with the \( C_p \) statistic as the application to the other criteria is reasonably straightforward.
Recall that the $C_p$ statistic is defined as:

$$C_p = \frac{\text{RSS}}{s^2} - (N-2)_p$$

where $s^2$ is the error mean square obtained by fitting the full model. Recall also that the use of the $C_p$ statistic is usually equivalent to choosing the model with the minimum value of $C_p/p$.

As with the AIC, one can write the $C_p/p$ for the $\lambda$th model as a quadratic form in $V(V \sim N(\mu, \Sigma))$ as:

$$C_{p\lambda}(V)/p_\lambda = V' A_\lambda V/s^2_{p\lambda}$$

where

$$A_\lambda = I - Z_\lambda Z_\lambda' Z_\lambda (Z_\lambda' Z_\lambda)^{-1} Z_\lambda' .$$

Given two $C_p$'s, say $C_{p_i}$ and $C_{p_j}$, we define the following decision rule:

$$a(C_{p_i}(V), C_{p_j}(V)) =$$

$$\begin{cases}
  a_1: & \text{conclude that the } i\text{th model is the "better" if } C_{p_i}(V)/p_i < C_{p_j}(V)/p_j . \\
  a_2: & \text{conclude that the } j\text{th model is the "better" if } C_{p_i}(V)/p_i \geq C_{p_j}(V)/p_j
\end{cases}$$

Again we could define a loss function and calculate an expected loss which depended on the unknown parameters $\theta$ and $\sigma^2$. However, since we cannot evaluate such a risk function, let us proceed directly to the application of the perturbation approach. Assume a perturbation of the data of the same form as in section 4.1 (i.e., $y + D$ where $D \sim N(0, s^2_1)$. Then on the basis of the perturbed data one may again ask Question 3 (i.e., How different would the data have had to have been in order to have changed the conclusion?; or what is
(assuming \( C_{p_i} / p_i < C_{p_j} / p_j \)):

\[
\Pr(D: C_{p_i} (y+D)/p_i \geq C_{p_j} (y+D)/p_j) .
\]

Again, we are able to express the above probability in terms of a conditional risk. The decision rule for the \( C \)'s based on the perturbed data is:

\[
a[C_{p_i} (y+D), C_{p_j} (y+D)] =
\]

\[
\begin{cases}
a_1: \text{conclude that the } i^{th} \text{ model is the better} \\
C_{p_i} (y+D)/p_i < C_{p_j} (y+D)/p_j
\end{cases}
\]

\[
a_2: \text{conclude that the } i^{th} \text{ model is the better} \\
C_{p_i} (y+D)/p_i \geq C_{p_j} (y+D)/p_j .
\]

Define the conditional loss function for the perturbed data as:

\[
L[C_{p_i} (y), C_{p_j} (y); a[C_{p_i} (y+D), C_{p_j} (y+D)]]
\]

<table>
<thead>
<tr>
<th>Actions</th>
<th>&quot;States of Nature&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( C_{p_i} (y)/p_i \leq C_{p_j} (y)/p_j )</td>
</tr>
<tr>
<td>( a_1: \text{assumed the } i^{th} \text{ model is better} )</td>
<td>0</td>
</tr>
<tr>
<td>( a_2: \text{assumed the } j^{th} \text{ model is better} )</td>
<td>1</td>
</tr>
</tbody>
</table>

Note that the "states of nature" for the conditional loss function contain the \( C \)'s computed from the original data, not parameters. This reflects the conditional (upon the data) nature of the perturbation inference. Based upon the preceding decision rule and loss
function, the conditional risk is:

\[
R(C_{p_i}(y), C_{p_j}(y)) =
\begin{cases}
\Pr[D: C_{p_i}(y+D)/p_i \geq C_{p_j}(y+D)/p_j] & \text{if } C_{p_i}(y)/p_i < C_{p_j}(y)/p_j \\
\Pr[D: C_{p_j}(y+D)/p_j < C_{p_i}(y+D)/p_i] & \text{if } C_{p_i}(y)/p_i \geq C_{p_j}(y)/p_j
\end{cases}
\]

For definiteness assume \( C_{p_i}(y)/p_i < C_{p_j}(y)/p_j \).

The conditional risk then becomes:

\[
R[C_{p_i}(y)/p_i < C_{p_j}(y)/p_j] = \Pr[C_{p_i}(y+D)/p_i \geq C_{p_j}(y+D)/p_j | Y].
\]

The following theory shows that we may again use Imhof's results to evaluate the conditional risk.

**Lemma 3.** The following equality holds:

\[
C_{p_i}(y)/p_i - C_{p_j}(y)/p_j = Y'AV/s^2 + b
\]

where \( A = (A_i/p_i - A_j/p_j) \) and \( b = (n-2p_i)/p_i - (n-2p_j)/p_j \).

Proof:

\[
C_{p_k}(y) = Y'AV/s^2 - (n-2p_k)
\]

thus

\[
C_{p_i}(y)/p_i - C_{p_j}(y)/p_j = Y'AV/s^2 p_i - Y'AV/s^2 p_j
\]

\[
- (n-2p_i)/p_i + (n-2p_j)/p_j
\]

\[
= Y'(A_i/p_i - A_j/p_j)V/s^2 - (N-2p_i)/p_i + (N-2p_j)/p_j
\]

\[
= Y'AV/s^2 + b .
\]

The following corollary shows that the probability mentioned above may be evaluated using Imhof's techniques.
Corollary 2. Let $V \sim N(u, cI)$. (In our application $c = s^2$ and we already have divided the quadratic form by $s^2$.) Then
\[
\Pr\{V: \frac{C_p(V)}{p_1} - \frac{C_p(V)}{p_j} \geq 0\} = \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \frac{\sin \theta(u)}{u \rho(u)} \, du
\]
where
\[
\theta(u) = \frac{1}{2} \sum_{r=1}^n \left[ \tan^{-1}(\lambda_r u) + \delta_r^2 \lambda_r u (1 + \lambda_r^2 u^2)^{-1} \right] + \frac{1}{2} \frac{b u}{u}
\]
\[
\rho(u) = \prod_{r=1}^n (1 + \lambda_r^2 u^2)^{1/4} \exp\left[ \frac{1}{2} \sum_{r=1}^n \delta_r^2 \lambda_r^2 u^2 (1 + \lambda_r^2 u^2)^{-1} \right].
\]

$\Lambda$ is the diagonal matrix of eigenvalues of the matrix $A$;

$A = P \Lambda P'$ where $P'P = PP' = I$

$A = A_1/p_1 - A_j/p_j$

$A_\lambda = I - Z_\lambda^T Z_\lambda^{-1} Z_\lambda^T$

$\delta = P'\mu/\sqrt{c}$

$P$ is the matrix eigenvectors of $A$

$b = -(N-2p_1)/p_1 + (N-2p_j)/p_j$

Proof:
By lemma 3
\[
\frac{C_1(V)}{p_1} - \frac{C_p(V)}{p_j} = \frac{s^2}{\sqrt{c}} + b.
\]
Now in our case where $s^2 = c$ this becomes
\[
\frac{s^2}{\sqrt{c}} + b = \frac{s^2}{\sqrt{c}} + b.
\]
Taking the spectral decomposition of $A$, i.e., $A = P \Lambda P'$ where

$P'P = PP' = I$ and $A$ is diagonal, and making the transformation

$w = P'v/\sqrt{c}$, we obtain

\[
\Pr\{V: \frac{C_1(V)}{p_1} - \frac{C_p(V)}{p_j} \geq 0\} = \Pr\{w' \Lambda w \geq -b\} \text{ where } \Lambda \text{ is}
\]
the matrix of eigenvalues of A.

Now by theorems 1, 2, and 3

\[ \Pr(\mathbf{w}' \Lambda \mathbf{w} = \sum_{i=1}^{n} \lambda_i w_i^2 \geq -b) = 1/2 + 1/\pi \int_{0}^{\infty} \frac{\sin \theta(u)}{u \rho(u)} \, du \]

where

\[ \theta(u) = 1/2 \sum_{r=1}^{n} \left[ \tan^{-1}(\lambda_r u) + \delta_r^2 \lambda_r u (1 + \lambda_r^2 u^2)^{-1} \right] + 1/2 \, bu \]

\[ \rho(u) = \prod_{r=1}^{n} (1 + \lambda_r^2 u^2)^{1/4} \exp\left[1/2 \sum_{r=1}^{n} \delta_r^2 \lambda_r^2 u^2 (1 + \lambda_r^2 u^2)^{-1}\right] \]

and \( \delta = \mu' \mu / \sqrt{c} \). In our case \( \mu = \mathbf{y} \) and \( c = s^2 \).

Thus we see that this probability may be evaluated in much the same manner as in the case of the AEV. The numerical integration uses the same procedure for both and the arguments with respect to stability are the same.
CHAPTER V

THE ALGORITHM USED FOR THE CALCULATION OF THE CONDITIONAL RISK

This chapter contains a discussion of the algorithm which was used in the evaluation of the conditional risk for the AEV. The first section contains an outline of the major steps of the algorithm; the second section a discussion of the techniques used in the implementation of the algorithm; and the third section contains a discussion of the accuracy of the results of the computation.

5.1 The Algorithm

In this section we will discuss the algorithm used in the evaluation of the conditional risk. The algorithm proceeds in the following steps:

1. Extract $Z_i$ and $Z_j$ from the X matrix.
2. Compute $Z_i'Z_i$ and $Z_j'Z_j$.
3. Compute $(Z_i'Z_i)^{-1}$ and $(Z_j'Z_j)^{-1}$ using the Beaton Sweep arrangement of Gauss-Jordan elimination.
4. Compute $T_i$ and $T_j$.
   (a) Extract $M_{ii}$ and $M_{jj}$ if necessary.
   (b) Compute $T_i = \text{Tr}[(Z_i'Z_i)^{-1} M_{ii}]$ and
       $T_j = \text{Tr}[(Z_j'Z_j)^{-1} M_{jj}]$. 
5. Compute $k_i = T_i/(n-p_i)$ and $k_j = T_j/(n-p_j)$.

6. Compute $A_i = I - Z_i(Z_i'Z_i)^{-1} Z_i'$ and $A_j = I - Z_j(Z_j'Z_j)^{-1} Z_j'$.

7. Compute $A = k_i A_i - k_j A_j$.

8. Compute the eigenvalues, $\Lambda$, and eigenvectors (columns of $P$) of $A$: $A = P \Lambda P'$, and for accuracy checking purposes,

$||A - P \Lambda P'||_{\infty}, ||PP' - I||_{\infty}$ where $||A||_{\infty} = \max_{i,j} |a_{ij}|$.

9. Compute $\delta = P'y$.

10. Compute an approximation to

$$\int_{0}^{\infty} \frac{\sin \theta(u)}{u \rho(u)} \, du$$

where $\theta(u)$ and $\rho(u)$ are defined in Chapter IV. The integral is approximated by numerically approximating the finite integral

$$\int_{0}^{u} \frac{\sin \theta(u)}{u \rho(u)} \, du.$$

(a) Select an upper limit, $U$, for the finite interval of integration.

(b) Perform the integration numerically using Romberg integration.

11. Compute $Pr[(y + D)' A(y + D) > 0]$ as

$$1/2 + 1/\pi \int_{0}^{\infty} \frac{\sin \theta(u)}{u \rho(u)} \, du.$$

The next section contains a discussion of the techniques used in the implementation of these steps.
5.2 Implementation of the Algorithm

The algorithm was implemented using the PL/I(F) compiler on the IBM 370/165 computer at the Triangle Universities Computation Center. All calculations were performed in double precision (Binary Float (53)).

Steps 1, 2, 4, 5, 6, 7, 9, and 11 are implemented using straightforward programming. In step 4b if \( M = (X'X)/n \), \( T_1 = P_1/n \), but if \( M \neq (X'X)/n \) the trace is computed without performing the matrix multiplication using the relationship:

\[
\text{Tr}(AB) = \sum \sum a_{ij} b_{ij}
\]

where \( A \) and \( B \) are symmetric.

There are three steps of the algorithm which require further elaboration. These are the calculation of the inverses in step 3, the calculation of the eigenvalues and eigenvectors in step 8, and the numerical integration in step 10.

The inversion of \((Z'_iZ_i)\) and \((Z'_jZ_j)\) was accomplished using the sweep operator mentioned in Chapter I. This method of inversion was chosen because it was readily available (the author had already programmed it) and because it is a form of Gaussian elimination which is known to be numerically stable except when the matrix to be inverted is nearly singular (in which case almost any inversion routine yields poor accuracy).

The method of computing the matrix factorization \( A = P \Lambda P' \) is the Jacobi method. This method was used because it was readily available and is known to produce numerically accurate results, particularly when the eigenvalues are well separated (no multiple eigenvalues). The \( A \) matrix in step 7 has multiple eigenvalues but the
error checks computed for the sample problems revealed good accuracy (see next section).

One problem encountered was the determination of which eigenvalues are zero; the integration procedure fails if zero eigenvalues are input to it. The rule used for this implementation: if the ratio of the maximum eigenvalue to any given is greater in absolute value than $10^8$, the given eigenvalue is considered to be effectively zero and is set to zero in the step 10 calculations. Near zero eigenvalues have no effect on the accuracy of the calculations as their effect is overwhelmed by the larger eigenvalues. The near zero eigenvalues are set to zero as a convenience to prevent underflows during the calculations.

In step 10(a) the upper limit, $U$, of the integration is selected as follows. Imhof has shown that the error of truncation arising from integration over a finite interval satisfies

$$\left| \int_0^\infty \frac{\sin\theta(u)}{u\rho(u)} \, du \right| \leq \left\{ \frac{\pi}{2} U^{1/2} \prod_{r=1}^{n} \lambda_r^{1/2} \exp \left[ \frac{1}{2} \sum_{r=1}^{k} \frac{\lambda_r}{\lambda_r^2 + 1} \right] \right\}^{-1}.$$

Successive values of $U = 1, 2, \ldots$ are attempted until the right hand side in the expression is less than $10^{-4}$.

In step 10(b) the integral

$$\int_0^U \frac{\sin\theta(u)}{u\rho(u)} \, du$$

is approximated using the Romberg technique of successive approximations (see Wilf (1967)). The technique basically calculates successive approximations by dividing the interval $[0, U]$ into $2^k$ equal subintervals
and (basically) applying the trapezoidal rule. The algorithm terminates when a computed upper bound on the absolute error of the kth approximation is less than a specified value, which was set at $10^{-4}$ in the computer program used. The packaged program DQATR from the IBM PL/I Scientific Subroutine Package (IBM, 1966) was used without modification.

As was previously mentioned all computations were done using double precision arithmetic in order to improve accuracy. Since the program was not designed for production use, little effort was made to produce a highly efficient computer program; instead, emphasis was placed on accuracy.

5.3 Accuracy of the Algorithm

As was noted in the previous section there are only three segments of the algorithm in which we must really be concerned about the accuracy of our computation. The first of these is in the inversion of the matrices $(Z_1^T Z_1)$ and $(Z_j^T Z_j)$. As was mentioned the accuracy of the Gaussian Elimination algorithm using double precision arithmetic is well known. However, checks were also written into the program. The AEV's were calculated on the basis of the $(I - X(X'X)^{-1}X')$ matrix and compared with the AEV's calculated with Efroymson's algorithm described in Chapter I. In the examples the AEV's agreed to four significant digits.

The accuracy of the Jacobi method of matrix decomposition was checked internally. The maximum deviations in each of $A - P \Lambda P'$, $P'P - I$, and $PP' - I$ were calculated and printed. In the examples these maximum errors were of the order $10^{-16}$, $10^{-15}$ and $10^{-15}$ respectively. Thus the accuracy of the matrix decomposition is established.
The final major, potential source of error to be considered is the error in the numerical integration. Imhof points out the two sources of error in the integration; the error in using a numerical method for evaluating the integral and the error of truncation due to the fact that one has substituted a finite value, \( U \), for \( \infty \) as the upper limit of integration. Imhof has shown that the error of truncation may be bounded above by

\[
\left\{ \frac{\pi}{2} U^{1/2} \prod_{r=1}^{n} \frac{|\lambda_r|^{1/2}}{\exp \left[ \frac{1}{2} (\delta_r \lambda_r U)^2 (1 + \lambda_r^2 U^2)^{-1} \right]} \right\}^{-1}.
\]

In the computation of the conditional risk examples an upper limit of \( 10^{-4} \) was set on each of these errors. The numerical integration routine was checked by testing it on several noncentral chi-square probabilities and comparing the results with values from the program CFTDIS written by Dr. Dana Quade of the Department of Biostatistics at University of North Carolina. The probabilities from CFTDIS rounded to four decimal places were in complete agreement with results obtained using the numerical integration algorithm (step 10). The integration routine was also used to compute some values given in Imhof's article; the computed values agreed with Imhof's "true values" to the four places given in the article. These examples included several distinct positive and negative eigenvalues.

In conclusion the examples in the next chapter are believed to be accurate to at least the number of digits given.
CHAPTER VI

SOME EXAMPLES ILLUSTRATING THE USE
OF THE CONDITIONAL RISK

In this chapter we will look at the application of the perturbation approach to two sets of data which have appeared extensively in the literature. These are the Hald data and a six variable example both of which were used by Gorman and Toman (1966) and Daniel and Wood (1971) for illustrating the use of the $C_p$ statistic. These data sets may be found in Chapter VI of Daniel and Wood (1971) and Appendices I and II in this paper.

6.1 Application to the Hald Data

The Hald data consists of 13 observations on four independent variables and a response. In this section we will use the following notation:

\[ y = \text{cumulative heat of hardening after 180 days in} \]

\[ \text{calories/gram of cement} \]

\[ X_0 = \text{the intercept term (} X_0 = 1) \]

\[ X_1 = \% \text{ Tricalcium Aluminate} \]

\[ X_2 = \% \text{ Tricalcium Ferrate} \]

\[ X_3 = \% \text{ Calcium Aluminum Ferrate} \]

\[ X_4 = \% \text{ Dicalcium Silicate}. \]

The AEV's for all possible models were computed using

\[ M = (X'X)/n \]

and the conditional risk, $P[AEV_i(y+D) > AEV_j(y+D)]$, was
calculated for all comparisons among the ten smallest AEV's. This led to 45 ordered pairs such that $AEV_i(y) < AEV_j(y)$. For the Hald data, using $M = (X'X)/n$, the ten models with the smallest AEV's and their AEV's are displayed in Table 6.1. The conditional risks for the above models are numbered in the order of computation by the Garside (1966) algorithm for efficient calculation of all possible regressions.

<table>
<thead>
<tr>
<th>Model Number</th>
<th>Rank (by AEV)</th>
<th>Variables in Model</th>
<th>AEV</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>1</td>
<td>0, 1, 2</td>
<td>1.33</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>0, 1, 2, 4</td>
<td>1.64</td>
</tr>
<tr>
<td>31</td>
<td>3</td>
<td>0, 1, 2, 3</td>
<td>1.64</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>0, 1, 3</td>
<td>1.73</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>0, 1, 3, 4</td>
<td>1.74</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1, 2, 3, 4</td>
<td>1.80</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>0, 1, 2, 3, 4</td>
<td>2.30</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0, 2, 3, 4</td>
<td>2.52</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>0, 3, 4</td>
<td>4.05</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>1, 2, 4</td>
<td>4.26</td>
</tr>
</tbody>
</table>

*The models are numbered in the order of computation by the Garside (1966) algorithm for efficient calculation of all possible regressions.

mentioned ordered pairs of AEV's are displayed in Table 6.2. As one can see model 17, containing variables 0, 1, and 2, is superior, in the sense that one can be quite confident it has a smaller expected AEV, to the other models with the possible exception of model 9, containing the variables 0, 1, and 4. The conditional risk for comparing models 17 and 9 is 0.371 indicating that adding random $N(0, s^2 I)$ perturbations to
### TABLE 6.2

**CONDITIONAL RISKS FOR THE HALD DATA AND M = (X'X)/n**

<table>
<thead>
<tr>
<th>Model Number</th>
<th>16</th>
<th>31</th>
<th>9</th>
<th>8</th>
<th>2</th>
<th>1</th>
<th>4</th>
<th>5</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Variables in Model</td>
<td>0,1,2,4</td>
<td>0,1,2,3</td>
<td>0,1,4</td>
<td>0,1,3,4</td>
<td>1,2,3,4</td>
<td>0,1,2,3,4</td>
<td>0,2,3,4</td>
<td>0,3,4</td>
<td>1,2,4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Model Number</th>
<th>Rank</th>
<th>Variables in Model</th>
<th>0.097</th>
<th>0.095</th>
<th>0.371</th>
<th>0.098</th>
<th>0.083</th>
<th>0.030</th>
<th>0.046</th>
<th>0.001</th>
<th>0.028</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>1</td>
<td>0,1,2</td>
<td>.097</td>
<td>.095</td>
<td>.371</td>
<td>.098</td>
<td>.083</td>
<td>.030</td>
<td>.046</td>
<td>.001</td>
<td>.028</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>0,1,2,4</td>
<td>.498</td>
<td>.721</td>
<td>.460</td>
<td>.436</td>
<td>.029</td>
<td>.205</td>
<td>.060</td>
<td>.045</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>3</td>
<td>0,1,2,3</td>
<td>.712</td>
<td>.456</td>
<td>.421</td>
<td>.031</td>
<td>.242</td>
<td>.064</td>
<td>.044</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>0,1,4</td>
<td>.236</td>
<td>.266</td>
<td>.096</td>
<td>.124</td>
<td>.095</td>
<td>.090</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>0,1,3,4</td>
<td>.475</td>
<td>.061</td>
<td>.229</td>
<td>.069</td>
<td>.074</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1,2,3,4</td>
<td>.475</td>
<td>.081</td>
<td>.305</td>
<td>.105</td>
<td>.058</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>0,1,2,3,4</td>
<td>.639</td>
<td>.307</td>
<td>.268</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0,2,3,4</td>
<td>.202</td>
<td>.219</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>0,3,4</td>
<td>.417</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
the data would change the conclusion that $E(\text{AEV}_{17}) < E(\text{AEV}_9)$ about 37% of the time. In all other cases the conditional risk for comparing model 17 with others is less than 0.10 indicating one can place considerable confidence in the conclusion that model 17 has a smaller expected AEV than the other models in the "top ten," model 9 excluded. Thus unless there is some pressing external reason to use model 9, the model containing variables 0, 1, and 2 should be used for predictive purposes. It is interesting that analyses based on $C_p$ and Stepwise Regression/Partial F also selected this model.

We may also compare the second, third and fourth, etc. ranked models with the other models considered. One can place considerable confidence in the decisions that the second ranked model is better (in the sense described above) than the seventh, eighth, ninth and tenth ranked models, because the conditional risks of those decisions are small. However, when compared with the third, fifth, and sixth ranked models the conditional risks are roughly 0.5 indicating approximately a 50-50 chance of changing the conclusions.

The conditional risk of the decision that model 16, ranked second, has smaller expected AEV than model 9, ranked fourth, is 0.721, i.e., approximately 72% of the data perturbed by $N(\mathbf{0}, \mathbf{s}^2 \mathbf{I})$ "errors" would result in a different conclusion. It seems surprising that the decision $E(\text{AEV}_{1}(\mathbf{y})) < E(\text{AEV}_{16}(\mathbf{y}))$ on the basis of $\text{AEV}_{16}(\mathbf{y}) < \text{AEV}_{9}(\mathbf{y})$ is a poor decision in this case. This unusual situation, which cannot happen in the simple example of Chapter IV, can be attributed to a joint distribution of $(\text{AEV}_{1}(\mathbf{y}+\mathbf{D}), \text{AEV}_{16}(\mathbf{y}+\mathbf{D}))$ similar to the one depicted in Figure 6.1. The curves represent "contour lines" of the
Figure 6.1. "Contour lines" representing sets of points where the joint density of $AEV_1(y+D)$ and $AEV_j(y+D)$ takes on equal values.
joint density; i.e., each curve in the figure represents a set of points for which the joint density function takes on equal values. The figure depicts the situation in which \( AEV_i(y) \) is slightly less than \( AEV_j(y) \) but \( \text{VAR}(AEV_i(y+D)) >> \text{Var}(AEV_j(y+D)) \) and the correlation between \( AEV_i(y+D), AEV_j(y+D) \) is negative, zero or slightly greater than zero.

In such a case the greatest "volume" under the density function surface will lie over the points such that \( AEV_i(y+D) > AEV_j(y+D), \) i.e., the lower right hand area, and these are the points at which the perturbed data would result in a changed decision.

This situation can arise with any \( M \) matrix, but the situation is simplest when \( M = (X'X)/n \). In this case the equation for the variance of \( AEV_\ell, \ell = 1,j \) is

\[
\text{Var}[AEV_\ell(y+D)] = \frac{2s^2 p_\ell^2}{(n-p_\ell)n^2} + \frac{4s^2 p_\ell}{(n-p_\ell)n} \cdot AEV_\ell(y).
\]

Within a particular problem \( n \) and \( s^2 \) are fixed. If \( AEV_i \) is only slightly less than \( AEV_j \) then \( p_\ell \) and \( n-p_\ell \) are the controlling terms: larger \( p_\ell \) (more variables in the model) means larger variance of \( AEV_\ell(y+D) \). From the figure, if \( AEV_i(y) \approx AEV_j(y) \) and \( \text{Var}[AEV_i(y+D)] > \text{Var}[AEV_j(y+D)] \) one expects the joint density of \( AEV_i(y+D), AEV_j(y+D) \) to be approximately symmetric with respect to the \( AEV_i = AEV_j \) line and the conditional risk (probability of changing the conclusion with perturbed data) to approximately 0.5. However, as in the situation depicted, if \( AEV_i(y) \) is slightly less than \( AEV_j(y) \) and \( \text{Var}[AEV_i(y+D)] >> \text{Var}[AEV_j(y+D)] \), one would expect the conditional risk to be greater than 0.5. This situation arises in the \( M = (X'X)/n \) case when two models with different numbers of variables have
approximately the same \( \text{AEV}(y) \); the \( \text{Var}[\text{AEV}(y+D)] \) is larger for the model with more variables, leading to the situation depicted in the figure.

The result has considerable intuitive appeal. In the \( M = (X'X)/n \) case the net effect is, when two models have essentially equal values of \( \text{AEV}(y) \) the model with the smaller number of variables is preferred. More generally, if the \( \text{AEV} \)'s are approximately equal, the model with the smaller \( \text{Var}[\text{AEV}(y+D)] \) is preferred.

In the example model 31 has a slightly smaller \( \text{AEV} \) than model 9 but a larger \( \text{Var}[\text{AEV}(y+D)] \); as a result, the probability of perturbed data changing the conclusion is high, 0.71. In this case, even though the third ranked model has a slightly smaller \( \text{AEV}(y) \), the fourth ranked model is preferred.

The fourth ranked model, as was the first, is apparently superior to all models with higher \( \text{AEV} \)'s. The maximum conditional risk (for comparison with models in the "top ten") is 0.266, the conditional risk of the decision that the fourth ranked model is better than the sixth ranked model. As a matter of fact, if the first ranking model were not present or were undesirable for other reasons, the fourth ranked (in \( \text{AEV} \)) model would be the model of choice for predictive purposes for any region of interest—weight function represented by the \( M = (X'X)/n \) matrix.

6.2 A Six Variable Example

This data set consists of 31 observations of six variables and a response which are as follows.
\[ x_0 = \text{intercept term} \quad (x_0 = 1) \]

\[ x_1 = \log \text{ (viscosity of asphalt)} \]

\[ x_2 = \text{percentage of asphalt in surface course} \]

\[ x_3 = \text{percentage of asphalt or base course} \]

\[ x_4 = \text{indicator variable to separate two sets of runs} \]

\[ (= -1 \text{ for 1st set, } +1 \text{ for second set}) \]

\[ x_5 = \text{percentage of fines in surface course} \]

\[ x_6 = \text{percentage of voids in surface course} \]

\[ y = \log \text{ (change of rut depth in inches per million wheel passes)} \]

In this case the AEV's were calculated using two different M matrices. The first of these was the data moment matrix, \( M = (X'X)/n \), and the second was the one displayed in Table 6.3. (As M is symmetric only the upper triangular portion is displayed.) This M, taken from Helms (1973), was constructed to represent a nonartificial M matrix. In both cases it was again decided to look in detail only at the ten smallest AEV's which are displayed in Tables 6.4 and 6.5. Upon examination of these two tables, it is immediately obvious that variable \( x_2 \) is an important variable, as it appears in every model in both tables. The conditional risks were calculated and are displayed in Tables 6.6 and 6.7.

Let us consider first the case in which \( M = (X'X)/n \). The model with the minimum AEV is the model 82, containing variables 1 and 2. For the weight function-region of interest combination represented by this M matrix, this model is clearly better than all but the other two variable models as the conditional risks (probability of changing the conclusion that model i is better than model j) are essentially
### TABLE 6.3

**M matrix for the six variable example**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>1.18</th>
<th>5</th>
<th>5</th>
<th>.0032</th>
<th>7.30</th>
<th>9.88</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.276</td>
<td>5.90</td>
<td>5.90</td>
<td>.006426</td>
<td>8.614</td>
<td>5.7584</td>
<td></td>
</tr>
<tr>
<td></td>
<td>25.3333</td>
<td>25</td>
<td>0.16</td>
<td>36.5</td>
<td>24.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25.3333</td>
<td>0.16</td>
<td>36.5</td>
<td>24.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.03201024</td>
<td>.02336</td>
<td>.01562</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>53.2953</td>
<td>36.624</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>23.81774</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### TABLE 6.4

**Models for the six variable example ranked by AEV for M = (X'X)/n**

<table>
<thead>
<tr>
<th>Model Number</th>
<th>Rank</th>
<th>Variables in Model</th>
<th>AEV</th>
</tr>
</thead>
<tbody>
<tr>
<td>82</td>
<td>1</td>
<td>1, 2</td>
<td>.00111</td>
</tr>
<tr>
<td>87</td>
<td>2</td>
<td>0, 1</td>
<td>.00135</td>
</tr>
<tr>
<td>47</td>
<td>3</td>
<td>1, 2, 3</td>
<td>.00153</td>
</tr>
<tr>
<td>89</td>
<td>4</td>
<td>1, 3</td>
<td>.00156</td>
</tr>
<tr>
<td>79</td>
<td>5</td>
<td>1, 2, 4</td>
<td>.00159</td>
</tr>
<tr>
<td>110</td>
<td>6</td>
<td>1, 2, 5</td>
<td>.00169</td>
</tr>
<tr>
<td>81</td>
<td>7</td>
<td>0, 1, 2</td>
<td>.00172</td>
</tr>
<tr>
<td>94</td>
<td>8</td>
<td>1, 2, 3</td>
<td>.00173</td>
</tr>
<tr>
<td>98</td>
<td>9</td>
<td>0, 1, 2, 3</td>
<td>.00175</td>
</tr>
<tr>
<td>105</td>
<td>10</td>
<td>1, 5</td>
<td>.00193</td>
</tr>
</tbody>
</table>
TABLE 6.5
MODELS FOR THE SIX VARIABLE EXAMPLE RANKED
BY AEV FOR M AS IN TABLE 6.3

<table>
<thead>
<tr>
<th>Model Number</th>
<th>Rank</th>
<th>Variables in Model</th>
<th>AEV</th>
</tr>
</thead>
<tbody>
<tr>
<td>82</td>
<td>1</td>
<td>1, 2</td>
<td>.00116</td>
</tr>
<tr>
<td>87</td>
<td>2</td>
<td>0, 1</td>
<td>.00136</td>
</tr>
<tr>
<td>104</td>
<td>3</td>
<td>0, 1, 5</td>
<td>.00147</td>
</tr>
<tr>
<td>41</td>
<td>4</td>
<td>0, 1, 6</td>
<td>.00154</td>
</tr>
<tr>
<td>47</td>
<td>5</td>
<td>1, 2, 6</td>
<td>.00156</td>
</tr>
<tr>
<td>89</td>
<td>6</td>
<td>1, 3</td>
<td>.00167</td>
</tr>
<tr>
<td>24</td>
<td>7</td>
<td>0, 1, 5, 6</td>
<td>.00178</td>
</tr>
<tr>
<td>105</td>
<td>8</td>
<td>1, 5</td>
<td>.00186</td>
</tr>
<tr>
<td>110</td>
<td>9</td>
<td>1, 2, 5</td>
<td>.00189</td>
</tr>
<tr>
<td>18</td>
<td>10</td>
<td>1, 2, 5, 6</td>
<td>.00211</td>
</tr>
</tbody>
</table>
TABLE 6.6

CONDITIONAL RISKS FOR THE SIX VARIABLE EXAMPLE AND \( m = (X'X)/n \)

<table>
<thead>
<tr>
<th>Model Number</th>
<th>87</th>
<th>47</th>
<th>89</th>
<th>79</th>
<th>110</th>
<th>81</th>
<th>94</th>
<th>48</th>
<th>105</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Variables in Model</td>
<td>0,1</td>
<td>1,2,6</td>
<td>1,3</td>
<td>1,2,4</td>
<td>1,2,5</td>
<td>0,1,2</td>
<td>1,2,3</td>
<td>0,1,2,6</td>
<td>1,5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model Number</th>
<th>Rank</th>
<th>Variables in Model</th>
<th>( p )</th>
<th>( \beta )</th>
<th>( \gamma )</th>
<th>( \delta )</th>
<th>( \epsilon )</th>
<th>( \zeta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>82</td>
<td>1,2</td>
<td>.488</td>
<td>.000</td>
<td>.478</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>87</td>
<td>0,1</td>
<td>.000</td>
<td>.490</td>
<td>.002</td>
<td>.000</td>
<td>.004</td>
<td>.003</td>
<td>.005</td>
</tr>
<tr>
<td>47</td>
<td>1,2,6</td>
<td>.946</td>
<td>.497</td>
<td>.492</td>
<td>.490</td>
<td>.490</td>
<td>.030</td>
<td>.498</td>
</tr>
<tr>
<td>89</td>
<td>1,3</td>
<td>.068</td>
<td>.020</td>
<td>.040</td>
<td>.039</td>
<td>.022</td>
<td>.482</td>
<td></td>
</tr>
<tr>
<td>79</td>
<td>1,2,4</td>
<td>.495</td>
<td>.494</td>
<td>.493</td>
<td>.088</td>
<td>.559</td>
<td></td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>1,2,5</td>
<td>.498</td>
<td>.498</td>
<td>.102</td>
<td>.685</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>0,1,2</td>
<td>.500</td>
<td>.113</td>
<td>.670</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>94</td>
<td>1,2,3</td>
<td>.117</td>
<td>.676</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>0,1,2,6</td>
<td>.871</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**TABLE 6.7**

<table>
<thead>
<tr>
<th>Model Number</th>
<th>87</th>
<th>104</th>
<th>41</th>
<th>47</th>
<th>89</th>
<th>24</th>
<th>105</th>
<th>110</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Rank</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>Variables in Model</td>
<td>0,1,5</td>
<td>0,1,5</td>
<td>0,1,6</td>
<td>1,2,6</td>
<td>1,3</td>
<td>0,1,5,6</td>
<td>0,1,5,6</td>
<td>1,5</td>
<td>1,2,5,1,2,5,6</td>
</tr>
</tbody>
</table>

| Number | 82 | 1   | 1,2 | .119 | .040 | .040 | .102 | .110 |
| Rank in Model | 104 | 2   | 0,1 | .001 | .021 | .053 | .380 | .873 |
| Variables in Model | 5   | 4   | 0,1 | .005 | .003 | .004 | .022 | .873 |

| Number | 47 | 5   | 1,2,6 | 6   | 89 | 6   | 1 | 3 |
| Rank in Model | 24 | 7   | 0,1 | 5 | 105 | 8   | 1 | 5 |
| Variables in Model | 5,6 | 110 | 9   | 1,2,5,1,2,5,6 | 5 | 110 | 9   | 1,2,5,1,2,5,6 |

Note: The table continues with more rows and columns with similar data. The values in the table are conditional risks for the six variable example and are as in Table 6.3.
zero. When comparing two-variable models with each other one observes conditional risks of the order of 0.5 implying that perturbing the data with $N(0, \sigma^2 I)$ "errors" is just as likely to change the conclusion as not. The same is true when comparing three-variable models with other three-variable models. There is a confusing aspect to Table 6.6 in that the tenth ranked model is "equivalent" (in the sense of having a conditional risk of about 0.5 when compared to another model) to the other two-variable models and some of the three-variable models as well. At the same time, the conditional risks indicate that the other two-variable models are superior to the three-variable models. Because of this confusion one would choose the model to be used from among the models in which the decision rule is supported by the conditional risk in a clear cut manner (i.e. one would choose either the first, second or fourth ranked model).

In the case where $M$ is as in Table 5.3 there is virtually no question that model 82 is the best model in the sense of having the smallest $E(AEV)$. The conditional risks indicate that one can have substantial confidence in the conclusion that the first four models (ranked by AEV) are in their "proper" order. Also the last two models are apparently in their "proper" position. There is little doubt that the model of choice for these sample data and the $M$ matrix of Table 6.6 is model 82, containing variables 1 and 2.

In addition to the conditional risk calculations, the two approximate probabilities arising from questions Q1 and Q2 of Chapter IV were calculated using the $M$ matrix of Table 6.3 and the computer program CPTDIS mentioned in the previous chapter. These two probabilities are essentially noncentral $\chi^2$'s with noncentrality parameters
<table>
<thead>
<tr>
<th>Model \textsuperscript{j}</th>
<th>Model Number</th>
<th>87</th>
<th>104</th>
<th>41</th>
<th>47</th>
<th>89</th>
<th>24</th>
<th>105</th>
<th>110</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Variables in Model</td>
<td>0,1</td>
<td>0,1,5</td>
<td>0,1,6</td>
<td>1,2,6</td>
<td>1,3</td>
<td>0,1,5,6</td>
<td>1,5</td>
<td>1,2,5</td>
<td>1,2,5,6</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model \textsuperscript{i}</th>
<th>Model Number</th>
<th>82</th>
<th>87</th>
<th>104</th>
<th>41</th>
<th>47</th>
<th>89</th>
<th>24</th>
<th>105</th>
<th>110</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variables in Model</td>
<td>1,2</td>
<td>.004</td>
<td>.002</td>
<td>.001</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Variables in Model</td>
<td>0,1</td>
<td>.012</td>
<td>.007</td>
<td>.003</td>
<td>.003</td>
<td>.002</td>
<td>.001</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Variables in Model</td>
<td>0,1,5</td>
<td>.014</td>
<td>.006</td>
<td>.008</td>
<td>.004</td>
<td>.003</td>
<td>.001</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Variables in Model</td>
<td>0,1,6</td>
<td>.009</td>
<td>.013</td>
<td>.007</td>
<td>.006</td>
<td>.002</td>
<td>.001</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Variables in Model</td>
<td>1,2,6</td>
<td>.016</td>
<td>.008</td>
<td>.009</td>
<td>.002</td>
<td>.001</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Variables in Model</td>
<td>1,3</td>
<td>.014</td>
<td>.014</td>
<td>.004</td>
<td>.001</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Variables in Model</td>
<td>0,1,5,6</td>
<td>.029</td>
<td>.008</td>
<td>.003</td>
<td>.012</td>
<td>.004</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Variables in Model</td>
<td>1,5</td>
<td>.012</td>
<td>.004</td>
<td>.003</td>
<td>.005</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Variables in Model</td>
<td>1,2,5</td>
<td>.005</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Variables in Model</td>
<td>1,2,5,6</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Model Number</td>
<td>Variables in Model</td>
<td>Variables in Model</td>
<td>Variables in Model</td>
<td>Variables in Model</td>
<td>Variables in Model</td>
<td>Variables in Model</td>
<td>Variables in Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>------------------</td>
<td>------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>87</td>
<td>1, 2</td>
<td>.947</td>
<td>.951</td>
<td>.956</td>
<td>.981</td>
<td>.966</td>
<td>.946</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>104</td>
<td>0, 1</td>
<td>.869</td>
<td>.909</td>
<td>.958</td>
<td>.974</td>
<td>.974</td>
<td>.948</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0, 1</td>
<td>.883</td>
<td>.929</td>
<td>.974</td>
<td>.981</td>
<td>.966</td>
<td>.946</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>89</td>
<td>1, 2</td>
<td>.855</td>
<td>.937</td>
<td>.963</td>
<td>.956</td>
<td>.923</td>
<td>.948</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>105</td>
<td>0, 1</td>
<td>.784</td>
<td>.819</td>
<td>.843</td>
<td>.895</td>
<td>.945</td>
<td>.962</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0, 1</td>
<td>.692</td>
<td>.764</td>
<td>.824</td>
<td>.881</td>
<td>.881</td>
<td>.898</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>89</td>
<td>1, 2</td>
<td>.592</td>
<td>.739</td>
<td>.812</td>
<td>.881</td>
<td>.945</td>
<td>.962</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>105</td>
<td>0, 1</td>
<td>.395</td>
<td>.555</td>
<td>.674</td>
<td>.764</td>
<td>.807</td>
<td>.862</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0, 1</td>
<td>.18</td>
<td>.10</td>
<td>.15</td>
<td>.20</td>
<td>.25</td>
<td>.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0, 1</td>
<td>.09</td>
<td>.07</td>
<td>.12</td>
<td>.15</td>
<td>.18</td>
<td>.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>105</td>
<td>0, 1</td>
<td>.05</td>
<td>.03</td>
<td>.07</td>
<td>.09</td>
<td>.12</td>
<td>.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>89</td>
<td>0, 1</td>
<td>.05</td>
<td>.04</td>
<td>.07</td>
<td>.09</td>
<td>.12</td>
<td>.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
$\frac{1}{2} \text{AEV}_j(y)/K_j s^2$ and $\frac{1}{2} \text{AEV}_1(y)/K_1 s^2$ respectively. However, the program CPTDIS uses a definition of the noncentrality parameter without the factor $1/2$. In order to reproduce the results in Tables 6.8 and 6.9 the noncentrality parameters actually furnished to the program are

$$\text{AEV}_j(y)/K_j s^2 \text{ and } \text{AEV}_1(y)/K_1 s^2.$$ 

As suspected the probabilities of Table 6.8 are quite small and those of Table 6.9 are quite large, which is to be expected from the discussion of Section 4.2.

These tables of probabilities shed some light on the nature of the distribution of "perturbed" AEV's, $\text{AEV}(y+D)$. Consider for example the first entry in Table 6.8, viz. 0.004. This value represents the probability that the perturbed $\text{AEV}_{87}(y+D)$ is perturbed from the value 0.00136 to a value which is less than $\text{AEV}_{82}(y) = 0.00116$. The point is that $\text{AEV}(y)$ is in the left tail of the distribution of $\text{AEV}(y+D)$. For the other side note the value in the second row and second column of Table 6.9, viz. 0.951; this is the probability that the same $\text{AEV}_{87}(y+D)$ exceeds the third ranked AEV, viz. $\text{AEV}_{104}(y) = 0.00147$. By looking at the $\text{AEV}_j(y)$ values in Table 6.5 and the probabilities in the second row of Table 6.9 one obtains values of $1 - \text{CDF}[\text{AEV}_{87}(y+D)]$:
<table>
<thead>
<tr>
<th>Probability from Table 5.9</th>
<th>$1 - \text{Probability from Table 6.9} = \text{CDF}[\text{AEV}_j(y+D)]$</th>
<th>Value of $\text{AEV}_j(y+D)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.951</td>
<td>0.049</td>
<td>0.00147</td>
</tr>
<tr>
<td>0.937</td>
<td>0.063</td>
<td>0.00154</td>
</tr>
<tr>
<td>0.909</td>
<td>0.091</td>
<td>0.00156</td>
</tr>
<tr>
<td>0.885</td>
<td>0.115</td>
<td>0.00167</td>
</tr>
<tr>
<td>0.819</td>
<td>0.181</td>
<td>0.00178</td>
</tr>
<tr>
<td>0.764</td>
<td>0.236</td>
<td>0.00186</td>
</tr>
<tr>
<td>0.739</td>
<td>0.261</td>
<td>0.00189</td>
</tr>
<tr>
<td>0.555</td>
<td>0.445</td>
<td>0.00211</td>
</tr>
</tbody>
</table>

Points from the marginal distributions of the other $\text{AEV}(y+D)$ can be similarly extracted from Tables 6.8 and 6.9.

In conclusion it seems fair to state that the conditional risk of the $\text{AEV}$'s is a valuable adjunct to the $\text{AEV}$ as a criterion for model selection as the conditional risk allows one to express in quantitative terms the confidence that one has in the selection of one model over another on the basis of $\text{AEV}$ statistics.
CHAPTER VII

SUMMARY AND SUGGESTIONS FOR FUTURE RESEARCH

7.1 Summary and Discussion

In this dissertation an attempt has been made to provide tools for evaluating the reliability of decisions based on the AEV criterion for model selection. We demonstrated that the traditional distribution theory approach to such an evaluation was intractable due to the complex nature of the joint distribution of the AEV's. As an alternative we proposed the perturbation/conditional risk approach. We described this approach in general terms using a simple example and then applied the technique to the more complex situation encountered in using the AEV. This resulted in a conditional risk which can be evaluated as the probability that an indefinite quadratic form in normal random variables is greater than zero. We then proved that the techniques due to Imhof were applicable to the AEV problem and demonstrated that the desired probabilities could be computed with reasonable accuracy.

Finally we gave several examples of the use of the conditional risk in conjunction with the AEV in the selection of models for predictive purposes. These examples indicated that the AEV should be used as a criterion for model selection only in conjunction with the conditional risk, as one definitely needs a quantitative measure of the confidence one can place in decisions, or model selections, based on the AEV.
7.2 Suggestions for Future Research

There seem to be many possible extensions of the material presented in this dissertation, especially with reference to the perturbation/conditional risk approach. It would be very interesting to examine the behavior of the unconditional risk using various choices of \( X, \beta, \sigma^2, \) and \( M(W(x), \) and \( R). \) The applications of the perturbation/conditional risk approach seem well suited to a number of problems with complicated probability structures. One of the applications which springs immediately to mind is an extension of the material in Chapter III to several means (i.e., one wishes to determine the ordering of several population means where the samples are normally distributed with different means and equal variances.) There are certainly other problems to which the perturbation/conditional risk approach might be profitably applied.
LIST OF REFERENCES


APPENDIX 1

THE HALD DATA

<table>
<thead>
<tr>
<th>Obs. No.</th>
<th>$x_0$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>7.0</td>
<td>26.0</td>
<td>6.0</td>
<td>60.0</td>
<td>78.5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1.0</td>
<td>29.0</td>
<td>15.0</td>
<td>52.0</td>
<td>74.3</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>11.0</td>
<td>56.0</td>
<td>8.0</td>
<td>20.0</td>
<td>104.3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>11.0</td>
<td>31.0</td>
<td>8.0</td>
<td>47.0</td>
<td>87.6</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>7.0</td>
<td>52.0</td>
<td>6.0</td>
<td>33.0</td>
<td>95.9</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>11.0</td>
<td>55.0</td>
<td>9.0</td>
<td>22.0</td>
<td>109.2</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>3.0</td>
<td>71.0</td>
<td>17.0</td>
<td>6.0</td>
<td>102.7</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1.0</td>
<td>31.0</td>
<td>22.0</td>
<td>44.0</td>
<td>72.5</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>2.0</td>
<td>54.0</td>
<td>18.0</td>
<td>22.0</td>
<td>93.1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>21.0</td>
<td>47.0</td>
<td>4.0</td>
<td>26.0</td>
<td>115.9</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1.0</td>
<td>40.0</td>
<td>23.0</td>
<td>34.0</td>
<td>83.8</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>11.0</td>
<td>66.0</td>
<td>9.0</td>
<td>12.0</td>
<td>113.3</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>10.0</td>
<td>68.0</td>
<td>8.0</td>
<td>12.0</td>
<td>109.4</td>
</tr>
</tbody>
</table>
APPENDIX II
DATA FOR THE SIX VARIABLE EXAMPLE

<table>
<thead>
<tr>
<th>Obs. No.</th>
<th>$x_0$</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2.800</td>
<td>4.680</td>
<td>4.870</td>
<td>-1.000</td>
<td>8.400</td>
<td>4.916</td>
<td>6.750</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1.400</td>
<td>5.190</td>
<td>4.500</td>
<td>-1.000</td>
<td>6.500</td>
<td>4.563</td>
<td>13.000</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1.400</td>
<td>4.820</td>
<td>4.730</td>
<td>-1.000</td>
<td>7.900</td>
<td>5.321</td>
<td>14.750</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3.300</td>
<td>4.850</td>
<td>4.760</td>
<td>-1.000</td>
<td>8.300</td>
<td>4.865</td>
<td>12.600</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1.700</td>
<td>4.860</td>
<td>4.950</td>
<td>-1.000</td>
<td>8.400</td>
<td>3.776</td>
<td>8.250</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2.900</td>
<td>5.160</td>
<td>4.450</td>
<td>-1.000</td>
<td>7.400</td>
<td>4.397</td>
<td>10.670</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>3.700</td>
<td>4.820</td>
<td>5.050</td>
<td>-1.000</td>
<td>6.800</td>
<td>4.867</td>
<td>7.280</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1.700</td>
<td>4.860</td>
<td>4.700</td>
<td>-1.000</td>
<td>8.600</td>
<td>4.828</td>
<td>12.670</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0.920</td>
<td>4.780</td>
<td>4.840</td>
<td>-1.000</td>
<td>6.700</td>
<td>4.865</td>
<td>12.580</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.680</td>
<td>5.160</td>
<td>4.760</td>
<td>-1.000</td>
<td>7.700</td>
<td>4.034</td>
<td>20.600</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>6.000</td>
<td>4.570</td>
<td>4.820</td>
<td>-1.000</td>
<td>7.400</td>
<td>5.450</td>
<td>3.580</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>4.300</td>
<td>4.610</td>
<td>4.650</td>
<td>-1.000</td>
<td>6.700</td>
<td>4.853</td>
<td>7.000</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0.600</td>
<td>5.070</td>
<td>5.100</td>
<td>-1.000</td>
<td>7.500</td>
<td>4.257</td>
<td>26.200</td>
</tr>
<tr>
<td>Obs. No.</td>
<td>$x_0$</td>
<td>$x_1$</td>
<td>$x_2$</td>
<td>$x_3$</td>
<td>$x_4$</td>
<td>$x_5$</td>
<td>$x_6$</td>
<td>$y$</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>1.800</td>
<td>4.660</td>
<td>5.090</td>
<td>-1.000</td>
<td>8.200</td>
<td>5.144</td>
<td>11.470</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>6.000</td>
<td>5.420</td>
<td>4.410</td>
<td>-1.000</td>
<td>5.800</td>
<td>3.718</td>
<td>7.670</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>4.400</td>
<td>5.010</td>
<td>4.740</td>
<td>-1.000</td>
<td>7.100</td>
<td>4.715</td>
<td>12.250</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>88.000</td>
<td>4.970</td>
<td>4.660</td>
<td>1.000</td>
<td>6.500</td>
<td>4.625</td>
<td>0.760</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>62.000</td>
<td>5.010</td>
<td>4.720</td>
<td>1.000</td>
<td>8.000</td>
<td>4.977</td>
<td>1.350</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>50.000</td>
<td>4.960</td>
<td>4.900</td>
<td>1.000</td>
<td>6.800</td>
<td>4.322</td>
<td>1.440</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>58.000</td>
<td>5.200</td>
<td>4.700</td>
<td>1.000</td>
<td>8.200</td>
<td>5.087</td>
<td>1.600</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>90.000</td>
<td>4.800</td>
<td>4.600</td>
<td>1.000</td>
<td>6.600</td>
<td>5.971</td>
<td>1.100</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>66.000</td>
<td>4.980</td>
<td>4.690</td>
<td>1.000</td>
<td>6.400</td>
<td>4.647</td>
<td>0.850</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>140.000</td>
<td>5.350</td>
<td>4.760</td>
<td>1.000</td>
<td>7.300</td>
<td>5.115</td>
<td>1.200</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
<td>240.000</td>
<td>5.040</td>
<td>4.800</td>
<td>1.000</td>
<td>7.800</td>
<td>5.939</td>
<td>0.560</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>420.000</td>
<td>4.800</td>
<td>4.800</td>
<td>1.000</td>
<td>7.400</td>
<td>5.916</td>
<td>0.720</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td>500.000</td>
<td>4.830</td>
<td>4.600</td>
<td>1.000</td>
<td>6.700</td>
<td>5.471</td>
<td>0.470</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>180.000</td>
<td>4.660</td>
<td>4.720</td>
<td>1.000</td>
<td>7.200</td>
<td>4.602</td>
<td>0.330</td>
</tr>
<tr>
<td>28</td>
<td>1</td>
<td>270.000</td>
<td>4.670</td>
<td>4.500</td>
<td>1.000</td>
<td>6.300</td>
<td>5.043</td>
<td>0.260</td>
</tr>
<tr>
<td>Obs. No.</td>
<td>$x_0$</td>
<td>$x_1$</td>
<td>$x_2$</td>
<td>$x_3$</td>
<td>$x_4$</td>
<td>$x_5$</td>
<td>$x_6$</td>
<td>$y$</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>29</td>
<td>1</td>
<td>170.00</td>
<td>4.720</td>
<td>4.700</td>
<td>1.000</td>
<td>6.800</td>
<td>5.075</td>
<td>0.760</td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td>98.00</td>
<td>5.000</td>
<td>5.070</td>
<td>1.000</td>
<td>7.200</td>
<td>4.334</td>
<td>0.800</td>
</tr>
<tr>
<td>31</td>
<td>1</td>
<td>35.00</td>
<td>4.700</td>
<td>4.800</td>
<td>1.000</td>
<td>7.700</td>
<td>5.705</td>
<td>2.000</td>
</tr>
</tbody>
</table>

Note: $x_2$ and $y$ were transformed by taking the $\log_{10}$ transformation prior to performing the analysis.