AN ABSTRACT OF THE THESIS OF


Title: Fisher and Logistic Discriminant Function Estimation in the Presence of Collinearity

Abstract approved: Redacted for Privacy

David R. Thomas

The relative merits of the Fisher linear discriminant function (Efron, 1975) and logistic regression procedure (Press and Wilson, 1978; McLachlan and Byth, 1979), applied to the two group discrimination problem under conditions of multivariate normality and common covariance, have been debated. In related research, DiPillo (1976, 1977, 1979) has argued that a biased Fisher linear discriminant function is preferable when one or more collinearities exist among the classifying variables.

This paper proposes a generalized ridge logistic regression (GRL) estimator as a logistic analog to DiPillo's biased alternative estimator. Ridge and Principal Component logistic estimators proposed by Schaefer et al. (1984) for conventional logistic regression are shown to be special cases of this generalized ridge logistic estimator.

Two Fisher estimators (Linear Discriminant Function (LDF) and Biased Linear Discriminant Function (BLDF)) and three logistic estimators (Linear Logistic Regression (LLR), Ridge Logistic Regression (RLR) and Principal Component Logistic Regression (PCLR)) are compared in a Monte Carlo simulation under varying conditions of distance between populations, training set size and degree of collinearity. A new approach to the selection of the ridge parameter in the BLDF method is proposed and evaluated.
The results of the simulation indicate that two of the biased estimators (BLDF, RLR) produce smaller MSE values and are more stable estimators (smaller standard deviations) than their unbiased counterparts. But the improved performance for MSE does not translate into equivalent improvement in error rates. The expected actual error rates are only marginally smaller for the biased estimators. The results suggest that small training set size, rather than strong collinearity, may produce the greatest classification advantage for the biased estimators.

The unbiased estimators (LDF, LLR) produce smaller average apparent error rates. The relative advantage of the Fisher estimators over the logistic estimators is maintained. But, given that the comparison is made under conditions most favorable to the Fisher estimators, the absolute advantage of the Fisher estimators is small. The new ridge parameter selection method for the BLDF estimator performs as well as, but no better than, the method used by DiPillo.

The PCLR estimator shows performance comparable to the other estimators when there is a high level of collinearity. However, the estimator gives up a significant degree of performance in conditions where collinearity is not a problem.
Fisher and Logistic Discriminant Function Estimation in the Presence of Collinearity

by

Robert P. O'Donnell

A THESIS submitted to Oregon State University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Completed September 27, 1990
Commencement June 1991
APPROVED:

Redacted for Privacy

Professor of Statistics in charge of major

Redacted for Privacy

Head of Department of Statistics

Redacted for Privacy

Dean of Graduate School

Date thesis is presented September 27, 1990

Typed by Robert P. O'Donnell for Robert P. O'Donnell
Acknowledgements

If invention arises from inspiration and perspiration, then it could also be said that dissertations consist of innovation and motivation. This work is the result of the contributions of several people as well as my own effort.

David Thomas always made time to discuss the problems that arose and gave me numerous suggestions for improving my work. Justus Seely shared several ideas that caused me to critically rethink some of the underlying premises of this problem. And Ken Rowe helped me to improve the analysis and interpretation of the Monte Carlo portion of the dissertation.

My wife Karren was a constant source of support and patience. This work would most certainly have never been completed without her companionship, her understanding, and her love. Paula Kanarek, Norbert Hartmann and Mike Hafner (my co-workers) helped me, in a very positive way, to remember that I had a second job to pursue during my evenings and weekends.

And finally, to my mother Clara and my sister Kathy who always believed that despite the time it took, I would stay the course.
# Table of Contents

1. **FISHER AND LOGISTIC DISCRIMINATION METHODS**  
   1.1 The Two Group Discrimination Problem  
   1.2 Fisher Linear Discriminant Function (LDF) Estimation  
   1.3 Linear Logistic Regression (LLR) Discriminant Function Estimation  
   1.4 Collinearity and Discriminant Function Estimation  
   1.5 Study Objectives  

2. **LITERATURE REVIEW**  
   2.1 Relative Performance of LDF and LLR Discriminant Estimation  
   2.2 Definition and Measurement of Collinearity  
   2.3 LDF Estimation and Collinearity  
   2.4 Generalized Ridge and Logistic Regression Estimation  
   2.5 Logistic Regression Estimation and Collinearity  
   2.6 Classification Error Rates  
   2.7 Summary  

3. **A GENERALIZED RIDGE LOGISTIC ESTIMATOR**  
   3.1 Development/Rationale of the Modified LLR Estimators  
   3.2 Generalized Ridge Regression Estimation  
   3.3 Generalized Ridge Logistic Estimation  
   3.4 Related Research  

4. **DEFINITION OF ESTIMATORS**  
   4.1 The Estimators  
   4.2 Linear Discriminant Function (LDF) Estimator  
   4.3 Biased Linear Discriminant Function (BLDF) Estimator  
   4.4 Linear Logistic Regression (LLR) Estimator  
   4.5 Ridge Logistic Regression (RLR) Estimator  
   4.6 Principal Components Logistic Regression (PCLR) Estimator  

5. **DESIGN OF MONTE CARLO ESTIMATOR EVALUATION**  
   5.1 Monte Carlo Analysis of Estimators  
   5.2 Evaluation of Condition Number  
   5.3 Definition of Estimation Method Metrics  

6. **ANALYSIS OF MONTE CARLO SIMULATION**  
   6.1 Expected Actual Error Rate  
   6.2 Average Apparent Error Rate  
   6.3 Mean Square Error  
   6.4 Standard Deviation of AER  
   6.5 Additional Analyses
7. ASYMPTOTIC PERFORMANCE OF GRL ESTIMATOR
   7.1 Asymptotic Behavior of the GLR Estimator

8. SUMMARY
   8.1 Classification Performance of the Discriminant Functions
   8.2 Stability of the Discriminant Functions
   8.3 Parameter Estimation Performance of the Discriminant Functions
   8.4 Concluding Remarks

Bibliography
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Condition Number and Collinearity Parameter</td>
<td>38</td>
</tr>
<tr>
<td>6.1</td>
<td>EAER Normal Plot</td>
<td>43</td>
</tr>
<tr>
<td>6.2</td>
<td>EAER Box-Cox Analysis</td>
<td>44</td>
</tr>
<tr>
<td>6.3</td>
<td>EAER - Distance<em>Collinearity</em>Method Interaction</td>
<td>45</td>
</tr>
<tr>
<td>6.4</td>
<td>EAER - Collinearity*Method Interaction</td>
<td>47</td>
</tr>
<tr>
<td>6.5</td>
<td>EAER - Distance*Method Interaction</td>
<td>49</td>
</tr>
<tr>
<td>6.6</td>
<td>EAER - Distance*Method Interaction (No PCLR)</td>
<td>50</td>
</tr>
<tr>
<td>6.7</td>
<td>EAER - Training Set*Method Interaction</td>
<td>52</td>
</tr>
<tr>
<td>6.8</td>
<td>EAER - Training Set*Method Interaction (No PCLR)</td>
<td>53</td>
</tr>
<tr>
<td>6.9</td>
<td>AAER Normal Plot</td>
<td>55</td>
</tr>
<tr>
<td>6.10</td>
<td>AAER Box-Cox Analysis</td>
<td>56</td>
</tr>
<tr>
<td>6.11</td>
<td>AAER - Distance<em>Collinearity</em>Method Interaction</td>
<td>57</td>
</tr>
<tr>
<td>6.12</td>
<td>AAER - Collinearity*Method Interaction</td>
<td>59</td>
</tr>
<tr>
<td>6.13</td>
<td>AAER - Distance<em>Training Set</em>Method Interaction</td>
<td>61</td>
</tr>
<tr>
<td>6.14</td>
<td>AAER - Distance*Method Interaction</td>
<td>62</td>
</tr>
<tr>
<td>6.15</td>
<td>MSE Normal Plot</td>
<td>65</td>
</tr>
<tr>
<td>6.16</td>
<td>MSE Box-Cox Analysis</td>
<td>66</td>
</tr>
<tr>
<td>6.17</td>
<td>MSE - Collinearity*Method Interaction</td>
<td>67</td>
</tr>
<tr>
<td>6.18</td>
<td>MSE - Distance*Method Interaction</td>
<td>70</td>
</tr>
<tr>
<td>6.19</td>
<td>MSE - Training Set*Method Interaction</td>
<td>72</td>
</tr>
<tr>
<td>6.20</td>
<td>SD(AER) Normal Plot</td>
<td>73</td>
</tr>
<tr>
<td>6.21</td>
<td>SD(AER) Box-Cox Analysis</td>
<td>75</td>
</tr>
<tr>
<td>6.22</td>
<td>SD(AER) - Distance<em>Collinearity</em>Method Interaction</td>
<td>76</td>
</tr>
<tr>
<td>6.23</td>
<td>SD(AER) - Collinearity*Method Interaction</td>
<td>77</td>
</tr>
<tr>
<td>6.24</td>
<td>SD(AER) - Distance*Method Interaction</td>
<td>80</td>
</tr>
<tr>
<td>6.25</td>
<td>SD(AER) - Training Set*Method Interaction</td>
<td>81</td>
</tr>
<tr>
<td>6.26</td>
<td>EAER - Effect of Training Set Size</td>
<td>83</td>
</tr>
<tr>
<td>6.27</td>
<td>EAER - Sensitivity of PCLR to Collinearity</td>
<td>86</td>
</tr>
</tbody>
</table>
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>EAER - Distance<em>Collinearity</em>Method Decomposition</td>
<td>46</td>
</tr>
<tr>
<td>6.2</td>
<td>EAER - Collinearity*Method Decomposition</td>
<td>48</td>
</tr>
<tr>
<td>6.3</td>
<td>EAER - Distance*Method Decomposition</td>
<td>48</td>
</tr>
<tr>
<td>6.4</td>
<td>EAER - Training Set*Method Decomposition</td>
<td>51</td>
</tr>
<tr>
<td>6.5</td>
<td>AAER - Distance<em>Collinearity</em>Method Decomposition</td>
<td>58</td>
</tr>
<tr>
<td>6.6</td>
<td>AAER - Collinearity*Method Decomposition</td>
<td>58</td>
</tr>
<tr>
<td>6.7</td>
<td>AAER - Distance<em>Training Set</em>Method Decomposition</td>
<td>60</td>
</tr>
<tr>
<td>6.8</td>
<td>AAER - Distance*Method Decomposition</td>
<td>63</td>
</tr>
<tr>
<td>6.9</td>
<td>MSE - Collinearity*Method Decomposition</td>
<td>68</td>
</tr>
<tr>
<td>6.10</td>
<td>MSE - Distance*Method Decomposition</td>
<td>69</td>
</tr>
<tr>
<td>6.11</td>
<td>MSE - Training Set*Method Decomposition</td>
<td>71</td>
</tr>
<tr>
<td>6.12</td>
<td>SD(AER) - Distance<em>Collinearity</em>Method Decomposition</td>
<td>74</td>
</tr>
<tr>
<td>6.13</td>
<td>SD(AER) - Collinearity*Method Decomposition</td>
<td>78</td>
</tr>
<tr>
<td>6.14</td>
<td>SD(AER) - Distance*Method Decomposition</td>
<td>79</td>
</tr>
<tr>
<td>6.15</td>
<td>Comparison with DiPillo (1976) Study</td>
<td>84</td>
</tr>
</tbody>
</table>
1.1. The Two Group Discrimination Problem

A classic problem in multivariate statistics is that of classifying vector observations into one of two populations. Suppose that a random vector \( x \) can arise from either of two \( p \)-dimensional normal distributions that have identical covariance structure but differing mean vectors:

\[
\begin{align*}
  x & \sim N_p (\mu_1, \Sigma) \text{ with probability } \pi_1 \\
  x & \sim N_p (\mu_2, \Sigma) \text{ with probability } \pi_2
\end{align*}
\]

and \( \pi_1 + \pi_2 = 1 \).

We wish to specify a function of the \( x \) vector (the classification variables) that can be used to classify the vector into one of the two populations and will yield the smallest possible expected probability of misclassification. When \( \mu_1, \mu_2, \Sigma, \pi_1, \) and \( \pi_2 \) are known, it has been shown that an optimal rule (minimizing the expected probability of misclassification) is obtained using the linear classification function:

\[
f(x) = \beta_0 + \beta'x
\]
with \[ \beta_0 = \ln(\pi_1/\pi_2) - 1/2 (\mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2) \] 1.4

and \[ \hat{\beta} = (\mu_1 - \mu_2)/\Sigma^{-1}. \] 1.5

The vector is assigned to population 1 if \( f(x) > 0 \) and to population 2 if \( f(x) \leq 0 \).

1.2 Fisher Linear Discriminant Function (LDF) Estimation

When one does not have knowledge of these parameters, the means of specifying an optimal rule is no longer so straightforward. The statistician will usually have a set (termed a training set) \((y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)\) of random members of the two populations with \( y_j \) indicating the population of origin for \( x_j \) such that

\[
y_j = 1 \quad \text{with probability } \pi_1
\]

\[
y_j = 0 \quad \text{with probability } \pi_2
\]

and \( x_j | y_j \sim N_p(\mu_{y_j}, \Sigma) \)

From this training set, estimates of \( \beta_0 \) and \( \hat{\beta} \) are derived and used to specify the classification rule. The first solution proposed for this problem was Fisher's Sample Linear Discriminant Function (LDF). The LDF method assumes that the training set is composed of independent observations \((n_1 \text{ from population 1 and } n_2 \text{ from population 2})\) and uses the maximum likelihood estimates of the unknown parameters:

\[
\hat{\mu}_1 = \sum_{\{x_j : y_j = 0\}} x_j / n_1
\] 1.6

\[
\hat{\mu}_2 = \sum_{\{x_j : y_j = 1\}} x_j / n_2
\] 1.7
\[ \hat{\Sigma} = \frac{1}{n} \left[ \sum_{\{x_j : y_j = 0\}} (x_j - \mu_1) (x_j - \mu_1)' + \sum_{\{x_j : y_j = 1\}} (x_j - \mu_2) (x_j - \mu_2)' \right] \]

and \( \hat{\pi}_1 = \frac{n_1}{n} , \quad \hat{\pi}_2 = \frac{n_2}{n} . \)

1.3 Linear Logistic Regression (LLR) Discriminant Function Estimation

A second method for deriving a classification rule is based on the result from Bayes' Theorem (described by Efron, 1975) that gives the a posteriori log odds ratio for population 1 versus population 2 producing the observed \( x \) as:

\[ f(x_j) = \ln \left[ \frac{\pi_1(x_j)}{\pi_2(x_j)} \right] \]

where \( \pi_1(x_j) = \Pr\{ y_j = 1 | x_j \} \) and \( \pi_2(x_j) = \Pr\{ y_j = 0 | x_j \} \)

Given \( x_1, x_2, \ldots, x_n \), the \( y_j \) are conditionally independent random variables and we may model the probability structure using the logistic function as:

\[ \Pr\{ y_j = 1 | x_j \} = \frac{\exp(\beta_0 + \beta'x_j)}{1 + \exp(\beta_0 + \beta'x_j)} \]

and

\[ \Pr\{ y_j = 0 | x_j \} = 1 - \Pr\{ y_j = 1 | x_j \} \]

The parameters \( \beta_0 \) and \( \beta \) are estimated using a conditional maximum likelihood estimation procedure. The classification rule is to assign vector \( x_j \) to population 1 if \( f(x_j) > 0 \) and to population 2 if \( f(x_j) \leq 0 \). This method is known as linear logistic regression.
classification (LLR).

1.4 Collinearity and Discriminant Function Estimation

A series of articles by DiPillo (1976, 1977 and 1979), which will be described in more detail in the next chapter, has examined the impact of collinearity between the classifying variables on the performance of the LDF estimator. His analysis suggests that such collinearities can degrade the ability of the LDF method to produce discriminant functions that can do a satisfactory job of correctly classifying members to the two populations. He goes on to propose a biased linear discriminant function (BLDF), which is analogous to the biased estimators used in ridge regression in classical regression, as a desirable alternative.

1.5 Study Objectives

The goals of this study are to:

1) Develop modified LLR based classification rules, which are analogs to the BLDF, in the logistic regression setting, and show them to be special cases of a generalized ridge logistic (GRL) estimator.

2) Develop and evaluate a sample based rule for choosing the ridge value \( k \), in order to extend DiPillo's work with the BLDF.

3) Compare the performance of these various classification rules under differing degrees of collinearity (ill conditioning of \( \Sigma \)), distance between populations and training set sizes.

4) Attempt to develop some insight into the asymptotic behavior of the GRL estimator.
Chapter 2
LITERATURE REVIEW

There are several lines of published research that are relevant to the present study. These are: 1) the relative performance of LDF and LLR estimation in the two population classification problem, 2) the definition and measurement of collinearity, 3) the development and performance of alternative, biased estimators to the LDF for conditions of collinearity, 4) the definition of a generalized ridge estimator for the problem of collinearity in classical regression, 5) the potential for extending this generalized ridge estimator to logistic regression estimation procedures in discrimination and 6) the choices of misclassification error rates used to evaluate the performance of the various discriminant functions. This chapter will review the research in each of these areas that contributes to the present study.

2.1 Relative Performance of LDF and LLR Discriminant Estimation

The initial work comparing LDF and LLR discriminant estimation focused on determining which method is superior. Efron (1975) computed, under conditions of multivariate normality and identical covariance structure, the asymptotic relative efficiency (ARE) of the LDF and LLR discrimination procedures. His analysis showed that the ARE is a function of the Mahalanobis Distance (Δ) between the two populations. And as the distance grows greater, since the LDF method is based on the full maximum likelihood estimator of λ(ξ), the LDF procedure becomes increasingly more efficient than the LLR method. In particular, Efron points out that when Δ is 2.5 or greater (and effective discrimination becomes possible) that the ARE of the LLR procedure drops off markedly.

However, several other writers have argued that Efron’s analysis is not necessarily the only criteria by which the procedures should be compared. McLachlan and Byth (1979) took Efron’s analysis one additional step. They argued that evaluating the merits of the
two methods by considering the ratio of their asymptotic error rates is potentially misleading. It is also important to consider the absolute sizes of these error rates, particularly for larger values of $\Delta$. The authors determined the asymptotic expansions for the expected error rates up to terms of the first order for both the LDF and LLR classification procedures. They found that even though the asymptotic error rates for the LLR procedure can be several times those for the LDF procedure, the largest absolute differences between the expected error rates is in the range of 0.02 to 0.03.

Press and Wilson (1978) took a different tack in comparing the LDF and LLR procedures. They presented a series of arguments for the advantages of the LLR approach. Some of these are:

1) Many types of underlying distributional assumptions about the classification variables lead to the same logistic formulation.

2) Violations of the multivariate normality-identical covariance assumptions result in the LDF estimators being inconsistent.

3) Use of LDF estimators may hide the presence of characteristics in a training set that suggest the investigator should exercise caution in model estimation and interpretation.

4) When multivariate normality is not satisfied, LLR estimators are functions of sufficient statistics while LDF estimators are not.

The relative performance of the LDF and LLR methods have also been investigated under different conditions than just the multivariate normal - common covariance setting. Bayne et al. (1983) investigated the performance of LDF and LLR procedures in the situation where a quadratic model is used for classification. They considered three types of bivariate distributions for the classifying variables: the bivariate normal distribution with equal and unequal covariance structure, a mixture distribution of Bernoulli and Normal
random variables, and a bivariate Bernoulli distribution. Training sets were generated from the distributions, the parameters for each discrimination procedure were estimated, and the expected misclassification probabilities were computed under each method. The authors concluded that the LDF discrimination rules were typically preferable (though not always by a large margin) in terms of minimizing the expected misclassification probability.

Another interesting study was reported by O'Neill (1980). He extended Efron’s work in the following manner. O'Neill determined the large sample distribution of an arbitrary estimator of the optimal classification rule. He also developed an expression for the asymptotic distribution of the logistic regression estimator. He computed relative efficiencies for several distributions other than the normal, comparing in each case the appropriate maximum likelihood estimator of the classification rule to the logistic regression estimator. He concluded that, compared to the maximum likelihood estimator, the logistic regression estimator is relatively inefficient and that whenever possible the maximum likelihood estimator should be used in preference to the logistic regression estimator.

2.2 Definition and Measurement of Collinearity

Before addressing the relationship of collinearity to discriminant analysis, it is important to consider exactly what is meant when it is stated that a collinearity exists, be it in classical regression or in discrimination. A book by Belsley, Kuh and Welsh (1980) addresses at length the impact of collinearity in ordinary least squares regression, the various definitions of collinearity that have been suggested and offers their approach to dealing with collinearity.

The authors note that, while the idea of collinearity can in a simple sense be defined as the presence of a strong correlation between two or more of the variates of the data matrix X, the most useful definition must be more complex to fully describe the
phenomena. They review the techniques that have been suggested for
the detection of collinearity: comparison of the hypothesized
coefficient signs to those estimated from the data, examination of
the correlation matrix of the explanatory variables (R),
consideration of the Variance Inflation Factors (VIF), a measure
based on the determinant of R, examination of bunch maps and the
study of the eigenvectors and eigenvalues of R. The strengths and
shortcomings of each approach are discussed.

The definition and metric they eventually settle on is the
condition number, defined to be the ratio of the largest to smallest
singular values of X (the squareroots of the eigenvalues of the X'X
matrix). This metric is proposed to be directly related to the
degree of collinearity (ill conditioning) in a given data set. The
larger the condition number, the greater the collinearity (the more
ill conditioned the data). An associated concept, the condition
indices (the ratios of the largest singular value to each of the
other singular values), is also introduced as a means of handling the
situation where there are two or more sources of collinearity in a
data set. And in addition, a method of relating the singular values
to the estimated variances of the regression coefficients provides a
means of describing the specific impact of the existing collinearity
on the performance of the regression estimation. The authors go on
to show that applications of this metric to real data sets indicate
that a condition number of approximately 30 or more is apt to result
in numerically unstable parameter estimation.

The authors comment on an interesting aspect of the
collinearity problem in the linear regression setting. They state
(pg. 86) : "... collinearity is a data problem, not a statistical
problem." They later go on to describe the statistical impact, as
opposed to the computational impact, of collinearity. The
introduction of new and well conditioned data is described as the
best solution to collinearity. When the problem of collinearity is
investigated in the context of discriminant function estimation, a
different situation exists. Here, the collinearity is a dependence
of some kind between variables that are random rather than fixed.
The collinearity directly impacts the estimation of the covariance structure and the discriminant function coefficients. To more fully understand the interrelation of collinearity and discriminant estimation, it is necessary to describe the impact of collinearity on the estimation of the covariance matrix, the estimation of the discriminant function coefficients and the performance of the discriminant function in the correct classification of members of the two populations.

The definition and measurement of collinearity espoused by Belsley et al. (1980) is by no means universally accepted. DiPillo (1976, 1977, 1979) uses a simpler definition specific to the discriminant problem and the simulation setting he examined. In his first study (1976), he defined an initial covariance matrix that is based on a real data set concerning shellfish. This initial structure was then augmented so that, for the two sample discriminant problem, one parameter controlled the level of collinearity, according to his definition. When the collinearity parameter (σ) equals zero the covariance matrix is singular and DiPillo would say that collinearity is maximized. As σ increases, the collinearity decreases. DiPillo does not state whether collinearity decreases without bound for increasing σ. DiPillo's reasoning is that singularity is equivalent to the most extreme degree of collinearity and that the greater the size of σ, the less the degree of collinearity.

For another perspective on collinearity we consider the work of Schaefer (1986). Schaefer was interested in developing alternatives to MLE logistic regression estimators under conditions of collinearity for standard applications of logistic regression. He defined collinearity in terms of the maximum R² values (coefficients of determination) among the independent variables. In his Monte Carlo study, the theoretical covariance matrix starts out as a simple identity matrix and is then modified as necessary to induce the desired R² value. He then computed and averaged the Squared Error, \((\hat{\beta} - \beta)'(\hat{\beta} - \beta)\), for each simulation run. The resulting MSE values were then compared.
2.3 LDF Estimation and Collinearity

Next, we consider what is known about the impact of collinearity on LDF and LLR estimation. DiPillo (1976, 1977, 1979) argued that collinearity could have a negative impact on discriminant analysis in a manner similar to what happens in classical regression. He cited a paper by Bartlett (1939) where the author referred to "the instability of the individual coefficients due to high correlations (real or apparent) between some of the variables." DiPillo went on to propose an alternative to Fisher's Linear Discriminant Function (LDF) that he argued would be superior to the LDF under conditions of collinearity. He made an analogy to ridge regression and hypothesized that the degraded performance of the LDF is due to the influence of the sample covariance matrix (S). He proposed a biased version of the LDF (called a BLDF) that substituted a ridge alternative \((S + kI)^{-1}\) for \(S^{-1}\).

DiPillo carried out a Monte Carlo study comparing the performance of the LDF and BLDF estimators under varying conditions of collinearity, training set size, and number of classifying variables. The results of the simulation indicated that the Probability of Misclassification (also referred to as the Actual Error Rate by Seber (1984)) tended to be smaller for the biased procedure. In the second of his papers, DiPillo (1977) extended the application of his BLDF estimator to the case where the covariance matrices of the two populations are not equal and to the case of more than two populations. In both cases, the results of his simulations continued to support a preference for the BLDF over the LDF under conditions of strong collinearity. In his first study, DiPillo (1976) selected the ridge parameter \(k\) to be 1.0 in all cases. In his third paper (DiPillo, 1979), he considered the problem of selecting a value for \(k\) that was optimal for minimizing the Probability of Misclassification. The value of \(k\) was incremented in small steps until the value that solved the expression

\[
\frac{\delta}{\delta k} \text{PMC}[k] = 0
\]
was obtained. His results indicated that a marked reduction in PMC could be gained by determining an optimal value for k. He briefly discussed several approaches to selecting k, such as: generalizations of the ridge parameter selection methods used in ridge regression in classical regression, choosing k in light of a decomposition of the sample Mahalanobis distance that shows the impact of k on the estimated distance, and a cross-validation procedure that is repeated for different values of k. DiPillo concluded that while the desirability of finding an optimal k was clear, the means for selecting it was not at all obvious.

At about the same time as DiPillo published his first paper describing the BLDF procedure, Smidt and McDonald (1976) were carrying out a similar line of investigation. Posing an argument not unlike DiPillo's, they also proposing using a ridge modification of the sample covariance matrix. They also used a Monte Carlo approach to evaluate the performance of their alternative ridge estimator. They generated an initial training set and an additional set of 500 observations from each of the two populations. They computed the standard Fisher discriminant function coefficients and the the number of misclassifications obtained using the rule on the training set and the additional data set. This was repeated for the biased discriminant functions obtained for 24 different values of k (and for the optimal discriminant function). This was done for population covariance structures that exhibited what the authors defined as near singularity (one small eigenvalue), perfect singularity and nonsingular (not specifically defined).

Like DiPillo, the authors found that, under conditions of near singularity and perfect singularity, the biased estimation procedure performed well when compared to the unbiased alternative. The biased estimation procedure could produce coefficient estimates that were closer to the optimal population based coefficients. In addition, while the unbiased discriminant function classified the new observations much worse than the population discriminant function, the ridge estimator regained some of this loss.

Also like DiPillo, Smidt and McDonald tried to develop a rule
for selecting an optimal ridge value. Some of their solutions were similar to those suggested by DiPillo. One method, which they called the First Relative Extremum method, showed some promise in their simulation trials. As increasing values of \( k \) are used to compute the discriminant function coefficients, the values of the coefficients change. Smidt and McDonald found that the minimum error rate for the new observations was obtained for a value of \( k \) that also accompanied or immediately preceded the first point at which one of the coefficients appeared to reach its limit and began to reverse its direction of change. They also noted that all the coefficients typically stabilized at about this value and that the ratios of the coefficients had also stabilized. This was of interest since discriminant function coefficients are unique only up to a multiplicative constant.

A few additional studies have looked at alternative, biased discriminant function estimators, although not necessarily because of problems of collinearity. Peck and Van Ness (1982) and Peck, Jennings and Young (1988) investigated using alternative estimators when the dimensionality of the covariance matrix is large relative to the sample size. Under these conditions, the estimates of the inverse of the covariance matrix, \( \Sigma^{-1} \), deteriorate. The first paper looked at the linear discriminant function case. The second paper considered the quadratic discriminant function case. In both cases, the results of the simulations supported a shrinkage estimator which is a special case of an empirical Bayes estimator attributed to Haff (1979).

### 2.4 Generalized Ridge and Logistic Regression Estimation

Two papers that discuss generalized ridge regression are of interest to us here. Goldstein and Smith (1974) developed a definition of a general class of ridge estimator that included both the ridge regression estimator originally formulated by Hoerl and Kennard (1970) and the principal component regression estimator. A paper by Campbell (1980) adopted their definition of a generalized
ridge regression estimator and argued that, in certain conditions related to small eigenvalues, a ridge type estimator will be more stable (smaller sampling variation) in problems involving both discriminant analysis and canonical variate analysis. In an interesting paper on diagnostics for logistic regression, Pregibon (1981) illustrates how the maximum likelihood estimation of the logistic regression coefficients can be expressed as the solution to an iteratively reweighted least squares computation. In the following chapter the results of these two lines of research are combined to produce an expression for a generalized ridge logistic estimator, with both a ridge logistic estimator and a principal components logistic estimator as special cases.

2.5 Logistic Regression Estimation and Collinearity

In considering whether a ridge alternative to logistic regression estimation might benefit the performance of a LLR discriminant model, it is encouraging to note research reported by Schaefer (1981, 1986). Schaefer first developed an argument that, as with conventional linear regression, collinearities among the independent variables in a logistic regression can result in unstable, imprecise estimators. He then investigated the behavior of four alternative logistic regression estimators under conditions of collinearity. Two of these estimators, a ridge estimator and a principal component estimator (defined in the next chapter) showed reduced bias and variability relative to the standard MLE under conditions of collinearity. This provides support for the proposition that the application of logistic regression to the classification problem could indeed benefit from the use of biased estimation, particularly under the presence of collinearity.

2.6 Classification Error Rates

A secondary, but none the less important, issue that must be addressed in studies of discriminant analysis methodology involves
the metrics used to measure the performance of various classification procedures. There are several metrics that have been promoted as being useful for this task when the distributions are known but the parameters are unknown. A concise summary of the most popular measures is given in Seber (1984) and will be briefly recapped here.

Let the $R_{0i}$ be the optimal regions for assignment to Population $i$ ($i = 1, 2$). The data in a training set are used to produce estimates $\hat{R}_{0i}$ of the $R_{0i}$. Let the $f_i(x|\theta_i)$ be the distribution density functions with the true parameter values and let the $f_i(x|\hat{\theta}_i)$ be the distribution density functions with estimates of the parameters (i.e. maximum likelihood parameter estimates).

The principal metrics which have been proposed are:

1) The optimum error rates -

$$e_{i,\text{opt}} = \int_{R_{0j}} f_i(x, \theta_i) \, dx$$

2.1

and

$$e_{\text{opt}} = \pi_1 e_{1,\text{opt}} + \pi_2 e_{2,\text{opt}}$$

2.2

(which is given primarily for comparison to the following metrics)

2) The "actual" error rates (conditional on the sample in the logistic case) -

$$e_{i,\text{act}} = \int_{\hat{R}_{0j}} f_i(x, \hat{\theta}_i) \, dx$$

2.3

and

$$e_{\text{act}} = \pi_1 e_{1,\text{act}} + \pi_2 e_{2,\text{act}}$$

2.4
(the quantity \( e_{\text{act}} \) is what DiPillo calls the Probability of Misclassification)

3) The expected "actual" error rates \( E[e_{1,\text{act}}] \) and

\[
E[e_{\text{act}}] = \pi_1 E[e_{1,\text{act}}] + \pi_2 E[e_{2,\text{act}}]
\]

Seber goes on to describe another group of metrics that he refers to as the intuitive estimates. These are:

1) The "plug in" error rate estimates

\[
\hat{e}_{i,\text{act}} = \int_{\tilde{R}_0} \hat{f}_i(x, \hat{\theta}_i) \, dx
\]

\text{and}

\[
\hat{e}_{\text{act}} = \pi_1 \hat{e}_{1,\text{act}} + \pi_2 \hat{e}_{2,\text{act}}
\]

2) The apparent error rates

\[
e_{i,\text{app}} = \frac{m_i}{n_i}
\]

\text{and}

\[
e_{\text{app}} = \pi_1 e_{1,\text{app}} + \pi_2 e_{2,\text{app}}
\]

where \( m_i \) of the \( n_i \) observations from group \( i \) are classified incorrectly.

3) The cross-validation method described by Lachenbruch and Mickey (1968) where the allocation rule is determined using the training set minus one observation and the rule is used to classify that one
observation. This is done for each observation in each group, giving

\[ e_{i,c} = \frac{a_i}{n_i} \quad 2.10 \]

where \( a_i \) is the number of misclassified observations in group \( i \)

and

\[ e_c = \pi_1 e_{1,c} + \pi_2 e_{2,c} \quad 2.11 \]

4) The bootstrap method described by Efron (1981) involves improving \( e_{i,\text{app}} \) by estimating and correcting its bias. The estimate of bias \( \overline{d} \) is obtained by repeated sampling with replacement from the groups of size \( n_i \). The bootstrap estimate is then

\[ e_{i,\text{boot}} = \left( \frac{m_i}{n_i} \right) + \overline{d} \quad 2.12 \]

and

\[ e_{\text{boot}} = \pi_1 e_{1,\text{boot}} + \pi_2 e_{2,\text{boot}} \quad 2.13 \]

To allow the best comparison with DiPillo's work, we will use a simulation estimate of \( E[e_{\text{act}}] \) (called the Expected Actual Error Rate (EAER)), as one of our performance metrics.

Classification rules can be effective at correctly classifying members of a specific data set, correctly classifying members of a larger population, both, or neither. Therefore, to obtain a picture of how the estimators perform in both senses, we will also use a simulation estimate of \( e_{\text{app}} \) (called the Average Apparent Error Rate - AAER).
2.7 Summary

Summarizing the key points of this chapter:

1) Both Fisher LDF and conditional maximum likelihood LLR procedures have been used to perform a two population linear discriminant function analysis.

2) Under conditions of multivariate normality and identical covariance structure, the LDF procedure is asymptotically more efficient than the LLR procedure, although the absolute difference in asymptotic error rate may be small enough to be insignificant, from a practical viewpoint.

3) Collinearity can produce coefficient estimates with inflated variance estimates in classical regression. Biased regression methods have been shown to reduce the impact of collinearity.

4) Collinearities among the classifying variables in a two population multivariate normal discrimination problem can result in classification rules with inflated probabilities of misclassification.

5) This loss of discrimination can be reduced through the use of biased estimators such as DiPillo's BLDF procedure.

6) The performance of logistic regression estimators (with regard to MSE and bias) can also deteriorate as a result of collinearities in the data.

7) This loss of logistic regression performance can also be improved by the use of biased estimators.

We are led to the hypothesis that collinearities among the classifying variables can result in increases in misclassification.
for both the logistic regression estimation and the Fisher LDF estimation of the discriminant function. Previous work with biased estimation in conventional regression and logistic regression suggests that logistic regression analogs to DiPillo's BLDF could bring about improved performance under conditions of collinearity.

The relative performance of Fisher (conventional and biased) and logistic (conventional and biased) estimators of the discriminant function, and the influence of associated factors (such as distance between populations) emerge as the key issues to examine in this study.
3.1 Development/Rationale of the Modified LLR Estimators

To develop modified LLR estimators which exhibit improved performance under conditions of collinearity, we will draw together results from a number of sources. First we develop a general form for biased estimators such as ridge and principal component estimators for the standard linear model. We then show how the mle estimation of the logistic regression estimators can be expressed in a "linear" form. Using this "linear" form, we arrive at a general expression for a biased logistic regression estimator which includes Ridge and Principal Component estimators as special cases.

3.2 Generalized Ridge Regression Estimation

Following the work of Campbell (1980), we develop the following general expression for linear model regression estimation. Let X be defined as follows:

\[ X = (1 \ x_1 \ \ldots \ x_g) \]

Utilizing the principle of spectral decomposition for symmetric matrices, we know that \((X'X)^{-1}\) can be equivalently expressed as

\[ (X'X)^{-1} = U E^{-1} U' \]  \hspace{1cm} 3.1

where \(E\) is a diagonal matrix of elements \(e_1, e_2, \ldots, e_g\) being the ordered (largest to smallest) eigenvalues of \((X'X)\) and \(U\) is a matrix consisting of the respective orthogonal eigenvectors of \((X'X)\). Extending this result to the biased estimator case, we get:

\[ (X'X + K)^{-1} = U (E + K^*)^{-1} U' \]  \hspace{1cm} 3.2
where

\[ K = \text{diag}(k_1, k_2, \ldots, k_{g+1}) \]

and

\[ K = U K^* U' \]  \hspace{1cm} (3.3)

The diagonal elements of \((E + K^*)^{-1}\) are

\[ \frac{1}{(e_i + k_i^*)} \quad i = 1, \ldots, g \]

A principal component estimator and a ridge estimator may then be considered as special cases of the general biased estimator:

\[ \hat{\beta}(K) = (X'X + K)^{-1} X'Y \]

3.4

1) Principal Components Estimator - Suppose we set \(k_i = 0\) for \(i \leq r\) and \(k_i = \infty\) for \(i > r\). This is equivalent to deciding that the \(g-r+1\) smallest eigenvalues are not notably different from zero and therefore setting them to zero.

Then our estimator is

\[ \hat{\beta}(K) = (X'X + K)^{-1} X'Y = (X'X)^+ X'Y \]

3.5

where

\[ (X'X)^+ = \sum_{i=1}^{r} e_i^{-1} u_i u_i' \]  \hspace{1cm} (3.6)
2) Ridge Estimator - If we set $k_i = k$ for all $i$, then our estimator is

$$\hat{\beta}(K) = (X'X + K)^{-1} X'Y = (X'X + kI)^{-1} X'Y$$  \hspace{1cm} 3.7

### 3.3 Generalized Ridge Logistic Estimation

Next we will formulate an analogous estimator for the logistic regression case. Following Pregibon (1981), letting $\Theta = \text{logit}(\pi)$ where $y$ is binomial$(n, \pi)$, we can write the probability function of $y$ as

$$f(y, \Theta) = \exp\{y\Theta - a(\Theta) + b(y)\}$$  \hspace{1cm} 3.8

where $a(\Theta) = n \ln(1 + e^\Theta)$ and $b(y) = \ln \left(\frac{n}{y}\right)$.

For a sample of $n$ independent binomial observations, the sample loglikelihood function can be written as

$$l(Q, y) = \sum_{i=1}^{n} l(\Theta_i, y_i) = \sum_{i=1}^{n} \{y_i\Theta_i - a(\Theta_i) + b(y_i)\}$$  \hspace{1cm} 3.9

The logistic regression model may be expressed as

$$\Theta_i = \text{logit}(\pi_i) = x_i'\hat{\beta} \hspace{1cm} i = 1, 2, \ldots, n$$  \hspace{1cm} 3.10

and $X = (\mathbf{1}, x_1, x_2, \ldots, x_g)$ is a set of explanatory variables. The loglikelihood function expressed in terms of $\hat{\beta}$ is

$$l(X, y) = \sum_{i=1}^{n} y_i x_i'\hat{\beta} - a(x_i'\hat{\beta}) + b(y_i)$$  \hspace{1cm} 3.11
The maximum likelihood estimator maximizes the above and is obtained as a solution to the system of equations

\[ \sum_{i=1}^{n} x_{ij} (y_i - \alpha'(x_i', \hat{\beta})) = 0 \quad \text{for } j = 1, 2, \ldots, g+1 \]  

or equivalently

\[ X'(y - \hat{\gamma}) = 0 \]  

The solution to this maximization problem is determined iteratively utilizing the Newton-Raphson method or some similar procedure. The vector is iterated to convergence by the process

\[ \hat{\beta}^{t+1} = \hat{\beta}^t + (X'X)^{-1} X' \hat{y}^t \]  

where

\[ \hat{\gamma}^t = y - \hat{\gamma}^t \]

and \( V \) and \( \hat{\gamma} \) are evaluated at \( t \) for \( t = 0, 1, \ldots \)

Pregibon points out that this process can be viewed as an iteratively reweighted least squares procedure. We define a pseudo observation vector \( z^t \) as

\[ z^t = X\hat{\gamma}^t + (V^t)^{-1} \hat{\gamma}^t \]

This yields the iterative equation
\[ \hat{\beta}^{t+1} = (X'V^tX)^{-1} X'V^tZ^t \]  

When the process converges, \( z \) has the form

\[ \hat{z} = X\hat{\beta} + V^{-1}s \]

and the mle of \( \beta \) can be written as

\[ \hat{\beta} = (X'VX)^{-1} X'V \hat{z} \]

Taking this expression for the logistic MLE of \( \hat{\beta} \), we can extend the idea of the generalized ridge estimator to the logistic case. The formulation of such an estimator would be

\[ \hat{\beta}_{GRL} = (X'VX + K)^{-1} X'V \hat{z} \quad K = \text{diag}(k_1, k_2, \ldots, k_{g+1}) \]

Now suppose we define \( K \) as we did above for the principal component regression estimator. This would give us

\[ \hat{\beta}_{PCLR} = (X'VX)^+ X'V \hat{z} \]

a little linear algebra yields

\[ \hat{\beta}_{PCLR} = (X'VX)^+ (X'VX) (X'VX)^{-1} X'V \hat{z} \]
and since \((X'VX)^{-1} X'Vz = \hat{\beta}\) (the mle) we have

\[
\hat{\beta}_{PCLR} = (X'VX)^+ (X'VX) \hat{\beta}
\]  

Similarly for the ridge case

\[
\hat{\beta}_{RLR} = (X'VX + kI)^{-1} X'Vz
\]  

\[
\hat{\beta}_{RLR} = (X'VX + kI)^{-1} (X'VX) (X'VX)^{-1} X'Vz
\]  

\[
\hat{\beta}_{RLR} = (X'VX + kI)^{-1} (X'VX) \hat{\beta}
\]

3.4 Related Research

What makes these special cases interesting is that they are in fact the same alternative logistic regression estimators proposed by Schaefer (1986) and Schaefer et al. (1984). Two arguments, both different from that proposed above, were developed to justify the alternative estimators in these papers.

In his 1986 paper, Schaefer makes his argument as follows. He begins with an expression of the logistic MLE as a sum of iterative steps leading to convergence.

\[
\hat{\beta} = \hat{\beta}_0 + \sum_{l=0}^{L} (X'V_1X)^{-1}X'Vz\ 1 = (X'X)^{-1}X'Vz + \sum_{l=0}^{L} (X'V_1X)^{-1}X'Vz\ 1
\]

He then replaces \((X'X)^{-1}\) and \((X'V_1X)^{-1}\) with their respective eigenvector-eigenvalue expressions. He goes on to argue that in the case where a single collinearity exists that there is no meaningful loss in deleting the first principle component. This results in a
principal component estimator expressed as

\[ \hat{\mu}_{pc} = (X'X)^+X'Y + \sum_{l=0}^{L} (X'V_1X)^+X'\delta_l \]

Next he suggests using several approximations which are permissible in the logistic setting due to the fact that small changes in \( \hat{\beta} \) result in even smaller changes in \( \hat{\gamma} \) and thus \( \hat{V} \). In addition \( \hat{\gamma} \) typically will not be near 0 or 1. This means that

\[ (X'V_1X) \simeq (X'X) \quad \text{and} \quad (X'V_1X)^+ \simeq (X'X)^+ \]

\[ \hat{V} = C^*I \]

\[ (X'\hat{V}X) \simeq C^*(X'X) \quad \text{and} \quad (X'\hat{V}X)^+ \simeq C^{-1}(X'X)^+ \]

Making these substitutions into the expression for \( \hat{\mu}_{pc} \), we obtain the \( \hat{\mu}_{PCLR} \) estimator given above.

In their 1984 paper, Schaefer, Roi and Wolfe use reasoning associated with linear model ridge estimation to support their logistic ridge estimator. When an estimator is constrained to increase the Weighted Sum of Squares Error (WSSE) by some small fixed amount and the estimator is shrunk, the \( \hat{\mu}_{RLR} \) estimator defined above is obtained. This occurs because the logistic estimator \( \hat{\mu} \) approximately minimizes the WSSE.

The argument given in this paper, based on Campbell's formulation, seems more direct than the arguments put forth by Schaefer et. al. It allows the treatment of biased estimators to be made in a more general context and does not need to appeal to additional approximations to be valid. But regardless of the
heuristic value (or lack of) of Schaefer's theoretical development, his simulation study using these estimators indicated that in situations where collinearities exist between the variates, these alternative estimators exhibit reduced bias and variability when compared to the standard MLE. These findings provide support for the concept of extending the generalized ridge estimator used in classical linear models to a form applicable to the logistic regression setting and classification models.
Chapter 4
DEFINITION OF ESTIMATORS

4.1 The Estimators

There are five estimators which will be evaluated in this study. In this section each estimator will be defined in terms of the computations needed to produce a classification rule from a sample training set. In addition, the various measures of estimator performance will be operationally defined.

4.2 Linear Discriminant Function (LDF) Estimator

This is the original estimator, proposed by Fisher, with mle estimates substituted for \( \Sigma, \mu_1 \) and \( \mu_2 \), assuming \( \pi_1 = \pi_2 = 1/2. \)

Given \( \bar{x}_1, \bar{x}_2 \) and \( S \)

\[
\hat{\beta}_0 = -\frac{1}{2} \left( \bar{x}_1' S^{-1} \bar{x}_1 - \bar{x}_2' S^{-1} \bar{x}_2 \right) \quad 4.1 
\]

\[
\hat{\beta}' = (\bar{x}_1 - \bar{x}_2)' S^{-1} \quad 4.2 
\]

Decision Rule - let \( \lambda(\bar{x}) = \hat{\beta}_0 + \hat{\beta}' \bar{x} \)

If \( \lambda(\bar{x}) > 0 \) then \( \bar{x} \) is assigned to Population 1

If \( \lambda(\bar{x}) < 0 \) then \( \bar{x} \) is assigned to Population 2

The Actual Error Rate (AER), conditional on \( \bar{x}_1, \bar{x}_2 \) and \( S \), is defined as

\[
\text{AER} = \frac{1}{2} \left[ 1 - \Phi(z_2) + \Phi(z_1) \right] \quad \text{(actual error rate)} \quad 4.3 
\]
\[ z_i = \frac{1}{2} \frac{(\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 + \bar{x}_2) - (\bar{x}_1 - \bar{x}_2)'S^{-1} \mu_i}{[((\bar{x}_1 - \bar{x}_2)'S^{-1} \Sigma S^{-1}(\bar{x}_1 - \bar{x}_2))]^{1/2}} \]

4.3 Biased Linear Discriminant Function (BLDF) Estimator

This is the biased version of Fisher's LDF, proposed by DiPillo (1976).

Given \( \bar{x}_1, \bar{x}_2 \) and \( S \)

\[ \hat{\beta}_0 = -1/2 (\bar{x}_1' (S + kI)^{-1} \bar{x}_1 - \bar{x}_2' (S + kI)^{-1} \bar{x}_2) \]

\[ \hat{\beta}' = (\bar{x}_1 - \bar{x}_2)'(S + kI)^{-1} \]

Decision Rule - Same as for LDF, but substituting the biased estimators defined above.

\[ AER = 1/2 [1 - \Phi(z_2^*) + \Phi(z_1^*)] \] (actual error rate)

where \( z_i^* = z_i \) with \((S + kI)\) substituted for \( S \)

This definition leaves one value unspecified, the value for \( k \). Resolving this issue has been a recurring theme in the work done in biased regression. What is desired is a value for \( k \) which is based on the sample data and is optimal for reducing the MSE. Many proposals have been made, none totally satisfying since it turns out that the optimal \( k \) is a function of \( \beta \), which is unknown.

This same difficulty exists for the discrimination problem.
DiPillo (1979) considers several procedures without settling on any one as preferable. A criteria for choosing $k$ for the discrimination problem, which has its roots in an analogous argument for the validity of ridge regression, will be offered here.

In conventional regression, perhaps the most common characterization of $\hat{\beta} = (X'X)^{-1}X'Y$ is that it is the least squares estimator of $\beta$. Following Hoerl and Kennard (1970), for any estimator $B$ of $\beta$, define $\psi$ as

$$\psi = (Y - XB)'(Y - XB)$$

$$= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (\hat{B} - \hat{\beta})'X'X(\hat{B} - \hat{\beta})$$

$$= \psi_{\text{min}} + \psi(B)$$

$\psi$ is the residual sum of squares. It is minimized when $B = \hat{\beta}$ (the least squares solution). Choosing any other $B$ will increase the residual sum of squares.

Hoerl and Kennard argue that when $X'X$ has a small eigenvalue that the distance between $\hat{\beta}$ and $\beta$ will tend to be large and that $\hat{\beta}$ will tend to be too "long" ($\hat{\beta}'\hat{\beta}$ too large). From this they suggest that in such a situation the correct procedure is to move away from the least squares solution (choose some $B \neq \hat{\beta}$) in a direction that will reduce the length of $B$ (reduce $B'B$). From this argument they go on to develop their ridge regression estimator:

$$\hat{\beta}(k) = (X'X + kI)^{-1}X'Y$$

In discriminant analysis, the optimal probability of misclassification is expressed as:
\[ e_{\text{opt}} = \Phi\left(-\frac{\delta}{2}\right) \quad 4.13 \]

where \( \delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2) \)

If one is estimating the population parameters from the data, then a natural estimate of \( e_{\text{opt}} \) is:

\[ \hat{e}_{\text{opt}} = \Phi\left(-\frac{\hat{\delta}}{2}\right) \quad 4.14 \]

where \( \hat{\delta}^2 = (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2) \)

It is widely agreed that this estimate tends to be optimistic, that \( \hat{e}_{\text{opt}} \) underestimates \( e_{\text{opt}} \). Let \( M \) be a matrix consisting of the orthogonal eigenvectors of \( S \) and let \( \Omega \) be a diagonal matrix composed of the eigenvalues of \( S \). As described by DiPillo (1979), \( \hat{\delta}^2 \) can be expressed as

\[ \hat{\delta}^2 = (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2) = (\bar{x}_1 - \bar{x}_2)' M \Omega^{-1} M' (\bar{x}_1 - \bar{x}_2) \]

\[ = c'\Omega^{-1}c = \sum_{i=1}^{\nu} \frac{c_i^2}{\tau_i} \quad \text{where} \quad c = M' (\bar{x}_1 - \bar{x}_2) \quad 4.15 \]

where \( \tau_j \) is the \( j \)th eigenvalue.

This means that when one or more of the eigenvalues are small, there is a greater likelihood that \( \hat{e}_{\text{opt}} \) will underestimate \( e_{\text{opt}} \) depending also on the value if the \( c_i \). The BLDF estimator is based on substituting \( (S+kI)^{-1} \) for \( S^{-1} \). When this substitution is made the estimated probability of misclassification is

\[ \hat{e}_{\text{opt}}(k) = \Phi\left(-\frac{\delta(k)}{2}\right) \quad 4.16 \]
where $\hat{\delta}^2(k) = (\bar{x}_1 - \bar{x}_2)' (S + kI)^{-1} (\bar{x}_1 - \bar{x}_2) = \sum_{i=1}^{\nu} \frac{c_i^2}{(\tau_i + k)}$

as is shown by DiPillo (1979).

Therefore, for allowable values of $k (k \geq 0)$, $\hat{e}_{opt}(0)$ minimizes $e_{opt}(k)$. Choosing some value of $k$ greater than zero will increase the value of $\hat{e}_{opt}$. But, since this estimator is quite possibly an underestimate of $e_{opt}$ due to a very small eigenvalue, such an inflation may not be unreasonable. The research by DiPillo indicates that use of some $k > 0$ can result in a distinct reduction in the actual error rate. In determining what value of $k$ to choose, it is helpful to look again at the components of $\hat{\delta}^2$ which contribute to the estimation of $e_{opt}$.

In the situation where collinearity is expressed as a single small eigenvalue, the component $\frac{c_g}{\tau_g}$ (where $\tau_g$ is the smallest eigenvalue) could be viewed as the prime contributor to the underestimation of $e_{opt}$. One could argue that

$$\hat{e}_{opt}^* = \Phi(\frac{-\hat{\delta}^*}{2})$$

where $\hat{\delta}^* = \sum_{i \neq g} \frac{c_i^2}{\tau_i}$

is a more realistic estimate of $e_{opt}$. Therefore, it would be reasonable to choose $k$ such that

$$\sum_{i=1}^{g} \frac{c_i^2}{(\tau_i + k)} = \sum_{i \neq g} \frac{c_i^2}{\tau_i}$$

The logic of this approach follows this line of thought. Due to collinearity, the estimate of $e_{opt}$ is optimistically small since the distance between the two populations is represented as being greater than it should be. By choosing some $k > 0$ the estimated
distance is reduced. But how much reduction is reasonable? The component with the most questionable contribution to the distance is that associated with the smallest eigenvalue. So reducing the distance and increasing the estimate of $e_{opt}$ by the magnitude of this component seems intuitively justified.

Therefore, the choice of $k$ that increases the estimate of $e_{opt}$ by this amount should result in a realistic inflation and should produce a BLDF estimator with improved performance when the data are collinear and the evidence points to one small eigenvalue. If investigation indicated that more than one eigenvalue is unreasonably small than the same logic would lead to a choice for $k$ that would again adjust the estimate of $e_{opt}$ by the amount represented by multiple components with small eigenvalues.

### 4.4 Linear Logistic Regression (LLR) Estimator

This is the classic maximum likelihood estimator of $\beta$ in the logistic parametrization. The Newton-Rapheson iterative algorithm is used to determine the $\hat{\beta}$ that maximizes the loglikelihood function.

Decision Rule - Same as LDF, but substituting the mle for $\beta$ in the logistic setting.

$$AER = \frac{1}{2} \left[ 1 - \Phi(z_{1,2}) + \Phi(z_{1,1}) \right] \quad \text{(actual error rate)}$$  \hspace{1cm} (4.19)

where $z_{1,i} = -\frac{(\hat{\beta}_0 + \hat{\beta}'\mu_i)}{[\hat{\beta}' \Sigma \hat{\beta}]^{1/2}}$ for $i = 1, 2$

### 4.5 Ridge Logistic Regression (RLR) Estimator

To obtain the RLR estimator we begin by obtaining the LLR estimate of $\beta$ as defined above. We then compute:
\[ \hat{R}_{RLR} = (X'VX + kI)^{-1} (X'VX) \hat{\beta} \]

where \( V \) is evaluated for \( \hat{\beta} = \hat{\beta} \)

Decision Rule - Same as LDF, but substituting \( \hat{R}_{RLR} \) for the LDF estimator.

\[ \text{AER} = \frac{1}{2} \left[ 1 - \Phi(z_{R,2}) + \Phi(z_{R,1}) \right] \quad \text{(actual error rate)} \]

where \( z_{R,i} = -\frac{(\hat{\beta}_{R0} + \hat{\beta}_{R}' \mu_i)}{[\hat{\beta}_{R}' \Sigma \hat{\beta}_{R}]^{1/2}} \) for \( i = 1, 2 \)

and \( \hat{R}_{RLR} = [ \hat{\beta}_{R0} | \hat{\beta}_{R} ] \)

As with the BLDF, there is the matter of choosing \( k \). In this case, our choice will be based on earlier work by Schaefer et al (1984). Their Monte Carlo analysis of several different procedures for choosing \( k \) in the logistic regression setting found that the most satisfactory performance came from:

\[ k = \frac{(g + 1)}{\hat{\beta}_{R}' \hat{\beta}_{R}} = \frac{6}{\hat{\beta}_{R}' \hat{\beta}_{R}} \quad (g = \# \text{ of classifying variables)} \]

4.6 Principal Components Logistic Regression (PCLR) Estimator

As with the RLR, this estimator begins with the LLR estimate of \( \hat{\beta} \). We then compute

\[ \hat{\beta}_{PCLR} = (X'VX)^+ (X'VX) \hat{\beta} \]
where \((X'\Sigma X)^+ = (X'\Sigma X)^{-1}\) with the smallest eigenvalue set to zero.

Decision Rule - Same as LDF, but substituting PC for the LDF estimates.

\[
\text{AER} = \frac{1}{2}[1 - \Phi(z_{pc,2}) + \Phi(z_{pc,1})] \quad \text{(actual error rate)}
\]

where \(z_{pc,i} = -\frac{(\hat{\beta}_{pc0} + \hat{\beta}_{pc}^i \mu_i)}{[\hat{\beta}_{pc}^i \Sigma \hat{\beta}_{pc}]^{1/2}}\) for \(i = 1, 2\)

and \(\hat{\beta}_{pCLR} = [\hat{\beta}_{pc0} \mid \hat{\beta}_{pc}]\)
5.1 Monte Carlo Analysis of Estimators

In order to assess the relative performance of the various estimators, a Monte Carlo study was undertaken. Following the format used by DiPillo (1976), the performance of the five estimators was compared under variation of three conditions. These were:

1) Size of Training Set = 50, 100 and 200 (half from each population)

2) Degree of Collinearity $\sigma = 0.001, 0.01, 1.25$ (see definition of Collinearity below)

3) Distance between Populations Mahalanobis Distance $\delta = 1, 2, 3, 4$

These choices were made to allow some degree of comparison of results obtained here with those reported by DiPillo (1976) and Efron (1975). The number of classifying variables for this study was set at 5.

The covariance matrix of the classification variables is specified in the following manner, consistent with DiPillo (1976).

$$
\Sigma = \begin{bmatrix}
A & \bar{a} \\
\bar{a}/A & \frac{\bar{a}/A}{A} + \sigma
\end{bmatrix} = \begin{bmatrix}
\begin{bmatrix}
I_4 & \bar{a} \\
\bar{a} & \frac{\bar{a}}{\bar{a}/A} + \sigma
\end{bmatrix}
\end{bmatrix}
$$

5.1
The matrix $A$, in this study, is an identity matrix $\left( I_4 \right)$. The column vector $a$ is defined as $(p - 1) \times 1$ with elements $\frac{1}{(p - 1)} (p = 5$ for this study). And $\sigma$ is a positive scalar value. When $\sigma = 0$, $\Sigma$ is a singular matrix. As $\sigma$ increases, the conditioning of $\Sigma$ improves according to DiPillo's perspective. One of the advantages of defining $\Sigma$ in this manner is that $\sigma$, representing the degree of collinearity, is independent of $\delta$, the distance between populations. This is achieved by defining the population mean vectors as follows. The initial mean vector $\mu_1^*$ is $4 \times 1$ with all entries equal to zero. The other mean vector $\mu_2^*$ is chosen with all equal values such that the Mahalanobis distance (here equal to $\mu_2^* \mu_2^*$) is the desired quantity. These initial mean vectors are then augmented in the following manner. We compute $a' \mu_i^*$ (for $i = 1, 2$) and the final mean vectors are $\mu_i = (\mu_i^*| a' \mu_i^*)$. When the mean vectors are defined in this way we obtain the specified Mahalanobis distance and can set $\sigma$ to achieve the desired degree of collinearity.

5.2 Evaluation of Condition Number

However, as has been noted earlier, the definition of collinearity as the condition number of the $XX$ matrix can be extended to the covariance matrix for this discrimination problem. We can evaluate the eigenvalues of $\Sigma$ as functions of $\sigma$ with the following result. Partitioning $\Sigma$ according to Rao (1973), we find that three eigenvalues are identically equal to 1 for all values of $\sigma$. The largest eigenvalue can be expressed as

$$\frac{\sigma}{2} + \frac{5}{8} + \sqrt{\left(\left(\frac{\sigma}{2} - \frac{5}{8}\right)^2 + \frac{\sigma}{4}\right)}$$ 5.2

and the smallest eigenvalue as

$$\frac{\sigma}{2} + \frac{5}{8} - \sqrt{\left(\left(\frac{\sigma}{2} - \frac{5}{8}\right)^2 + \frac{\sigma}{4}\right)}$$ 5.3
Examination of these eigenvalues produces the following conclusions.

1) As $\sigma \to 0$, the smaller eigenvalue goes to zero and the larger eigenvalue goes to $10/8$; and

2) as $\sigma \to +\infty$, the smaller eigenvalue goes to 1.0 and the larger eigenvalue goes to $+\infty$.

Therefore, when we look at the behavior of the condition number graphically (Figure 5.1) we find that as $\sigma$ gets either very small or very large, the value of the condition number gets very large. According to DiPillo large values of $\sigma$ indicate a small degree of collinearity. From the perspective of the condition number metric, the degree of collinearity is great for large values of $\sigma$ as well as small values.

By looking at the structure of $\Sigma$ additional insight can be obtained concerning the behavior of the classifying variables for various values of $\sigma$. For small values of $\sigma$ the classifying variables all contribute approximately equally to the overall variability. For large values of $\sigma$ the fifth classifying variable begins to dominate the overall variability and as a result dominates the discriminant function as well.

Further consideration of Figure 5.1 indicates that the condition number of $\Sigma$ reaches a minimum for some value of $\sigma$ in the neighborhood of 1. The ratio of the largest eigenvalue (5.2) to the smallest eigenvalue (5.3) is the square of the condition number and the value of $\sigma$ that minimizes this ratio will minimize the condition number. By taking the derivative of the ratio and solving for zero it is determined that the condition number is minimized for $\sigma = 1.25$. Therefore, to assess the impact of collinearity on the estimators, it was decided to use values of $\sigma = 0.001$, 0.01 and 1.25. The smallest value (0.001) produces a condition number of approximately 39.5, larger than the value of 30 mentioned by Belsley, Kuh and Welsh as being indicative of strong collinearity; and the largest value (1.25)
Condition Number and Collinearity Parameter

Figure 5.1
produces the smallest possible condition number, 1.6, for this covariance matrix, a value that should not cause problems for any of the estimators.

In order to test the effect of larger values of \( \sigma \), two supplemental Monte Carlo runs with \( \sigma = 10 \) and 100 were made. The results of these runs supported the contention that larger values of \( \sigma \) within the present structure of \( \Sigma \) result in greater collinearity rather than less collinearity. Increasing values of the Expected Actual Error Rate were produced in runs with these larger values of \( \sigma \). No additional analysis was made of these runs and the formal analysis described in the next chapter did not include any values of \( \sigma \) larger than 1.25.

The relationship of the covariance structure of the classifying variables and the resulting collinearity of a data set as described above shows, as Belsley et. al. point out, that it is difficult to relate in a simple manner the connection between correlated variables and the degree of collinearity.

Each trial of the simulation proceeds in the following way. The indicator vector (\( y \)) of the appropriate dimension is generated. The matrix \( X \) is generated such that the first partition follows the first multivariate normal population specification and the second partition follows the second multivariate normal population specification. The \( \beta \) vector is then estimated according to the definition for each estimation procedure (given in the prior section).

5.3 Definition of Estimation Method Metrics

These estimates are used to obtain several measures of estimator performance. First, the actual error rate (PMC as defined by DiPillo (1976)) was computed as described in the previous section. Next, the apparent error rate was computed, according to Seber's definition. Third, the squared error is computed for each estimator. This is defined as \( (\hat{\beta} - \beta)'(\hat{\beta} - \beta) \), where \( \hat{\beta} \) is, in turn, each of the
proposed estimators.

In the event that the iterative search procedure for the logistic \( \hat{\theta} \) produced perfect separation of the two segments of the training set, the mle estimate of \( \hat{\theta} \) is not uniquely defined and the iteration procedure was halted. The occurrence of perfect separation was noted and the trial was terminated.

At the completion of 100 trials, a summary report for the run was printed. It contained the following information for each of the estimation procedures:

1) EAER (The expected actual error rate for the run) - This measure estimates the expected actual error rate for the estimator relative to the larger population from which the samples are drawn. Smaller values indicate that the estimator will do a better job of correctly classifying members of each population.

2) AAER (The average apparent error rate for the run) - This measure is considered to belong to the same general class of performance metrics as the EAER. However, while the AAER is considered a overly optimistic estimate of the general performance of a classification rule, it does have the characteristic of being an appropriate measure of the ability of a rule to adapt to a specific data set. By comparing an estimator's performance for both of these metrics, it is possible to gauge the tendency of an estimator to work well for a particular data set, possibly at the expense of working less well for the parent population.

3) MSE (The sample average Squared Error for the run) - This measure evaluates another aspect of the performance of the estimators. While the EAER measures the probability of a rule misclassifying cases, the MSE measures the specific degree to which the parameter estimates deviate from the true values. There will be a relationship between the two metrics, but it is quite possible that one estimator which outperforms another with respect to one metric shows no meaningful difference for the other metric.
Mean Squared Error = \frac{\sum_{i=1}^{R} (\hat{\beta}_i - \beta)'(\hat{\beta}_i - \beta)}{n}

4) SD(AER) (The sample standard deviation of the AER values for the run) - This measure is meant to represent the relative stability of the estimator. Smaller values indicate that the estimator is more resistant to sampling variation.

The first four summary measures defined above were computed for the trials on which perfect separation did not occur. So each measure for a given simulation run is based on the same number of trials. One such simulation run was made for each combination of conditions (Training Set Size = 3, Collinearity = 3, Mahalanobis Distance = 4) for a total of 48 runs. The following chapter details the analysis of the data obtained from these runs.
Chapter 6
ANALYSIS OF MONTE CARLO SIMULATION

6.1 Expected Actual Error Rate

An Expected Actual Error Rate (EAER) was computed for each simulation run. These data were then analyzed using a Split-Plot design Analysis of Variance with the estimation method being the subplot factor. The four factor interaction component of the summary table was used as the error term for evaluation of the effects of interest. These effects were the main effect and interactions involving estimation method.

The residuals for this model were examined graphically and numerically for violations of the normality assumption. Both the normal plot (Figure 6.1) and the computed test statistic confirmed that the residuals exhibited kurtosis.

An array of Box-Cox power transformations were carried out on the data along with a likelihood analysis of the results. As indicated by the graph (Figure 6.2), the optimal power transformation of the data would appear to be in the range of $\lambda = 0.5$ to 1.0. Since the sample kurtosis was large for both values of lambda and the skewness statistic achieved significant magnitude for $\lambda = 0.5$, it was decided to proceed with the analysis using $\lambda = 1.0$ (the original scale).

Examination of the ANOVA results indicated the presence of significant interactions for Collinearity*Distance*Method ($F = 85.46$, df1 = 24, df2 = 48, $p < .001$) and Training Set*Method ($F = 4.75$, df1 = 8, df2 = 48, $p < .001$). The Collinearity*Distance*Method interaction is graphically displayed as a Log Collinearity (to improve readability) by Method plot for each of the four levels of Distance (Figure 6.3).

Examination of these plots suggest that the three factor interaction may be due entirely to the performance of the PCLR estimator versus the others across the levels of collinearity and distance. A set of contrasts were constructed to partition the
Figure 6.1
EAER Box–Cox Analysis

Figure 6.2
EAER – Distance*Collinearity*Method Interaction

Distance = 1.0

Distance = 2.0

Distance = 3.0

Distance = 4.0

Figure 6.3
interaction and test this conjecture. The results (Table 6.1) indicate that the significant interaction is in fact due entirely to the PCLR estimator. The PCLR estimator achieves the same level for EAER as the other estimators for the two levels of increased collinearity (Fisher Protected LSD comparisons), but shows a significant degradation of performance, relative to the other estimators, when collinearity is minimized. And while increasing distance between the populations results in decreasing values of EAER for these other estimators, the PCLR estimator achieves no improvement for increases in distance over 2.0 for the minimum collinearity conditions.

Table 6.1 EAER - Distance*Collinearity*Method Decomposition

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dist<em>Coll</em>Meth</td>
<td>0.04922</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCLR vs Others</td>
<td>0.04913</td>
<td>6</td>
<td>0.008188</td>
<td>341.17</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Remainder</td>
<td>0.00009</td>
<td>18</td>
<td>0.000005</td>
<td>0.21</td>
<td>0.999</td>
</tr>
<tr>
<td>Error</td>
<td>0.000024</td>
<td>48</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Having identified the three factor interaction with the PCLR estimator, it remains to examine the two factor interactions involving Method, embedded within the three factor interaction, for the other estimators. Looking at a plot of the Collinearity*Method interaction (Figure 6.4), it appears that PCLR may be the sole contributor to this interaction as well. Construction of the appropriate contrasts (Table 6.2), decomposing the two factor interaction, shows that there is in fact no interaction between Collinearity and Method for these other estimators.
EAER – Collinearity * Method Interaction

Figure 6.4
Next we turn our attention to the Distance*Method interaction for the other estimators (Figure 6.5). Decomposition of this interaction (Table 6.3) shows that after extracting the impact due to PCLR there is still evidence of further interaction. Examination of the interaction plot with PCLR eliminated (Figure 6.6) suggests that a difference between one group consisting of LDF and BLDF (the Fisher estimators) on one hand and another group consisting of LLR and RLR on the other may be present. Further decomposition of the interaction shows this to be the case. The two groups of estimators perform equivalently at the Distance = 1.0 and 2.0 levels, but as the distance continues to increase between the populations, the LDF/BLDF group begins to show a smaller EAER than the LLR/RLR group, achieving significance at Distance = 4.0. However, the magnitude of this difference is in the range of .012, not a terribly large difference.
EAER – Distance * Method Interaction

Figure 6.5
EAER – Distance * Method Interaction (No PCLR)

Figure 6.6
Concentrating now only on the LDF, BLDF, LLR and RLR estimators, we examine the Training Set*Method interaction (Figure 6.7). Decomposition of the interaction (Table 6.4) indicates that after eliminating the PCLR versus Others component, there are still additional significant effects in the data to be explained. Consideration of the Training Set * Method (without PCLR) plot (Figure 6.8) leads us to suspect that once again there is a split in performance between the Fisher estimators and the LLR and RLR estimators. Further decomposition of the interaction confirms that the remainder of the interaction is due to this difference between groups. The Fisher Protected LSD comparisons indicate that there is a significant difference between the two groups at the Training Set = 50 level but not at the 100 or 200 level.

Table 6.4 EAER - Training Set*Method Decomposition

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-Set*Meth</td>
<td>0.00090</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCLR vs Others</td>
<td>0.00039</td>
<td>2</td>
<td>0.000195</td>
<td>8.13</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>LDF/BLDF vs LLR/RLR</td>
<td>0.00035</td>
<td>2</td>
<td>0.000175</td>
<td>7.29</td>
<td>0.002</td>
</tr>
<tr>
<td>Remainder</td>
<td>0.00016</td>
<td>4</td>
<td>0.000040</td>
<td>1.67</td>
<td>0.172</td>
</tr>
<tr>
<td>Error</td>
<td>48</td>
<td></td>
<td>0.000024</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The final comparisons are for LDF versus BLDF and LLR versus RLR. The LDF and BLDF means are .1509 and .1488 ($\Delta = .0021$) respectively, with sample size equal to 36. The LLR and RLR means are .1561 and .1532 ($\Delta = .0029$) respectively, with the same sample sizes. The Fisher Protected LSD critical difference is .0038. While both mean differences are in the direction suggesting a small improvement for the biased alternatives, the difference did not attain statistical significance.

To summarize the findings for EAER, the principal conclusion is that while the PCLR estimator may perform as well as the other
EAER – Training Set * Method Interaction

Figure 6.7
EAER – Training Set * Method Interaction (No PCLR)

Figure 6.8
estimators when the degree of collinearity is strong, it's performance under conditions of no collinearity is markedly inferior. In addition, the PCLR estimator also failed to realize any benefit from increasing distance between the two populations or from increasing training set size. Considering the other estimators, the LDF and BLDF estimators show evidence of superior performance over the LLR and RLR estimators for small training set sizes and larger distances between populations.

6.2 Average Apparent Error Rate

An Average Apparent Error Rate (AAER) was computed for each simulation run. As with EAER, a Split-Plot design Analysis of Variance was carried out on these mean values. The normal plot of the residuals (Figure 6.9) and the computed test statistics showed evidence of significant kurtosis and skewness. The Box-Cox procedure was again used to suggest a suitable transformation. The loglikelihood plot (Figure 6.10) indicates that a power transformation in the range of 0.5 to 1.0 is advised. There was virtually no difference in sample kurtosis statistics between the two transformations. Since the sample skewness was markedly smaller (not significant) for the square root transformed data than for the untransformed data it was decided to continue the analysis with the data in the square root scale.

The ANOVA results indicate that both the Distance*Collinearity*Method ($F = 14.29, df1 = 24, df2 = 48, p < .001$) and the Distance*Training Set*Method ($F = 16.02, df1 = 24, df2 = 48, p < .001$) interactions are significant.

The Distance*Collinearity*Method interaction is graphically displayed in Figure 6.11. Examination of the graphs suggest that, as with EAER, the PCLR estimator behaves differently from the other estimators as distance increases. At high levels of collinearity the PCLR estimator performs as well as the LLR and RLR estimators and better than the LDF and BLDF estimators for distances equal to 2, 3 and 4. But at low levels of collinearity, while the other estimators
AAER Normal Plot

Figure 6.9
AAER Box–Cox Analysis

Figure 6.10
AAER – Distance * Collinearity * Method Interaction

Figure 6.11
show a progressively lower level of misclassification as distance increases, the performance of the PCLR estimator shows virtually no improvement. Decomposition of the interaction (Table 6.5) shows that this conclusion is reasonable. The PCLR versus Others component contains the entirety of the significance of the three factor interaction.

Table 6.5 AAER - Distance*Collinearity*Method Decomposition

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dist<em>Coll</em>Meth</td>
<td>0.09501</td>
<td>24</td>
<td>0.0003906</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCLR vs Others</td>
<td>0.09236</td>
<td>6</td>
<td>0.015393</td>
<td>55.57</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Remainder</td>
<td>0.00265</td>
<td>18</td>
<td>0.000147</td>
<td>0.53</td>
<td>0.93</td>
</tr>
<tr>
<td>Error</td>
<td></td>
<td>48</td>
<td>0.000277</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Collinearity*Method interaction was then investigated. The plot of the interaction (Figure 6.12) suggests that for the other estimators that there may in fact be no interaction between Collinearity and Method. Computation of the PCLR versus Others contrast for this interaction (Table 6.6) shows this to be the case. What also becomes apparent in examining this plot is that the LLR and RLR estimators consistently show smaller mean AAER over all levels of collinearity.

Table 6.6 AAER - Collinearity*Method Decomposition

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coll*Meth</td>
<td>1.34393</td>
<td>8</td>
<td>0.1679875</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCLR vs Others</td>
<td>1.34217</td>
<td>2</td>
<td>0.671085</td>
<td>606.22</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Remainder</td>
<td>0.00177</td>
<td>6</td>
<td>0.000295</td>
<td>0.27</td>
<td>0.948</td>
</tr>
<tr>
<td>Error</td>
<td></td>
<td>48</td>
<td>0.001107</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
AAER – Collinearity * Method Interaction

Figure 6.12
For the Distance*Training Set*Method interaction, the plot (Figure 6.13) suggests that the performance of the estimators can be separated into three groups - the PCLR estimator, the LDF and BLDF estimators, and the LLR and RLR estimators. This judgment is confirmed by the contrasts shown in Table 6.7. The AAER for PCLR does decrease going from Distance = 1.0 to 2.0. However, no further improvement results for any additional increase in distance. As distance increases both the LDF/BLDF and the LLR/RLR groups show decreases in AAER. And at the 100 and 200 levels of Training Set Size no significant difference between the two groups occurs for AAER. But for Training Set Size = 50 the LLR/RLR group has consistently smaller mean AAER values than the LDF/BLDF group.

Table 6.7 AAER - Distance*Training Set*Method Decomposition

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dist<em>T-Set</em>Meth</td>
<td>0.10649</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCLR vs Others</td>
<td>0.02541</td>
<td>6</td>
<td>0.004235</td>
<td>15.29</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>LDF/BLDF vs LLR/RLR</td>
<td>0.07903</td>
<td>6</td>
<td>0.013172</td>
<td>47.55</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Remainder</td>
<td>0.00204</td>
<td>12</td>
<td>0.000170</td>
<td>0.61</td>
<td>0.82</td>
</tr>
<tr>
<td>Error</td>
<td></td>
<td>48</td>
<td>0.000277</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Next we consider the Distance*Method interaction (Figure 6.14), ignoring PCLR. The plot shows that while the other four estimators perform essentially the same at Distance = 1.0, as distance increases the LLR and RLR estimators show a more rapid decrease in mean AAER than do the LDF and BLDF estimators. This difference achieves significance for all distances greater than 1.0. The two logistic estimators do not differ significantly from one another, as is also the case for the Fisher estimators. This interpretation is confirmed by the decomposition of the interaction shown in Table 6.8.
AAER – Distance * Training Set * Method Interaction

Distance = 1.0

Distance = 2.0

Distance = 3.0

Distance = 4.0

Figure 6.13
AAER - Distance * Method Interaction

Figure 6.14
Table 6.8 AAER - Distance\*Method Decomposition

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dist*Meth</td>
<td>0.07567</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCLR vs Others</td>
<td>0.03379</td>
<td>3</td>
<td>0.011263</td>
<td>40.66</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>LDF/BLDF vs LLR/RLR</td>
<td>0.04058</td>
<td>3</td>
<td>0.013527</td>
<td>48.83</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Remainder</td>
<td>0.00130</td>
<td>6</td>
<td>0.000217</td>
<td>0.78</td>
<td>0.59</td>
</tr>
<tr>
<td>Error</td>
<td>48</td>
<td></td>
<td>0.000277</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Looking more closely at the LDF and BLDF estimators, the mean difference (.0001) fails to achieve significance (critical LSD difference is .0079). The same is true for the LLR and RLR estimators (mean difference = .0057). It is interesting to note, comparing the EAER results to the AAER results, that we now find the direction of the sample mean differences are reversed. Where the biased estimators in each pair (BLDF vs LDF and RLR vs LLR) had the smaller values for EAER, the unbiased estimators have the smaller values for AAER.

There are two principal conclusions to be drawn regarding the AAER data. First is that, like the EAER data, the PCLR estimator performs differently depending upon the level of collinearity. When collinearity is high, the PCLR estimator achieves values of AAER as low as the other logistic estimators and lower than the Fisher estimators. But when collinearity is low, the PCLR estimator's performance degrades markedly relative to the other estimators.

Second, the LLR and RLR estimators show superior performance compared to the LDF and BLDF estimators under conditions of small training set size, large distance and across all levels of collinearity.
6.3 Mean Square Error

As with EAER and AAER, a split-plot Analysis of Variance was utilized to examine the Mean Square Error (MSE) data. The normal plot of the residuals (Figure 6.15) and the computed test statistics showed evidence of both positive skewness and kurtosis. A Box-Cox analysis was again carried out on the data. The loglikelihood plot (Figure 6.16) indicates that the optimal transformation power is in the range of -0.5 to 0.0 (log transformation). Since the log transformation is widely considered to be the desirable alternative for data of this type, it was chosen for use in this situation.

Inspection of the ANOVA results show that all three of the two factor interactions involving Method tested to be statistically significant (Collinearity*Method - \( F = 65.22, \text{df1} = 8, \text{df2} = 48, p < .001 \); Training Set*Method - \( F = 4.75, \text{df1} = 8, \text{df2} = 48, p < .001 \); Distance*Method - \( F = 7.73, \text{df1} = 12, \text{df2} = 48, p < .001 \)).

The Collinearity*Method interaction is graphically illustrated in Figure 6.17. Looking at the performance of each estimator across all three levels of collinearity, the computation of the two contrasts which compare the most similar estimators - PCLR vs BLDF and LDF vs LLR (Table 6.9) indicate that no two estimators behave the same. Closer examination of the plot suggests in the range of high (\( \sigma = .001 \)) to moderate (\( \sigma = .01 \)) collinearity that the performance of the estimators is segregated into two groups, BLDF/PCLR and LDF/LLR/RLR. The contrasts listed in the second part of Table 6.9 show this to be true. The BLDF and PCLR estimators both show no significant change in mean ln(MSE) (\( \Delta = 0.0002 \), averaged over both estimators), while the three other estimators show a significant reduction in mean ln(MSE) (\( \Delta = 2.2465 \), averaged over the three estimators).
MSE Normal Plot

Figure 6.15
Figure 6.16
Figure 6.17

MSE – Collinearity * Method Interaction
Table 6.9 MSE - Collinearity*Method Decomposition

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coll*Meth</td>
<td>180.4704</td>
<td>8</td>
<td>22.5586</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCLR vs BLDF</td>
<td>2.4911</td>
<td>2</td>
<td>1.2456</td>
<td>3.60</td>
<td>0.03</td>
</tr>
<tr>
<td>LDF vs LLR</td>
<td>3.7081</td>
<td>2</td>
<td>1.8541</td>
<td>5.36</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Remainder</td>
<td>172.2712</td>
<td>4</td>
<td>43.0678</td>
<td>124.51</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Coll*Meth (Hi&amp;Mod Coll only)</td>
<td>37.0710</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCLR/BLDF vs Others</td>
<td>36.3421</td>
<td>1</td>
<td>36.3421</td>
<td>105.07</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Remainder</td>
<td>0.7289</td>
<td>3</td>
<td>0.2430</td>
<td>0.70</td>
<td>0.56</td>
</tr>
<tr>
<td>Error</td>
<td>0.3459</td>
<td>48</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

However, going from moderate collinearity to minimal collinearity ($\sigma = 1.25$) we find that the relative performance of the estimators change. The BLDF estimator shows no significant change while the PCLR estimator shows a significant increase in mean $\ln(\text{MSE})$. The other three estimators show significant reductions in $\ln(\text{MSE})$, the LDF estimator showing the greatest decrease, then the LLR and last the RLR.

Looking at the absolute differences between the estimators, we find that at the high collinearity level that the BLDF estimator has a significantly lower mean $\ln(\text{MSE})$ than any of the other estimators. The PCLR estimator in turn has a lower mean $\ln(\text{MSE})$ than the remaining estimators. At the moderate collinearity level, the same relative order of performance is found, but the difference between the PCLR and the RLR estimators is no longer significant ($p = .14$). At the low collinearity level the ordering of the estimators has notably changed. The LDF estimator now has a significantly smaller mean $\ln(\text{MSE})$ than the other estimators, followed by the RLR and BLDF estimators and last by the PCLR and LLR estimators (which show no evidence of differing).

There are several general conclusions about Collinearity and
Method we can draw from this data. First, the BLDF was superior at the high and moderate levels of collinearity as well as showing the most consistent behavior across levels of collinearity. Second, the PCLR estimator, while showing good performance at the high and moderate collinearity levels, showed significant deterioration under conditions of minimal collinearity. Third, the biased logistic estimators performed as well as or better than the conventional logistic estimators over all levels of collinearity. And fourth, the RLR estimator shows significantly better performance than the LDF estimator at the high and moderate collinearity levels, and came in second overall behind the LDF estimator at the minimal collinearity level.

Turning next to the Distance*Method interaction (Figure 6.18), we find that the means ln(MSE) values for all the estimators increase significantly with distance. The biased estimators tend to perform consistently better than their unbiased counterparts. The LDF estimator increases more slowly than the other estimators (Table 6.10). The BLDF estimator has mean ln(MSE) values as small or smaller than the other estimators at all distances. At Distance = 1.0, the PCLR and BLDF do not differ significantly, but as distance increases, the PCLR estimator degrades in performance more quickly than the BLDF estimator. Both the PCLR and RLR estimators do better than the LDF estimator at the two smallest distances, but these three estimators do not differ significantly at Distance = 3.0 or 4.0.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dist*Meth</td>
<td>32.0728</td>
<td>12</td>
<td>2.6728</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDF vs Others</td>
<td>19.8078</td>
<td>3</td>
<td>6.6026</td>
<td>19.09</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>BLDF vs RLR</td>
<td>0.6505</td>
<td>3</td>
<td>0.2168</td>
<td>0.63</td>
<td>0.98</td>
</tr>
<tr>
<td>PCLR vs LLR</td>
<td>2.7112</td>
<td>3</td>
<td>0.9037</td>
<td>2.61</td>
<td>0.06</td>
</tr>
<tr>
<td>Error</td>
<td>48</td>
<td></td>
<td>0.3459</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
MSE - Distance * Method Interaction

Figure 6.18
Next to consider is the Training Set-Method interaction (Figure 6.19). The general pattern here is for the mean ln(MSE) values of all the estimators to decrease with increasing training set size. Decomposition of this interaction into a contrast between LDF/LLR/RLR and BLDF/PCLR (Table 6.11) shows that these two group's patterns of change over training set size constitute the entirety of the interaction. The significant contrast implies that the first group is showing a faster decrease in mean ln(MSE) values than the second group as training set size increases.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-Set*Meth</td>
<td>12.2985</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDF/RLR/LLR vs BLDF/PCLR</td>
<td>10.4256</td>
<td>2</td>
<td>5.2128</td>
<td>15.07</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Remainder</td>
<td>1.8729</td>
<td>4</td>
<td>0.9365</td>
<td>2.71</td>
<td>0.08</td>
</tr>
<tr>
<td>Error</td>
<td></td>
<td>48</td>
<td>0.3459</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

At all levels of training set size the mean ln(MSE) value of the BLDF estimator is significantly smaller than any other estimator. At Training Set = 50, the PCLR estimator is significantly superior to the remaining estimators. But as training set size continues to increase, the LDF and RLR estimators improve to the level of the PCLR estimator. As was the case for the Distance*Method interaction and in part for the Collinearity*Method interaction, the biased estimators did better than their unbiased counterparts.

6.4 Standard Deviation of AER

The split-plot analysis of the Standard Deviation of the Actual Error Rate (SD(AER)) data produced residuals that were examined as before. And as before, the normal plot (Figure 6.20) showed evidence of significant positive skew and kurtosis. The Box-Cox analysis
MSE – Training Set * Method Interaction

Figure 6.19
SD(AER) Normal Plot

Figure 6.20
(Figure 6.21) indicated that a power transformation in the range of 1.0 to 0.0 would be optimal. For the same reasons that applied to the MSE measure, it was decided to again apply a log transformation to the data and proceed with the analysis.

The analysis of the log-transformed data found a significant three factor interaction between Distance, Collinearity and Method \((F = 3.32, df1 = 24, df2 = 48, p < .001)\). The Training Set factor did not interact with Method. The means plots of the three factor interaction are shown in Figure 6.22. Visual inspection of the plots suggests that the performance of the estimators split them into three groups. The three factor interaction is due to the PCLR and RLR estimators performing differently from one another and differently from the LDF, BLDF and LLR estimators. But these three estimators appear to perform similarly over the levels of Distance.

Examination of the decomposition of this interaction (Table 6.12) shows that this separation does in fact account for the entirety of the significant interaction. Therefore, the next step is to examine the Collinearity*Method and Distance*Method interactions, keeping these groupings in mind.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dist<em>Coll</em>Meth</td>
<td>5.47934</td>
<td>24</td>
<td>0.2289</td>
<td>5.93</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>PCLR/RLR vs LDF/BLDF/LLR</td>
<td>2.44970</td>
<td>6</td>
<td>0.4083</td>
<td>5.93</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>PCLR vs RLR</td>
<td>2.62384</td>
<td>6</td>
<td>0.4373</td>
<td>6.36</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Remainder</td>
<td>0.40580</td>
<td>12</td>
<td>0.0338</td>
<td>0.49</td>
<td>0.91</td>
</tr>
<tr>
<td>Error</td>
<td>0.0688</td>
<td>48</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Looking first at the Collinearity*Method interaction (Figure 6.23), we see the pattern of the three factor interaction plots repeated. The LDF, BLDF and LLR estimators appear not to interact with Collinearity. Conversely, the PCLR and RLR estimators do appear
SD(AER) Box–Cox Analysis

Figure 6.21
SD(AER) – Distance * Collinearity * Method Interaction

Figure 6.22
SD(AER) – Collinearity * Method Interaction

Figure 6.23
to interaction with Collinearity and with the other three estimators. Decomposition of this interaction (Table 6.13) shows that this interpretation is consistent with the quantitative analysis. Additional comparisons of group means show that the LDF, BLDF and LLR estimators show no significant change in ln(SD(AER)) over the three levels of collinearity. However, over the entire range of collinearity investigated, the LDF and BLDF estimators showed significantly lower levels of ln(SD(AER)) than did the LLR estimator. But these two estimators did not differ significantly from one another.

Table 6.13 SD(AER) - Collinearity*Method Decomposition

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coll*Meth</td>
<td>13.4810</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDF/BLDF/LLR vs RLR/PCLR</td>
<td>8.3249</td>
<td>2</td>
<td>4.1625</td>
<td>60.50</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>RLR vs PCLR</td>
<td>5.1121</td>
<td>2</td>
<td>2.5561</td>
<td>37.15</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Remainder</td>
<td>0.0440</td>
<td>4</td>
<td>0.0110</td>
<td>0.16</td>
<td>0.96</td>
</tr>
<tr>
<td>Error</td>
<td>48</td>
<td>0.0688</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For the RLR estimator, there was no significant change in mean ln(SD(AER)) going from the highest level of collinearity to the moderate level. But in going to the minimal level of collinearity, the mean ln(SD(AER)) increased significantly for this estimator. At the first two levels of collinearity the RLR estimator had a marginally lower mean ln(SD(AER)) than did the LLR estimator. But at the third level of collinearity the mean ln(SD(AER)) for the RLR estimator became larger than that of the LLR estimator, although the difference did not attain significance. Looking at the PCLR estimator, we find that it did not differ significantly from the other logistic estimators at the level of greatest collinearity. But the mean ln(SD(AER)) of this estimator increased significantly with the change to each of the other levels of collinearity, resulting in
the PCLR estimator performing the worst of any estimator when the level of collinearity was minimized.

Looking next at the Distance*Method interaction (Figure 6.24), we find an interesting pattern of change as Distance increases from 1.0 to 4.0. Examination of the decomposition of this interaction (Table 6.14) indicates that the estimators can be separated into three groups, one being the LDF and BLDF estimators, another being the LLR and RLR estimators (despite the crossing of the two estimators between 3.0 and 4.0) and the last being the PCLR estimator. Further comparisons of means show that while the LDF and BLDF estimators never differ at any level of distance, they both show significant decreases in mean \( \ln(\text{SD(AER)}) \) with each increase in distance. The LLR and RLR estimators, which likewise do not differ at any level of distance, show marginal evidence of a decrease in mean \( \ln(\text{SD(AER)}) \) going from Distance = 1.0 to Distance = 2.0. And last, the PCLR estimator shows evidence of an increasing mean \( \ln(\text{SD(AER)}) \) as distance goes from 2.0 to 3.0 and 4.0 (where the increase is significant.

Table 6.14 SD(AER) - Distance*Method Decomposition

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dist*Meth</td>
<td>24.9930</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LDF/BLDF vs LLR/RLR/PCLR</td>
<td>20.4107</td>
<td>3</td>
<td>6.8036</td>
<td>98.89</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>LLR/RLR vs PCLR</td>
<td>4.3050</td>
<td>3</td>
<td>1.4350</td>
<td>20.86</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>LLR vs RLR</td>
<td>0.2650</td>
<td>3</td>
<td>0.0883</td>
<td>1.28</td>
<td>0.29</td>
</tr>
<tr>
<td>Remainder</td>
<td>0.0123</td>
<td>3</td>
<td>0.0041</td>
<td>0.06</td>
<td>0.98</td>
</tr>
<tr>
<td>Error</td>
<td>48</td>
<td>0.0688</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Last, even though there was no significant interaction between Method and Training Set Size, it is informative to examine an interaction plot for these two factors (Figure 6.25). It shows that all the estimators produce smaller \( \ln(\text{SD(AER)}) \) values as training
Figure 6.24
SD(AER) – Training Set * Method Interaction

Figure 6.25
set size increases. The Fisher estimators tend to have smaller values than the others across the entire range.

6.5 Additional Analyses

When analyzing the results for the EAER metric, it is interesting to compare the results obtained here for the LDF and BLDF estimators with the results reported by DiPillo (1976). The Monte Carlo study described in the sections above used sample sizes from each population of 25 at the smallest. In contrast, DiPillo's largest sample sizes were 25. In order to allow for more of a comparison between the two studies, some additional simulation runs were carried out with a smaller sample size (10), for the high collinearity condition, for values of the Mahalanobis distance equal to 1, 2 and 3. The resulting average EAER values have been used to augment the original data in Figures 6.26 (Collinearity Parameter = .001 and Distances = 1, 2 and 3 respectively). The graphs show the presence of a tendency for the gap between LDF and BLDF to increase as the training set size decreases. This is also evident in the results cited by DiPillo (Table I, 1976).

The two studies have their simulation results based on fundamentally different population covariance matrices. This limits the ability to accurately relate a "high" collinearity condition in the DiPillo study to one in the present study. Bearing this limitation in mind, Table 6.15 compares the Proportion Improvement results for selected factor combinations of the DiPillo study to comparable values determined from this study. Specifically, the high collinearity condition in the DiPillo study (σ=0.01) was compared to the high collinearity condition in the present study (σ=0.001). An examination of the table shows that the improvement found in the present study is very consistent with that reported by DiPillo. The agreement between the two studies indicates that the method used in this study to estimate the ridge parameter, even if not optimal, is evidently a reasonable method. The correspondence also provides both reconfirmation of the basic conclusion of the DiPillo study and
EAER – Effect of Training Set Size

Distance = 1.0

Distance = 2.0

Distance = 3.0

Figure 6.26
evidence that the results cited by DiPillo are not necessarily just an artifact of the particular covariance matrix used to generate the different conditions of collinearity in his study.

Table 6.15 Comparison with DiPillo (1976) Study

Proportion Improvement (%)

Collinearity = High

\( \sigma = .001 \) (Present Study) and \( \sigma = .01 \) (DiPillo Study)

<table>
<thead>
<tr>
<th>Distance</th>
<th>Training Set = 20</th>
<th>Training Set = 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>19.2</td>
<td>26.1</td>
</tr>
<tr>
<td></td>
<td>(12.5)</td>
<td>(19.1)</td>
</tr>
<tr>
<td>2.0</td>
<td>28.4</td>
<td>27.8</td>
</tr>
<tr>
<td></td>
<td>(29.5)</td>
<td>(26.9)</td>
</tr>
<tr>
<td>3.0</td>
<td>39.5</td>
<td>33.5</td>
</tr>
<tr>
<td></td>
<td>(39.8)</td>
<td>(18.3)</td>
</tr>
</tbody>
</table>

Note: DiPillo results in parentheses.

Examination of this table also gives us some insight into the performance of the ridge parameter selection method proposed and used in this study. The similarity of the results from the DiPillo study and the present study suggests that the new method of ridge parameter selection can do as well, but in this case does no better than, the simple choice of \( k = 1.0 \) used by DiPillo. It is not clear from this limited comparison whether either method would show an advantage when systematically compared over a variety of common mean-covariance conditions.

The second issue that arose during the analysis and interpretation of the EAER data concerned the degraded performance of the PCLR estimator under conditions of weak or no collinearity. At the \( \sigma = .01 \) level the PCLR estimator still performed comparably to
the other logistic estimators. But at the $\sigma = 1.25$ level it did very poorly. This raises the question: Does the poor performance of the PCLR estimator occur over a wide range of values of $\sigma$, or a narrow range? In order to address this question, additional runs were made for a wider range of $\sigma$ values, specifically $\sigma = 0.1, 0.5, 1.0, 5.0$ and 10.0. A graph of the results is shown in Figure 6.27. The average EAER values in this figure are for Distance = 2.0, Training Set Size = 50 and for values of $\sigma$ equal to 0.001, 0.01, 0.1, 0.5, 1.0, 1.25, 5.0 and 10.0 and are typical of the results found for other values of Distance and Training Set Size. As can be seen, the performance of the PCLR estimator is comparable to the other estimators for $\sigma$ equal to 0.001 and 0.01, conditions of relatively high collinearity. However, at $\sigma$ equal to 0.1 the average EAER value for PCLR begins to deviate from the other estimators. This deterioration increases until $\sigma$ exceeds 1.25 and then appears to flatten out. This result leads us to conclude that even though the PCLR estimator can perform satisfactorily for conditions of high collinearity, when the degree of collinearity is weak or non-existent the estimator does a particularly poor job of classification. The PCLR estimator simply seems to lack the adaptive characteristic of the other estimators.
EAER - Sensitivity of PCLR to Collinearity

Figure 6.27
Chapter 7
ASYMPTOTIC PERFORMANCE OF GRL ESTIMATOR

7.1 Asymptotic Behavior of the GRL Estimator

While the performance of the GRL estimator (in each of its variations) has been assessed through the method of Monte Carlo simulation, it would be desirable to make some determination of the asymptotic behavior of the GRL estimator, if possible relative to the LDF estimator.

There are two approaches that have been taken to address this task. In Efron’s (1975) paper, the asymptotic relative efficiency of the standard logistic regression estimator was developed. It was defined to be the ratio of the terms of order 1/n in the expansions of the overall expected error rates for the normal and logistic procedures. The results of this approach, as outlined in chapter 2, indicate that the normal procedure (LDF) looks progressively better as the distance between the two populations increases.

An alternative approach to the examination of the asymptotic behavior of the LLR estimator was developed by McLachlan and Byth (1979). They determined the asymptotic expansions of the expected error rates for the LLR procedure and the LDF procedure separately (up to terms of order 1/n). Their analysis suggests that Efron’s asymptotic relative efficiency may not adequately summarize the relative performance of the two methods. While the ratio of the asymptotic LLR error rate to the asymptotic LDF error rate may be large for large values of the Mahalanobis Distance, the absolute magnitude, as well as the absolute difference, of these error rates tend to be relatively small in these circumstances.

Therefore, the most promising approach seems to be in trying to determine the asymptotic error rate in a manner analogous to that used by McLachlan and Byth.

The goal is to determine the asymptotic error rate of the GRL estimator. Following McLachlan and Byth, we begin by developing an expansion of the probability of making a misclassification (assigning
a case to Population 2 when the correct choice was Population 1) of
the first type. This probability is expressed as

\[ P_1 \approx G\left(-\frac{1}{2}b^*\right) + q_i(\hat{\beta})(b_i - \beta_i) + \frac{1}{2} q_{ij}(b^*) (b_i - \beta_i)(b_j - \beta_j) \]  \tag{7.1}

where \( b^* \) is a point in the interval joining \( b \) and \( \hat{\beta} \) and \( b \) is the GRL estimate of \( \beta \).

We then take the expectation in order to obtain the expected error rate, yielding

\[ E(P_1) = G\left(-\frac{1}{2}b^*\right) + q_i(\hat{\beta}) \text{Bias}(b_i) + \frac{1}{2} q_{ij}(\hat{\beta}) \text{Cov}(b_i, b_j) + o(1/n) \]  \tag{7.2}

In a separate paper (Byth and McLachlan, 1978), the details of evaluating this expression are given. A key condition that allowed the authors to determine the bias and covariance terms was the status of the LLR estimator as a maximum likelihood estimator. In the present case the GRL estimator is a function (of sorts) of the LLR mle. However, the GRL estimator is a (random) function of the data as well as the mle. Therefore, the GRL estimator is not a maximum likelihood estimator. Without being able to take advantage of the results available to the class of mle estimators, we are unable to proceed to evaluate the bias and covariance terms. While the form of the GRL estimator is perhaps intuitively appealing and plausible, its lack of statistical rigor prevents us from arriving at any more general statements about its performance relative to the LLR or LDF estimators.
Chapter 8
SUMMARY

8.1 Classification Performance of the Discriminant Functions

Both the Expected Actual Error Rate (EAER) and the Average Apparent Error Rate (AAER) tell us something about the classification capability of a discriminant function estimator. The key indicator of the relative capabilities of the various estimators for correctly classifying the members of two populations, under conditions of multivariate normality with identical covariance structure, is the EAER. While the Apparent Error Rate is also typically described as a measure of estimator performance relative to the parent populations, it is well recognized that it tends to be overly optimistic when used for this purpose. Despite this shortcoming, it is still quite reasonable to use the AAER as a measure of an estimator's ability to correctly separate the training set, to correctly classify the members of two samples. By comparing not only the performance of the estimators for each error rate, but also comparing the performance of the estimators between the two error rates, we can better understand when and how biased estimators are preferred to unbiased estimators for statistical discrimination.

With respect to the EAER measure, the relative performance of the Fisher and the logistic estimators, and the inconsistent performance of the PCLR estimator are the key results. The preferred performance for an estimator would be to show smaller values of the EAER as distance and training set size increases and to show little or no increase of the EAER as the degree of collinearity increases.

The Fisher estimators and the LLR and RLR estimators all improved as distance increased. However, the Fisher estimators improved at a slightly faster rate than did the LLR and RLR estimators. This first result is consistent with Efron's work regarding the relative performance of the two classes of estimators. But the absolute size of the difference in EAER performance between the Fisher and logistic classes, as McLachlan and Byth argued, is relatively small and could be inconsequential in most practical
situations.

All estimators, except PCLR, showed shrinking EAER values as training set size increased. But the principle effect of increasing training set size was to make the estimators behave more alike. Only for the smallest training set size did the Fisher estimators show a slight advantage.

For collinearity, the Fisher estimators maintained a small, consistent advantage over the LLR and RLR estimators over all tested levels.

Collinearity proved to be the Achilles' Heel for the PCLR estimator. The relatively poor performance of this estimator for moderate and low levels of collinearity eliminates it from serious consideration. The other estimators match the performance of the PCLR estimator when collinearity is present without demonstrating it's deterioration of performance when collinearity is not present. When collinearity is not a factor, it appears that the PCLR estimator gives up too much information in discarding the one eigenvalue/eigenvector component.

The BLDF estimator did not give significant evidence of superiority over the LDF estimator for the range of simulation conditions investigated here (as the work of DiPillo would suggest). However, the simulation differences are in the direction predicted by DiPillo. And for the runs with training set size most similar to his, the differences are of a magnitude similar to those reported by him. It is likely that the degree of differential performance between the unbiased and biased estimators will vary widely depending on specific conditions concerning the two populations - such as distance, training set size and covariance structure. While the BLDF estimator may always do better under conditions of collinearity, the difference might not always be great.

Comparing the LLR and RLR estimators, the simulation results show some indication of an advantage for the biased RLR estimator. But the lack of significance for the small difference in EAER between the two estimators again suggests that characteristics of the parent
multivariate normal populations will strongly influence the relative advantage (if any) in performance of a biased estimator (RLR) over an unbiased estimator (LLR).

For the AAER measure, the preferred performance for an estimator again would be to show smaller values of the AAER as distance and training set size increase and to show no degradation of the AAER as the degree of collinearity increases.

Perhaps the most interesting facet of the AAER analysis is the reversal between the logistic and Fisher estimators. While the Fisher estimators had the performance advantage for the EAER, now the LLR and RLR estimators exhibit the superior behavior. These logistic estimators show smaller AAER values when distance is greater and when the training set size is smaller. And they have consistently smaller AAER values over all levels of collinearity. The logistic estimators seem more able to conform to the available training set (and produce small numbers of misclassifications), but at the same time they sacrifice performance with respect to their ability to classify the parent multivariate normal populations.

Another interesting result for this measure was that the unbiased estimators of each pair (LDF and LLR) had the slightly smaller (but not significantly different) mean values. This is the reverse of the result seen for the EAER. And once more, the PCLR estimator shows evidence of giving up too much information in eliminating the influence of the smallest eigenvalue and it's associated eigenvector.

The larger pattern of results for EAER and AAER suggests a particular relationship between the combinations of Fisher/Logistic and Unbiased/Biased estimators. First, it seems these estimators do not do a superior job of classifying both the training set data and the populations as a whole. When multivariate normality holds, the Fisher estimators produce classification rules that are superior for the underlying populations (as evidenced by the EAER) at the expense of performing well on the training set data (as evidenced by the AAER). Conversely, the logistic estimators produce rules that do
well for the training set data, but not for the parent distributions. Given that the Fisher estimators are maximum likelihood estimators for the parent populations, such a finding is plausible.

Second, for either group of estimators the biased estimator seems to produce classification rules better suited to the underlying population while the unbiased estimator seems to produce rules that perform better for the training set data. Since the biased estimators are designed to "discount", in various ways, the information contained in the training set, it is reasonable that they might do a poorer job of classifying these observations.

8.2 Stability of the Discriminant Functions

The Standard Deviation of the Actual Error Rate (SD(AER)) reflects the stability of a given estimator. An estimator that produces Actual Error Rates, from a variety of samples from the same population, with less variation is preferable. Again, a desirable estimator would show smaller values as training set size increases and should not react poorly to conditions of high collinearity.

All the estimators displayed smaller SD(AER) values with increasing training set size. Looking across levels of collinearity, we find that the standard deviations for the LDF, BLDF and the LLR estimators change the least. However, the BLDF and LDF estimators tended to produce smaller SD(AER) values than the LLR estimator. With respect to distance, the Fisher estimators showed a tendency to produce smaller SD(AER) values as the populations were further apart. The LLR and RLR estimators show only a small tendency to improve with greater distance, and the PCLR estimator tends to do worse as distance increases.

Among all the estimators, for the SD(AER) measure, the Fisher estimators appear to be the estimators of choice. Among the logistic estimators, the RLR estimator does as well or better than the LLR estimator. Again the PCLR estimator appears least desirable, showing the most variability across the different simulation conditions
studied. This leads us to conclude, again excepting PCLR, that the biased estimators are preferable over their unbiased counterparts with respect to minimizing sampling variation.

8.3 Parameter Estimation Performance of the Discriminant Functions

The Mean Square Error (MSE) measures estimator performance on a more elementary level than the EAER and AAER measures. By doing so it alerts us to differences between the estimators that might otherwise be overlooked. Once again, we would like to see the MSE decrease with increasing training set size and be indifferent to high levels of collinearity.

All the estimators showed decreasing values of MSE as the training set size increases. The BLDF estimator consistently had the smallest values among the estimators. All the estimators tended to show increasing values of MSE as the distance increased. In the entire group, the BLDF estimator again tended to have the smallest values of MSE. Only the BLDF estimator showed clearly stable performance over the entire range of collinearity. The PCLR estimator, as was the case for the earlier metrics, performance satisfactorily for the high and moderate collinearity conditions, but degraded at the low collinearity condition, just when the LDF and RLR estimators were catching up to the BLDF estimator.

The performance of the BLDF estimator was a clear step ahead of the other estimators. The RLR estimator shows a clear edge over the LLR estimator and does as well or better than the LDF estimator for this metric over the levels of Distance, Training Set Size and Collinearity.

The significant differences found between estimators for MSE, contrasted with the nonsignificant differences between estimators found for the EAER measure imply that poor parameter estimation may not necessarily translate into poor classification performance. One coefficient estimate that is markedly far from the true value, while producing a large MSE, could have virtually no impact on the
discriminant functions ability to correctly classify members of the two populations.

8.4 Concluding Remarks

Looking at the Monte Carlo results as a whole, we conclude that the Fisher estimators typically perform as well or better than the logistic estimators. This is not surprising, since the comparison conducted here was under conditions most favorable to the Fisher estimators - multivariate normality with common covariance structure. In fact, the relatively small absolute differences in EAER between the Fisher estimators and the LLR and RLR estimators indicate that using the logistic estimators (even in conditions of multivariate normality) may result in little deterioration in misclassification error rates. And when other distributional assumptions apply, it has been argued by others (e.g. Press and Wilson, 1978) that the logistic estimators will be preferable to the Fisher estimators.

Earlier research by DiPillo argued that the unbiased Fisher estimator could be improved upon by a biased alternative like the BLDF under conditions of collinearity. The present study goes on to suggest that at least one biased logistic alternatives (RLR) can also improve on it's unbiased counterpart. Within either group (Fisher or logistic), the biased estimators (excepting PCLR) compared favorably to the biased estimators, weakly for the EAER and more strongly for SD(AER) and for MSE. However, the differences in MSE found between estimators in this study do not appear to translate into equally large differences in error rate.

The small differences in the EAER also draw attention to another aspect of these results. For the EAER there was no interaction between Method and Collinearity (ignoring PCLR). Any difference between estimators did not depend on the level of collinearity. This goes against the argument that biased estimators provide relief from poor classification due to collinearity. Instead, the condition that produced the most marked differences between biased estimators and their unbiased alternatives was small
training set size. In fact, when we look at DiPillo's results (i.e., 1976) we find that the relationship between level of collinearity and relative performance of the LDF and BLDF estimators is not a consistent one. What is consistently found is that the largest differences between estimators occur for the smallest training set sizes. It could be argued that the major benefit of the biased estimators, with respect to minimizing misclassification, is their ability to compensate for the instability of estimation arising from small training sets.

DiPillo also considered, without satisfactory conclusion, the nature of a good ridge parameter estimation method. The sample based method proposed here has given some evidence of performing satisfactorily under certain conditions of collinearity. More general conclusions about the method's desirability must await further research.
Bibliography


