Categorising variables in medical contexts

Douglas Campbell Watt B.Sc.(Hons)

A Dissertation submitted to the
University of Glasgow
for the degree of
Doctor of Philosophy

Department of Statistics

October, 1997

To my wife Linda and my mum and dad

Abstract

Many medical studies involve modelling the relationship between an outcome variable and a series of one or more continuous/interval scaled discrete explanatory variables. It is common practice in many of these studies for some, or indeed all, of the continuous/interval scaled discrete explanatory factors to be incorporated into the analysis in a categorised or grouped form.

One of the main reasons for adopting this methodology is that it will simplify the interpretation of results for clinicians and hopefully patients. It is often easier to interpret conclusions based on an explanatory variable with two or three levels (i.e. categorisations) than from a continuous/interval scaled discrete explanatory. The main drawback with this technique is in identifying the categorisation points. Often preconceived and/or historical grounds are the determining factor used to decide the location of these categorisation points. However, this may not give rise to sensible or justifiable locations for such points for a given application.

This thesis will consider the analysis of data from various types of medical study and, by applying non-parametric statistical methodology, provide alternative, more logical rationale for identifying categorisation points. The analysis will concentrate on data from three specific types of medical study - a cohort study with a binary outcome, a matched case/control study and survival analysis.

In a ***cohort study with a binary response*** the standard methodology of logistic regression will be applied and extended using a non-parametric logistic approach to identify potential categorisation points. As a further extension consideration will be given to the more formal methodology of examining the first derivative of the

resultant non-parametric logistic regression to provide the location of categorisation points.

In ***matched case/control studies*** the standard technique used for analysis is conditional logistic regression. The theory and application of this model will be discussed before considering two new, alternative, non-parametric approaches to analysing matched case/control studies with an ***interval scaled discrete*** explanatory variable. The proposed non-parametric approaches will be tested to investigate their usefulness in identification of categorisations for the explanatory variable. Possible extensions to these approaches to incorporate a single continuous explanatory variable will be discussed. In order to compare the two non-parametric approaches a simulation study will be carried out to investigate the power of these approaches.

Finally, consideration will be given to the ***analysis of survival data***. Initially, the standard methodologies of the Kaplan and Meier estimator in the absence of explanatory variables and Cox's Proportional Hazards model to incorporate explanatory variables will be discussed. A more detailed examination of three alternative methods for analysing survival data in the presence of a single continuous explanatory will be carried out. Each of the methods will be applied in turn to a survival analysis problem to investigate if any categorisations can be identified for a single continuous explanatory variable. Further simulations will be undertaken to compare the three methods across a variety of scenarios.

Contents

Chapter 1: Introduction 1

1.1 Background 1

1.2 Non-parametric Statistical Methodology 8

1.3 Summary of Chapters 15

Chapter 2: Cohort Studies 16

2.1 Introduction 16

2.2 The Linear Logistic Model 18

2.3 Prognostic factors for surviving stage 2 melanoma: An illustration of linear logistic regression 21

2.3.1 Introduction 21

2.3.2 Univariate analysis 22

2.3.3 Multivariate analysis 28

2.4 Non-parametric Logistic Regression 31

2.5 Prognostic factors for surviving stage 2 melanoma (revisited): An application of non-parametric logistic regression 34

2.5.1 Introduction 34

2.5.2 Univariate analysis 34

2.5.3 Multivariate analysis 38

2.5.4 Formal identification of categorisation points 43

2.5.4.1 One explanatory - the use of function derivatives 43

2.5.4.2 Two explanatories - the use of directional derivatives 47

2.5.4.3 Relevance of the number of malignant nodes for five year survival 53

2.6 Conclusions 57

Chapter 3: Case / Control Studies 59

3.1 Introduction 59

3.2 Conditional linear logistic model 61

3.2.1 The model 61

3.2.2 Conditional likelihood 62

3.2.3 Relative risk 64

3.3 Cutaneous malignant melanoma: An illustration of a case / control study 68

3.4 Non-parametric approaches to analysing data from a matched case / control study 74

3.4.1 Introduction 74

3.4.2 Pairwise cells comparison 74

3.4.2.1 Binary risk factor 75

3.4.2.2 Extention to a single interval scaled discrete risk factor 78

3.4.2.3 Confidence intervals for the relative risk 83

3.4.2.4 Inclusion of covariance terms 85

3.4.3 Conditional likelihood method 87

3.4.4 Nearest neighbour smoothing 92

3.5 Cutaneous malignant melanoma revisited: An application of non-parametric methods to analysing data from a case / control study 95

3.5.1 Introduction 95

3.5.2 Pairwise cells comparison 95

3.5.3 Conditional likelihood method 99

3.5.4 Summary 102

3.6 Isotonic regression 103

3.6.1 Introduction 103

3.6.2 Isotonic regression 103

3.6.3 Isotonic regression in practice 105

3.6.4 Summary 109

3.7 Extension to a continuous risk factor 110

3.7.1 Introduction 110

3.7.2 Extension to a continuous risk factor 110

3.7.3 Sun exposure and cancer risk 111

3.7.4 Summary 120

3.8 Simulation study 121

3.8.1 Introduction and examples 121

3.8.2 Summary of the results from the simulation study 165

3.9 Conclusions 168

Chapter 4: Non-parametric approaches to the analysis of survival data 170

4.1 Introduction 170

4.2 Standard approaches to the analysis of survival data 176

4.3 Analysis of a subset of Scottish melanoma group database 183

4.4 Non-parametric approaches to the analysis of survival data 191

4.4.1 Kaplan-Meier based approach 191

4.4.2 Non-parametric hazard based approach 196

4.4.3 Non-parametric logistic based approach 204

4.5 Illustration of the Non-parametric approaches to survival analysis 209

4.5.1 Melanoma example: Kaplan-Meier based approach 210

4.5.2 Melanoma example: Non-parametric hazard approach 214

4.5.3 Melanoma example: Non-parametric logistic survival approach 222

4.6 Simulation study 227

4.6.1 Scenario 1: Simulated data with no covariate effect 230

4.6.2 Scenario 2: Simulated data from a proportional hazards model 246

4.6.3 Scenario 3: Simulated data with a single categorisation point 258

4.6.4 Summary of the results from the simulation study 268

4.7 Conclusions 270

Chapter 5: Conclusions and future work 272

5.1 Conclusions 272

5.2 Future work 276

Appendices 278

A: Calculation of covariance terms in $\hat{V}_2(\hat{\underline{\beta}})$ 278

B: Derivation of the variance of $\hat{h}(t; z)$ 282

List of Tables

2.3.1 Number of Malignant nodes removed 26

3.8.1 Sample sizes and levels of smoothing used in simulation study 124

3.8.2 Summary of simulation scenarios 125

Simulation Study: Scenario 1

4.6.1 Parameter values used in simulations 232

4.6.2 Summary of observed follow-up times 232

Simulation Study: Scenario 2

4.6.3 Parameter values used in simulations 247

4.6.4 Summary of observed follow-up times 247

Simulation Study: Scenario 3

4.6.5 Parameter values used in simulations 259

4.6.6 Summary of observed follow-up times 259

List of Figures

2.3.1 Boxplot of Age at diagnosis 23

2.3.2 Plot of $\hat{p}(z)$ against age 24

2.3.3 Normal plot of Standardised Residuals with Simulated envelope 24

2.3.4 Plot of $\hat{p}(z)$ against Number of Malignant nodes 27

2.3.5 Plot of age against Number of Malignant nodes 30

2.3.6 5 year survival contours based on linear logistic model 30

2.5.1 Plots of $\hat{p}(z)$ against age for a selection of smoothing parameters 36

2.5.2 Plots of $\hat{p}(z)$ against Number of Malignant Nodes for a selection of smoothing parameters 38

2.5.3 Bivariate survival contours for a selection of smoothing parameters 40

2.5.4 Non-parametric logistic contours with Linear logistic contours superimposed 41

2.5.5 Three dimensional perspective plot of non-parametric logistic regression 41

2.5.6 Simultaneous plots of survival curves and 1st derivatives for age 46

2.5.7 Simultaneous plots of survival curves and 1st derivatives for Number of Malignant nodes 48

2.5.8 Three dimensional perspective plot of the grad function 50

3.3.0 Bivariate plot of the number of naevi for matched case/control pairs 69

3.3.1 Plot of Relative Risk against Number of naevi for males 72

3.3.2 Plot of Relative Risk against Number of naevi for females 72

3.4.1 Grid representation of case/control data 78

3.4.2 Display of neighbourhood sizes 93

Pairwise cells method:

3.5.1 Plot of Relative Risk against Number of naevi for males 96

3.5.2 Plot of Relative Risk against Number of naevi for females 96

3.5.3 Plot of Relative Risk against Number of naevi for males with kernel smoother added 98

3.5.4 Plot of Relative Risk against Number of naevi for females with kernel smoother added 98

Conditional Likelihood method:

3.5.5 Plot of Relative Risk against Number of naevi for males 100

3.3.6 Plot of Relative Risk against Number of naevi for females 100

3.5.7 Plot of Relative Risk against Number of naevi for males with kernel smoother added 101

3.5.8 Plot of Relative Risk against Number of naevi for females with kernel smoother added 101

Pairwise Cells Method

3.6.1 Isotonic regression of Relative Risk for males 107

3.6.2 Isotonic regression of Relative Risk for females 107

Conditional Likelihood Method

3.6.3 Isotonic regression of Relative Risk for males 108

3.6.4 Isotonic regression of Relative Risk for females 108

3.7.1 Boxplot of sun exposure 113

3.7.2 Bivariate plot of sun exposure 113

Pairwise Cells Method

3.7.3 Plot of Relative Risk against sun exposure - Category size = 5 hours 116

3.7.4 Plot of Relative Risk against sun exposure - Category size = 10 hours 117

Conditional Likelihood Method

3.7.5 Plot of Relative Risk against sun exposure - Category size = 5 hours 118

3.7.6 Plot of Relative Risk against sun exposure - Category size = 10 hours 119

3.8.1 Plot of the Relative Risk and logarithm of the Relative Risk 126

Simulation Study: Scenario 1

3.8.2 Plots of the average mean square error and empirical standard deviation of the mean square error 129

3.8.3 Plots of the average bias and empirical standard deviation of the bias 131

3.8.4 Plots of the coverage for the Conditional Likelihood method 133

3.8.5 Plots of the width of the interval for the Conditional Likelihood method 134

3.8.6 Plots of the coverage for the Pairwise Cells method 135

3.8.7 Plots of the width of the interval for the Pairwise Cells method 136

Simulation Study: Scenario 2

3.8.8 Plots of the average mean square error and empirical standard deviation of the mean square error 139

3.8.9 Plots of the average bias and empirical standard deviation of the bias 140

3.8.10 Plots of the coverage for the Conditional Likelihood method 142

3.8.11 Plots of the width of the interval for the Conditional Likelihood method 143

3.8.12 Plots of the coverage for the Pairwise Cells method 144

3.8.13 Plots of the width of the interval for the Pairwise Cells method 145

Simulation Study: Scenario 3

3.8.14 Plots of the average mean square error and empirical standard deviation of the mean square error 148

3.8.15 Plots of the average bias and empirical standard deviation of the bias 149

3.8.16 Plots of the coverage for the Conditional Likelihood method 152

3.8.17 Plots of the width of the interval for the Conditional Likelihood method 153

3.8.18 Plots of the coverage for the Pairwise Cells method 154

3.8.19 Plots of the width of the interval for the Pairwise Cells method 155

Simulation Study: Scenario 4

3.8.20 Plots of the average mean square error and empirical standard deviation of the mean square error 158

3.8.21 Plots of the average bias and empirical standard deviation of the bias 159

3.8.22 Plots of the coverage for the Conditional Likelihood method 160

3.8.23 Plots of the width of the interval for the Conditional Likelihood method 161

3.8.24 Plots of the coverage for the Pairwise Cells method 162

3.8.25 Plots of the width of the interval for the Pairwise Cells method 163

4.3.1 Kaplan Meier estimate of survival 184

4.3.2 Plots of Tumour thickness against follow-up time 186

4.3.3 Survival curves based on the Proportional hazards model 188

4.3.4 Plots of Deviance Residuals 189

Kaplan Meier based approach

4.5.1 Survival curves for a selection of smoothing parameters 211

4.5.2 Contour plot of the Survivor function 213

4.5.3 Confidence bands for the Survivor function 215

Hazard based approach

4.5.4 Survival curves for a selection of smoothing parameters 216

4.5.5 Contour plot of the Survivor function 220

4.5.6 Confidence bands for the Survivor function 221

Logistic based based approach

4.5.7 Survival curves for a selection of smoothing parameters 223

4.5.8 Contour plot of the Survivor function 224

4.5.9 Confidence bands for the Survivor function 225

Simulation study: Scenario 1

4.6.1 Three dimensional perspective plot of true surface 231

4.6.2 Sample plots of precision and bias for the Kaplan Meier based approach 234

4.6.3 Plots of precision against sample size 235

4.6.4 Plots of bias against sample size 237

4.6.5 Plots of coverage against sample size (Kaplan Meier based approach) 240

4.6.6 Plots of coverage against sample size (Hazard based approach) 241

4.6.7 Plots of coverage against sample size (Logistic based approach) 242

4.6.8 Hazard based approach: Confidence bands for simulated data 245

Simulation study: Scenario 2

4.6.9 Three dimensional perspective plot of true surface 248

4.6.10 Plots of precision against sample size 250

4.6.11 Plots of bias against sample size 252

4.6.12 Plots of coverage against sample size (Kaplan Meier based approach) 255

4.6.13 Plots of coverage against sample size (Hazard based approach) 256

4.6.14 Plots of coverage against sample size (Logistic based approach) 257

Simulation study: Scenario 3

4.6.15 Plots of precision against sample size 261

4.6.16 Plots of bias against sample size 262

4.6.17 Plots of coverage against sample size (Kaplan Meier based approach) 264

4.6.18 Plots of coverage against sample size (Hazard based approach) 265

4.6.19 Plots of coverage against sample size (Logistic based approach) 266

Acknowledgements

I wish to thank Tom Aitchison for his supervision, insight, friendship and above all, long-suffering throughout the duration of my Ph.D. I also acknowledge the financial support of the ESPRC who funded me throughout my degree.

I would also like to thank all the members of staff at the Statistics department in Glasgow who have given me so much support during my time here. I would particularly like to thank Dr. Marian Scott and Dr. Jim Kay for their invaluable and much appreciated assistance. I would also like to thank Dr. Stuart Young for sharing the "Ph.D. experience" with me. A special thanks to Mary Nisbet and Myra Smith for their wonderful friendship throughout my entire time at the University.

Finally, I would like to thank my wife Linda, my mum and dad and my Aunt Cathie. They have all been so supportive of me during the course of my Ph.D. The completition of this thesis is undoubtedly due to their love and faith in me.

Chapter 1

Introduction

Section 1.1. Background

In most fields of scientific research, but especially in medical research, observational and designed experiments often generate large data sets of quite complex structure. In the past the application of appropriate statistical methodology to the analysis of such data has often been neglected in that ***little*** or ***no statistical analysis*** has been carried out on the data collected in such studies. Yates and Healey (1964) stated "It is depressing to find how much good biological work is in danger of being wasted through incompetent and misleading analysis of numerical results". With the advent of more rigorous guidelines on the publication of results in medical journals (Altman et al (1983), Evans (1989)) the use of statistical techniques to analyse data from medical studies have become the "norm" and "incompetent" analyses are unlikely to be found in the current medical literature. In fact, in many medical journals it is now almost essential that a full statistical analysis is carried out before a piece of work can be published. For example, in the New England Journal of Medicine, Shepherd et al (1995) provided a detailed statistical analysis of the results

from the West of Scotland Coronary Prevention Study (WOSCOPS) and in the British Medical Journal Harper et al (1994) carried out an in depth statistical analysis of the effects of the dose of bendrofluazide on the levels of hypertension present in a group of diabetics in Northern Ireland. However, it remains the case that some of the statistical techniques employed in the analysis of the data in many current medical publications are still not as ***rigorous*** as the statistical world would desire (see Murray (1991a) and Altman (1994)). On some occasions the presentation of results are not as clear as statisticians would desire and, more worryingly, the techniques employed would sometimes appear to be inadequate to meet the needs of the study. In a review article in the British Journal of Surgery, Murray (1991b) outlined the statistical aspects which should be considered in any study before submission to a journal. This article covered all aspects of a study from presentation of results through to the consideration of methodological issues. This thesis will focus on ***three*** particular medical study frameworks and initially examine the standard statistical techniques which ***should be used*** on data collected from these studies. It will then consider ***alternative*** non-standard statistical techniques which may prove even more appropriate in highlighting particular features of any set of data.

The types of study under consideration here are as follows:

(i) ***Cohort study with a binary response***: In this type of study individuals are followed up over a period of time to identify individuals who develop the "disease of interest" in order to ascertain factors which may be of prognostic significance. At a specified point

in time the status of each individual (e.g. developed/not developed the disease of interest) is established and factors which affect this difference in individual status can then be investigated. This allows factors which may be important for ***prognostic outcome*** to be highlighted.

(ii) ***Case/Control Study***: In this type of study a group of individuals known to be suffering from the disease of interest are identified (the cases) and compared to a group of disease free individuals (the controls). An examination of factors which differ between the cases and the controls makes it possible to identify factors which influence the risk of disease. The precision of this type of study is often increased by pre-matching individual cases with individual controls for known or established risk factors (e.g. sex, age). The case/control study, matched or unmatched, allows factors which may be important for determining the ***risk*** of a disease to be highlighted.

(iii) ***Survival Analysis:*** Here interest is in determining factors which may have an effect on prognostic survival. A group of individuals known to be suffering from the disease of interest are followed up through time. Consideration is given to each individual's ***full*** survival profile across time. This will essentially lead to a rather complex analysis as the pattern across time must be considered for each individual. By giving consideration to the full survival profile it is

again possible to highlight factors which may be relevant for ***prognostic survival.***

In most medical studies, the methods used to analyse the data focus on applying standard statistical methodology. These will often involve the use of some form of parametric modelling of the relationship between the outcome variable (e.g. alive/dead, case/control) and any potentially important prognostic factors. In the ***cohort study with a binary response*** the standard model used to examine any underlying relationship between the response and the explanatories is often the ***linear logistic model*** (Cox (1970)). In a study into the development of toxoplasmic encephalitis in AIDS patients, Raffi et al (1997) fitted a linear logistic model to the data from the 186 patients in the study to identify risk factors for development. With ***case/control studies*** involving a continuous potential risk factor, the standard model is the ***conditional linear logistic model*** (Mantel (1973)). Schneider et al (1997) used a conditional linear logistic model in a study of factors influencing the incidence of Parkinson's disease where the cases were paired with age matched controls. Finally, in ***survival problems*** various models have been proposed to attempt to explain the underlying relationship between survival and any potential explanatory factors including ***the Cox Proportional Hazards model*** (Cox (1972)) and the ***accelerated failure time model*** (Cox and Oakes (1984)). Karpf et al (1997) carried out a meta-analysis on factors influencing the development of osteoporosis in postmenopausal women. The women were followed up for a period of time and proportional hazards models were used to identify factors which affect the development of osteoporosis among these women. Accelerated failure time models were used as an alternative to

proportional hazards models by Deredita et al (1996) in a survival study, in order to highlight prognostic factors for survival from colorectal cancer.

There is nothing inherently wrong with ***only*** giving consideration to the use of standard parametric statistical techniques for analysis, and it is reassuring to see the use of formal statistical methodology in the papers mentioned in the previous paragraph. However, the idea of forcing the use of a specific parametric model to explain any underlying relationship is rather restrictive and does not allow consideration to be given to the vast number of possible relationships which may exist. It may be that the development of sophisticated non-parametric techniques will offer valuable and perhaps complementary alternative forms of analysis. This thesis will attempt to move away from this rather constricting parametric framework and consider the use of non-parametric models to attempt to explain any underlying relationships. Non-parametric modelling of any underlying relationships will allow far more flexible and varied relationships to be considered. Such models are data driven and hence allow the data itself to indicate the nature of any existing relationships. The use of this wider class of non-parametric models will, by definition, produce far more "open ended" solutions (Simonoff (1996)) than can be produced using the corresponding parametric models. Therefore, care must be taken to ensure that both a ***sensible*** and ***appropriate*** non-parametric model is chosen to explain any underlying relationship between the covariate and the explanatories.

Many medical studies involve modelling the relationship between the outcome variable and a series of one or more ***continuous*** or ***interval scaled discrete*** explanatory factors. It is common practice in many of these studies for some, or indeed all, of the continuous explanatory factors to be incorporated into the analysis in a ***categorised or grouped form.*** In a study into causes of high blood cholesterol, Grundy and Vega (1990) grouped the continuous explanatory factor age into 3 categories (20-29 years, 30-39 years, >39 years) and even the response factor, cholesterol level, was grouped into 3 categories (desirable, borderline high, high). Doll et al (1994) carried out an investigation into the effect of alcohol consumption on mortality where alcohol consumption was grouped into 4 categories (1-14 units/week, 15-28 units/week, 29-42 units/week, >42 units/week). In these publications categorisation points are therefore chosen for the continuous explanatories ***before*** any subsequent analysis is carried out. One of the main reasons for adopting this methodology is that it will simplify the interpretation of results for clinicians and hopefully patients. It is often easier to interpret conclusions based on an explanatory variable with two or three levels (i.e. categorisations) than from a continuous explanatory. The main drawback with this technique is in identifying appropriate categorisation points. If this methodology is indeed to be employed then it would appear logical that these categorisation points should be chosen at values of the explanatory at which there is a marked change in the effect that the explanatory has on the response. For example, in cancer studies there may be a marked change in the incidence of the disease at a particular value (or over a short range of values) of a continuous explanatory. This would therefore appear to be a sensible location for a categorisation point. In many studies, however, the data itself provides little

justification for the categorisations used. One possible explanation for this is that preconceived and/or historical grounds are often the determining factor used to decide the location of these categorisation points. However, this may not give rise to sensible locations for such points from a particular data set. In a large scale study into the association of blood pressure with cancer incidence and mortality Grove et al (1991) considered two ***different***, historically motivated, categorisations for each of systolic and diastolic blood pressure. Conclusions were then drawn based on ***both*** categorisations. A more sensible, alternative method for deciding upon the location of these categorisation points would be to allow the data itself to determine the location of the categorisation points. Areas of the data where there appears to be a ***change*** in the effect the explanatory has on the response would seem to imply the location of a categorisation point. The use ***of non-parametric methodology*** as mentioned previously provides a possible approach to allow the data itself to determine the number and location of any categorisation points.

The ***major aim of this thesis*** is to consider the analysis of data in various medical contexts and, by applying non-parametric statistical methodology, ***to highlight the location of any potential categorisation points***. Such a choice of categorisation points will surely have more credence than any which have been chosen without giving due consideration to the data itself.

The next section will give a brief introduction to the use of non-parametric statistical techniques.

Section 1.2: Non-parametric statistical methodology

The use of non-parametric statistical techniques came into prominence after the end of the Second World War. Wilcoxon (1945) and Tukey (1949) devised distribution free tests for examining a single sample location problem whilst Mann and Whitney (1947) derived a solution for the two sample location problem. A possible non-parametric solution for Analysis of Variance was proposed by Kruskal (1952) and Wallis (1952). Regression problems were first addressed by Thiel (1950) who introduced non-parametric tests and confidence intervals for the slope in a simple linear regression model with a continuous response and one continuous explanatory. Hollander (1970) considered a non-parametric test for parallelism in simple linear regression problems. These were developments for ***specific*** tests in linear regression problems but in a much more general sense the predominant emphasis of the current work in non-parametric statistics considers more detailed solutions to the general regression problem. One area which has attracted considerable research is the use of ***data smoothing techniques*** in both density estimation and non-parametric regression.

Rosenblatt (1956) and Parzen (1962) produced the first fully ***non-parametric regression model*** (The Rosenblatt-Parzen kernel density estimator) which fits a smooth regression curve (not necessarily a straight line) to a set of data. The proposal of a smooth regression curve with no parametric constraints was totally unique at this time. However, since this first model was derived, many other alternative smooth regression models based on the kernel density approach (see later) of Rosenblatt and

Parzen have been suggested as a solution to this problem. Priestley and Chao (1971) and Gasser and Muller (1979) provided other, more complex, alternative structures for the form of the smooth regression curve. One of the models most in common usage today was independently derived by Nadaraya (1964) and Watson (1964). Similar to the Rosenblatt-Parzen kernel density estimator it produces an estimate of the mean response based on fitting a smooth regression curve across the explanatory variable.

Formally, Nadaraya and Watson proposed the following non-parametric estimate of the mean response, y, based on a single continuous explanatory, z.

$$\hat{y} = \frac{\sum_{i=1}^{n} y_i \, \Delta_h(z, z_i)}{\sum_{i=1}^{n} \Delta_h(z, z_i)} \qquad -(1.1)$$

where

The weighting function, $\Delta_h(z, z_i) = K(\frac{z - z_i}{h})$

n is the number of observations

K is a smooth probability density function

h is the smoothing parameter

z_i is the continuous explanatory value for i^{th} observation

y_i is the continuous response value for i^{th} observation

This model non-parametrically smooths across the explanatory variable in order to determine the estimated value of the response at each value of the explanatory. It allows each observation's response to have an influence in a neighbourhood of its explanatory value; an influence which decreases as you move away from the value of the explanatory. The nature and degree of the influence are determined by the form of the weighting, or ***kernel***, function and the value of the smoothing parameter.

In essence this method uses the data to produce a weighted average of the response at each value of the explanatory. This technique will be used extensively throughout the work presented in this thesis as it allows the data itself to control the pattern in the response across the values of the explanatory. Hence, any unusual patterns in the data will become immediately obvious. In the determination of categorisation points (as discussed in Section 1.1), this technique will clearly prove very useful as categorisation points ***should*** be located at areas of the explanatory where there is a ***marked change*** in the value of the response (i.e. an ***unusual pattern***).

Although the model suggested by Nadaraya and Watson in (1.1) will not be directly applied in this thesis, the ideas contained within it provide basic building blocks which will be heavily relied upon in any non-parametric models which are used here. In each of the three study frameworks discussed here, non-parametric estimators will be proposed to describe the relationship between the response variable and any explanatory variables. These estimators will provide smooth estimates of the response across the values of the explanatory variable along the lines of Nadaraya and Watson. By a close examination of the resulting estimates any potential

categorisation points can be highlighted at values of the explanatory where to be a dramatic change is suggested in the value of the response.

Any smooth non-parametric estimator of the underlying relationship between the response and a continuous explanatory will involve the use of some form of weighting function; in essence similar to the function defined in (1.1). In this thesis the use of both ***kernel weighting functions*** (Rosenblatt (1956)) and ***nearest neighbour weighting functions*** (Loftsgaarden and Quesenberry (1965)) will be considered where appropriate.

A vast number of possible ***kernel weighting functions*** have been proposed. These range from the Epanechnikov parabolic kernel suggested by Epanechnikov (1969) and Barlett (1963) to the multidimensional product kernel function (Cacoullos (1966)). Here a standard Gaussian kernel (Silvermann (1986)) will be used as it has been shown that the choice of kernel has remarkably little effect on the estimates obtained (Hardle (1990)). Kernel weighting functions produce a weighted average of the response in a ***fixed*** neighbourhood around the explanatory value. An alternative weighting function is to consider a weighted average of the response in a ***varying*** neighbourhood around the explanatory value. This type of weighting function is known as a ***nearest neighbour weighting function***. Again, various forms of nearest neighbour weighting functions have been proposed. Yang (1981) proposed a symmetrized nearest neighbour weighting function, whilst Stone (1977) suggested both triangular and quadratic nearest neighbour weighting functions. Here, an adapted version of the most straightforward nearest neighbour weighting function, the

"uniform" weight, as suggested in density estimation by Loftsgaarden and Quesenberry (1965) will be used. The method suggested by Loftsgaarden and Quesenberry will be adapted slightly to reflect the data structure(s) being considered in the work presented here.

Non-parametric estimators are strongly influenced by the underlying pattern in the data. Estimates of the response at a value of the explanatory are calculated by smoothing the observed values of the response in a neighbourhood of the explanatory value. Both kernel and nearest neighbour weighting functions incorporate a parameter which determines the ***level*** of smoothing carried out on the data. On many occasions these estimators will be used in situations where the data itself is very sparse. This situation is particularly common in the analysis of data from case/control studies as these are often used in the study of rare diseases (i.e. relatively few observations will be present). In situations where data is very sparse this issue of smoothing the data becomes of particular importance. The less data that is present, the greater the degree of smoothing that is required to obtain a clear picture of any underlying relationships. The degree of smoothing is controlled by the smoothing parameter which is represented by the value h in (1.1). However there is then concern with establishing the "best" smoothing parameter for any given data set. If the data is undersmoothed (i.e. the smoothing parameter is too small) then a very jagged picture of any underlying relationship between the response and the explanatories will be produced. Conversely, if the data is oversmoothed (i.e. the smoothing parameter is too large) a clearer picture of the relationship will be produced, but with the possible consequence of smoothing out potentially important local features of the data (e.g. possible

categorisation points). Hence, it is crucial that the correct smoothing parameter is chosen on each occasion. Various methods have been suggested for obtaining the ***optimal smoothing parameter.***

The most common methods are based on cross-validation (Clark (1975)), penalizing functions (Shibita (1981)) and plug-in methods (Gasser et al (1991)). There are theoretical justifications in terms of the degree of differentiability of the final smoothed curve (Hardle (1990)) to prefer either the cross-validation or penalizing functions methods. In the work presented here, the cross-validation and approach to choice of smoothing parameter was given due consideration. Unfortunately this method tended to produce choices for the smoothing parameter which were too large and hence oversmoothed the results. As the main aim of this thesis is concerned with local features of the data it is essential not to oversmooth any underlying relationship. Therefore, this method was rejected and instead a simple ***subjective search method*** will be used to identify the optimal smoothing parameter. In other words, plots of the smooth estimate of the underlying relationship between the response and the explanatory will be produced for a ***range*** of sensible smoothing parameters and an appropriate value for the smoothing parameter will be chosen in light of these graphs. Once a final solution has been found, the graph of the results will be examined to investigate if any possible categorisation points can be found for the potential explanatory variable.

In summary, for each of the study frameworks under consideration here, non-parametric methodology will be used to examine the relationship between the

response and explanatory variables. Use of such methodology will make it possible to identify categorisations, if any exist, for the explanatory variable.

Section 1.3 Summary of Chapters

CHAPTER 2 considers the analysis of data from cohort studies with a binary response. In this chapter both the linear logistic model and non-parametric logistic model (Copas (1983)) will be applied to a problem within the field of medical research. Consideration is also given to the use of function derivatives in order to identify categorisation points.

CHAPTER 3 examines the case/control study and initially outlines the standard methodology based on the conditional linear logistic approach. Two less restrictive non-parametric approaches to the analysis of such data are also proposed. Attention is given to highlighting any possible categorisations for ***interval scaled discrete*** explanatories. A comparison of the two non-parametric methods is presented, based on a simulation study.

CHAPTER 4 gives full details of the standard analysis of data from a survival study. Three possible non-parametric approaches to the analysis of survival data are then presented. Notice is taken of any possible categorisations for ***continuous*** explanatories. The non-parametric methods are compared in terms of how they perform with both real and simulated data.

Finally, CHAPTER 5 presents a summary of the findings of the previous three chapters.

Chapter 2

Cohort Studies

Section 2.1: Introduction

Many types of observational study exist but by far the most common is the cohort study since it is the easiest to both design and organise.

A ***cohort*** is basically a group of individuals who are traced over a period of time (Campbell and Machin (1993)). A ***cohort study*** involves following a group of individuals over a period of time and recording various pieces of information on them before and through time. Subsets of the cohort under study can be identified who have been exposed to various factors which may influence the probability of occurrence of the disease under study. This makes it possible to obtain information about the occurrence of the disease under study and also to identify potential risk factors for the disease.

There are two separate and distinct classes of cohort study, these being

(1) Historical cohort study: Analysis is carried out upon data obtained from historical records and then it is possible to examine how

certain characteristics affect the occurrence of the disease of interest. A major advantage of this type of cohort study is that the results are available almost immediately. However since the data are obtained from large historical databases one possible disadvantage is that a lot of superfluous information will also be used in the analysis.

(2) Prospective cohort study: Here information is collected on subjects in the present and they are then followed up in to the future. Here a major advantage is that it is then possible to collect exactly the information thought to be required. The major drawback is that it may take years for any potential results to become available.

This chapter will demonstrate various ways to analyse data from a cohort study initially outlining, in sections 2.2 and 2.3, the approach of using the linear logistic model.

The emphasis of the work in this thesis is to develop methods of analysis which will allow the highlighting of potential categorisations for variables. To this end consideration will be given to a non-parametric method first devised by Copas(1983) which will be explained in section 2.4 while an example will be outlined in section 2.5.

Section 2.2: The Linear Logistic model

Regression methods are one of the most important techniques when considering any data analysis which involves describing the relationship between a response variable and a set of one or more potential explanatory variables. Often, as in the case of cohort studies involving the identification of risk factors for a particular disease, this response is discrete taking two values (e.g. did/did not develop disease). The standard model used for analysis in this situation is the binary linear logistic model (Cox 1970) which is defined as follows:

Let y_i represent the response variable indicating whether ($y_i = 1$) or not ($y_i=0$) the i^{th} individual develops the disease during the study period and let the p explanatory variables $z_{i1}, \ldots, z_{ip}$ be a set of p characteristics for each subject such as age, height, sex, etc. Then

$$p_{\underline{z_i}} = pr(y_i = 1 / \underline{z_i}) = \frac{\exp(\underline{\beta}^T \underline{z_i})}{1 + \exp(\underline{\beta}^T \underline{z_i})} \qquad - (2.1)$$

where

$$\underline{\beta}^T = (\beta_0 \quad \beta_1 \quad \ldots \quad \beta_p)$$

$$\underline{z_i}^T = (1 \quad z_{i1} \quad \ldots \quad z_{ip})$$

or equivalently the log odds on having the disease is linear in the explanatories

$$\text{i.e.} \quad \text{logit}(p_{\underline{z}_i}) = \log\left(\frac{p_{\underline{z}_i}}{1-p_{\underline{z}_i}}\right) = \underline{\beta}^T \underline{z}_i \qquad \text{- (2.2)}$$

Interest is then in estimating $\underline{\beta}$ and hence allowing modelling of $p_{\underline{z}}$.

The standard approach would be to maximise the likelihood function $\text{Lik}\left(\underline{\beta}; \underline{y}, \underline{z}\right)$ to produce estimates $\hat{\underline{\beta}}^T$ for $\underline{\beta}^T$ (Cox (1970)) and then use the logistic transformation in (2.1) to obtain an estimate $\hat{p}_{\underline{z}}$ for $p_{\underline{z}}$.

Further, approximate confidence bands for $p_{\underline{z}}$, for any $\underline{z}$, can be obtained by using the inverse of the information matrix $I\left(\underline{\beta}\right)_{\underline{\beta}=\hat{\underline{\beta}}}$, $\left(\text{i.e.} \quad I^{-1}\left(\hat{\beta}\right)\right)$, as an approximation to $\text{cov}(\hat{\underline{\beta}})$ (Kalbfleisch (1985)) to produce $100*(1-\alpha)\%$ confidence bands for $\underline{b}^T\underline{\beta}$ (and hence any component of $\underline{\beta}$) of the form

$$\underline{b}^T\hat{\underline{\beta}} \pm z_{\alpha/2}\sqrt{\underline{b}^T I^{-1}\left(\hat{\beta}\right)\underline{b}}$$

where

$$\underline{b}^T = (1 \quad z_1 \quad \ldots \quad z_p)$$

$z_{\alpha/2}$ is the $100*\left(1-\alpha/2\right)$ percentage point of the standard normal.

The logistic transformation can then be used to produce induced approximate confidence bands for $p_{\underline{z}}$.

Having fitted a linear logistic model some method of assessing the fit of the model is required. The assessment of the fit of the model could be broken into two separate sections, firstly to look at any individual observations which may be causing problems and secondly some formal test of the overall fit of the model. For simple linear regression the most common method used to examine model inadequacies is to look at residuals. There are two residuals which are commonly used to assess the fit of the ***linear logistic model.***

(i) The Pearson residual (McCullagh & Nelder (1990))

(ii) The Deviance residual(Pregibon (1981))

Calculation of these residuals allows any unusual observations to be identified and χ^2 tests based on the residuals (Hosmer and Lemeshow (1989)) can be constructed to test the overall fit of the model.

This section has outlined some of the simple theory involved in the use of the linear logistic model. More aspects of the inference involved in this type of modelling can be found in Breslow and Day(1980), Mike and Stanley(1982), Carter et al(1983), and Hosmer and Lemeshow(1989). For present purposes however the theory presented above will suffice as it gives sufficient background for the following examples.

Section 2.3: Prognostic factors for surviving stage 2 melanoma: An Illustration of linear logistic regression.

Section 2.3.1: Introduction

A useful data set to illustrate the techniques discussed in section 2.2 can be found in a paper by Tillman et al (1991). In this paper interest lies in identifying potential prognostic factors for surviving stage 2 melanoma. The data came from a prospective cohort study where the outcome of 109 patients undergoing therapeutic lymphadenectomy for clinical stage 2 malignant melanoma was assessed. The outcome chosen was whether or not the patient was alive five years after being identified as a stage 2 melanoma. Note that this choice of five years involves a rather arbitrary cutpoint in order to simplify presentation. Chapter 4, which deals specifically with survival analysis, will examine ways of justifying such a cutpoint. However, when considering outcome after five years, Tilmann et al identified 2 main prognostic factors, these being

(i) The age of the subject on being identified as a stage 2 melanoma

(ii) The number of malignant nodes the subject had surgically removed

Within the next section (section 2.3.2) a full univariate analysis of both of these factors will be carried out utilising the techniques described in section 2.2 whilst a later section (section 2.3.3) will give a slightly briefer outline of the multivariate analysis.

Section 2.3.2: Univariate analysis

This section will consider firstly any possible effect on outcome after five years of the age that the subject was when diagnosed stage 2 melanoma.

To gain an initial feel for the data consider a simple boxplot of the data and some summary measures. Of the 109 patients in the study only 24 were still alive five years after being diagnosed as stage 2 melanoma with the remaining 85 having died at some point within five years of diagnosis. Figure 2.3.1 presents a boxplot of the age of each subject when diagnosed stage 2 melanoma against their outcome after five years. The Figure suggests that although there are far more subjects who did not survive five years those who ***did survive five years*** appear to have been diagnosed stage 2 melanoma at a ***younger age***. This impression is backed up by the fact that that the mean age for those who did not survive five years was approximately 52 whereas for those who did survive the mean age was some 11 years younger at approximately 41 years of age. This seems to imply some difference between the two groups with the logical conclusion being that the younger the subject is when diagnosed stage 2 melanoma the better their prospects of surviving five years appear to be.

A linear logistic model was then fitted, and this confirmed the subjective impression with age having a significant effect (p-value = 0.0006). In Figure 2.3.2 the continuous curve provides a plot of $\hat{p}_z$ vs age with approximate 95% confidence bands for p_z shown by the dotted lines. This plot clearly shows how the probability of surviving five years ***decreases as the age of the subject increases***, confirming the subjective impression.

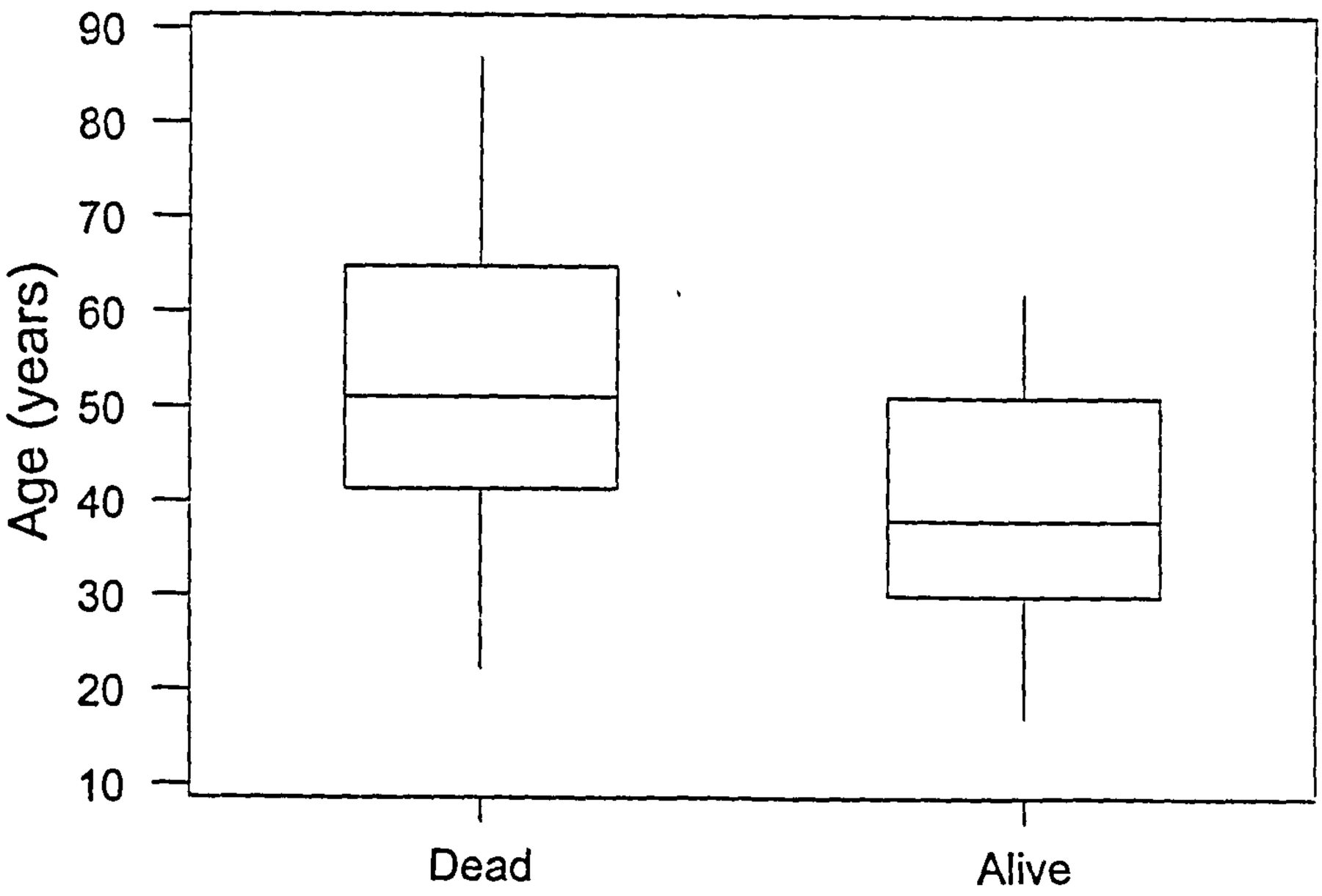

90
80
70
60
50
40
30
20
10
Age (years)
Dead
Alive
Status after 5 years

Figure 2.3.1

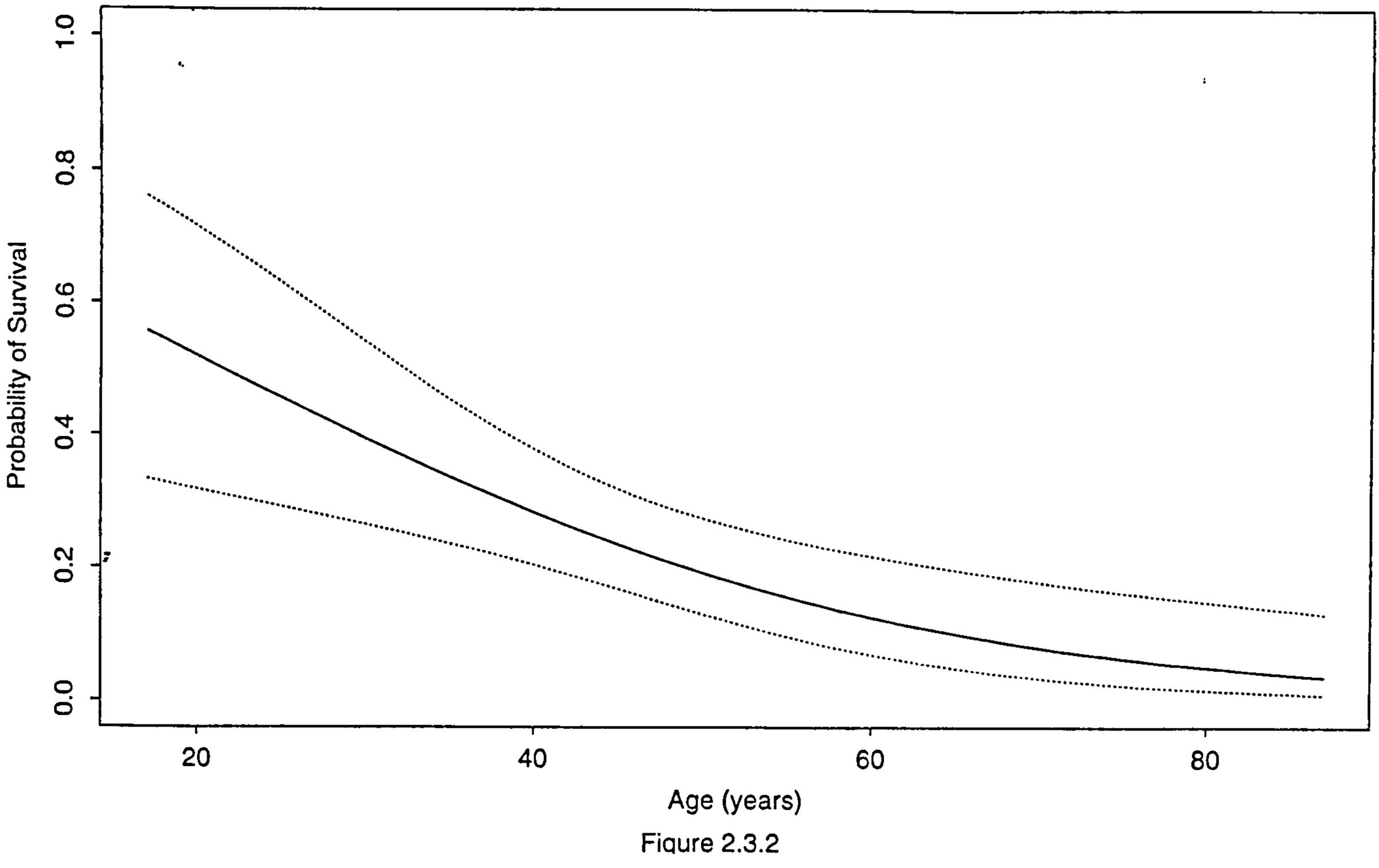


Figure 2.3.2

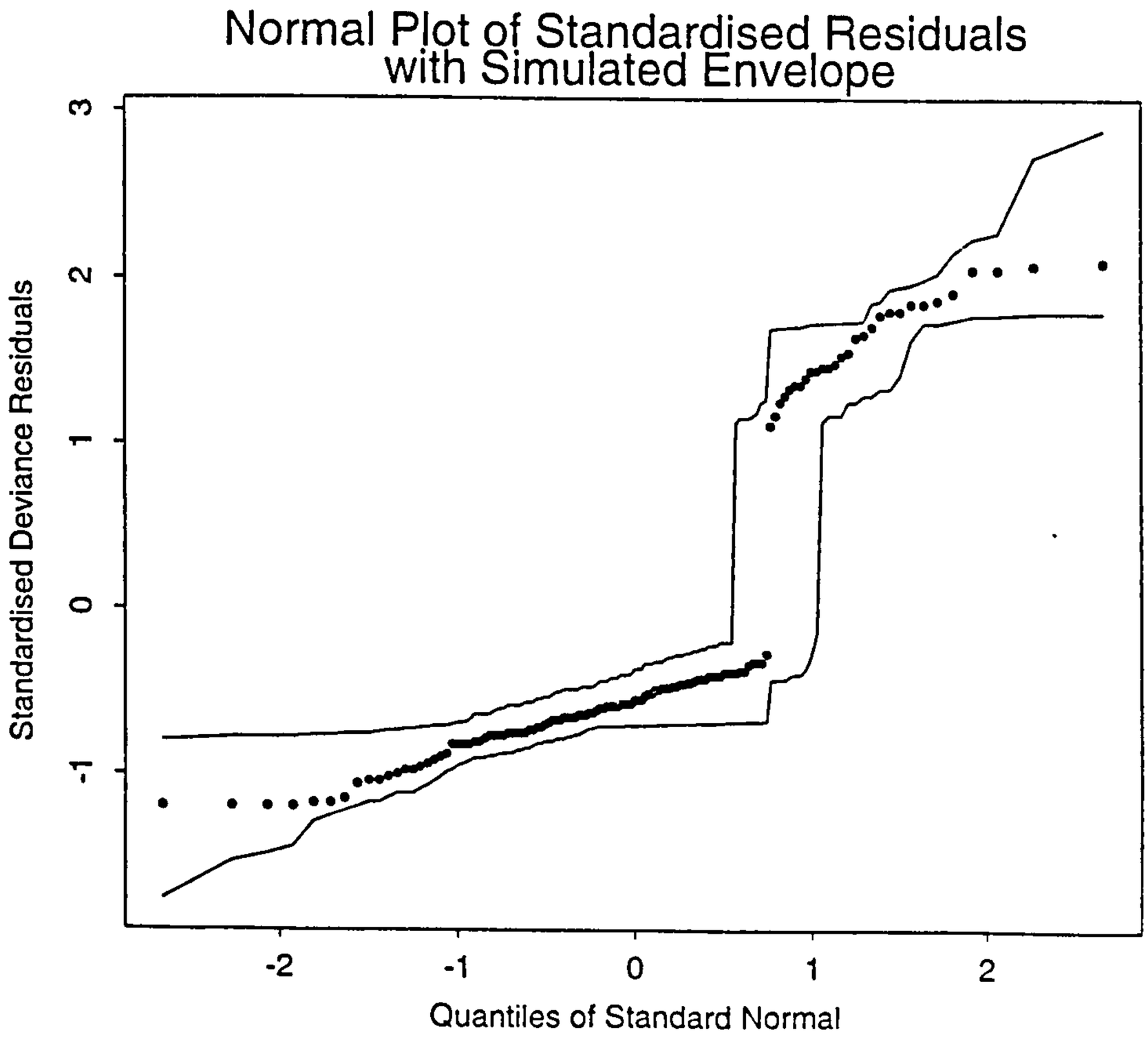


Figure 2.3.3

In order to assess the fit of the model the ***Deviance*** residuals were calculated. Consider Figure 2.3.3 which is a probability plot of the standardised Deviance residuals. The Figure incorporates a simulation envelope which gives appropriate values for such a plot based on simulating residuals from appropriate binomial distributions (Everitt (1994)). If any points lie outside this envelope then these may be potential outliers and, if a number of points lie outside the envelope then the assumptions underlying the model may be doubtful.

As ***none*** of the points lie outside the envelope it would appear that the logistic model gives a reasonable fit to the data and also that there is little evidence of any outliers. As a formal test of the fit of the model the Hosmer and Lemeshow test (Hosmer and Lemeshow (1989)) produces a p-value of 0.232 confirming that the logistic model gives a good fit to the data.

The second important prognostic factor was the number of malignant nodes the subject had surgically removed. It should probably be noted that this is more of an interval scaled discrete variable, having only 11 distinct categories, compared with age which was clearly continuous with 51 distinct ages among the 109 patients. Subjectively it would seem that the ***more nodes a subject had removed*** the more likely the ***disease was 'widespread'*** in the patient. Thus it may seem realistic to expect this variable to have an effect with subjects still alive after 5 years likely to have had less nodes removed.

Table 2.3.1 shows the number of nodes removed for each of the two groups. For those with 1 node removed the odds of a subject dying seem to be roughly 3 to 1 against (27 dead, 9 survivors) whereas if one considers those with more than 5 nodes removed the

odds of a subject dying appear to rise sharply to 18 to 1. This seems to imply that number of nodes removed does indeed have a ***detrimental effect*** on the probability of being alive after five years.

		Number of subjects		
		Dead	Survivors	Total
Number of nodes removed	1	27	9	36
	2	17	7	24
	3	17	4	21
	4	4	2	6
	5	2	1	3
	6	5	0	5
	7	4	1	5
	8	3	0	3
	10	3	0	3
	13	2	0	2
	20	1	0	1
	Total	85	24	109

Table 2.3.1

These tentative conclusions are ***partially*** backed up by the linear logistic model with the number of malignant nodes having an effect, although possibly marginal (p-value = 0.0933. The coefficient for nodes in this model was negative indicating poorer chances of being alive as the number of nodes increases. The continuous curve in Figure 2.3.4

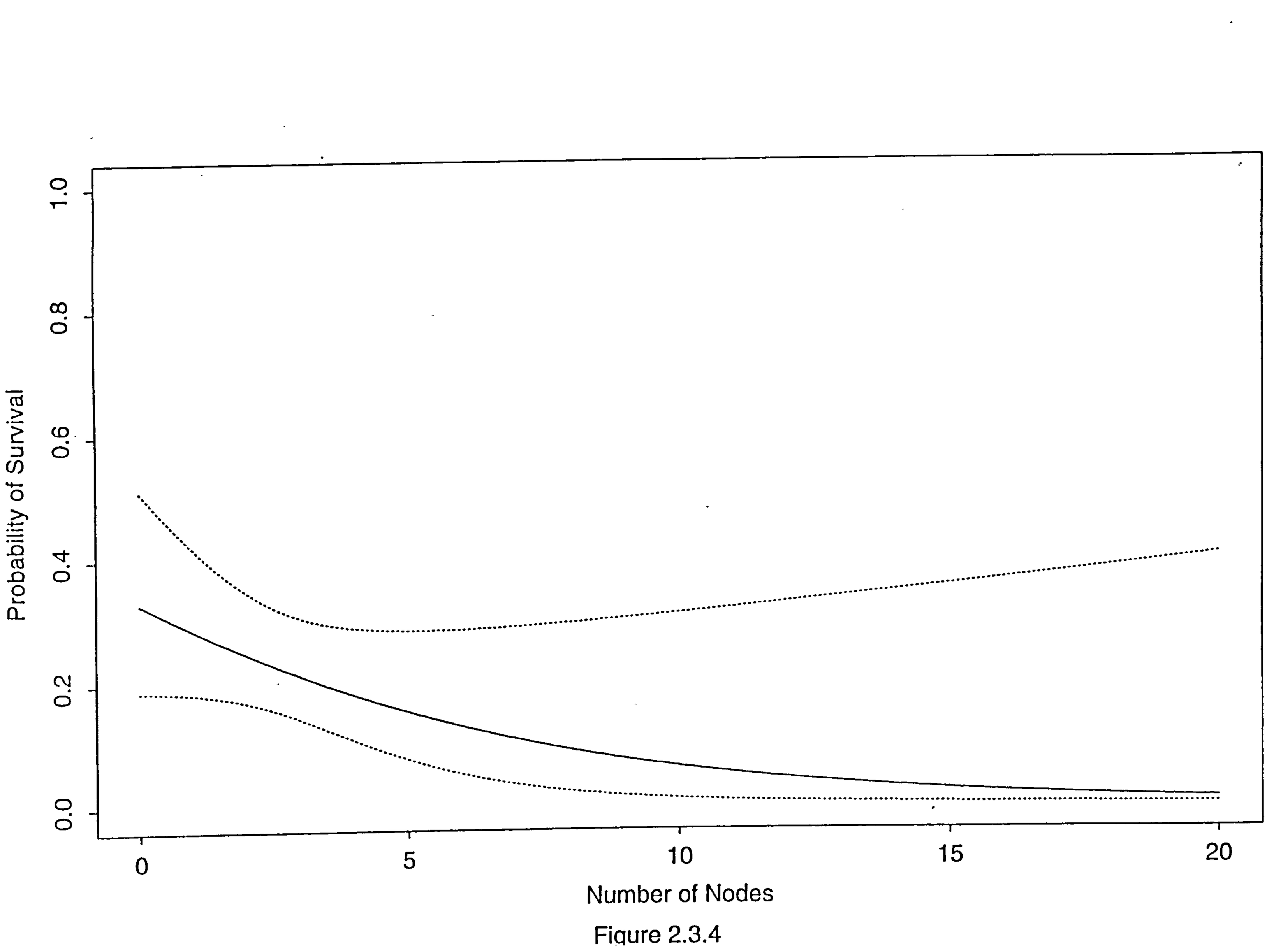

Probability of Survival
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
Number of Nodes

Figure 2.3.4

provides a plot of $\hat{p}_z$ against number of nodes with approximate 95% confidence bands for p_z shown by the dotted lines. Notice that the confidence bands become increasingly wide for more than 8 nodes removed due to the lack of data in these areas suggesting that any inferences should be very tentative in these data-sparse areas.

The Deviance residuals were examined and a formal Hosmer and Lemeshow test carried out and these both confirmed the adequacy of the fit of the logistic model.

Section 2.3.3: Multivariate Analysis

This section will give a brief description of a multivariate analysis of the data set. Figure 2.3.5 shows a plot of age against number of malignant nodes labelled by the status of the subject five years after being diagnosed stage 2 melanoma. Of the 109 subjects in the study 85 were dead after five years (those subjects marked with a D) and 24 survived at least five years (those subjects marked with an A). The majority of the subjects who survived at least five years (i.e. the A's) are located towards the bottom left of the plot. This plot appears to suggest that only those subjects who have had a ***few nodes surgically removed*** and were ***relatively young*** on being diagnosed stage 2 melanoma have ***any realistic chance of being alive after five years***. Note that there are also only 6 subjects who have had 10 or more nodes removed with the majority of subjects having had between 1 and 7 nodes removed. All subjects who had 10 or more nodes removed failed to survive five years suggesting very poor, if any, prospects of five year survival for subjects who have had many nodes surgically removed.

When ***only*** five year outcome was examined and a multivariate logistic model was fitted only ***age was significant*** and the ***number of nodes did not quite prove significant.*** Although number of nodes did not prove significant it will still be included with age in the model and a later section (Section 2.5) will return to this issue and give a discussion of the significance of number of nodes in terms of predicting five year outcome. If number of nodes is included it is then possible to produce estimates of the probability of surviving at least five years for this bivariate model. Using the fitted model various contours of the probability of surviving at least five years were constructed and are displayed in Figure 2.3.6.

On this plot contours are drawn at 15, 25 and 35% probability of five year survival for the fitted bivariate model and this shows that the overall prospects for subjects are not particularly good especially as the subject gets older. According to the linear logistic model once a subject is diagnosed as a stage 2 melanoma at older than approximately 40 years of age then they have a less than 25% chance of surviving five years regardless of the number of nodes.

Although the analysis carried out in this section is useful it does not provide a very simple explanation of how five year survival from a stage 2 melanoma depends upon age and the number of nodes removed. A more easily digestible conclusion might be that 'reasonable survival' only occurs for, say, those younger than 40 years of age and with fewer than 3 nodes removed. This however necessitates 'categorisation' of both variables and justification of such. The linear logistic model is a very restrictive model and as such cannot highlight any unusual patterns in the data which may suggest areas for categorisation. Hence the next section will introduce a data fitting method which ***will allow*** possible categorisations to be highlighted.

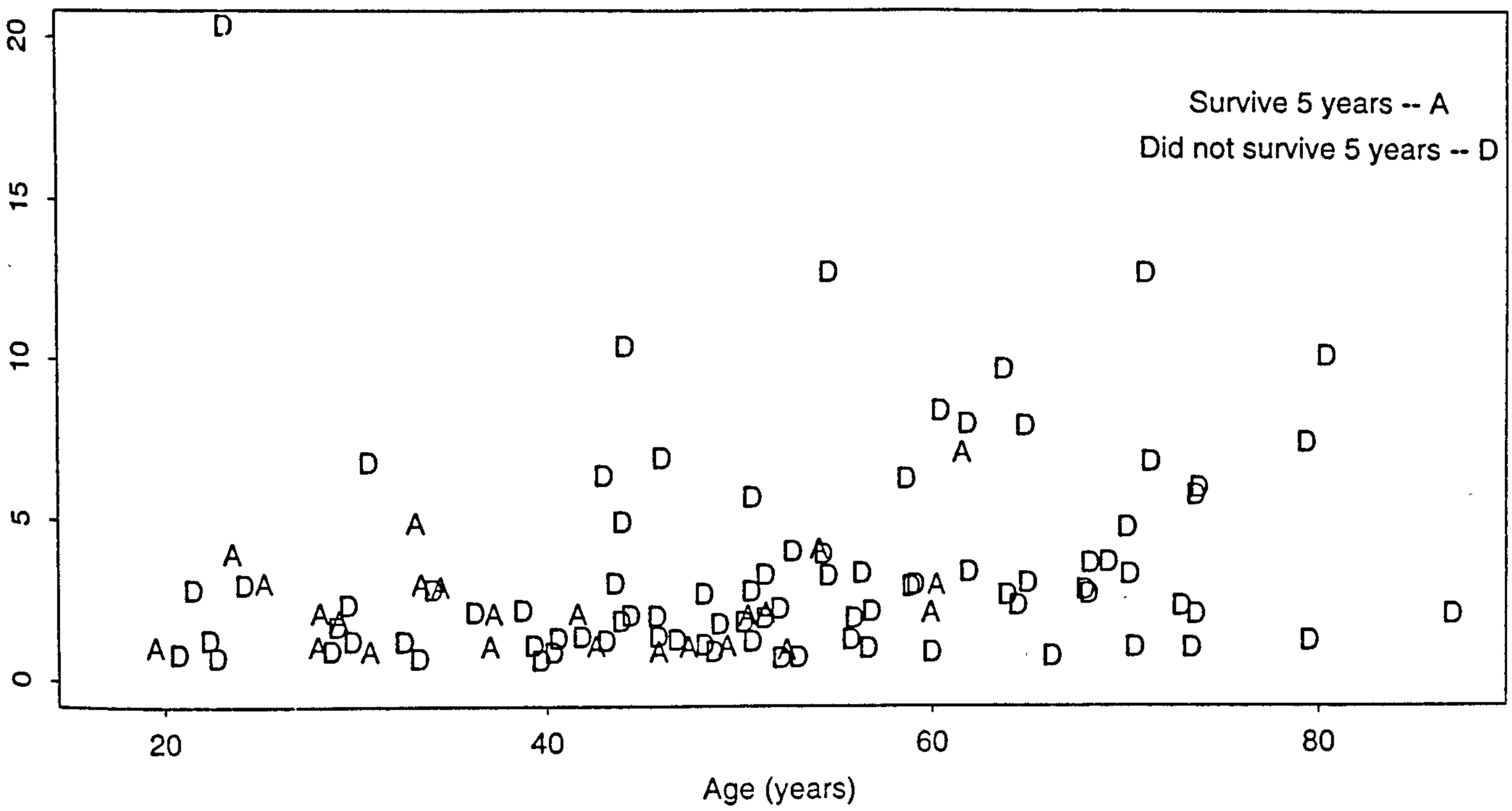


Figure 2.3.5

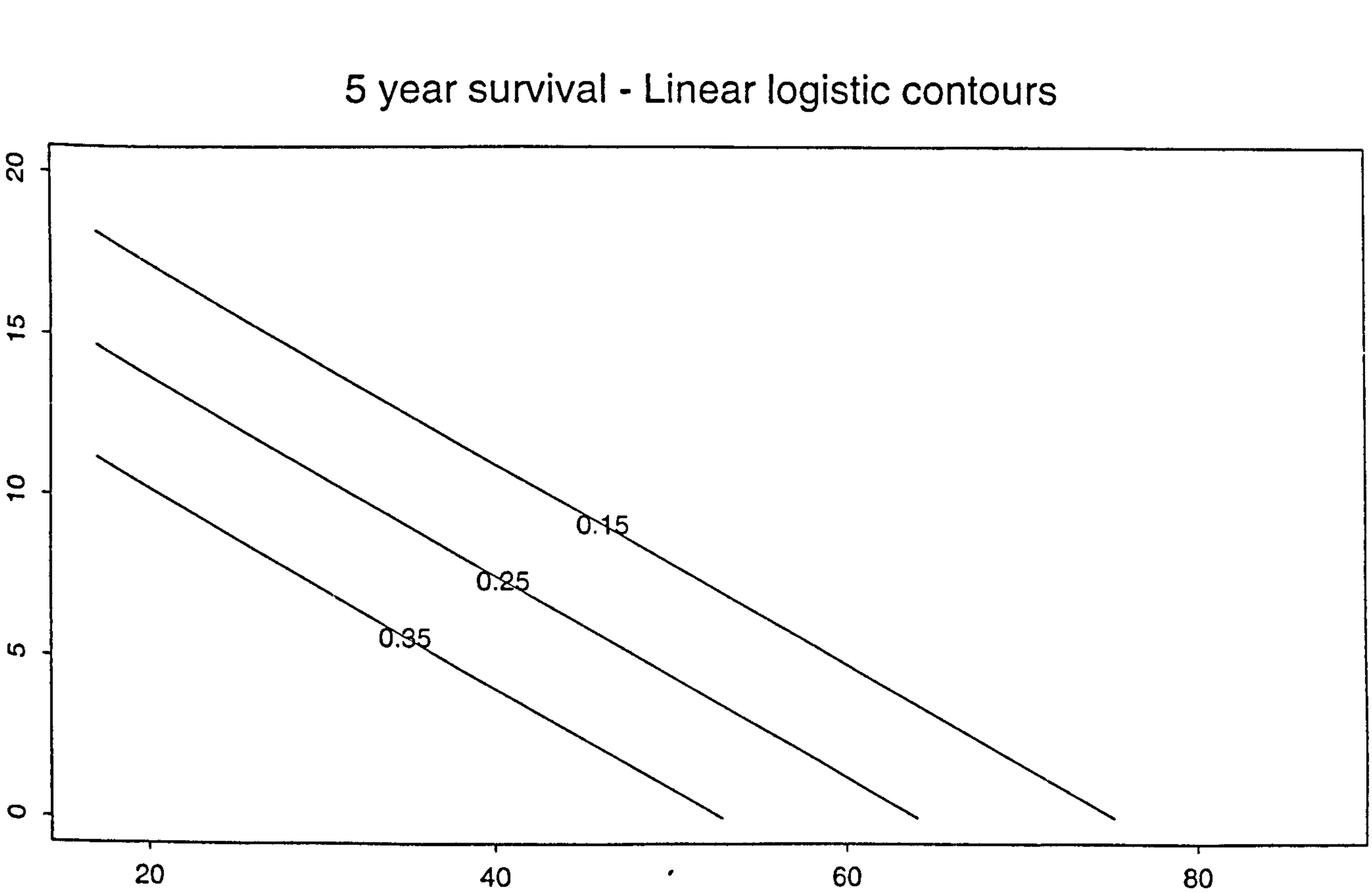


Figure 2.3.6

Section 2.4: Non-parametric logistic regression

One of the aims of this thesis is to investigate methods for identifying potential categorisations for continuous explanatories in logistic regression models. In the present context such categories would be defined as sections / areas of the explanatory where the probability of response appears to be roughly constant. The problems of how many categories to provide and where the appropriate cut points should be are the reasons why non-parametric logistic (binary) regression is now considered. This technique ***evolved*** from the standard idea of ***non-parametric regression for a continuous response*** first introduced by Nadaraya (1964) and Watson (1964). The ***non-parametric logistic regression*** concept is an adaptation of the standard case and is defined as follows.

$$\hat{p}_z = \hat{pr}(y = 1 / z) = \frac{\sum_{i=1}^{n} y_i \, \Delta_h(z, z_i)}{\sum_{i=1}^{n} \Delta_h(z, z_i)} \qquad \text{- (2.3)}$$

where

The weighting function, $\Delta_h(z, z_i) = K(\frac{z - z_i}{h})$

n is the number of subjects

K is a smooth probability density function

h is a smoothing parameter

z_i is the continuous explanatory value for i^{th} subject

y_i is the discrete response for i^{th} subject (coded 0 or 1)

The general method uses the observed data to compute a smoothed value for the response at each possible value of the continuous explanatory variable based on creating a weighted average of the values of the response variable over all subjects. The weighting attributed to each individual subject is a ***continuous decreasing function*** of the distance of the value of the explanatory for that particular subject from the value of the explanatory under consideration. The degree of smoothing is controlled by a smoothing parameter where ***small values*** provide ***minimal smoothing*** i.e. only subjects whose value of the explanatory variable(s) are close to the value of the explanatory under consideration will have much influence. As the ***smoothing parameter increases*** the amount of ***smoothing increases proportionally***. It is common to find that the more sparse the data the greater the degree of smoothing required to obtain any meaningful results.

Copas(1983) was the first to introduce this idea of non-parametric (logistic) regression with a ***binary response***, y_i. It has been shown (Hardle (1990)) that, in practice, the choice of smooth probability density function K(u) has remarkably little effect on the resulting estimate so, for convenience, Copas took K(u) to be proportional to the standard Normal density

$$\text{i.e.} \qquad K(u) = \exp(-\frac{1}{2}u^2)$$

Although the choice of kernel is not important the choice of smoothing parameter often is. As a result many methods for choosing an optimal smoothing parameter exist including several versions of cross-validation (Hardle (1990)) and / or the use of a penalty function (Rice(1984)). These methods do not always produce sensible answers and indeed

the work presented here suggests that the cross-validation method tended to produce values for the smoothing parameter which, appear to, grossly over-smooth the data. Therefore, in this chapter, a simple 'subjective search' method will be used to 'choose' an appropriate value for the smoothing parameter. A suitable value will be chosen which also ensures that the resulting estimates are, essentially, monotonic in nature. In practice the use of this technique tended to produce simple and easily interpretable results based on examination of data plots.

This then gives a simple method for producing point estimates of p_z when the response is binary. It would again be more helpful to produce interval estimates for p_z. Copas(1983) provided the following approximate variance for $\hat{p}_z$.

$$\hat{var}(\hat{p}_z) \approx \hat{p}_z(1-\hat{p}_z)\frac{\sum_{i=1}^{n} K\left(\frac{\sqrt{2}(z-z_i)}{h}\right)}{\left(\sum_{i=1}^{n} K\left(\frac{z-z_i}{h}\right)\right)^2} \quad \text{- (2.4)}$$

and corresponding approximate pivotal function

$$\frac{p_z-\hat{p}_z}{\sqrt{\hat{var}(\hat{p}_z)}} \dot{\sim} N(0,1)$$

This allows the derivation of an approximate 100*(1-α)% confidence interval for p_z of the form

$$\hat{p}_x \pm z_{\alpha/2}\sqrt{\hat{var}(\hat{p}_x)}$$

where $z_{\alpha/2}$ is the $100*\left(1-\alpha/2\right)$ percentage point of the standard normal.

Section 2.5: Prognostic factors for surviving stage 2 melanoma (revisited): An application of non-parametric logistic regression.

Section 2.5.1: Introduction

In a clinical context where potentially important prognostic factors are measured on a continuous scale it is often desirable to categorise such factors. The primary reason for this is that it facilitates interpretation for both clinicians and patients. In the analysis of five year survival from stage 2 melanoma, Tillman et al were keen to provide categorisations for any important prognostic factors. In section 2.3 this data set was considered and it was concluded that, in a univariate context, there were two potentially important prognostic factors, age on diagnosis of stage 2 melanoma and the number of nodes surgically removed although the effect of nodes was of borderline significance. In this section an examination will be made of both these important prognostic factors with plausible categorisations based on the non-parametric method outlined in section 2.4.

Section 2.5.2: Univariate analysis

Initially, consideration will be given to how five year outcome is affected by the age of the patient upon diagnosis of stage 2 melanoma. Interest is primarily in identifying potential categorisations for this variable. In order to identify such categorisations a non-parametric logistic regression will be fitted to the data and categorisations will be imposed

at covariate values where there is a clear and definite change in the probability of surviving five years.

Figure 2.5.1 provides a plot of $\hat{p}_z$ vs age for a selection of smoothing parameters. Point estimates $\hat{p}_z$ are represented by the continuous curve with approximate 95% confidence bands for p_z shown by the dotted lines.

A sensible choice of smoothing parameter should be a compromise between one which allows the fitted response curve to show too many dramatic (and hence spurious) changes in shape and one which completely smooths out any features of the data. Frames 1 and 2 of Figure 2.5.1 show far too many 'spurious' changes in shape while frames 7 through 9 appear to have smoothed out any features of the data. Frames 5 and 6 represent, in the author's opinion, a reasonably sensible choice of smoothing parameter as providing a nice balance between the conditions mentioned above.

Frames 5 and 6 of Figure 2.5.1 suggest a two-step categorisation to be appropriate. From these plots values of the covariate can clearly be identified where there are quite marked changes in the probability of surviving five years. Indeed there appear to be 3 categories :-

Category (1)	Less than 40 years of age
Category (2)	40-60
Category (3)	More than 60 years of age

Survival curves from univariate non-parametric logistic regression

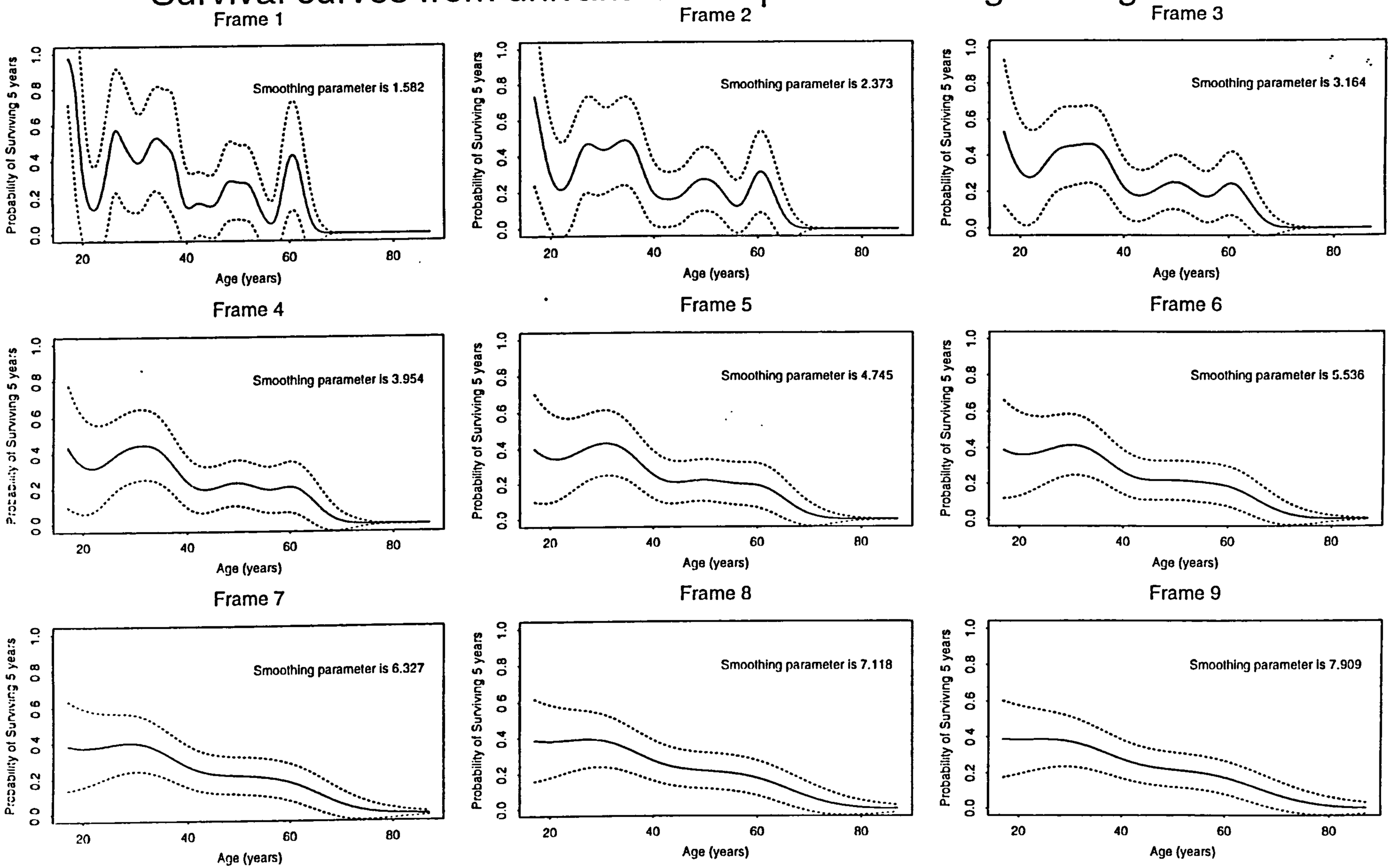


Figure 2.5.1

These categories seem sensible due to the clear dips in the fitted curve between 30 and 40 and then later on between 60 and 70. The probability of surviving five years initially drops very rapidly from approximately 0.4 at age 30 to about 0.22 at age 40, it then remains relatively stable between 40 and 60 and then drops again, although less rapidly than before, after 60 years of age to effectively 0.

Now consider the number of malignant nodes as a risk factor for five year survival. Tilmann et al proposed 2 categories for this variable, these being ***less than or equal to 3 nodes*** removed and ***more than 3 nodes*** removed. One would like to justify or indeed refute this categorisation by fitting an appropriate non-parametric logistic regression model to the data.

It is possible to investigate possible categorisations for the number of nodes by again using the technique of non-parametric logistic regression. Figure 2.5.2 provides a plot of $\hat{p}_z$ against number of nodes for a selection of smoothing parameters. Frame 4 of this Figure appears to represent the most plausible choice of smoothing parameter for this example. On this occasion the non-parametric logistic regression model seems to give some credence to the suggestion by Tillman et al that this variable should be split into two categories. However their choice of placing the cutpoint at 3 nodes as opposed to 4 or 5 seems somewhat arbitrary in this instance. There is clear evidence in Figure 2.5.2 of a dramatic change in the probability of surviving five years at around 4 or 5 nodes. The probability of survival remains relatively constant at approximately 0.25 until 4 or 5 nodes but drops rapidly from this point onwards.

Survival curves from univariate non-parametric logistic regression

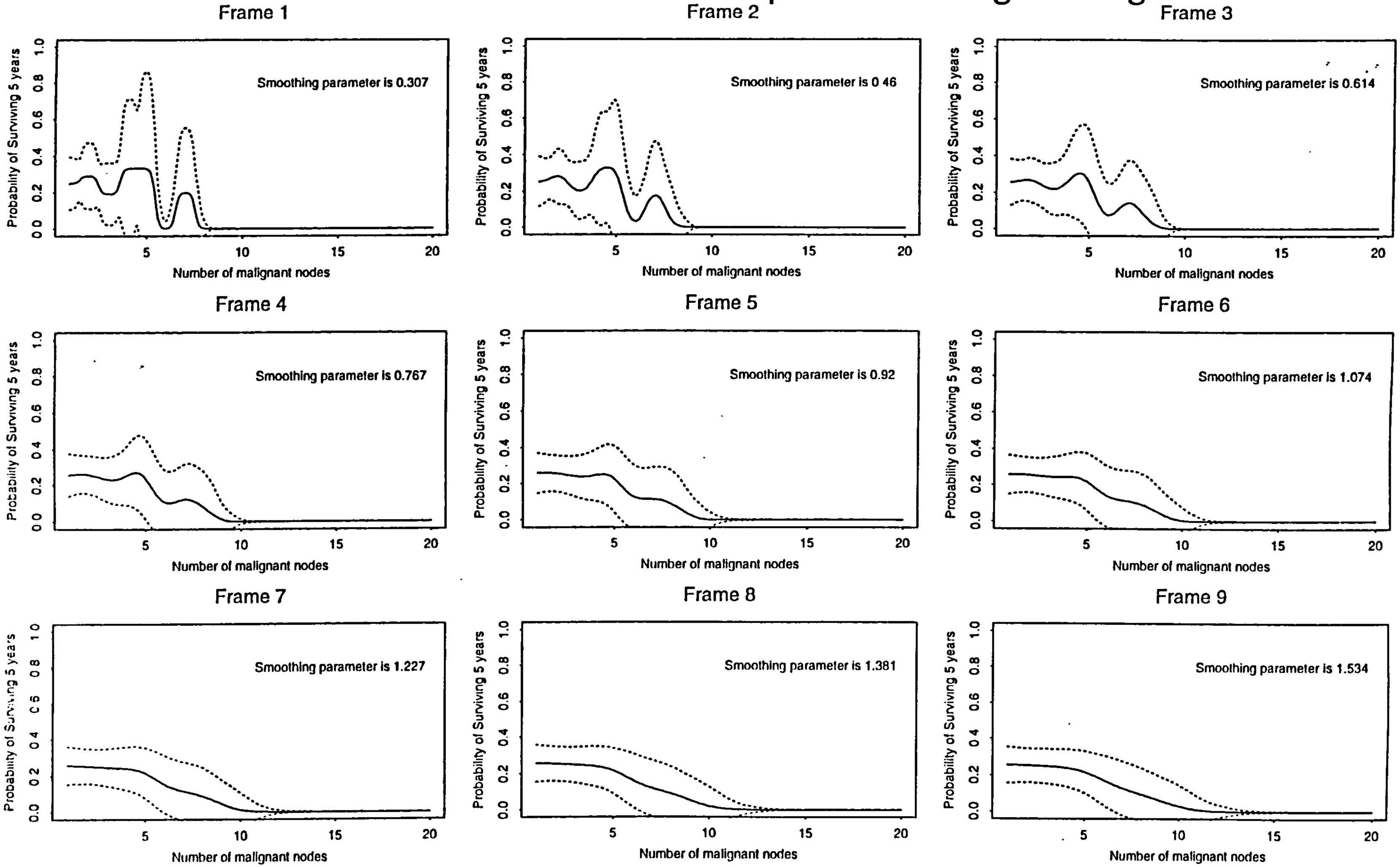


Figure 2.5.2

Section 2.5.3: Multivariate analysis

Section 2.3.3 outlined the multivariate analysis of this data set based on a linear logistic model. Here the multivariate analysis fitting a non-parametric logistic model based on the two continuous explanatory variables, age of the patient on diagnosis of stage 2 melanoma and the number of malignant nodes the patient had surgically removed will be examined. The multivariate non-parametric logistic regression model used here is an extension of the univariate model described in section 2.4 to incorporate a vector of continuous covariates $\underline{z}$.

When consideration is given to fitting a non-parametric logistic model with 2 continuous explanatories the situation becomes slightly more complicated than the univariate case illustrated in section 2.5.2 as two smoothing parameters now have to be chosen. For simplicity the same technique used in section 2.5.2 of choosing these parameters through a subjective search will again be used.

Figure 2.5.3 shows the 15, 25 and 35% probability contours for various combinations of the two smoothing parameters. A sensible combination of smoothing parameters is again one which removes any spurious changes in the probability of surviving five years without completely smoothing out the features of the data and this is obtained in Frame 5 of Figure 2.5.3. Figure 2.5.4 concentrates on Frame 5 of Figure 2.5.3 and also superimposes the corresponding linear logistic contours obtained in section 2.3. Figure 2.5.5 gives a 3 dimensional representation of this chosen non-parametric model.

Survival contours from bivariate non-parametric logistic regression

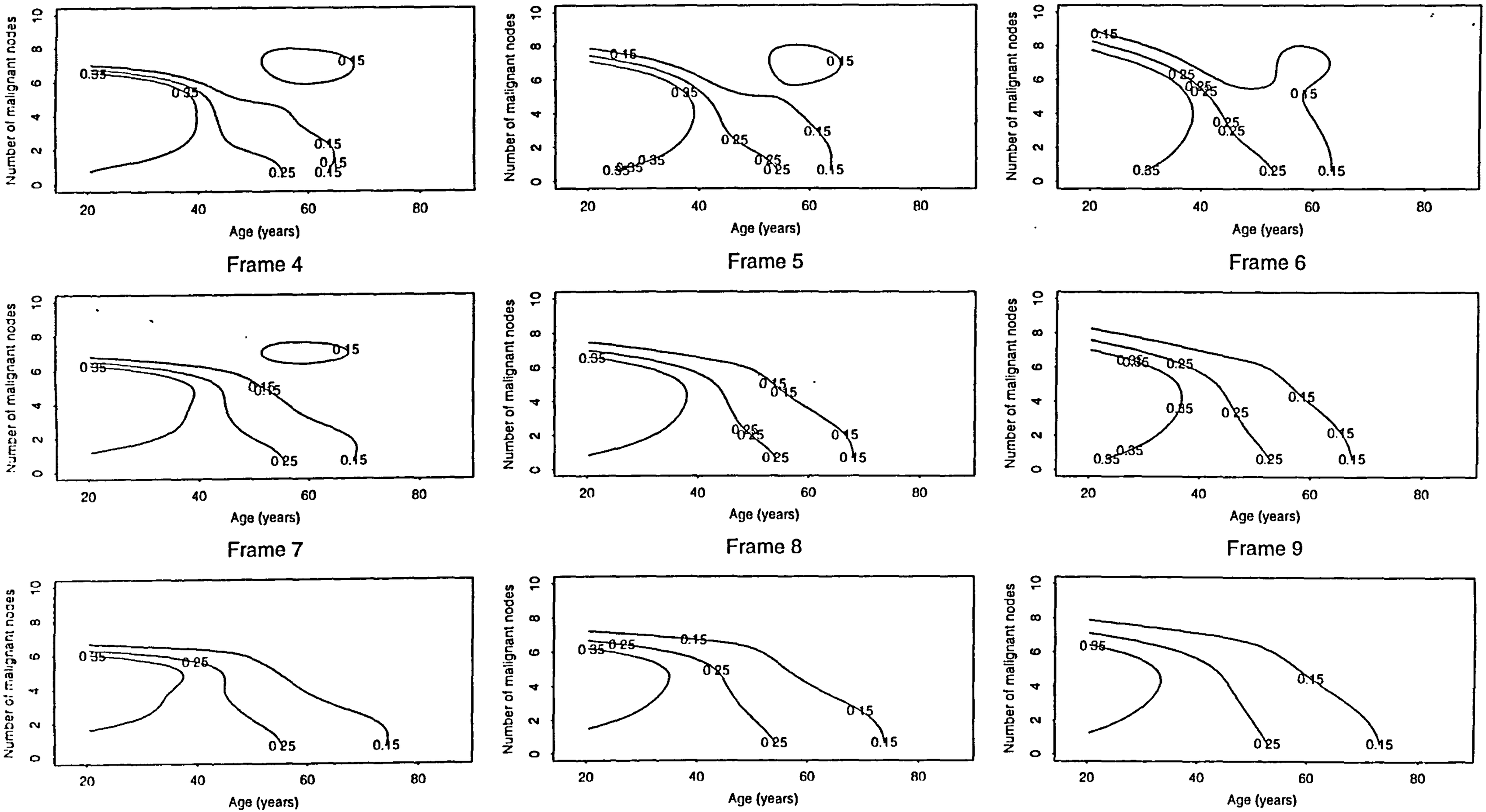


Figure 2.5.3

Plot of Non-parametric logistic contours with Linear logistic contours superimposed

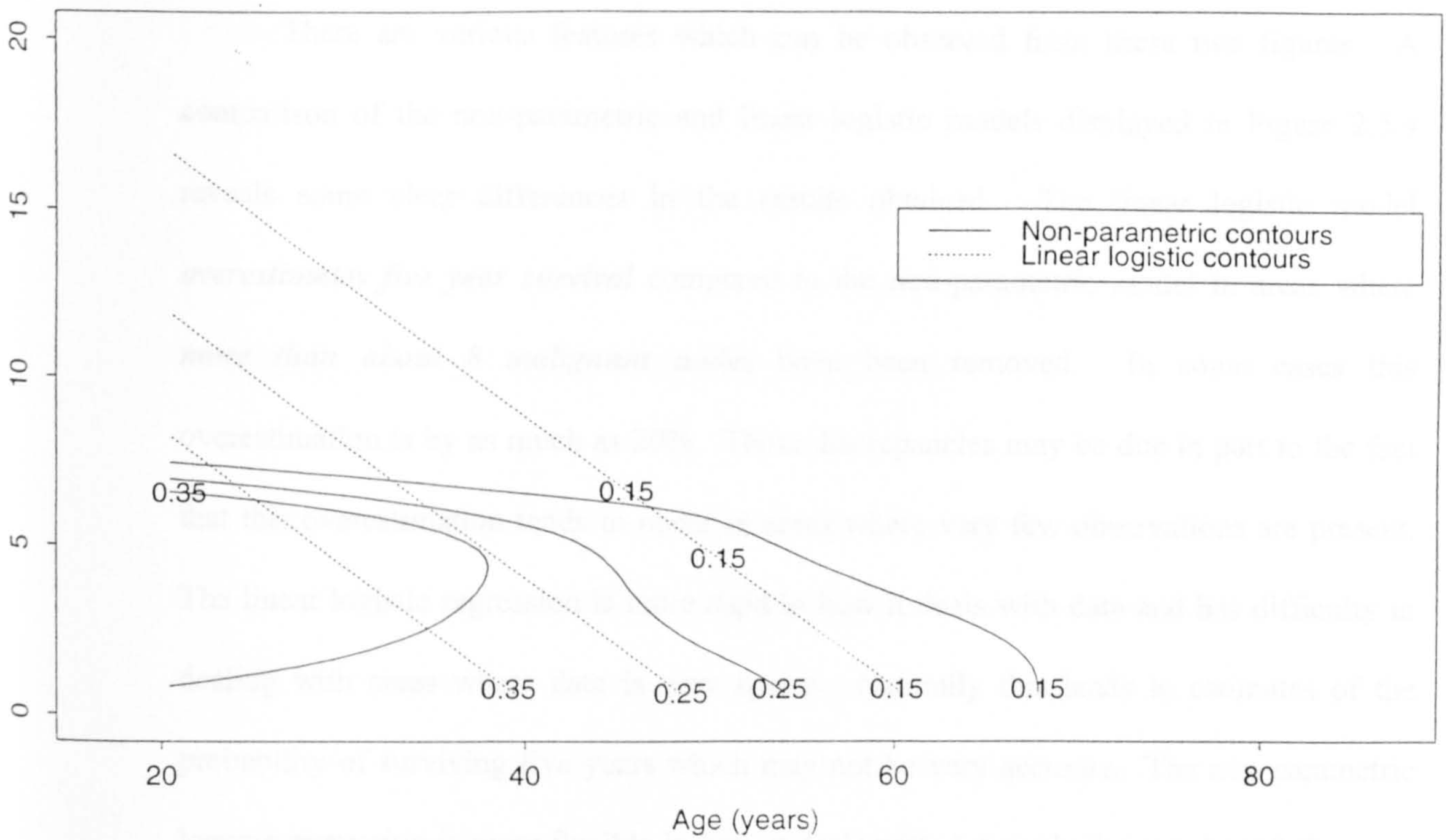


Figure 2.5.4

3 Dimensional perspective plot of non-parametric logistic regression

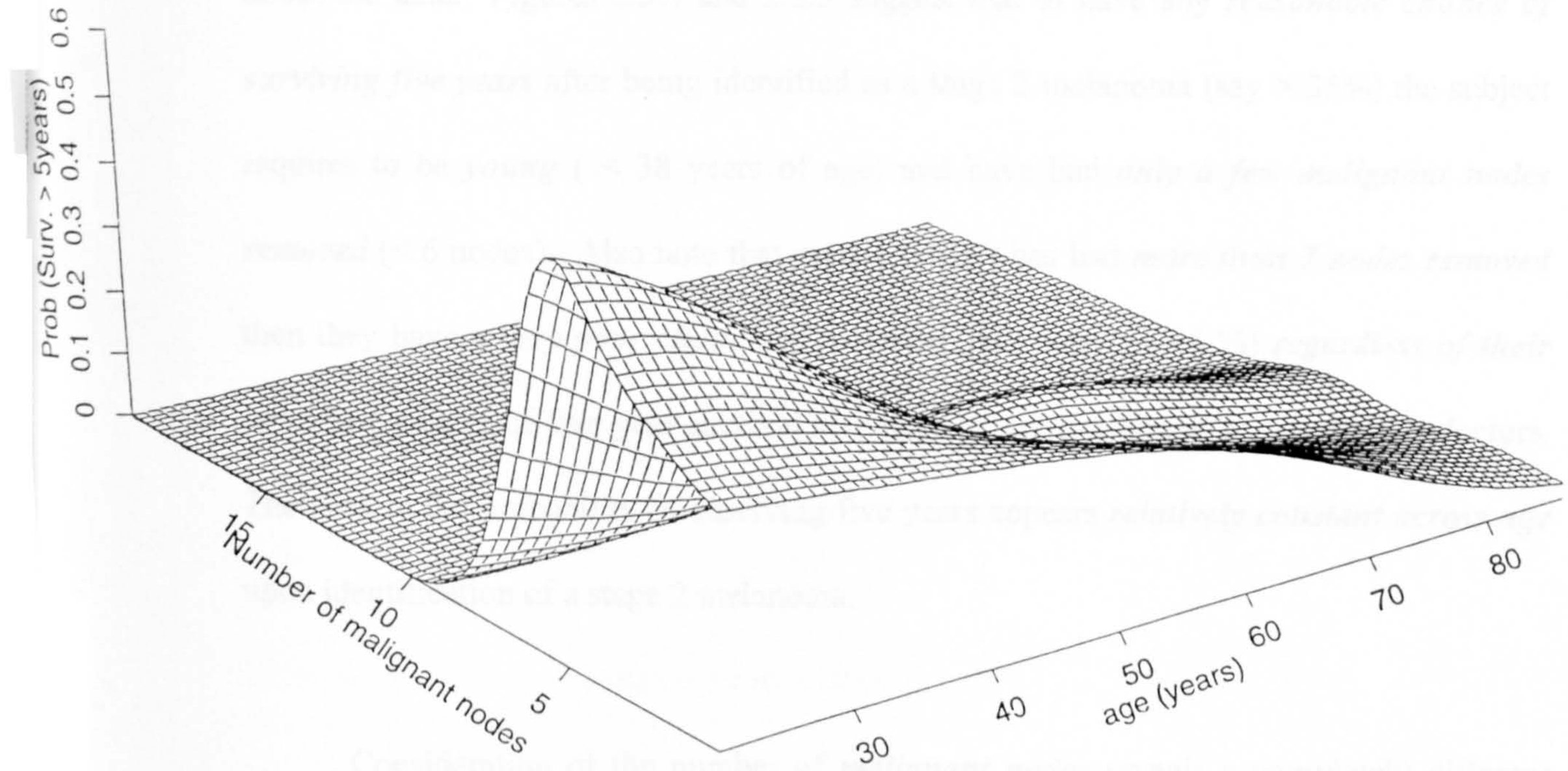


Figure 2.5.5

There are various features which can be observed from these two figures. A comparison of the non-parametric and linear logistic models displayed in Figure 2.5.4 reveals some clear differences in the results obtained. The linear logistic model ***overestimates five year survival*** compared to the non-parametric model in areas where ***more than about 8 malignant node***s have been removed. In some cases this overestimation is by as much as 20%. These discrepancies may be due in part to the fact that this overestimation tends to occur in areas where very few observations are present. The linear logistic regression is more rigid in how it deals with data and has difficulty in dealing with areas where data is very sparse. Naturally this leads to estimates of the probability of surviving five years which may not be very accurate. The non-parametric logistic regression is more flexible in how it deals with data and allows a, hopefully, truer (and more flexible) estimate of five year survival to be produced.

The results obtained from the non-parametric model reveal some interesting points about the data. Figures 2.5.4 and 2.5.5 suggest that to have any ***reasonable chance of surviving five years*** after being identified as a stage 2 melanoma (say > 35%) the subject requires to be ***young*** (< 38 years of age) and have had ***only a few malignant nodes removed*** (< 6 nodes). Also note that once a subject has had ***more than 7 nodes removed*** then they have a ***very poor chance of surviving five years*** (< 15 %) ***regardless of their age***. The pattern of five year survival also appears quite different across the two factors. The drop in the probability of surviving five years appears ***relatively constant across age*** upon identification of a stage 2 melanoma.

Consideration of the number of ***malignant nodes*** reveals a completely different pattern in the changes in survival however. The changes appear ***relatively minor from 1 to***

5 nodes removed. However there is then a ***very sharp drop*** in survival prospects ***from 5 to about 8 nodes*** removed. The survival prospects remain ***reasonably constant for more than 8 nodes removed*** but unfortunately these prospects are very poor (< 15%).

Section 2.5.4 Formal Identification of Categorisation Points

In section 2.4 non-parametric logistic regression was introduced and its application to a specific data set illustrated in section 2.5. This technique was used to highlight possible categorisation points. Categorisation points were chosen by examining plots of the probability of response and identifying points where there were marked changes in the pattern of the probability of response. However it is possible to construct a more formal approach to the identification of categorisation points.

Section 2.5.4.1: One Explanatory - The Use of Function Derivatives

The first derivative of a function, f '(z), is the slope of the tangent line to the original function, f(z) (Hunter (1972). Clearly values of z where there are dramatic changes in f '(z) correspond to areas where the function f(z) is changing most rapidly. Similarly values of z where f '(z) is relatively stable indicate areas where f(z) is relatively stable. Therefore plotting f '(z) against z and looking for values of z where there is a dramatic change in f '(z) may allow possible categorisation points to be identified.

In the non-parametric logistic regression situation

$$f(z) \quad = \quad \hat{p}_z \quad = \quad \hat{pr}(y=1/z) \quad = \quad \frac{\sum_{i=1}^{n} y_i \, \Delta_h(z, z_i)}{\sum_{i=1}^{n} \Delta_h(z, z_i)}$$

(see (2.3) for definitions of terms)

and

$$f'(z) = \frac{\left[\sum_{i=1}^{n} y_i \Delta_h(z, z_i) * \sum_{i=1}^{n} \left(\Delta_h(z, z_i) * \frac{(z - z_i)}{h^2} \right)\right] \; - \; \left[\sum_{i=1}^{n} \Delta_h(z, z_i) * \sum_{i=1}^{n} \left(y_i * \Delta_h(z, z_i) * \frac{(z - z_i)}{h^2} \right)\right]}{\left[\sum_{i=1}^{n} \Delta_h(z, z_i)\right]^2}$$

In order to identify possible categorisations for any explanatory variable z it is reasonable to plot f '(z) over the range of z and look for areas of rapid change in this function as this will highlight areas where f(z) changes most rapidly.

Illustration

In section 2.5.2 separate univariate non-parametric analyses of how the probability of surviving 5 years after being diagnosed stage 2 dependent upon the age at diagnosis of stage 2 melanoma and the number of malignant nodes surgically removed were carried out and possible categorisations suggested for each variable.

Dealing initially with age at diagnosis of stage 2 melanoma Figure 2.5.6 gives a series of plots, with varying choices of smoothing parameter, of the probability of surviving five years against age. Superimposed on these plots with a ***dotted line*** are the equivalent first derivatives of p_Z. Note that 10% has been cut off either end of the derivative to attempt to remove or limit edge effects on the derivative distorting the picture. In the author's opinion the most 'sensible' picture was given by frames 5 and 6 of Figure 2.5.1 and so here the derivatives presented in frames 5 and 6 of Figure 2.5.6 will be concentrated upon.

A close examination of these derivatives suggests that there is a major change occurring in the probability of surviving 5 years until a subject's age reaches about 40. The probability then gradually levels of till about the age of 60. This is followed by the probability dropping off again but at an apparently slower rate than the drop observed before 40 years of age.

This conclusion is very reassuring in that it produces results which are the same as those obtained in section 2.5.2 by consideration of the non-parametric logistic regression estimate itself. Both methods produce results which suggest very similar patterns across the probability of surviving 5 years in terms of the location of any cutpoints.

Next, consider separately the effect of the number of malignant nodes a subject had surgically removed has on the probability of surviving 5 years after diagnosis of stage 2 melanoma. In section 2.5.2 plots of the probability of surviving five years after entering stage 2 melanoma against the number of malignant nodes removed were displayed in Figure 2.5.2. These suggested that the probability of surviving five years remained

Simultaneous plots of survival curves and 1st derivatives

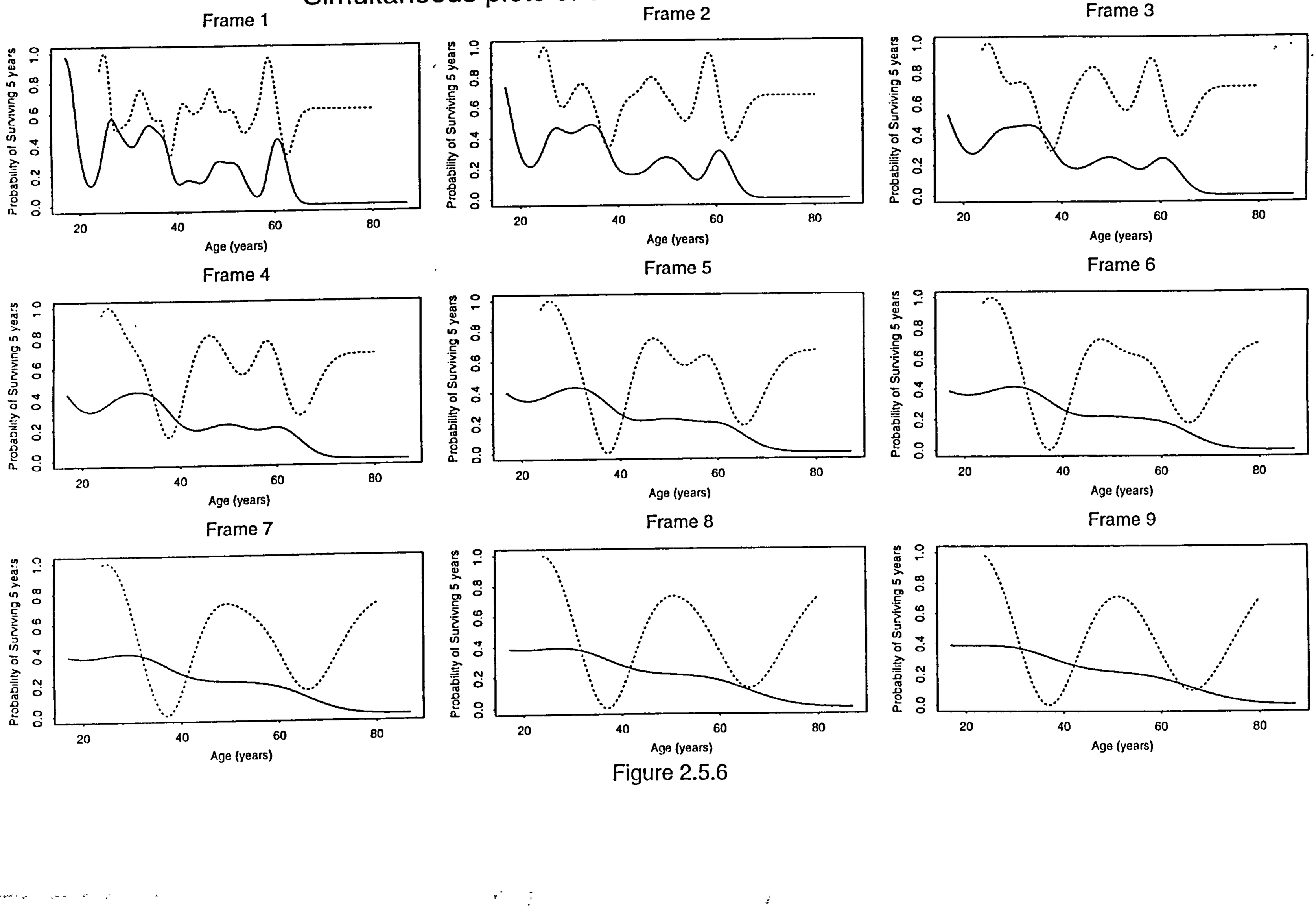


Figure 2.5.6

relatively constant up till about 4 or 5 nodes removed before dropping off very rapidly from 5 nodes onwards. Here, in each frame of Figure 2.5.7, a dotted line which indicates the ***gradient*** of the fitted non-parametric logistic regression curve is again superimposed on top of the actual fitted non-parametric logistic regression curve. The dotted line in frames 5 and 6 of this plot show that the probability of surviving five years does indeed drop off rapidly (gradient increasing very sharply) between 3 and 5 nodes before levelling off till about 7 nodes where the probability again appears to drop off although not as rapidly as observed in the earlier sharp drop. These results are again very similar to those obtained in section 2.5.2 when only the fitted non-parametric logistic regression curve was examined.

Section 2.5.4.2: Two Explanatories - The Use of Directional Derivatives

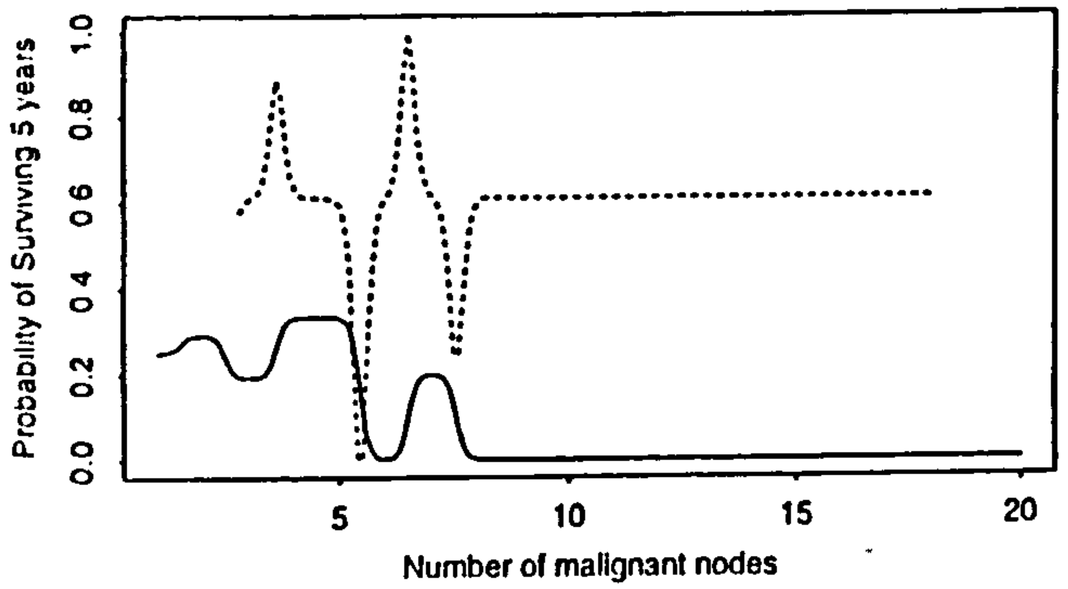
The theory previously discussed concerning one variable z can be extended to consider two variables x and z. Here if there exists a function $f(x,z)\ x \in R_x, z \in R_z$ which is at least once continuously differentiable then $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial z}$ both exist and the theory of directional derivatives (Spiegel (1974)) show that the maximum value of the directional derivative occurs in the direction normal to the surface of f(x,z) and is given by the function $|\nabla f|$ where

$$|\nabla f| \equiv \text{grad } f$$

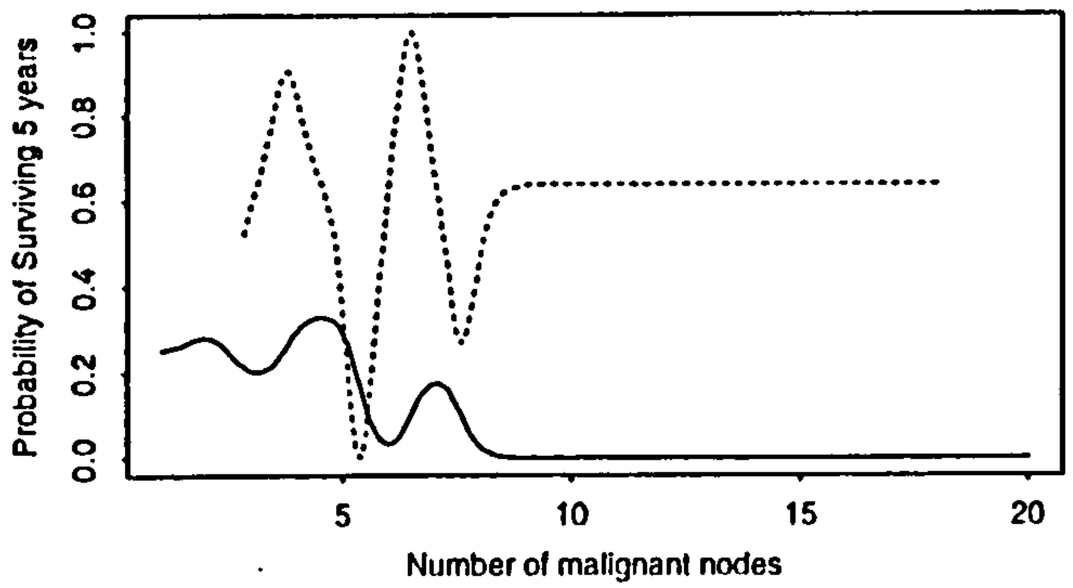
$$= \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial z}\right)^2} \qquad \text{- (2.5)}$$

Simultaneous plots of survival curves and 1st derivatives

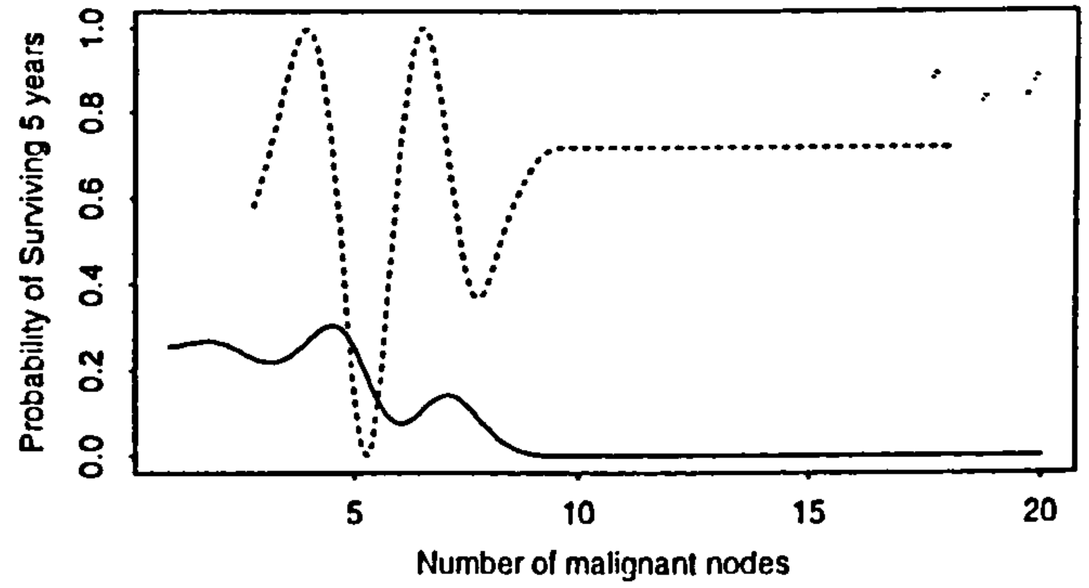
Frame 1



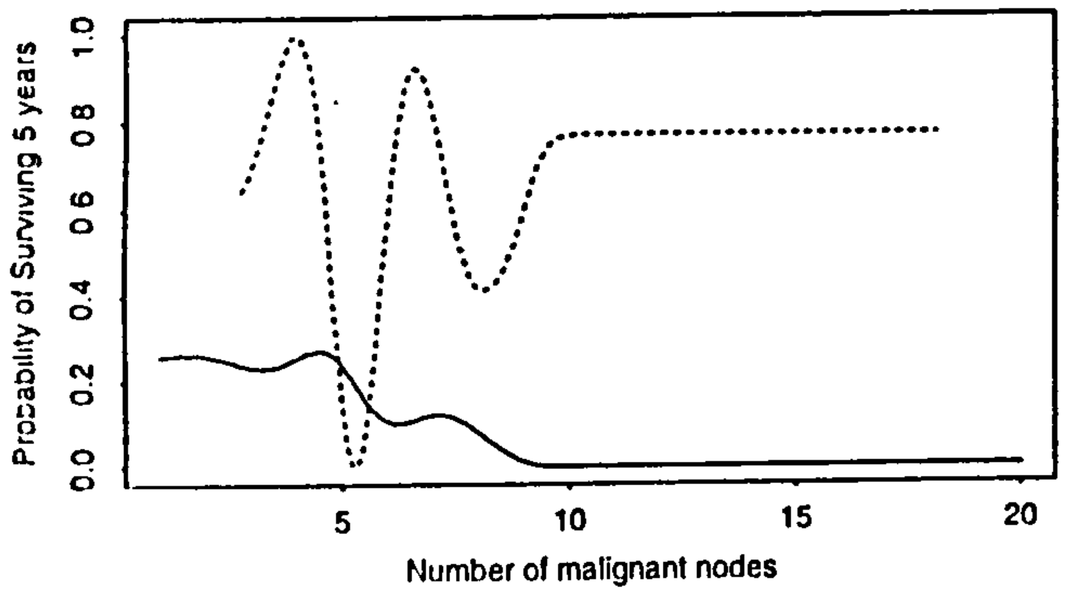
Frame 2



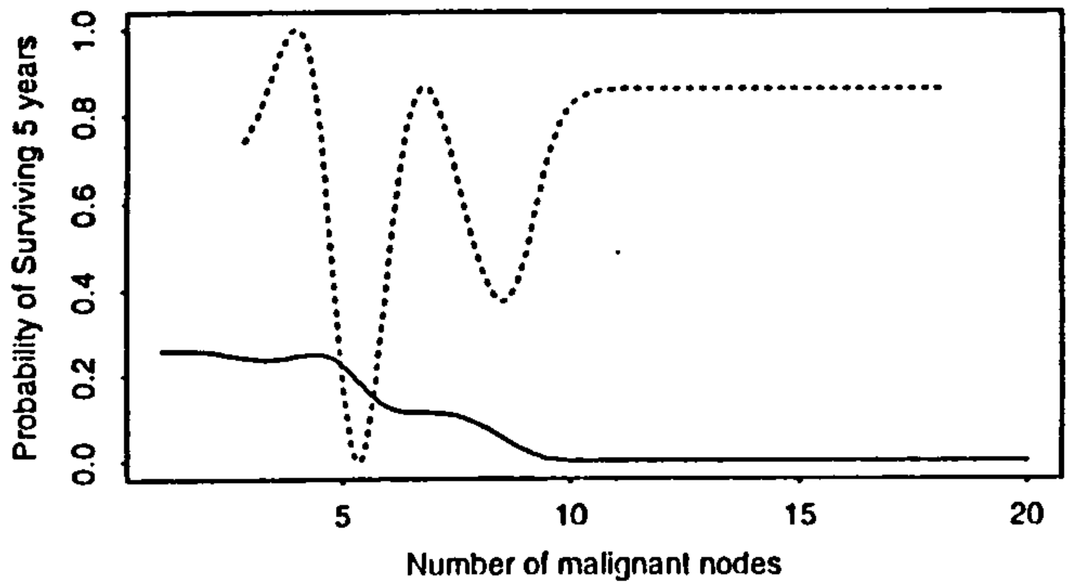
Frame 3



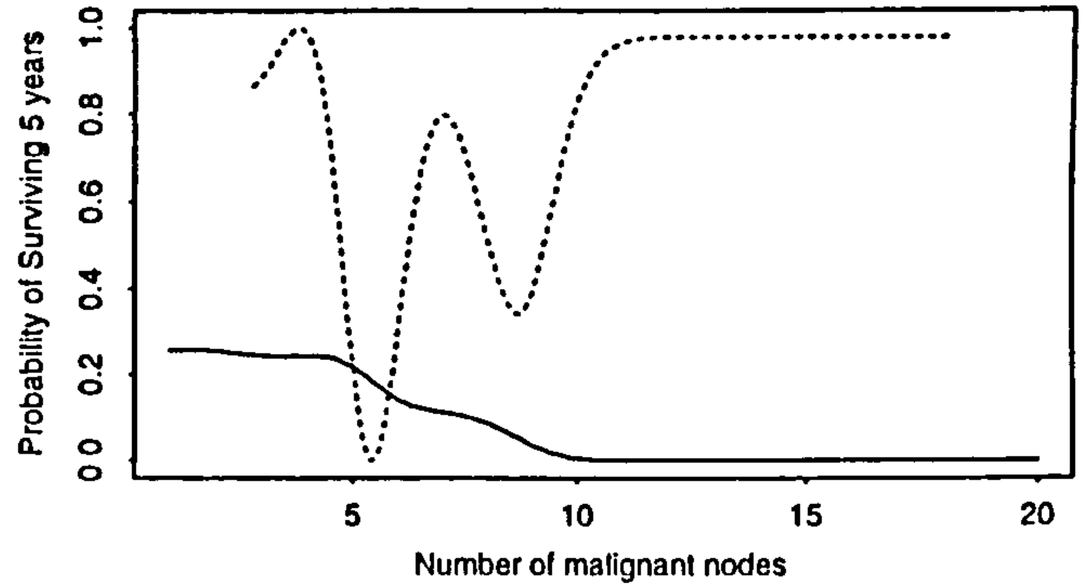
Frame 4



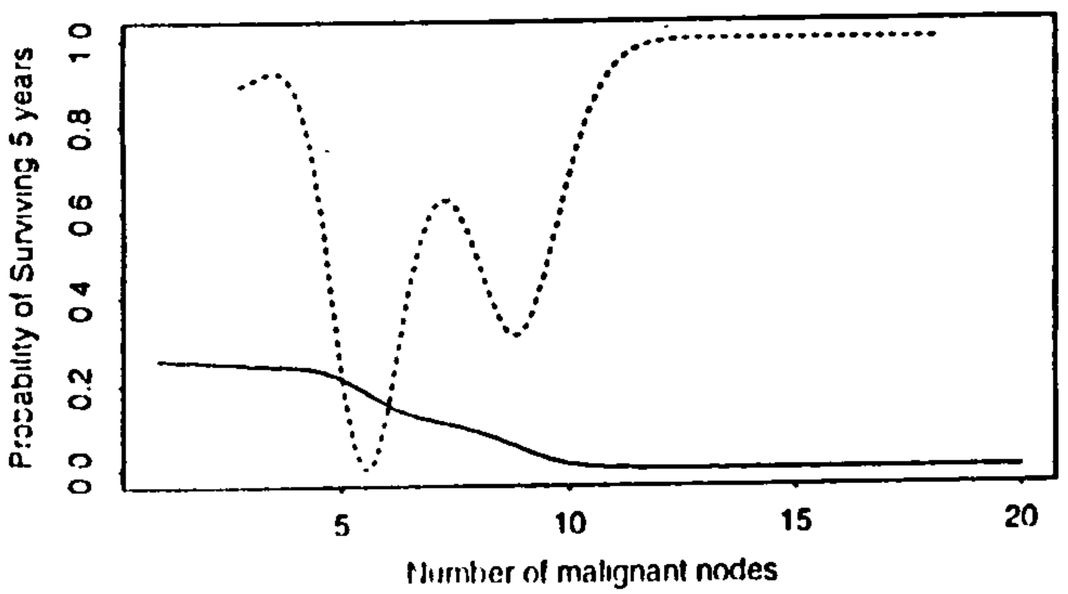
Frame 5



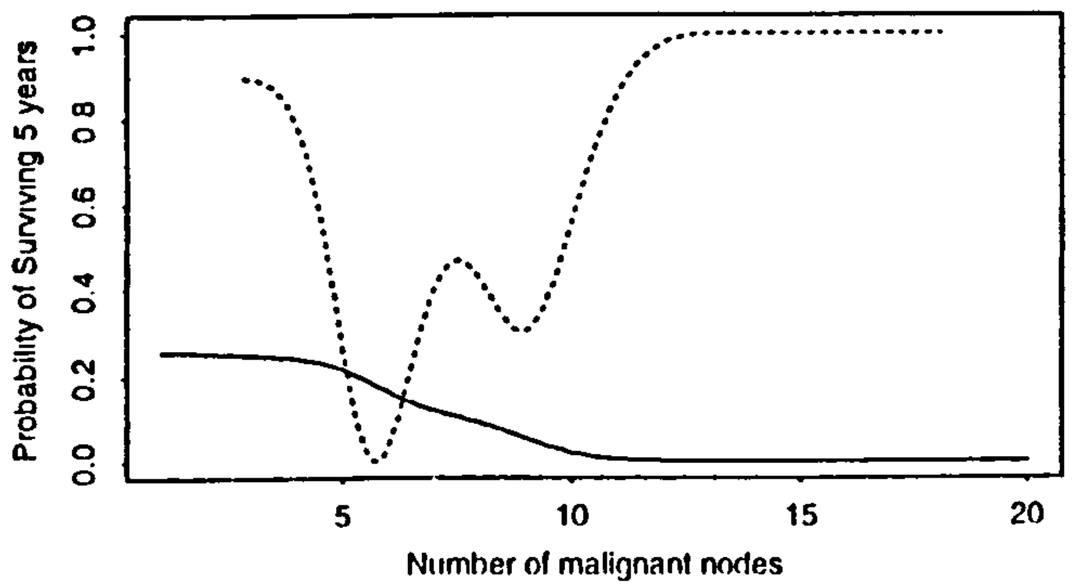
Frame 6



Frame 7



Frame 8



Frame 9

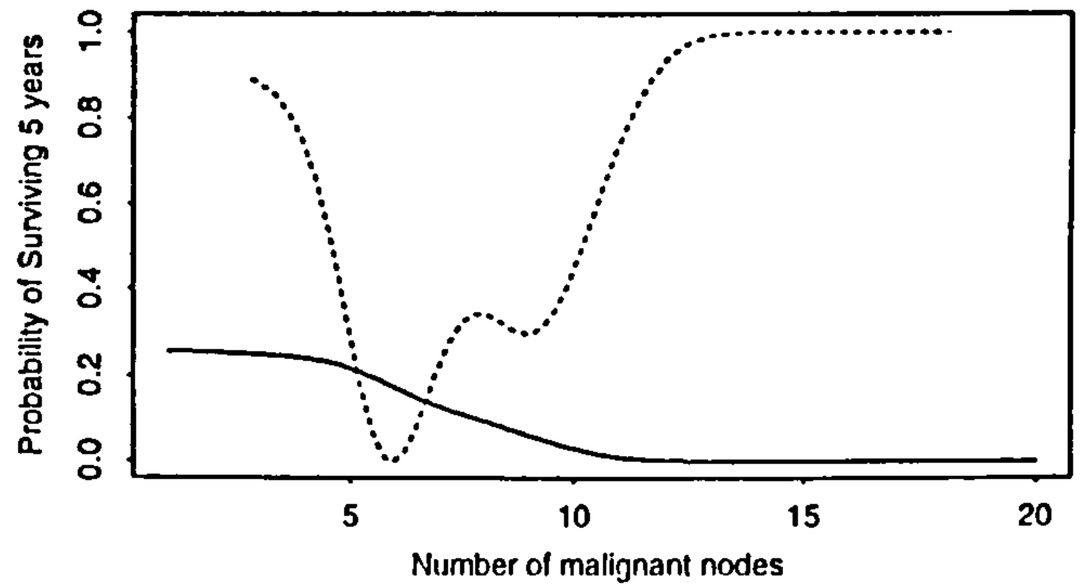


Figure 2.5.7

This may prove useful in identifying possible 2 dimensional 'cutpoints'. Combinations of x and z where the function $|\nabla f|$ takes markedly high values indicate areas where the function f(x,z) is changing most rapidly in both directions and combinations of x and z where $|\nabla f|$ takes lower values indicate areas where the surface of the function f(x,z) is relatively stable. A surface/contour plot of $|\nabla f|$ across the range of x and z will then make it possible to identify potential cutpoints. Areas of rapid change should be looked for as these will highlight areas of change in f(x,z) (i.e changes in the probability of five year survival).

Illustration

In section 2.5.3 a multivariate non-parametric analysis of the data set concerning five year survival from stage 2 melanoma was carried out displaying joint non-parametric contours of the probability of surviving at least 5 years for the 2 prognostically valuable variables age when diagnosed stage 2 melanoma and number of nodes surgically removed. The contours shown in Figure 2.5.3 indicated that the probability of surviving at least five years decreased at a relatively constant rate across the age variable. However across the nodes variable the pattern of the probability of surviving at least five years appeared quite different with the probability appearing relatively constant till about 5 nodes then dropping off rapidly between 5 and 8 nodes before again remaining constant, although fairly poor, for more than 8 nodes.

Figure 2.5.8 displays the corresponding series of 3-dimensional perspective plots of the ***grad*** function defined in (2.5) with 10% again cut off either end for each variable. As this technique is primarily looking for 'joint' categorisation points these results in

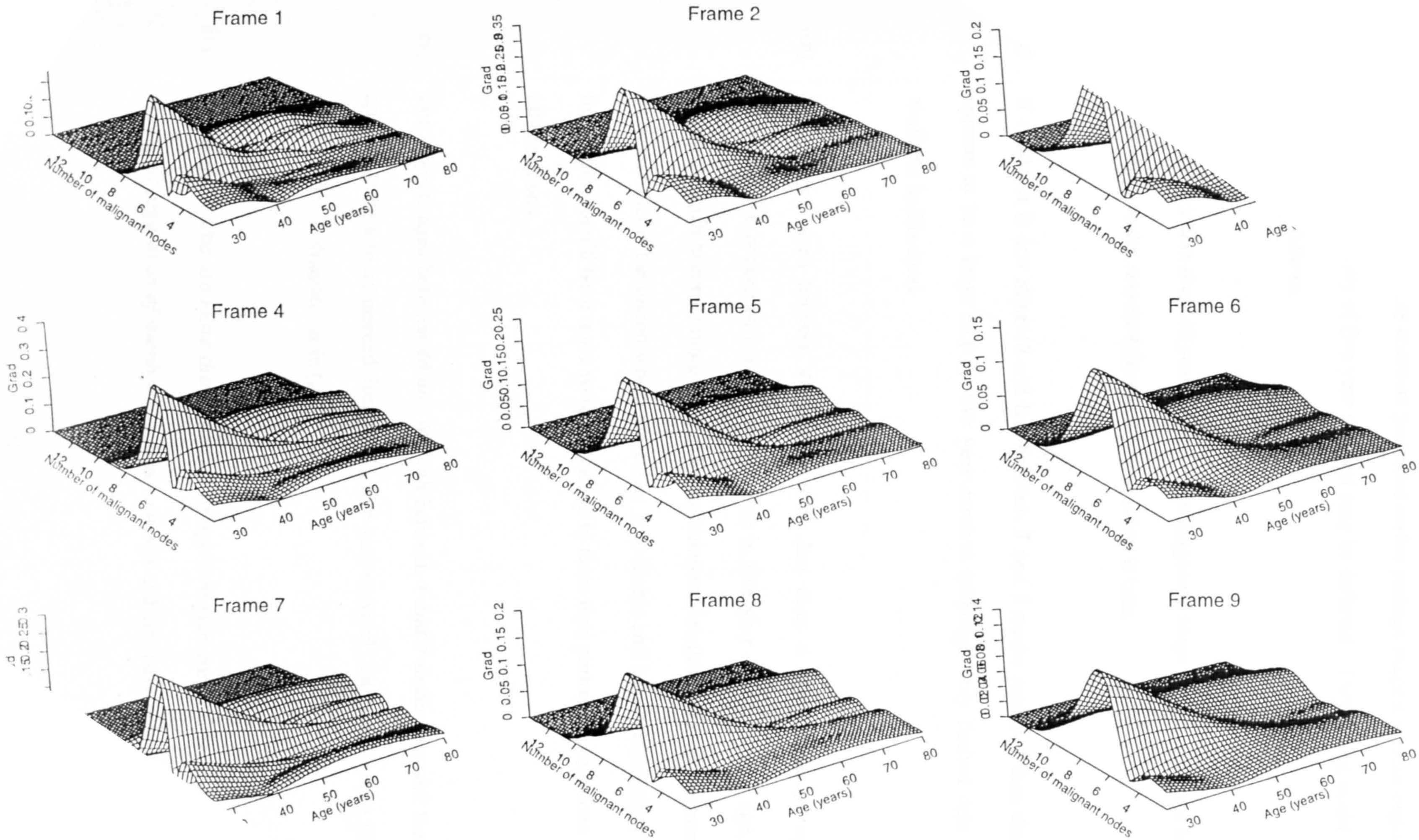

Frame 1
Frame 2
Frame 4
Frame 5
Frame 6
Frame 7
Frame 8
Frame 9
Grad
Number of malignant nodes
Age (years)

Figure 2.5.8

conjunction with the probability contours presented earlier perhaps suggest various regions / areas where the probability of five year survival may be different. These areas could be ***roughly*** described as follows.

(i) If a subject is ***less than 40*** years of age when diagnosed stage 2 melanoma and has ***less than 5 nodes removed*** then five year survival is best.

(ii) If a subject is ***less than 40*** and has between ***5 and 9 nodes removed*** then there appears to be a large drop in five year survival (although only limited data is available in this area).

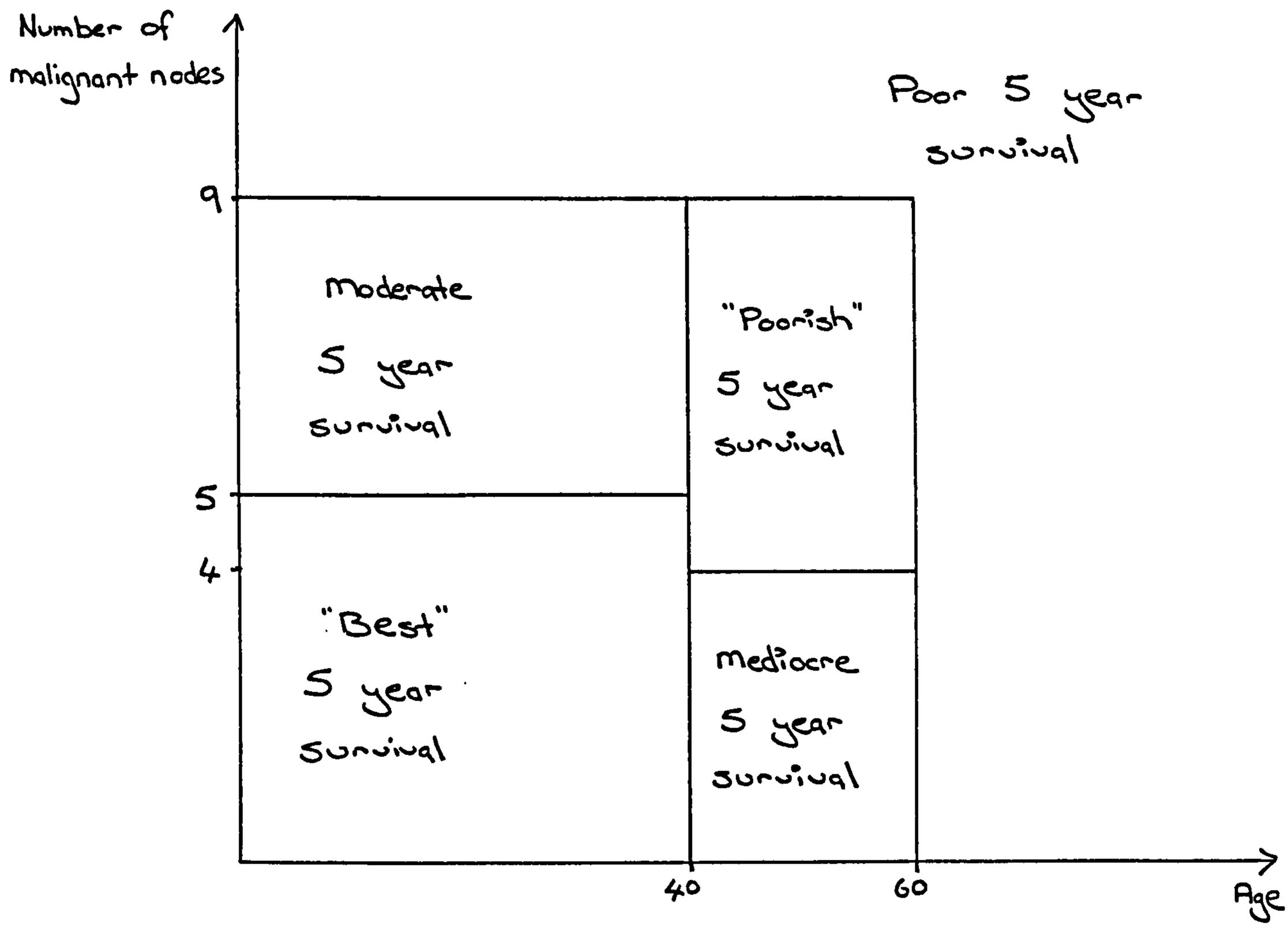
(iii) For subjects aged between ***40 and 60*** with ***less than 4 nodes removed*** age dominates the pattern with a gradual decrease in five year survival across age, a decrease which becomes more marked as the number of nodes removed increases. Note that this is the region where most data is available and hence conclusions in this region should have more weight attached to them than perhaps conclusions in other regions.

(iv) For subjects aged between ***40 and 60*** with between ***4 and 9 nodes*** removed there appears to be a more marked decrease in five year survival than is present in (iii) although not as dramatic as in (ii).

(v) For subjects who are ***more than 60*** years of age five year survival prospects are very poor ***regardless of number of nodes removed*** and similarly for subjects with

more than 9 nodes removed five year survival prospects are ***very poor regardless of age.***

These results are summarised in the diagram below



Section 2.5.4.3: Relevance of the Number of Malignant Nodes for Five Year Survival

A final point to consider is the importance of number of malignant nodes as a prognostic factor for five year survival. In Section 2.3.3 it was observed that in a multivariate linear logistic model which included both age at diagnosis of stage 2 melanoma and number of nodes surgically removed, the latter proved to be non-significant *in addition* to age in terms of five year survival. In this section non-parametric logistic models have been fitted to both the univariate and bivariate data. In conjunction with the first derivatives of these models, categorisations have been suggested in both the univariate and bivariate cases. In order to investigate if the number of nodes has anything significant to add in terms of five year survival it is necessary to compare the categorised model involving *age alone* with the categorised model which incorporates *both age and number of malignant nodes*. In Section 2.5.4.1 the following model was suggested based on using a categorised version of age alone:

Model A:

Category (1)	Less than 40 years of age
Category (2)	41-60 years of age
Category (3)	More than 60 years of age

In Section 2.5.4.2 the following model was suggested based on using categorised versions of both age and number of malignant nodes:

Model B:

Category (1)	Less than 40 years of age and less than 5 malignant nodes
Category (2)	Less than 40 years of age and 5-9 malignant nodes
Category (3)	40-60 years of age and less than 4 malignant nodes
Category (4)	40-60 years of age and 4-9 malignant nodes
Category (5)	More than 60 years of age or more than 9 malignant nodes

These models can be formally compared to each other since model A is a sub-model of B. The most general form of the model will have 12 separate categories and this is represented graphically in Figure (c) below, whilst Figures (a) and (b) below display the simplifications of this general model suggested by models A and B respectively.

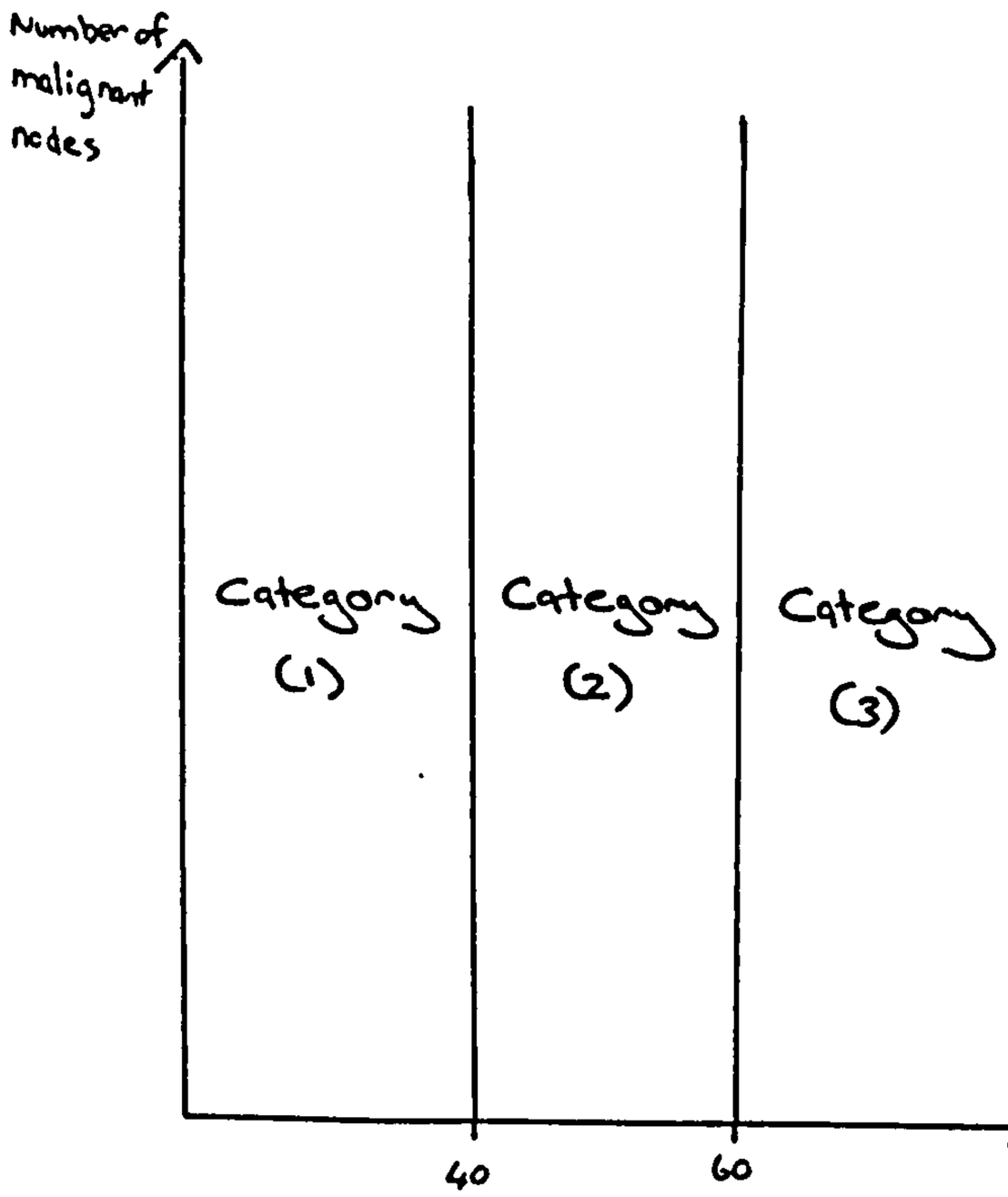


Figure (a)

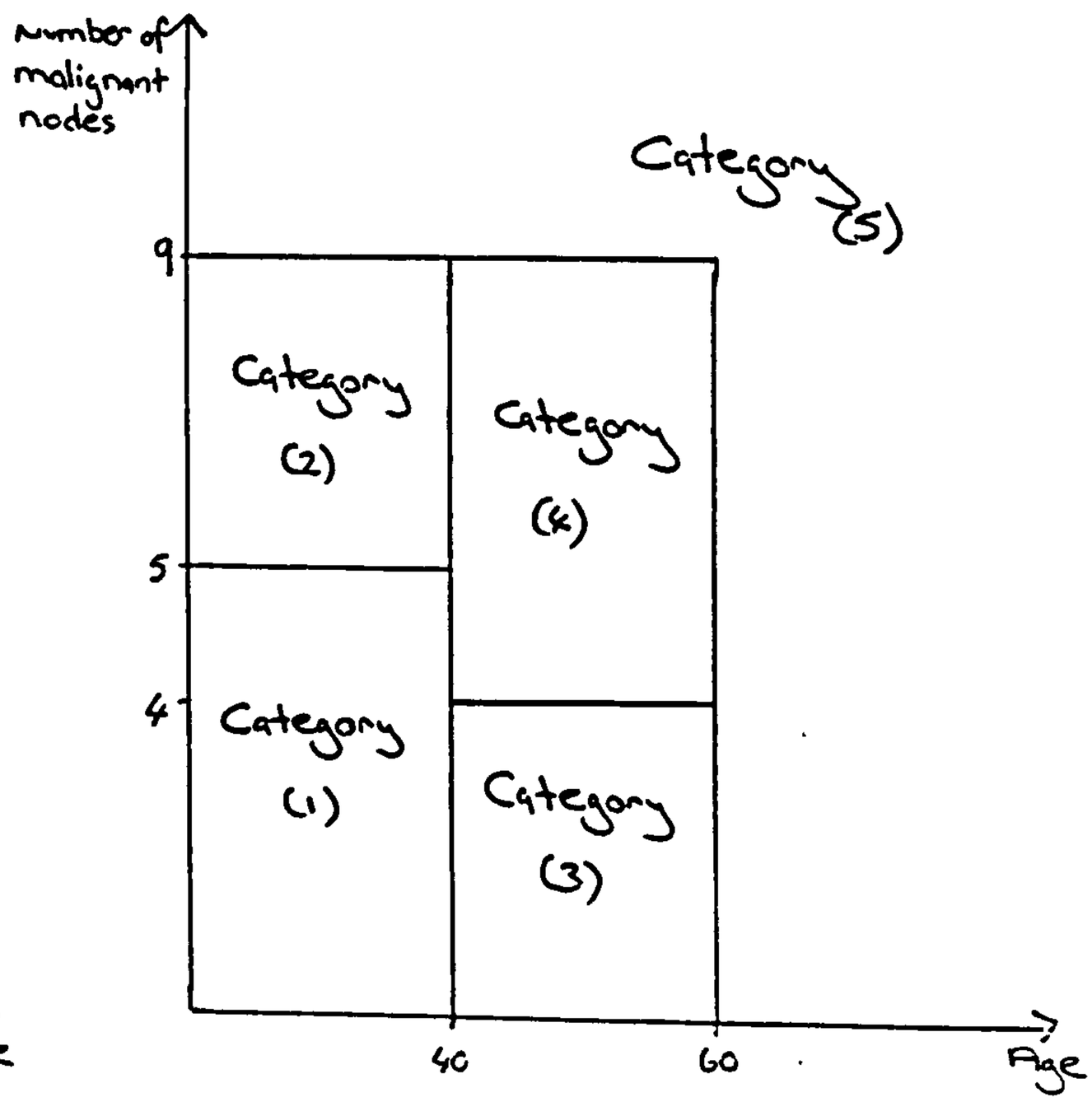


Figure (b)

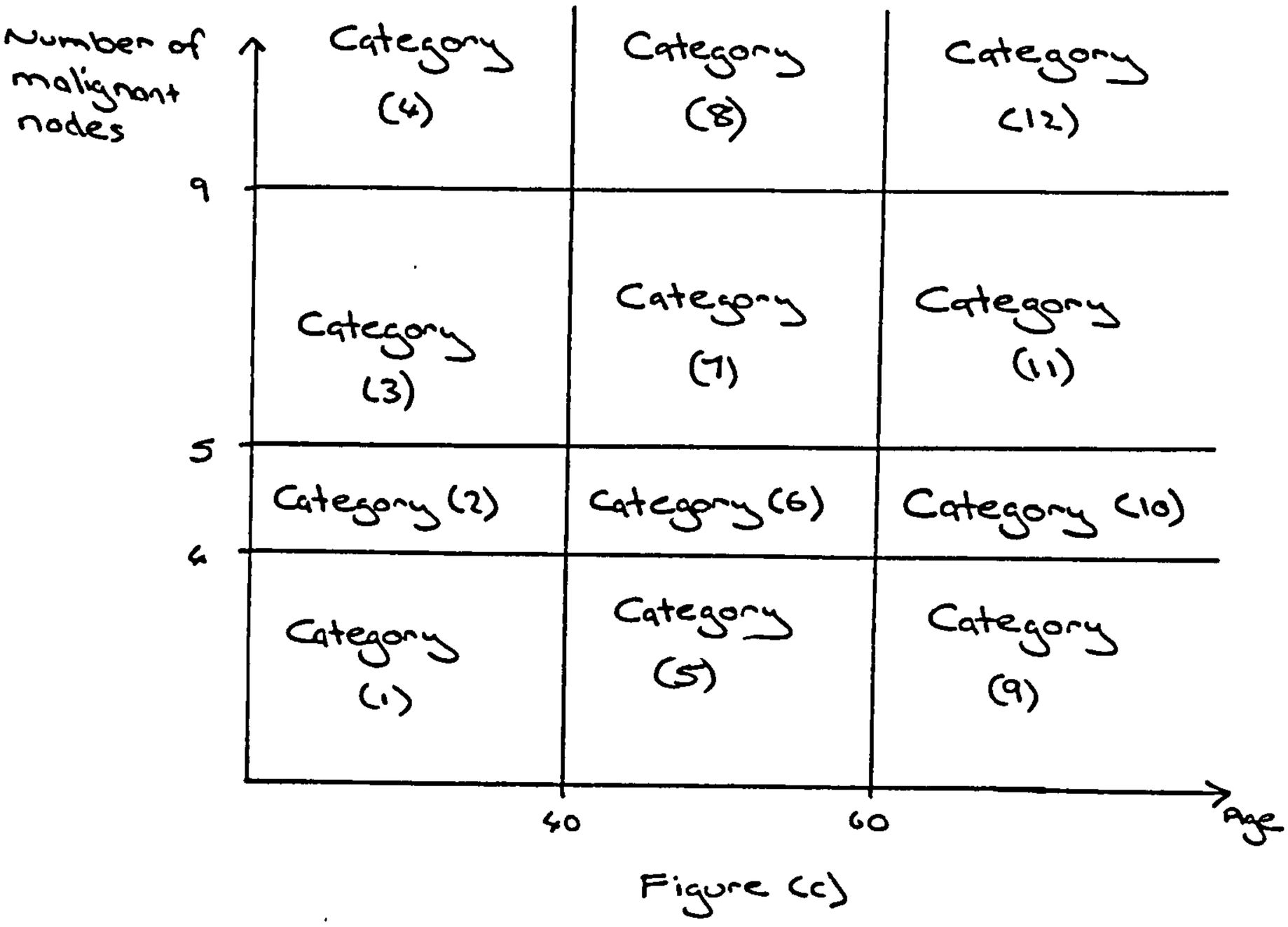


Figure (c)

To compare the categorisations suggested by Models A and B an hypothesis test based on likelihood was constructed to see if the categorisation suggested by Model B gives a significant improvement over the categorisation suggested by Model A. Initially a generalised likelihood ratio test of the categorisation based on Model A against no categorisation was carried out and produced a test statistic of 12.3 with 2 degrees of freedom (p-value = 0.002). Hence there is clear evidence that the categorisation suggested by Model A does have an effect (i.e. There is evidence of a difference in the probability of surviving five years between some or all of the categories). A generalised likelihood ratio test of Model A within Model B produced a test statistic of 5.4 with 2 degrees of freedom

(p-value = 0.067). Therefore, there is borderline evidence to suggest that Model A be rejected in favour of Model B. i.e. There is at least some evidence to suggest that age alone is not satisfactory in fully explaining the pattern of the probability of five year survival. When these two variables are considered in a categorised form there is evidence that incorporating malignant nodes ***in addition to age*** will produce a better prediction of the pattern of the probability of five year survival than using age alone. This allows the suggestion that, although number of malignant nodes was non-significant in a multivariate linear logistic model (Section 2.3.3), it is of prognostic value in addition to age when considered in a categorised form.

Section 2.6: Conclusions

This chapter has examined the analysis of data from a cohort study with a binary outcome. Consideration has been given to two possible methods of analysis, firstly the standard method of fitting a (parametric) linear logistic regression model and an alternative non-parametric logistic regression model. The linear logistic model imposes a linear constraint on the log odds of the fitted model whereas the non-parametric technique is a purely data fitting technique which does not impose any formal constraints on the final model. However the non-parametric method involves some degree of subjectivity in the choice of an appropriate data smoothing parameter.

The main aim of the work here is to identify possible categorisations for any explanatory variables in logistic regression analyses with 1 or 2 explanatory variables. The flexibility of the non-parametric technique in dealing with unusual data patterns makes it particularly appealing as a tool for allowing categorisations to be highlighted. The parametric technique, on the other hand, cannot highlight any potential categorisations as it involves a rigid assumption which cannot be influenced by unusual data patterns. Since a sharp change in survival prospects will clearly be identified as an unusual data pattern the non-parametric method becomes increasingly useful in identifying such features.

In this chapter these two methods were applied firstly to a data set concerning the probability of being alive 5 years after entering stage 2 malignant melanoma. Some differences in the estimates of the probability of being alive after 5 years were observed

between the linear and the non-parametric logistic regression models. In the bivariate case it was also clear that the linear model does not cope all that well with situations where data is very sparse.

The non-parametric logistic regression model allowed identification of potential categorisations and illustration was made of how it could be used to suggest such categorisations. It was proposed that categorisations should be placed at areas where there are marked changes in the pattern of the probability of surviving 5 years. This allowed various categorisations to be suggested both in the univariate and bivariate cases.

The idea of functional derivatives was used to provide an alternative viewpoint for suggesting categorisations. On the whole this technique produced results which were not dissimilar to those obtained by looking at the plots of the fitted non-parametric logistic curve. In general they suggested similar patterns for the probability of surviving 5 years but on occasions suggested slightly different locations for any actual cutpoints.

It may also be possible to use these techniques for more than two explanatories but visual representation of the results becomes increasingly complicated as the number of explanatories increases and hence demonstration has only been given here to results for two explanatories.

Chapter 3

Case / Control Studies

Section 3.1: Introduction

Case / control studies, often effectively "retrospective" studies, provide a research method for investigating potential risk factors for a specific disease. In this type of study a group of individuals known to be suffering from a particular condition or disease are obtained (the cases) and then compared with another group of individuals who are condition / disease free (the controls). The resulting analysis involves comparing the cases with the controls to identify ***factors*** that may differ between the two groups and hence which may in some way ***explain*** the occurrence of disease among cases. The work in this chapter will concentrate ***specifically*** on ***matched case / control studies*** where the cases are matched to specific controls by some variable(s) known to have an effect on the occurrence of disease; for example sex, age or social class.

Case / control studies are often used in the context of rare diseases since although they can be difficult to organise, especially when matched, they require less time and effort than prospective studies. This is due to the fact that with a case / control study the cases have already been identified whereas with other types of study a large sample may be

required to obtain a sufficient number of cases. The use of cohort studies (see chapter 2) for rare diseases would be impractical as a large amount of time and resources would be concentrated on following up individuals who would remain free of the disease (Schlesselman(1982)). Case / control studies are relatively quick to set up and conduct and, as a consequence of this, tend to be reasonably inexpensive. There are however some disadvantages with the use of case / control studies. One of the most common disadvantages with case / control studies is that selection of an appropriate comparison group can often prove problematic. Further, due to the design of case / control studies, rates of disease in exposed and unexposed individuals cannot be determined. All that can be obtained is an estimate of the ***Relative Risk of disease*** given a potential risk factor.

The next few sections will consider standard methods used in the analysis of data from case / control studies with particular emphasis on the conditional linear logistic model (section 3.2) and its application to the Relative Risk associated with ordinal explanatory risk factors in case / control studies (section 3.3).

The main impetus of this thesis is to consider whether methods of categorising variables can be established within various types of study framework. Therefore, in order to identify appropriate cut-points for an ***ordinal explanatory risk factor*** two non-parametric methods of analysis are developed and illustrated in sections 3.4 and 3.5. In section 3.6 consideration will be given to how to adapt these techniques to incorporate ***order*** restrictions on the ordinal explanatory risk factor. Finally, in section 3.7, a brief mention will be made of possible extensions to these non-parametric techniques to deal with a ***continuous explanatory risk factor.***

Section 3.2: Conditional Linear Logistic Model

Section 3.2.1: The model

In analysing multivariate data from a case / control study the standard model used is the Conditional Linear Logistic Model which is defined as follows. If the i^{th} subject has a p-dimensional set of potential risk factors $\underline{z}_i$ then

$$\Pr(\text{subject has the disease} / \underline{z}_i) = \frac{\exp(\underline{\beta}^T \underline{z}_i)}{1+\exp(\underline{\beta}^T \underline{z}_i)} \quad \text{- (3.1)}$$

where $\underline{z}_i^T = (1 \;\; z_{i1} \;\; z_{i2} \;\; z_{i3} \;\; \ldots \;\; z_{ip})$

$$\underline{\beta}^T = (\beta_0^* \;\; \beta_1 \;\; \beta_2 \;\; \ldots \;\; \beta_p)$$

$$\beta_0^* = \beta_0 + \log(\pi_1/\pi_0)$$

Notes: (i) The parameter π_1 is the probability that a ***diseased*** person is included in the study as a ***case*** and is known as the ***case sampling fraction***.

(ii) The parameter π_0 is the probability that a ***non-diseased*** person is included in the study as a ***control*** and is known as the ***control sampling fraction***.

In essence this model is very similar to the linear logistic model discussed in chapter 2 with the exception of the term π_1/π_0. This term is difficult to estimate due to the fact that in many practical contexts the ***sampling fraction among controls***, π_0, is often ***unknown***. Therefore it is almost impossible to be able to estimate the intercept β_0^* from a case / control study. Fortunately this is not of major concern as interest usually lies in estimating the Relative Risk of disease which does not involve π_1/π_0. (see section 3.2.3)

Section 3.2.2: Conditional Likelihood

In order to ***estimate all the unknown parameters*** in the Conditional Linear Logistic model it is necessary to ***maximise*** the appropriate ***Conditional Likelihood.*** The following definition of the Conditional Likelihood applies to the situation where there is a ***1 to 1 matching*** of I pairs of cases and controls.

Let $\underline{x}_i$ be the p-dimensional characteristic vector for the i^{th} case (i = 1, ..., I)

Let $\underline{y}_i$ be the p-dimensional characteristic vector for the i^{th} control (i = 1, ..., I)

Let **A** represent the event that the case has the disease and **A'** its complement

Let **B** represent the event that the control has the disease and **B'** its complement

Then the Conditional Likelihood for this model (Hosmer & Lemeshow (1989)) is given by

$$\prod_{i=1}^{I} \Pr(\mathbf{A} \text{ and } \mathbf{B}' \;/\; \underline{x}_i, \underline{y}_i \text{ and that one of the two has the disease})$$

$$= \prod_{i=1}^{I} \frac{\Pr(A / \underline{x}_i)\Pr(B' / \underline{y}_i)}{\left\{\Pr(A / \underline{x}_i)\Pr(B' / \underline{y}_i)\right\} + \left\{\Pr(A' / \underline{x}_i)\Pr(B / \underline{y}_i)\right\}}$$

which can easily be shown (Hosmer & Lemeshow (1989)) to reduce to

$$= \prod_{i=1}^{I} \frac{\exp(\underline{\beta}^T \underline{x}_i)}{\exp(\underline{\beta}^T \underline{x}_i) + \exp(\underline{\beta}^T \underline{y}_i)} \qquad \text{- (3.2)}$$

where $\underline{\beta}^T = (\beta_1 \;\; \beta_2 \;\; \beta_3 \;\; \ldots \;\; \beta_p)$

and so the conditional likelihood does not involve π_1/π_0 .

Note that these results are for the case where there is a 1-1 matching of cases and controls but simple extensions exist for other situations.

Estimates of $\underline{\beta}^T$ can be found by directly ***maximising the conditional likelihood*** or more commonly the ***logarithm of the conditional likelihood.***

Confidence intervals for any of the individual coefficients, β_i, can be obtained based on using the information matrix $I(\underline{\beta})_{|\underline{\beta}=\hat{\underline{\beta}}}$ as an approximation to $\text{cov}(\hat{\underline{\beta}})$ (Kalbfleisch(1985)).

Therefore let

$$\hat{Q} = \left(\hat{q}_{ij}\right) = I^{-1}(\underline{\beta})_{|\underline{\beta}=\hat{\underline{\beta}}}$$

This can then be used in conjunction with the approximate pivotal result

$$\frac{\underline{b}^T\underline{\beta} - \underline{b}^T\hat{\underline{\beta}}}{\underline{b}^T\hat{Q}\underline{b}} \quad \dot{\sim} \quad N(0,1)$$

to produce confidence bands for $\underline{b}^T\underline{\beta}$ and hence an approximate 100*(1-α)% confidence interval for any individual coefficient, β_i, is given by

$$\hat{\beta}_i \pm z_{\alpha/2}\sqrt{\hat{q}_{ii}}$$

where $z_{\alpha/2}$ is the $100*\left(1-\alpha/2\right)$ percentage point of the standard normal.

Section 3.2.3: Relative Risk

In most case / control studies interest is primarily in estimating the ***Relative Risk*** of disease. If two groups of subjects are present who differ only in the presence or absence of exposure to some

study factor then the Relative Risk is a measure of how many times more (or less) likely it is that disease occurs in the ***exposed*** group than in the ***unexposed*** group.

A formal definition of the Relative Risk is the ***ratio*** of the ***incidence rate*** (proportion of new cases) among the ***exposed group*** to the ***incidence rate*** among the ***unexposed group***.

One problem with the Relative Risk is that, in general, it can only be evaluated from a cohort study. Fortunately an approximation to the Relative Risk can be calculated for a case-control study. This measure is known as the ***odds-ratio*** or ***relative odds*** and for rare diseases it closely ***approximates*** the Relative Risk (Schlesselman 1982)).

The ***odds ratio*** is the ratio of the odds of disease in exposed individuals relative to the unexposed individuals and is one of the most common estimators of Relative Risk. A formal definition of the odds ratio for $\underline{z}$ compared to $\underline{z}^*$ is as follows:

$$\text{Odds ratio} = \psi(\underline{z}:\underline{z}^*) = \frac{\Pr(\text{diseased}/\ \underline{z}) \,/\, \Pr(\text{not diseased}/\ \underline{z})}{\Pr(\text{diseased}/\ \underline{z}^*) \,/\, \Pr(\text{not diseased}/\ \underline{z}^*)}$$

$$= \exp\left[\sum_{j=1}^{p} \beta_j (z_j^* - z_j)\right]$$

$$= \prod_{j=1}^{p} \exp\left[\beta_j (z_j^* - z_j)\right] \qquad \text{- (3.3)}$$

which is the odds ratio for a subject with vector of risk factors $\underline{z}$ compared to a subject with vector of risk factors $\underline{z}^*$. Often these vectors will only differ with respect to one risk factor. This then makes it possible to see how that particular risk factor affects the risk of disease. The β_j's are estimated using the techniques of section 3.2.2 and the odds ratio is then estimated as

$$\hat{\psi}(\underline{z}:\underline{z}^*) \quad = \quad \prod_{j=1}^{p} \exp\left[\hat{\beta}_j(z_j^* - z_j)\right]$$

With this model the effect of any particular risk factor will be multiplicative through the term $\exp\left\{\hat{\beta}_j\left(z_j^* - z_j\right)\right\}$.

As always an interval estimate for the odds ratio would be more informative than a simple point estimate and this can be derived (Breslow & Day(1980)) as follows:

The variance of $\log(\hat{\psi}(\underline{z}:\underline{z}^*))$ (i.e the log of the odds ratio) is estimated by

$$\hat{\upsilon} \quad = \quad \sum_{j=1}^{p} \hat{q}_{jj}(z_j^* - z_j)^2 + \sum_{\substack{j=1 \\ j \neq r}}^{p} \sum_{r=1}^{p} \hat{q}_{jr}(z_j^* - z_j)(z_r^* - z_r)$$

where the matrix $\hat{Q} = \left(\hat{q}_{ij}\right)$ is as defined in Section 3.2.2.

An approximate 100*(1-α)% induced confidence interval for the odds-ratio is given by

$$\hat{\psi}(\underline{z}:\underline{z}^{*})^{\exp\pm z_{\alpha/2}\sqrt{\hat{\upsilon}}}$$

where $z_{\alpha/2}$ is the $100*\left(1-\alpha/2\right)$ percentage point of the standard normal.

Section 3.3: Cutaneous Malignant melanoma: An illustration of a case / control study.

In the last few years both clinicians and the general public have become concerned with the rapid increase in the incidence of skin cancer in the United Kingdom. This has led to increasing investigation of factors which may affect an individual's risk of contracting skin cancer. MacKie et al (1989) carried out a matched case / control study where cases and controls were matched by age and sex in an attempt to identify personal risk factors for cutaneous malignant melanoma, the most severe form of skin cancer. One important potential interval scaled discrete risk factor for cutaneous malignant melanoma was thought to be a subject's number of naevi (i.e. moles). In their paper MacKie et al produced personal risk factor charts for cutaneous malignant melanoma for both males ***and*** females. Here a separate univariate analysis for males and females will be carried out on how the number of naevi affects the risk of melanoma.

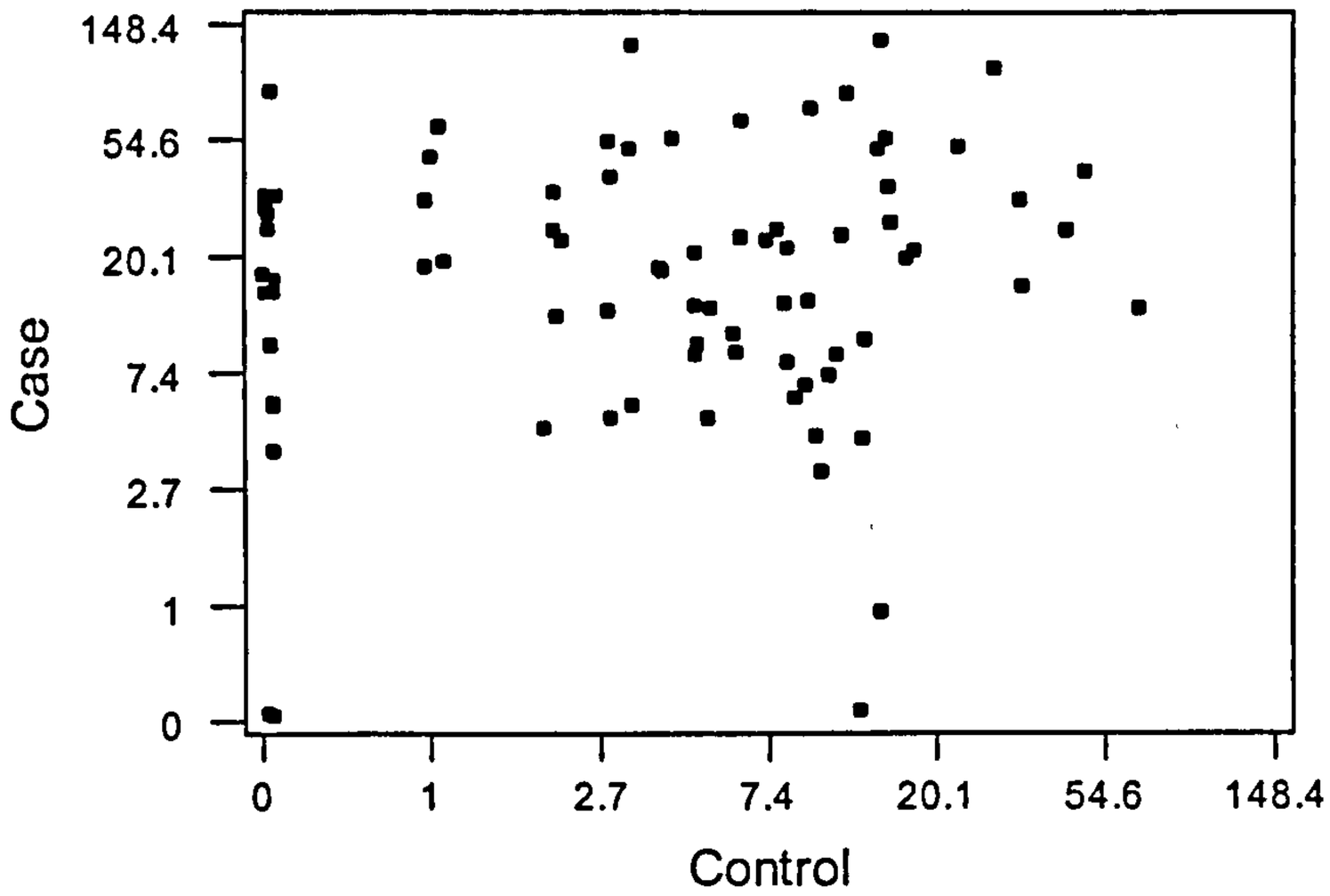
Figure 3.3.0 displays a bivariate plot of the number of naevi for the matched case/control pairs separately for males and females. These plots have been drawn on a log scale (values displayed are of $\log_e$(naevi+1)) with the original naevi values retained on the axes. The plots have also been "jittered" to separate out multiple observations. These plots clearly indicate that, in general, the controls appear to have less naevi than the cases, for both sexes. This is particularly noticeable when the control has zero naevi, as there are a vast number of matched cases who clearly have far more than zero naevi. Conversely when the case has zero naevi there are only a small number of matched controls with more than zero naevi.

Males: The results of fitting a conditional linear logistic regression were

$$\hat{\beta} = 0.087$$

$$\text{ese}(\hat{\beta}) = 0.021$$

Bivariate plot of naevi for MALES (log scale)



Bivariate plot of naevi for FEMALES (log scale)

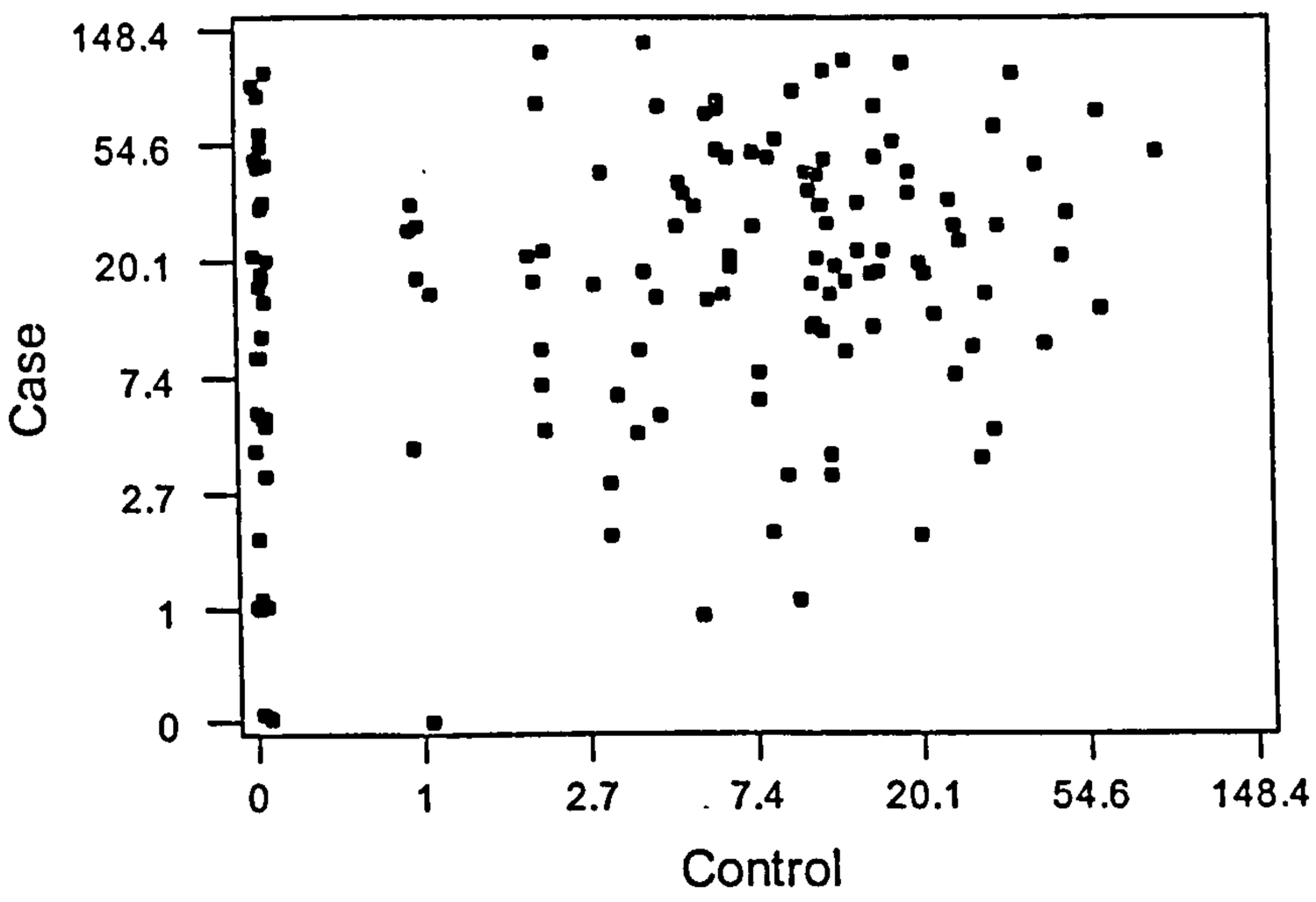


Fig 3.3.0

These estimates allow point estimates and confidence intervals for the Relative Risk to be calculated for any number of naevi using the formulas discussed in section 3.2.3. Figure 3.3.1 shows a plot of the Relative Risk vs the number of naevi the subject had, for males only. The full line on the graph is the point estimate of Relative Risk whereas the dotted lines indicate approximate 95% confidence bands.

This plot shows how the Relative Risk increases on an exponential scale due to the linear logistic assumption and also illustrates the widening of the confidence bands as the number of naevi increases and the data becomes more sparse (see Figure 3.3.0).

Now, Mackie et al suggest that this risk factor may be adequately categorised into two categories namely less than or equal to 20 naevi or greater than 20 naevi. Relative Risks were then calculated for this simple *categorisation*. This gave the following estimates of Relative Risk.

Category	Point estimate of Relative Risk	Confidence Interval for Relative Risk
≤ 20 naevi	1.0	–
> 20 naevi	13.9	(2.7, 71)

Notice that this categorisation changes the baseline from being 0 naevi as used in the conditional linear logistic model to a baseline of less than or equal to 20 naevi. Using less than or equal to 20 naevi as the baseline (and unknown) risk category, then, a subject with 20 or more naevi has an estimated Relative Risk of around 14 compared to a subject with less than or equal to 20 naevi. Later sections within this chapter will discuss methods of justifying such a choice of categorisation. If the linear logistic model displayed in Figure 3.3.1 is adequate then the idea of

having a massive "jump" in the Relative Risk of contracting the disease from 1 to 14 at around 20 naevi seems rather dubious. However, if the model is inadequate then some form of categorisation *might* be possible. It would appear a method of producing a sensible categorisation is required.

Females: The results of fitting a conditional linear logistic regression were

$$\beta = 0.073$$

$$\text{ese}(\hat{\beta}) = 0.013$$

From these estimates Figure 3.3.2 was produced.

Again the categorisation applied to males was also applied to females leading to the following estimates of Relative Risk

Category	Point estimate of Relative Risk	Confidence Interval for Relative Risk
≤ 20 naevi	1.0	–
> 20 naevi	6.7	(2.9,15)

The point esimate for the Relative Risk for a female with more than 20 naevi compared to a female with less than or equal to 20 naevi is about half the value obtained for males in the same comparison. Also the confidence interval is far narrower. The more precise confidence interval is due in part to the fact the there were almost twice as many females in the study as there were males.

It is well documented that females are more likely to contract malignant melanoma than males (Mackie et al (1985), Holman et al (1987), Schreiber et al (1981)) with the ***absolute risk*** of

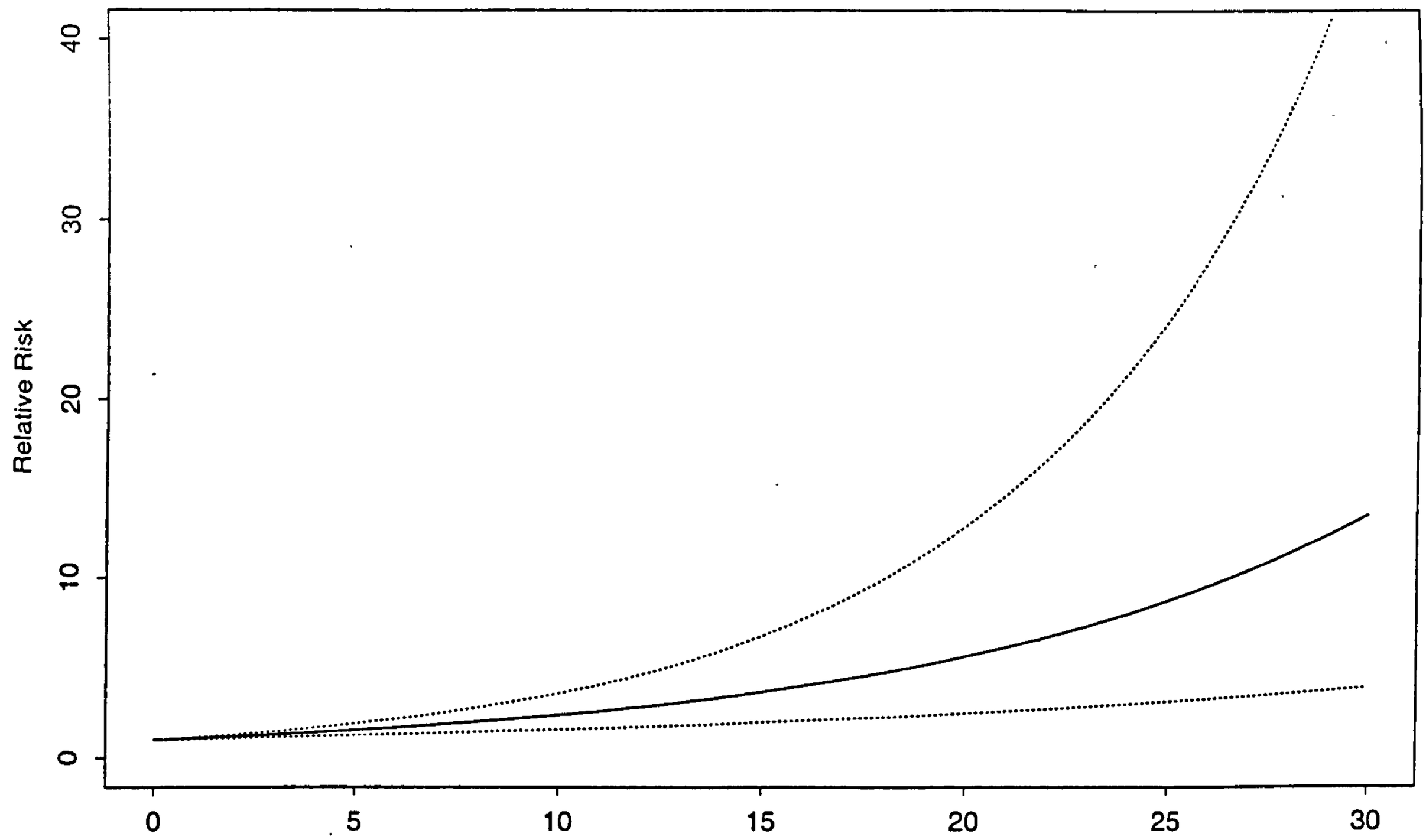


Figure 3.3.1

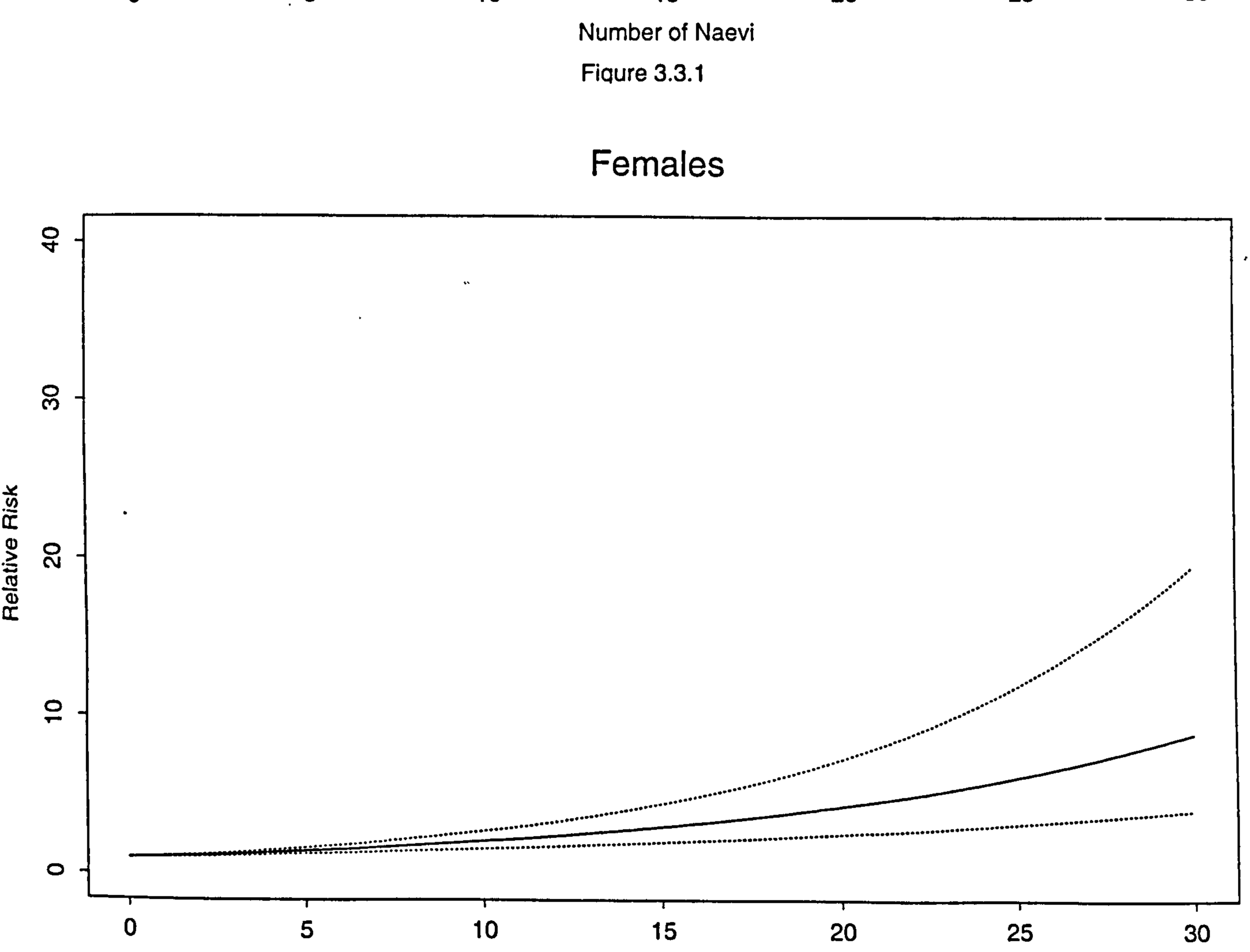


Figure 3.3.2

contracting malignant melanoma likely to be twice as high for females than males. However with this particular risk factor the ***Relative Risk*** of contracting the disease increases more dramatically for males than females (The point estimate of β is ***larger*** for males than females). In other words the ***relative difference*** in the risk of contracting the disease for ***two females***, for example one with 23 naevi and one with 7 naevi, is ***less noticeable*** than for ***two equivalent males***. This would imply that although females are in general more at risk than males in terms of ***contracting*** cutaneous malignant melanoma the ***increasing presence*** of a ***particular*** risk factor has a more pronounced effect for males.

Section 3.4: Non-parametric approaches to analysing data from a matched case/control study

Section 3.4.1: Introduction

Sections 3.2 and 3.3 outlined a standard analysis of a matched case / control study. One of the aims of this thesis is to identify possible categorisations for interval scaled discrete or continuous risk factors in matched case/control studies. In section 3.3 categorisations employed by MacKie et al of an interval scaled discrete potential risk factor were illustrated but it is essential to find some way to justify these choices. In order to ***identify*** potential categorisations for an interval scaled discrete risk factor non-parametric approaches will be used to produce estimates of Relative Risk. Any regions in a plot of the Relative Risk against the interval scaled discrete risk factor where there are rapid changes in the Relative Risk will highlight potential categorisations. Here two possible non-parametric approaches will be discussed.

(1) ***Pairwise Cells Comparison Method*** (Section 3.4.2)

(2) ***Conditional Likelihood Method*** (Section 3.4.3)

Section 3.4.2: Pairwise Cells Comparison

Consider first the case of a single potential risk factor. If the linearity assumption inherent in (3.1) is dropped and βz is replaced instead by an arbitrary smooth function $f(z)$ then the ***model*** is

$$\Pr(\text{subject has the disease} / z_i) = \frac{\exp(f(z_i))}{1 + \exp(f(z_i))} \qquad \text{- (3.4)}$$

where z_i is the value of the risk factor for the i^{th} subject

with the ***conditional likelihood*** being

$$\prod_{i=1}^{I} \frac{\exp(f(x_i))}{\exp(f(x_i)) + \exp(f(y_i))} \qquad \text{- (3.5)}$$

where x_i and y_i are the values of the risk factor for the i^{th} case and the i^{th} control.

Now define the ***odds ratio*** for a subject with a value of the risk factor x compared to a subject with a value of the risk factor y as

$$\psi(x:y) = \exp\{f(x) - f(y)\} \qquad \text{- (3.6)}$$

The motivation for the first non-parametric approach comes from considering what happens with a ***single binary risk factor*** in the linear logistic situation.

Section 3.4.2.1: Binary risk factor

If one returns to the conditional linear logistic model with a single binary risk factor then the conditional likelihood (3.2) simplifies to

$$\prod_{i=1}^{I} \frac{\exp(\beta x_i)}{\exp(\beta x_i) + \exp(\beta y_i)}$$

where

$$x_i , y_i = \begin{cases} 0 & \text{if the risk factor is absent} \\ 1 & \text{if the risk factor is present} \end{cases}$$

A frequency table of the risk factor "of pairs" would look thus

		Number of cases: 0	Number of cases: 1
Number of controls	0	n_{00}	n_{10}
	1	n_{01}	n_{11}

Now the odds-ratio (3.3) for the presence of the risk factor (i.e. x = 1) is defined to be

$$\psi(x=1:y=0) = \psi = \frac{\exp(\beta)}{\exp(0)} \equiv \exp(\beta) \quad \text{- (3.7)}$$

From this it can be seen that for a single binary risk factor, β is the ***true relative log odds of disease*** for an individual in whom the risk factor is ***present*** compared to an individual in whom the risk factor is ***absent.***

The conditional likelihood is now

$$\left(\frac{1}{2}\right)^{n_{00}+n_{11}}\left(\frac{\exp(\beta)}{1+\exp(\beta)}\right)^{n_{10}}\left(\frac{1}{1+\exp(\beta)}\right)^{n_{01}}$$

i.e. effectively

$$\left(\frac{\exp(\beta)}{1+\exp(\beta)}\right)^{n_{10}}\left(\frac{1}{1+\exp(\beta)}\right)^{n_{01}}$$

which on being maximised gives

$$\hat{\beta} \quad = \quad \log\left(\frac{n_{10}}{n_{01}}\right) \qquad \text{- (3.8)}$$

where $\hat{\beta}$ is the ***estimated relative log odds of disease***

and hence the ***odds-ratio*** is estimated by

$$\hat{\psi} \quad = \quad \frac{n_{10}}{n_{01}}$$

This estimate of the odds-ratio gives a point estimate of the Relative Risk of disease caused by the ***presence*** of a single binary risk factor.

Section 3.4.2.2: Extension to a single interval scaled discrete risk factor.

In a matched case/control study with a single interval scaled discrete risk factor with (k+1) levels the data can be easily displayed in a grid form as follows

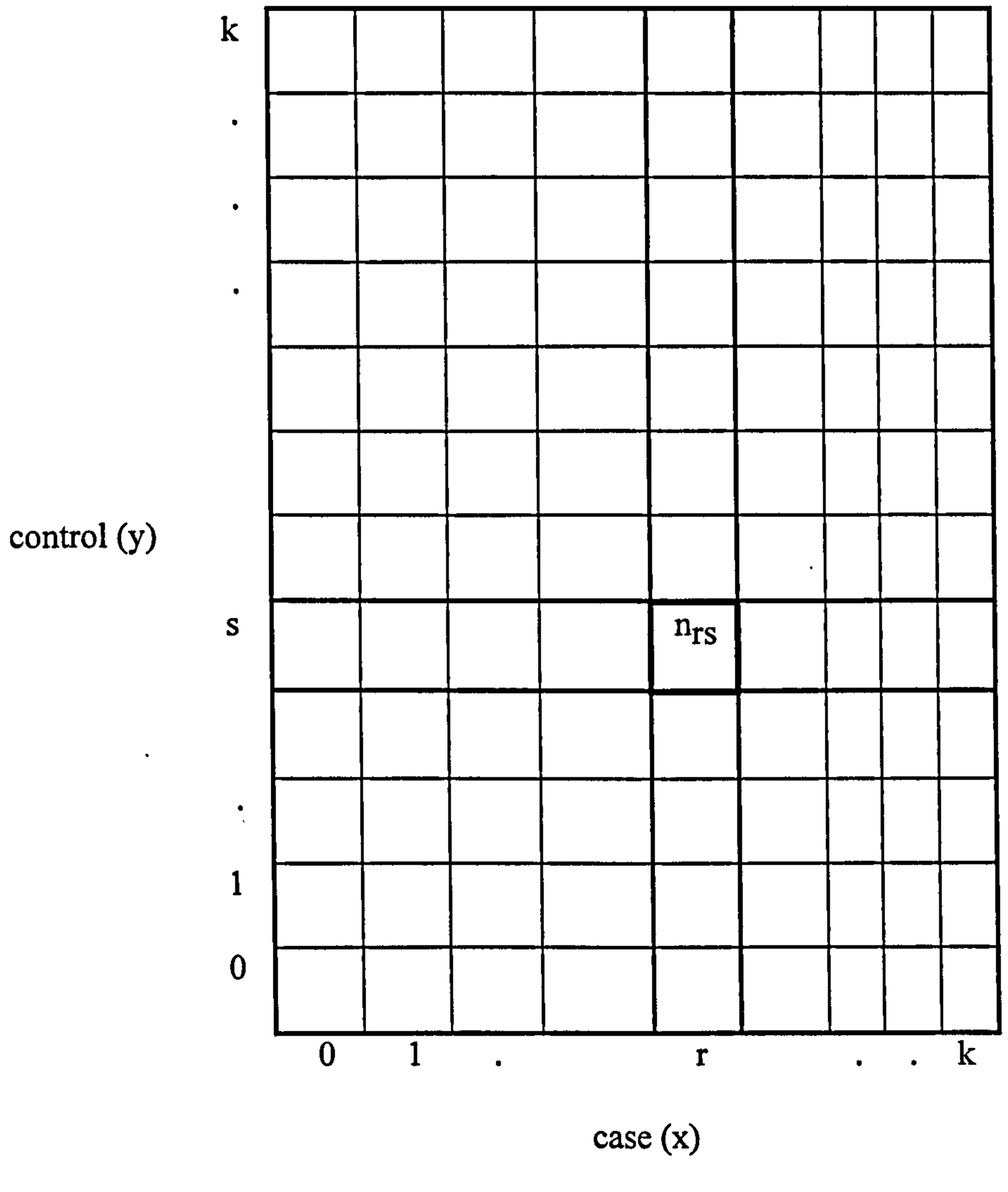


Figure 3.4.1

where

cell (r,s) is the cell where the case has value r and the control has value s

and

n_{rs} is the number of pairs of data in cell (r,s).

The above table now provides a complete summary of the data from a matched case/control study with one interval scaled discrete risk factor present since it displays the number of pairs present in each cell.

If consideration is given to any 2 x 2 subtable of this full grid then the problem could be approached in exactly the same manner as in section 3.4.2.1.

In (3.7) β was the ***true relative log odds of disease*** for an individual in whom the risk factor is ***present*** compared to an individual in whom the risk factor is ***absent***. In (3.7) let β_{10} represent β then in ***general*** β_{xy} will be the true relative log odds of disease for an individual with risk factor equal to x compared to an individual with risk factor equal to y.

From (3.8) let $\hat{\beta}_{10}$ represent $\hat{\beta}$ then for each sub-table separately (3.8) will generalise to give

$$\hat{\beta}_{xy} = \log\left(\frac{n_{xy}}{n_{yx}}\right) \text{ for } x>y \text{ and } x,y = 0, 1, ..., k$$

where $\hat{\beta}_{xy}$ is now the ***estimated relative log odds of disease*** for an individual with risk factor equal to x compared to an individual with risk factor equal to y.

The corresponding ***estimate*** of the ***odds ratio*** is

$$\hat{\psi}(x:y) = \frac{n_{xy}}{n_{yx}} \text{ for } x>y \text{ and } x,y = 0, 1, ..., k \qquad \text{- (3.9)}$$

From (3.6) and the generalised model the ***true odds ratio*** for any level x compared to any level y is

$$\psi(x:y) = \exp\{f(x) - f(y)\}$$

i.e.

$$\log(\psi(x:y)) = f(x) - f(y) \qquad \text{- (3.10)}$$

Therefore in order to produce ***estimates of the Relative Risk*** of disease for any level, x, of the risk factor compared to another level, y, it is necessary to firstly produce estimates of

$f(x)$, $x = 0, 1, ..., k$.

Now if the true odds ratio $\psi(x:y)$ is estimated by $\hat{\psi}(x:y)$ then (3.9) and (3.10) give

$$f(x) - f(y) = \log\left(\frac{n_{xy}}{n_{yx}}\right)$$

for $x>y$ and $x,y = 0, 1, ..., k$

or, equivalently,

$$\hat{\beta}_{xy} = f(x) - f(y) \qquad \text{- (3.11)}$$

for $x>y$ and $x,y = 0, 1, ..., k$

Now, if (3.11) can be solved to produce estimates of $f(x)$, $x = 0, 1, ..., k$ then, by plugging these estimates into (3.6), estimates of the Relative Risk can be produced for any level, x, of the interval scaled discrete risk factor compared to any other level, y.

One possible approach to solving (3.11) is to use the following least squares analogue.

Rewriting equation (3.11) in vector notation and defining the baseline value, $f(0)$, to be equal to 0 then

$$\underline{\hat{\beta}} = A\underline{f}$$

where

$$\underline{\hat{\beta}} = \begin{pmatrix} \hat{\beta}_{10} \\ \hat{\beta}_{20} \\ \hat{\beta}_{21} \\ \hat{\beta}_{30} \\ \hat{\beta}_{31} \\ . \\ . \\ . \end{pmatrix} \qquad A = \begin{pmatrix} 1 & 0 & 0 & . & . & . & . \\ 0 & 1 & 0 & . & . & . & . \\ -1 & 1 & 0 & . & . & . & . \\ 0 & 0 & 1 & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \end{pmatrix} \qquad \underline{f} = \begin{pmatrix} f(1) \\ f(2) \\ f(3) \\ f(4) \\ f(5) \\ . \\ . \\ . \end{pmatrix}$$

and

$\underline{\hat{\beta}}$ is a vector of length $\frac{1}{2}k(k+1)$

$\underline{f}$ is a vector of length k

A is a matrix of dimensions $\frac{1}{2}k(k+1)$ by k

If 'least squares' were appropriate to such a relationship then, treating $\underline{f}$ as unknown parameters, one would have

$$\underline{\hat{f}} = (A^T A)^{-1} A^T \underline{\hat{\beta}} \qquad \text{- (3.12)}$$

which can be written in the form

$$\hat{f}(z) = \frac{\hat{\beta}_{z+} - \hat{\beta}_{z-} + K}{1 + k} \qquad \text{for } z = 1, \ldots, k$$

where $\hat{\beta}_{z+} = \sum_{y=0}^{z-1} \hat{\beta}_{zy}$

$$\hat{\beta}_{z-} = \sum_{y=z+1}^{k} \hat{\beta}_{yz}$$

$$K = \sum_{x=0}^{k} (\hat{\beta}_{x+} - \hat{\beta}_{x-})$$

$k+1$ = Number of levels of risk factor

If it is then possible to solve these equations, estimates of f(z) can be produced for any z.

Hence, through (3.6), estimates of the odds ratio between any two levels of the variable can be calculated. In particular if each level is compared to a chosen baseline (i.e. z = 0) where the risk is equal to 1 then the odds ratio

$$\hat{\psi}(z:0) \quad = \quad \exp(\hat{f}(z))/1 \equiv \quad \exp(\hat{f}(z))$$

This gives a ***point estimate*** for the Relative Risk of disease for any level of a variable compared to a chosen baseline.

Section 3.4.2.3: Confidence Intervals for the Relative Risk

To produce a confidence interval for the Relative Risk, it is necessary to first provide an estimate of the variance of $f(z)$. This could be done by considering a simple Taylor expansion.

Recall the definition of $\hat{\beta}_{xy}$ from section 3.4.2.2

$$\hat{\beta}_{xy} \quad = \quad \log\left(\frac{n_{xy}}{n_{yx}}\right) \qquad x, y = 1, ..., k \quad x > y$$

Conditional on the pairwise totals (i.e. $n_{xy} + n_{yx}$), the counts, n_{xy} $x > y$, can be assumed independent of each other and distributed as $Bi(n_{xy} + n_{yx}, \theta_{xy})$. An application of a 1st order Taylor expansion to $\hat{\beta}_{xy}$ provides

$$\hat{V}(\hat{\beta}_{xy}) \quad \approx \quad \frac{1}{n_{xy}} + \frac{1}{n_{yx}}$$

Also, the variance-covariance matrix of $\underline{\hat{\beta}}$, $\hat{V}_1\left(\underline{\hat{\beta}}\right)$, is of the form

$$\hat{V}_1\left(\underline{\hat{\beta}}\right) = \begin{bmatrix} \hat{V}\left(\hat{\beta}_{10}\right) & \hat{cov}\left(\hat{\beta}_{10},\hat{\beta}_{20}\right) & \hat{cov}\left(\hat{\beta}_{10},\hat{\beta}_{21}\right) & \cdots & \hat{cov}\left(\hat{\beta}_{10},\hat{\beta}_{k,k-1}\right) \\ \hat{cov}\left(\hat{\beta}_{20},\hat{\beta}_{10}\right) & \hat{V}\left(\hat{\beta}_{20}\right) & \hat{cov}\left(\hat{\beta}_{20},\hat{\beta}_{21}\right) & \cdots & \hat{cov}\left(\hat{\beta}_{20},\hat{\beta}_{k,k-1}\right) \\ \hat{cov}\left(\hat{\beta}_{21},\hat{\beta}_{10}\right) & \hat{cov}\left(\hat{\beta}_{21},\hat{\beta}_{20}\right) & \hat{V}\left(\hat{\beta}_{21}\right) & \cdots & \hat{cov}\left(\hat{\beta}_{21},\hat{\beta}_{k,k-1}\right) \\ \vdots & \vdots & \vdots & & \vdots \\ \hat{cov}\left(\hat{\beta}_{k,k-1},\hat{\beta}_{10}\right) & \hat{cov}\left(\hat{\beta}_{k,k-1},\hat{\beta}_{20}\right) & \hat{cov}\left(\hat{\beta}_{k,k-1},\hat{\beta}_{21}\right) & \cdots & \hat{V}\left(\hat{\beta}_{k,k-1}\right) \end{bmatrix}$$

with all off diagonal covariance terms in $\hat{V}_1\left(\underline{\hat{\beta}}\right)$ being equal to zero due to the independence of the counts n_{xy}.

Using these results the corresponding analogue to (3.12) gives an approximate variance matrix for $\underline{f}$

$$\hat{V}_1\left(\underline{\hat{f}}\right) = \left(A^T A\right)^{-1} A^T \hat{V}_1\left(\underline{\hat{\beta}}\right) A \left(A^T A\right)^{-1}$$

Based on the approximate normality of $\underline{\hat{f}}$ one can provide an approximate $100*(1-\alpha)\%$ confidence interval for each level of the Relative Risk of the form

$$\left(\hat{f}(z)\right)^{\exp\left[\pm z_{\alpha/2}\sqrt{\hat{V}_1\left(\hat{f}(z)\right)}\right]}$$

where $z_{\alpha/2}$ is the $100*\left(1-\alpha/2\right)$ percentage point of the standard normal.

Section 3.4.2.4: Inclusion of Covariance terms

The use of non-parametric approaches to the analysis of small data sets often results in some form of data smoothing having to be used. In this chapter the data will be smoothed via a nearest neighbour smoothing technique (see Section 3.4.4 for a full definition). When smoothing is present the counts n_{xy} will no longer be independent. The inclusion of the covariance between the counts may lead to more accurate estimates of the Relative Risk being produced. The distribution of the counts can be adequately described as

$$\underline{n} \sim Mu(N, \underline{\theta}) \qquad \text{where} \quad \underline{n} = (n_{00}, n_{01}, n_{02}, \ldots, n_{kk})$$

$$N = \sum_{i=0}^{k} \sum_{j=0}^{k} n_{ij}$$

$$\underline{\theta} = (\theta_{00}, \theta_{01}, \theta_{02}, \ldots, \theta_{kk})$$

As before,

$$\underline{\hat{\beta}} = (\hat{\beta}_{10}, \hat{\beta}_{20}, \hat{\beta}_{21}, \hat{\beta}_{30}, \hat{\beta}_{31}, \ldots)^T$$

with

$$\hat{\beta}_{xy} = \log\left(\frac{n_{xy}}{n_{yx}}\right) \qquad x, y = 1, \ldots, k \qquad x > y$$

and

$$\hat{V}_2\left(\underline{\hat{\beta}}\right) = \begin{bmatrix} \hat{V}\left(\hat{\beta}_{10}\right) & \hat{cov}\left(\hat{\beta}_{10},\hat{\beta}_{20}\right) & \hat{cov}\left(\hat{\beta}_{10},\hat{\beta}_{21}\right) & \cdots & \hat{cov}\left(\hat{\beta}_{10},\hat{\beta}_{k,k-1}\right) \\ \hat{cov}\left(\hat{\beta}_{20},\hat{\beta}_{10}\right) & \hat{V}\left(\hat{\beta}_{20}\right) & \hat{cov}\left(\hat{\beta}_{20},\hat{\beta}_{21}\right) & \cdots & \hat{cov}\left(\hat{\beta}_{20},\hat{\beta}_{k,k-1}\right) \\ \hat{cov}\left(\hat{\beta}_{21},\hat{\beta}_{10}\right) & \hat{cov}\left(\hat{\beta}_{21},\hat{\beta}_{20}\right) & \hat{V}\left(\hat{\beta}_{21}\right) & \cdots & \hat{cov}\left(\hat{\beta}_{21},\hat{\beta}_{k,k-1}\right) \\ \vdots & \vdots & \vdots & & \vdots \\ \hat{cov}\left(\hat{\beta}_{k,k-1},\hat{\beta}_{10}\right) & \hat{cov}\left(\hat{\beta}_{k,k-1},\hat{\beta}_{20}\right) & \hat{cov}\left(\hat{\beta}_{k,k-1},\hat{\beta}_{21}\right) & \cdots & \hat{V}\left(\hat{\beta}_{k,k-1}\right) \end{bmatrix}$$

$\hat{V}_2\left(\underline{\hat{\beta}}\right)$ is a more complex variance-covariance matrix than $\hat{V}_1\left(\underline{\hat{\beta}}\right)$ as, due to the covariance between the counts n_{xy}, the off diagonal covariance terms are not equal to zero.

However, a 1st order Taylor expansion can be used to obtain

$$\hat{cov}\left(\hat{\beta}_{xy},\hat{\beta}_{x^*y^*}\right) = \hat{cov}\left(\log\left(\frac{n_{xy}}{n_{yx}}\right),\log\left(\frac{n_{x^*y^*}}{n_{y^*x^*}}\right)\right) \qquad \text{for} \quad x \neq y, \quad x \leq x^*, \quad y \leq y^*$$

An example of calculating these covariance terms with ***a first order neighbourhood of smoothing*** is given in Appendix A.

From here one can obtain

$$\hat{\underline{f}} = \left(A^T \hat{V}_2\left(\underline{\hat{\beta}}\right)^{-1} A\right)^{-1} A^T \hat{V}_2\left(\underline{\hat{\beta}}\right)^{-1} \underline{\hat{\beta}}$$

The approximation also results

$$\hat{V}_2\left(\hat{\underline{f}}\right) = \left(A^T \hat{V}_2\left(\underline{\hat{\beta}}\right)^{-1} A\right)^{-1} A^T \hat{V}_2\left(\underline{\hat{\beta}}\right)^{-1} \hat{V}_2\left(\underline{\hat{\beta}}\right)\left[\hat{V}_2\left(\underline{\hat{\beta}}\right)^{-1}\right]^T A\left(A^T \hat{V}_2\left(\underline{\hat{\beta}}\right)^{-1} A\right)^{-1}$$

Now, since $\hat{V}_2\left(\hat{\underline{\beta}}\right)$ is symmetric, then $\left[\hat{V}_2\left(\hat{\underline{\beta}}\right)^{-1}\right]^T = \hat{V}_2\left(\hat{\underline{\beta}}\right)^{-1}$.

Therefore the above simplifies to give

$$\hat{V}_2\left(\hat{\underline{f}}\right) = \left(A^T \hat{V}_2\left(\hat{\underline{\beta}}\right)^{-1} A\right)^{-1}$$

Based on the approximate normality of $\hat{\underline{f}}$ one can provide an approximate 100*(1-α)% confidence interval for each level of the Relative Risk of the form

$$\left(\hat{f}(z)\right)^{\exp\left[\pm z_{\alpha/2}\sqrt{\hat{V}_2\left(\hat{f}(z)\right)}\right]}$$

where $z_{\alpha/2}$ is the $100*\left(1-\alpha/2\right)$ percentage point of the standard normal.

Section 3.4.3: Conditional Likelihood Method

The ***conditional likelihood*** based on the conditional linear logistic model was defined in (3.2). If the linear assumption inherent in this model is dropped then (3.5) gave the conditional likelihood as

$$\prod_{i=1}^{I} \frac{\exp(f(x_i))}{\exp(f(x_i)) + \exp(f(y_i))}$$

In general let $p_i = \exp(f(z_i))$ and consider all possible case / control pairs. For a single interval scaled discrete variable with (k+1) levels the conditional likelihood then becomes a product over $(k+1)^2$ cells and is of the form

$$\prod_{i=0}^{k}\prod_{j=0}^{k}\left(\frac{p_i}{p_i+p_j}\right)^{n_{ij}} \qquad \text{- (3.13)}$$

where

p_i = Relative Risk of category i compared to the baseline, (i.e. 0)

for i = 1, .., k

(i.e. p_i = $\exp(f(z_i))$)

n_{ij} = number of case/control pairs in cell (i,j)

k+1 = Number of levels of risk factor

Then maximise (3.13) to obtain estimates of p_i and hence directly obtain estimates of the Relative Risk. Unfortunately this problem cannot always be solved analytically and numerical methods are often required to solve it.

The use of the Newton-Raphson method to solve this not only provides point estimates for the Relative Risk but also allows one to construct interval estimates by producing an estimate of the variance of the Relative Risk through the information matrix.

Let $T = \left[t_{ij} \right]$ be the k by k information matrix with

$$t_{ij} = -\frac{\partial^2 L}{\partial p_i \partial p_j} = \begin{cases} \dfrac{n_{ij} + n_{ji}}{(p_i + p_j)^2} & \text{for} \quad i \neq j \\ \left(\dfrac{1}{(p_i)^2}\right) \sum\limits_{\substack{w=0 \\ w \neq i}}^{k} n_{iw} + \sum\limits_{m=0}^{k} \dfrac{(n_{im} + n_{mi})}{(p_i + p_m)^2} & \text{for} \quad i = j \end{cases}$$

Then $\hat{T}^{-1}$ is the ***asymptotic*** variance-covariance matrix of $\underline{\hat{p}}$.

A marginal approximate 100*(1-α)% confidence interval for p_i is then given by

$$\hat{p}_i \pm z_{\alpha/2} \sqrt{\hat{w}_{ii}} \qquad \text{- (3.14)}$$

where

$\hat{w}_{ii}$ is the i'th diagonal element of $\hat{W} = \hat{T}^{-1}$

$z_{\alpha/2}$ is the $100*\left(1-\alpha/2\right)$ percentage point of the standard normal.

However $p_i = \exp(f(z_i))$ which implies that $\hat{p}_i$ will be constrained to take only positive values on the real line. Therefore, instead of assuming $\hat{p}_i$ to be asymptotically normal, it seems more logical to produce confidence intervals for p_i based on using the function $\log(\hat{p}_i)$ as a pivotal function. This leads to the following approximate 100*(1-α)% interval for p_i.

$$\hat{p}_i^{\exp\left[\pm z_{\alpha/2}\sqrt{\left(\frac{1}{\hat{p}_i}\right)^2 \hat{w}_{ii}}\right]} \qquad \text{- (3.15)}$$

Illustration for a risk factor with (a) **2 levels** and (b) **3 levels**

(a) If a risk factor is present with ***2 levels*** then (3.13) gives the conditional likelihood to be

$$\text{Conditional likelihood} = \prod_{i=0}^{1}\prod_{j=0}^{1}\left(\frac{p_i}{p_i + p_j}\right)^{n_{ij}}$$

$$\alpha \quad \left(\frac{1}{p_1+1}\right)^{n_{01}}\left(\frac{p_1}{p_1+1}\right)^{n_{10}}$$

$$= \quad \left(\frac{1}{p_1+1}\right)^{n_{01}+n_{10}} p_1^{\,n_{10}}$$

Maximise this by taking logarithms and then solving the first derivative equal to zero

$$\log(\text{Conditional Likelihood}) = \quad -(n_{01}+n_{10})\log(p_1+1) + n_{10}\log(p_1)$$

Therefore

$$\frac{dL}{dp_1} = \frac{n_{10}}{p_1} - \frac{(n_{01}+n_{10})}{p_1+1}$$

giving as an estimate of the Relative Risk for the ***presence*** of the risk factor compared to the ***absence*** of the risk factor as

$$\hat{p}_1 = \frac{n_{10}}{n_{01}}$$

As expected this produces the standard result for a binary risk factor (i.e. one with 2 levels) given in section 3.4.2.1.

(b) If a risk factor is present with ***3 levels*** then (3.13) gives the conditional likelihood to be

$$\text{Conditional likelihood} = \prod_{i=0}^{2}\prod_{j=0}^{2}\left(\frac{p_i}{p_i + p_j}\right)^{n_{ij}}$$

$$= \left(\frac{1}{p_1+1}\right)^{n01}\left(\frac{1}{p_2+1}\right)^{n02}\left(\frac{p_1}{p_1+1}\right)^{n10}\left(\frac{p_1}{p_1+p_2}\right)^{n12}\left(\frac{p_2}{p_2+1}\right)^{n20}\left(\frac{p_2}{p_1+p_2}\right)^{n21}$$

$$= \left(\frac{1}{p_1+1}\right)^{n01+n10}\left(\frac{1}{p_2+1}\right)^{n02+n20} p_1^{n10+n12} p_2^{n20+n21}\left(\frac{1}{p_1+p_2}\right)^{n12+n21}$$

Maximise this by taking logarithms and then solving the relevant ***partial derivatives*** equal to zero

$$\log(\text{Conditional Likelihood}) = -(n_{01}+n_{10})\log(p_1+1) - (n_{02}+n_{20})\log(p_2+1) + (n_{10}+n_{12})\log(p_1) + (n_{20}+n_{21})\log(p_2) - (n_{12}+n_{21})\log(p_1+p_2)$$

Therefore

$$\frac{\partial L}{\partial p_1} = -\frac{(n_{01}+n_{10})}{p_1+1} + \frac{(n_{10}+n_{12})}{p_1} - \frac{(n_{12}+n_{21})}{p_1+p_2}$$

$$\frac{\partial L}{\partial p_2} = -\frac{(n_{02}+n_{20})}{p_2+1} + \frac{(n_{20}+n_{21})}{p_2} - \frac{(n_{12}+n_{21})}{p_1+p_2}$$

To obtain estimates $\hat{p}_1, \hat{p}_2$ for p_1, p_2 it is necessary to solve these equations simultaneously . No simple analytical solution exists and hence numerical methods such as Newton Raphson are required to solve this problem.

Section 3.4.4: Nearest Neighbour Smoothing

The techniques described in sections 3.4.2 and 3.4.3 both require use of the observed number of case/control pairs, n_{ij}, in cell (i,j). If the sample size were large then there would always be a reasonable number of case/control pairs in each cell. However, since case control studies are primarily used in the study of rare diseases (Section 3.1) it is often the case that the data will be very sparse. For example even in a relatively large case/control study with, say, 200 subjects if a risk factor is being studied which has more than 15 levels it will be impossible to have even one observation in each of the 225 possible case/control cells.

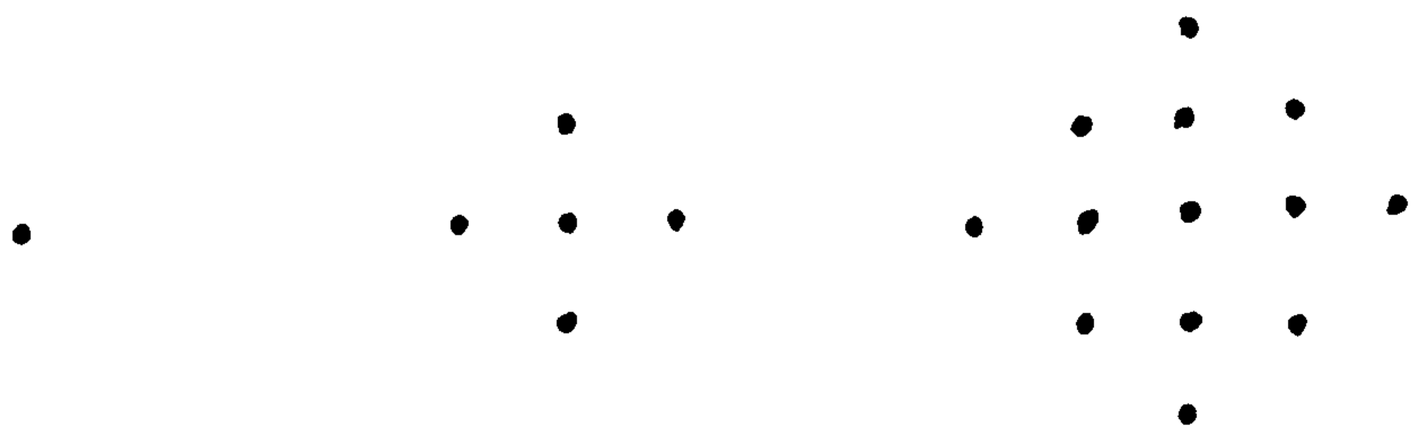
In order to try and get a clearer picture of the pattern across cells where there is little or no data it is potentially useful to introduce some form of smoothing across neighbouring cells. This allows more information to be gleaned about any cell by considering what is occurring in a neighbourhood of the cell. When smoothing the data each case/control pair has an influence on all possible cells which decreases as one moves away from their particular cell.

Instead of using the raw count n_{xy} in each cell define the neighbourhood count, n^{+}_{xy}, of a cell as follows

$$n^{+}_{xy} = \begin{cases} \sum_{L(x,y):\ x<y} n_{xy} & \text{if } x < y \\ \sum_{L(x,y)} n_{xy} & \text{if } x = y \\ \sum_{L(x,y):\ x>y} n_{xy} & \text{if } x > y \end{cases}$$

The neighbourhood count n^{+}_{xy} is the count obtained by summing the count obtained from *all* the cells in a local neighbourhood L(x,y) of the cell (x,y). Note that the whole area of interest is divided into two regions by the line x = y and only cells in the ***same region*** as the cell (x,y) are used in calculating the neighbourhood count for cell (x,y).

This will produce a set of neighbourhood counts rather than raw counts which will hopefully give a clearer picture of what is happening in the neighbourhood of a cell than can be obtained purely from the raw cell count. Various sizes of neighbourhoods are possible and Figure 3.4.2 illustrates three of these.



No smoothing First order neighbourhood Second order neighbourhood

Figure 3.4.2

The larger the neighbourhood the more the data is smoothed. In general the more sparse a data set is the larger the degree of smoothing required to obtain a useful and hopefully still informative picture of the underlying data pattern. In the specific examples here an appropriate size of neighbourhood is chosen on the basis of a compromise between producing estimates of Relative Risk which did not fluctuate wildly but also trying to avoid completely smoothing out any underlying patterns / trends in the Relative Risk.

Section 3.5: Cutaneous malignant melanoma revisited: An application of non-parametric methods to analysing data from a case/control study.

Section 3.5.1: Introduction

This section will illustrate both of the non-parametric methods discussed in Section 3.4. A comparison of the two methods will be made in the context of whether they highlight similar cut-points for a particular interval scaled discrete risk factor. Once again the cutaneous malignant melanoma data set from section 3.3 will be examined and the Relative Risk associated with number of naevi will be discussed. MacKie et al suggested that this risk factor be split into two categories by choosing a somewhat arbritary cut-point at 20 naevi. This section will attempt to justify such a choice of cut-point by the techniques introduced in Section 3.4.

Section 3.5.2: Pairwise Cells Comparison

Figures 3.5.1 and 3.5.2 provide plots of the estimates of Relative Risk based on the Pairwise cells approach plotted against the number of naevi separately for males and females. Both of these figures are based on a first order neighbourhood of smoothing. The full line on these plots represents the best point estimate of Relative Risk while the dotted lines represent confidence bands for the Relative Risk. The point estimates and confidence intervals are based on the formulae given in Section 3.4.2.3 and do not include any covariance terms.

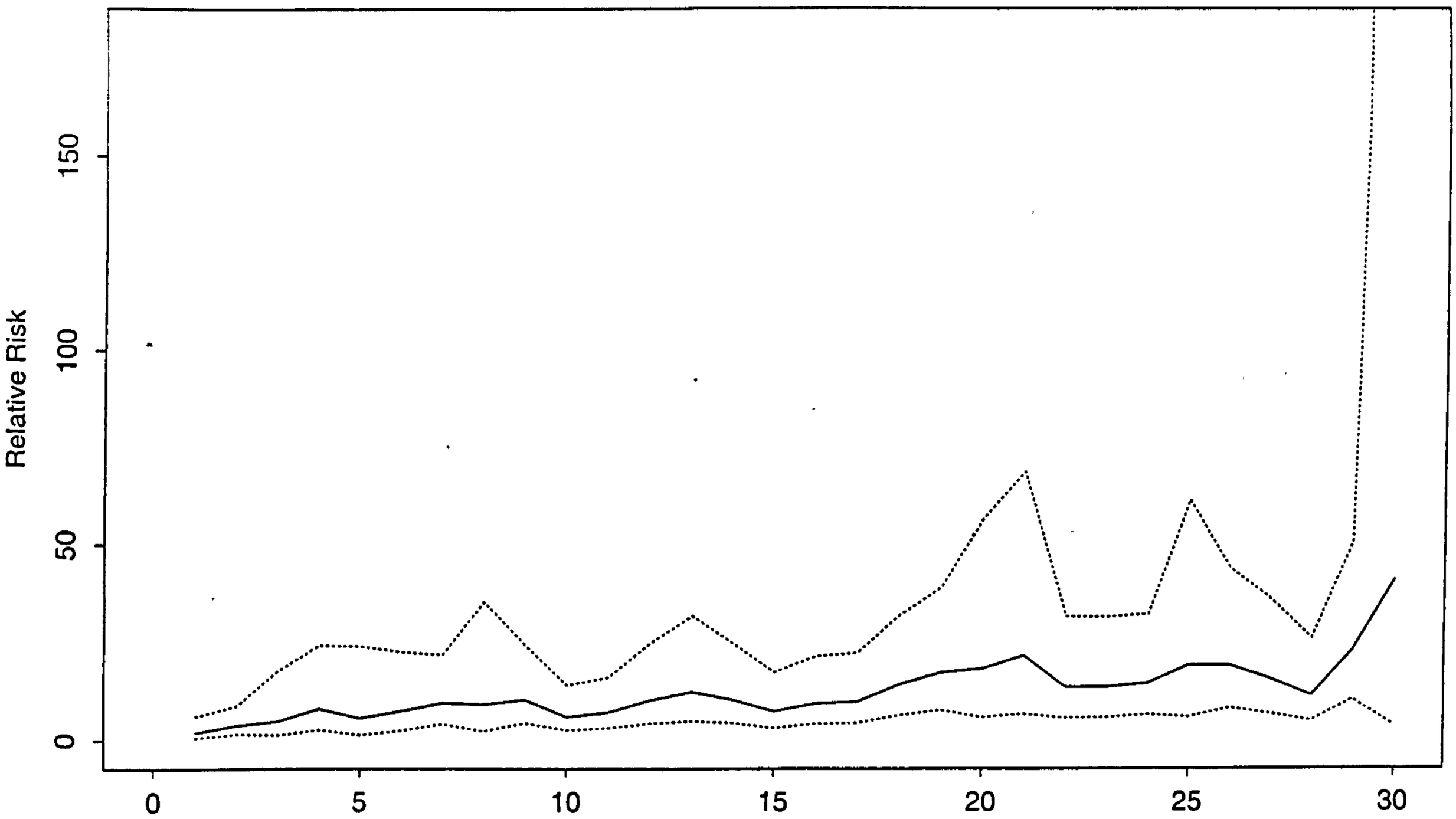


Figure 3.5.1

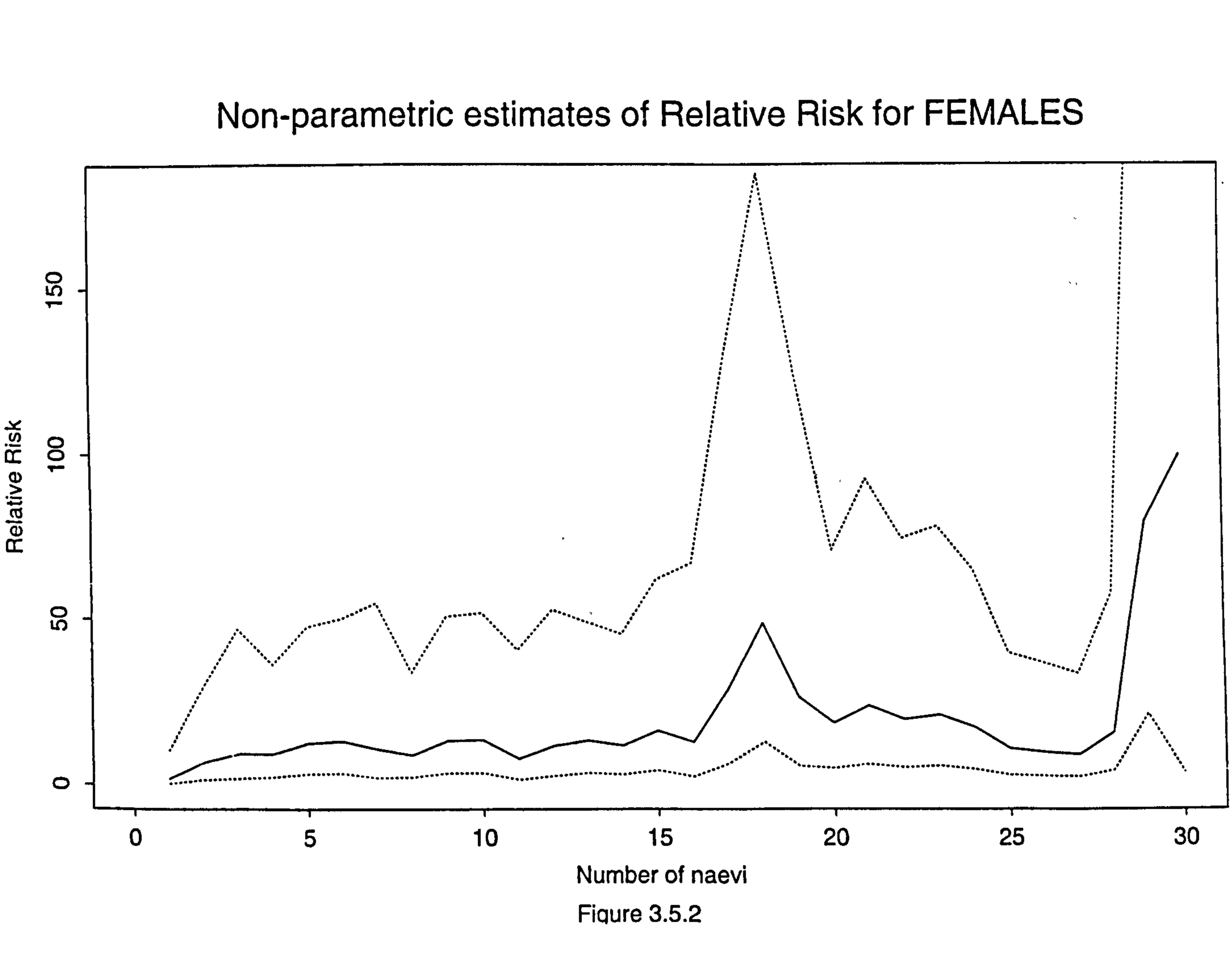


Figure 3.5.2

The main point of using this technique is to attempt to identify potential cutpoints for the interval scaled discrete risk factor of number of naevi. In order to discover if any cut-points are appropriate it is necessary to examine these plots in greater detail. Figures 3.5.3 and 3.5.4 show separately for males and females, plots of the ***point estimates*** of Relative Risk. These are indicated by the dots on the Figures. In order to give a clearer picture of any patterns in the Relative Risk the bold line is the estimate obtained after running a simple kernel regression smoother (Nadaraya (1964) and Watson (1964)) through these original values. This technique is in essence similar to the method discussed in section 4 of chapter 2, the difference being that the response here is continuous/interval scaled discrete compared with binary in chapter 2.

With respect to any possible categorisations, Figure 3.5.3 seems to suggest that if cutpoints are desired for males then perhaps only one is necessary and that it should be somewhere around 17 or 18 naevi since this is where the change in Relative Risk appears most dramatic.

For females Figure 3.5.4 would again suggest a cutpoint around about 17 naevi but notice here that something unusual appears to be happening after 17 naevi as the risk appears to drop back down. This is something which one would not expect but may be a quirk of this particular data set perhaps due to a lack of data in this area.

The application of the conditional linear logistic model in section 3.3 to this data set resulted in the conclusion that the Relative Risk of contracting malignant melanoma increased more dramatically among males than among females as the number of naevi increased. This was shown by the higher parameter estimate in the fitted model for males leading to a steeper gradient on the Relative Risk curve. The graphs of Relative Risk presented in this section are ***not*** in agreement with these results as they give a ***different picture*** of the pattern in the Relative Risk. Figures 3.5.3 and

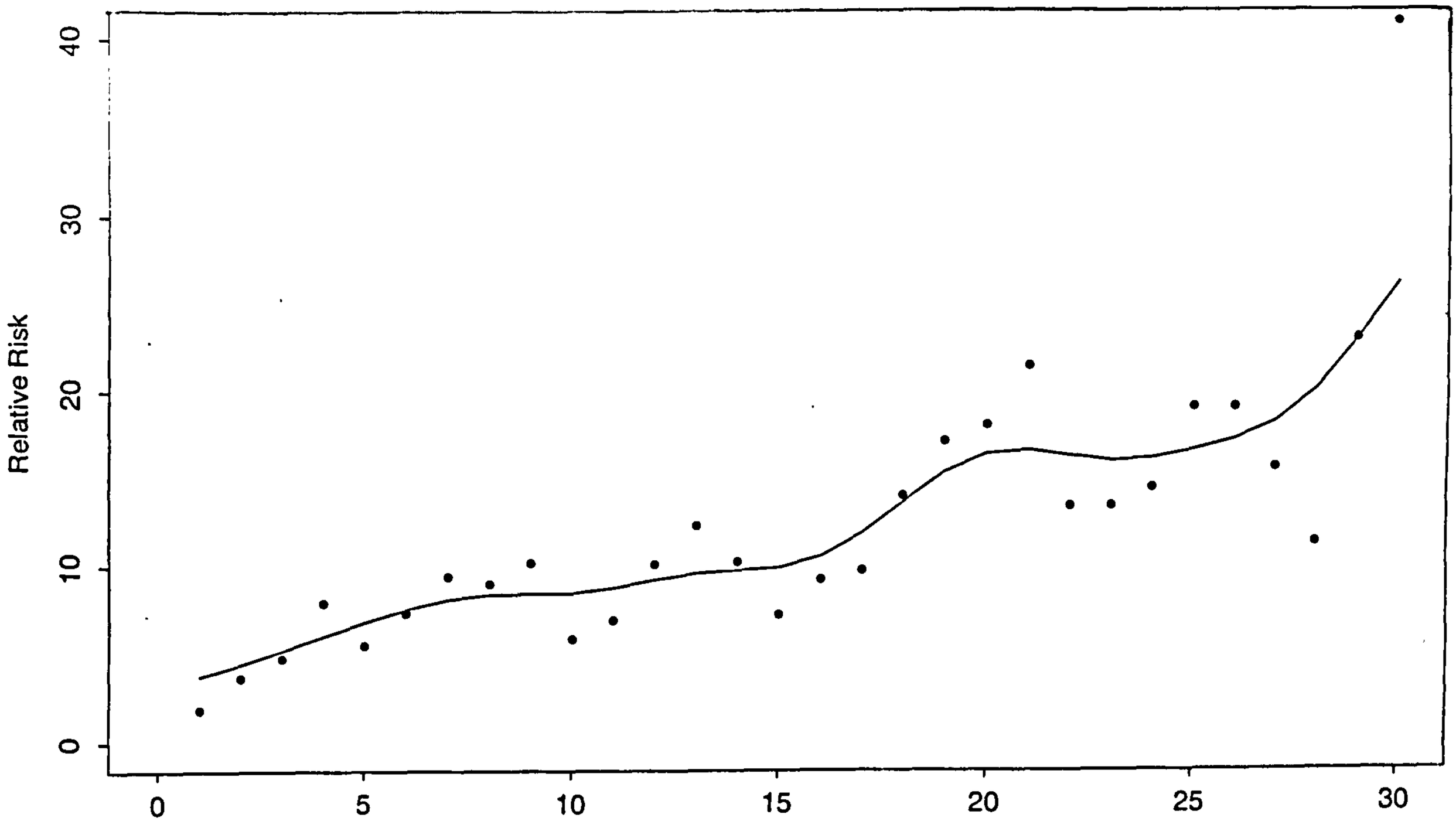


Figure 3.5.3

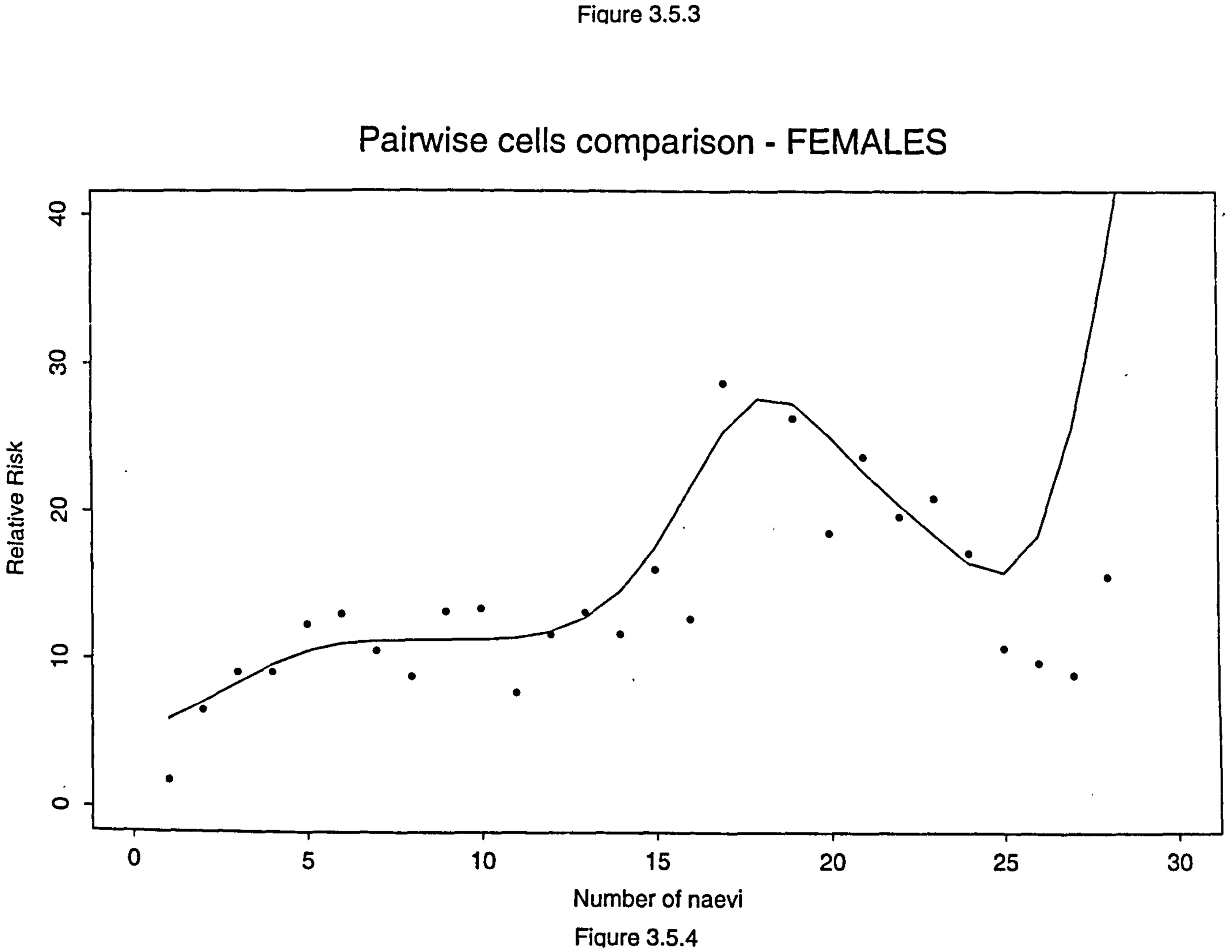


Figure 3.5.4

3.5.4 do not give the impression that the Relative Risk increases more dramatically among males. In fact it appears as if the gradient of the female curve is, if anything, steeper than that of the male curve suggesting that the risk ***increases more dramatically*** among females. This is particularly the case between 10 and 20 naevi although the "fall-back" for females for higher numbers of naevi is probably the reason why the linear logistic model appears flatter for females.

Section 3.5.3: Conditional Likelihood Method

Figure 3.5.5 shows a plot of the non-parametric estimate of Relative Risk against number of naevi for males based on the Conditional Likelihood approach of Section 3.4.3 while Figure 3.5.6 shows a similar plot for females. Both of these figures are again based on a first order neighbourhood of smoothing.

As in section 3.5.2 a kernel smoother was again run through these original point estimates to obtain Figures 3.5.7 and 3.5.8. These two plots are similar in shape and scale to those obtained by the pairwise cells comparison method which is reassuring. Again they seem to imply a categorisation taking place around about 17 naevi for both males and females, with something strange appearing to occur later on in females.

Figures 3.5.7 and 3.5.8 are again in disagreement with the results given by the conditional linear logistic model in section 3.3 as they also do not give the impression that the Relative Risk increases more for males than females as the number of naevi increases. As with the pairwise cells method in section 3.5.2 the estimates of Relative Risk obtained by the conditional likelihood method perhaps suggest that the risk increases more for females between 10 and 20 naevi. This lack of agreement between the parametric and nonparametric approaches would imply that the use of the conditional linear logistic model for this particular data set is somewhat dubious.

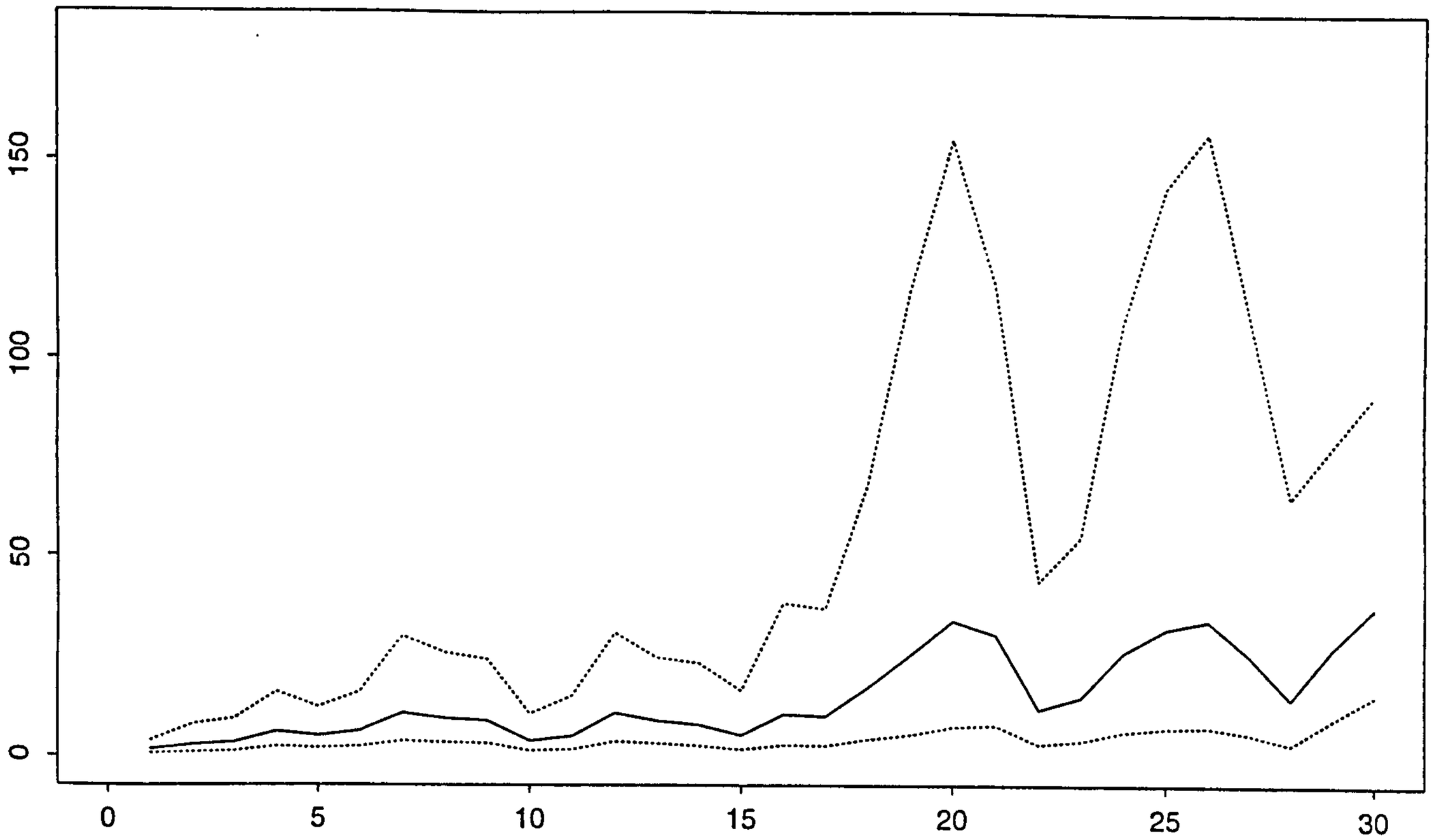


Figure 3.5.5

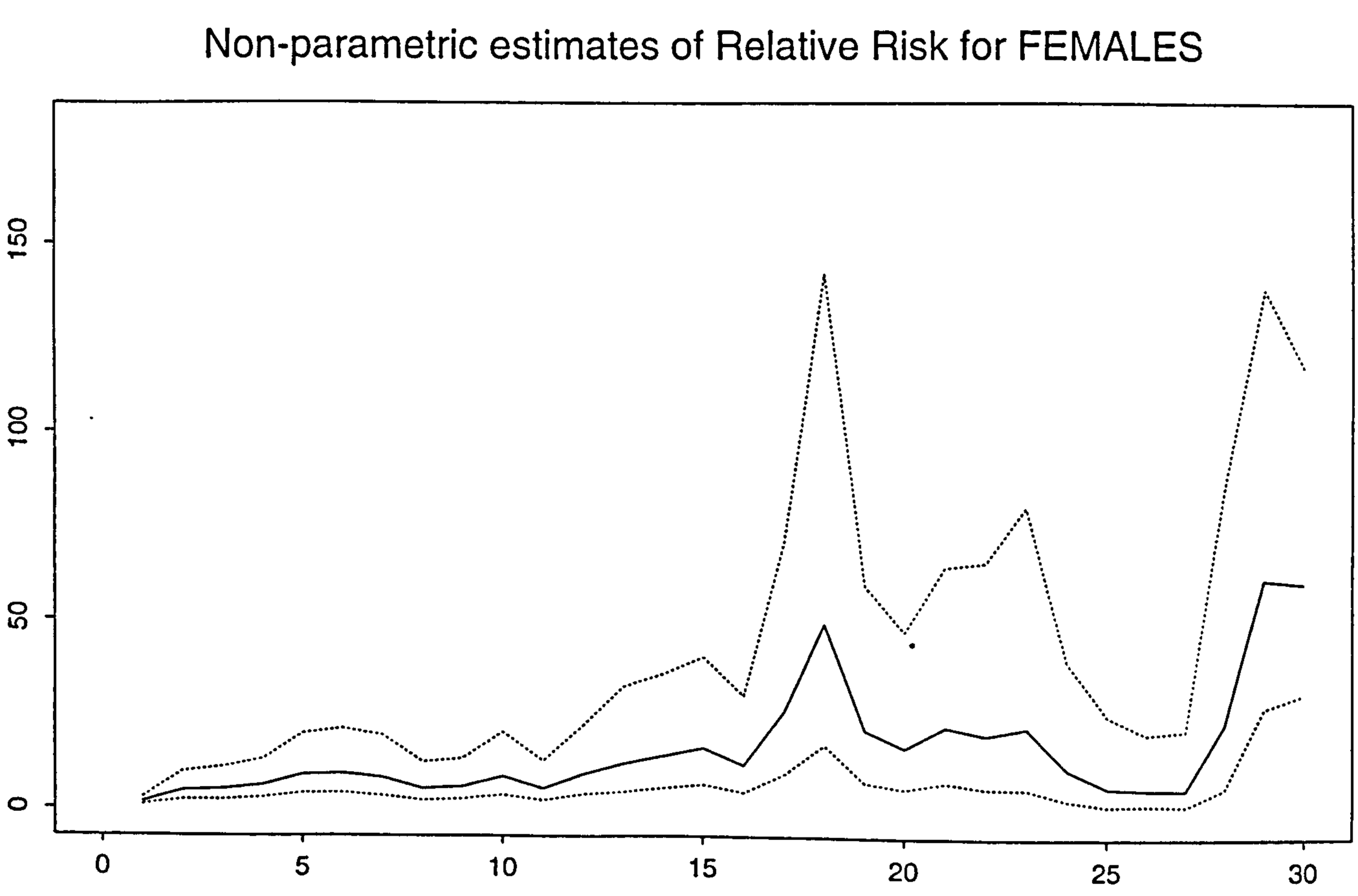


Figure 3.5.6

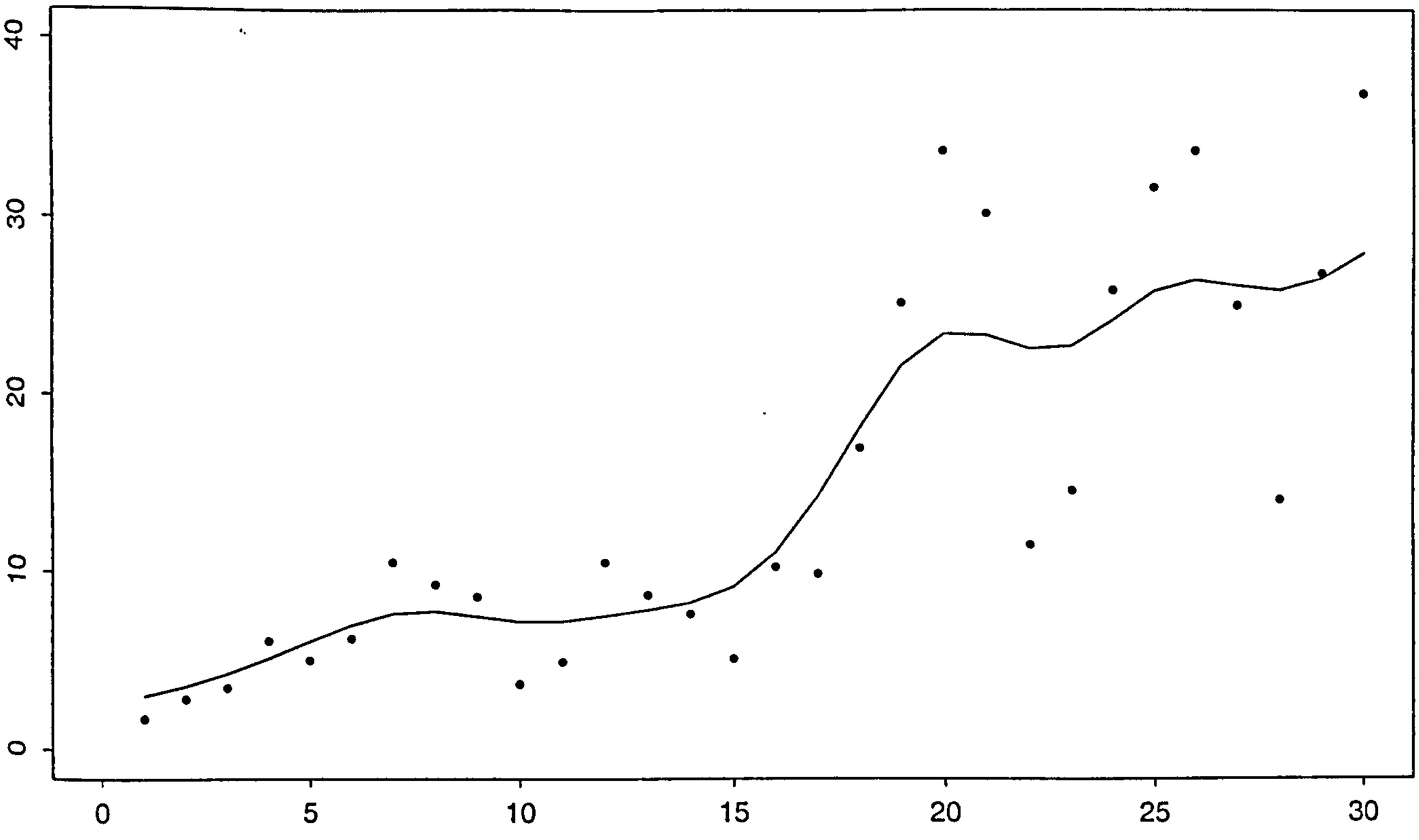


Figure 3.5.7

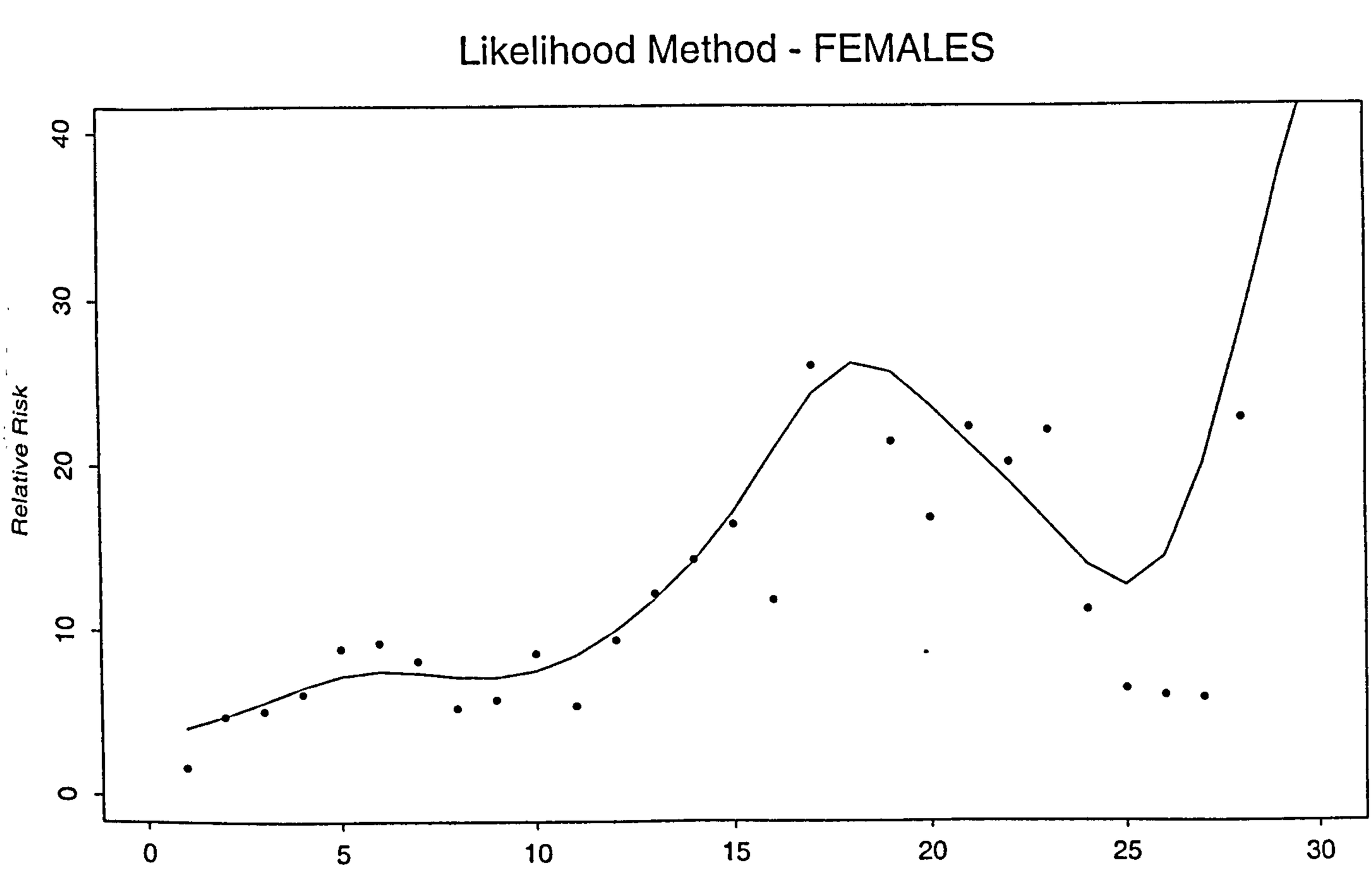


Figure 3.5.8

Section 3.5.4: Summary

Both of these methods have made it possible to identify possible categorisations and have tended to agree on categorisations. They also suggest that the categorisation employed by MacKie et al was reasonably accurate.

The estimates of Relative Risk obtained by the nonparametric approaches and those obtained by the parametric approach were shown to be quite different perhaps bringing into doubt the use of the conditional linear logistic model for this particular data set.

Section 3.6: Isotonic regression

Section 3.6.1: Introduction

The non-parametric analyses presented in sections 3.5.2 and 3.5.3 are both useful in identifying potential categorisations of an interval scaled discrete risk factor in a matched case/control study. Unfortunately they ignore one important constraint implicit in this type of study, namely that the Relative Risk is a monotonically ***increasing*** or ***decreasing*** function of the potential risk factor. Both methods described earlier have allowed the Relative Risk to fluctuate both up and down as the level of the risk factor increases. The effect of these fluctuations was dampened by running a kernel smoother through the original values, but this still did not require the final estimate to be monotonic in nature. This section will use isotonic regression in order to produce monotonic estimates of the Relative Risk which satisfy the above monotonic restriction.

Section 3.6.2: Isotonic regression

Isotonic regression (Barlow et al(1972)) is used to produce sensible estimates of a function which is constrained to be monotically increasing or decreasing. The following definitions briefly describe an ***isotonic function*** and formally outline the constraints which would be present when ***isotonic regression*** is used and also give an outline of the methodology involved in applying this technique.

Definitions

Let X be the finite ordered set $\{x_1, x_2 \ldots x_k\}$. A real valued ***function*** f on x is ***monotonic increasing*** if $x,y \in X$ and $x < y \Rightarrow f(x) \leq f(y)$.

Let g be a given function on X and w a given positive function on X. An isotonic function g^* on X is an ***isotonic regression*** of g with weights w with respect to the simple ordering $x_1 < x_2 < \ldots < x_k$ if it minimises the sum of squares

$$\sum_{x \in X} [g(x) - f(x)]^2 \, w(x) \qquad \text{- (3.16)}$$

over all possible functions f on X.

Isotonic regression therefore provides a method of producing an estimator which minimises the sum of squares function (3.16) under an ***order restriction***. This chapter is concerned with producing estimates of ***Relative Risk*** under the constraint that these estimates are ***monotonic*** across the levels of a ***single risk factor***.

Various algorithms exist for finding the relevant g* to minimise (3.16) and the one which will be used in this chapter is the ***"pool adjacent violators"*** algorithm. This is essentially a very simple ***algorithm*** and is as follows:

Assume one has function values $g(x_1), g(x_2), \ldots, g(x_k)$ at points $x_1, x_2, \ldots, x_k$.

It is necessary to satisfy the constraint $g(x_1) \leq g(x_2) \leq \ldots \leq g(x_k)$.

Initially if $g(x_1) \leq g(x_2) \leq \ldots \leq g(x_k)$ then the initial partition is final partition, and

$g^*(x_i) = g(x_i)$ $i = 1, \ldots, k$.

If not, however, select any of the pairs that violate the ordering i.e. select an i such that

$$g(x_i) > g(x_{i+1})$$

Join the two points x_i and x_{i+1} in a "new block" $\{x_i, x_{i+1}\}$ with associated average value

$$\frac{w(x_i)g(x_i) + w(x_{i+1})g(x_{i+1})}{\left[w(x_i) + w(x_{i+1})\right]}$$

and associated weight $(w(x_i) + w(x_{i+1}))$.

After each step in the algorithm, the new, average, values $g^*(x_i)$ i = 1, ..., k, associated with the blocks are examined to see whether they are in the required order. If so the final partition has been reached and the value of g* at each point of block is the "pooled" value associated with the block. If not, a pair of adjacent violating blocks is selected, and pooled to form a single block, with associated weight the sum of their weights and associated average value the weighted average of their average values, completing another step of the algorithm. The algorithm continues in this manner until the initial constraints are satisfied giving the final solution g*.

Section 3.6.3: Isotonic regression in practice

In the cutaneous malignant melanoma example the number of naevi had 30 distinct levels with associated estimated Relative Risks $\hat{f}(z_1), \hat{f}(z_2), \ldots, \hat{f}(z_{30})$. Initially no constraint was placed on these function values but now it seems logical that $\hat{f}(z_1) \leq \hat{f}(z_2) \leq \ldots \leq \hat{f}(z_{30})$. This places us

within the framework of isotonic regression. Hence isotonic regression will be used to find new estimates $\hat{f}^*(z_i)$, i = 1 , ... , 30. Both non-parametric approaches to estimation of the relative risk will be considered and the weighting function used in both situations will be taken to be

$$w(z_i) = 1 / \hat{V}(\hat{f}(z_i))$$

The sum of squares (3.16) will then be minimised, using the "pool adjacent violators" algorithm to obtain a set of monotonic estimates $\hat{f}^*(z_i)$, i = 1, ... ,30.

(a) Pairwise cells comparison

For the malignant melanoma example the estimates of Relative Risk for the number of naevi should be constrained to be monotonically increasing. Figures 3.6.1 and 3.6.2 display the ***monotonic estimates*** of Relative Risk separately for males and females.

For both data sets a categorisation around about 17 naevi is once again suggested. However this approach seems to perhaps highlight another potential point for males at somewhere around 10 or 11 naevi which was not detected with the use of the kernel smoother alone.

(b) Conditional Likelihood Method

An isotonic regression of the results obtained from the conditional likelihood method was carried out producing Figure 3.6.3 for males and Figure 3.6.4 for females.

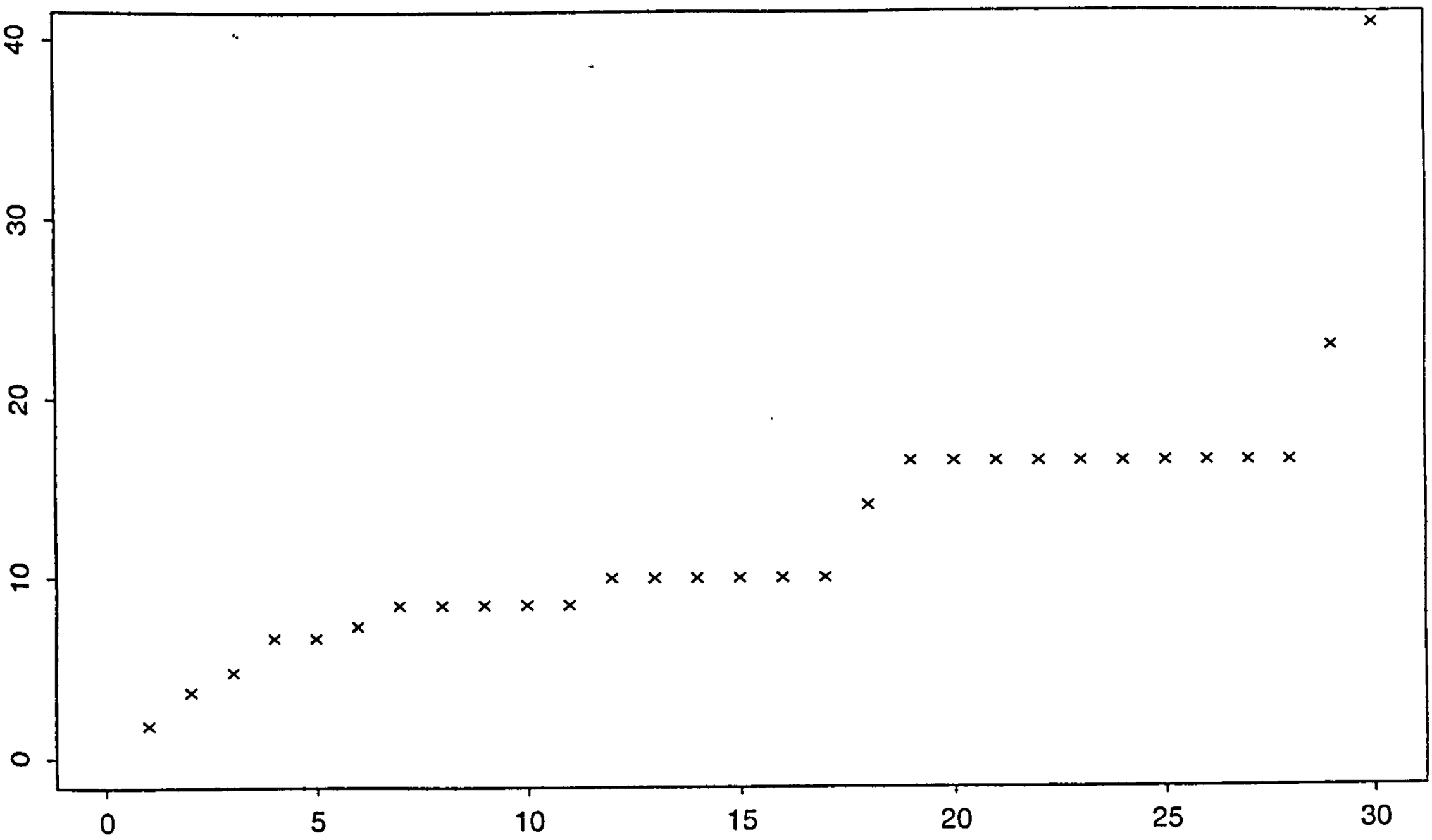


Figure 3.6.1

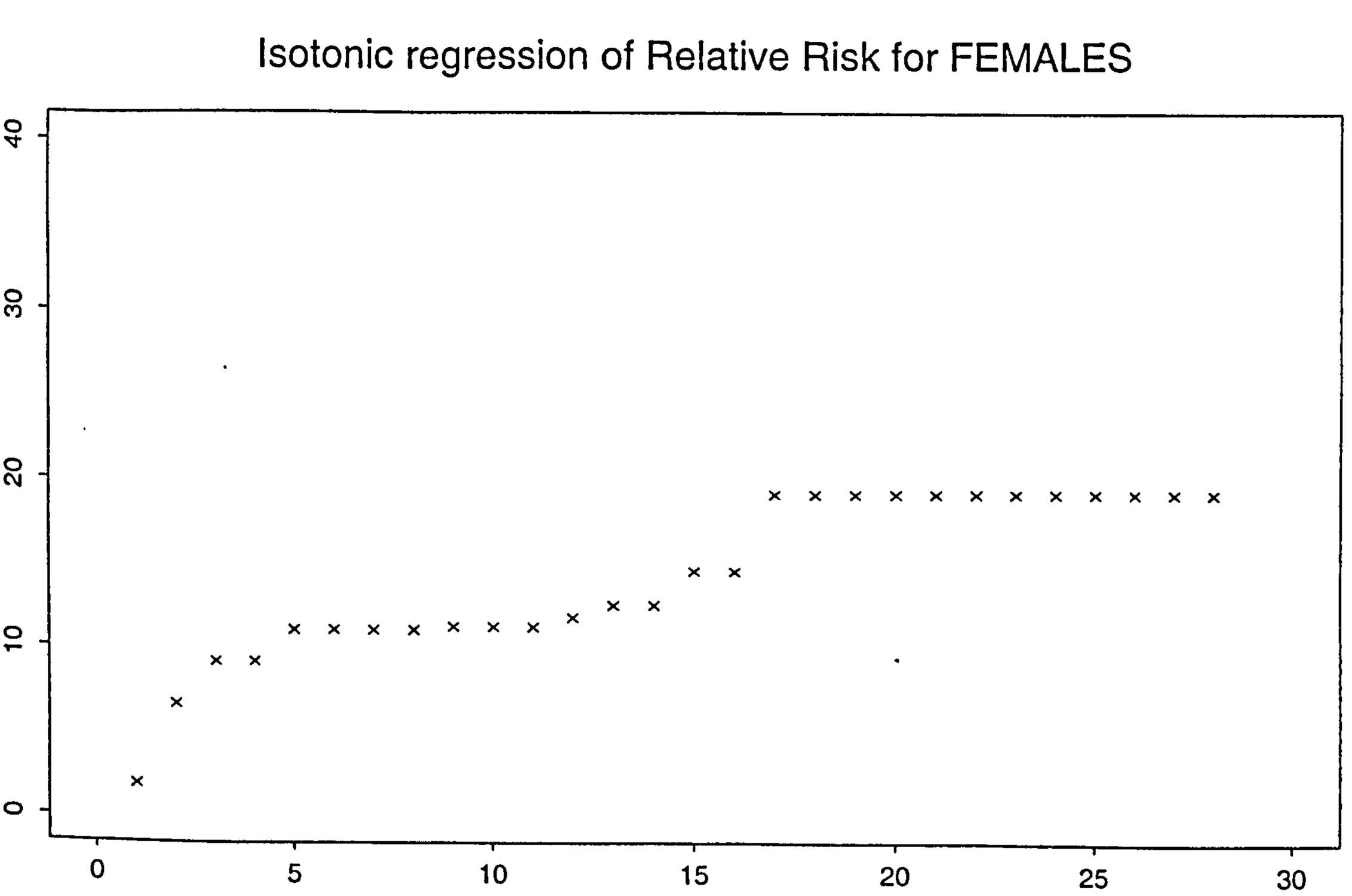


Figure 3.6.2

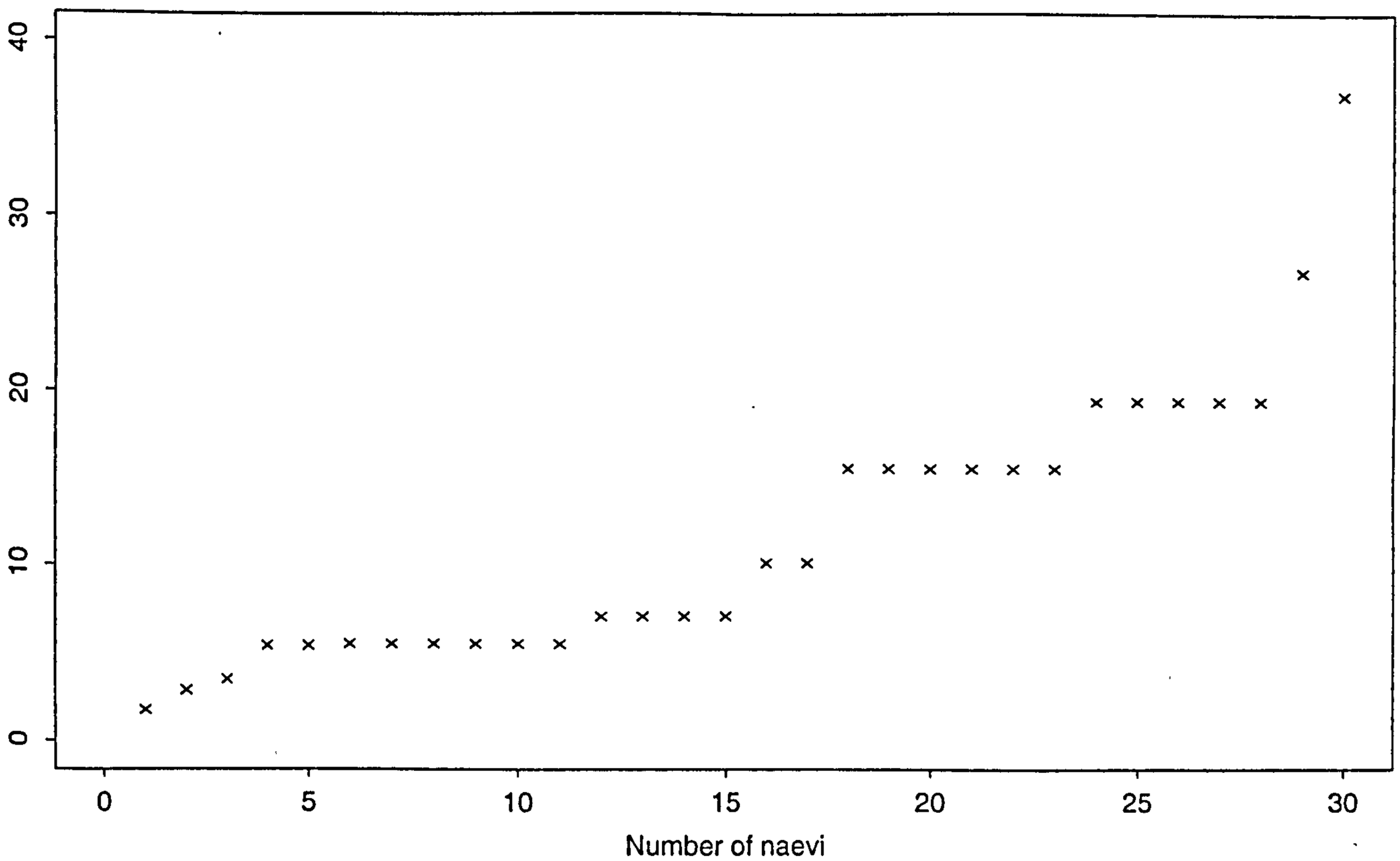


Figure 3.6.3

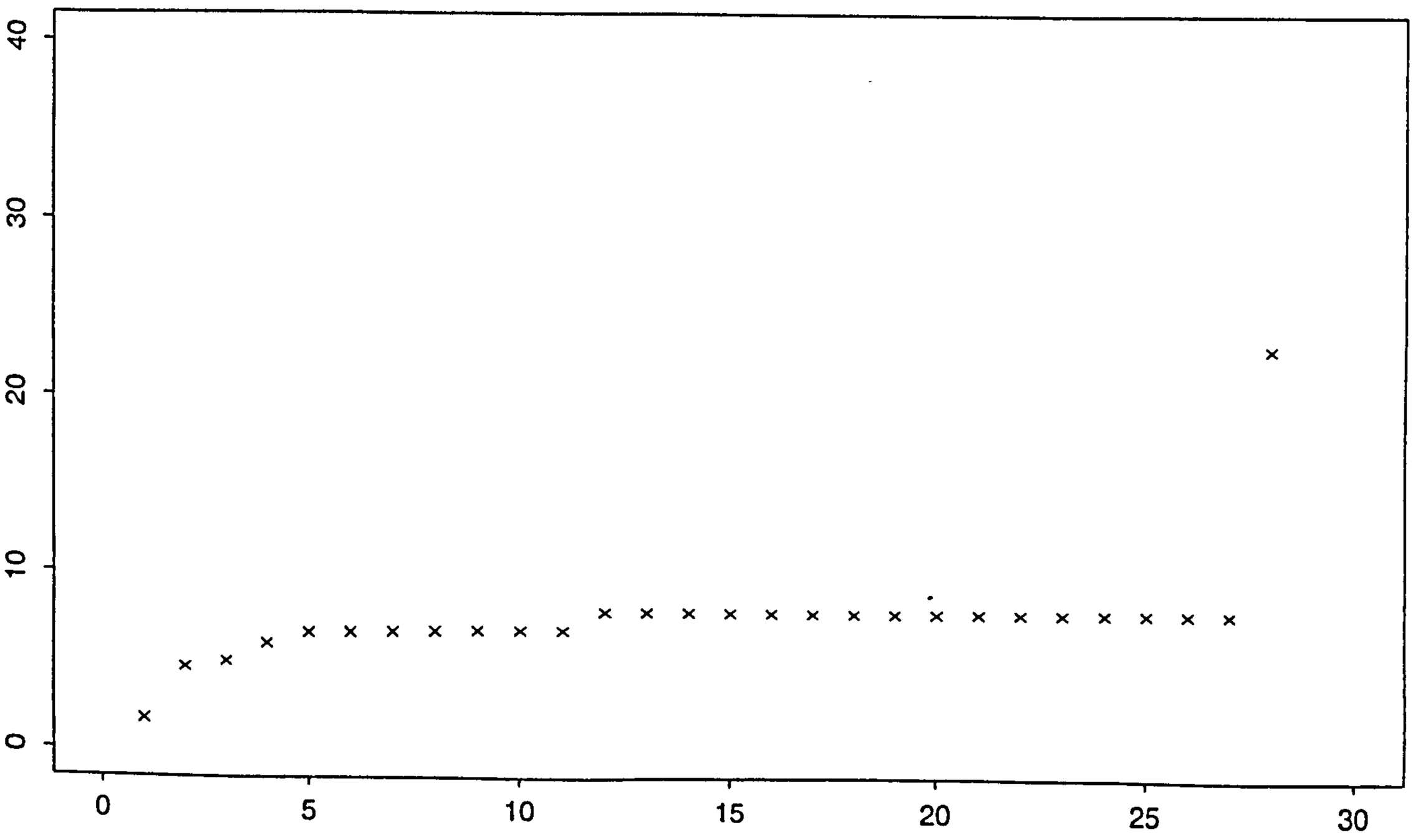


Figure 3.6.4

There is strong evidence for males of a cutpoint at around 18 naevi with perhaps slightly weaker evidence of another potential cutpoint slightly later at about 23 naevi. It is harder to pinpoint a clear cutpoint for females as there appears little evidence of any dramatic changes in the estimates of Relative Risk across number of naevi. The only area where there is a marginal change in the risk is around 12 naevi.

Section 3.6.4: Summary

The use of isotonic regression can remove potentially unreasonable fluctuations in Relative Risk and produces a clearer impression of where categorisation points, if any, exist. However a consequence of using isotonic regression is to produce "flatter" estimates of Relative Risk than were obtained previously. On some occasions this technique produces a slightly different picture than is obtained with kernel smoothing alone. This was noticeable for the pairwise cells method where isotonic regression seems to suggest two potential categorisation points in comparison to the single point highlighted by kernel smoothing. When the two non-parametric methods of estimation are compared there is some evidence that isotonic regression produces different conclusions. This is particularly noticeable for females where the pairwise cells method (Figure 3.6.2) suggested a clear jump at around 15 naevi whereas a far "flatter" estimate of Relative Risk is produced by the conditional likelihood approach (Figure 3.6.4).

Section 3.7: Extension to a continuous risk factor

Section 3.7.1: Introduction

In the last 3 sections, methods for producing non-parametric estimates of Relative Risk from an interval scaled discrete risk factor have been introduced. In this section a possible extension of these methods to a continuous risk factor will be examined. The non-parametric techniques discussed earlier for an interval scaled discrete risk factor cannot be directly applied to a continuous risk factor. The primary reason for this is that with a continuous variable there are a potentially infinite ***number of levels*** of the risk factor to be considered. There does not appear to be a straightforward adaptation of the techniques discussed in section 3.4 to deal with this. The following section examines a possible method to adapt the techniques introduced previously to cater for the case of a continuous risk factor.

Section 3.7.2: Extension to a continuous risk factor

In order to produce estimates of Relative Risk for a continuous risk factor one possible technique is simply to initially categorise the continuous risk factor to create a "pseudo" interval scaled discrete risk factor. Then the technique of ***pairwise cells comparison*** (section 3.4.2) or indeed the ***conditional likelihood method*** (section 3.4.3) could be applied to the categorised data. The technique used here to initially categorise a continuous risk factor creates "bins" into which observations are placed dependent upon their value. The table below illustrates the idea

Value of risk factor	[m,m+k)	[m+k,m+2k)	[m+2k,m+3k)	.....	[m+nk,∞)
"bin" / category number	0	1	2	.....	n

Usually m will be the minimum value of the continuous risk factor under consideration. The values of k and n will be specific to each particular continuous risk factor and will depend upon the range of the risk factor and also the size of the sample.

Since one of the aims of the work in this thesis is to look for possible cutpoints for risk factors this technique of arbitrarily categorising a continuous risk factor may seem somewhat self defeating. However if it is borne in mind that this technique of creating "bins" for the observations only provides ***rough initial categorisations*** then applying the pairwise cells method or conditional likelihood method will hopefully improve on these rough categorisations and produce a clearer picture of where any potential points for such categorisations lie.

From here the method proceeds in a similar fashion to the techniques discussed for an interval scaled discrete variable to produce estimates of Relative Risk for each category compared to the baseline category. It must however be remembered that each estimate of Relative Risk is comparing two ***ranges*** of values as opposed to two ***specific*** values. In this situation the ***baseline*** category will always be values of the continuous risk factor between the minimum possible value, m, and the value m+k. The next section will look at the application of this technique to an example from the medical field.

Section 3.7.3: Sun exposure and cancer risk

McHenry et al (1994) carried out a large-scale study of malignant melanoma in the West of Scotland. They collected information on a large number of cases and their age/sex matched controls. One of the potential risk factors considered was the average number of hours of exposure to United Kingdom sun per year. Many studies have shown that for British subjects the risk of contracting malignant melanoma increases with

number of hours of exposure to United Kingdom sun per year. Many studies have shown that for British subjects the risk of contracting malignant melanoma increases with exposure to foreign sun. Less work has been done to examine whether these same types of subjects are more at risk if they have been exposed to larger amounts of United Kingdom sun. Here the relationship between contracting malignant melanoma and exposure to United Kingdom sun will be examined.

Parametric analysis

Figure 3.7.1 displays boxplots of the average number of hours of exposure to sun per year for both the cases and controls. Since there are large areas of overlap between these two boxplots there appears little evidence to suggest that those subjects who contract malignant melanoma have experienced higher levels of United Kingdom sun exposure than those subjects who do not contract the disease. Since this is a matched case/control study, a bivariate plot of the case/control pair values may help to reveal a clearer pattern. The bivariate plot is displayed in Figure 3.7.2 with the line of equality superimposed. Again there seems little evidence to distinguish between the cases and controls.

However, before any conclusions can be drawn, a formal analysis should be carried out. The results of fitting a univariate conditional linear logistic model were

$$\hat{\beta} = 0.288$$

$$\hat{se}\left(\hat{\beta}\right) = 0.764$$

Therefore

$$\frac{\hat{\beta}}{\hat{se}\left(\hat{\beta}\right)} = 0.377$$

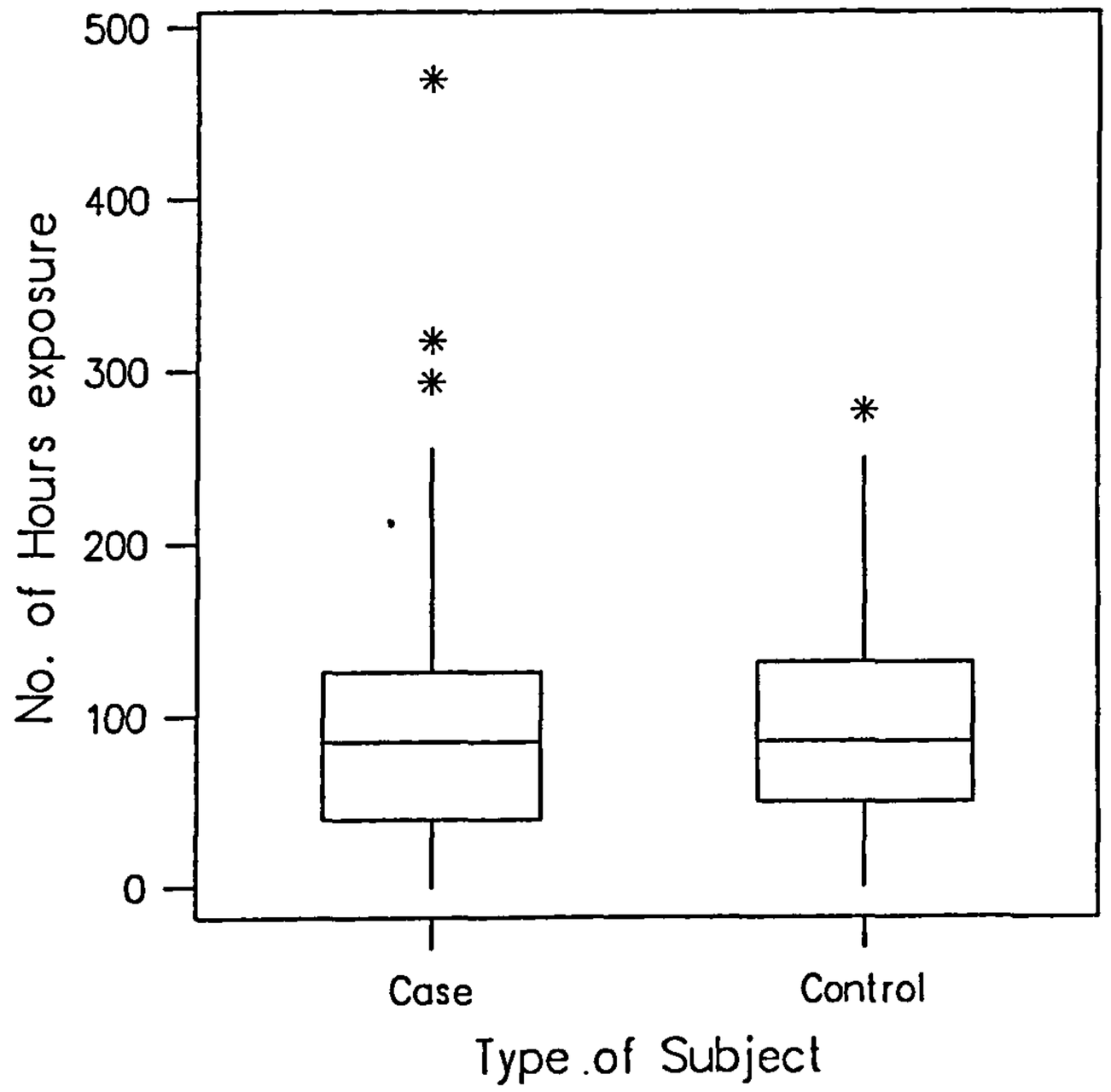


Figure 3.7.1

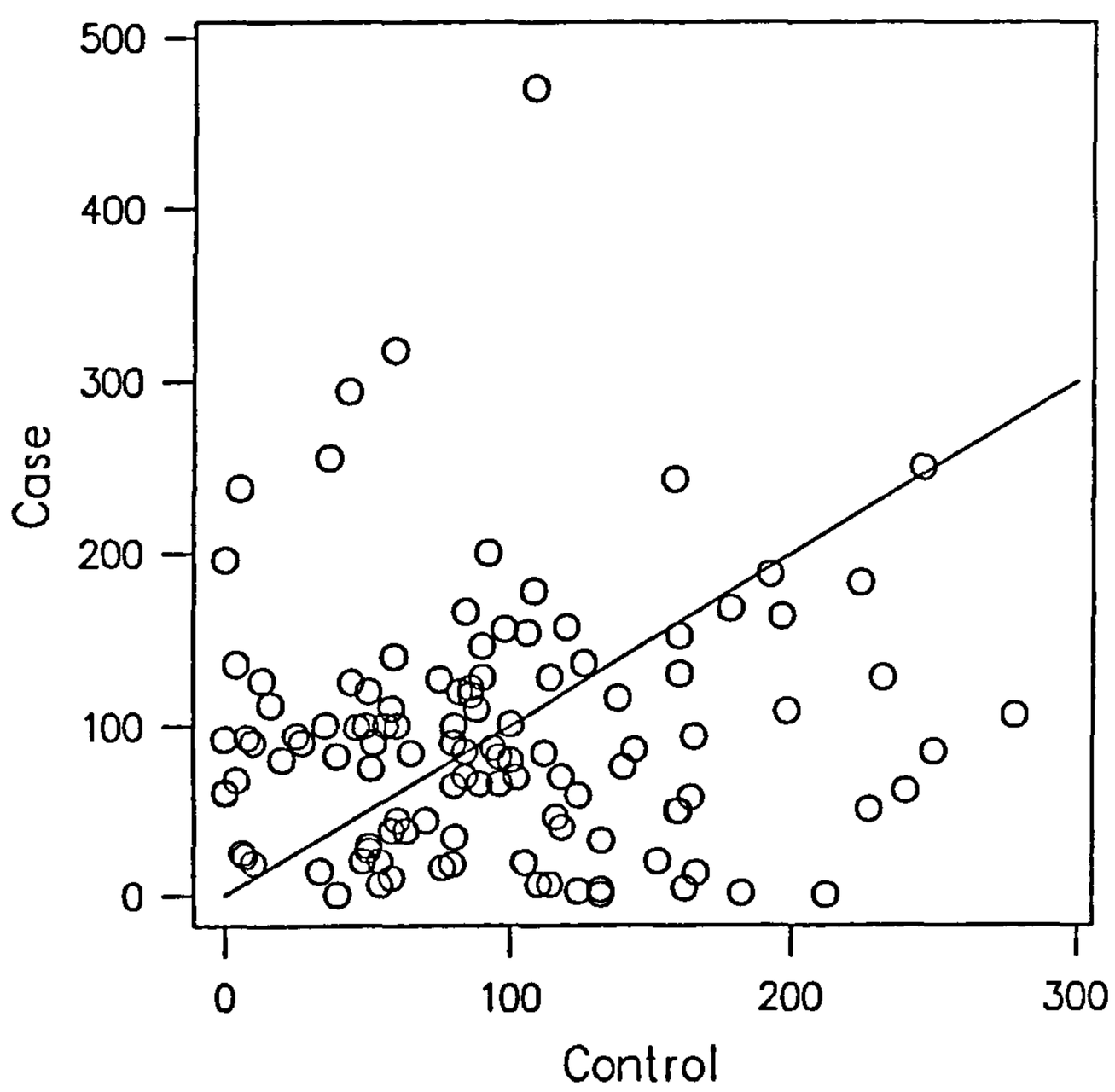


Figure 3.7.2

Since this is quite clearly a non-significant ratio confirmation, this confirms the subjective impression that the average number of hours of exposure to United Kingdom sun for these subjects has little effect on their chances of contracting malignant melanoma.

Non-parametric analysis

One way of examining whether the method discussed in section 3.7.2 appears to produce sensible estimates of Relative Risk is to apply the technique to a data set and compare the results to those obtained from a corresponding parametric analysis. If the conclusions produced by the non-parametric analysis are ***overall*** not markedly different from the parametric analysis then it seems reasonable to assume that the non-parametric technique will ***in general*** produce plausible estimates of Relative Risk.

The first step in using the non-parametric method is to choose an appropriate "bin" width or category size. For this particular data set the average number of hours of exposure to the sun range from 0 hours to approximately 250 hours. This quite large range of values in conjunction with a relatively small number of 114 case/control pairs suggest that the category sizes considered should be reasonably large. Therefore two specific category sizes of width 5 hours and 10 hours respectively will be studied.

Pairwise cells comparison:

Figure 3.7.3 displays plots of estimates of Relative Risk versus average number of hours of United Kingdom sun exposure per year for a category size of 5 hours exposure

for the method of pairwise cells comparison. Superimposed on these plots are confidence bands for the Relative Risk. The individual frames of the figure represent differing levels of data smoothing as discussed in section 3.4.4. Figure 3.7.4 displays a similar plot for a category size of 10 hours exposure. Both of these figures suggest very similar conclusions. These results appear to be in very good agreement with those obtained from the parametric analysis as they show little if any effect of the number of hours of exposure on contracting malignant melanoma. This can be seen since, without exception, each frame clearly shows the Relative Risk fluctuating reasonably randomly around the value 1 (i.e. No effect). It would also appear that a "bin" width of 5 hours is more relevant for this data set than 10 hours as any possible features of the data appear to be rapidly smoothed out for the larger "bin" size.

Likelihood Method:

Figures 3.7.5 and 3.7.6 show the corresponding plots to Figures 3.7.3 and 3.7.4 for the likelihood method. These figures permit the same conclusion that exposure to United Kingdom sun has little effect on chances of contracting malignant melanoma. They do however produce confidence bands for the Relative Risk which are more precise than those obtained by the pairwise cells comparison methods perhaps suggesting that in general the likelihood method may produce slightly more reliable results then the pairwise cells comparison method.

PAGE
NUMBER
CUT OFF
IN
ORIGINAL

Pairwise cells method: Category size = 5 Hours

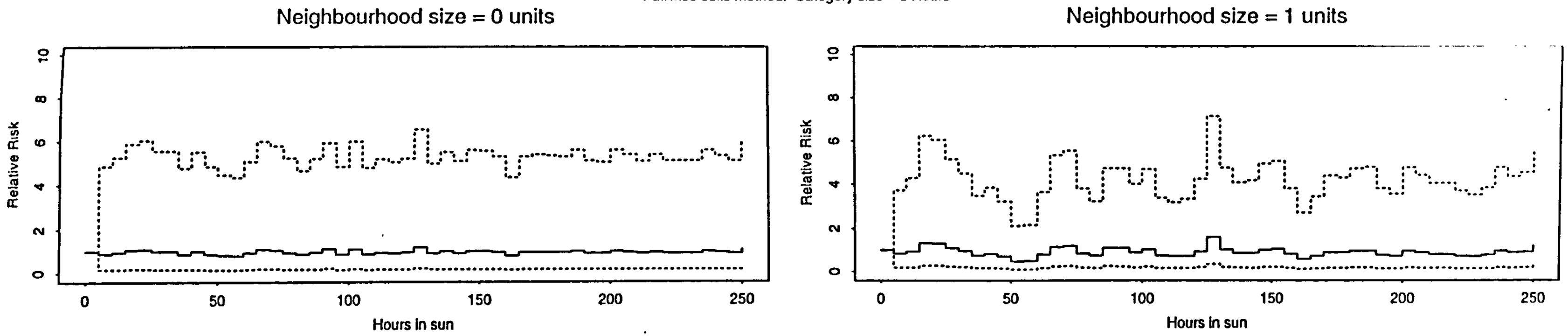
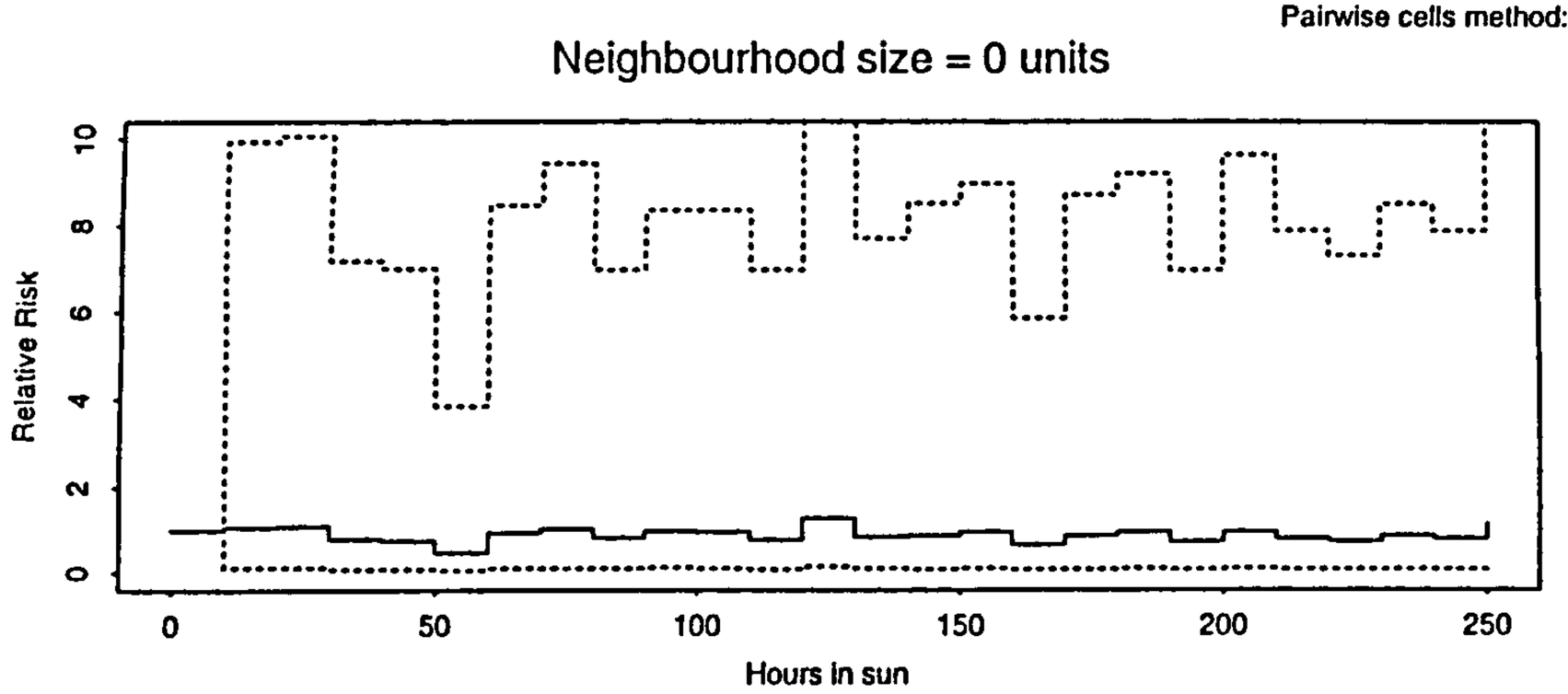
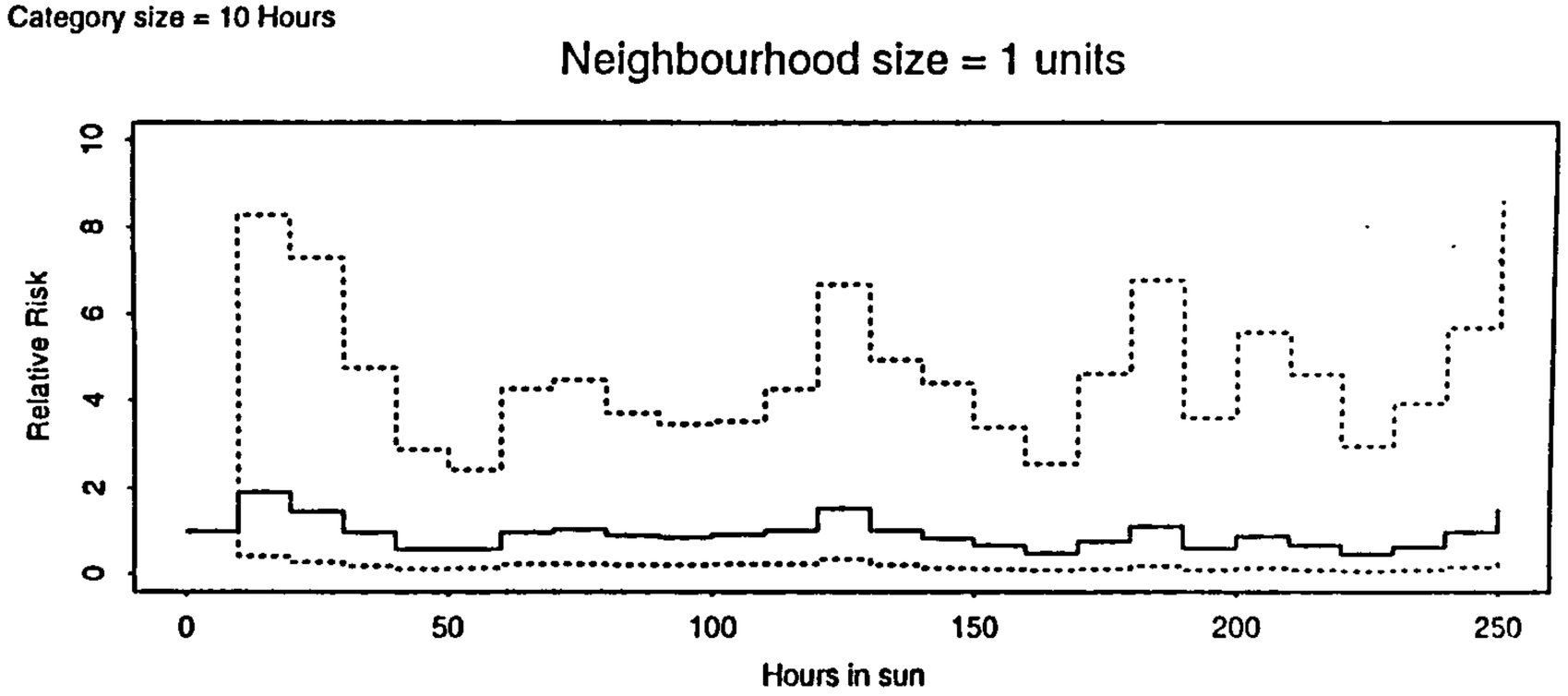
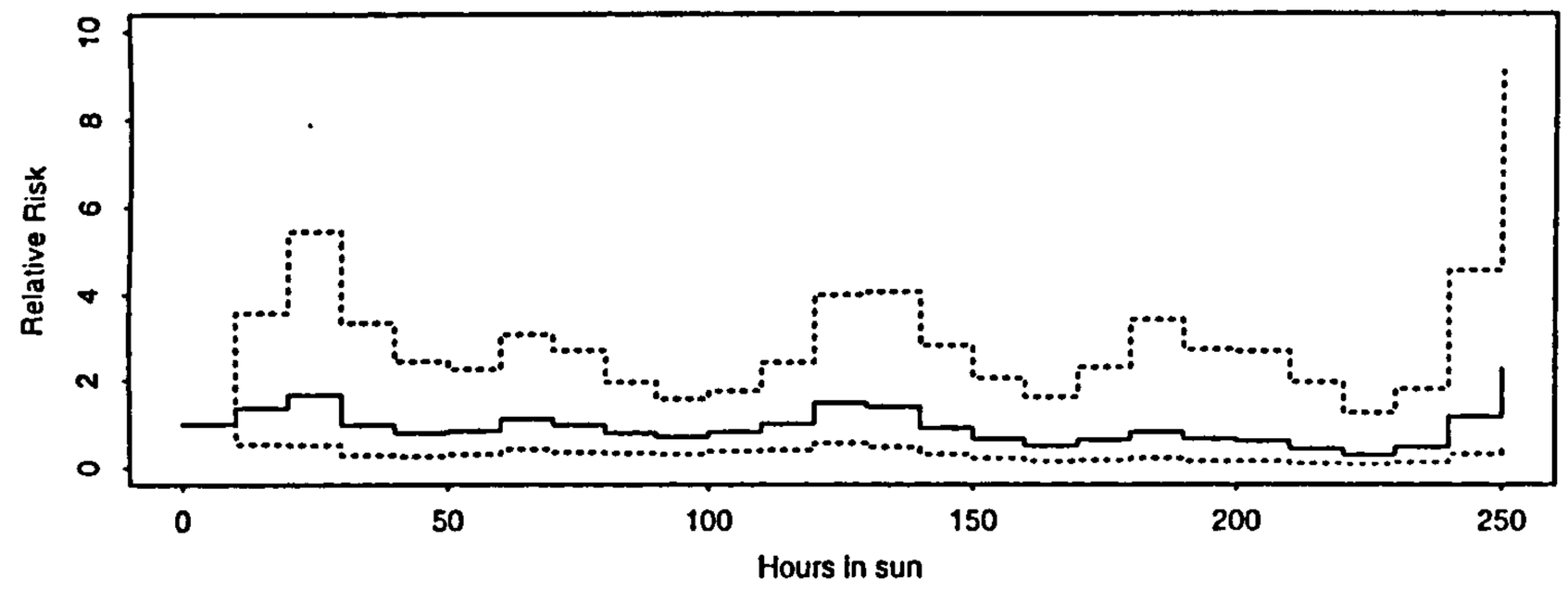


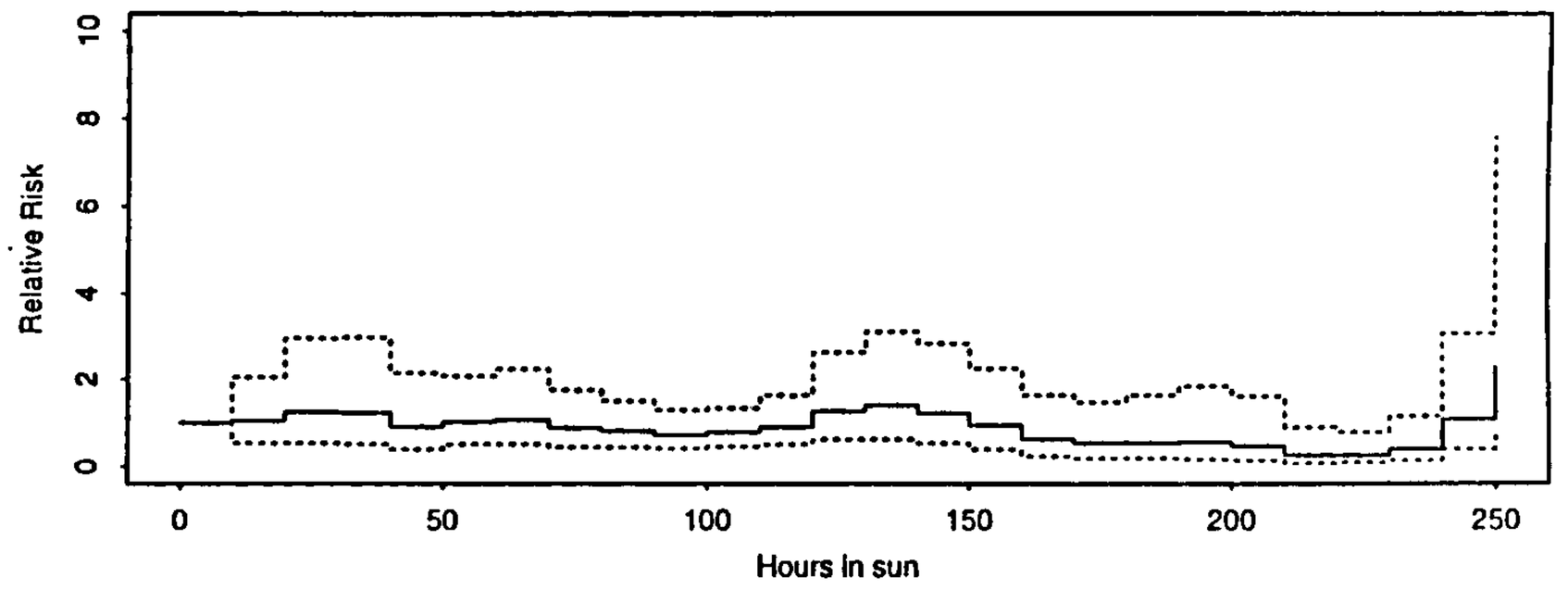
Figure 3.7.3

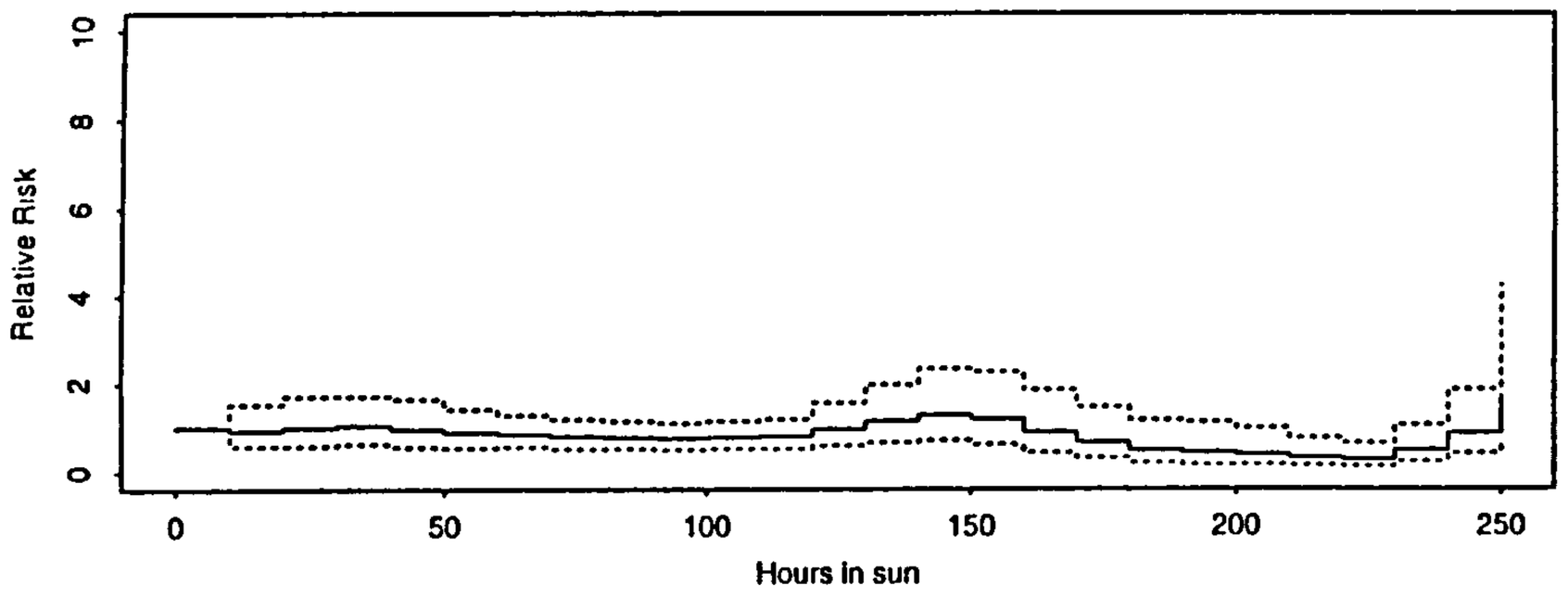
Pairwise cells method: Category size = 10 Hours











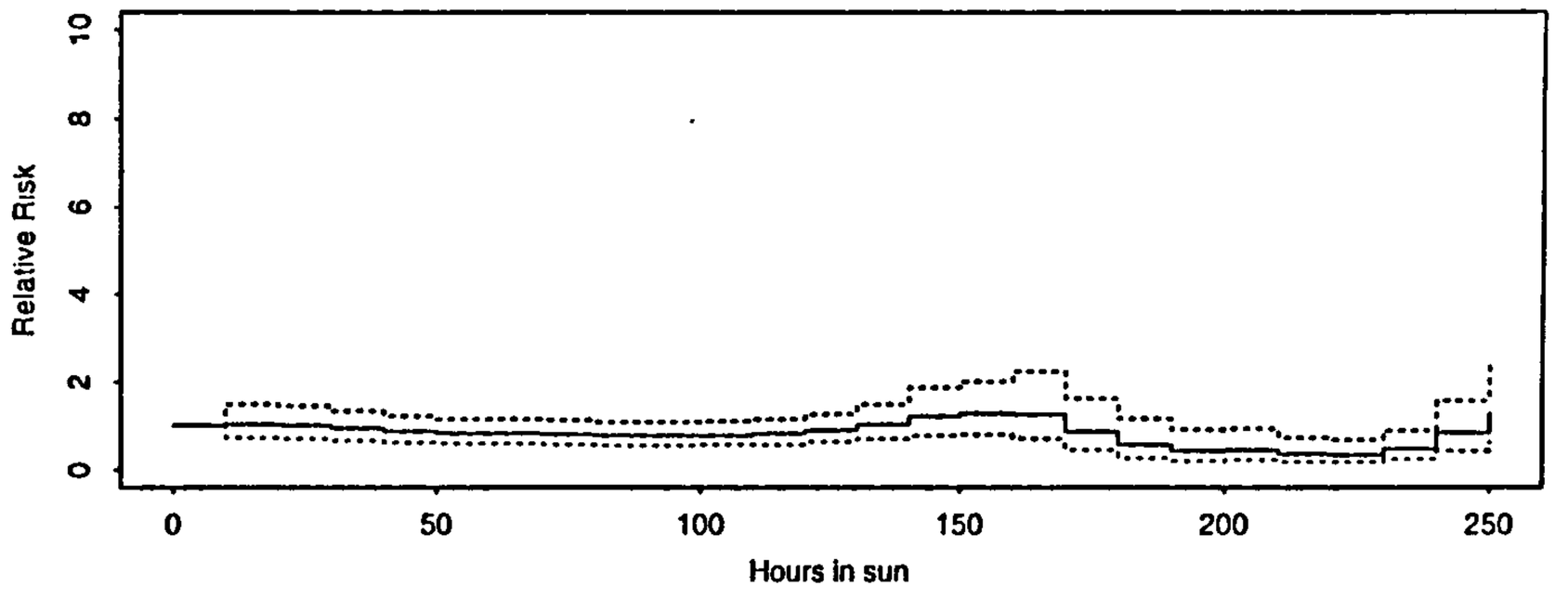


Figure 3.7.4

Likelihood method: Category size = 5 Hours

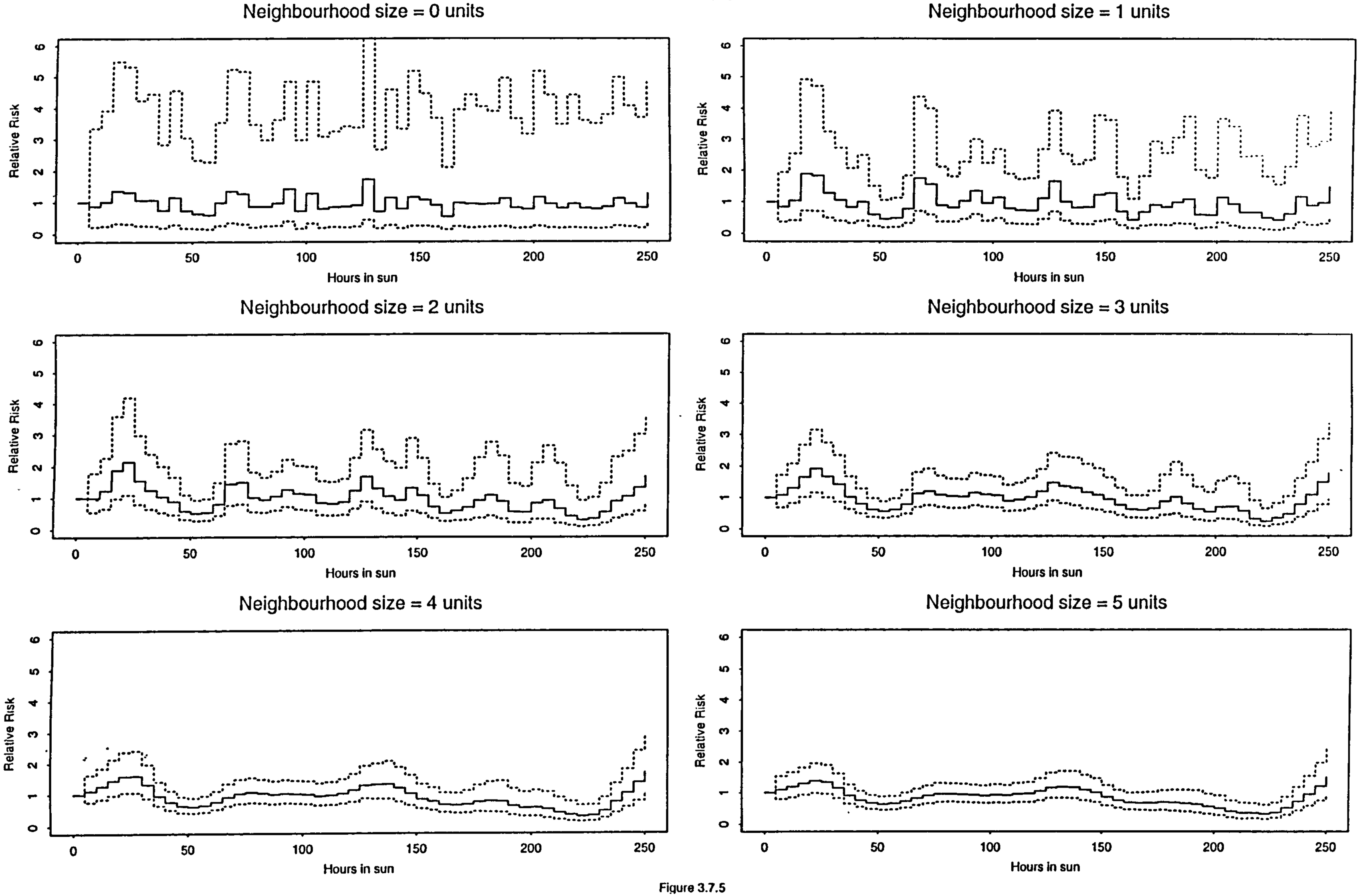
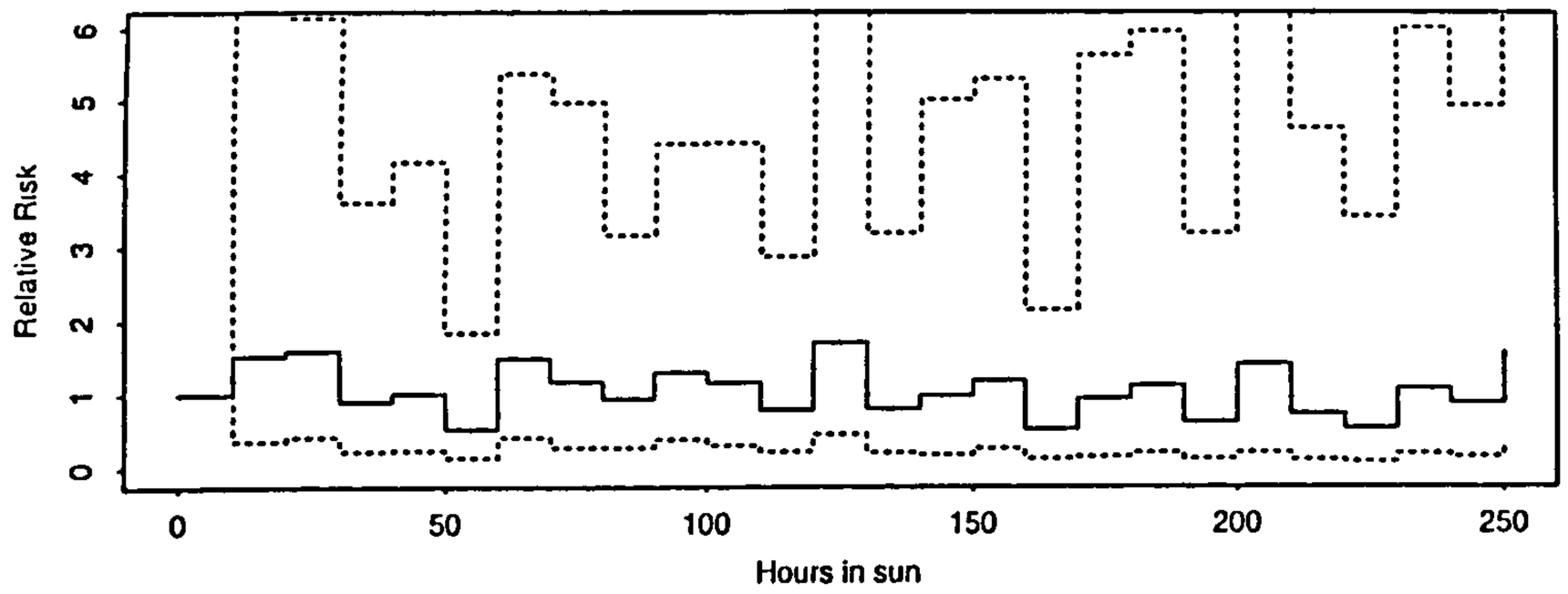
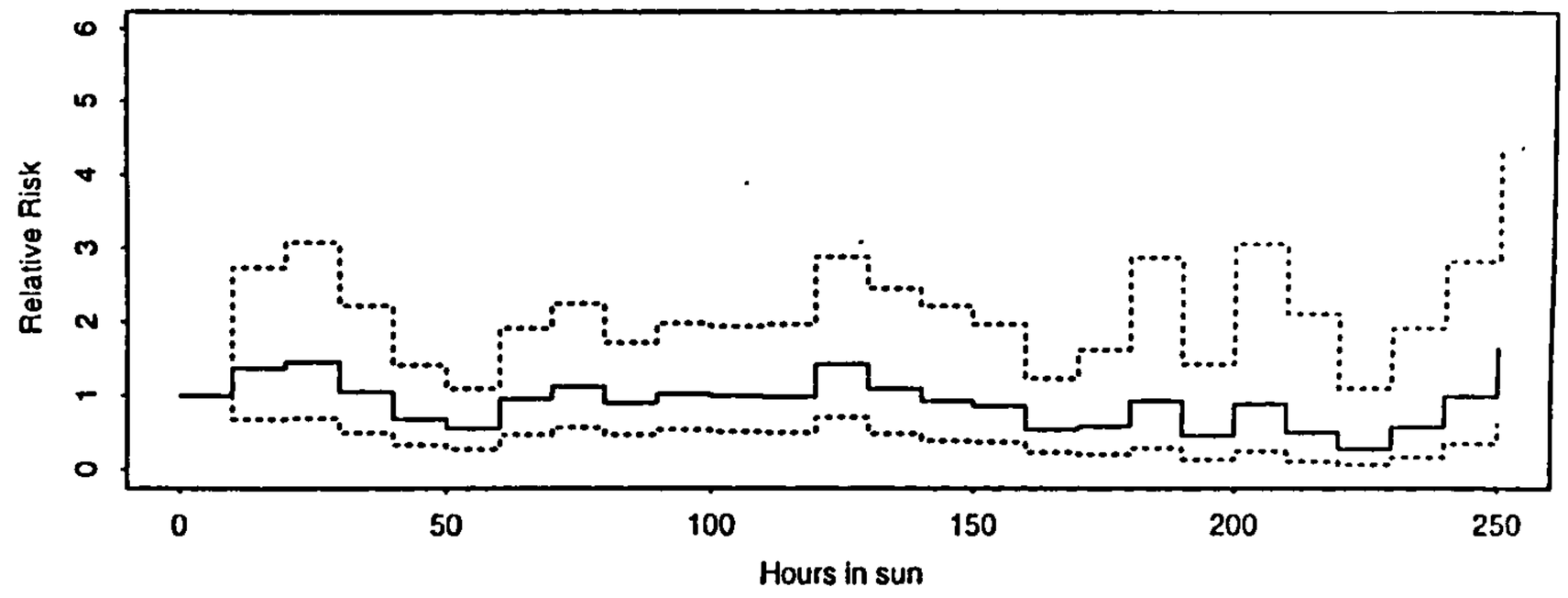
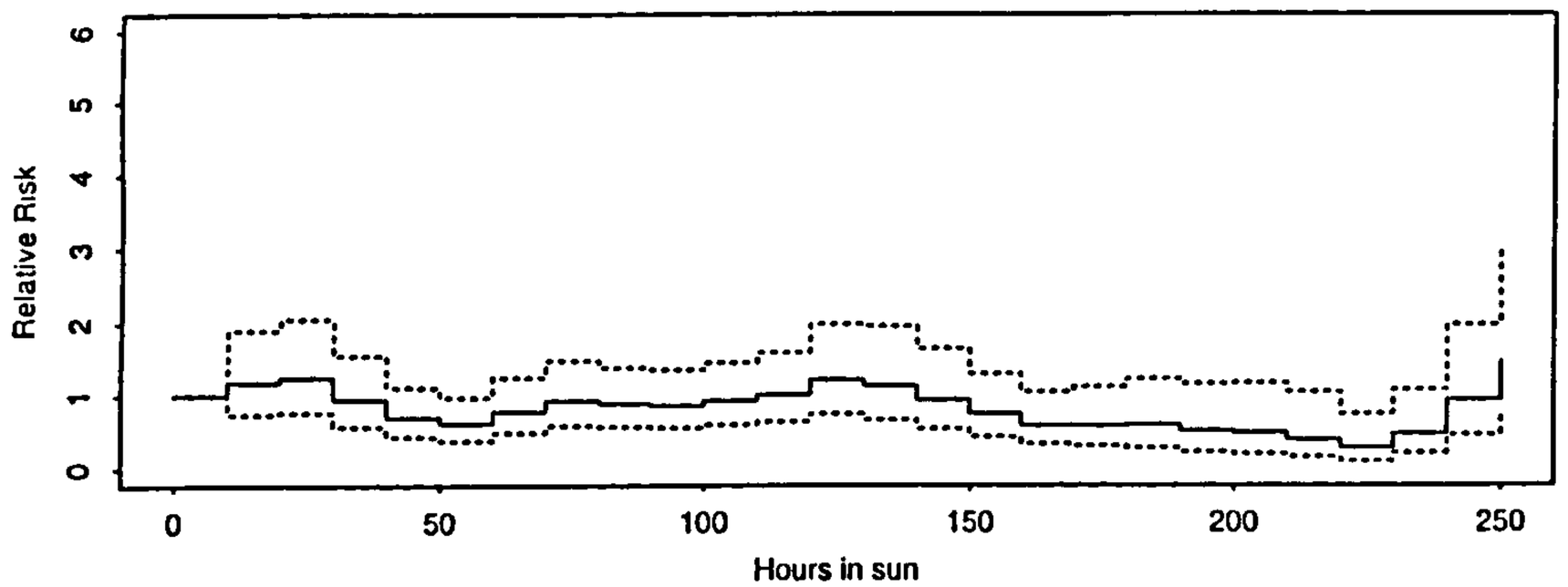


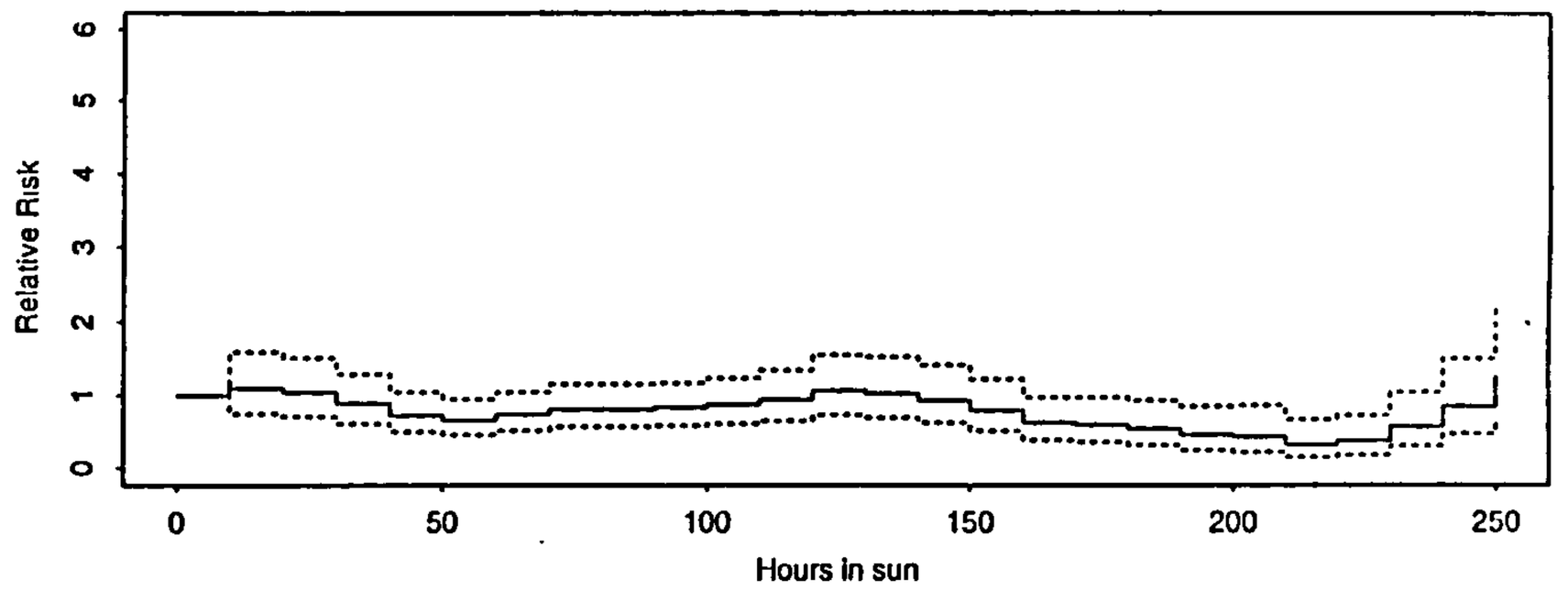
Figure 3.7.5

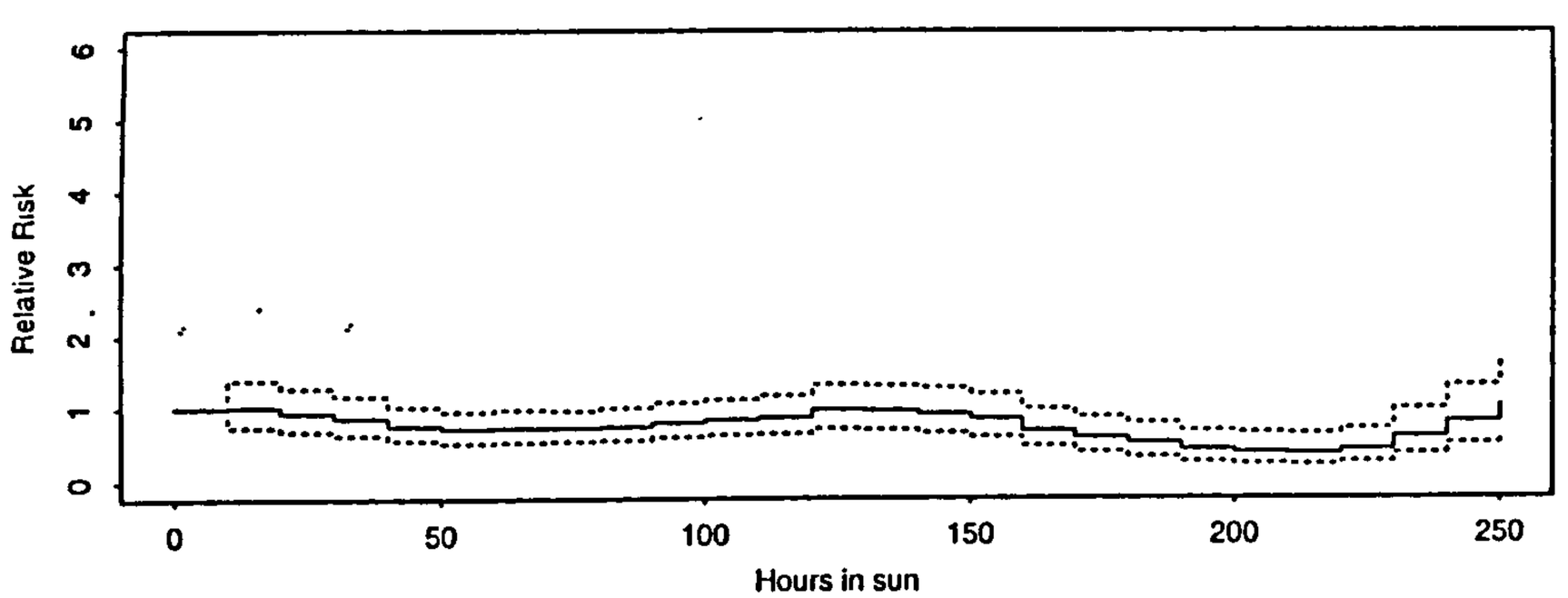
Likelihood method: Category size = 10 Hours











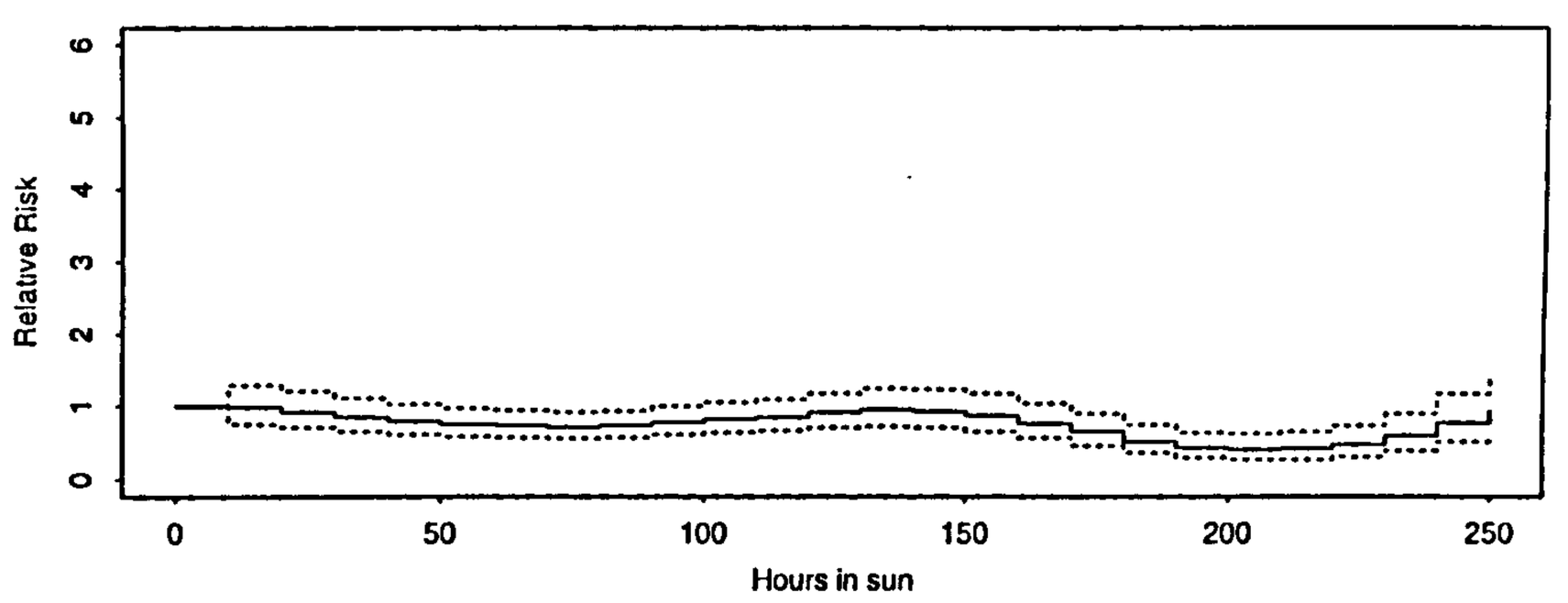


Figure 3.7.6

Section 3.7.4: Summary

This idea of producing "bins" for a continuous risk factor has allowed non-parametric estimates of Relative Risk to be produced for a continuous risk factor. For the specific example under consideration in this section the estimates of Relative Risk obtained by using the technique of "binning" the continuous risk factor produced estimates which in general agreed with those obtained from a parametric model. In both cases there was little evidence to suggest that exposure to United Kingdom sun has any effect on the chances of contracting malignant melanoma. In conjunction with this the non-parametric analyses did not highlight any potential cut-points in terms of changes in Relative Risk. One interesting aim would be to attempt to create some ***automatic*** procedure for choosing the "bin" width which took account of both the sample size and the range of possible values for the risk factor.

Section 3.8: Simulation Study

Section 3.8.1: Introduction and examples

Section 3.4 introduced two non-parametric methods for producing estimates of Relative Risk for an interval scaled discrete risk factor in a matched case/control study; the "pairwise cells" method and the "conditional likelihood" method. It is essential to examine if either of these proposed methods of estimation produces "better" estimates of Relative Risk. This can be investigated by simulating data from a known, underlying, situation and investigating which method performs better in terms of precision, bias and coverage.

Numerous studies (Neuhauser and Becher (1997), Aickin et al (1994), Commenges and Moreau (1991)) have carried out simulations from an ***unmatched*** case/control study but the literature on simulating from a ***matched*** case/control study is fairly limited. The crucial difference between the unmatched and matched scenario is that in the matched scenario the distribution of the risk factor is dependent upon the distribution of the matching variable (Schlesselman (1982)), a dependency which is not present in the unmatched situation. In matched case/control studies this dependency must be incorporated into in any simulations which are carried out. It has also been shown Cox(1970) and Egijuo & McHugh(1977) that in a matched case/control study the pattern of ***Relative Risk*** is assumed constant across the levels of the matching variable. Hence the Relative Risk will be "independent" of the matching variable. In this section data will be simulated from a ***matched*** case/control study based on specifying

(i) An underlying distribution for the risk factor among ***the non-diseased (control) population*** which will be dependent on the value of the matching variable.

(ii) A known Relative Risk function which will be independent of the matching variable.

The specification of (i) and (ii) make it possible to obtain the distribution of the risk factor among the ***diseased (case) population*** as follows:

Let z_1 represent the risk factor under consideration and z_2 represent the factor on which the cases and controls have been matched.

Under the assumption of a multivariate ***conditional linear logistic*** model with only additive main effects the odds of being diseased (i.e. a case) given specified levels of z_1 and z_2 are then

$$\frac{p(\text{diseased}/z_1, z_2)}{p(\text{not diseased}/z_1, z_2)} = \exp(\alpha + \beta z_1 + \gamma z_2) \qquad \text{- (3.17)}$$

Based on the above model the Relative Risk for any level, z_1, of the risk factor compared to an arbitrary baseline, 0, is given by

$$\text{Relative Risk}(z_1:0) = \frac{p(\text{diseased}/z_1, z_2)/p(\text{not diseased}/z_1, z_2)}{p(\text{diseased}/0, z_2)/p(\text{not diseased}/0, z_2)} = \exp(\beta z_1) \qquad \text{- (3.18)}$$

A simple application of Bayes' Theorem produces

$$\text{Relative Risk}(z_1{:}0) = \frac{p(z_1/\text{diseased}, z_2)/p(z_1/\text{not diseased}, z_2)}{p(0/\text{diseased}, z_2)/p(0/\text{not diseased}, z_2)} \quad \text{- (3.19)}$$

Using (3.19) in conjunction with (3.17) and (3.18) gives

$$p(z_1/\text{diseased}, z_2) = p(z_1/\text{not diseased}, z_2) * \exp(\beta z_1) * \frac{p(0/\text{diseased}, z_2)}{p(0/\text{not diseased}, z_2)} \quad \text{- (3.20)}$$

Therefore, if the distribution of the risk factor among the controls given a level of the matching variable (i.e. $p(z_1/\text{not diseased}, z_2)$) is specified along with a known Relative Risk function, then (3.20) shows it is possible to obtain the distribution of the risk factor among the cases given a level of the matching variable (i.e. $p(z_1/\text{diseased}, z_2)$)

One problem with simulating data from a matched case/control study is that the assumption of a known, underlying, Relative Risk function is based on comparing the risk for any level of the risk factor with the risk at the baseline. This suggests it is essential that in any simulation enough data is generated at the baseline to allow adequate estimation of the Relative Risk. Therefore in the simulations presented in this section a proportion of controls will be generated at the baseline level to guarantee that enough information will be available at the baseline.

In the example concerning number of naevi as a potential risk factor for malignant melanoma discussed in Sections 3.3 and 3.5, approximately 35% of the controls were at the baseline (i.e. 35% of the controls had zero naevi). In an attempt to mirror a real

situation as closely as possible the scenarios in this section will also be generated on the basis of a population with 35% of the controls at the baseline.

In the simulations discussed here two possible underlying ***interval scaled discrete*** distributions for the risk factor among the controls will be considered; a ***Poisson*** distribution and a ***discrete uniform*** distribution. In conjunction with these, two known Relative Risk functions will be incorporated into the simulations; a ***linear*** Relative Risk function and a Relative Risk function with a ***single, large step*** in the Relative Risk at a pre-assigned value of the interval scaled discrete risk factor.

As each simulation is from a ***matched*** case/control study an underlying distribution must be assumed for the matching variable. In the scenarios presented here the matching variable will be generated from a Uniform distribution.

Table (3.8.1) details the ***sample sizes*** and ***levels of smoothing*** which will be used in each scenario and table (3.8.2) presents a summary of the scenarios which will be considered in the simulations.

Sample Sizes	25,50,75,100,150,200,250,300
Levels of smoothing	No smoothing, first order neighbourhood, second order neighbourhood
Number of simulations	1000 of each sample size with each level of smoothing

Table (3.8.1)

	Distribution of the matching variable, $p(z_2)$	Distribution of the risk factor among the controls, $p(z_1$ / not diseased, $z_2)$	Relative Risk function
Scenario 1	Un(1,12)	Po(z_2)	exp($\beta * z_1$)
Scenario 2	Un(1,12)	Po(z_2)	Step function
Scenario 3	Un(1,20)	Un(0,z_2)	exp($\beta * z_1$)
Scenario 4	Un(1,20)	Un(0,z_2)	Step function

Table (3.8.2)

Notes: (i) The distribution of the matching variable has been chosen to produce a distribution of the risk factor among the controls which will have, in each scenario, approximately 20 levels.

(ii) In column 4 of Table (3.8.2) the underlying linear and step Relative Risk functions have been chosen to produce values of the Relative Risk which are in the same "ball-park" as one another to allow direct comparisons to be made across the four scenarios. In order to achieve this the value selected for β was 0.15 and the step function was chosen as

$$\text{Relative Risk}(z_1:0) = \begin{cases} 1 & 1 \le z_1 \le 10 \\ 10 & 10 \le z_1 \le 20 \end{cases}$$

The graphs of Relative Risk and log(Relative Risk) for these two choices are shown in Figure 3.8.1.

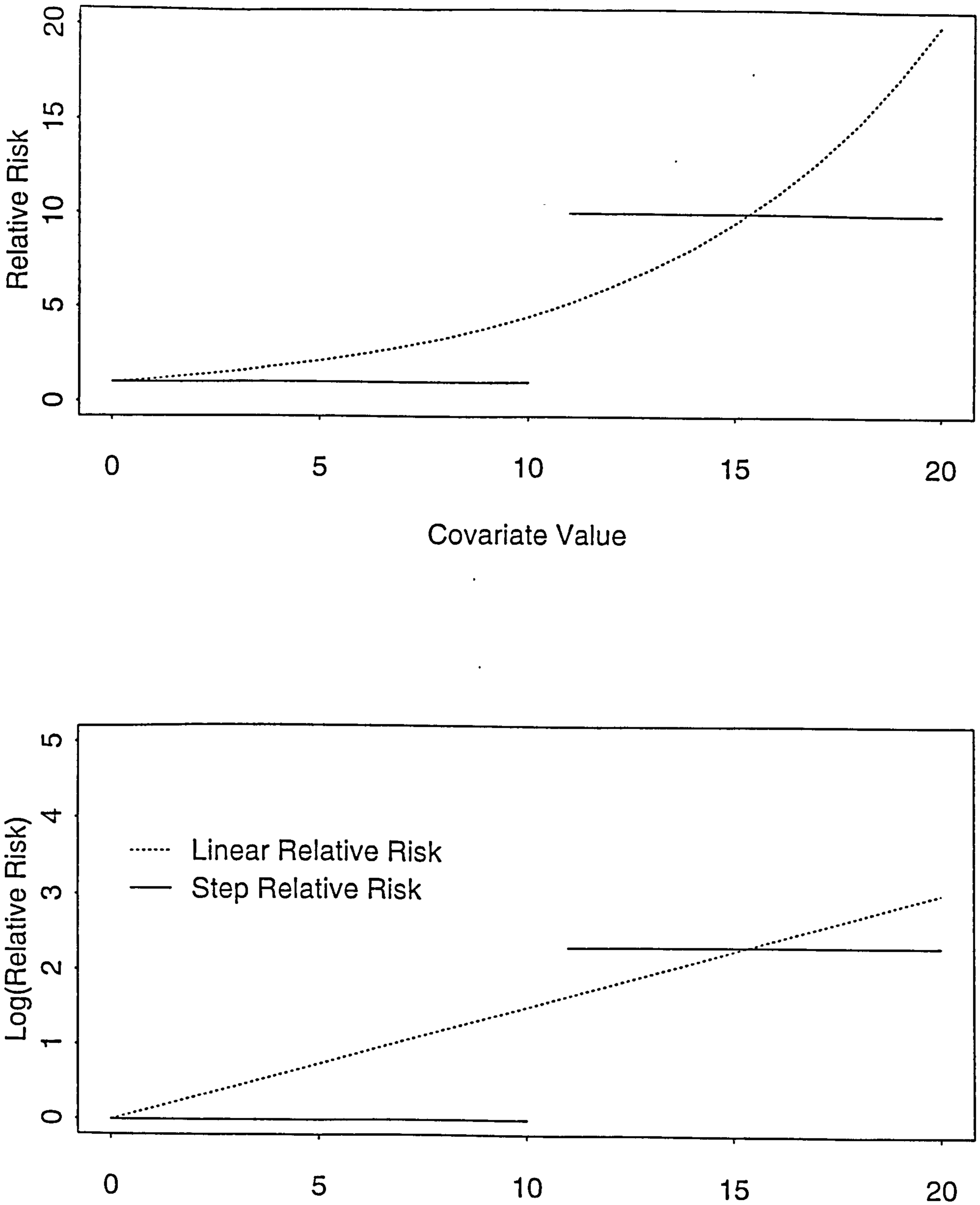

Relative Risk
0
5
10
15
20
0
5
10
15
20
Covariate Value
Log(Relative Risk)
0
1
2
3
4
5
Linear Relative Risk
Step Relative Risk
0
5
10
15
20
Covariate value

Figure 3.8.1

To assess how well the two non-parametric estimators perform under each underlying, known, situation the comparison criteria will be ***mean square error*** (a measure of precision), ***bias*** and ***coverage*** defined as follows:

(i) Mean square error $= \dfrac{\sum_{i=1}^{\text{number of levels}} \left(\hat{\log}(RR_i) - \log(RR_i)\right)^2}{\text{number of levels}}$

where RR = Relative Risk

This measures the ***precision*** of the estimates, with smaller values of the measure indicating a greater precision in the resultant estimates. The average mean square error (MSE) and empirical standard deviation (ESD) of the mean square error across all 1000 simulations for each scenario will be used as an objective measure of precision.

(ii) Bias $= \dfrac{\sum_{i=1}^{\text{number of levels}} \left(\hat{\log}(RR_i) - \log(RR_i)\right)}{\text{number of levels}}$

where RR = Relative Risk

This measures the ***bias*** present in the estimates, with smaller values indicating the presence of less bias in the resultant estimates. The average bias and empirical standard deviation of the bias across all 1000 simulations for each scenario will be used as an objective measure of bias.

(iii) Coverage = Proportion of intervals containing the true value of the Relative Risk. The intervals will be constructed based on a ***nominal coverage of 95%.*** The coverage will be evaluated ***separately*** at ***each*** level of the risk factor.

Scenario 1: **Poisson distribution for $p(z_1 / \text{not diseased}, z_2)$, linear Relative Risk function.**

Figures 3.8.2 - 3.8.7 show the results for this set of simulations for both the proposed non-parametric methods of estimating Relative Risk. Figure 3.8.2 displays plots of the average mean square error and empirical standard deviation of the mean square error across all simulations against sample size for both methods of estimation. Each frame in the figure refers to the simulation results for a different level of smoothing. Frames 1 to 3 of Figure 3.8.2 suggest that, regardless of sample size and level of smoothing, the conditional likelihood method will produce slightly more precise estimates than the method based on pairwise cells whilst frames 4 to 6 indicate that there is, in general, marginally more variability in the average precision based on the conditional likelihood method. Under this scenario the values for the log of the true Relative Risk range, on average, from 0 to approximately 3 (see Figure 3.8.1). The values obtained for the average mean square error in frame 1 of Figure 3.8.2 suggest that, for sample sizes of 100 pairs or more, both methods of estimation perform reasonably well in the absence of smoothing. It is only with smaller sample sizes (25 to 75 pairs) that the methods appear to have some difficulty in producing precise estimates of Relative Risk. This is perhaps to be

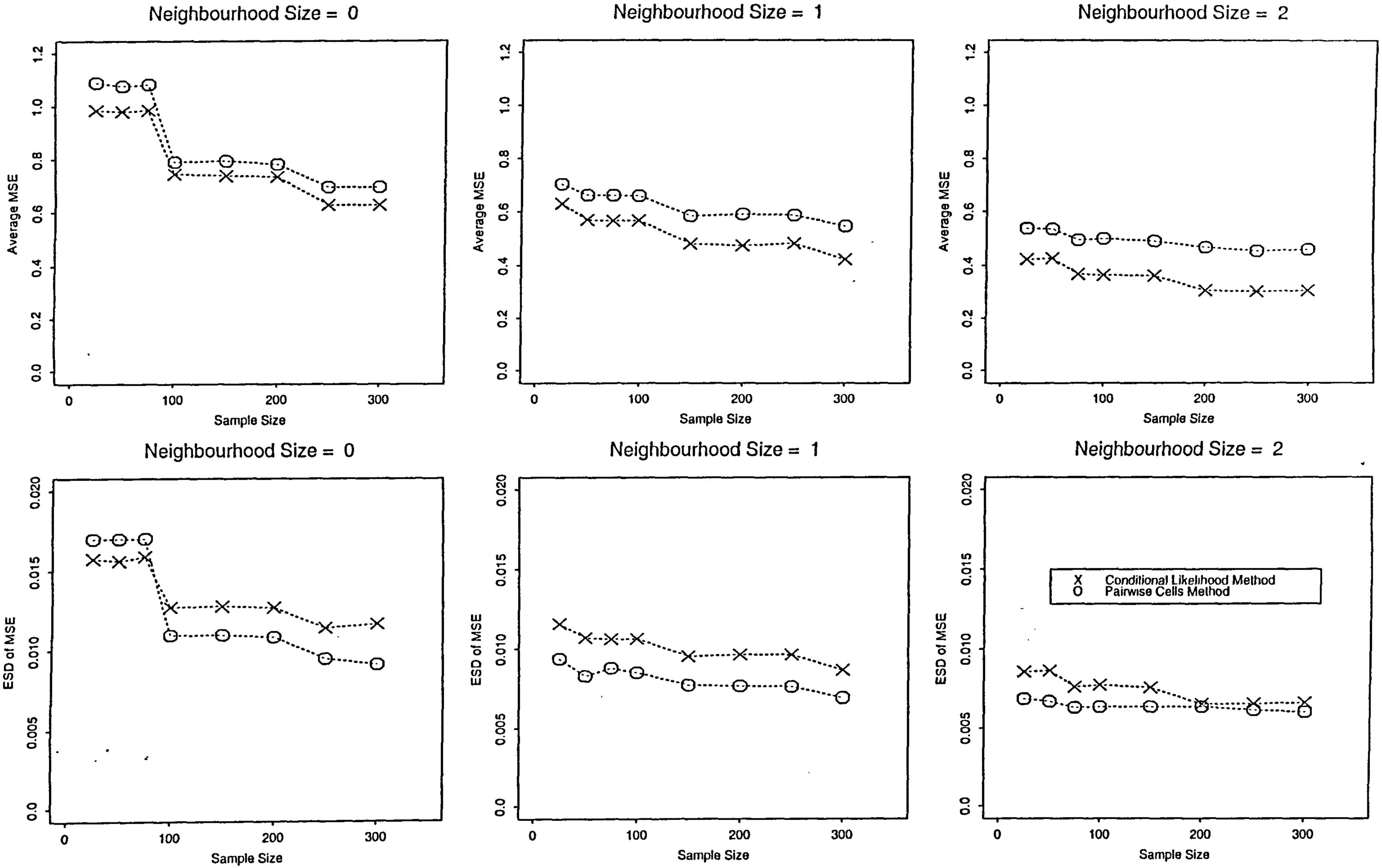

Neighbourhood Size = 0
Neighbourhood Size = 1
Neighbourhood Size = 2
Average MSE
ESD of MSE
Sample Size
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.005
0.010
0.015
0.020
0
100
200
300
X Conditional Likelihood Method
O Pairwise Cells Method

Figure 3.8.2

expected as the methods are attempting to produce non-parametric estimates over a 20 by 20 grid, on average. With a total of 400 cells available it is clearly impractical to expect good estimation with smaller sample sizes. In these circumstances it is clear that some degree of data smoothing will be required before sensible estimates can be produced. This is confirmed by frames 2 and 3 of Figure 3.8.2 which clearly demonstrate that once smoothing is introduced both methods produce precise estimates of the Relative Risk even for small sample sizes.

Figure 3.8.3 displays plots of the average bias and empirical standard error of the bias across all simulations against sample size for both methods of estimation. Frames 1 to 3 suggest that the method based on pairwise cells will produce less biased estimates. The sole exception is with small sample sizes and no smoothing when both methods appear almost identical in terms of bias. Frames 4 to 6 of Figure 3.8.3 reveal that there is slightly more variability in the average bias based on the pairwise cells method. Given the range of true values of the log Relative Risk in this scenario it is clear that in the absence of smoothing both methods appear to substantially ***underestimate*** the Relative Risk particularly for smaller sample sizes. However, even with large sample sizes (i.e. at least 250 pairs of observations) there is still evidence of a significant presence of bias. The introduction of smoothing has the effect of reducing the degree of this underestimation, particularly for small sample sizes. However, even the introduction of smoothing never entirely removes the bias.

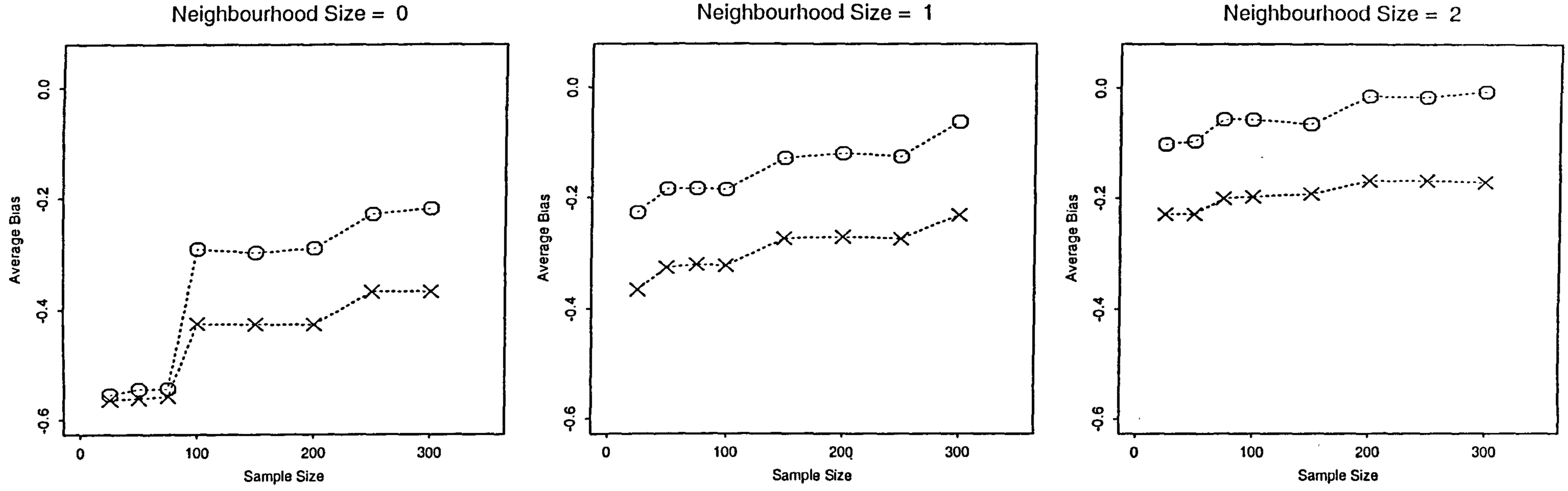

Neighbourhood Size = 0
Neighbourhood Size = 1
Neighbourhood Size = 2
Average Bias
0.0
-0.2
-0.4
-0.6
0
100
200
300
Sample Size
Neighbourhood Size = 0
Neighbourhood Size = 1
Neighbourhood Size = 2
ESD of Bias
0.020
0.015
0.010
0.005
0.0
X Conditional Likelihood Method
O Pairwise Cells Method

Figure 3.8.3

For the conditional likelihood method Figures 3.8.4 and 3.8.5 display plots of the coverage and average width of each nominal 95% interval against each level of the risk factor separately for the different levels of smoothing. Figures 3.8.6 and 3.8.7 display the corresponding plots for the pairwise cells method. In each figure the individual frames represent the results for a different sample size. Figures 3.8.4 and 3.8.6 reveal that both methods appear to exhibit similar patterns in terms of coverage. With each method the coverage ***reduces*** as the level of the risk factor increases and as the level of smoothing increases. The coverage appears ***unrealistically high*** when ***no smoothing*** is present (more than 99%) especially for smaller sample sizes. This is particularly evident for the pairwise cells method. The explanation for this can be seen from Figures 3.8.5 and 3.8.7 where, in general, with no smoothing, the widths of the confidence intervals are larger than are obtained when smoothing is introduced resulting in intervals which will, necessarily, have very high, ***unrealistic***, levels of coverage. Regardless of the level of the risk factor the confidence intervals produced by the pairwise cells method are invariably wider than those produced by the conditional likelihood method. The introduction of smoothing, particularly a first order neighbourhood, produces far more acceptable levels of coverage. In general, with reasonable sample sizes , say 75 - 200 observations, and a first or second order neighbourhood of smoothing, both methods of estimation produce plausible/realistic levels of coverage of between about 85 and 95%. With the exception of some of the higher values of the risk factor, Figures 3.8.5 and 3.8.7 illustrate that, in general, the width of the confidence intervals will decrease as the level of smoothing is increased. This in turn leads to more realistic levels of coverage being attained (i.e. closer to the nominal value of 95%). A final interesting point to observe is that, with both methods, the coverage is higher and the width of the confidence intervals narrower for smaller values of the risk factor. This is to be expected, as, with a Relative Risk of a

Conditional Likelihood Method - Linear Relative Risk

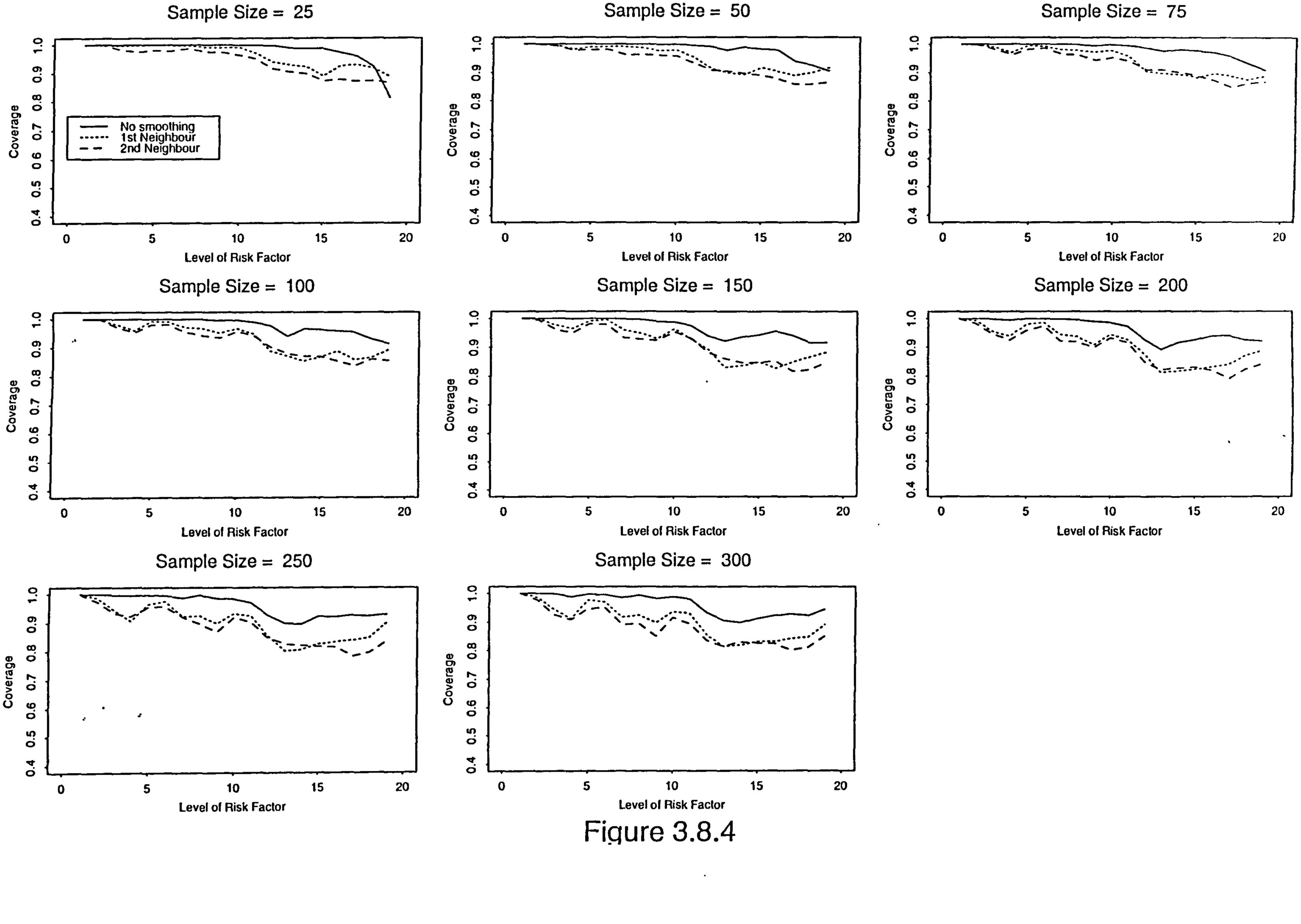
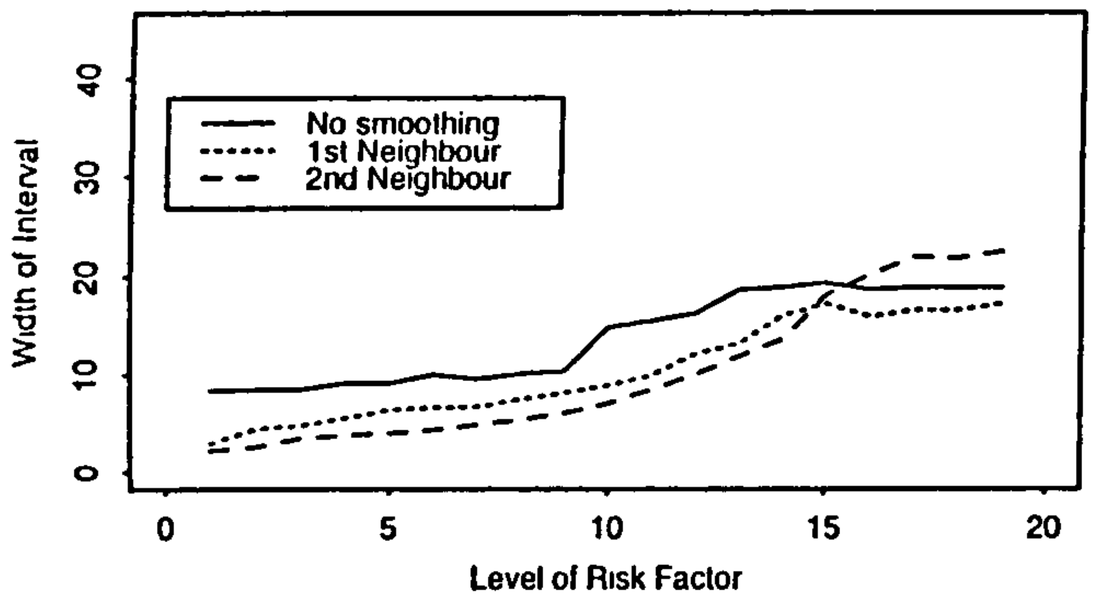


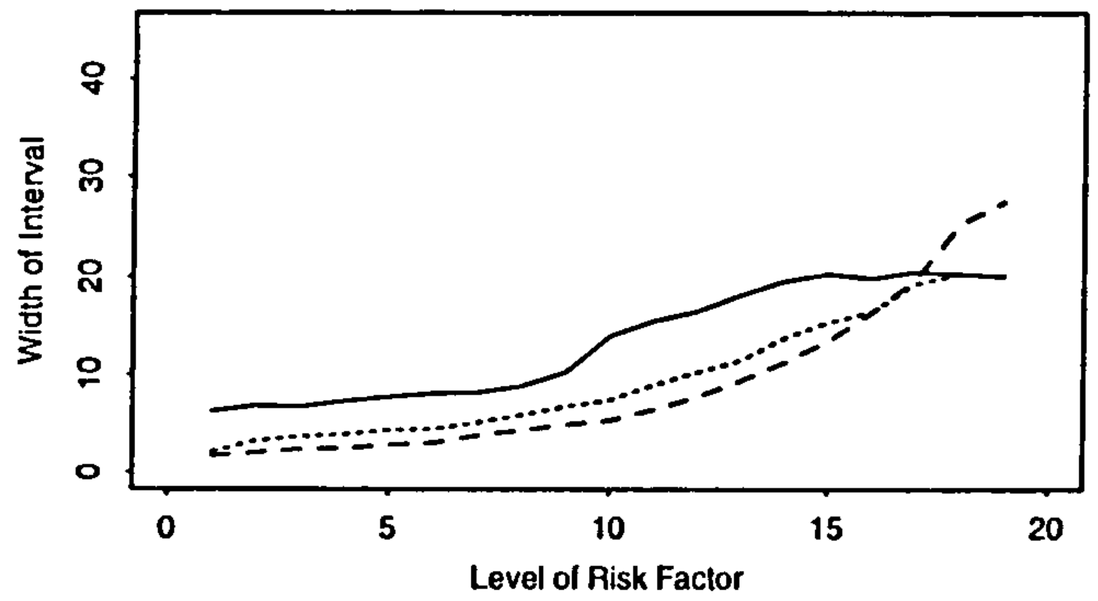
Figure 3.8.4

Conditional Likelihood Method - Linear Relative Risk

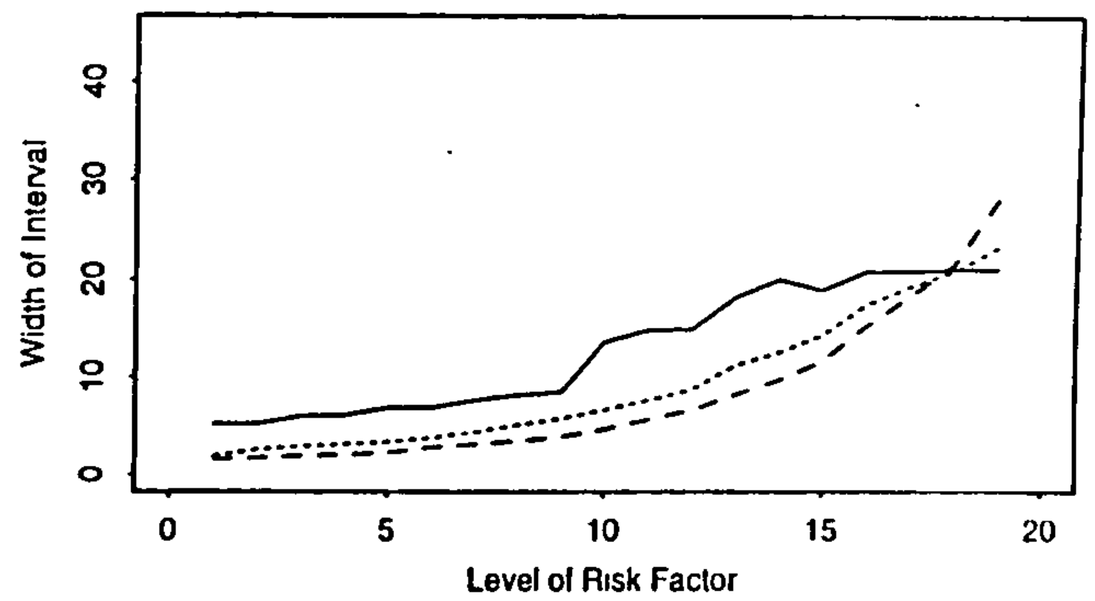
Sample Size = 25



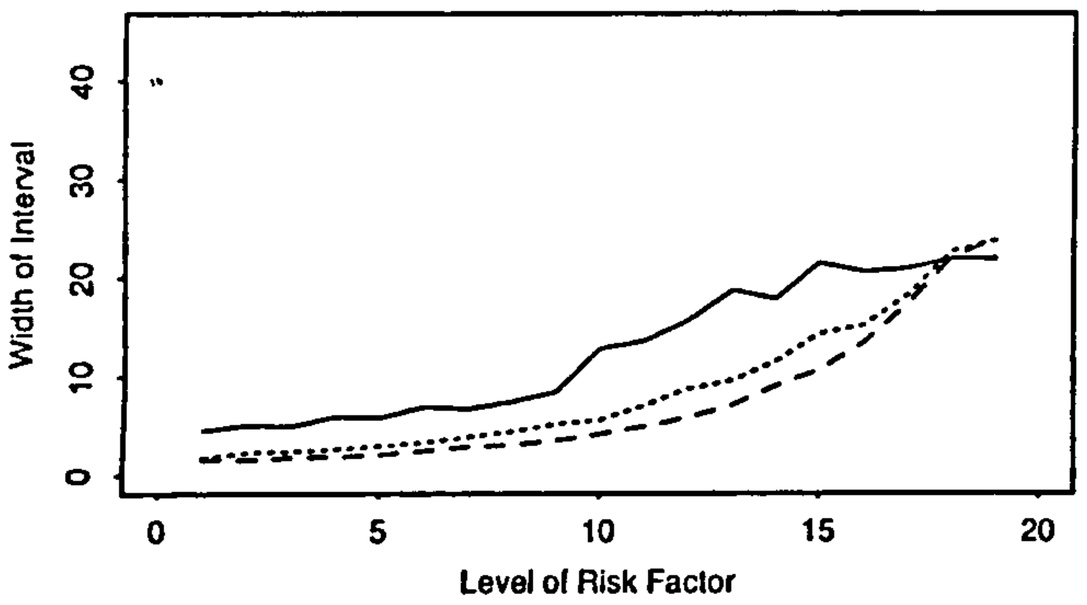
Sample Size = 50



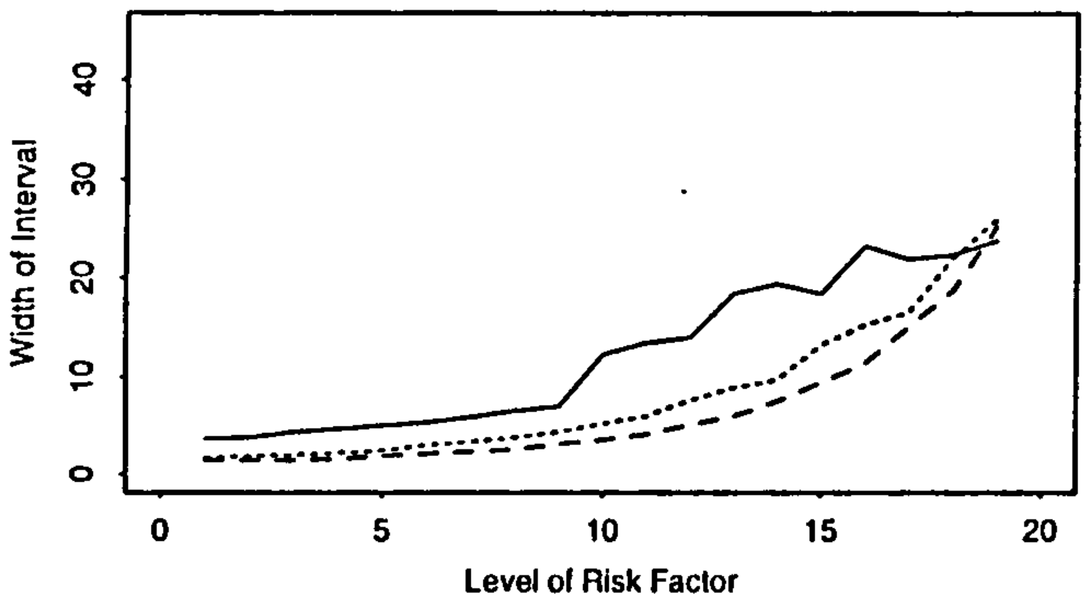
Sample Size = 75



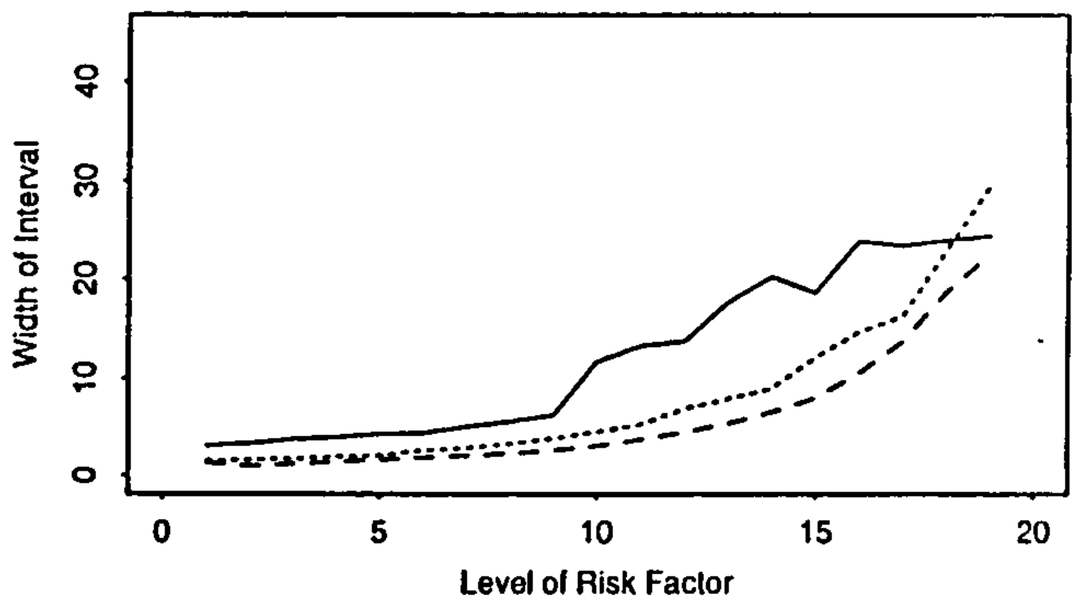
Sample Size = 100



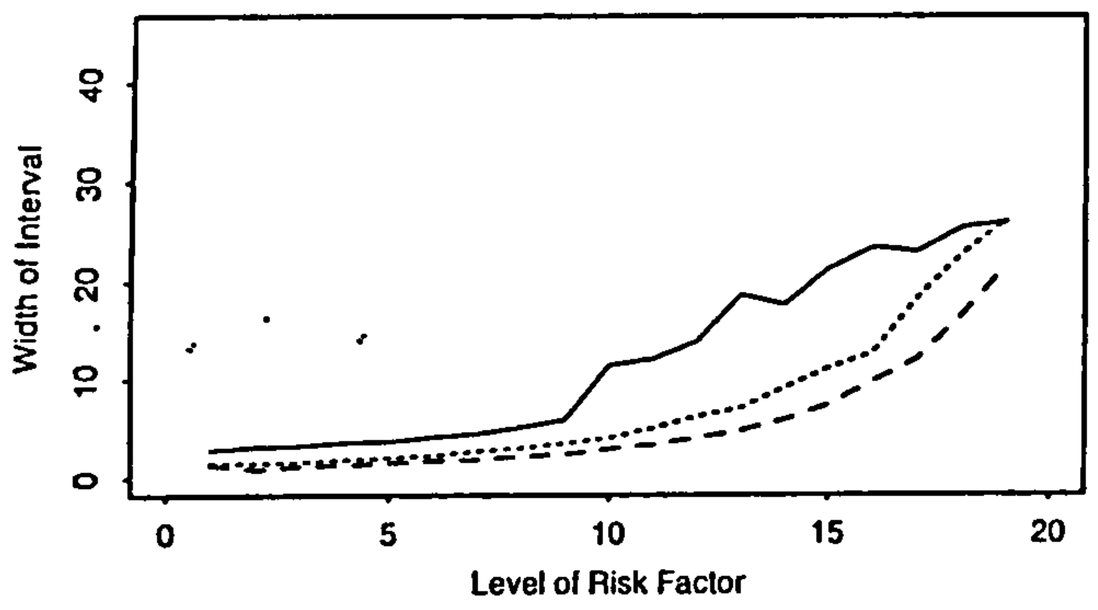
Sample Size = 150



Sample Size = 200



Sample Size = 250



Sample Size = 300

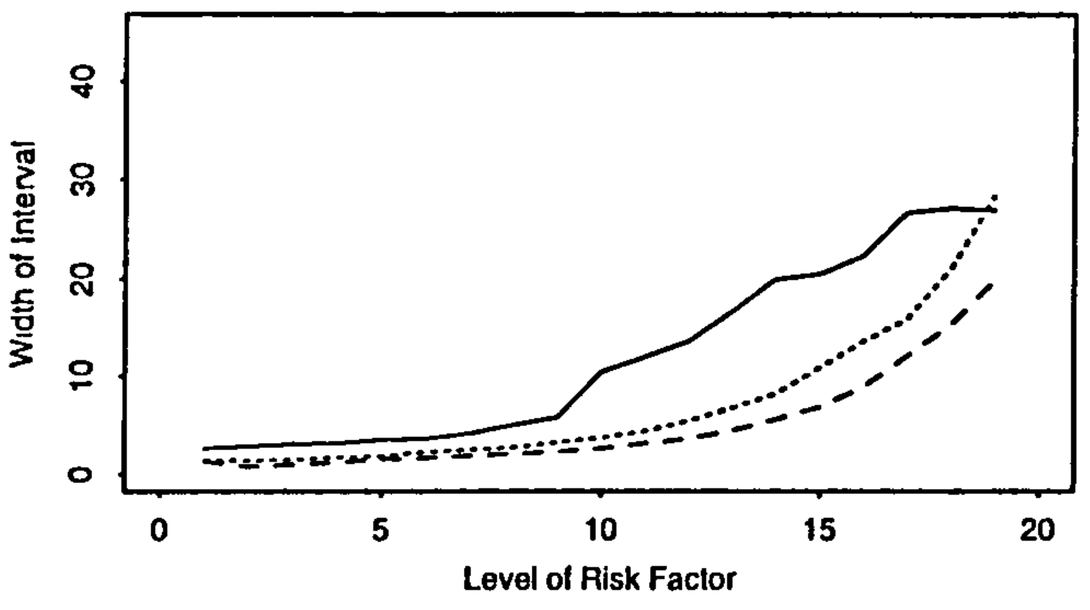


Figure 3.8.5

Pairwise Cells Method - Linear Relative Risk

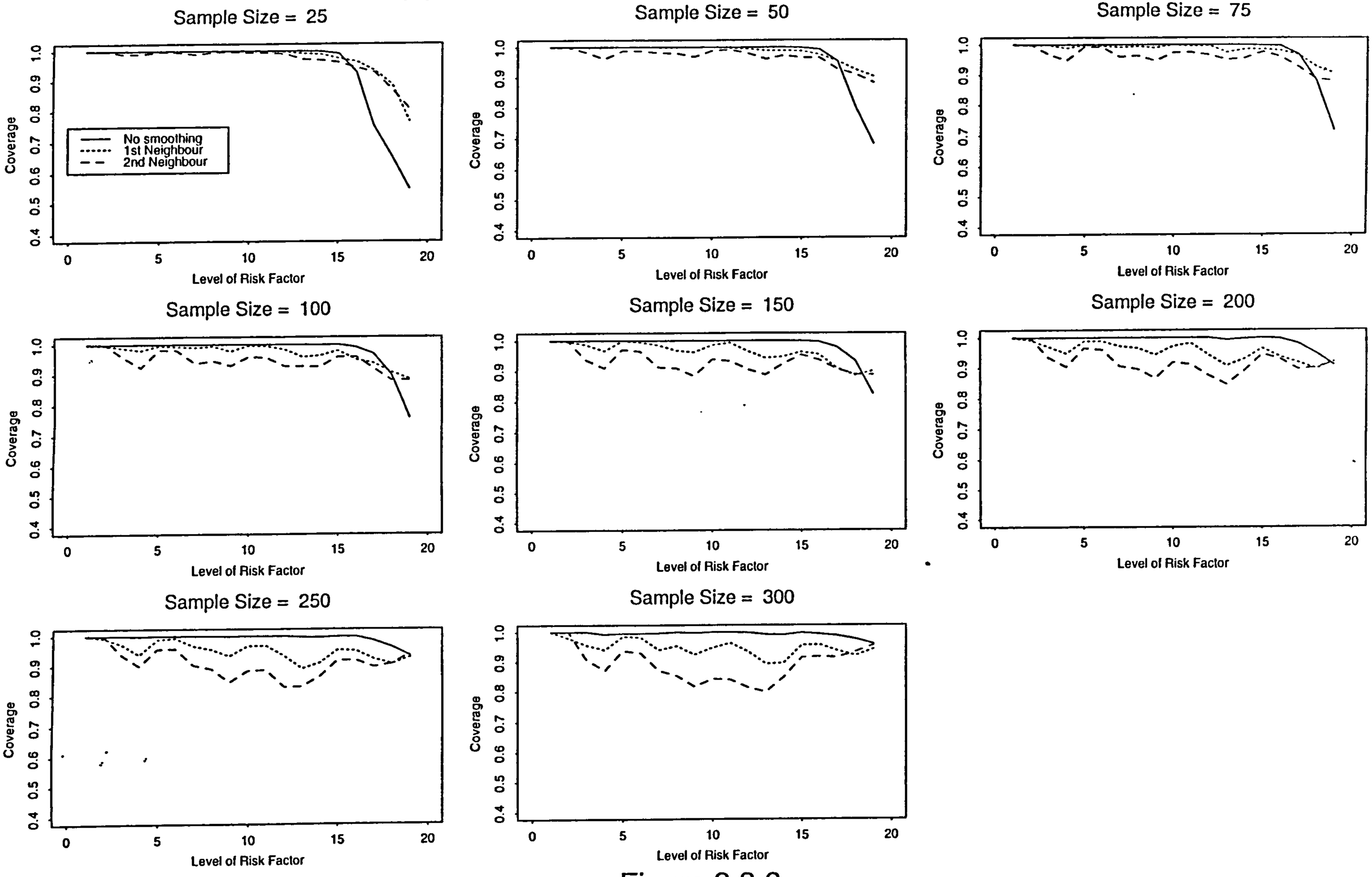


Figure 3.8.6

Pairwise Cells Method - Linear Relative Risk

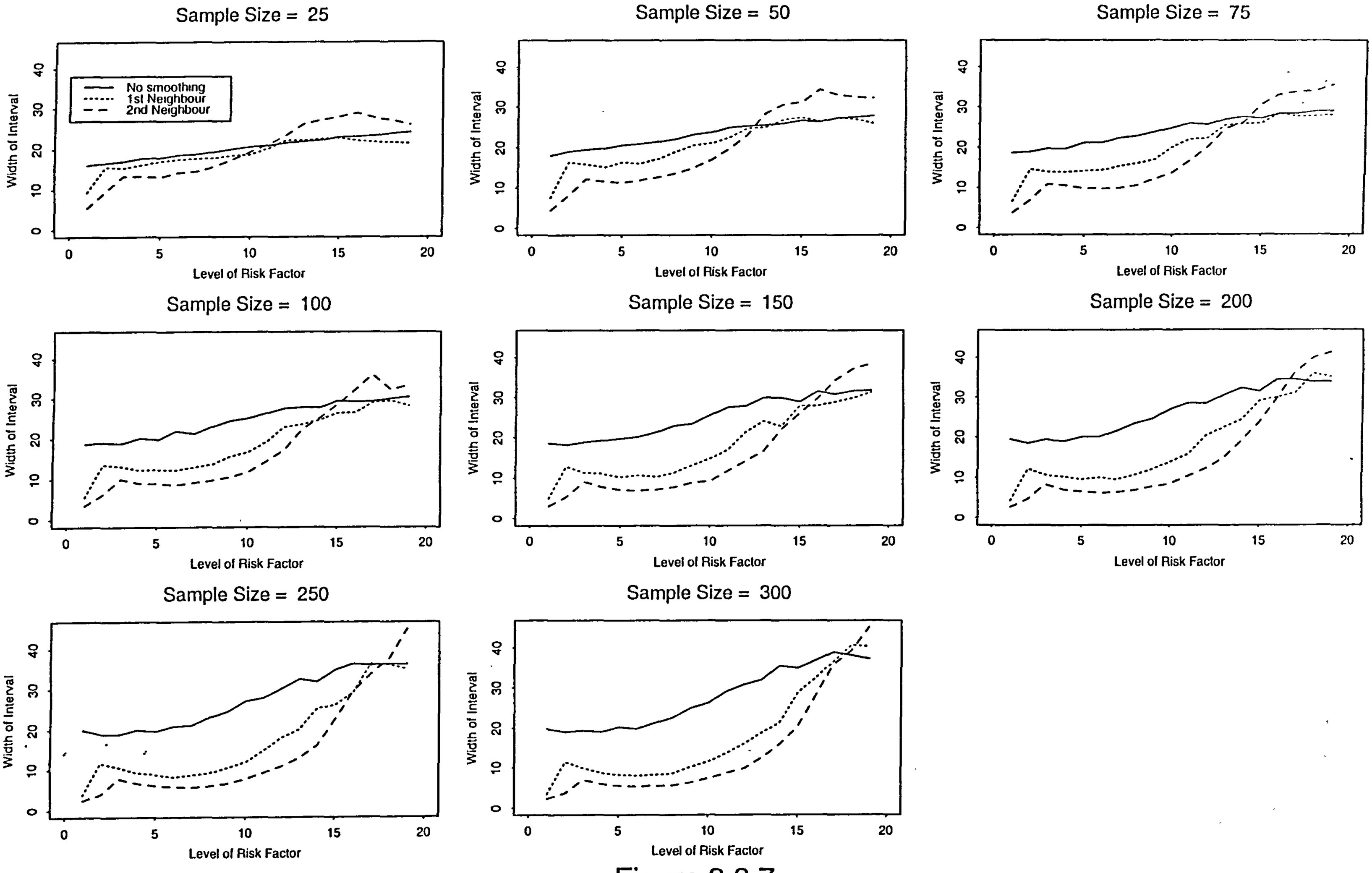


Figure 3.8.7

linear nature and a reasonable amount of data at the baseline, it is clear that more precise estimation will be available closer to the baseline (i.e. at smaller values of the risk factor). In general the coverage is quite poor and the confidence intervals relatively wide for *larger* values of the risk factor. One explanation for this is that since these values of the risk factor are quite far removed from the baseline and in data sparse areas, then it is inevitable that there will be less precise estimation at these values.

In summary it appears as though both methods perform reasonably well under this scenario, in terms of precision, bias and coverage. There is evidence that the conditional likelihood method produces slightly more precise estimates particularly for smaller sample sizes whilst the pairwise cells method results in estimates which display marginally less bias. In the absence of smoothing both methods require relatively large sample sizes before estimates can be produced which are both precise and display little bias. However once smoothing is introduced both methods quickly become more precise and display much less bias even for small sample sizes. There is little to choose between the two methods in terms of coverage, although it should be noted that both methods, particularly the pairwise cells method, produce unrealistic levels of coverage with no smoothing and smaller sample sizes.

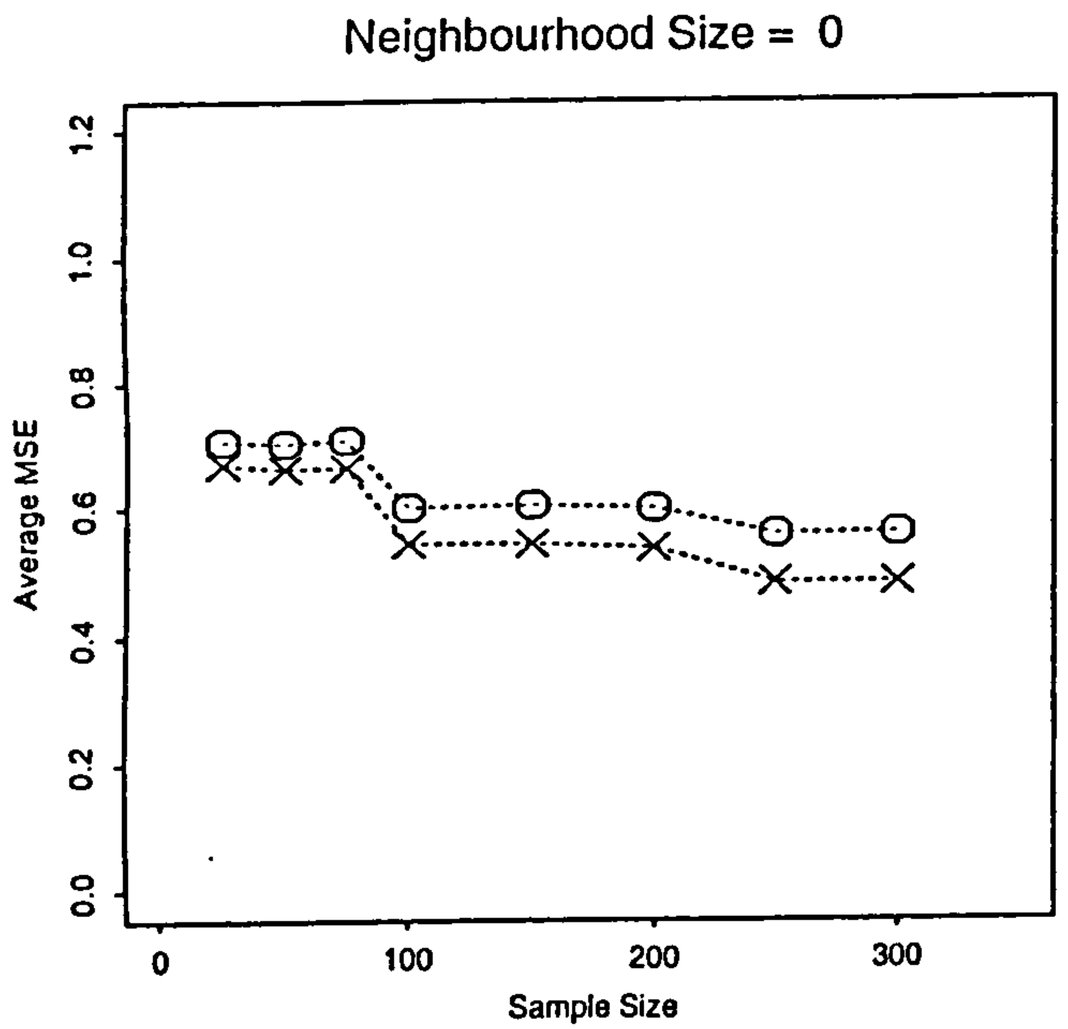
Scenario 2: Poisson distribution for $p(z_1$ / not diseased, $z_2)$, a step Relative Risk function.

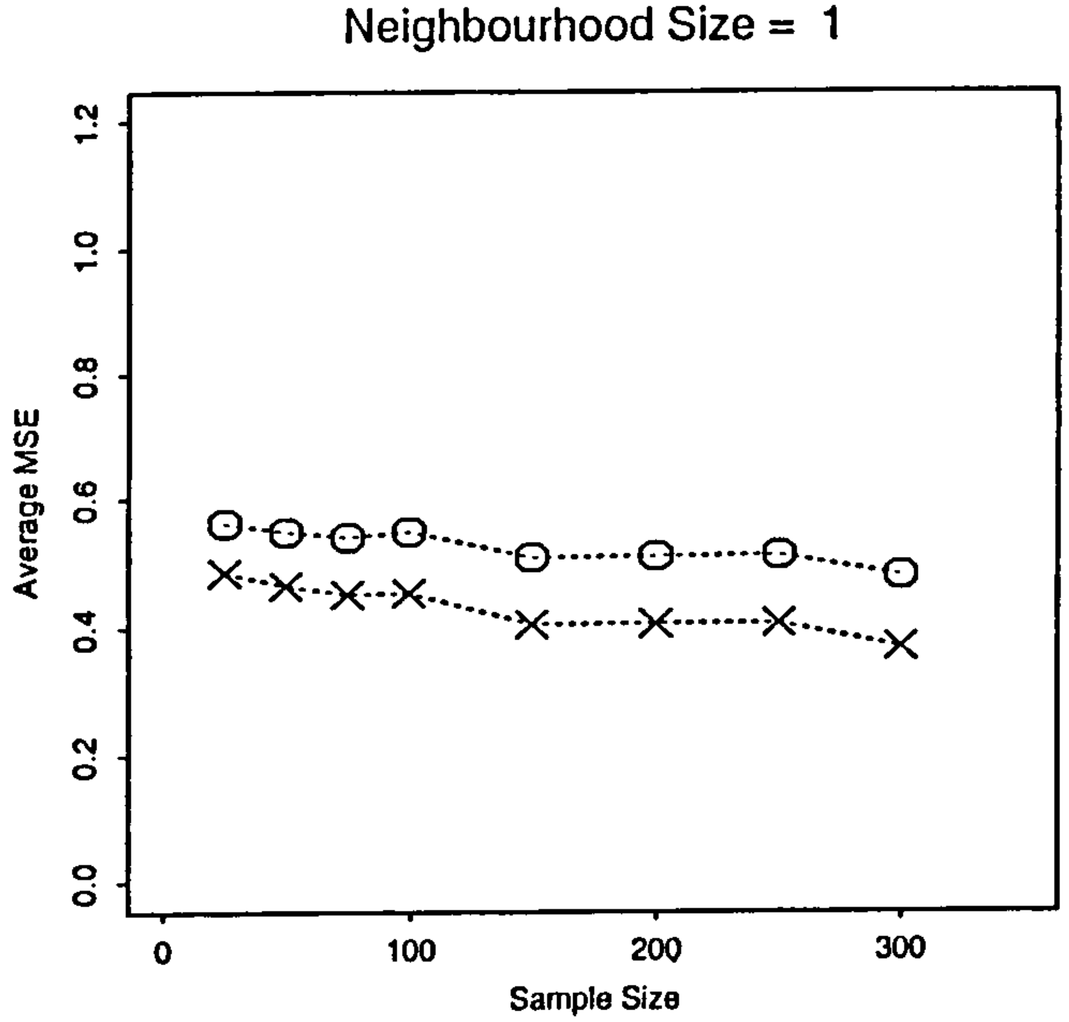
Here the same underlying distribution has been used for the control population as in scenario 1. In this scenario, however, a Relative Risk function has been incorporated

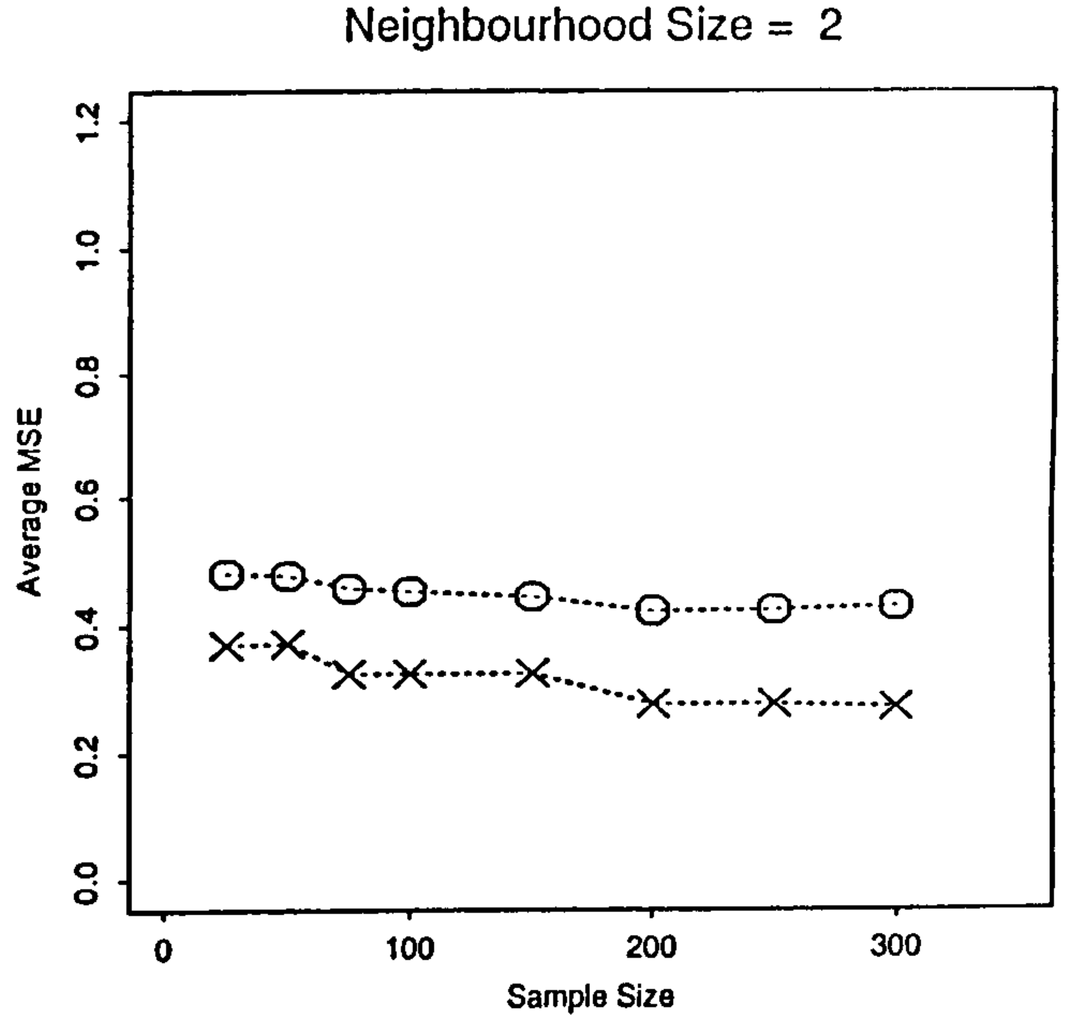
which exhibits only ***one large step*** in the Relative Risk compared to scenario 1 where the Relative Risk changed at ***every level*** of the risk factor, albeit by a smaller amount. Therefore it is perhaps sensible to think that the non-parametric methods should find this scenario easier to reproduce than the situation where the Relative Risk was of a linear nature.

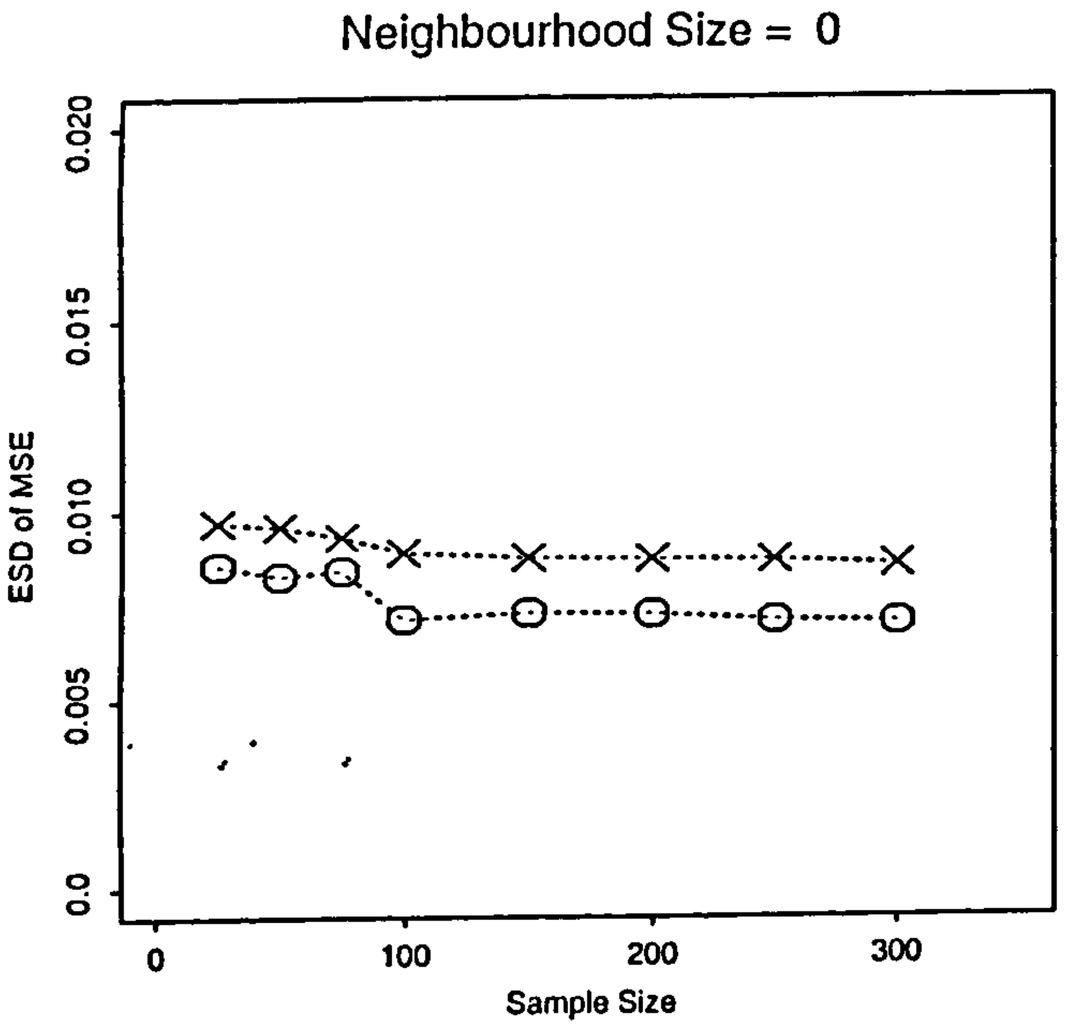
Figures 3.8.8 - 3.8.13 show the results for this simulation study. A comparison of Figures 3.8.8 and 3.8.9 with Figures 3.8.2 and 3.8.3 backs up the suggestion that this scenario is easier to reproduce both in terms of precision ***and*** bias. Both non-parametric methods produce estimates which are, in general, moderately more precise and exhibit slightly less bias when the Relative Risk function is of a step nature. This is particularly evident when no smoothing has been used. However, although slight differences exist between the two scenarios in terms of the ***degree*** of precision and bias, the actual ***patterns*** produced across both scenarios are very similar. Here, as in scenario 1, there is evidence that the conditional likelihood method produces slightly more precise estimates (frames 1-3 of Figure 3.8.8) while the pairwise cells method produces estimates which are marginally less biased (frames 1-3 of Figure 3.8.9). There is again evidence that both methods are less precise and more biased with ***smaller sample sizes***, particularly when ***no smoothing*** has been used. The introduction of smoothing produces a ***marked improvement*** in the precision and bias of the resultant estimates, especially for smaller sample sizes.

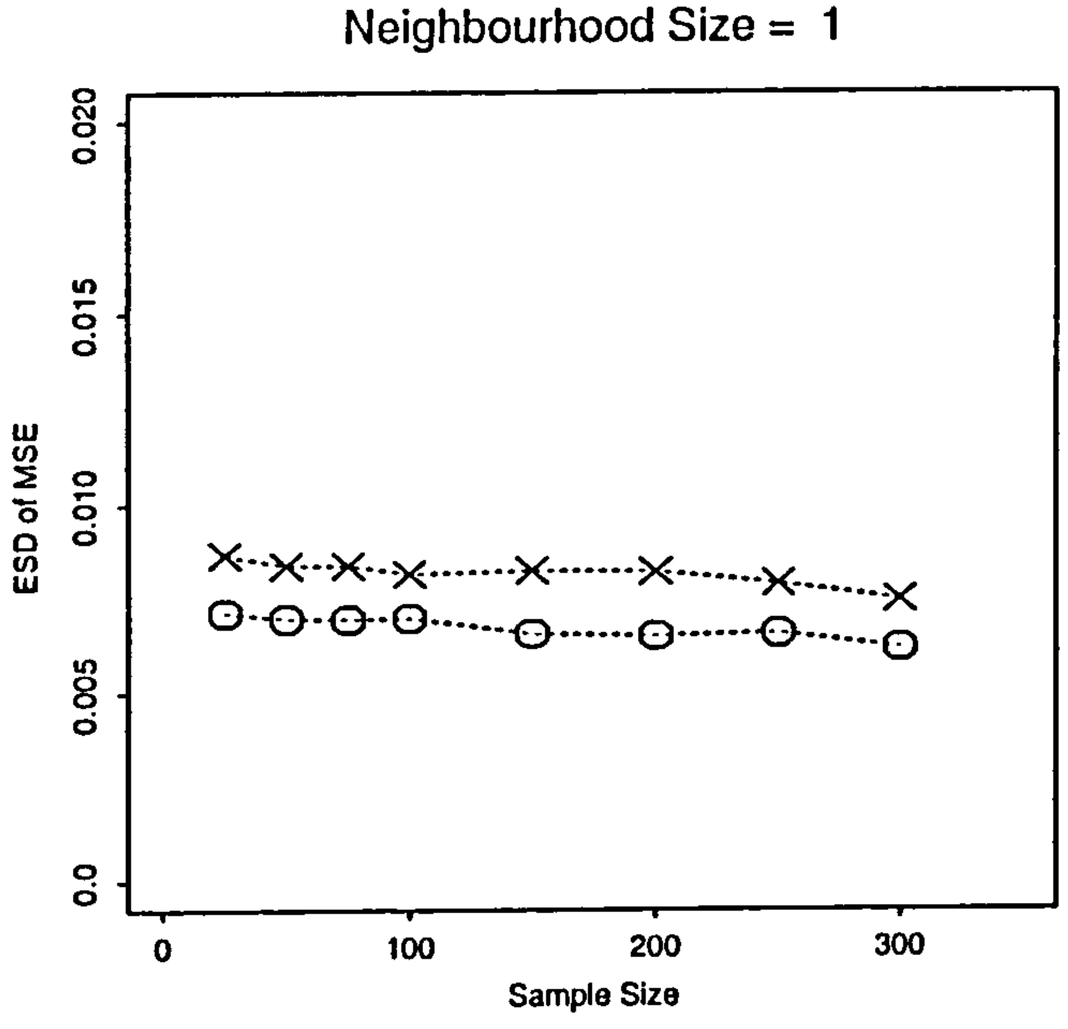
In this scenario, the true value for the log of the relative Risk for values of the risk factor between 0 and 9 is 1 and jumps to approximately 2.3 for values of the risk factor between 10 and 19 (See Figure 3.8.1). Given this range of values Figures 3.8.8 and 3.8.9


Neighbourhood Size = 0
Average MSE
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
100
200
300
Sample Size


Neighbourhood Size = 1
Average MSE
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
100
200
300
Sample Size


Neighbourhood Size = 2
Average MSE
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
100
200
300
Sample Size


Neighbourhood Size = 0
ESD of MSE
0.0
0.005
0.010
0.015
0.020
0
100
200
300
Sample Size


Neighbourhood Size = 1
ESD of MSE
0.0
0.005
0.010
0.015
0.020
0
100
200
300
Sample Size

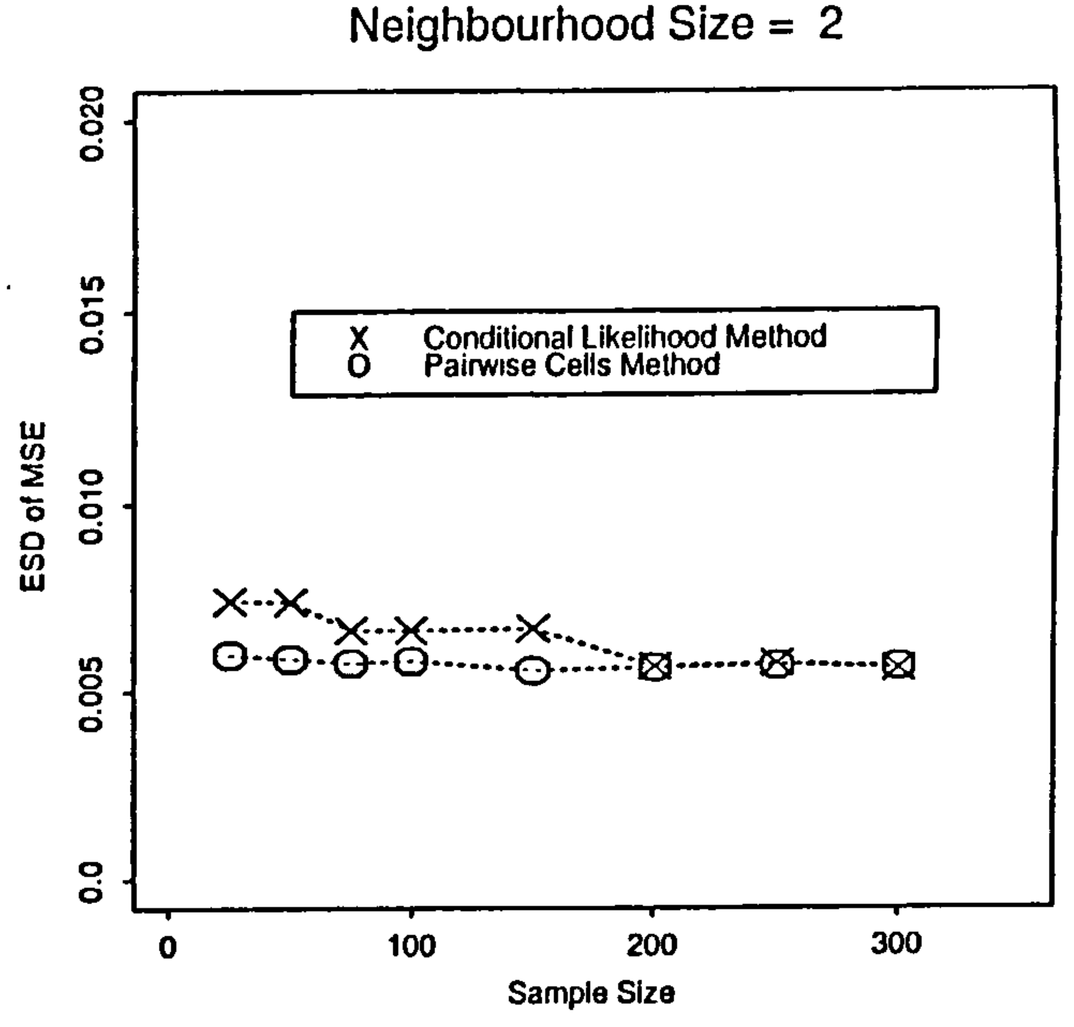

Neighbourhood Size = 2
ESD of MSE
0.0
0.005
0.010
0.015
0.020
X Conditional Likelihood Method
O Pairwise Cells Method
0
100
200
300
Sample Size

Figure 3.8.8

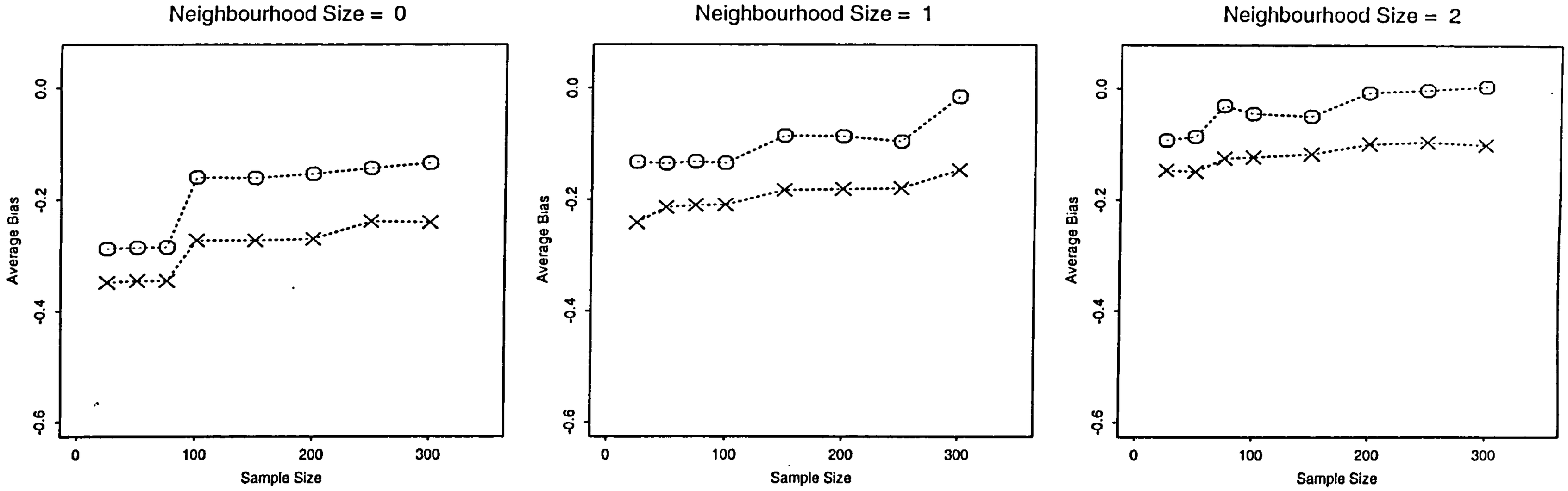

Neighbourhood Size = 0
Neighbourhood Size = 1
Neighbourhood Size = 2
Average Bias
0.0
-0.2
-0.4
-0.6
0
100
200
300
Sample Size
Neighbourhood Size = 0
Neighbourhood Size = 1
Neighbourhood Size = 2
ESD of Bias
0.020
0.015
0.010
0.005
0.0
X Conditional Likelihood Method
O Pairwise Cells Method

Figure 3.8.9

suggest that, in the ***absence of smoothing,*** a sample size of at least 100 pairs is required before the magnitudes of the average mean square error and average bias reduce to a reasonable level. Regardless of sample size, when ***a first order neighbourhood of smoothing*** is used, the average mean square error ranges from approximately 0.4 to approximately 0.6 with average bias between -0.05 and -0.25 suggesting that both methods are reasonably precise and exhibit only minor levels of bias once smoothing is introduced. With ***a second order neighbourhood of smoothing***, a further, smaller, improvement is observed in both the average mean square error and the average bias. However, as in scenario 1, it is clear that both methods again ***underestimate*** the true log Relative Risk regardless of sample size and level of smoothing.

In terms of coverage and the average width of the nominal 95% confidence intervals, Figures 3.8.10 - 3.8.13 display the results for the two methods. As with scenario 1 it is clear that for both methods, the coverage is ***unrealistically high*** for small sample sizes and no smoothing. This is particularly noticeable for values of the risk factor at the baseline, where the true Relative Risk is equal to 1 (i.e. values of the risk factor of 0 to 9). The primary reason for this is that, at the baseline, there is less potential for the estimates to deviate from their true Relative Risk. Since the width of the intervals at the baseline are of a reasonable magnitude when no smoothing is present, particularly for the pairwise cells method (see Figures 3.8.11 and 3.8.13), then it becomes increasingly likely that, regardless of the point estimate of Relative Risk, a very high percentage of intervals will contain the true value (i.e. a Relative Risk of 1). For both methods the coverage drops dramatically and the width of the interval increases in the ***vicinity of the cut-point,*** suggesting that this is the area where it is most difficult to obtain accurate estimates of the true Relative Risk. This is not unexpected as, at this point, there is a ***marked change*** in

Conditional Likelihood Method - Step Relative Risk

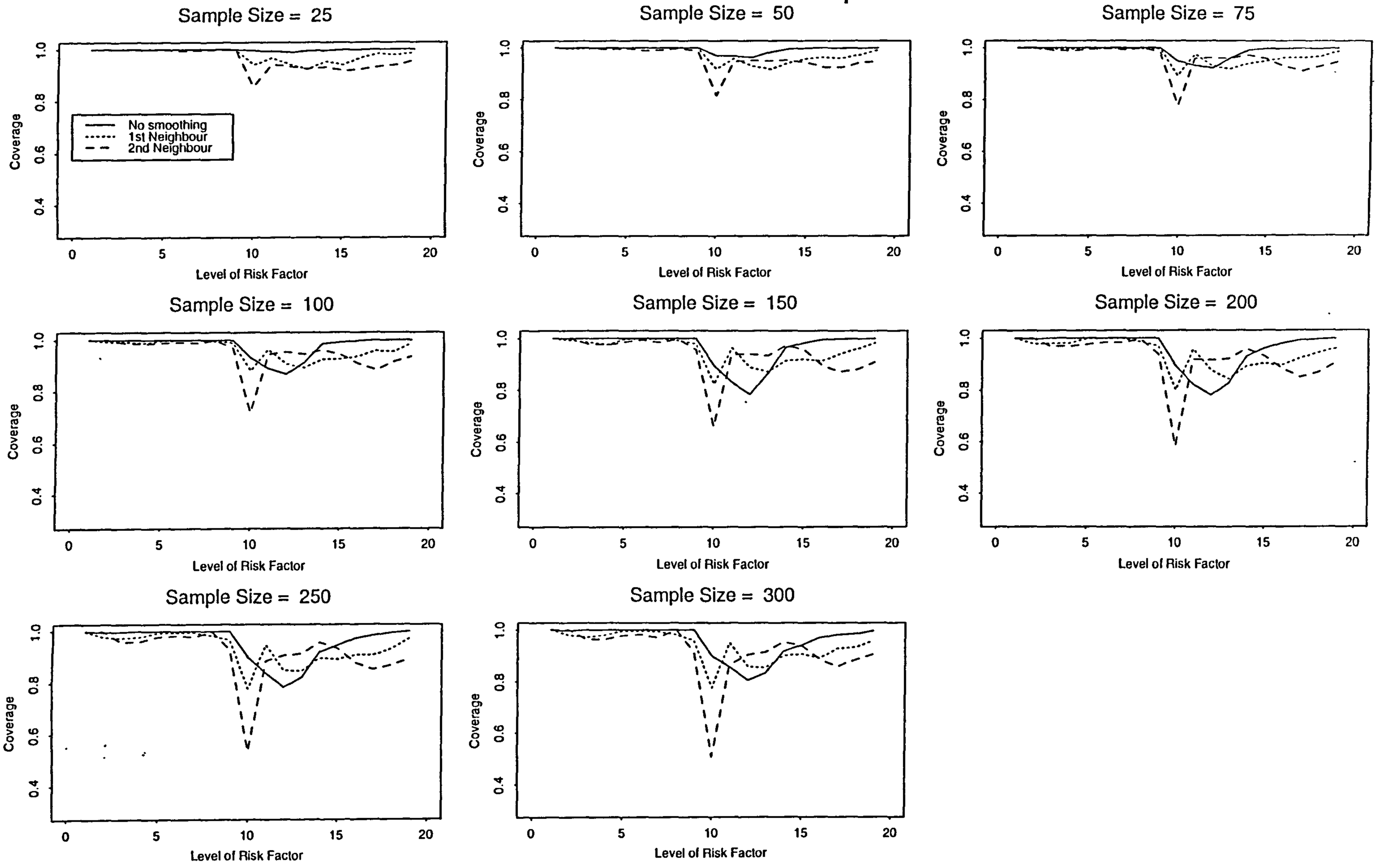
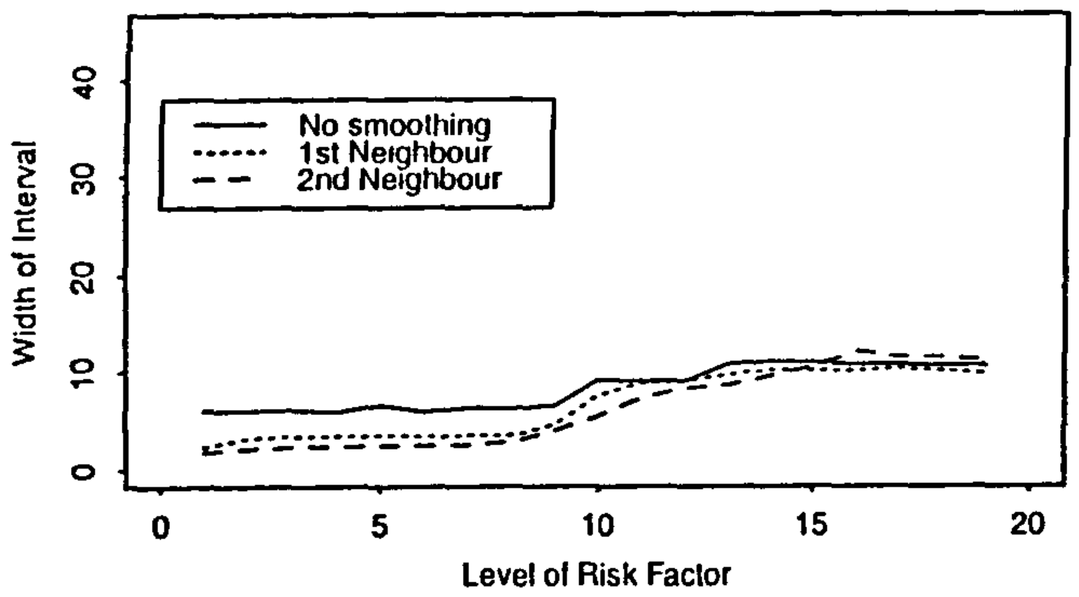


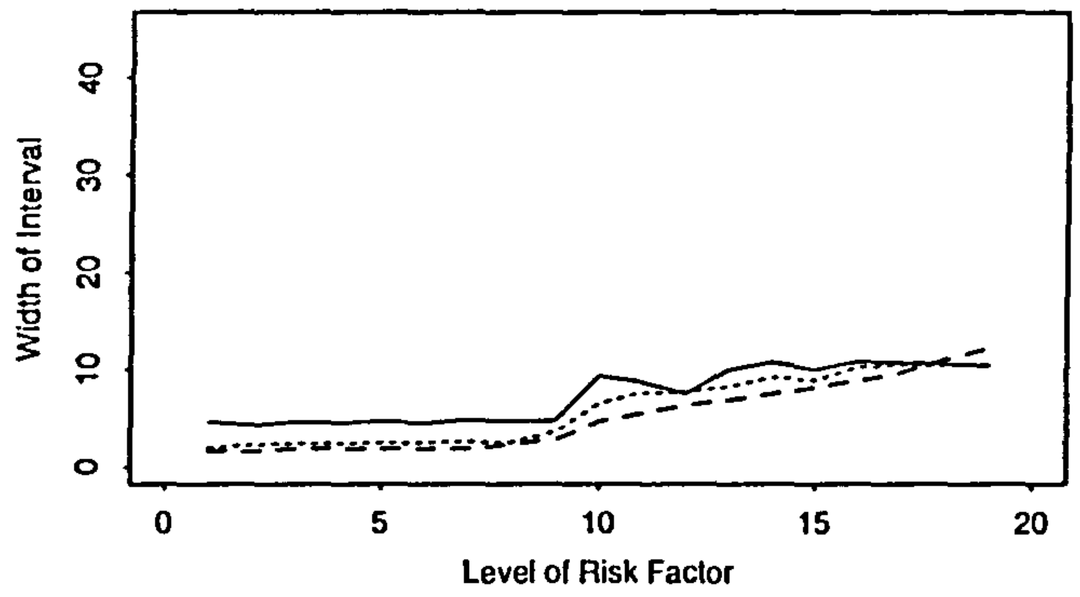
Figure 3.8.10

Conditional Likelihood Method - Step Relative Risk

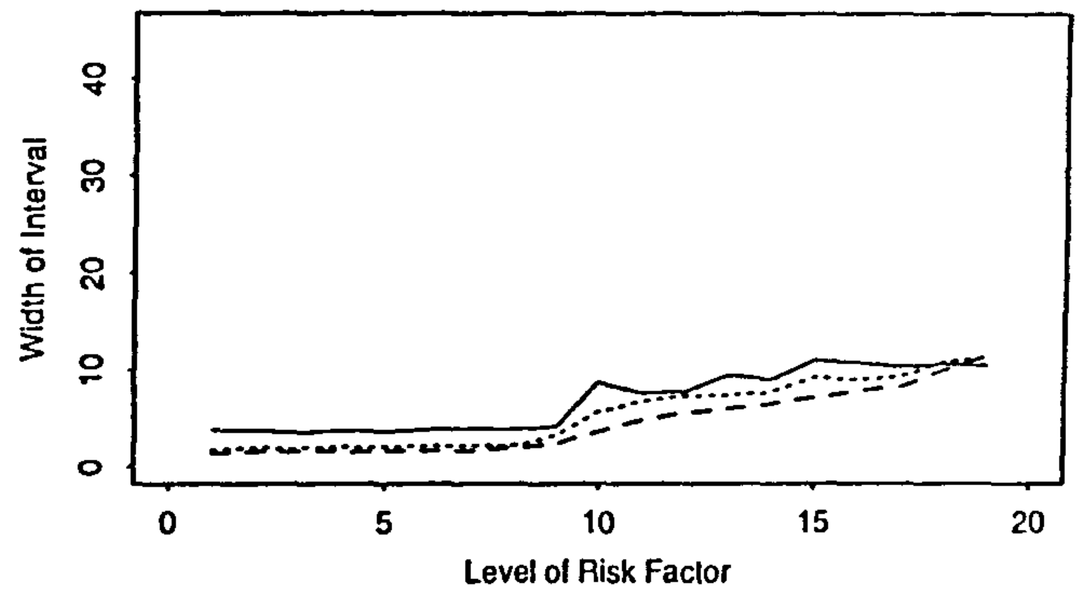
Sample Size = 25



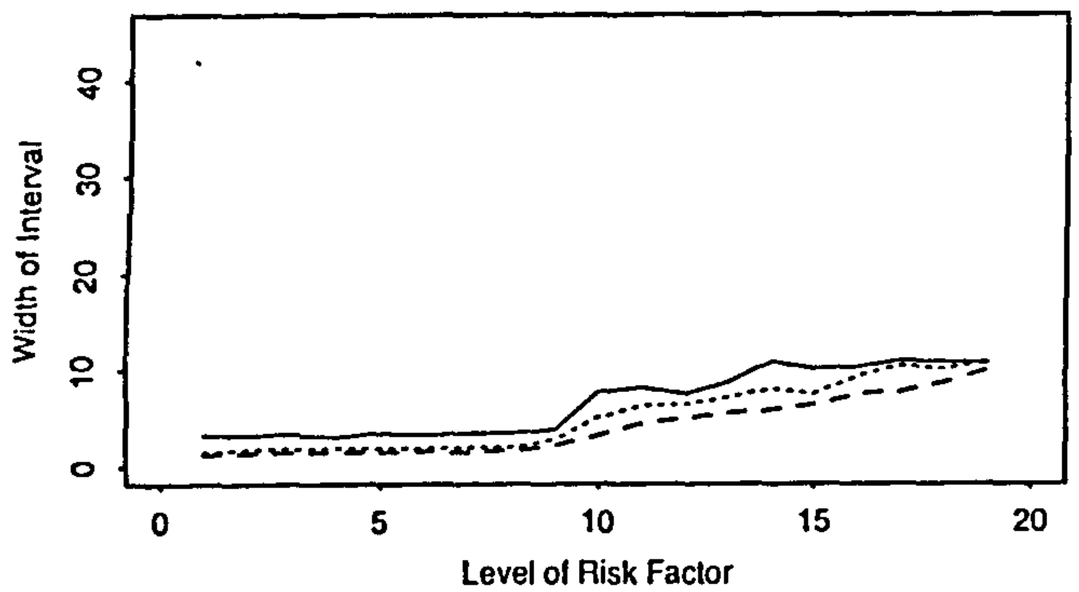
Sample Size = 50



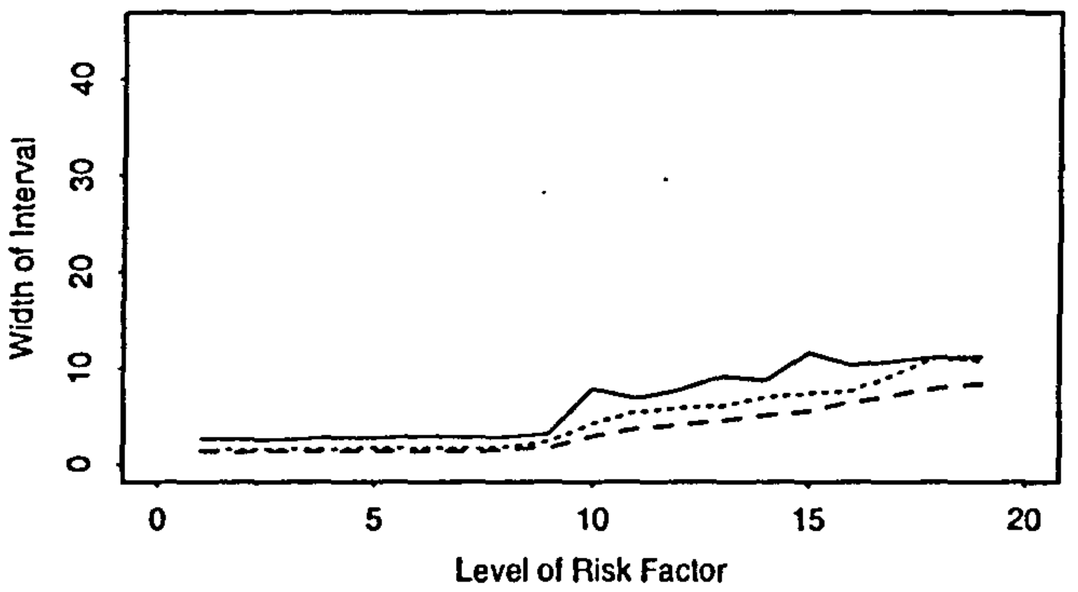
Sample Size = 75



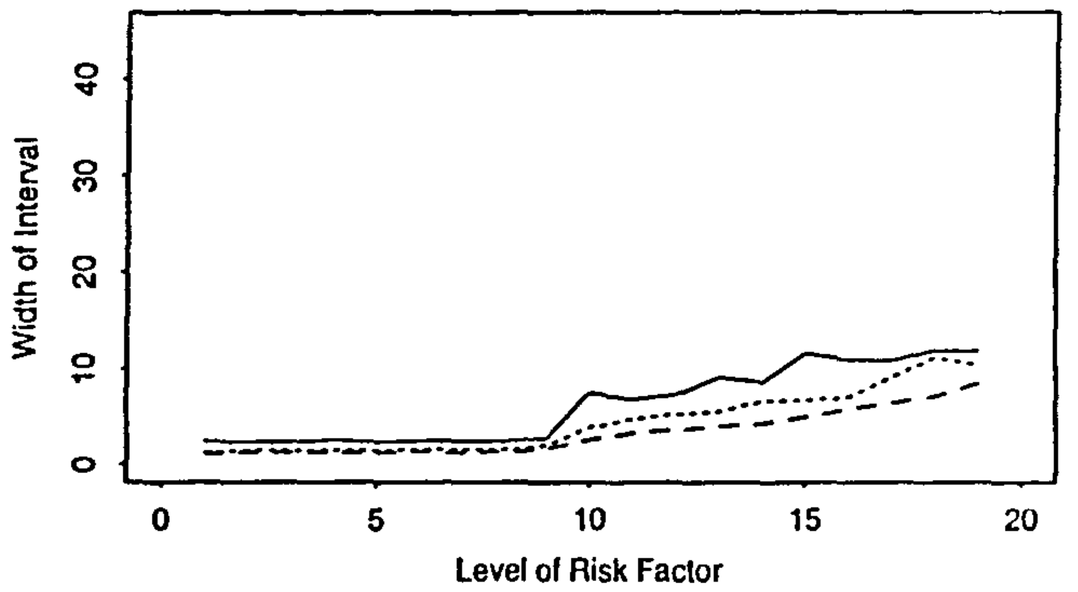
Sample Size = 100



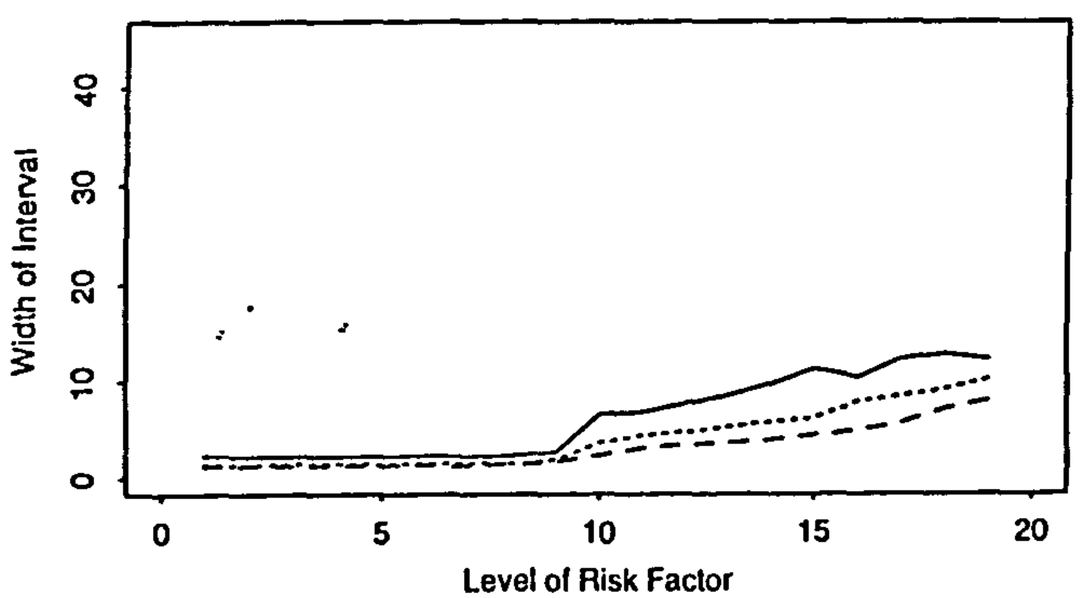
Sample Size = 150



Sample Size = 200



Sample Size = 250



Sample Size = 300

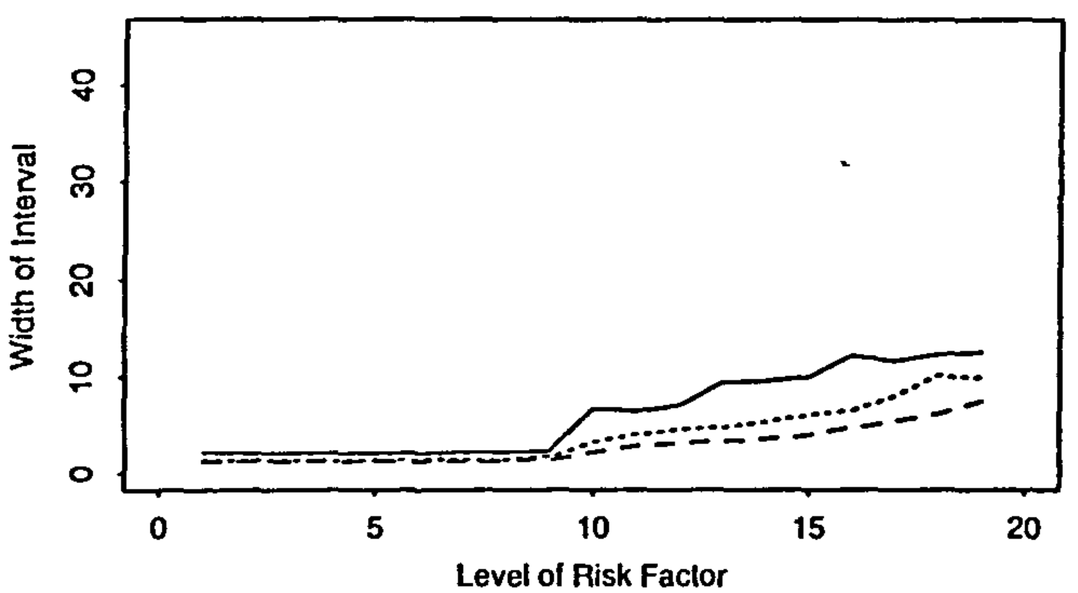


Figure 3.8.11

Pairwise Cells Method - Step Relative Risk

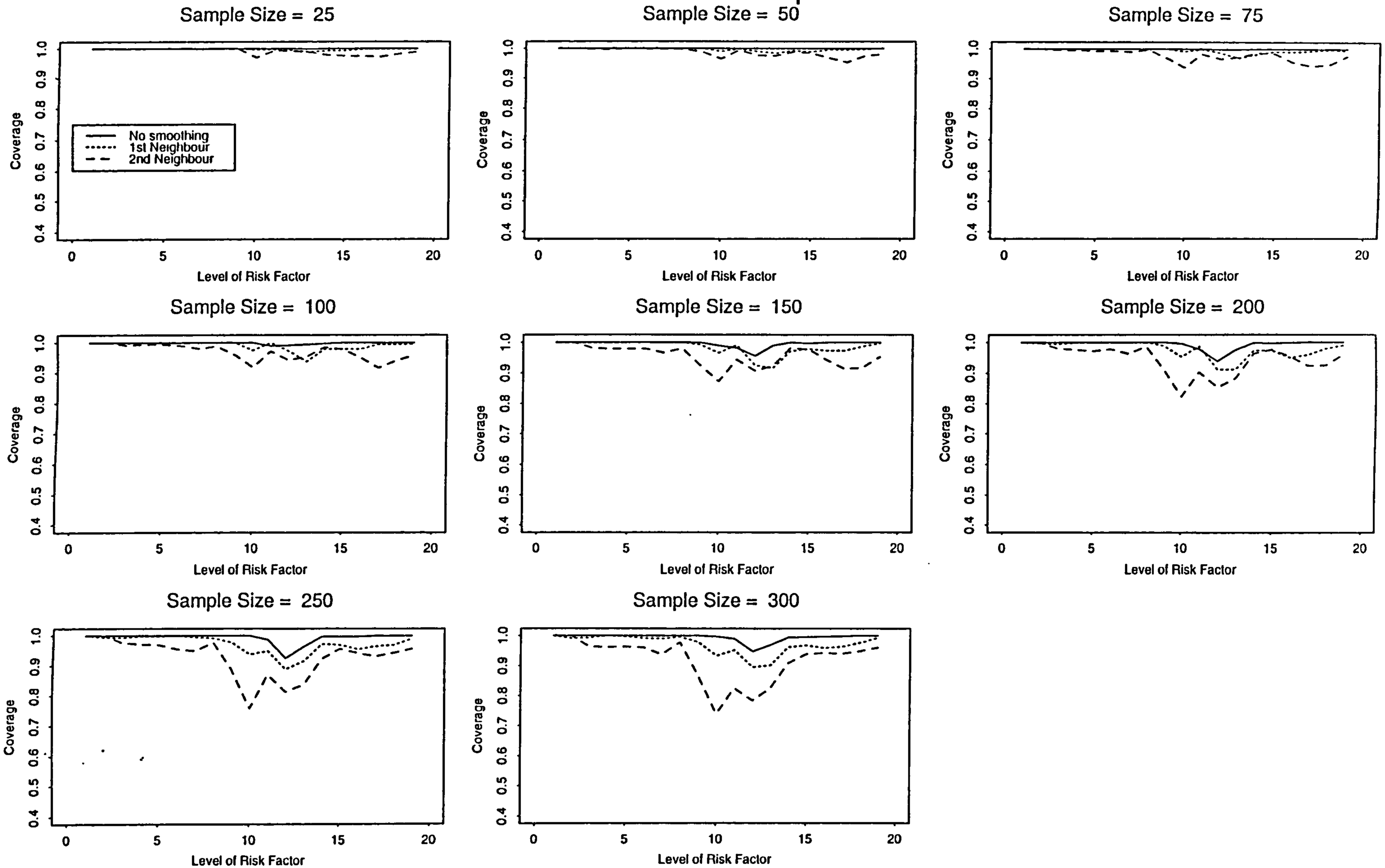
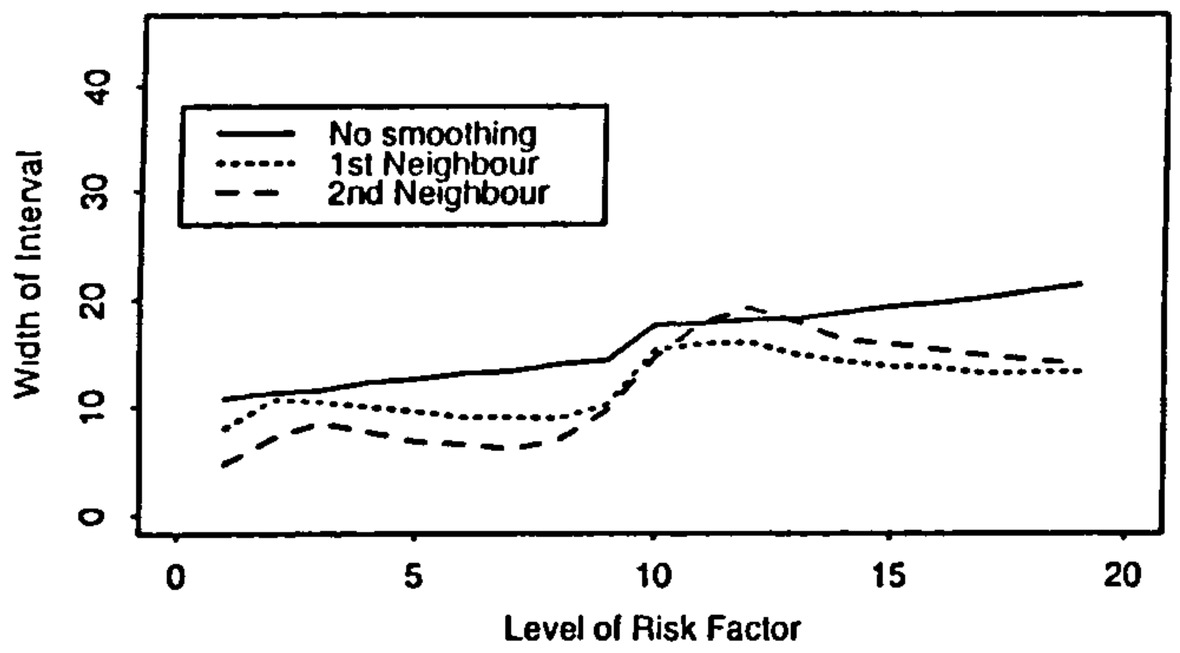


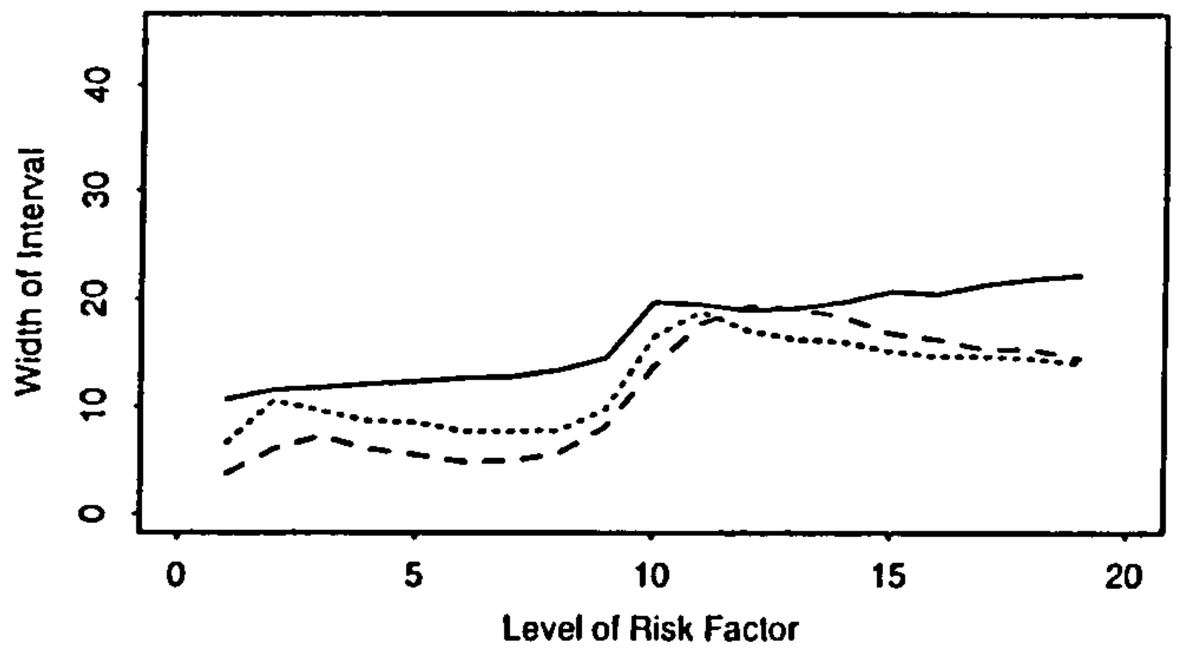
Figure 3.8.12

Pairwise Cells Method - Step Relative Risk

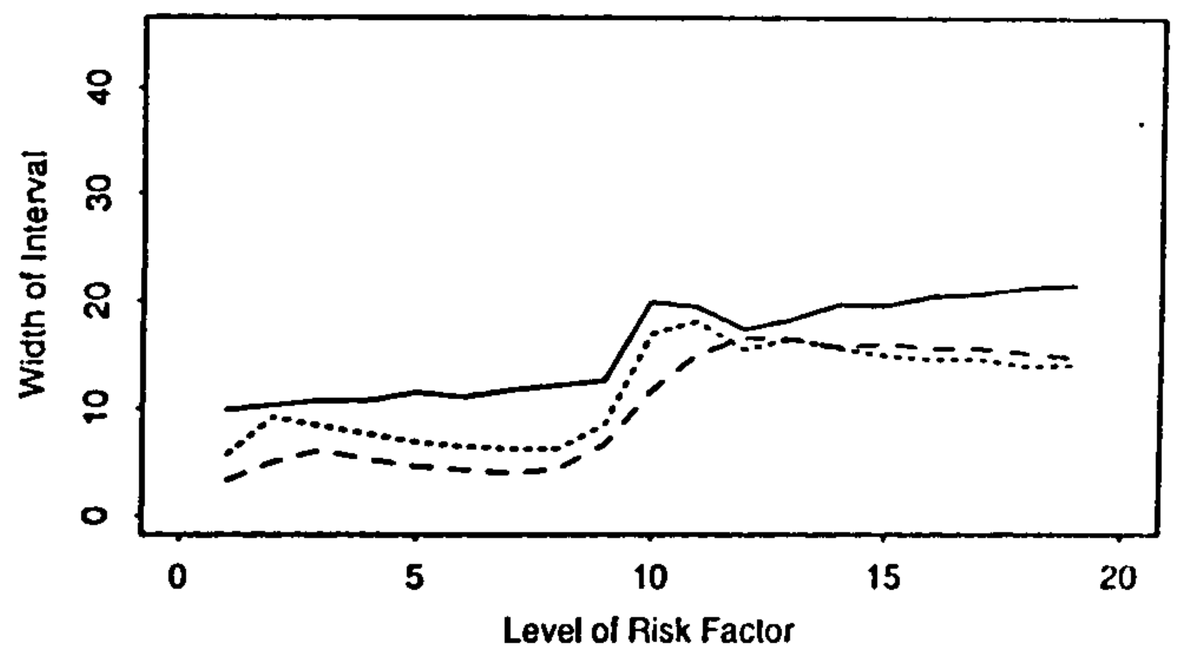
Sample Size = 25



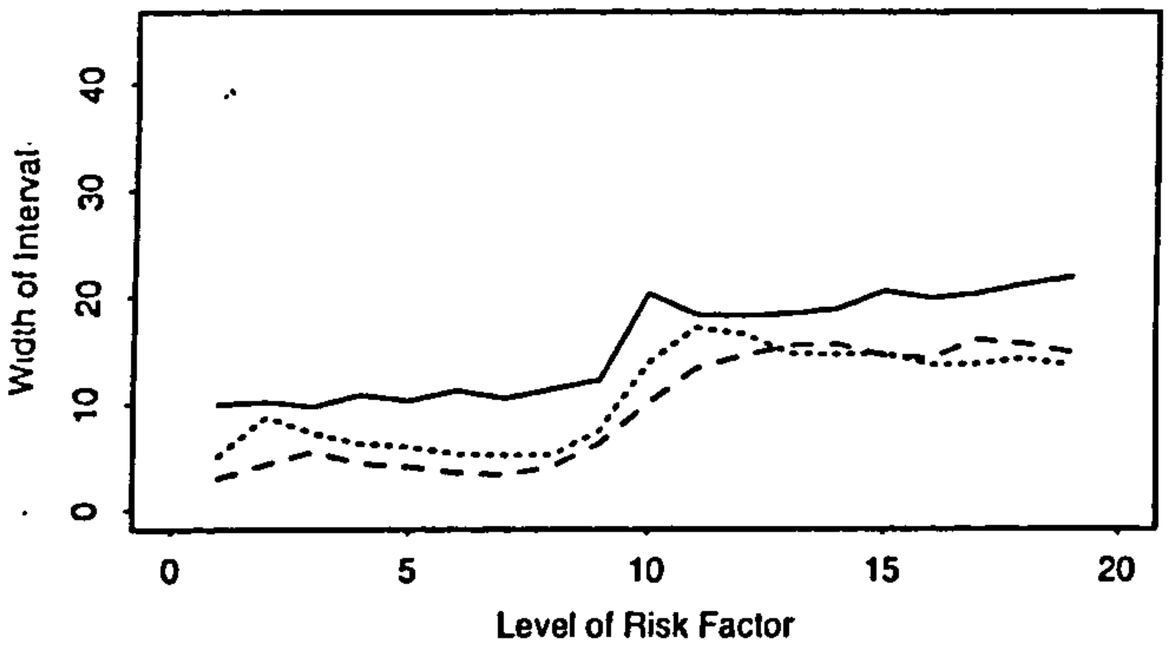
Sample Size = 50



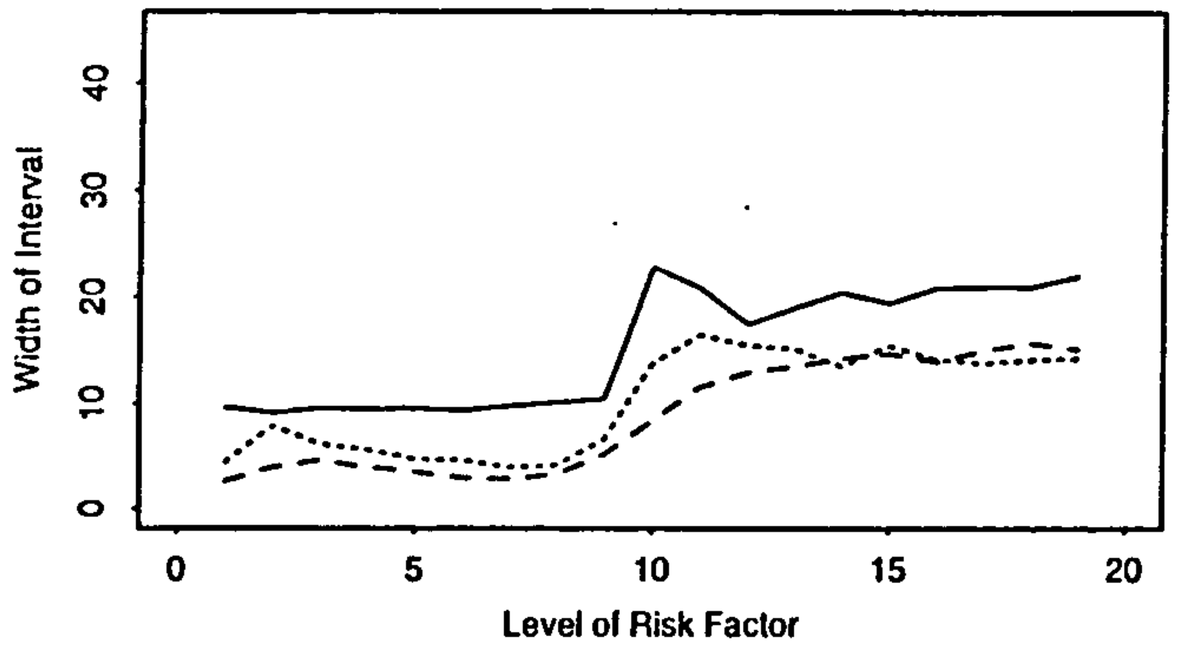
Sample Size = 75



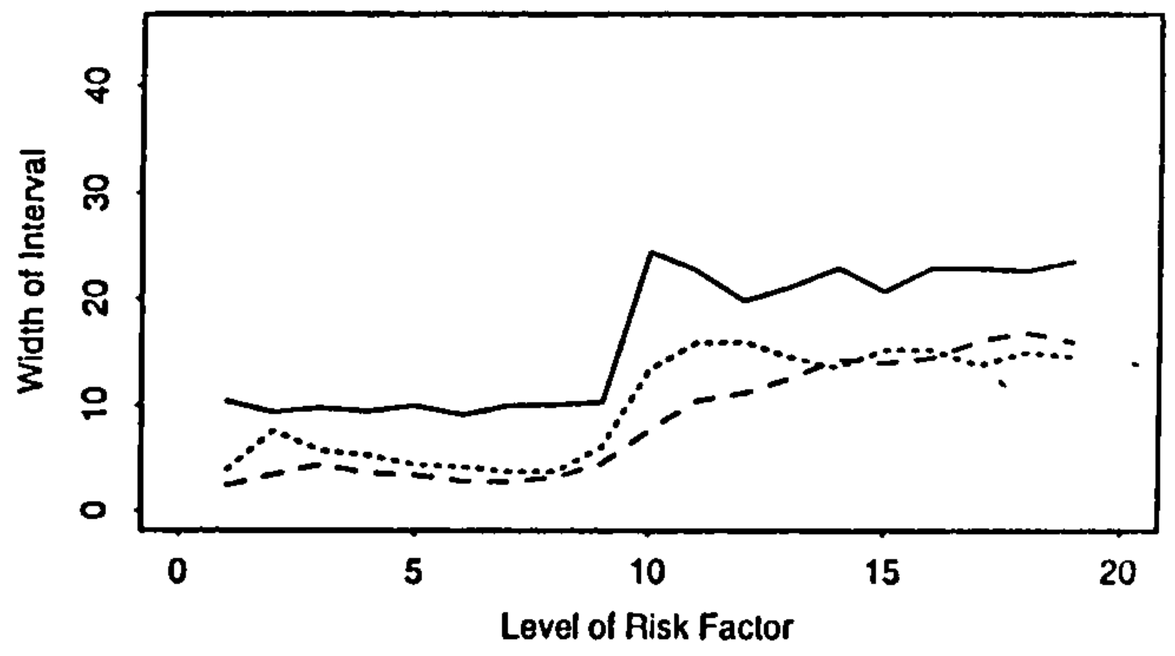
Sample Size = 100



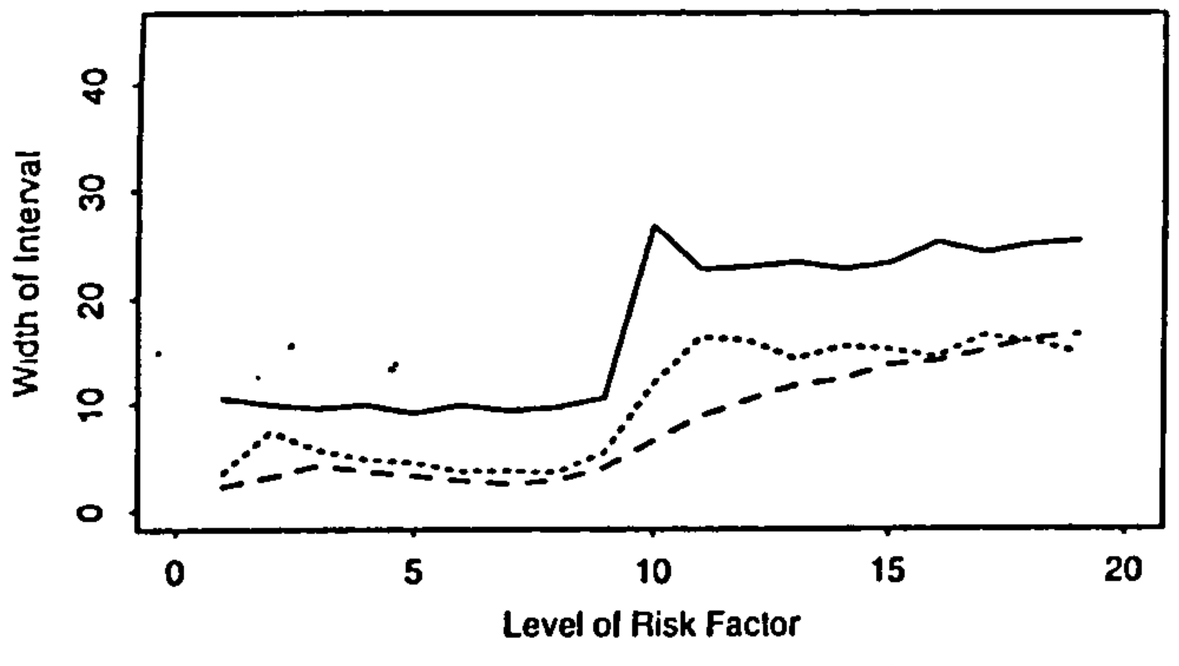
Sample Size = 150



Sample Size = 200



Sample Size = 250



Sample Size = 300

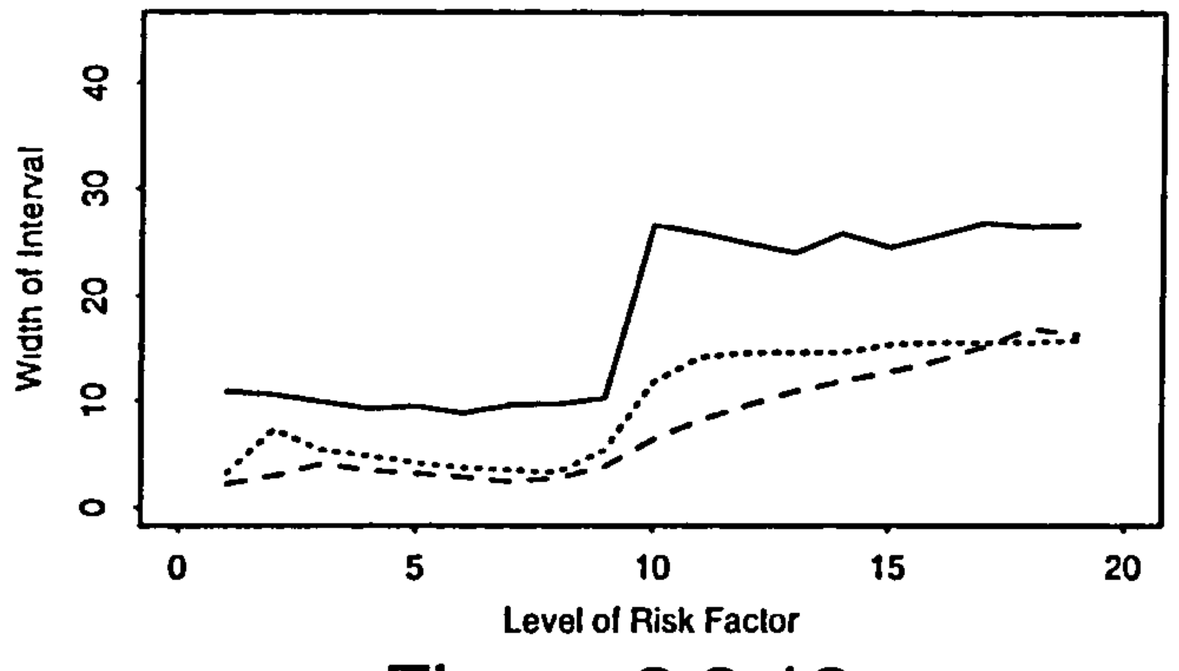


Figure 3.8.13

the true Relative Risk suggesting that this will be the most difficult point at which to obtain good estimation. After the cut-point the average width of the intervals gradually levels off leading to the coverage again increasing. This suggests that estimation is more precise as the value of the risk factor moves away from the location of the cut-point since there is no further change in the true Relative Risk. Comparing Figures 3.8.10 and 3.8.12 it is clear that, in general, the pairwise cells method appears to produce values which are closer to the nominal levels of coverage. With the conditional likelihood method the pattern of coverage is often erratic in the vicinity of the cut-point whereas the pairwise cells method produces a far more stable pattern. On the other hand Figures 3.8.11 and 3.8.13 reveal that the 95% confidence intervals are clearly wider, on average, with the pairwise cells method regardless of the value of the risk factor. For both methods of estimation, the levels of coverage would appear to be ***closest to the nominal value of95%*** when a ***neighbourhood of size 1*** is used. For reasonable sample sizes (i.e. 75 pairs or more) and a neighbourhood of size 1, the pairwise cells method results, on average, in a coverage of between 90% and 98% whilst the conditional likelihood method results, on average, in a coverage of between 85% and 95%. With a neighbourhood of size 2 there is evidence that for larger sample sizes of 200 pairs or more, both methods produce levels of coverage which are ***very low*** at the cut-point. This is due to a ***large amount*** of smoothing being carried out ***across the location of the cut-point*** resulting in the estimation of the Relative Risk being very imprecise at this point. This, in conjunction with the larger sample sizes producing intervals with smaller widths leads to a larger percentage of intervals not containing the true value. For both methods these results suggest that, regardless of sample size, a first order neighbourhood will produce the best ***combination*** of the most acceptable levels of coverage in conjunction with intervals which are not excessively wide. Further, the pairwise cells method appears, in general, to produce

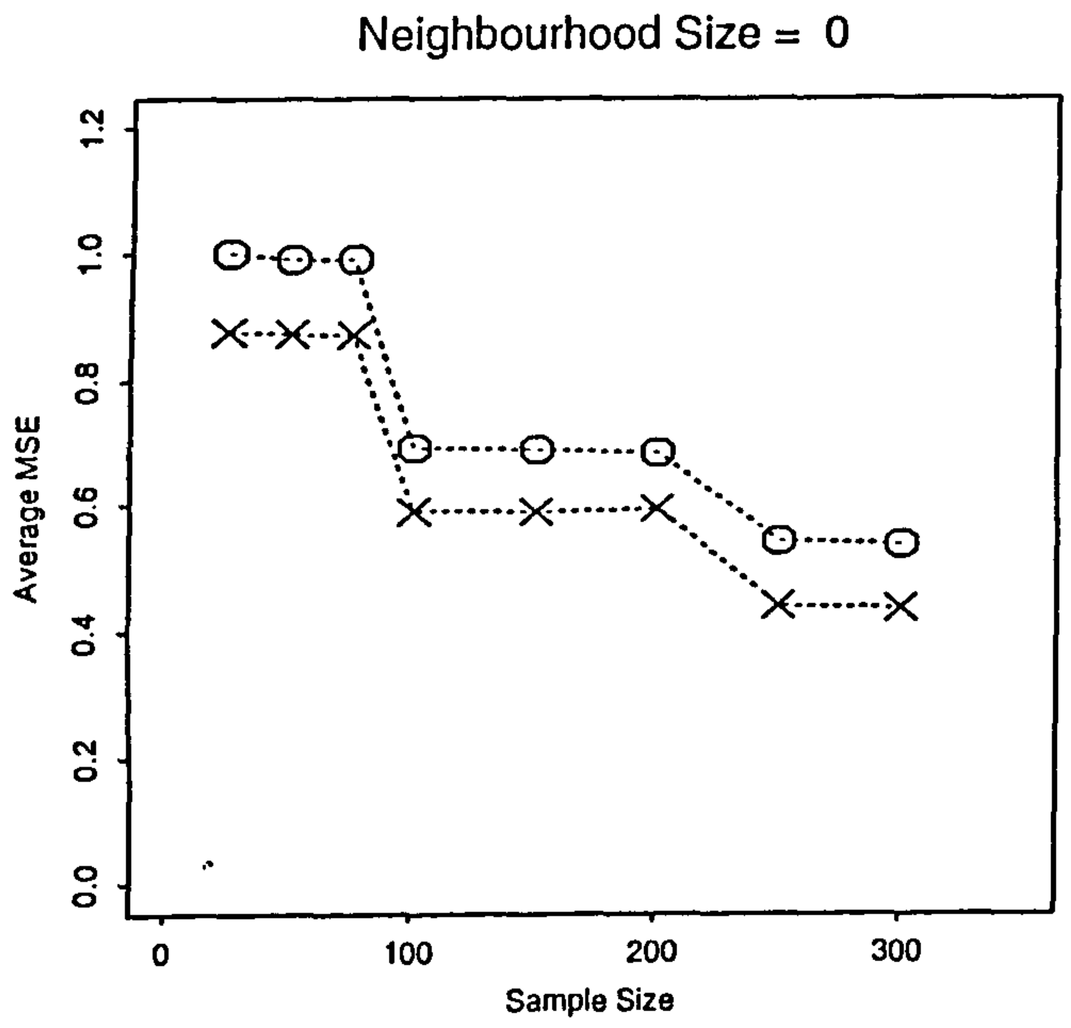
higher levels of coverage but this may be due, at least in part, to the wider intervals which are generally produced by this method. However, one genuine advantage with the pairwise cells method is that it clearly produces ***less erratic*** patters of coverage around the true location of the cutpoint.

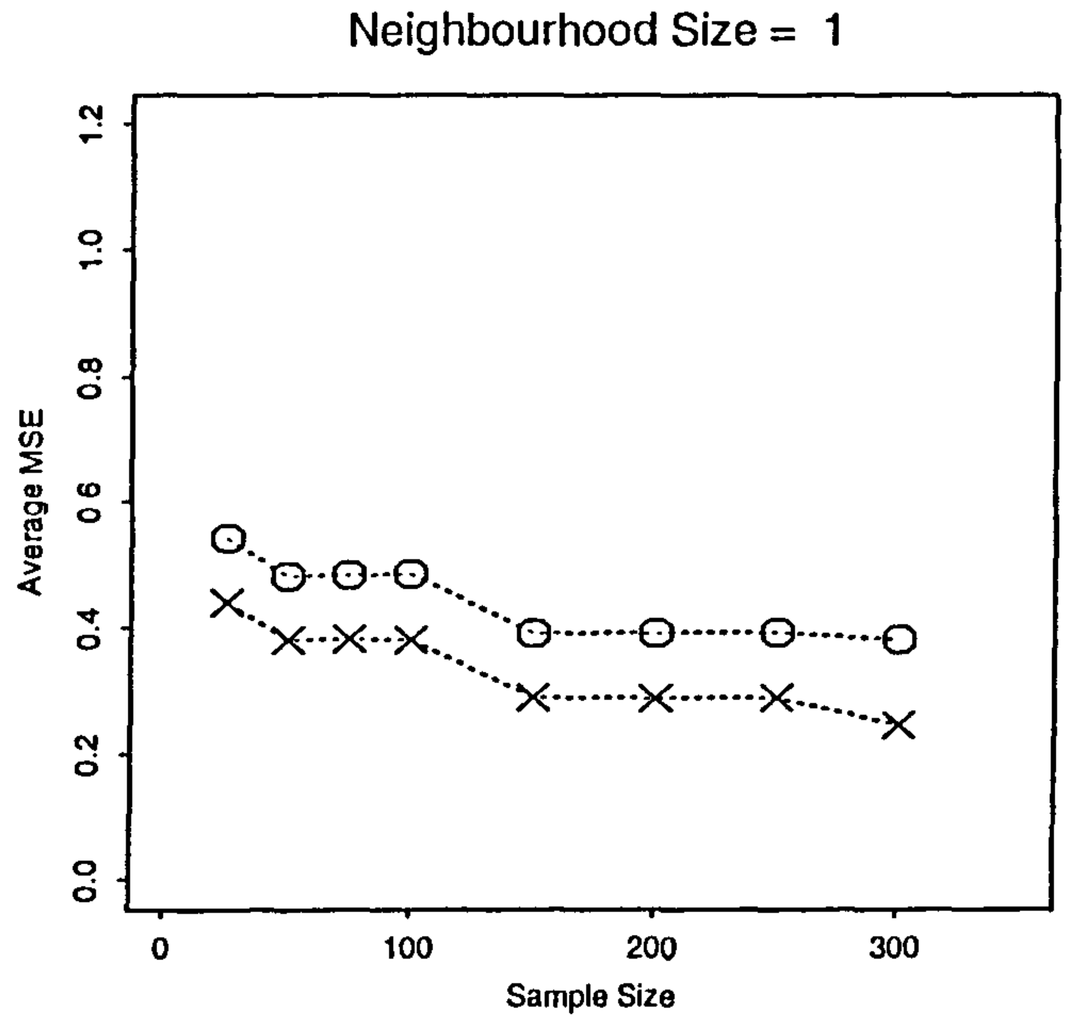
In summary, it appears that both methods of estimation produce good estimates in this scenario in terms of both precision and bias. The conditional likelihood method appears to be marginally more precise but the pairwise cells method appears to exhibit less bias. In terms of coverage the results are not as promising and there is clear evidence that the ***choice of smoothing parameter*** may have a dramatic effect on the levels of coverage, particularly for the conditional likelihood method. It is also worrying that, regardless of the method of estimation, the coverage drops considerably at the location of the cut-point.

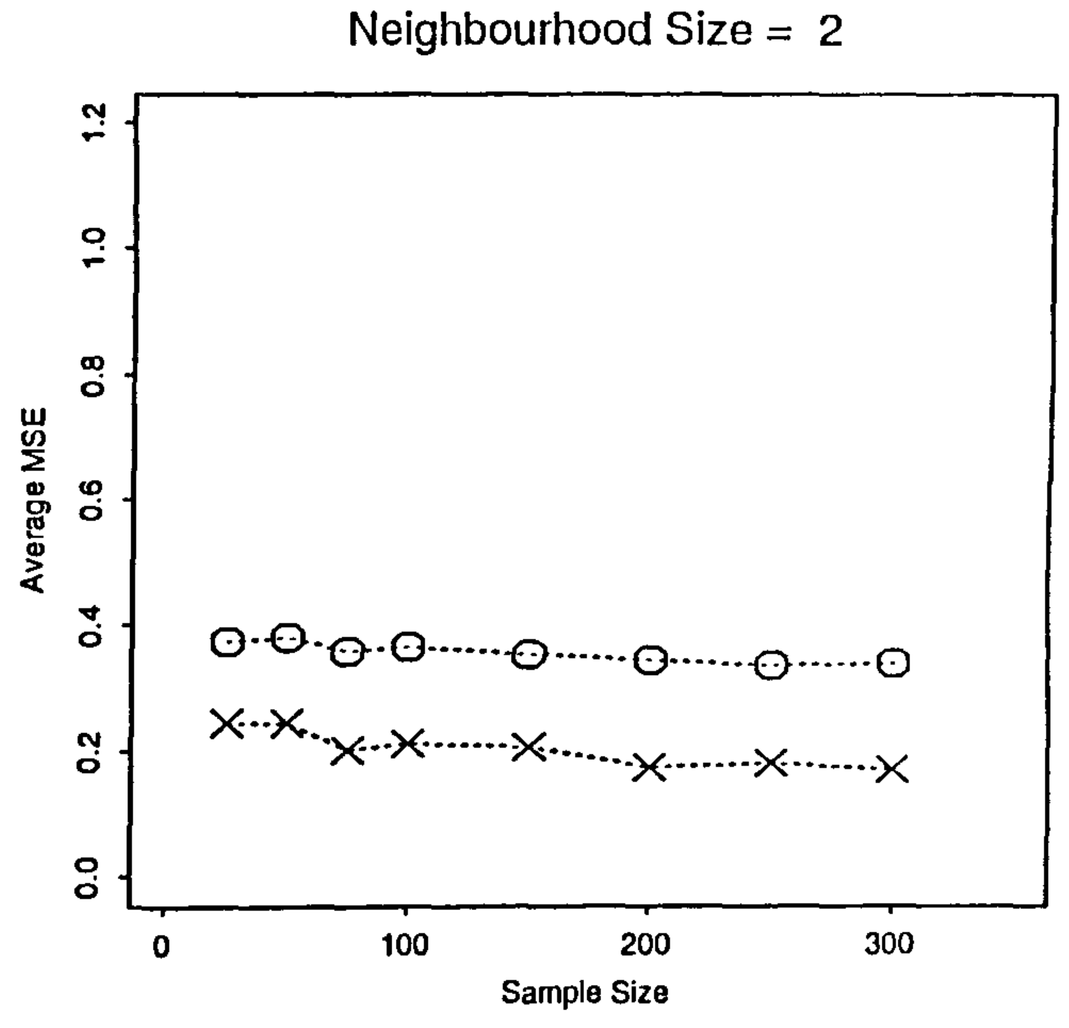
Scenario 3: **Uniform distribution for $p(z_1$ / not diseased, $z_2)$, linear Relative Risk function**

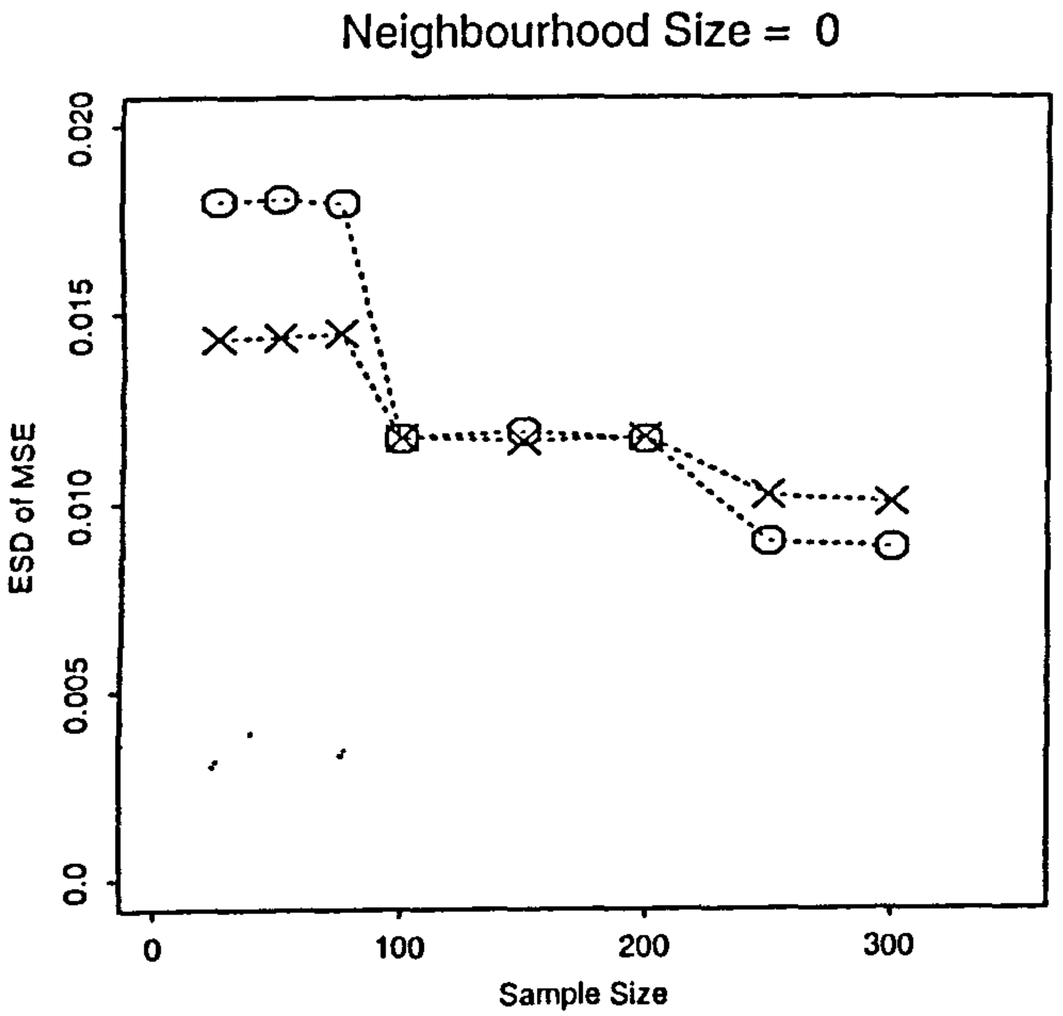
Here the Relative Risk function is of a linear nature and the underlying distribution of the risk factor among the controls is based on a Uniform distribution.

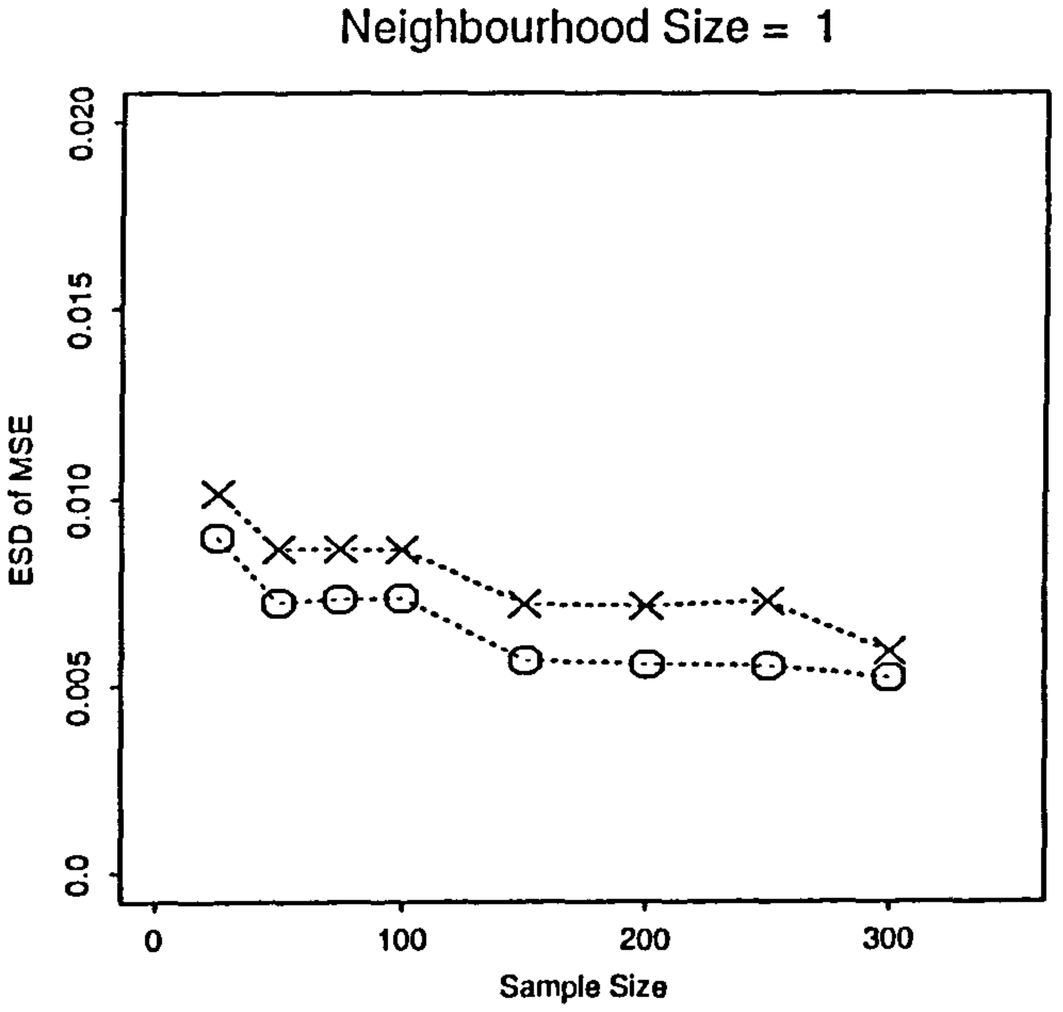
Figures 3.8.14 - 3.8.19 show the results of this set of simulations. The first major point to observe is that if Figures 3.8.14 and 3.8.15 are compared with Figures 3.8.2 and 3.8.3 it can be seen that the ***underlying distribution of the cases/controls*** appears to have ***very little effect*** on the resultant values of precision and bias. Regardless of the underlying


Neighbourhood Size = 0
Average MSE
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
100
200
300
Sample Size


Neighbourhood Size = 1
Average MSE
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
100
200
300
Sample Size


Neighbourhood Size = 2
Average MSE
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
100
200
300
Sample Size


Neighbourhood Size = 0
ESD of MSE
0.0
0.005
0.010
0.015
0.020
0
100
200
300
Sample Size


Neighbourhood Size = 1
ESD of MSE
0.0
0.005
0.010
0.015
0.020
0
100
200
300
Sample Size

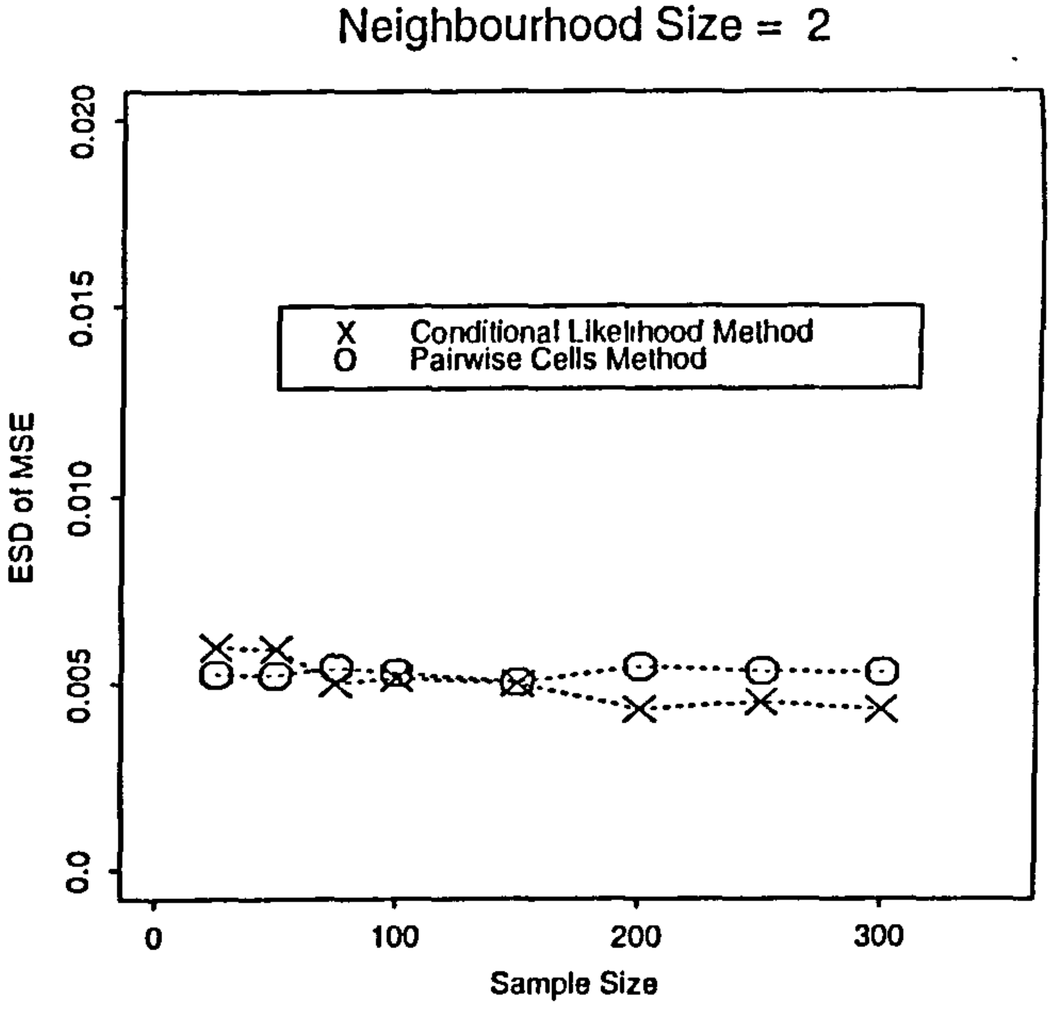

Neighbourhood Size = 2
ESD of MSE
0.0
0.005
0.010
0.015
0.020
X Conditional Likelihood Method
O Pairwise Cells Method
0
100
200
300
Sample Size

Figure 3.8.14

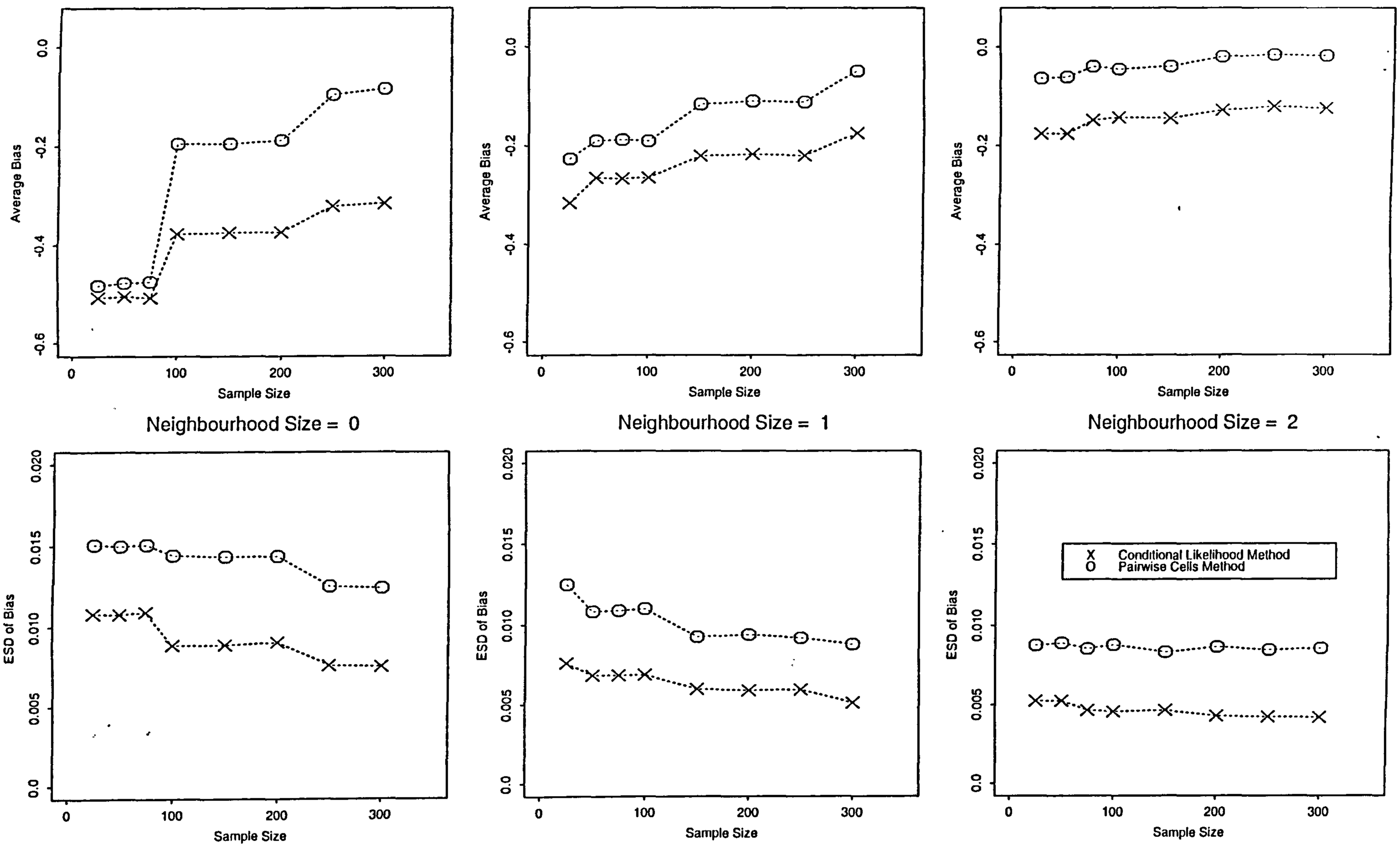

Neighbourhood Size = 0
Neighbourhood Size = 1
Neighbourhood Size = 2
Average Bias
0.0
-0.2
-0.4
-0.6
0
100
200
300
Sample Size
ESD of Bias
0.020
0.015
0.010
0.005
0.0
X Conditional Likelihood Method
O Pairwise Cells Method

Figure 3.8.15

distribution very similar results are produced both in terms of precision and bias. In this scenario frames 1 to 3 of Figure 3.8.14 reveal that the conditional likelihood method, in general, produces slightly more precise estimates than the pairwise cells method across the three levels of smoothing. In terms of the spread of the estimates of mean square error, there is little to choose between the two methods in terms of the values for the empirical standard deviation of the mean square error displayed in frames 4 to 6 of Figure 3.8.14. Regardless of the method used the estimates are less precise when small sample sizes are used in conjunction with no smoothing. There is, in general, an improvement in precision as the sample size increases although the degree of improvement appears ***more noticeable*** when there is no smoothing. In terms of bias, an examination of Figure 3.8.15 reveals that the pairwise cells method appears to be less biased than the conditional likelihood method. The only exception to this appears to be with no smoothing and sample sizes of 75 observations or less, where the bias is almost identical for both methods. For sample sizes of more than 75 observations and no smoothing there is quite a large difference in bias between the two methods of estimation, a difference which, to a certain extent, reduces with the introduction of smoothing. The presence of bias is, again, greatest when the combination of small sample sizes and no smoothing is used. There is a ***marked decrease*** in bias with increasing sample size and when moving from no smoothing to a neighbourhood of size 1. As in scenarios 1 and 2, one worrying point to observe is the nature of the bias. Regardless of sample size and level of smoothing the resultant estimates are always underestimates of the true Relative Risk. However with large data sets and/or smoothing the degree of underestimation is not especially large.

Moving on to consider coverage and average width of the nominal 95% confidence intervals Figures 3.8.16 to 3.8.19 reveal a very similar pattern in terms of coverage and

average width to that observed in scenario 1. Again the coverage seems to be unrealistically high when no smoothing is present, particularly for smaller sample sizes. The use of smoothing, particularly a neighbourhood of size 1, leads, on the whole, to more reasonable levels of coverage and slightly narrower confidence intervals. This is particularly evident for larger sample sizes. Regardless of the method used, when consideration is given to values of the risk factor between 0 and approximately 10 then, with smoothing, the coverage is better for ***lower*** values of the risk factor. For these values the average width of the confidence intervals also increases as the value of the risk factor moves away from the baseline and into areas where less data is present. However, with smoothing and ***larger*** values of the risk factor (i.e. between 11 and 19) there is evidence that the coverage ***increases*** as the value of the risk factor increases. This is particularly noticeable for larger sample sizes. This is possibly due to the sharp increase in the corresponding width of the confidence intervals over the range 11 to 19 (see Figures 3.1.17 and 3.1.19) again leading to, perhaps, unrealistic levels of coverage. One, perhaps surprising, observation from these figures is the pattern of the width of the confidence intervals for the pairwise cells method (Figure 3.8.19). Here there is evidence of something unusual happening when no smoothing is used in combination with smaller sample sizes (i.e. 75 observations or less). With these smaller sample sizes, it appears that, on average, the width of the confidence intervals ***increases*** as the level of smoothing increases. One possible explanation for this may be that these sample sizes are simply too small to obtain a true representation of this scenario, as, when larger sample sizes are used the plots begin to exhibit similar patterns to those observed elsewhere in this section.

Conditional Likelihood Method - Linear Relative Risk

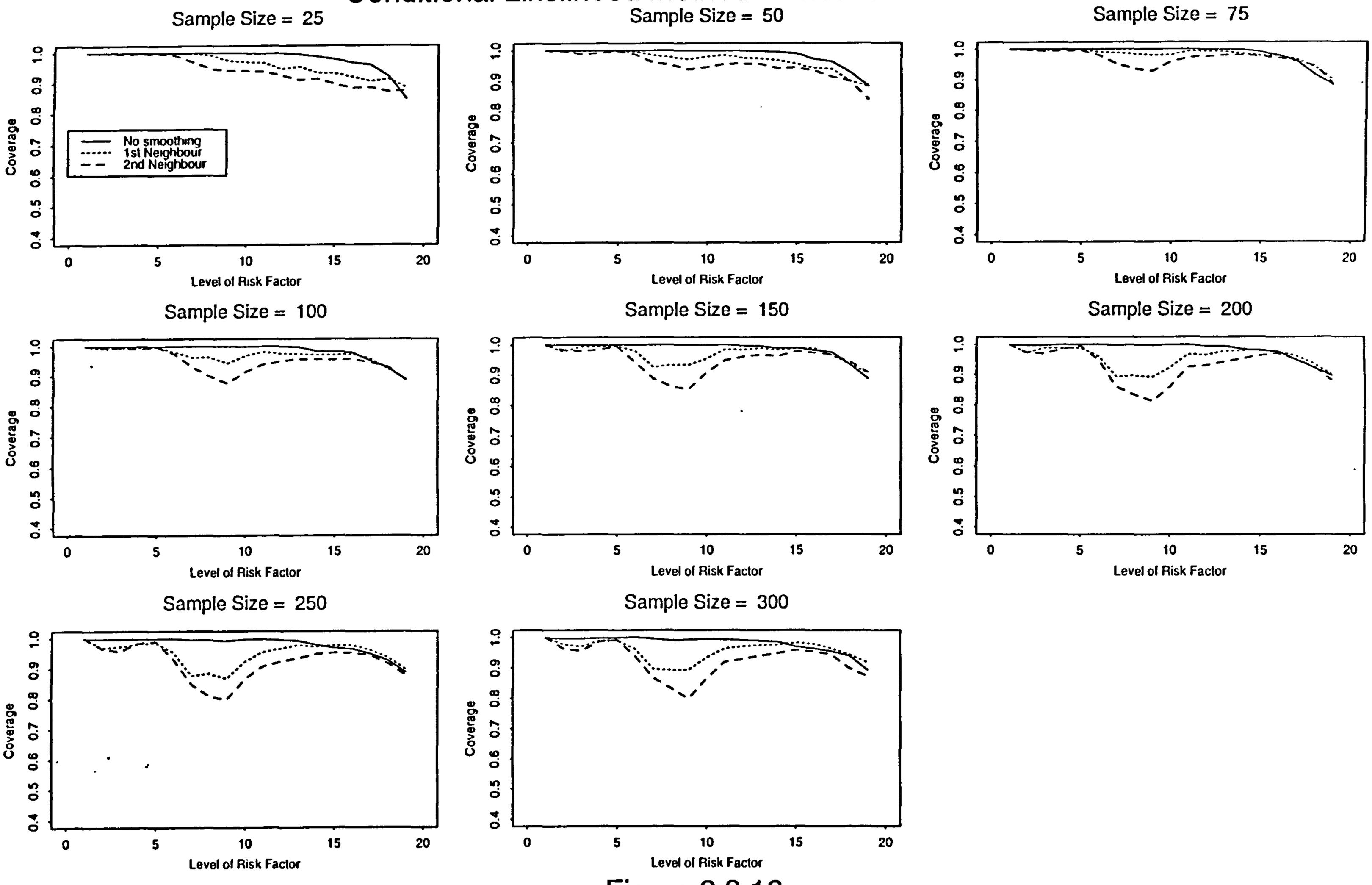
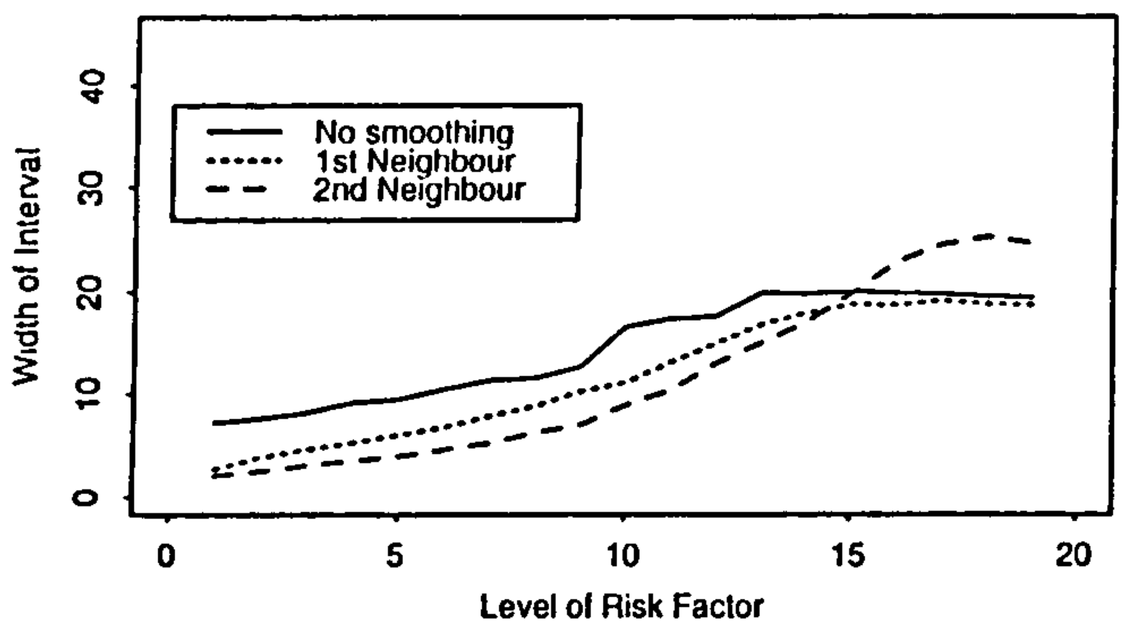


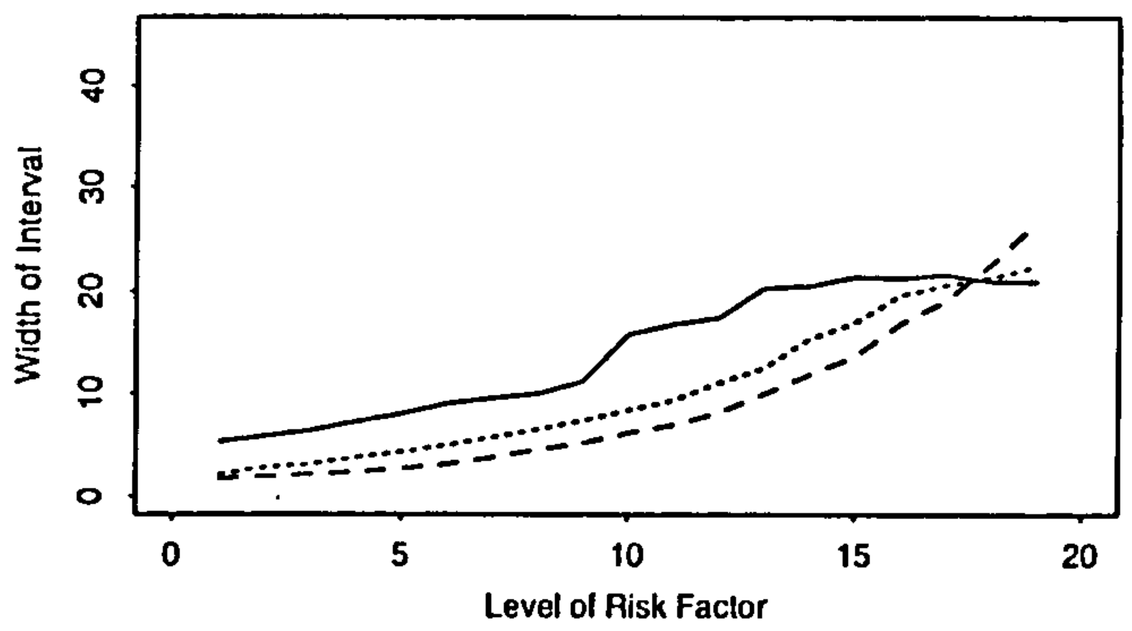
Figure 3.8.16

Conditional Likelihood Method - Linear Relative Risk

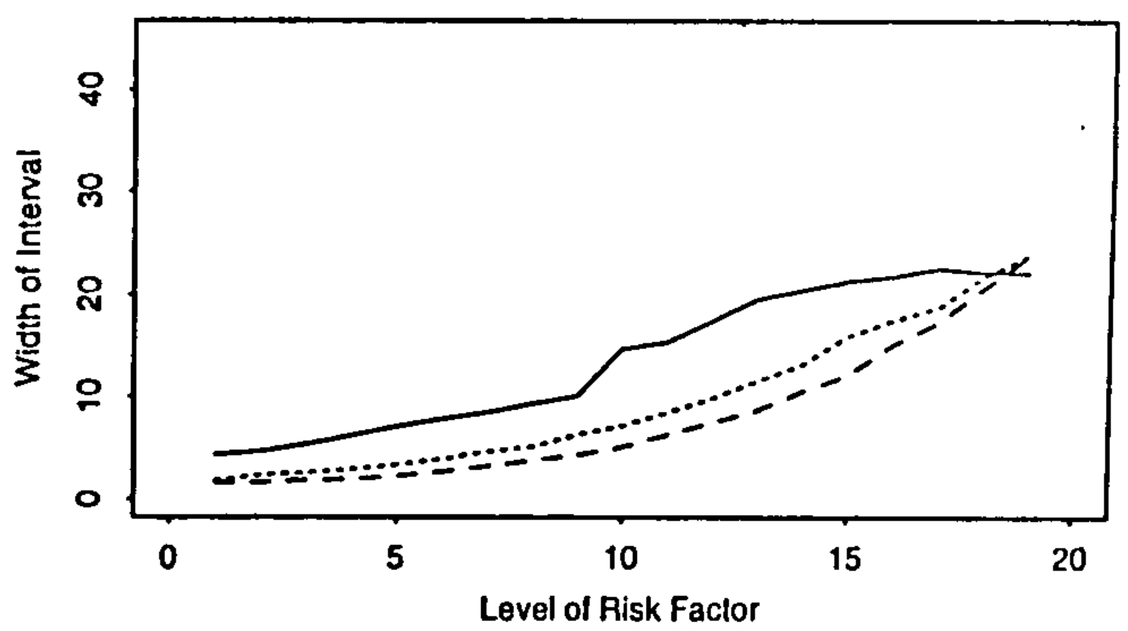
Sample Size = 25



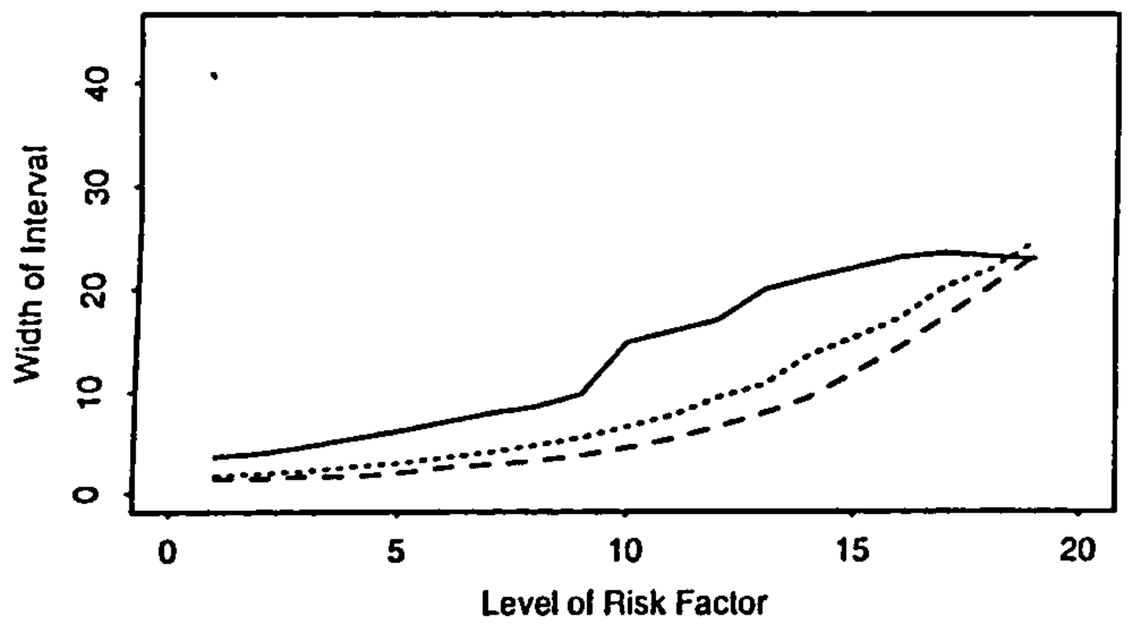
Sample Size = 50



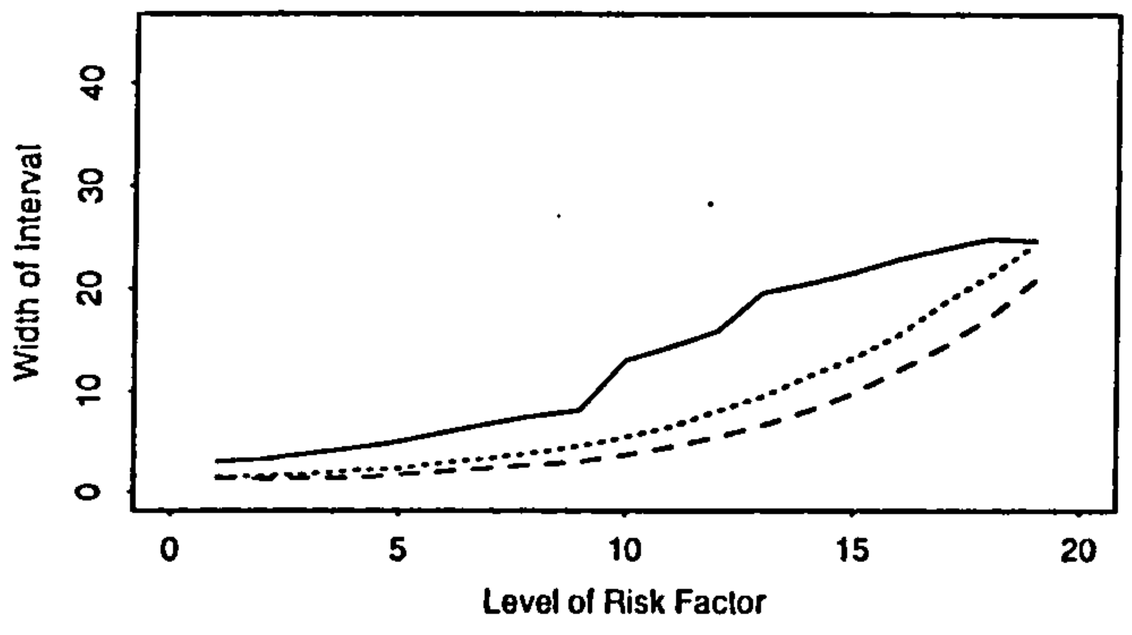
Sample Size = 75



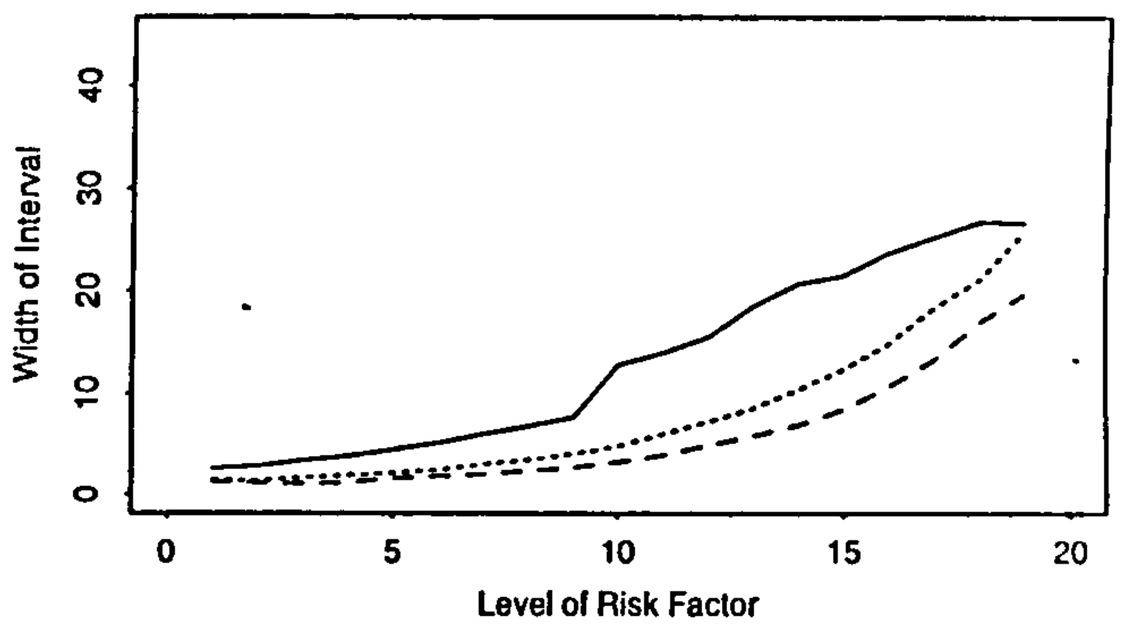
Sample Size = 100



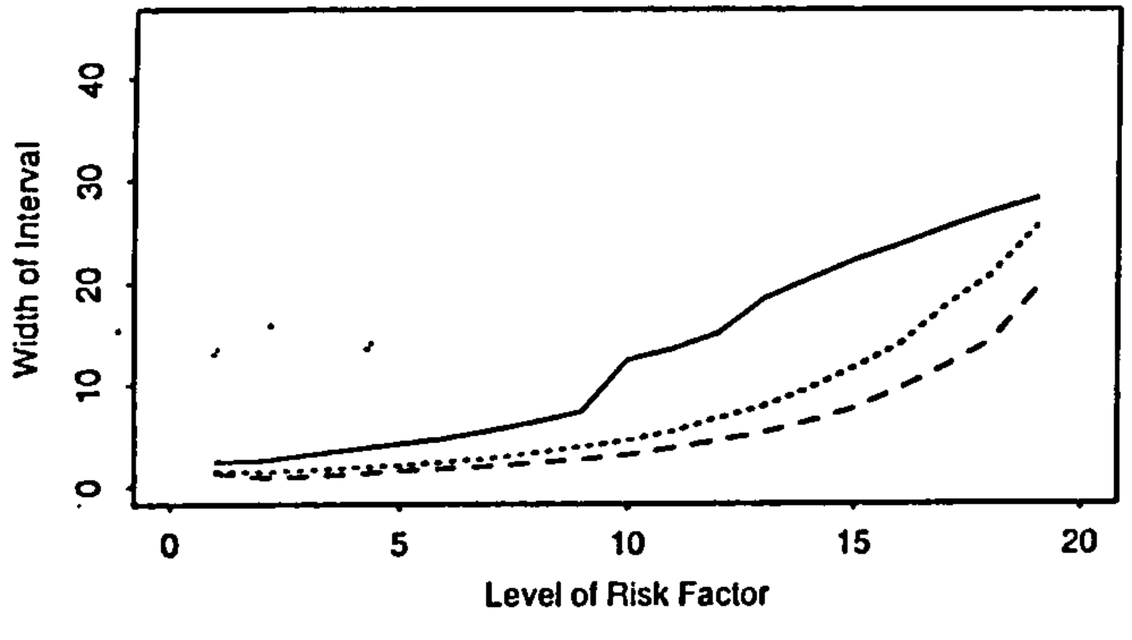
Sample Size = 150



Sample Size = 200



Sample Size = 250



Sample Size = 300

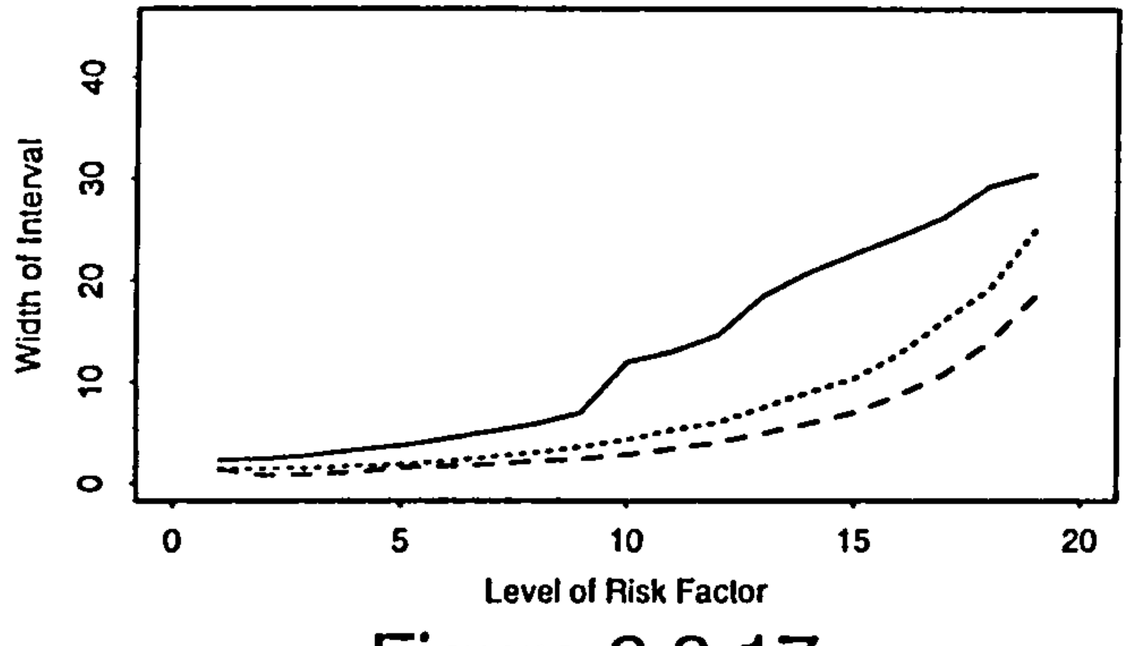


Figure 3.8.17

Pairwise cells Method - Linear Relative Risk

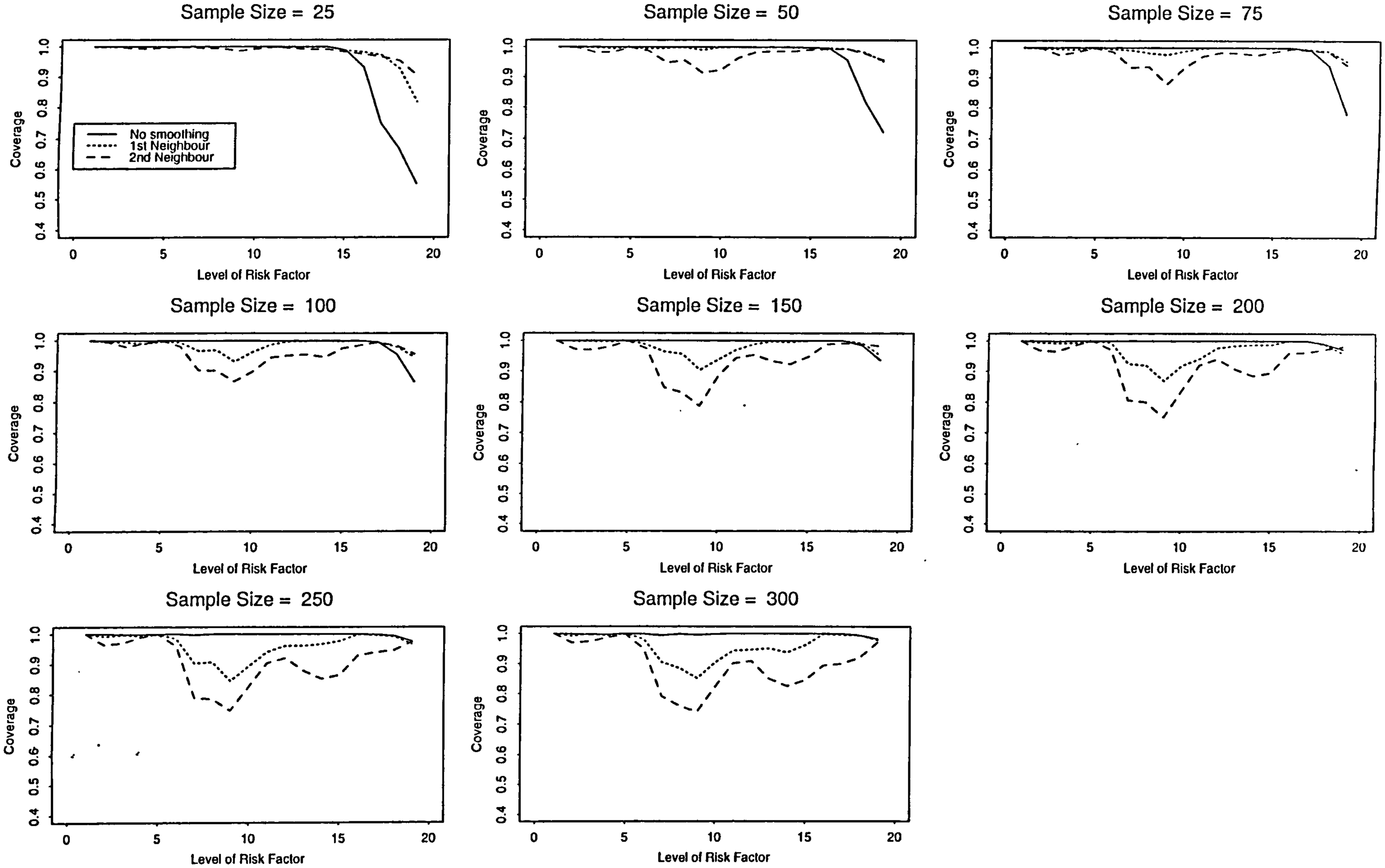


Figure 3.8.18

Pairwise Cells Method - Linear Relative Risk

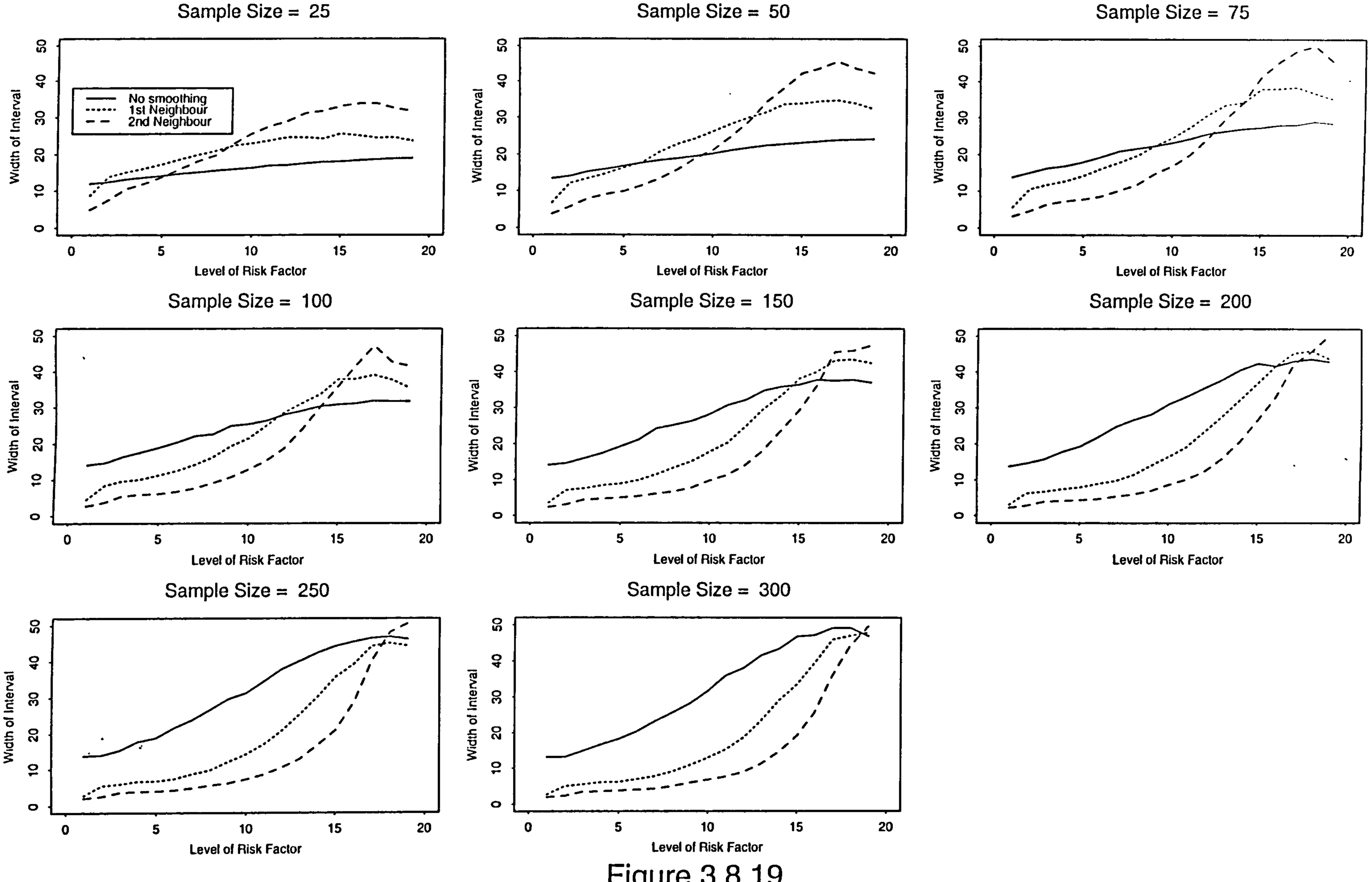


Figure 3.8.19

In summary, there is very little to choose between the two methods in terms of precision and bias in this scenario. With smaller sample sizes and no smoothing both methods do not appear very precise and exhibit some bias. However it can be seen that both methods produce reasonable estimates of the Relative Risk when larger sample sizes are used and/or smoothing is introduced. Examination of the coverage and corresponding width of the confidence intervals reveals that reasonable sample sizes or smoothing ***must*** be used before sensible estimates of Relative Risk can be obtained in this scenario. This is particularly evident when the pairwise cells method is used.

Scenario 4: **Uniform distribution for $p(z_1$ / not diseased, $z_2)$, a step Relative Risk function.**

The final scenario under consideration is again based on an underlying Uniform distribution for the controls, as in scenario 3, but on this occasion, instead of a being of a linear nature, the Relative Risk function exhibits one large step.

Figures 3.8.20 - 3.8.25 show the resultant plots for this set of simulations. A comparison of these with Figures 3.8.8 - 3.8.13 which were obtained in scenario 2 suggest that the underlying distribution of the cases/controls again has little effect on the results which are produced confirming the observation made during scenario 3. There is again evidence however that the nature of the Relative Risk (i.e. a linear Relative Risk or a step Relative Risk) has some effect on the results which are produced. In scenarios 1 and 2 when the underlying distributions of the cases/controls were based on a Poisson distribution there was some evidence that the estimates produced were moderately more

precise and slightly less biased when the Relative Risk had a single, large, step compared to a Relative Risk of a linear nature. A comparison of Figures 3.8.20 and 3.8.21 with Figures 3.8.14 and 3.8.15 reveals that when the underlying distribution of the cases/controls is based on a Uniform distribution the estimates are again more precise/less biased when the relative risk function is of a step nature. In terms of this specific scenario, Figure 3.8.20 reveals that both methods produce reasonably precise estimates with precision increasing with both sample size and neighbourhood size. In this scenario, the change in precision as the sample sizes increase is not as dramatic as the change observed in the first three scenarios. This is possibly due to estimation being, in general, more precise in this scenario, regardless of sample size. For all sample sizes and all levels of smoothing, the conditional likelihood method appears marginally more precise with little difference between the two methods in terms of the empirical standard deviation of the mean square error. The levels of bias present in Figure 3.8.21 suggest that, with smoothing, both methods exhibit relatively small amounts of bias in this scenario. In the absence of smoothing the bias is slightly higher and decreases with increasing sample size but the introduction of smoothing leads to the sample size having little effect on the level of bias. Also the pairwise cells method appears, in general, to be less biased particularly when there is no smoothing and larger sample sizes.

An examination of the coverage and average width of the nominal 95% confidence intervals for both methods (Figures 3.8.22 - 3.8.25) suggests that the levels of coverage are, once more, unrealistically high when no smoothing is used. This conclusion is especially evident for smaller sample sizes. For both methods, the levels of coverage reduce to values which are closer to the nominal value once a neighbourhood of size 1 is considered with a corresponding decrease in the average width of the 95% confidence

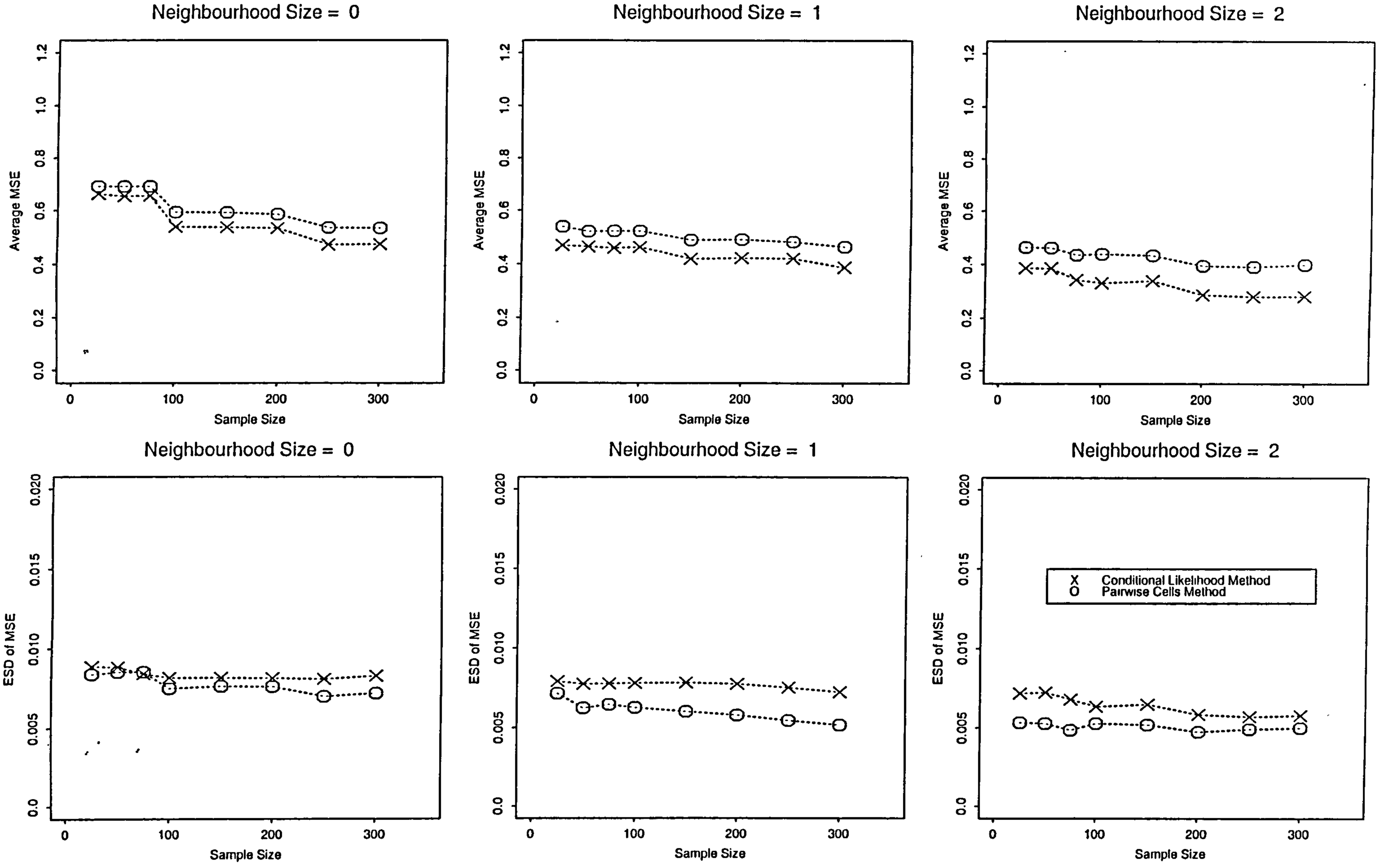
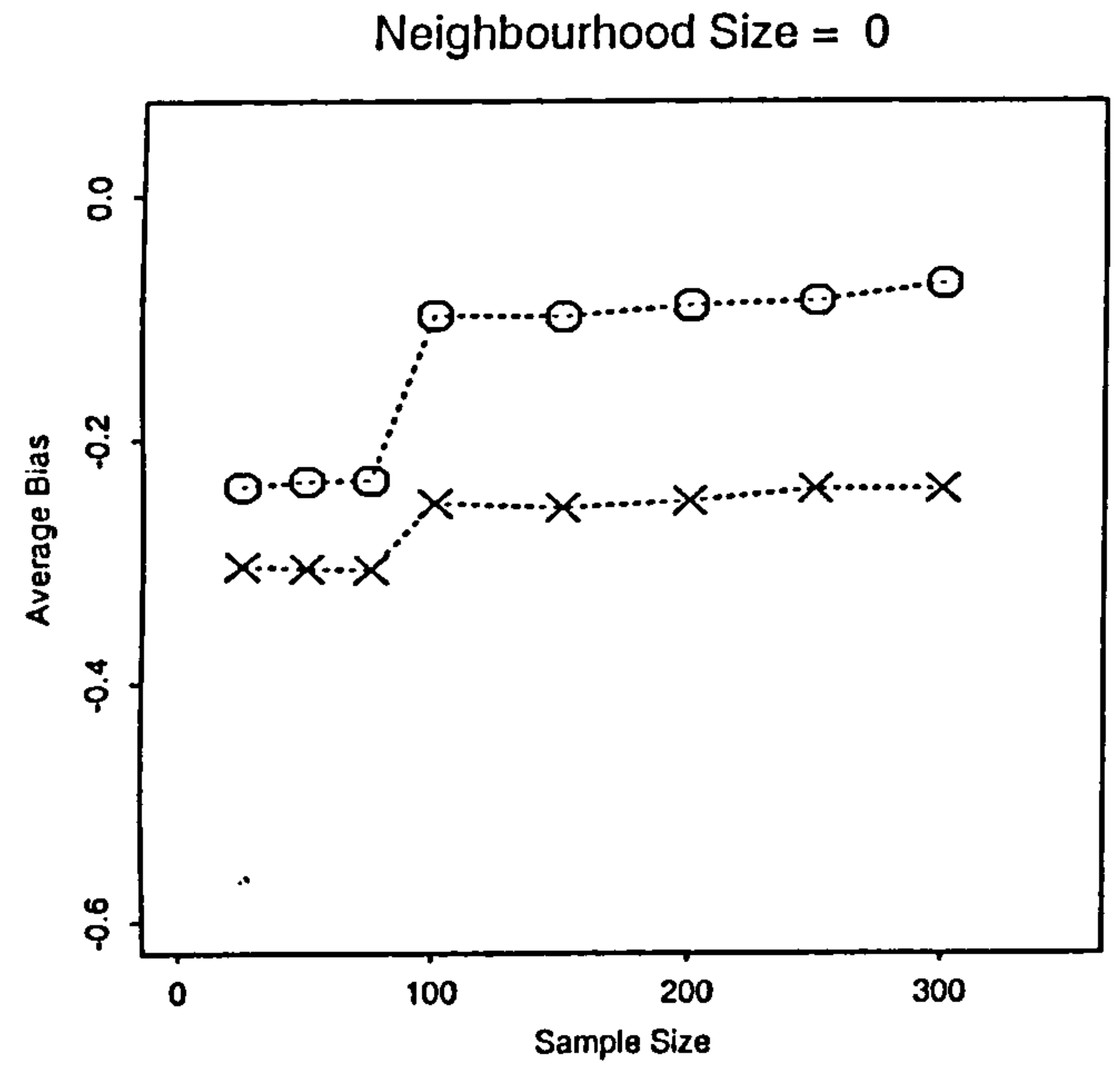
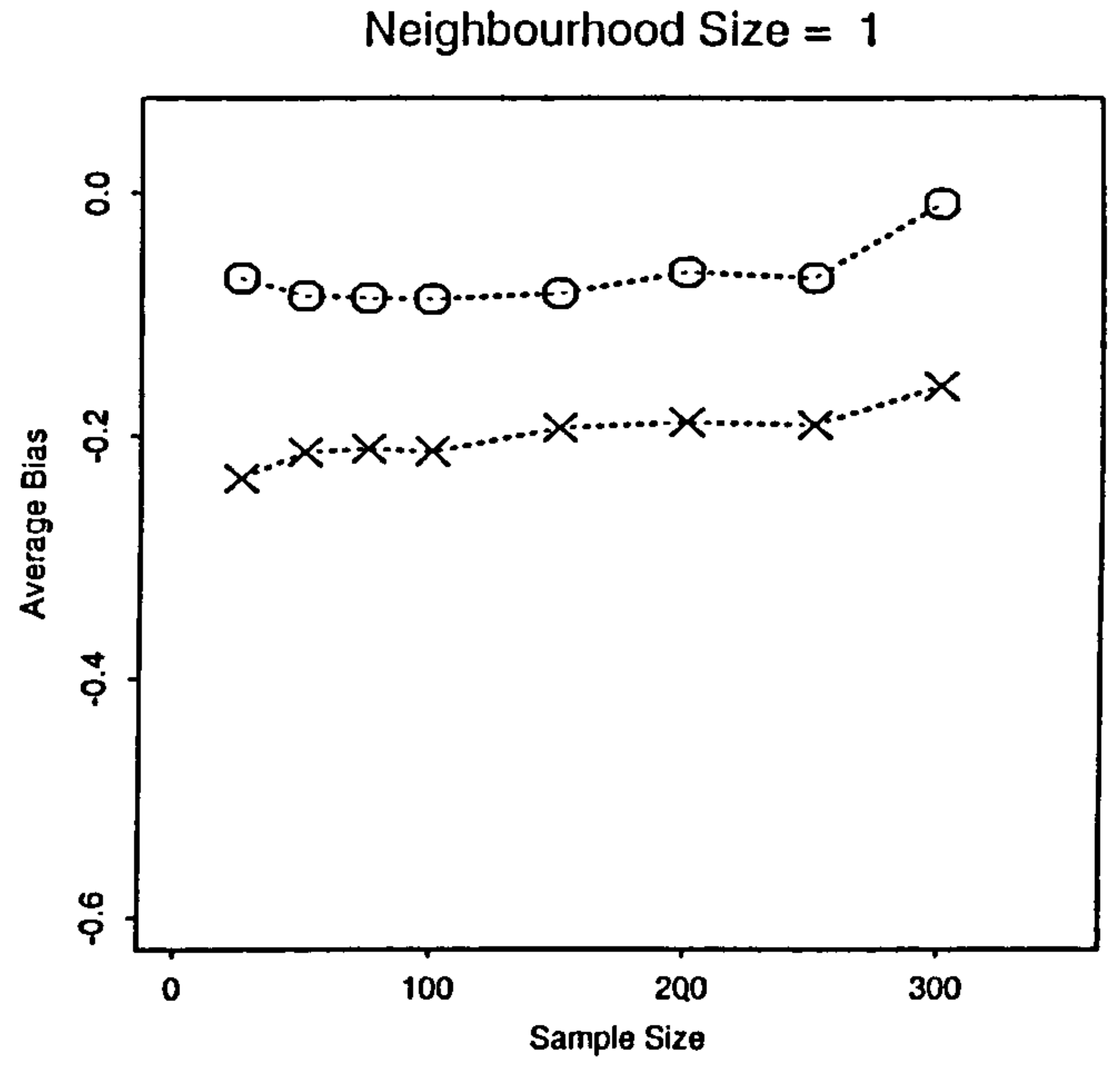
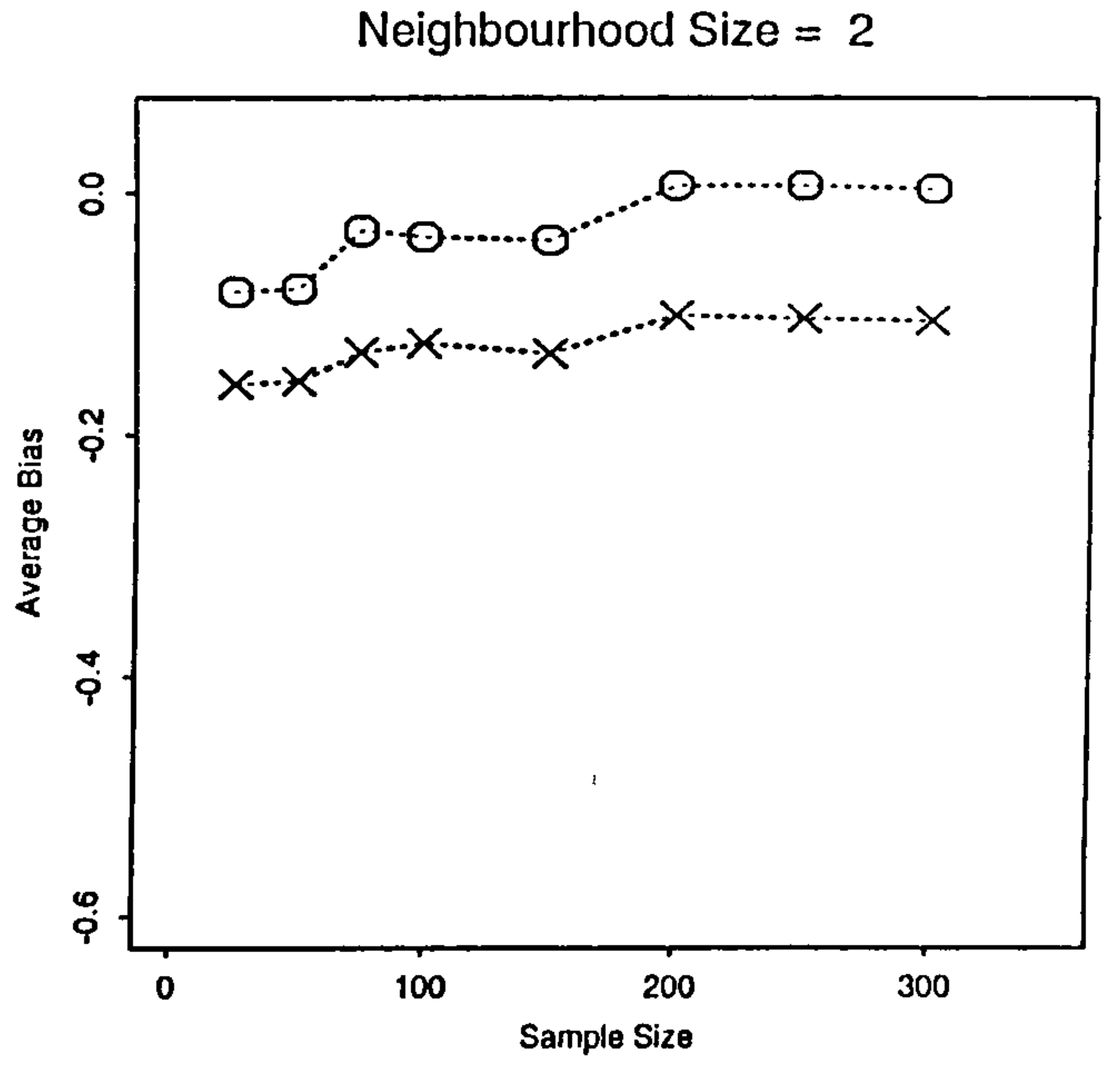
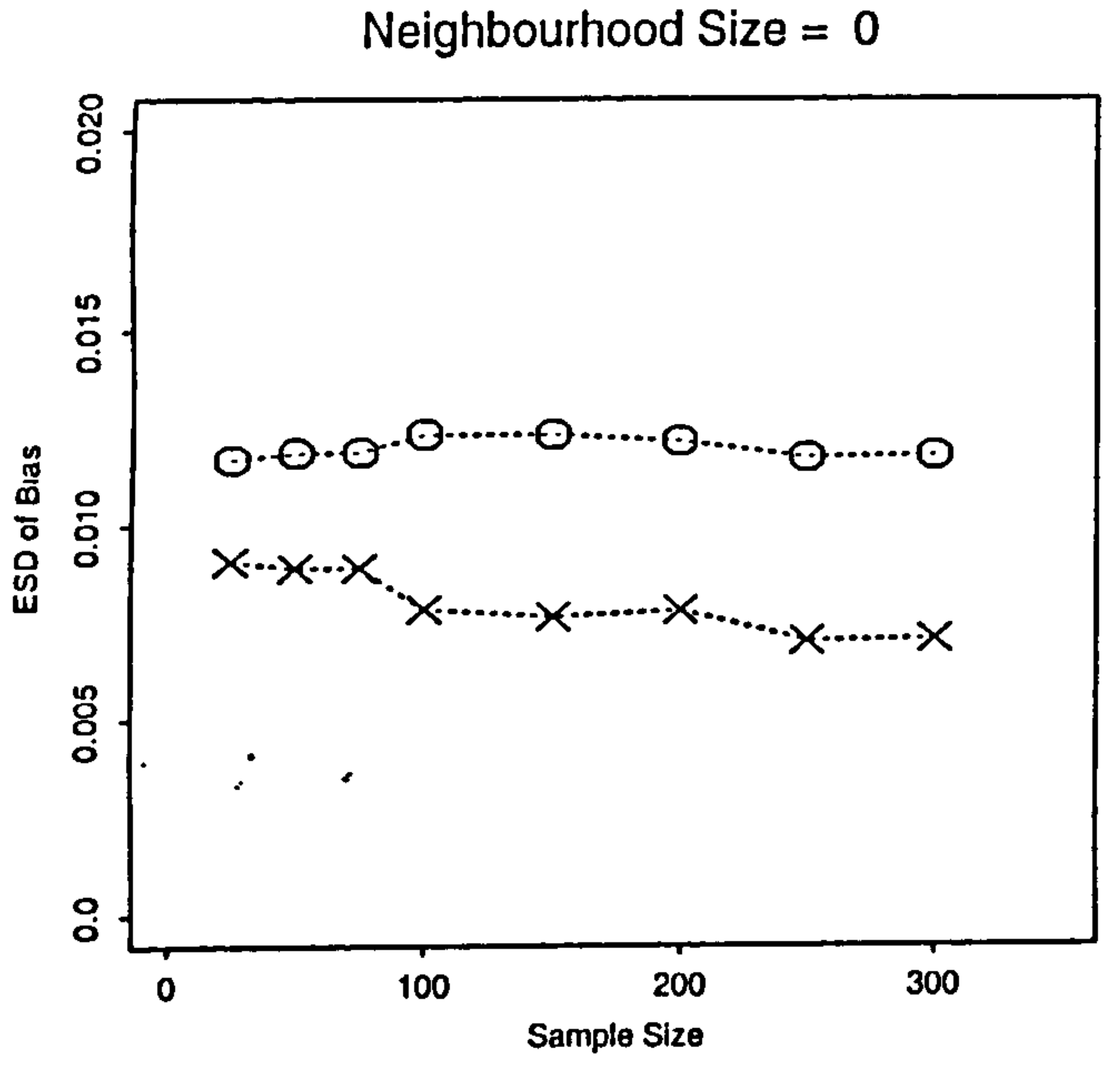

Neighbourhood Size = 0
Neighbourhood Size = 1
Neighbourhood Size = 2
Average MSE
ESD of MSE
Sample Size
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.005
0.010
0.015
0.020
0
100
200
300
X Conditional Likelihood Method
O Pairwise Cells Method

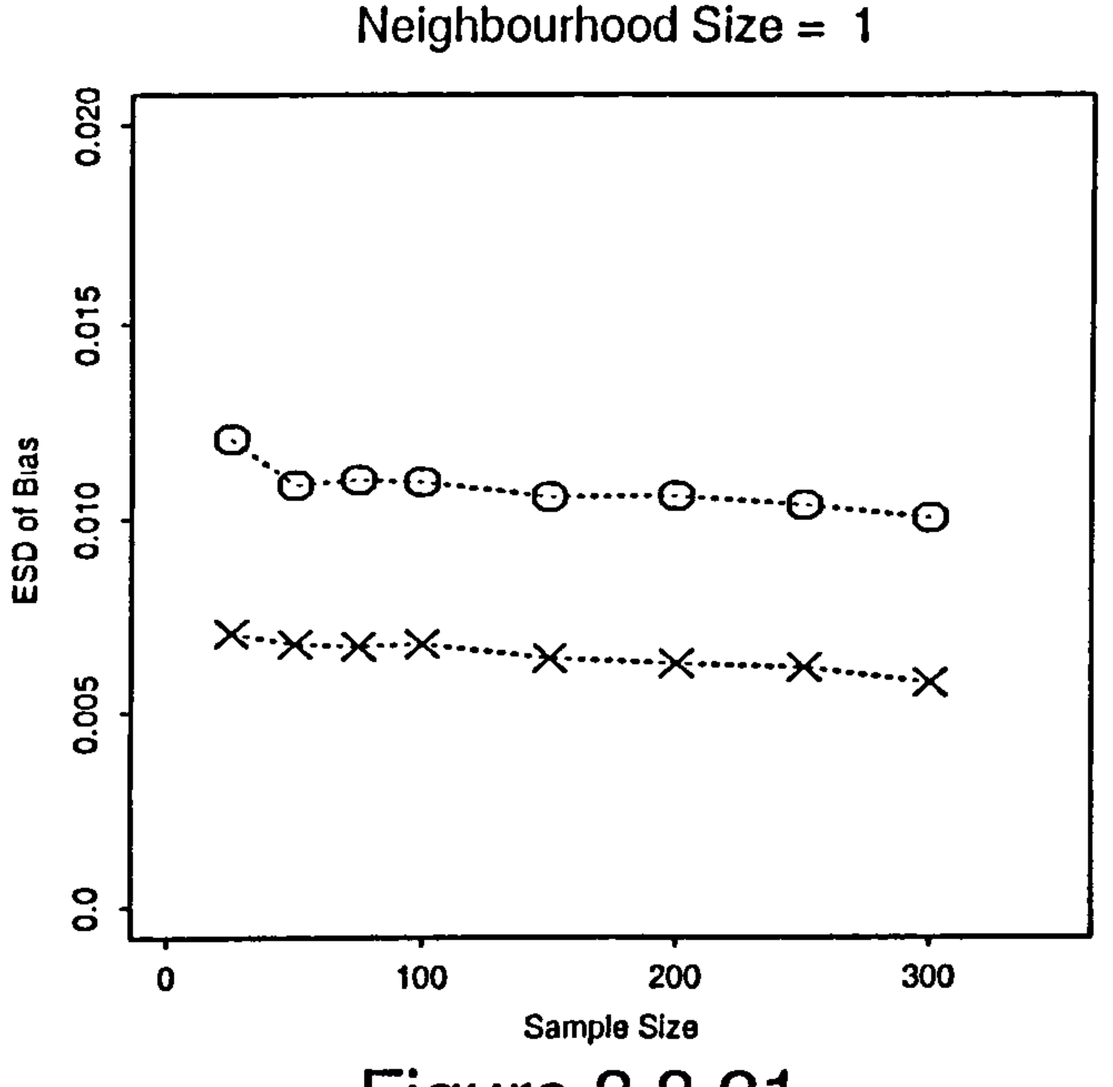
Figure 3.8.20


Neighbourhood Size = 0
Average Bias
0.0
-0.2
-0.4
-0.6
0
100
200
300
Sample Size


Neighbourhood Size = 1
Average Bias
0.0
-0.2
-0.4
-0.6
0
100
200
300
Sample Size


Neighbourhood Size = 2
Average Bias
0.0
-0.2
-0.4
-0.6
0
100
200
300
Sample Size


Neighbourhood Size = 0
ESD of Bias
0.020
0.015
0.010
0.005
0.0
0
100
200
300
Sample Size


Neighbourhood Size = 1
ESD of Bias
0.020
0.015
0.010
0.005
0.0
0
100
200
300
Sample Size

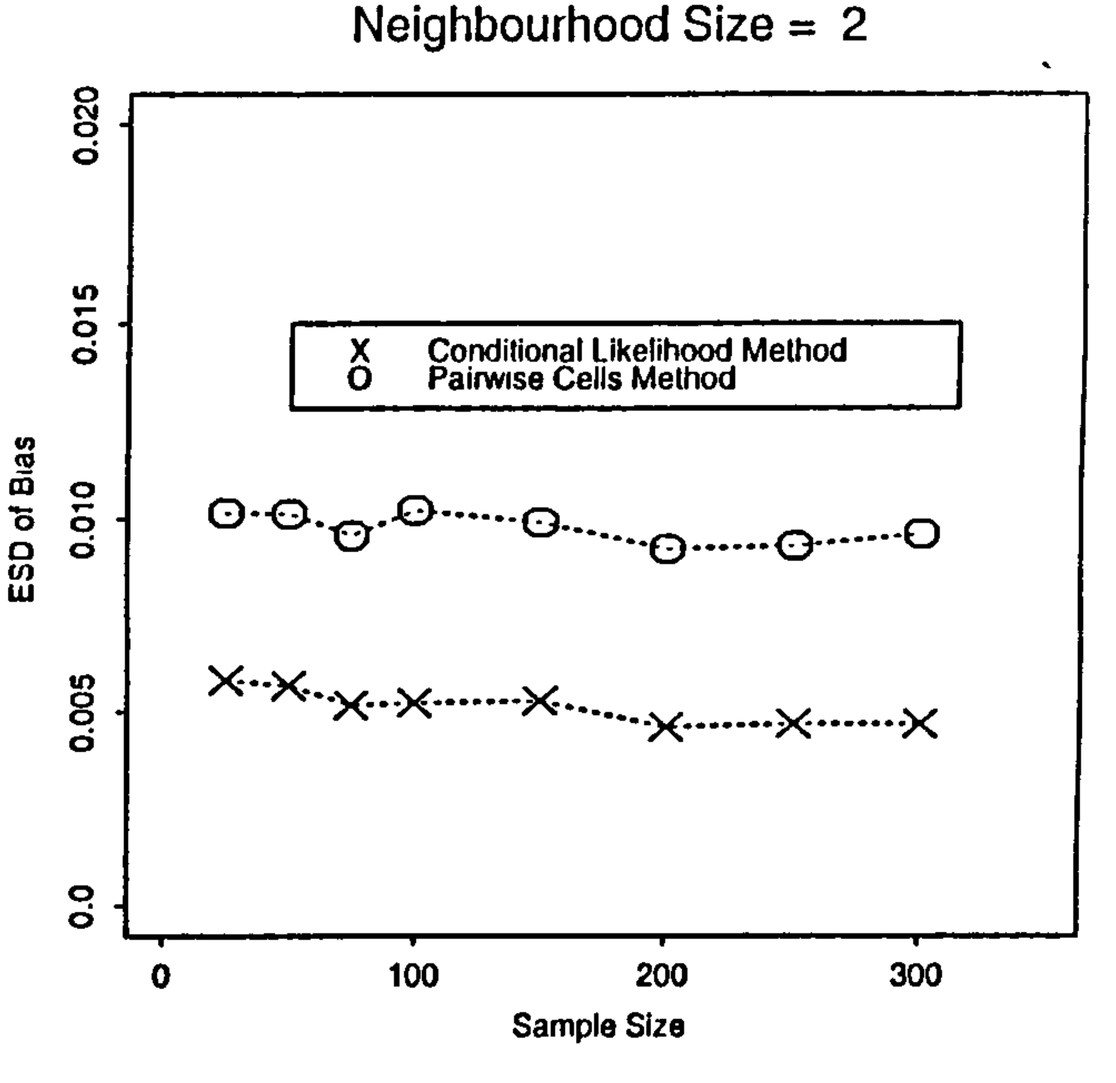

Neighbourhood Size = 2
ESD of Bias
0.020
0.015
0.010
0.005
0.0
X Conditional Likelihood Method
O Pairwise Cells Method
0
100
200
300
Sample Size

Figure 3.8.21

Conditional Likelihood Method - Step Relative Risk

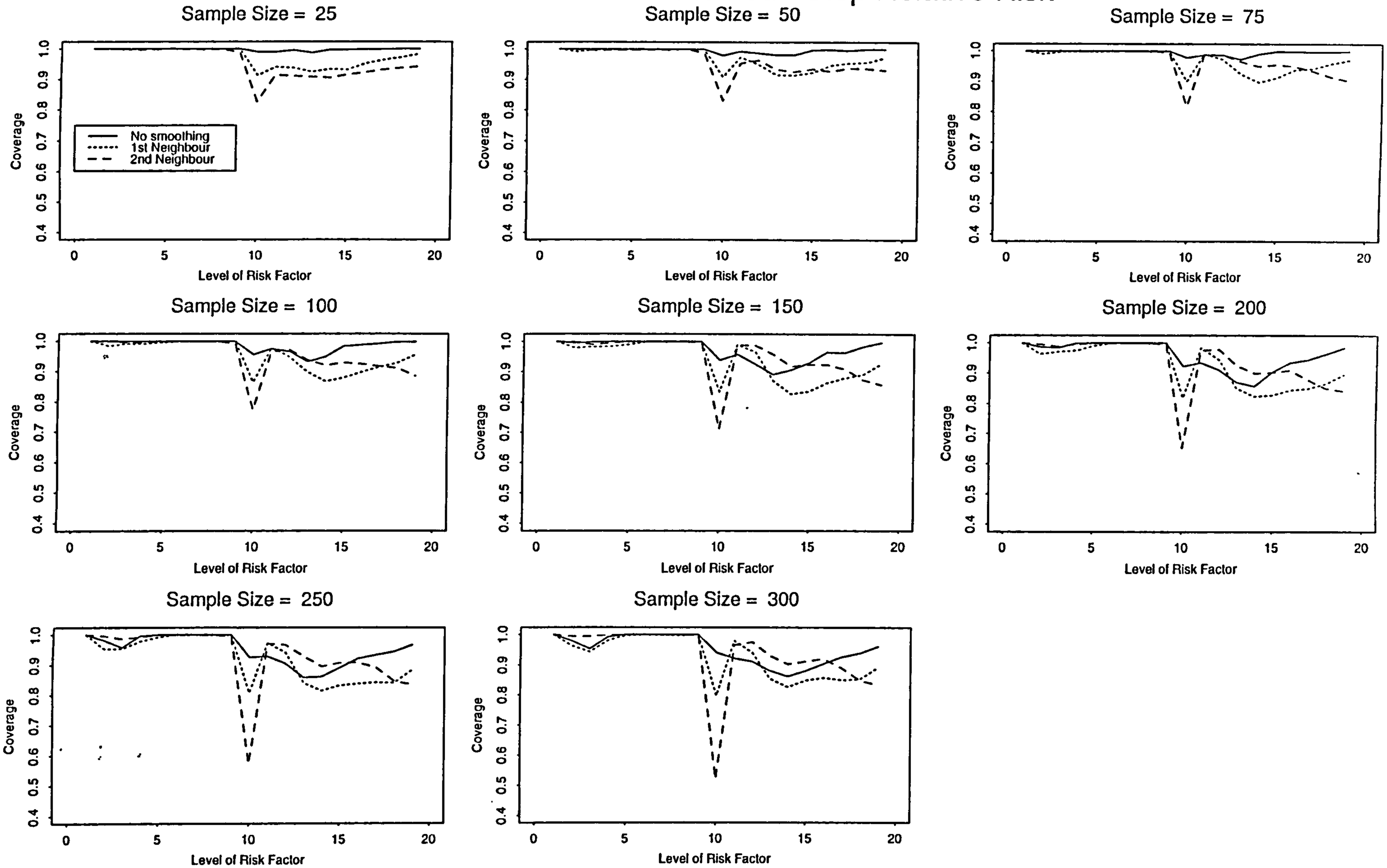


Figure 3.8.22

Conditional Likelihood Method - Step Relative Risk

Figure 3.8.23

Pairwise Cells Method - Step Relative Risk

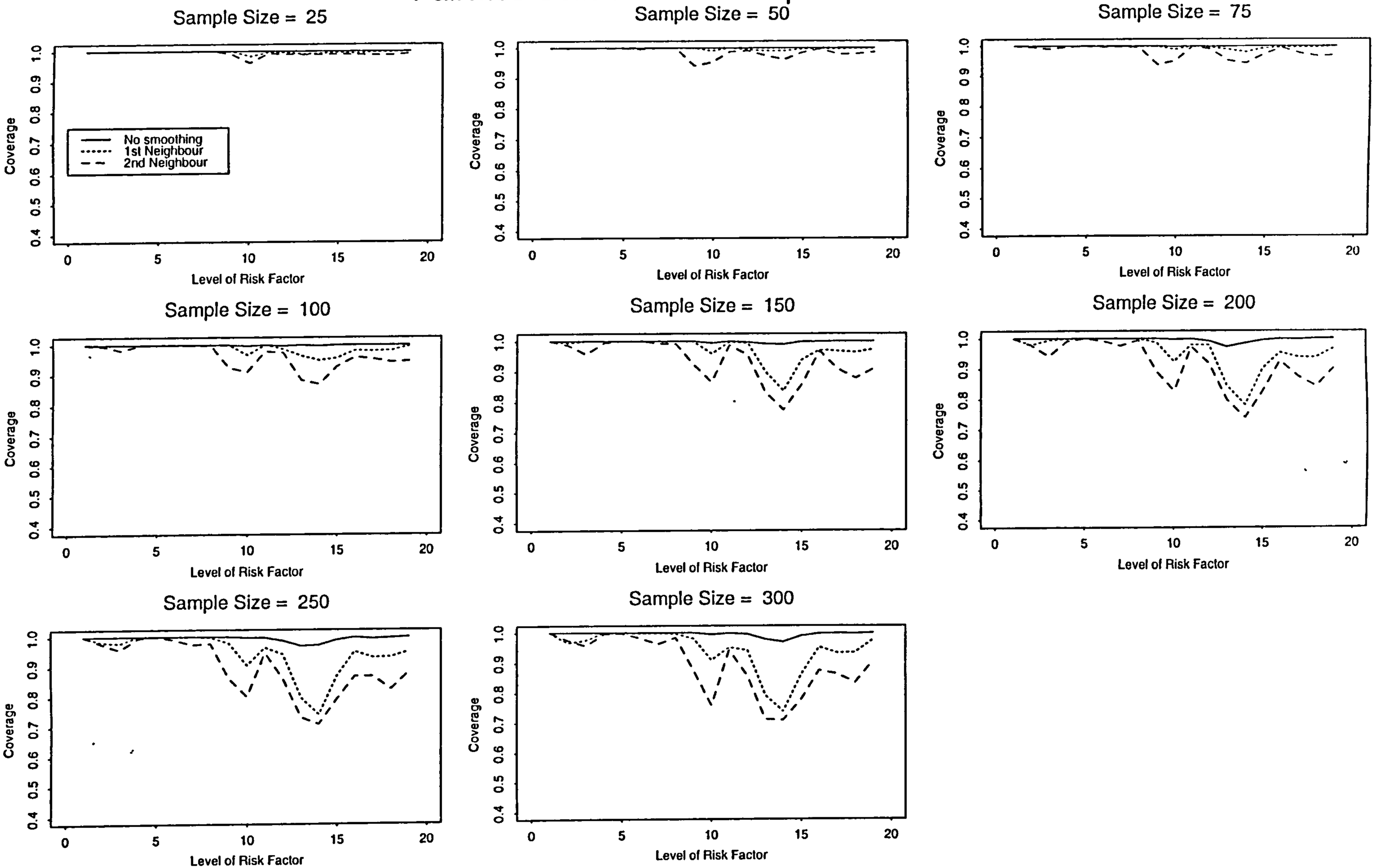
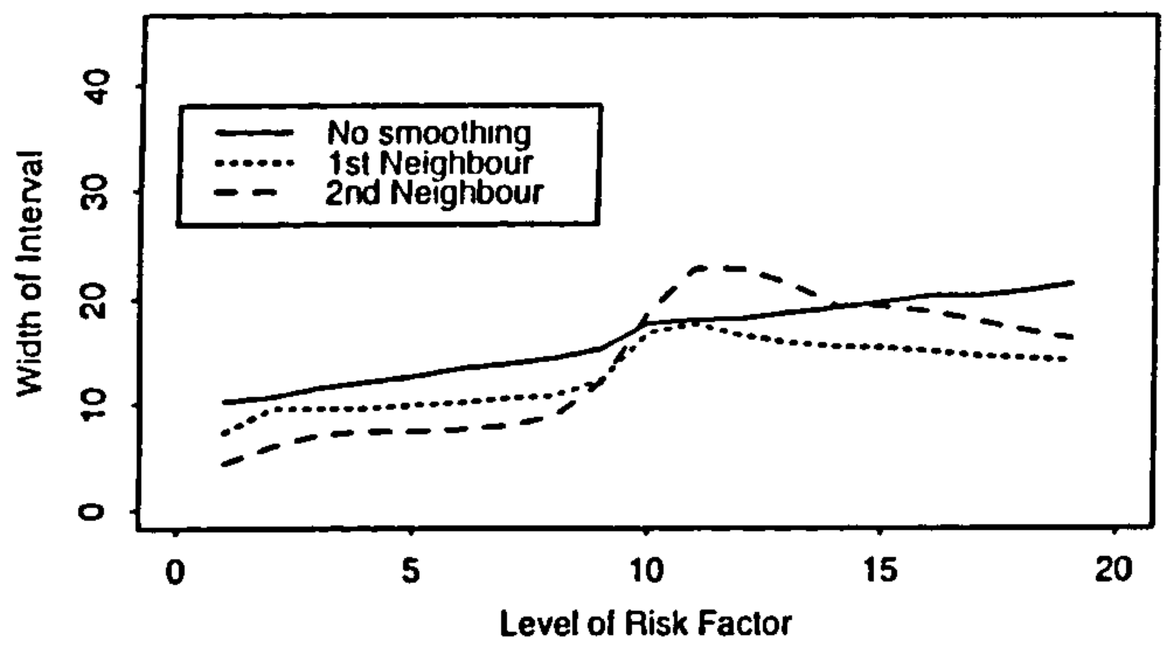


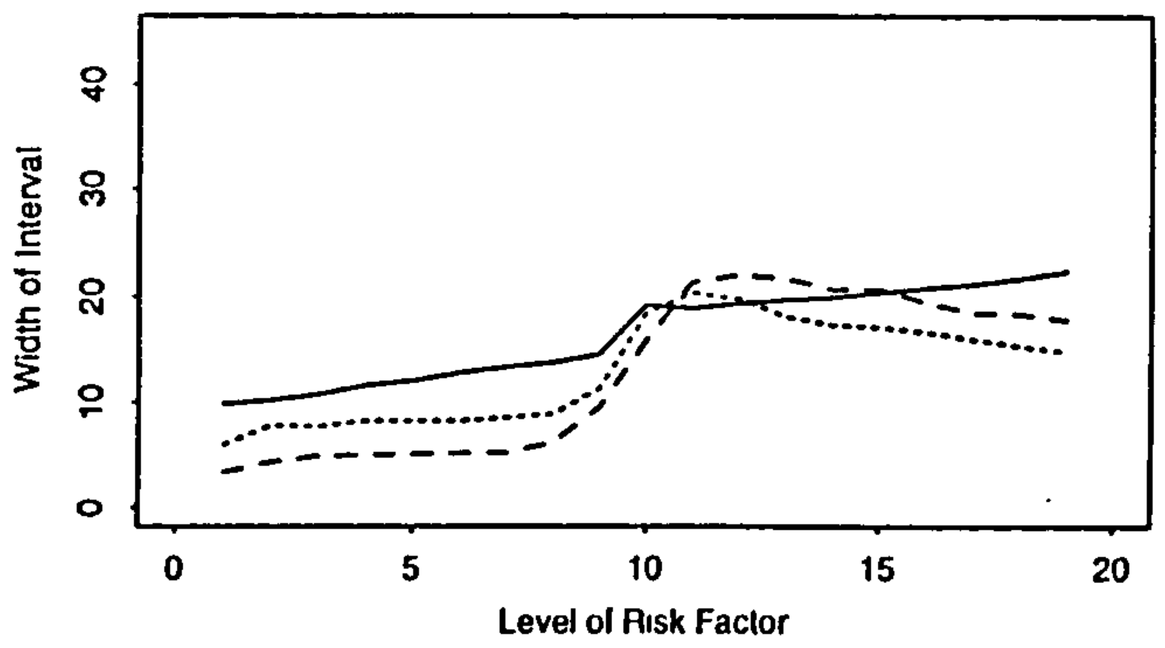
Figure 3.8.24

Pairwise Cells Method - Step Relative Risk

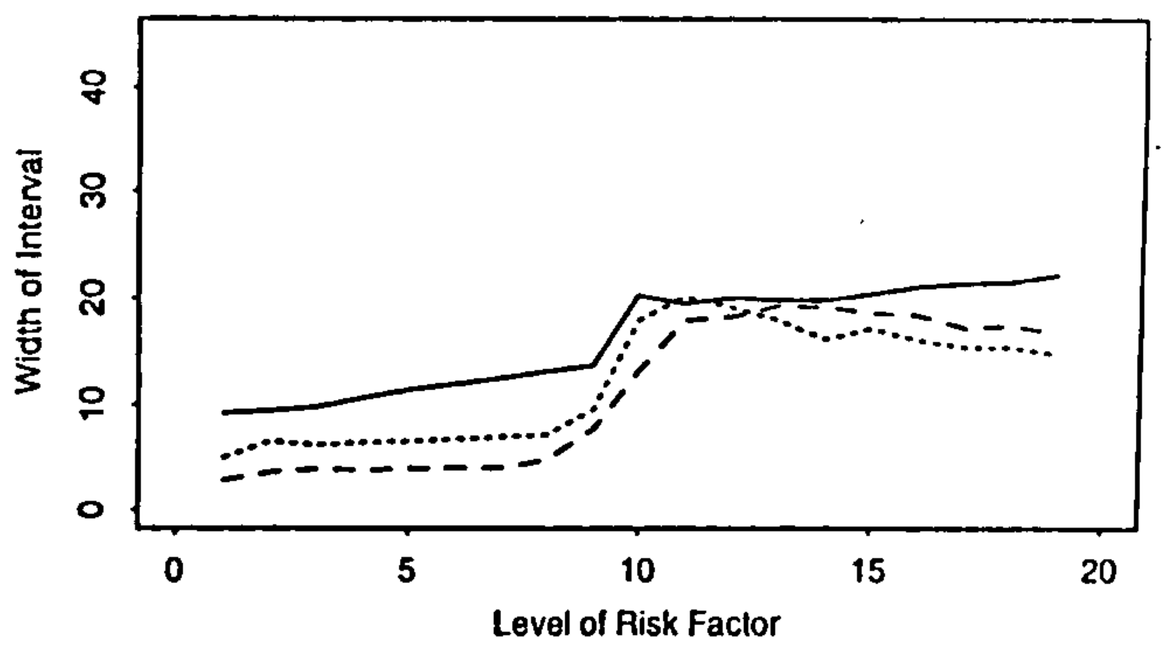
Sample Size = 25



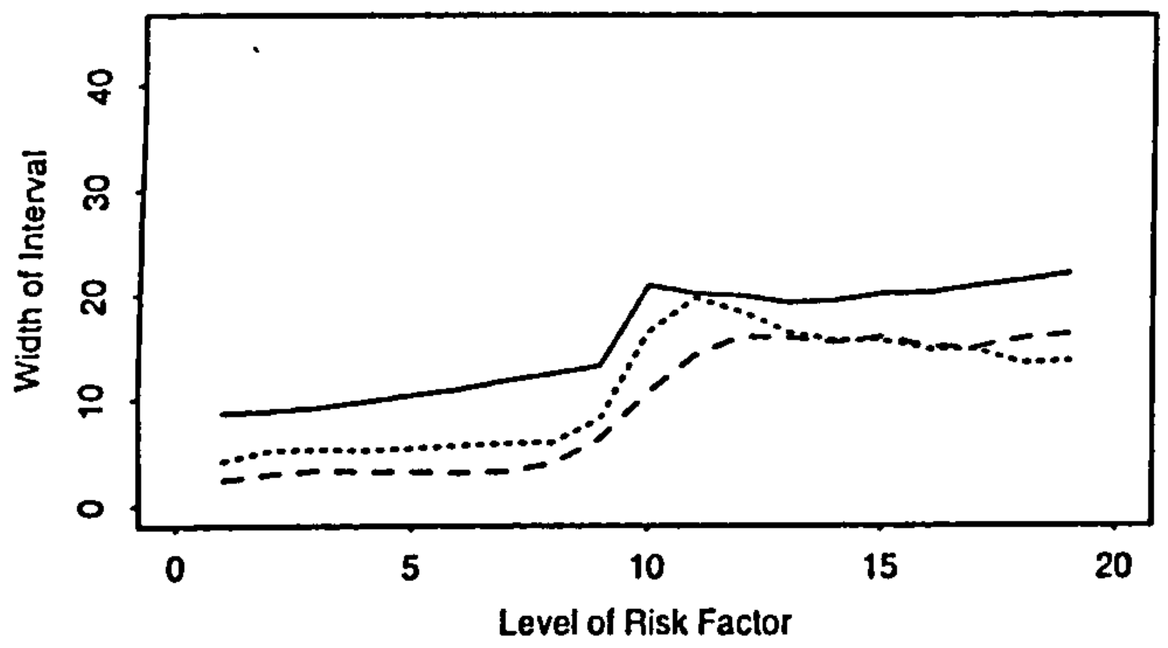
Sample Size = 50



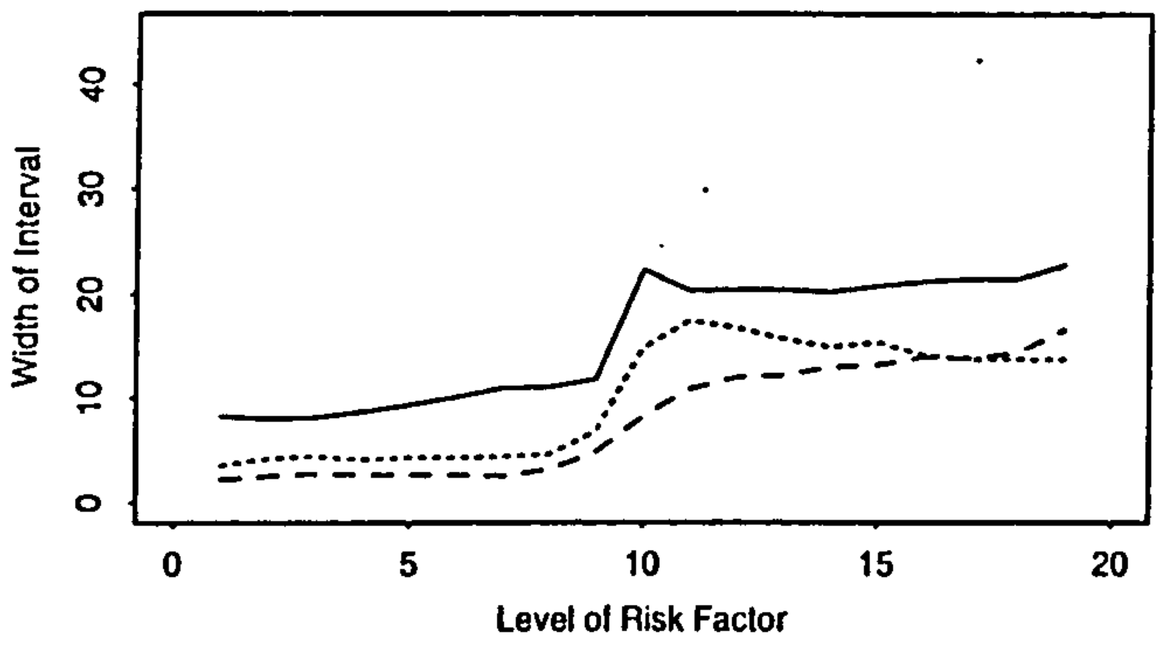
Sample Size = 75



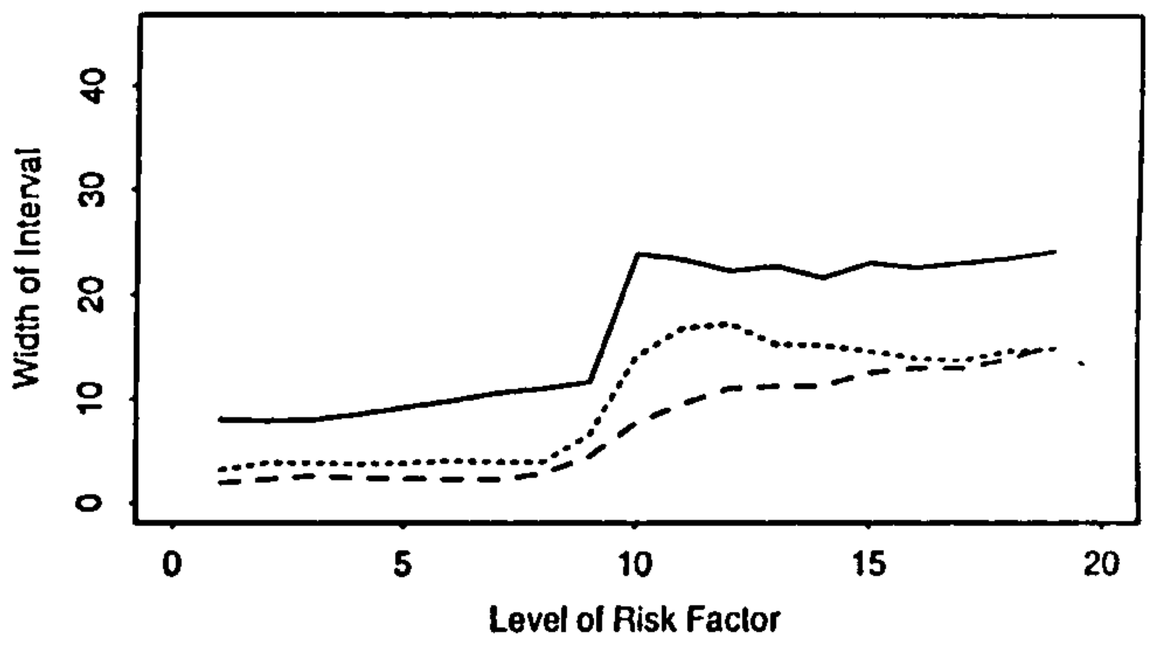
Sample Size = 100



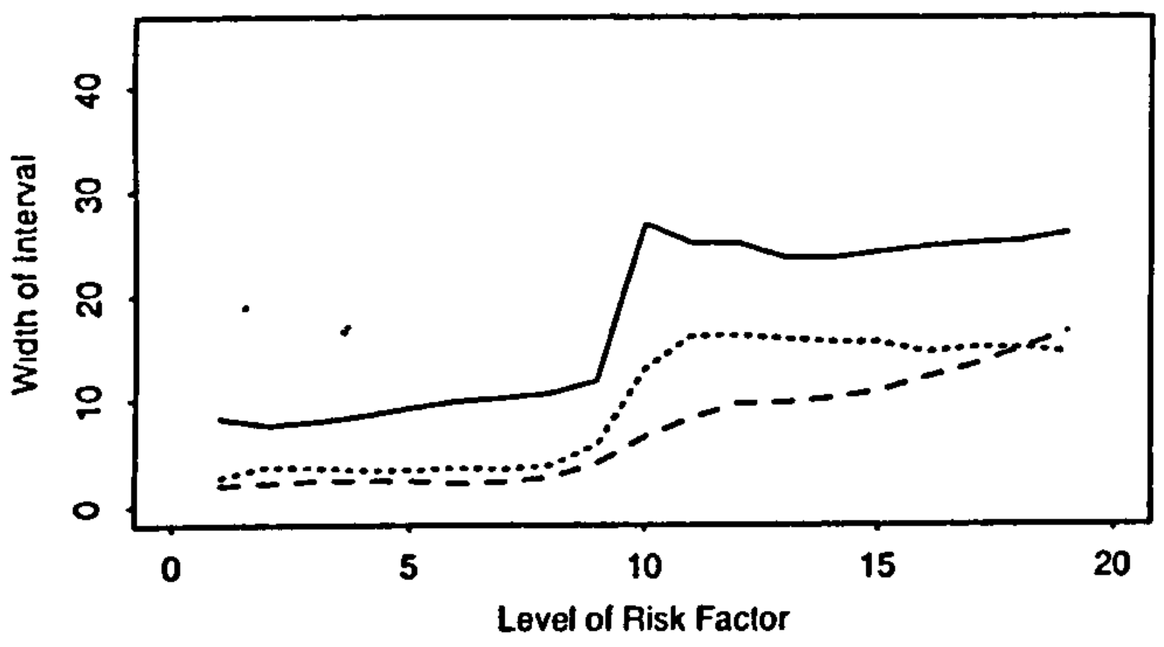
Sample Size = 150



Sample Size = 200



Sample Size = 250



Sample Size = 300

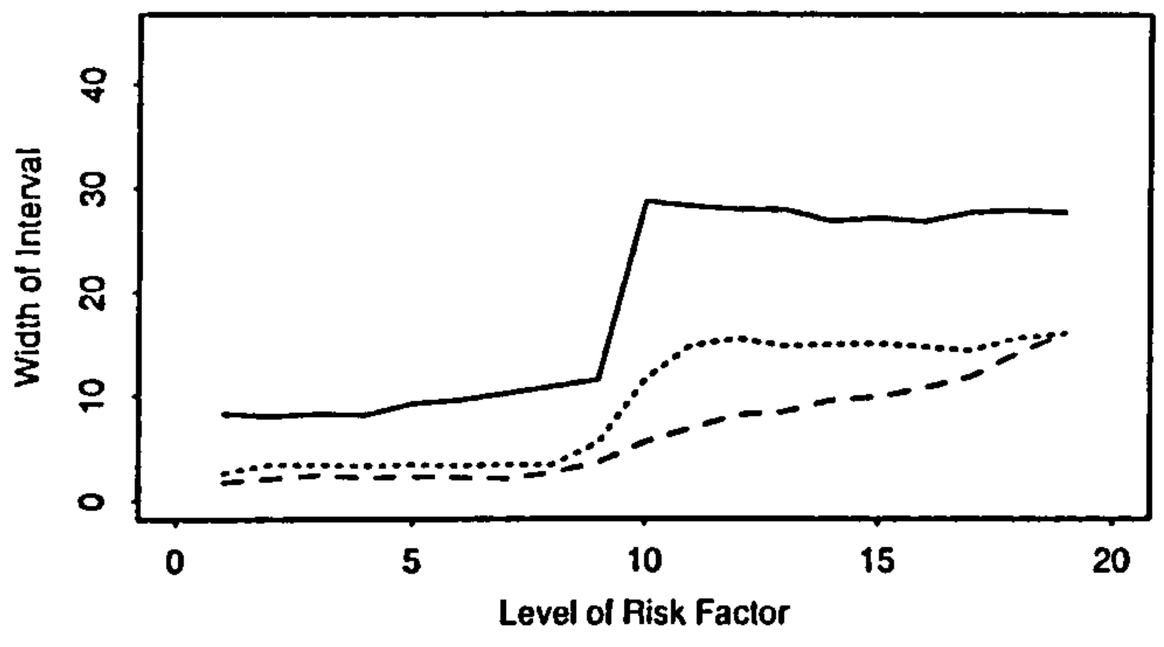


Figure 3.8.25

intervals. The coverage is again higher at values of the risk factor where the Relative Risk is equal to the baseline value and drops by quite a large amount at the location of the cut-point. The coverage gradually rises again as the level of the risk factor moves away from the location of the cut-point. In scenario 2 there was a suggestion that the pairwise cells method produced a more stable pattern of coverage around the cut-point compared to the rather erratic pattern displayed with the conditional likelihood method. There is little evidence to suggest that this is the case in this scenario, as, ***both methods*** actually appear quite erratic around the cut-point. Since the average width of the corresponding intervals are clearly narrower with the conditional likelihood method, it seems logical to suggest that the conditional likelihood method performs better in this scenario, in terms of the ***combination*** of coverage and width of interval. Finally, regardless of method, sample size or neighbourhood size, Figures 3.8.23 and 3.8.25 illustrate quite clearly that estimation will be better at lower values of the risk factor since the width of the 95% confidence intervals clearly increase as the level of the risk factor increases.

In summary, in this scenario, both methods produce very precise estimates which exhibit very little bias. This is particularly the case when larger sample sizes and/or smoothing are used. The conditional likelihood method produces estimates which are more precise whereas the pairwise comparisons method produces estimates which are slightly less biased. The results also suggest that the conditional likelihood method produces superior results in terms of the coverage and the average width of the corresponding 95% confidence intervals.

Section 3.8.2: Summary of the Results from the Simulation Study

The simulations carried out here suggest that both the conditional likelihood method and the pairwise cells method would be plausible methods to use to produce non-parametric estimates of Relative Risk in the presence of an interval scaled discrete risk factor. Although neither method produced "perfect" results in terms of precision, bias or coverage they did produce results which were reasonably promising. It is clear from considering these four scenarios that if sample sizes of a practical nature are being used (i.e. less than 300 pairs of observations) then some form of data smoothing will be required before acceptable solutions can be obtained. In the absence of smoothing both of the non-parametric methods under consideration here struggled to obtain "good" estimates of the true, underlying, pattern of Relative Risk. However when smoothing was introduced there was a clear improvement in precision, bias and coverage. There was also evidence that if the sample sizes are reasonably large (i.e. more than 200 pairs of observations) then care must be taken not to oversmooth the data particularly if the aim of the work is to identify cutpoints. In terms of the scenarios here the underlying distribution of the cases/controls had little effect on the precision, bias or coverage present with either method. The factor which appears most to influence the results is the underlying Relative Risk function. In this section two possible Relative Risk functions were considered; a linear and a step Relative Risk function. These were chosen in such a manner as to allow a direct comparison between the two functions. Both methods produced slightly more precise and marginally less biased estimates when attempting to reproduce the scenarios involving the step Relative Risk function. This is perhaps to be expected as the non-parametric methods of estimation proposed here are data fitting techniques and should therefore find it easier to identify only one major change in the Relative Risk function as opposed to a Relative Risk which changes, albeit linearly, at each level of the risk factor.

It is also somewhat reassuring that these methods perform reasonably well when dealing with a situation which involves a step in the Relative Risk: i.e. a categorisation point. The main reason for developing these techniques is to deal primarily with the scenario of identifying categorisation points for risk factors. It is encouraging to see that the methods can clearly identify such a point, if one does exist.

Perhaps the only, slightly worrying conclusion is in terms of the bias which is clearly present. There is evidence that the estimates of Relative Risk produced are invariably ***underestimates*** of the true Relative Risk. The use of larger sample sizes and/or the introduction of smoothing reduces the level of this underestimation but never entirely removes it.

In summary, both methods appear ***reasonably satisfactory*** in dealing with the scenarios considered here. It must be borne in mind that many more scenarios could have been considered but those observed here are fairly representative of the scenarios which may be of interest in a practical context. There is little evidence to suggest that one method is vastly superior to the other in terms of precision and bias as any differences which exist between the two appear relatively minor. The conditional likelihood method performs slightly better in terms of precision but the pairwise cells method performs better in terms of bias. The deciding factor between the two methods may come in terms of coverage and the width of the corresponding 95% confidence intervals. Here there is a clear suggestion that the conditional likelihood method is superior as, in general, this method produced more acceptable levels of coverage. More importantly, the conditional likelihood approach clearly produced narrower intervals regardless of sample size and level of smoothing. A final point to observe is that the methods do not perform especially

well with small sample sizes and no smoothing, suggesting that when the methods are applied to real data examples smoothing will invariably have to be used with smaller sample sizes. For example, the results presented in Section 3.5 concerning the Relative Risk of malignant melanoma associated with the presence of naevi were all based on data which had to be smoothed to a lesser or greater degree.

Section 3.9: Conclusions

In this chapter consideration has been given to both the theory and application of techniques for analysing data from a ***matched case-control study***. The standard conditional linear logistic model and two possible non-parametric approaches were considered. Both of the non-parametric methods produced estimates of the Relative Risk in the presence of an interval scaled discrete risk factor and helped to identify potential categorisations for such a risk factor.

An adaptation of both these non-parametric approaches involved "smoothing" these raw estimates of Relative Risk by either using a kernel smoother or carrying out an isotonic regression. The two methods of "smoothing" the estimates of Relative Risk occasionally produced different conclusions as to where potential categorisations may exist. The use of isotonic regression will, necessarily, produce flatter estimates of Relative Risk whilst the kernel smoother is likely to produce estimates which exhibit greater degrees of fluctuation. These differences in approach may lead to slightly different conclusions being obtained at points where any changes in the estimates of Relative Risk are of a relatively small nature. With the exception of these minor disparities there is however general agreement between the two smoothing techniques.

A simulation study was carried out and this revealed that both non-parametric approaches performed reasonably well across a number of scenarios. The methods were compared for three criteria; precision, bias and coverage. For moderate sample sizes, both methods appear reasonably precise and display levels of bias which are not particularly excessive. There is some evidence to prefer the likelihood based method as it produced better levels of precision and coverage whilst the pairwise cells method appeared slightly better only in terms of bias.

Future work should give consideration to extending this problem to investigate if a suitable non-parametric approach can be found to incorporate a continuous risk factor. Section 3.7 touched briefly on this problem by roughly categorising the continuous risk factor and then using the estimators discussed in Section 3.4 to produce estimates of Relative Risk. Further work on this method may allow some simple adaptation of the existing techniques for an interval scaled discrete risk factor to be used for the continuous case. Future work should also include extensions to deal with multiple risk factors as opposed to the univariate problem investigated here.

Chapter 4

Non-parametric Approaches to the Analysis of Survival Data

Section 4.1: Introduction

Survival data usually consists of observations on individuals for each of whom there is a well-defined point event of interest (e.g. failure/death) which occurs after a period of time. The unique ingredient of survival data is that it will typically contain some censored data (i.e. observation on some individuals in the study having had to cease before the event of interest (i.e failure/death) has occurred). This censoring of data will necessarily complicate any analysis but cannot just be ignored as this throws away information and will lead to the introduction of bias in conclusions on the distribution of failure/death times. In a survival data problem each individual subject will have ***both*** a failure/death time and possibly a censored time ***but*** only one of these will have been observed.

In the analysis of survival data interest is primarily in modelling the distribution of failure/death times (referred to as the failure time distribution from now on) possibly in the presence of important covariates. When the distribution of failure times is being modelled there are two related functions which are of particular interest.

(1) The SURVIVOR function, S(t), which is defined to be the probability that a randomly selected individual survives beyond the time point, t.

(2) The HAZARD function, h(t), which is defined as the probability that a randomly selected individual dies at time t, conditional on the individual having survived up till the time, t. The hazard function is also known as the instantaneous rate of failure at the time, t.

Kaplan and Meier (1958) introduced a simple non-parametric method of estimating the survivor function when no covariates are present using a maximum likelihood estimator. The problem of estimating the survivor function becomes more complicated when covariates are introduced and various methods of modelling the effect of covariates on the distribution of failure times have been suggested. One

common method is to use the Cox proportional hazards model (Everitt (1989)). This method models the hazard function with a non-parametric baseline hazard and incorporates a log-linear function to introduce the covariates. From this estimate of the hazard function an estimate of the survivor function is produced. Models such as exponential (Elandt-Johnson & Johnson (1980)) and Weibull (Cox & Oakes (1984)) regression models along with the accelerated failure time models (Cox & Oakes(1984)) are also in common usage.

Within some survival data problems where covariates are present interest is not specifically in modelling the full failure time distribution but instead an appropriate summary of survival is considered. Relevant summaries which are often used are to consider the probability of surviving a specified length of time given a covariate value (e.g. What is the probability of an individual surviving 5 years after a heart operation given their age on having the operation?). This would lead to modelling a binary response (does / does not survive 5 years). The common approach to modelling a binary response in the presence of covariates is to use the linear logistic model (Breslow & Day (1980)). One drawback to this technique when analysing survival data is that it may ignore the presence of censored observations and this often leads to an underestimate of the true probability of , say, 5 year survival. This particular problem of the bias incurred in survival analysis if censored observations are ignored is discussed in detail by Watt et al (1996).

Survival data occurs in many fields of study ranging from the analysis of failure times of components fitted in jet engines to the study of mortality intensities in animal experiments but is particularly prevalent in the area of medical research. The nature of medical research often leads to the production of survival data as many medical studies involve following up patients until they die of the disease of interest or are censored. Patients become censored through either dying of some cause other than the disease of interest or simply being lost to follow up. Many studies take place over an extended time period resulting in large amounts of censored data. In a study of Peripherial Arterial Disease, Criqui et al (1992) attempted to follow up patients for 10 years leading to large amounts of censored data. Of the 67 patients identified approximately 52% of them remained alive (i.e. censored) after the 10 year study period. Similarly in a study of survival from Hepatitis Seeff et al (1992) looked at 18 year survival again creating the potential for a large presence of censoring with an overall average of 49% censoring for mortality.

The work presented in this chapter will consider the analysis of survival data in the presence of a covariate and examine various non-parametric alternatives to the log-linear component in the survivor function. These techniques will be used to assist in identifying possible categorisation points for a single continuous covariate. These categorisations should be chosen at points where there appear to be marked changes in the prognosis of survival.

An examination of recent medical literature reveals that in the analysis of survival data in the presence of covariates/prognostic factors the proportional hazards model is the model most commonly used. For example in analysing survival from melanoma, Soong et al (1992) used a proportional hazards model to predict 5 and 10 year survival while the International Non-Hodgkin's Lymphoma Prognostic Factors Project (1993) carried out their analysis of 5 year survival from the disease by incorporating covariates/prognostic factors via the proportional hazards model. However although this model is often a sensible model to use in the analysis of survival data it is not particularly useful in identifying potential categorisation points for any important prognostic factors. The proportional hazards model incorporates parametric constraints on the relationship between the survival time distribution and the covariate(s) of interest. This rigid parametric framework often proves insufficient to deal with survival problems as it is inflexible in modelling this relationship. A wider range of relationships can be considered if the log-linear constraint is relaxed/removed and some form of non-parametric component is incorporated. Any non-parametric technique used will be essentially data-driven and hence will allow a very flexible development of the relationship between the failure time distribution and the covariate(s). Non-parametric techniques may not only illustrate any such important relationship but, due to their flexibility, allow any unusual features of the data to be highlighted. This latter point demonstrates that non-parametric techniques may prove useful in the identification of possible categorisation points.

Section 4.2 will consider the theory behind the standard methods of estimating failure time distributions and these will be applied to a set of data from the field of medical research in section 4.3. Section 4.4 will introduce non-parametric approaches to the analysis of survival data in the presence of a covariate and these approaches will be demonstrated in section 4.5. Section 4.6 will present some simulation studies discussing the results obtained by using the non-parametric approaches to reproduce a known situation.

Section 4.2: Standard approaches to the analysis of survival data

The survival time, t, of a randomly selected individual can be defined by some, unknown, underlying distribution function, F(t). The distribution function of the associated random variable, T, is given by

$$F(t) = \text{Prob}(T < t)$$

However when censoring is present, it is often more relevant to consider the survivor function at the time, t, denoted by S(t). If no covariates are present the survivor function at the time, t, is defined by

$$S(t) = 1 - F(t) = P(T \geq t) \qquad \text{- (4.0)}$$

(i.e. Probability of surviving beyond time t)

Interest here is often in estimating this function in the presence of censoring. In 1958 Kaplan and Meier devised an essentially non-parametric method of estimating the survivor function (when no covariates are present) based on data

including censored observations. This estimator is commonly known as the product-limit estimator and a brief description of it is as follows.

Let the data consist of observations on n subjects and assume that each subject has a failure time $\{t_i ; i = 1, \ldots n\}$ and a censored time $\{c_i ; i = 1, \ldots, n\}$. However, for each subject only one of these times will actually be observed and hence the data could be summarised as

$$X_i = \min(t_i, c_i) ; i = 1, \ldots, n$$

Further, denote by r_j the number of items at risk throughout the period $(t_{j-1}, t_j]$ (i.e. between the $(j-1)^{th}$ and $(j)^{th}$ failure times) and by s_j the number who survive beyond t_j.

Now the survivor function, S(t), is defined to be the probability of surviving beyond time t. Therefore for any particular failure time, t_i

$$\begin{aligned} S(t_i) &= P\{T > t_i\} \\ &= P\{T > t_i / T > t_{i-1}\} P\{T > t_{i-1}\} \end{aligned}$$

etc.

$$= \prod_{j=1}^{i} P\{T > t_j \ / \ T > t_{j-1}\}$$

Kaplan and Meier estimated each of these conditional probabilities separately by s_j/r_j to give

$$\hat{S}(t_i) = \prod_{j=1}^{i} s_j/r_j \qquad \text{- (4.1)}$$

and

$$\hat{S}(t) = \hat{S}(t_i) \qquad \text{for } t_i < t < t_{i+1}$$

This is a simple, logical estimate of the survival function when no covariates are present. Note that the estimate of survival, $\hat{S}(t)$, only changes at the observed failure times. Hence this estimator will be a step function which changes at each observed failure time. Further, techniques originally devised by Greenwood (1926) allowed confidence bands for S(t) at any value of t to be derived and these are given by

$$\left[\hat{S}(t)\right]^{\exp\left[\pm 1.96\sqrt{\hat{var}(\hat{v}(t))}\right]} \qquad \text{- (4.2)}$$

where

$$v(t) = \log_e\{-\log_e(S(t))\}$$

and

$$\hat{var}(\hat{v}(t)) = \frac{1}{\left[\log_e(\hat{S}(t))\right]^2}\hat{var}\left(\log_e(\hat{S}(t))\right)$$

$$\hat{var}\left(\log_e(\hat{S}(t))\right) = \sum_{j=1}^{i}\frac{r_j - s_j}{s_j r_j}$$

With the introduction of covariates the problem becomes more complicated and section 4.1 mentioned various methods of modelling the relationship between the failure time distribution and covariates. The key method in common usage is the Cox Proportional Hazards model which is defined as follows.

Let $h(t;\underline{z})$ be the hazard function at any time t for an item with p-dimensional covariate vector $\underline{z} = (z_1, z_2, \ldots, z_p)^T$. The proportional hazards model is then defined as

$$h(t;\underline{z}) = h_0(t)\exp(\underline{\beta}^T\underline{z}) \qquad \text{- (4.3)}$$

for an arbitrary baseline hazard function $h_0(t)$ (i.e. hazard at $\underline{z} = \underline{0}$) where $\underline{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ is a set of unknown parameters. This model is essentially semi-parametric as it contains both a non-parametric component via the distribution free baseline hazard and a parametric component through the exponential function. From this definition the survival function can be estimated in the presence of covariates since a simple relationship exists between the hazard and survivor functions of the following form

$$S(t) = \exp\left\{-\int_0^t h(u)du\right\}$$

Under the Cox Proportional Hazards model the following relationship between the survivor function and the hazard function at any value of the covariate z can readily be observed (Cox and Oakes (1984))

$$S(t / z) = [S_0(t)]^{\exp\{\underline{\beta}^T \underline{z}\}} \qquad -(4.4)$$

where

$$S_0(t) = \exp\left\{-\int_0^t h_0(u)du\right\}$$

This model can then be fitted to an appropriate set of data, parameter estimates $\underline{\hat{\beta}} = \left(\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p\right)^T$ for $\underline{\beta} = \left(\beta_1, \beta_2, ..., \beta_p\right)^T$ found and hence an estimate of the survival function using the relationship in (4.4) to give

$$\hat{S}(t / z) = [S_0(t)]^{\exp\{\underline{\hat{\beta}}^T \underline{z}\}}$$

In order to assess whether the resultant fitted Cox Proportional Hazards model is appropriate residual plots can be carried out. Various residuals have been proposed in the analysis of survival data. These include the Cox-Snell residuals (Cox and Snell (1968)), Score residuals (Schoenfeld (1982)) and Martingale residuals (Fleming and Harrington (1991)). Here the deviance residuals (Therneau, Granbsch and Fleming (1990)) will be plotted against the corresponding follow-up time in order to assess if the fitted proportional hazards model is appropriate and to highlight potential outliers.

Another situation which may be of interest is to take a ***fixed*** point in time and examine the probability of surviving beyond that time. If a fixed point in time is chosen then the problem essentially becomes one of modelling a ***binary response*** (does / does not survive past the fixed time point). Hence this situation is sometimes analysed via the use of the linear logistic model. Section 2.2 of this thesis examined the linear logistic model in great detail and (2.1) gave a formula for relating the probability of a success (e.g. alive or survival) to a series of covariates. Here, a

linear logistic model would be fitted to the data in order to identify factors which may be of prognostic significance in terms of predicting survival ***past*** the specified, fixed, point in time. One point to observe is that this method will ***ignore*** any observations who are alive ***but*** have ***not*** been followed up for at least the specified time. The importance of this issue will be discussed later in this chapter.

Section 4.3: Analysis of a subset of the Scottish Melanoma Group database.

The Scottish Melanoma Group has collected data on 4399 patients first presenting in Scotland with cutaneous melanoma (i.e. stage 1 melanoma) between 1979 and 1990. This database contains detailed clinical, pathological, surgical and follow-up data on all these patients. The data has already been analysed by MacKie et al (1995) to identify important prognostic factors for survival from this severe form of skin cancer. MacKie et al used both the technique of Kaplan and Meier and the proportional hazards model in their analysis in order to predict survival for various subgroups of patients.

In order to simplify the illustrations which follow, a small representative subset of the full database will be considered. Therefore, consideration will here be given to ***females with ulcerated lesions on an axial site*** who have been followed up for a minimum of 5 years. This leaves a total of 108 subjects for study of whom 63 had a complete follow-up time (i.e. failure time) and 45 had an incomplete follow-up time (i.e. censored time).

From these data the Kaplan-Meier estimate of survival was calculated and is displayed in Figure 4.3.1. This shows that overall survival for these females is relatively poor with a probability of surviving at least 2 years of about 80% (95%

Females with Ulcerated Lesions on an axial site

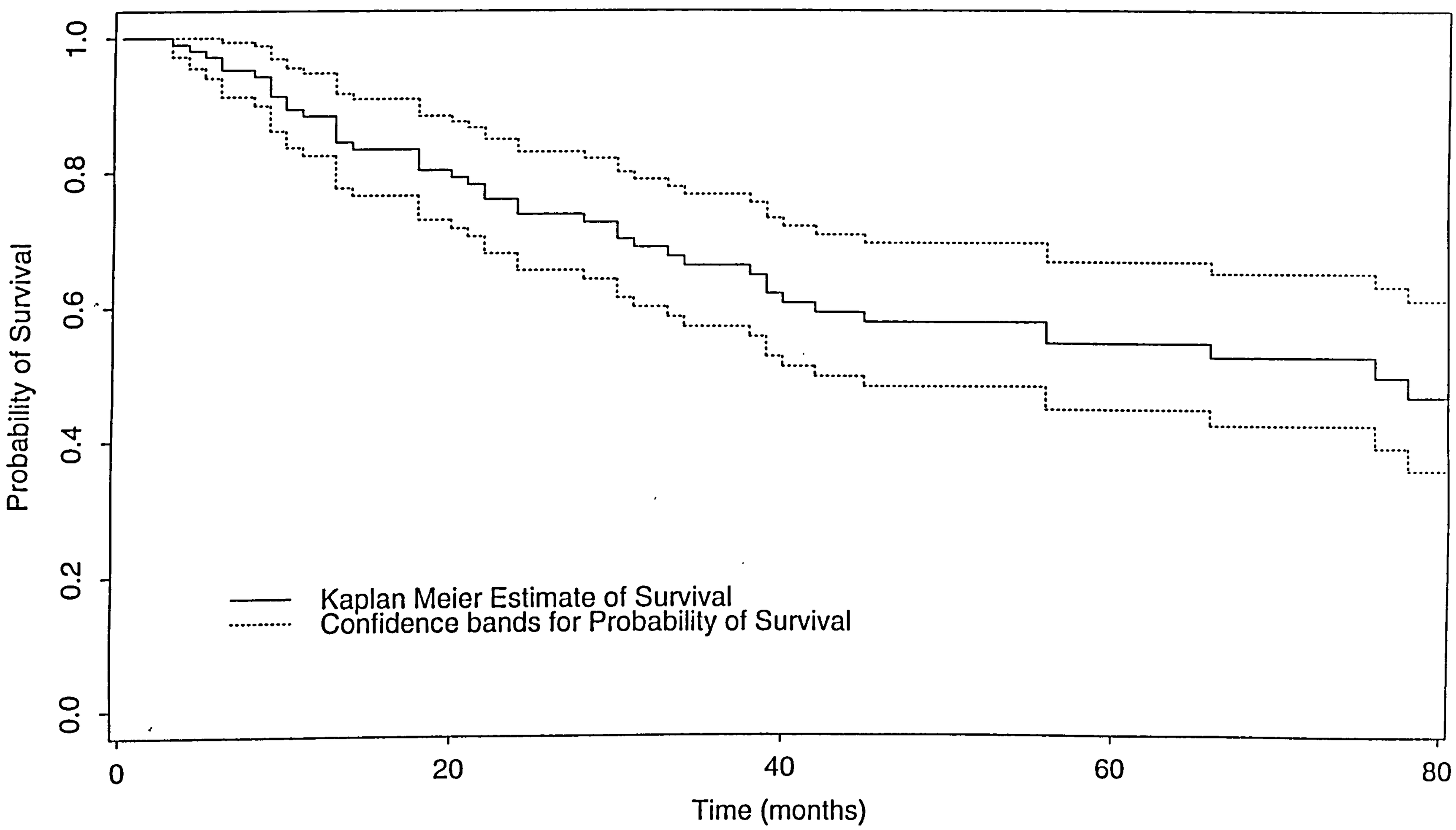


Figure 4.3.1

confidence interval of 75% to 85%) and a probability of surviving at least 5 years of 50% (95% confidence interval of 42% to 62%). This drops below 40% when the probability of surviving at least 10 years (not on Figure) is considered (95% confidence interval of 33% to 58%).

Numerous studies (Szymik and Woosley (1993), Rigel et al (1991) & Ronan et al (1988)) have shown that possibly the most important factor in survival from stage 1 melanoma is the tumour/Breslow thickness on diagnosis. In order to incorporate this factor into any analysis it is necessary to use a model which relates the failure time distribution to a covariate.

MacKie et al fitted proportional hazards models to show that for the Scottish Melanoma Group database ***tumour thickness*** has a significant effect on survival. To illustrate the effect tumour thickness has on survival for the small subset of the database under examination here firstly consider Figure 4.3.2 which shows a plot of the distribution of the tumour thickness by status (i.e. censored or dead) for this group of females. From Figure 4.3.2 it can be seen that among the subjects who have complete observations (i.e. a clearly defined endpoint; death due to melanoma) there appears to be a higher proportion with thicker tumours. The subjects who have died due to melanoma also appear to exhibit a larger amount of variability in their tumour thickness than is present among the censored observations.

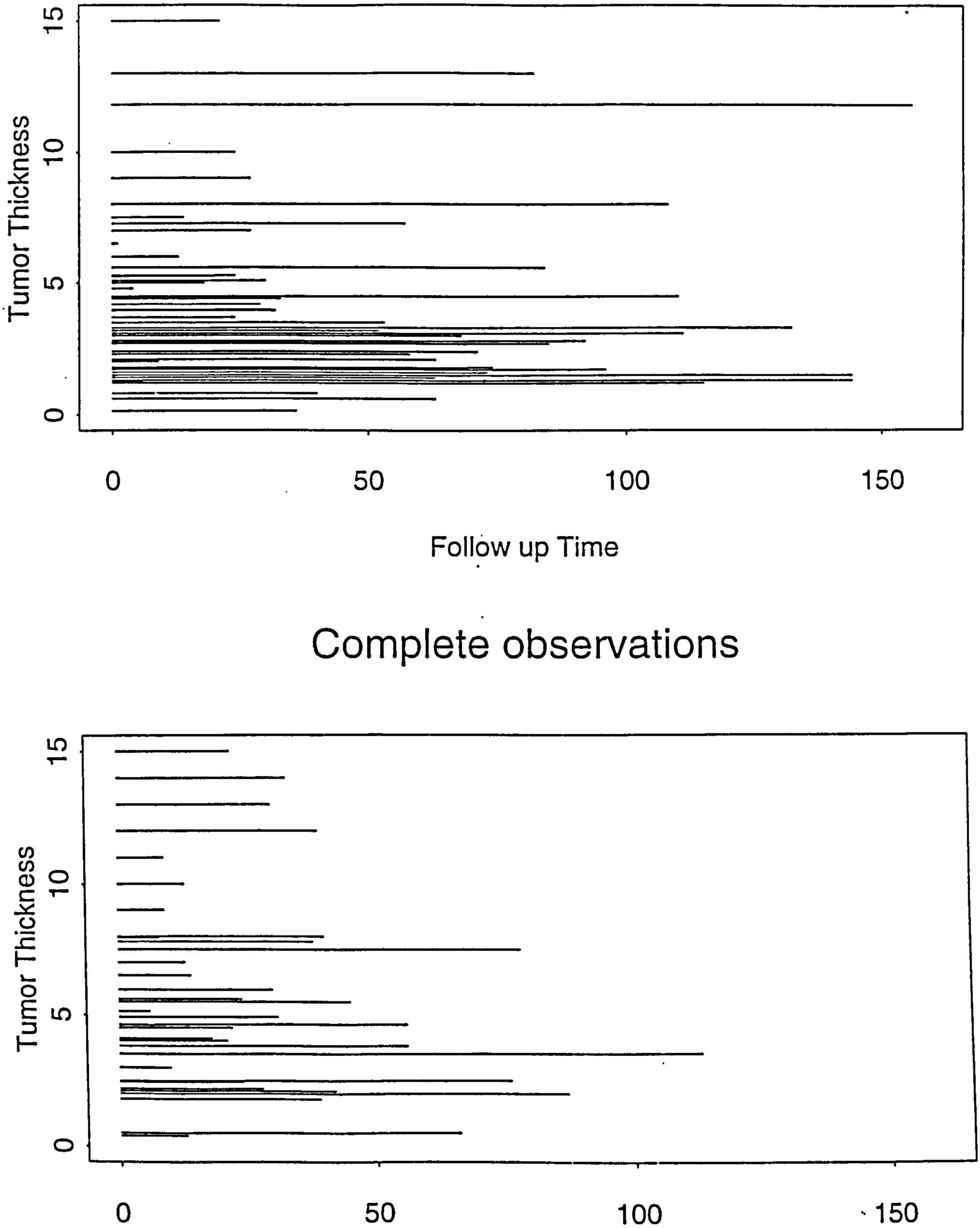

Censored observations
Tumor Thickness
0
5
10
15
0
50
100
150
Follow up Time
Complete observations
Tumor Thickness
0
5
10
15
0
50
100
150
Follow up Time

Figure 4.3.2

Figure 4.3.3 shows a graph of the fitted proportional hazards model for this data with tumour thickness incorporated as a covariate. It shows estimated survival curves at a selection of the possible tumour thicknesses and demonstrates that under this model, as expected, survival prospects decrease ***both*** through time and as the thickness of the tumour increases. Given the severity of the disease present in these subjects (i..e. subjects with ***ulcerated*** lesions) survival prospects appear reasonably good for those subjects with a tumour thickness of less than 3 mm. More specifically subjects with a tumour thickness of 1 mm have 5 year survival of approximately 75% and even 7 year survival of over 60%. However the survival prospects are very poor for subjects who have a tumour thickness of greater than about 7 mm and in particular those with a tumour thickness of 9 mm only have about a 32% chance of surviving 5 years dropping to just over 20% by the time 7 year survival is considered.

In order to assess the fit of the model the Deviance residuals were calculated. Figure 4.3.4 displays a plot of the standardised Deviance residuals against both the follow up time and the included explanatory, tumour thickness. As neither of these plots display any suggestion of a trend, or, indeed evidence of any outliers, it is reasonable to assume that the fitted proportional hazards model gives an adequate fit to the data.

One drawback with using the proportional hazards model is that it cannot be used to highlight potential categorisations for the covariate under study. The

PAGE
NUMBER
CUT OFF
IN
ORIGINAL

Females with Ulcerated Lesions on an axial site

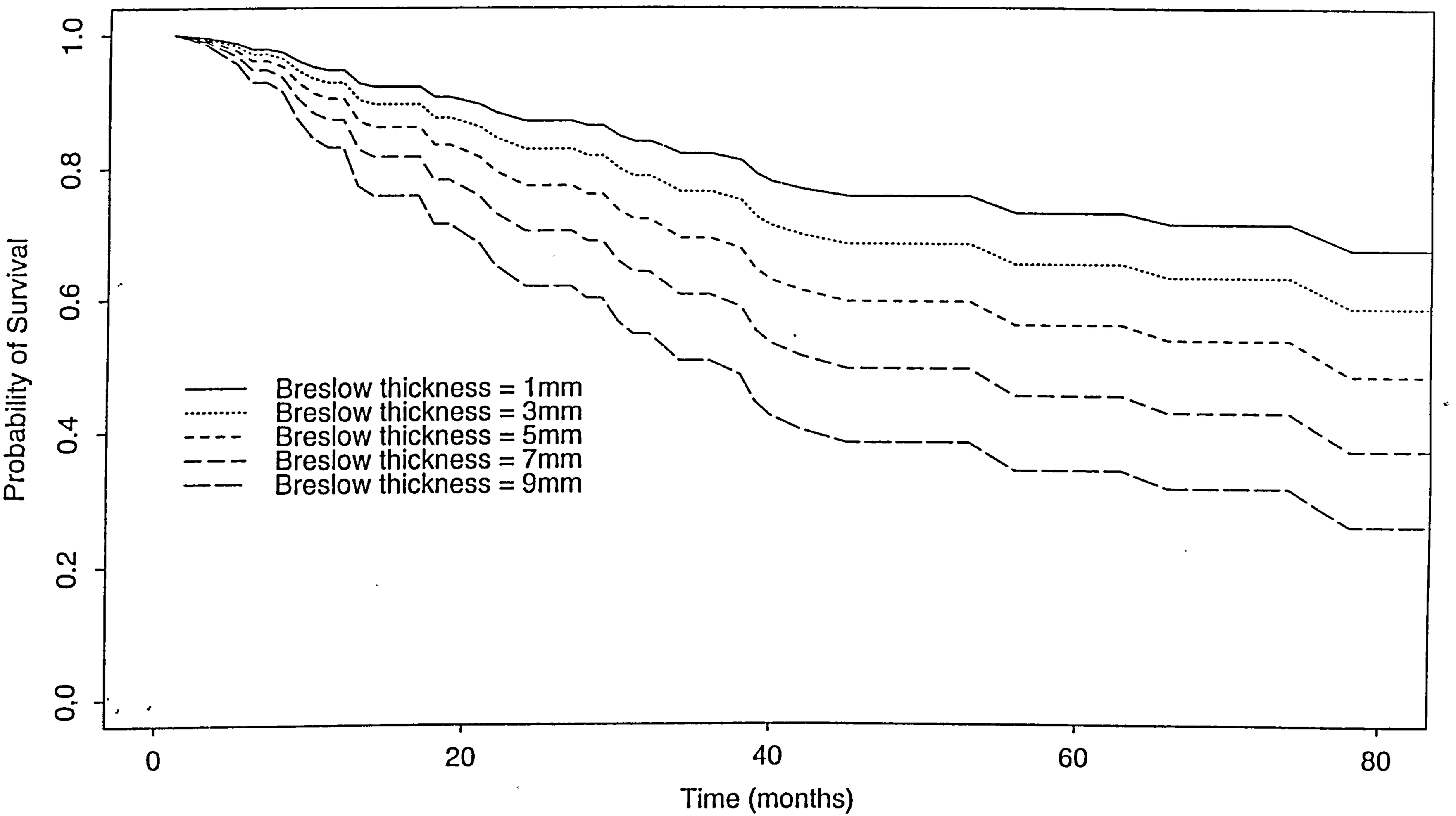


Figure 4.3.3

Females with Ulcerated Lesions on an axial site

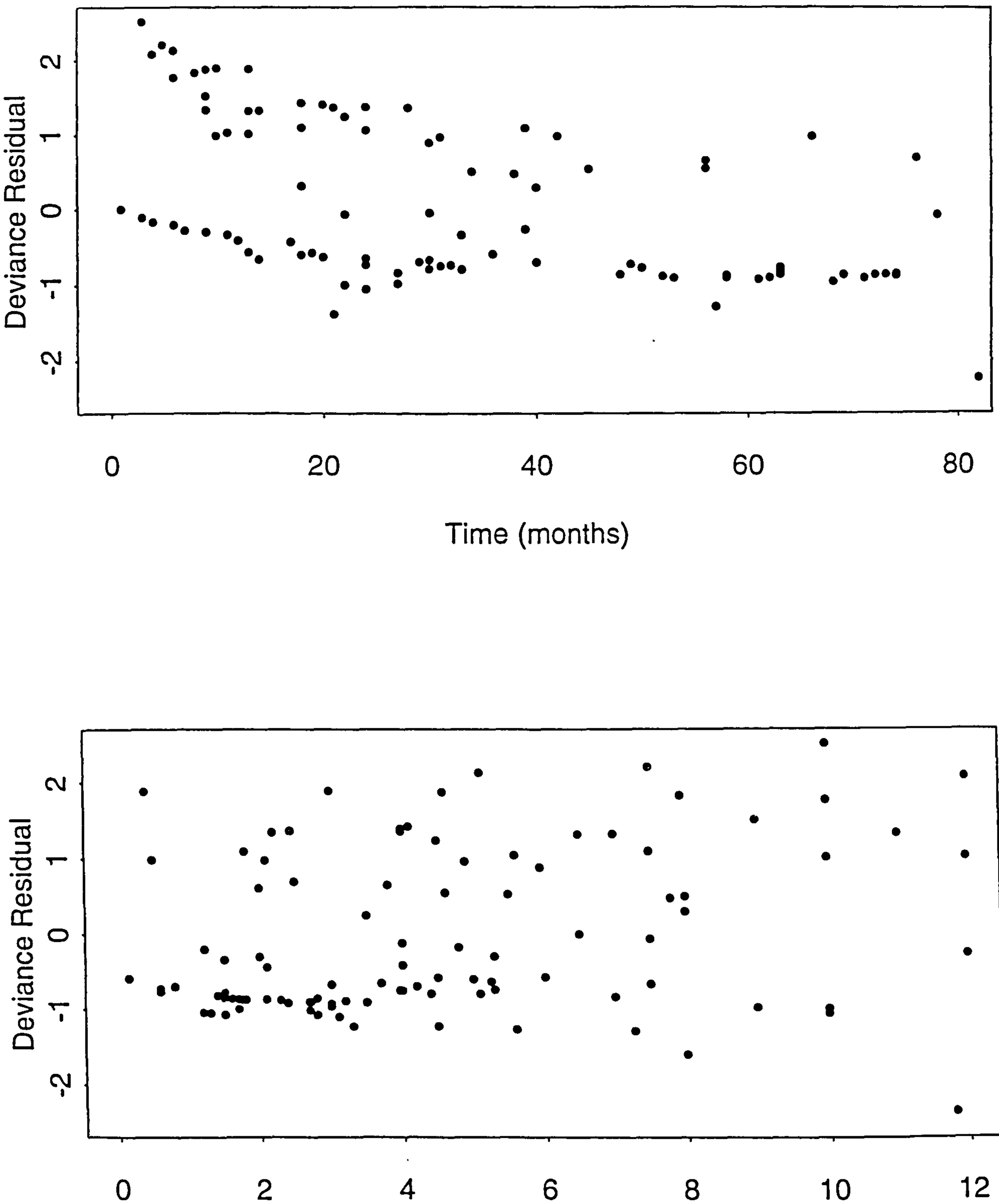


Figure 4.3.4

proportional hazards model forces a "parallel lines pattern" across the different levels of the covariate (illustrated in Figure 4.3.3) through the use of the exponential power function in (4.3). In order to look for possible categorisations for a potentially important covariate it is necessary to adopt a method which is less rigid in how it models the relationship between survival and the covariate. This leads to the idea of using a non-parametric approach which drops the log-linear assumption inherent in the proportional hazards model. In the next section some consideration will be given to non-parametric techniques to hopefully give a more flexible but still sensible solution to the problem.

Section 4.4: Non-parametric approaches to the analysis of Survival Data

One of the main aims of this study is to examine possible ways of producing non-parametric estimators of survival in the presence of a continuous covariate as this will allow a "flexible" relationship between survival and the covariate to be examined. The use of these estimators will hopefully allow possible categorisations for the covariate, if any exist, to be highlighted. Section 4.2 outlined various approaches to survival analysis, with or without a covariate, but none of these meet the criterion required. This section attempts to extend each of the standard approaches described in Section 4.2 to allow one to attempt to identify meaningful categorisations of a single explanatory variable.

Section 4.4.1: Kaplan-Meier based approach

The Kaplan Meier approach (Section 4.3) is a non-parametric approach to survival analysis but it does not incorporate a covariate. If an extension to this method can be found which incorporates a covariate it may produce sensible non-parametric estimates of survival in the presence of a covariate. If the sample size were 'infinite' the natural extension to the Kaplan Meier estimate of Survival incorporating a covariate would be to produce as an estimator of survival

$$\hat{S}(t_i / z) = \prod_{j=1}^{i} s_j(z) \Big/ r_j(z) \qquad \text{- (4.5)}$$

where

t_i represents the i'th failure time; $i = 1,..., m$;

m is the number of distinct failure times;

$s_j(z)$ are the number of subjects with covariate value z who survive past t_j;

$j = 1,..., i$;

$r_j(z)$ are the number of subjects with covariate value z who are at risk at t_j;

$j = 1,..., i$.

This estimator is simply a separate Kaplan Meier estimate of survival for each level of the covariate. In order to use this estimator large amounts of data would have to be present at each level of the covariate. In practice however it is extremely unlikely that large enough data sets will be available. A natural and practical solution therefore is to 'smooth' the data across the covariate space and consider the following estimator

$$\hat{S}(t_i / z) = \prod_{j=1}^{i} s_j^*(z) \Big/ r_j^*(z) \qquad \text{- (4.6)}$$

where

$$s_j^*(z) = \sum_{k=1}^{n}(1 - x_k(t_j))R_k(t_j)\Delta_h(z, z_k)$$

$$r_j^*(z) = \sum_{k=1}^{n}R_k(t_j)W_k(z)$$

and

$$R_k(t) = \begin{cases} 1 \text{ if } t_k \geq t \\ 0 \quad \text{else} \end{cases}$$

$$I_k(t) = \begin{cases} 1 & \text{if person k is dead at time t} \\ 0 & \text{else} \end{cases}$$

n = number of observations

$$\Delta_h(z, z_k) = K\left(\frac{z - z_k}{h}\right)$$

Here, $\Delta_h(z, z_k)$ is a smooth kernel function with the parameter h controlling the amount of smoothing. This kernel function will put more weight on the k'th subject's covariate value which is close to the value, z, of the covariate of immediate interest and exponentially less on those whose values are further away.

Point estimates, $\hat{S}(t / z)$, for $S(t / z)$ for any value of t can then be provided by linearly interpolating between failure times as follows

$$\hat{S}(t / z) = \frac{(t_i - t) * \hat{S}(t_{i-1} / z) + (t - t_{i-1}) * \hat{S}(t_i / z)}{t_i - t_{i-1}} \quad \text{for } t_{i-1} < t < t_i$$

Confidence bands for $S(t / z)$can also be produced by using results analagous to those for the simple Kaplan Meier (i.e. without a covariate). In the absence of a covariate (4.2) gave confidence bands for S(t) of the form

$$\left[\hat{S}(t)\right]^{\exp\left[\pm 1.96\sqrt{\hat{var}(\hat{v}(t))}\right]}$$

(See section 4.2 for definitions of $\hat{v}(t)$)

In the presence of a covariate, confidence bands can be produced in a similar fashion except that $\hat{S}(t)$ is replaced by $\hat{S}(t / z)$ as follows:-

$$\left[\hat{S}(t/z)\right]^{\exp\left[\pm 1.96\sqrt{\hat{var}\left(\hat{v}(t/z)\right)}\right]} \qquad - (4.7)$$

where

$$v(t/z) = \log_e\left\{-\log_e\left(S(t/z)\right)\right\}$$

and

$$\hat{var}\left(\hat{v}(t/z)\right) = \frac{1}{\left[\log_e\left(\hat{S}(t/z)\right)\right]^2}\hat{var}\left(\log_e\left(\hat{S}(t/z)\right)\right)$$

with

$$\hat{var}\left(\log_e\left(\hat{S}(t/z)\right)\right) = \sum_{j=1}^{i}\frac{r_j^*(z) - s_j^*(z)}{s_j^*(z)r_j^*(z)}$$

In essence, this estimator of survival in the presence of a covariate incorporates the covariate by basically smoothing the Kaplan Meier estimate of survival across the values of the covariate. Therefore, in general, $\hat{S}(t/z)$ will exhibit a "Kaplan Meier type of profile" but, at values of the covariate where survival is poorer, $\hat{S}(t/z)$ will exhibit a sharper rate of descent than at values where survival is better. Further, by linearly interpolating between failure times a smoother

estimate of survival will be produced across time than is produced with the standard Kaplan Meier.

Section 4.4.2: Non-parametric Hazard based approach

A second approach to producing a non-parametric estimate for the survivor function is to produce a ***non-parametric*** estimate for the hazard function and hence for the survivor function. The proportional hazards model (section 4.3) incorporates a ***semi-parametric*** estimator of the hazard function h(t;z) into the estimator of the survivor function. Tanner and Wong (1983) suggested the following completely non-parametric estimator of the hazard function when ***no*** covariates are present.

$$\hat{h}(t) = \sum_{j=1}^{n} \frac{\delta_{(j)}}{n-j+1} K_h(t - t_{(j)}) \qquad \text{- (4.8)}$$

where

$t_{(1)}, \ldots t_{(n)}$ are the ordered t_j's (i.e includes both censored times and failure times);

$\delta_{(1)}, \ldots \delta_{(n)}$ are the corresponding indicators of whether a failure time or censored time has been observed;

and $K_h(t - t_{(j)})$ is a symmetric non-negative kernel usually taken to be a Normal Kernel.

$$\text{i.e.} \quad K_h(t - X_{(j)}) = \frac{1}{h} K\left(\frac{t - t_{(j)}}{h}\right)$$

$$\text{with} \quad K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$

Hence the estimate of S(t) based on this is

$$\hat{S}(t) = \exp\left[-\int_0^t \hat{h}(t)_{dt}\right]$$

$$= \exp\left[-\sum_{j=1}^{n} \frac{\delta_{(j)}}{n-j+1} \int_0^t K_h(u - X_{(j)})du\right] \qquad \text{- (4.9)}$$

$$= \exp\left[-\sum_{j=1}^{n} \frac{\delta_{(j)}}{n-j+1} \Phi\left(\frac{t - X_{(j)}}{h}\right)\right]$$

where $\Phi(\cdot)$ is the cdf of the standard Normal distribution

Tanner and Wong derived the following formula for the variance of $\hat{h}(t)$:-

$$\text{var}\left(\hat{h}(t)\right) = \int_y I_n\left(F(y)\right) h(y) K_h^2(t - y)\, dy$$

$$+ 2 \iint_{y \le x} \left\{ \left(F(x)\right)^n - \left(F(y)\right)^n \left(F(x)\right)^n - \frac{1 - F(y)}{F(x) - F(y)} \left[\left(F(x)\right)^n - \left(F(y)\right)^n \right] \right\}$$

$$h(y)\, h(x)\, K_h(t - y)\, K_h(t - x)\, dy\, dx \qquad \text{- (4.10)}$$

where

$$I_n\left(F(.)\right) = \sum_{j=0}^{n-1} (n - j)^{-1} \binom{n}{j} F(.)^j \left(1 - F(.)\right)^{n-j}$$

$h(t)$ = 'true, unknown, hazard'

and

$F(.)$ = the cumulative distribution function of the observed times

(i.e. of both the censored and failure times)

By using a dominated convergence argument, Tanner and Wong showed that $\mathrm{var}\left(\hat{h}(t)\right)$ simplifies to

$$\mathrm{var}\left(\hat{h}(t)\right) = \frac{1}{nh} \left[\int_y K^2(y)\, dy \right] h(t) \left(1 - F(t)\right)^{-1} + o\left((nh)^{-1}\right) \qquad \text{- (4.11)}$$

Finally, Tanner and Wong used the projection method (Hajek (1968)) to justify asymptotic normality of $\hat{h}(t)$ and hence asymptotic confidence for the hazard function. From these induced confidence intervals can be provided for the survivor function by using the standard relationship between hazard and survivor functions.

More recent papers within this field have considered various properties of the estimator proposed by Tanner and Wong. Muller and Wang (1990) discussed the use of the first derivative of the hazard in (4.8) to identify points of most rapid change in the hazard. Also, Muller and Wang (1994) considered a modified version of (4.8) which incorporated variable degrees of smoothing to assist with boundary effects. Various papers (Marron and Padgett (1987), Sarda and Vieu (1990), Patil (1990)) have examined methods for choosing the optimal smoothing parameter in the hazard in (4.8). However, very little work has been done on incorporating a covariate into the problem. In their book on local polynomial modelling, Fan and Gijbels (1996) briefly discussed the idea of considering neighbourhoods of covariate values and, within these neighbourhoods, fitting local proportional hazards models. The parameter estimates will be different within each local proportional hazards model and, when the models are all joined together, a smooth estimate of the hazard will be produced across both time and the covariate.

However, by extending the method of Tanner and Wang, a simple alternative method exists to incorporate a ***single*** covariate z into the hazard function. Consider

the following non-parametric estimator of the hazard function allowing for a single covariate

$$\hat{h}(t;z) = \sum_{j=1}^{n} \frac{\delta_{(j)}}{n-j+1} K_{h_1}(t - X_{(j)}) K_{h_2}(z - z_{(j)}) \qquad \text{- (4.12)}$$

where

$z_{(j)}$ is the covariate value corresponding to $t_{(j)}$

and

$K_{h_2}\left(z - z_{(j)}\right)$ is a symmetric non-negative kernel as defined earlier in the section.

The corresponding estimated survivor function is

$$\hat{S}(t / z) = \exp\left[-\int_0^t \hat{h}(t;z)_{dt} \right]$$

$$= \exp\left[-\sum_{j=1}^{n} \frac{\delta_{(j)}}{n-j+1} \int_0^t K_{h_1}(u - X_{(j)})_{du} K_{h_2}(z - z_{(j)}) \right] \qquad \text{- (4.13)}$$

$$= \exp\left[-\sum_{j=1}^{n} \frac{\delta_{(j)}}{n-j+1} \Phi\left(\frac{t - X_{(j)}}{h_1} \right) K_{h_2}(z - z_{(j)}) \right]$$

The variance of $\hat{h}(t;z)$ is as follows:

$$\operatorname{var}\left(\hat{h}(t;z)\right) = \left[E\left(\hat{h}(t;z)^2\right)\right] - \left[E\left(\hat{h}(t;z)\right)\right]^2$$

with

$$E\left(\hat{h}(t;z)\right) = \int_y G_1(y/z)\, h(y)\, K_{h_1}(t-y)\, dy$$

and

$$E\left(\hat{h}^2(t;z)\right) = \int_y G_2(y/z)\, h(y)\, K_{h_1}^2(t-y)\, dy$$

$$+2 \iint_{y<x} \sum\sum_{r<s} \frac{K_{h_2}\left(z-z_{(r)}\right) K_{h_2}\left(z-z_{(s)}\right)}{(n-r+1)(n-s+1)} \frac{n!}{(r-1)!\,(s-r-1)!\,(n-s)!}$$

$$F(y)^{r-1}\left[1-F(y)\right]\left[F(x)-F(y)\right]^{s-r-1}\left[1-F(x)\right]^{n-s+1}$$

$$h(y)\, h(x)\, K_{h_1}(t-y) K_{h_1}(t-x)\, dydx$$

where

$$G_1(y/z) = \sum_{j=1}^{n} \frac{n!}{(j-1)!\,(n-j+1)!} K_{h_2}\left(z-z_{(j)}\right) F(y)^{j-1}\left(1-F(y)\right)^{n-j+1}$$

$$G_2(y/z) = \sum_{j=1}^{n} \frac{n!}{(j-1)!\,(n-j)!} \frac{K_{h_2}^2\left(z - z_{(j)}\right)}{(n-j+1)^2} F(y)^{j-1}\left[1-F(y)\right]^{n-j+1}$$

See Appendix 2 for full derivation of these results.

Unlike the variance of the hazard function derived by Tanner and Wong in (4.10) there appears to be no obvious simplification of the variance here.

These results give the exact mean and variance for the hazard function. In order to provide confidence intervals for the hazard function it may be possible to adapt the results on the asymptotic normality of the hazard function to produce approximate confidence intervals for the hazard function and hence the survivor function. However, in practice, due to the complexity of computing the variance term, calculation of any confidence intervals is impractical for larger sample sizes. In practice therefore confidence intervals for S(t/z) will be produced using the following approximate pivotal result based on the proportional odds model (Collett (1991))

$$\frac{\log_e\left(\frac{S(t/z)}{1-S(t/z)}\right) - \log_e\left(\frac{\hat{S}(t/z)}{1-\hat{S}(t/z)}\right)}{\mathrm{var}\left(\log_e\left(\frac{\hat{S}(t/z)}{1-\hat{S}(t/z)}\right)\right)} \;\dot{\sim}\; N(0,1)$$

with the asymptotic variance of $\log_e\left(\frac{\hat{S}(t/z)}{1-\hat{S}(t/z)}\right)$ being

$$\frac{1}{n*\hat{S}(t/z)}+\frac{1}{n*\left(1-\hat{S}(t/z)\right)}$$

giving an approximate 95% confidence interval for $\log\left(\frac{S(t/z)}{1-S(t/z)}\right)$ of the form

$$\log\left(\frac{\hat{S}(t/z)}{1-\hat{S}(t/z)}\right) \pm 1.96*\text{sqrt}\left(\frac{1}{n*\hat{S}(t/z)}+\frac{1}{n*\left(1-\hat{S}(t/z)\right)}\right) \quad \text{- (4.14)}$$

$$= [a \ , \ b]$$

Hence an induced approximate 95% confidence interval for S(t/z) is of the form

$$\left[\frac{\exp(a)}{1+\exp(a)}, \frac{\exp(b)}{1+\exp(b)}\right] \quad \text{- (4.15)}$$

One final point to observe here is that, in practice, this estimator may prove rather problematic to implement due to the presence of two levels of smoothing; one across the covariate and one across time.

Section 4.4.3: Non-parametric logistic based approach

The technique of Logistic regression (Breslow & Day (1980)) is used to examine the relationship of a binary response (e.g. dead/alive etc.) on one or more potentially important covariates. The major difference of logistic regression with survival analysis is that data is being modelled ***through time*** rather than at ***a fixed point in time***. Hence the binary response for an individual will ***change*** at some point in time (e.g. from alive to dead). Copas (1983) suggested a non-parametric logistic approach to relating a binary response, y, to a single covariate, z. Chapter 2 of this thesis and in particular section 2.3 gave a detailed discussion of this methodology with (2.3) giving the following formula for relating y to z.

$$\hat{p}_z = \hat{P}(Y=1/z) = \frac{\sum_{j=1}^{n} \Delta_h(z,z_j) y_j}{\sum_{j=1}^{n} \Delta_h(z,z_j)}$$

where

z_j is the continuous explanatory variable for the jth subject

y_j is the discrete outcome with 2 levels (e.g. response/non-response, dead/alive), for the jth subject.

$$\left(\text{i.e.}\quad y_j = \begin{cases} 0 \text{ for a 'non-response'} \\ 1 \text{ for a 'response'} \end{cases}\right)$$

$\Delta_h(z, z_j)$ is a smooth kernel function as defined in section 4.4.1.

The asymptotic variance of this estimator was given in (2.4) as

$$\hat{\text{var}}(\hat{p}_z) \approx \hat{p}_z(1-\hat{p}_z)\frac{\sum_{j=1}^{n} K\left(\frac{\sqrt{2}(z-z_j)}{h}\right)}{\left(\sum_{j=1}^{n} K\left(\frac{z-z_j}{h}\right)\right)^2}$$

This estimator deals with a ***fixed*** point in time and hence does not allow for the fact that in survival problems the status (i.e. the binary outcome) of each subject will ***change*** through time. Therefore the effective difference in a survival problem is that the probability of a response (i.e. y = 1) depends on ***both*** the covariate, z, ***and*** time, t. The following approach attempts to derive time dependent estimates of survival based on the idea devised by Copas. Let the probability of a subject with covariate value z having a response (i.e. y = 1) at time t be estimated as follows

$$\hat{P}(Y=1/z,t) \;=\; \frac{\sum_{j=1}^{n}\Delta_h(z,z_j)y_j(t)}{\sum_{j=1}^{n}\Delta_h(z,z_j)} \qquad \text{- (4.16)}$$

where

$$y_j(t) = \begin{cases} 1 & \text{if subject j has a response at time t} \\ 0 & \text{if subject j has a non - response at time t} \end{cases}$$

(i.e. $y_j(t)$ is the status of the j'th subject ***at*** time t)

In this section the aim has been the provision of non-parametric estimates of the survivor function in the presence of a continuous covariate, i.e. estimates of S(t/z). In (4.16) a non-parametric estimate of the probability of being alive at a particular point in time given a covariate value is proposed. However, it seems logical that the probability of being alive at a particular point in time should be equivalent to the probability of surviving past that point in time which, by (4.0), is the definition of the survivor function. Therefore (4.16) may present a sensible alternative method of estimating S(t/z) as follows:-

$$\hat{S}(t/z) \;\equiv\; \hat{P}(Y=1/z,t) \;=\; \frac{\sum_{j=1}^{n}\Delta_h(z,z_j)y_j(t)}{\sum_{j=1}^{n}\Delta_h(z,z_j)}$$

The asymptotic variance of this estimator can be obtained by direct comparison with the asymptotic variance of the estimator derived by Copas giving

$$\hat{var}\left(\hat{S}(t/z)\right) = \hat{S}(t/z)\left(1-\hat{S}(t/z)\right)\frac{\sum_{j=1}^{n} K\left(\frac{\sqrt{2}\left(z-z_j\right)}{h}\right)}{\left(\sum_{j=1}^{n} K\left(\frac{z-z_j}{h}\right)\right)^2}$$

To produce confidence intervals for S(t/z/), assume $\log_e\{-\log_e(S(t/z))\}$ to be asymptotically normal. A Taylor expansion then produces the following approximate variance for $\log_e\{-\log_e(S(t/z))\}$

$$\hat{var}\left\{\log_e\left(-\log_e\left(\hat{S}(t/z)\right)\right)\right\} \approx \left[\frac{1}{\log_e \hat{S}(t/z) * \hat{S}(t/z)}\right]^2 * \hat{var}\left(\hat{S}(t/z)\right)$$

giving an approximate interval for $\log_e\{-\log_e(S(t/z))\}$ of the form

$$\log_e\left(-\log_e\left(\hat{S}(t/z)\right)\right) \pm 1.96 * \left[\frac{1}{\log_e \hat{S}(t/z) * \hat{S}(t/z)}\right] * \sqrt{\hat{var}\left(\hat{S}(t/z)\right)}$$

$$= [a \ , \ b]$$

Hence an induced approximate 95% confidence interval for S(t/z) is of the form

$$\left[\exp(-\exp(b)) \quad , \quad \exp(-\exp(a))\right]$$

This non-parametric logistic survival approach will therefore produce an estimate of survival based on extending the non-parametric logistic model to allow for the pattern of survival through time.

Section 4.5: Illustration of the Non-Parametric Approaches to Survival Analysis.

In this section further analysis will be carried out on the subgroup of data from the Scottish Melanoma Group database described in section 4.3. Here interest is primarily in examining the data to see if sensible simplifying categorisations can be found for any continuous covariates which have a significant effect on survival prognosis. Initially the three non-parametric methods of estimating survival in the presence of a covariate outlined in section 4.4 will be produced for this data set.

When examining each of the three suggested methods of producing non-parametric estimates of survival in the presence of a covariate consideration will be given to the effect that ***tumour thickness*** has on survival prospects. In studies into survival from stage 1 melanoma the continuous covariate, tumour thickness, is sometimes categorised before any analysis. This is only sensible if the categorisation employed had meaningful implications for prognosis. However the discrepancies between studies as to the actual location of these categorisation point(s) suggest some problems with the reasoning used to locate sensible choices for such point(s). For example Szymik and Woolley (1992) categorised tumour thickness into two groups (<1.70 mm and ≥ 1.70 mm) whereas Rigel et al (1991) had 4 groups (0-0.85mm, 0.85-1.69 mm, 1.7 -3.59 mm and >3.6 mm) whilst Keefe and MacKie (1991) went as far as to discuss 8 different groupings. The three techniques for analysing

survival data presented here will use purely data fitting techniques to allow any such categorisations to be highlighted. These methods will hopefully allow sensible choices for categorisation point(s) by relying on the data itself to highlight any areas where there are marked changes in survival indicating the location of a potential categorisation point.

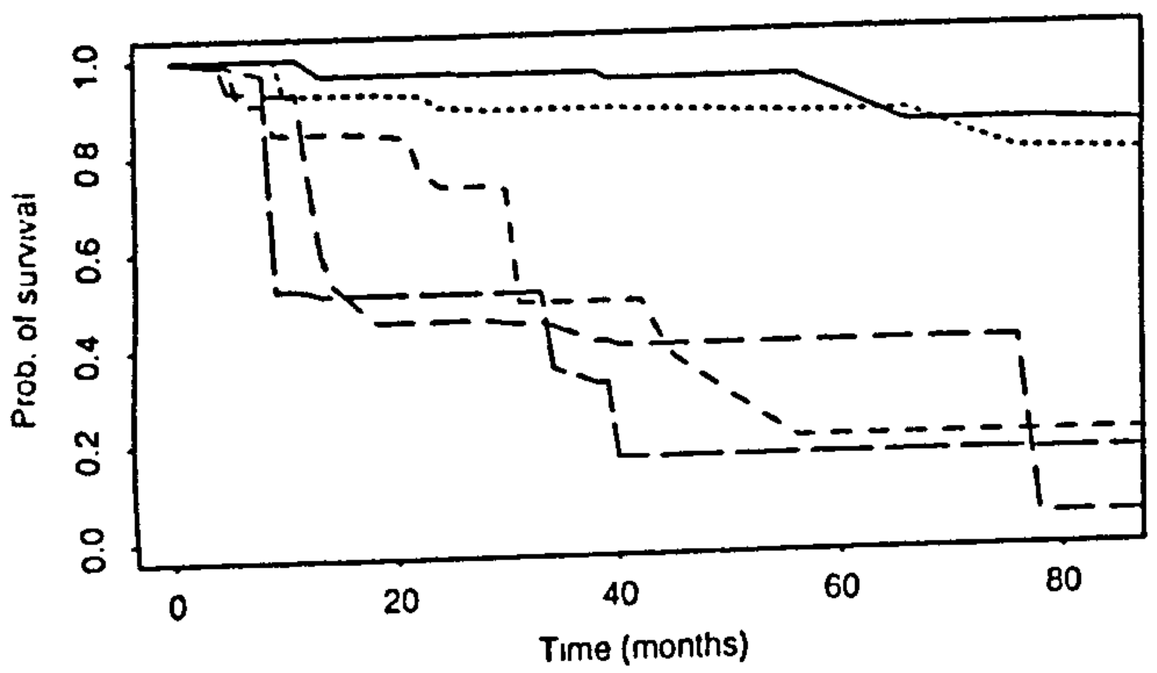
Section 4.5.1: Melanoma Example: Kaplan-Meier based approach

Recall that the data set being considered consists of ***females with ulcerated lesions on an axial site***. In Section 4.3 a Cox proportional hazards model was fitted to incorporate the effect of tumour thickness on survival prospects. Figure 4.3.3 presented a graphical display of the fitted model at the selected tumour thicknesses of 1,3,5,7 and 9 mm. Here the Kaplan Meier based approach incorporating a covariate will be applied to the data set.

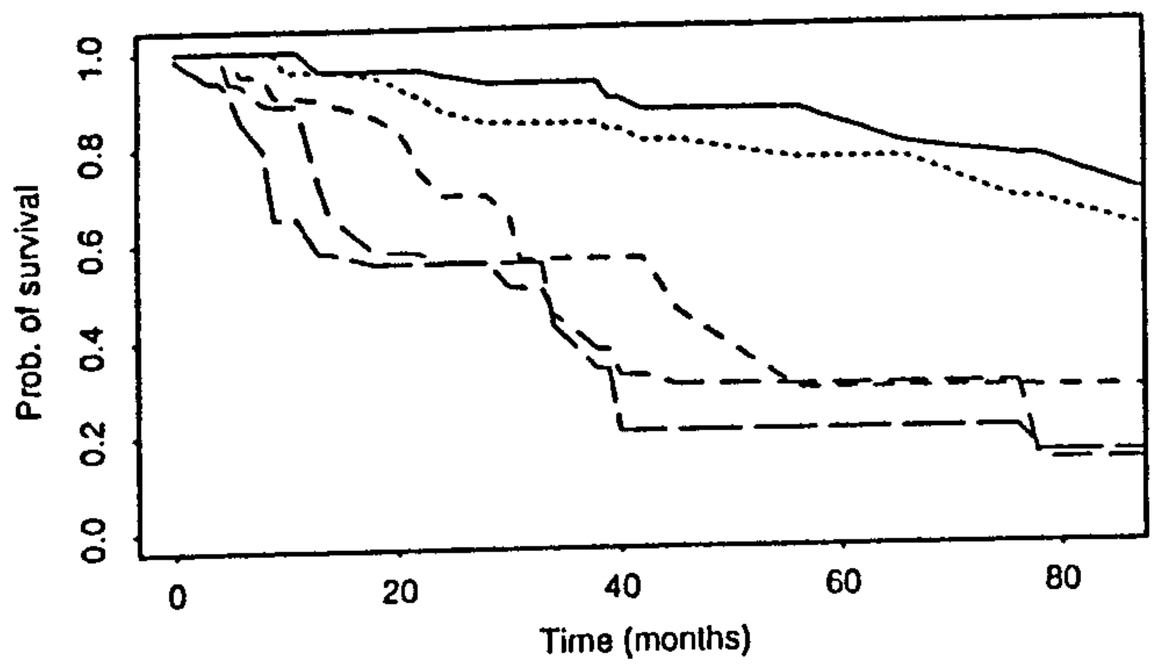
Figure 4.5.1 displays estimates of the survivor function based on a selection of values for the smoothing parameter in (4.6). This figure iillustrates the effect the choice of smoothing parameter can have on the interpretation of the results. In frame 1 of Figure 4.5.1 there is gross ***undersmoothing*** of the data leading to a very confused picture of what is happening across the levels of the covariate. Conversely frame 9 demonstrates what happens if the data is ***oversmoothed***. A comparison of frame 9 of Figure 4.5.1 with Figure 4.3.3 leads to the conclusion that if this 'Kaplan

Females with Ulcerated Lesions on an axial site

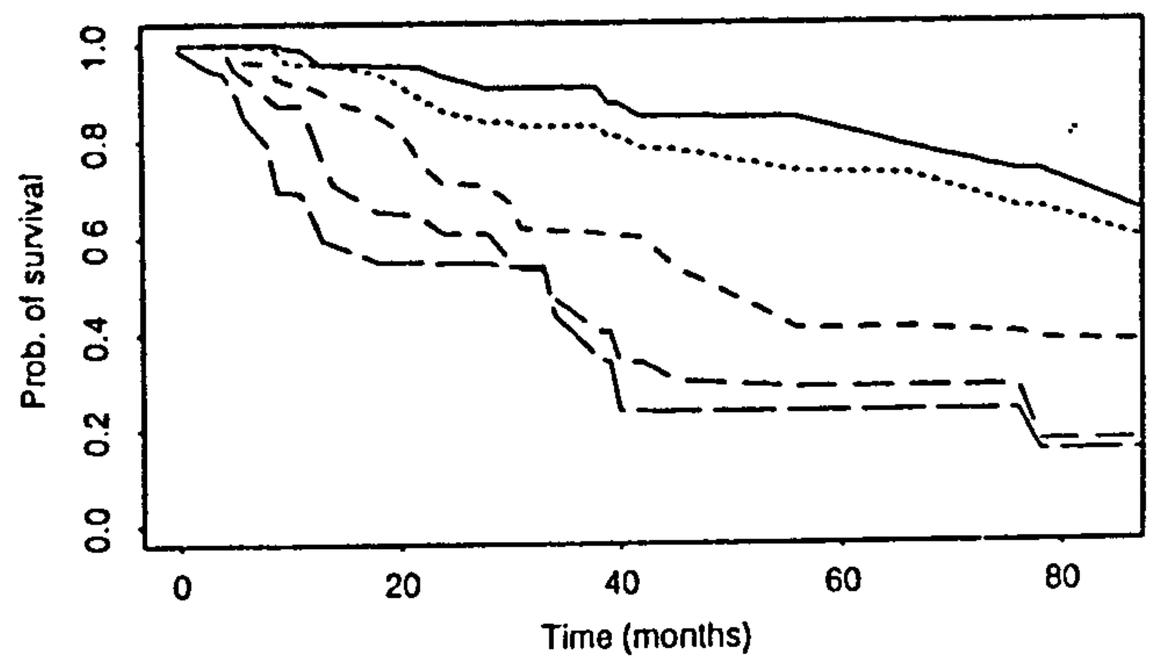
Smoothing parameter is 0.37



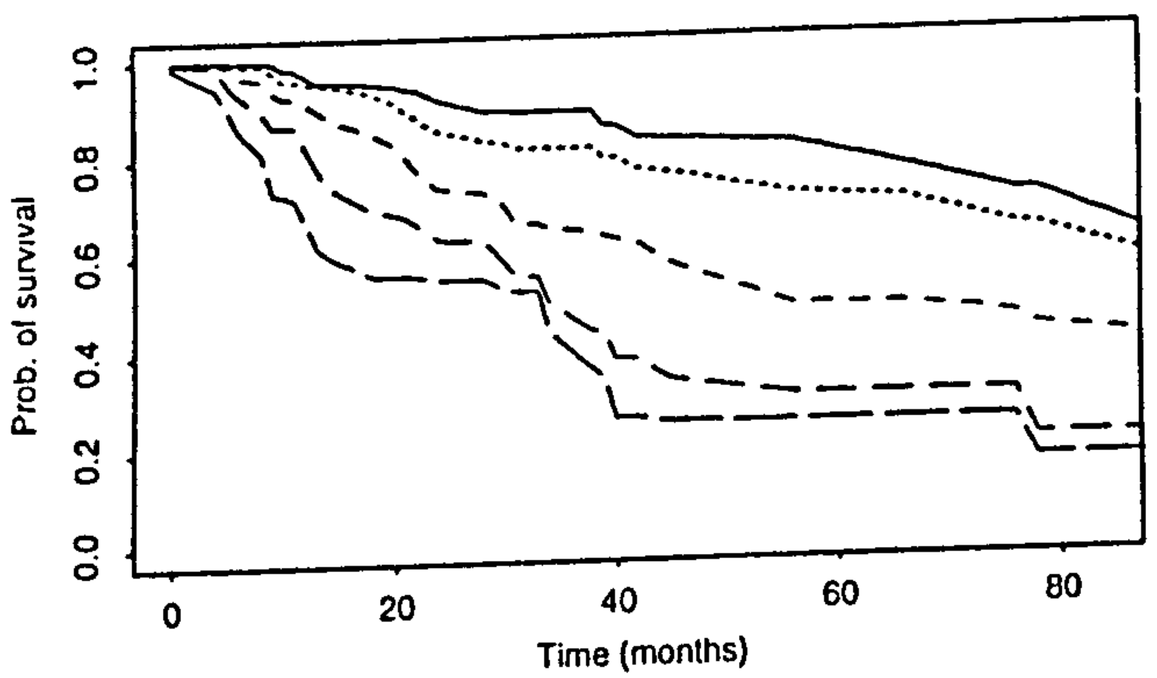
Smoothing parameter is 0.79



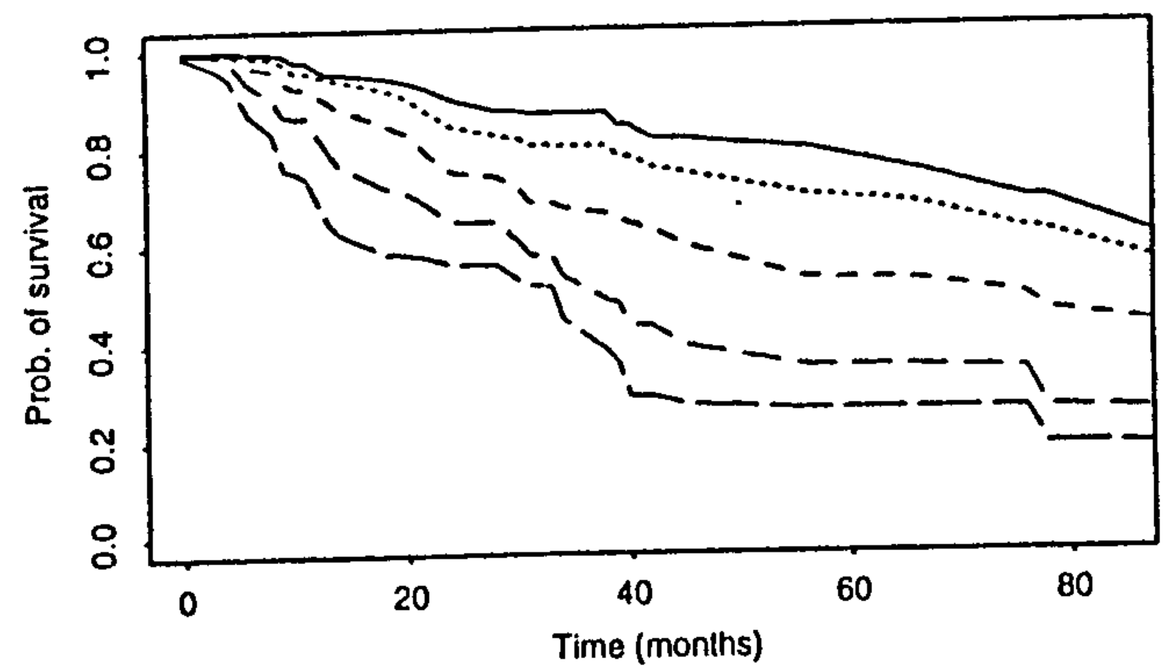
Smoothing parameter is 1.2



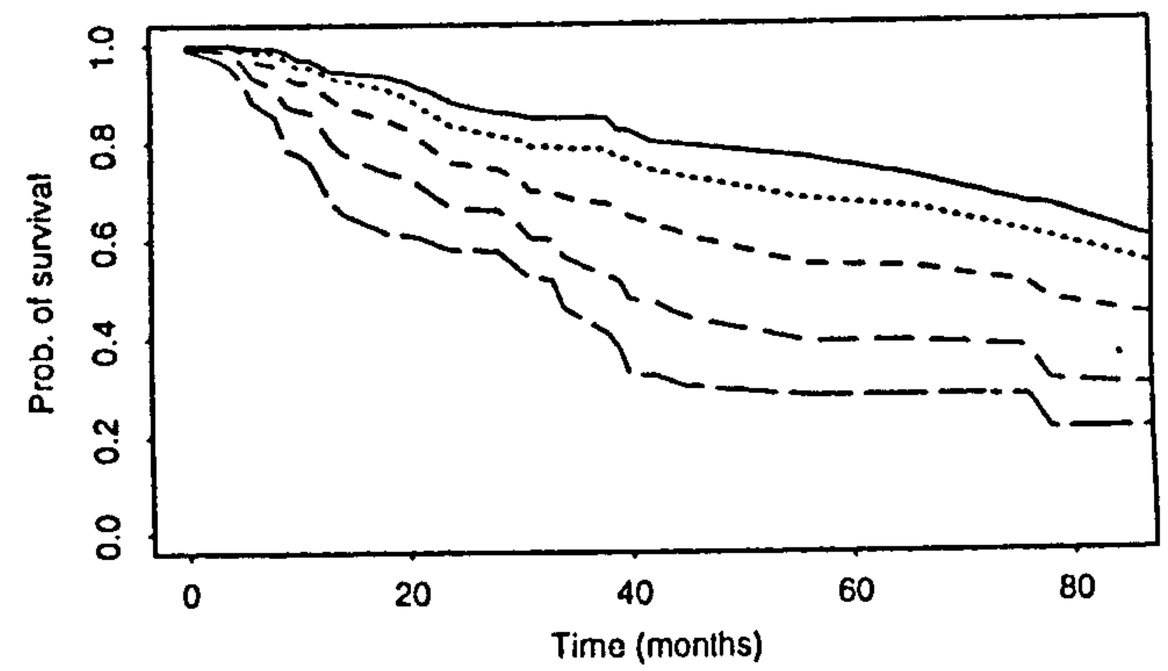
Smoothing parameter is 1.6



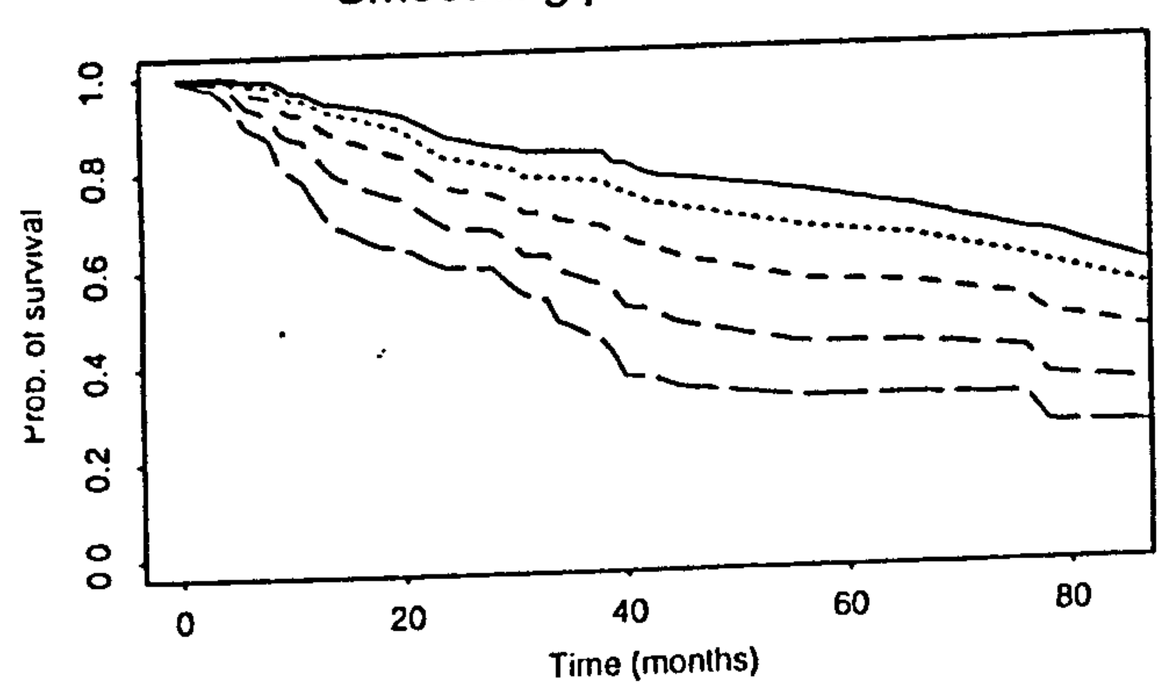
Smoothing parameter is 2



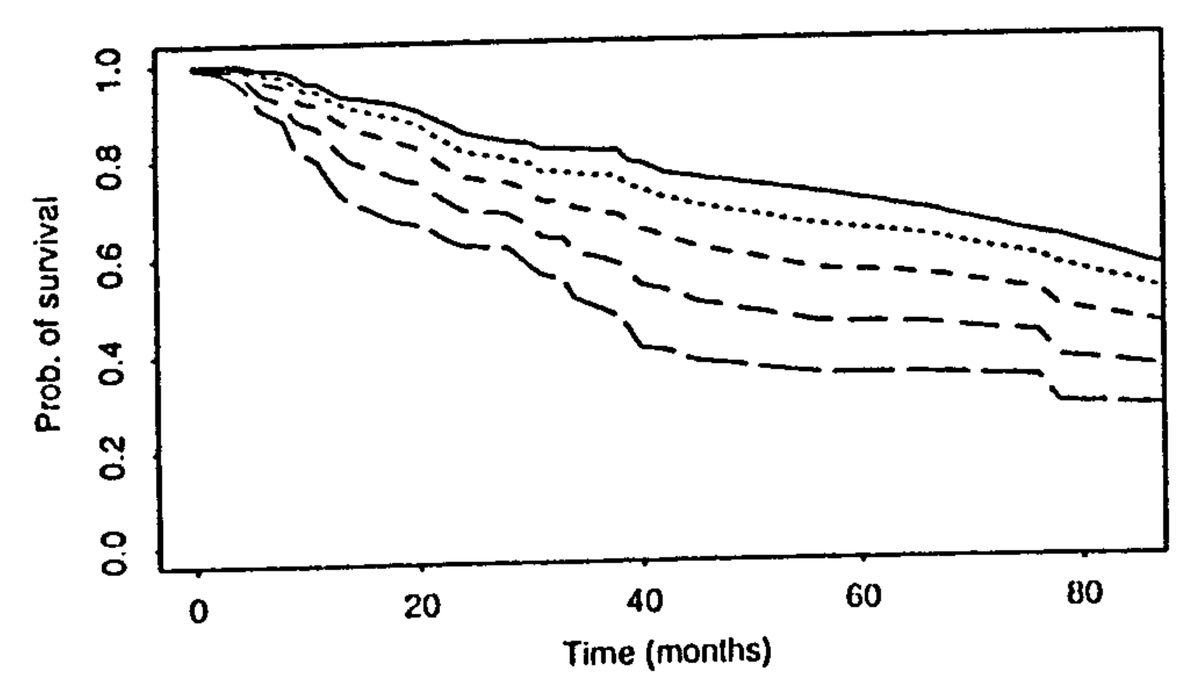
Smoothing parameter is 2.5



Smoothing parameter is 2.9



Smoothing parameter is 3.3



Smoothing parameter is 3.7

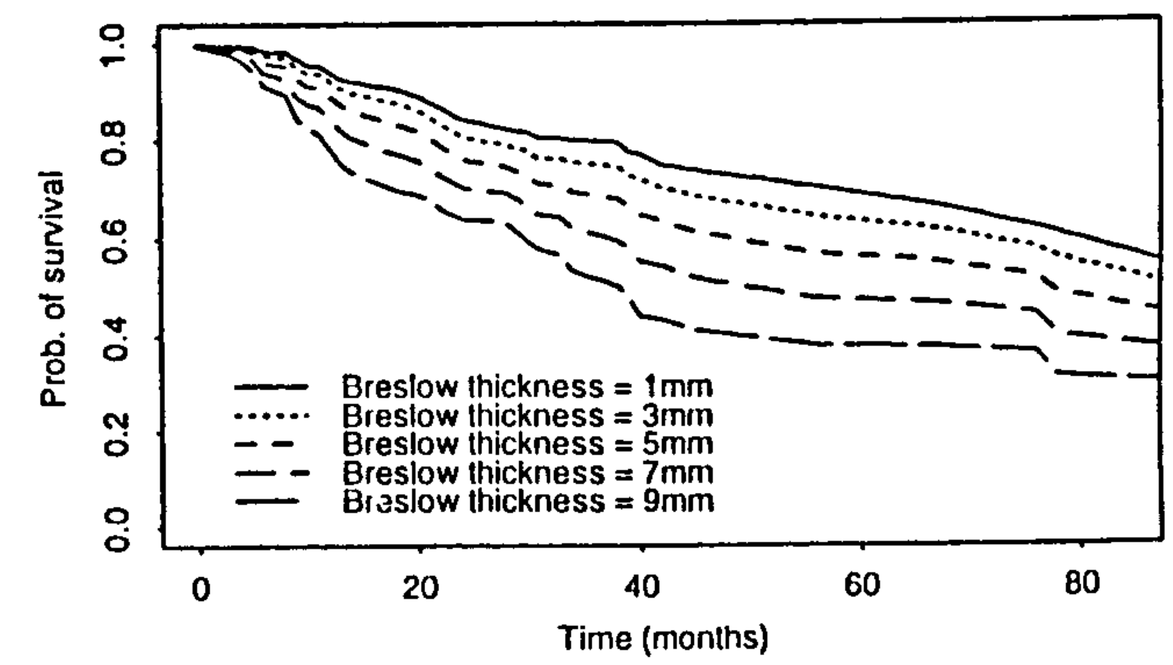


Figure 4.5.1

Meier style' estimator is oversmoothed it will produce estimates of survival which are similar to those produced by the proportional hazards model. A 'sensible' choice of smoothing parameter appears to be around 1.5 to 2 (i.e. frame 4 or frame 5 of Figure 4.5.1). This choice of smoothing parameter leads to the conclusion that the proportionality assumption inherent in the proportional hazards model fitted in section 4.4 may not be reasonable given the representation of the underlying trend in this data set. It appears that instead of a steady decrease in survival prospects across the tumour thickness as implied by the proportional hazards model there is in fact a sharp drop in survival between thicknesses of 3 and 7 mm. Estimates of survival obtained from the proportional hazards model seem reasonable up to about 3 mm but greater than 3 mm the proportional hazards model appears to be overestimating the probability of survival.

Figure 4.5.2 displays a contour plot for the survivor function given by choosing a value for the smoothing parameter of 2 as suggested by Figure 4.5.1. This allows a clearer picture of the pattern of survival to be obtained. The contour plot suggests that for tumour thicknesses less than about 3 mm survival prospects drop off at a ***relatively slow rate*** through time and, in general, survival is reasonably good for these values of tumour thickness. However, for tumour thicknesses between 3 mm and, about 7 mm there is evidence that survival drops off ***more rapidly*** in the earlier months before levelling off. Finally, for tumour thicknesses greater than 7 mm survival drops off ***very rapidly*** in the early months and, overall, survival is relatively poor for these values of tumour thickness. In terms of

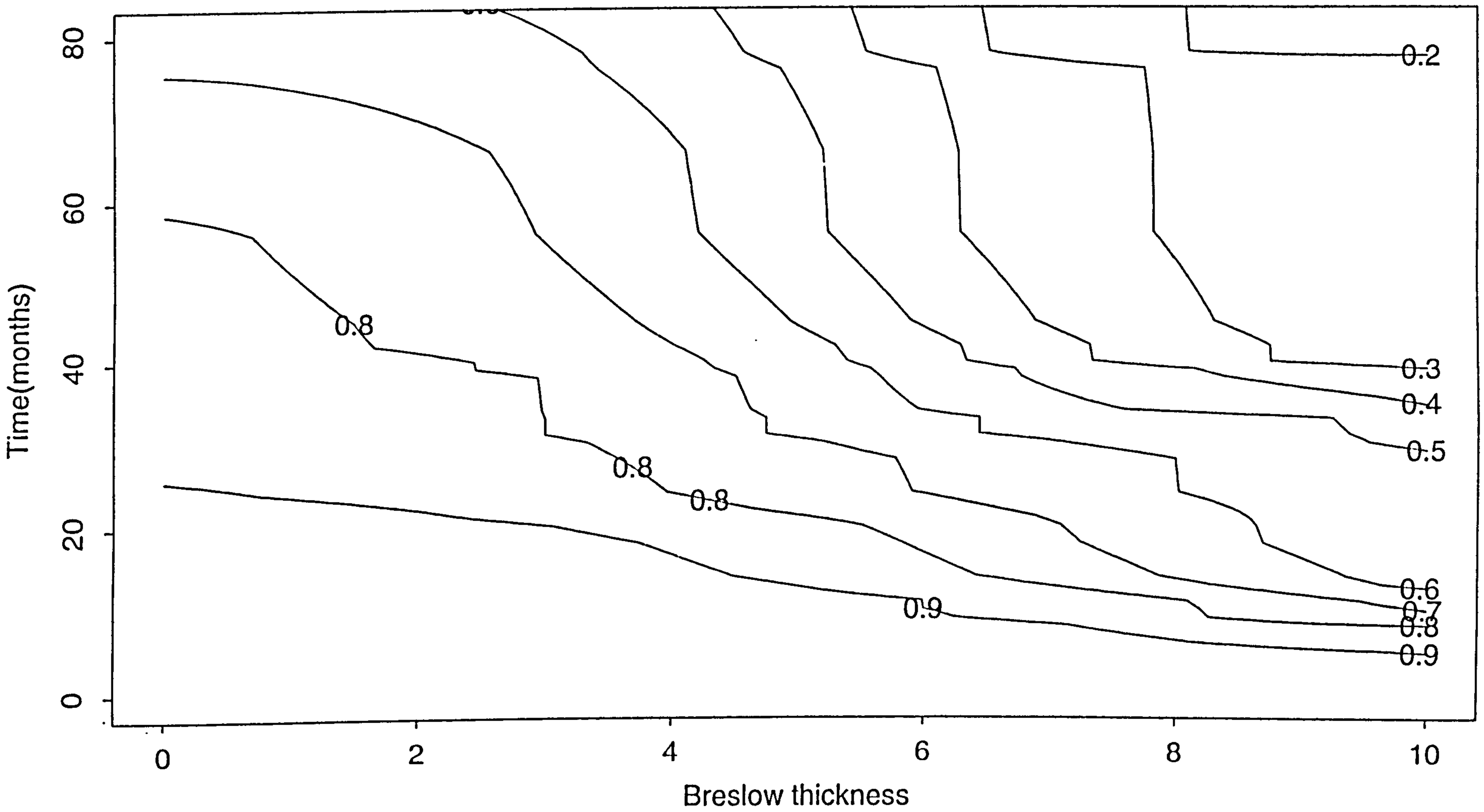

Females with Ulcerated Lesions on an axial site - Survival Contours
Time(months)
0
20
40
60
80
0
2
4
6
8
10
Breslow thickness
0.8
0.8
0.8
0.9
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

Figure 4.5.2

producing categorisations for this variable these results would perhaps suggest that two categorisation points exist; firstly at a tumour thickness of approximately 3 mm and then later at around 7 mm.

Confidence bands can also be produced for the individual values of tumour thickness for a specified smoothing parameter. Figure 4.5.3 shows separate confidence bands for each of the 5 important values of tumour thickness for the chosen smoothing parameter of 2. These confidence bands indicate the precision in the estimates of survival and they demonstrate that the most precise estimates of survival are obtained within the first two to three years of follow up and for tumour thicknesses of up to about 5 mm. This corresponds to the five-year follow-up pattern of the SMG and the sad fact that few patients with "thick" tumours survive long after diagnosis.

Section 4.5.2 Melanoma example: Non-parametric Hazard approach

Here the approach based on the hazard function (Section 4.4.2) is applied to the melanoma data set. One important point to notice with this technique is that there are *two* distinct levels of smoothing present: ***one across time and one across the covariate.*** This leads to the diagrammatic representation of results being somewhat complicated. However Figures 4.5.4 (a) and (b) illustrate the effect of ***both*** levels of smoothing on estimates of survival.

Females with Ulcerated Lesions on an axial site

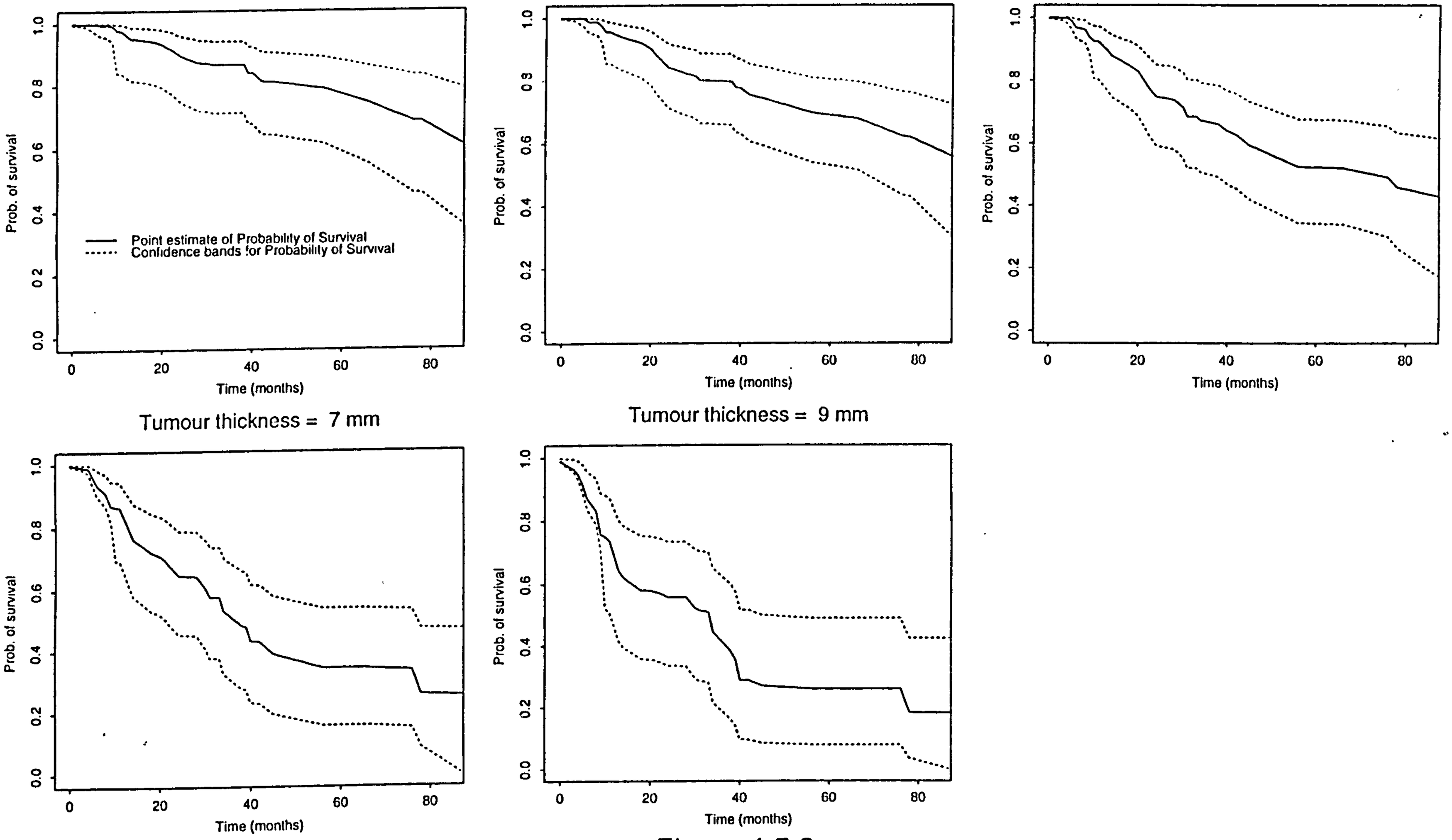


Figure 4.5.3

Females with Ulcerated Lesions on an axial site

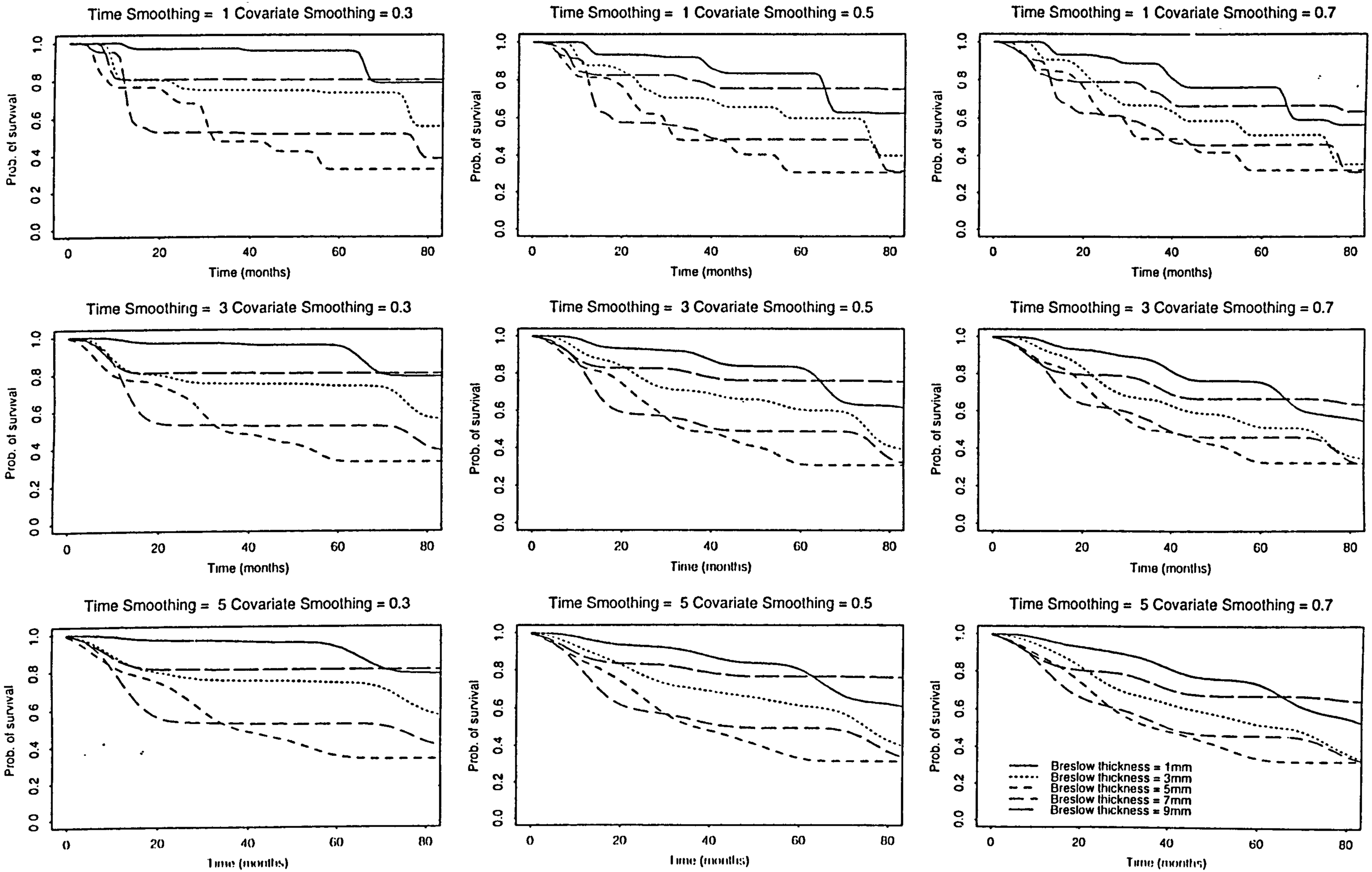


Figure 4.5.4(a)

Females with Ulcerated Lesions on an axial site

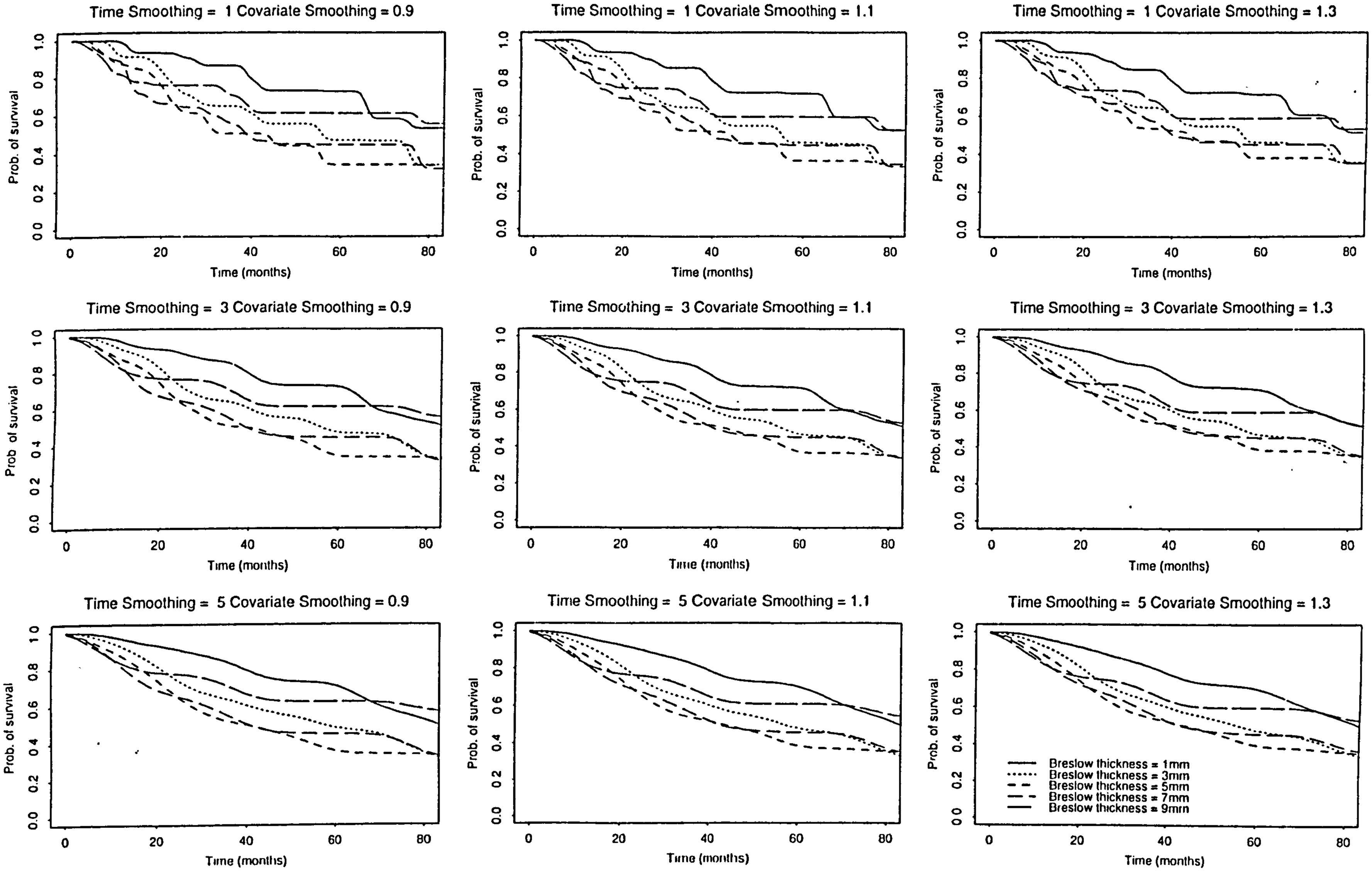


Figure 4.5.4(b)

The smoothing across time does not seem to have a major effect on the estimate of survival as a reasonably similar picture for the pattern of survival is obtained regardless of the "time" smoothing parameter. However the level of smoothing across the covariate has a more marked effect on the estimates of survival. As the level of smoothing across the covariate increases the estimates of survival become much flatter. The undersmoothing present in frame 1 of Figure 4.5.4(a) shows estimates of survival which appear to exhibit a step pattern in nature but by the time frame 9 of Figure 4.5.4(b) is considered the curves representing the different levels of the covariate have all become very flat in nature. In general for larger values of the covariate this method of estimating survival produces estimates which appear slightly higher than those obtained by the method discussed in section 4.5.1. In particular it produces estimates of survival at a Breslow thickness of 9 mm which appear remarkably high. A closer examination of the data plot in Figure 4.3.2 demonstrates that this may be due to a couple of observations which have large values of Breslow thickness (12 mm and 13mm) and have both a long follow-up time and are still alive. These observations may be having undue influence on the estimate of survival. This unusual pattern can be removed but at the expense of smoothing out other features of the data. This would suggest that this technique is quite sensitive to the presence of unusual observations in areas where there is little data present. However, the method appears reasonably satisfactory in areas where the majority of the data is found.

A logical combination of smoothing parameters appears to be around frame 4 of Figure 4.5.4(b) (i.e. a time smoothing value of 3 in conjunction with a covariate

smoothing value of 0.9). Figure 4.5.5 shows a contour plot for this chosen combination whilst Figure 4.5.6 displays the corresponding approximate confidence bands for the survivor function at the five previously selected tumour thicknesses. These plots back up the impression given in section 4.5.1 that the proportionality assumption may not be sufficient to describe the underlying pattern in this data set. This method does however give a slightly different pattern to survival than that observed in section 4.5.1. There is again evidence that survival drops off relatively slowly for smaller values of tumour thickness. However, here there is evidence that for tumour thicknesses between 2.5 mm and 5 mm the drop in survival is steeper than was suggested by the Kaplan Meier method. Between 5 and 8 mm the two methods again suggest similar patterns of survival. Finally, for tumour thicknesses greater than 8 mm the Hazard based approach suggests that survival prospects drop off at a much slower rate and are, in fact, similar to those for tumour thickesses less than about 2.5 mm. Apart from the unexpectedly slow rate of the drop in survival for greater than 8 mm which may be explained by a few "unexpected" observations a comparison of the estimates of survival with the data plot in Figure 4.3.2 does give some credence to these results. Figure 4.3.2 shows few deaths and a large presence of high censored values up to 2 mm suggesting survival prospects will be reasonably good for such patients. This is followed by a number of early deaths among subjects whose tumours are between 2 and 5 mm thick suggesting a steeper decline in survival prospects. The existence of some reasonably high death and censored times

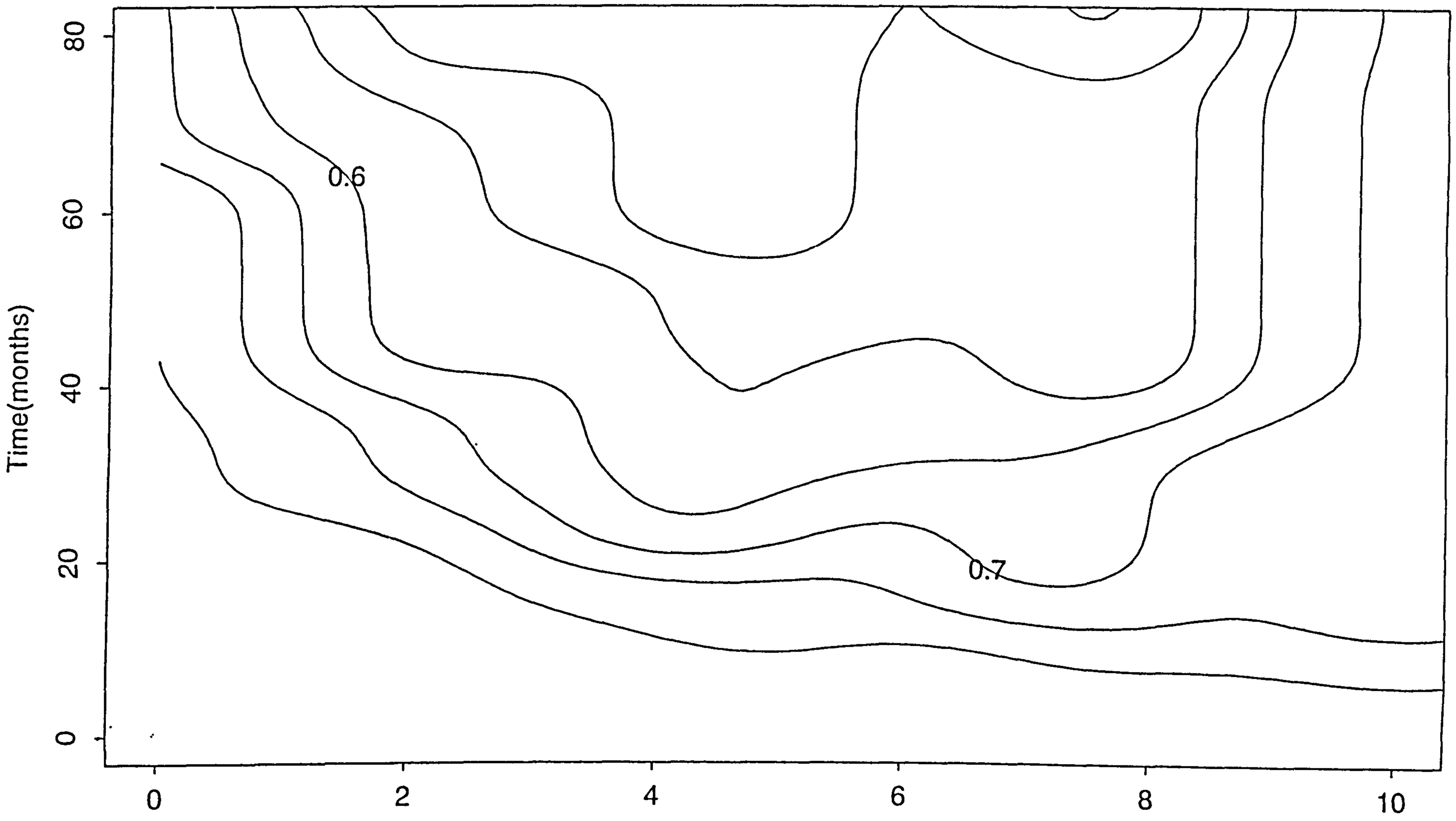
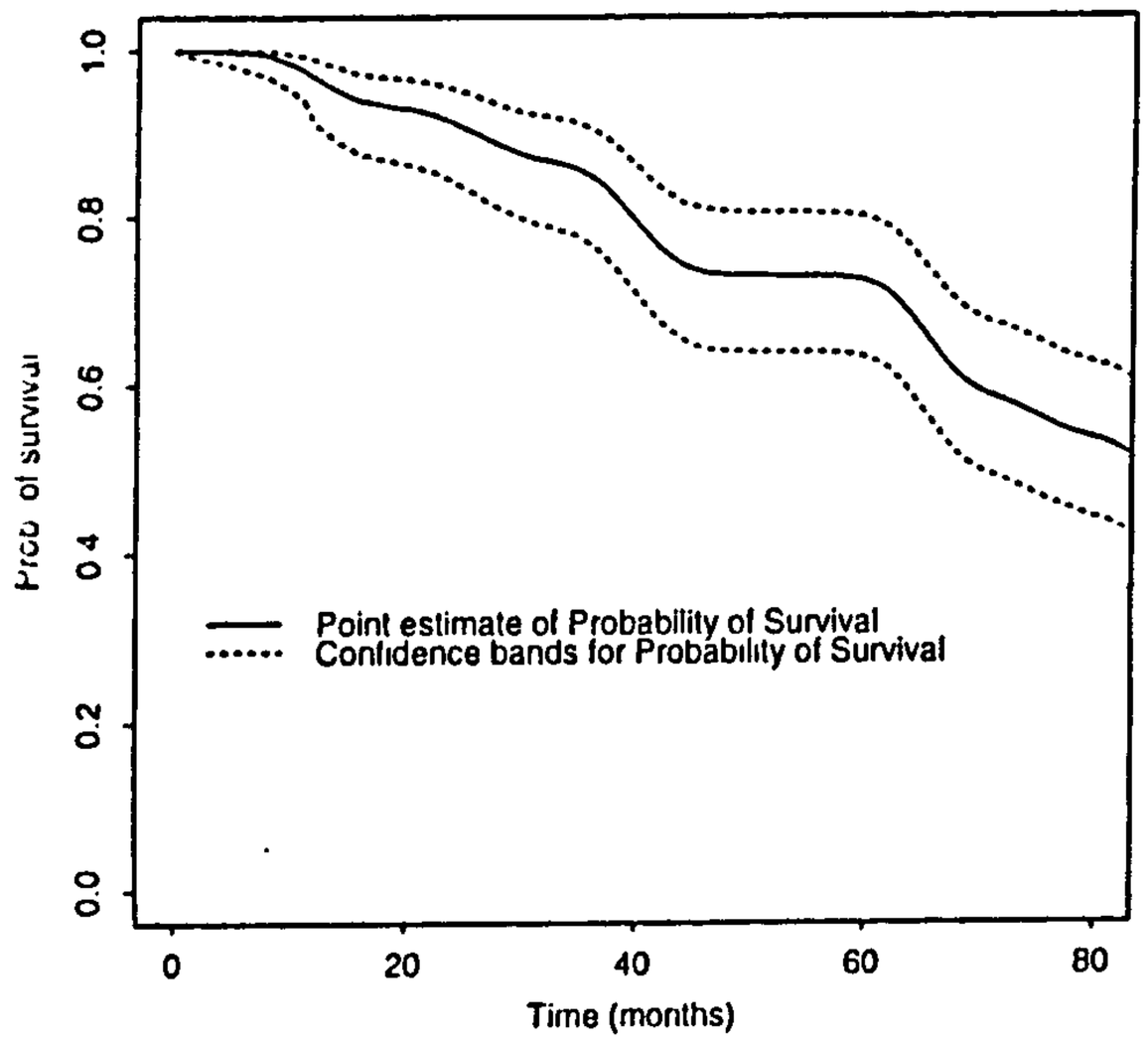

Females with Ulcerated Lesions on an axial site - Survival Contours
Time(months)
0
20
40
60
80
0.6
0.7
0
2
4
6
8
10
Breslow thickness

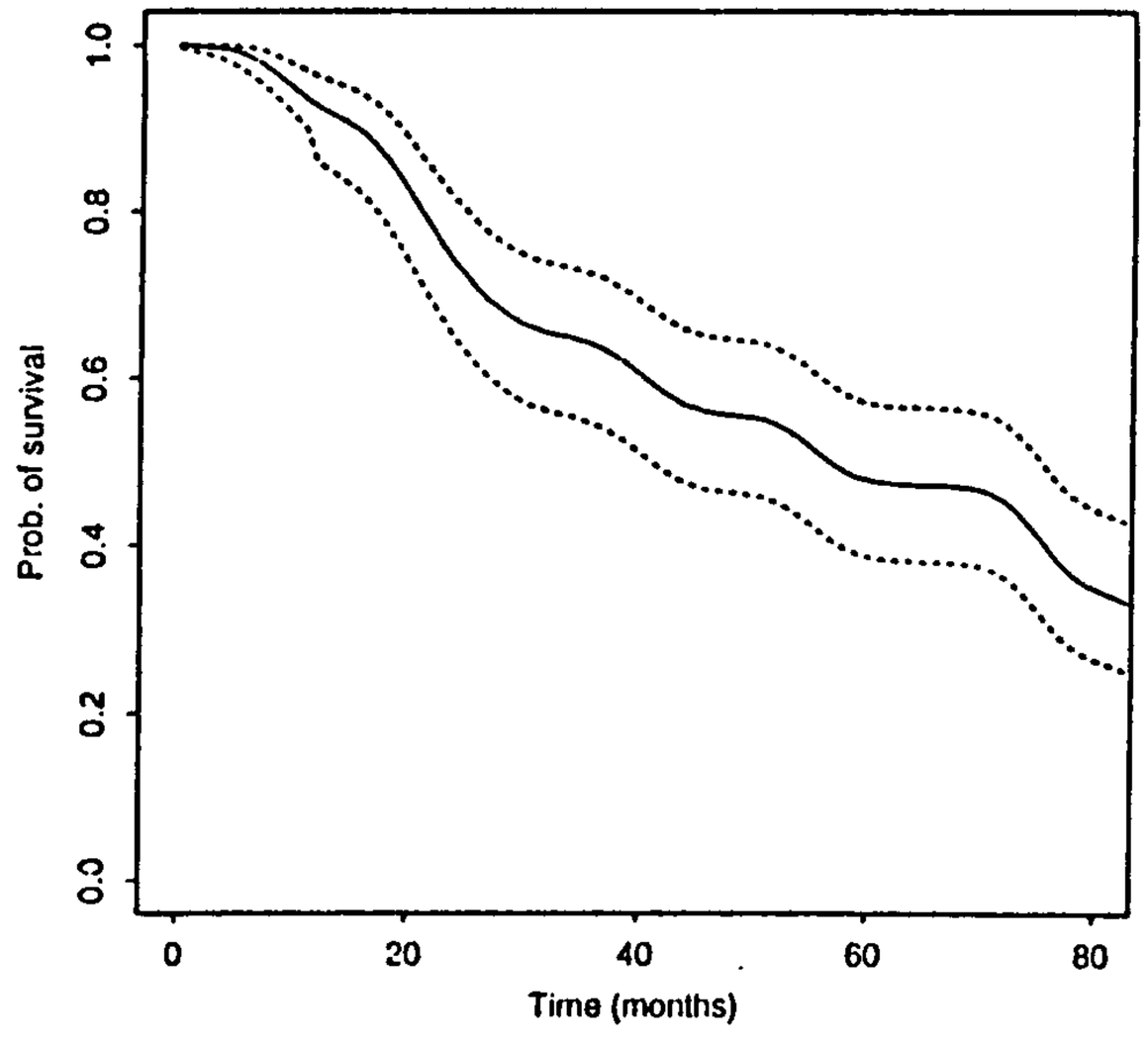
Figure 4 5 5

Females with Ulcerated Lesions on an axial site

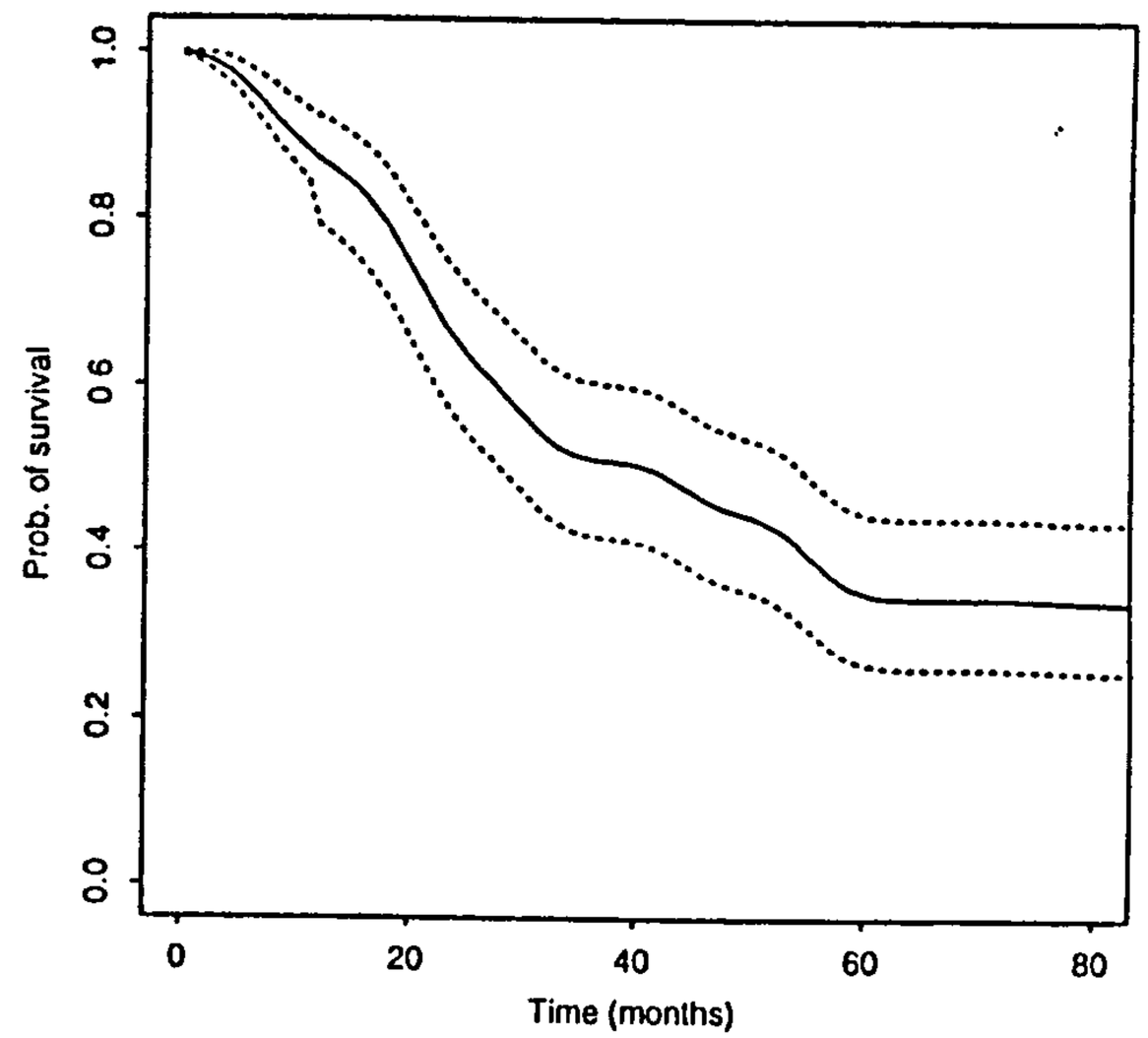
Tumour thickness = 1 mm



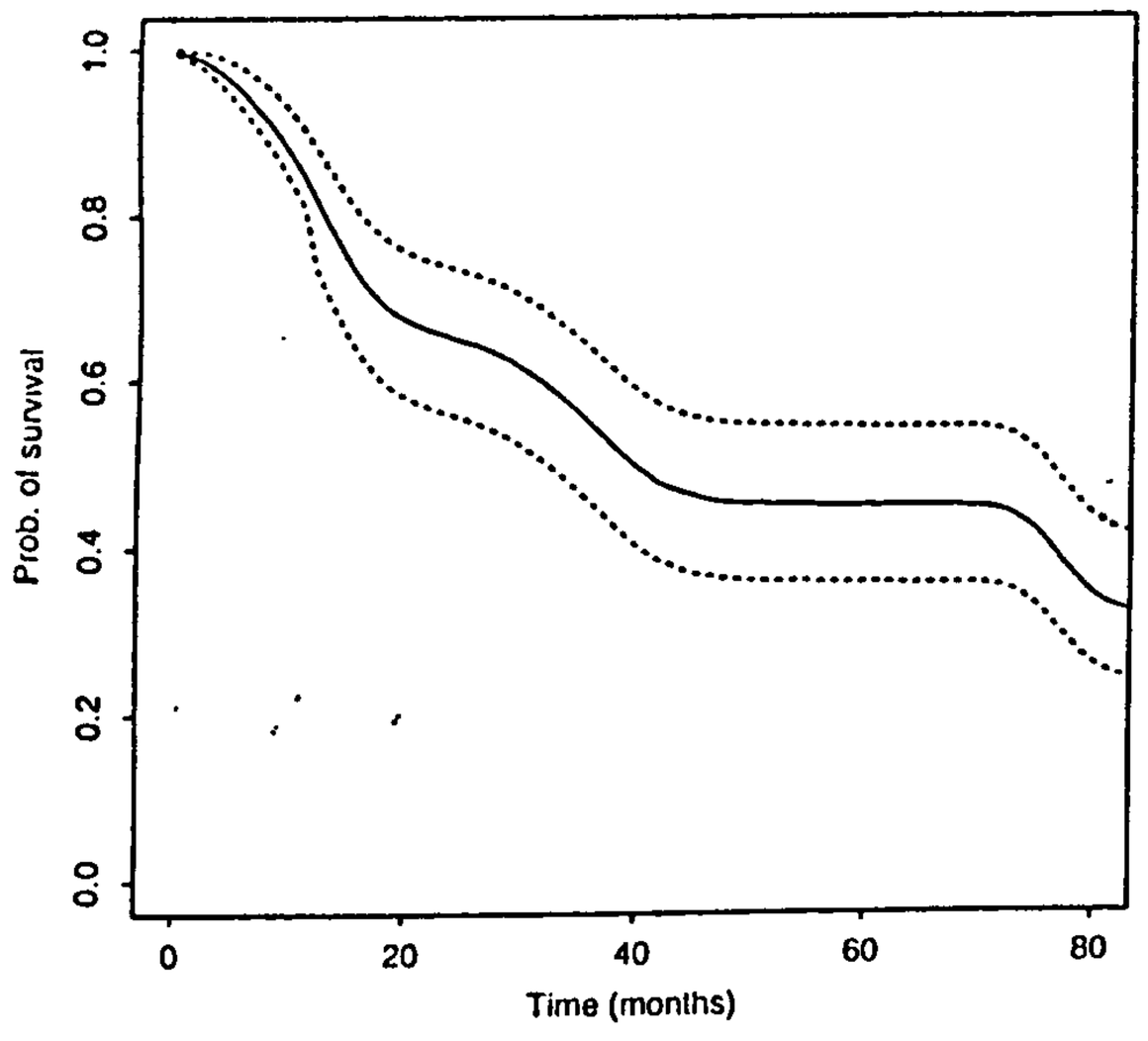
Tumour thickness = 3 mm



Tumour thickness = 5 mm



Tumour thickness = 7 mm



Tumour thickness = 9 mm

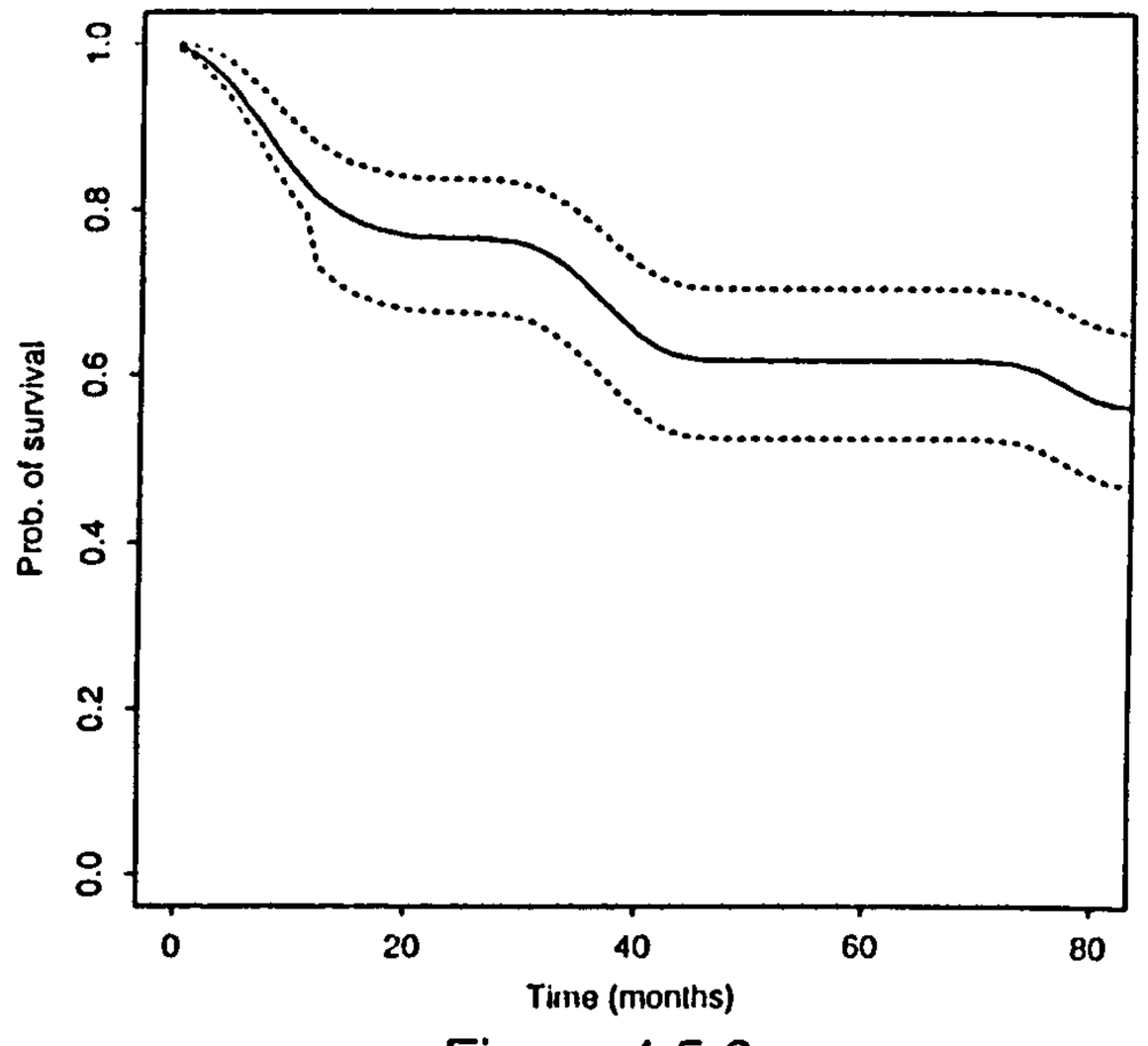


Figure 4.5.6

around 7-8 mm may also suggest that the levelling off of survival prospects is also plausible here.

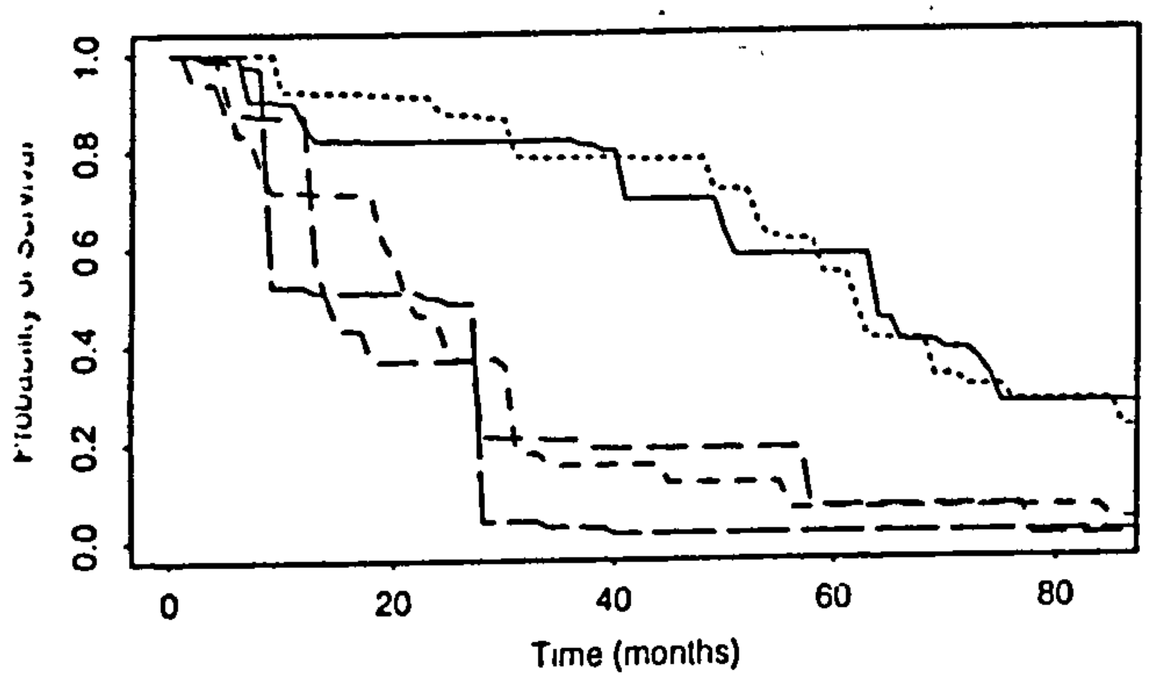
In terms of categorisation points these results would suggest that three categories perhaps exist; the first category being from 0-2.5 mm, the second from 2.5-8mm and the third being >8mm. In the first and third categories estimates of survival exist which suggest little change in survival across the values of the covariate whereas in the second category the estimates of survival appear to drop quite markedly.

Section 4.5.3: Melanoma Example: Non-parametric logistic survival approach

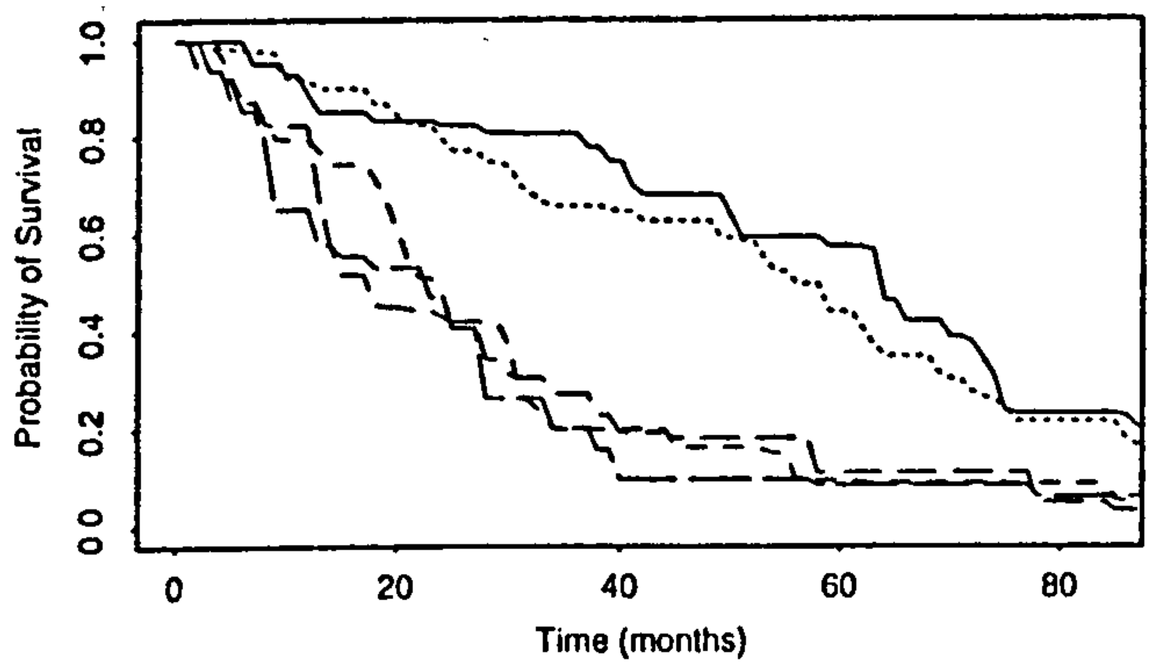
A third non-parametric method of estimating survival was described in section 4.4.3. This method adapted the standard non-parametric logistic approach to incorporate the time-dependent survival. This method also involves the use of one smoothing parameter and a subjective search will again be used to choose an appropriate smoothing value. Figure 4.5.7 shows survival estimates obtained for a range of smoothing values while Figures 4.5.8 and 4.5.9 display, respectively, a contour plot and confidence bands for the resulting 'optimal' choice of smoothing parameter. This method again highlights the slower drop in survival prospects that is present for smaller tumour thicknesses. It also shows that survival prospects drop

Females with Ulcerated Lesions on an axial site

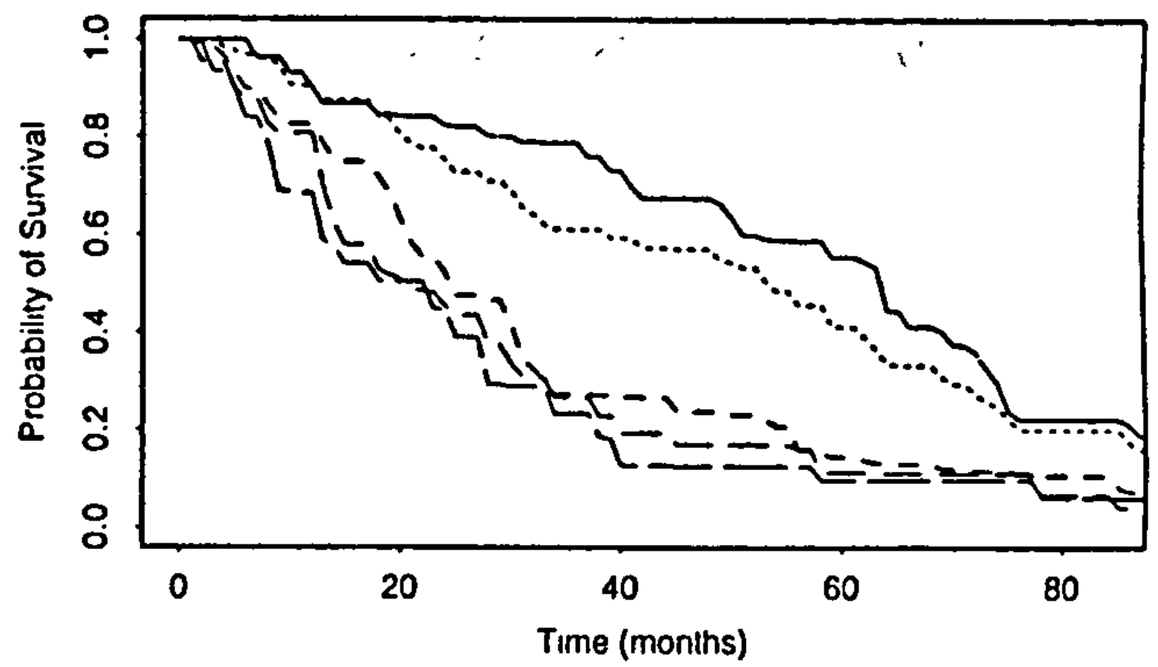
Smoothing parameter is 0.37



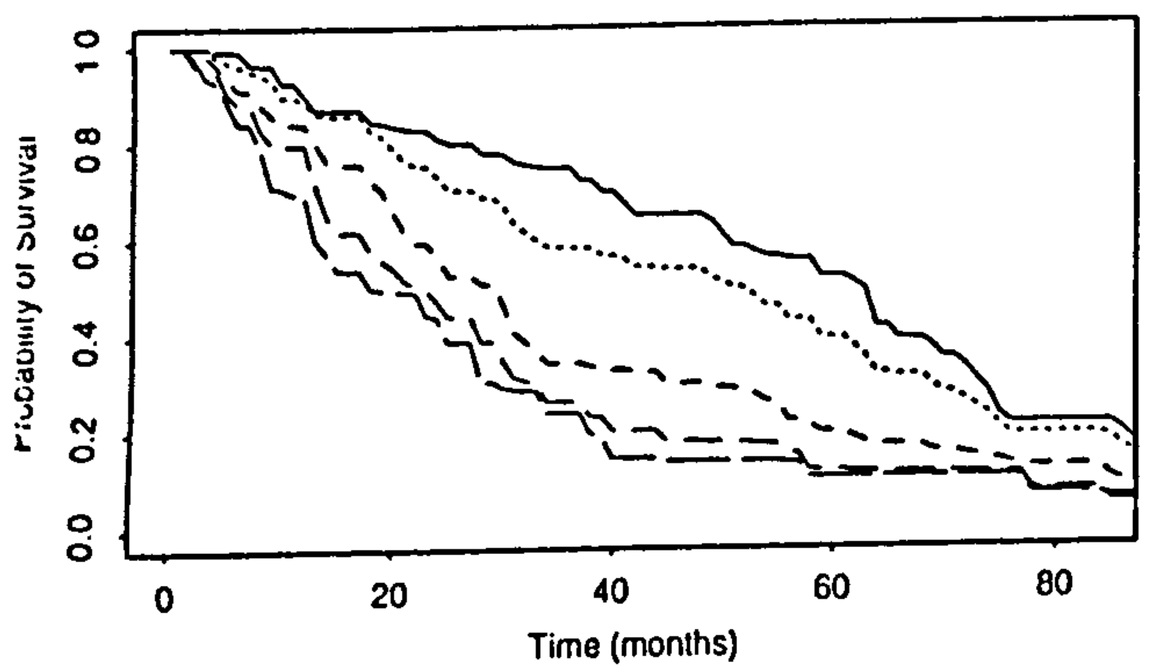
Smoothing parameter is 0.79



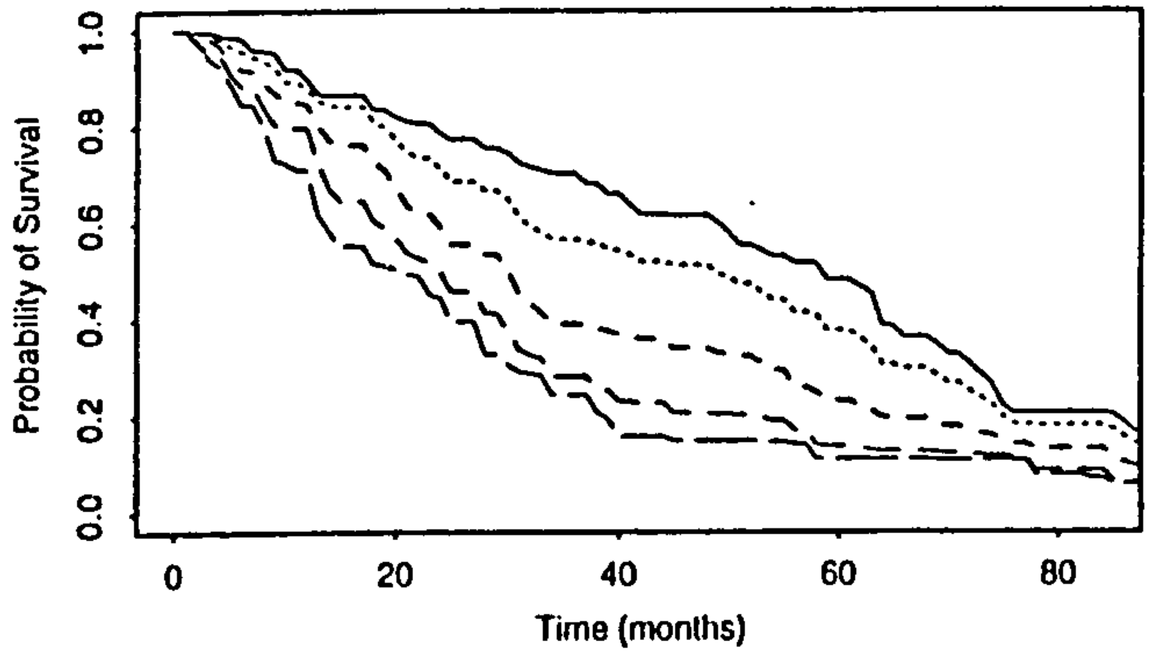
Smoothing parameter is 1.2



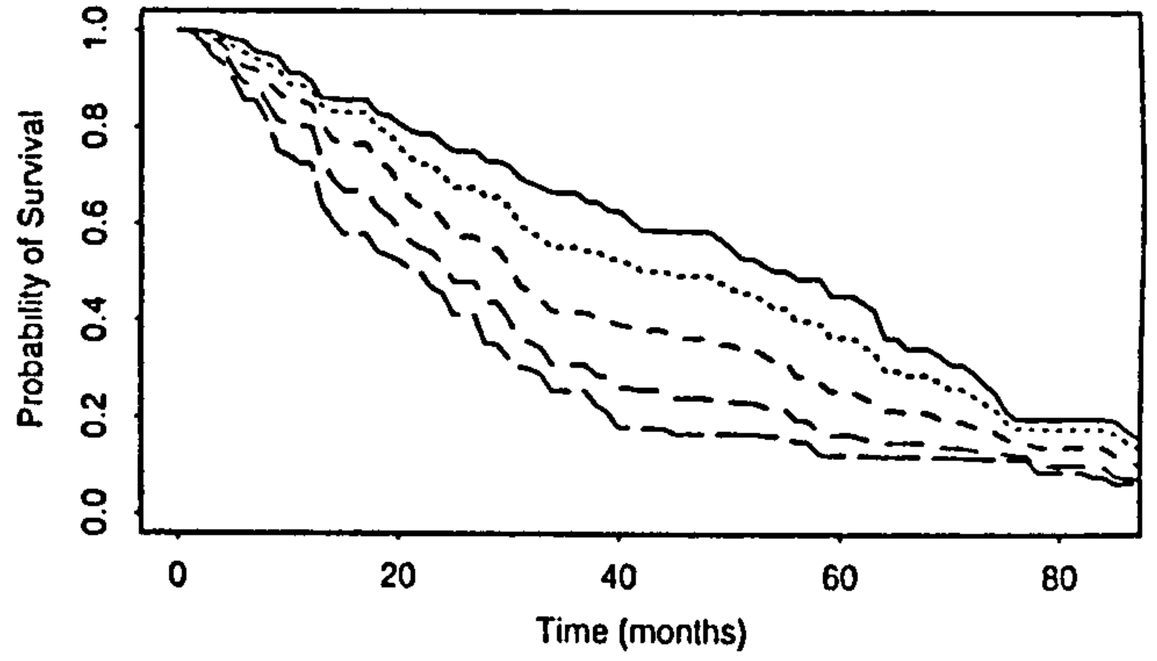
Smoothing parameter is 1.6



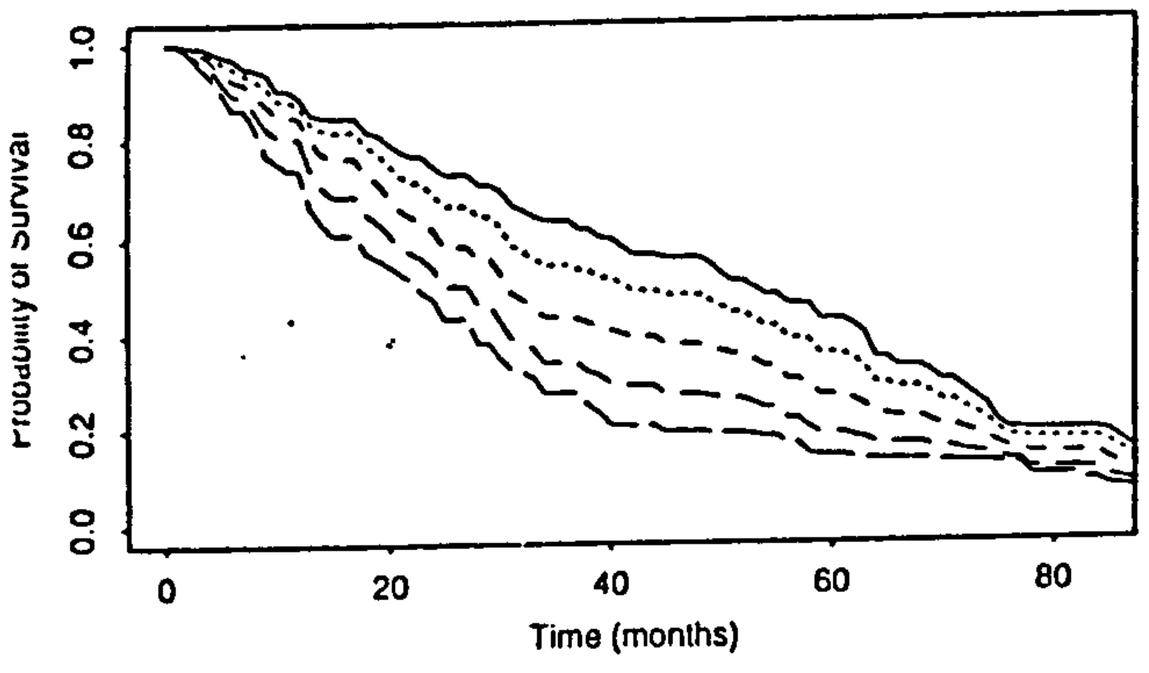
Smoothing parameter is 2



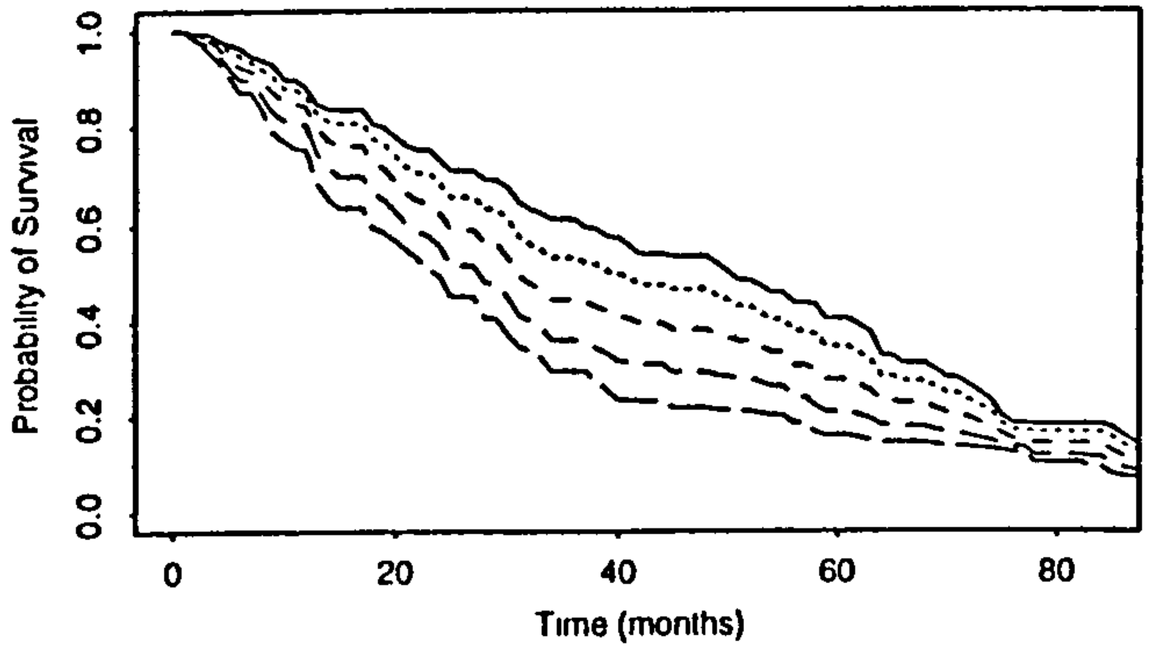
Smoothing parameter is 2.5



Smoothing parameter is 2.9



Smoothing parameter is 3.3



Smoothing parameter is 3.7

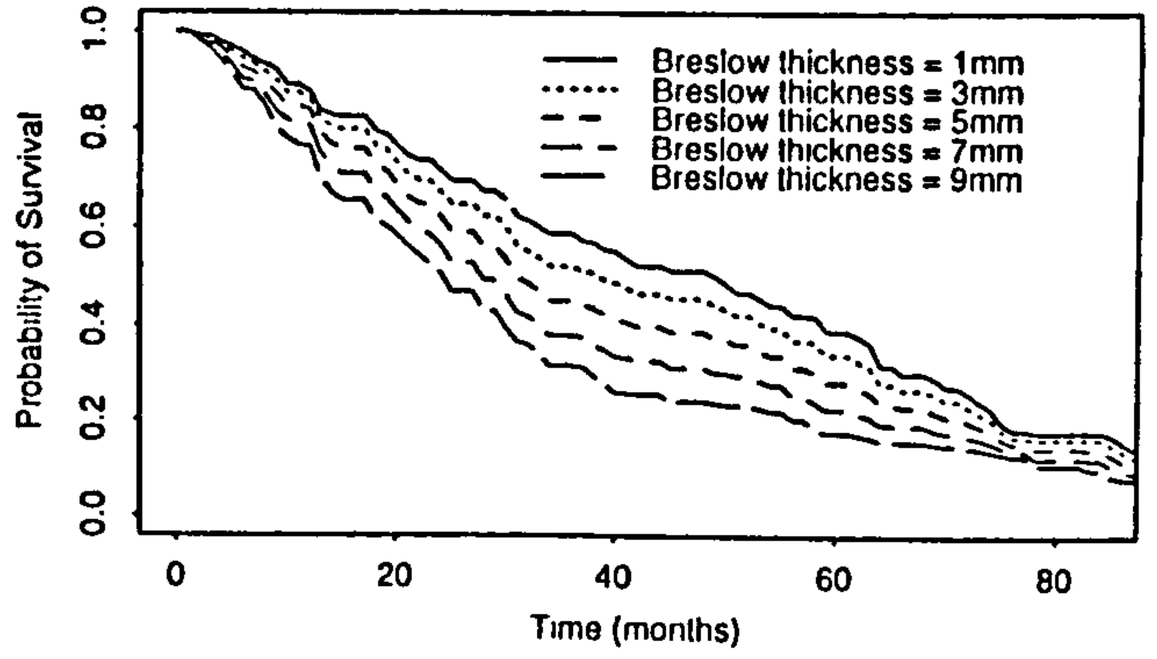


Figure 4.5.7

Females with Ulcerated Lesions on an axial site - Survival Contours

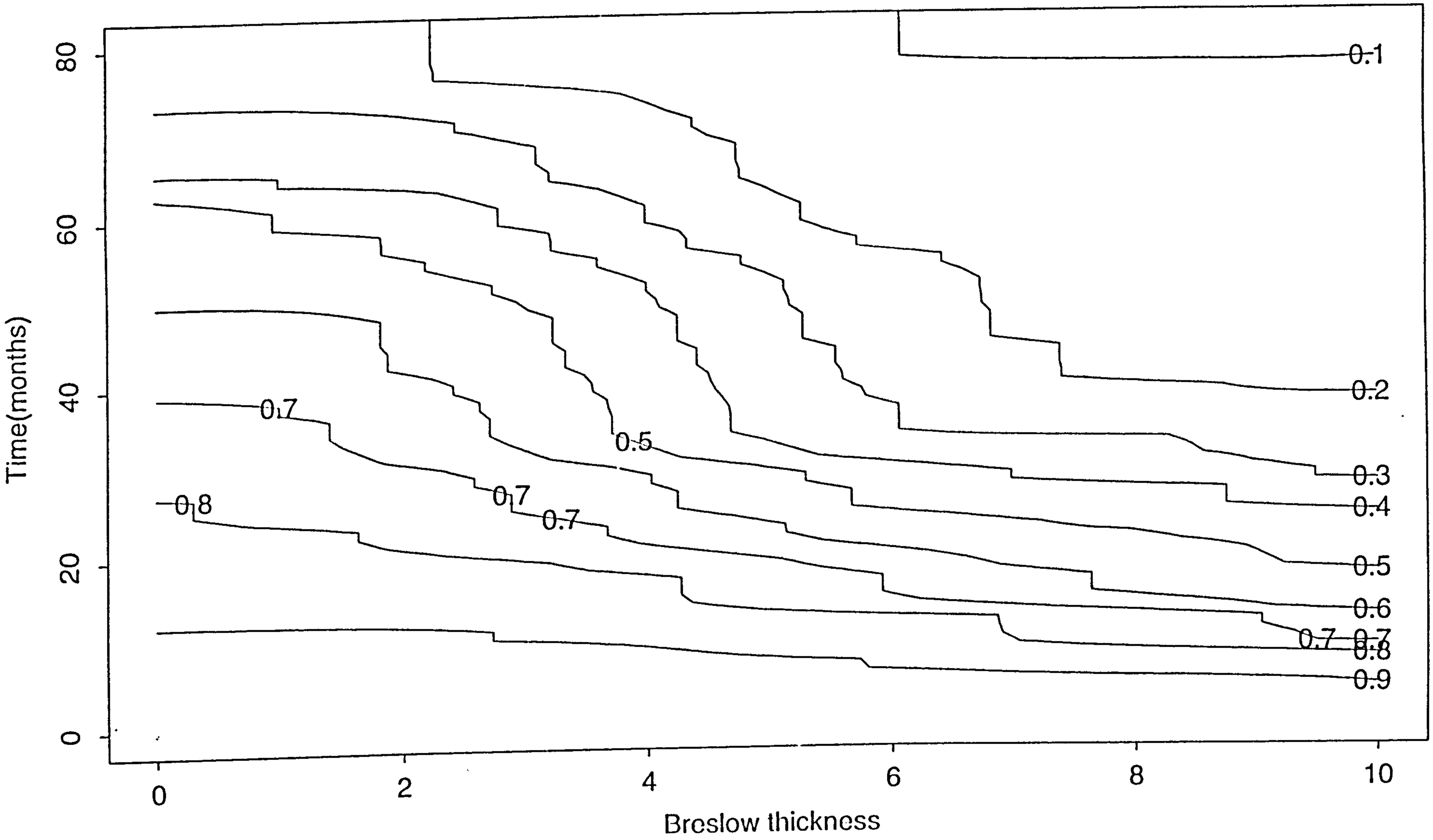
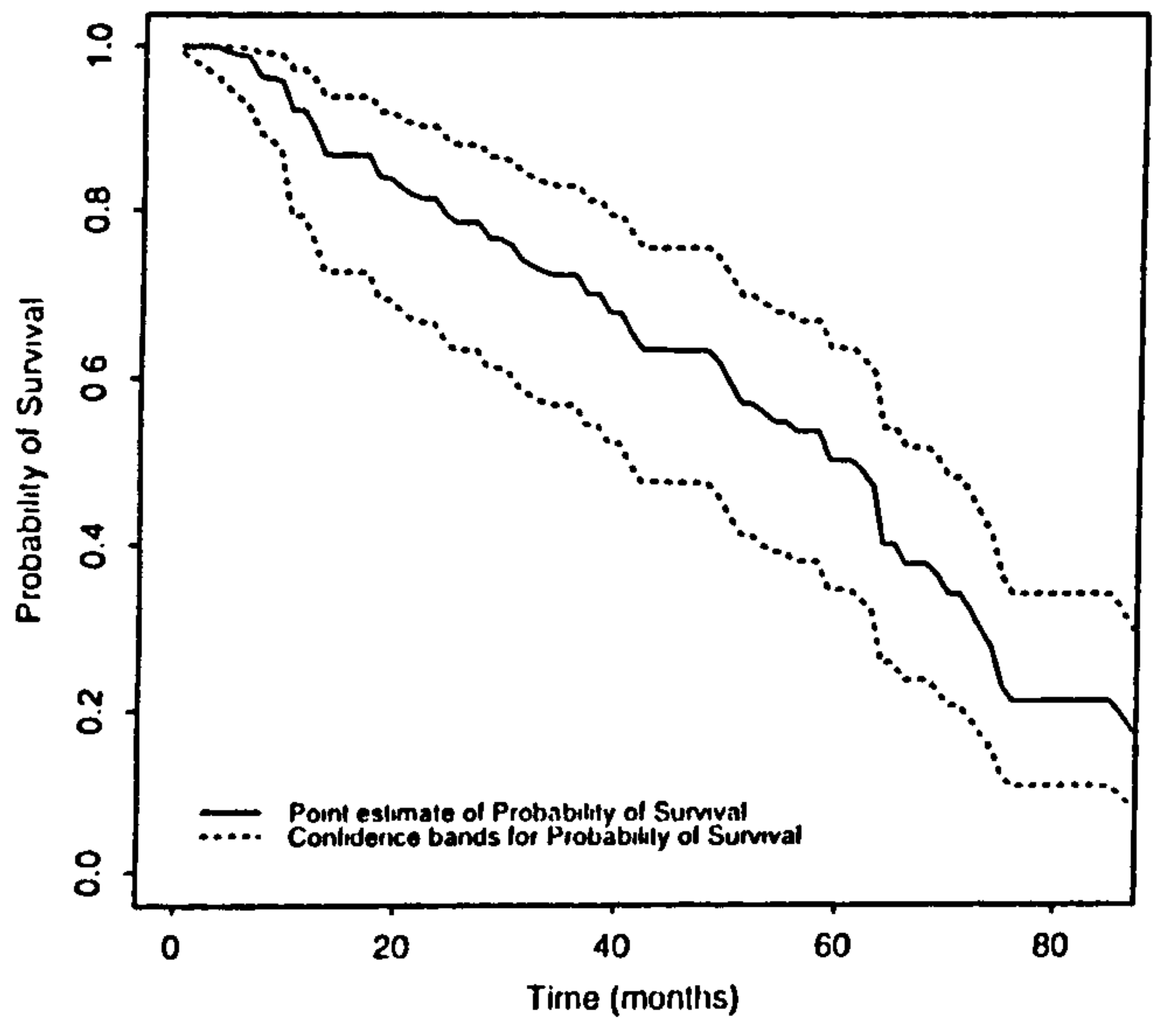


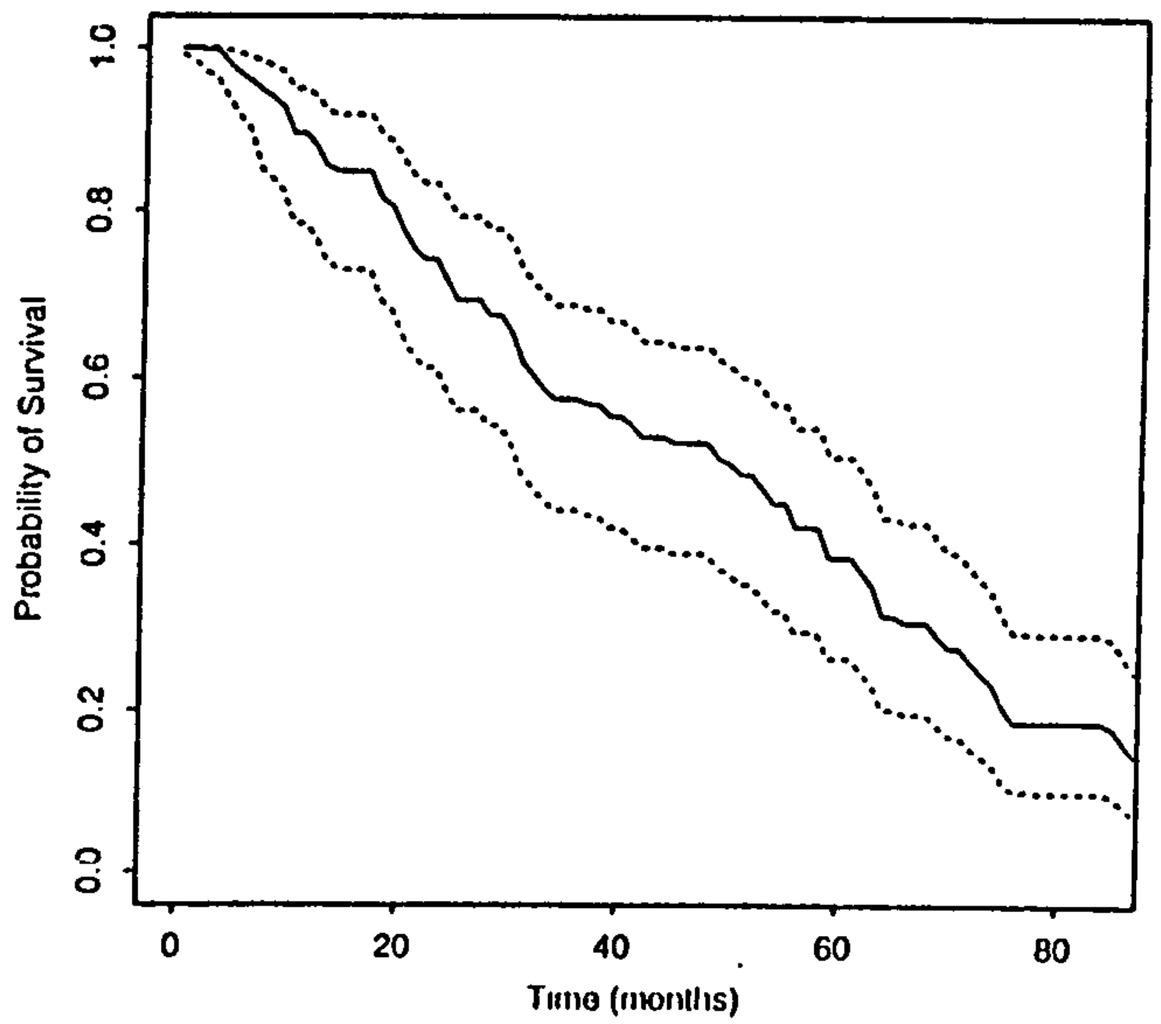
Figure 4.5.8

Females with Ulcerated Lesions on an axial site

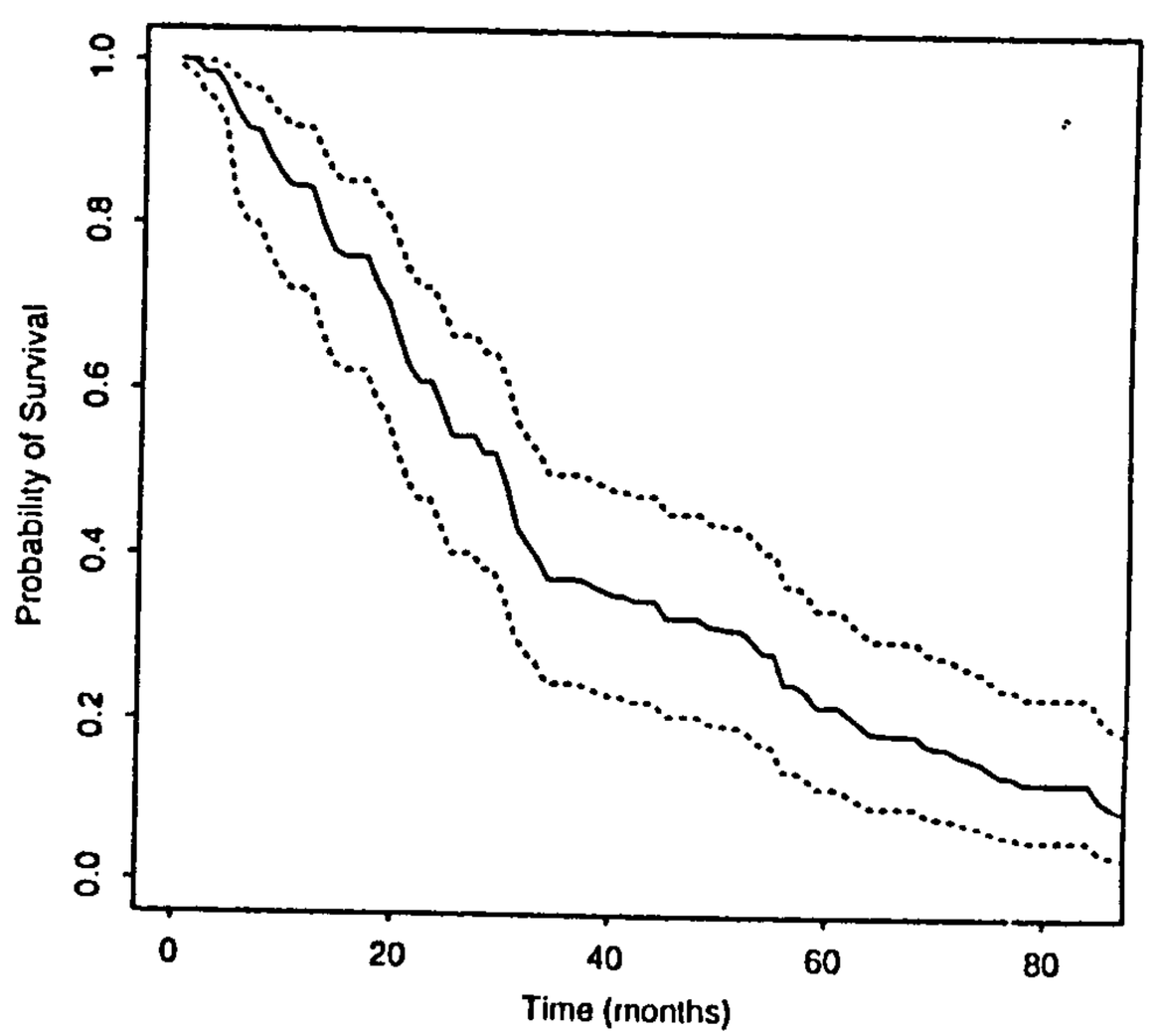
Tumour thickness = 1 mm



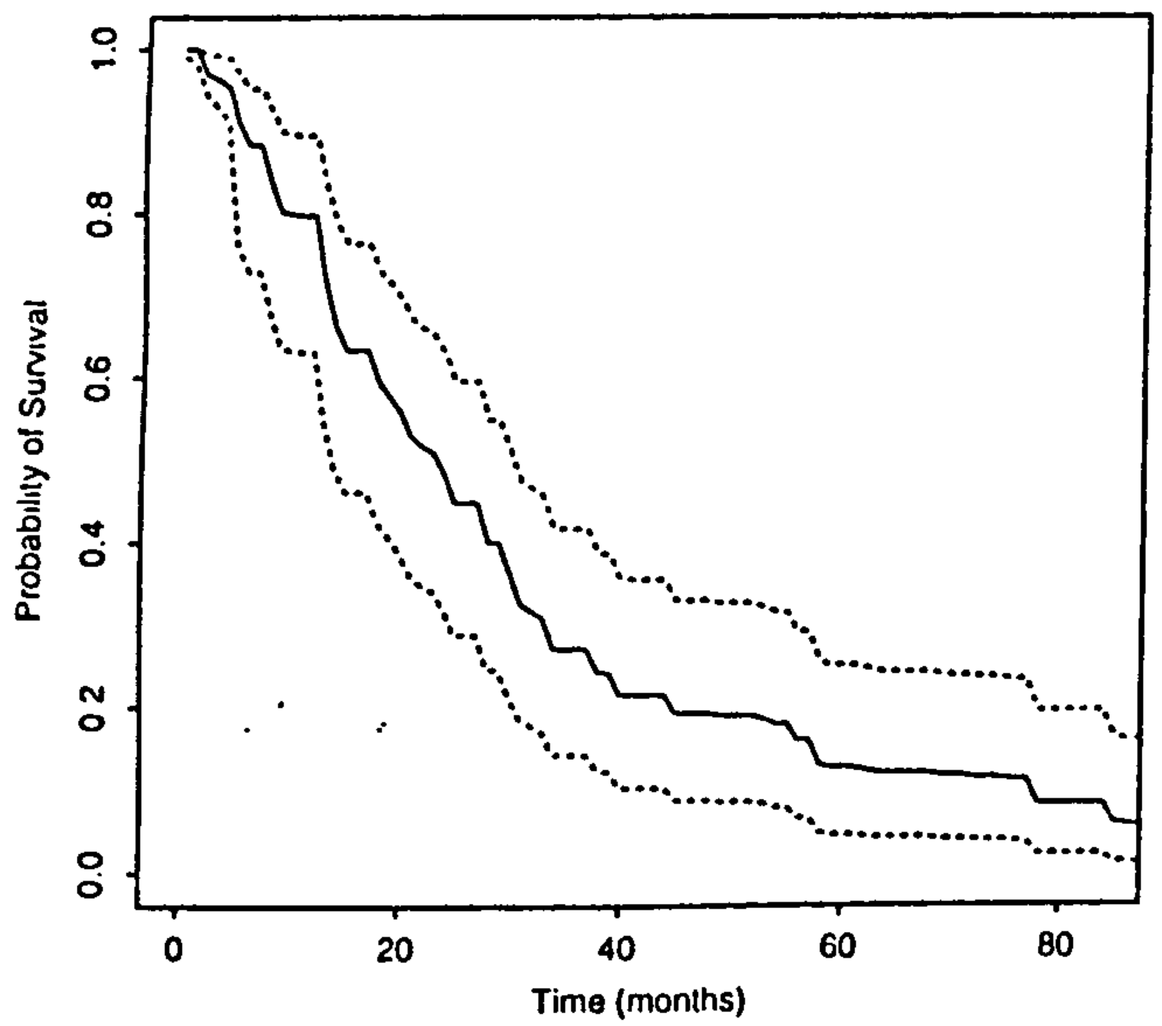
Tumour thickness = 3 mm



Tumour thickness = 5 mm



Tumour thickness = 7 mm



Tumour thickness = 9 mm

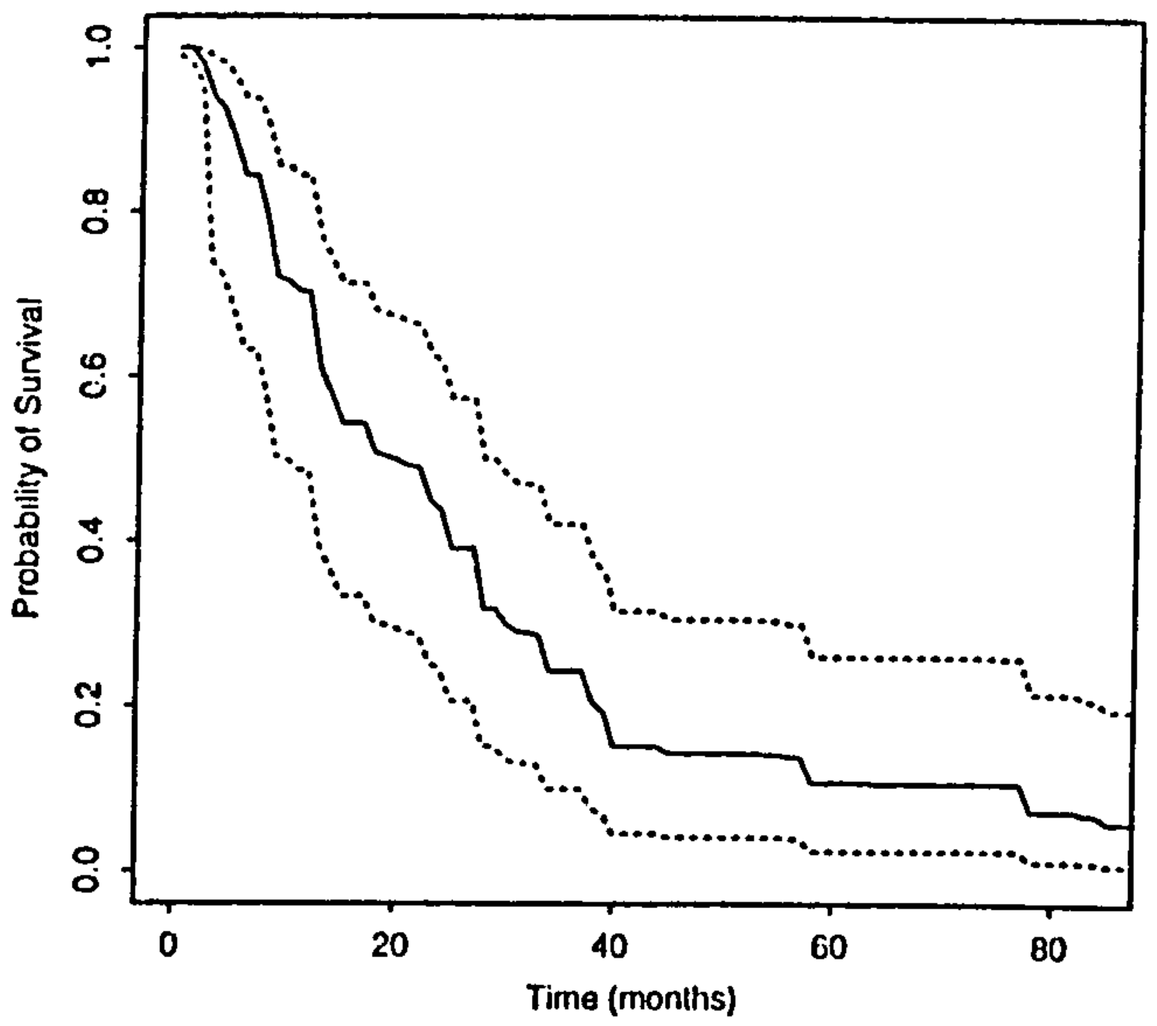


Figure 4.5.9

off far more rapidly for patients with larger tumour thicknesses. However there is an inherent problem with this method which can clearly be seen by comparing *each* frame of Figure 4.5.7 with either Figure 4.3.3 or indeed Figures 4.5.1 and 4.5.4. This method of estimating survival produces biased estimates of survival regardless of the tumour thickness or the smoothing parameter. The method produces ***underestimates*** of the probability of survival. Comparison of frame 9 of Figure 4.5.7 with the proportional hazards estimates shown in Figure 4.3.3 show that on average the estimates of survival produced by this method are approximately 20% lower than the proportional hazards estimates and can even be as much as 40% lower. In the ***absence of a covariate*** Watt et al (1996) compared the estimates of survival produced by the method of Kaplan and Meier to a simple estimator which ignored the presence of censored observations. Their findings suggested that, by ignoring the censored observations, the simple estimator will underestimate survival compared to the Kaplan Meier with the degree of underestimation increasing both through time and as the proportion of censoring increases. Similarly, here, at any ***specific point in time***, the logistic regression based approach to survival analysis also ***ignores*** any censored observations, again producing biased underestimates of survival. By definition, the proportion of censored observations increases through time and hence there will be a corresponding increase through time in the degree of underestimation present with this method.

Section 4.6: Simulation Study

Three possible non-parametric approaches for producing estimates of survival in the presence of a single, continuous covariate were introduced in Section 4.4 and illustrated in Section 4.5. To consider and compare the results of these three approaches it is advisable to carry out a simulation study across a variety of contexts and models likely to be similar to those met in practice. Here, survival data will be simulated from known situations and the estimates of survival obtained using the different approaches will be compared with the underlying, known survival. Clearly there are a multitude of possible survival scenarios which could be simulated and ideally all these scenarios should be given the relevant deliberation. However the practicalities involved in carrying out lengthy simulations make it inevitable that only a small subset of such can be considered. In this section three appropriate scenarios will be concentrated on; firstly the situation where the potential covariate has no effect on survival, secondly where the proportional hazards model is a suitable model to explain the effect of the covariate and thirdly the situation where a single categorisation point is present in the covariate. The first two scenarios are relatively self explanatory but the third perhaps requires some explanation. The third scenario corresponds to there being two specific, different hazard rates present. These two hazard rates will lead to 2 separate survival curves, one survival curve for covariate values less than the categorisation point and a different survival curve for covariate values greater than the categorisation point.

Various measures exist to quantify how reliable an estimator is at reproducing an underlying, known situation. In the context of a survival problem where interest lies in producing estimates of survival across both time and a covariate, each method of estimation will produce an estimate of the true surface. In this situation there are three obvious questions which may be asked about any particular method of producing an estimate of the true surface; firstly there is the question of the precision of the estimated surface when compared to the true surface, secondly the issue of whether the estimated surface exhibits any inherent bias and thirdly what levels of coverage are attained by the method of estimation (i.e. how often do the confidence bands capture the true surface).

It is difficult to describe a complete surface so here summary measures will be used to invetigate precision, bias and coverage. One such summary measure to quantify the precision of the estimated surface compared to the true surface would be to consider the difference in total squared area beneath the two surfaces. In this section the ***average of this difference in total squared area*** across all simulations will be used as an objective measure of ***precision***.

Gasser and Muller (1979) and Hardle (1990) showed that many of the standard smooth non-parametric estimators exhibit inherent bias. The aim in this section is to discover which, if any, of these non-parametric approaches to survival analysis here show bias. Again summary measures are required and hence, in order to examine aspects of ***bias*** the ***difference in total area averaged*** across all

simulations will be used as a measure of how much bias is present with each of the approaches. The closer this value is to zero, the less bias that is present.

Finally, for each method of estimation, rather than considering the ***overall coverage*** function across time and the covariate, the summary measure of ***coverage*** used here will consider ***specified points*** across both time and the covariate. In the simulations which follow confidence intervals for the true survival will be calculated at three time points; the lower quartile, the median and the upper quartile of the ***observed times*** (i.e. includes both failure and censoring times), and at two covariate values; the lower quartile and the upper quartile. This allows the coverage to be evaluated at six time/covariate combinations. The confidence intervals will be constructed based on a ***nominal coverage of 95%***. One point to notice is that the ***specific values*** for the aforementioned lower quartile, median and upper quartile ***time*** values will change as the proportion of censoring changes.

In the simulations which follow survival data have been simulated with three levels of censoring; approximately 15%, 30% and 45%. A range of sample sizes have been considered as follows; 25, 50, 75, 100, 250 and 500 observations. The results presented are based on carrying out 500 simulations of each sample size with each proportion of censoring.

It is of interest to compare the three methods of estimation both ***within*** and ***across*** scenarios in terms of precision, bias and coverage. To allow direct comparison across scenarios follow-up times have been generated in ***each scenario*** to produce, on average, with 30% of censoring, a lower quartile follow up time of 2 years, a median of 5 years and an upper quartile of 10 years. These values will hopefully correspond to follow-up times which are similar to those met in practice.

Section 4.6.1: Scenario 1: Simulated data with no covariate effect

The simplest model of interest is where the survival time is unaffected by a measured covariate. One way to simulate data of this form is to generate covariate values from a distribution which is independent of both the survival and censoring times. Survival times are simulated from an Ex(θ) distribution and censoring times from an Ex(ϕ) distribution where ϕ can be varied to alter the proportion of censoring. The actual observed time is taken to be the minimum of the survival and censored times. Generating the survival times from an exponential distribution implies that the hazard rate will be uniform. The covariate values are simulated independently from a simple Un(0,1) distribution. This will produce data from a model where the covariate has no effect on survival prospects. Figure 4.6.1 displays a three dimensional perspective plot of the true underlying surface for this scenario. Table 4.6.1 details the parameter values used in the simulations to produce the required proportion of censoring. Table 4.6.2 provides a summary of the corresponding

No Covariate Effect - Three Dimensional Plot of True Surface

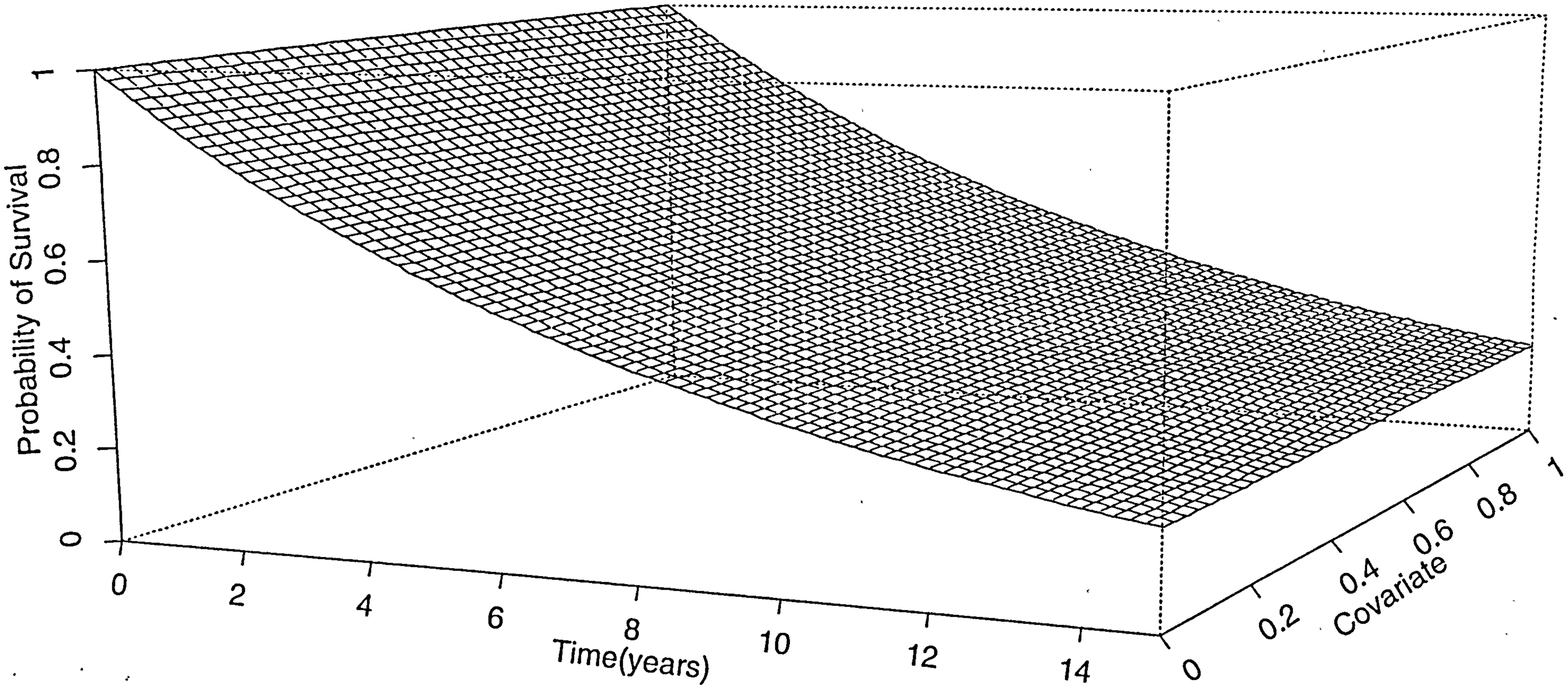


Figure 4.6.1

observed follow-up times. In this scenario the coverage function will be displayed at the time values specified in table 4.6.2.

Survival times: θ = 0.0909
Censoring times: ϕ = 0.0185, 0.0435, 0.0820
corresponding to 15%, 30%, 45% censoring

Table 4.6.1

Censoring Proportion		Observed follow-up times	
	Lower quartile	Median	Upper quartile
15%	2.5 years	5.9 years	11.9 years
30%	2 years	5 years	10 years
45%	1.6 years	3.8 years	7.6 years

Table 4.6.2

For each non-parametric approach the resultant estimates of survival will depend on a suitable choice of smoothing parameter(s). Under scenario 1, Figure

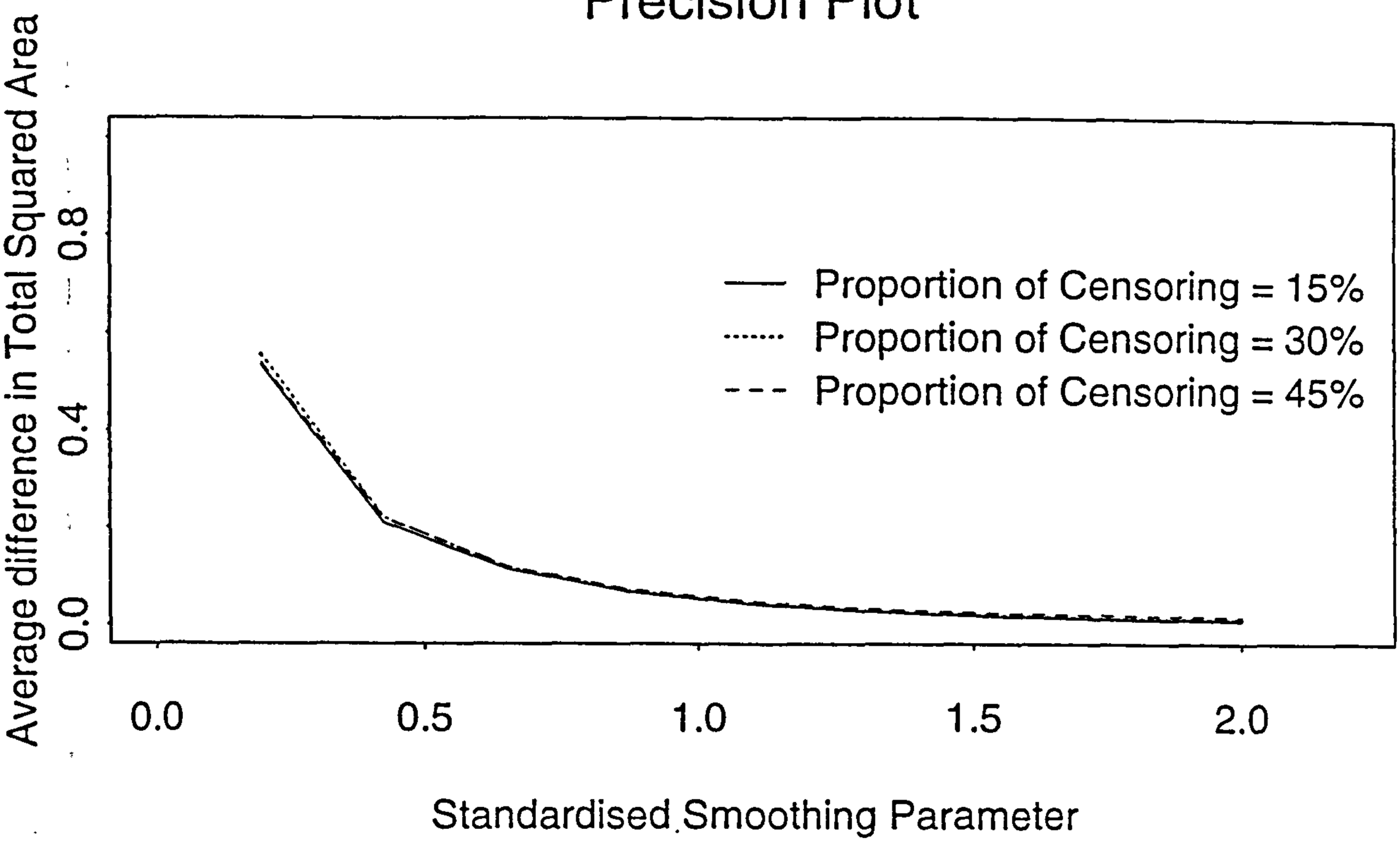
4.6.2 shows the patterns of precision and bias across different choices of smoothing parameter for the Kaplan Meier based approach based on 500 simulations of 50 observations. Similar patterns were obtained across all sample sizes for each of the three different non-parametric methods of estimation. Therefore, in the rest of this section, rather than considering a range of smoothing parameters, the results will be based on an "optimal" choice of smoothing parameter. The smoothing parameter is optimal in the sense that in a particular simulation the "average difference in total squared area" is minimised with this value of the smoothing parameter.

Figure 4.6.3 allows a comparison to be made across the three approaches in terms of the degree of precision produced (Note that the scale in frame 1 of Figure 4.6.3 is massively different from the scale in frames 2 and 3). These results clearly suggest that, regardless of sample size and proportion of censoring, the Kaplan Meier based approach produces the lowest values for the "average difference in total squared area". This would suggest that the Kaplan Meier based approach is the **most precise** of the three methods of estimation.

With the Kaplan Meier based approach (frame 1 of Figure 4.6.3) there is evidence that both the sample size and the proportion of censoring have an effect on precision for sample sizes of ***less than 100*** with greater precision being achieved with smaller proportions of censoring. Once a sample size of 100 is reached there is little effect of either the sample size or the proportion of censoring on precision.

Kaplan Meier Based Approach Sample Size = 50

Precision Plot



Bias Plot

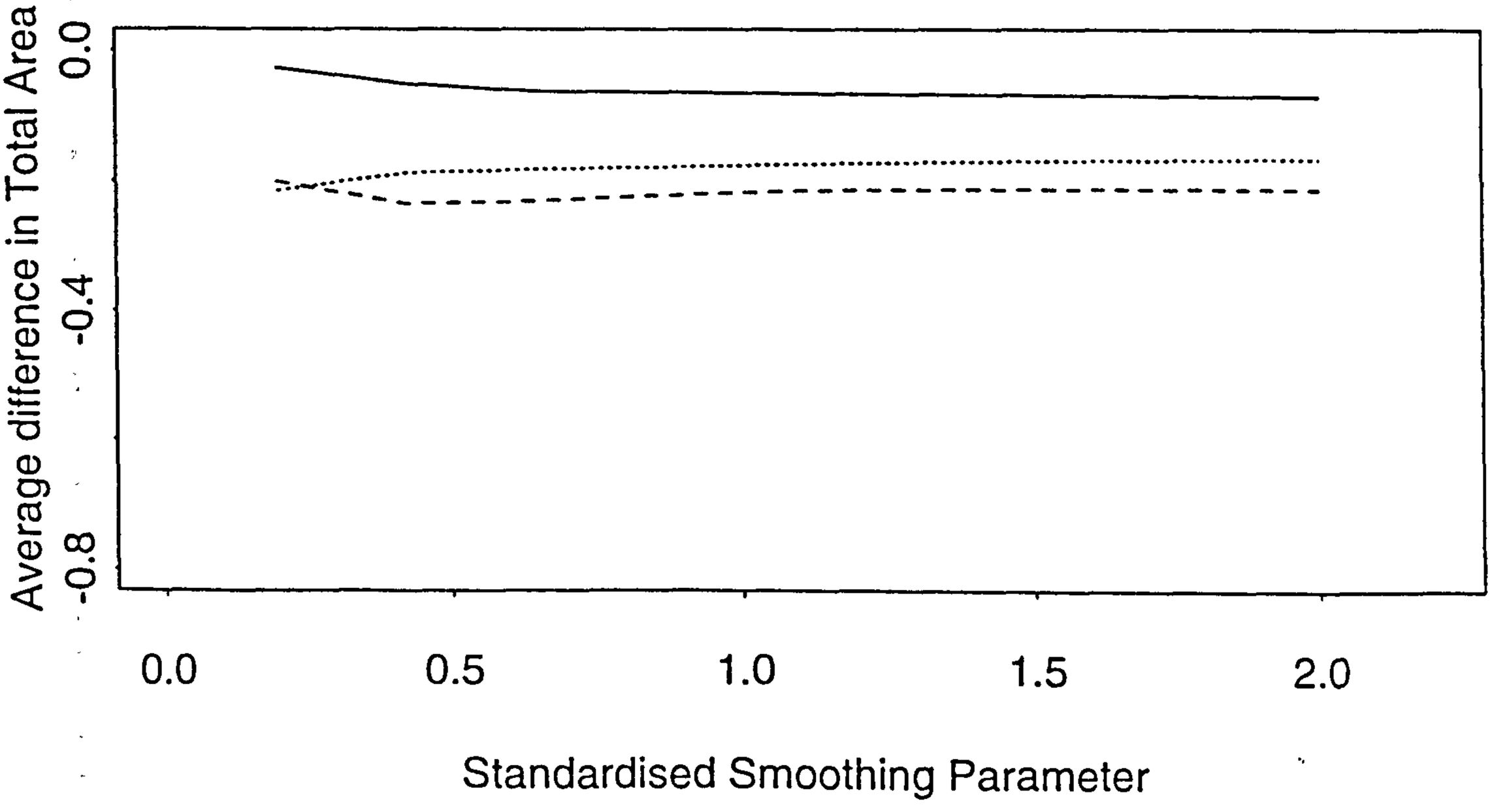
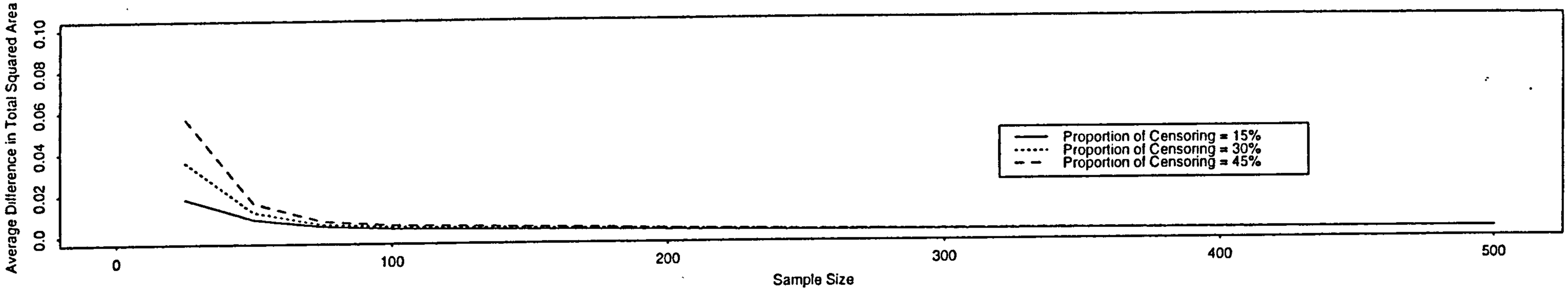
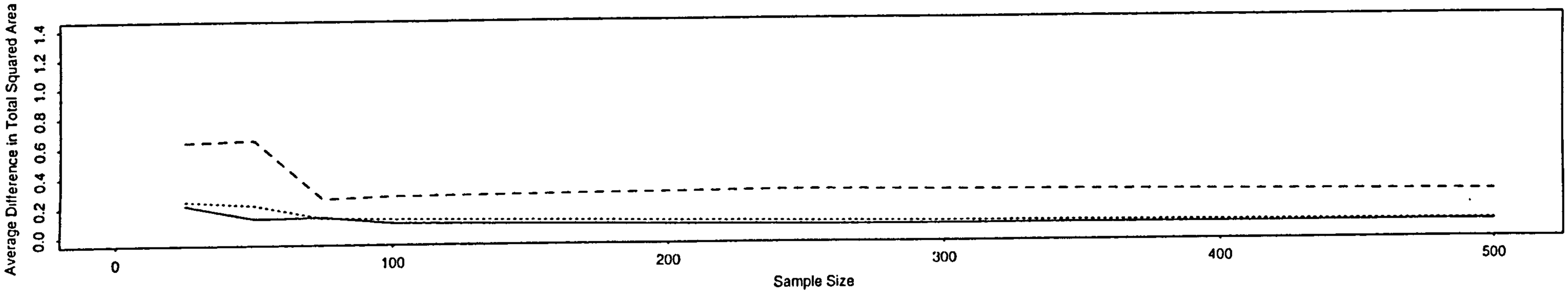


Figure 4.6.2


Kaplan Meier based approach
Average Difference in Total Squared Area
0.0 0.02 0.04 0.06 0.08 0.10
0
100
200
300
400
500
Sample Size
Proportion of Censoring = 15%
Proportion of Censoring = 30%
Proportion of Censoring = 45%


Hazard based approach
Average Difference in Total Squared Area
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4
0
100
200
300
400
500
Sample Size

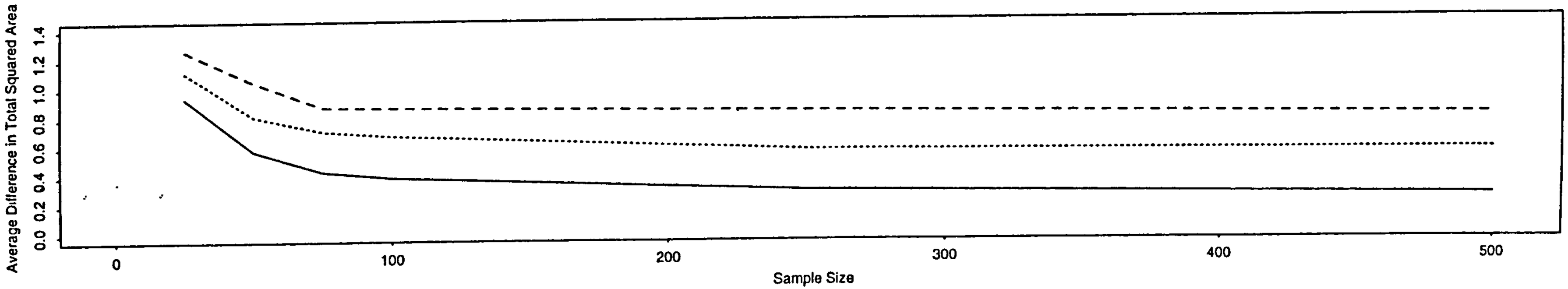
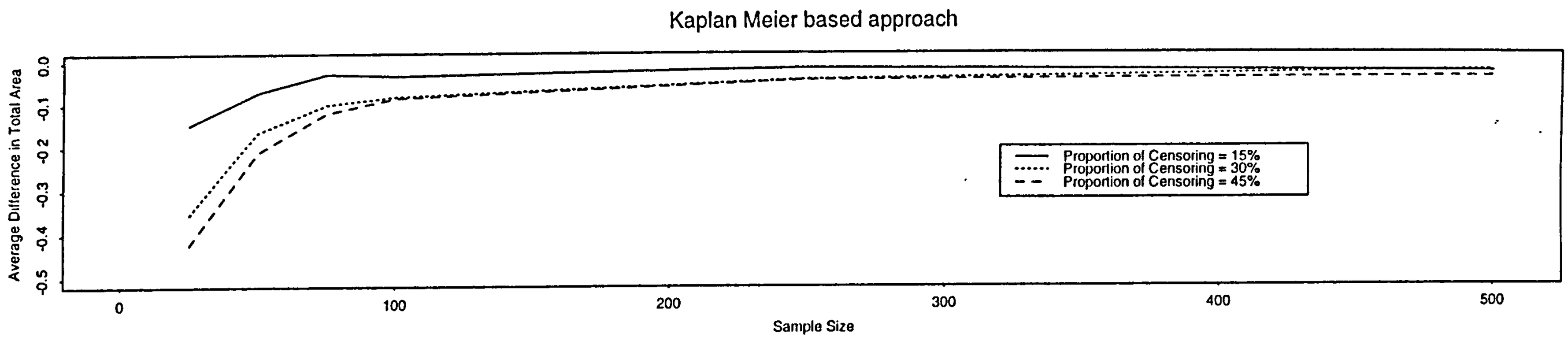

Logistic based approach
Average Difference in Total Squared Area
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4
0
100
200
300
400
500
Sample Size

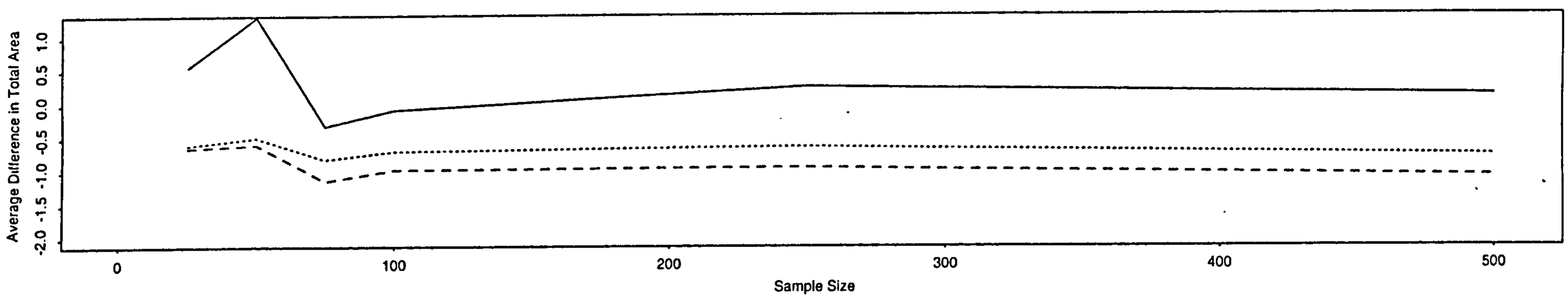
Figure 4.6.3

The hazard based approach (frame 2 of Figure 4.6.3) produces estimates which are less precise than those produced by the Kaplan Meier approach but clearly more precise than those obtained by the logistic based approach. An increase in the sample size initially leads to a small increase in the precision of the estimates. However, once a sample of about 75 observations is reached any increase in the sample size appears to have little effect on the precision of the estimates. Regardless of sample size, there is a clear decline in precision when a 45% proportion of censoring is present but very little difference between 15% and 30% censoring.

The estimates produced by the logistic based approach exhibit levels of precision which are clearly poorer than those of the other two methods. The results displayed in frame 3 of Figure 4.6.3 also show a slightly different pattern than those obtained with both the Kaplan Meier and hazard based approaches. A similar pattern is observed with regard to sample size where an initial increase in precision is obtained as the sample size increases before this improvement levels off. However the proportion of censoring appears to have a far more marked effect on the levels of precision. An increase in censoring here leads to a ***far more marked decrease*** in precision than was observed with the other two methods of estimation.

In terms of bias, Figure 4.6.4 illustrates quite clearly that the Kaplan Meier based approach shows the least bias (Note that the scales in each of the three frames in Figure 4.6.4 are different). The hazard based approach produces estimates which exhibit greater levels of bias than the Kaplan Meier based approach but much less


Kaplan Meier based approach
Average Difference in Total Area
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
0
100
200
300
400
500
Sample Size
Proportion of Censoring = 15%
Proportion of Censoring = 30%
Proportion of Censoring = 45%


Hazard based approach
Average Difference in Total Area
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
0
100
200
300
400
500
Sample Size

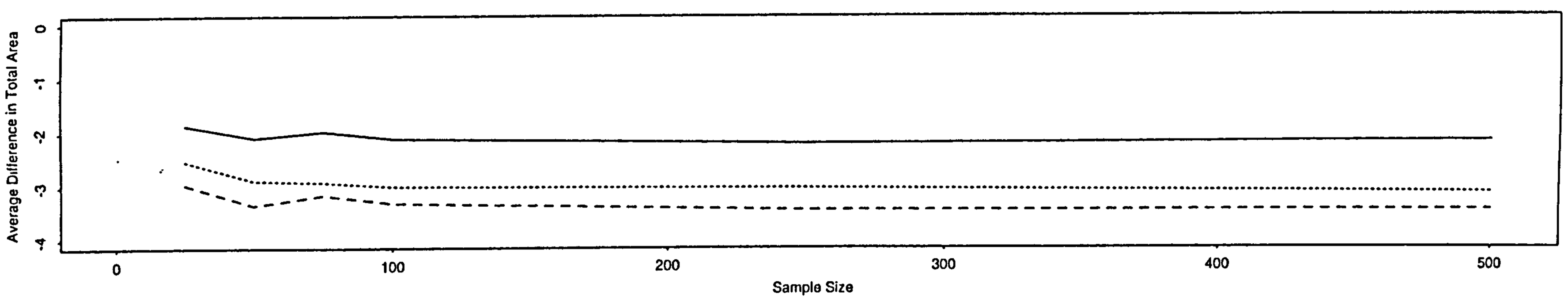

Logistic based approach
Average Difference in Total Area
0
-1
-2
-3
-4
0
100
200
300
400
500
Sample Size

Figure 4.6.4

than the logistic based approach. The logistic based approach clearly produces estimates which exhibit very high levels of bias, and this method clearly underestimates the true survival.

The Kaplan Meier based approach (frame 1 of Figure 4.6.4) actually exhibits very little bias irrespective of sample size or proportion of censoring. In general the method produces slight underestimates of the true survival, an underestimation which increases, slightly, as the proportion of censoring increases. This is particularly true for smaller sample sizes, but, once larger sample sizes are used (> 250 observations) the degree of bias is almost negligible, regardless of the proportion of censoring.

The hazard based approach (frame 2 of Figure 4.6.4) shows more bias than the Kaplan Meier approach, a bias which increases as the proportion of censoring increases. Regardless of sample size, the method appears to produce underestimates of the true surface with 30% and 45% censoring. However, with 15% censoring, the method produces overestimates for smaller sample sizes but exhibits very little bias, if any, for larger sample sizes. The levels of bias displayed here, although not as good as with the Kaplan Meier based approach, still do not appear particularly excessive.

The logistic based approach (frame 3 of Figure 4.6.4) produces levels of bias which would appear to be unacceptably high, particularly for higher proportions of censoring. Regardless of sample size this method produces estimates which underestimate the true survival by a reasonably large margin. The degree of bias is influenced heavily by the proportion of censoring with an increase in the amount of censoring leading to a corresponding increase in bias.

In terms of coverage, Figures 4.6.5 to 4.6.7 display the results for the three methods of estimation respectively. In each figure there are six frames representing the six combinations of the time and covariate values discussed in Section 4.6.1. The covariate lower quartile value equals 0.25 with the covariate upper quartile value being 0.75. These figures clearly demonstrate that the coverage achieved by the Kaplan Meier based approach is superior to the coverage with either of the other two methods. In turn, the hazard based approach displays levels of coverage which are superior to those achieved with the logistic based approach. The Kaplan Meier based approach is actually the only method which appears to achieve the nominal levels of 95% coverage.

With the Kaplan Meier based approach (Figure 4.6.5) the coverage increases through time but appears very similar at each of the two covariate values. This is perhaps as expected since the width of the intervals will, in general, increase through time leading to higher levels of coverage being achieved. Also, in this scenario, the covariate has no effect on the pattern of survival and, therefore, the levels of

Kaplan Meier Based Approach

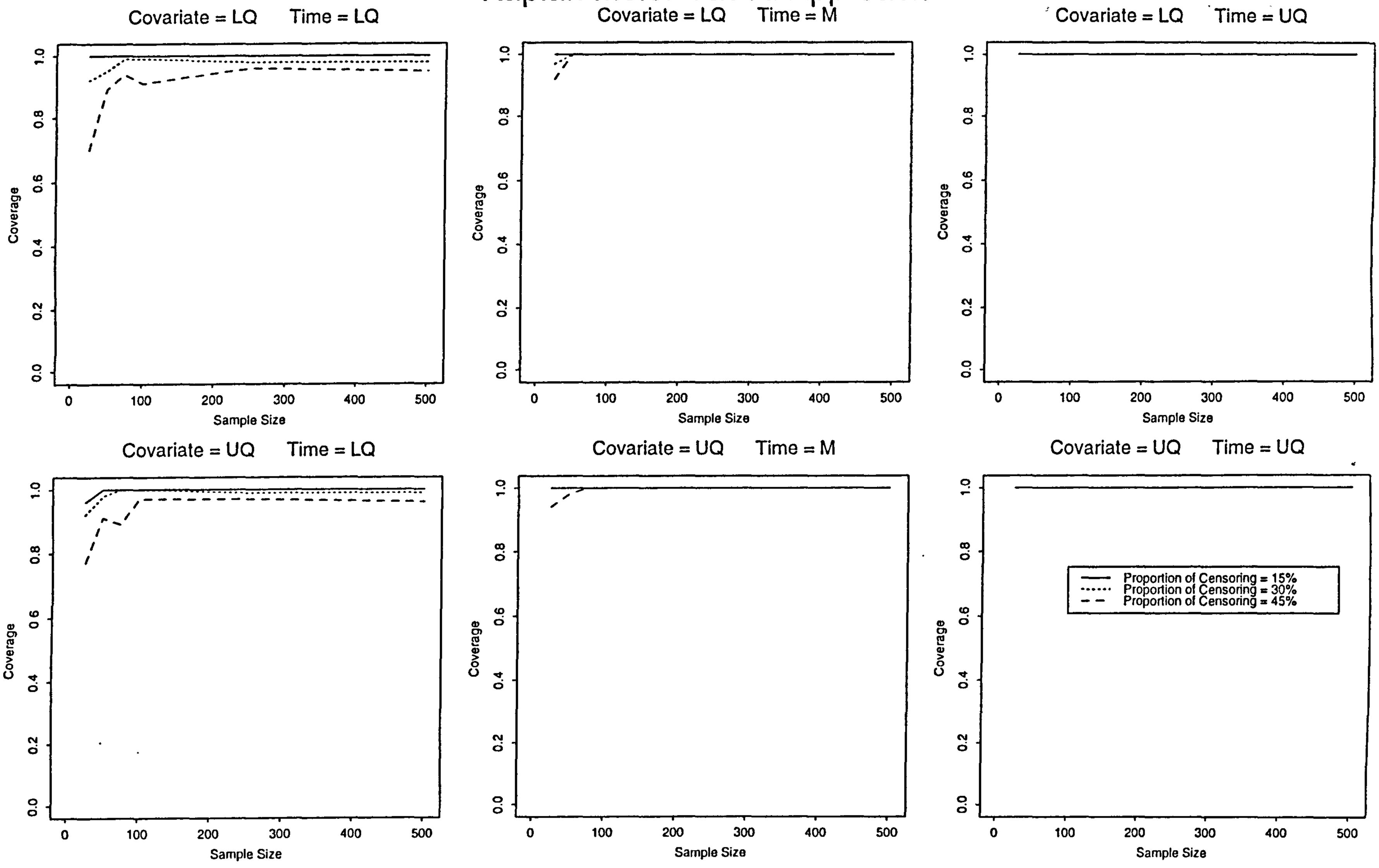


Figure 4.6.5

Hazard Based Approach

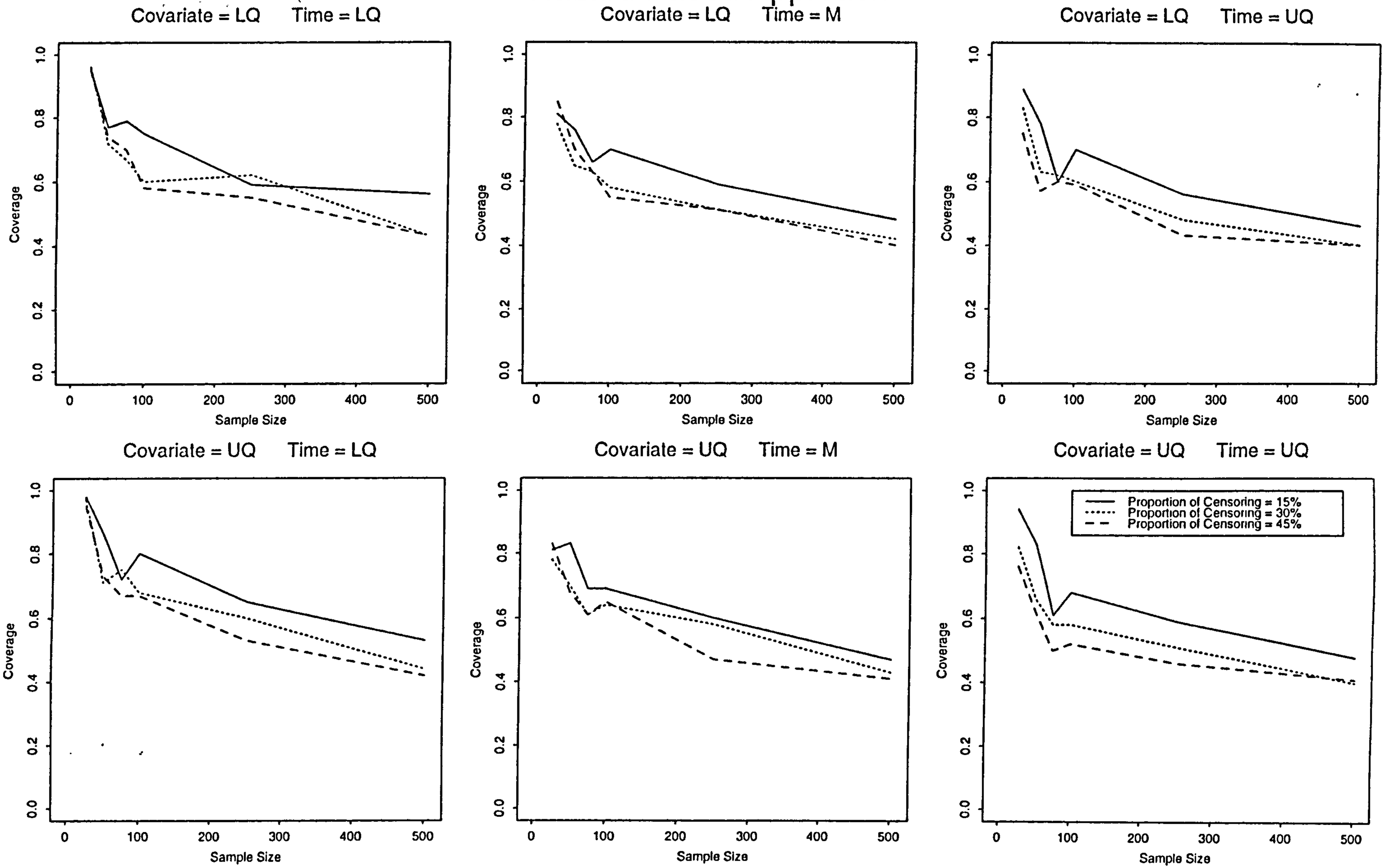


Figure 4.6.6

Logistic Based Approach

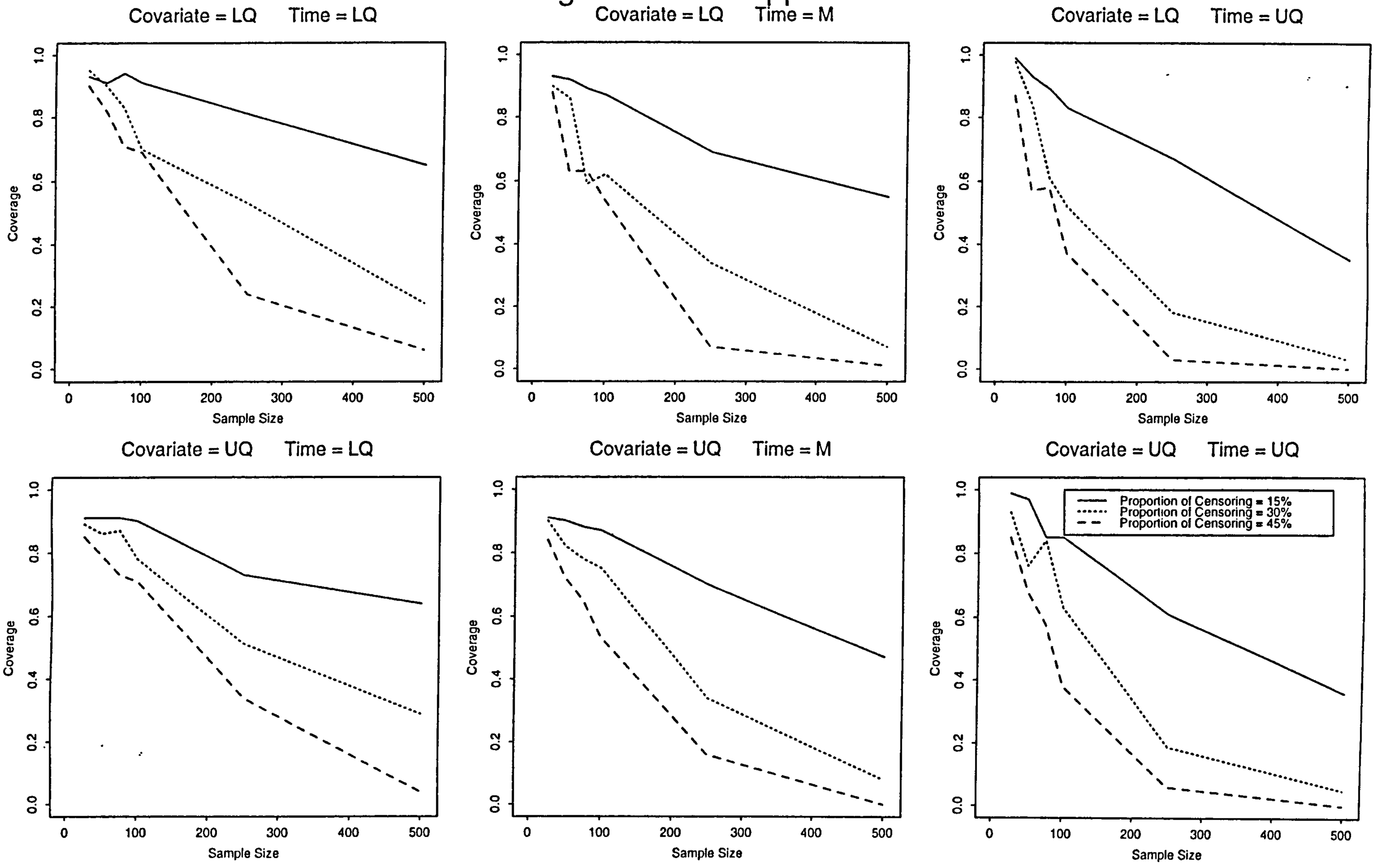


Figure 4.6.7

coverage should not depend on the value of the covariate. For the lower quartile of the observed times (i.e. frames 1 and 4) there is evidence that the coverage increases with increasing sample size and decreases as the proportion of censoring increases. However, for the median and lower quartile of the observed times, neither the sample size or proportion of censoring appear to have any effect on the levels of coverage. The main reason for this is that, regardless of sample size and proportion of censoring, the coverage is invariably very good at these observed times.

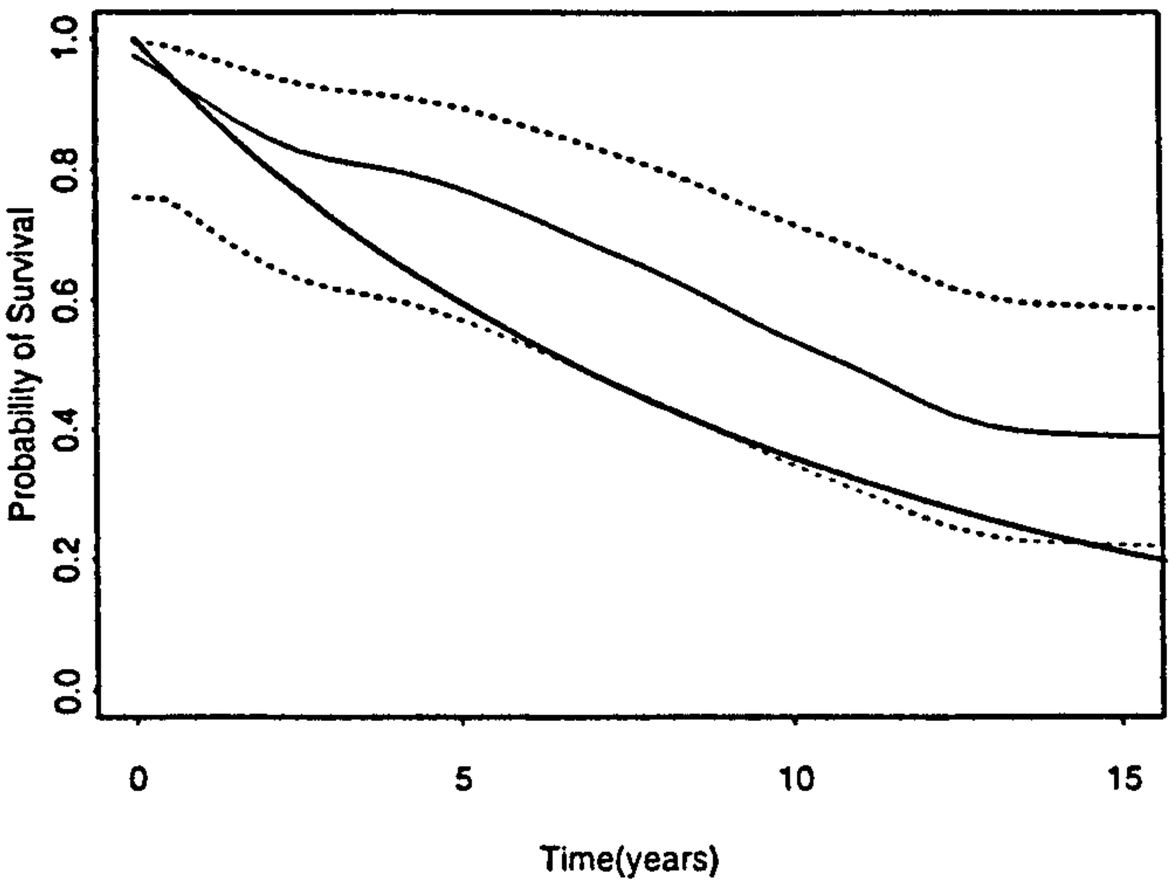
With the hazard based approach (Figure 4.6.6), far lower levels of coverage are, in general, achieved than with the Kaplan Meier based approach. Here, neither the value of the observed time or the value of the covariate seem to have any effect on the coverage, as the same picture is essentially observed in each frame of Figure 4.6.6. Regardless of the observed time or the value of the covariate the coverage decreases both with sample size and proportion of censoring. As an increase in censoring indicates the presence of less "complete" information it is perhaps sensible to expect the coverage to decrease as the proportion of censoring increases. However, it is less obvious why the coverage should decrease as the sample size increases. There appear to be two factors which may be contributing to this decrease in coverage with increasing sample size. Firstly, the levels of precision and bias (Figures 4.6.3 and 4.6.4) show little change with sample size, particularly when larger sample sizes are being considered. As the width of the confidence intervals will decrease with increasing sample size, it is clear that if the bias does not show a corresponding decrease, then, for larger sample sizes, less intervals are likely to contain the true value. This will invariably lead to poorer coverage with larger

sample sizes. Secondly, these intervals are approximate intervals based on the log odds (see Section 4.4.2) and are heavily dependent on the sample size (see 4.15). Therefore, these intervals may, in general, be too narrow for larger sample sizes leading to lower levels of coverage being achieved. These conclusions are confirmed, to a certain extent, by Figure 4.6.8 which shows the results for three specific simulations of sample sizes 25, 100 and 250 observations respectively and 15% censoring. The figure displays the estimated survival curve with the solid line (confidence bands as dotted lines) and the true survival curve with the thicker solid line. Figure 4.6.8 clearly shows that sample size has a large effect on the width of the confidence intervals with the intervals becoming increasingly narrow as the sample size increases. However the sample size does not appear to have much, if any, effect on either precision or bias. Although these conclusions are only based on only one simulation of each sample size they do give some back-up to the somewhat unexpected results obtained from Figure 4.6.6. An examination of results from further simulations suggests that, as in Figure 4.6.8 and particularly for larger sample sizes, these approximate intervals, based on the hazard approach, regularly ***just*** fail to capture the true survival curve. In future work it may therefore be necessary to give further consideration to using the ***exact form*** of the variance detailed in Section 4.4.2 as the approximate intervals used here clearly appear to be too narrow for larger sample sizes.

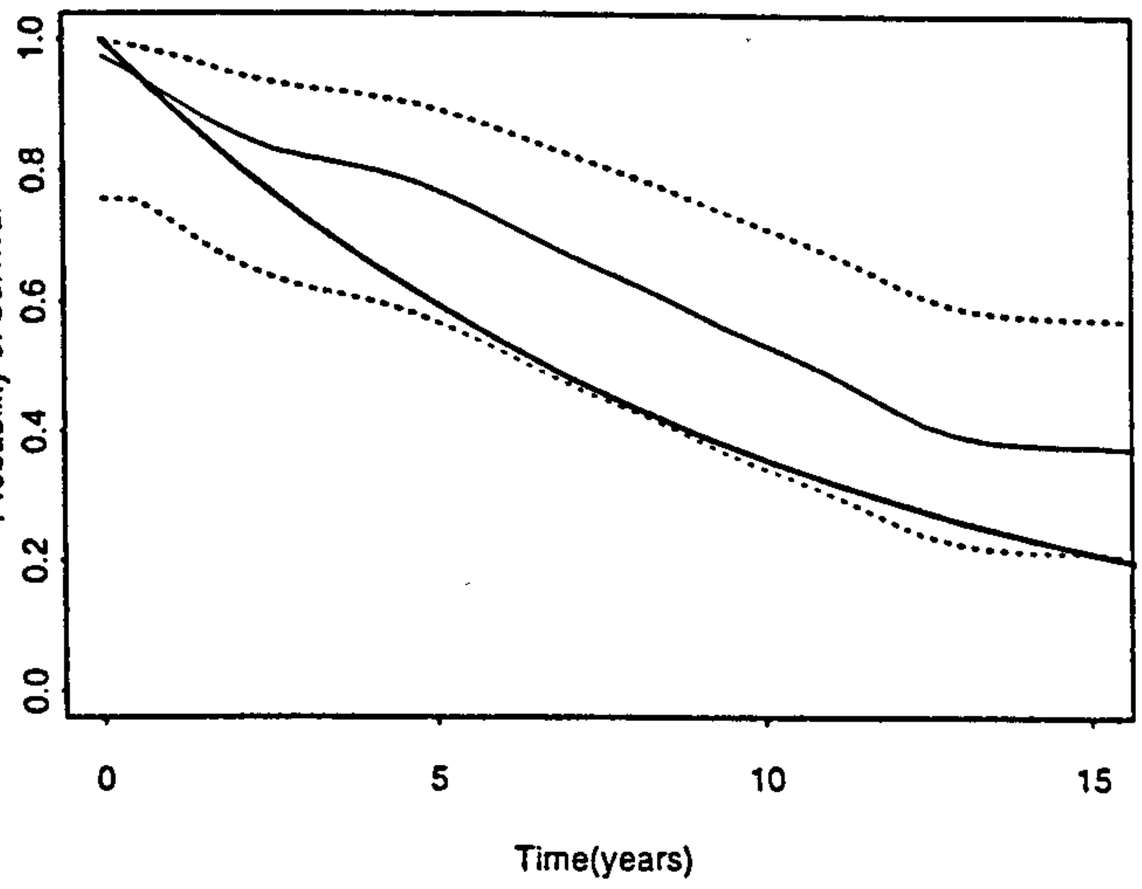
Finally, the logistic based approach shows very poor levels of coverage,

Hazard based approach - Confidence bands for simulated data

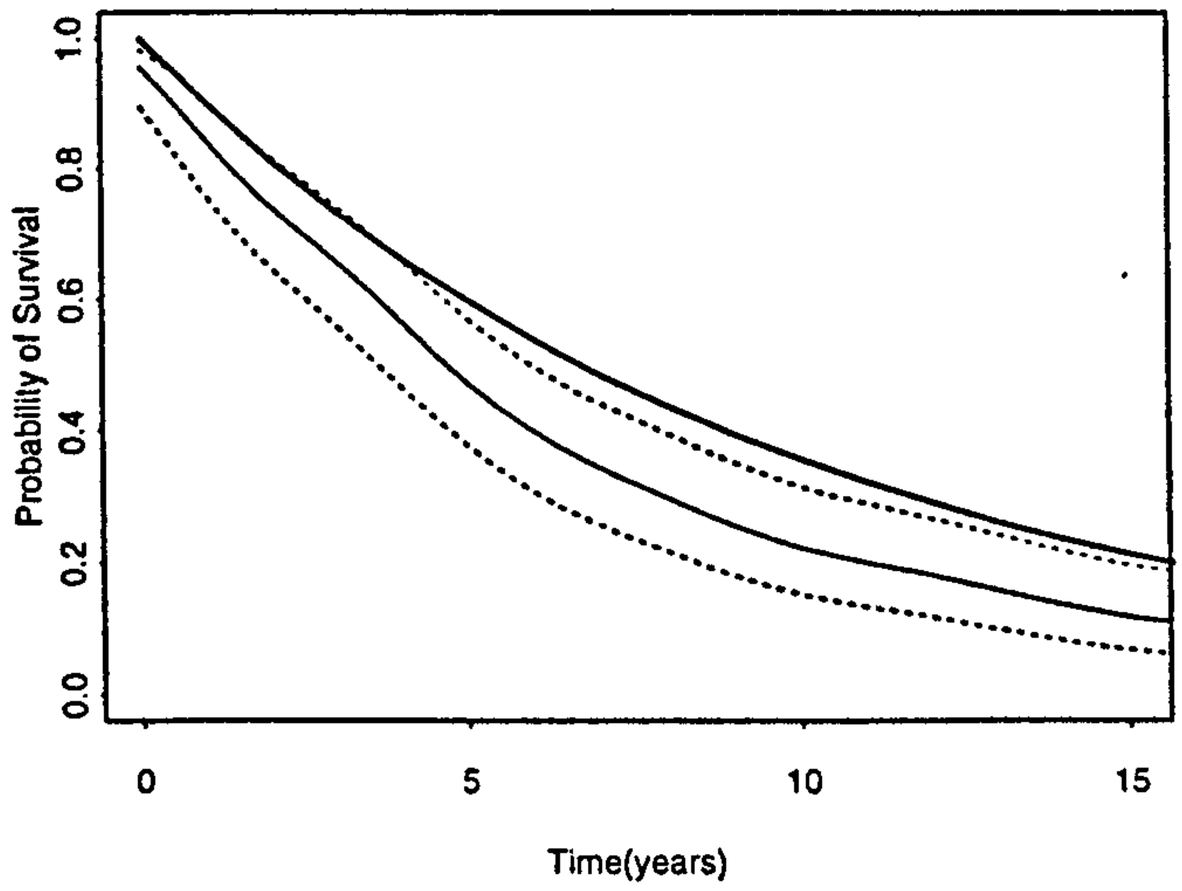
Sample size = 25 , Covariate Lower Quartile



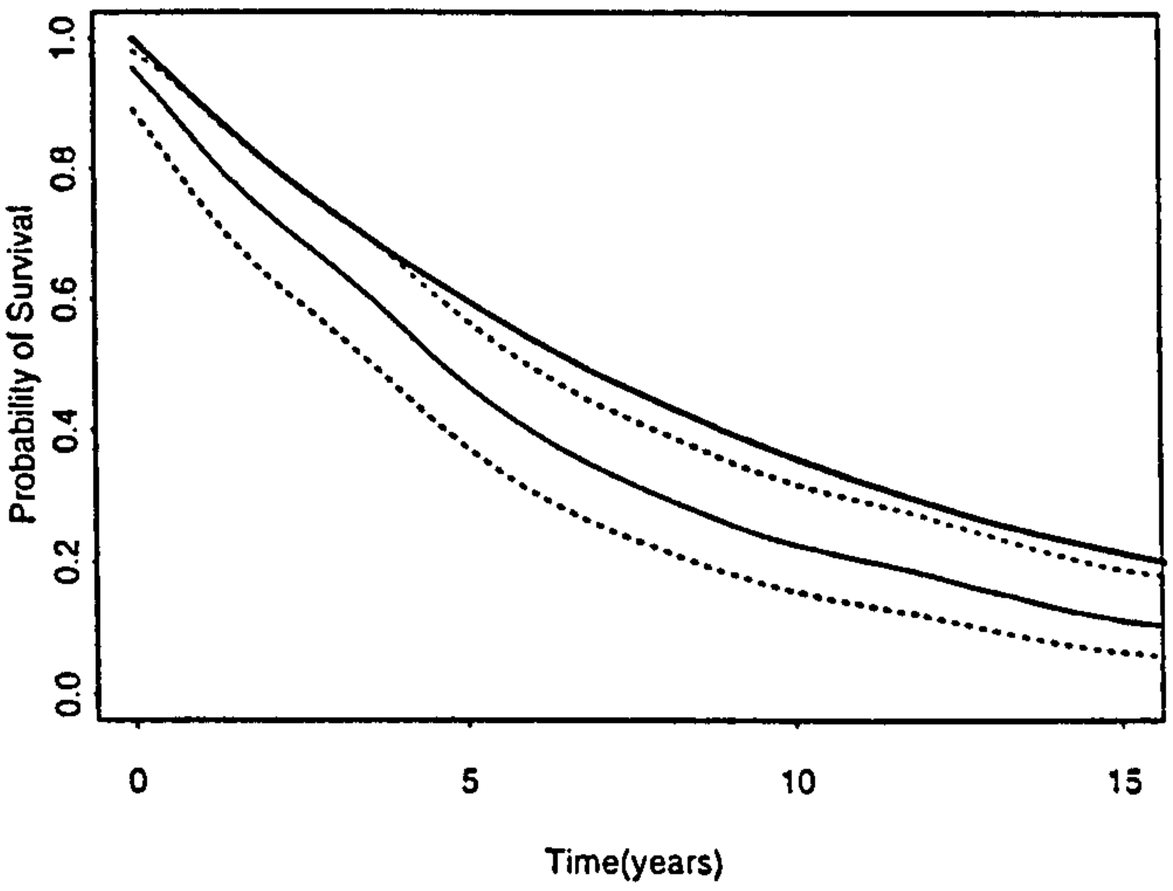
Sample size = 25 , Covariate Upper Quartile



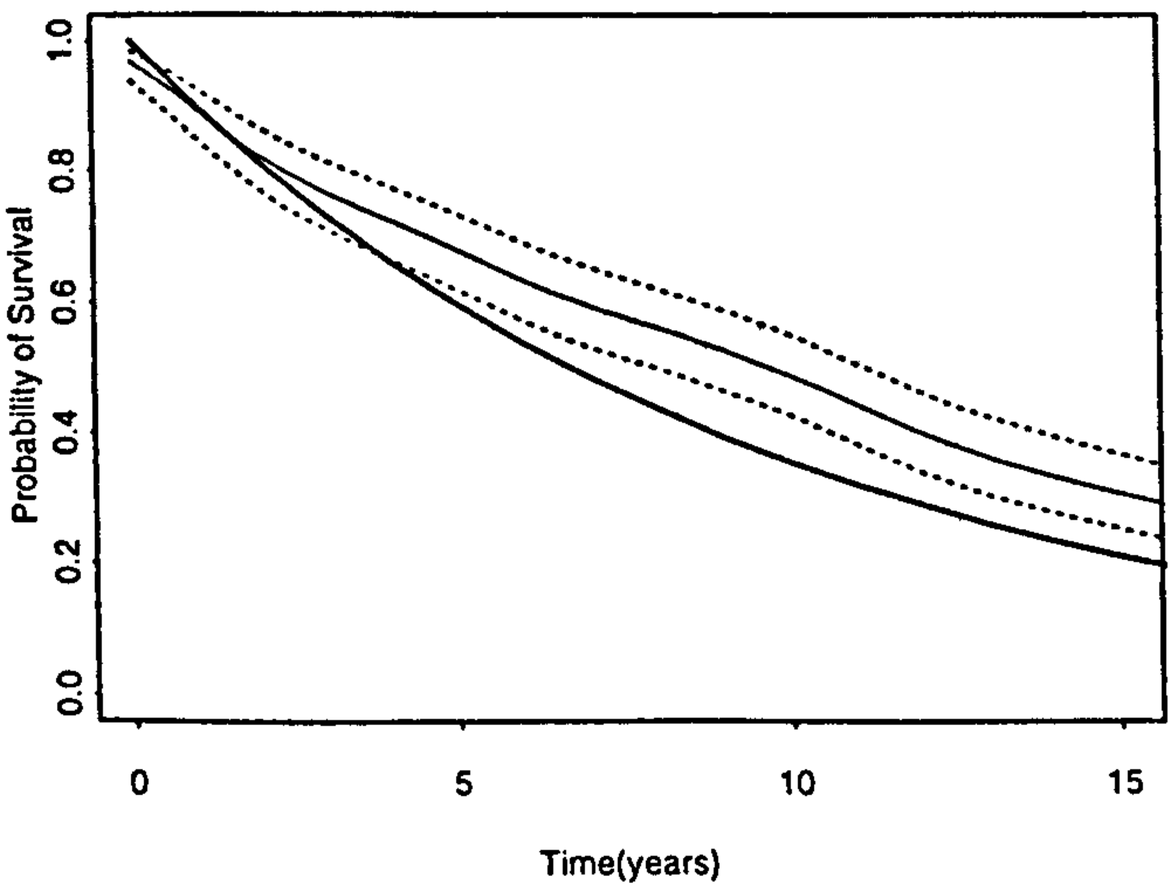
Sample size = 100 , Covariate Lower Quartile



Sample size = 100 , Covariate Upper Quartile



Sample size = 250 , Covariate Lower Quartile



Sample size = 250 , Covariate Upper Quartile

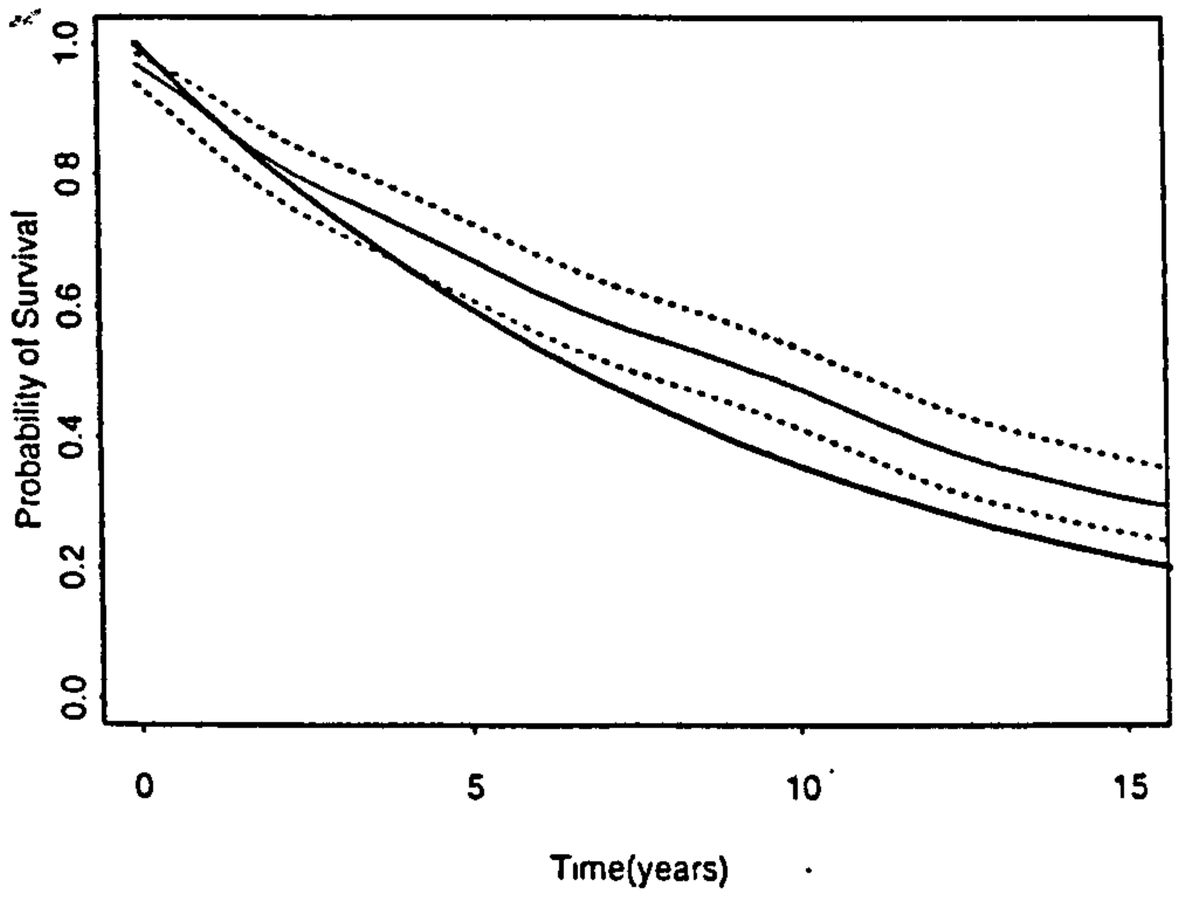


Figure 4.6.8

particularly for 30% and 45% censoring and larger sample sizes. Here the main reason for the poor levels of coverage is that this method produces biased underestimates of the true survival (see frame 3 of Figure 4.6.4) with the level of underestimation increasing as the proportion of censoring increases. Again the decrease in coverage with larger sample sizes is due to a combination of similar levels of bias being present, regardless of sample size, and narrower intervals being produced with larger sample sizes.

In summary, the Kaplan Meier approach can fairly confidently be used to produce estimates of survival in the situation where the covariate of interest has no effect on survival. The Kaplan Meier based approach has been shown to be clearly superior to both the other methods in terms of precision, bias and coverage. The hazard based approach produced reasonable estimates in terms of the levels of precision and bias. However, it is clear that the ***approximate*** confidence intervals based on the hazard approach should be used with caution, particularly if larger sample sizes are being considered. Finally the logistic based approach does not appear satisfactory as it produces far poorer levels of precision in the estimates coupled with unacceptable amounts of bias and low levels of coverage.

Section 4.6.2: Scenario 2: Simulated data from a proportional hazards model.

In this example, the data are generated from a proportional hazards model. This allows comparison of the three proposed non-parametric methods by comparing them to the true, underlying, proportional hazards curve. Here, the data are

generated from a proportional hazards model with regression coefficient, β, equal to -1. For a given covariate z generated under a Un(0,1) distribution, to generate observed follow-up times from the above model, survival times are simulated from an $Ex(\theta e^{-z})$ distribution and censoring times from an $Ex(\phi e^{-z})$ distribution where ϕ can again be varied to alter the proportion of censoring. Figure 4.6.9 displays a three dimensional perspective plot of the true underlying surface. Under this scenario, table 4.6.3 details the parameter values used in the simulations whilst table 4.6.4 summarises the corresponding observed follow-up times.

Survival times:	θ =	0.1602		
Censoring times:	ϕ =	0.0281,	0.0711,	0.1321
corresponding to		15%,	30%,	45% censoring

Table 4.6.3

Censoring Proportion		Observed follow-up times	
	Lower quartile	Median	Upper quartile
15%	2.4 years	6.1 years	12.7 years
30%	2 years	5 years	10 years
45%	1.6 years	3.9 years	8.2 years

Table 4.6.4

Proportional hazards model - Three Dimensional Plot of True Surface

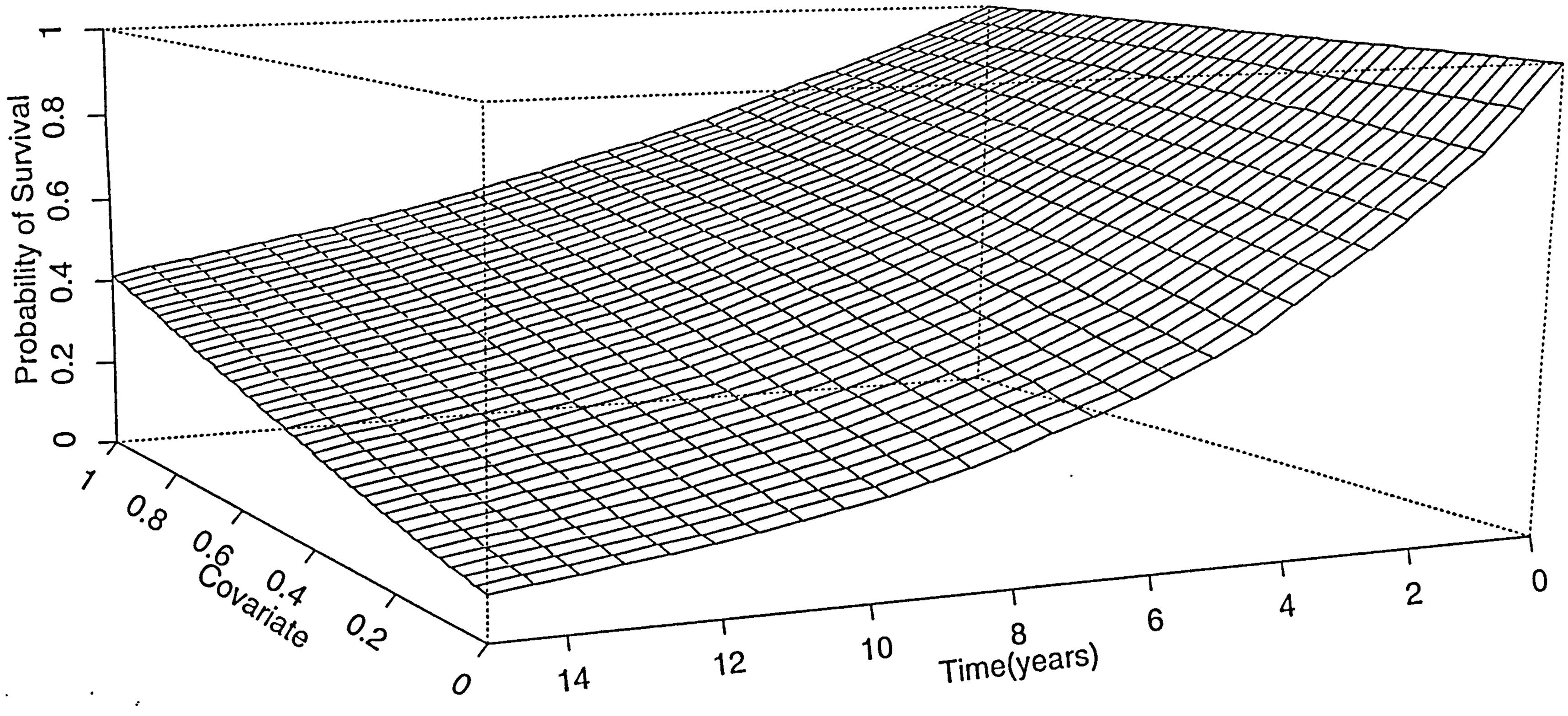


Figure 4.6.9

Figures 4.6.10 to 4.6.14 show the results for this simulation study based on the use of an "optimal" smoothing parameter as outlined in Section 4.6.1. Figure 4.6.10 displays plots of the average difference ***in total squared area*** across all simulations against sample size separately for each proportion of censoring. Again each frame of the figure refers to the simulation results for a different approach. Figure 4.6.11 shows equivalent plots for the average difference in **total area.** Figures 4.6.12 to 4.6.14 display plots of the ***coverage*** against sample size for each of the three non-parametric approaches respectively.

In terms of precision a comparison of the three frames of Figure 4.6.10 shows that the non-parametric Kaplan Meier based approach appears to produce the most precise results regardless of sample size and proportion of censoring (Note that the scale in frame 1 is different to the scales in frames 2 and 3). The levels of precision based on the Kaplan Meier approach appear to improve as the sample size increases and as the proportion of censoring decreases. Any changes which occur in the levels of precision are more noticeable for smaller sample sizes. The hazard based approach appears to produce estimates which are less precise than those produced by the Kaplan Meier based approach, with the levels of precision again decreasing as the proportion of censoring increases. The effect of increasing sample size is less noticeable with the hazard based approach than with the Kaplan Meier based approach. The logistic based approach produces estimates which clearly display the poorest levels of precision. As in scenario 1 there is clear evidence with the logistic approach that the proportion of censoring has the most noticeable effect

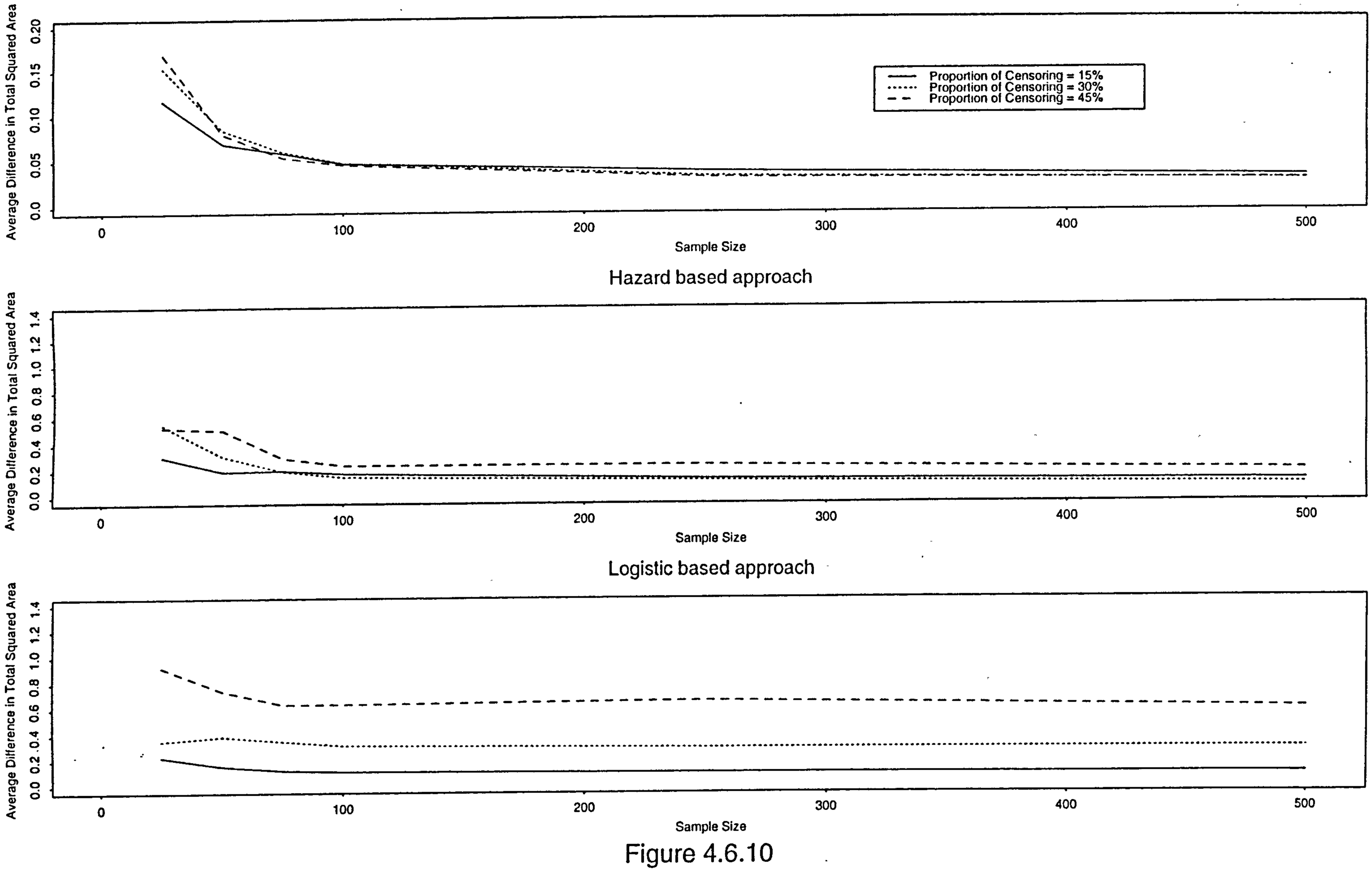
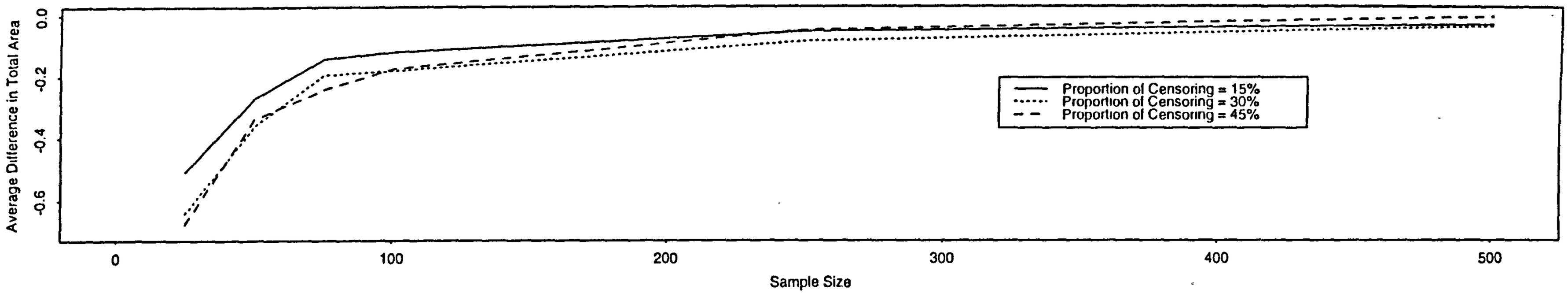

Kaplan Meier based approach
Average Difference in Total Squared Area
0.0
0.05
0.10
0.15
0.20
0
100
200
300
400
500
Sample Size
Proportion of Censoring = 15%
Proportion of Censoring = 30%
Proportion of Censoring = 45%
Hazard based approach
Average Difference in Total Squared Area
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4
0
100
200
300
400
500
Sample Size
Logistic based approach
Average Difference in Total Squared Area
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4
0
100
200
300
400
500
Sample Size

Figure 4.6.10

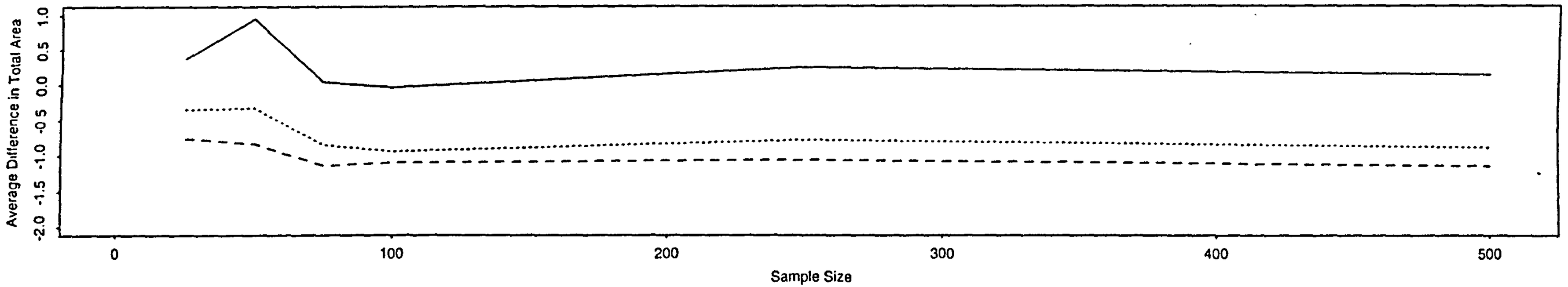
on the levels of precision. A large decrease in precision is observed as the proportion of censoring increases. A general point about the effect of the sample size is that it is interesting to observe that there is an initial increase in the precision of the estimates with sample size for each of the approaches but this precision does not appear to change after a sample of about 75 observations is obtained. This is reassuring as it implies that the estimators will be reasonably stable even for moderate sample sizes. A comparison of the levels of precision obtained in scenario 1 (Figure 4.6.3) with the levels of precision obtained in scenario 2 reveals that the Kaplan Meier based approach performs slightly better in the scenario where the covariate has no effect on survival rather than when the proportional hazards model is a suitable model to explain the underlying relationship. In contrast, the hazard based approach shows little difference across the two scenarios whilst the logistic based approach perform better when the proportional hazards model is an appropriate underlying model.

In terms of bias Figure 4.6.11 demonstrates that overall the Kaplan Meier based approach exhibits the least bias followed by the hazard based approach whilst the logistic based approach shows quite large bias in the estimates of survival (Note that the scales are different in each of the three frames). There is evidence from frame 1 of Figure 4.6.11 that for the Kaplan Meier based approach an increase in sample size will lead to a corresponding slight drop in the presence of bias and, for large sample sizes (greater than or equal to 250), the bias present with the Kaplan Meier approach appears negligible. However frame 2 of Figure 4.6.11 suggests that

Kaplan Meier based approach



Hazard based approach



Logistic based approach

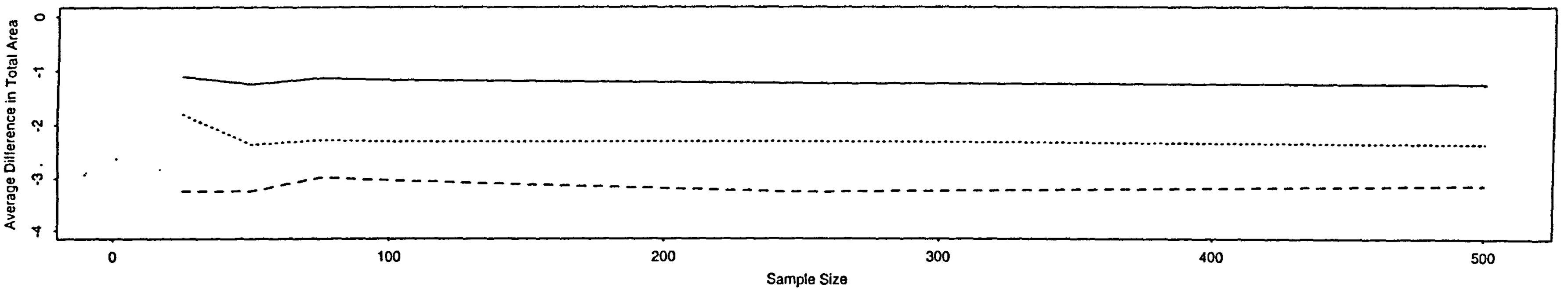


Figure 4.6.11

an increase in sample size will not necessarily lead to a decrease in bias for the hazard based approach and may in fact lead to a slight increase in bias for higher proportions of censoring (i.e. 30% and 45% censoring). The sample size has relatively little effect on the bias present with the logistic based approach. Also, for all three approaches, there is evidence to suggest that a change in the proportion of censoring will lead to a change in the amount of bias. For each method there is a larger presence of bias with the higher proportions of censoring. In general, the Kaplan Meier based approach produces results which show only a relatively small level of bias to be present. However, the bias present with the hazard based approach is more noticeable, particularly for 30% and 45% censoring. Finally the logistic approach produces very poor results in terms of bias. It clearly produces estimates of survival which are always less than the true survival regardless of sample size and proportion of censoring. It is also heavily influenced by the proportion of censoring with, in general, an increase in censoring leading to a corresponding increase in bias. Comparing the bias from scenario 1 (Figure 4.6.4) with scenario 2 it can be seen that the Kaplan Meier based approach exhibits marginally more bias when the proportional hazards model is appropriate. The hazard approach shows no difference across the two scenarios in terms of bias and the logistic based approach displays slightly less bias when the proportional hazards model is appropriate.

Figures 4.6.12 to 4.6.14 display plots of the coverage for each of the three methods of estimation. As in scenario 1, the covariate lower quartile value equals

0.25 with the covariate upper quartile value being 0.75. These indicate that the Kaplan Meier based approach produces the "best" coverage, followed by the hazard based approach with the logistic based approach again exhibiting poor levels of coverage. Regardless of the method used, the patterns of coverage are very similar to those observed in scenario 1 with the slight indication that, in general, the coverage is marginally better under scenario 1. Again, the Kaplan Meier based approach appears to be the only method which achieves the nominal, 95% level, of coverage.

The Kaplan Meier based approach (Figure 4.6.12) exhibits relatively high levels of coverage particularly for the later time values. The coverage is again higher with smaller proportions of censoring and the sample size has little, if any, effect on the coverage. With the hazard based approach (Figure 4.6.13) the coverage is, generally, not as high as would be anticipated. This is particularly the case for larger sample sizes and larger proportions of censoring. Again, the use of approximate intervals is possibly the main contributory factor to the poor levels of coverage. As in scenario 1 the logistic based approach (Figure 4.6.14) produces very poor levels of coverage, and when a combination of a large sample size and a large proportion of censoring are present it is doubtful if the true surface will be captured at any combination of time/covariate values.

These results would suggest that the Kaplan Meier based approach is a reasonably satisfactory method at reproducing estimates of survival from an

Kaplan Meier Based Approach

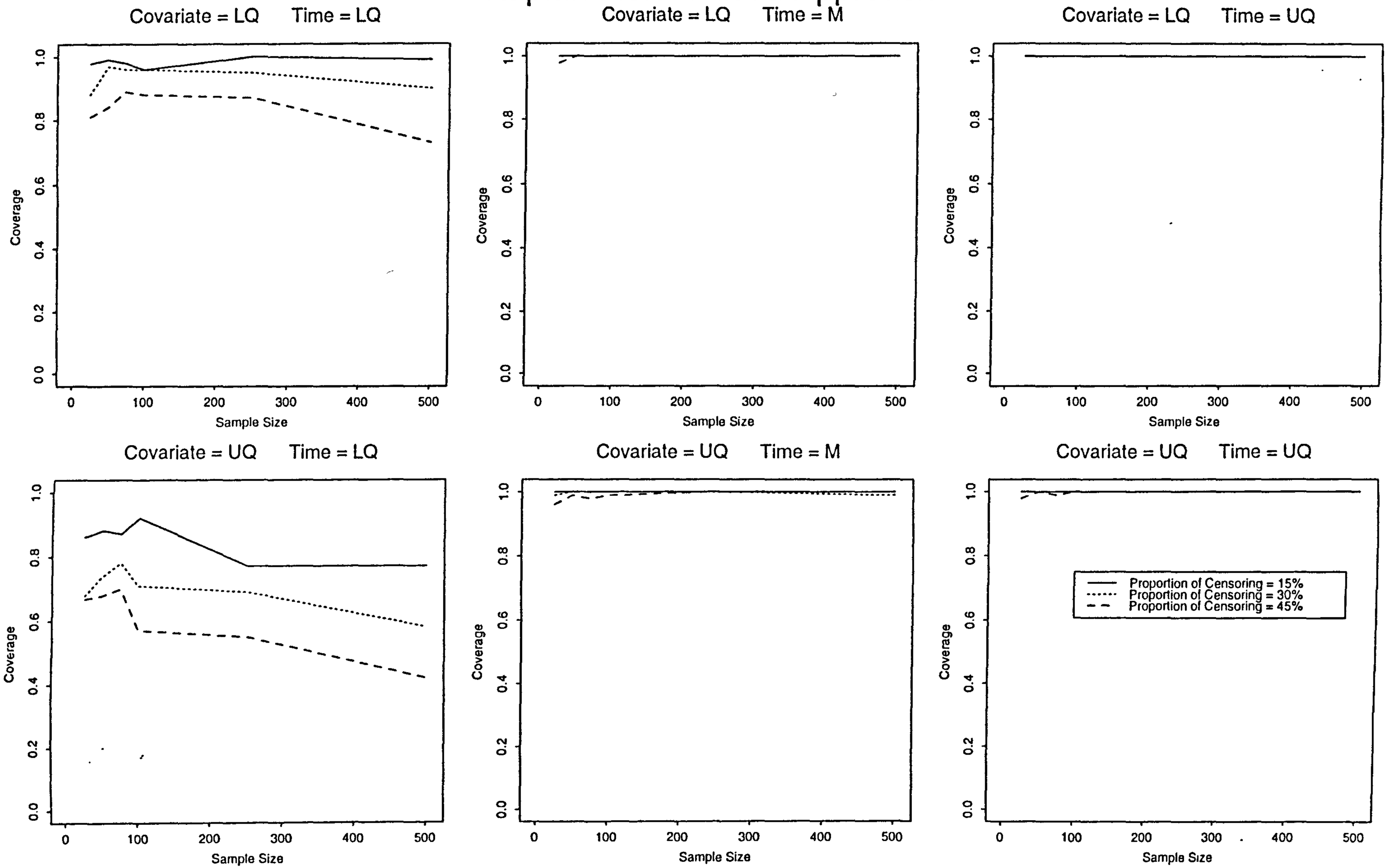


Figure 4.6.12

Hazard Based Approach

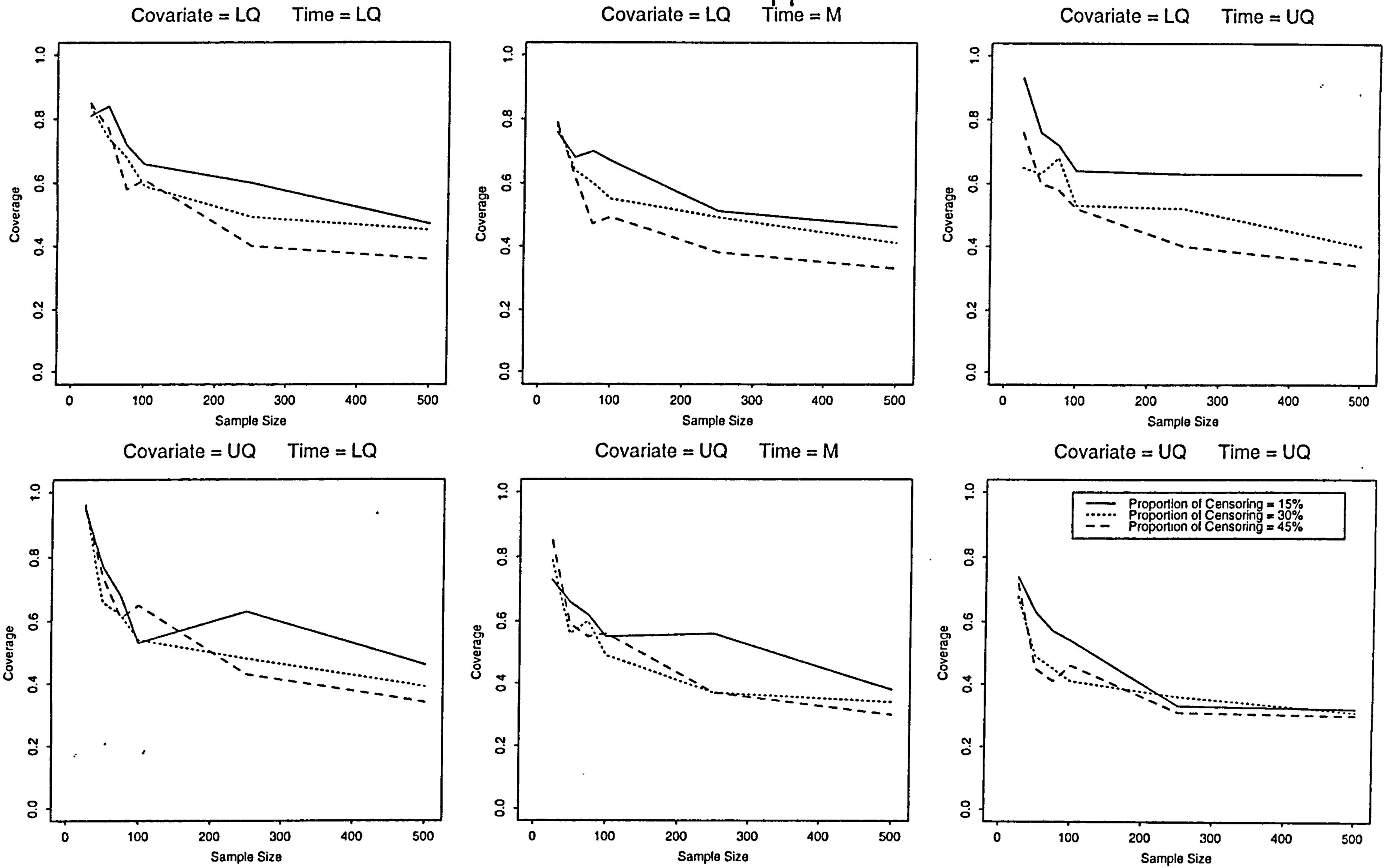


Figure 4.6.13

Logistic Based Approach

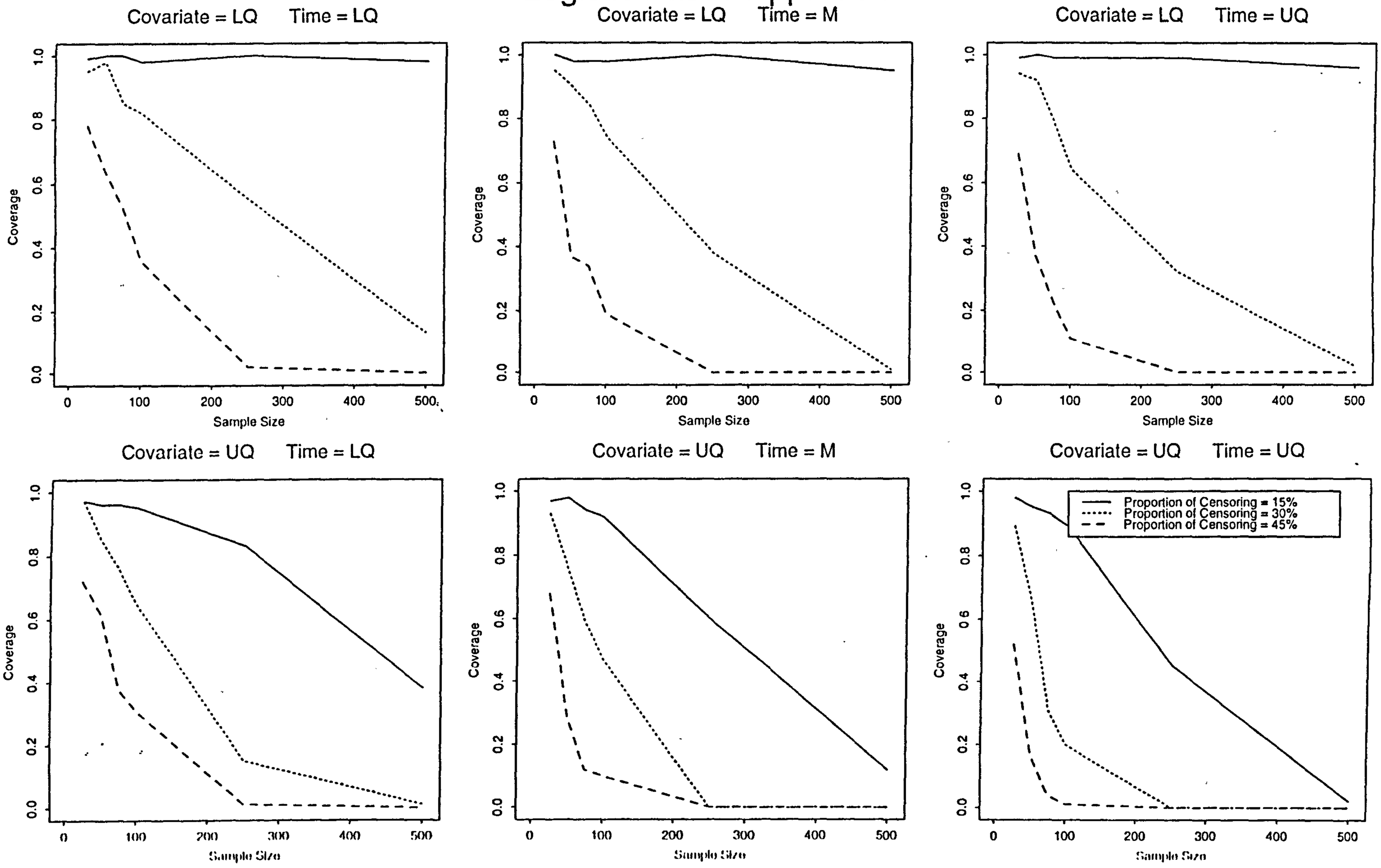


Figure 4.6.14

underlying, known, proportional hazards model. This method exhibited levels of precision and bias which were not excessive coupled with reasonable levels of coverage suggesting that this method would be the best method to use in practice. Although it is clearly not as good as the Kaplan Meier based approach, the hazard based approach performs reasonably well under this scenario in terms of precision and bias. However it must be pointed out that it produces levels of coverage which are lower than would be anticipated. Future work with the hazard based approach should again consider the use of an exact variance term in order to improve the levels of coverage. Finally reservations must be held about the logistic approach. It clearly produces estimates of survival which are less precise than those produced by either of the other two methods. On a more worrying note the estimates produced by the logistic based approach were clearly biased, producing underestimates of survival, with the degree of underestimation increasing as the proportion of censoring increases.

Section 4.6.3: Scenario 3: Simulated data with a single categorisation point

Here data have been simulated from a model where there is a single change in the hazard rate at a specified covariate point. Initially, in order to simulate data from this model where two distinct hazard rates are present, the covariate values are simulated from a simple Un(0,1) distribution. Then, for values of the covariate less than 0.5, survival times are simulated from an Ex(θ) distribution and censoring times from an Ex(ϕ) distribution. For values of the covariate greater than or equal to 0.5,

survival times are simulated from an Ex(kθ) distribution and censored times from an Ex(kϕ) distribution. In the simulations presented here k was chosen as 2.7. The parameter ϕ can then be varied to alter the proportion of censoring. Here, generating the survival times from two different exponential distributions implies that the two hazard rates will be from two ***different*** uniform distributions. Under this scenario, table 4.6.5 details the parameter values used in the simulations whilst table 4.6.6 summarises the corresponding observed follow-up times.

Survival times:	θ	=	0.1551	
Censoring times:	ϕ	=	0.0275, 0.0713, 0.1251	
corresponding to			15%, 30%, 45%	censoring

Table 4.6.5

Censoring Proportion		Observed follow-up times	
	Lower quartile	Median	Upper quartile
15%	2.4 years	6.0 years	13.0 years
30%	2 years	5 years	10 years
45%	1.6 years	4.0 years	8.7 years

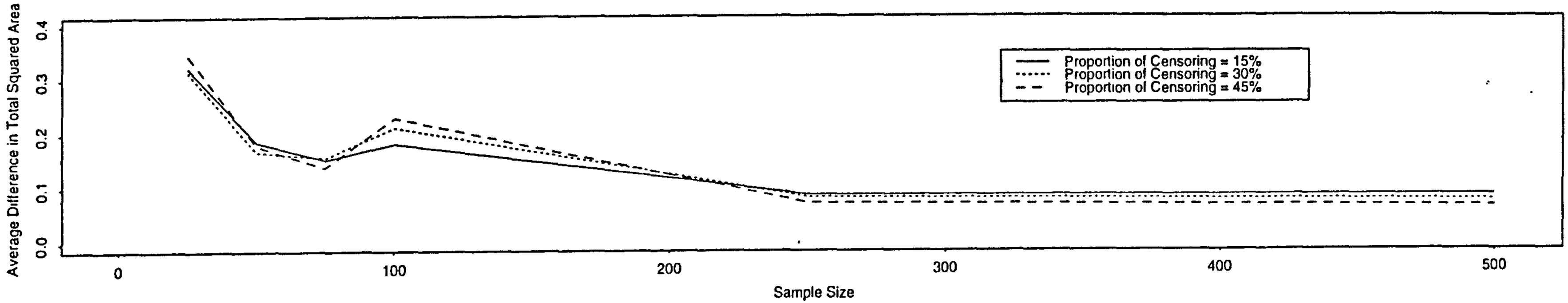
Table 4.6.6

Figures 4.6.15 to 4.6.19 show the results for this simulation study based on the use of an "optimal" smoothing parameter as outlined in Section 4.6.1. In terms

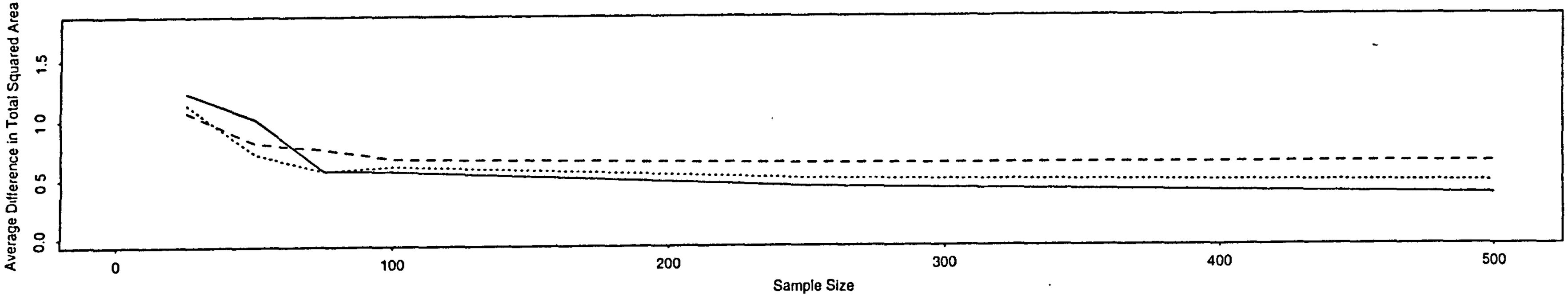
of precision Figure 4.6.15 indicates that the Kaplan Meier based approach again produces the most precise estimates (Note that the scale in frame 1 is different to the scales in frames 2 and 3). Comparing the hazard based approach and the logistic based approach, it appears that the hazard based approach produces more precise estimates although any difference between the two methods appears to only be present for smaller sample sizes. Regardless of the method used, the precision increases, in general, with increasing sample size. However, in contrast with scenarios 1 and 2, the precision does not necessarily decrease as the proportion of censoring increases. Also, unlike in scenarios 1 and 2, the effect of sample size on precision is more marked for both the hazard and logistic based approaches. In scenarios 1 and 2 increasing the sample size only led to a minor improvement in precision with the hazard and logistic based approaches. Here, increasing the sample size produces a clear improvement in the levels of precision with all 3 non-parametric approaches. Finally, regardless of method used, the precision achieved in this scenario is slightly poorer than obtained under scenarios 1 and 2.

Figure 4.6.16 displays plots of the bias against sample size for each of the three methods of estimation (Note that the scales are different in each of the three frames). The Kaplan Meier based approach exhibits the least bias followed by the hazard based approach then the logistic based approach. In general, the Kaplan Meier based approach displays relatively minor levels of bias which tend to decrease with increasing sample size. However, there is one exception to this, where the bias actually appears to increase when moving from 75 to 100 observations. For smaller

Kaplan Meier based approach



Hazard based approach



Logistic based approach

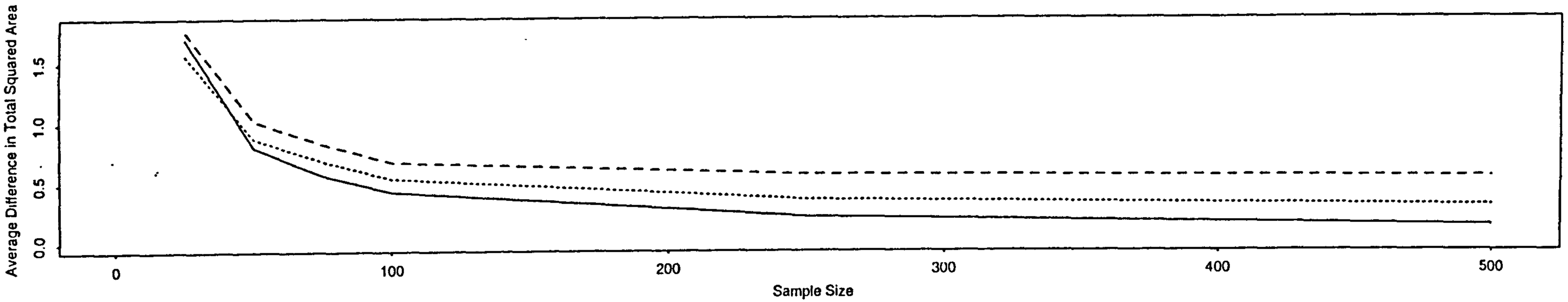
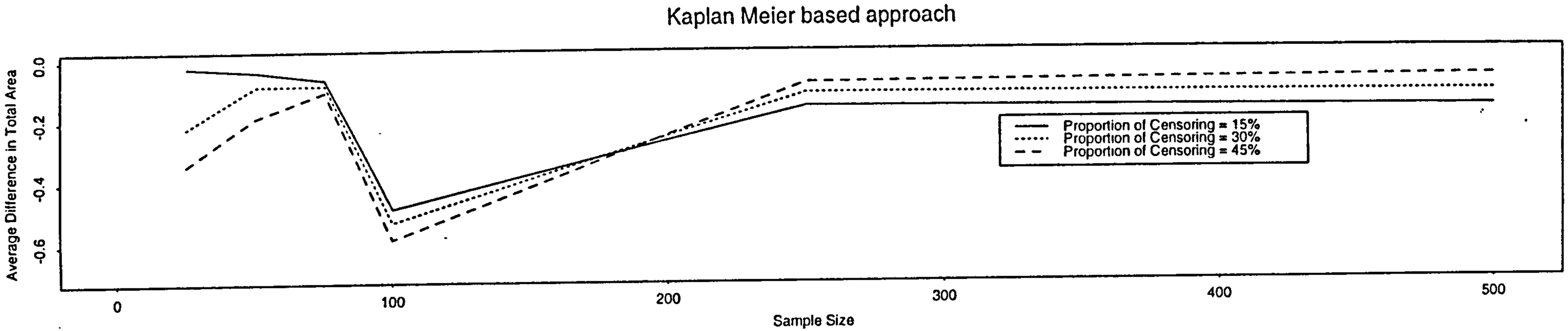
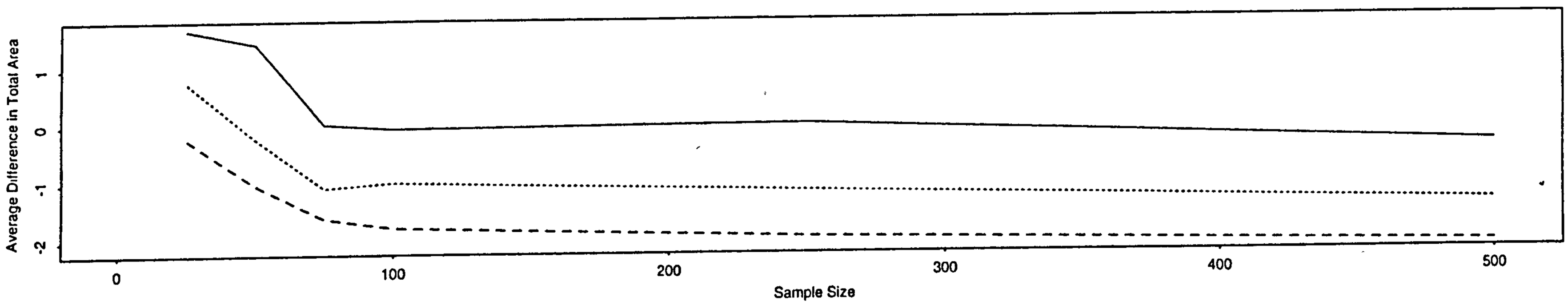


Figure 4.6.15


Kaplan Meier based approach
Average Difference in Total Area
0.0
-0.2
-0.4
-0.6
0
100
200
300
400
500
Sample Size
Proportion of Censoring = 15%
Proportion of Censoring = 30%
Proportion of Censoring = 45%


Hazard based approach
Average Difference in Total Area
1
0
-1
-2
0
100
200
300
400
500
Sample Size

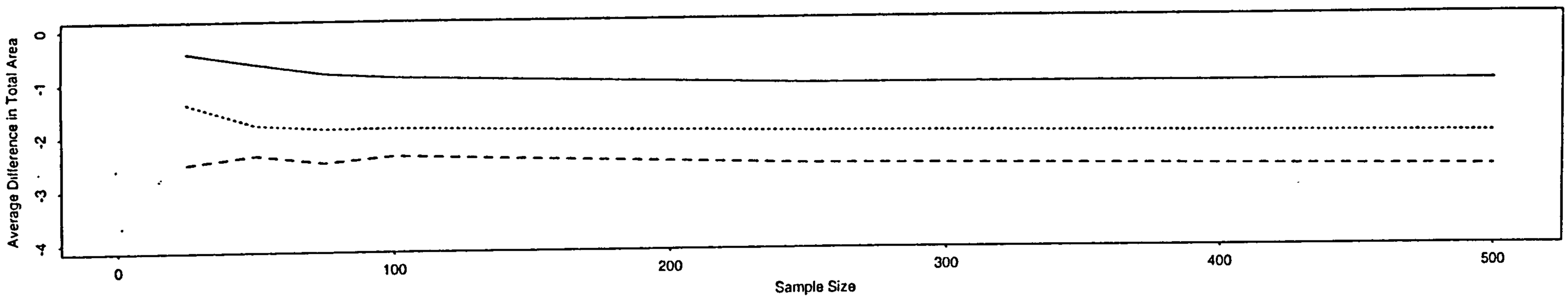

Logistic based approach
Average Difference in Total Area
0
-1
-2
-3
-4
0
100
200
300
400
500
Sample Size

Figure 4.6.16

sample sizes the bias appears to increase as the proportion of censoring increases. However, for larger sample sizes the bias seems to decrease as the proportion of censoring increases although any apparent differences are almost negligible. In this scenario, the levels of bias displayed by the Kaplan Meier based approach are similar to those displayed in scenarios 1 and 2. When consideration is given to the hazard based approach it is evident that it produces relatively minor levels of bias with 15% censoring. However with 30% and 45% censoring the hazard based approach will clearly underestimate the true survival, with the degree of underestimation appearing to be greater for larger sample sizes. Compared with scenarios 1 and 2, the levels of bias for the hazard based approach are slightly higher under this scenario, particularly for 30% and 45% censoring. Finally, the logistic based approach shows very little effect of the sample size on the levels of bias with the proportion of censoring dominating the pattern of bias. The bias is clearly greater with larger proportions of censoring. Regardless of the proportion of censoring or sample size, the logistic based method displays slightly less bias in this scenario compared to scenarios 1 and 2.

Figures 4.6.17 to 4.6.19 display plots of the coverage for each of the three methods of estimation. Notice that, in this scenario, the coverage has been evaluated at a third covariate value; at the actual location of the single categorisation point. Therefore the three chosen covariate values were as follows: 0.25 (the lower quartile), 0.5 (the median - ***location of the cutpoint***) and 0.75 (the upper quartile) These figures indicate that the Kaplan Meier based approach produces the "best"

Kaplan Meier Based Approach

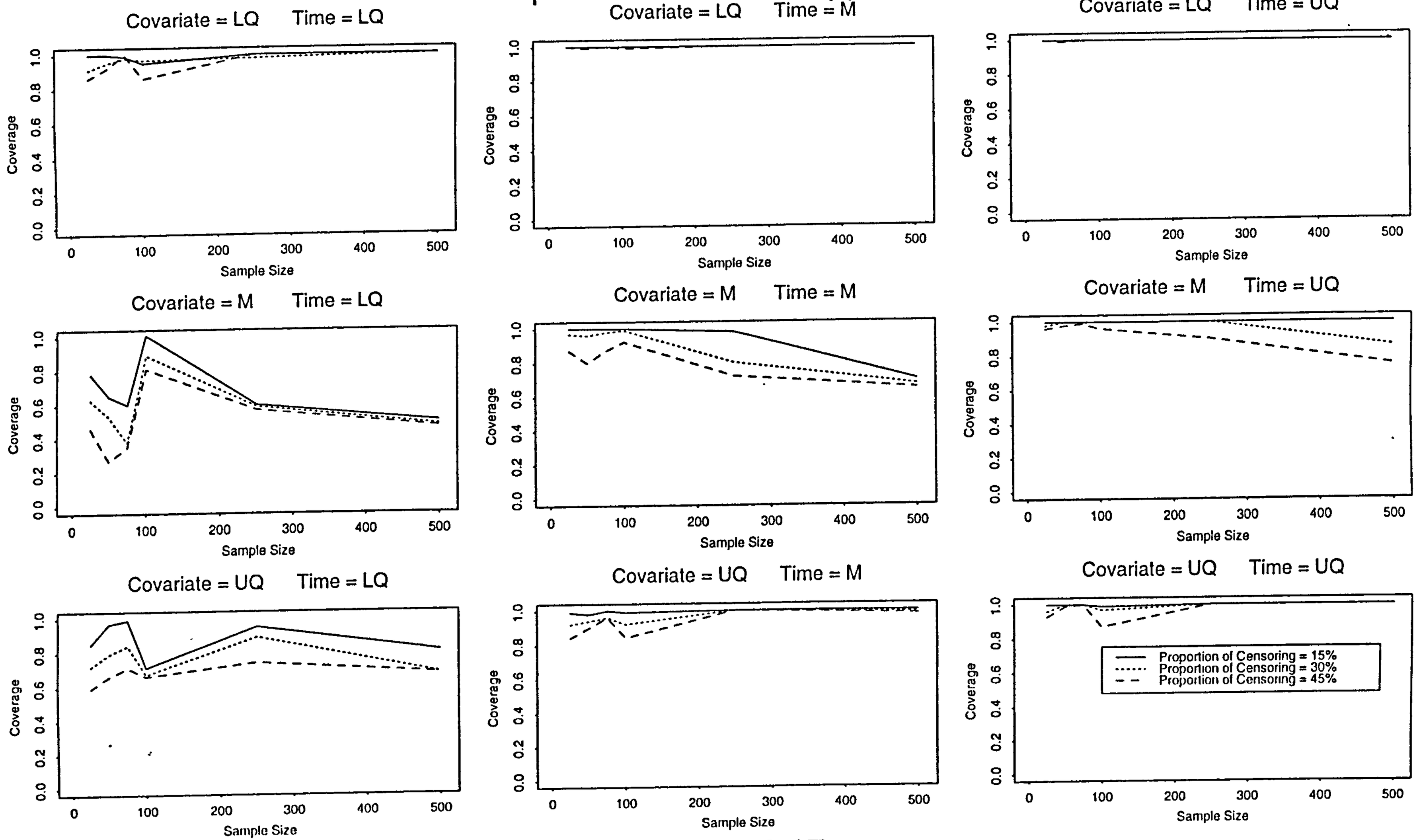


Figure 4.6.17

Hazard Based Approach

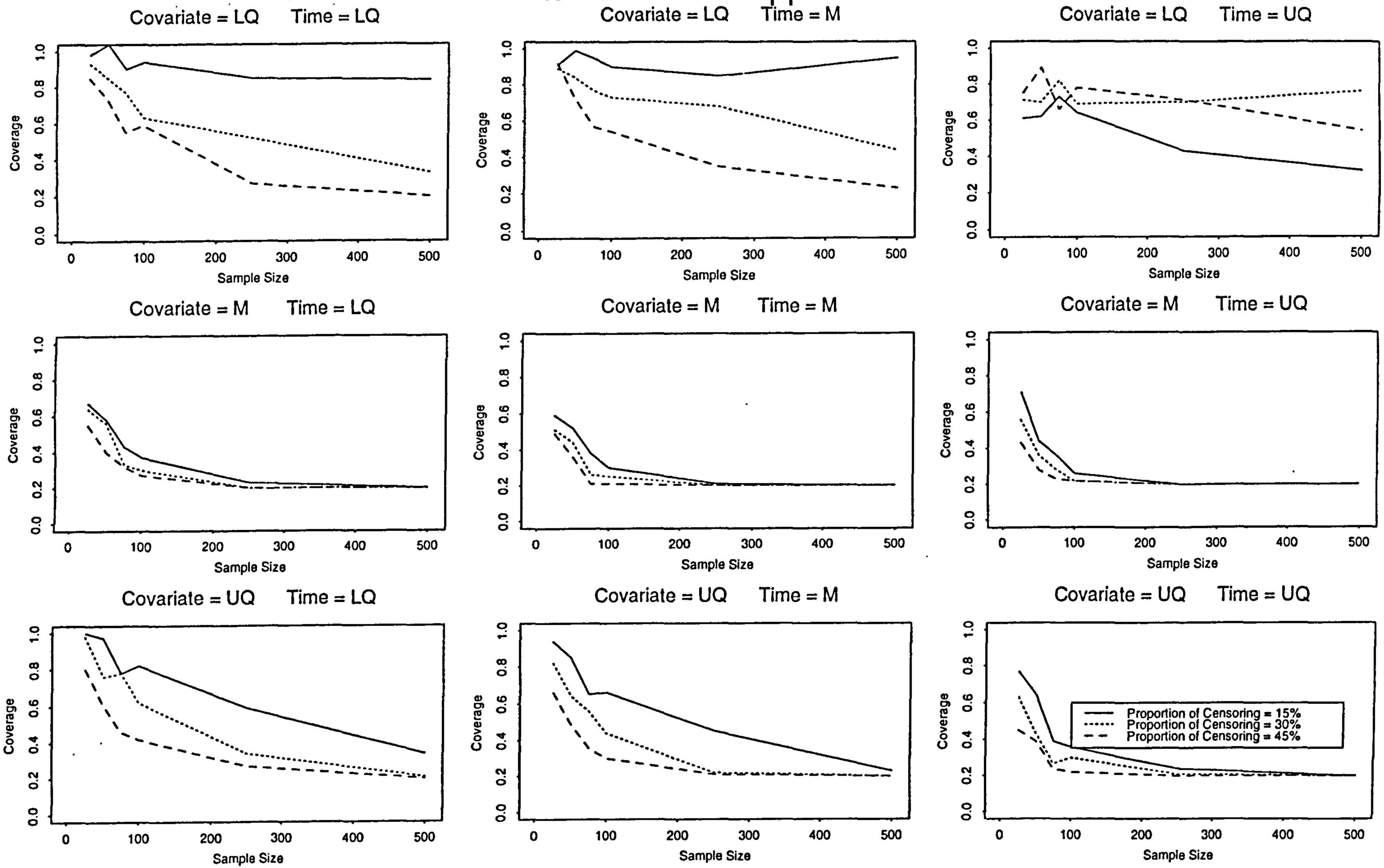


Figure 4.6.18

Logistic Based Approach

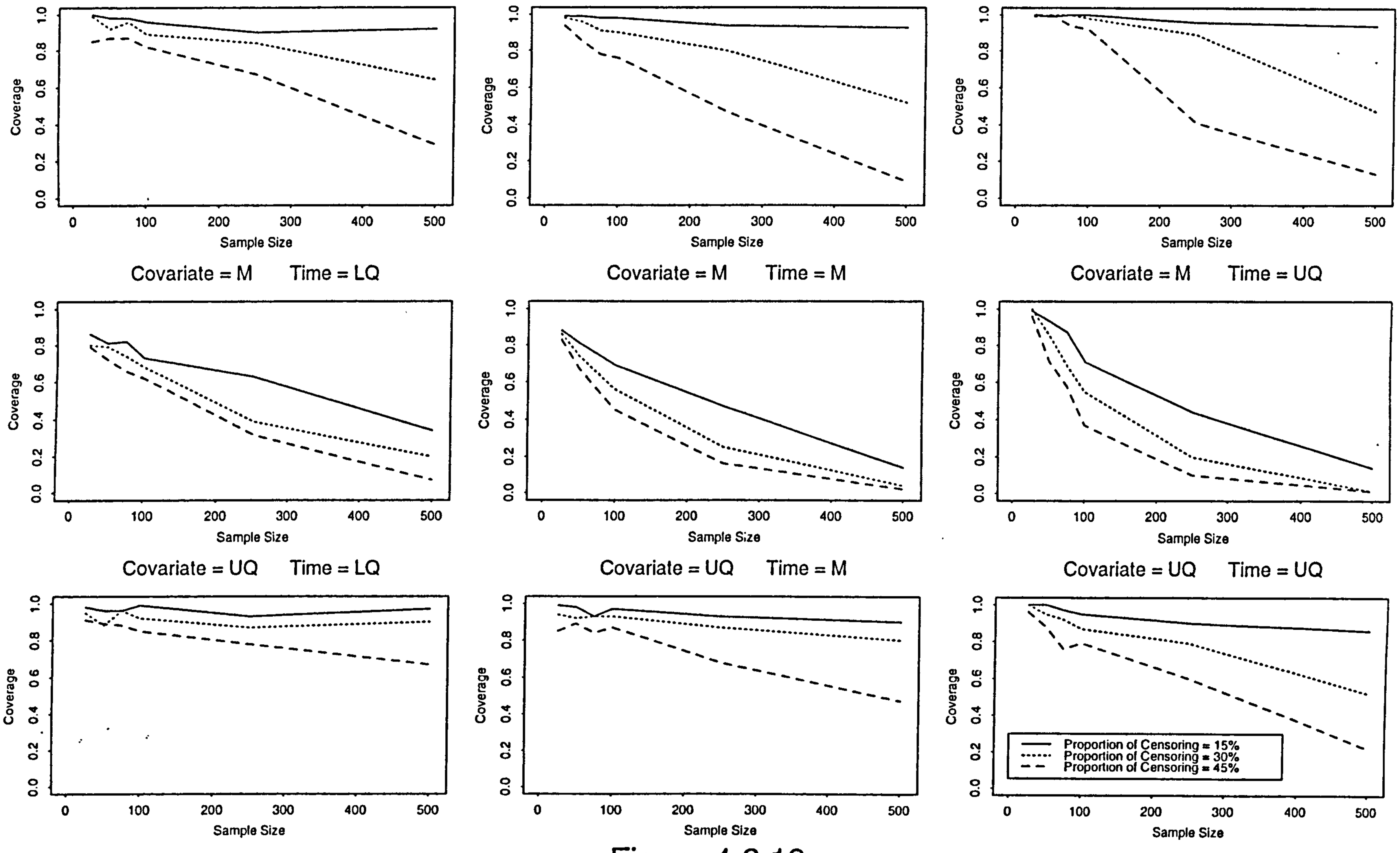


Figure 4.6.19

coverage, with both the hazard based approach and the logistic based approach exhibiting poor levels of coverage. The Kaplan Meier based approach is again the only method which comes close to achieving the nominal 95% level of coverage. Compared to scenarios 1 and 2, the levels of coverage displayed by the Kaplan Meier and particularly the hazard based approach are poorer in this scenario whereas the logistic based approach actually performs better in this scenario. In general, for each of the three methods of estimation, the coverage appears to drop at the location of the categorisation point (i.e. at the median covariate value), regardless of the time value. This is perhaps to be expected as this is the point where a distinct change in the pattern of survival exists and, as such, should prove the most difficult point at which to obtain "good" estimation. Each of the three methods show a decrease in coverage as the proportion of censoring increases and both the hazard and logistic based approaches again show a decrease in coverage as the sample size increases.

These results suggest that the Kaplan Meier based approach will produce the "best" results in terms of precision, bias and coverage under this scenario. An important point to make is that the hazard based approach does not perform particularly well in this scenario. A possible explanation for this may be obtained by comparison with the data example in section 4.5. During the analysis in section 4.5.2 it was observed that the hazard based approach was sensitive to unusual observations/data patterns. In this simulated scenario data has been generated which exhibits a rather unusual/unexpected pattern - a single , clear categorisation point. The hazard based approach may therefore have some difficulty in identifying this

"unusual" pattern. Perhaps, if the change in the pattern of survival were not as marked, the hazard based approach may perform better. Finally, the logistic based approach again produces clear underestimates of the true survival.

Section 4.6.4: Summary of the Results from the Simulation Study

The results of the simulations carried out here suggest that the Kaplan Meier based approach would be the best method to use to produce non-parametric estimates of survival in the presence of a single covariate. In each of the simulated scenarios this method produced estimates of survival which were, on average, reasonably precise when compared to the true survival, displayed only minimal bias and provided the target 95% levels of coverage. Although not performing as well as the Kaplan Meier based approach, the hazard based approach did produce reasonable estimates under the first two scenarios. However, the hazard based approach did not perform well for the case of a single categorisation point. This may be due, in part. to a lack of robustness with the hazard based approach as demonstrated in section 4.5.2.

The simulations also suggest that the logistic based approach is not satisfactory as it clearly underestimates survival regardless of sample size and scenario simulated with the bias present in these estimates increasing as the proportion of censoring increases. Regardless of the scenario under consideration

the logistic based approach also produces estimates which are, in general, not very precise. Finally this approach produces confidence intervals which provide very poor levels of coverage.

Section 4.7: Conclusions

In this chapter a selection of non-parametric methods for analysing survival data in the presence of a covariate have been introduced. The standard methodology was considered briefly and three fully non-parametric estimation methods proposed in an attempt to fit survival models which do not impose specific patterns (i.e. Proportional Hazards model) across the covariate. A major aim is also to fit models which allow possible categorisation points across the covariate to be detected. These categorisation points should be chosen at points where there are ***clear changes*** in the ***pattern*** of survival.

Three non-parametric methods were considered here; firstly a method which adapted the technique originally devised by Kaplan and Meier to incorporate a covariate (KMA); secondly an extension of the fully non-parametric hazard function due to Tanner and Wong to incorporate a covariate (TWA); thirdly an approach based on adapting the standard non-parametric logistic regression methodology with a fixed time point to consider time dependent survival (LRA).

The work presented here suggests that the (KMA) approach will produce the most sensible estimates of survival and will also produce results which allow issues of categorisation to be examined. The (TWA) approach also produces reasonably sensible estimates of survival but is likely to have difficulty in areas where unusual

observations are found, particularly if little data is available in these areas. This may imply that the (TWA) approach is likely to be of less use when examining issues of categorisation. Finally, there appear to be inherent problems with the (LRA) approach. The (LRA) approach produces a clear bias in terms of underestimating survival prospects regardless of the level of smoothing across the covariate.

In conclusion, it seems reasonable to use either the (KMA) or the (TWA) approach as the most sensible fully non-parametric estimators of survival in the presence of a covariate. However it should be borne in mind that the (TWA) approach does appear to be more sensitive to the presence of unusual observations particularly with small data sets.

Chapter 5

Conclusions and Future Work

Section 5.1: Conclusions

This thesis has considered the analysis of data within three basic medical contexts: a cohort study, a case/control study and a survival analysis. In each of these contexts the main aim has been to consider new, non-parametric methods of modelling the relationship between the response and explanatory variables. The primary reason for developing these new methods of analysis has been to allow "categorisations" for any explanatory variables to be highlighted. Categorisations for explanatory variables should be chosen at locations where there is a change in the effect the explanatory variable has on the response. Non-parametric methods of analysis are particularly appealing as they allow data to indicate the nature of any underlying relationship and hence highlight any potential categorisation points.

Chapter 2 considered the *cohort study* with a binary response. Firstly, the standard methodology of using a linear logistic model to explain any underlying relationship between the binary response and the explanatory was outlined. In order

to consider more general models, use was made of existing work on non-parametric modelling of the relationship between a binary response and one or more continuous explanatories (Copas (1983)). This methodology was applied in the context of a cohort study with a binary response to examine issues of categorisation both with one and two continuous explanatories. Here, the main innovation was to apply function derivatives as a more formal method for highlighting possible categorisations. The main finding from this chapter was that the use of function derivatives in conjunction with the non-parametric logistic model gave a clearer picture of the location of categorisation points than could be obtained by only giving consideration to the non-parametric logistic model.

Chapter 3 considered the risk associated with an interval scaled discrete risk factor in ***case/control studies***. Initially the standard methodology of using the conditional linear logistic model as a method for analysis of such data was outlined. Two new ***non-parametric*** methods of analysing data from case/control studies with an ***interval scaled discrete*** risk factor were presented; one based on a "pairwise cells" approach and the other based on considering the conditional likelihood. Both methods were applied to a case study in order to highlight potential categorisations for an interval scaled discrete explanatory variable. In the case study the methods identified similar, although not identical, locations for any categorisation points. They were also in general agreement as to the number of categorisation points that should be imposed. The two methods were then compared through a simulation study. In terms of degree of precision and level of bias ***both*** methods were found to provide reasonable estimation in each of the scenarios in the simulation study. In the

simulation study, the conditional likelihood method appeared to be superior both in terms of ***precision*** and ***coverage*** whilst the pairwise cells method appeared slightly superior in terms of ***bias***. Possible extensions of both methods to model the relationship between the response variable and a ***continuous*** explanatory variable were presented. In a brief case study these extensions to incorporate a continuous explanatory appeared to produce logical estimates of the relationship between the response and the explanatory.

Finally, Chapter 4 examined the ***analysis of survival data*** with one continuous explanatory variable. It considered the standard analysis which uses the proportional hazards model to describe the effect of a single continuous explanatory variable on survival. Three non-parametric approaches for modelling the underlying relationship between a continuous explanatory and survival were proposed: an extension of the method of Kaplan and Meier (1958) to incorporate a continuous explanatory, a method based on extending the idea of Tanner and Wong (1983) of non-parametrically estimating the hazard function to include a continuous explanatory and an attempt to adapt non-parametric logistic modelling to incorporate a time dependent binary response. Each of the methods was applied to an example from the medical field and both the Kaplan Meier approach and the hazard based approach produced reasonable solutions. The two methods were also used to highlight potential categorisations for a continuous explanatory. However, the logistic based approach produced estimates of survival which appeared to underestimate the pattern of survival. These findings were confirmed by a simulation study which suggested that both the Kaplan Meier and hazard based approaches were plausible methods of

estimation. Both of the methods were able to reproduce given scenarios with reasonable precision and acceptable levels of bias. However, the Kaplan Meier based approach proved superior to the hazard based approach in terms of both precision and bias. It also proved ***far superior*** in terms of coverage. Therefore, there was clear evidence from both the "real data" example and, particularly, the simulation study to favour the Kaplan Meier based approach. In areas where the data was quite sparse, unusual observations occasionally had a large effect on the estimates of survival produced by the hazard based approach whereas the Kaplan Meier approach appeared more robust. Again, in the simulation study, the logistic based approach produced underestimates of the true survival.

In summary, in each chapter/context suitable non-parametric methods have been found to model the relationship between the response of interest and a single continuous / interval scaled discrete explanatory variable. These methods have been used to examine the primary aim of this work; to suggest, in each context, data based methods which can be used to highlight potential categorisations for explanatory variables.

Section 5.2 Future work

Although suitable methods have been found in each of the three contexts there is still further work to be carried out. The non-parametric methods of analysis proposed for use in ***case/control studies*** and ***survival analysis*** only deal with one explanatory. Future work is required to extend or adapt these methods to deal with more than one explanatory.

Further, in each of the three contexts presented here the non-parametric methods rely heavily on the use of smoothing techniques; these involve the choice of a smoothing parameter. On each occasion the smoothing parameter has been chosen based on a subjective search method. A more automatic method for choosing the smoothing parameter is essential. Current work in this field tends to focus on the use of the plug-in methods for choice of smoothing parameter as mentioned in Chapter 1. These methods must be given due consideration here as an alternative to the subjective search method.

Finally, in the analysis of both case/control studies and survival data, the choice of location of categorisation points was essentially based on a large degree of subjectivity. In the analysis of cohort studies more formal techniques based on function derivatives were used to highlight potential categorisation points. More formal methodology should be applied in order to highlight categorisation points in

both the analysis of case/control studies and survival data to remove the subjectivity involved in the choice of these categorisation points.

In conclusion, within each chapter of this thesis, consideration has been given to the analysis of data from different medical frameworks. In each situation non-parametric methods have been proposed for modelling the relationship between the response of interest and the explanatory variable. These non-parametric methods have been used to highlight the locations of categorisation points for a single explanatory variable. Work still needs to be carried out in this area to extend the methods presented here to deal with more than one explanatory variable and to remove the degree of subjectivity involved in the choice of both the smoothing parameter and in the location of any categorisation points.

Appendix A: Calculation of Covariance Terms in $\hat{V}_2\left(\underline{\hat{\beta}}\right)$

The covariance matrix is of the form given below

$$\hat{V}_2\left(\underline{\hat{\beta}}\right) = \begin{bmatrix} \hat{V}\left(\hat{\beta}_{10}\right) & \hat{cov}\left(\hat{\beta}_{10}\hat{\beta}_{20}\right) & \hat{cov}\left(\hat{\beta}_{10}\hat{\beta}_{21}\right) & \cdots & \hat{cov}\left(\hat{\beta}_{10}\hat{\beta}_{k,k-1}\right) \\ \hat{cov}\left(\hat{\beta}_{20}\hat{\beta}_{10}\right) & \hat{V}\left(\hat{\beta}_{20}\right) & \hat{cov}\left(\hat{\beta}_{20}\hat{\beta}_{21}\right) & \cdots & \hat{cov}\left(\hat{\beta}_{20}\hat{\beta}_{k,k-1}\right) \\ \hat{cov}\left(\hat{\beta}_{21}\hat{\beta}_{10}\right) & \hat{cov}\left(\hat{\beta}_{21}\hat{\beta}_{20}\right) & \hat{V}\left(\hat{\beta}_{21}\right) & \cdots & \hat{cov}\left(\hat{\beta}_{21}\hat{\beta}_{k,k-1}\right) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{cov}\left(\hat{\beta}_{k,k-1}\hat{\beta}_{10}\right) & \hat{cov}\left(\hat{\beta}_{k,k-1}\hat{\beta}_{20}\right) & \hat{cov}\left(\hat{\beta}_{k,k-1}\hat{\beta}_{21}\right) & \cdots & \hat{V}\left(\hat{\beta}_{k,k-1}\right) \end{bmatrix}$$

Each of the terms in the matrix is produced in the same manner. As a simple example. consider calculation of $\hat{cov}\left(\hat{\beta}_{10},\hat{\beta}_{20}\right)$ when a ***first order neighbourhood of smoothing*** is present. In this situation the relevant ***neighbourhood counts*** are

$$\dot{n}_{10} = n_{10} + n_{20} + n_{21} \qquad \dot{n}_{01} = n_{01} + n_{02} + n_{12}$$
$$\dot{n}_{20} = n_{10} + n_{20} + n_{21} + n_{30} \qquad \dot{n}_{02} = n_{01} + n_{02} + n_{12} + n_{03}$$

with

$$\hat{cov}\left(\hat{\beta}_{10},\hat{\beta}_{20}\right) = cov\left\{\log\left(\frac{\dot{n}_{10}}{\dot{n}_{01}}\right),\log\left(\frac{\dot{n}_{20}}{\dot{n}_{02}}\right)\right\}$$

$$= cov\left\{\log\left(\dot{n}_{10}\right),\log\left(\dot{n}_{20}\right)\right\} - cov\left\{\log\left(\dot{n}_{10}\right),\log\left(\dot{n}_{02}\right)\right\} - cov\left\{\log\left(\dot{n}_{01}\right),\log\left(\dot{n}_{20}\right)\right\} + cov\left\{\log\left(\dot{n}_{01}\right),\log\left(\dot{n}_{02}\right)\right\}$$

$$= A - B - C + D \qquad -(*)$$

Consider each of the terms A, B, C and D separately

A:

An application of a first order Taylor expansion provides

$$\hat{cov}\left\{\log\left(n_{10}^{*}\right),\log\left(n_{20}^{*}\right)\right\} \approx \frac{1}{n_{10}^{*}} * \frac{1}{n_{20}^{*}} * \hat{cov}\left(n_{10}^{*}, n_{20}^{*}\right)$$

Now,

$$\hat{cov}\left(n_{10}^{*}, n_{20}^{*}\right) = \hat{cov}\left(n_{10}+n_{20}+n_{21}, n_{10}+n_{20}+n_{21}+n_{30}\right)$$

$$= \hat{V}\left(n_{10}\right)+2\hat{cov}\left(n_{10},n_{20}\right)+2\hat{cov}\left(n_{10},n_{21}\right)+\hat{cov}\left(n_{10},n_{30}\right)+\hat{V}\left(n_{20}\right)$$

$$+2\hat{cov}\left(n_{20},n_{21}\right)+\hat{cov}\left(n_{20},n_{30}\right)+\hat{cov}\left(n_{21},n_{30}\right)+\hat{V}\left(n_{21}\right)$$

$$= n_{10}\left(\frac{N-n_{10}}{N}\right)-2\frac{n_{10}n_{20}}{N}-2\frac{n_{10}n_{21}}{N}-\frac{n_{10}n_{30}}{N}+n_{20}\left(\frac{N-n_{20}}{N}\right)$$

$$-2\frac{n_{20}n_{21}}{N}-\frac{n_{20}n_{30}}{N}-\frac{n_{21}n_{30}}{N}+n_{21}\left(\frac{N-n_{21}}{N}\right)$$

Hence,

$$\hat{cov}\left\{\log\left(n_{10}^{*}\right),\log\left(n_{20}^{*}\right)\right\} = \frac{1}{n_{10}+n_{20}+n_{21}} * \frac{1}{n_{10}+n_{20}+n_{21}+n_{30}}$$

$$* \left[n_{10}\left(\frac{N-n_{10}}{N}\right)-2\frac{n_{10}n_{20}}{N}-2\frac{n_{10}n_{21}}{N}-\frac{n_{10}n_{30}}{N}+n_{20}\left(\frac{N-n_{20}}{N}\right)\right.$$

$$\left.-2\frac{n_{20}n_{21}}{N}-\frac{n_{20}n_{30}}{N}-\frac{n_{21}n_{30}}{N}+n_{21}\left(\frac{N-n_{21}}{N}\right)\right]$$

B:

An application of a first order Taylor expansion provides

$$\hat{cov}\left\{\log\left(n_{10}^{*}\right),\log\left(n_{02}^{*}\right)\right\} \approx \frac{1}{n_{10}^{*}} * \frac{1}{n_{02}^{*}} * \hat{cov}\left(n_{10}^{*}, n_{02}^{*}\right)$$

Now,

$$\hat{cov}\left(n_{10}^{*}, n_{02}^{*}\right) = \hat{cov}\left(n_{10} + n_{20} + n_{21}, n_{01} + n_{02} + n_{12} + n_{03}\right)$$

$$= \hat{cov}\left(n_{10}, n_{01}\right) + \hat{cov}\left(n_{10}, n_{02}\right) + \hat{cov}\left(n_{10}, n_{12}\right) + \hat{cov}\left(n_{10}, n_{03}\right) + \hat{cov}\left(n_{20}, n_{01}\right) + \hat{cov}\left(n_{20}, n_{02}\right)$$

$$+ \hat{cov}\left(n_{20}, n_{12}\right) + \hat{cov}\left(n_{20}, n_{03}\right) + \hat{cov}\left(n_{21}, n_{01}\right) + \hat{cov}\left(n_{21}, n_{02}\right) + \hat{cov}\left(n_{21}, n_{12}\right) + \hat{cov}\left(n_{21}, n_{03}\right)$$

$$= -\frac{n_{01}n_{10}}{N} - \frac{n_{10}n_{02}}{N} - \frac{n_{10}n_{12}}{N} - \frac{n_{10}n_{03}}{N} - \frac{n_{20}n_{01}}{N} - \frac{n_{20}n_{02}}{N}$$

$$- \frac{n_{20}n_{12}}{N} - \frac{n_{20}n_{03}}{N} - \frac{n_{21}n_{01}}{N} - \frac{n_{21}n_{02}}{N} - \frac{n_{21}n_{12}}{N} - \frac{n_{21}n_{03}}{N}$$

Hence,

$$\hat{cov}\left\{\log\left(n_{10}^{*}\right),\log\left(n_{02}^{*}\right)\right\} = \frac{1}{n_{10} + n_{20} + n_{21}} * \frac{1}{n_{01} + n_{02} + n_{12} + n_{03}}$$

$$* \left[-\frac{n_{10}n_{01}}{N} - \frac{n_{10}n_{02}}{N} - \frac{n_{10}n_{12}}{N} - \frac{n_{10}n_{03}}{N} - \frac{n_{20}n_{01}}{N} - \frac{n_{20}n_{02}}{N}\right.$$

$$\left. - \frac{n_{20}n_{12}}{N} - \frac{n_{20}n_{03}}{N} - \frac{n_{21}n_{01}}{N} - \frac{n_{21}n_{02}}{N} - \frac{n_{21}n_{12}}{N} - \frac{n_{21}n_{03}}{N}\right]$$

C:

A Similar argument to **B** provides

$$\hat{cov}\left\{\log\left(n_{01}^{*}\right),\log\left(n_{20}^{*}\right)\right\} = \frac{1}{n_{01}+n_{02}+n_{12}} * \frac{1}{n_{10}+n_{20}+n_{21}+n_{30}}$$

$$* \left[-\frac{n_{01}n_{10}}{N} - \frac{n_{01}n_{20}}{N} - \frac{n_{01}n_{21}}{N} - \frac{n_{01}n_{30}}{N} - \frac{n_{02}n_{10}}{N} - \frac{n_{02}n_{20}}{N} \right.$$

$$\left. - \frac{n_{02}n_{21}}{N} - \frac{n_{02}n_{30}}{N} - \frac{n_{12}n_{10}}{N} - \frac{n_{12}n_{20}}{N} - \frac{n_{12}n_{21}}{N} - \frac{n_{12}n_{30}}{N} \right]$$

D:

A similar argument to **A** provides

$$\hat{cov}\left\{\log\left(n_{01}^{*}\right),\log\left(n_{02}^{*}\right)\right\} = \frac{1}{n_{01}+n_{02}+n_{12}} * \frac{1}{n_{01}+n_{02}+n_{12}+n_{03}}$$

$$* \left[n_{01}\left(\frac{N-n_{01}}{N}\right) - 2\frac{n_{01}n_{02}}{N} - 2\frac{n_{01}n_{12}}{N} - \frac{n_{01}n_{03}}{N} + n_{02}\left(\frac{N-n_{02}}{N}\right) \right.$$

$$\left. - 2\frac{n_{02}n_{12}}{N} - \frac{n_{02}n_{03}}{N} - \frac{n_{12}n_{03}}{N} + n_{12}\left(\frac{N-n_{12}}{N}\right) \right]$$

Hence, from (*), $\hat{cov}\left(\hat{\beta}_{10},\hat{\beta}_{20}\right)$ is calculated as A-B-C+D

Appendix B: Derivation of the Variance of $\hat{h}(t;z)$

Definitions

Let $L_1, ..., L_n$ represent the lifetimes of the n items under study

$C_1, ..., C_n$ represent the corresponding censored times

$T_i = \min(L_i, C_i)$ and $\delta_i = I_{L_i < C_i}$

$L_1, ..., L_n$ are iid with cdf F_L and density function f_L

$C_1, ..., C_n$ are iid with cdf F_C and density function f_C

Denote the cdf and density function of $T_1, ..., T_n$ by F and f without any subscript.

From Tanner and Wong (1983) let

$$m(y) = f_L(y)\left(1 - F_C(y)\right) / f(y) \qquad \text{for } f(y) > 0 \qquad \text{- (i)}$$

$$E\left(\delta_{(i)} / t_{(i)} = y\right) = m(y) \qquad \forall\, i \qquad \text{- (ii)}$$

$$E\left(\delta_{(r)}\delta_{(s)} / t_{(r)} = y, t_{(s)} = x\right) = m(y)m(x) \qquad \forall\, r < s \text{ and } \forall\, y < x \qquad \text{- (iii)}$$

Also, by definition

$$h(y) = \frac{f_L(y)}{1 - F_L(y)} \qquad \text{- (iv)}$$

$$\delta_i^2 = \delta_i \quad \Rightarrow \quad \delta_{(i)}^2 = \delta_{(i)} \qquad \text{- (v)}$$

Now,

$$\hat{h}(t;z) = \sum_{j=1}^{n} \frac{\delta_{(j)}}{n-j+1} K_{h_1}\left(t - t_{(j)}\right) K_{h_2}\left(z - z_{(j)}\right)$$

and

$$\text{var}\left(\hat{h}(t;z)\right) = \left[E\left(\hat{h}(t;z)^2\right)\right] - \left[E\left(\hat{h}(t;z)\right)\right]^2$$

(A) Calculation of $E\left(\hat{h}(t;z)\right)$

$$E\left(\hat{h}(t;z)\right) = E_{t_{(j)}} \; E\left\{\hat{h}(t;z) / t_{(j)} = y\right\} \text{ by the law of iterated expectation}$$

$$= E_{t_{(j)}}\left[\sum_{j=1}^{n} \frac{1}{n-j+1} \; E\left\{\delta_{(j)} K_{h_1}\left(t - t_j\right) / t_{(j)} = y\right\} K_{h_2}\left(z - z_{(j)}\right)\right]$$

$$= E_{t_{(j)}}\left[\sum_{j=1}^{n} \frac{1}{n-j+1} K_{h_1}(t-y)\, m(y)\, K_{h_2}\left(z - z_{(j)}\right)\right] \qquad \text{by (ii)}$$

To complete the calculation, the density of $t_{(j)}$, $f_{t_{(j)}}(y)$, is required.

By standard calculation, this is

$$\frac{n!}{(j-1)!(n-j)!} F(y)^{j-1}\left[1 - F(y)\right]^{n-j} f(y) \qquad \text{- (vi)}$$

Hence,

$$E\left(\hat{h}(t;z)\right) = \sum_{j=1}^{n} \frac{1}{n-j+1} \frac{n!}{(j-1)!(n-1)!} \int_{y} K_{h_1}(t-y)\, K_{h_2}\left(z-z_{(j)}\right) m(y) f(y) \left(F(y)\right)^{j-i} \left(1-F(y)\right)^{n-j} dy$$

Now,

$$m(y) f(y) = f_L(y)\left(1-F_C(y)\right) \qquad \text{by (i)}$$

$$= h(y)\left(1-F_L(y)\right)\left(1-F_C(y)\right) \qquad \text{by (iv)}$$

Now, since

$$T_i = \min\left(L_i, C_i\right)$$

it is immediately obvious that

$$m(y) f(y) = h(y)\left(1-F(y)\right)$$

Therefore,

$$E\left(\hat{h}(t;z)\right) = \sum_{j=1}^{n} \frac{n!}{(j-1)!(n-j+1)!} K_{h_2}\left(z-z_{(j)}\right) \int_{y} h(y) \left(F(y)\right)^{j-1} \left(1-F(y)\right)^{n-j+1} K_{h_1}(t-y)\, dy$$

$$= \int_{y} h(y)\, K_{h_1}(t-y)\, G_1(y/z)\, dy$$

where

$$G_1(y/z) = \sum_{j=1}^{n} \frac{n!}{(j-1)!(n-j+1)!} K_{h_2}\left(z - z_{(j)}\right)\left(F(y)\right)^{j-1}\left(1-F(y)\right)^{n-j+1}$$

(B) Calculation of $E\left(\hat{h}^2(t;z)\right)$

$$\hat{h}(t;z) = \sum_{j=1}^{n} \frac{\delta_{(j)}}{n-j+1} K_{h_1}\left(t - t_{(j)}\right) K_{h_2}\left(z - z_{(j)}\right)$$

$$= \sum_{j=1}^{n} c_j \delta_{(j)} K_{h1}\left(t - t_{(j)}\right)$$

where

$$c_j = \frac{1}{n-j+1} K_{h_2}\left(z - z_{(j)}\right)$$

Therefore, by definition

$$E\left(\hat{h}^2(t;z)\right) = \sum_{j=1}^{n} c_j^2 \, E\left(\delta_{(j)}^2 K_{h1}^2\left(t - t_{(j)}\right)\right)$$

$$+ \; 2\sum_{r<s}\sum c_r c_s E\left(\delta_{(r)}\delta(s) K_{h1}\left(t - t_{(r)}\right) K_{h2}\left(t - t_{(s)}\right)\right)$$

symmetric in r and s

Now,

$E\left(\delta_{(j)}^2 K_{h_1}^2\left(t - t_{(j)}\right)\right) = E_{\left(t_{(j)}\right)} E\left(\delta_{(j)}^2 K_{h_1}^2\left(t - t_{(j)}\right) / t_{(j)} = y\right)$ by the law of iterated expectation

$$= \; E_{t_{(j)}}\left(m(y)\,K_{h_1}^2(t-y)\right) \qquad \text{by (ii) and (v)}$$

$$= \int_y \frac{n!}{(j-1)!\,(n-j)!}\,h(y)\left(F(y)\right)^{j-1}\;\left(1-F(y)\right)^{n-j+1}\,K_{h1}^2(t-y)\,dy$$

analogous to $E\left(\hat{h}(t;z)\right)$

Further,

$$E\left(\delta_{(r)}\delta(s)\,K_{h_1}\left(t-t_{(r)}\right)K_{h_1}\left(t-t_{(s)}\right)\right) = E_{t_{(r)}t_{(s)}}\left(m(y)\,m(x)\,K_{h_1}(t-y)\,K_{h_1}(t-x)\right) \quad \text{by (iii)}$$

Now, by standard calculation, the joint pdf of $\left(t_{(r)},t_{(s)}\right)$ is

$$\frac{n!}{(r-1)!\,(s-r-1)!\,(n-s)!}\left(F(y)\right)^{r-1}\left(F(x)-F(y)\right)^{s-r-1}\left(1-F(x)\right)^{n-s}\,f(y)\,f(x) \qquad \text{for } y<x$$

Therefore,

$$E_{t_{(r)}t_{(s)}}\left(m(y)\,m(x)\,K_{h_1}(t-y)\,K_{h_1}(t-x)\right)$$

$$= \iint_{y<x} m(y)\,f(y)\,m(x)\,f(x)\,\frac{n!}{(r-1)!\,(s-r-1)!\,(n-s)!}\left(F(y)\right)^{r-1}\left(F(x)-F(y)\right)^{s-r-1}\left(1-F(x)\right)^{n-s}$$

$$K_{h_1}(t-y)\,K_{h_1}(t-x)\,dy\,dx$$

$$= \iint_{y<x} \frac{n!}{(r-1)!\,(s-r-1)!\,(n-s)!}\left(F(y)\right)^{r-1}\left(1-F(y)\right)\left(F(x)-F(y)\right)^{s-r-1}\left(1-F(x)\right)^{n-s+1}$$

$$h(y)\,h(x)\,K_{h_1}(t-y)\,K_{h_1}(t-x)\,dy\,dx$$

using

$$m(y)\,f(y) \;=\; h(y)\left(1-F_L(y)\right)\left(1-F_C(y)\right)$$

$$= h(y)\left(1-F(y)\right) \quad \text{as before}$$

Hence,

$$E\left(\hat{h}^2(t;z)\right) = \int_y \sum_{j=1}^{n} \frac{K_{h_2}^2\left(z-z_{(j)}\right)}{(n-j+1)^2} \frac{n!}{(j-1)!(n-j)!} h(y)\left(F(y)\right)^{j-1}\left(1-F(y)\right)^{n-j+1} K_{h_1}^2(t-y)\,dy$$

$$+ \; 2 \iint_{y<x} \sum\sum_{r<s} \frac{K_{h_2}\left(z-z_{(r)}\right) K_{h_2}\left(z-z_{(s)}\right)}{(n-r+1)(n-s+1)} \frac{n!}{(r-1)!(s-r-1)!(n-s)!}$$

$$\left(F(y)\right)^{r-1}\left(1-F(y)\right)\left(F(x)-F(y)\right)^{s-r-1}\left(1-F(x)\right)^{n-s+1}$$

$$h(y)\,h(x)\,K_{h_1}(t-y)\,K_{h_1}(t-x)\,dy\,dx$$

Finally, $\text{var}\left(\hat{h}(t;z)\right)$ is calculated as $\left[E\left(\hat{h}^2(t;z)\right)\right] - \left[E\left(\hat{h}(t;z)\right)\right]^2$

References

Aickin M, Ritenbaugh C, Surwit E, Meyskens F. *Comparative exposure ratios - A non-parametric, multifactor technique for case-control studies.* Statistics in Medicine, 1994, **13:** 245-260

Altman DG, Gore SM, Gardner MJ, Pocock SJ. *Statistical Guidelines for Contributors to Medical Journals.* Br. Med. J. 1983; **286:** 1489-1493.

Altman DG. *The scandal of poor medical research.* British Medical Journal 1994; **308:** 283-284.

Barlow RE, Bartholomew DJ, Bremner JM, Brunk HD. *Statistical inference under order restrictions.* London: John Wiley & Sons 1972.

Bartlett MS. *Statistical Estimation of Density Functions.* Sankhya Series A 1963; **25:** 245-254.

Breslow NE and Day NE. *Statistical Methods in Cancer Research. Vol 1 - The Analysis of Case-Control Studies.* Oxford: IARC Scientific Publications 1980.

Breslow NE and Day NE. *Statistical Methods in Cancer Research, Vol 2, The analysis of Cohort Studies.* WHO International Agency for Research in Cancer. Sci. Pub. No. 32 Lyon, France, 1980.

Cacoullus T. *Estimation of a Multivariate Density.* Ann. Inst. Statist. Math 1966; 18:179-181.

Campbell MJ, Machin D. *Medical statistics : A Commonsense approach.* Chichester. Wiley 1993

Carter WH, Wampler GL & Stablein DM. *Regression Analysis of Survival Data in Cancer Chemotherapy.* New York: Marcel Decker Inc. 1983

Clark RM. *A Calibration Curve for Radiocarbon Dates.* Antiquity 1975; **49**: 251-266.

Clark WH, Elder DE, Guerry D et al. *Model predicting survival in stage melanoma based on tumour progression.* J. Natl Cancer Institute 1989; **81**: 1893-1904.

Collett, D. *Modelling Binary Data.* London: Chapman and Hall, 1991.

Commenges D, Moreau T. *Comparative efficiency of a survival-based case-control design and a random selection cohort design.* Statistic in Medicine, 1991, **10**: 1775-1782.

Copas JB. *Plotting p against x.* Journal of Applied Statistics 1983; **32**, No 1: 25-31.

Cox DR, Oakes D. *Analysis of Survival Data.* London: Chapman and Hall, 1984.

Cox DR. *Analysis of Binary Data.* London: Methuen & Co. Ltd. 1970.

Cox DR. *Regression models and Life Tables.* Journal of the Royal Statistical Society Series B 1972; **34**: 187-220.

Cox DR, Snell EJ. *A general definition of residuals (with discussion).* J.R. Statistical Society, B 1968; **30**: 248-275.

Criqui MH., Langer RD, Froneck A et al. *Mortality Over a Period of 10 years in Patients with Peripheral Arterial Disease.* New England Journal of Medicine 1992; **326**: 381-386.

Deredita G, Serio G, Neri V, Polizzi RA, Barberio G, Losacco T. *A survival regression-analysis of prognostic factors in colorectal cancer.* Australian and New Zealand Journal of Surgery 1996; **66**: 445-451.

Doll R, Peto R, Hall E, Wheatley K, Gray R. *Mortality in Relation to Consumption of Alcohol - 13 years Observations on Male British Doctors.* British Medical Journal 1994; **309**: 911-918.

Egijou A, McHugh R. *Estimation of Relative Risk for matched pairs in Epidemiological research.* Biometrics 1977; **33**: 552-556.

Elandt-Johnson RC, Johnson NL. Survival Models and Data Analysis. New York: John Wiley and Sons, 1980.

Epanechnikov V. *Nonparametric Estimates of a Multivariate Probability Density.* Theory of Probability and its Applications 1969; **14**: 153-158.

Evans M. *Presentation of manuscripts for publication in the British Journal of Surgery.* Br. J. Surgery 1989; **76**: 1311-1315.

Everitt BS. *Statistical Methods for Medical Investigations.* New York: Oxford University Press, 1989: 83-98.

Everitt BS. *Statistical methods in medical investigations.* London. Edward Arnold 1994

Fan J, Gijbels I. *Local polynomial modelling and its applications.* London. Chapman and Hall 1996.

Fleming TR, Harrington DP. *Counting Processes and Survival Analysis.* New York. Wiley 1991.

Gasser T, Kneip A, Kohler W. *A Flexible and Fast Method for Automatic Smoothing.* Journal of the American Statistical Association 1991; **86**: 643-652.

Gasser, T and Muller, HG. *Kernel Estimation of Regression Functions.* In Lecture Notes in Mathematics 1979; **757**: 23-68; eds. T. Gasser and M. Rosenblatt. Springer-Verlag: New York.

Greenwood M. *The natural duration of Cancer.* Reports on Public Health and Medical Subjects 1926; **33**: 1-26.

Grove JS, Nomura A, Severson RK, Stemmermann GN. *The association of blood-pressure with cancer incidence in a prospective-study.* American Journal of Epidemiology 1991; **134**: 942-947.

Grundy SM, Vega GL. *Causes of High Blood Cholesterol.* Circulation 1990; **81**: 412-427.

Hardle W. *Applied Nonparametric Regression.* Cambridge. Cambridge University Press 1990.

Harper R, Ennis CV, Sheridan B, Atkinson AB, Johnston GD, Bell PM. *Effects of low dose versus conventional dose thiazide diuretic on insulin action in essential hypertension.* British Medical Journal 1994; **309**: 226-230.

Hollander M. *A distribution-free test for parallelism.* Journal of the American Statistical Association 1970; **65**: 387-394.

Holman CDJ, James IR, Gattey PH, Armstrong BK. *An analysis of trends in mortality from malignant melanoma of the skin in Australia.* Int J Cancer 1980; **26**: 703-709.

Hosmer DW and Lemeshow S. *Applied Logistic Regression.* New York: John Wiley & Sons 1989.

Hunter J. *Calculus.* Glasgow and London: Blackie and Chambers 1972.

Kalbfleisch JG. *Probability and Statistical Inference Volume 2: Statistical Inference.* New York: Springer-Verlag 1985.

Kaplan EL and Meier P. Nonparametric estimation from incomplete observations. *J. American Statistical Association* 1958; **53**: 457-481.

Karpf DB, Shapiro DR, Seeman E, Ensrud KE, Johnston CC, et al. *Prevention of nonvertebral fractures by alendronate - A meta-analysis.* Journal of the American Medical Association 1997; **277**: 1159-1164.

Keefe M and Mackie RM. *The Relationship between Risk of Death from Cutaneous Melanoma and Thickness of Primary Tumour: No evidence for steps in risk.* British J. of Cancer 1991; **64**: 598-602.

Kruskal WH. *A Nonparametric Test for the Several Sample Problem.* Annals of Mathematical Statistics 1952; **23**: 525-540.

Loftsgaarden DO, Quesenberry GP. *A Nonparametric estimate of a Multivariate Density Function.* Annals of Mathematical Statistics 1965; **36**: 1049-1051.

Mackie RM, Aitchison TC, Sirel JM et al. *Clinicopathological Predictors of Prognosis in Subsets of Melanoma Patients.* British J. of Cancer 1995; **71**: 173-176.

MacKie RM, Freudenberger T and Aitchison TC. *Personal risk-factor chart for cutaneous melanoma.* Lancet 1989:487-490.

MacKie RM, Smyth JF, Soutar DS, et al. *Malignant melanoma in Scotland 1979-1983.* Lancet 1985;ii 859-862.

Mann HB, Whitney DR. *On a Test of Whether One of Two Random variables is Statistically Larger than the other.* Journal of the American Statistical Association 1947; **47**: 583-621.

Mantel N. *Synthetic Retrospective Studies and Related Topics.* Biometrics 1973; **29**: 479-486.

Marron JS, Padgett WJ. *Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples.* Annals of Statistics 1987; **15**: 1520-1535.

McCullagh P & Nelder JA. *Generalized Linear Models.* London: Chapman and Hall 1990.

Mike V and Stanley KE. *Statistics in Medical Research.* New York: John Wiley & Sons 1982.

Muller HG, Wang JL. *Nonparametric analysis of changes in hazard rates for censored survival data: An alternative to change-point models.* Biometrika 1990; **77**: 305-314.

Muller HG, Wang JL. *Hazard rate estimation under random censoring with varying kernels and bandwidths.* Biometrics 1994; **50**: 61-76.

Murray GD. *Statistical Guidelines for the British Journal of Surgery.* Br. J. Surgery 1991a; **78**: 782-784.

Murray GD. *Statistical Aspects of Research Methodology.* Br. J. Surgery 1991b; **78**: 777-781.

Nadaraya EA. *On Estimating Regression.* Theory Probability Applications 1964; **9**: 141-142.

Neuhauser M, Becher H. *Improved odds ratio estimation by post hoc stratification of case-control data.* Statistics in Medicine, 1997, **16**: 993-1004.

Parzen E. *On Estimation of a Probability Density and Mode.* Annals of Mathematical Statistics 1962; **35**: 1065-1076.

Patil PN. *On the least squares cross-validation bandwidth in hazard rate estimation.* The annals of Statistics 1993; **21**: 1792-1810.

Pregibon D. *Logisitic Regression Diagnostics.* Annals of Statistics 1981; **9**: 705-724.

Priestley MB, Chao MT. *Nonparametric Function Fitting.* Journal of the Royal Statistical Society, Series B 1971; **34**: 385-392.

Raffi F, Aboulker JP, Michelet C, Reliquet V, Pelloux H, et al. *A prospective study of criteria for the diagnosis of toxoplasmic encephalitis in 186 AIDS patients.* AIDS 1997; **11**: 177-184.

Rice JA. *Bandwidth choice for nonparametric regression.* Annals of Statistics 1984, **12**: 1215-30.

Rigel DS, Friedman MD, Kopf AW et al. *Factors influencing survival in melanoma.* Dermatologic Clinics Vol 9, No 4 Oct 1991.

Ronan SG, Han MC, Das Gupta TK. *Histologic prognostic indicators in cutaneous melanoma.* Semin Oncol 1988; 15: 558-65.

Rosenblatt M. *Remarks on some nonparametric estimates of a density function.* Annals of Mathematical Statistics 1956; **27**: 642-669.

Sarda P, Vieu P. *Smoothing parameter selection in hazard estimation.* Statistics & Probability Letters 1991; **11**: 429-434.

Scheider WL, Hershey LA, Vena JE, Holmlund T, Marshall JR, Freudenheim JL. *Dietary antioxidants and other dietary factors in the etiology of Parkinson's disease.* Movement Disorders 1997; **12**: 190-196.

Schlesselman JJ. *Case-Control Studies.* New York: Oxford University Press 1982.

Schoenfeld D. *Partial residuals for the proportional hazards regression-model* Biometrika 1982; **69**: 239-241.

Schreiber MM, Bozzo PD, Moon TE. *Malignant melanoma in southern Arizona.* Arch Dermatol 1981; **117**: 6-11.

Seeff LB, Buskell-Bales Z, Wright EC et al. *Long-Term Mortality after Transfusion-Associated Non-A Non-B Hepatitis.* New England Journal of Medicine 1992; **327**: 1906-1911.

Shepherd J, Cobbe SM, Ford I, Isles G, Lorimer AR, MacFarlane PW et al. *Prevention of Coronary Heart Disease with Pravastatin in Men with Hypercholesterolemia.* New England J. of Medicine 1995; **333**: 1301-1307.

Shibata R. *An optimal selection of regression variables.* Biometrika, 1981; **68**: 45-54.

Silvermann BW. *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall 1986.

Simonoff JS. *Smoothing Methods in Statistics.* New York: Springer-Verlag 1996.

Soong S-J, Shaw HM, Balch JM et al. *Predicting Survival and Recurrence in Localized Melanoma: A Multivariate Approach.* World J. Surgery 1992; **16**: 191-195.

Spiegel MR. *Advanced Calculus.* London: McGraw-Hill 1974.

Stone CJ. *Consistent Nonparametric Regression.* Annals of Statistics 1977; **5**: 595-645.

Szymik B and Woosley JT. *Further Validation of the Prognostic Model for Stage1 Malignant Melanoma based on Tumour Progression.* J. Cutaneous Pathology 1993; **20**: 50-53.

Tanner M and Wong WH. *The estimation of the hazard function from randomly censored data by the kernel method.* Annals of Statistics 1983; **11**, 989-93

The International Non-Hodkin's Lymphoma Prognostic Factors Project. *A Predictive Model for Aggressive Non-Hodkin's Lymphoma.* New England J. of Medicine 1993; **329**: 987-994.

Therneau TM, Grambsch PM, Fleming TR. *Martingale-based residuals for survival models.* Biometrika 1990; **77**: 147-160.

Thiel H. *A Rank-invariant Method of Linear and Polynomial Regression Analysis.* Nederl. Akad. Wetensch Proc. Ser. A 1950; **53**: 386-392.

Tillman DM, Aitchison TC, Watt DC et al. *Stage 11 Melanoma in the West of Scotland 1976-1985: Prognostic Factors for Survival.* European J. of Cancer 1991; **27**: 870-876.

Tukey JW. *The Simplest Signed-Ranks Tests.* Princeton University Stat. Res. Group Memo. Report 17 1949; 149,150,160,173,174.

Wallis WA. *Use of Ranks in One-Criterion variance Analysis.* Journal American Statistical Association 1952; 47: 583-612.

watson GS. *Smooth Regression Analysis.* Sankhya Series A 1964; **26**: 359-372.

Watt DC, Aitchison TC, MacKie RM, Sirel JM. *Survival Analysis: The importance of censored observations.* Melanoma Research 1996; **5**: 379-385.

Wilcoxon F. *Individual Comparisons by Ranking Methods.* Biometrics Bulletin 1 1945; **150**: 80-83.

Yang S. *Linear Functions of Concomitants of Order Statistics with Application to Nonparametric Estimation of a Regression Function.* Journal of the American Statistical Association 1981; **76**: 658-662.

Yates F, Healey MJR. *How should we reform the teaching of statistics?* J. R. Statistical Society, Series A 1964; 127: 199-210.