Accuracy of Logistic Models and Receiver Operating Characteristic Curves

Philip Corbett

This thesis is submitted for the degree of Doctor of Philosophy

Department of Statistics
University of Warwick
Coventry
CV4 7AL

August 2001

Contents

1 Introduction 1

2 ROC Curves, Logistic Regression and Predictive Accuracy 9

2.1 Diagnostic testing and ROC curves 10

2.1.1 The area under the ROC curve 19

2.2 Classification methods 23

2.2.1 Logistic regression 24

2.2.2 Discriminant analysis 26

2.2.3 Comparison between discriminant analysis and logistic regression 27

2.3 Assessment of statistical predictions and adequacy of models . . . 28

2.3.1 Model accuracy 29

2.3.2 Discrimination error 30

2.3.3 Calibration error and sampling techniques 35

2.4 Common issues 38

3 Overestimation of the ROC Curve and Area for Logistic Regression 41

3.1 Retrospective and prospective ROC 42

3.2 Expected overestimation of the ROC curve 45

3.3 Overestimation of the area under the ROC curve 54

3.4 Further approximations 61

3.5 Comments 65

4 The Corrected ROC Curve: Examples and Simulation Study 69

4.1 Melanoma case-control study 70
4.2 Simulation study 74
4.3 Breast Cancer study 83
4.4 Simulation study 86
4.5 Overview 91

5 Shrinkage in Logistic Models for Categorical Data 93

5.1 Background 97
5.2 Categorical data and logistic regression 100
5.3 Covariate model 107
5.4 Example 110
5.4.1 Simulation study 121
5.4.2 Some comments 132

6 A General Approach to Shrinkage in Models for Categorical Data 134

6.1 Methodology for more than one categorical variable 136
6.2 Example 145
6.3 Simulation studies 149
6.3.1 Artificial data 149
6.3.2 Prospective ROC area and model deviance 157
6.3.3 Ear infection data 159

7 Summary and Concluding Remarks 165

Bibliography 171

List of Figures

2.1 ROC curve for FEV_1 data 14

2.2 Distributions of populations with and without the attribute of interest 17

4.1 ROC Curve for the melanoma logistic regression 72

4.2 Original and 'corrected' ROC curves for the melanoma score . . . 73

4.3 Predicted and simulated overestimation of the ROC curve for the melanoma score 76

4.4 Retrospective, 'corrected' and cross-validated 'leave-one out' ROC curves for the melanoma score 78

4.5 Parametric bootstrapped prospective and cross-validated 'leave-one out' ROC curves for the melanoma score 79

4.6 Predicted overestimation of the ROC curve by $S(u)$ and by the sample splitting cross-validation approach (splitting fraction $= \frac{1}{2}$) for the melanoma score 81

4.7 ROC curve for the breast cancer logistic regression 84

4.8 Original and 'corrected' ROC curves for the breast cancer score . 85

4.9 Predicted and simulated overestimation of the ROC curve for the breast cancer score 87

4.10 Retrospective, 'corrected' and cross-validated 'leave-one out' ROC curves for the breast cancer score 88

4.11 Parametric bootstrapped prospective and cross-validated 'leave-one out' ROC curves for the breast cancer score 89

4.12 Predicted overestimation of the ROC curve by $S(u)$ and by the sample splitting cross-validation approach (splitting fraction = $\frac{1}{2}$) for the breast cancer score 90

5.1 Shrinkage effect of $\tau^2 = \hat{\tau^2}$ for the credit defaulting data (R is the identity matrix) 114

5.2 Values of the shrunk estimates for different values of τ^2 holding α fixed 115

5.3 Original model fitted probabilities and shrunk estimates against the actual proportions in the data 116

5.4 The effect on the ROC areas for the shrunk estimates of varying the correlations within occupational groups 120

5.5 Retrospective and prospective ROC areas for the model and the shrunk estimates, by varying correlations within the four occupational groups 123

5.6 Retrospective and prospective ROC areas for the model and the shrunk estimates, by varying correlations within the four occupational groups (with continuity correction applied) 125

5.7 Retrospective and prospective ROC areas for the model and the shrunk estimates - parametric bootstrap resampling with continuity correction 128

5.8 Results of the analysis applied to the data set D_{100} in order to test the properties of the occupational groupings 130

5.9 Results of the analysis applied to the data set D_{100} with occupations grouped according to their defaulting proportions in D_{100} . . 132

6.1 Scatter plot of difference in prospective shrunk and model ROC areas against the deviance of the model 158

6.2 Scatter plot smoothers of differences in prospective empirical, model and shrunk ROC areas against the deviance of the model (fitted probabilities are used as the prospective probabilities) 161

6.3 Scatter plot smoothers of differences in prospective empirical, model and shrunk ROC areas against the deviance of the model (proportion infected used as prospective probabilities) 162

6.4 Scatter plot smoothers of differences in prospective empirical, model and shrunk ROC areas against the deviance of the model (midpoints used as prospective probabilities) 164

List of Tables

2.1 Table summarising the results from a diagnostic test using a threshold u 11

2.2 Actual and predicted status for the FEV_1 data using the cutoff $u = 80\%$ 13

4.1 Description of the variables in the melanoma case-control data . 71

4.2 Overestimation of the area under the ROC curve for the melanoma score, using (3.38) with different class intervals 74

4.3 Overestimation of the area under the ROC curve for the melanoma score, using the sample splitting approach with two different sampling fractions 82

4.4 Description of the variables in the breast cancer study 83

4.5 Overestimation of the area under the ROC curve for the breast cancer score, using (3.38) with different class intervals 86

4.6 Overestimation of the area under the ROC curve for the breast cancer score, using the sample splitting approach with two different sampling fractions 91

5.1 Contingency table arising from considering x_i's as observed proportion of 'successes' amongst n_i individuals. 100

5.2 Description of the occupations in the credit defaulting data set . . 112

5.3 Effect of continuity correction applied to the resampled data . . . 126

6.1 Description of the variables in the Pilot Surf/Health Study of NSW Water Board 146

6.2 Empirical, model and shrunk estimates for the simple main effects model fitted to the ear infection data 147
6.3 Examples to verify the approximation of the deviance in calculating $\hat{\tau}^2$ 149
6.4 Results of the simulation procedure using probabilities $p1$ and cell frequencies $n1$ 152
6.5 Results of the simulation procedure using probabilities $p2$ and cell frequencies $n2$ 154
6.6 Results of the simulation procedure using probabilities $p3$ and cell frequencies $n3$ 155
6.7 Results of the final simulation procedure to assess the effect of the doubled cell numbers and cell totals. 156

Acknowledgements

First and foremost, this work would not have been possible without the guidance and wisdom of Professor John Copas. His encouragement and friendship will always be remembered and I could not have wished for a better supervisor.

This work is dedicated to my family, especially my parents who have always been there with love and support.

Thanks to the members of the Department of Statistics for the encouragement and friendship, especially Jonathan, Roberto and Ron.

To all the friends I have made at Cardiff and Warwick, but in particular

- Guy, Ian, Henry, David, James and Kevin for the many games of TS and constantly reminding me how old I am.
- Dr Jim Shields, for the football and the Tocil experience.
- Matt and Jill, for their friendship.

I am grateful to Dr Colin Begg for access to the Melanoma data set used in Chapter 4.

I would also like to thank the Engineering and Physical Sciences Research Council for funding this work.

Declaration

I declare that this thesis is my own work. Parts of this thesis appear in:

- Copas, J. B., and Corbett, P. J. Overestimation of ROC for Logistic Regression. *Biometrika*, (to be published).

Summary

The accuracy of prediction is a commonly studied topic in modern statistics. The performance of a predictor is becoming increasingly more important as real-life decisions are made on the basis of prediction. In this thesis we investigate the prediction accuracy of logistic models from two different approaches.

Logistic regression is often used to discriminate between two groups or populations based on a number of covariates. The receiver operating characteristic (ROC) curve is a commonly used tool (especially in medical statistics) to assess the performance of such a score or test. By using the same data to fit the logistic regression and calculate the ROC curve we overestimate the performance that the score would give if validated on a sample of future cases. This overestimation is studied and we propose a correction for the ROC curve and the area under the curve. The methods are illustrated through way of two medical examples and a simulation study, and we show that the overestimation can be quite substantial for small sample sizes.

The idea of shrinkage pertains to the notion that by including some prior information about the data under study we can improve prediction. Until now, the study of shrinkage has almost exclusively been concentrated on continuous measurements. We propose a methodology to study shrinkage for logistic regression modelling of categorical data with a binary response. Categorical data with a large number of levels is often grouped for modelling purposes, which discards useful information about the data. By using this information we can apply Bayesian methods to update model parameters and show through examples and simulations that in some circumstances the updated estimates are better predictors than the model.

Chapter 1

Introduction

As is evident from its title, this thesis is concerned with the accuracy of ROC curves and the accuracy of logistic models for analysing categorical data. Accuracy is a concept that has a different interpretation to many people. An individual may be concerned with how accurately their model fits the data or how accurate the model is in its predictions. In the course of this thesis we will be solely focusing on predictive accuracy, i.e. the performance of a model or some kind of scoring mechanism in predicting some characteristic of an individual or group.

Predictive accuracy is an important topic in modern statistics. As the understanding of statistical methods and their number of applications increases it is essential to make sure that those who utilize such methods are aware of the problems inherent in prediction. This is no more evident than in the use of statis-

tics in medical applications, where regression analysis is commonplace in trying to relate a number of prognostic factors to an individual characteristic of a patient. In turn, the results of these analyses may be used to predict or identify individuals or groups who maybe be at risk of a certain disease or characteristic for preventative action. In this sense, prediction accuracy is extremely important as we may end up identifying individuals or groups who actually do not need pre-emptive action, or perhaps even worse, exclude those who do.

The situation often arises, especially in medical applications where we wish to assign an individual to either of two groups based on a number of observable characteristics of the patient. These groups could be alive or dead, diseased or non-diseased or could be achieved by dichotomizing a continuous measurement. This problem is commonly termed the *discrimination* problem and its relationship with predictive accuracy is simple. If we form a decision rule based on some statistical procedure, how well does it predict individuals with a certain group membership as actually belonging to that group? This question will be the main topic of this thesis, although we shall briefly discuss predictive accuracy in terms of another interpretation, *calibration*, later on.

The discrimination problem is a common problem in modern statistics and consequently a number of methods exist specifically to address it. Of these, logistic regression and discriminant analysis are perhaps the best known. Logistic

regression models the probability of having a certain characteristic (for sake of a consistent explanation we shall think of larger probabilities in terms of indicating 'positive' status, whatever this maybe) based on available covariates, and discriminant analysis classifies an individual to a group on the basis of a discriminant function $\lambda(\underline{x})$. We use logistic regression in this thesis for a number of reasons, which are described in Chapter 2.

Once we have settled on our scoring mechanism we would like to assess its discriminatory power. A number of indices are available to quantify this, for example the positive and negative predictive values and the error rate but perhaps the most widely used are the true positive and true negative rates, otherwise known as the sensitivity and specificity of a particular score or test. The Receiver Operating Characteristic (ROC) Curve is plot of sensitivity against 1 - specificity and is primarily used as graphical representation of the overall discriminatory power of a score or test. The ROC curve is a regularly utilized tool, especially in medical studies. Zweig and Campbell (1993) report that in the first six issues of the journal *Clinical Chemistry* in 1991 at least 18 studies involved questions about test performance and of these 18, 5 included ROC analysis. The curve's visual nature is appealing, especially to those without a thorough grounding in statistics as it gives an overall graphical description of the usefulness of a score or test without having to explain complicated underlying statistical concepts. The

area under the ROC curve is often used as a single measure of the discriminatory power of the score.

Unfortunately, the simplicity of the ROC curve can mean that it is used and interpreted incorrectly. As mentioned above it is often used by investigators without an adequate knowledge of the underlying concepts of the model involved. For example, it is widely known that if you use the same data to calculate a logistic regression and assess its performance, you will overestimate how well the score discriminates if it were to be assessed on a future data set. Using the same data in constructing and assessing the score is often termed a *retrospective* assessment of performance. Ideally, we would wish to have a new independent set of data from the same population on which to perform a *prospective* assessment of the usefulness of the score. Nearly always in practice such a set of data doesn't exist and we are reduced to using the retrospective assessment as a basis for decisions about the score. It naturally follows that a retrospective assessment will result in both a retrospective ROC curve and area. Herein lies the root of the problem and the one of the main topics of this thesis - it is misleading to use the retrospective ROC curve or area as a true assessment of the performance of the score or test. This is clearly of interest as if the prospective ROC area of a particular score in use is considerably lower than that of the retrospective ROC area, we may be using a score that discriminates less well in practice than other available tests. In

this thesis we aim to explore and estimate the difference between the retrospective and prospective ROC curves and areas.

Another concept of prediction accuracy is *shrinkage*. Before the modelling process we often have some information about the process or topic of study. For example, in categorical data we might have some idea about relationships between 'success' rates between categories, but more often than not this information would not be used in the modelling process. The use of prior information of this kind in the analysis of an estimate or predictor is termed *shrinkage*. Krebs-Brown (2000) discusses shrinkage in terms of the *calibration* of estimates. Calibration is another aspect of predictive accuracy and Krebs-Brown defines a predictor as being well calibrated if 'it gives estimates that are, at least on average close to what we would like to predict'. This naturally leads to the theory of shrinkage correction where we can pre-multiply predictors by a shrinkage parameter to obtain the property of being well-calibrated. Previously, the theory of shrinkage has concentrated on continuous measurements, instead we will look at shrinkage in terms of discrimination, the idea being that by including information from the model and the data we can derive 'shrunk' predictors that will give better discriminatory performance than from the model alone.

Chapter 2 introduces the ROC curve by studying the predictive accuracy of diagnostic tests. A brief history of the origins of the ROC curve is followed

by a formal definition. ROC curves can be studied in a parametric or non-parametric setting and both of these are discussed along with the underlying distributional assumptions. The area under the ROC curve is reviewed and we show how it can be calculated directly from the rank sum statistic. We then include a more in depth examination of the discriminant problem, compare and contrast logistic regression and discriminant analysis and give reasons why logistic regression is the preferred choice in this work. We review the wide ranging area of the statistical assessment of predictions and model adequacy and examine a recently published argument that states that the empirical misclassification rates of a logistic regression are not the maximum likelihood estimates of the true misclassification rates. Finally, we introduce the idea of retrospective and prospective assessments of predictive accuracy and give a brief introduction to shrinkage through the concept of calibration error.

In Chapter 3 we distinguish between the retrospective ROC curve with the distributions of the true positive and false positive (1 - True Negative Rate) rates taken as the empirical distributions over the sample, and the prospective ROC curve with the true population distributions of the rates described above. We define the overestimation of the ROC curve as the difference in true positive rates between the prospective and retrospective curves for common values of the false positive rates. We derive a closed form expression for the expectation of this

difference and this is shown to be quite substantial for small sample sizes, particularly when the dimension of the covariate vector used in the logistic regression is not small relative to the size of the data. We also derive a closed form expression for the corresponding overestimation in the area. By a series of further approximations examining the conditional distribution of the covariate vector given the score we establish an approximation to the overestimation in the area which is a much easier and transparent calculation.

The formulae derived in Chapter 3 are used to examine the overestimation of ROC in two examples in Chapter 4. The first example concerns a case-control study of melanoma (a type of skin cancer) and the second a smaller study of prognosis in breast cancer. Then, to test the validity of the asymptotic formulae for the overestimation of the curve and area we perform a series of bootstrap simulations and briefly discuss the results.

Chapter 5 introduces a methodology for improving prediction accuracy through shrinkage in the modelling categorical data. We introduce the methodology by reviewing a standard result in Bayesian inference which allows us to update parameter estimates by including prior beliefs about random variables. We then extend this argument by considering the random variables to be the observed proportions of data in a contingency table. The methodology can be expanded to include the empirical logistic transform and consequently logistic regression

to derive a set of 'shrunk' estimates. These take the form of a weighted average incorporating R, the matrix of correlations between 'success' rates in categories and a shrinkage parameter τ^2 which is shown to be related to the *deviance*, the goodness-of-fit statistic for logistic regression. We test the properties of the theory by introducing an example studying a data set consisting of individuals' occupations and credit default status. Finally, by performing a number of bootstrap simulations on this data we can examine the effect of varying the correlations between the defaulting rates of a number of predefined occupational groupings.

In Chapter 6 we generalise the methodology introduced in Chapter 5 to any modelling procedure. The meaning of shrinkage in this context is explained by the way of a short example to show that we are now concerned with the misspecification of the model fitted to the data. The logistic regression model for a number of categorical covariates is studied in depth by way of an artificial example. We are able to use the retrospective and prospective ROC curves to show that the 'shrunk' estimates can produce predictors which have better discriminatory power than the model or data alone. Finally, we postulate a decision rule based on the deviance of the logistic regression model for choosing between the model and 'shrunk' estimates for prediction purposes.

Finally in Chapter 7 we summarise this work and reflect upon the possibilities for its extension.

Chapter 2

ROC Curves, Logistic Regression and Predictive Accuracy

In this chapter we will discuss and investigate the background and some aspects of ROC curves, modelling and classification procedures related to binary outcomes and the predictive accuracy of models in a wide-ranging context. Each topic is studied individually and some parallels are drawn between them at the end of the chapter.

2.1 Diagnostic testing and ROC curves

The accuracy of diagnostic scores or tests is a commonly studied topic, especially in medical situations where we might want to test the effectiveness of a new method over a current scheme. Although medical applications provide a natural focus, such 'scores' are commonplace in other areas e.g. finance (Hand and Henley (1997) and criminology (Copas *et al* (1996)). Diagnostic tests can take the form of a single measurement e.g. blood pressure, a function of a number of measurements on a patient or qualitative data on a rating scale. From now on we shall focus on the class of tests or scoring mechanisms that produce a binary (dichotomous) outcome e.g. diseased and non-diseased (the so called two class case - see the recent article by Hand (2001) for more details). Call the output from any kind of diagnostic testing procedure that produces a score s.

We assume that once an individual has been assigned a test value then the actual status of the patient is determined via some examination or 'gold standard', call this status E (positive outcome) and $\overline{E}$ (negative outcome). We shall assume from now on that higher values of a test or score are indicative of a positive outcome E. Once the test result and actual status of the individual are known then the ultimate aim is to assess the ability of the test to discriminate between the groups E and $\overline{E}$. If the test itself produces an s that directly states group membership we can summarise the results of the test in the 2 x 2 contingency

table in Table 2.1. If s is a score on a categorical or continuous scale then by choice of a suitable threshold value u we can also represent the information in the test by Table 2.1.

		Predicted Status		
		$\overline{E}$	E	
Actual Status	$\overline{E}$	a	b	$a+b$
	E	c	d	$c+d$
		$a+c$	$b+d$	N

Table 2.1: Table summarising the results from a diagnostic test using a threshold u

From Table 2.1 we can use a number of measures to assess the accuracy of the diagnostic test. It is obvious to see that if entries b and c are both zero then the diagnostic test is a perfect discriminator/classifier. If the score is continuous or categorical then the entries in the table change with the choice of threshold u. Hand (1994) gives an overview of a number of methods that can be used to assess a diagnostic test that produces a 2x2 contingency table.

Perhaps the two most widely known indices of diagnostic accuracy are *sensitivity* and *specificity*. They are defined as follows :-

- the sensitivity (also called True Positive Rate) of a diagnostic test is the proportion of those who actually have E who are predicted as having E ($= \frac{d}{c+d}$ from Table 2.1)

- the specificity of a diagnostic test is the proportion of those who actually have $\overline{E}$ who are predicted as having $\overline{E} (= \frac{a}{a+b})$

The quantity 1 - specificity is often referred to as the False Positive Rate. Although the terms sensitivity and specificity are commonly used we shall use the terms True Positive Rate (abbreviated to TPR) and False Positive Rate (abbreviated to FPR) to determine the performance of a diagnostic test as by their very names they are easier to understand what they are measuring.

As we mentioned u is an arbitrary threshold value chosen to dichotomise s. So for a particular test we can have a wide range of threshold values that will produce pairs of True Positive and False Positive values (TPR,FPR). When these values are plotted for all thresholds we can establish a graphical representation of the effectiveness of a diagnostic test. Such a plot is called a Receiver (or Relative) Operating Characteristic (ROC) Curve and is the focus of much of this thesis. Zweig and Campbell (1993) give a comprehensive overview of ROC curves and associated topics, many of which we will touch upon in this chapter.

An example of a ROC curve for the two-class diagnostic test that we have introduced above is given by Campbell and Machin (1993). They report a study

(see cited reference for full details of the accompanying paper) where Forced Expiratory Volume (FEV_1) was measured in 40 non-smoking individuals with and without a condition called coal-workers pnuemoconiosis. The purpose of the test is to determine whether FEV_1 is a good predictor of the disease. The authors note that a commonly used threshold value in this context is $u = 80\%$ of the FEV_1 you would expect to find in a healthy individual of same height and age. Using this cutoff provides us with the 2 x 2 contingency table in Table 2.2.

		Predicted Status		
		$\overline{E}$	E	
Actual Status	$\overline{E}$	22	5	27
	E	5	8	13
		27	13	40

Table 2.2: Actual and predicted status for the FEV_1 data using the cutoff $u = 80\%$

From Table 2.2 we can easily calculate the True Positive Rate (proportion correctly classified as E) to be $\frac{8}{13} = 0.615$ and the False Positive Rate to be (proportion incorrectly classified as positive) to be $\frac{5}{27} = 0.185$ for this particular u. If we calculate the TPR and FPR over a number of values of u then we can summarise the results in the ROC curve in Figure 2.1. (By convention the points on an ROC curve are joined together by straight lines.)

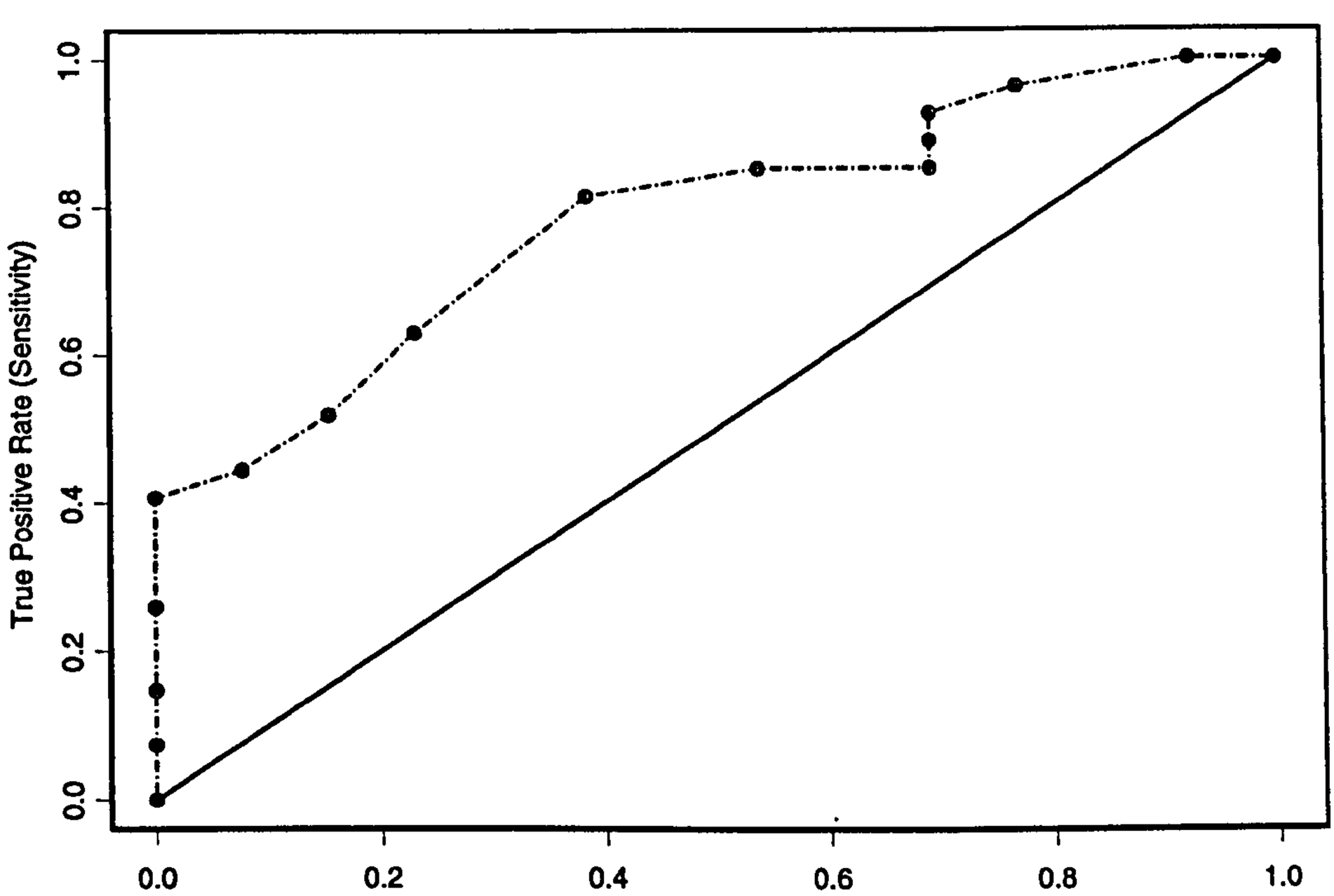

Figure 2.1: ROC curve for FEV_1 data

The initial purpose of the ROC curve is to enable us to view the overall effectiveness of the diagnostic test and make a decision on an optimal True Positive Rate or False Positive Rate. A test that perfectly separates the populations of interest will have an ROC curve that will pass through the points (0,0), (0,1) and (1,1). The ROC curve for a test that has no discriminatory power will lie along the diagonal from (0,0) to (1,1), i.e. True Positive Rate = False Positive Rate over all thresholds u. If the ROC curve lies below this diagonal then we simply reverse

the threshold value, for example $> u$ becomes $\leq u$ etc. Depending on the nature of the study we may wish to achieve as high a True Positive Rate as possible or minimize the False Positive Rate. ROC curves have the useful properties of being invariant to any monotonic transformations of the score s and being independent of the prevalence of the characteristic of interest (E). This removes the problem of selecting samples with a prevalence representative of the population at large.

The origins of ROC curves lie within the realm of signal detection theory, Green and Swets (1966) and Egan (1975) giving good introductions. Lusted (1971) was one of the first to make use of the ROC framework in the area of medical decision making with applications to radiographic chest films. The seminal and oft-quoted paper of Metz (1978) drew together the strands of diagnostic accuracy, sensitivity, specificity and ROC curves and illustrated the relationship between ROC and the cost/benefit analysis that naturally arises from decisions in diagnostic testing.

We will now provide a formal definition of the ROC curve. If s is a diagnostic score produced by some method, and we have two groups or populations indexed by the binary indicator $y = 0$ or 1 (we assume that higher values of s imply 'positive' status or $y = 1$), the threshold u gives the *false positive* rate

$$F_0(u) = \mathrm{P}(s \geq u | y = 0)$$

and the *true positive* rate

$$F_1(u) = \mathrm{P}(s \geq u | y = 1).$$

Definition 2.1 *The ROC curve, $\mathcal{C}$, is the graph of $F_1(u)$ against $F_0(u)$ as u ranges over all possible values,*

$$\mathcal{C} = \{(F_1(u), F_0(u)) : -\infty < u < +\infty\}. \tag{2.1}$$

So far we have looked at a strictly non-parametric approach to the analysis and construction of the ROC curve i.e. the values of the True Positive rate and the False Positive rates are calculated directly from the 2 x 2 table. At the opposite end of the spectrum there exists a large amount of work on parametric approaches to ROC estimation (see Hanley (1996), for example). For both non-parametric and parametric approaches, it is assumed that there are two underlying overlapping distributions of the populations of those with E and $\overline{E}$ (see Figure 2.2.).

As the threshold u moves along the decision axis we obtain differing True Positive rates and False Positive rates which are plotted to form the ROC curve. In the non-parametric case these distributions are estimated by their empirical analogues. In the parametric case we assume the form of these distributions and then use some procedure to estimate their parameters. By far the most common case is when both distributions are assumed to be normal (the so called *binormal*

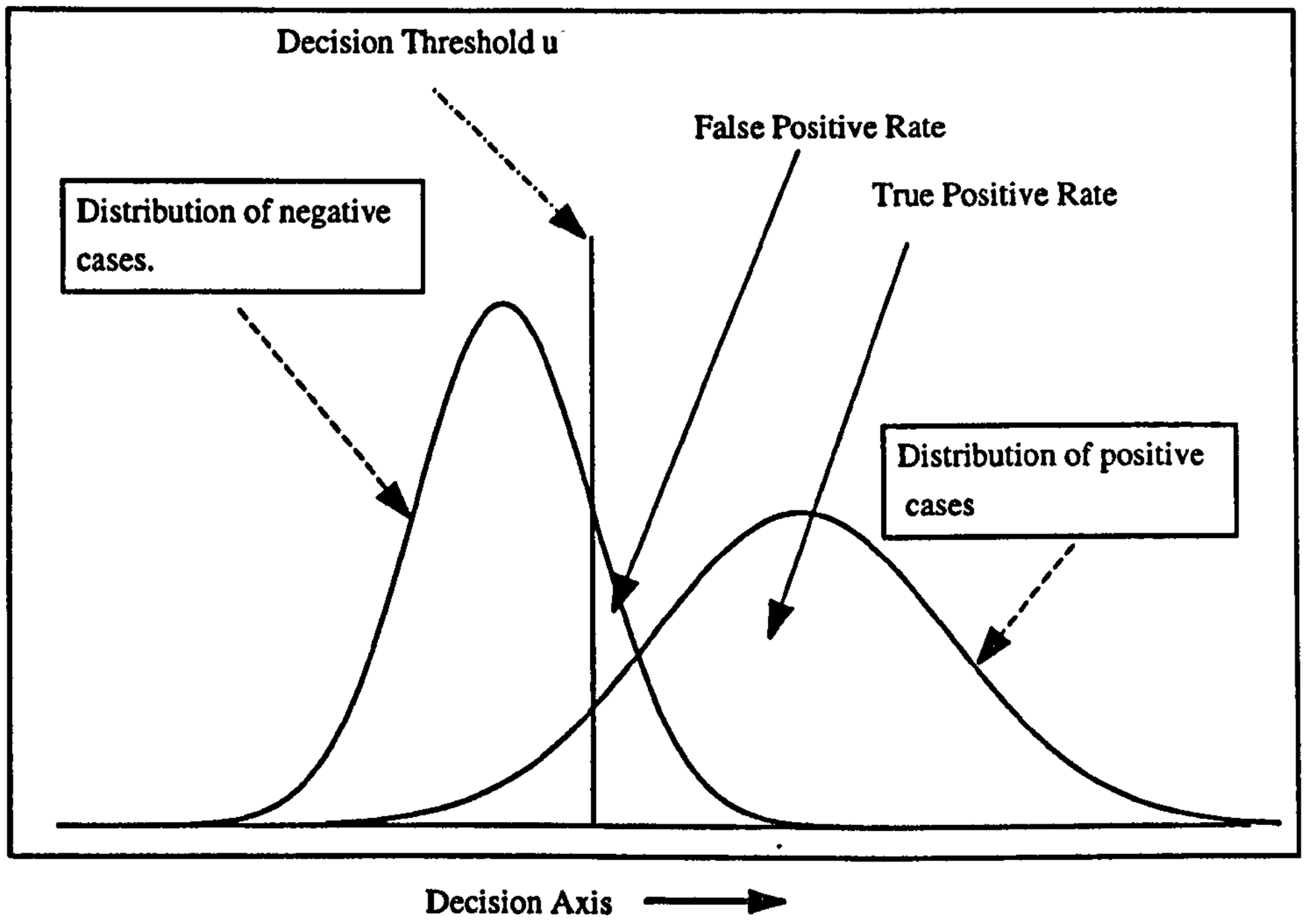

Figure 2.2: Distributions of populations with and without the attribute of interest

case), see Metz and Kronman (1980) and Hsieh and Turnbull (1996) for good introductions.

The binormal ROC curve is fully described by two parameters based upon the underlying Gaussian distributions, the difference in means and the ratio of the standard deviations. These parameters can be measured by maximum likelihood estimation, see Dorfman and Alf (1968) for the full algorithm. Hanley (1988) justifies the choice of the binormal form, especially in the context of ROC curves

based on rating data. Rating data occurs frequently in medical studies when an investigator is asked to make a judgement on whether a particular characteristic lies within one of a number of categories. These categories can be interpreted as class intervals whose frequencies are the number of decisions made. This data can be dichotomised by grouping one or more categories together.

Of course, the binormal model is not the only parameterisation of the ROC curve we could postulate. Pairs of logistic (*bilogistic*) and negative exponential distributions are compared and contrasted with the binormal model by Hanley (1988). Metz and Kronman (1980) give significance tests for comparing differences between or among parametric ROC curves. Recently, Lang and Aspelund (1999) have proposed a more flexible class of binormal models to assess the differences between two ROC curves and Lloyd (1998) has advocated the method of kernel density estimation in the construction and comparison of smoothed ROC curves. Lloyd states in the justification of the semi-parametric method that the non-parametric method suffers from the contamination of statistical variation and the parametric method is inherently biased.

Zweig and Campbell (1993) give a table of advantages and disadvantages for both the parametric and non-parametric approaches. They note that for continuous data, which we are most concerned with the non-parametric ROC is the preferred choice as it passes through all the observed points and provides

unbiased estimates of the True Positive Rate, False Positive Rate and the area under the ROC curve, which we shall study next.

2.1.1 The area under the ROC curve

The ROC curve is a powerful method to 'globally' represent the effectiveness of a diagnostic test but it may be more useful to summarise this overall effectiveness in a single number. Calculating the area under the ROC curve enables us to do this. As mentioned earlier, a test that is a perfect discriminator will produce a curve that will follow the left and upper edges of the unit square and hence produce an area under the curve of 1. Conversely, a test with no discriminatory power will have an area of 0.5. For example, the curve in Figure 2.1. for the FEV_1 data has an area of approximately 0.79 indicating that FEV_1 provides good discrimination between the diseased and non-diseased. A transformation of the ROC area on to the scale 0-1 is given by 2(Area - $\frac{1}{2}$), sometimes called the Gini Coefficient, which is often mentioned in financial applications.

The area under the ROC curve has been extensively studied - see Hanley and McNeil (1982) and Hilden (1991) for example. It can be used as a measure of diagnostic accuracy by itself or can be used to compare a number of ROC curves, possibly derived from the same sample. The calculation of the area can be done in a number of ways - Bamber (1975) was one of the first to relate the

equivalence between the area under the curve and the two sample Wilcoxon rank sum statistic. Hanley (1982) studies this relationship in detail and provides an algorithm for calculating the area under the curve and its standard error for a rating experiment using the Wilcoxon statistic. Another method of calculating the area is by using the fact that the area under the ROC curve is equivalent to the probability of an observation from the distribution of E and an observation from the distribution of $\overline{E}$ being in the correct order (Green and Swets (1966)) i.e.

$$\text{Area under the curve } (A) = P(S_E > S_{\overline{E}})$$

Now suppose that $P(E) = p$ and that we take a random sample of size n from the marginal distribution of S. Let $y_i = 1$ if E happens for the ith observation and $y_i = 0$, if it does not, and let

$$n_1 = \sum_i y_i = n - n_0.$$

Define $s_i^{(1)}$ be the ith largest of the n_1 observations with $y = 1$, and suppose that this is the j_ith largest amongst the order statistic of all the observed values of S. Then in the notation above

$$P(S \leq s_i^{(1)}|\overline{E}) = P(S_E \geq s_{\overline{E}}|S_E = s_i^{(1)})$$

is estimated by

$$\frac{j_i - i}{n_0}$$

Hence A is estimated by

$$\widehat{A} = \frac{1}{n_1}\sum_i \frac{j_i - i}{n_0} = \frac{\sum j_i}{n_0} - \frac{n_1 + 1}{2n_0}$$

Let r_i be the rank in the overall order statistic of the observation s_i. Then

$$\sum_{i=1}^{n_1} j_i = \sum_{i=1}^{n} y_i r_i$$

The term on the right hand side of the above expression is just the rank sum statistic i.e.

$$R = \sum_{i=1}^{n} y_i r_i$$

So as described above, an estimate of the ROC area can be calculated directly from R.

$$\widehat{A} = \frac{R}{n_0} - \frac{n_1 + 1}{2n_0} \tag{2.2}$$

The area under the ROC curve can be used to compare diagnostic tests carried out on different samples. When differing diagnostic tests are applied to the same sample then the areas under the respective ROC curves cannot be compared directly as they will have a positive correlation (see Hanley and McNeil (1983) for a full description/solution to this problem). Although the area under the ROC curve has a direct interpretation, its variance and the covariances between different areas have been difficult to quantify. Hanley and Hajian-Tilaki (1997) describe these problems in detail and compare the non-parametric methods of Delong *et al* (1988) and Wieand *et al* (1989) for computing variances and

covariances. Finally, Beck and Shultz (1986) show it is possible to calculate non-parametric confidence intervals for the area under the ROC curve

Although we will be focusing mainly on the ROC curve and the area underneath it there are numerous other areas of interest in ROC methodology. For example, Campbell (1994) gives a number of non-parametric confidence intervals based on the ROC curve rather than the area. Tosteson and Begg (1988) present a method for estimating ROC curves through generalised ordinal regression models. These models have the benefit of allowing the adjustment of ROC curve parameters for covariates of interest. Smith *et al* (1996) present an analogous method using generalised linear models that removes some of the problems associated with the Tosteson and Begg method. Raubertas *et al* (1994) propose a method for calculating the sensitivity and specificity for classification trees and hence plot the ROC curve. Choi (1994) studies the relationship between the differing slopes of the ROC curve and the likelihood ratios of a diagnostic test while Hanley (1991) studies the effect of verification bias, one of the many types of bias in diagnostic testing on ROC curves.

Finally, ROC curves are not the only graphical representation of the information in a diagnostic test. Adams and Hand (1999) present a loss based method for comparing diagnostic tests that unlike the area under the curve takes into account misclassification costs. Taube (1986) provides a illustration of sensitivity

and specificity that usefully incorporates prevalence of the characteristic of interest whilst Copas (1999a) has proposed the Logit Rank Plot, a new method for evaluating the effectiveness of risk scores. Copas' analogous measure to the area under the ROC curve, the *logit rank slope* is shown to be more sensitive to the distribution of s than the area measure. All of the ROC curves in this thesis can be generated using the S-code for ROC curves in Appendix A of Copas (1999a).

2.2 Classification methods

We have described how to assess the diagnostic accuracy of a score s, but the question remains how to arrive at s in the first place. Constructing s for the two-class case involves the **classification problem**, which concerns assigning an individual to one of two categories depending on the information available for the individual. There are many different types of classification procedures (see Hand (1997) for an overview) that can be used but we will focus on only two, logistic regression and discriminant analysis. We will briefly review the two approaches and explain why logistic regression is the preferred choice in this thesis.

2.2.1 Logistic regression

Logistic regression is a commonly used tool in modern statistics and has its origins in the field of generalized linear models (McCullagh and Nelder (1983)). It is especially useful in the context of ROC curves in that for each individual it produces an estimate of the probability of group membership, which along with the binary indicator can be used to easily construct the associated empirical ROC curve.

Consider the situation where we have independent observations from N populations where each is distributed $Y_i \sim B(n_i, p_i)$. The proportion of successes in each of the N populations (Y_i/n_i) can be modelled by

$$f(p_i) = \beta^T \mathbf{x_i}$$

where f is commonly called the link function, β is the vector of coefficients for the ith individual and $\mathbf{x_i}$ is the vector of covariates and dummy variables for factors.

For logistic regression, the link function is taken to be the **logit function** i.e.

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta^T \mathbf{x_i} \tag{2.3}$$

which is simply specifying a linear structure for the log of the odds, $p_i/1 - p_i$.

The general form for logistic regression is widely used for the analysis of multivariate data involving a binary outcome. All logistic regressions in this thesis were carried out within the *S-Plus* (Mathsoft Inc.) statistical package which

estimates the maximum likelihood estimates of the parameters β by iteratively reweighted least squares (IRLS). Pregibon (1981) analyses the effect outlying responses and extreme points can have on the maximum likelihood estimates.

Of interest is how well a particular model fits the data. This is expressed by the *deviance* or log-likelihood ratio statistic of the model, which is defined to be twice the difference between the saturated model and the model of interest i.e.

$$\text{Deviance} = D = 2[l(\mathbf{p}; y) - l(\hat{p}\,; y)] \tag{2.4}$$

where $\mathbf{p}$ is the vector of maximum likelihood estimates of the saturated model and $\hat{p}$ is the vector of estimates of the model of interest.

Hypotheses about lack of fit can be assessed using the relationship (see Dobson (1990) for a full proof) that

$$D \sim \chi^2_{N-m}$$

for a particular significance level, where N is the number of observations and m the number of parameters in the model. We use the Wald test (see Hosmer and Lemeshow (1989)) for testing hypothesis about individual variables and methods such as *stepwise logistic regression* to sequentially add or remove variables to obtain the most parsimonious model.

2.2.2 Discriminant analysis

As with logistic regression, discriminant analysis is a widely known and researched area in statistics (see Hand (1997) for a comprehensive overview). Discriminant analysis approaches the classification problem from a slightly different viewpoint to logistic regression. Essentially, given two distinct populations indexed by $y = 0$ and $y = 1$ say, with densities f_0 and f_1, how do the distributions differ most markedly?

Suppose we have a new measurement vector $\mathbf{x}'$ with unknown classification y' and this measurement could arise from f_0 or f_1. Let both f_0 and f_1 be multivariate normal with different means and equal covariance matrices i.e.

$$f_0 \sim \mathrm{N}(\mu_0, \Sigma)$$

$$f_1 \sim \mathrm{N}(\mu_1, \Sigma)$$

and the prior estimates of belonging to class 0 or 1 are given by π_0 and π_1. Then Fisher's linear discriminant function is given by

$$\lambda(x) = \mathbf{x}'\Sigma^{-1}(\mu_1 - \mu_0) + \log(\frac{\pi_1}{\pi_0}) - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 \qquad (2.5)$$

We then classify according to this function i.e. assign to class 1 if $\lambda(x) > 0$ and to class 0 if $\lambda(x) < 0$. Hand(1997) states that this rule is optimal for elliptical distributions (e.g. multivariate normal) but does not assume multivariate normality.

Further extensions to the discriminant function can include higher order terms such as products and squares of variables (quadratic discriminant analysis) and take in more than two classes which involves the introduction of the sample covariance matrix. Like logistic regression, discriminant analysis is available in all major statistical analysis packages.

2.2.3 Comparison between discriminant analysis and logistic regression

Cox and Snell (1970) include a discussion on the relationship between discriminant analysis and logistic regression. From the linear discriminant function and use of Bayes theorem they arrive at (2.3), the logistic regression equation which illustrates the near equivalence of the two methods under the assumption of multivariate normality. They conclude that logistic regression is slightly preferable as we do not have to make any distributional assumptions about $\mathbf{x}$ within the two populations. But if we can assume multivariate normality then the discriminant analysis approach is usually more efficient. Efron (1975) compares the asymptotic efficiency of the two procedures and determines that logistic regression is between a half and two thirds as effective as normal discriminant analysis. Press and Wilson (1978) advocate the use of the logistic regression form for classifying individuals into populations in the presence of qualitative data which rules out

multivariate normality. This argument is the reasoning behind the use of logistic regression throughout this paper.

In Chapter 4 we will look at examples where at least one of the variables will be qualitative and in Chapter 5 we look at logistic models where all of the variables are categorical, ruling out multivariate normality completely. Although we choose not to use it in practice, discriminant analysis should not be dismissed outright. Under certain assumptions it is a powerful method of multivariate analysis and is closely related to some of the ideas and methods we shall discuss in the next section and in Chapter 3.

2.3 Assessment of statistical predictions and adequacy of models

Once we have decided upon a model, such as logistic regression for classifying individuals into one of two populations it follows naturally that we should want to assess its performance e.g. in medical terminology is the rule or model better than the current standard or are we simply aiming for a certain accuracy. Assessing the accuracy of models and predictions covers a large area of theory and so for simplification we discuss the topic in three sections.

2.3.1 Model accuracy

We may wish to assess the accuracy of a particular model in the context of model choice e.g. comparing a number of models for some desirable properties. For multiple linear regression, we have established methods like Mallows C_P (Mallows (1973)) where

$$C_P = \frac{\text{RSS}}{\sigma^2} - (n - 2p) \tag{2.6}$$

where RSS is the residual sum of squares for the model, p the dimensionality of the model, σ^2 the known variance and n the sample size.

A more general formulation for model choice is given by Akaike's Information Criterion (see Stone (1977) for a brief discussion), where if m indexes the model, choose m to maximize

$$L(m, \hat{\theta}_m) - p_m \tag{2.7}$$

where $L(m, \theta_m)$ is the log-likelihood function, $\hat{\theta}_m$ is the maximum likelihood estimate of θ_m and p_m is the dimensionality of model. As Stone notes, p_m is a correction term that penalises the model m with high dimensionality, and maximizing (2.7) is equivalent to minimizing (2.6). Allen (1971) advocates the use of the mean square error prediction as a criterion for selecting variables in multiple regression whilst for logistic regression, we could use the deviance measure described in (2.4) to make a comparison between different models. The reader is referred to Harrell *et al* (1984 and 1996) for a full discussion of model building

and selection.

From our viewpoint, model accuracy is better described in terms of two concepts, *reliability* and *discriminability* ((Hand (1994) also mentions *separability* which is of less interest in the context of this discussion). Hence we define these two terms

- **reliability** or **calibration** is concerned with the agreement between the predicted probabilities and the true probabilities of occurrence.

- **discriminability** is concerned with how well the rule assigns an individual to its correct status.

Pearce and Ferrier (2000) give a useful introduction to the predictive performance of models in the context of ecology using the two terms defined above and utilize the ROC curve in their discussion of discrimination. We will investigate the above terms further in the context of error to show they are intimately linked with the work in Chapters 3 and 5.

2.3.2 Discrimination error

Discrimination error is concerned with the misclassification rate or error rate of a prediction rule i.e. what is the probability of predicting a future observation incorrectly (a comprehensive introduction on the estimation of misclassification

can be found in Hand (1997)). Van Houwelingen and Le Cessie (1990) introduce error rates for a wide range of modelling procedures (including logistic regression) and propose some strategies for estimating this rate and correcting the prediction rule. In an early paper, Lauchenbruch and Mickey (1968) present and compare several methods of estimating error rates in discriminant analysis.

Before we look at error rates for logistic regression in detail it is important to note a recent paper by Lloyd (2000), who mentions that the empirical estimates of the misclassification rates of a logistic regression are not the maximum likelihood estimates of the true rates for a given threshold if the logistic regression model is assumed to be true. It is worth studying this assertion in greater detail. In Lloyd's notation, let individuals belong to group G_0 or G_1, let $\pi_1(x)$ be the probability of being in group 1 given covariate values x and $\hat{\pi}_1(x)$ be it's estimator. The accuracy of the classification rule is summarised by the error rates

$$\alpha_0(c) = P(\hat{\pi}_1(x) > c|G_0) \quad \text{and} \quad \alpha_1(c) = P(\hat{\pi}_1(x) \leq c|G_1)$$

for a given threshold c. The ROC curve is a plot of 1 - $\alpha_1(c)$ against $\alpha_0(c)$ and empirical estimates of $\alpha_i(c)$ are easily calculated. Rename $\hat{\pi}_1(x)$ as Z and denote the distribution function of Z conditional on G_i as $F_i(z)$ and densities by $f_i(z)$. The core of the paper is the fact that the empirical estimates of the F_i's directly contradict the model underlying the estimated classification rule $\hat{\pi}_1(x)$.

Specifically under the model, the paper shows that

$$\frac{f_1(z)}{f_0(z)} = \frac{z}{1-z} \quad \text{or} \quad f_1(z) = \phi(z) f_0(x) \tag{2.8}$$

where $\phi(z) = z/(1-z)$ or the odds of membership of G_1. The fully empirical estimators of the F_i's are

$$\hat{F}_0(z) = \frac{1}{n_0} \sum_j (1 - Y_j) I_{z_j \leq z} \quad \text{and} \quad \hat{F}_1(z) = \frac{1}{n_1} \sum_j Y_j I_{z_j \leq z}$$

but these estimators ignore the restriction on the densities in (2.8). Lloyd goes on to estimate the maximum likelihood estimates of the distribution functions, $F_j(z)$ in terms of their probability functions $\hat{p}_j(k)$. It follows that

$$\hat{\alpha}_0(c) = \sum_{k:z_k > c} \hat{p}_0(k) \quad \text{and} \quad \hat{\alpha}_1(c) = \sum_{k:z_k \leq c} \hat{p}_1(k) \tag{2.9}$$

These estimators are maximum likelihood under the logistic regression model. In an example, Lloyd shows that these estimators provide smoothed estimates of the empirical error rates and a bootstrap simulation shows that the bias of the maximum likelihood estimates is very small.

Suppose we have a sample of n patients with each individual having a vector of explanatory variables and a binary classification of true group status i.e. (x_i, y_i). From this we can realise a particular logistic regression on the explanatory variables $\mathbf{x}$, call this $\mathrm{LR}_{\mathbf{x}}$ with predicted probabilities $\hat{p}_i$. Define predicted class membership by the prediction rule

$$\hat{\gamma}_i \;=\; 1 \text{ if } \hat{p}_i > 0.5$$

$$= \quad 0 \text{ if } \hat{p}_i \leq 0.5 \tag{2.10}$$

Now Efron (1986) suggests that the *apparent error rate* is given by the number of cases ($y_i = 1$) in the **original data** that are misclassified by $\mathrm{LR}_{\mathbf{x}}$ i.e.

$$\text{Apparent Error rate} \;=\; \frac{\sum_i^n I(c_i)}{n}$$

where $I(c_i) = 1$ if $\hat{\gamma}_i \neq y_i$. Now this apparent error rate is usually biased because we have used the original data in constructing and assessing $\mathrm{LR}_{\mathbf{x}}$. We will call any fit of a prediction rule on the data from which it is constructed the **retrospective fit.** Ideally we would wish to have some new independent data vector y' to calculate the *true error rate* of the prediction rule. Conversely, we call any fit of a prediction rule on a future independent data vector the **prospective fit.** Efron states that the analogue of (2.10) for the true error rate is the expected proportion of incorrect predictions i.e.

$$\text{True Error rate} \;=\; \mathrm{E}\left[\frac{\sum_i^n I(c_i')}{n}\right]$$

where $I(c_i') = 1$ if $\hat{\gamma}_i \neq y_i'$. He defines the difference between the true and apparent error rates to be the *optimism* and gives an estimate of the optimism for logistic regression as being

$$\text{Optimism} = \omega(\hat{p}) = \frac{2}{n}\sum_1^n \hat{p}_i(1-\hat{p}_i)\phi\left(\frac{\hat{t}_i}{\sqrt{\hat{d}_i}}\right)\sqrt{\hat{d}_i}. \tag{2.11}$$

Here $\phi(z)$ is the standard normal density,

$$\hat{t}_i = \log\left(\frac{C_0}{1-C_0}\right) - \hat{\beta}^T x_i,$$

C_0 is the cutoff for classifying individuals into one of two groups (taken to be 0.5 in (2.10)), $\hat{\beta}$ is the vector of parameter estimates from the logistic regression and $\hat{d}_i = x_i^T \hat{\Sigma}^{-1} x_i$ where $\hat{\Sigma}^{-1}$ is the usual estimate of the covariance matrix of the parameter estimates $\hat{\beta}^T$. The expression in (2.11) is quite similar to the expression for overestimation of ROC which we shall present in Chapter 3, although we do not condition on the classification rate C_0.

Efron goes onto extend the procedure to generalised linear models and investigates the properties of these error rates under various resampling schemes such as cross-validation. As shown by an earlier paper (Efron (1983)), the choice of resampling scheme is important for deriving the properties of estimators for error rates of prediction rules as cross-validation is shown to give a nearly unbiased estimate of the true error rate. Bootstrap resampling is shown to be a more efficient method for smaller sample sizes. Gong (1986) compares three resampling schemes, the bootstrap, cross-validation and the jackknife for estimating the difference between the true and apparent error rates and states that the bootstrap has the best performance of the three. Cheng and Hseuh (1999) adopt a slightly different method in correcting bias from misclassification in logistic regression models. Their bias correction methods are heavily dependent on the choice of validation sub-sample and the misclassification probabilities.

2.3.3 Calibration error and sampling techniques

As described above, calibration error occurs when the probabilities generated from a prediction rule do not agree with the true probabilities of occurrence. When a prediction rule is applied retrospectively (e.g. to the data from which it was derived) then the fit is nearly always better than a prospective fit (e.g. to a future set of independent data). A solution to this involves the notion of *shrinkage* (Copas (1983)) which is defined to be the amount of bias induced by the retrospective fit over the prospective fit. Copas introduces the argument of 'pre-shrunk' predictors where in a linear multiple regression we scale the least squares estimate of the vector of regression coefficients.

To avoid confusion with the binary indicator notation $y = 0$ or 1 with which we have been working so far we define the continuous outcome in linear regression to be y_{CONT}. Copas (1983) considers a preshrunk predictor, $(\tilde{y}_{CONT})$ of y_{CONT} given by

$$\tilde{y}_{CONT} = \hat{\alpha} + \hat{K}(k_c)\hat{\beta}^T x$$

where $K(k_c)$ is an estimated shrinkage factor with parameter k_c given by

$$\hat{K}(k_c) = 1 - \frac{k_c \hat{\sigma}^2}{n\hat{\beta}^T V \hat{\beta}} \tag{2.12}$$

where $\hat{\sigma}^2$ is the residual mean square estimate, $\hat{\beta}$ are the estimates of the parameters β, n is the sample size and $V = n^{-1}X^T X$ where X is the data ma-

trix. As we mentioned above the predictive mean square error ($PMSE$) can be used to assess the accuracy of a predictor. For the usual least squares estimator $\hat{y}_{CONT} = \alpha + \hat{\beta}^T x$, Copas (1983) shows that

$$PMSE(\tilde{y}_{CONT}) < PMSE(\hat{y}_{CONT}) \text{ for all } O < k < \frac{2(p-2)}{r}$$

where $r = \frac{2}{n-p-1} + 1$.

Van Houwelingen and Le Cessie (1990) propose a type of heuristic shrinkage parameter:

$$\hat{c}_{heur} = \frac{\text{model}\,\chi^2 - p}{\text{model}\,\chi^2}$$

which depends on the fit of the model. This isn't that remarkably different from the shrinkage parameter suggested by Copas. Indeed, the idea of predictions being scaled down by factor which depends upon the model fit is investigated further in Copas (1999b). In proposing the shrinkage estimator in (2.12), distributional assumptions about the sample from which the prediction rule is constructed are made. Copas and Jones (1986) address the robustness of the shrinkage estimators to the departure from these assumptions and conclude that even large differences between past and future samples have little impact on the validity of the estimator. A new non-parametric version of the shrinkage estimator has been proposed by Copas (1987) when assessing the cross-validation shrinkage of predictors, and Copas (1999a) discusses shrinkage of the logit rank slope, a measure of discrimination analogous to the area under the ROC curve.

One of the first examples of a methodology to make gains on classification accuracy using a shrinkage type procedure is given by Friedman (1989). In his paper he considers linear and quadratic analysis in the small-sample, high-dimensional setting (similar to the situation studied in Chapters 5 and 6). He studies an alternative to the usual maximum likelihood estimation of the covariance matrices where parameters are customised to individual problems by jointly minimising a sample-based estimate of the future misclassification risk.

In the preceeding chapter we have mentioned the terms cross-validation, bootstrap and to a lesser extent, jackknife. Stone (1974) gives a useful introduction to cross-validation in the assessment of statistical predictions and introductions to the bootstrap and jackknife are given by Efron (1975). Essentially they are resampling plans, enabling us to reuse our original data many times to achieve 'unbiased' estimates of error rates, areas under the ROC etc. In its basic form cross-validation can take the form of sample splitting (split the original sample into a *construction* or *training* set on which the prediction rule is realised and a *validation* set upon which we can test the prediction rule). Stone (1974) studies the special case of the 'leave one out' case where we have a construction set of size $(n-1)$ individuals and a validation set of size 1. This is repeated n possible ways. Bootstrapping at its basic level takes the form of simply resampling without replacement from the original data set (the reader is referred to Efron

(1975) for technical details). We have already mentioned above that in some specialised instances bootstrap out performs cross-validation, and for ease of use we will utilise it in simulations throughout this thesis.

Finally, it was mentioned in the introduction to the thesis that in Chapters 5 and 6 we use a Bayesian approach to update the estimates of (contingency table) cell probabilities in an attempt to gain discriminatory power. One of the attractive properties of the method is that is applicable in situations where there are many cells and few observations per cell. An analogous approach is given by Bishop *et al* (1975) where psuedo-Bayes estimates of cell probabilities are calculated by establishing a weighting factor (based on the data itself) and using this factor to update the probabilities.

2.4 Common issues

So far we have discussed ROC, logistic regression and predictive accuracy separately, but how do they all fit together? It is clear that the ROC curve can be used to assess the discriminatory ability of a particular logistic regression but how does this relate to the future fit of the logistic regression for a new independent set of observations?

After calculating a particular score using logistic regression, what we are really

interested in is how the score would perform in practice (i.e. in the future). In the terminology above this would involve a prospective assessment of the score. It naturally follows that we can have retrospective and prospective ROC curves, i.e. a ROC curve calculated from the score assessed on the original data and a ROC curve calculated on the score assessed on future data. Let a general score calculated from logistic regression be $s = \hat{\beta}^T x$ and call the retrospective ROC curve $\hat{C}(\hat{\beta})$ with the distributions of $\overline{E}$ and E, F_0 and F_1 being estimated by their empirical analogues. Call the prospective ROC curve $C(\widehat{\beta})$ (notice the absence of the hat on C denoting this is the ROC curve for the true population distributions F_0 and F_1). In Chapter 3 we discuss the *overestimation* of the ROC for logistic regression as the difference

$$\hat{C}(\hat{\beta}) - C(\widehat{\beta}), \tag{2.13}$$

Typically (2.13) is positive, i.e the retrospective ROC gives an inflated assessment of the true performance of the score. In Chapter 3 we will derive a closed form expression for the expectation of (2.13). It is natural to extend these ideas to the overestimation of the area under the ROC curve by a retrospective fit. In Chapter 3 we shall also discuss and suggest corrections for this overestimation in the area.

We can view this overestimation in the sense of shrinkage discussed above. Copas's (1983) shrinkage estimator is designed to give a well calibrated predictor.

So far we have not discussed shrinkage in terms of discrimination - essentially this is what we are doing here - the corrected ROC curve, under certain assumptions gives an unbiased estimate of the discrimination of the score s, if it were to be validated on a large future sample of patients.

Retrospective and prospective ROC curves and areas are again utilized in Chapters 5 and 6, but this time in a different context. We use prior subjective knowledge of relationships between 'success' rates in categorical data with a binary response to produce 'shrinkage' estimators of the logistic transform of the 'success' probability. We then use the retrospective and prospective ROC areas to assess the discriminatory performance of the 'shrunk' estimates and model estimates to assess whether the 'shrinkage' procedure has produced any gain in discriminatory power over the model.

Chapter 3

Overestimation of the ROC Curve and Area for Logistic Regression

In Chapter 2 we defined the retrospective and prospective ROC curves, $\hat{C}(\hat{\beta})$ and $C(\hat{\beta})$. We will set up the model and notation for $C(\hat{\beta})$ and $\hat{C}(\hat{\beta})$ and derive expressions for calculating the overestimation in the ROC curve and area under the curve. Finally, some approximations are given that enable us to present the calculation of the overestimation of the area under the curve in a simpler and more transparent form.

3.1 Retrospective and prospective ROC

Suppose we have a total sample size of n individuals with data (x_i, y_i), where y denotes the group membership as before. There are m components in the covariate vector x, including the intercept term $x_1 \equiv 1$. We assume throughout that the data are modelled by the logistic regression

$$\mathrm{P}(y = 1|x) = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}} \ . \tag{3.1}$$

Let $\hat{\beta}$ be the maximum likelihood estimate of β. Then if we apply the scores $\hat{u}_i = \hat{\beta}^T x_i$ to the data against a threshold u, the observed proportions of false and true positives will be respectively

$$\hat{F}_0(u) = \frac{1}{n(1-\bar{y})} \sum_i H(\hat{u}_i - u)(1 - y_i) \tag{3.2}$$

$$\hat{F}_1(u) = \frac{1}{n\bar{y}} \sum_i H(\hat{u}_i - u) y_i, \tag{3.3}$$

where $\bar{y} = \sum y_i/n$ and H is the Heaviside function, $H(z)$ equals 1 if $z \geq 0$ and 0 if $z < 0$. The ROC curve $\hat{C} = \hat{C}(\hat{\beta})$ is the graph of (3.3) against (3.2).

For the prospective ROC, suppose the random y's in these data are replicated a large number of times. Then at each x_i, we will expect a proportion p_i of replicated cases to have $y = 1$, where

$$p_i = \frac{e^{u_i}}{1 + e^{u_i}}, \quad u_i = \beta^T x_i.$$

If the scores $\hat{u}_i$ are applied to the replicated data against a threshold v, the future proportions of false and true positives are

$$F_0(v) = \frac{1}{n(1-\bar{p})} \sum_i H(\hat{u}_i - v)(1-p_i) \tag{3.4}$$

$$F_1(v) = \frac{1}{n\bar{p}} \sum_i H(\hat{u}_i - v)p_i, \tag{3.5}$$

where $\bar{p} = \sum p_i/n$. Then $\mathcal{C} = \mathcal{C}(\widehat{\beta})$ is the graph of (3.5) against (3.4).

When a screening method is used in practice, its performance will be affected not only by random variation in the y's but also by the distribution of the values of x amongst the cases to which it is applied. Since we cannot hope to find a method of assessment which will work well under all possible uncontrolled changes in the population, a minimalist approach is to find a method which works when relevant properties of the target population match as closely as possible the corresponding properties observed in the data. The above development does this by taking the distribution of the future xs to be the empirical distribution of the sample covariates $x_1, \cdots, x_n$.

To simplify the notation, let $\epsilon_i = y_i - p_i$, $H_i = H(\hat{u}_i - u)$ and $H_i^* = H(\hat{u}_i - v)$. Then

$$\bar{y} = \bar{p} + \bar{\epsilon},$$

and, from (3.2) and (3.4),

$$
\begin{aligned}
n(\hat{F}_0(u) - F_0(v)) &= \frac{1}{n(1-\bar{y})}\sum_i H_i(1-y_i) - \frac{1}{n(1-\bar{p})}\sum_i H_i^*(1-p_i) \\
&= \frac{1}{(1-\bar{p}-\bar{\epsilon})}\sum_i H_i(1-p_i-\epsilon_i) - \frac{1}{(1-\bar{p})}\sum_i H_i^*(1-p_i) \\
&= \frac{1}{(1-\bar{p}-\bar{\epsilon})}\sum_i H_i(1-p_i-\epsilon_i) - \frac{1}{(1-\bar{p})}\sum_i H_i(1-p_i) \\
&+ \frac{1}{(1-\bar{p})}\sum_i H_i(1-p_i) - \frac{1}{(1-\bar{p})}\sum_i H_i^*(1-p_i) \\
&= \sum_i H_i\left(\frac{1-p_i-\epsilon_i}{1-\bar{p}-\bar{\epsilon}} - \frac{1-p_i}{1-\bar{p}}\right) - \frac{1}{1-\bar{p}}\sum_i (H_i^* - H_i)(1-p_i)
\end{aligned}
\tag{3.6}
$$

For a given threshold u, we want to choose v to make this zero, *i.e.* to match the false positive rates. As the first term in (3.6) tends to zero as n increases, the second term must become small also, and so v will converge to u. But the only non-zero terms in the second summation are for those values of i for which $\hat{u}_i$ is sandwiched between u and v, and hence the corresponding values of p_i will, for large n, be close to p_u, the true value of $\mathrm{P}(y=1|\beta^T x = u)$. Hence

$$
n(\hat{F}_0(u) - F_0(v)) \simeq \sum H_i\left(\frac{1-p_i-\epsilon_i}{1-\bar{p}-\bar{\epsilon}} - \frac{1-p_i}{1-\bar{p}}\right) - \frac{1-p_u}{1-\bar{p}}\sum (H_i^* - H_i). \tag{3.7}
$$

Using the same method as above

$$
n(\hat{F}_1(u) - F_1(v)) \simeq \sum H_i\left(\frac{p_i+\epsilon_i}{\bar{p}+\bar{\epsilon}} - \frac{p_i}{\bar{p}}\right) - \frac{p_u}{\bar{p}}\sum (H_i^* - H_i). \tag{3.8}
$$

Defining v_u to be the value of v which makes (3.7) equal to zero, we have

$$
n(\hat{F}_0(u) - F_0(v_u)) = \sum_i H_i\left(\frac{1-p_i-\epsilon_i}{1-\bar{p}-\bar{\epsilon}} - \frac{1-p_i}{1-\bar{p}}\right) - \frac{1-p_u}{1-\bar{p}}\sum_i (H_i^* - H_i) = 0
$$

and so

$$\frac{1-p_u}{1-\bar{p}}\sum_i(H_i^*-H_i)=\sum_i H_i\left(\frac{1-p_i-\epsilon_i}{1-\bar{p}-\bar{\epsilon}}-\frac{1-p_i}{1-\bar{p}}\right),$$

or

$$\frac{1-p_u}{1-\bar{p}}\sum_i H_i^*=\left\{\frac{1-p_u}{1-\bar{p}}+\left(\frac{1-p_i-\epsilon_i}{1-\bar{p}-\bar{\epsilon}}-\frac{1-p_i}{1-\bar{p}}\right)\right\}\sum_i H_i.$$

Thus

$$\sum_i H_i^*=\frac{1-\bar{p}}{1-p_u}\left\{\frac{1-p_u}{1-\bar{p}}+\left(\frac{1-p_i-\epsilon_i}{1-\bar{p}-\bar{\epsilon}}-\frac{1-p_i}{1-\bar{p}}\right)\right\}\sum_i H_i \qquad (3.9)$$

and so, substituting for $\sum H_i^*$ in (3.8),

$$\begin{aligned}\hat{F}_1(u)-F_1(v_u) &\simeq \frac{1}{n}\sum_i H_i\left(\frac{p_i+\epsilon_i}{\bar{p}+\bar{\epsilon}}-\frac{p_i}{\bar{p}}\right)\\ &- \frac{p_u(1-\bar{p})}{n\bar{p}(1-p_u)}\sum_i H_i\left(\frac{1-p_i-\epsilon_i}{1-\bar{p}-\bar{\epsilon}}-\frac{1-p_i}{1-\bar{p}}\right)\\ &= \frac{1}{n}\sum_i H_i A(i,u) \qquad (3.10)\end{aligned}$$

where

$$A(i,u)=\left(\frac{p_i+\epsilon_i}{\bar{p}+\bar{\epsilon}}-\frac{p_i}{\bar{p}}\right)-\frac{(1-\bar{p})p_u}{\bar{p}(1-p_u)}\left(\frac{1-p_i-\epsilon_i}{1-\bar{p}-\bar{\epsilon}}-\frac{1-p_i}{1-\bar{p}}\right). \qquad (3.11)$$

3.2 Expected overestimation of the ROC curve

We want to find the asymptotic expectation of (3.10), the distance between the true positive rates of the retrospective and prospective ROC curves for the cutoff u. We note that $\bar{\epsilon}=O_p(n^{-\frac{1}{2}})$ so that the fractions in (3.11) can be expanded into

a power series in $\bar{\epsilon}$. Firstly

$$\begin{aligned}\frac{1}{\bar{p}+\bar{\epsilon}} &= (\bar{p}+\bar{\epsilon})^{-1} \\ &= \frac{1}{\bar{p}}\left(1+\frac{\bar{\epsilon}}{\bar{p}}\right)^{-1} \\ &= \frac{1}{\bar{p}}\left(1-\frac{\bar{\epsilon}}{\bar{p}}+\cdots\right) \qquad (3.12)\end{aligned}$$

ignoring terms in $\bar{\epsilon}^2$ and higher. So

$$\begin{aligned}(p_i+\epsilon_i)(\bar{p}+\bar{\epsilon})^{-1} &= \frac{1}{\bar{p}}(p_i+\epsilon_i)\left(1-\frac{\bar{\epsilon}}{\bar{p}}\right) \\ &= \frac{1}{\bar{p}}\left(p_i+\epsilon_i+\frac{p_i\bar{\epsilon}}{\bar{p}}\right) \\ &= \frac{p_i}{\bar{p}}+\frac{\epsilon_i}{\bar{p}}+\frac{p_i\epsilon}{\bar{p}^2} \qquad (3.13)\end{aligned}$$

ignoring the term in $\bar{\epsilon}\epsilon_i$. Similarly

$$\begin{aligned}\frac{1}{1-\bar{p}-\bar{\epsilon}} &= (1-\bar{p}-\bar{\epsilon})^{-1} \\ &= \frac{1}{1-\bar{p}}\left(1-\frac{\bar{\epsilon}}{1-\bar{p}}\right)^{-1} \\ &= \frac{1}{\bar{p}}\left(1+\frac{\bar{\epsilon}}{1-\bar{p}}+\cdots\right) \qquad (3.14)\end{aligned}$$

(ignoring terms in $\bar{\epsilon}^2$ and higher) and

$$\begin{aligned}(1-p_i-\epsilon_i)(1-\bar{p}-\bar{\epsilon})^{-1} &= \frac{1}{1-\bar{p}}(1-p_i-\epsilon_i)\left(1+\frac{\bar{\epsilon}}{1-\bar{p}}\right) \\ &\simeq \frac{1}{1-\bar{p}}\left(1-p_i-\epsilon_i+\frac{(1-p_i)\bar{\epsilon}}{1-\bar{p}}\right) \\ &= \frac{1-p_i}{1-\bar{p}}-\frac{\epsilon_i}{1-\bar{p}}+\frac{(1-p_i)\bar{\epsilon}}{(1-\bar{p})^2} \qquad (3.15)\end{aligned}$$

again ignoring higher order terms. Now substituting (3.13) and (3.15) into (3.11) gives

$$\begin{aligned} A(i,u) &\simeq \left(\frac{\epsilon_i}{\bar{p}} + \frac{p_i\bar{\epsilon}}{\bar{p}^2}\right) - \frac{(1-\bar{p})p_u}{\bar{p}(1-p_u)}\left(\frac{\epsilon_i}{1-\bar{p}} + \frac{(1-p_i)\bar{\epsilon}}{(1-\bar{p})^2}\right) \\ &= \left(\frac{1}{\bar{p}} - \frac{p_u}{\bar{p}(1-p_u)}\right)\epsilon_i + \left(\frac{p_i}{\bar{p}} - \frac{p_u(1-p_i)}{\bar{p}(1-p_u)(1-\bar{p})}\right)\bar{\epsilon} \\ &= \left(\frac{1}{\bar{p}(1-p_u)}\right)\epsilon_i - \left(\frac{(1-p_u)p_i(1-\bar{p}) + p_u(1-p_i)\bar{p}}{(1-p_u)\bar{p}^2(1-\bar{p})}\right)\bar{\epsilon} \end{aligned} \quad (3.16)$$

We now give a lemma which gives closed form expressions for the asymptotic expectation of terms involving ϵ_i and $\bar{\epsilon}$. Let

$$\Omega = \frac{1}{n}\sum_i p_i(1-p_i)x_i x_i^T \quad (3.17)$$

and

$$d_i^2 = x_i^T \Omega^{-1} x_i.$$

Lemma 1 *To an accuracy of $O(n^{-\frac{3}{2}})$, the asymptotic expectations of $H_i\epsilon_i$ and $H_i\bar{\epsilon}$ are given by*

$$E(H_i\epsilon_i) = n^{-\frac{1}{2}} p_i(1-p_i)d_i\phi\left(\frac{n^{\frac{1}{2}}(u_i - u)}{d_i}\right) \quad (3.18)$$

and

$$E(H_i\bar{\epsilon}) = n^{-\frac{1}{2}} d_i^{-1}\phi\left(\frac{n^{\frac{1}{2}}(u_i - u)}{d_i}\right) \quad (3.19)$$

where ϕ is the standard normal density function.

Proof: From the maximum likelihood equations for logistic regression we have

$$\begin{aligned}\hat{\beta} - \beta &\simeq (n\Omega)^{-1} \sum_j (y_j - p_j) x_j \\ &\simeq (n\Omega)^{-1} \sum_j \epsilon_j x_j\end{aligned}$$

and so the scalar product of this with x_i gives

$$\hat{u}_i - u_i \simeq \frac{1}{n} \sum_j \epsilon_j \gamma_{ij} \tag{3.20}$$

where $\gamma_{ij} = x_i^T \Omega^{-1} x_j$ and so $\gamma_{ii} = d_i^2$. Now

$$\begin{aligned}\text{Var}(\hat{u}_i) &\simeq \text{Var}\left(u_i + \frac{1}{n} \sum_j \epsilon_j \gamma_{ij}\right) \\ &\simeq \frac{1}{n^2} \sum_j \text{Var}(\epsilon_j) \gamma_{ij}^2\end{aligned}$$

Now $\epsilon_j = y_j - p_j$ so $\text{Var}(\epsilon_j) = \text{Var}(y_j - p_j) = p_j(1 - p_j)$.

So

$$\begin{aligned}\text{Var}(\hat{u}_i) &\simeq \frac{1}{n^2} \sum_j p_j(1 - p_j) \gamma_{ij}^2 \\ &\simeq \frac{1}{n^2} \sum_j p_j(1 - p_j) x_i^T \Omega^{-1} x_j x_j^T \Omega^{-1} x_i \\ &\simeq \frac{1}{n^2} x_i^T \Omega^{-1} \left[\sum_j p_j(1 - p_j) x_j x_j^T\right] \Omega^{-1} x_i \\ &\simeq \frac{1}{n^2} x_i^T \Omega^{-1} [n\Omega] \Omega^{-1} x_i \\ &\simeq \frac{1}{n} x_i^T \Omega^{-1} x_i \\ &\simeq \frac{d_i^2}{n}\end{aligned}$$

and

$$\begin{aligned} \mathrm{E}(\hat{u}_i) &\simeq \mathrm{E}\left(u_i + \frac{1}{n}\sum_j \epsilon_j \gamma_{ij}\right) \\ &\simeq u_i + \frac{\epsilon_i d_i^2}{n} \end{aligned}$$

Now condition on ϵ_i to give approximately

$$\hat{u}_i|\epsilon_i \sim N\left(u_i + \frac{\epsilon_i d_i^2}{n}, \frac{d_i^2}{n}\right).$$

Thus, if Φ is the standard normal distribution function,

$$\begin{aligned} \mathrm{E}\left(\epsilon_i H(\hat{u}_i - u)\right) &= \mathrm{E}\left\{\epsilon_i \mathrm{P}(\hat{u}_i \geq u|\epsilon_i)\right\} \\ &\simeq \mathrm{E}\left\{\epsilon_i \Phi\left(\frac{n^{\frac{1}{2}}(u_i - u)}{d_i} + \frac{\epsilon_i d_i}{n^{\frac{1}{2}}}\right)\right\} \end{aligned}$$

Using a Taylor expansion

$$\Phi\left(\frac{n^{\frac{1}{2}}(u_i - u)}{d_i} + \frac{\epsilon_i d_i}{n^{\frac{1}{2}}}\right) = \Phi\left(\frac{n^{\frac{1}{2}}(u_i - u)}{d_i}\right) + \frac{\epsilon_i d_i}{n^{\frac{1}{2}}}\phi\left(\frac{n^{\frac{1}{2}}(u_i - u)}{d_i}\right)$$

and substituting into the above expression for the expectation

$$\begin{aligned} \mathrm{E}\left(\epsilon_i H(\hat{u}_i - u)\right) &\simeq \mathrm{E}\left\{\epsilon_i \Phi\left(\frac{n^{\frac{1}{2}}(u_i - u)}{d_i}\right) + \frac{\epsilon_i^2 d_i}{n^{\frac{1}{2}}}\phi\left(\frac{n^{\frac{1}{2}}(u_i - u)}{d_i}\right)\right\} \\ &\simeq \mathrm{E}\left\{\frac{\epsilon_i^2 d_i^2}{n^{\frac{1}{2}}}\phi\left(\frac{n^{\frac{1}{2}}(u_i - u)}{d_i}\right)\right\} \\ &\simeq \frac{p_i(1-p_i)d_i}{n^{\frac{1}{2}}}\phi\left(\frac{n^{\frac{1}{2}}(u_i - u)}{d_i}\right) \end{aligned}$$

which equals expression (3.18) as expectations of linear terms involving ϵ_i are 0 and $\mathrm{E}(\epsilon_i^2) = \mathrm{Var}(y_i) = p_i(1-p_i)$.

To find $\mathrm{E}(\bar{\epsilon}H_i)$ we condition on ϵ_j instead of ϵ_i, and add over j, i.e.

$$\mathrm{E}(\bar{\epsilon}H_i) = \frac{1}{n}\sum_j \mathrm{E}(\epsilon_j H_i)$$

Now

$$\hat{u}_i|\epsilon_j \sim N\left(u_i + \frac{\epsilon_j\gamma_{ij}}{n}, \frac{d_i^2}{n}\right)$$

and using the same method above with the Taylor expansion

$$\begin{aligned}
\mathrm{E}\left(\epsilon_j H(\hat{u}_i - u)\right) &= \mathrm{E}\left\{\epsilon_j \mathrm{P}(\hat{u}_i \geq u|\epsilon_j)\right\} \\
&\simeq \mathrm{E}\left\{\epsilon_j \Phi\left(\frac{n^{\frac{1}{2}}(u_i - u)}{d_i} + \frac{\epsilon_j\gamma_{ij}}{n^{\frac{1}{2}}d_i}\right)\right\} \\
&\simeq \mathrm{E}\left\{\frac{\epsilon_j^2\gamma_{ij}}{n^{\frac{1}{2}}d_i}\phi\left(\frac{n^{\frac{1}{2}}(u_i - u)}{d_i}\right)\right\} \\
&\simeq \frac{p_j(1-p_j)\gamma_{ij}}{n^{\frac{1}{2}}d_i}\,\phi\left(\frac{n^{\frac{1}{2}}(u_i - u)}{d_i}\right)
\end{aligned}$$

So

$$\mathrm{E}(\bar{\epsilon}H_i) = \frac{1}{n}\sum_j \frac{p_j(1-p_j)\gamma_{ij}}{n^{\frac{1}{2}}d_i}\,\phi\left(\frac{n^{\frac{1}{2}}(u_i - u)}{d_i}\right)$$

But

$$\sum_j p_j(1-p_j)\gamma_{ij} = n\left[\frac{1}{n}\sum_j p_j(1-p_j)x_j\right]\Omega^{-1}x_i = n$$

as the term in square brackets in the centre expression is just the first row of Ω, and the first element of x_i is 1. Therefore

$$\mathrm{E}(\bar{\epsilon}H_i) = n^{-\frac{1}{2}}d_i^{-1}\,\phi\left(\frac{n^{\frac{1}{2}}(u_i - u)}{d_i}\right)$$

which equals (3.19).

Following the same method as in the proof of the lemma, it can also be shown that the expectations of the terms left out in (3.16), namely $\mathrm{E}(H_i\epsilon_i\bar{\epsilon})$, $\mathrm{E}(H_i\bar{\epsilon}^2)$, and so on, are all $O(n^{-\frac{3}{2}})$ or less. So substituting (3.18) and (3.19) in (3.10) and consequently into (3.16), we find

$$\begin{aligned}
S(u) &= \mathrm{E}(\hat{F}_1(u) - F_1(v_u)) \\
&\simeq \frac{1}{n}\sum_i \mathrm{E}\left(H_i A(i,u)\right) \\
&\simeq \frac{1}{n}\sum_i \left\{ \frac{1}{\bar{p}(1-p_u)}\mathrm{E}(H_i\epsilon_i) - \left(\frac{(1-p_u)p_i(1-\bar{p}) + p_u(1-p_i)\bar{p}}{(1-p_u)\bar{p}^2(1-\bar{p})} \right) \mathrm{E}(\bar{\epsilon}H_i) \right\} \\
&\simeq \frac{1}{n^{\frac{3}{2}}\bar{p}}\sum_i \left[\left\{ \frac{p_i(1-p_i)}{1-p_u}\, d_i - \left(\tfrac{(1-p_u)p_i(1-\bar{p}) + p_u(1-p_i)\bar{p}}{(1-p_u)\bar{p}(1-\bar{p})} \right) \frac{1}{d_i} \right\} \phi\left(\frac{n^{\frac{1}{2}}(u_i - u)}{d_i} \right) \right] \\
&\simeq \frac{p_u}{n^{\frac{3}{2}}\bar{p}}\sum_i \left(d_i - \frac{1}{\bar{p}(1-\bar{p})d_i} \right) \phi\left(\frac{n^{\frac{1}{2}}(u_i - u)}{d_i} \right) \qquad (3.21)
\end{aligned}$$

The last step in the above set of equations may not immediately follow for the reader but it is accomplished by substituting p_u for p_i. The reasoning behind this is that the argument of ϕ in the above equations only gives weight to scores close to u. Therefore we can approximate p_i by p_u and the last equation follows by cancellation of terms. The terms p_u and u_i in the right hand side of (3.21) are functions of the true parameter vector β, and the terms d_i and $\bar{p}$ depend on p_i and hence are also functions of β. Estimating these in the obvious way by calculating a logistic regression to obtain $\hat{\beta}$ gives the corresponding estimate $\hat{S}(u)$. The corrected ROC curve is the plot of the false positive rates against the

true positive rate minus the correction i.e.

$$C^* = \{\hat{F}_1(u) - \hat{S}(u), \hat{F}_0(u)\}.$$

C^* is an estimate of the ROC curve that would be obtained if the fitted score $\hat{\beta}^T x$ were to be validated on a large replicated sample. The sum in (3.21) can be written as an expectation over the empirical distribution of the sample covariate vectors $x_1, \cdots, x_n$. Denoting such empirical expectations by $\bar{\mathrm{E}}(\cdot)$, and defining $U = \beta^T X$ and $D^2 = X^T \Omega^{-1} X$, (3.21) is

$$\frac{p_u}{n^{\frac{1}{2}} \bar{p}} \bar{\mathrm{E}} \left\{ \left(D^2 - \frac{1}{\bar{p}(1-\bar{p})} \right) \frac{1}{D} \phi \left(\frac{n^{\frac{1}{2}}(U-u)}{D} \right) \right\} \tag{3.22}$$

The factor $n^{\frac{1}{2}}$ in the argument of the standard normal density in (3.22) means that, as $n \to \infty$, the expectation puts increasing weight on values of X with U close to u, and so essentially depends on conditional expectations of the relevant quantities given $U = u$. We now introduce a lemma which enables us to investigate this relationship further.

Lemma 2 *Let (X, Y) be two continuous random variables with $Y > 0$. Let $f(x)$ be the marginal probability density function of X, $g(Y)$ be any smooth function of Y, and a be a constant. Then under smooth regularity conditions*

$$E \left\{ \frac{g(Y)}{Y} \phi \left(\frac{n^{\frac{1}{2}}(X-a)}{Y} \right) \right\} = n^{-\frac{1}{2}} E(g(Y)|X=a)\, f(a) + O(n^{-\frac{3}{2}}).$$

Proof: Let $f(x,y)$ be the joint probability density function of (X,Y), so

$$\mathrm{E}\left\{\frac{g(Y)}{Y}\phi\left(\frac{n^{\frac{1}{2}}(X-a)}{Y}\right)\right\} = \int\int y^{-1}g(y)\,\phi\left(\frac{n^{\frac{1}{2}}(x-a)}{y}\right) f(x,y)\,dx\,dy$$

and let

$$t = \frac{n^{\frac{1}{2}}(x-a)}{y} \quad \text{so that} \quad x = a + \frac{ty}{n^{\frac{1}{2}}} \quad \text{and} \quad dx = \frac{y\,dt}{n^{\frac{1}{2}}}.$$

Then substituting for x the required expectation can be written

$$n^{-\frac{1}{2}}\int\int g(y)\phi(t)f(a+\frac{ty}{n^{\frac{1}{2}}},y)\,dt\,dy$$

Using a Taylor Expansion

$$f(a+\frac{ty}{n^{\frac{1}{2}}},y) \simeq f(a,y) + \frac{ty}{n^{\frac{1}{2}}}\frac{\delta}{\delta x}f(a,y) + \frac{t^2y^2}{n}\frac{\delta^2}{\delta x^2}f(a,y) + \cdots$$

The required expectation is now

$$\int\int g(y)f(a,y)\phi(t)\,dt\,dy + n^{-\frac{1}{2}}\int\int tyg(y)\phi(t)\frac{\delta}{\delta x}f(a,y) + O(n^{-\frac{3}{2}})$$

$$= \int\int g(y)f(a,y)\phi(t)\,dt\,dy + O(n^{-\frac{3}{2}})$$

as $\int t\phi(t) = \mathrm{E}(t) = 0$.

If we assume that the empirical distribution of X converges to a continuous distribution in which $U = \beta^T X$ has probability density function $f(u)$, a direct application of Lemma 2 to (3.22) gives

$$\frac{p_u f(u)}{n\bar{p}}\left(\bar{\mathrm{E}}(D^2|U=u) - \frac{1}{\bar{p}(1-\bar{p})}\right). \tag{3.23}$$

This expression shows that, asymptotically, the overestimation of ROC at each threshold u is of the order $O(n^{-1})$. Values of D^2, however, depend on the dimension m, and tend to be large if m is large. We derive a further approximation in Section 3.4, suggesting that if m is large as well as n, then the expected overestimation is proportional to the ratio m/n.

3.3 Overestimation of the area under the ROC curve

The area under the retrospective ROC curve $\hat{C}$ is

$$\int \hat{F}_1(u)d\hat{F}_0(u) = \frac{1}{n(1-\bar{y})}\sum_{j=1}^{n}(1-y_j)\hat{F}_1(\hat{u}_j)$$

and the area under the prospective ROC curve C is

$$\int \hat{F}_1(v)dF_0(v) = \frac{1}{n(1-\bar{p})}\sum_{j=1}^{n}(1-p_j)\hat{F}_1(v_j).$$

Using the same method presented in Sections 3.1 and 3.2 we calculate the overestimation in the area by integrating over the differences in true positive rates with respect to the false positive rates i.e. the overestimation is

$$\int(\hat{F}_1(u) - F_1(v_u))d\hat{F}_0(u)$$

where as before, v_u is the value of v which matches the false positive rates.

From (3.10) we have

$$\hat{F}_1(u) - F_1(v_u) \simeq \frac{1}{n}\sum_i H_i A(i,u).$$

Substituting in above we have

$$\begin{aligned}\int (\hat{F}_1(u) - F_1(v_u))d\hat{F}_0(u) &= \frac{1}{n(1-\bar{y})}\sum_{j=1}^{n}(1-y_j)\left(\hat{F}_1(\hat{u}_j) - F_1(v_{\hat{u}_j})\right) \\ &\simeq \frac{1}{n^2(1-\bar{y})}\sum_{i,j} H(\hat{u}_i - \hat{u}_j)(1-y_j)A(i,\hat{u}_j)\end{aligned} \tag{3.24}$$

where $v_{\hat{u}_j}$ is the value of v that corresponds to the jth value of $\hat{u}$ which matches the false positive rates.

As before, we approximate the expectation of (3.24) by expanding in terms of the ϵ_j's, noting that the variances of $\bar{\epsilon}$ and $\hat{u}_j$ (the only random terms in the denominators) are both of the order $O(n^{-1})$. In $A(i,\hat{u}_j), p_{\hat{u}_j}$ is just the value p_j so

$$A(i,\hat{u}_j) = \left(\frac{1}{\bar{p}(1-p_j)}\right)\epsilon_i - \left(\frac{(1-p_j)p_i(1-\bar{p}) + p_j(1-p_i)\bar{p}}{(1-p_j)\bar{p}^2(1-\bar{p})}\right)\bar{\epsilon}$$

ignoring higher order terms.

We let $y_j = p_j + \epsilon_j$ and expand power series such that

$$
\begin{aligned}
\frac{1-y_j}{1-\bar{y}}A(i,\hat{u}_j) &= \frac{1-p_j-\epsilon_j}{1-\bar{p}-\bar{\epsilon}}A(i,\hat{u}_j) \\
&= (1-p_j-\epsilon_j)(1-\bar{p}-\bar{\epsilon})^{-1}A(i,\hat{u}_j) \\
&= \frac{1}{(1-\bar{p})}(1-p_j-\epsilon_j)\left(1-\frac{\bar{\epsilon}}{(1-\bar{p})}\right)^{-1}A(i,\hat{u}_j) \\
&= \frac{1}{(1-\bar{p})}(1-p_j-\epsilon_j)\left(1+\frac{\bar{\epsilon}}{(1-\bar{p})}+\cdots\right)A(i,\hat{u}_j) \\
&= \left(\frac{1-p_j}{1-\bar{p}}-\frac{\epsilon_j}{1-\bar{p}}+\frac{(1-p_j)\bar{\epsilon}}{(1-\bar{p})^2}+\cdots\right)A(i,\hat{u}_j) \\
&= \frac{(1-p_j)}{\bar{p}(1-\bar{p})(1-p_j)}\epsilon_i+\cdots \\
&= \frac{1}{\bar{p}(1-\bar{p})}\epsilon_i+\cdots \qquad (3.25)
\end{aligned}
$$

plus terms in $\epsilon_i\epsilon_j$, $\bar{\epsilon}$, $\epsilon_i\bar{\epsilon}$, and so on. We now introduce another Lemma, very similar in fact to Lemma 1 which gives an expression for the expectation involving the term ϵ_i in the overestimation of the area.

Lemma 3 *When $i \neq j$ the asymptotic expectation of $\epsilon_i H(\hat{u}_j - \hat{u}_i)$ is given by*

$$E(\epsilon_i H(\hat{u}_j - \hat{u}_i)) \simeq \frac{p_i(1-p_i)a_{ij}}{n^{\frac{1}{2}}b_{ij}}\phi\left(\frac{n^{\frac{1}{2}}(u_j-u_i)}{b_{ij}}\right), \qquad (3.26)$$

where

$$a_{ij} = (x_i - x_j)^T\Omega^{-1}x_i$$

and

$$b_{ij}^2 = a_{ij} + a_{ji} = (x_i - x_j)^T\Omega^{-1}(x_i - x_j).$$

Proof: Again from (3.20), the equations for maximum likelihood estimation in logistic regression we have

$$\begin{aligned}\hat{\beta}_j - \beta &\simeq \frac{1}{n}\sum_k \epsilon_k \gamma_{jk} \\ \hat{\beta}_i - \beta &\simeq \frac{1}{n}\sum_k \epsilon_k \gamma_{ik}\end{aligned}$$

so taking the scalar product with x_j and x_i respectively and subtracting we have

$$\hat{u}_j - \hat{u}_i \simeq u_j - u_i + \frac{1}{n}\sum_k \epsilon_k(\gamma_{jk} - \gamma_{ik}).$$

Now

$$\begin{aligned}\text{Var}\,(\hat{u}_j - \hat{u}_i) &\simeq \frac{1}{n^2}\sum_k V(\epsilon_k)(\gamma_{jk} - \gamma_{ik})^2 \\ &\simeq \frac{1}{n^2}\sum_k p_k(1-p_k)(\gamma_{jk} - \gamma_{ik})^2 \\ &\simeq \frac{1}{n^2}\sum_k p_k(1-p_k)(\gamma_{jk}^2 - \gamma_{ik}^2 - 2\gamma_{jk}\gamma_{ik})\end{aligned}$$

Also

$$\begin{aligned}\sum_k p_k(1-p_k)\gamma_{jk}^2 &= \sum_k p_k(1-p_k) x_j^T \Omega^{-1} x_k x_k^T \Omega^{-1} x_j \\ &= x_j^T \Omega^{-1} \left[\sum_k p_k(1-p_k) x_k x_k^T\right] \Omega^{-1} x_j \\ &= x_j \Omega^{-1} [n\Omega] \Omega^{-1} x_j \\ &= n x_j^T \Omega^{-1} x_j \\ &= n\gamma_{jj}\end{aligned}$$

Subsequently

$$\begin{aligned}\sum_k p_k(1-p_k)\gamma_{ik}^2 &= \sum_k p_k(1-p_k)x_i^T\Omega^{-1}x_k x_k^T\Omega^{-1}x_i \\ &= nx_i^T\Omega^{-1}x_i \\ &= n\gamma_{ii}\end{aligned}$$

and

$$\begin{aligned}\sum_k p_k(1-p_k)\gamma_{jk}\gamma_{ik} &= \sum_k p_k(1-p_k)x_j^T\Omega^{-1}x_k x_k^T\Omega^{-1}x_i \\ &= x_j^T\Omega^{-1}\left[\sum_k p_k(1-p_k)x_k x_k^T\right]\Omega^{-1}x_i \\ &= x_j\Omega^{-1}\left[n\Omega\right]\Omega^{-1}x_i \\ &= nx_j^T\Omega^{-1}x_i \\ &= n\gamma_{ji}\end{aligned}$$

So substituting these terms into the expression for the variance above gives

$$V(\hat{u}_j - \hat{u}_i) = \frac{1}{n}(\gamma_{jj} + \gamma_{ii} - 2\gamma_{ji}) = \frac{b_{ij}^2}{n}$$

as

$$a_{ij} = (x_i - x_j)^T\Omega^{-1}x_i = x_i^T\Omega^{-1}x_i - x_j^T\Omega^{-1}x_i = \gamma_{ii} - \gamma_{ij}$$

so

$$b_{ij}^2 = a_{ji} + a_{ij} = \gamma_{ii} + \gamma_{jj} - 2\gamma_{ij}.$$

Also

$$\mathrm{E}(\hat{u}_j - \hat{u}_i) = u_j - u_i + \frac{\epsilon_i(\gamma_{ii} - \gamma_{ij})}{n} + \frac{\epsilon_j(\gamma_{jj} - \gamma_{ji})}{n}$$

so conditioning on ϵ_i to give approximately

$$\hat{u}_j - \hat{u}_i | \epsilon_i \sim N\left(u_j - u_i + \frac{\epsilon_i(\gamma_{ii} - \gamma_{ij})}{n} + \frac{\epsilon_j(\gamma_{jj} - \gamma_{ji})}{n}, \frac{b_{ij}^2}{n}\right)$$

From now on we should strictly include all terms in ϵ_j, but this will give rise to expectations involving the product $\epsilon_i\epsilon_j$ which will be zero, so we exclude these for the sake of brevity. Then

$$\begin{aligned} \mathrm{E}\left(\epsilon_i H(\hat{u}_j - \hat{u}_i)\right) &= \mathrm{E}\left\{\epsilon_i \mathrm{P}(\hat{u}_j - \hat{u}_i \geq 0 | \epsilon_i)\right\} \\ &= \mathrm{E}\left\{\epsilon_i \Phi\left(\frac{(u_j - u_i)n^{\frac{1}{2}}}{b_{ij}} + \frac{\epsilon_i a_{ij}}{n^{\frac{1}{2}} b_{ij}}\right)\right\} \end{aligned}$$

Again using a Taylor Expansion

$$\Phi\left(\frac{(u_j - u_i)n^{\frac{1}{2}}}{b_{ij}} + \frac{\epsilon_i a_{ij}}{n^{\frac{1}{2}} b_{ij}}\right) \simeq \Phi\left(\frac{(u_j - u_i)n^{\frac{1}{2}}}{b_{ij}}\right) + \frac{\epsilon_i a_{ij}}{n^{\frac{1}{2}} b_{ij}} \phi\left(\frac{(u_j - u_i)n^{\frac{1}{2}}}{b_{ij}}\right) + \cdots$$

and substituting into the above expression

$$\begin{aligned} \mathrm{E}\left(\epsilon_i H(\hat{u}_j - \hat{u}_i)\right) &\simeq \mathrm{E}\left\{\epsilon_i \Phi\left(\frac{(u_j - u_i)n^{\frac{1}{2}}}{b_{ij}}\right) + \frac{\epsilon_i^2 a_{ij}}{n^{\frac{1}{2}} b_{ij}} \phi\left(\frac{(u_j - u_i)n^{\frac{1}{2}}}{b_{ij}}\right)\right\} \\ &\simeq \mathrm{E}\left\{\frac{\epsilon_i^2 a_{ij}}{n^{\frac{1}{2}} b_{ij}} \phi\left(\frac{(u_j - u_i)n^{\frac{1}{2}}}{b_{ij}}\right)\right\} \\ &\simeq \frac{p_i(1 - p_i)a_{ij}}{n^{\frac{1}{2}} b_{ij}} \phi\left(\frac{(u_j - u_i)n^{\frac{1}{2}}}{b_{ij}}\right) \end{aligned}$$

When $i = j$ the expectation in (3.26) is zero. Extending Lemma 3 it can also be shown that the expectation of the product of $H(\hat{u}_j - \hat{u}_i)$ with each of the terms

left out in (3.25) is of a lower order of magnitude and so, for large n, can be omitted.

Using (3.26) in (3.25) and (3.24) gives

$$\int (\hat{F}_1(u) - F_1(u_v)) d\hat{F}_0(u) \simeq \frac{1}{n^{\frac{5}{2}}\bar{p}(1-\bar{p})} \sum_{i \neq j} \frac{p_j(1-p_j)a_{ij}}{b_{ij}} \phi\left(\frac{n^{\frac{1}{2}}(u_j - u_i)}{b_{ij}}\right) \tag{3.27}$$

$$\simeq \frac{1}{n^{\frac{5}{2}}\bar{p}(1-\bar{p})} \sum_{i \neq j} \frac{p_i(1-p_i)a_{ij}}{b_{ij}} \phi\left(\frac{n^{\frac{1}{2}}(u_j - u_i)}{b_{ij}}\right) \tag{3.28}$$

$$\simeq \frac{1}{2n^{\frac{5}{2}}\bar{p}(1-\bar{p})} \sum_{i \neq j} p_i(1-p_i)b_{ij} \phi\left(\frac{n^{\frac{1}{2}}(u_j - u_i)}{b_{ij}}\right) \tag{3.29}$$

The factor $n^{\frac{1}{2}}$ in the argument of ϕ means that only pairs with values of u_i and u_j close to each other are given weight in these double sums. This allows p_j in (3.27) to be replaced by p_i in (3.28) to this order of approximation. Equation (3.29) follows by taking (3.27), reversing i and j, then adding to equation (3.28), and then dividing the result by two. This makes the calculation easier as we only have to calculate b_{ij} and not a_{ij}. As before, the value of this can be estimated in the natural way by evaluating the terms b_{ij} using $\hat{\beta}$.

Now using Lemma 2, for any fixed suffix i,

$$\frac{1}{n}\sum_j b_{ij}\phi\left(\frac{n^{\frac{1}{2}}(u_j - u_i)}{b_{ij}}\right) \simeq n^{-\frac{1}{2}}\bar{\mathrm{E}}\left((X - x_i)^T\Omega^{-1}(X - x_i)|U = u_i\right) f(u_i)$$

and so the overestimation in area is approximately

$$\begin{aligned}
&\frac{1}{2n^2\bar{p}(1-\bar{p})}\sum_i p_i(1-p_i)\bar{\mathrm{E}}\left((X-x_i)^T\Omega^{-1}(X-x_i)|\beta^T X=\beta^T x_i=u_i\right)f(u_i)\\
\simeq\ &\frac{1}{2n\bar{p}(1-\bar{p})}\bar{\mathrm{E}}\left\{p_U(1-p_U)\bar{\mathrm{E}}\left((Y-X)^T\Omega^{-1}(Y-X)|\beta^T X=\beta^T Y=U\right)f(U)\right\}\\
=\ &\frac{1}{n\bar{p}(1-\bar{p})}\bar{\mathrm{E}}\left\{p_U(1-p_U)f(U)\ \mathrm{tr}\left(\Omega^{-1}\overline{\mathrm{Var}}(X|U)\right)\right\}\ , \qquad (3.30)
\end{aligned}$$

where Y is an independent replication of X. Note that the overestimation is again of the order $O(n^{-1})$.

3.4 Further approximations

Formulae (3.22) and (3.30) involve moments of the conditional distribution of covariate vector X given the score $U = \beta^T X$. We can gain a greater understanding of the magnitude of these quantities by approximating this conditional distribution by the residual distribution in a linear regression of X on U. The approximation is exact under an assumption of multivariate normality.

In Section 3.2 we defined $\bar{\mathrm{E}}$ to denote empirical expectation over the finite sample values $x_1, \cdots, x_n$. Extending this notation, let $\mathrm{E}^*(\cdot)$ denote the corresponding weighted expectation given by

$$\mathrm{E}^*(\cdot) = \frac{\sum_i p_i(1-p_i)(\cdot)}{\sum_i p_i(1-p_i)}.$$

Let

$$\mu = \mathrm{E}^*(X) \text{ and } V = \mathrm{Var}^*(X) = \mathrm{E}^*(XX^T) - \mu\mu^T.$$

Then

$$\mathrm{E}^*(\beta^T X) = \mathrm{E}^*(U) = \bar{u} = \beta^T \mu \text{ and } \mathrm{Var}^*(U) = \sigma_u^2 = \beta^T V \beta$$

Now consider the following linear regression of vector X on the scalar U,

$$X = \mu + a(U - \bar{u}) + R \tag{3.31}$$

where R is the corresponding vector of residuals. Then if this regression is fitted by least squares under the expectation operator E^*, the normal equations are

$$\mathrm{E}^*(R) = \mathrm{E}^*\left(R(U - \bar{u})\right) = 0.$$

Hence the regression vector is

$$a = \frac{V\beta}{\beta^T V \beta}$$

and the residual variance is

$$\mathrm{Var}^*(R) = V - \frac{V\beta\beta^T V}{\beta^T V \beta}.$$

Had the (weighted) distribution of X been multivariate normal, the right hand side of (3.31) would give a complete description of the conditional distribution of X given U. Also note that as p_u is a function of u, conditional expectations given U under E^* are exactly the same as conditional expectations under $\bar{\mathrm{E}}$. Hence we would then have

$$\bar{\mathrm{E}}(X|U = u) = \mathrm{E}^*(X|U = u) = \mu + \frac{V\beta}{\beta^T V \beta}(u - \bar{u}) \tag{3.32}$$

and

$$\overline{\mathrm{Var}}(X|U=u) = \mathrm{Var}^*(X|U=u) = V - \frac{V\beta\beta^T V}{\beta^T V \beta}. \tag{3.33}$$

Of course X is not actually multivariate normal, and so we can only use (3.32) and (3.33) as approximations to the first two conditional moments of X given U.

From (3.17) we have

$$\Omega = \frac{1}{n}\sum_i p_i(1-p_i)x_i x_i^T = \frac{\sum p_i(1-p_i)}{n}(V + \mu\mu^T).$$

Also, as the first element of vector x_i is always 1, the first element of μ is also 1, all elements in the first row and the first column of V are zero, and the first row or column of Ω is just $(\sum p_i(1-p_i)/n)$ times the vector μ. It follows that

$$\mu^T\Omega^{-1}\mu = \frac{n}{\sum p_i(1-p_i)}$$

and so

$$\mathrm{tr}(\Omega^{-1}V) = \frac{n(m-1)}{\sum p_i(1-p_i)}.$$

Hence, from (3.33),

$$\mathrm{tr}\left(\Omega^{-1}\overline{\mathrm{Var}}(X|U)\right) \simeq \frac{n(m-2)}{\sum p_i(1-p_i)}. \tag{3.34}$$

Also, using (3.32), we find

$$\bar{\mathrm{E}}(D^2|U=u) = \bar{\mathrm{E}}(X^T\Omega^{-1}X|U=u) \simeq \frac{n}{\sum p_i(1-p_i)}\left(m-1+\frac{(u-\bar{u})^2}{\sigma_u^2}\right). \tag{3.35}$$

Applying (3.35) to (3.22), the corresponding approximation to the overestimation in ROC is

$$S(u) \simeq \frac{p_u f(u)}{\bar{p}\sum p_i(1-p_i)}\left[(m-1) - \frac{\sum p_i(1-p_i)}{n\bar{p}(1-\bar{p})} - \frac{(u-\bar{u})^2}{\sigma_u^2}\right]. \tag{3.36}$$

Note that the term in square brackets is the sum of three components: $(m-1)$, the number of non-intercept covariates in the logistic regression, the centre term which lies between 0 and 1 according to the distribution of the true probabilities, and the standardized squared distance between u and the overall mean $\bar{u}$.

Applying (3.34) to (3.30), the corresponding approximation to the overesti-mation in the area under ROC is .

$$\frac{m-2}{n\bar{p}(1-\bar{p})}\mathrm{E}^* f(U) \simeq \frac{m-2}{n\bar{p}(1-\bar{p})} \cdot \frac{\bar{\mathrm{E}}\left(p_U(1-p_U)f(U)\right)}{\bar{\mathrm{E}}\left(p_U(1-p_U)\right)}. \tag{3.37}$$

Whereas (3.29) is a cumbersome expression, the corresponding approximation (3.37) is quite simple. In our examples we have simply fitted the logistic regression to give the estimates of the probabilities p_i and their logits u_i, and then sorted the u_is into class intervals. Suppose that the jth interval has class mid-point $u^{(j)}$, class width $d^{(j)}$ and class frequency $f^{(j)}$. Let $p^{(j)}$ be the probability whose logit is $u^{(j)}$ and let there be n_c class intervals in total. Then (3.37) is estimated by

$$\frac{m-2}{n^2\tilde{p}(1-\tilde{p})} \cdot \frac{\sum_j^{n_c} p^{(j)}(1-p^{(j)})(d^{(j)})^{-1}(f^{(j)})^2}{\sum_j^{n_c} p^{(j)}(1-p^{(j)})f^{(j)}}, \tag{3.38}$$

where

$$\tilde{p} = \frac{1}{n} \sum_{j}^{n_c} p^{(j)} f^{(j)}.$$

These formulae are based on the assumption that the appropriately weighted empirical distribution of the x's is approximately Gaussian. The fact that only marginal expectations over all possible values of the conditioning variate U are involved may suggest that these approximations are robust to this distributional assumption, but there are no theoretical results to justify this. It has been shown, however, that in all the bootstrap simulations studied (some of these are reported in Chapter 4), the value of (3.38) provides a good estimate of the overestimation in the area under the retrospective ROC, even in cases where the x's are clearly far from Gaussian, for example when the covariates are categorical.

3.5 Comments

The reader may wonder why the word overestimation is used rather than the more usual term bias. In the context of ROC, bias would refer to the difference between the expected value of $\hat{C}$ and the 'true' population ROC curve, i.e. the ROC for the true score $\beta^T x$ acting on the true population distributions of y and x. We are comparing two random quantities, since both $\hat{C}$ and C are functions of the sample estimate $\hat{\beta}$. For the practical validation of risk scores we are interested

in the future performance of the actual score $\hat{\beta}^T x$ and not in the properties of the hypothetical score $\beta^T x$.

In Section 2.3.2 we examined an argument by Lloyd, which stated that the empirical estimates of the misclassification rates of a logistic regression are not maximum likelihood if the logistic regression model is assumed to be true. In this Chapter we take the usual non-parametric estimate of the ROC curve and work out how big the overestimation would be if the logistic model were true. The bootstrap and cross-validation procedures suggest that the overestimation doesn't seem to depend very sensitively on whether the logistic model is actually correct. If the logistic model is false then Lloyd's ROC is wrong by an amount O(1) whereas the overestimation calculation would only be wrong by an amount $O(1/n)$. The relationship between overestimation and Lloyd's ROC is of little relevance in the context of this thesis as we are studying the overestimation of the ROC curves that people are actually using in practice.

The approximations in Section 3.4 suggest that asymptotic overestimation in ROC is $O(n^{-1})$, whereas the sampling standard deviation of ROC is $O(n^{-\frac{1}{2}})$. However the role of the number of covariates is important, (3.37) giving the overestimation as proportional to the ratio $(m-2)/n$. The orders of magnitude of these quantities are similar to those mentioned in Chapter 2 in the discussion of over-fitting and shrinkage in linear regression models (see Copas (1983)). As

also mentioned in Chapter 2 the future accuracy of linear predictions might be measured by prediction mean squared error ($PMSE$), which in the simplest situation, and using the usual notation, is like $\sigma^2(1+m/n)$. For a given strength of the covariates (for a given residual variance σ^2), the prediction mean squared error inflates by adding a term proportional to m/n.

The results of the simulations for large sample sizes, and to a lesser extent for smaller sizes seem to agree very well with the overestimation predicted by the formulae in (3.21) and (3.29). One may ask what is the use of the formulae at all if in practice we could just use the cross-validation procedure to assess the overestimation? As well as being able to express the overestimation in convenient analytical expressions, we were able to use the formulae to give an insight into the size of the overestimation and what it depends upon (see discussion in previous paragraph). Cross-validation would give an estimate of the overestimation, but it would just be a number and we would have no idea why it would be big in one case and small in another.

It should be noted that the overestimation in area is not the same as the appropriate integral under the function $S(u)$. This is because the calculations of overestimation in the true positive rates is conditional on a fixed threshold u, whereas overestimation in the area involves integrating with respect to the false positive rate which is also random. Technically, the difference is seen in the fact

that the term $\bar{\epsilon}H_i$ in the development in Section 3.2 is $O(n^{-\frac{1}{2}})$, but the analogous term in the development in Section 3.3 is of lower order of magnitude.

As mentioned in Chapter 2, the ROC is invariant under monotonic transformations in the score, and so if there is just a single covariate ($m = 2$) any sampling error in $\hat{\beta}$ will not affect the ROC curves (provided the non-intercept coefficient has the correct sign). This is seen in formula (3.37) which is zero when $m = 2$.

Finally, as described in Chapter 2, another way of assessing the effectiveness of risk scores is given by the so called 'logit rank slope' (Copas (1999)). Appendix B of this paper derives an asymptotic approximation to the covariance between y_i and the sample rank of $\hat{u}_i$. We have shown that the area under the ROC curve can be represented as a function of the rank sum statistic (2.2). Using this relationship and reworking the approximation in the cited paper above gives an expression for the bias in the rank sum statistic, and this is another way of deriving approximation (3.30).

Chapter 4

The Corrected ROC Curve: Examples and Simulation Study

In this chapter we shall present a number of examples to demonstrate the overestimation of the ROC curve and area for logistic regression as derived in Chapter 3. We shall focus on two medical examples, a study on melanoma (a type of skin cancer) and a study on advanced breast cancer. A simulation study to assess the properties of some of the formulas presented in Chapter 3 is also included, using some of the sampling techniques described in Section 2.3.3.

4.1 Melanoma case-control study

For our first example we take data from a case-control study of melanoma reported by Berwick *et al* (1996) (see the cited paper for full details of the study). This study has the common aim of assessing the risk of disease in terms of a number of measurable risk factors. A number of the individuals in this study had incomplete data (and one duplicate record), for this reason 872 records out of an original 999 were available for analysis and the subjects are roughly divided between cases and controls. These data are also studied by Begg *et al* (1998), who discussed a number of approaches to measuring the strength of risk factors in predicting risk. Figure 1 of the cited Begg paper shows the ROC curve for the linear discriminant function calculated from the data. There are twelve covariates and a binary indicator of case/control status - when the categorical variables are replaced with dummy variables we have $m = 15$ (including an intercept term) and $n = 872$. A description of the variables is presented in Table 4.1.

A logistic regression of these data was performed and using the Wald test we see that 10 out of the 14 covariates significantly contribute to the logistic regression, the strongest contributions being from age, tendency to tan, family history of skin cancer and total number of nevi on the subject's arms and back. The ROC curve for this logistic regression, $\hat{C}$ can be seen in Fig. 4.1.

Name	Description	Variable Type
SEX	Self-explanatory	Categorical
AGE	Self-explanatory	Numeric
SKINCOL	Skin colour	Categorical
SKINTYP1	Tendency to Burn	Categorical
SKINTYP3	Tendency to Tan	Categorical
BURNPAIN	Pain due to Sunburn	Categorical
FRECK25	Develops Freckles before age 25	Categorical
HAIR	Hair colour type	Categorical
EYES	Eye Colour	Categorical
SUNTOTG	Level of Sun Exposure	Categorical
CASE	Binary indicator of status	Categorical
RELSCKA	Family history of skin cancer	Categorical
TOTNEVI	Total no. of nevi (moles)	Numeric

Table 4.1: Description of the variables in the melanoma case-control data

When comparing the ROC curve in Fig 4.1 with the ROC curve in Figure 1 of the cited Begg paper we see that they are very similar, even though one has been calculated from a linear discriminant function and one from a logistic regression. This might suggest near equality of the two methods under the distributions of

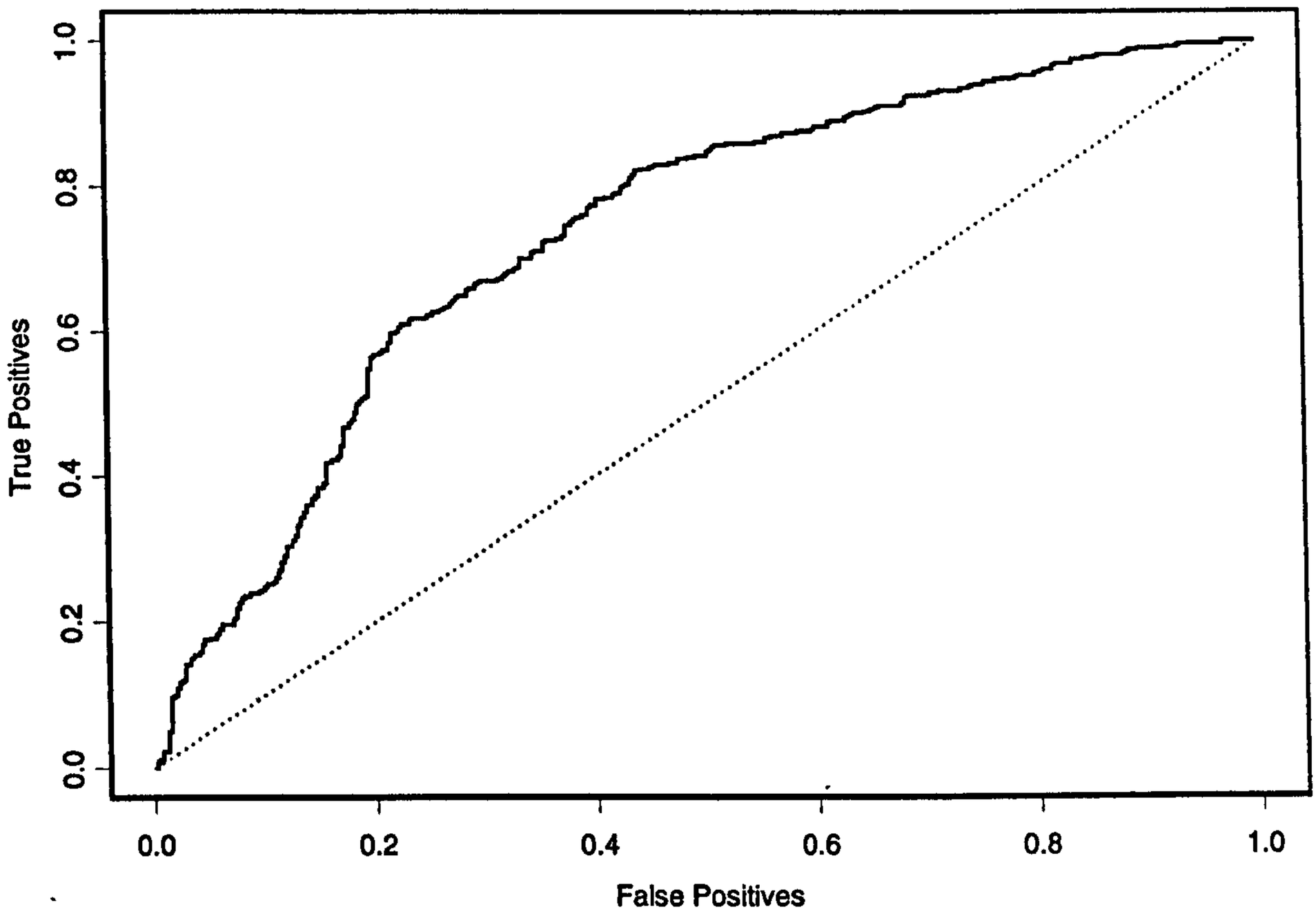

Figure 4.1: ROC Curve for the melanoma logistic regression

the covariates.

Recall the correction in the ROC curve can be calculated by way of (3.21)

$$S(u) = \mathrm{E}(\hat{F}_1(u) - F_1(v_u)) \simeq \frac{p_u}{n^{\frac{3}{2}}\bar{p}} \sum_i \left(d_i - \frac{1}{\bar{p}(1-\bar{p})d_i} \right) \phi \left(\frac{n^{\frac{1}{2}}(u_i - u)}{d_i} \right).$$

and the corrected ROC curve is given by

$$C^* = \{\hat{F}_1(u) - \hat{S}(u), \hat{F}_0(u)\}.$$

The terms in the above expression are easily calculated from the logistic regression

described above. In particular, Ω in the calculation of d_i can be calculated from the *S-Plus* function `vcov.glm` - see Venables and Ripley (1997) p228. In calculating $S(u)$ we can take the predicted values of the logistic regression themselves as the cutoffs for ease of calculation. The 'corrected' ROC curve for the logistic regression (score) from the melanoma study is shown as the stippled curve in Fig. 4.2. The area under the ROC curve, $\hat{C}$ is 0.7370 giving the model only adequate to good discriminatory power even though two thirds of the included covariates

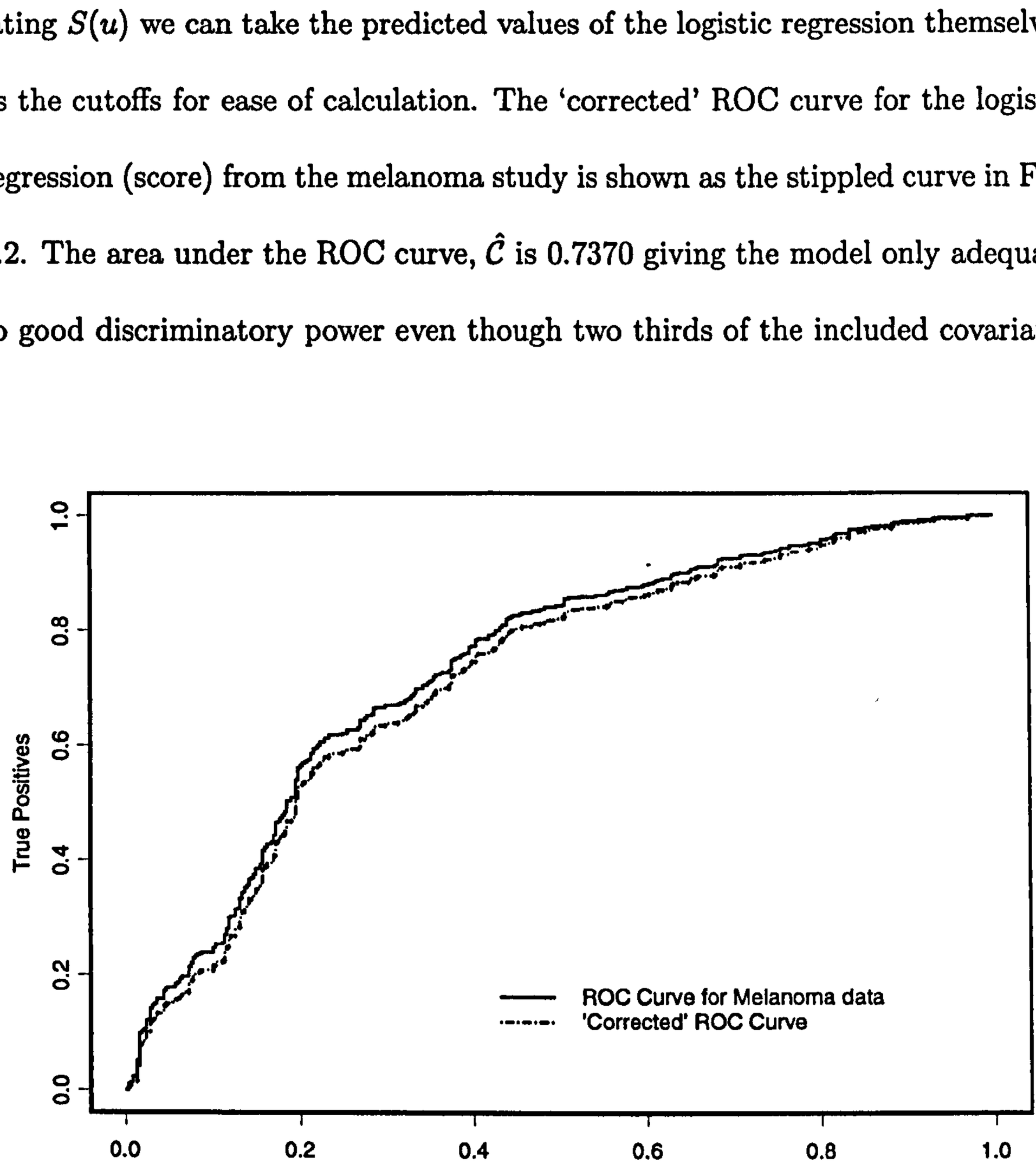

Figure 4.2: Original and 'corrected' ROC curves for the melanoma score

are significant. The overestimation in the area calculated by the expression in (3.29) is 0.01933. The approximation to (3.29) in (3.38) can be calculated by using the class intervals and frequencies from a histogram of the predicted values $\hat{u}_i$ from the logistic regression (other density estimates can also be used to calculate the frequencies and class intervals). Different values of (3.38) can be obtained by varying the number of class intervals - these are presented in Table 4.2. As we can see from the table, (3.38) gives a very good approximation to (3.29).

Number of class intervals	Value of (3.38)
10	0.0191
20	0.0198
30	0.0203

Table 4.2: Overestimation of the area under the ROC curve for the melanoma score, using (3.38) with different class intervals

4.2 Simulation study

To assess the properties and accuracy of some of the estimates in Section 4.1 we carried out a simulation study involving a parametric bootstrap simulation and two differing types of cross-validation to directly examine the ROC curves $\hat{C}$ (retrospective) and C (prospective).

The parametric bootstrap procedure is described by the following algorithm

1. Fit a logistic regression to obtain a set of fitted probabilities p_i.

2. Generate a replicate set of y_i's, call these y_i^* randomly from the p_i's.

3. Fit a new logistic regression using y_i^* on the original covariates x_i to give a new regression vector, call this β^*.

4. Calculate the ROC curve C_1 using score $u_i^* = \beta^{*T} x_i$ and data y_i^* ('retrospective').

5. Calculate the ROC curve C_2 using score $u_i^* = \beta^{*T} x_i$ and original fitted probabilities p_i ('prospective').

6. For a fine grid of values along the horizontal axis of the unit square we use linear interpolation to find the vertical heights between the two ROC curves described in 4. and 5. and subtract to give C_1 - C_2, the overestimate of the true positives.

7. Repeat this procedure n_{sim} times and average the difference to give an estimate of the expected overestimation.

It is important to distinguish between the parametric (as used in this example) and the non-parametric bootstrap. Typically, the non-parametric bootstrap would involve randomly sampling from the data to create a new data set from

which we calculate a logistic regression. We then return to the original data to test the prediction rule. This type of bootstrap sampling procedure ignores the distribution of the covariates, whereas in the parametric case we are implicitly assuming that the model is correct when we generate the vector y_i^*. The above procedure was calculate for $n_{sim} = 1000$ repetitions and the result can be seen as the stippled line in Fig. 4.3 , to be compared with the solid line which is the

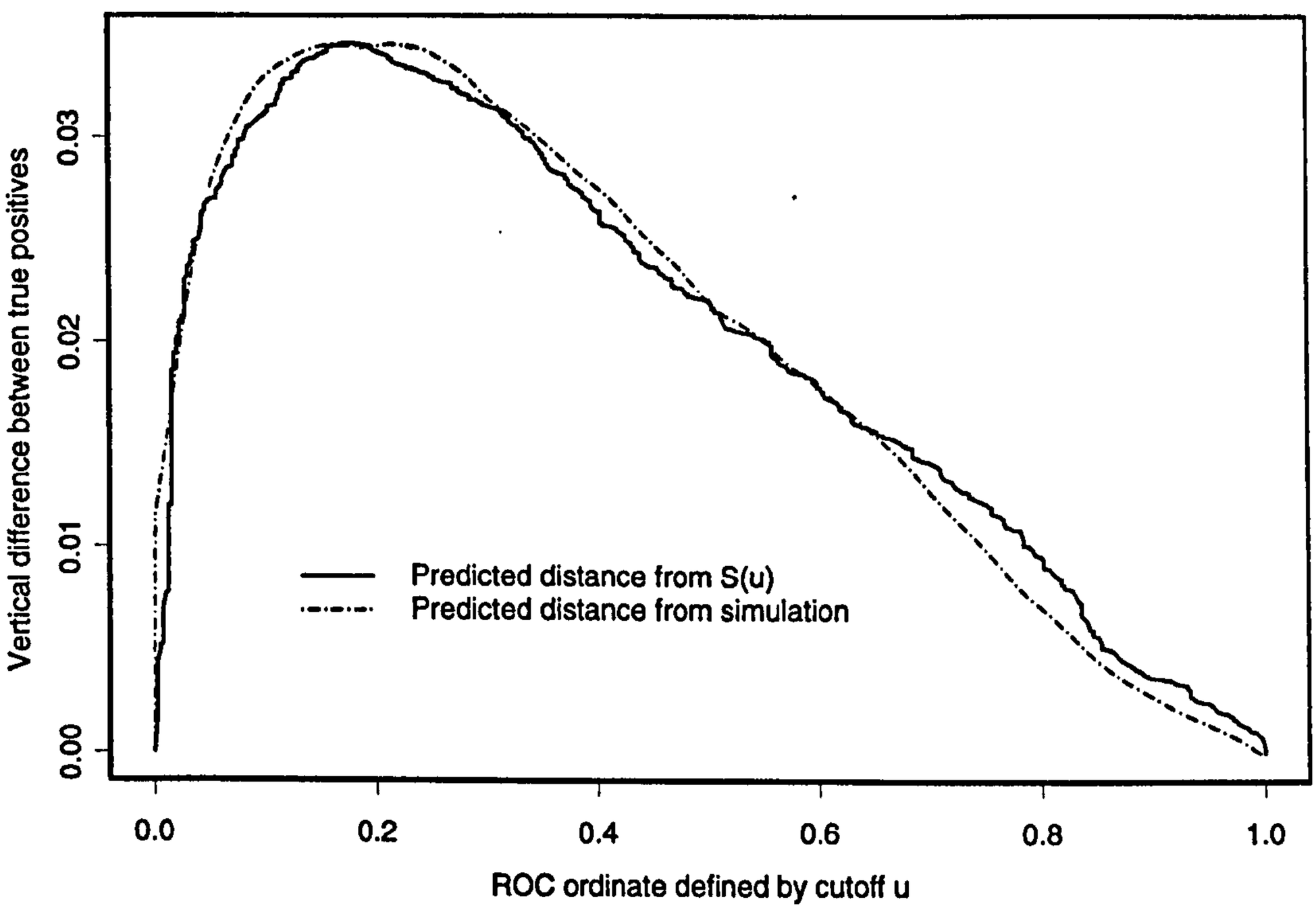

Figure 4.3: Predicted and simulated overestimation of the ROC curve for the melanoma score

estimate of $S(u)$ from (3.21). The agreement is particularly good. It is important to note that the overestimation of ROC is greatest towards the left of the curve, which is usually the region of interest, corresponding to values of the threshold for which the false negative rate is small.

We can also use the above procedure to test the formulae for the overestimation in the area. For each of the simulations we calculate the areas of C_1 and C_2 and subtract to find the difference. After $n_{sim} = 1000$ simulations we find that the average difference in the area is 0.0197 ($\pm$ 0.0003) which compares favorably with the value of (3.29) 0.0193 and the value of (3.38) 0.0191 (on 10 class intervals).

The first of our cross-validation approaches is the 'leave one out' method described in Section 2.3.3. The 'leave one out' method can be described by the following procedure

1. Omit record i from the data set comprising of n individuals.

2. Calculate a logistic regression from the remaining (n - 1) individuals, call the regression vector $\beta_{(-i)}$.

3. Use $\hat{\beta}_{(-i)}$ and the omitted record from individual i to create the score $\hat{u}_{(i)}$.

4. Repeat the procedure until we have n pairs of values $(\hat{u}_{(i)}, y_i)$, from which we can construct the ROC curve.

This procedure was carried out for each of the $n = 872$ individuals in the Melanoma study. The results can be seen in Fig. 4.4, where the solid line is the original ROC curve $\hat{C}$, the stippled curve is the 'corrected' ROC curve and

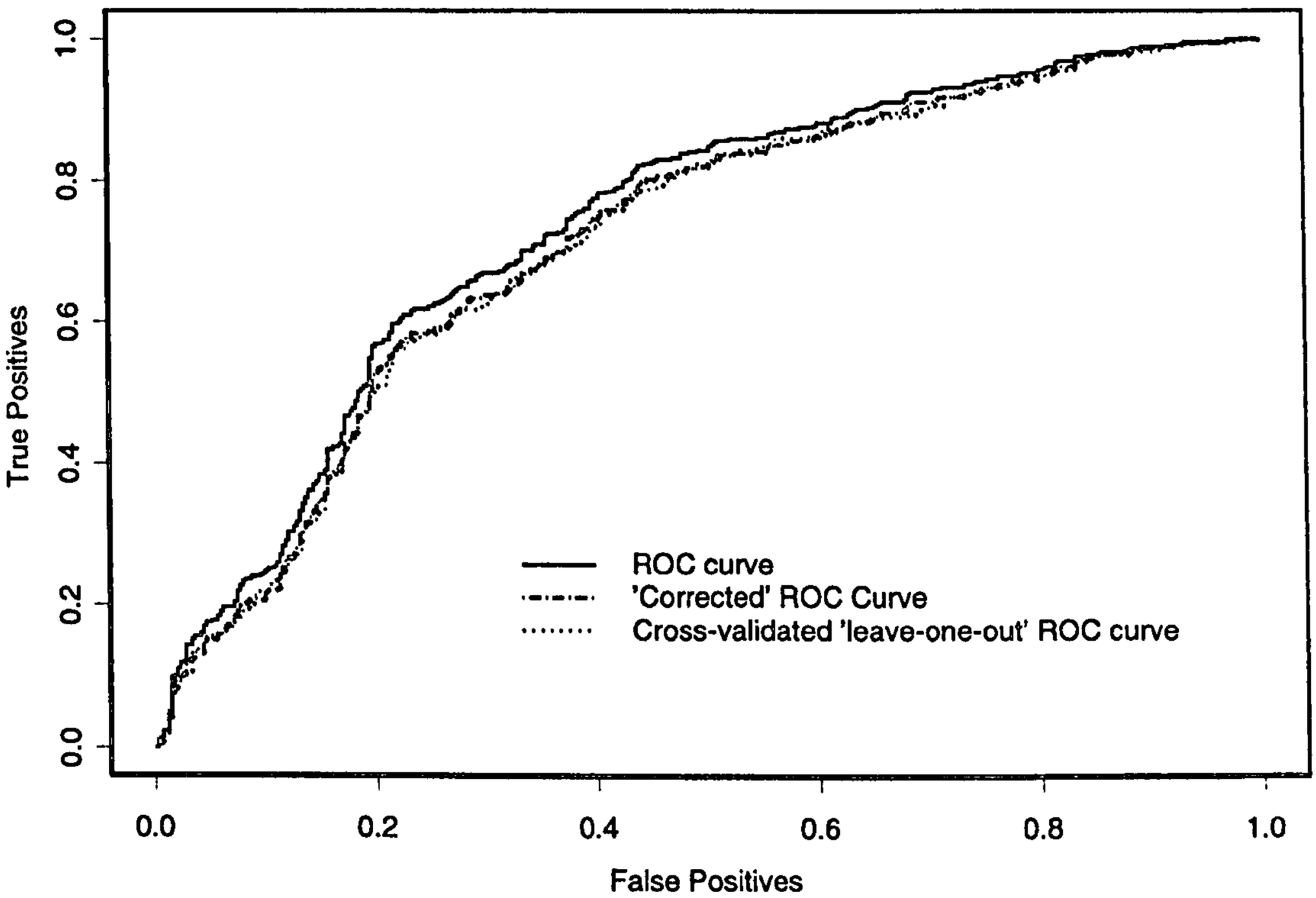

Figure 4.4: Retrospective, 'corrected' and cross-validated 'leave-one out' ROC curves for the melanoma score

the dotted curve is the cross-validated 'leave one out' ROC curve. As we can see, the cross-validated curve is in close agreement with the 'corrected' curve. Fig. 4.5 shows the plot of the 'leave one out' ROC curve against the parametric bootstrap prospective ROC curve C_2 (as described in the algorithm for the parametric bootstrap). The curves are broadly in line with each other, except for a deviation in the lower part of the curve. This deviation could arise from the

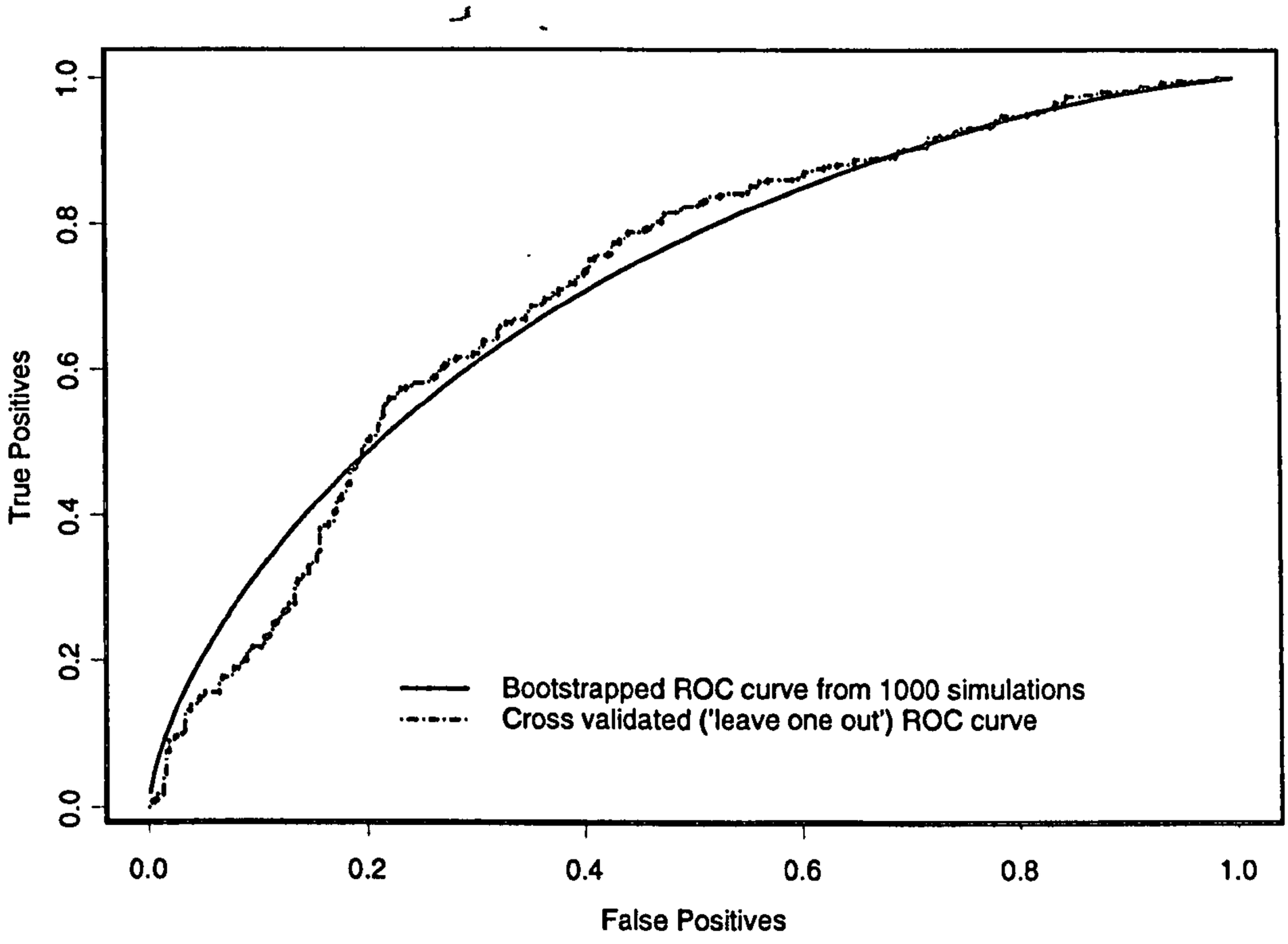

Figure 4.5: Parametric bootstrapped prospective and cross-validated 'leave-one out' ROC curves for the melanoma score

particular model that we are using to calculate the ROC curve. The 'smoothed' nature of the bootstrapped ROC curve is most likely due to averaging over the large number of simulations performed, as well as the parametric nature of the bootstrap simulation which places emphasis on the model. The area under the cross-validated curve is 0.7145, a difference of 0.0225 from the area under $\hat{C}$. This difference is in quite close agreement to the values predicted by (3.29) and (3.38).

The second method of cross-validation comprises sample splitting. In a sense the 'leave one out' method is a specialized case of sample splitting, as we are splitting the sample into a training set of $(n-1)$ individuals, a validation set of 1 individual and repeating the procedure on all n combinations of individuals. For our purposes we use the following algorithm for sample splitting:-

1. Choose a sampling fraction f_{SS} to split the data.

2. Split the data into a training set T of n_{SS} individuals with binary indicator y_i and a validation set V of $n - n_{SS}$ individuals with binary indicator y_i^*.

3. Calculate a logistic regression on the data T to obtain a regression vector β and scores $\hat{u}_i$.

4. Use $\hat{\beta}$ and the validation data V to calculate risk scores u_i^*.

5. Calculate the ROC curve C_1 using the scores $\hat{u}_i$ and y_i ('retrospective').

6. Calculate the ROC curve C_2 using scores u_i^* and y_i^* ('prospective').

7. Repeat the same linear interpolation procedure for a fine grid of values as described in the parametric bootstrap algorithm above to obtain the difference C_1 - C_2.

8. Repeat the procedure n_{sim} times and average the difference to obtain the expected overestimation in the ROC curve.

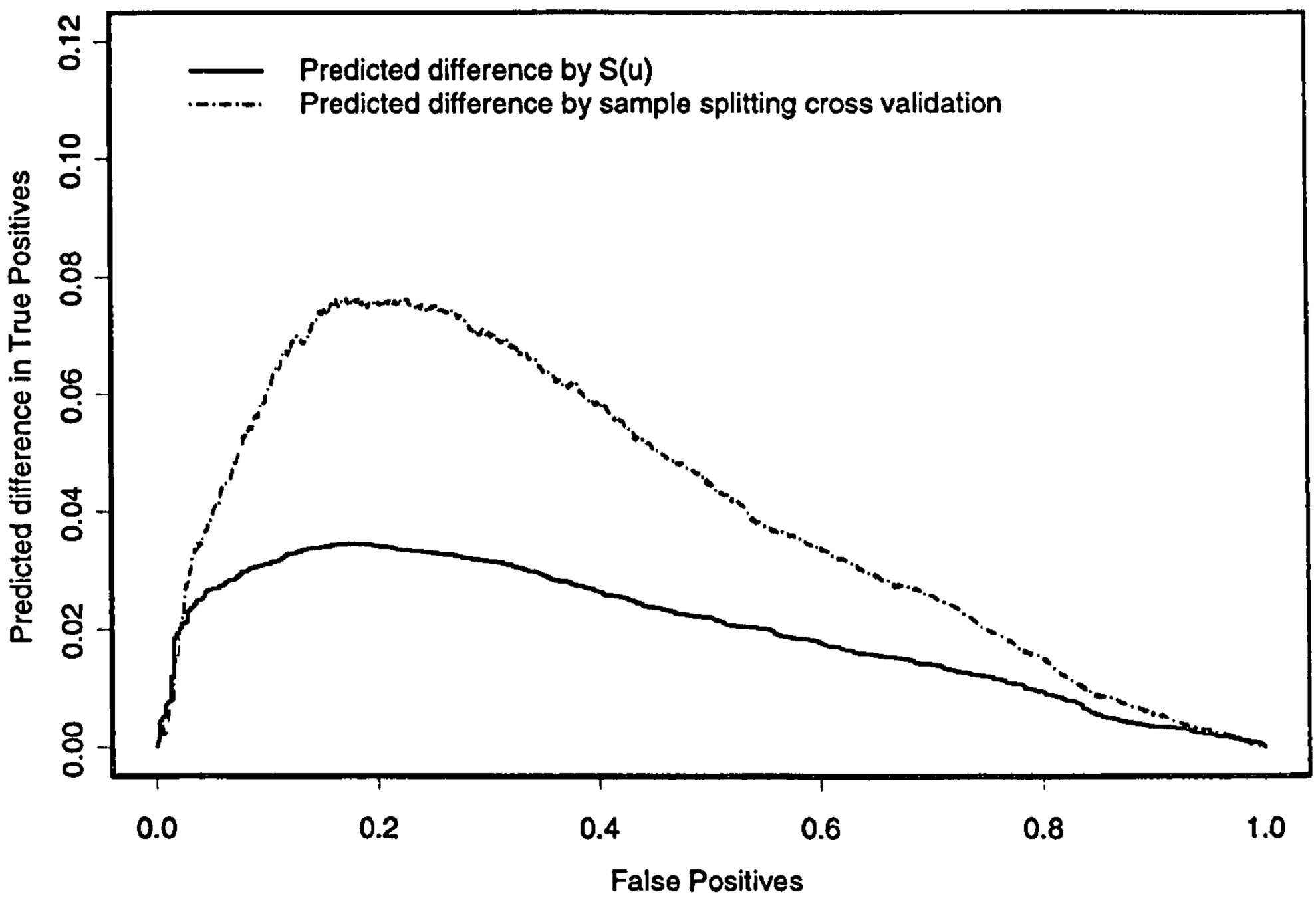

Figure 4.6: Predicted overestimation of the ROC curve by $S(u)$ and by the sample splitting cross-validation approach (splitting fraction = $\frac{1}{2}$) for the melanoma score

The above sampling splitting procedure was completed for $n_{sim} = 500$ repetitions, each of which had $f_{SS} = \frac{1}{2}$ i.e. the data were split in half. The result can be seen in Fig. 4.6 on the previous page. At first glance we see that the predicted difference is not in close agreement and from a rough glance it appears that the sample splitting approach has calculated the overestimation as about double what it should be. From (3.23) we see that the overestimation of the ROC area is $O(n^{-1})$. So if we calculate the overestimation by the sample splitting approach with $f_{SS} = \frac{b}{c}$ then we have to scale the resulting overestimation by $\frac{b}{c}$ to get the correct result. An illustration of this is by looking at the mean difference in areas over the 500 simulations and comparing with the value predicted by (3.29).

As we can see from Table 4.3, by multiplying the average area from the sample splitting simulation by the sampling fraction we obtain another estimate of the overestimation of the retrospective ROC area. Both the figures in the table are in close agreement with the value of (3.29) and its approximation (3.38).

f_{SS}	Size of T	Size of V	Average area	Average area x f_{SS}
$\frac{1}{2}$	436	436	0.03980	0.01990
$\frac{2}{3}$	582	291	0.03076	0.02051

Table 4.3: Overestimation of the area under the ROC curve for the melanoma score, using the sample splitting approach with two different sampling fractions

4.3 Breast Cancer study

We now focus our attention on a study with a smaller sample size. In this medical example, we take the study of prognosis in breast cancer reported in Armitage *et al* (1966) where a sample of n = 187 patients were followed up after surgery. The aim of the study is to see whether clinical characteristics measured before surgery could be effective in discriminating in advance between patients who responded to surgery (y=1) and patients who did not (y=0). The number of individuals responding to surgery was 47 (leaving 140 non-responders) and for the analysis we take four prognostic indicators (see Table 4.4) which results in a

Name	Description	Variable Type
Free Period	Time between primary treatment and recurrence of disease	Numeric
Age	Self-Explanatory	Numeric
Operation	Type of operation	Categorical
Disc	Discriminant score calculated from a linear function of two steroids	Numeric

Table 4.4: Description of the variables in the breast cancer study

value of $m = 5$. A logistic regression of these data on the outcome was performed (none of the covariates were transformed prior to modelling) and of the covariates,

only Disc significantly contributed to the logistic regression. This agrees with an assertion in the accompanying literature to the data set that stated that response to surgery is closely related to levels of steroids in the blood.

The ROC curve for the breast cancer data is shown in Fig. 4.7. The 'step-like' appearance of the non-parametric ROC curve is highlighted here because of the reduced sample size.

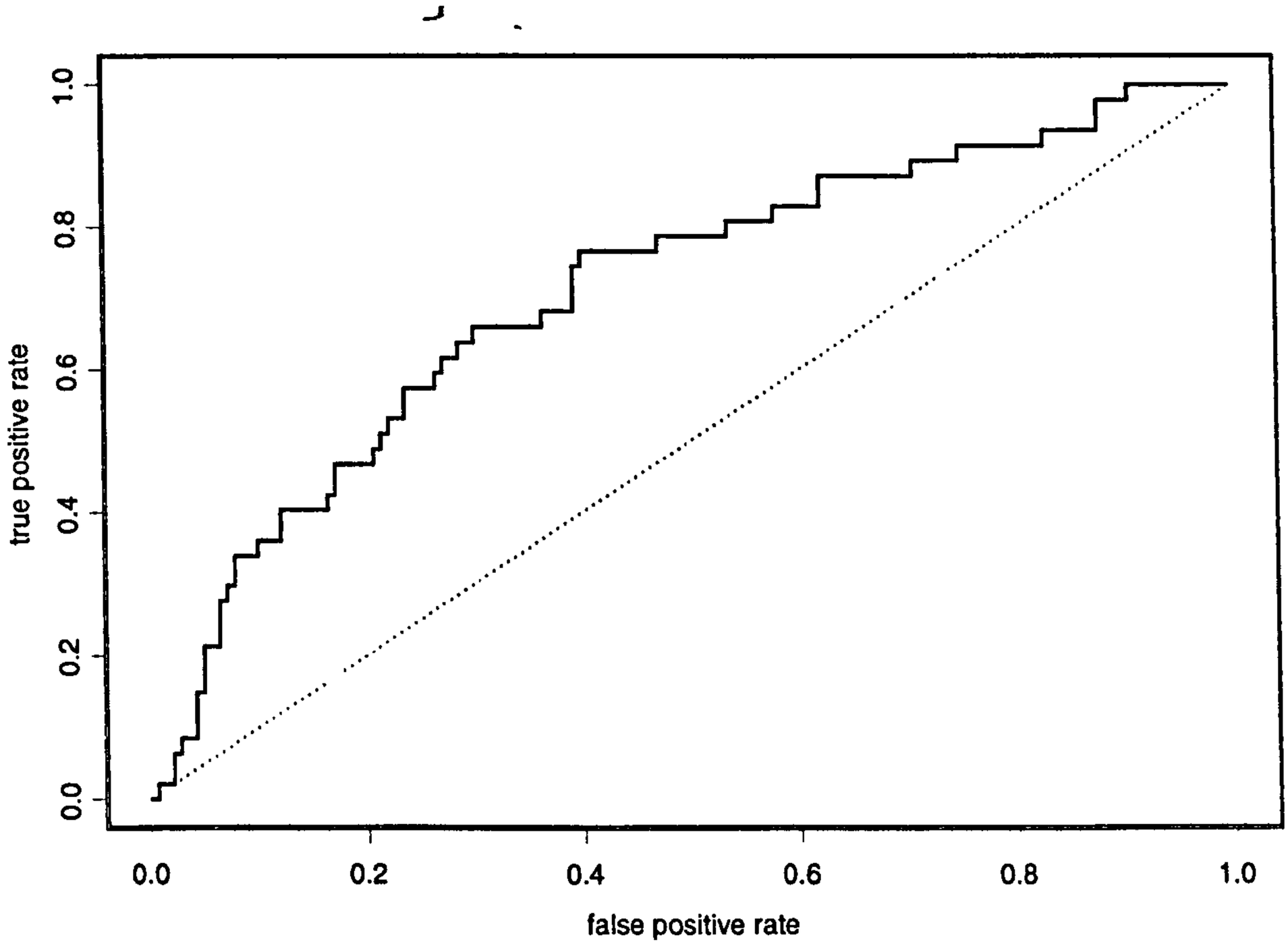

Figure 4.7: ROC curve for the breast cancer logistic regression

As before the overestimation in the ROC curve can be calculated by way of (3.21), and the resulting 'corrected' ROC curve is shown in Fig 4.8.

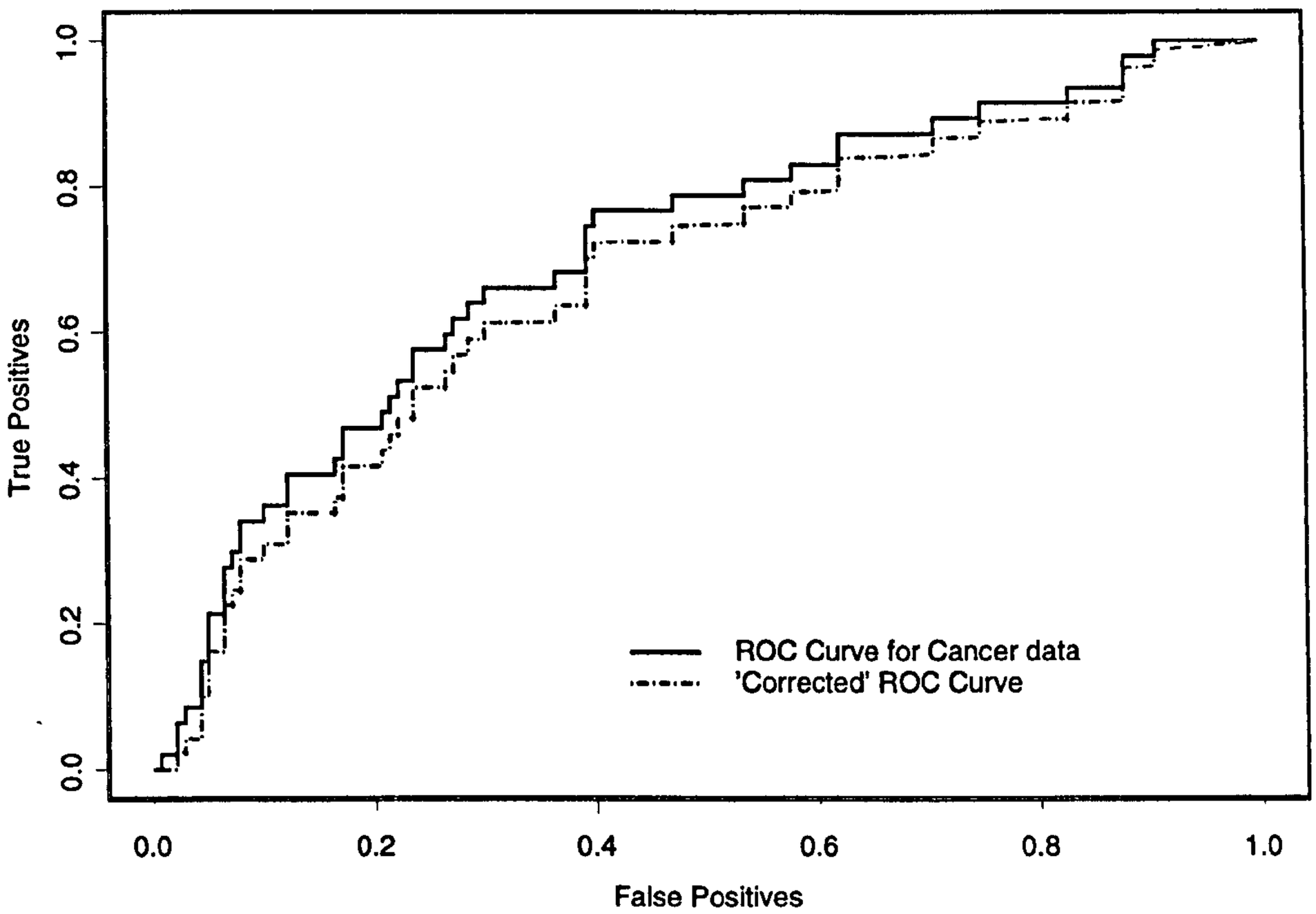

Figure 4.8: Original and 'corrected' ROC curves for the breast cancer score

The area under the ROC curve $\hat{C}$ is 0.7109 giving the model adequate discriminatory power. The overestimation in the area calculated by the expression in (3.29) is 0.0295, suggesting that the area shrinks to approximately 0.681. In terms of the Gini Coefficient this equates to a drop from 0.422 to 0.362, suggesting that the retrospective analysis overestimates the performance of the logistic

regression by about 16%. The approximation to (3.29) in (3.38) for various class intervals can be seen in Table 4.5. As with the Melanoma example, (3.38) gives a very good approximation to (3.29).

Number of class intervals	Value of (3.38)
10	0.0305
20	0.0319
30	0.0319

Table 4.5: Overestimation of the area under the ROC curve for the breast cancer score, using (3.38) with different class intervals

4.4 Simulation study

We use the same resampling procedures described in Section 4.2 to test the properties of the overestimation of the ROC curve and area for the breast cancer data.

The parametric bootstrap was calculated for $n_{sim} = 1000$ simulations and the results of the average of the simulations can be seen in Fig. 4.9. The agreement between the curves is not so close as the Melanoma case study as a result of the smaller sample size but still serves as a good indication and warning of the over-optimism in the retrospective ROC area. The average area under C_1 is 0.715

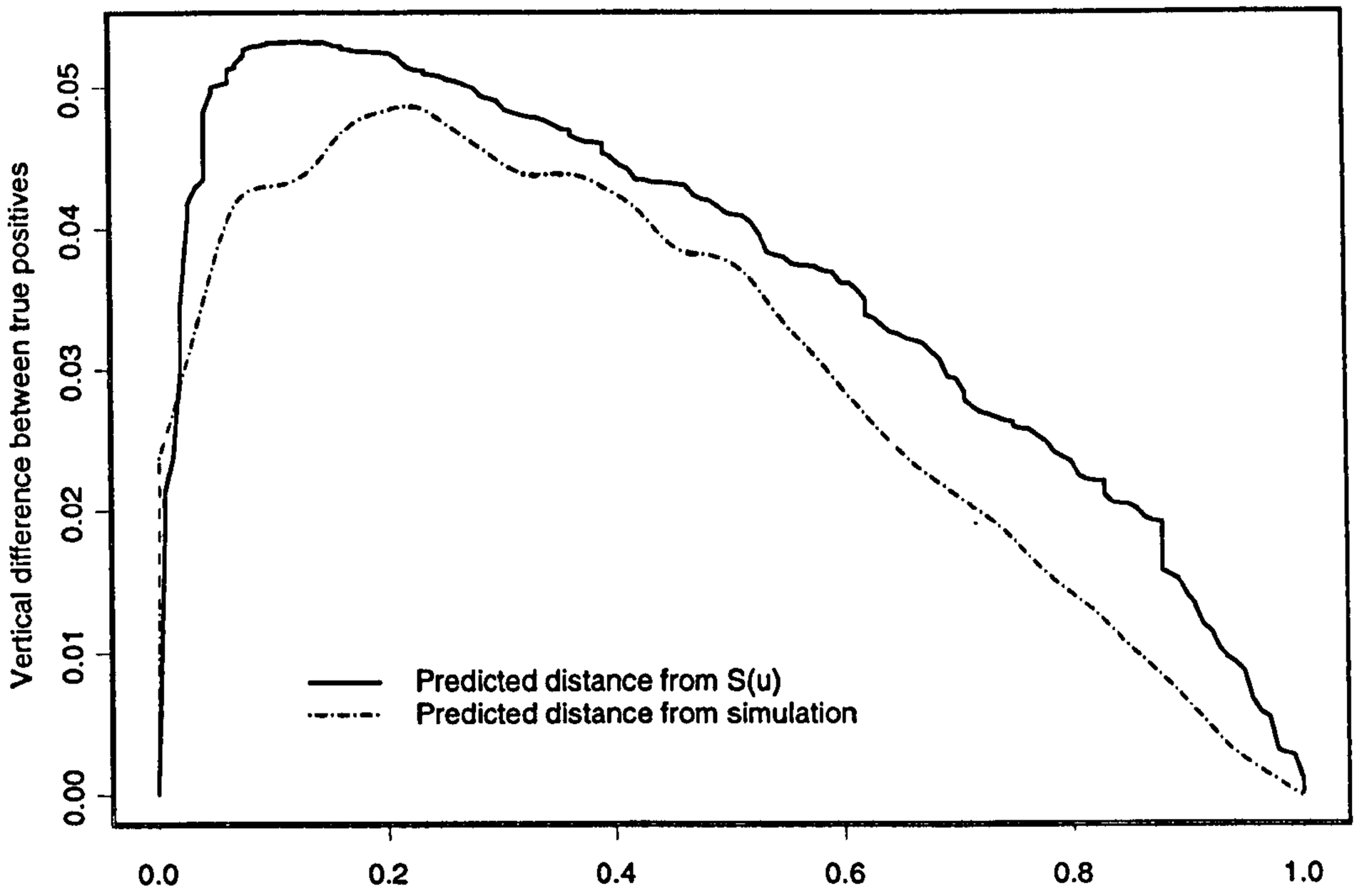

Figure 4.9: Predicted and simulated overestimation of the ROC curve for the breast cancer score

and the average area under $\mathcal{C}_2$ is 0.685, a shrinkage in area of 0.300 ($\pm$ 0.002) which compares favorably with the value of (3.29) 0.0295 and the value of (3.38) 0.0305 (on 10 class intervals).

The results of the 'leave one out' cross-validation procedure can be seen in Fig. 4.10. As we can see, the agreement between the 'leave one out' curve and

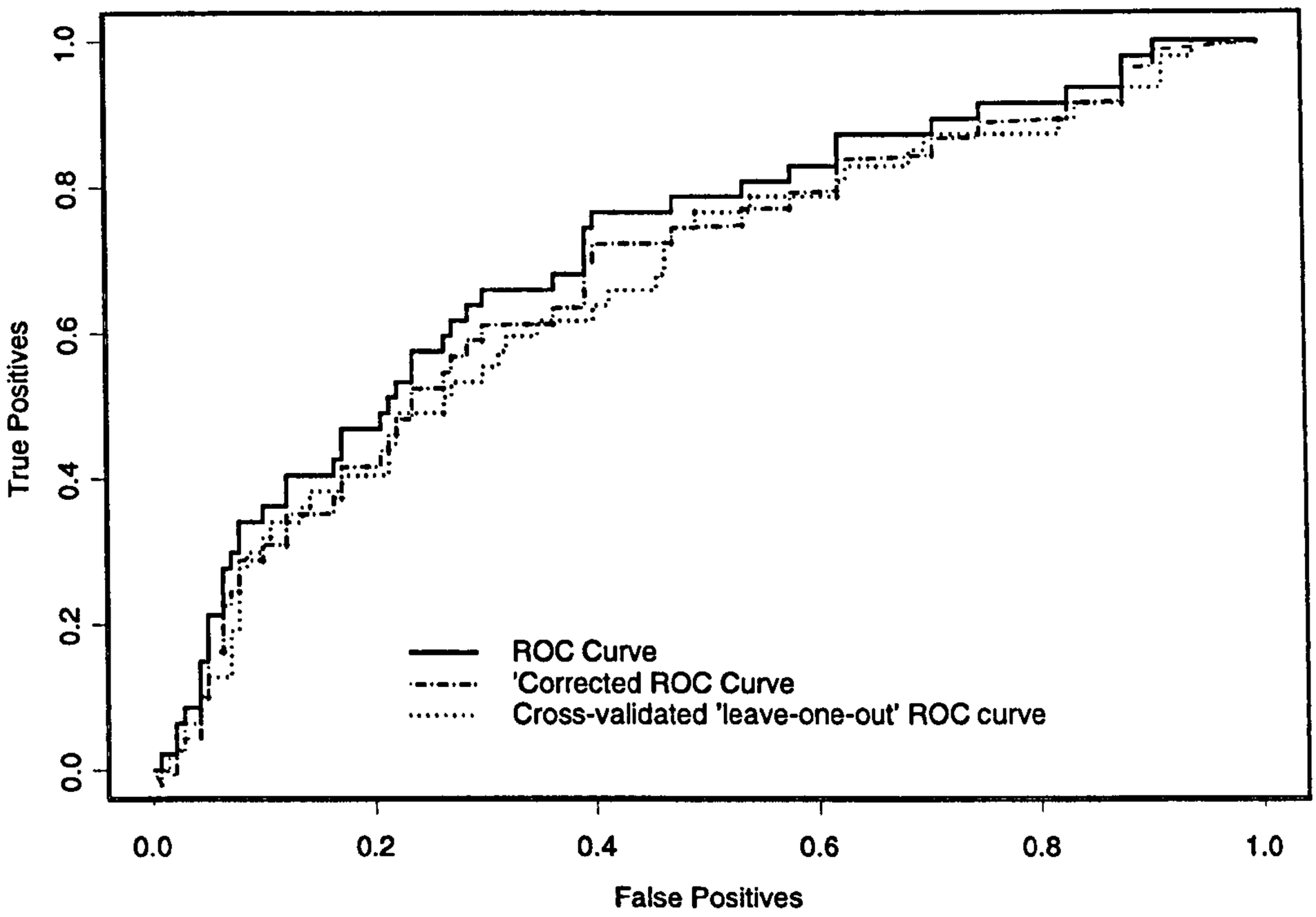

Figure 4.10: Retrospective, 'corrected' and cross-validated 'leave-one out' ROC curves for the breast cancer score

the 'corrected' curve is still good, but not as good as the melanoma data set, again this is most likely due to the reduced sample size.

Fig. 4.11 shows the plot of the 'leave one out' ROC curve against the parametric bootstrap prospective ROC curve as described above. The smooth parametric bootstrap ROC curve closely follows the path of the cross-validated curve,

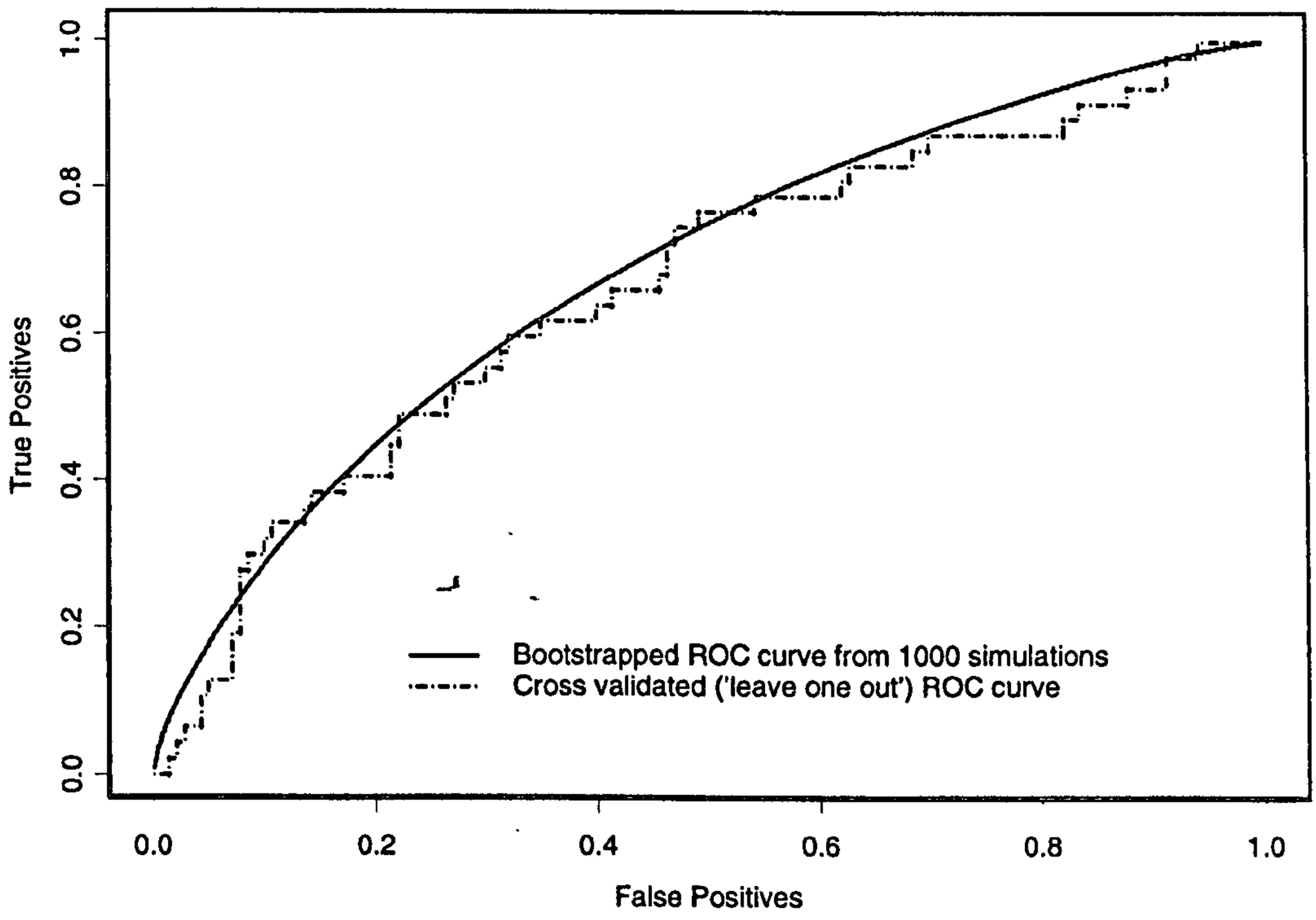

Figure 4.11: Parametric bootstrapped prospective and cross-validated 'leave-one out' ROC curves for the breast cancer score

suggesting in this sense the assumption that the model is correct is justified as opposed to the deviations seen in the melanoma study.

The area under the cross-validated curve is 0.664, a difference of 0.047 from the area under $\hat{C}$. This is not particularly close to the estimate of 0.0295 from (3.29), but we might expect this as the estimate of the area for a 'stepped' ROC

curve from a smaller data set can be significantly affected by the way the area is defined (evaluating trapeziums under the curve, for example). The sample splitting approach can be seen in Fig. 4.12. As with the previous example, splitting the data in half appears to have doubled the overestimation of the curve. We can take an analogous approach to before and study the overestimation in

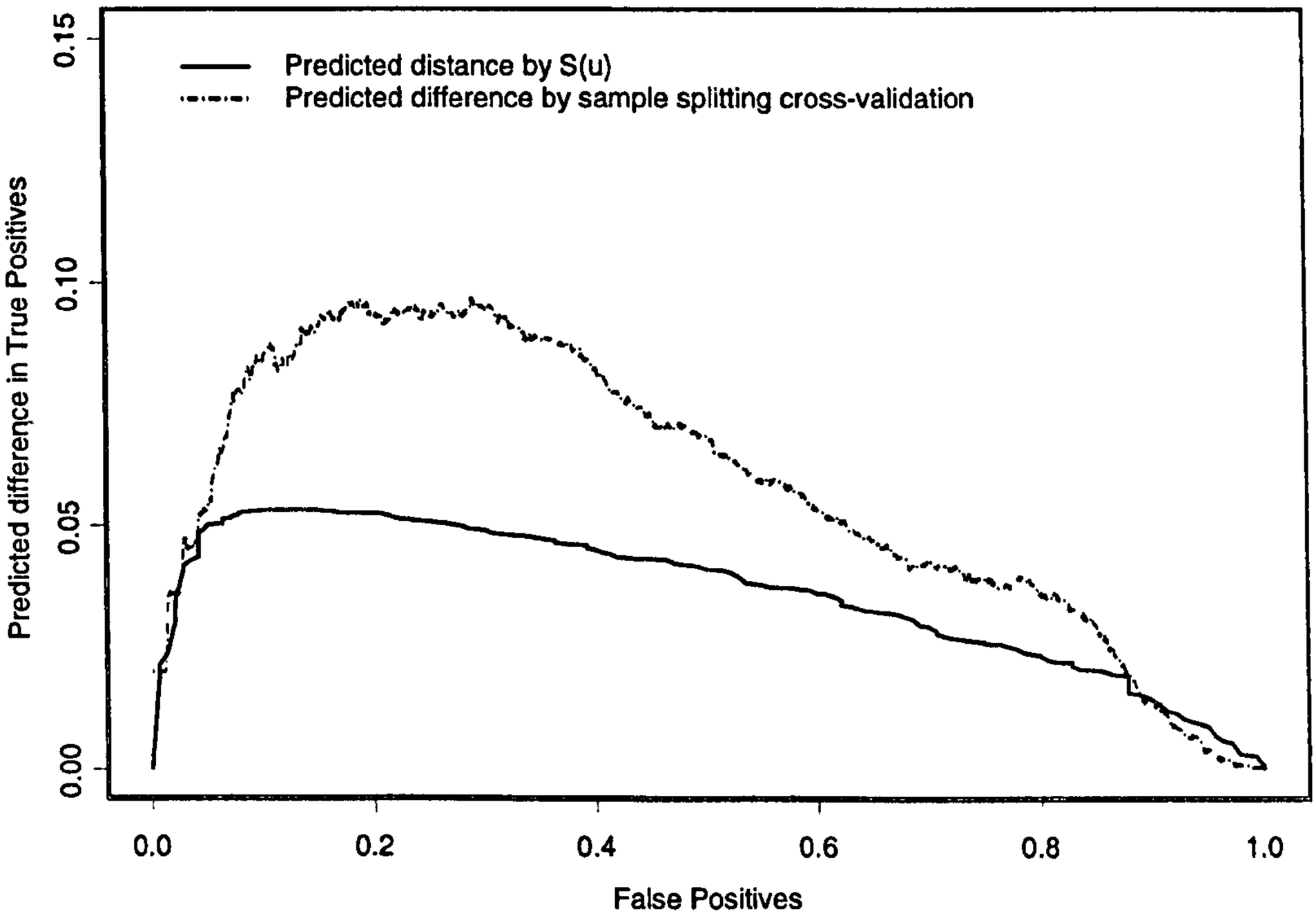

Figure 4.12: Predicted overestimation of the ROC curve by $S(u)$ and by the sample splitting cross-validation approach (splitting fraction = $\frac{1}{2}$) for the breast cancer score

areas produced by this approach. From Table 4.6, as before by multiplying the average area from the sample splitting simulation by the sampling fraction we obtain another estimate of the overestimation of the retrospective ROC curve. The figure for the $f_{SS} = \frac{1}{2}$ is close to the value produced by (3.29). The figure for $f_{SS} = \frac{2}{3}$ does not correspond closely to the predicted value - the reason for this is most likely the small size of the validation sample.

f_{SS}	Size of T	Size of V	Average area	Average area x f_{SS}
$\frac{1}{2}$	94	93	0.05802	0.02901
$\frac{2}{3}$	126	61	0.05907	0.03938

Table 4.6: Overestimation of the area under the ROC curve for the breast cancer score, using the sample splitting approach with two different sampling fractions

4.5 Overview

In this chapter, through two medical examples we have illustrated the overestimation of the ROC curve and area for a retrospective assessment of a logistic regression. In the larger of the two studies, involving the melanoma data, the amount of overestimation in the ROC curve and area is verified particularly well by the three sampling procedures we have undertaken. The breast cancer study illustrates the methodology for a smaller sample size and though not being as in

close agreement as the melanoma example, still serves to illustrate how significant the overestimation can be. Many medical studies involving the utilization of the ROC curve for discriminatory purposes often have small sample sizes due to the difficulty in selecting similar patients. There is an approximate 16% overestimation in the Gini Coefficient for the Breast Cancer study on $n = 187$ individuals so the order of overestimation for much smaller studies could be much larger.

As mentioned in Chapter 3 the formulae for the approximation to the overestimation in the area given by (3.38) are based on the assumption that the distribution of the covariates are approximately Gaussian. We mention that we have no results to justify our assumption that the approximations are robust to departures from these distributional assumptions. As can be seen from the example involving the Melanoma data, where eleven of the covariates are categorical the value of (3.38) is a good estimate of the overestimation, providing evidence in favour of the robustness assumption.

Chapter 5

Shrinkage in Logistic Models for Categorical Data

In this chapter we give a new framework for the modelling of categorical data when the outcome variable is dichotomous. The study of categorical data has become more important in recent years as more measurements are made on this scale (see Agresti (1990) for a comprehensive introduction to categorical data analysis). For example, a recent report by Copas *et al* (1996) introduces the OGRS, a risk score which estimates the probability of a convicted offender being reconvicted at least once within two years. One of the covariates used in construction of the score is offence type, a categorical variable recording the principal offence for the current conviction. As we would expect, offence types do not nat-

urally fall into a manageable number of predefined categories, and therefore for the purposes of modelling the reconviction data Copas *et al* coded the offences into nine different categories. This grouping is practically necessary, as the data in its raw form will contain many categories with small numbers of offenders in each which is inadequate for modelling purposes. But if we code offence into a few categories with a large number of individuals in each we are losing information about particular offence types. This chapter aims to address this grouping problem by incorporating prior information about 'success' rates between categories in the analysis.

To illustrate this concept we study the OGRS example further. Originally, there were approximately 1000 offence categories which is patently too many for modelling purposes. These were grouped into the nine general offence categories briefly mentioned above. For the purposes of our example, suppose three of the original 1000 offences were:

Theft from a Car

Theft from a Van

Treason

Looking at these offence types, we would expect that categories Theft from a Car and Theft from a Van could be grouped together with others into a general category entitled Theft. Of course, we would expect that Treason would not

be included in the Theft category due to the great difference in nature of the offences. In this example we are primarily concerned with reconviction rates within offence categories, and it would be fair to assume that the reconviction rates for individuals in the Theft from Car and Theft from Van categories are similar and they are not related to the reconviction rate for the Treason category. We could describe this information in terms of a correlation between reconviction rates, for example the correlation between the rate for Theft from a Car and Treason would be extremely small, perhaps even 0. Conversely we would expect the rates for categories Theft from Van or Theft from Car to be highly correlated, perhaps 0.9. We could also argue that a van is a commercial vehicle, and therefore more likely to be a target for offenders than a private vehicle. We could use this information to justify a change in the correlation between the two Theft rates. We can then summarise the correlations between reconviction rates in the form of a matrix R where (strictly, R is the correlation matrix between the logit of the reconviction rates)

$$R = \begin{bmatrix} 1 & 0.8 & 0.0 \\ 0.8 & 1 & 0.0 \\ 0.0 & 0.0 & 1 \end{bmatrix}$$

where $R_{(1,2)}$ is the correlation between the reconviction rates for categories Theft from a Car and Theft from a Van, while $R_{(1,3)}$ is the correlation between the reconviction rates for categories Theft from a Car and Treason and so on.

By use of this correlation matrix R we can include the information about reconviction rates between offence categories that would otherwise have been discarded in a statistical analysis. Then, by including R in an assumed prior distribution on the logit of the reconviction rates, we can use an Empirical Bayes approach to estimate the parameters of this prior. One of these parameters, τ^2 can be interpreted as the overall variation within the offence categories and using this and the correlation matrix we can use standard Bayesian inference to 'update' our estimates of reconviction rates to reflect our prior beliefs.

We may ask what is the benefit of this procedure apart from reducing the problem of losing information by grouping categorical data? The idea of 'shrinkage' as described Chapter 1 alludes to the notion that incorporating prior information about variables in an analysis gives better predictive accuracy than the model without shrinkage. As well as removing unnecessary grouping, improved prediction by shrinkage is also our aim, although with a subtle difference. To this point, shrinkage in terms of better prediction has been primarily concerned with continuous measurements and *calibration* i.e. does the predicted probability of an event agree well with the observed frequency of an event? For binary outcomes, prediction accuracy is normally described in terms of discrimination error (i.e. how accurately are individuals assigned to one of two groups on the basis of a scoring system) instead of calibration error. As mentioned in previous chapters,

discrimination can be measured by the ROC curve and its area. These topics have been covered in depth and will not be studied again here but we will recap the important distinction between *prospective* and *retrospective* ROC curves and areas. The retrospective ROC area is an assessment of how well a model or score discriminates applied to the data it is constructed from, as has been mentioned before this is an overestimate of how well a score actually discriminates. To obtain a true estimate of the discriminatory power we must assess the model or score on how well it performs in the future (prospectively), ideally through way of an independent set of data. We will use these ideas in the next two chapters to show that in certain circumstances, a prospective ROC assessment indicates that the discriminatory power of the 'shrunk' estimates is better than that of the associated model estimates.

5.1 Background

To explain our methodology, we first review a standard result in Bayesian statistics. We observe a set of independent observations $x_1, \ldots, x_k$ where

$$x_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$$

and the problem is to estimate μ_i. Put $\Sigma = \mathrm{diag}(\sigma_i^2)$.

Suppose the prior on $\underline{\mu}$ is

$$\underline{\mu} \sim \mathrm{N}(\alpha \underline{1}, \tau^2 R)$$

where R is the prior correlation matrix. Assume R is given but fit α and τ^2 to the overall mean of the x_i's. As

$$\mathrm{E}[\overline{x}] = \alpha \qquad \text{and} \qquad \mathrm{E}[s^2] = \tau^2 + \frac{k}{\sum \frac{1}{\sigma_i^2}}$$

where

$$\overline{x} = \frac{\sum \frac{x_i}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}} \qquad \text{and} \qquad s^2 = \frac{\sum \frac{(x_i - \overline{x})^2}{\sigma_i^2}}{\sum \frac{1}{\sigma_i^2}}$$

we can estimate

$$\alpha \text{ by } \hat{\alpha} = \overline{x} \quad \text{and} \quad \tau^2 \text{ by } \hat{\tau}^2 = s^2 - \frac{k}{\sum \frac{1}{\sigma_i^2}}. \tag{5.1}$$

Standard formula for normal Bayesian inference gives the posterior mean as

$$\mathrm{E}[\underline{\mu} \mid \underline{x}] = (\Sigma^{-1} + \tau^{-2} R^{-1})^{-1} (\Sigma^{-1} \underline{x} + \alpha \tau^{-2} R^{-1} \underline{1}) \tag{5.2}$$

and we can estimate this expression by substituting in the estimates of α and τ^2 in (5.1). In principal we would expect R to be available from some expert source.

Clearly R must be positive definite, which is a fundamental property of variance-covariance matrices. If we elicit the correlation matrix from an expert source it is entirely possible that R will be not positive definite and therefore an incorrect correlation matrix. A way around this problem is to replace the negative eigenvalues with 0 and then reconstruct R with the new set of eigenvalues and the original eigenvectors. To do this let the eigenvalues of R be

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$$

and let A be the k x k matrix whose columns are the corresponding normalized eigenvectors. Then $R = A\Lambda A^T$ where $\Lambda = \text{diag}(\lambda_i)$ and $A^T R A = \Lambda$. Put

$$\underline{y} = A^T \underline{x} \quad and \quad \underline{v} = A^T \underline{\mu}$$

Then

$$\underline{y} \sim \text{N}(\underline{v}\,,\ A^T \Sigma A) \qquad and \qquad \underline{v} \sim \text{N}(\alpha A^T \underline{1}\,, \tau^2 \Lambda)$$

Arrange the eigenvalues such that

$$\lambda_1 \geq \cdots \geq \lambda_s > 0 \geq \lambda_{s+1} \geq \cdots \geq \lambda_k$$

and let

$$\underline{y}^* = (y_1, \ldots, y_s)^T \qquad , \qquad \underline{v}^* = (v_1, \ldots, v_s)^T$$

Also, partition $A = (A_1 \vdots A_2)$ and let $\Lambda_1 = \text{diag}(\lambda_1, \ldots, \lambda_s)$. Then

$$\underline{y}^* \sim \text{N}(\underline{v}^*, A_1^T \Sigma A_1) \qquad \text{and} \qquad \underline{v}^* \sim \text{N}(\alpha A_1^T \underline{1}, \tau^2 \Lambda_1)$$

and so $\underline{v}^*$ is estimated by

$$\underline{v}^* = \left(\tau^2 \Lambda_1^{-1} + \left(A_1^T \Sigma A_1\right)^{-1}\right)^{-1} \left(\alpha \tau^{-2} \Lambda_1^{-1} A_1^T \underline{1} + \left(A_1^T\right)^{-1} \underline{y}^*\right)$$

We now take $\lambda_{s+1} = \cdots = \lambda_k = 0$, so that $(v_{s+1}, \ldots, v_k)^T$ is estimated by the

prior mean $\alpha A_2^T \underline{1}$. But $\underline{\mu} = A\underline{v}$ and so this gives us the final estimate of μ as

$$\underline{\hat{\mu}} = \begin{pmatrix} \underline{\hat{v}^*} \\ \cdots \\ \alpha A_2^T \underline{1} \end{pmatrix}$$

5.2 Categorical data and logistic regression

We now apply these ideas to data of the form of the 2 x k contingency table in Table 5.1, where we define the x_i's introduced in Section 5.1 to be the observed proportion of 'successes' in each category (we say that an individual with binary indicator 1 corresponds to a 'successful' outcome). The assumption that these

	1	2	3	⋯	k	
0	$n_1(1-x_1)$	$n_2(1-x_2)$	$n_3(1-x_3)$	⋯	$n_k(1-x_k)$	$\sum n_i(1-x_i)$
1	n_1x_1	n_2x_2	n_3x_3	⋯	n_kx_k	$\sum n_ix_i$
	n_1	n_2	n_3	⋯	n_k	$\sum n_i = N$

Table 5.1: Contingency table arising from considering x_i's as observed proportion of 'successes' amongst n_i individuals.

proportions are normally distributed follows by the Central Limit Theorem and the subsequent results in Section 5.1 follow. In particular, if we assume that the

proportions do not greatly vary between categories we can approximate σ_i^2 by

$$\sigma_i^2 \simeq \frac{\alpha(1-\alpha)}{n_i}. \tag{5.3}$$

We can estimate α by $\hat{\alpha}$ where

$$\hat{\alpha} = \overline{x} = \text{overall proportion} = \frac{\sum n_i x_i}{\sum n_i}$$

Substituting $\hat{\alpha}$ for α in (5.3) and in turn substituting for σ_i^2 in the expression for τ^2 in (5.1) we have

$$\begin{aligned}
\hat{\tau^2} &= s^2 - \frac{k}{\sum \frac{1}{\sigma_i^2}} \\
&= \frac{\sum n_i(x_i-\overline{x})^2}{N} - \frac{k\overline{x}(1-\overline{x})}{N} \\
&= \frac{\overline{x}(1-\overline{x})}{N}\left\{\frac{\sum n_i(x_i-\overline{x})^2}{\overline{x}(1-\overline{x})} - k\right\} \\
&= \frac{\overline{x}(1-\overline{x})}{N}\left\{\frac{\sum (n_i x_i - n_i\overline{x})^2}{n_i\overline{x}(1-\overline{x})} - k\right\}
\end{aligned}$$

But by rearrangement

$$\frac{\sum (n_i x_i - n_i\overline{x})^2}{n_i\overline{x}(1-\overline{x})} = \sum\left[\frac{(n_i x_i - n_i\bar{x})^2}{n_i\bar{x}} + \frac{(n_i(1-x_i) - n_i(1-\bar{x}))^2}{n_i(1-\bar{x})}\right] = \chi^2$$

and hence τ^2 can be estimated by $\hat{\tau^2}$ where

$$\hat{\tau^2} = \frac{\overline{x}(1-\overline{x})}{N}(\chi^2 - k) \tag{5.4}$$

and χ^2 is the usual χ^2 statistic for testing independence in the 2 x k contingency table.

In previous chapters we have introduced the concept of 'shrinkage' as incorporating prior information about data in an analysis to improve predictive accuracy. We can consider τ^2 as a 'shrinkage' parameter because it can be estimated by taking information from the data, in this case the χ^2 statistic, and be used to update the original parameter estimates using (5.2). The χ^2 statistic represents how different the observed proportions in the different categories are. The more different the proportions are, the greater the value of χ^2 and $\hat{\tau^2}$. Hence from (5.2) there will be less shrinkage of the original parameter estimates.

As is the case in the OGRS example, we may have other covariates as well as the categorical variable. Logistic regression is the natural modelling approach for relating the probability of 'success' of an individual to a number of covariates. As a preliminary to this we show that a very similar analysis to the above can be applied to the empirical logits of the number of successes in each category. Define y_i to be the number of successes in the ith category and so $y_i \sim \text{Bin}(n_i, p_i)$ where p_i is the probability of a success at the ith level.

The Empirical Logistic Transform

$$z_i = \log \frac{y_i + \frac{1}{2}}{n_i - y_i + \frac{1}{2}} \tag{5.5}$$

is simply the logistic transform of the observed proportion, adjusted by a constant so that finite values are obtained when $y_i = 0$ or $y_i = n_i$.

Define

$$\mathrm{E}[z_i] \simeq \log \frac{p_i}{(1-p_i)} = \mu_i^*$$

and

$$\mathrm{Var}[z_i] \simeq \frac{1}{n_i p_i (1-p_i)} = \sigma_i^{*2}$$

A well known property of the empirical logistic transformation is that it is approximately normally distributed. So

$$z_i \sim \mathrm{N}(\mu_i^*, \sigma_i^{*2})$$

Again we assume a prior on $\underline{\mu}^*$ as

$$\underline{\mu}^* \sim \mathrm{N}(\alpha^* \underline{1}, \tau^{*2} R).$$

The natural unbiased estimator for α^* is the weighted mean of the z_i's i.e.

$$\widehat{\alpha^*} = \frac{\sum n_i z_i}{\sum n_i} \tag{5.6}$$

We can express α^* and τ^{*2} in terms of α and τ^2 by using the assumption from (5.3) that the proportions of successes do not vary greatly between categories. So assuming that $p_i - \bar{p}$ is small we can use a Taylor series expansion to show that

$$\mu_i^* = \log \frac{p_i}{1-p_i} \simeq \log \frac{\bar{p}}{1-\bar{p}} + \frac{p-\bar{p}}{\bar{p}(1-\bar{p})}.$$

Therefore

$$\begin{aligned}
\alpha^* = \mathrm{E}(\mu_i^*) &= \mathrm{E}\left(\log\frac{p_i}{1-p_i}\right) \\
&\simeq \mathrm{E}\left(\log\frac{\bar{p}}{1-\bar{p}} + \frac{p_i-\bar{p}}{\bar{p}(1-\bar{p})}\right) \\
&\simeq \mathrm{E}\left(\log\frac{\bar{p}}{1-\bar{p}}\right) \\
&\simeq \log\frac{\alpha}{1-\alpha} \qquad (5.7)
\end{aligned}$$

and

$$\begin{aligned}
\tau^{*2} = \mathrm{Var}(\mu_i^*) &= \mathrm{Var}\left(\log\frac{p_i}{1-p_i}\right) \\
&\simeq \mathrm{Var}\left(\log\frac{\bar{p}}{1-\bar{p}} + \frac{p_i-\bar{p}}{\bar{p}(1-\bar{p})}\right) \\
&\simeq \mathrm{Var}\left(\frac{p_i-\bar{p}}{\bar{p}(1-\bar{p})}\right) \\
&\simeq \frac{\mathrm{Var}(p_i)}{\bar{p}^2(1-\bar{p})^2} \\
&\simeq \frac{\tau^2}{\bar{p}^2(1-\bar{p})^2} \qquad (5.8)
\end{aligned}$$

In the context of logistic regression, τ^{*2} can be interpreted as the overall variability in the model for observed proportions. If we estimate τ^2 in (5.8) by the expression in (5.4), this will give an estimate of τ^{*2} in terms of the χ^2 statistic for the 2 x k contingency table.

When analysing logits, the more natural measure of 'goodness of fit' is the *deviance* or log-likelihood ratio statistic. Dobson (1990) shows that the deviance

of the logistic regression model can be expressed as

$$D = 2\sum O\log\frac{O}{E} \tag{5.9}$$

where O denotes the observed frequencies $y_i = n_i x_i$ and $(n_i - y_i) = n_i(1-x_i)$, and E denotes the expected frequencies or fitted values $n_i\bar{x}$ and $n_i(1-\bar{x})$. Substituting these values into (5.9) we have

$$D = 2\left(\sum n_i x_i \log\frac{x_i}{\bar{x}} + \sum n_i(1-x_i)\log\frac{(1-x_i)}{(1-\bar{x})}\right) \tag{5.10}$$

Letting $x_i = p_i + \varepsilon_i$, we have $\bar{x} = \bar{p} + \bar{\varepsilon}$ and so

$$\begin{aligned} D = {} & 2\sum n_i\,(p_i+\varepsilon_i)\,(\log\,(p_i+\varepsilon_i) - \log\,(\bar{p}-\bar{\varepsilon})) \\ & + 2\sum n_i\,(1-p_i-\varepsilon_i)\,(\log\,(1-p_i-\varepsilon_i) - \log\,(1-\bar{p}-\bar{\varepsilon})) \end{aligned}$$

Assuming the n_i's are large, we can use Taylor series expansions

$$\begin{aligned} (p_i+\epsilon_i)\log(p_i+\epsilon_i) &\simeq (p_i+\epsilon_i)\left(\log p_i + \frac{\epsilon_i}{p_i} - \frac{\epsilon_i^2}{2p_i^2} + \cdots\right) \\ &\simeq p_i\log p_i + \epsilon_i(1+\log p_i) + \frac{\epsilon_i^2}{2p_i} + \cdots \\ (p_i+\epsilon_i)\log(\bar{p}+\bar{\epsilon_i}) &\simeq (p_i+\epsilon_i)\left(\log\bar{p} + \frac{\bar{\epsilon}}{\bar{p}} - \frac{\bar{\epsilon}^2}{2\bar{p}^2} + \cdots\right) \\ &\simeq p_i\log\bar{p} + \cdots \\ (1-p_i-\epsilon_i)\log(1-p_i-\epsilon_i) &\simeq (1-p_i)\log(1-p_i) - \\ & \qquad \epsilon_i(1+\log(1-p_i)) + \frac{\epsilon_i^2}{2(1-p_i)^2} + \cdots \\ (1-p_i-\epsilon_i)\log(1-\bar{p}-\bar{\epsilon}) &\simeq (1-p_i)\log(1-\bar{p}) - \cdots \end{aligned}$$

(terms in $\epsilon_i\bar{\epsilon}$ are omitted as when we take expectations they will be zero).

As before we assume that the p_i's don't vary too much so that if we let $p_i = \alpha + \delta_i$

we can use Taylor Expansions for small δ_i. Now

$$\mathrm{E}(\delta_i) = 0 \text{ and } V(\delta_i) = \mathrm{E}(\delta_i^2) = \tau^2.$$

so the above expressions then become

$$\begin{aligned}
(p_i + \epsilon_i)\log(p_i + \epsilon_i) &\simeq (\alpha + \delta_i)\log(\alpha + \delta_i) + \frac{\epsilon_i^2}{2p_i} \\
(p_i + \epsilon_i)\log(\bar{p} + \bar{\epsilon_i}) &\simeq (\alpha + \delta_i)\log(\bar{\alpha} + \bar{\delta}) \\
(1 - p_i - \epsilon_i)\log(1 - p_i - \epsilon_i) &\simeq (1 - \alpha - \delta_i)\log(1 - \alpha - \delta_i) + \frac{\epsilon_i^2}{2(1 - p_i)} \\
(1 - p_i - \epsilon_i)\log(1 - \bar{p} - \bar{\epsilon}) &\simeq (1 - \alpha - \delta_i)\log(1 - \bar{\alpha} + \bar{\delta})
\end{aligned}$$

Using Taylor expansions for small δ_i we have

$$\begin{aligned}
\mathrm{E}\{(p_i + \epsilon_i)\log(p_i + \epsilon_i)\} &\simeq \alpha\log\alpha + \frac{\tau^2}{\alpha} + \frac{1 - \alpha}{2n_i} \\
\mathrm{E}\{(p_i + \epsilon_i)\log(\bar{p} + \bar{\epsilon_i})\} &\simeq \alpha\log\alpha \\
\mathrm{E}\{(1 - p_i - \epsilon_i)\log(1 - p_i - \epsilon_i)\} &\simeq (1 - \alpha)\log(1 - \alpha) + \frac{\tau^2}{(1 - \alpha)} + \frac{\alpha}{2n_i} \\
\mathrm{E}\{(1 - p_i - \epsilon_i)\log(1 - \bar{p} - \bar{\epsilon})\} &\simeq (1 - \alpha)\log(1 - \alpha)
\end{aligned}$$

Substituting these expectations into (5.10) gives

$$\mathrm{E}(D) = \frac{N\tau^2}{\alpha} + k(1 - \alpha) + \frac{N\tau^2}{(1 - \alpha)} + k\alpha = \frac{N\tau^2}{\alpha(1 - \alpha)} + k$$

Rearranging we have

$$\tau^2 = \frac{\alpha(1 - \alpha)}{N}(\mathrm{E}(D) - k). \tag{5.11}$$

Estimating α by $\hat{\alpha}$ and substituting in (5.11) we obtain another expression for $\hat{\tau^2}$. Comparing this $\hat{\tau^2}$ with (5.4), we see that the two expressions are almost identical, except that the χ^2 statistic is replaced by the deviance D. This is intuitively reasonable as it is simply a confirmation of the widely known result that the deviance is approximately χ^2 distributed.

Substituting the expression for τ^2 in (5.11) into (5.8) gives

$$\tau^{*2} = \frac{\alpha(1-\alpha)}{N\bar{p}^2(1-\bar{p})^2}\,(\mathrm{E}(D)-k)$$

Then, expressing α in terms of α^* from (5.7) and estimating α^* by $\widehat{\alpha^*}$ from (5.6), we can estimate by τ^{*2} by $\widehat{\tau^{*2}}$ where

$$\widehat{\tau^{*2}} = \frac{e^{\widehat{\alpha^*}}}{N(1+e^{\widehat{\alpha^*}})^2\,\bar{p}^2\,(1-\bar{p})^2}(D-k)$$

5.3 Covariate model

In the preceding work we have set up the methodology for the simple case involving proportions and extended it to the logistic model. We now look at the general model involving categorical and non-categorical covariates. We wish to express the logit of the response probability of each individual as a linear function of the explanatory variables. We imagine we have n_j individuals in category j,

$j = 1, \ldots, k$, with responses

$$y_{ij} = \left\{ \begin{array}{ll} 1 & \text{'Success'} \\ 0 & \text{'Failure'} \end{array} \right\}$$

and let $p_{ij} = P(Y_{ij} = 1)$ (note that we are now defining y_{ij} to be the response for the ijth individual, not the number of successes as in the previous section) . Each individual has a vector of covariates so our proposed model is of the form

$$\text{logit}(p_{ij}) = \underline{\beta^T}\, \underline{z_{ij}} + \underline{\theta_j}$$

where β^T is the vector of parameters for the model, z_{ij} is the vector of explanatory variables for the ith individual in the jth category and θ_j is the effect due to the jth category. Let $\underline{\theta} = (\theta_1, \ldots, \theta_k)$ and denote the maximum likelihood estimators of $(\underline{\beta}, \underline{\theta})$ as $(\underline{\hat{\beta}}, \underline{\hat{\theta}})$ with

$$\text{Var}\left(\underline{\hat{\beta}}, \underline{\hat{\theta}}\right) = \left(\begin{array}{ll} V_{\beta\beta} & V_{\beta\theta} \\ V_{\theta\beta} & V_{\theta\theta} \end{array} \right)$$

Let the prior on $\underline{\theta}$ be $\underline{\theta} \sim \text{N}(\alpha \underline{1}, \tau^2 R)$ and the prior on $\underline{\beta}$ be vague i.e. $\underline{\beta} \sim \text{N}(0, \infty)$. As before we can find expressions for $\hat{\alpha}$ and $\hat{\tau^2}$ by using the method of moments technique which then can be substituted in (5.2) to estimate the posterior parameters for the model.

Now asymptotically,

$$\underline{\hat{\theta}} \sim \text{N}(\underline{\theta}, \text{V}_{\theta\theta})$$

so

$$\mathrm{E}(\underline{\hat{\theta}}) = \mathrm{E}(\underline{\theta}) = \alpha\underline{1} \quad \text{and} \quad \mathrm{E}\left(\frac{\sum_1^k n_i\theta_i}{\sum_1^k n_i}\right) = \alpha.$$

Therefore we can estimate α by

$$\hat{\alpha} = \frac{\sum_1^k n_i\hat{\theta}_i}{\sum_1^k n_i} \tag{5.12}$$

By the method of moments the global variance of $\underline{\hat{\theta}}$ is

$$\begin{aligned}\mathrm{Var}(\underline{\hat{\theta}}) &= \mathrm{E}\left(\mathrm{Var}(\underline{\hat{\theta}}\,|\,\underline{\theta})\right) + \mathrm{Var}\left(\mathrm{E}(\underline{\hat{\theta}}\,|\,\underline{\theta})\right) \\ &= \mathrm{V}_{\theta\theta} + \tau^2 R\end{aligned}$$

So for any weight matrix W

$$\mathrm{E}\left((\underline{\hat{\theta}} - \alpha\underline{1})^T W(\underline{\hat{\theta}} - \alpha\underline{1})\right) = \mathrm{tr}\left(W\left(V_{\theta\theta} + \tau^2 R\right)\right) = \mathrm{tr}(WV_{\theta\theta}) + \tau^2\mathrm{tr}(WR)$$

Rearranging, we can therefore estimate τ^2 by

$$\hat{\tau^2} = \frac{(\underline{\hat{\theta}} - \hat{\alpha}\underline{1})^T W(\underline{\hat{\theta}} - \hat{\alpha}\underline{1}) - \mathrm{tr}(WV_{\theta\theta})}{\mathrm{tr}(WR)} \tag{5.13}$$

When we are combining estimates together, it is a standard statistical concept to weight with the inverse of the variances. Hence take W to be $\mathrm{V}_{\theta\theta}^{-1}$, then

$$\hat{\tau^2} = \frac{(\underline{\hat{\theta}} - \hat{\alpha}\underline{1})^T V_{\theta\theta}^{-1}(\underline{\hat{\theta}} - \hat{\alpha}\underline{1}) - k}{\mathrm{tr}(V_{\theta\theta}^{-1}R)} \tag{5.14}$$

The first part of the numerator in (5.13) can be interpreted as the part of the deviance of the model explained by the categorical covariate. This is easily calculated from the analysis of variance table output by most statistical packages.

So just as before with the simple case, $\hat{\tau}^2$ is related to how well the model fits the data.

In practice, in defining $\hat{\tau}^2$ we truncate to zero as it is perfectly possible that $\hat{\tau}^2$ could be negative. (A negative value of τ^2 would conflict with our interpretation of τ^2 as the overall variation in the data.)

Using (5.2), the posterior for the general model are

$$\mathrm{E}\left[\left(\begin{array}{c}\beta\\ \underline{\theta}\end{array}\right)\middle|\left(\begin{array}{c}\hat{\beta}\\ \underline{\hat{\theta}}\end{array}\right)\right] =$$

$$\left[\left(\begin{array}{cc} V_{\beta\beta} & V_{\beta\theta}\\ V_{\theta\beta} & V_{\theta\theta}\end{array}\right)^{-1} + \left(\begin{array}{cc} 0 & 0\\ 0 & \tau^{-2}R^{-1}\end{array}\right)\right]^{-1}\left[\left(\begin{array}{cc} V_{\beta\beta} & V_{\beta\theta}\\ V_{\theta\beta} & V_{\theta\theta}\end{array}\right)^{-1}\left(\begin{array}{c}\hat{\beta}\\ \underline{\hat{\theta}}\end{array}\right) + \left(\begin{array}{c} 0\\ \alpha\tau^{-2}R^{-1}\end{array}\right)\right] \tag{5.15}$$

and these can be estimated by substituting in the estimates of α and τ^2 in (5.12) and (5.13) respectively. Note that the parameter estimates of the non-categorical covariates are not changed by the methodology described above. This is of course, correct because we are only focusing on the relationships between categories given by R.

5.4 Example

To gain an understanding of the framework presented above we present an example based upon a set of data with its origins in credit defaulting (data supplied

by a colleague). The data for an individual consists of occupational name and a binary indicator denoting whether they have defaulted on credit payments. The original size of the data was $n = 2792$ individuals with 71 different occupational categories covering a diverse range of employment such as Actor/Actress, Footballer, Chemist, Architect, Company Director,Pilot, Military Employees, Nurse, Student, Government Employee, Engineer and Teacher. The largest two of the occupational categories concerned no occupational information (No Information, (845)) and those in the military (Military Employee (875)). These categories have been omitted from the analysis because of their size and in the case of the No Information category, it would be unwise to include such a large group of individuals of whom we have no clue to occupational status.

The data now consists of 69 occupational categories, some of which have a very small number of individuals. Later, we will be performing a series of simulations which will mean sampling from the data. It is quite likely that categories with a small number of individuals could be missed out altogether in the sampling procedures. For this reason we have decided to group the remaining 1072 individuals into 18 wide-ranging occupational categories which are presented in Table 5.2 along with the number and proportion of defaulters in that category.

A logistic regression was calculated using the following model

$$\log \frac{p_i}{1 - p_i} = \theta_i$$

Occupation	Number	Prop	Occupation	Number	Prop
Professional/Entertainment	153	0.0196	Business Management	13	0.1538
Manual Work/Student	28	0.25	Housewife	29	0.1034
Academic/Skilled	30	0.2	Architect/Engineer	37	0.189
Shopkeeper/Sales	177	0.271	Public Accountant	15	0.3333
Director	93	0.1613	General Employee	80	0.113
Company Employee	182	0.0714	Businessman/woman	23	0.217
Health/Teaching	33	0.182	Government Employee	11	0.182
Executive	129	0.248	Executive Manager	12	0.25
Doctor	12	0.1667	Self-Employed	15	0.2667

Table 5.2: Description of the occupations in the credit defaulting data set

where p_i is the probability of defaulting for the ith category and θ_i is the effect due to the ith category. The fitted probabilities will just be the proportion of defaulters in the ith category as the model is saturated - there are the same number of parameters as there are observations. From (5.15) we can calculate the posterior estimates of the parameters as

$$\mathrm{E}(\underline{\theta} \mid \underline{\hat{\theta}}) = \left[V_{\theta\theta}^{-1} + \tau^{-2}R^{-1}\right]^{-1}\left[V_{\theta\theta}^{-1}\underline{\hat{\theta}} + \alpha\tau^{-2}R^{-1}\right] \tag{5.16}$$

where α and τ^2 can be estimated by

$$\hat{\alpha} = \frac{\sum_1^k n_i\hat{\theta}_i}{\sum_1^k n_i} \text{ and } \hat{\tau^2} = \frac{(\underline{\hat{\theta}} - \hat{\alpha}\underline{1})^T V_{\theta\theta}^{-1}(\underline{\hat{\theta}} - \hat{\alpha}\underline{1}) - k}{\mathrm{tr}(V_{\theta\theta}^{-1}R)}$$

At this point we have to make a choice for R, the correlation matrix. We have no

information supplied with the data to enable us to form R, so for the purposes of this example we shall take R to be the identity matrix which we shall call R_I i.e. there are no relationships between the defaulting rates in occupational categories. For this particular model (i.e one categorical covariate with no intercept) we can deduce from the expression for $\hat{\tau^2}$ that $\hat{\tau^2}$ will be the same for all different R matrices. This reason for this is $\hat{\tau^2}$ is only affected by R in the denominator, and because $V_{\theta\theta}$ is diagonal (for this model $\text{Var}(\theta_i) \simeq (n_i p_i(1-p_i))^{-1}$ and $\text{Cov}(\theta_i, \theta_j) = 0$) the trace term will effectively just sum the reciprocals of the variances i.e.

$$\text{tr}(V_{\theta\theta}^{-1} R) \simeq \sum_i n_i p_i (1 - p_i).$$

Figure 5.1 shows a graphical representation of the estimates of $\text{E}(\theta_i \mid \hat{\theta}_i)$ obtained by substituting $\hat{\alpha}$ and $\hat{\tau^2}$ for α and τ^2 in (5.16). These estimates are the points plotted on the right side of the plot whilst the original estimates from the model are plotted on the left (as illustrated below, the model estimates correspond to the theoretical case where $\tau^2 = \hat{\tau^2} = \infty$). By connecting these points with lines we can illustrate the phenomenon of the model estimates being shrunk towards the mean $\hat{\alpha}$ (this idea was briefly described in Chapter 1). The lines also express the idea that shrinkage is not uniform - some parameter estimates shrink faster than others because of differing sample sizes of the categories.

We can investigate the limits of the posterior estimates or *shrunk* estimates (as we shall now call them) by studying the posterior expectation in (5.16).

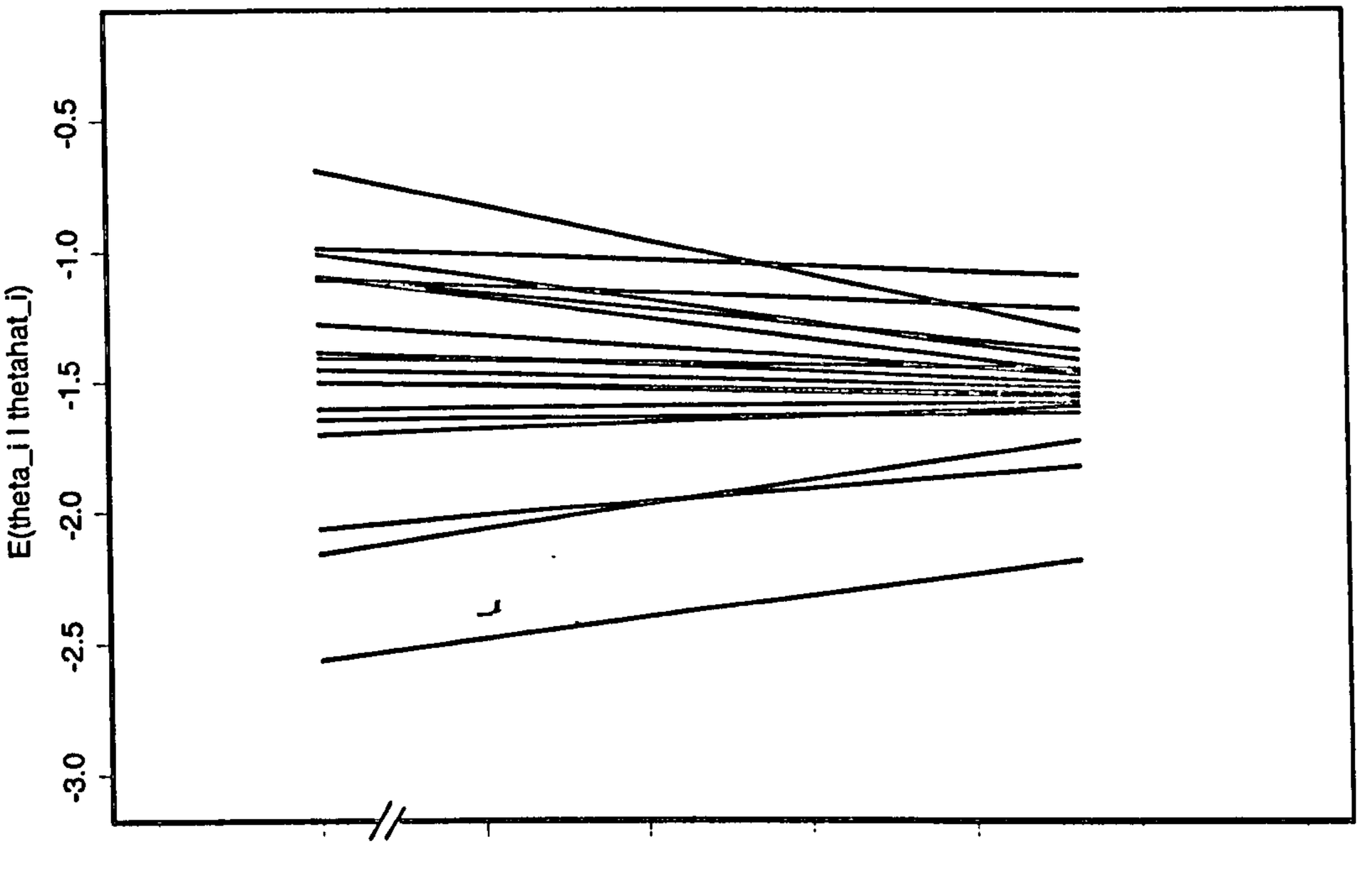

Figure 5.1: Shrinkage effect of $\tau^2 = \hat{\tau^2}$ for the credit defaulting data (R is the identity matrix)

As $\tau^2 \to 0$, $\tau^{-2} \to \infty$

$$\mathrm{E}(\theta \,|\, \hat{\theta}) \to \left[\tau^{-2}R^{-1}\right]^{-1}\left[\alpha\tau^{-2}R^{-1}\right] = \alpha$$

and as $\tau^2 \to \infty$, $\tau^{-2} \to 0$

$$\mathrm{E}(\theta|\hat{\theta}) \to \left[V_{\theta\theta}^{-1}\right]^{-1}\left[V_{\theta\theta}^{-1}\hat{\theta}\right] = \hat{\theta}$$

We can represent this graphically by holding $\alpha = \hat{\alpha}$ fixed and varying τ^2. Figure

5.2 demonstrates how as τ^2 approaches 0, the shrunk estimates converge to $\hat{\alpha}$, illustrating the shrinkage to the mean phenomenon. It is interesting to note the rapid convergence of those posterior estimates that are furthest away from $\hat{\alpha}$ as τ^2 gets closer to 0.

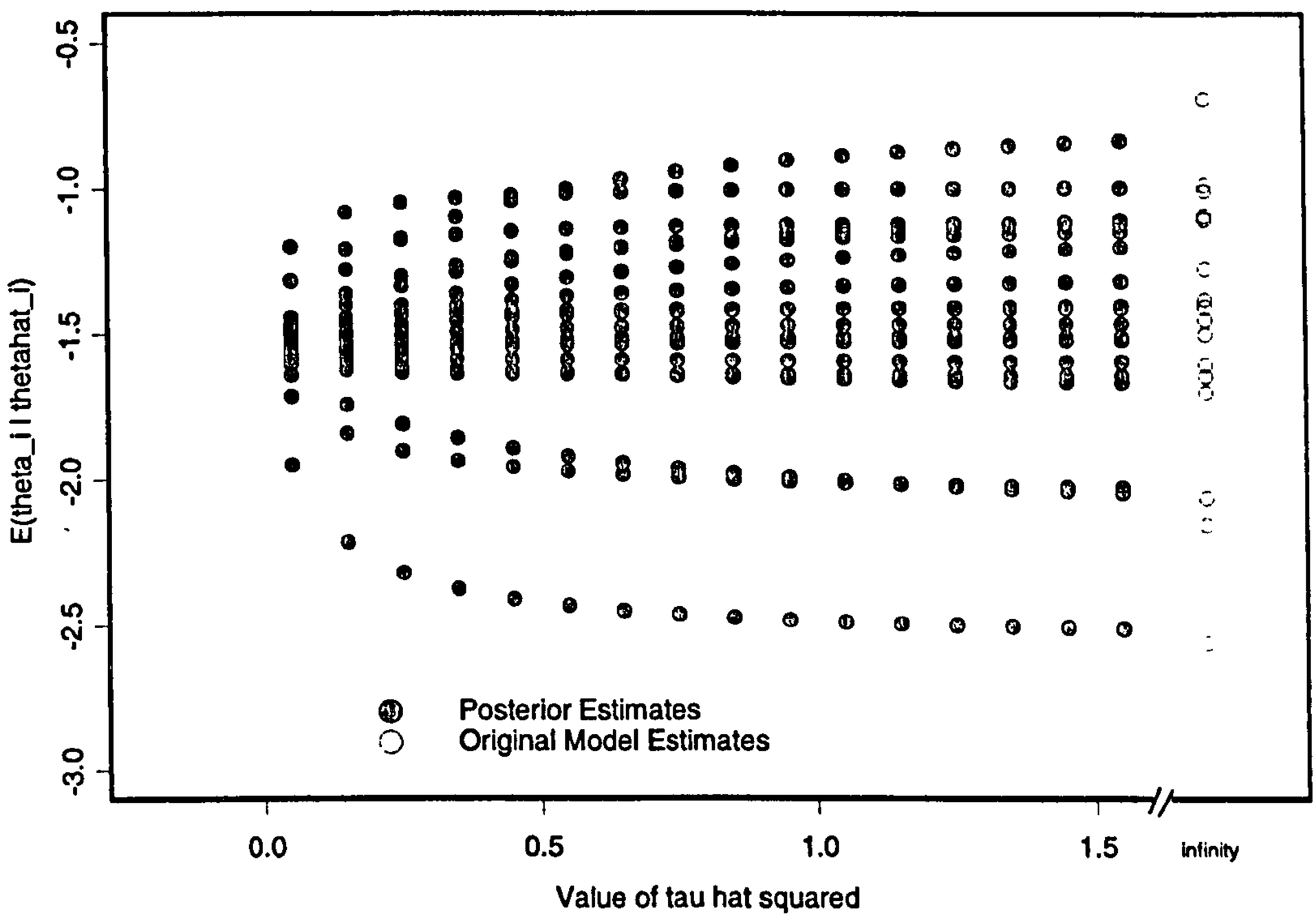

Figure 5.2: Values of the shrunk estimates for different values of τ^2 holding α fixed

The plot on the left of Fig 5.3 graphs the fitted probabilities from the logistic regression against the actual proportion of defaulters within a category (with

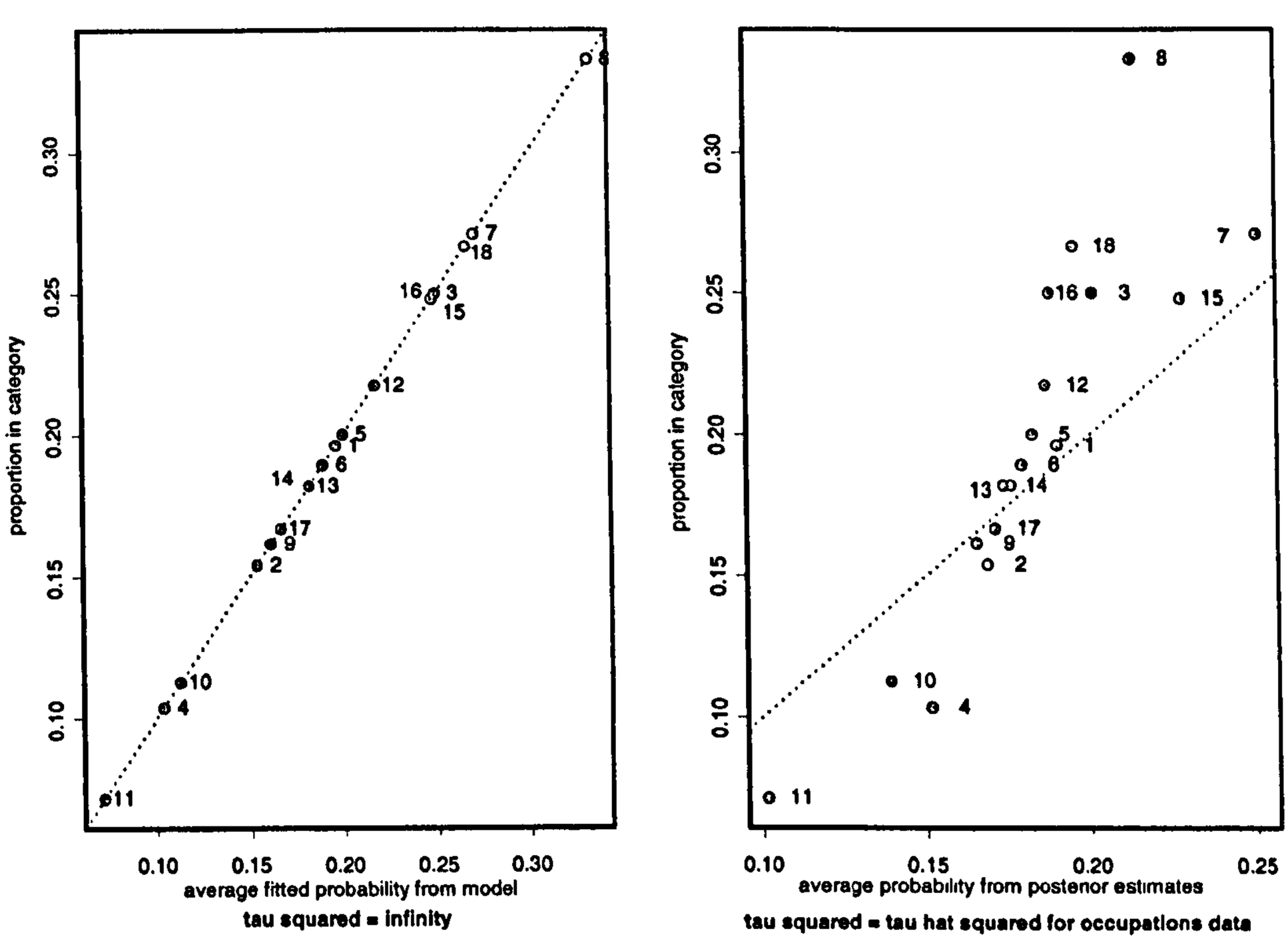

Figure 5.3: Original model fitted probabilities and shrunk estimates against the actual proportions in the data

the line $y = x$ drawn for reference). As we would expect points lie on the line $y = x$ because the model is saturated and the fitted probabilities just reflect the proportion of defaulters in each category. The plot on the right shows an analogous graph of the inverse logit of the shrunk estimates against the actual proportions of defaulters in each category. Again we can see evidence of shrinkage here, as the inverse logits of the shrunk estimates are contained within limits from

0.1 to 0.25, whilst the fitted probabilities from the model are contained within limits from approximately 0.05 to 0.35. This fits in with our idea of shrinkage as the posterior estimates are being concentrated into a tighter 'bandwidth' as they are shrunk towards the mean.

We now to turn our attention to the impact of correlations on the occupational groupings given above in Table 5.2. It is feasible that defaulting habits might be closely related in a number of occupations. For example, we may consider that one of the main reasons for defaulting on a credit payment may be a low salary or lack of available funds for a particular month due to some other payments. Then, individuals with occupations such as Doctors and Executives might not default as often as others because we would perceive them to be well salaried, hence inducing a correlation between the defaulting rates for Doctors and Executives. We take this idea as a basis for studying correlations between defaulting rates. We group the 18 categories in Table 5.2 into 4 purely subjective groups along the lines of perceived occupational status. The groupings are as follows,

- **Group 1** Academic/Skilled, Director, Executive, Doctor, Executive Manager

- **Group 2** Government Employee, Professional/Entertainment, Architect/Engineer, Health/Teaching, Self-Employed,Business Management, Businessman/woman, Accountant

- **Group 3** Shopkeeper/Sales, General Employee, Company Employee

- **Group 4** Housewife, Manual Work/Student

the idea being all of the credit defaulting rates within a group of categories should be correlated with all of the other categories in that group but not correlated with the defaulting rates in any of the other groups. This has the effect of giving R a 'block' appearance with clusters of correlations corresponding to the groups above. Of course, we could just assign various correlations to the original 18 categories but we would have no guarantee whether this matrix would be positive definite. If R was not positive definite we could then use the principal component method described in Section 5.1, but for the purposes of this analysis we shall concern ourselves with the situation where the original R is positive definite.

Once we have calculated the shrunk estimates for a certain correlation matrix R, how will we compare them with corresponding estimates for other correlations? We could represent them graphically, as in Figure 5.1 but graphing more than one set of shrunk estimates corresponding to different R matrices would produce a plot difficult to interpret. Instead we choose to study the area under the ROC curve.

For any correlation matrix R and for any grouping of categories, we will assume that the defaulting rate for an offence category will have correlation ρ with all the other defaulting rates in that group. So any particular R matrix

will either have its (i, j)th entry equal to ρ or zero. Therefore, for any R we can calculate the shrunk estimates and construct the ROC curve from the shrunk estimates and the binary indicator of default. We can then plot the area under the ROC curve against the correlation ρ. By varying ρ we can calculate ROC areas for different R matrices and obtain a plot to show how the discriminatory power of the shrunk estimates depends on the correlation between defaulting rates in groups of categories.

Figure 5.4 shows the **retrospective** ROC areas calculated from the shrunk estimates by varying the correlation ρ between the occupational groups for the credit defaulting data. There is a gradual downward trend as the correlation increases until we have a sharp fall at a correlation of approximately 0.8. We would expect this decreasing trend for a large data set, because when the correlation between defaulting rates within groups is large it is analogous to interpreting the data set as consisting of only the four occupational status groups. For a data set of this size, eighteen occupational groups will be much better for discriminatory purposes than four, hence the reduction in ROC area and discriminatory power for large ρ.

As has been mentioned before, a retrospective ROC assessment will overestimate the discriminatory power of the shrunk estimates. We wish to know how well the shrunk estimates perform in the future (prospectively), for which we

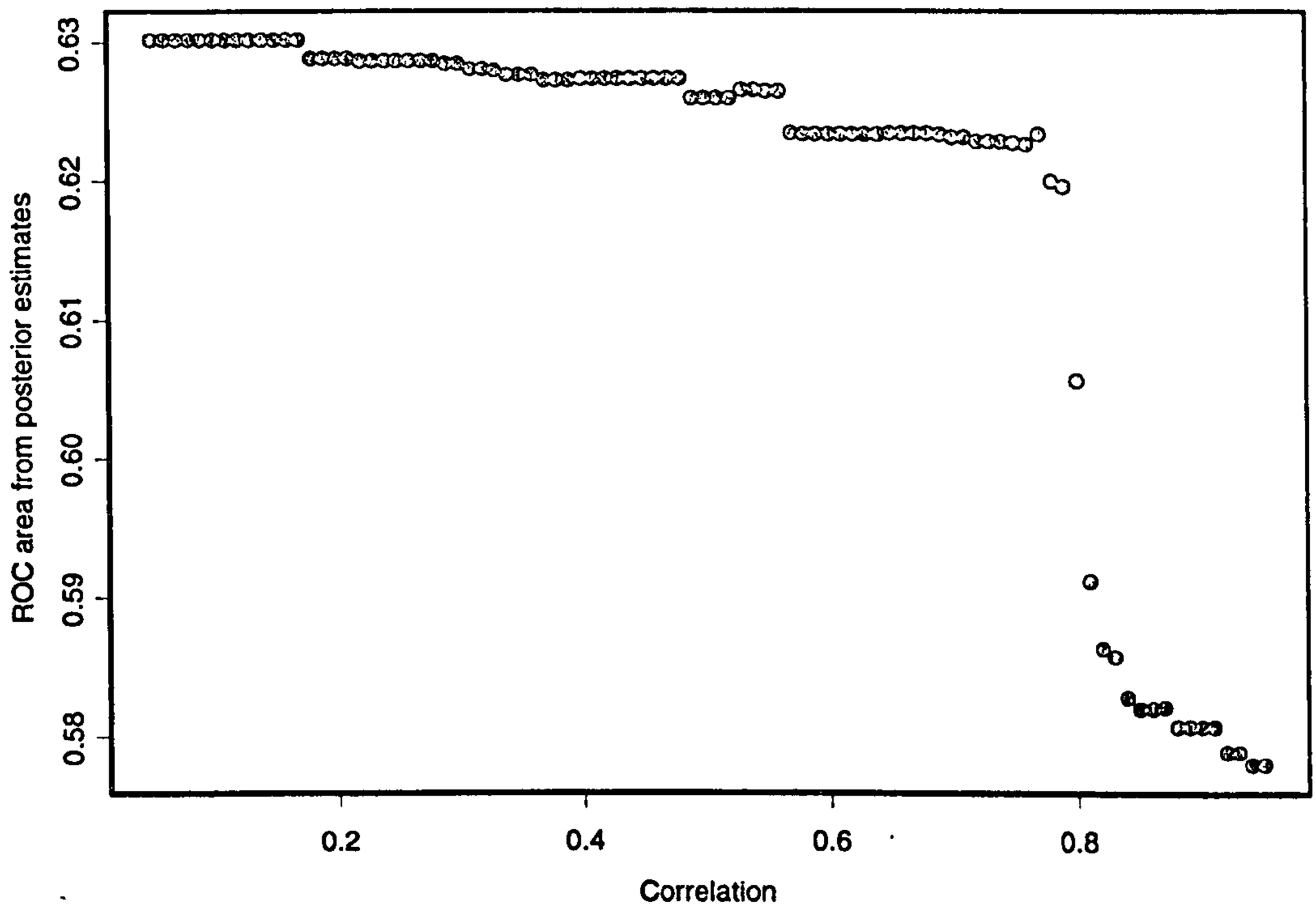

Figure 5.4: The effect on the ROC areas for the shrunk estimates of varying the correlations within occupational groups

would require a new independent set of data. Typically, such data is not available and instead we try to mimic a prospective assessment through a simulation study, which we shall introduce next.

5.4.1 Simulation study

To assess whether the pattern illustrated in Figure 5.4 (especially the sudden loss of discriminatory power for correlations of 0.8 and greater) is an accurate portrayal of the effect of varying the correlation between defaulting rates we perform a simulation study. Generally, we will use a bootstrap procedure to resample but introduce a refinement due to the nature of the data. We present two bootstrap simulations, one parametric and one non-parametric.

As mentioned in Chapter 4 the non-parametric bootstrap simply resamples the data randomly with replacement. Unfortunately, because of the nature of this data set if we randomly resample we could miss out an occupational category completely, or end up with a category containing all defaulters or non-defaulters, hence giving infinite logits when fitting the logistic regression. A solution to the first problem is to simply perform a check on whether the resampled data is 'usable' in the sense that all categories have been sampled and no categories exist with all defaulters or non-defaulters.

From the data set described above, 2000 bootstrap resamples were created. After passing these through a filter, 1144 were 'usable' in the sense described above. We use 1000 of these 'usable' resamples in the following algorithm:-

1. Calculate a logistic regression on the resampled data set with default status y_i^* producing a regression vector β^*.

2. Calculate the **retrospective** ROC area using y_i^* and β^*.

3. Calculate the **prospective** ROC area by using the defaulting vector y_i from the original data set and β^*.

4. By varying the correlation ρ, calculate sets of shrunk estimates for the resampled data.

5. For each set of shrunk estimates indexed by ρ calculate the **retrospective** 'shrunk' ROC area using y_i^*.

6. For each set of shrunk estimates indexed by ρ calculate the **prospective** 'shrunk' ROC area using y_i.

7. After this procedure has been carried out for all 1000 resamples we can produce average retrospective and prospective ROC areas from the model and vectors of average retrospective and prospective 'shrunk' ROC areas corresponding to the range of correlations used in step 4.

(Since, for this model the fitted probabilities are just the proportions of defaulters in the categories, step 1. in the algorithm is not strictly needed but is included for completeness). The simulation results can be seen in Figure 5.5. On this plot, the plus and cross signs denote the retrospective (calculated using y_i^*) and prospective (calculated using y_i) ROC areas respectively from the model. The black dots represent the retrospective shrunk areas for different ρ calculated by

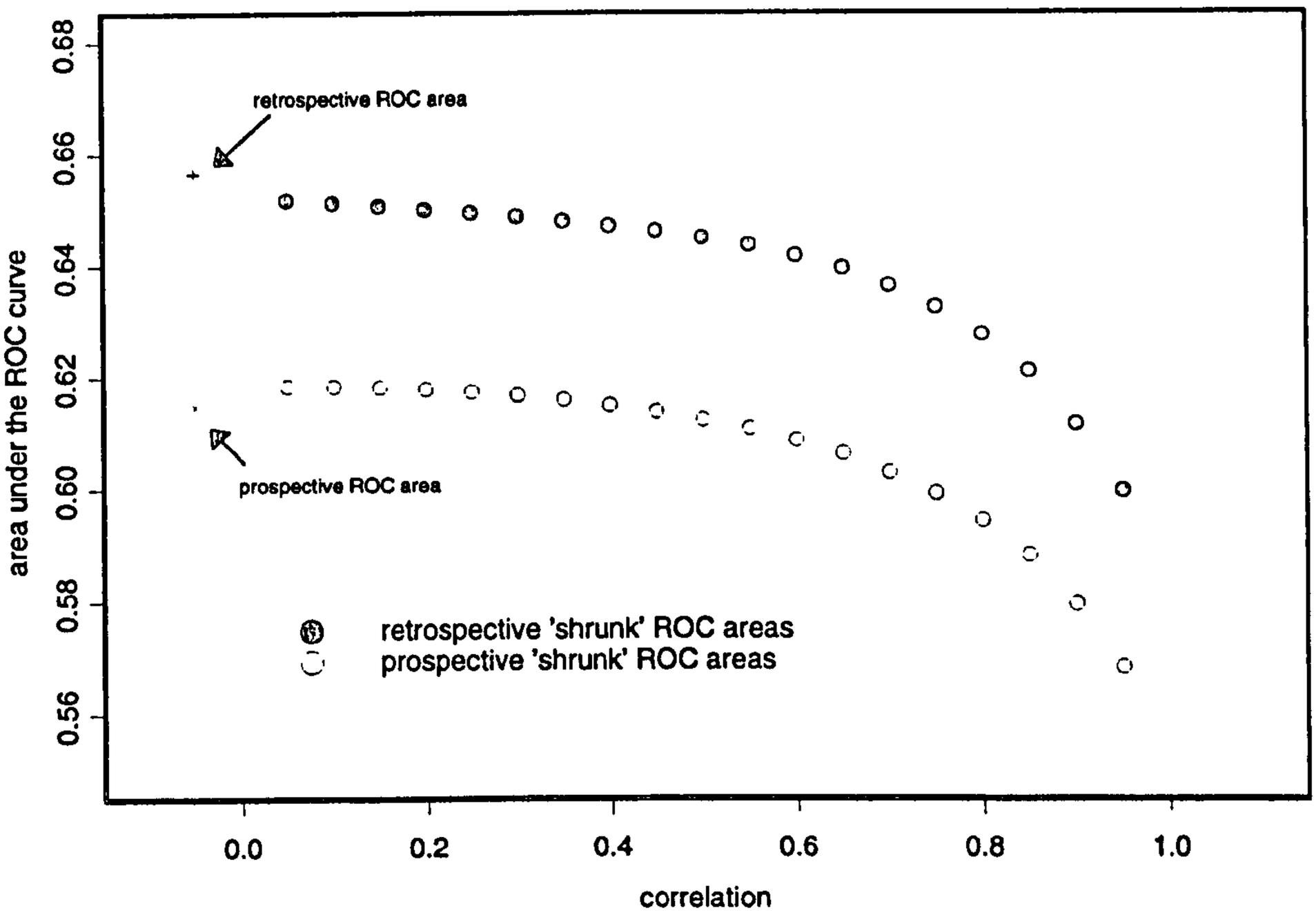

Figure 5.5: Retrospective and prospective ROC areas for the model and the shrunk estimates, by varying correlations within the four occupational groups

step 5 in the algorithm, whilst the white dots represent the prospective shrunk areas calculated by step 6. No gain in discriminatory power is achieved retrospectively by the shrunk estimates but for low values of ρ, the prospective shrunk ROC areas are greater than the prospective ROC area from the model. As we are primarily interested in how a score acts prospectively, Figure 5.5 illustrates the shrinkage procedure by improving discriminatory power of a score by the pro-

cess of including prior information. It is also interesting to note that as ρ nears 1 the performance of the shrunk estimates (both retrospective and prospective) decreases. As was mentioned above, we would expect this because as the correlation increases, we are mimicking the situation where there are only 4 categories, namely the 4 groupings of occupation described above. For a data set with $n = 1072$ individuals, it is sensible to suggest that we are losing information when grouping into 4 categories instead of 18 and hence we lose discriminatory power.

Another approach to resampling the data which addresses both the missing category and infinite logit problems is to apply a continuity correction to the observed proportion of defaulters in a category after resampling. So, if after resampling we have category c with x_c defaulters out of a total n_c individuals, let the proportion of defaulters be $\tilde{p}_c$ where

$$\tilde{p}_c = \frac{x_c + \frac{1}{2}}{n_c + 1}$$

If a category is not sampled at all then $n_c = x_c = 0$ and $\tilde{p}_c = 0.5$. If we have all defaulters or non-defaulters in a category then $x_c = 0$ or n_c and $\tilde{p}_c = \frac{1}{2(n_c+1)}$ or $\frac{n_c+\frac{1}{2}}{n_c+1}$ respectively. We can also express the variance of the logit of $\tilde{p}_c$ as

$$\text{Var}\left[\log\left(\frac{\tilde{p}_c}{1-\tilde{p}_c}\right)\right] \simeq \frac{1}{x_c + \frac{1}{2}} + \frac{1}{n_c - x_c + \frac{1}{2}}$$

The advantage of using this method is that we do not have to worry about the nature of the resampled data, as we did in the non-parametric bootstrap when

we 'filtered' the data before use. The simulation procedure for the data involving the continuity correction is quite similar to the procedure described above for the non-parametric bootstrap, except for a few minor technical details in the calculation of the ROC area which are omitted here.

Figure 5.6 shows the result of the simulation incorporating the continuity

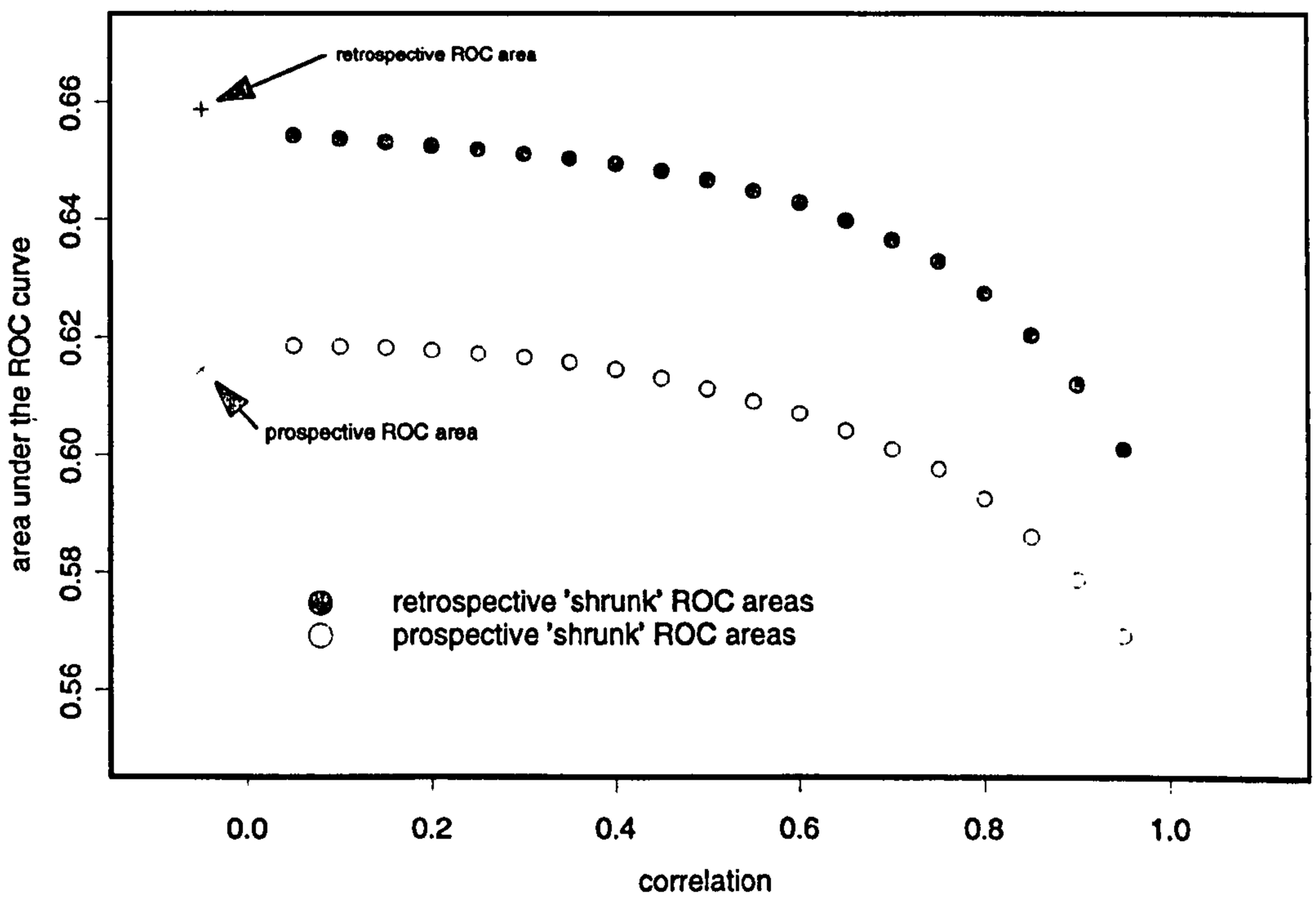

Figure 5.6: Retrospective and prospective ROC areas for the model and the shrunk estimates, by varying correlations within the four occupational groups (with continuity correction applied)

correction applied to 1000 non-parametric bootstrap resamples. The plot is practically identical to Figure 5.5 for resampling without the continuity correction and the discussion of the increased discriminatory power in the prospective 'shrunk' estimates follows from before.

The effect of the continuity correction is seen in Table 5.3 where ROC areas were calculated for a set of 50 bootstrap resamples with and without the continuity correction. As is evident from the table, using the continuity correction has little effect on the calculation of the ROC areas and because of it's usefulness in dealing with the resampling problems outlined above, we shall employ the continuity correction method from now on.

	Retrospective	Prospective
With c.c.	0.5567($\pm$0.01)	0.5055($\pm$0.01)
Without c.c	0.5571($\pm$0.02)	0.5054($\pm$0.02)

Table 5.3: Effect of continuity correction applied to the resampled data

The parametric bootstrap procedure we shall use here involves an analogous method to that used in Chapter 4. For each category c with proportion of defaulters p_c and number of individuals n_c, we simulate a new binary vector from a binomial distribution with parameters n_c and p_c. Combining these vectors gives a new vector of default status, y' say. As the proportion of defaulters in the data as

a whole is low (0.228), if we were to simulate y' from the original binary indicator of default status we would obtain many categories containing all non-defaulters, again creating infinite logits in the logistic regression.

To solve this problem we form a new set of data of 398 individuals by taking the 199 defaulters from the original data and randomly sample 199 non-defaulters from the remaining 873. Call this new set of data D_{HALF}. Therefore the overall proportion of defaulters is now 0.5. Although this procedure alters the distributional characteristics of the data, the ROC areas should on average remain comparable with the original data because the ROC curve and area depend on the distribution of the 'score' (shrunk values etc.) rather than the proportion of defaulters in the data. So we can change the proportion of defaulters in the data (as we have done here) and still compare the results with those obtained on the original data. We now simulate the new default vector y' according to the procedure outlined above, apply the continuity correction (the problem of unsampled categories and infinite logits still remains) and calculate the ROC areas.

The results of the above approach can be seen in Fig 5.7. These are much the same as the non-parametric resampling simulations with and without the continuity correction, except for an increase in ROC area which is likely to be attributed due to the structure of the new data D_{HALF}.

So far we have seen that as the correlation within groups approaches 1 there

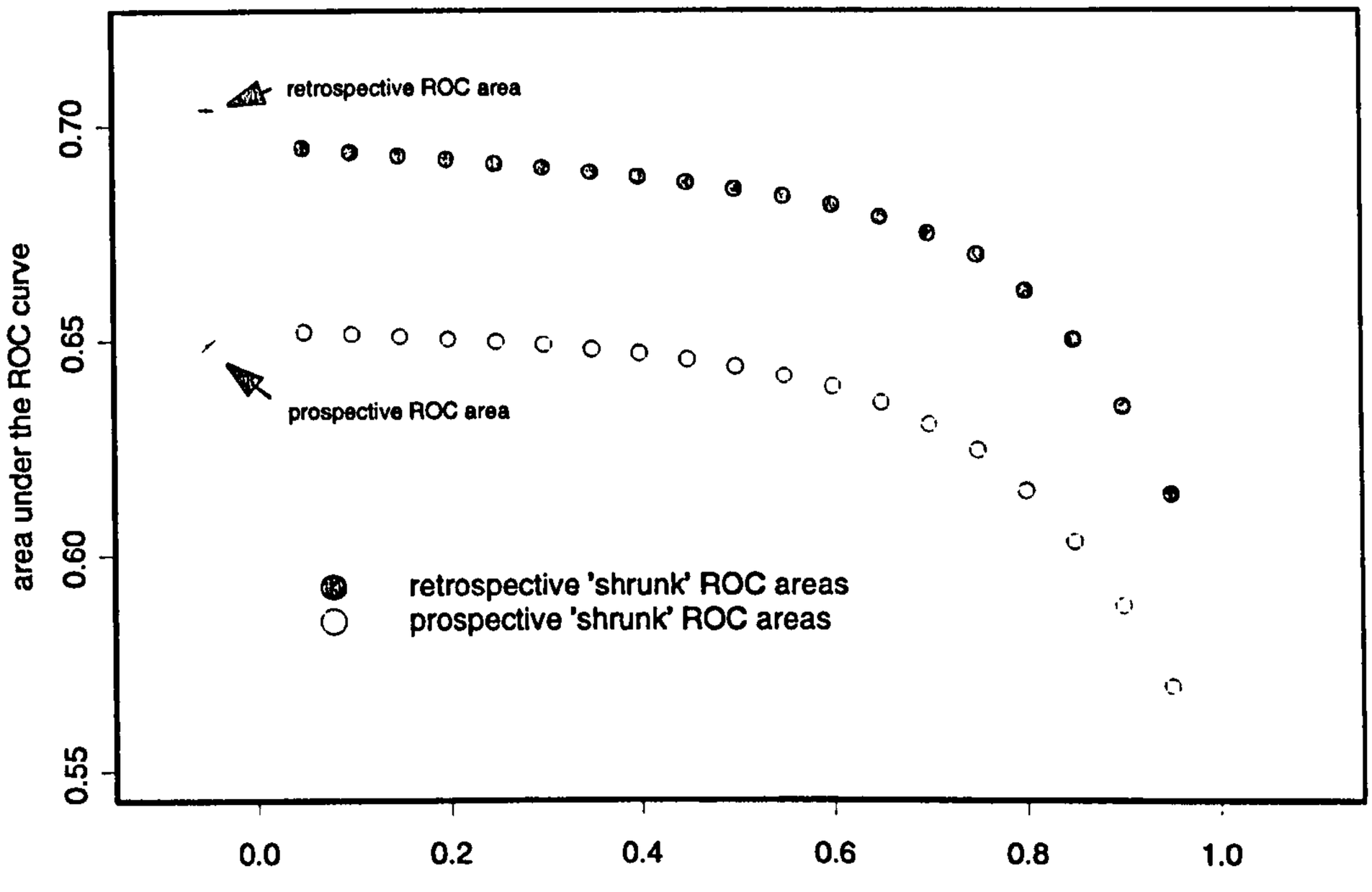

Figure 5.7: Retrospective and prospective ROC areas for the model and the shrunk estimates - parametric bootstrap resampling with continuity correction

is a decline in the discriminatory ability of the score as we lose information due to grouping into a smaller number of categories. It would be interesting to see if the reverse is true i.e. for a small data set with a large number of categories, grouping the data together into a number of smaller categories might improve the discriminatory performance of the shrunk estimates.

To test this assumption, we take at random 50 defaulters and 50 non-defaulters

from the original data to form the data set D_{100}. From D_{100} we use the parametric bootstrap to simulate 500 new data sets. Of course, because there are still 18 categories and now only 100 individuals in the data there is a high chance now that we will experience the sampling problems described above. Therefore we use the continuity correction again and calculate the ROC areas as before. The results can be seen in Figure 5.8.

There is now a pronounced change as the correlations tend to 1. The discriminatory power of the scores still decrease but nowhere near as markedly with the previous simulations. As mentioned above, with such a small data set we might gain some predictive accuracy from grouping together a number of categories. The reason why we do not seem to gain any discriminatory power could be because of the poor predictive accuracy of the shrunk estimates (see following discussion). It is also of interest to note that the retrospective and prospective ROC areas are much greater than their respective shrunk areas. A reason for this could be that because of the continuity correction a large number of categories will have a defaulting proportion of 0.5. If so, then in obtaining the shrunk estimates we are likely to be shrinking to a mean around 0.5. With many of the estimates near 0.5 we cannot discriminate between categories with an actually high or low proportion of defaulters, consequently the ROC area itself will be quite poor and near 0.5.

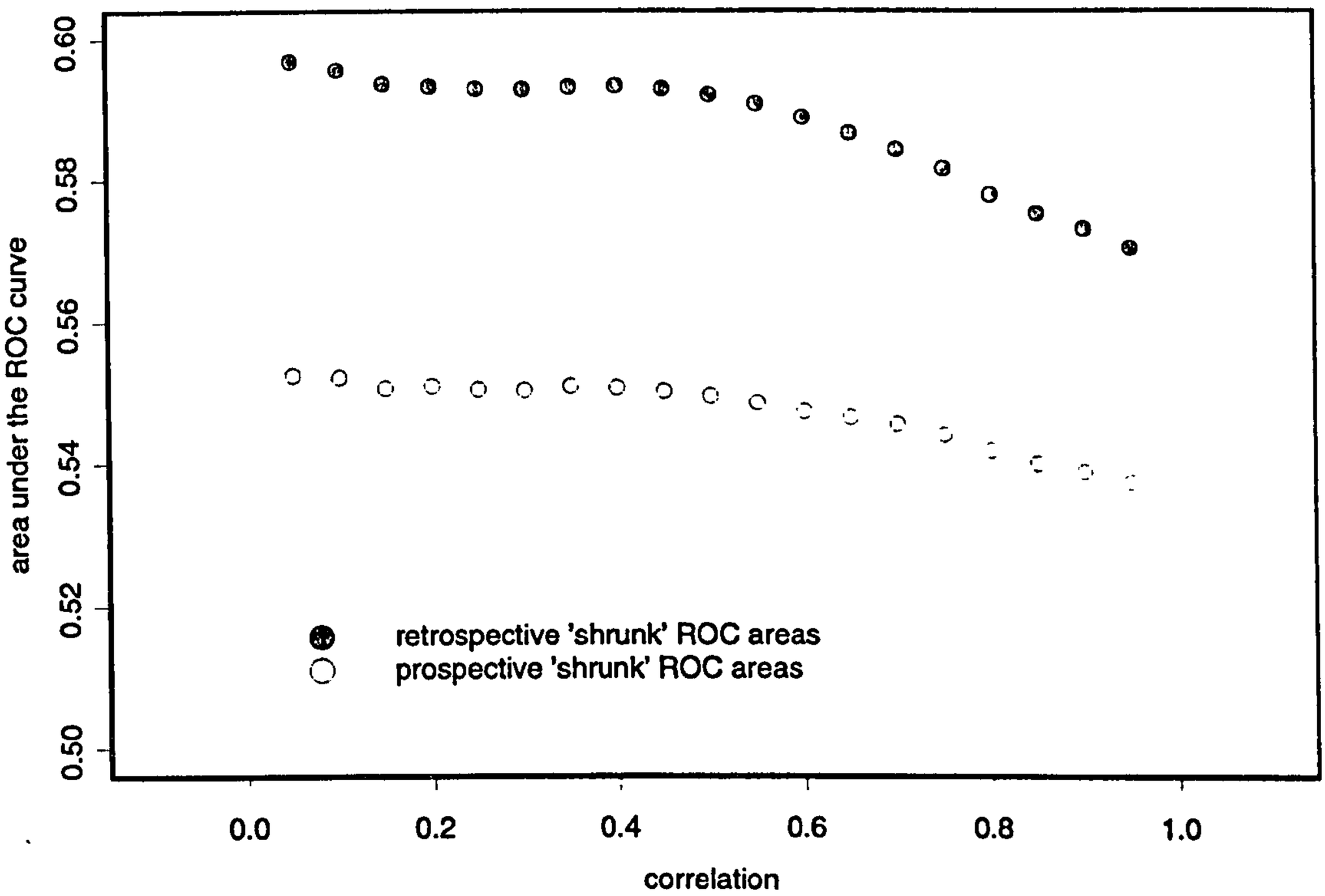

Figure 5.8: Results of the analysis applied to the data set D_{100} in order to test the properties of the occupational groupings

All through this simulation we have been using the occupational status groupings defined earlier in this chapter. These of course were subjective choices but what would happen if these were in fact not a good choice of grouping? To test the effect of the occupational status groupings we can order the categories according to the proportion of defaulters within them. Then, by taking the first five

of these ordered categories and grouping them together, taking the next five and grouping them together and so on we can artificially create the situation where our subjective information has informed us of the best four groupings available i.e all the categories with a high or low proportion of defaulters are grouped together. In the case of the last example with the data set D_{100}, as the correlation between defaulting rates within groups approaches 1 we should hope to see some increase in the ROC area as we have created the best four occupational groups for discriminatory purposes. Combining the categories into four groups according to their proportions in D_{100} and repeating the simulation procedure again with the continuity correction gives Figure 5.9.

The prospective shrunk areas in Figure 5.9 are quite similar to those in Figure 5.8 but show evidence of not decreasing as quickly over the range of correlations. This slight change is probably due to categories being grouped in the best possible way. We may well have expected some evidence of a significant increase in ROC area as correlations tend to 1 but it is worth remembering that as described above, the poor discrimination given by the shrunk estimates close to 0.5 may be having a negative effect. Also, the 'optimal' grouping according to the data, on closer inspection is close to the grouping that we originally gave earlier in this chapter. As mentioned above, it could well be the case that the continuity correction is 'cancelling out' any discriminatory benefit the optimal grouping may have had.

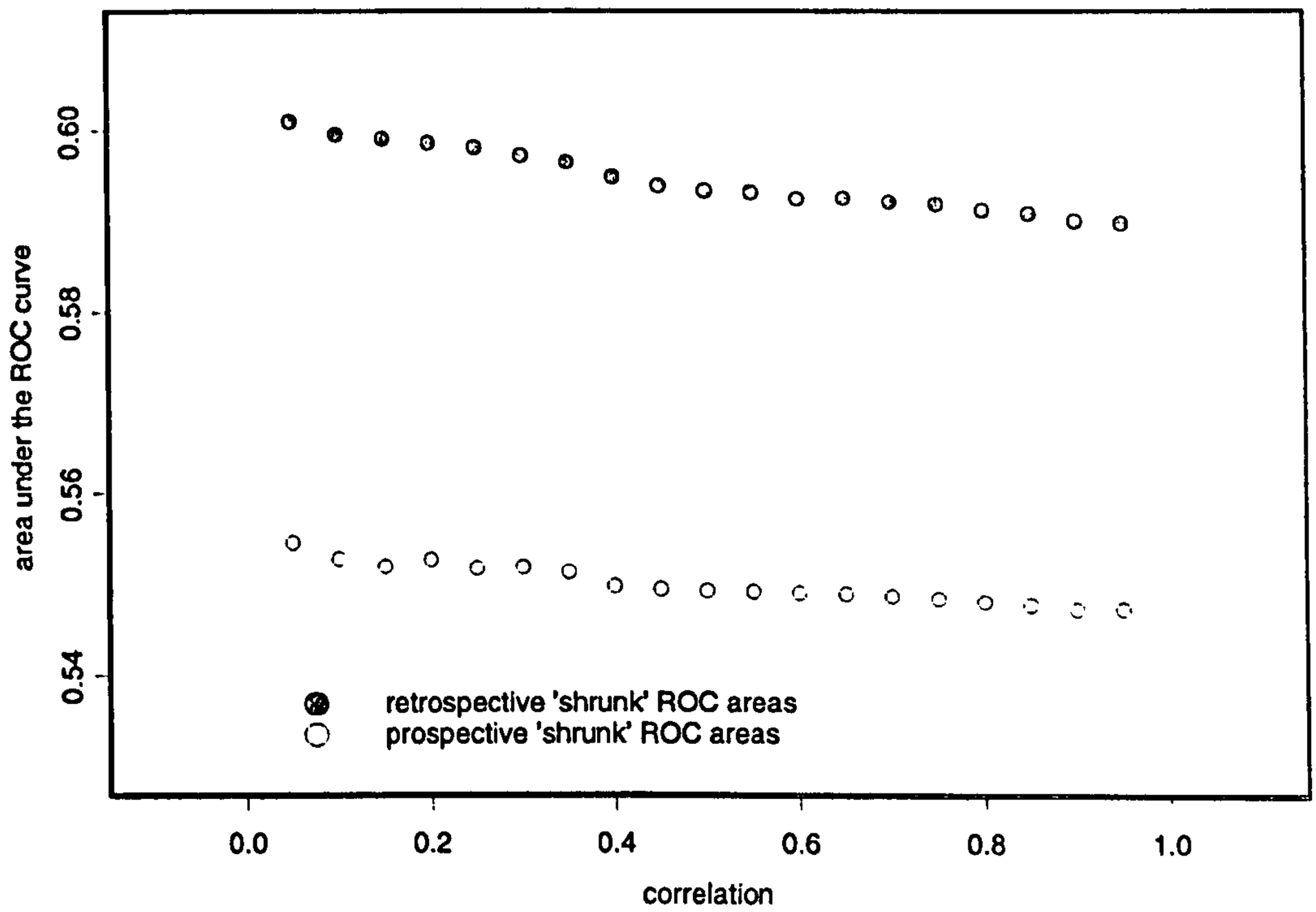

Figure 5.9: Results of the analysis applied to the data set D_{100} with occupations grouped according to their defaulting proportions in D_{100}

5.4.2 Some comments

The credit defaulting example and the simulations that followed it, although being of a fairly simple nature has shown some of the properties of the shrinkage method developed in this chapter. We take account of the prior information, specifically R, the correlation matrix and τ^2, the contribution of a categorical

covariate to a model to calculate 'shrunk' estimates of the $\hat{\theta}_i$'s, the parameter estimates from a logistic regression. From Figure 5.1 we have evidence of the shrinkage to the mean phenomenon and Figure 5.2 verifies that as $\tau^2 \rightarrow 0$, i.e. the 'contribution' of the categorical covariate to the model decreases the shrunk estimates tend to the mean $\hat{\alpha}$.

It has been mentioned before that the study of shrinkage estimators have been concentrated in the context of the *calibration* problem. Figures 5.5 and 5.6 seem to add more credibility to the idea that we may be able to form 'shrunk' scores that gives better discriminatory power than the usual predicted values from logistic regression.

In Chapter 6 we extend the basic methodology for shrinkage estimators to the general model involving a number of categorical covariates. We also propose that by studying the deviance of the model we could form a basic decision rule to choose which method, the model or the shrunk estimates gives better discriminatory power.

Chapter 6

A General Approach to Shrinkage in Models for Categorical Data

In this chapter we shall extend and generalise the work of Chapter 5. Previously we have introduced the idea of shrinkage as a way of improving prediction in logistic regression for categorical data by incorporating prior information. Before, we expressed a prior mean and variance on the logistic transform of the number of 'successes' in the categories. Generalising this method we can now express a prior model instead of a prior mean on the logit of the successes (we assume that we use logistic regression modelling throughout). Consequently we can now

talk of 'shrinking to the model' instead of shrinking to the mean as discussed previously.

For example, we may erroneously fit a logistic regression with just the main effects to contingency table data when in fact the model with first, second etc. order interactions is a much better fit to the data. In this case, the shrinkage procedure will produce estimates that reflect the empirical logits of the proportions in the data as the model produces a poor fit to the data. If the main effects model were in fact an excellent fit to the data, the shrinkage procedure would almost reproduce the model estimates, hence 'shrinking to the model'. Another interpretation to this is that we are allowing the data and our prior beliefs in R to 'inform' us if we are misspecifying the model for the data.

In this chapter we extend the methodology introduced in Chapter 5 and apply it to a data set concerning ear infection rates amongst beach users. We then present the results of a simulation study. The study shows that when we mimic the situation where neither the model or the empirical logits are an excellent fit to the data (we would expect this situation to occur most often), the shrunk estimates provide a greater prospective ROC area and hence greater discriminatory power than the model or empirical logits.

6.1 Methodology for more than one categorical variable

Let there be N categorical variables $X_1, \ldots, X_N$. Cross-classify these variables into a N dimensional contingency table and number the cells in this table $i = 1, \ldots, m$. Define y_i to be the number of successes in the ith cell. Then, as before, $y_i \sim \text{Bin}(n_i, p_i)$ where n_i is the number of individuals in the ith cell and p_i is the probability of success in the ith cell.

Let θ_i be the logistic transform of the probability of success in the ith cell. So as before we can estimate θ_i by the empirical logistic transform of the y_i's. Define

$$\hat{\theta}_i = \log \frac{y_i + \frac{1}{2}}{n_i - y_i + \frac{1}{2}}$$

and

$$\hat{\underline{\theta}} \mid \underline{\theta} \sim \text{N}(\underline{\theta}, V_{\theta\theta}).$$

where $V_{\theta\theta}$ is the variance matrix of the $\hat{\theta}_i$'s and $\text{Var}(\hat{\theta}_i) = (n_i p_i (1 - p_i))^{-1}$. Alternatively we could assume that $\underline{\theta}$ follows a logistic model $\underline{f}(\beta)$, such that

$$\theta_i = f_i(\beta) = \frac{e^{\beta^T t_i}}{1 + e^{\beta^T t_i}}$$

where t_i is the covariate vector for the ith individual, so that an alternative estimate of θ_i is $\tilde{\theta}_i$ where

$$\tilde{\theta}_i = f_i(\hat{\beta})$$

and $\hat{\beta}$ is an estimate of β. We assume a prior on $\underline{\theta}$ to be

$$\underline{\theta} \sim \mathrm{N}(\underline{f}(\beta), \tau^2 R)$$

and we have

$$\mathrm{E}(\hat{\theta}) = \mathrm{E}(\underline{\theta}) = \underline{f}(\beta)$$

so $\underline{\theta}$ can be estimated by $\hat{\underline{\theta}}$ (the empirical logits) or $\underline{f}(\hat{\beta})$ (the predicted values of the model). In order to estimate τ^2 we firstly need to evaluate the expression $\mathrm{Var}\left[\hat{\theta}_i - f_i(\hat{\beta})\right]$, the variance of the difference of the two estimates.

Theorem 6.1 *If $\hat{\theta}_i$ are the empirical logits from the contingency table and $f_i(\hat{\beta})$ are the fitted logits from a logistic regression with estimated coefficients $\hat{\beta}$ then:*

$$\mathrm{Var}\left[\hat{\theta}_i - f_i(\hat{\beta})\right] \simeq \frac{1}{n_i p_i (1-p_i)} - \underline{t}_i^T \Omega^{-1} \underline{t}_i \tag{6.1}$$

$$\mathrm{Cov}\left[\hat{\theta}_i - f_i(\hat{\beta}), \hat{\theta}_j - f_j(\hat{\beta})\right] \simeq -\underline{t}_j^T \Omega^{-1} \underline{t}_i$$

where

$$\Omega = \sum_i \underline{t}_i \underline{t}_i^T n_i p_i (1-p_i)$$

p_i is the probability of success in ith cell

n_i is the total number in the ith cell

t_i is the covariate vector for the ith individual

Proof

Let

$$\hat{\theta}_i = \log\left(\frac{y_i}{n_i - y_i}\right)$$

where

$$y_i \sim \text{Bin}(n_i, p_i)$$

be the empirical logit estimated directly from the data. As we have already noted, in practice we add a continuity correction of a $\frac{1}{2}$ to the denominator and numerator of the logit. We omit the correction here as asymptotically it has no effect.

Asymptotically, assuming the n_i's are large (so that $(\frac{y_i}{n_i} - p_i)$ is small) we can use a first order Taylor expansion about p_i to show

$$\hat{\theta}_i \simeq \log\left(\frac{p_i}{1-p_i}\right) + \left(\frac{y_i}{n_i} - p_i\right)\frac{1}{p_i(1-p_i)} \tag{6.2}$$

This is the 'linearisation' of the logit: note that the variance of $\hat{\theta}_i$ is derived from the second term in this expansion. Now, for the ith cell we call our estimate of the logit from the model $f_i(\hat{\beta})$ where $\hat{\beta}$ are the usual maximum likelihood estimates.

The estimating equations for logistic regression are

$$\sum_i \underline{t}_i n_i \left(\frac{y_i}{n_i} - p_i(\beta)\right) \tag{6.3}$$

We estimate β by equating (6.3) to 0 and solving i.e.

$$\sum_i \underline{t}_i n_i \left(\frac{y_i}{n_i} - p_i(\hat{\beta})\right) = 0 \tag{6.4}$$

By Taylor expansion about $\hat{\beta}$, (assuming that $\hat{\beta} - \beta$ is small) we have

$$p_i(\hat{\beta}) = p_i(\beta) - p_i'(\beta)^T(\hat{\beta} - \beta) \tag{6.5}$$

and by substituting into (6.4) we have

$$\begin{aligned} \sum_i \underline{t}_i n_i \left(\frac{y_i}{n_i} - p_i(\beta) - p_i'(\beta)^T(\hat{\beta} - \beta) \right) &= 0 \\ \sum_i \underline{t}_i n_i \left(\frac{y_i}{n_i} - p_i(\beta) \right) - \sum_i \underline{t}_i n_i p_i'(\beta)^T(\hat{\beta} - \beta) &= 0 \end{aligned}$$

So

$$(\hat{\beta} - \beta) \simeq \left[\sum_i \underline{t}_i n_i p_i'(\beta)^T \right]^{-1} \left(\sum_i \underline{t}_i n_i \left(\frac{y_i}{n_i} - p_i(\beta) \right) \right) \tag{6.6}$$

Now

$$\begin{aligned} f_i(\hat{\beta}) &= \log \frac{p_i(\hat{\beta})}{1 - p_i(\hat{\beta})} \\ &= \log \frac{p_i(\beta)}{1 - p_i(\beta)} - \frac{p_i(\hat{\beta}) - p_i(\beta)}{p_i(\beta)(1 - p_i(\beta))} \end{aligned} \tag{6.7}$$

again 'linearising' the logit through a Taylor expansion about $\hat{\beta}$.

Using the identity

$$p_i'(\beta) = p_i(\beta)(1 - p_i(\beta))\underline{t}_i \tag{6.8}$$

we have from (6.5) and (6.7)

$$\begin{aligned} f_i(\hat{\beta}) &= \log \frac{p_i(\beta)}{1 - p_i(\beta)} + \frac{p_i(\beta)(1 - p_i(\beta))\underline{t}_i^T(\hat{\beta} - \beta)}{p_i(\beta)(1 - p_i(\beta))} \\ f_i(\hat{\beta}) &= \log \frac{p_i(\beta)}{1 - p_i(\beta)} + \underline{t}_i^T(\hat{\beta} - \beta) \end{aligned} \tag{6.9}$$

As we are assuming that the model is true, we can write $p_i(\beta) = p_i$, the true 'success' probability in the ith cell.

We wish to evaluate $V\left[\hat{\theta}_i - f_i(\hat{\beta})\right]$, the variance of the difference between the empirical logit and fitted logit from the model. From (6.2) and (6.9) we have

$$\begin{aligned}\text{Var}\left[\hat{\theta}_i - f_i(\hat{\beta})\right] &= \text{Var}\left[\log\frac{p_i}{1-p_i} + \left(\frac{y_i}{n_i} - p_i\right)\frac{1}{p_i(1-p_i)} - \log\frac{p_i}{1-p_i} - \underline{t}_i^T(\hat{\beta}-\beta)\right] \\ &= \text{Var}\left[\left(\frac{y_i}{n_i} - p_i\right)\frac{1}{p_i(1-p_i)} - \underline{t}_i^T(\hat{\beta}-\beta)\right] \qquad (6.10)\end{aligned}$$

Substituting (6.8) in (6.6) we have

$$(\hat{\beta}-\beta) \simeq \left[\sum_i \underline{t}_i\underline{t}_i^T n_i p_i(1-p_i)\right]^{-1}\left(\sum_i \underline{t}_i n_i\left(\frac{y_i}{n_i} - p_i\right)\right)$$

Let

$$\Omega = \left[\sum_i \underline{t}_i\underline{t}_i^T n_i p_i(1-p_i)\right]$$

and substituting for $(\hat{\beta}-\beta)$ in (6.10) gives

$$\text{Var}\left[\hat{\theta}_i - f_i(\hat{\beta})\right] = \text{Var}\left[\left(\frac{y_i}{n_i} - p_i\right)\frac{1}{p_i(1-p_i)} - \Omega^{-1}\underline{t}_i^T\left(\sum_j \underline{t}_j n_j\left(\frac{y_j}{n_j} - p_j\right)\right)\right] \qquad (6.11)$$

From (6.11) the variance consists of three parts

$$\begin{aligned}\text{Var}\left[\left(\frac{y_i}{n_i} - p_i\right)\frac{1}{p_i(1-p_i)}\right] &= \frac{1}{n_i p_i(1-p_i)} \\ \text{Var}\left[\left(\sum_j \underline{t}_j n_j\left(\frac{y_j}{n_j} - p_j\right)\right)\right] &= \sum_j n_j p_j(1-p_j)\underline{t}_j\underline{t}_j^T = \Omega \\ \text{so Var}\left[\underline{t}_i^T\Omega^{-1}\left(\sum_j \underline{t}_j n_j\left(\frac{y_j}{n_j} - p_j\right)\right)\right] &= \underline{t}_i^T\Omega^{-1}\Omega\Omega^{-1}\underline{t}_i = \underline{t}_i^T\Omega^{-1}\underline{t}_i\end{aligned}$$

and

$$\begin{aligned}
&\text{Cov}\left[\left(\frac{y_i}{n_i}-p_i\right)\frac{1}{p_i(1-p_i)}\,,\ \underline{t}_i^T\Omega^{-1}\left(\sum_j \underline{t}_j n_j\left(\frac{y_j}{n_j}-p_j\right)\right)\right]\\
&= \underline{t}_i^T\Omega^{-1}\underline{t}_i n_i\frac{1}{p_i(1-p_i)}\text{Var}\left[\frac{y_i}{n_i}-p_i\right]\\
&= \underline{t}_i^T\Omega^{-1}\underline{t}_i
\end{aligned}$$

Therefore

$$\text{Var}\left[\hat{\theta}_i - f_i(\hat{\beta})\right] \simeq \frac{1}{n_i p_i(1-p_i)} + \underline{t}_i^T\Omega^{-1}\underline{t}_i - 2\underline{t}_i^T\Omega^{-1}\underline{t}_i = \frac{1}{n_i p_i(1-p_i)} - \underline{t}_i^T\Omega^{-1}\underline{t}_i$$

We also need to find the expression for the covariance term

$$\text{Cov}(\hat{\theta}_i - f_i(\hat{\beta})\,,\ \hat{\theta}_j - f_j(\hat{\beta}))$$

The covariance term consists of four parts.

$$\begin{aligned}
\text{Cov}\left[\left(\frac{y_i}{n_i}-p_i\right)\frac{1}{p_i(1-p_i)}\,,\ \left(\frac{y_j}{n_j}-p_j\right)\frac{1}{p_j(1-p_j)}\right] &= 0 \quad \forall\, i\neq j\\
\text{Cov}\left[\left(\frac{y_i}{n_i}-p_i\right)\frac{1}{p_i(1-p_i)}\,,\ \underline{t}_j^T\Omega^{-1}\left(\sum_s \underline{t}_s n_s\left(\frac{y_s}{n_s}-p_s\right)\right)\right] &= \underline{t}_j^T\Omega^{-1}\underline{t}_i\\
\text{Cov}\left[\left(\frac{y_j}{n_j}-p_j\right)\frac{1}{p_j(1-p_j)}\,,\ \underline{t}_i^T\Omega^{-1}\left(\sum_r \underline{t}_r n_r\left(\frac{y_r}{n_r}-p_r\right)\right)\right] &= \underline{t}_i^T\Omega^{-1}\underline{t}_j\\
\text{Cov}\left[\underline{t}_i^T\Omega^{-1}\left(\sum_s \underline{t}_s n_s\left(\frac{y_s}{n_s}-p_s\right)\right)\,,\ \underline{t}_j^T\Omega^{-1}\left(\sum_r \underline{t}_r n_r\left(\frac{y_r}{n_r}-p_r\right)\right)\right] &= \underline{t}_i^T\Omega^{-1}\underline{t}_j
\end{aligned}$$

The last term in the covariance calculation follows as

$$\text{Cov}\left[\sum_s \underline{t}_s n_s\left(\frac{y_s}{n_s}-p_s\right)\,,\ \sum_r \underline{t}_r n_r\left(\frac{y_r}{n_r}-p_r\right)\right] = \sum_s \underline{t}_s\underline{t}_s^T n_s p_s(1-p_s) = \Omega$$

So combining all terms we have

$$\text{Cov}(\hat{\theta}_i - f_i(\hat{\beta})\ ,\ \hat{\theta}_j - f_j(\hat{\beta})) = -\underline{t}_i^T\Omega^{-1}\underline{t}_j - \underline{t}_j^T\Omega^{-1}\underline{t}_i + \underline{t}_i^T\Omega^{-1}\underline{t}_j = -\underline{t}_j^T\Omega^{-1}\underline{t}_i$$

We know that

$$\text{Var}(\hat{\theta}_i) \simeq \frac{1}{n_i p_i(1-p_i)}$$

Now using (6.1), the global variance of $\underline{\hat{\theta}} - \underline{f}(\hat{\beta})$ is

$$\begin{aligned}\text{Var}\left[\underline{\hat{\theta}} - \underline{f}(\hat{\beta})\right] &= \text{E}_{\underline{\theta}}\left\{\text{Var}\left[\underline{\hat{\theta}} - \underline{f}(\hat{\beta})\,|\,\underline{\theta}\right]\right\} + \text{Var}_{\underline{\theta}}\left\{\text{E}\left[\underline{\hat{\theta}} - \underline{f}(\hat{\beta})\,|\,\underline{\theta}\right]\right\} \\ &= V_{\theta\theta} - T^T\Omega^{-1}T + \tau^2 R\end{aligned}$$

where T is the matrix of covariate vectors. As before, the natural choice of weight matrix is $W = V_{\theta\theta}^{-1}$ so

$$\begin{aligned}\text{E}[(\hat{\underline{\theta}} - \underline{f}(\hat{\beta}))^T V_{\theta\theta}^{-1}(\hat{\underline{\theta}} - \underline{f}(\hat{\beta}))] &= \text{tr}\left[V_{\theta\theta}^{-1}(V_{\theta\theta} - T^T\Omega^{-1}T + \tau^2 R)\right] \\ &= \text{tr}\left[V_{\theta\theta}^{-1}V_{\theta\theta} - V_{\theta\theta}^{-1}T^T\Omega^{-1}T + V_{\theta\theta}^{-1}\tau^2 R\right] \\ &= m - \text{tr}(V_{\theta\theta}^{-1}T^T\Omega^{-1}T) + \tau^2\text{tr}(V_{\theta\theta}^{-1}R) \\ &= m - k + \tau^2\text{tr}(V_{\theta\theta}^{-1}R)\end{aligned}$$

as $\text{tr}(V_{\theta\theta}^{-1}T^T\Omega^{-1}T) = \text{tr}(\Omega^{-1}T^T V_{\theta\theta}^{-1}T) = \text{tr}(\Omega^{-1}\Omega) = k$

So τ^2 can be estimated by $\hat{\tau}^2$ where

$$\hat{\tau^2} = \frac{(\hat{\underline{\theta}} - \underline{f}(\hat{\beta}))^T V_{\theta\theta}^{-1} (\hat{\underline{\theta}} - \underline{f}(\hat{\beta})) - (m-k)}{\mathrm{tr}(V_{\theta\theta}^{-1} R)}$$

$$= \frac{(\hat{\underline{\theta}} - \underline{f}(\hat{\beta}))^T V_{\theta\theta}^{-1} (\hat{\underline{\theta}} - \underline{f}(\hat{\beta})) - (m-k)}{\sum_i n_i p_i (1-p_i)}$$

as $\mathrm{tr}(V_{\theta\theta}^{-1} R) = \sum_i n_i p_i (1-p_i)$ because $V_{\theta\theta}$ is diagonal and $\mathrm{Var}(\hat{\theta}_i) = (n_i p_i (1-p_i))^{-1}$. Again we truncate $\hat{\tau^2}$ at zero.

The behaviour of $\hat{\tau}^2$ is principally related to the term in the numerator

$$(\hat{\underline{\theta}} - \underline{f}(\hat{\beta}))^T V_{\theta\theta}^{-1} (\hat{\underline{\theta}} - \underline{f}(\hat{\beta}))$$

which we shall call M. We shall now show that M is approximately the deviance of the model.

Theorem 6.2 *If DEV is the deviance of the model under study and*

$$M = (\hat{\underline{\theta}} - \underline{f}(\hat{\beta}))^T V_{\theta\theta}^{-1} (\hat{\underline{\theta}} - \underline{f}(\hat{\beta}))$$

then $DEV \simeq M$

Proof

The likelihood, L of the model f under study is

$$L = \sum_i y_i \log p_i + (n_i - y_i) \log(1-p_i)$$

and the likelihood of the maximal (saturated) model is

$$\underset{f}{MaxL} = \sum_i y_i \log \frac{y_i}{n_i} + (n_i - y_i) \log \left(1 - \frac{y_i}{n_i}\right)$$

Therefore the deviance of the model f is

$$DEV(f) = 2 \sum_i \left[y_i \log \frac{y_i}{n_i p_i} + (n_i - y_i) \log \frac{n_i - y_i}{n_i (1 - p_i)} \right]$$

Now let

$$\frac{y_i}{n_i} = p_i + \epsilon_i$$

for small ϵ_i. Then using power series expansions we have

$$\begin{aligned}
y_i \log \frac{y_i}{n_i p_i} + (n_i - y_i) \log \frac{n_i - y_i}{n_i (1 - p_i)} = & \\
& n_i \left[(p_i + \epsilon_i) \log \left(\frac{p_i + \epsilon_i}{p_i} \right) + (1 - p_i - \epsilon_i) \log \frac{1 - p_i - \epsilon_i}{1 - p_i} \right] \\
= \quad & n_i \left[p_i \left(1 + \frac{\epsilon_i}{p_i}\right) \left(\frac{\epsilon_i}{p_i} - \frac{1}{2} \frac{\epsilon_i^2}{p_i^2} \right) + (1 - p_i) \left(1 - \frac{\epsilon_i}{1 - p_i}\right) \left(-\frac{\epsilon_i}{1 - p_i} - \frac{1}{2} \frac{\epsilon_i^2}{(1 - p_i)^2} \right) \right] \\
= \quad & n_i \left[\epsilon_i \left(1 + \frac{\epsilon_i}{p_i}\right) \left(1 - \frac{1}{2} \frac{\epsilon_i}{p_i}\right) - \epsilon_i \left(1 - \frac{\epsilon_i}{1 - p_i}\right) \left(1 + \frac{1}{2} \frac{\epsilon_i}{1 - p_i}\right) \right] \\
= \quad & n_i \epsilon_i \left(\frac{1}{2} \frac{\epsilon_i}{p_i} + \frac{1}{2} \frac{\epsilon_i}{1 - p_i} \right) \\
= \quad & \frac{1}{2} n_i \frac{\epsilon_i^2}{p_i (1 - p_i)}
\end{aligned}$$

Hence

$$DEV(f) \simeq \sum n_i \frac{\epsilon_i^2}{p_i (1 - p_i)}$$

Now

$$\theta_i = \log \frac{p_i}{1 - p_i} = f_i(\beta) \text{ and } \hat{\theta}_i = \log \frac{y_i}{n_i - y_i}$$

So using the same substitution as above

$$(\hat{\theta}_i - \theta_i) = \log \frac{y_i(1-p_i)}{(n_i - y_i)p_i} = \log \frac{1 + \frac{\epsilon_i}{p_i}}{1 - \frac{\epsilon_i}{1-p_i}} \simeq \frac{\epsilon_i}{p_i} + \frac{\epsilon_i}{1-p_i} = \frac{\epsilon_i}{p_i(1-p_i)}$$

Therefore

$$\sum \frac{(\hat{\theta}_i - \theta_i)^2}{\mathrm{Var}(\hat{\theta}_i)} = \sum (\hat{\theta}_i - \theta_i)^2 n_i p_i (1-p_i) = \sum n_i \frac{\epsilon_i^2}{p_i(1-p_i)} = DEV(f).$$

Then the deviance of the fitted model is

$$DEV = DEV(\hat{f}) = \sum \frac{(\hat{\theta}_i - f_i(\hat{\beta}))^2}{\mathrm{Var}(\hat{\theta}_i)} = (\underline{\hat{\theta}} - \underline{f}(\hat{\beta}))^T V_{\theta\theta}^{-1} (\underline{\hat{\theta}} - \underline{f}(\hat{\beta}))$$

Having evaluated τ^2 by the above formula, the shrunk estimates of $\underline{\theta}$ can be evaluated by

$$\mathrm{E}(\underline{\theta} \mid \underline{\hat{\theta}}) = \left[V_{\theta\theta}^{-1} + \tau^{-2} R^{-1}\right]^{-1} \left[V_{\theta\theta}^{-1} \underline{\hat{\theta}} + \underline{f}(\beta) \tau^{-2} R^{-1}\right].$$

As before these can be estimated by substituting in $\hat{\beta}$ and $\hat{\tau^2}$, the estimates of β and τ^2.

6.2 Example

To investigate the properties of the shrinkage methodology for the general case described above, we take as an example data from the 1990 Pilot Surf/Health

Study of the New South Wales Water Board (taken from Hand *et al* (1994)).

Measurements were taken on 287 individuals and the objective is to produce a score to identify which individuals have a greater risk of developing ear infections. A main effects model was fitted to the four covariates in the data using logistic regression. The results of this logistic regression and a description of the four covariates can be found in Table 6.1. To illustrate our generalised shrinkage

Short Name	Full name	VALUES TAKEN	t value
Intercept	-	-	-0.871
FreqOS	Frequent Ocean Swimmer	1 (if they are), 2 (if they're not)	1.674
Loc	Swimming Location	1 (non-beach), 4 (beach)	2.951
AgeGrp	Age Group	2 (15-19), 3 (20-25), 4 (25-29)	(2) 0.531, (3) -0.404
Gen	Gender	1 (male), 2 (female)	-0.112
EarInf (Response)	Ear Infection Suffered	0 (no), 1 (yes)	-

Table 6.1: Description of the variables in the Pilot Surf/Health Study of NSW Water Board

approach, we fit a main effects model to Location, FreqOS and Gender using logistic regression. Location is the only significant covariate and the area under the ROC curve for this model was 0.610, indicating poor discriminatory power of this particular model.

As there are only eight cells in the contingency table, we can illustrate the empirical, model and shrunk values in a table. These are presented in Table 6.2 ($\hat{\tau}^2 = 0.02903$ for this model). We can see the shrunk estimates are fairly close to

the model although the data is having an impact on some of the estimates (in this example we have used the identity matrix for R). The area under the ROC curve

Empirical Logit	Model	Shrunk Values
0.829	0.467	0.560
0.100	0.465	0.423
-0.609	-0.262	-0.344
-0.121	-0.264	-0.238
0.037	0.062	0.055
-0.427	0.060	0.004
-0.676	-0.667	-0.669
-0.260	-0.669	-0.602

Table 6.2: Empirical, model and shrunk estimates for the simple main effects model fitted to the ear infection data

for these shrunk values is 0.606, a slight worsening of discriminatory ability than the model. Two of the covariates in the model are Location and Frequency Of Ocean Swimming. We would expect that the ear infection rates between frequent ocean swimmers and individuals who use the beach as a swimming location to be quite similar, hence inducing a correlation of 0.9 say. Conversely, we might expect that the infection rates between individuals who are not frequent ocean swimmers and beach users to be quite different, perhaps inducing a high negative

correlation, say -0.9 and vice versa for frequent ocean swimmers and non-beach users. Using this information to produce R we can calculate $\hat{\tau}^2$ and the shrunk estimates as before. Calculating the ROC for this R gives an area of 0.610, implying slightly better discrimination. Although the rise in ROC area is very small, by including some subjective information about the data we have improved on discrimination using the shrinkage method. Including more correlations in R by gathering more information about relationships between beach use and age group etc. might help improve on the model estimates further.

To check Theorem 6.2, that the principal term in the calculation of $\hat{\tau}^2$ is approximately equal to the deviance we can calculate each separately for a number of models and compare in Table 6.3 (G = Gender, L = Location, F = Frequent Ocean Swimmer, A = Age Group). The figures in the table are in close agreement implying the approximation is particularly good. The model and empirical deviances for the model G + L + A are not particularly close, but on closer inspection it appears that one of the cells in the contingency table consists of individuals who all have ear infections. The logit in this cell is large relative to the estimate from the model, and therefore the value of $(\hat{\underline{\theta}} - \underline{f}(\hat{\underline{\beta}}))$ is large relative to all the other cells, overestimating the deviance.

Model	$(\hat{\underline{\theta}} - \underline{f}(\hat{\underline{\beta}}))^T V_{\theta\theta}^{-1} (\hat{\underline{\theta}} - \underline{f}(\hat{\underline{\beta}}))$	Model Deviance
G + L	3.95	4.07
L + F	1.54	1.58
G + L + A	8.59	9.38
G + L + F	5.79	5.99
G + L + F + G*F	1.75	1.82
G + L + F + L*F	4.14	4.42
G + L + F + A	18.31	18.26

Table 6.3: Examples to verify the approximation of the deviance in calculating $\hat{\tau}^2$

6.3 Simulation studies

6.3.1 Artificial data

To test out the properties of the general methodology further we perform a number of simulation studies. Our first simulations will involve creating a number of artificial data sets to investigate the discriminatory power of the empirical logits and the model and shrunk estimates. We purposefully create situations in which the model fits well and badly to see if this is borne out by the shrunk estimates. The simulation procedure is as follows:-

1. Let p_j act as the 'true' probability of occurrence in the jth cell in the contingency table and let n_j be the number of individuals in that cell.

2. From p_j simulate new observed proportions p_j^* from a binomial distribution with parameters p_j and n_j, to act as the data on which we model.

3. Use the observed proportions p_j^* to fit the linear trend model, $\log \frac{p_j}{1-p_j} = \alpha + \beta j$

4. Use the shrinkage methodology to calculate the shrunk estimates.

5. Use the observed proportions p_j^* to calculate the retrospective ROC areas for the empirical logits, model and shrunk estimates.

6. Use the probabilities p_j to calculate the prospective ROC areas for the empirical logits, model and shrunk estimates.

7. Repeat the above n_{sim} times and average the ROC areas.

Until now we have been using a binary vector of occurrence and the score for each individual in calculating the ROC area. To calculate the ROC area from the scores (empirical, model or shrunk), the probabilities p_j (prospective) or p_j^* (retrospective) and the cell sizes n_j's, we use the following theorem.

Theorem 6.3 *For ranked scores s_i, associated probabilities p_i and weights $q_i = n_i / \sum n_i, i = 1, \ldots, n$ the area under the ROC curve is*

$$A = \sum_i^n \left\{ \sum_{j=i+1}^n \alpha_j + \frac{\alpha_i}{2} \right\} \beta_i$$

where

$$\alpha_i = \frac{p_i q_i}{\sum p_i q_i} \text{ and } \beta_i = \frac{(1-p_i)q_i}{\sum (1-p_i)q_i}$$

Proof

If y is the binary indicator of status then

$$\mathrm{P}(Y = 1 | s = s_i) = p_i \text{ and } \mathrm{P}(s = s_i) = q_i$$

therefore using a simple application of Bayes Theorem

$$\mathrm{P}(s = s_i | Y = 1) = \frac{p_i \mathrm{P}(s = s_i)}{\mathrm{Pr}(Y = 1)} = \frac{p_i q_i}{\sum p_i q_i} = \alpha_i$$

and

$$\mathrm{P}(s = s_i | Y = 0) = \frac{(1-p_i)\mathrm{P}(s = s_i)}{\mathrm{P}(Y = 1)} = \frac{(1-p_i)q_i}{\sum (1-p_i)q_i} = \beta_i$$

Let $s^{(1)}$ and $s^{(0)}$ be random variables where $\mathrm{P}(s^{(1)} = s_i) = \alpha_i$ and $\mathrm{P}(s^{(0)} = s_i) = \beta_i$. The area under the ROC curve is then

$$\mathrm{P}(s^{(1)} \geq s^{(0)}) + \frac{1}{2}\mathrm{P}(s^{(1)} = s^{(0)})$$

Now

$$\mathrm{P}(s^{(1)} \geq s^{(0)}) = \sum_i^n \left(\sum_{j=i+1}^n \alpha_j \right) \beta_i$$

and

$$\mathrm{P}(s^{(1)} = s^{(0)}) = \mathrm{P}(s^{(1)} = s^{(0)} = s_i) = \sum_i^n \alpha_i \beta_i$$

and the result follows.

For our first simulation we choose our set of prospective probabilities and totals as

$$p1 = (0.1, 0.2, 0.3, 0.4, 0.3, 0.6, 0.8\ 0.8, 0.9, 0.8)$$

$$n1 = (10, 10, 10, 10, 10, 10, 10, 10, 10, 10)$$

We would expect that, on average the linear trend model would fit these probabilities well. The results from $n_{sim} = 1000$ simulations are shown in Table 6.4 where A_R and A_P are the retrospective and prospective ROC areas. In this simulation

	A_R	A_P
Empirical	0.842($\pm$0.04)	0.788 $\pm$0.02)
Model	0.807($\pm$0.04)	0.804 ($\pm$0)
Shrunk	0.823($\pm$0.04)	0.804 ($\pm$0.01)

Table 6.4: Results of the simulation procedure using probabilities $p1$ and cell frequencies $n1$

and those to follow it is obvious that retrospectively, the empirical logits will have the highest ROC area as they reflect the proportions in the different cells in the contingency table. Prospectively we have the model as the best performing of the three but it has an area only 0.001 greater than the shrunk estimates. The retrospective areas are more variable as we are creating a new set of probabilities for the retrospective analysis in each simulation, whilst holding the prospective probabilities constant. The model prospective ROC area is constant because of the linear nature of the model - the model estimates will always be in the same order and hence the prospective ROC area will be the same for every simulation. This simulation serves to show that when the model fits well, the shrinkage method is weighted heavily towards the model estimates and consequently have a similar prospective ROC area.

For the second simulation we have

$$p2 = (0.5, 0.9, 0.5, 0.9, 0.5, 0.9, 0.5\ 0.9, 0.5, 0.9)$$

$$n2 = (10, 10, 10, 10, 10, 10, 10, 10, 10, 10)$$

The resulting ROC areas are shown in Table 6.5. In this case the linear trend model should be a poor fit on average and this is borne out by the results. Prospectively the empirical logits are the best discriminator and the shrunk values are weighted heavily in favour of these empirical values. This is intuitively reasonable as the deviance is large for a poor model fit, hence a large value of

	A_R	A_P
Empirical	0.808($\pm$0.04)	0.729 ($\pm$0.02)
Model	0.560($\pm$0.04)	0.529 ($\pm$0.04)
Shrunk	0.805($\pm$0.05)	0.724($\pm$0.03)

Table 6.5: Results of the simulation procedure using probabilities $p2$ and cell frequencies $n2$

$\hat{\tau}^2$ which weights in favour of the empirical values. The standard deviation is non zero for the prospective model ROC area in this simulation because the poor fit of the model fit means that a negative sign could occur on the coefficient in the linear trend model, hence ranking the scores in the opposite direction and producing an area of less than 0.5.

For the third simulation we have

$$p3 = (0.1, 0.7, 0.4, 0.9, 0.5, 0.8, 0.2, 0.6, 0.7, 0.9)$$

$$n3 = (10, 10, 10, 10, 10, 10, 10, 10, 10, 10)$$

The resulting ROC areas are shown in Table 6.6. For this simulation we have deliberately chosen the probabilities to be roughly 'half way' between the extremes of the first two simulations. The rationale behind this is that the shrinkage method should extract information from the data in the form of the empirical logits, and information from the model to give as good if not better prospective

	A_R	A_P
Empirical	0.833(±0.04)	0.773 (±0.02)
Model	0.631(±0.05)	0.631 (±0.001)
Shrunk	0.831(±0.04)	0.774(±0.02)

Table 6.6: Results of the simulation procedure using probabilities $p3$ and cell frequencies $n3$

estimator. From Table 6.6, the shrunk estimates are slightly better than the empirical logits and roughly 20% better than the model estimates. Throughout these simulations we have taken R to be the identity matrix.

Our final simulation in this series assesses the effect of a larger number of cells combined with larger cell totals. We can do this by comparing with simulation 1 where the model was a good fit to the probabilities but extending the number of cells to 20 with 20 in each cell. The probabilities are similarly changed i.e.

$$p4 = (0.1,0.1,0.2,0.2,0.3,0.3, \ldots, 0.8,0.8)$$

$$n4 = (20,20,\ldots,20)$$

The results are shown in Table 6.7 where we notice a marked difference. Both the empirical and model have quite a high prospective ROC area, as the probabilities fit the model quite well (due to the overall linear trend) and the empirical logits exploit the fact that the probabilities are grouped together i.e. 0.1,0.1.

	A_R	A_P
Empirical	0.832(±0.02)	0.800 (±0.006)
Model	0.804(±0.02)	0.804 (±0)
Shrunk	0.820(±0.02)	0.806(±0.003)

Table 6.7: Results of the final simulation procedure to assess the effect of the doubled cell numbers and cell totals.

The shrunk estimates combine these two pieces of information to give a larger prospective ROC area. Notice also that the standard errors have decreased approximately by a factor of two as we have increased the number of cells by two.

In three of the situations above, the shrunk predictors have produced the largest or equal largest prospective ROC area. In the situation where the empirical logits were the best predictor, the difference in the prospective ROC areas for the shrunk estimates and the empirical logits was 0.005, which is almost negligible. Therefore, we can conclude that if the shrunk predictors were used instead of the model or empirical logits we would not have lost any discriminatory power, even in situations where the model or empirical logits are clearly a good fit to the data.

6.3.2 Prospective ROC area and model deviance

We have mentioned previously that the prospective shrunk ROC areas are indirectly related to the deviance of the model through τ^2. For any of the simulations above we can represent this graphically by plotting the difference between the model and shrunk prospective ROC areas against the deviance. Instead of keeping the prospective probabilities fixed for every simulation run of $n_{sim} = 1000$, we now randomly create a new set of probabilities for each simulation. Some of these probabilities will fit the model well, some won't and most will be between the two extremes. This will enable us to visualise fully the dependence of the prospective ROC areas on the deviance.

Figure 6.1 illustrates the relationship over 1000 simulations for the situation where there are 10 cells and 10 individuals in each cell (note that there are no points plotted for the deviance less than $(p - k)$, when $\hat{\tau}^2 \leq 0$ and the shrunk estimates are equal to the model estimates). The solid line is a scatter plot smoother based on splines fitted to this data. It is fairly obvious that there is a positive trend, that is as the deviance increases the difference between the shrunk and model ROC areas also increases. We would expect this, because a higher deviance implies a poorly fitting model which in turn implies a greater value of $\hat{\tau}^2$. This results in more weight on the empirical logits and enhances the difference between the ROC areas.

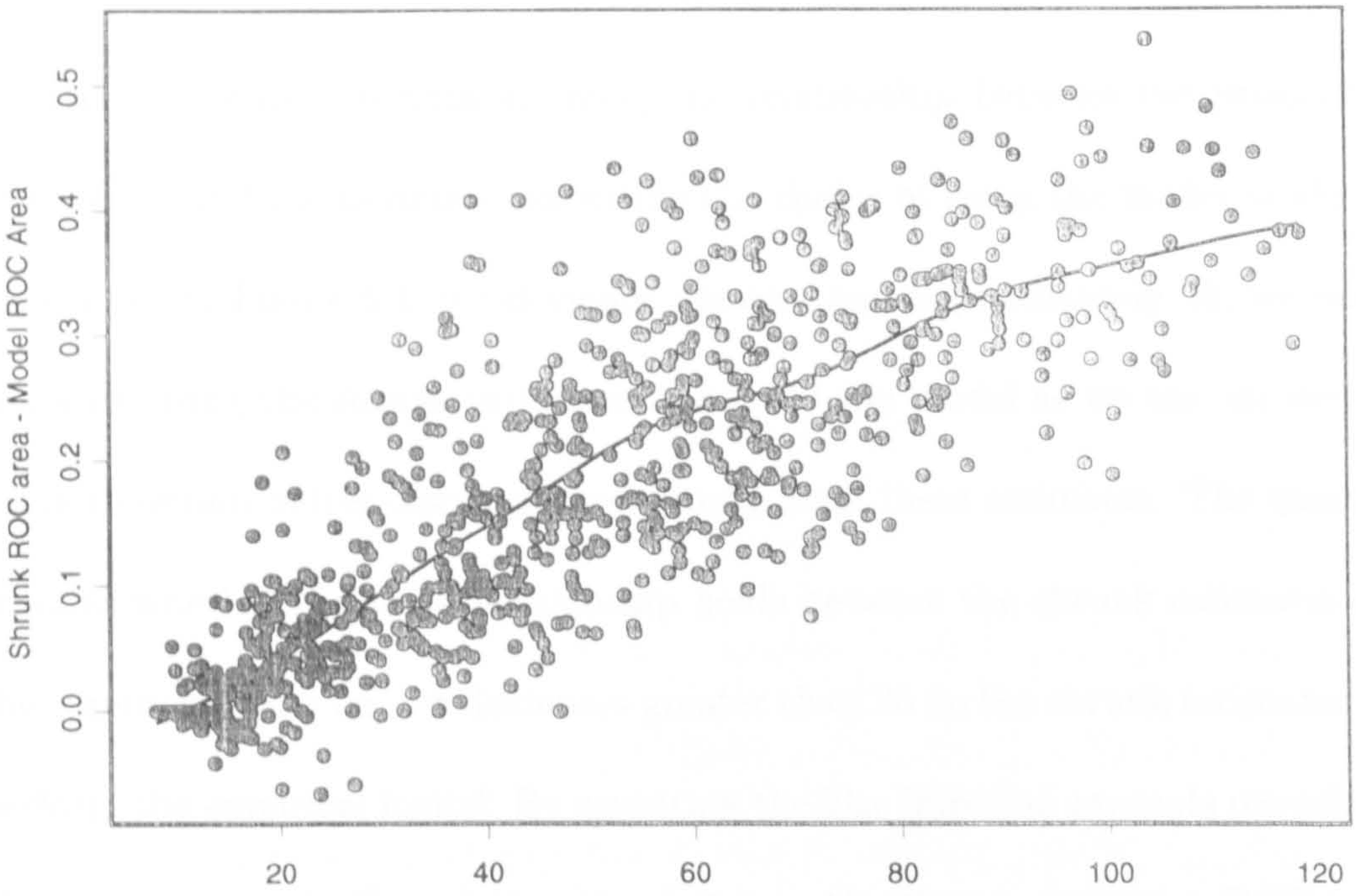

Figure 6.1: Scatter plot of difference in prospective shrunk and model ROC areas against the deviance of the model

The phenomenon of a negative difference i.e. model ROC area greater than shrunk ROC area when the deviance is small deserves explanation. A small deviance implies a well fitting model and it would be very difficult to do better than the model at this point. Therefore, for low deviances the model will nearly always be better than the shrunk estimates which results in the difference in ROC areas being negative. Repeating the above procedure for an increased number of

cells and cell sizes results in an almost identical plot to Figure 6.1, except that the deviance of the model increases roughly with the number of cells.

The justification behind studying the relationship between the prospective ROC areas and the deviance lies within the choice of using the model or shrunk estimates. In Figure 6.1, for deviance greater than approximately 20, we would advocate using the shrunk estimates instead of the model as we are, on average likely to obtain better discriminatory power from these estimates. The question remains whether the same relationship holds between the shrunk estimates and the empirical logits i.e. for deviances greater than 20 do the shrunk estimates out perform the empirical logits? By revisiting the Ear Infection example introduced above we can study the relationships between the three prospective ROC areas and the model deviance.

6.3.3 Ear infection data

To simulate from the Ear Infection data we simply count the number of cells in the contingency table ($m = 32$) and the number of individuals within them. We then have to decide what to use as the prospective probabilities in the simulation. We can either:

- use the fitted probabilities from the model, to mimic the situation where the model fits the data particularly well.

- use the observed proportion of infected individuals in each cell to mimic the situation where the empirical logits fit the data particularly well.

- use the midpoint between the fitted probabilities from the model and observed proportions to mimic the situation where neither the model or empirical logits fit the data well.

For the following simulations we study the simple additive model of the covariates described in Table 6.3.

Figure 6.2 illustrates the relationship between the ROC areas and the deviance when we simulate the prospective probabilities from the fitted probabilities from the model. To make the interpretation easier, we have omitted the individual points and plotted the scatter plot smoothers instead.

From Figure 6.2 we can see that prospective ROC area for the model is always larger than the empirical estimates (which we would expect as we are simulating from the fitted probabilities) but only slightly larger than then shrunk ROC areas over the range of the deviance. As the deviance increases, the difference between the model and the shrunk areas gets slightly larger and then decreases. This is most likely an indication of a phenomenon we mentioned above, in that if we alter the model estimates for relatively small deviance we are likely to get a slightly worse prospective ROC area. This is cancelled out as the deviance gets larger and the model becomes increasingly a poor fit to the data.

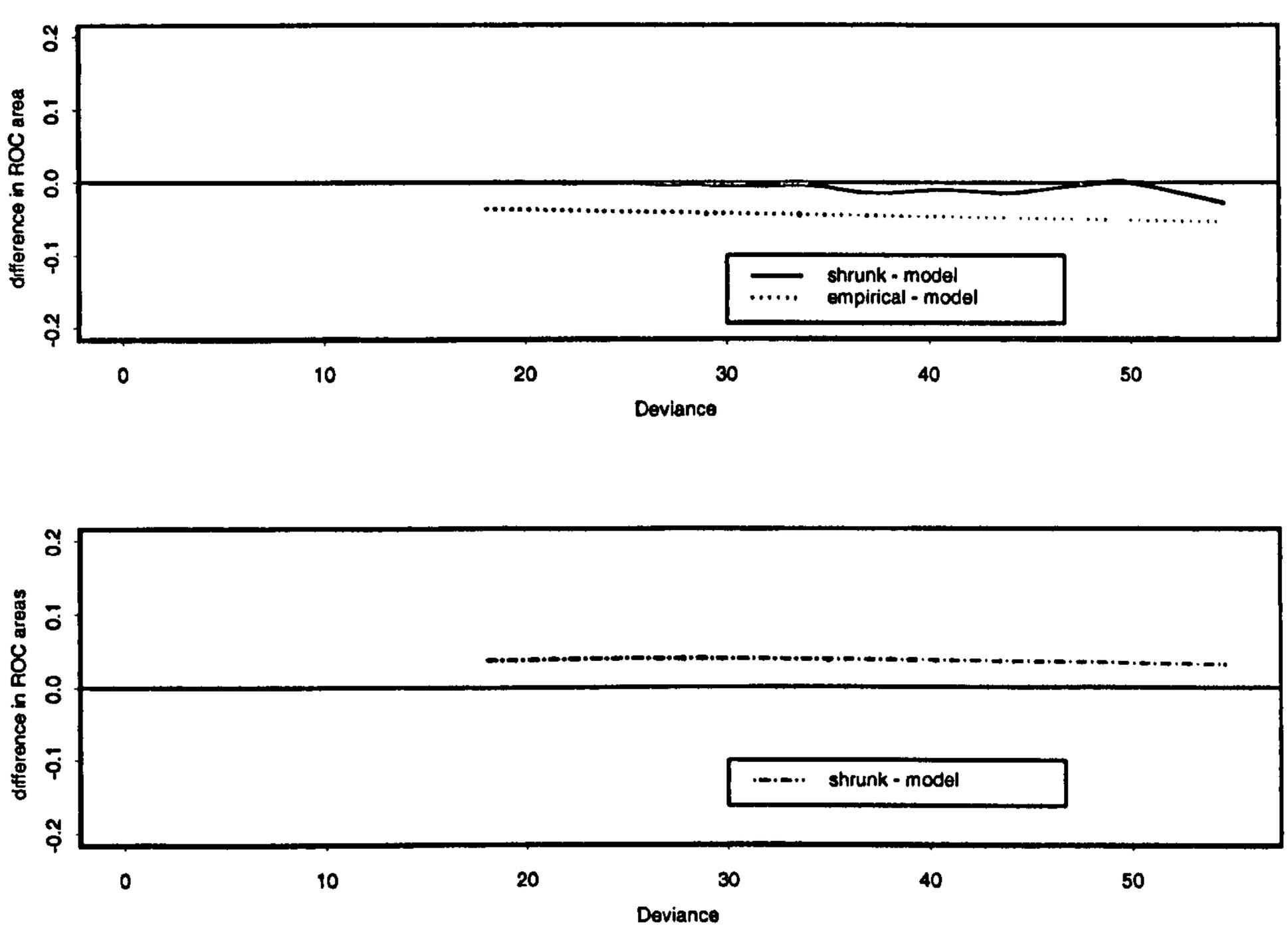

Figure 6.2: Scatter plot smoothers of differences in prospective empirical, model and shrunk ROC areas against the deviance of the model (fitted probabilities are used as the prospective probabilities)

Figure 6.3 illustrates the relationship between the three prospective ROC areas if we simulate from the cell proportions. Again, as we would expect the empirical estimates perform best prospectively because they are reflecting the cell proportions. The shrunk estimates and empirical estimates are both significantly

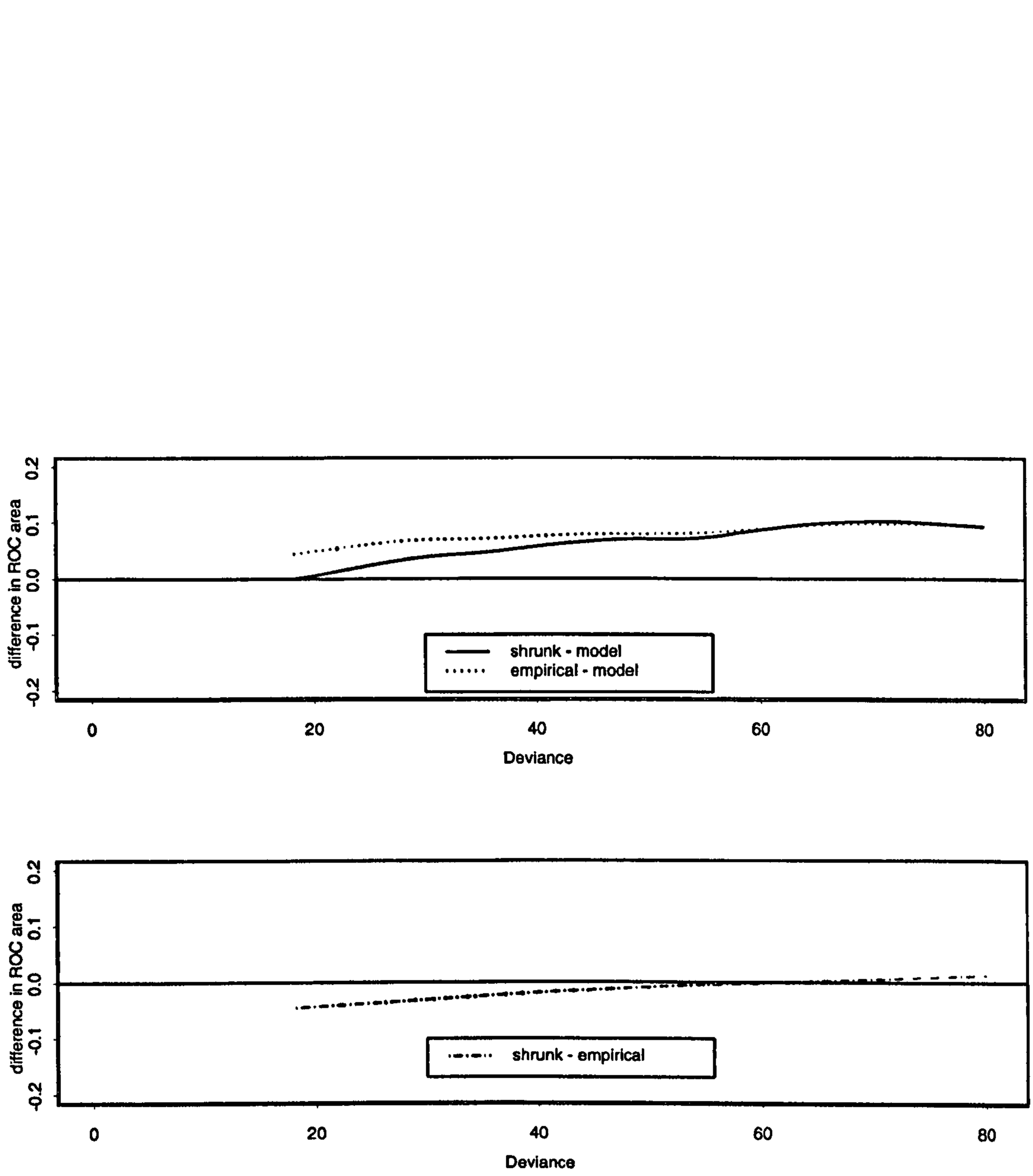

Figure 6.3: Scatter plot smoothers of differences in prospective empirical, model and shrunk ROC areas against the deviance of the model (proportion infected used as prospective probabilities)

better than the model as it fits the data poorly. It is important to notice that in both Figure 6.2 and Figure 6.3 for the extreme situations, the shrunk estimates are never the worst predictor.

This agrees with the summary of the results from the artificial simulation, that using the shrunk predictors would result in either the best performance or a very small loss of discriminatory power.

Of course we would not expect either of the above situations to occur with great frequency. More often than not, we would expect the model and the empirical logits to provide an adequate fit to the data.

We can investigate this scenario by taking the prospective probabilities to be the midpoint between the fitted probabilities from the model and the observed proportions. Using these midpoints as the prospective probabilities produces the results in Figure 6.4. It is clear that the shrunk estimates are the best predictor here over nearly the entire range of deviances. In this setting the model is the worst of the predictors, even though the number of cells we are simulating on is not particularly small ($m = 32$). This leads us to the conclusion that if we expect an adequate model fit to the data most of the time, then using the shrinkage methodology and the shrunk estimates is preferable because they will give greater discriminatory power.

We could envisage an example where the model is not so poorly fitting as in Figure 6.4, therefore the line representing the difference in ROC areas for the shrunk and model estimates would cross the x-axis at some value of the deviance, DEV_0 say. At this point we could postulate a decision rule for the model with

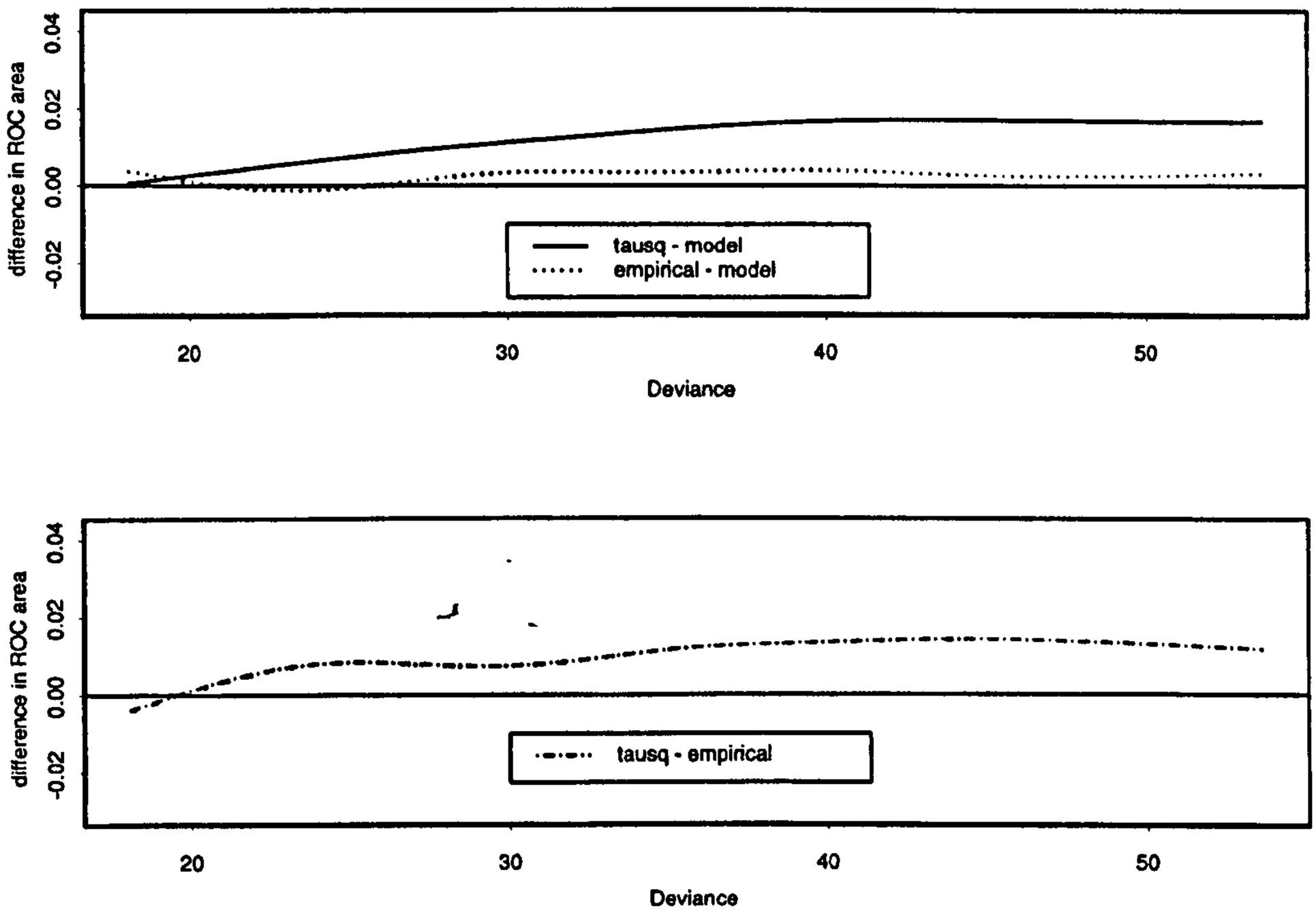

Figure 6.4: Scatter plot smoothers of differences in prospective empirical, model and shrunk ROC areas against the deviance of the model (midpoints used as prospective probabilities)

deviance DEV_M, which could take the form of choosing the model for prediction purposes if $DEV_M \leq DEV_0$ and the shrinkage method if $DEV_M \geq DEV_0$.

Chapter 7

Summary and Concluding Remarks

This thesis has presented two new approaches to the predictive accuracy of logistic regression in terms of discriminatory ability. Clearly this is an important area, especially in medical applications. For example, formulating a score that perfectly separates individuals into diseased and non-diseased on the basis of some observable measurements would have huge benefits and public health connotations. Therefore it is important to know *exactly* how well a score discriminates, preferably in practice.

In Chapter 3 we have derived formulae for the overestimation of the ROC curve and area for logistic regression by calculating the difference between the

retrospective and prospective true positive rates for fixed false positive rates. By way of a number of approximations we arrive at a simpler approximation for the overestimation of the area that only necessitates the calculation of a histogram on the scores from the logistic regression. The formulae, especially (3.38) are reasonably easy to implement and we could envisage the situation where both retrospective ROC area and the subsequent overestimation are reported together much like an estimate with accompanying standard error.

In Chapter 4, an example studying the prognosis of breast cancer patients is shown to overestimate the area under the ROC curve by approximately 16%. It would be a great concern if this score were used in practice on the basis of a retrospective analysis which overestimated the ability of the score to discriminate by such an amount. Of course, this would never happen as such scores or clinical tests go through many stages of testing before being used in practice but it serves to highlight the dangers of over optimism. We also performed a number of simulation experiments to test the validity of the overestimation formulae. The agreement is good, even for the breast cancer study with the smaller sample size.

We have also shown that the overestimation is related to the sample size and number of covariates in the study. This is clearly of relevance in medical studies as the sample sizes of study groups are often quite small due to cost implications or complexity of the study. Using and basing results on a retrospective ROC

analysis in such situations could, as we have shown be misleading.

The overestimation in ROC presented here can only be applied for logistic regression. This work is clearly related to the underestimation of error rates in discriminant analysis and further work could generalise the overestimation methodology to discriminant analysis and other classification methods. Also we have ignored sampling issues in the construction of the overestimation formulae. A logistic regression analysis can be calculated on data from a prospective, retrospective or cross-sectional study and it remains unclear whether it is reasonable to use the empirical distributions used in Chapter 3 to calculate the area under the ROC curve. Preliminary work on the sampling problem has been carried out and results suggest that the overestimation formulae hold with a few minor alterations.

We have introduced the concept of shrinkage as the process of incorporating prior information about the data into the estimates from the model. Shrinkage is usually discussed in terms of calibration, another aspect of predictive accuracy concerned with the agreement between the probability of an event and its observed frequency. Usually, shrinkage is considered in terms of continuous measurements resulting in pre-shrunk predictors but in Chapter 5 we introduce a shrinkage formulation where we have a number of categorical covariates and a binary outcome. We introduce a shrinkage parameter τ^2 which takes into ac-

count the fit of a model to the data (the deviance in a logistic regression) and the information from R, the correlation matrix between success rates in different categories. The methodology was developed for a model containing a single covariate and we investigated its properties by considering a simple example based on credit scoring. The methodology illustrates the phenomenon of shrinking the model estimates to their mean (in the case of a single covariate), by using a weighted average of the model estimates to reduce the effect of more variable information. A simulation study on credit defaulting data studied the effect of various correlations between success rates on the prospective ROC area. For small correlation values, the prospective ROC values for the shrunk values were larger than the prospective model ROC, thus enhancing discriminatory ability. Also, for a small subset of the data, high correlations could be seen to be improving the discriminating ability of the shrunk estimates.

In Chapter 6 we generalised the shrinkage methodology in Chapter 5 to the a model containing any number of categorical covariates and again showed that τ^2 is related to the deviance of the model of interest. In an example concerning Ear Infection in beach users, we saw that by including a number of subjective correlations in R we increased the discriminatory power of the model, although only by a small amount. Then in a simulation study, the properties of the shrinkage formulation were explored by simulating artificial data and then returning to

the Ear Infection data. The results show that in the situation we would expect to occur most of the time, where the model adequately fits the data the shrinkage estimates are better than the model or empirical estimates. This leads to the suggestion of a possible decision rule, based on the deviance of the model to indicate which set of estimates, model or shrunk to use.

The methodology described above is hopefully of some interest as it describes a technique to enhance the discriminatory power of a particular model by including prior information about the context of the data. We should note however that gains in predictive accuracy through the shrinkage approach are not substantial in the examples we have studied. For example, in the breast cancer study presented in Chapter 4, the retrospective ROC curve overestimates the discriminatory power of the score by 16%, while for the ear infection data in Chapter 6 we see that by including some reasonable correlations between ear infection rates in different categories we increase the discriminatory power by 3.6% (both figures in terms of the Gini Coefficient). This could be due in part to the fact that for all of the examples presented we have been unable to elicit a genuine prior correlation matrix from experts in the respective areas. We anticipate the application of the shrinkage methodology is likely to be useful in situations analogous to the OGRS score, where there are a large number of categories of interest (approximately 1000 in the OGRS case) and there is a realistic chance of the provision of the correlation

matrix. Finally, the shrinkage method is currently only defined for categorical variables with a binary outcome. We could look to generalise the methodology to include continuous measurements by altering some of the techniques that already exist for shrinkage in multiple regression.

Bibliography

Adams, N. M. and Hand, D. J (2000). An improved measure for comparing diagnostic tests. *Computers in Biology and Medicine*, 80 89 - 96.

Agresti, A. (1990). *Categorical Data Analysis.* Wiley

Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, **13** 469 - 475.

Angelos-Tosteson, A. N. and Begg, C. B. (1988). A general regression methodology for ROC curve estimation. *Medical Decision Making*, 8 204 - 215.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the the receiver operating characteristic curve. *Journal of Mathematical Psychology*, **12** 387 - 415.

Beck, J. R. and Shultz, E. K. (1986). The use of relative operating characteristic (ROC) curves in test performance evaluation. *Archives of pathology and laboratory medicine*, **110** 13 - 20.

Begg, C. B., Satogopan, J. M. and Berwick, M. (1998). A new strategy for evaluating the impact of epidemiologic risk factors for cancer with respect to melanoma. *Journal of the American Statistical Association*, **93** 415 - 426.

Berwick, M., Begg, C. B, Fine, J. A. and Roush, G. C. and Barnhill, R. L. (1996). Screening for cutaneous melanoma by self-skin examination. *Journal of the National Cancer Institute*, **88** 17 - 23.

Bishop, Y. M., Feinberg, S. E. and Holland, P. W. (1975) *Discrete Multivariate Analysis*. MIT Press

Campbell, G. (1994). Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine*, **13** 499 - 508.

Campbell, M and Machin, D. (1993). *Medical statistics, a commonsense approach*. Wiley.

Cheng, K. F and Hsieh, H. M. (1999). Correcting bias due to misclassification in the estimation of logistic regression models. *Statistics and Probability Letters*, **44** 229 - 240.

Choi, B. C. K. (1998). Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test. *American Journal of Epidemiology*, 148 1127 - 1132.

Copas, J. B. (1983). Regression, prediction and shrinkage (with discussion). *Journal of the Royal Statistical Society B*, 45 311 - 354.

Copas, J. B. (1987). Cross-validation shrinkage of regression predictors. *Journal of the Royal Statistical Society B*, 49 175 - 183.

Copas, J. B. (1999a). The effectiveness of risk scores: The logit rank plot. *Journal of the Royal Statistical Society C*, 48 165 - 184.

Copas, J. B. (1999b). Overfitting and the Stein factor. *Proceedings of the International Conference on Statistical Modelling*, 1 - 9.

Copas, J. B. and Stone, M. C. (1986). On the robustness of shrinkage predictors in regression to differences between past and future data. *Journal of the Royal Statistical Society B*, 48 223 - 237.

Copas, J. B, Marshall, P. and Tarling R. (1996). Predicting reoffending for discretionary conditional release. *Home Office Research and Statistics Directorate Report 150.*

Cox, D. R. and Snell, E. (1970). *Analysis of Binary Data.* Chapman and Hall.

Delong, E. R., Delong, D. M. and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a non-parametric approach. *Biometrics*, 44 837 - 845.

Dobson, A. J. (1990). *An introduction to generalized linear models.* Chapman and Hall.

Dorfman, D. D. and Alf, E. (1968). Maximum likelihood estimation of the parameters of signal detection theory - a direct solution. *Psychometrika*, **33** 117 - 124.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, **70** 893 - 898.

Efron, B. (1982). *The jackknife, bootstrap and other resampling plans.* Society for Industrial and Applied Mathematics.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement of cross-validation. *Journal of the American Statistical Association*, 78 316 - 331.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81 461 - 470.

Egan, J. P. (1975). *Signal detection theory and ROC analysis.* Academic Press.

Friedman, J. H. (1989). Regularized Discriminant Analysis. *Journal of the American Statistical Association*, **84** 165 - 175.

Gong, G. (1986). Cross-validation, the jackknife and the bootstrap, excess error estimation in forward logistic regression. *Journal of the American Statistical Association*, **81** 108 - 113.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics.* Wiley.

Hand, D. J. (1994a). Assessing classification rules. *Journal of Applied Statistics*, **21** 1- 16.

Hand, D. J. (1997). *Construction and Assessment of Classification Rules.* Wiley.

Hand, D. J. (2001). Measuring diagnostic accuracy of statistical prediction rules. *Statistica Neerlandica*, **55** 3 - 16.

Hand, D. J., Daly, F., McConway, K., Lunn, D. and Ostrowski, E. (1994b). *The Handbook of Small Data Sets.* Chapman and Hall.

Hand, D. J. and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society A*, **160** 523 - 541.

Hanley, J. A. (1988). The robustness of 'binormal' assumptions used in fitting ROC curves. *Medical Decision Making*, **8** 197 - 203.

Hanley, J. A. (1991). Verification bias and the one parameter logistic ROC curve - some clarifications. *Medical Decision Making*, **11** 203 - 207.

Hanley, J. A. (1996). The use of the 'binormal' model for parametric ROC analysis of quantitative diagnostic tests. *Statistics in Medicene*, **15** 1575 - 1585.

Hanley, J. A. and Hajan-Tilaki, K. (1997). Sampling variability of non-parametric estimates of the areas under the receiver operating characteristic curves: an update. *Academic Radiology*, **4** 49 - 58.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic curve (ROC) curve. *Radiology*, **143** 29 - 36.

Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, **148** 839 - 843.

Harrell, F. E., Lee, K. L., Mark, D. B., Califf, R. M., Pryor, D. B. and Rosati, R. A. (1984). Regression modelling strategies for improved prognostic prediction. *Statistics in Medicene*, **3** 143 - 152.

Harrell, F. E., Lee, K. L. and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy and measuring and reducing errors. *Statistics in Medicene*, **15** 361 - 387.

Hilden, J. (1991). The area under the ROC curve and its competitors. *Medical Decision Making*, **11** 95 - 101.

Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression.* Wiley.

Hsieh, F. and Turnbull, T. W. (1996). Nonparametric and semi-parametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*, **24** 25 - 40 .

Krebs-Brown, A. J. (2000). *Shrinkage and Calibration in Multiple Regression.* PhD Thesis. Warwick University.

Lang, B. J. and Aspelund, T. (1999). Binormal association - marginal models for ROC analysis. *Proceedings of the 11th International Conference on Statistical Modelling*, 251 - 258.

Lloyd, C. J. (1998). Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association*, **93** 1356 - 1364.

Lloyd, C. J. (2000). Maximum likelihood estimation of misclassification rates of binomial regression. *Biometrika*, **87** 700 - 705.

Lusted, L. B. (1971). Signal detectability and medical decision making. *Science*, **171** 1217 - 1219.

Mallows, C. L. (1973). Some comments on c_P. *Technometrics*, **15** 661 - 675.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models.* Chapman and Hall.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicene*, 8 283 - 298.

Metz, C. E. and Kronman, H. B. (1980). Statistical significance tests for binormal ROC curves. *Journal of Mathematical Psychology*, **22** 218 - 243.

Pearce, J. and Ferrier, S. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, **133** 225 - 245.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, **9** 705 - 724.

Press, S. J. and Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, **73** 699 - 705.

Raubertas, R. F., Rodewald, L. E., Humiston, S. G. and Szilagyi, P. G. (1994). ROC curves for classification trees. *Medical Decision Making*, **14** 169 - 174.

Smith, P. J., Thompson, T. J., Engelgau, M. M. and Herman, W. H. (1996). A generalized linear model approach for analysing receiver operating characteristic curves. *Statistics in Medicene*, **15** 323 - 333.

Stone, M. (1977). An asymptotic equivalence of choice by model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society B*, **39** 44 - 47.

Taube A. (1986). Sensitivity, specificity and predictive values: a graphical approach. *Statistics in Medicene*, **5** 585 - 591 .

Van Houwelingen, J. C. and Le Cessie, S. (1990). Predictive value of statistical models. *Statistics in Medicene*, **9** 1303 - 1325 .

Venables, W. N. and Ripley, B. D. (1997). *Modern Applied Statistics with S-PLUS*. Springer.

Wieand, S., Gail, M. H., James, B. R., James, K. L. (1989). A family of non-parametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, **76** 585 - 592 .

Zweig, M. H. and Campbell, G. (1993). Receiver operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicene. *Clinical Chemistry*, **39** 561 - 577.