

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

**A Thesis Submitted for the Degree of PhD at the University of Warwick**

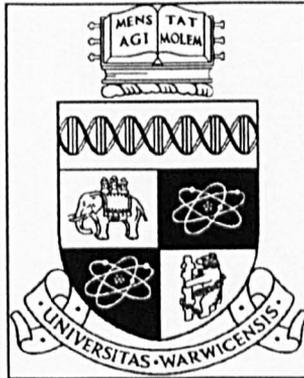
<http://go.warwick.ac.uk/wrap/74554>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

In the Name of the Almighty



**Theory and Applications of  
Delayed Censoring Models  
in Survival Analysis**

Fariborz Heydari

Thesis submitted for the degree of Doctor of Philosophy

University of Warwick

Department of Statistics

Coventry, England

March 1997

# **In the Name of the Almighty**

**To my wife Soghra and our daughter Soulmaz**

# Contents

<b>1</b>	<b>Introduction</b>	<b>10</b>
1.1	Survival Analysis . . . . .	10
1.2	Delayed Censoring . . . . .	14
1.3	Choice of Length of Follow-up . . . . .	15
1.4	Truncation . . . . .	16
1.5	Data on Reoffending . . . . .	17
1.6	Weighted Hazards Model . . . . .	19
<b>2</b>	<b>Preliminary Analysis of Survival Data</b>	<b>21</b>
2.1	Introduction . . . . .	21
2.2	Basic Survival Distributions . . . . .	22
2.3	Split Population Model (SPM) . . . . .	25
2.4	Model for Reoffending Prediction Score . . . . .	27
2.5	Model for Reconviction Prediction Score . . . . .	43
<b>3</b>	<b>Delayed Censoring Analysis (DCA)</b>	<b>55</b>
3.1	Statistical Analysis . . . . .	55

3.2	Diagnostics for Reoffence Time ( $X$ ) . . . . .	58
3.3	Split Population Model for DCA . . . . .	61
3.4	Diagnostics for Delay Time ( $Y$ ) . . . . .	65
3.5	Imputation of Delay Time . . . . .	65
3.6	Diagnostics With Imputed Data . . . . .	70
3.6.1	Goodness of Fit of the Distribution of $Y$ . . . . .	70
3.6.2	Goodness of Fit of the Distribution of $X$ . . . . .	72
3.7	Diagnostics for $X$ in Restricted Model . . . . .	75
3.8	Comparison of the Models by Hypothesis Testing . . . . .	78
3.9	Diagnostics for Reconviction Time ( $Z$ ) . . . . .	79
3.10	Standard Errors . . . . .	82
3.10.1	The Marginal Model . . . . .	82
3.10.2	The Full Model . . . . .	85
3.10.3	The Restricted Model . . . . .	92
3.11	Comparison of the Score Coefficients . . . . .	95
<b>4</b>	<b>Choice of Follow-Up and Fairer Assessment of Risk Scores</b>	<b>97</b>
4.1	Introduction . . . . .	97
4.2	Estimation of the Marginal Model . . . . .	99
4.3	Diagnostics for $X$ in Marginal Model . . . . .	100
4.4	The Observed and Expected Number of Reoffence . . . . .	102
4.5	Kaplan-Meier and Fitted Survival Plots . . . . .	105
4.6	Diagnostics for $X$ in Restricted Model . . . . .	111

4.7	Subset Kaplan-Meier and Fitted survival Plots . . . . .	113
4.7.1	Survival Plots for $X$ . . . . .	113
4.7.2	Survival Plots for $Z$ . . . . .	117
4.8	Split-Sample and Simulation Methods . . . . .	120
4.9	Standardized Estimates . . . . .	124
4.10	Correlation Matrix . . . . .	127
4.11	Probability of $X$ Within a Time Period $t$ . . . . .	127
<b>5</b>	<b>Independence of Delay and Reoffence Times</b>	<b>131</b>
5.1	Introduction . . . . .	131
5.2	Analysis of Truncated Data Without Covariates . . . . .	133
5.2.1	Parametric Analysis . . . . .	133
5.2.2	Semi-parametric Analysis . . . . .	145
5.2.3	Nonparametric Analysis . . . . .	155
<b>6</b>	<b>Nonparametric Analysis of Truncated Data with Covariates</b>	<b>166</b>
6.1	Introduction . . . . .	166
6.2	Backward Regression Model . . . . .	167
6.3	Partial Likelihood . . . . .	169
6.4	Estimating Distribution of Failure Time . . . . .	171
6.5	Analysis in Discrete Time . . . . .	173
6.6	Diagnostic Plots . . . . .	175
6.7	Application of the Regression Model . . . . .	178

<b>7</b>	<b>A Simple DCA Model with Correlated Offence and Delay Times</b>	<b>188</b>
7.1	Statistical Analysis . . . . .	188
7.2	Standard Errors . . . . .	191
<b>8</b>	<b>Delayed Censoring Modification to the Cox Model</b>	<b>198</b>
8.1	Introduction . . . . .	198
8.2	Statistical Theory for Weighted Hazards Model . . . . .	199
8.3	Estimation of the Weights . . . . .	203
8.4	Estimating the Regression Parameters . . . . .	205
8.5	Generalized Weighted Hazards Model . . . . .	209
8.6	Survival Curves by Risk Group . . . . .	212
<b>9</b>	<b>Summary and Conclusions</b>	<b>220</b>
9.1	Results with Ordinary Censored Survival Data . . . . .	220
9.2	Results with Delayed Censoring Analysis . . . . .	222
9.3	Results on Independence Analysis . . . . .	224
9.4	Results from Further Delayed Censoring Analysis (DCA) . . . . .	233
9.5	Further Applications . . . . .	238
	<b>Bibliography</b>	<b>240</b>

# List of Tables

1.1	Variables in Reoffending Data . . . . .	19
2.1	Basic Survival Functions . . . . .	27
2.2	Relative Importance of Covariates in Reoffence Models . . . . .	41
2.3	Relative Importance of Covariates in Reconviction Models . . . . .	53
3.1	Estimates of log-likelihood functions . . . . .	78
3.2	Standard errors for $p$ -part of full model . . . . .	90
3.3	Standard errors for $\lambda$ -part of full model . . . . .	91
3.4	Standard errors for $\theta$ -part of full model . . . . .	91
3.5	Standard errors under restricted model . . . . .	94
3.6	Score coefficients for actual and imputed data . . . . .	96
4.1	Parameter estimates of marginal model . . . . .	99
4.2	Expected no. of reoffence and reconviction, F/U=1 yr . . . . .	103
4.3	Expected no. of reoffence and reconviction, F/U=2 yr . . . . .	104
4.4	Expected no. of reoffence and reconviction, full F/U . . . . .	104
4.5	Committed, Convicted and Censored no. of reoffences . . . . .	108

4.6	Pattern of Diagnostics with Simulation . . . . .	123
4.7	Score coefficients for different follow up . . . . .	125
4.8	Standard errors of score coefficients . . . . .	125
4.9	Standardized estimates . . . . .	126
8.1	Parameter estimates (Cox Model) . . . . .	208
8.2	Parameter estimates (Weighted Model) . . . . .	212

## Acknowledgements

I am very grateful to my supervisor Professor John B. Copas for all his valuable comments, advice and unceasing encouragement during my studies at Warwick University.

I would also like to thank Professor J. Behboodian, Professor A. R. Soltani and Professor A. Parssian from Shiraz University (Iran), for their helpful and kind recommendations.

In addition, I want to express my gratitude to the Ministry of Culture and Higher Education of the Islamic Republic of Iran and the University of Shiraz (Iran), for their financial support.

Many thanks to the Staff of the Department of Statistics at Warwick University and to all my fellow research students. Their friendship and support have been remarkable. I am very grateful to them all, and to all my Iranian friends in England who have made these years very enjoyable.

I wish also to thank Mr. Ali Jahangiri, Dr. Reza Kamali (USA) and my brother Hossein for their encouragement and support.

Finally, I must appreciate Soghra and Soulmaz for their enthusiastic love, patience and understanding. Also many thanks to God for giving me this marvellous opportunity.

## Declaration

I hereby declare that this thesis is entirely the result of my own work during the past three years.

## Summary

The objective of this thesis is to develop new statistical models for the analysis of censored survival data, particularly for the study of *recidivism data*, such as the reoffence data used in the analysis here. This has been an area of great interest in criminology in recent years. There is a growing literature on survival analysis in criminology, where interest centres on the time from an offender's conviction, or release from prison, to the first reconviction or reimprisonment. In deciding whether to release a prisoner on parole, the Parole Board is provided with a statistical score which estimates the chance that the prisoner will reoffend within the period of time that he or she would otherwise be in prison. This score is based on a survival analysis of data on a sample of releases from long-term prison sentences. To capture most reoffences which occur within 2 years of release, follow-up must continue for at least 3 years to allow for the delay between offence and conviction. We reanalyse the data by using a model which explicitly allows for this delay. We refer to this as 'delayed censoring model'. The new analysis can be applied to data with a substantially shorter length of follow-up. This means that risk scores can be constructed from more up-to-date data and at less cost.

It is models of this kind that we shall be concerned with in this thesis, and this is the principal motivation of the work done. The statistical models that this thesis provides bring in a number of new ideas which are undoubtedly useful both at a theoretical level and in applications.

Other major divisions of the work include:

(i) Assessing the possibility of an association between the delay and reoffence times by studying truncated distributions fitted to these data, by parametric, semi-parametric and nonparametric models. With the nonparametric approach we have developed a 'backward regression model' which is similar to the Cox model.

(ii) We have also discussed delayed censoring modification to the Cox model, and developed a more general semi-parametric model for all the data including both observed and censored cases. In this model the delay and reoffence times are allowed to be correlated. We refer to this as the 'generalized weighted hazards model'.

(iii) Finally, we have compared the results by applying all these models to the data. Although the parametric models give a good fit to the data, the semi-parametric and nonparametric models give a slightly better fit, as expected.

# Chapter 1

## Introduction

### 1.1 Survival Analysis

Survival methods, originally developed by actuaries for studying the distribution of length of life, are widely used in medical and engineering applications of statistics. The data measure the times from a defined starting point to the occurrence of some terminal event: in a typical medical application it may be the time from onset of a chronic disease to death, in engineering it may be the time from the installation of a piece of equipment to its first failure. The power of survival analysis is also being increasingly recognised in the social sciences. There is a growing literature on survival analysis in criminology, where interest centres on the time from an offender's conviction, or release from prison, to the first reconviction or reimprisonment. Among other papers, parametric survival models are considered in Stollmark and Harris (1974), Harris and Moitra (1978) and Maltz (1984). The Cox proportional hazards model (Cox, 1972)

is discussed in Barton and Turnbull (1978, 1981) and Allison (1984). A good review and discussion is in Schmidt and Witte (1988). There is also a very readable chapter on survival methods in the recent book by Tarling (1993), chapter 6.

The importance of survival analysis for risk assessment has been brought into focus by the 1991 Criminal Justice Act in England and Wales. This legislation has made major changes to the administration of parole, the arrangement by which prisoners can be granted early release from prison. Under the Act, all prisoners with a prison sentence of 4 years or more are automatically released two thirds of the way through their sentence, but are eligible for discretionary release on parole half way through their sentence. The Act is explicit about the way that this parole decision should be made. The primary consideration is that of risk, the risk that the prisoner will reoffend during the time he or she would otherwise be in prison. This assessment of risk is to be systematic, taking into account the prisoner's criminal background as well as other factors which may be indicative of reoffending. The essentially statistical nature of this task was recognised by the Carlisle Committee (1988), whose report formed the basis of the sections of the Act which deal with parole. Their advice was that the Parole Board, the body which is entrusted with parole decisions, should use a statistical risk score to help them measure the likelihood of reoffending.

The Parole Board has in fact used a statistical risk score for many years, the Reconviction Prediction Score developed by Nuttall (1977) and reassessed

by Ward (1987). This score estimates the probability of a *reconviction* within a fixed period of 2 years from release from prison. A literal interpretation of the Act means that we now need the probability of a *reoffence* occurring within a time period  $t$ ,  $t$  being the length of potential parole, which is not fixed but varies from one prisoner to another, up to a maximum of around 2 years. This probability is just 1 minus the survival curve

$$S_v(x) = P(\text{no offence within time } x|v), \quad (1.1)$$

where the probability is conditional on  $v$ , a vector of covariates or risk factors.

Of course the true incidence of offending can not be measured, as a reoffence can only be attributed to a particular prisoner if he or she is subsequently convicted for that offence. Inevitably, conviction has to be used as a proxy for offending. Thus in estimating equation (1.1) we interpret 'offence' to mean 'offence which leads to a conviction'. With this proviso, Copas *et al.* (1996) reported the estimation of equation (1.1) using a survival analysis of data on a large sample of prisoners recently released from prison. The analysis led to the formulation of the new risk score now being used by the Parole Board.

Section 1.5 presents a brief description of the Home Office study about the reoffence data used in the analysis, reported in Copas *et al.* (1996). Chapter 2 presents the basic concepts, such as survival curve, split population model and censoring, in survival analysis. Diagnostic plots which play a key role in the development of model selection methodology are being used throughout in our statistical analysis. These include partly some residual plots and mainly

some order-based diagnostics, depending on the conditional distributions of relevant quantities. Essentially the order-based diagnostics are just quantile-quantile plots. Of course, these are not the only plots which could give useful diagnostic information about the models being considered, and although intuitively appealing they need to be used with caution. Barlow and Prentice (1988) comment on the difficulties of testing proportional hazards assumptions using Cox-Snell residuals, which are somewhat akin to those suggested here. A major problem is that misspecification of the models can be masked when only the marginal distributions of relevant quantities are examined, and a more sensitive approach would be to repeat these plots using subsets of the data subdivided by values of the covariate vector  $v$ . In practice, however, the scope for dividing the sample into any more than a few subsets is limited by the availability of data.

A more honest test of predictive fit of a survival model, overcoming the problem of overfitting, is to estimate the parameters from a randomly chosen half of the data (the training sample), and then to validate the survival curves on the remaining cases (the validation sample)—this is the so called split-sample method. In Chapter 4 we have applied this method to the criminological data and the results are satisfactory.

In view of the comments in Barlow and Prentice (1988) referred to previously, we have carried out some simulation experiments to check the sensitivity of these diagnostic plots to model misspecification. In particular we have studied the behaviour of these plots when the exponential distributions are

replaced by Weibull distributions with a range of different shape parameters. We find that the plots are noticeably non-linear for simulated data with the same sample size and with the same set of covariates  $v$  but with shape parameters differing from the exponential value of 1 by more than about 0.2. This analysis is also discussed in Chapter 4.

## 1.2 Delayed Censoring

Although an offence can only be captured in the data if there is a conviction, it is important to distinguish between the date of offence and the date of the conviction, since there is often a substantial delay between the two. Even if there is no recorded reoffence for a particular prisoner, there may in fact be a reoffence which is awaiting court proceedings and which would be captured in the data had the period of follow-up been long enough to cover the date of the trial. Thus if times of reoffence are modelled as ordinary censored survival data there will be substantial under-reporting towards the end of follow-up, and hence underestimation of the rate of decrease of the survival curve. We refer to this as the problem of ‘delayed censoring’. Many of the references cited earlier avoid the problem altogether by modelling the time of reconviction rather than reoffence, but modelling offence times has some important advantages. It is clearly more relevant to the incidence of crime. It avoids the problem of ‘pseudoconvictions’, early convictions which in fact relate to offending committed before the start of the study—this can be a

major problem, for example when rates of reoffending after different types of short non-custodial sentences are being compared. And thirdly, as we shall argue, the resulting models can be simpler and easier to interpret.

The problem of delayed censoring had not previously been solved in the explicit form which has been done in this thesis, and I think such a useful treatment provided here will hopefully be helpful and fill an important gap in the literature. The statistical analysis necessary for this new approach in survival analysis is discussed at length in Chapter 3.

### 1.3 Choice of Length of Follow-up

As estimation of model (1.1) is required for  $x$  up to a maximum of 2 years, the sample used in Copas *et al.* (1996) was followed up for a sufficiently long period of time to ensure that most of the actual reoffences in the first 2 years had resulted in a conviction by the end of the study. The sample was of prisoners released from prison in 1987, and follow-up continued to the end of 1990. But reliable longitudinal data is difficult and expensive to collect and, more importantly, the longer the follow-up the earlier the sample has to be and so the less relevant is the study to current judicial and social conditions. Thus, there is a strong incentive to use the most recent data possible, and hence the incentive to use as short a length of follow-up as possible. We suggest that, by making proper allowance for delayed censoring in the statistical method, it is possible to fit a survival model to reoffence data (by maximum likelihood

fitting) with a substantially shorter length of follow-up. Chapter 4 contains an extensive mathematical and applied statistical discussion of the proposed model.

## 1.4 Truncation

Truncation induced by censoring is sometimes used in survival analysis. Usually there are two kinds of truncation, left-truncation and right-truncation. Left-truncation in survival analysis means that an individual is included in the study sample if his lifetime (failure time) is larger than some value (censoring point), whereas by right-truncation, we mean that the individual is included only when his lifetime is smaller than some value (censoring point).

Now in the criminological data, let  $X$  be the time from release to first reoffence and  $Y$  be the time from this reoffence to the corresponding first reconviction which occurs a time  $Z$  after release, that is,  $Y = Z - X$  (the delay between  $Z$  and  $X$ ). In the delayed censoring analysis model we are assuming that  $X$  and  $Y$  are independent which is a crucial assumption and needs to be justified. However, due to the censoring,  $X$  and  $Y$  are truncated from the right at  $T - Y$  and  $T - X$  respectively,  $T$  being the time to follow-up. In order to assess the possibility of a relationship between  $X$  and  $Y$  we study truncated distributions fitted to these data. In Chapter 5, we provide a rather detailed discussion of the specification of the independence between delay and reoffence times, through parametric, semi-parametric and nonparametric analysis of

truncated data (induced by the censoring), firstly in the simple case excluding covariates. Chapter 6 is an exposition of nonparametric analysis of truncated data including covariates, and we develop a ‘backward regression model’, which is based on proportional hazards assumptions similar to the Cox proportional hazards model (Cox, 1972). However, in Chapter 7, we go on to consider a general model in which  $X$  and  $Y$  are allowed to be correlated. This model extends to all the data including both observed and censored cases.

## 1.5 Data on Reoffending

The data used in developing the new Parole Board prediction score, and re-analysed in this thesis, consisted of a sample of 1179 male prisoners released from prison in 1987 and with sentences of 4 years or more. The methods of data collection, and details of which variables were measured, are discussed at length in Copas *et al.* (1996). Briefly, a large number of covariates covering social, demographical and criminal history factors were recorded for each subject, and the follow-up data consisted of the time to first reconviction, if any, and the time of the principal offence leading to that conviction. Trivial offences such as parking or minor motoring offences were not included. A parallel analysis has also been done for the time to the first serious offence, defined for this purpose as an offence leading to reimprisonment, but for brevity only the ‘all-reoffences’ is discussed here.

A preliminary analysis of the data shows that several of the covariates are

associated with reoffending, but there is a strong practical incentive to keep the model simple, and only to use covariates which would be readily available to the Parole Board and which would be unambiguously measured in the course of routine administration. In the event only a small number of criminal history variables were used—the analysis here is based on age, total number of convictions, number of juvenile custodial sentences and the number of previous adult custodial sentences. Once these basic variables are included very little is gained by adding further covariates. The previous parole score (Nuttall, 1977) used 17 variables, but with little or no improvement in prediction. The covariates are correlated with each other, and so many different subsets of them will provide an equally good fit to the data. Great care is therefore needed in assigning any causative role to the covariates chosen. Table 1.1 represents the variables included in the data. It is worth emphasizing that the object of the analysis is description of the past data as an aid to prediction, and not explanation of why some offenders reoffend and some do not. For more background to data of this kind, see the comprehensive picture of crime and justice in England and Wales in the recent book by Gordon C Barclay *et al.* (1995).

**Remark:**

- (i) The variables **tfo**, **tfc** and **tconv** are coded 9999 if there was no conviction by the end of follow-up.
- (ii) In the main analysis the covariates **age**, **ac**, **pre** and **jc** are used for prediction scores.

Table 1.1: Variables in Reoffending Data

variable	name	meaning
1	ID	identification number
2	age	age (years)
3	ac	number of adult custodial sentences
4	pre	number of previous convictions
5	jc	number of juvenile custodial sentences
6	tfo	time from release to first reoffence (days)
7	tfc	time from release to first reoffence leading to a custodial sentence (days)
8	of	offence code (1 to 7)
9	tconv	time from release to first reconviction (days)
10	tal	time from release to end of follow-up (days)

(iii) In the data there are 486 uncensored and 693 censored observations.

## 1.6 Weighted Hazards Model

In Chapters 5 and 6 we provide an overview of both the theory and applications of some truncated models, models in which the data are limited and can be observed only in certain ranges because of some stochastic mechanism such as censoring. In Chapter 8 we discuss delayed censoring modification to the Cox

proportional hazards regression model and develop a semi-parametric model for all the data including both observed and censored observations, firstly for the case that the delay and reoffence times are assumed to independent. In this model the hazard function of observed failure time is expressed in terms of the hazard function of actual failure time multiplied by a weight function, which can be estimated from one of the parametric models developed in this thesis. We refer to this model as the 'weighted hazards model'. This model is then extended to a more general case in which the delay and reoffence times are allowed to be correlated. We refer to this as the 'generalized weighted hazards model'.

Finally, in Chapter 9 the results obtained in the thesis are discussed.

# Chapter 2

## Preliminary Analysis of Survival Data

### 2.1 Introduction

One important class of models in scientific applications is survival analysis for failure data. This chapter introduces some basic notions in survival analysis, such as survival curves, censoring and split population model. Also, some statistical models for ordinary censored survival data are discussed, and some graphical methods of model checking are derived. Finally, a measure for comparing the coefficients of the covariates in different models is also defined.

## 2.2 Basic Survival Distributions

Let  $X$  be a nonnegative random variable with density function  $f_X$  and probability distribution function  $F_X$  then we consider the following distributions as a part of our statistical models.

### Exponential Distribution

**Definition:** If the random variable  $X$  has a density function given by

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0, \lambda > 0, \quad (2.1)$$

then  $X$  is defined to have an exponential distribution with parameter  $\lambda$ .

#### Properties:

- (i)  $F_X(x) = P(X \leq x) = 1 - e^{-\lambda x}$ ,  $P(X > x) = e^{-\lambda x}$
- (ii)  $E(X) = 1/\lambda$  and  $\text{Var}(X) = 1/\lambda^2$
- (iii) If  $U = \lambda X$ , then  $U$  has exponential distribution with parameter 1 and  $P(U > u) = e^{-u}$ .

### Weibull Distribution

**Definition:** The random variable  $X$  is said to have a Weibull distribution with parameters  $b$  and  $\lambda$  if the density function of  $X$  is given by

$$f_X(x) = \lambda b x^{b-1} e^{-\lambda x^b}, \quad x > 0, \lambda > 0, b > 0. \quad (2.2)$$

The parameters  $b$  and  $\lambda$  are referred as the shape and scale parameter respectively. The Weibull model is a generalization of the exponential model,

because for  $b = 1$  the Weibull distribution reduces to the exponential distribution.

**Properties:**

(i)  $F_X(x) = P(X \leq x) = 1 - e^{-\lambda x^b}$ ,  $P(X > x) = e^{-\lambda x^b}$

(ii)  $E(X) = (1/\lambda)^{(1/b)}\Gamma(1 + b^{-1})$ , where

$$\Gamma(t) = \int_0^{+\infty} x^{t-1} e^{-x} dx, \quad t > 0$$

is the gamma (or complete gamma) function.

(iii) If  $U = \lambda X^b$ , then  $P(U > u) = e^{-u}$ .

**Gamma Distribution**

**Definition:** If the random variable  $X$  has a density function given by

$$f_X(x) = \frac{\lambda^b}{\Gamma(b)} x^{b-1} e^{-\lambda x}, \quad x \geq 0, b > 0, \lambda > 0, \quad (2.3)$$

then  $X$  is defined to have a gamma distribution with parameters  $b$  and  $\lambda$ . The parameters  $b$  and  $\lambda$  are called the shape and scale parameter respectively. The gamma model is another generalization of the exponential model, since for  $b = 1$  the gamma density reduces to the exponential density.

**Properties:**

(i)  $F_X(x) = P(X \leq x) = \int_0^x f(t) dt$ , or

$$F_X(x) = \frac{\int_0^{\lambda x} v^{b-1} e^{-v} dv}{\Gamma(b)} = \Gamma^*(\lambda x, b), \text{ say.}$$

(ii)  $P(X > x) = 1 - \Gamma^*(\lambda x, b)$

(iii)  $E(X) = b/\lambda$ ,  $\text{Var}(X) = b/\lambda^2$

(iv) If  $U = \lambda X$ , then  $P(U > u) = 1 - \Gamma^*(u, b)$ .

## Censoring

One aspect which distinguishes survival analysis from other field of statistics is censoring. Vaguely speaking, a censored observation contains only partial information about the random variable of interest. More precisely, as explained in Chapter one, in the classical survival analysis a collection of individuals are observed from some entry time until a particular event (such as death) happens. Often it is impossible to wait for the event to happen for all individuals; for some, it is only known that the event had not yet happened at some specific time and in this case the observation of the time to the occurrence of the event is censored (right-censored), in other words, the time at which an observation ceases is called censoring time of this observation. The number of censored observations may be a high proportion of the total observations in the study sample. For our criminological data, the number of censored cases is nearly 59% of the total. This indicates the fact that in practical applications censoring cannot be ignored.

The following type of censoring is frequently used.

### Random Censoring

Let  $X_1, \dots, X_N$  be independent identically distributed random variables (failure times) and  $T_1, \dots, T_N$  be the corresponding censoring times,  $X_i$  the  $i$ th failure time and  $T_i$  the censoring time associated with  $X_i$ . Here, each  $T_i$  is a random variable with some distribution. Instead of observing  $X_i$ 's, the ran-

dom variables of interest, we can only observe  $(W_i, \delta_i), i = 1, \dots, N$ , where  $W_i = \min(X_i, T_i)$  and

$$\delta_i = \begin{cases} 1 & \text{if } X_i \leq T_i, \text{ that is } X_i \text{ is not censored} \\ 0 & \text{if } X_i > T_i, \text{ that is } X_i \text{ is censored.} \end{cases}$$

Note that  $W_i$ 's are independent identically distributed and the variable  $\delta_i$  is known as a *censoring indicator*. In the special case if  $T_1 = T_2 = \dots = T_N = C$ , a fixed time, then we have fixed censoring. In this case  $C$  is determined at the start of the study.

### Survival Curve for Reoffence time

Let  $X$  be the time from release to first reoffence with probability density function  $f_X$ , probability distribution function  $F_X$  and reoffending rate  $\lambda$ . The survival curve  $S^o(x)$  is defined to be

$$S^o(x) = P(\text{first reoffence occurs after time } x \text{ from release}).$$

The value of  $S^o(x) = P(X > x)$  leads to the chance of reoffending within any given time  $x$ , since

$$P(\text{reoffence within time } x) = P(X \leq x) = 1 - S^o(x)$$

## 2.3 Split Population Model (SPM)

Split population model assumes that from a population of subjects, only a proportion  $p$  will ever reoffend and a proportion  $(1 - p)$  will never reoffend

at all. The survival curve for the actual reoffence times, including the cases where there may be a reoffence which is not recorded in the data because the subsequent reconviction occurs after the follow-up period is given by the basic split population model

$$S(x) = 1 - p + p S^o(x), \quad (2.4)$$

where  $S^o(x)$  is the survival curve for those who will eventually reoffend. Note that for  $p = 1$  the SPM survival curve reduces to the conventional survival curve

$$S^o(x) = P(X > x) = 1 - F_X(x)$$

An unattractive feature of exponential, Weibull, gamma and other conventional survival models is that  $S^o(x)$  tends to zero as  $x$  increases, implying that every subject is bound to reoffend at some time in the future. However, under the split population model the survival curve  $S(x)$  tends to  $(1 - p)$  rather than zero. This is the reason for considering the split population model—an approach advocated by several previous researchers in this area (Maltz (1984), Schmidt and Witte (1984, 1988), Tarling (1993)). Table 2.1 represents the survival functions for the basic survival distributions.

Several papers on survival modelling of medical data have studied mixture models similar to equation (2.4). Mallor and Zhou (1992, 1996) allowed for ‘long term survivors’ in the population, by assuming that there is a proportion  $(1 - p)$  in the population who will never succumb to the disease being studied. Farewell (1982) discussed parameter estimation in a Weibull generalization

Table 2.1: Basic Survival Functions

survival curve	Exponential	Weibull	Gamma
$S^o(x)$	$e^{-\lambda x}$	$e^{-\lambda x^b}$	$1 - \Gamma^*(\lambda x, b)$
$S(x)$	$1 - p + pe^{-\lambda x}$	$1 - p + pe^{-\lambda x^b}$	$1 - p \Gamma^*(\lambda x, b)$

of model (2.4). Further references to related work are also listed in Farewell (1982). By making proper allowance for delayed censoring in the statistical method, we fit the mixture survival model (2.4) to reoffence data. This is discussed at length in Chapter 4.

## 2.4 Model for Reoffending Prediction Score

Initial exploration of data and full analysis with delayed censoring is complicated and initially we examine the data as if these were ordinary censored survival data. We will explore each of the above models with and without split population model. Later chapters will give a full analysis making proper allowance for delayed censoring. Note that we examined exponential distribution without split population model, but diagnostics showed that this is not a suitable choice for the reoffence times.

## Exponential Distribution With SPM

We begin by assuming that the time from release to first reoffence  $X$ , with the rate of reoffending  $\lambda$ , has an exponential distribution given by equation (2.1).

Let  $X = (X_1, \dots, X_N)$ ,  $\lambda = (\lambda_1, \dots, \lambda_N)$ , where  $X_1, \dots, X_N$  are independent exponential random variables and  $X_i$  is the time from release to first reoffence for the  $i$ th subject with parameter  $\lambda_i$  and density function  $f_{X_i}$ , and  $N$  is the total sample size. Let  $T = (T_1, \dots, T_N)$  with  $T_i, i = 1, \dots, N$ , being the follow-up period for the  $i$ th subject. Since  $X_i$ 's are censored if no event happened by end of follow-up period  $T_i$ , instead of observing  $X_1, \dots, X_N$  (the random variables of interest) we can only observe  $W_1, \dots, W_N$  where

$$W_i = \begin{cases} X_i & \text{if } X_i \leq T_i, \text{ that is, } X_i \text{ is not censored} \\ T_i & \text{if } X_i > T_i, \text{ that is, } X_i \text{ is censored} \end{cases}$$

for  $i = 1, \dots, N$ . Using the split population model given by  $p$  and  $\lambda$ , the likelihood function of the full sample is given by

$$L = \prod_{i=1}^n p f_{X_i}(x_i) \prod_{i=n+1}^N S(T_i)$$

or

$$L = \prod_{i=1}^n (p \lambda_i e^{-\lambda_i x_i}) \prod_{i=n+1}^N (1 - p + p e^{-\lambda_i T_i})$$

where  $\prod_{i=1}^n$  and  $\prod_{i=n+1}^N$  are the products over the uncensored and censored observations respectively. The log-likelihood function, used to obtain the maximum likelihood estimates of the model parameters is therefore given by

$$\ell = \sum_{i=1}^n (\log p + \log \lambda_i - \lambda_i x_i) + \sum_{i=n+1}^N \log(1 - p + p e^{-\lambda_i T_i}).$$

We are assuming that if an observation  $X_i$  is censored with  $T_i$ , then the contribution to the likelihood is  $P(X_i > T_i) = S(T_i)$ . This is not strictly correct, as an observation is censored if its time of conviction exceeds  $T_i$ . A more careful model, allowing for the delay between offence and conviction is considered later. This delay is here being ignored, that is the times of reoffence are assumed to be modelled as ordinary censored survival data. If the  $i$ th observation in the study sample has the vector of covariates

$$v_i = (1, age_i, ac_i, pre_i, jc_i)^T \quad (2.5)$$

then the natural parameterizations for  $\lambda_i$  and  $p$  are

$$\text{Model (1)} : \begin{cases} \log \lambda_i = a^T v_i \\ \log \frac{p}{1-p} = A. \end{cases}$$

This assumes that  $p$  is constant; we will later study a more general model in which  $p$  also depends on the covariates.

Our approach is to fit this model by maximum likelihood and then to examine the fit by appropriate diagnostic plot. Using model (1), the log-likelihood function is given by

$$\ell = \ell_1 + \ell_2$$

where

$$\ell_1 = \sum_{i=1}^n \{A - \log(1 + e^A) + a^T v_i - x_i e^{a^T v_i}\}$$

$$\ell_2 = \sum_{i=n+1}^N \log \left\{ \frac{1 + e^{(A - T_i) e^{a^T v_i}}}{1 + e^A} \right\}.$$

The maximum likelihood estimates of  $A$  and  $a^T = (a_1, a_2, a_3, a_4, a_5)$  can be found by numerical maximization of the function  $\ell$ . But maximizing  $\ell$  is

equivalent to minimizing  $(-\ell)$ , which is a non-linear function of the parameters. So a non-linear minimization method is required. This can be done by using `nlinb` function which is available in the statistical computer package **S-PLUS** for Windows (StatSci, 1992). After getting the estimates of the parameters we can check the goodness of fit of the model as follows.

### Goodness of fit of Model (1)

To check the validity of the model we use the following basic method of plotting. Let  $U = \lambda X$ , where  $X$  has exponential distribution with mean  $1/\lambda$ .

Under the split population model given by  $p$  and  $\lambda$  we have

$$P(X > x) = 1 - p + p e^{-\lambda x}$$

$$P(U > u) = 1 - p + p e^{-u}.$$

Therefore

$$\log\left\{\frac{P(U > u) - (1 - p)}{p}\right\} = -u.$$

We now order  $u_1, \dots, u_n$ , the sample values of  $U$ , into the order statistics  $u_{(1)} \leq u_{(2)} \leq \dots \leq u_{(n)}$ .  $P(U > u)$  is estimated from the sample values  $u_1, \dots, u_n$  and the censoring times  $\lambda_i T_i$  by the Kaplan-Meier survival curve, which is easily done in **S-PLUS** by using the function `surv.fit`. Denoting this estimated survival curve by  $\hat{P}(U > u)$ , and replacing the parameter  $p$  by its maximum likelihood estimate  $\hat{p}$ , we have

$$\log\left\{\frac{\hat{P}(U > u_{(i)}) - (1 - \hat{p})}{\hat{p}}\right\} \approx -u_{(i)}.$$

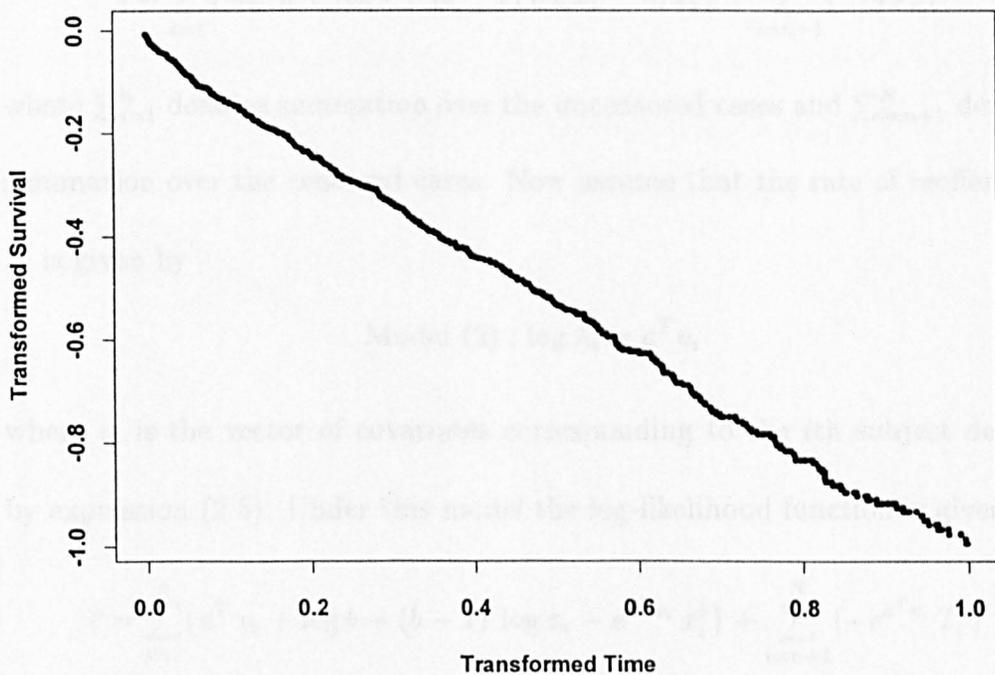
Therefore, we can plot

$$\log\left\{\frac{\hat{P}(U > u_{(i)}) - (1 - \hat{p})}{\hat{p}}\right\} \text{ against } u_{(i)}.$$

If the model holds the plot should resemble a straight line. The plot is depicted in Figure 2.1 which is close to a straight line, showing the validity of the model.

The estimated value of  $p$  is  $\hat{p}=0.80$ .

Fig. 2.1 Transformed Survival Plot for SPM (X:Exponential)



### Weibull Distribution

Consider those values of reoffence times which are not zero and denote these nonzero values by  $X$ . Now assume that the time from release to first reoffence  $X$  has Weibull distribution with density given by equation (2.2). The

likelihood function is now given by

$$L = \prod_{i=1}^n f_X(x_i) \prod_{i=n+1}^N S(T_i)$$

or

$$L = \prod_{i=1}^n (\lambda_i b x_i^{b-1} e^{-\lambda_i x_i^b}) \prod_{i=n+1}^N e^{-\lambda_i T_i^b}$$

and the log-likelihood function is given by

$$\ell = \sum_{i=1}^n \{\log \lambda_i + \log b + (b-1) \log x_i - \lambda_i x_i^b\} + \sum_{i=n+1}^N (-\lambda_i T_i^b).$$

where  $\sum_{i=1}^n$  denotes summation over the uncensored cases and  $\sum_{i=n+1}^N$  denotes summation over the censored cases. Now assume that the rate of reoffending  $\lambda_i$  is given by

$$\text{Model (2) : } \log \lambda_i = a^T v_i$$

where  $v_i$  is the vector of covariates corresponding to the  $i$ th subject defined by expression (2.5). Under this model the log-likelihood function is given by

$$\ell = \sum_{i=1}^n \{a^T v_i + \log b + (b-1) \log x_i - e^{a^T v_i} x_i^b\} + \sum_{i=n+1}^N (-e^{a^T v_i} T_i^b)$$

The estimates of the parameters can be obtained as before and to check the goodness of fit the model we proceed as follows.

### Goodness of fit of Model (2)

Let  $U = \lambda X^b$ . Then

$$P(X > x) = e^{-\lambda x^b}$$

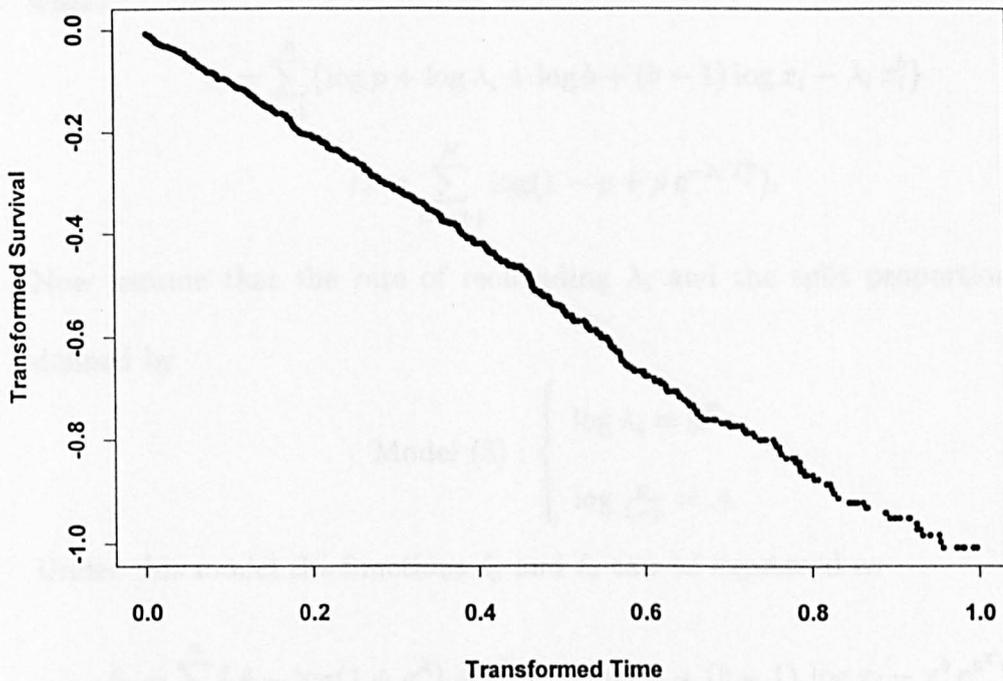
$$P(U > u) = e^{-u},$$

therefore,

$$\log\{P(U > u)\} = -u.$$

Now order  $u_1, \dots, u_n$  as  $u_{(1)} \leq u_{(2)} \leq \dots \leq u_{(n)}$ . Here,  $P(U > u)$  is estimated from the sample values  $u_1, \dots, u_n$  and the censoring times  $\lambda_i T_i^b$  by the Kaplan Meier survival curve. Denoting this estimated curve by  $\hat{P}(U > u)$ , we now need to plot  $\log\{\hat{P}(U > u_{(i)})\}$  against  $u_{(i)}$  for checking the goodness of fit of the model. This gives Figure 2.2 which is fairly close to a straight line, indicating the validity of the model. The estimated value of  $b$  is  $\hat{b} = 0.755$ .

Fig. 2.2 Transformed Survival Plot (X:Weibull)



## Weibull Distribution With SPM

Suppose that the reoffence time  $X$  has density function defined by equation (2.2). Then using the split population model given by  $p$  and  $\lambda$ , the likelihood function is given by

$$L = \prod_{i=1}^n \{p \lambda_i b x_i^{b-1} e^{-\lambda_i x_i^b}\} \prod_{i=n+1}^N \{1 - p + p e^{-\lambda_i T_i^b}\}$$

and the log-likelihood function is given by

$$\ell = \ell_3 + \ell_4$$

where

$$\ell_3 = \sum_{i=1}^n \{\log p + \log \lambda_i + \log b + (b-1) \log x_i - \lambda_i x_i^b\}$$

$$\ell_4 = \sum_{i=n+1}^N \log(1 - p + p e^{-\lambda_i T_i^b}).$$

Now assume that the rate of reoffending  $\lambda_i$  and the split proportion  $p$  are defined by

$$\text{Model (3)} : \begin{cases} \log \lambda_i = a^T v_i \\ \log \frac{p}{1-p} = A. \end{cases}$$

Under this model the functions  $\ell_3$  and  $\ell_4$  can be expressed as

$$\ell_3 = \sum_{i=1}^n \{A - \log(1 + e^A) + a^T v_i + \log(b) + (b-1) \log x_i - x_i^b e^{a^T v_i}\}$$

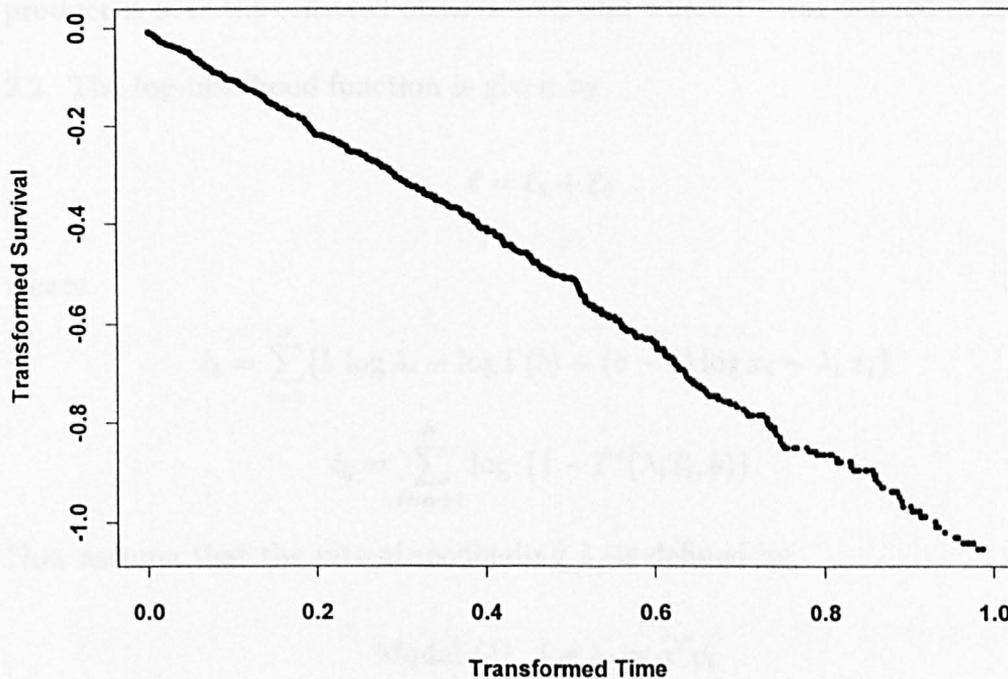
$$\ell_4 = \sum_{i=n+1}^N \log\left\{\frac{1 + e^{(A - T_i^b e^{a^T v_i})}}{1 + e^A}\right\}.$$

After finding the estimates of the parameters we can assess the model by the following diagnostic procedure.

### Goodness of fit of Model (3)

By using the transformation  $U = \lambda X^b$  and the split population model it can be shown that in order to check the goodness of fit of the model we need to plot  $\log\{(\hat{P}(U > u_{(i)}) - (1 - p))/p\}$  against  $u_{(i)}$  for  $i = 1, \dots, n$ . This gives Figure 2.3 which is fairly close to a straight line, showing the validity of the model. Here the estimate of  $b$  is  $\hat{b} = 0.7811$  and the estimate of  $p$  is  $\hat{p} = 0.930$ . Note that the estimated value of  $p$  is reasonably close to 1; we would expect this as model (2) showed that the Weibull model was adequate.

Fig. 2.3 Transformed Survival Plot for SPM (X:Weibull)



## Gamma Distribution

Suppose that the time from release to first reoffence  $X$  ( $X > 0$ ) has gamma distribution with density function given by equation (2.3). The likelihood function is now given by

$$L = \prod_{i=1}^n f_{X_i}(x_i) \prod_{i=n+1}^N S(T_i)$$

or

$$L = \prod_{i=1}^n \left\{ \frac{\lambda_i^b}{\Gamma(b)} x_i^{b-1} e^{-\lambda_i x_i} \right\} \prod_{i=n+1}^N \{1 - \Gamma^*(\lambda_i T_i, b)\}$$

where the first product is over the uncensored observations and the second product is over the censored observations and where  $\Gamma^*$  was defined in section 2.2. The log-likelihood function is given by

$$\ell = \ell_5 + \ell_6$$

where

$$\ell_5 = \sum_{i=1}^n \{b \log \lambda_i - \log \Gamma(b) + (b-1) \log x_i - \lambda_i x_i\}$$

$$\ell_6 = \sum_{i=n+1}^N \log \{1 - \Gamma^*(\lambda_i T_i, b)\}.$$

Now assume that the rate of reoffending  $\lambda_i$  is defined by

$$\text{Model (4) : } \log \lambda_i = a^T v_i$$

where  $v_i, i = 1, \dots, N$ , denotes the vector of covariates specific to the  $i$ th subject in the study sample and is defined by expression (2.5). Under this model the functions  $\ell_5$  and  $\ell_6$  can be expressed as

$$\ell_5 = \sum_{i=1}^n \{b a^T v_i - \log \Gamma(b) + (b-1) \log x_i - e^{a^T v_i} x_i\}$$

$$\ell_6 = \sum_{i=n+1}^N \log\{1 - \Gamma^*(e^{aT} v_i T_i, b)\}.$$

After finding the maximum likelihood estimates of the components of  $a$  and  $b$  the goodness of fit of the model can be verified as follows.

### Goodness of fit of Model (4)

To check the validity of the model, let  $U = \lambda X$ . This gives

$$P(U > u) = 1 - \Gamma^*(u, b).$$

Now order  $u_1, u_{(2)}, \dots, u_n$  as  $u_{(1)} \leq u_{(2)} \leq \dots \leq u_{(n)}$ . Thus we have

$$P(U > u_{(i)}) = 1 - \Gamma^*(u_{(i)}, b).$$

Therefore

$$\log\{\hat{P}(U > u_{(i)})\} \approx \log\{1 - \Gamma^*(u_{(i)}, b)\}.$$

So in order to check the validity of the model we can plot

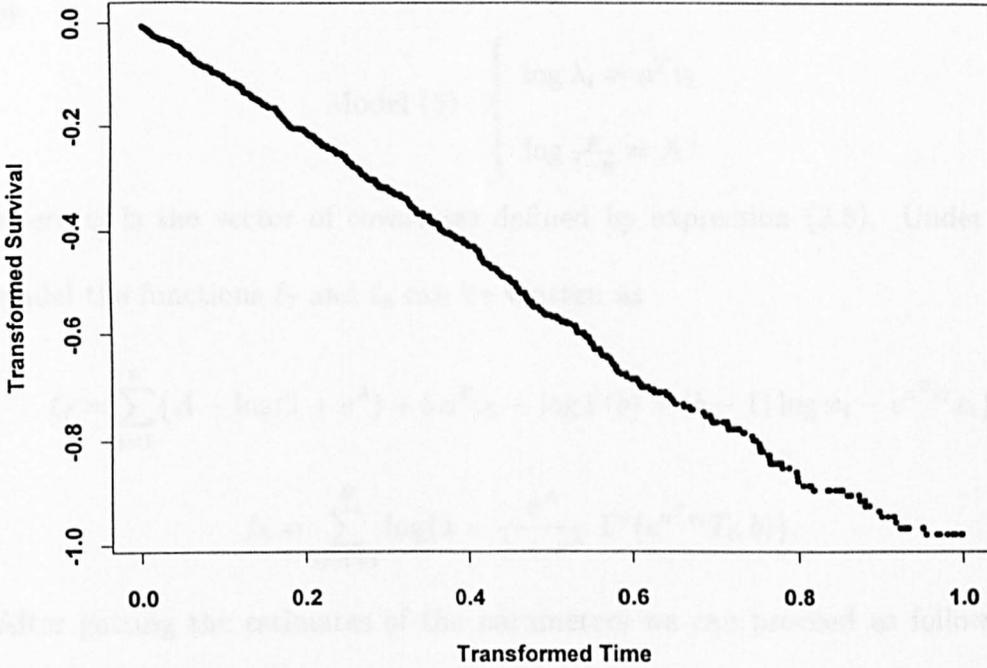
$$\log\{\hat{P}(U > u_{(i)})\} \text{ against } -\log\{1 - \Gamma^*(u_{(i)}, b)\}.$$

If the model holds the plot should resemble a straight line. The plot is depicted in Figure 2.4 which is reasonably linear, substantiating the goodness of fit of the model. The estimated value of  $b$  is  $\hat{b} = 0.7136$ .

### Gamma Distribution With SPM

Suppose that the reoffence time  $X$  has density function defined by equation (2.3). Then using the split population model given by  $p$  and  $\lambda$ , the likelihood

Fig. 2.4 Transformed Survival Plot (X:gamma)



function is given by

$$L = \prod_{i=1}^n \left\{ p \frac{\lambda_i^b}{\Gamma(b)} x_i^{b-1} e^{-\lambda_i x_i} \right\} \prod_{i=n+1}^N \{1 - p \Gamma^*(\lambda_i T_i, b)\}$$

and the log-likelihood function is given by

$$\ell = \ell_7 + \ell_8$$

where

$$\ell_7 = \sum_{i=1}^n \{ \log p + b \log \lambda_i - \log \Gamma(b) + (b-1) \log x_i - \lambda_i x_i \}$$

$$\ell_8 = \sum_{i=n+1}^N \log \{1 - p \Gamma^*(\lambda_i T_i, b)\}.$$

Now assume that the rate of reoffending  $\lambda_i$  and the split proportion  $p$  are given by

$$\text{Model (5) : } \begin{cases} \log \lambda_i = a^T v_i \\ \log \frac{p}{1-p} = A \end{cases}$$

where  $v_i$  is the vector of covariates defined by expression (2.5). Under this model the functions  $\ell_7$  and  $\ell_8$  can be written as

$$\ell_7 = \sum_{i=1}^n \{A - \log(1 + e^A) + b a^T v_i - \log \Gamma(b) + (b - 1) \log x_i - e^{a^T v_i} x_i\}$$

$$\ell_8 = \sum_{i=n+1}^N \log \left\{ 1 - \frac{e^A}{1 + e^A} \Gamma^*(e^{a^T v_i} T_i, b) \right\}.$$

After getting the estimates of the parameters we can proceed as follows for the goodness of fit of the model.

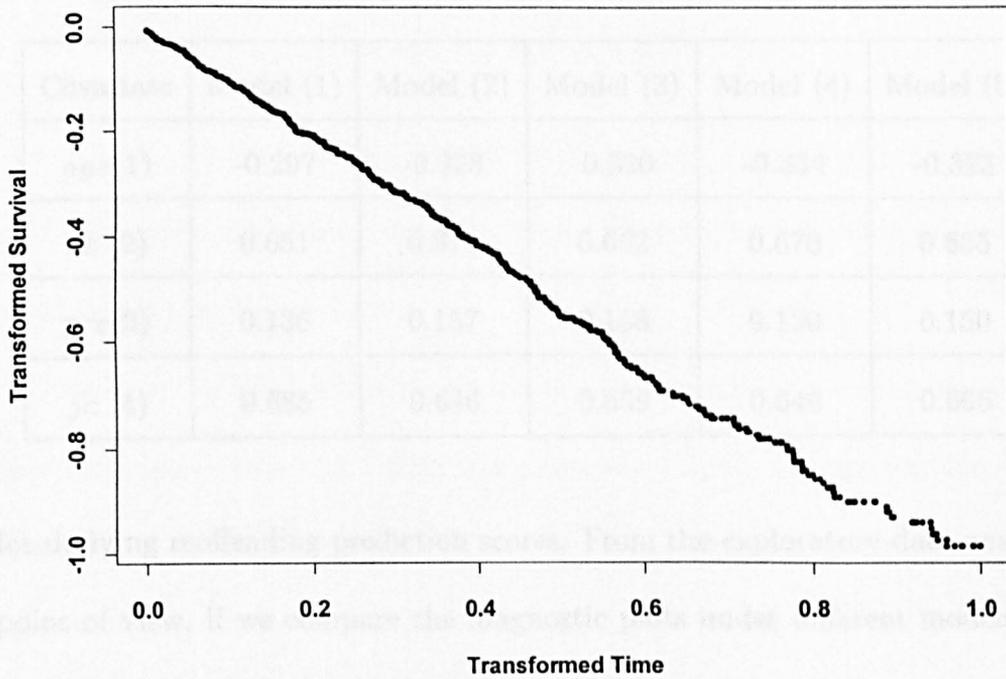
### Goodness of fit of Model (5)

Let  $U = \lambda X$  then using the split population model it can be shown that in order to check the goodness of fit of the model we need to plot

$$\log\{\hat{P}(U > u_{(i)})\} \text{ against } -\log\{1 - p \Gamma^*(u_{(i)}, b)\}.$$

for  $i = 1, \dots, n$ . If the model holds the plot should resemble a straight line. The plot is illustrated in Figure 2.5 which is reasonably linear, substantiating the goodness of fit of the model. Here the maximum likelihood estimate of  $b$  is  $\hat{b} = 0.752$  and that of  $p$  is  $\hat{p} = 0.89$ . Again note that the estimate of  $p$  is reasonably close to 1.

Fig. 2.5 Transformed Survival Plot for SPM (X:gamma)



### Comparison of the Reoffence Models (1)–(5)

Let  $C_{jk}$  be the coefficient of covariate  $j$  in model  $k$ ,  $j = 1, \dots, J$ ,  $k = 1, \dots, K$ . Unfortunately, the values of  $\lambda$  in the different models are not directly comparable because of the different statistical distributions being used. One way of comparing the  $C_{jk}$ 's is to define the 'relative importance' of covariate  $j$  in model  $k$  by

$$\frac{C_{jk}}{\sqrt{\sum_{j=1}^J C_{jk}^2}}, \quad j = 1, \dots, J, \quad k = 1, \dots, K. \quad (2.6)$$

These values are summarized in Table 2.2.

From this table we conclude that the models, in particular the models (2), (3), (4) and (5), provide similar results for the contribution of any given covariate. In other words, we can use each of them as a statistical model

Table 2.2: Relative Importance of Covariates in Reoffence Models

Covariate	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
<i>age</i> (1)	-0.297	-0.328	-0.320	-0.334	-0.323
<i>ac</i> (2)	0.651	0.671	0.662	0.670	0.655
<i>pre</i> (3)	0.136	0.157	0.156	0.150	0.150
<i>jc</i> (4)	0.685	0.646	0.659	0.646	0.666

for deriving reoffending prediction scores. From the exploratory data analysis point of view, if we compare the diagnostic plots under different models, we see that all the models give a good fit to the data.

Let

$$S_{ik} = \log \lambda_i \quad (2.7)$$

for the  $k$ th model. The numerical values of  $S_{ik}$  can be thought of as prediction scores. Now define

$$S_k = \{S_{ik} | i = 1, \dots, N\} \quad (2.8)$$

for  $k = 2, 3, 4, 5$ . The correlation matrix of  $S_k$ 's is

$$\begin{pmatrix} 1.00000 & 0.99994 & 0.99979 & 0.99995 \\ 0.99994 & 1.00000 & 0.99954 & 0.99987 \\ 0.99979 & 0.99954 & 1.00000 & 0.99986 \\ 0.99995 & 0.99987 & 0.99986 & 1.00000 \end{pmatrix}$$

From this correlation matrix it is obvious that the  $S_k$ 's are highly correlated with each other, implying that under different models considered the scores

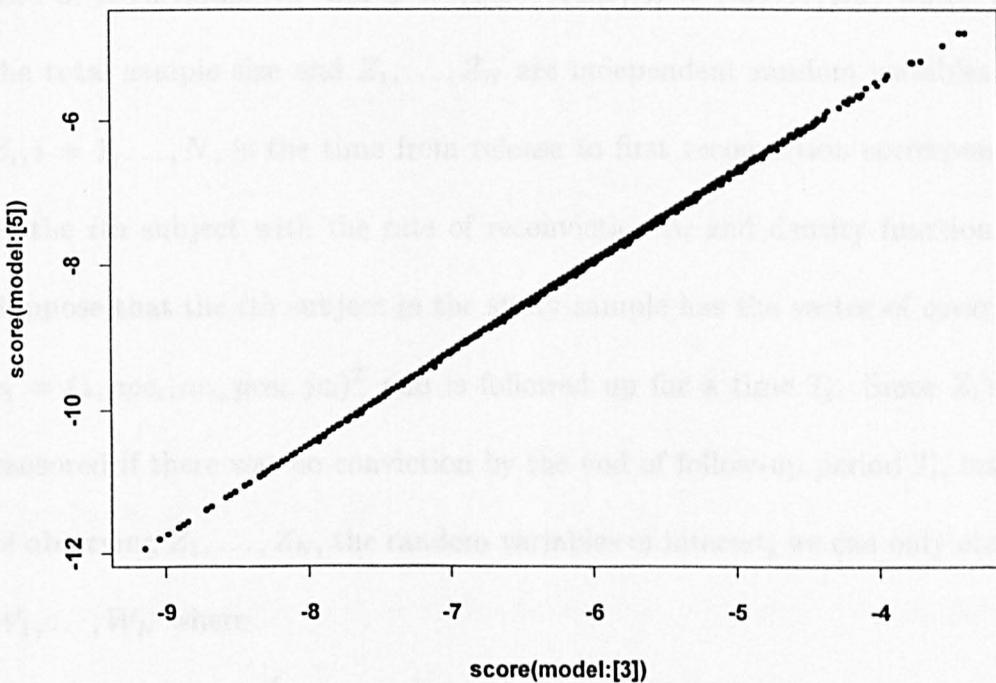
are almost the same. Another approach to this result is to look at the plots of

$$\{(S_k, S_\ell), k, \ell = 2, 3, 4, 5, k \neq \ell\}.$$

It can be shown that all the plots are linear and very similar, and thus showing that the scores are almost the same under different models. One typical plot is illustrated in Figure 2.6.

We have explored different models but the exponential is simple and also interpretable in terms of  $\lambda$  (the rate parameter) and  $p$  (the proportion of the population who will reoffend). This is the model we choose to develop in later chapters.

Fig. 2.6 Plot of Scores for Reoffence Models



## 2.5 Model for Reconviction Prediction Score

Most studies in the past have studied the time to reconviction rather than the time to reoffence. For completeness we complete this chapter by examining which if any of these models can also be used for examining the reconviction times. This is simpler in the sense that the problem of delayed censoring does not arise—data on reconviction times take the form of standard censored survival data.

Everything we have talked about so far for the reoffence times can be repeated exactly in the same way for the reconviction times. Assume that the time from release to first reconviction is a random variable  $Z$  with the rate of reconviction  $\lambda$ . Let  $Z = (Z_1, \dots, Z_N)$ ,  $\lambda = (\lambda_1, \dots, \lambda_N)$  where  $N$  is the total sample size and  $Z_1, \dots, Z_N$  are independent random variables and  $Z_i, i = 1, \dots, N$ , is the time from release to first reconviction corresponding to the  $i$ th subject with the rate of reconviction  $\lambda_i$  and density function  $f_{Z_i}$ . Suppose that the  $i$ th subject in the study sample has the vector of covariates  $v_i = (1, age_i, ac_i, pre_i, jc_i)^T$  and is followed up for a time  $T_i$ . Since  $Z_i$ 's are censored if there was no conviction by the end of follow-up period  $T_i$ , instead of observing  $Z_1, \dots, Z_N$ , the random variables of interest, we can only observe  $W_1, \dots, W_N$  where

$$W_i = \begin{cases} Z_i & \text{if } Z_i \leq T_i, \text{ that is, } Z_i \text{ is not censored} \\ T_i & \text{if } Z_i > T_i, \text{ that is, } Z_i \text{ is censored.} \end{cases}$$

The number of uncensored and censored observations in the sample are  $n$

and  $N - n$  respectively. We shall now consider the following models for the reconviction time  $Z$ .

### Exponential Distribution With SPM

Suppose that  $Z_i, i = 1, \dots, N$ , has exponential distribution with mean  $1/\lambda_i$  where  $\lambda_i$  and the split proportion  $p$  are given by

$$\text{Model (1)} : \begin{cases} \log \lambda_i = a^T v_i \\ \log \frac{p}{1-p} = A. \end{cases}$$

Under this model the log-likelihood function is given by

$$\ell = \ell_9 + \ell_{10}$$

where

$$\ell_9 = \sum_{i=1}^n \{A - \log(1 + e^A) + a^T v_i - z_i e^{a^T v_i}\}$$

$$\ell_{10} = \sum_{i=n+1}^N \log\left\{\frac{1 + e^{(A - T_i e^{a^T v_i})}}{1 + e^A}\right\}.$$

After finding the maximum likelihood estimates of  $A$ ,  $a^T = (a_1, a_2, a_3, a_4, a_5)$ ,  $\lambda$  and  $p$  we can check the goodness of fit of the model as follows.

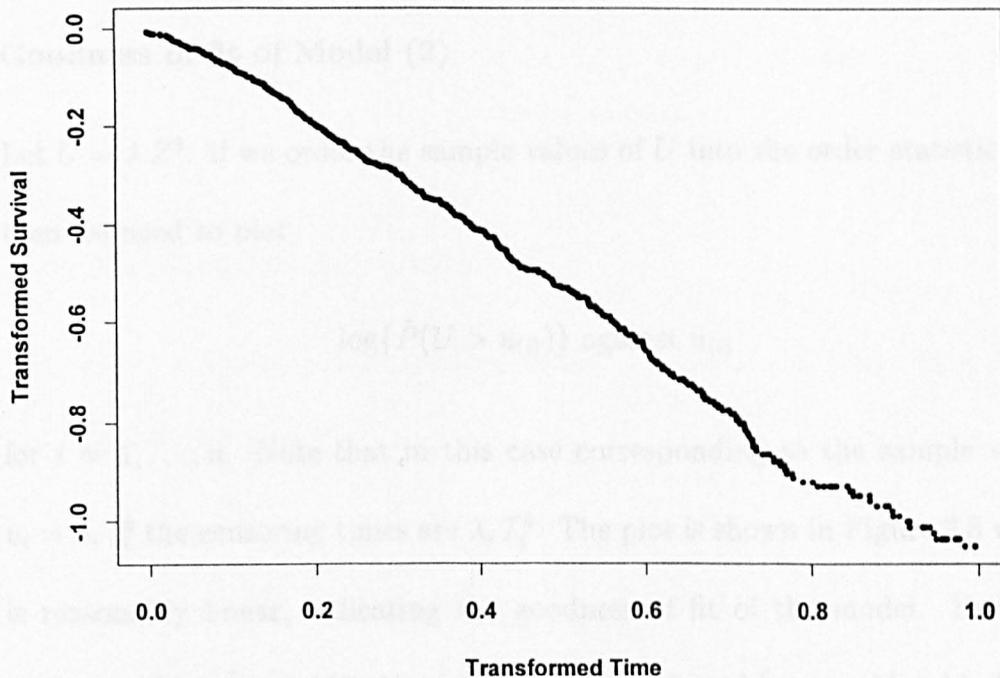
### Goodness of fit of Model (1)

To check the goodness of fit of the model let  $U = \lambda Z$ , where  $Z$  has exponential distribution with mean  $1/\lambda$  and  $U$  is unit exponential. We can now order the sample values of  $U$  into the order statistics  $u_{(i)}$  and plot

$$\log\left\{\frac{\hat{P}(U > u_{(i)}) - (1 - p)}{p}\right\} \text{ against } u_{(i)}$$

for  $i = 1, 2, \dots, n$ . Here  $\hat{P}(U > u_{(i)})$  is the estimated survival curve of  $P(U > u_{(i)})$  which is obtained from the sample values of  $u_{(1)}, \dots, u_{(n)}$  and the censoring times  $\lambda_i T_i$  by the Kaplan Meier survival curve. If the model holds the plot should resemble a straight line. The plot is shown in Figure 2.7 which is clearly not linear, suggesting that the exponential distribution is not an appropriate model assumption for the reconviction time  $Z$ , also as noted by earlier researchers in this area. Here the estimated value of  $p$  is  $\hat{p}=0.99$ .

Fig. 2.7 Transformed Survival Plot for SPM (Z:Exponential)



### Weibull Distribution

Assume that  $Z$  has Weibull distribution with density function

$$f_Z(z) = \lambda b z^{b-1} e^{-\lambda z^b}, \quad \lambda > 0, b > 0, z > 0.$$

If the rate of reconviction  $\lambda_i$  is defined by

$$\text{Model (2) : } \log \lambda_i = a^T v_i$$

then under this model the log-likelihood function is given by

$$\ell = \sum_{i=1}^n \{\log \lambda_i + \log b + (b - 1) \log z_i - \lambda_i z_i^b\} + \sum_{i=n+1}^N (-\lambda_i T_i^b)$$

from which we can obtain the maximum likelihood estimates of the parameters.

To check the goodness of fit of the model we proceed as follows.

### Goodness of fit of Model (2)

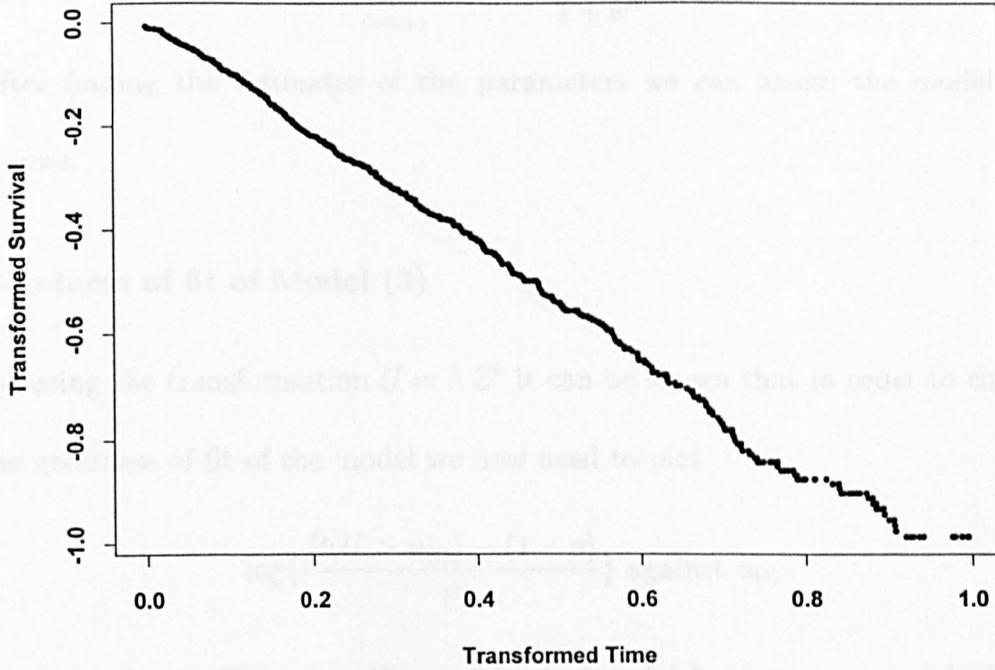
Let  $U = \lambda Z^b$ . If we order the sample values of  $U$  into the order statistics  $u_{(i)}$

then we need to plot

$$\log\{\hat{P}(U > u_{(i)})\} \text{ against } u_{(i)}$$

for  $i = 1, \dots, n$ . Note that in this case corresponding to the sample values  $u_i = \lambda_i z_i^b$  the censoring times are  $\lambda_i T_i^b$ . The plot is shown in Figure 2.8 which is reasonably linear, indicating the goodness of fit of the model. Here the estimate of  $b$  is  $\hat{b} = 1.127$ . Note this that the value of  $\hat{b}$  is considerably larger than the estimate found in the analysis of the time to first reoffence; we would expect this as the reconviction time is equal to the sum of reoffence time and the delay time between the offence and conviction.

Fig. 2.8 Transformed Survival Plot (Z:Weibull)



### Weibull Distribution With SPM

Consider the same assumptions as before for the reconviction time  $Z$ . If  $\lambda_i$  and  $p$  are defined by

$$\text{Model (3) : } \begin{cases} \log \lambda_i = a^T v_i \\ \log \frac{p}{1-p} = A \end{cases}$$

then the log-likelihood function is given by

$$\ell = \ell_{11} + \ell_{12}$$

where

$$\ell_{11} = \sum_{i=1}^n \{A - \log(1 + e^A) + a^T v_i + \log(b) + (b - 1) \log z_i - z_i^b e^{a^T v_i}\}$$

$$\ell_{12} = \sum_{i=n+1}^N \log\left\{\frac{1 + e^{(A - T_i^b e^{a^T v_i})}}{1 + e^A}\right\}.$$

After finding the estimates of the parameters we can assess the model as follows.

### Goodness of fit of Model (3)

By using the transformation  $U = \lambda Z^b$  it can be shown that in order to check the goodness of fit of the model we now need to plot

$$\log\left\{\frac{\hat{P}(U > u_{(i)}) - (1 - p)}{p}\right\} \text{ against } u_{(i)}$$

for  $i = 1, \dots, n$ . This gives Figure 2.9 which is fairly close to a straight line, substantiating the goodness of fit of the model. This model fits the data very well as compared with the exponential model. Here the estimated value of  $b$  is  $\hat{b} = 1.19$  and that of  $p$  is  $\hat{p} = 0.89$ . Note that the estimated value of  $p$  is reasonably close to 1 and the value of  $\hat{b}$  is considerably large again.

### Gamma Distribution

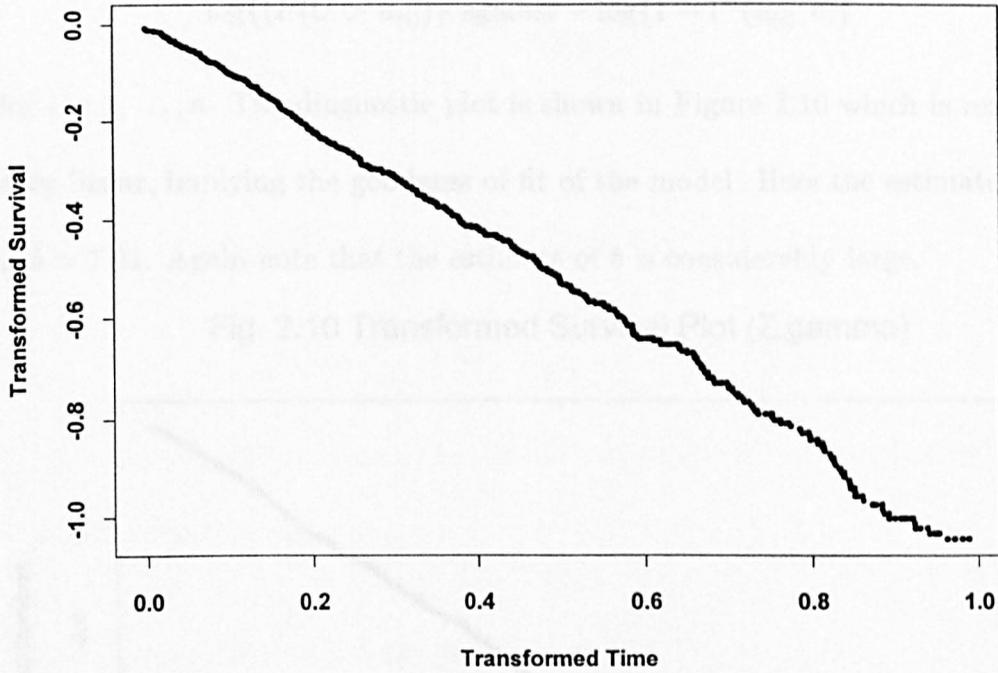
Assume that the reconviction time  $Z$  ( $Z > 0$ ) has gamma distribution with density function

$$f_Z(z) = \frac{\lambda^b}{\Gamma(b)} z^{b-1} e^{-\lambda z}, \quad z > 0, \lambda > 0, b > 0$$

and  $\lambda_i, i = 1, \dots, N$ , is defined by

$$\text{Model (4) : } \log \lambda_i = a^T v_i.$$

Fig. 2.9 Transformed Survival Plot for SPM (Z:Weibull)



Using this model the log-likelihood function is given by

$$\ell = \ell_{13} + \ell_{14}$$

where

$$\ell_{13} = \sum_{i=1}^n \{b a^T v_i - \log \Gamma(b) + (b-1) \log z_i - e^{a^T v_i} z_i\}$$

$$\ell_{14} = \sum_{i=n+1}^N \log \{1 - \Gamma^*(e^{a^T v_i} T_i, b)\}.$$

First we find the maximum likelihood estimates of the parameters and then we can assess the model as follows.

#### Goodness of fit of Model (4)

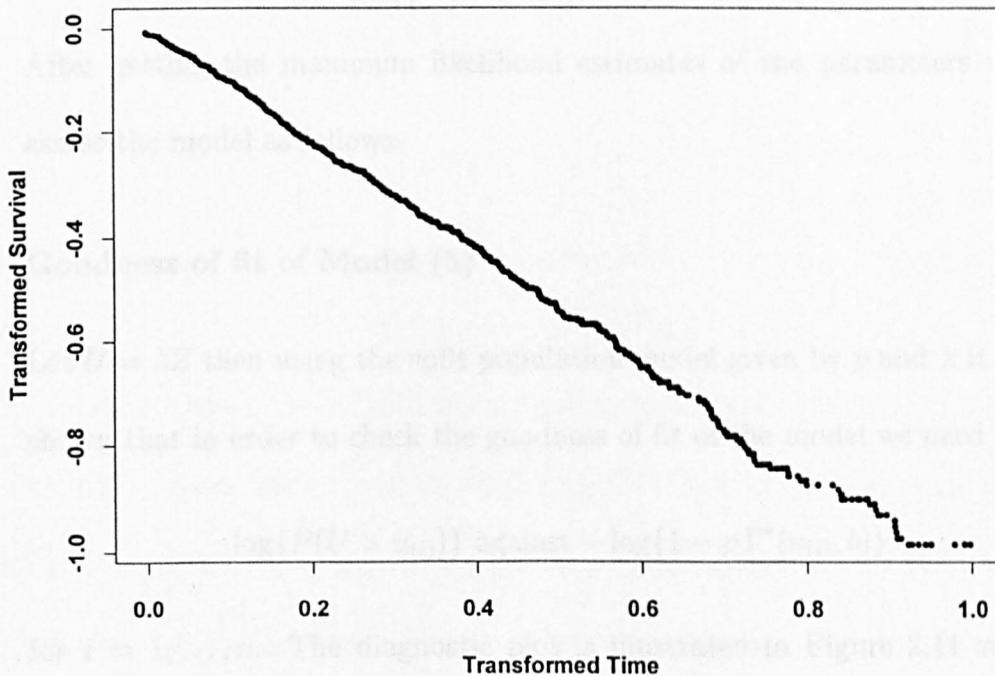
Let  $U = \lambda Z$  and order the sample values of  $U$  into the order statistics  $u_{(i)}$ . It can be shown that in order to check the goodness of fit of the model we can

plot

$$\log\{\{\hat{P}(U > u_{(i)})\} \text{ against } -\log\{1 - \Gamma^*(u_{(i)}, b)\}$$

for  $i = 1, \dots, n$ . The diagnostic plot is shown in Figure 2.10 which is reasonably linear, implying the goodness of fit of the model. Here the estimate of  $b$  is  $\hat{b} = 1.21$ . Again note that the estimate of  $b$  is considerably large.

Fig. 2.10 Transformed Survival Plot (Z:gamma)



### Gamma Distribution With SPM

Suppose that the reconviction time  $Z$  has gamma distribution as before. If  $\lambda_i$  and the split proportion  $p$  are defined by

$$\text{Model (5) : } \begin{cases} \log \lambda_i = a^T v_i \\ \log \frac{p}{1-p} = A \end{cases}$$

then the log-likelihood function is given by

$$\ell = \ell_{15} + \ell_{16}$$

where

$$\ell_{15} = \sum_{i=1}^n \{A - \log(1 + e^A) + b a^T v_i - \log \Gamma(b) + (b - 1) \log z_i - e^{a^T v_i} z_i\}$$

$$\ell_{16} = \sum_{i=n+1}^N \log\left\{1 - \frac{e^A}{1 + e^A} \Gamma^*(e^{a^T v_i} T_i, b)\right\}.$$

After getting the maximum likelihood estimates of the parameters we can assess the model as follows.

### Goodness of fit of Model (5)

Let  $U = \lambda Z$  then using the split population model given by  $p$  and  $\lambda$  it can be shown that in order to check the goodness of fit of the model we need to plot

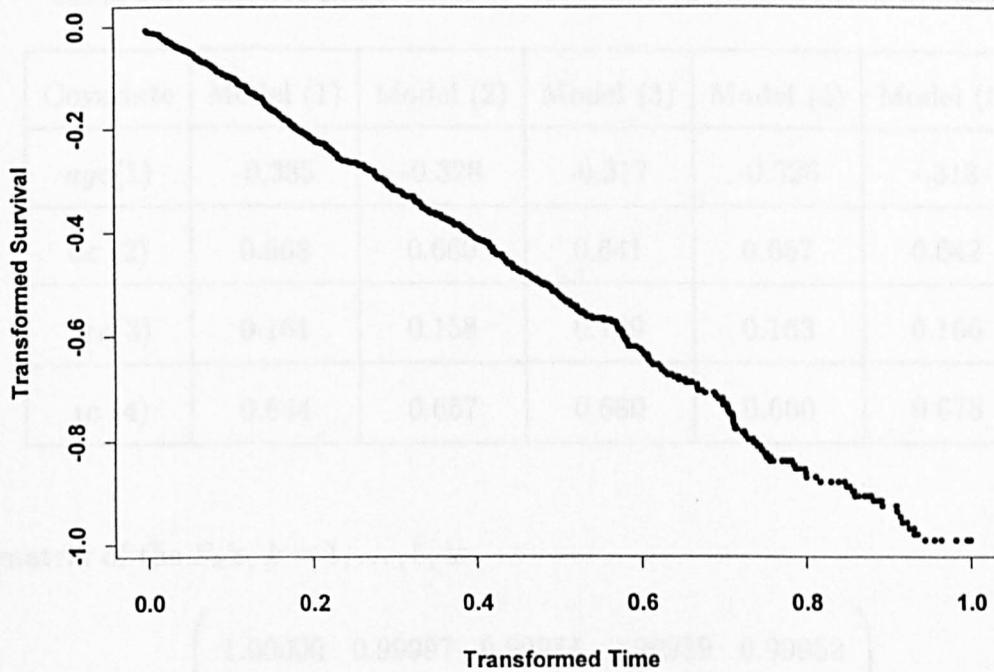
$$\log\{\hat{P}(U > u_{(i)})\} \text{ against } -\log\{1 - p \Gamma^*(u_{(i)}, b)\}$$

for  $i = 1, \dots, n$ . The diagnostic plot is illustrated in Figure 2.11 which is reasonably linear, indicating the goodness of fit of the model. This model also gives a good fit to the data. Here the estimate of  $b$  is  $\hat{b}=1.28$  and that of  $p$  is  $\hat{p}=0.91$ . Again note that the estimate of  $p$  is reasonably close to 1 and that of  $b$  is considerably large.

### Comparison of the Reconviction Models (1)–(5)

Using the same definition as before, expression (2.6), the relative importance of covariate  $j$  in model  $k$ ,  $j = 1, \dots, 4$ ,  $k = 1, \dots, 5$ , is given in Table 2.3.

Fig. 2.11 Transformed Survival Plot for SPM (Z:gamma)



From this table we see that the models, especially the models (2), (3), (4) and (5), provide similar results for the contribution of any given covariate. Thus each of the models can be used as a statistical model for deriving reconviction prediction scores. On the other hand, if we compare the diagnostic plots under different models, then it is obvious that the models (2), (3), (4) and (5) are more appropriate than model (1) for these data. Now using the formulae (2.7) and (2.8) for the reconviction prediction scores, the correlation

Table 2.3: Relative Importance of Covariates in Reconviction Models

Covariate	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)
<i>age</i> (1)	-0.335	-0.328	-0.317	-0.326	-.318
<i>ac</i> (2)	0.668	0.660	0.641	0.657	0.642
<i>pre</i> (3)	0.161	0.158	0.159	0.163	0.166
<i>jc</i> (4)	0.644	0.657	0.680	0.660	0.678

matrix of the  $S_k$ 's,  $k = 1, \dots, 5$ , is

$$\begin{pmatrix} 1.00000 & 0.99997 & 0.99974 & 0.99989 & 0.99952 \\ 0.99997 & 1.00000 & 0.99986 & 0.99995 & 0.99968 \\ 0.99974 & 0.99986 & 1.00000 & 0.99995 & 0.99994 \\ 0.99989 & 0.99995 & 0.99995 & 1.00000 & 0.99987 \\ 0.99952 & 0.99968 & 0.99994 & 0.99987 & 1.00000 \end{pmatrix}$$

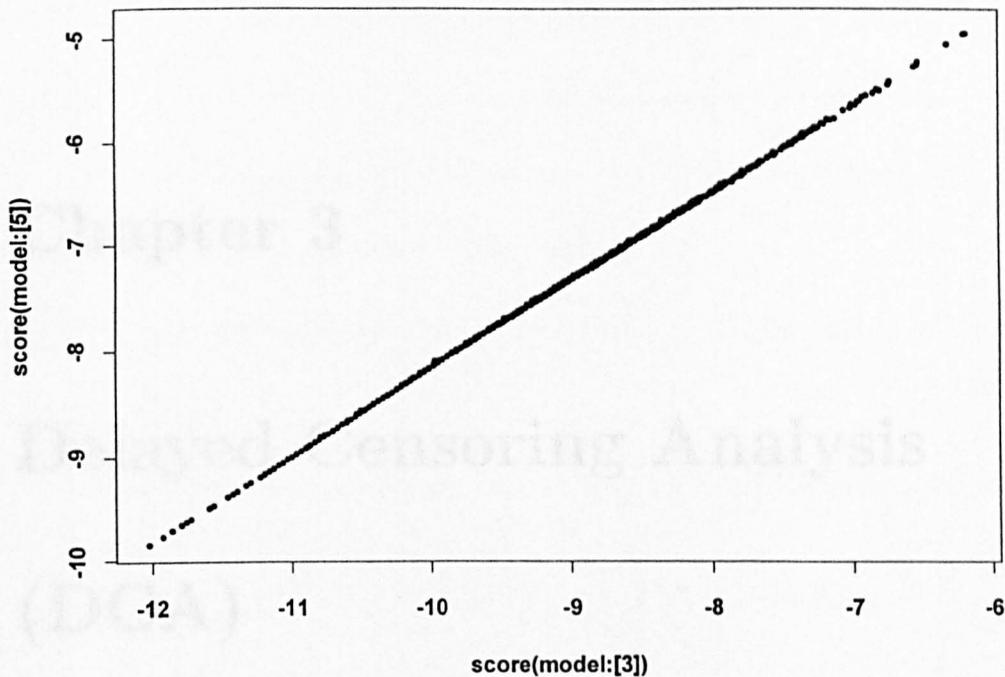
From this correlation matrix we observe that the  $S_k$ 's are highly correlated with each other, implying that the scores are almost the same under different models. An alternative approach to this result is to look at the plots of

$$\{(S_k, S_\ell), k, \ell = 1, \dots, 5, k \neq \ell\}.$$

Here again it can be shown that all the plots are linear and very similar, indicating that the scores are almost the same under different models. One typical plot is shown in Figure 2.12.

Comparison of the Tables 2.2 and 2.3 shows that the reconviction and

Fig. 2.12 Plot of Scores for Reconviction Models



reoffending scores are almost the same under different models considered for the offence and conviction times. Consequently, the statistical scores obtained in the prediction models for reconviction can be used for reoffence as well.

# Chapter 3

## Delayed Censoring Analysis (DCA)

### 3.1 Statistical Analysis

Let  $X$ ,  $Z$  and  $T$  be the reoffence, reconviction and follow up times respectively. The delay  $Y$  is then defined to be  $Y = Z - X$ . We start with the simplest possible model, to assume that  $X$  and  $Y$  are independent and exponentially distributed with parameters  $\lambda$  and  $\theta$  respectively. We have two cases for our observations:

- (1) observe both  $X$  and  $Y$  with probability  $\lambda e^{-\lambda x} \theta e^{-\theta y}$
- (2) observe nothing if and only if  $X + Y > T$  with probability  $P(X + Y > T)$ .

Note that if the values of  $X$  are modelled as ordinary censored survival data, then  $Y = 0$  or equivalently  $X = Z$ . In this case an observation is assumed to be censored if  $X > T$ . But this is an invalid assumption, as an observation

is censored if  $X + Y > T$ , and this does not necessarily imply that  $X > T$ . However, in delayed censoring, an offence may in fact have occurred by time  $X$ ,  $X < T$ , but not observed as the case has not yet come to court.

With delayed censoring our data now consist of the following form:

$(x_i, y_i)$ ,  $i = 1, 2, \dots, n$  (observed cases)

$(N - n)$  censored cases with follow-up times  $t_{n+1}, \dots, t_N$ , where  $N$  is the total sample size.

To estimate the parameters, we use the likelihood function as follows:

$$L = \prod_{i=1}^n \lambda e^{-\lambda x_i} \theta e^{-\theta y_i} \prod_{i=n+1}^N P(X + Y > t_i).$$

To compute  $P(X + Y > t)$ , let  $Z = X + Y$  then

$$P(X + Y > t) = \int_t^{+\infty} f_Z(z) dz$$

where  $f_Z(z)$  is the probability density function of  $Z$ . In general, we have

$$f_Z(z) = \int_{-\infty}^{+\infty} f_{X,Y}(x, z - x) dx$$

which, under our assumptions, can be written as

$$f_Z(z) = \int_0^z f_Y(z - x) f_X(x) dx = \int_0^z \theta e^{-\theta(z-x)} \lambda e^{-\lambda x} dx$$

Now consider two cases:

**case 1:**  $\lambda = \theta$ , **case 2:**  $\lambda \neq \theta$ .

For case 1, we find that

$$f_Z(z) = \lambda^2 z e^{-\lambda z}$$

which is a gamma density with shape parameter 2, and

$$P(Z > t) = P(X + Y > t) = (1 + \lambda t)e^{-\lambda t}. \quad (3.1)$$

The log-likelihood function is then given by

$$\ell = \sum_{i=1}^n (2 \log \lambda - \lambda z_i) + \sum_{i=n+1}^N \{\log(1 + \lambda t_i) - \lambda t_i\}$$

or

$$\ell = \sum_u (2 \log \lambda - \lambda z) + \sum_c \{\log(1 + \lambda t) - \lambda t\}$$

where the first sum is taken over the uncensored cases and second sum is taken over the censored cases.

Similarly, for case 2, it can be shown that

$$f_Z(z) = \frac{\theta \lambda (e^{-\lambda z} - e^{-\theta z})}{\theta - \lambda}$$

and

$$P(Z > t) = P(X + Y > t) = \frac{(\theta e^{-\lambda t} - \lambda e^{-\theta t})}{\theta - \lambda} \quad (3.2)$$

so the log-likelihood function is given by

$$\ell = \sum_u (\log \lambda - \lambda x + \log \theta - \theta y) + \sum_c \log \frac{\theta e^{-\lambda t} - \lambda e^{-\theta t}}{\theta - \lambda}.$$

Note that

$$\frac{\theta e^{-\lambda t} - \lambda e^{-\theta t}}{\theta - \lambda} \rightarrow (1 + \lambda t)e^{-\lambda t}$$

and

$$\frac{\theta \lambda (e^{-\lambda z} - e^{-\theta z})}{\theta - \lambda} \rightarrow \lambda^2 z e^{-\lambda z}$$

as  $\theta \rightarrow \lambda$ , as expected.

Combining these results, the general log-likelihood function is given by

$$\ell(\lambda, \theta) = \begin{cases} \ell_1 + \ell_2 & \text{if } \lambda = \theta \\ \ell_3 + \ell_4 & \text{if } \lambda \neq \theta \end{cases} \quad (3.3)$$

where

$$\ell_1 = \sum_{\mathbf{u}} (2 \log \lambda - \lambda z)$$

$$\ell_2 = \sum_c (\log(1 + \lambda t) - \lambda t)$$

$$\ell_3 = \sum_{\mathbf{u}} (\log \lambda - \lambda x + \log \theta - \theta y)$$

$$\ell_4 = \sum_c \log \frac{\theta e^{-\lambda t} - \lambda e^{-\theta t}}{\theta - \lambda}$$

and as before,  $\sum_{\mathbf{u}}$  denotes the sum over uncensored and  $\sum_c$  denotes the sum over censored observations. Note that in maximizing  $\ell(\lambda, \theta)$ , it will be helpful to use the transformations  $\log \lambda = a_1$ ,  $\log \theta = a_2$ .

Fitting this model to the data gave the maximum likelihood estimates of the parameters to be  $\hat{\lambda} = 0.0005$  and  $\hat{\theta} = 0.0066$ .

## 3.2 Diagnostics for Reoffence Time ( $X$ )

Let  $U = \lambda X$ , then we observe  $u_1, \dots, u_n$  only if  $x + y < t$  (uncensored cases) and do not observe  $u_1, \dots, u_n$  if  $x + y > t$  (censored cases). Note that to allow for the censoring we need to condition on the event that  $X$  or equivalently  $U$  is observed. To assess the assumptions of the model, we now suggest a diagnostic plot based on the observed values of  $U$ .

Consider

$$P(U > u | X + Y < t) = P(X > \frac{u}{\lambda} | X + Y < t). \quad (3.4)$$

First we find density function of  $X$  conditioned on  $X + Y < t$  which is defined to be

$$f_{X|X+Y<t}(x, t) = \frac{P(X, X + Y < t)}{P(X + Y < t)}. \quad (3.5)$$

After some algebra it can be shown that

$$f_{X|X+Y<t}(x, t) = \begin{cases} \frac{\lambda\{e^{-\lambda x} - e^{-\lambda t}\}}{1 - (1 + \lambda t)e^{-\lambda t}} & \text{if } \lambda = \theta \\ \frac{\lambda\{e^{-\lambda x} - e^{-(\theta - \lambda)x - \theta t}\}}{1 - \frac{\theta e^{-\lambda t} - \lambda e^{-\theta t}}{\theta - \lambda}} & \text{if } \lambda \neq \theta \end{cases}$$

and so

$$\begin{aligned} P(U > u | X + Y < t) &= \int_{u/\lambda}^t f_{X|X+Y<t}(x, t) dx \\ &= \begin{cases} F_1(\lambda, u, t) & \text{if } \lambda = \theta \\ F_2(\lambda, \theta, u, t) & \text{if } \lambda \neq \theta \end{cases} \end{aligned}$$

where

$$F_1(\lambda, u, t) = \frac{e^{-u} + (u - \lambda t - 1)e^{-\lambda t}}{1 - (1 + \lambda t)e^{-\lambda t}} \quad (3.6)$$

and

$$F_2(\lambda, \theta, u, t) = \frac{(\theta - \lambda)(e^{-u} - e^{-\lambda t}) - \lambda e^{-\theta t}\{e^{(\theta - \lambda)t} - e^{(\theta - \lambda)u/\lambda}\}}{\theta(1 - e^{-\lambda t}) - \lambda(1 - e^{-\theta t})} \quad (3.7)$$

whenever  $u < \lambda t$ . The value of  $F_2(\lambda, \theta, u, t)$  is taken as zero if  $u \geq \lambda t$ . Note that if  $t \rightarrow +\infty$ , then

$$P(U > u | X + Y < t) \rightarrow e^{-u} = P(U > u)$$

which is the survival function of the unit exponential distribution, as expected. In section 3.1 it was found that the estimates of the parameters are different and so only equation (3.7) is needed for plotting.

Now consider the uncensored subjects. Averaging equation (3.4) over these subjects gives

$$E(\text{proportion of } U_i\text{'s} > u | X_i + Y_i < t_i) = \frac{1}{n} \sum_{i=1}^n F_2(\lambda, \theta, u, t_i).$$

Thus we expect the proportion of  $U_i$ 's  $> u$  to be approximately

$$\frac{1}{n} \sum_{i=1}^n F_2(\lambda, \theta, u, t_i).$$

Now order  $u_1, u_2, \dots, u_n$  as  $u_{(1)} \leq u_{(2)} \leq \dots \leq u_{(n)}$ , then

$$\text{proportion of } U_i\text{'s} > u_{(j)} \approx \frac{1}{n} \sum_{i=1}^n F_2(\lambda, \theta, u_{(j)}, t_i).$$

But on the other hand, we have

$$\text{proportion of } U_i\text{'s} > u_{(j)} \approx \frac{n+1-j}{n+1}.$$

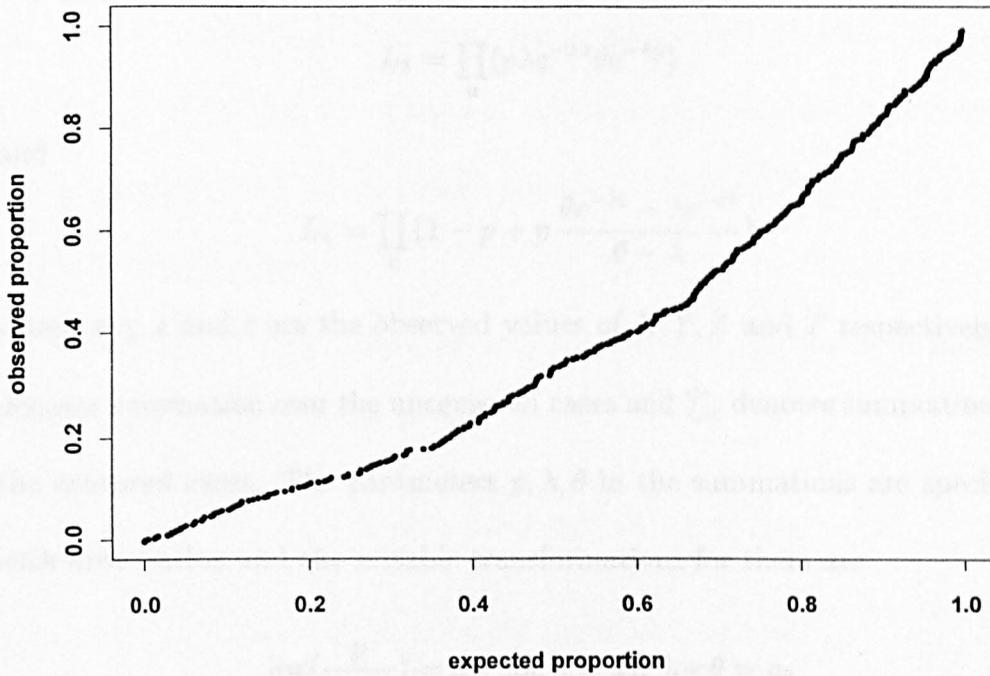
Thus, in order to check the validity of the model we plot

$$\frac{n+1-j}{n+1} \text{ against } \frac{1}{n} \sum_{i=1}^n F_2(\lambda, \theta, u_{(j)}, t_i), \text{ for } j = 1, \dots, n.$$

If the model holds, the plot should resemble a straight line. The plot is depicted in Figure 3.1 which is clearly not linear, implying that the model does not fit the data well. In order to get a possible improvement in the goodness of fit of the model we now go on to consider the split population model which is described in section 3.3.

**Remark:** The formulae in this section are symmetric with respect to  $X$  and  $Y$  and symmetric with respect to  $\lambda$  and  $\theta$  as well.

Fig. 3.1 Diagnostics for X in DCA model (Actual Data)



### 3.3 Split Population Model for DCA

If  $t$  is the time to follow-up, then the value of  $X$ , and hence also the delay  $Y$ , is observed if and only if  $Z \leq t$ . The total probability of a case being censored is therefore  $1 - p + pP(Z > t)$  in which  $p$  is the split proportion used in the split population model, defined in section 2.2, and  $P(Z > t)$  is given by equation (3.1) if  $\lambda = \theta$  and by equation (3.2) if  $\lambda \neq \theta$ . Thus the likelihood function is given by

$$L = \begin{cases} L_1 L_2 & \text{if } \lambda = \theta \\ L_3 L_4 & \text{if } \lambda \neq \theta \end{cases}$$

where

$$L_1 = \prod_u p \lambda^2 e^{-\lambda z}$$

$$L_2 = \prod_c \{1 - p + p(1 + \lambda t) e^{-\lambda t}\}$$

$$L_3 = \prod_u (p\lambda e^{-\lambda x} \theta e^{-\theta y})$$

and

$$L_4 = \prod_c \left\{1 - p + p \frac{\theta e^{-\lambda t} - \lambda e^{-\theta t}}{\theta - \lambda}\right\}$$

where  $x, y, z$  and  $t$  are the observed values of  $X, Y, Z$  and  $T$  respectively,  $\sum_u$  denotes summation over the uncensored cases and  $\sum_c$  denotes summation over the censored cases. The parameters  $p, \lambda, \theta$  in the summations are specific to each observation and the suitable transformations for them are

$$\log\left(\frac{p}{1-p}\right) = a_1, \quad \log \lambda = a_2, \quad \log \theta = a_3$$

where  $a_1, a_2$  and  $a_3$  scalars to be estimated. The log-likelihood function is therefore given by

$$\ell = \begin{cases} \ell_1 + \ell_2 & \text{if } \lambda = \theta \\ \ell_3 + \ell_4 & \text{if } \lambda \neq \theta \end{cases} \quad (3.8)$$

where

$$\ell_1 = \sum_u (\log p + 2 \log \lambda - \lambda z)$$

$$\ell_2 = \sum_c \log\{1 - p + p(1 + \lambda t) e^{-\lambda t}\}$$

$$\ell_3 = \sum_u (\log p + \log \lambda - \lambda x + \log \theta - \theta y)$$

and

$$\ell_4 = \sum_c \log\left\{1 - p + p \frac{\theta e^{-\lambda t} - \lambda e^{-\theta t}}{\theta - \lambda}\right\}.$$

Analogous to section 3.2, we consider equation (3.4) as a basis for the goodness of fit of the model and follow the same procedures developed before.

Under the split population model given by  $p$ ,  $\lambda$  and  $\theta$  we get exactly the same formulae as in section 3.2 because it can be shown that the split proportion  $p$  will be cancelled from the numerator and denominator of the right hand side of the equation (3.5). Consequently, equations (3.6) and (3.7) are still valid in the same form as before.

Fitting the split population model to the data gave the maximum likelihood estimates of the parameters to be

$$\hat{p} = 0.475, \quad \hat{\lambda} = 0.002, \quad \hat{\theta} = 0.007.$$

Therefore, to check the goodness of fit the model, we plot

$$\frac{n+1-j}{n+1} \text{ against } \frac{1}{n} \sum_{i=1}^n F_2(\lambda, \theta, u_{(j)}, t_i)$$

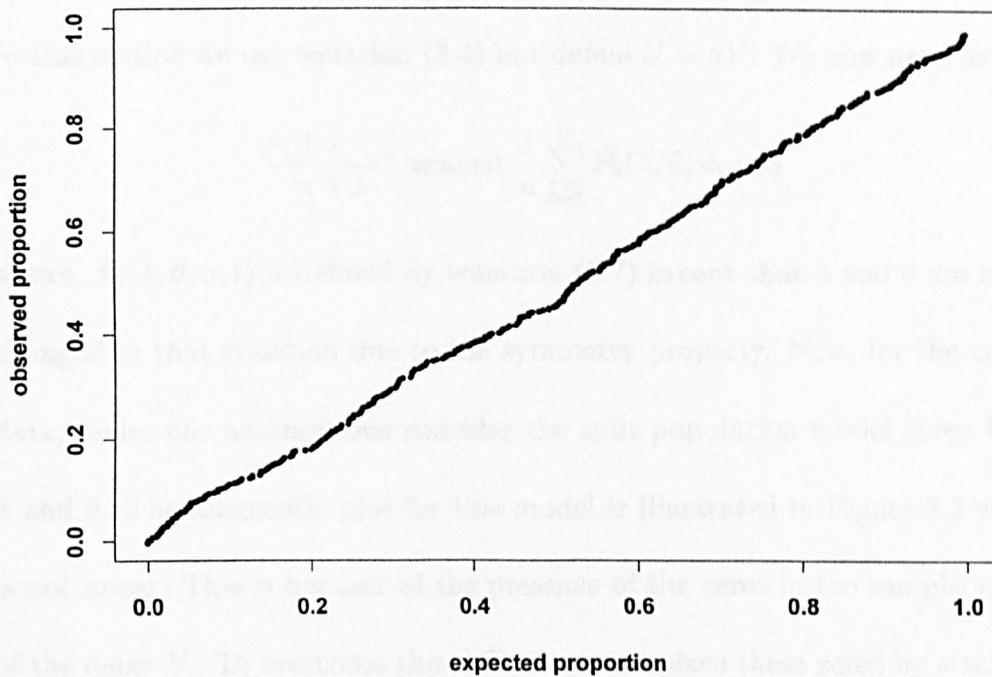
for  $j = 1, \dots, n$ . If the model fits the data well, the plot should resemble a straight line. The plot is illustrated in Figure 3.2 which is reasonably linear, substantiating the validity of the model.

**Remark:**

(i) In sections 3.2 and 3.3 we assumed that  $p$ ,  $\lambda$  and  $\theta$  to be constants. But, in general these parameters can be functions of the covariates. In such cases by a similar argument as in section 3.2 it can be shown that in order to check the goodness of fit of the model we need to plot the observed proportion of  $U$ 's  $> u$  against the expected proportion of  $U$ 's  $> u$  and examine the straightness of the plot. That is, for  $j = 1, \dots, n$ , we plot

$$\frac{n+1-j}{n+1} \text{ against } \frac{1}{n} \sum_{i=1}^n F_2(\lambda_i, \theta_i, u_{(j)}, t_i)$$

Fig. 3.2 Split Population Model Version of Fig. 3.1



where  $n$  is the number of uncensored subjects and the function  $F_2(\lambda, \theta, u, t)$  is defined by equation (3.7). This is of particular importance and is referred to several times in the later diagnostic plots.

(ii) In the following sections where the split population model is not used, i.e.  $p = 1$ , the log-likelihood function is given by equation (3.3) and in the situation where the split population model is considered as well, the log-likelihood function is defined by equation (3.8).

### 3.4 Diagnostics for Delay Time ( $Y$ )

In this section we use equation (3.4) but define  $U = \theta Y$ . We now need to plot

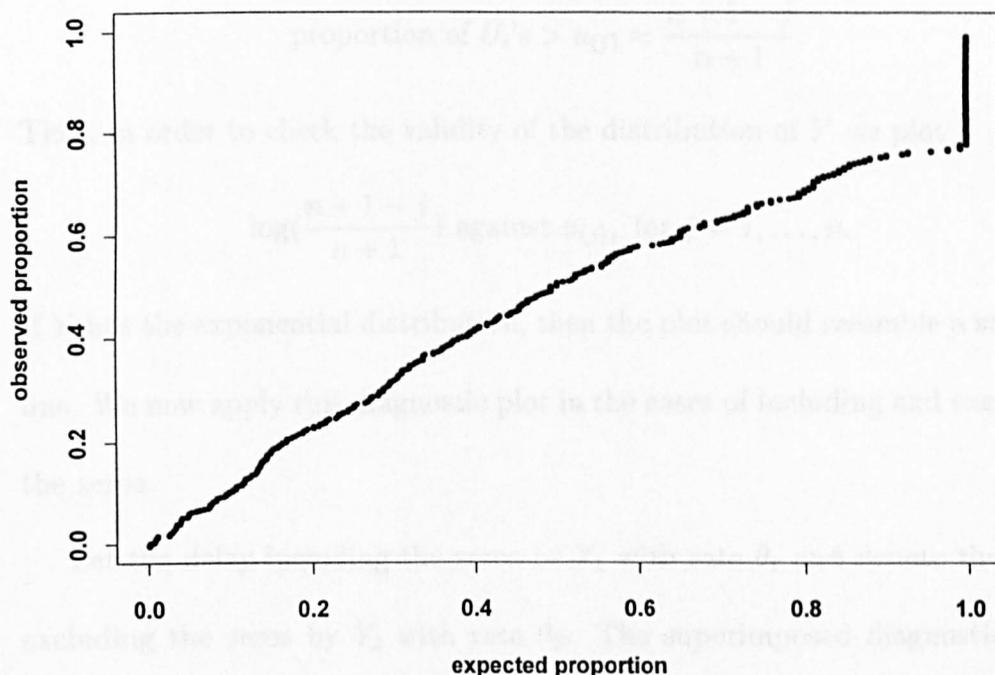
$$\frac{n+1-j}{n+1} \text{ against } \frac{1}{n} \sum_{i=1}^n F_3(\lambda, \theta, u_{(j)}, t)$$

where,  $F_3(\lambda, \theta, u, t)$  is defined by equation (3.7) except that  $\lambda$  and  $\theta$  are interchanged in that equation due to the symmetry property. Now, for the actual data, under our assumptions consider the split population model given by  $p$ ,  $\lambda$  and  $\theta$ . The diagnostic plot for this model is illustrated in Figure 3.3 which is not linear. This is because of the presence of the zeros in the sample values of the delay  $Y$ . To overcome this difficulty we replace these zeros by a sample of small nonzero values of an exponential distribution which is discussed more fully in the following section.

### 3.5 Imputation of Delay Time

The data collection involved linking together different data sources. Data on reconvictions were reliably recorded, but in a substantial minority of cases the date of the corresponding reoffence could not be traced. Sometimes this could be estimated from a known date of arrest, but in about 100 cases this was not possible and the reoffence date was simply recorded as the reconviction date. Consequently, in the sample values of the delay  $Y$  there are about 100 zeros. Removing the zeros from the observed delay times would seriously bias the sample, since we would be removing subjects who are known to have

Fig. 3.3 Diagnostics for Y with SPM (Actual Data)



reoffended, and so we replace the zeros in the data by imputing random observations from an exponential distribution as suggested in this section. In fact, as shown in section 3.11, this imputation makes very little difference to the fitted models, but the presence of the zeros in the delay distribution upsets some of the diagnostic plots to be discussed (such as Figure 3.3).

We shall now consider an alternative diagnostic plot for checking the distribution of  $Y$  which gives us an idea regarding our discussion.

Let  $U = \theta Y$ , then

$$P(Y > y) = e^{-\theta y} \text{ and } P(U > u) = e^{-u}.$$

Now order  $u_1, u_2, \dots, u_n$  as  $u_{(1)} \leq u_{(2)} \leq \dots \leq u_{(n)}$ , then we have

$$\text{proportion of } U_i\text{'s} > u_{(j)} \approx e^{-u_{(j)}}.$$

But on the other hand, we have

$$\text{proportion of } U_i\text{'s } > u_{(j)} \approx \frac{n+1-j}{n+1}.$$

Thus, in order to check the validity of the distribution of  $Y$  we plot

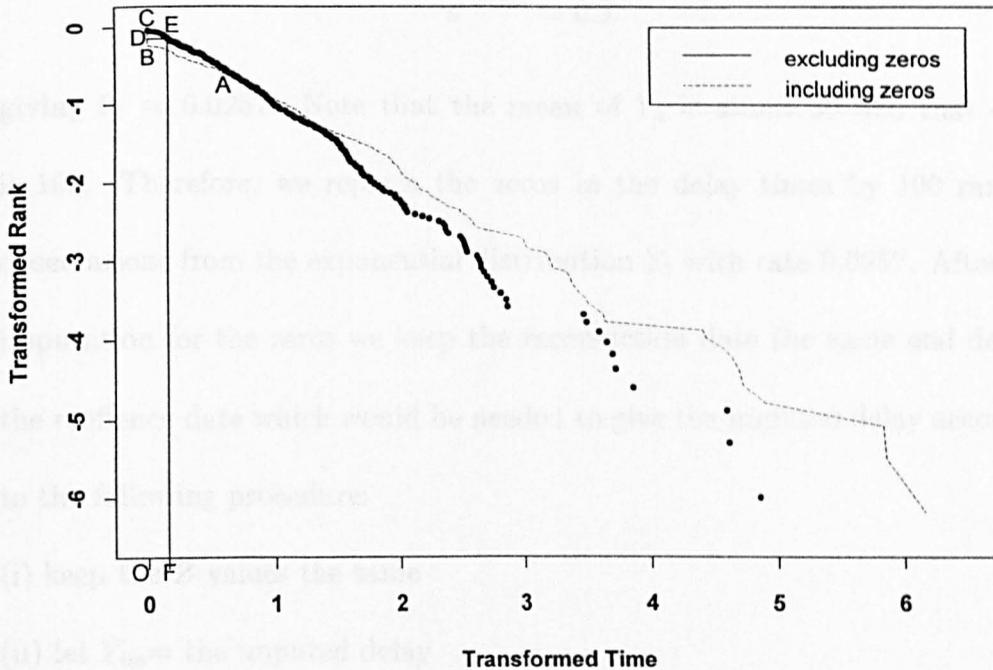
$$\log\left(\frac{n+1-j}{n+1}\right) \text{ against } u_{(j)}, \text{ for } j = 1, \dots, n.$$

If  $Y$  has the exponential distribution, then the plot should resemble a straight line. We now apply this diagnostic plot in the cases of including and excluding the zeros.

Let the delay including the zeros be  $Y_1$  with rate  $\theta_1$  and denote the delay excluding the zeros by  $Y_2$  with rate  $\theta_2$ . The superimposed diagnostic plots for  $Y_1$  and  $Y_2$  are illustrated in Figure 3.4, from which we see that the plot corresponding to  $Y_2$  is fairly close to a straight line, suggesting the validity of an exponential distribution for the delay excluding the zeros. The deviations from linearity in the diagnostic plot of  $Y_2$  are not surprising: the shallower start to the plot corresponds to the fact that some genuine zeros (or short delay times) will have been removed, and the steeper right tail corresponds to the fact that only very unusually large delay times are likely to be censored. The plot for  $Y_1$  is clearly not linear on that part corresponding to the zeros, implying that the exponential distribution for the delay including the zeros fails. However, the plot for  $Y_1$  is fairly linear on the other part corresponding to the nonzero values of  $Y_1$ .

Now suppose we had an exponential distribution  $Y_3$  with the property that if we replaced the zeros in the delay times by a sample of random observations

Fig. 3.4 Diagnostics for Delay Time with Actual Data



from  $Y_3$ , while keeping the nonzero values of the delay times the same, then the segment  $ABC$  shown in Figure 3.4 would be shifted to the segment  $AC$ . This suggests that if we can find such a distribution  $Y_3$ , then the diagnostic plot for the overall data consisting of the original nonzero values supplemented by these imputed values, will be more or less similar to that of  $Y_2$ . To do this we proceed as follows.

We keep the nonzero values of the delay times the same, but replace the zeros by a random sample of observations from an exponential distribution  $Y_3$  with median  $m_3$  and mean  $1/\theta_3$ , where

$$\theta_2 m_3 \approx 0.16$$

as seen from Figure 3.4 ( $OF \approx 0.16$ ). This gives  $m_3 \approx 27$ , since  $\theta_2 = e^{-5.13}$ .

Also, we have

$$e^{-\theta_3 m_3} = 0.5$$

giving  $\theta_3 \approx 0.0257$ . Note that the mean of  $Y_3$  is about 39 and that of  $Y_2$  is 169. Therefore, we replace the zeros in the delay times by 100 random observations from the exponential distribution  $Y_3$  with rate 0.0257. After this imputation for the zeros we keep the reconviction date the same and deduce the reoffence date which would be needed to give the imputed delay according to the following procedure:

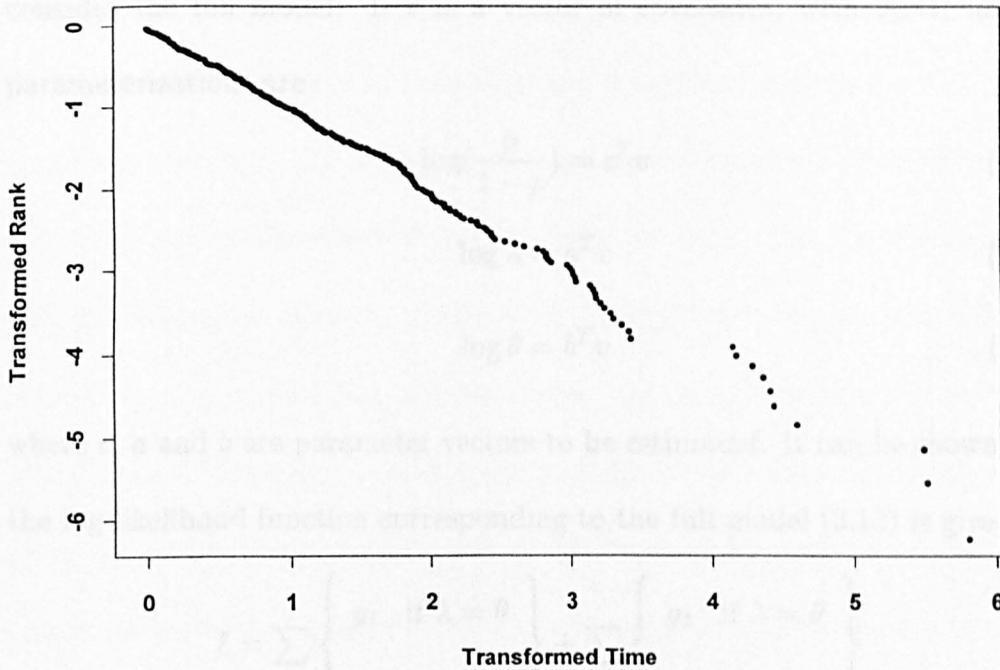
- (i) keep the  $Z$  values the same
- (ii) let  $Y_{im}$  = the imputed delay
- (iii) let  $X_{im} = Z - Y_{im}$  = the imputed reoffence.

Note that it might be possible to have a few negative values for  $X_{im}$  after this imputation process, but this happened in only two cases. We therefore define

$$X_{im} = \max(0, Z - Y_{im}).$$

From now on we will assume throughout that the data under consideration are the imputed data and for these data the diagnostic plot corresponding to the delay is illustrated in Figure 3.5 which is fairly close to a straight line, suggesting an exponential distribution for the delay.

Fig. 3.5 Diagnostics for Delay Time (Imputed Data)



## 3.6 Diagnostics With Imputed Data

### 3.6.1 Goodness of Fit of the Distribution of $Y$

In this section we consider the distribution of  $Y$  for the imputed data by using the method of checking the goodness of fit of the model, described in section 3.4. Applying this method, the following models are fitted to these data.

$$\lambda \text{ and } \theta \text{ constants} \quad (3.9)$$

$$\text{SPM with } p, \lambda \text{ and } \theta \text{ constants—the marginal model} \quad (3.10)$$

$$\lambda \text{ and } \theta \text{ functions of covariates} \quad (3.11)$$

$$\text{SPM with } p \text{ constant but } \lambda \text{ and } \theta \text{ functions of covariates} \quad (3.12)$$

$$\text{SPM with } p, \lambda \text{ and } \theta \text{ functions of covariates—the full model.} \quad (3.13)$$

The models (3.9)–(3.12) are special cases of the full model (3.13). First we consider the full model. If  $v$  is a vector of covariates, with  $v_1=1$ , natural parameterizations are

$$\log\left(\frac{p}{1-p}\right) = c^T v \quad (3.14)$$

$$\log \lambda = a^T v \quad (3.15)$$

$$\log \theta = b^T v \quad (3.16)$$

where  $c$ ,  $a$  and  $b$  are parameter vectors to be estimated. It can be shown that the log-likelihood function corresponding to the full model (3.13) is given by

$$\ell = \sum_u \left\{ \begin{array}{l} g_1 \text{ if } \lambda = \theta \\ g_2 \text{ if } \lambda \neq \theta \end{array} \right\} + \sum_c \left\{ \begin{array}{l} g_3 \text{ if } \lambda = \theta \\ g_4 \text{ if } \lambda \neq \theta \end{array} \right\}$$

where

$$g_1 = c^T v - \log(1 + e^{c^T v}) + 2a^T v - ze^{a^T v}$$

$$g_2 = c^T v - \log(1 + e^{c^T v}) + a^T v - xe^{a^T v} + b^T v - ye^{b^T v}$$

$$g_3 = -\log(1 + e^{c^T v}) + \log\{1 + (1 + te^{a^T v}) e^{(c^T v - te^{a^T v})}\}$$

$$g_4 = -\log(1 + e^{c^T v}) + \log(1 + g_5/g_6)$$

$$g_5 = e^{(c^T v + b^T v - te^{a^T v})} - e^{(c^T v + a^T v - te^{b^T v})}$$

$$g_6 = e^{b^T v} - e^{a^T v},$$

and as before everything in the sums,  $\sum_u$  and  $\sum_c$ , are specific to each observation and the parameters  $p$ ,  $\lambda$  and  $\theta$  are given by substituting the vector of covariates  $v$  into equations (3.14)–(3.16). Here, the first element of the vector  $v$  consists of one, so that the first components of the vectors  $c$ ,  $a$  and  $b$  are the

intercept terms corresponding to the  $p$ -part,  $\lambda$ -part and  $\theta$ -part of the model respectively.

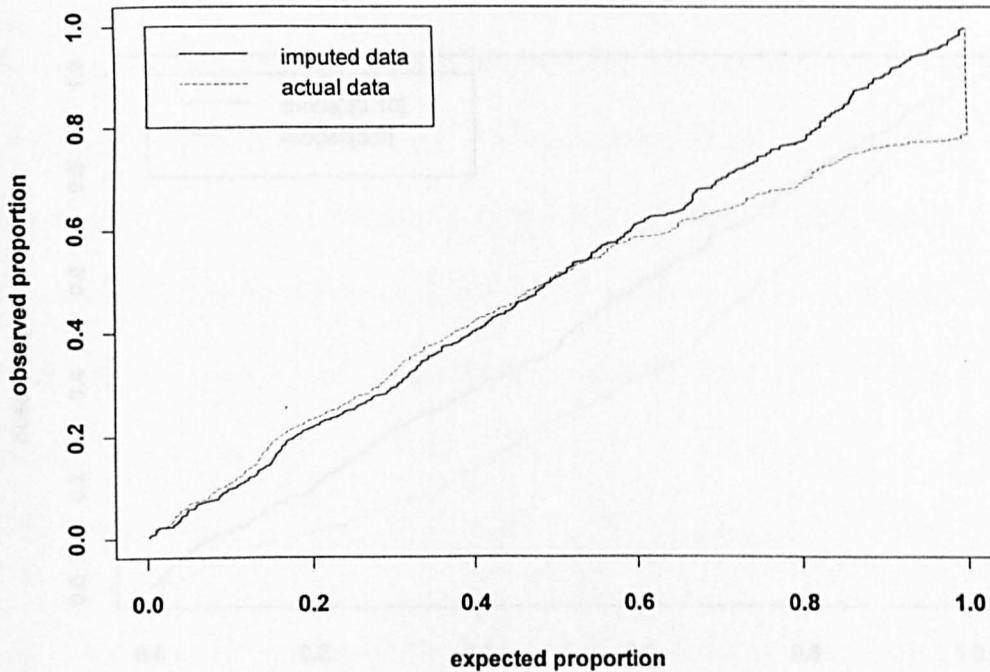
It can be shown that the corresponding diagnostic plots for models (3.9)–(3.13) are all reasonably linear, substantiating the goodness of fit of the distribution of  $Y$  in these models for the imputed data. But, for convenience, only the diagnostic plot for model (3.10) is depicted in Figure 3.6. Also, in order to have a comparison, the diagnostic plot for the same model but with actual data is illustrated in this figure as well. So, the difference between the goodness of fit of the model with actual and imputed data can be monitored by this superimposed diagnostic plots. Note that if we constrain the model (3.13) to have  $\log(\frac{p}{1-p})=c_1$  = the intercept term of  $c^T v$ , that is  $p$  is constant, then this model reduces to the model (3.12). If in the model (3.12) we put  $p = 1$ , then this model reduces to the model (3.11) and if in the model (3.12) we let  $\log \lambda = a_1$  = the intercept term of  $a^T v$ , and  $\log \theta = b_1$  = the intercept term of  $b^T v$ , then this model reduces to the model (3.10). And finally by putting  $p = 1$  in the model (3.10) we can get the model (3.9).

Fitting the model (3.10) to the imputed data, the maximum likelihood estimates of the parameters are found to be  $\hat{p}=0.472$ ,  $\hat{\lambda}=0.00191$ ,  $\hat{\theta}=0.00668$ .

### 3.6.2 Goodness of Fit of the Distribution of $X$

In this section by using  $U = \lambda X$  rather than  $U = \theta Y$  we shall investigate the goodness of fit of the models listed in section 3.6. It can be shown that

Fig. 3.6 Diagnostic Plot Comparing Two Data for Y  
(Marginal Model)



among different diagnostic plots those corresponding to the models (3.10) and (3.13) are reasonably linear, indicating the validity of these two models. The other diagnostic plots are not linear, implying that the models (3.9), (3.11) and (3.12) do not fit the data well. The superimposed diagnostic plots for models (3.9) and (3.10) are pictured in Figure 3.7 and the diagnostic plot corresponding to the model (3.13) is illustrated in Figure 3.8.

Here, for the model (3.9) the estimated values of  $\lambda$  and  $\theta$  are  $\hat{\lambda}=0.0005$ ,  $\hat{\theta}=0.0062$  and for model (3.10) the estimates are  $\hat{p}=0.472$ ,  $\hat{\lambda}=0.00191$ , and  $\hat{\theta}=0.00668$ .

Fig. 3.7 Diagnostic Plot Comparing Two Models for X

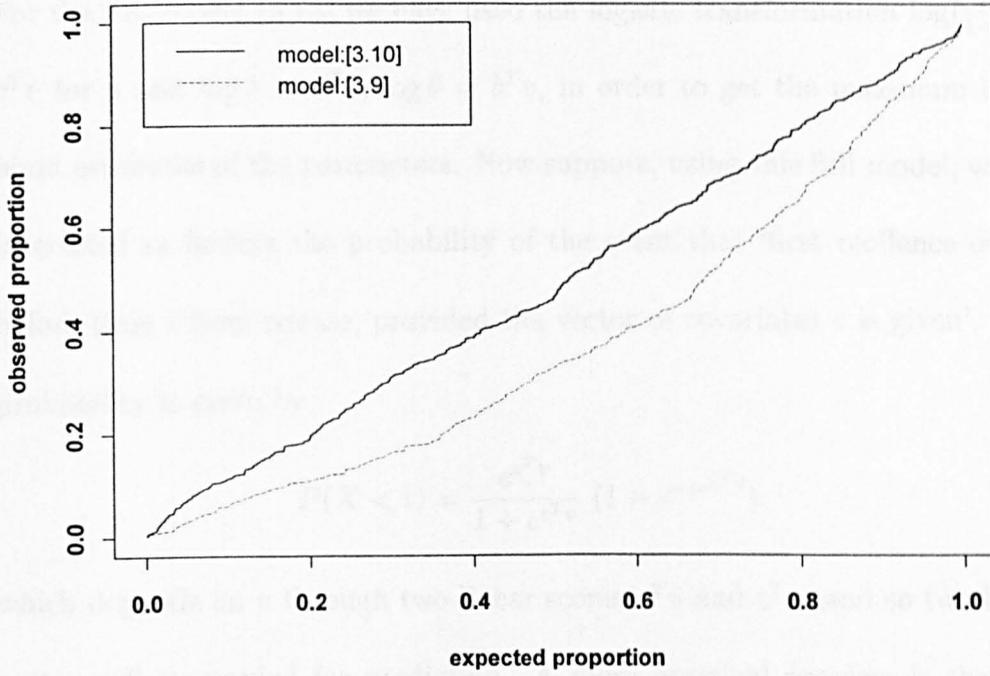
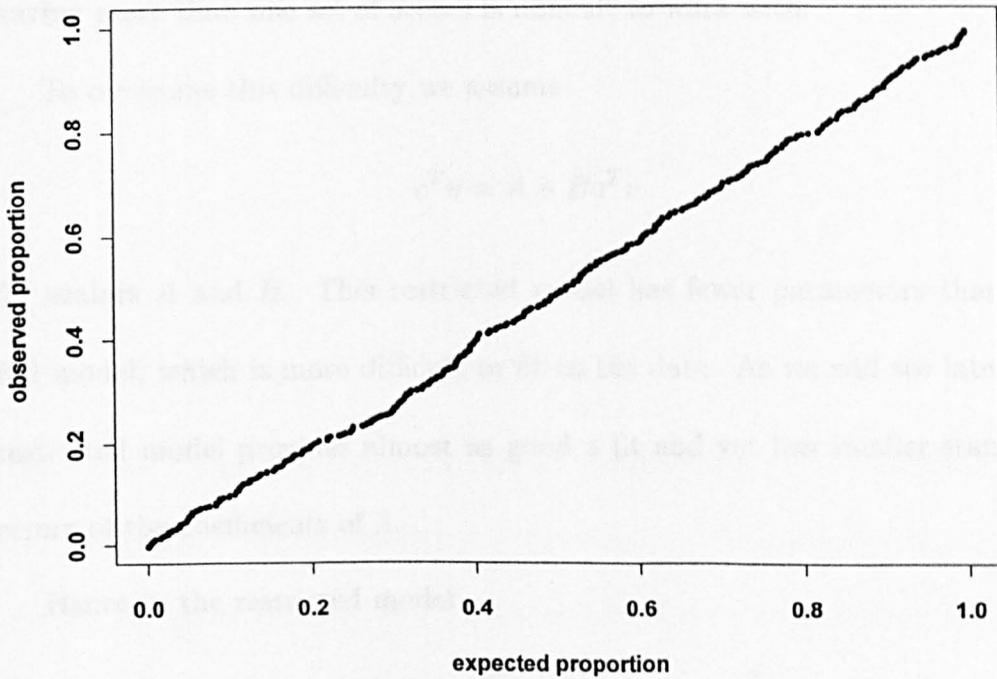


Fig. 3.8 Diagnostic Plot for X (Full Model)



### 3.7 Diagnostics for $X$ in Restricted Model

For the full model (3.13) we have used the logistic transformation  $\log\left(\frac{p}{1-p}\right) = c^T v$  for  $p$  and  $\log \lambda = a^T v$ ,  $\log \theta = b^T v$ , in order to get the maximum likelihood estimates of the parameters. Now suppose, using this full model, we are interested in finding the probability of the event that 'first reoffence occurs before time  $t$  from release, provided the vector of covariates  $v$  is given'. This probability is given by

$$P(X < t) = \frac{e^{c^T v}}{1 + e^{c^T v}} (1 - e^{-te^{a^T v}})$$

which depends on  $v$  through two linear scores  $a^T v$  and  $c^T v$ , and so two linear scores will be needed for prediction. A more practical problem is that the survival function of  $X$ , which is 1 minus this probability, also depends on  $v$  through the two linear scores  $a^T v$  and  $c^T v$ . In such practical applications having more than one set of scores is difficult to work with.

To overcome this difficulty we assume

$$c^T v = A + B a^T v$$

for scalars  $A$  and  $B$ . This restricted model has fewer parameters than the full model, which is more difficult to fit to the data. As we will see later the restricted model provides almost as good a fit and yet has smaller standard errors of the coefficients of  $\lambda$ .

Hence in the restricted model

$$P(X < t) = \frac{e^{A + B a^T v}}{1 + e^{A + B a^T v}} (1 - e^{-te^{a^T v}})$$

for which there is only one set of scores  $a^T v$ , and the survival function of  $X$  now depends on only this single risk score. Therefore, we define a model:

$$\log\left(\frac{p}{1-p}\right) = A + Ba^T v, \log \lambda = a^T v, \log \theta = b^T v \quad (3.17)$$

with parameters  $(A, B, a, b)$ . This model, which we call the restricted model, is a special case of the full model (3.13). It can be shown that the log-likelihood function of the restricted model is given by

$$\ell = \sum_u \left\{ \begin{array}{l} g_1 \text{ if } \lambda = \theta \\ g_2 \text{ if } \lambda \neq \theta \end{array} \right\} + \sum_c \left\{ \begin{array}{l} g_3 \text{ if } \lambda = \theta \\ g_4 \text{ if } \lambda \neq \theta \end{array} \right\} \quad (3.18)$$

where

$$g_1 = A + Ba^T v - \log(1 + e^{A+Ba^T v}) + 2a^T v - ze^{a^T v}$$

$$g_2 = A + Ba^T v - \log(1 + e^{A+Ba^T v}) + a^T v - xe^{a^T v} + b^T v - ye^{b^T v}$$

$$g_3 = -\log(1 + e^{A+Ba^T v}) + \log\{1 + (1 + te^{a^T v}) e^{(A+Ba^T v - te^{a^T v})}\}$$

$$g_4 = -\log(1 + e^{A+Ba^T v}) + \log\left(1 + \frac{h_1 - h_2}{h_3}\right)$$

and

$$h_1 = e^{(A+Ba^T v + b^T v - te^{a^T v})}$$

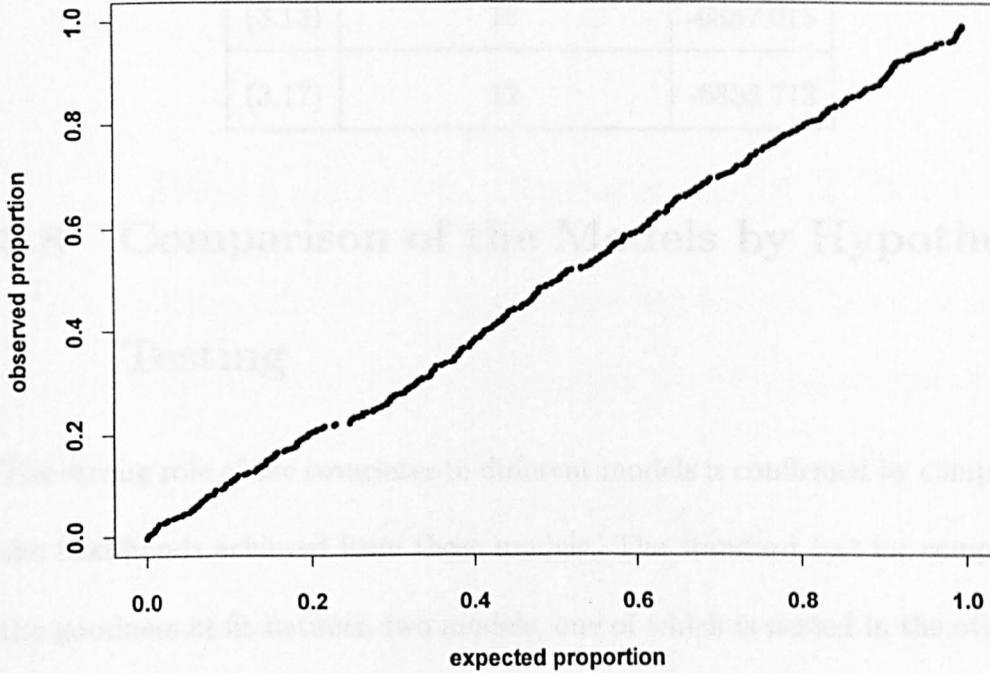
$$h_2 = e^{(A+Ba^T v + a^T v - te^{b^T v})}$$

$$h_3 = e^{b^T v} - e^{a^T v}.$$

The diagnostic plot corresponding to the restricted model is depicted in Figure 3.9 which is reasonably linear, substantiating the goodness of fit of the model. By considering several statistical checks and in particular the diagnostic plots, which play the key role in the development of a model selection

methodology, we shall see that this model provides a good description of the data. The restricted model consists of 12 parameters which is simpler than the full model (3.13), so it is of great practical interest.

Fig. 3.9 Diagnostic Plot for X (Restricted Model)



The maximum likelihood estimates of the parameters under this model are

$$\hat{A} = 20.203, \hat{B} = 3.111, \hat{a}_1 = -6.217, \hat{a}_2 = -0.020$$

$$\hat{a}_3 = 0.041, \hat{a}_4 = 0.018, \hat{a}_5 = 0.058, \hat{b}_1 = -4.580$$

$$\hat{b}_2 = -0.015, \hat{b}_3 = -0.030, \hat{b}_4 = 0.008, \hat{b}_5 = -0.041.$$

Note that in the  $\lambda$ -part of the model,  $a_1$  is the constant term and  $a_2, a_3, a_4$  and  $a_5$  are the coefficients of the covariates *age*, *ac*, *pre* and *jc* respectively. Also in the  $\theta$ -part of the model,  $b_1$  is the constant term and  $b_2, b_3, b_4$  and  $b_5$  are the coefficients of the same covariates.

Table 3.1: Estimates of log-likelihood functions

model	number of parameters	estimate
(3.10)	3	-7016.729
(3.13)	15	-6857.015
(3.17)	12	-6858.713

### 3.8 Comparison of the Models by Hypothesis

#### Testing

The strong role of the covariates in different models is confirmed by comparing the likelihoods achieved from these models. The standard test for comparing the goodness of fit between two models, one of which is nested in the other, is to use hypothesis testing based on the log-likelihood ratio statistic. In practical applications we estimate  $\Delta D$ , the log-likelihood ratio statistic, and compare the estimated value with the appropriate chi-squared distribution. We shall now apply this method of comparison to the models (3.10), (3.13) and (3.17) by using the calculations summarized in Table 3.1.

Using the models (3.10) and (3.13),  $\Delta D = 319.428$  and using the models (3.10) and (3.17),  $\Delta D = 316.032$ . In either case the value of  $\Delta D$  is very significant when compared with the corresponding  $\chi_{12}^2$  and  $\chi_9^2$  distributions. This indicates that each of the models (3.13) and (3.17) provides a significantly better description of the data than model (3.10), suggesting that the covariates

are strongly predictive of reoffending. Using the full model (3.13) and the restricted model (3.17),  $\Delta D = 3.396$  which is not statistically significant when compared with the  $\chi_3^2$  distribution. Thus, the data do not provide evidence against choosing the restricted model (3.17) and we have reason for preferring the simpler model (3.17).

### 3.9 Diagnostics for Reconviction Time ( $Z$ )

To get a diagnostic plot for visually inspecting the distribution of  $Z$  in our model, described in section 3.1, we consider the following conditional probability

$$P(Z > z|Z < t) = \frac{P(z < Z < t)}{P(Z < t)}, \quad 0 \leq z < t.$$

Using equations (3.1) and (3.2) the probability distribution function of  $Z$  is given by

$$F_Z(\lambda, \theta, z) = P(Z < z) = \begin{cases} 1 - (1 + \lambda z)e^{-\lambda z} & \text{if } \lambda = \theta \\ 1 - \frac{\theta e^{-\lambda z} - \lambda e^{-\theta z}}{\theta - \lambda} & \text{if } \lambda \neq \theta \end{cases} \quad (3.19)$$

which leads to

$$P(Z > z|Z < t) = \begin{cases} \frac{(1 + \lambda z)e^{-\lambda z} - (1 + \lambda t)e^{-\lambda t}}{1 - (1 + \lambda t)e^{-\lambda t}} & \text{if } \lambda = \theta \\ \frac{\theta(e^{-\lambda z} - e^{-\lambda t}) - \lambda(e^{-\theta z} - e^{-\theta t})}{\theta(1 - e^{-\lambda t}) - \lambda(1 - e^{-\theta t})} & \text{if } \lambda \neq \theta. \end{cases}$$

Hence, for the uncensored subjects we have

$$E(\text{proportion of } Z_i\text{'s} > z) = \frac{1}{n} \sum_{i=1}^n P(Z_i > z|Z_i < t_i)$$

where  $n$  is the number of uncensored observations. Therefore, we expect that

$$\text{proportion of } Z_i\text{'s } > z \approx \frac{1}{n} \sum_{i=1}^n P(Z_i > z | Z_i < t_i).$$

Now order  $z_1, \dots, z_n$  as  $z_{(1)} \leq z_{(2)} \leq \dots \leq z_{(n)}$ , then

$$\text{proportion of } Z_i\text{'s } > z_{(j)} \approx \frac{n+1-j}{n+1}$$

consequently

$$\frac{n+1-j}{n+1} \approx \frac{1}{n} \sum_{i=1}^n P(Z_i > z_{(j)} | Z_i < t_i).$$

Thus, in order to check the validity of the distribution of  $Z$  in the model we require to plot

$$\frac{n+1-j}{n+1} \text{ against } \frac{1}{n} \sum_{i=1}^n P(Z_i > z_{(j)} | Z_i < t_i)$$

for  $j = 1, \dots, n$  and examine the straightness of the outcome. If the model holds, the plot should resemble a straight line. We shall now apply this method to the models (3.13) and (3.17) to see if  $Z$  has the probability distribution function defined by equation (3.19). The diagnostic plots corresponding to these models are illustrated in Figures 3.10 and 3.11 respectively which are reasonably linear, substantiating the validity of the distribution of  $Z$  in these two models.

Fig. 3.10 Diagnostic Plot for Z (Full Model)

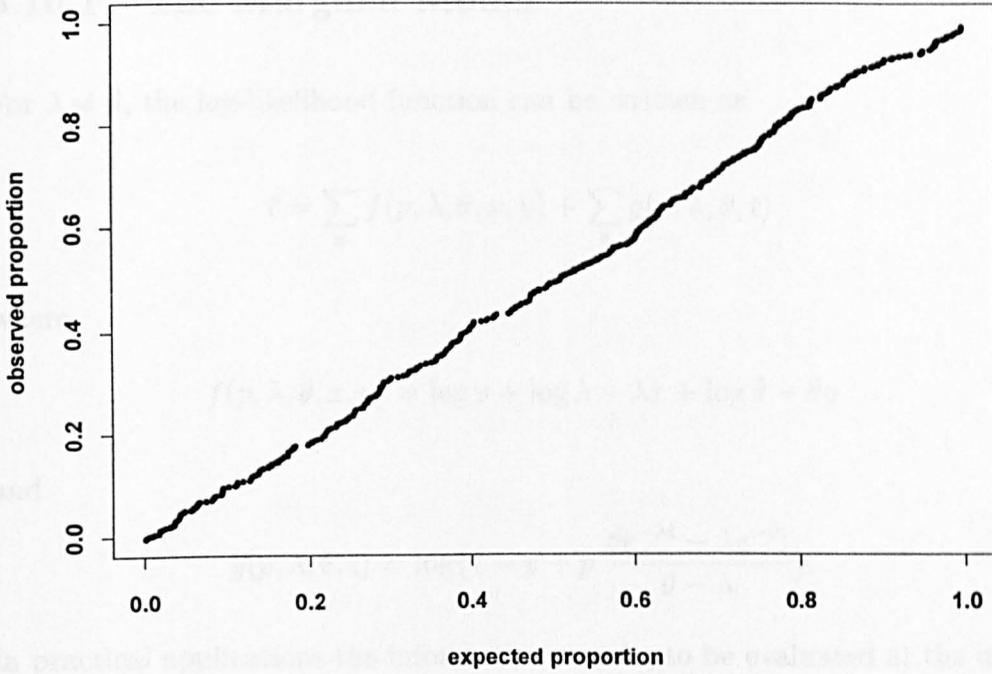
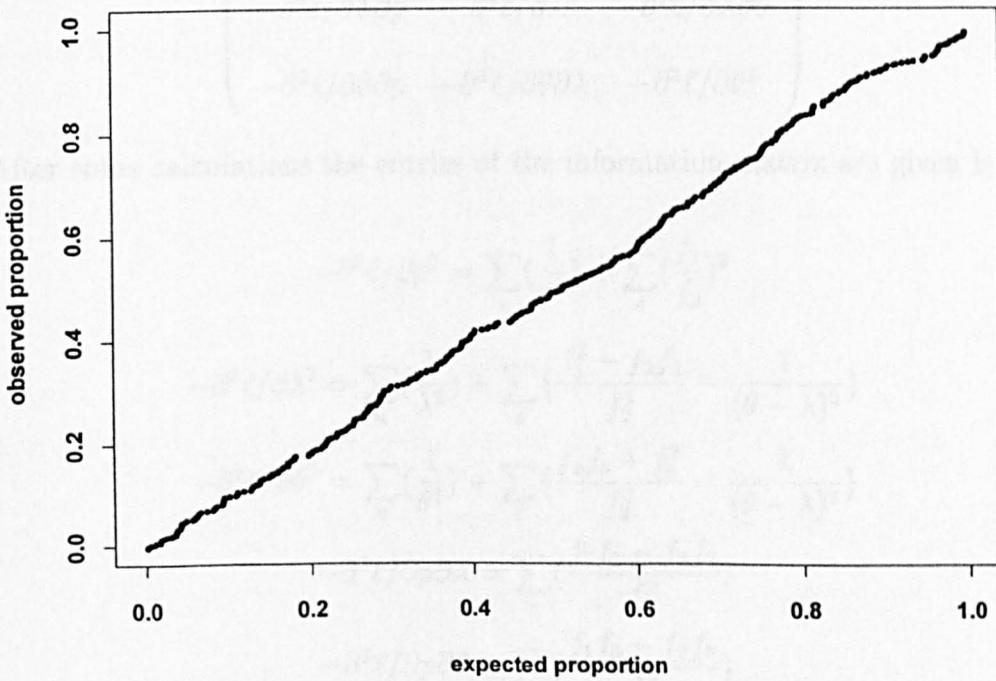


Fig. 3.11 Diagnostic Plot for Z (Restricted Model)



## 3.10 Standard Errors

### 3.10.1 The Marginal Model

For  $\lambda \neq \theta$ , the log-likelihood function can be written as

$$\ell = \sum_{\mathbf{u}} f(p, \lambda, \theta, x, y) + \sum_{\mathbf{c}} g(p, \lambda, \theta, t)$$

where

$$f(p, \lambda, \theta, x, y) = \log p + \log \lambda - \lambda x + \log \theta - \theta y$$

and

$$g(p, \lambda, \theta, t) = \log\left\{1 - p + p \frac{\theta e^{-\lambda t} - \lambda e^{-\theta t}}{\theta - \lambda}\right\}.$$

In practical applications the information matrix, to be evaluated at the maximum likelihood estimates of the parameters, is given by the symmetric matrix

$$\begin{pmatrix} -\partial^2 \ell / \partial p^2 & -\partial^2 \ell / \partial p \partial \lambda & -\partial^2 \ell / \partial p \partial \theta \\ -\partial^2 \ell / \partial \lambda \partial p & -\partial^2 \ell / \partial \lambda^2 & -\partial^2 \ell / \partial \lambda \partial \theta \\ -\partial^2 \ell / \partial \theta \partial p & -\partial^2 \ell / \partial \theta \partial \lambda & -\partial^2 \ell / \partial \theta^2 \end{pmatrix}.$$

After some calculations the entries of the information matrix are given by

$$\begin{aligned} -\partial^2 \ell / \partial p^2 &= \sum_{\mathbf{u}} \left(\frac{1}{p^2}\right) + \sum_{\mathbf{c}} \left(\frac{f_1}{f_2}\right)^2 \\ -\partial^2 \ell / \partial \lambda^2 &= \sum_{\mathbf{u}} \left(\frac{1}{\lambda^2}\right) + \sum_{\mathbf{c}} \left\{ \frac{f_3^2 - f_2 f_4}{f_2^2} - \frac{1}{(\theta - \lambda)^2} \right\} \\ -\partial^2 \ell / \partial \theta^2 &= \sum_{\mathbf{u}} \left(\frac{1}{\theta^2}\right) + \sum_{\mathbf{c}} \left\{ \frac{f_2 f_5 + f_6^2}{f_2^2} - \frac{1}{(\theta - \lambda)^2} \right\} \\ -\partial^2 \ell / \partial p \partial \lambda &= \sum_{\mathbf{c}} \left(\frac{f_1 f_3 - f_2 f_7}{f_2^2}\right) \\ -\partial^2 \ell / \partial p \partial \theta &= \sum_{\mathbf{c}} \left(\frac{f_1 f_6 - f_2 f_8}{f_2^2}\right) \end{aligned}$$

$$-\partial^2 \ell / \partial \lambda \partial \theta = \sum_c \left\{ \frac{f_3 f_6 - f_2 f_9}{f_2^2} + \frac{1}{(\theta - \lambda)^2} \right\}$$

where

$$f_1(p, \lambda, \theta, t) = \lambda(1 - e^{-\theta t}) - \theta(1 - e^{-\lambda t})$$

$$f_2(p, \lambda, \theta, t) = (\theta - \lambda)(1 - p) + p(\theta e^{-\lambda t} - \lambda e^{-\theta t})$$

$$f_3(p, \lambda, \theta, t) = p(1 - \theta t e^{-\lambda t} - e^{-\theta t}) - 1$$

$$f_4(p, \lambda, \theta, t) = p \theta t^2 e^{-\lambda t}$$

$$f_5(p, \lambda, \theta, t) = p \lambda t^2 e^{-\theta t}$$

$$f_6(p, \lambda, \theta, t) = p(\lambda t e^{-\theta t} + e^{-\lambda t} - 1) + 1$$

$$f_7(p, \lambda, \theta, t) = 1 - \theta t e^{-\lambda t} - e^{-\theta t}$$

$$f_8(p, \lambda, \theta, t) = \lambda t e^{-\theta t} + e^{-\lambda t} - 1$$

$$f_9(p, \lambda, \theta, t) = p t (e^{-\theta t} - e^{-\lambda t}).$$

We evaluate the entries of the information matrix directly using the package S-PLUS (StatSci, 1992). Consequently the computer programme provides the estimated information matrix

$$\begin{pmatrix} 3718.57 & 263157.69 & 14781.83 \\ 263157.69 & 61640238.80 & -907650.80 \\ 14781.83 & -907650.80 & 9292087.24 \end{pmatrix}$$

and hence the variance-covariance matrix of the parameter estimates

$$\begin{pmatrix} 3.909960 \times 10^{-4} & -1.680837 \times 10^{-6} & -7.861793 \times 10^{-7} \\ -1.680837 \times 10^{-6} & 2.347222 \times 10^{-8} & 4.966636 \times 10^{-9} \\ -7.861793 \times 10^{-7} & 4.966636 \times 10^{-9} & 1.093542 \times 10^{-7} \end{pmatrix}$$

which is the inverse of the information matrix. Therefore, the standard errors of the parameter estimates, which are the square roots of the main diagonal elements of the variance-covariance matrix, are  $se(\hat{p})=0.020$ ,  $se(\hat{\lambda})=0.00015$  and  $se(\hat{\theta})=0.00033$ . Note that for the marginal model (3.10), the maximum likelihood estimates of the parameters are  $\hat{p}=0.472$ ,  $\hat{\lambda}=0.00191$  and  $\hat{\theta}=0.00668$ . From the magnitudes of the standard errors we can assess the accuracy of the parameter estimates and construct confidence intervals for the model parameters, for instance an approximate 95% confidence interval for  $\lambda$  is given by  $\hat{\lambda} \pm 1.96 \times se(\hat{\lambda})$ , i.e. (0.0016, 0.0022). Also, with the same confidence level the confidence intervals corresponding to  $p$  and  $\theta$  are given by  $\hat{p} \pm 1.96 \times se(\hat{p})$  and  $\hat{\theta} \pm 1.96 \times se(\hat{\theta})$ , i.e. (0.43, 0.51) and (0.0060, 0.0073) respectively.

Recall that,

$p =$  split proportion in SPM

$\lambda =$  rate of time from release to first reoffence

$1/\lambda =$  mean time to first reoffence

$\theta =$  rate of delay

$1/\theta =$  average delay.

Therefore, under the marginal model (3.10) these confidence intervals imply that the split proportion  $p$  is in the ranges 0.43 to 0.51 and the mean time to first reoffence (days) and the average delay (days) will be in the intervals (454, 625) and (136, 166) respectively.

### 3.10.2 The Full Model

Our objective in the following two sections is to measure the importance of the covariates included in the model of interest. The usual statistical measure of the importance of a covariate is the statistical significance of the coefficient of that covariate in the model. A conventional procedure is to fit several covariates together, retain those which are statistically significant at some nominal level (such as 5%) and discard those which are not. From the magnitude of each standardized estimate we can measure the importance of the corresponding covariate in the model.

In the full model (3.13) the parameters  $p$ ,  $\lambda$  and  $\theta$  are assumed to be functions of the covariates. Let  $V$  be the covariates matrix which is of dimension  $5 \times 1179$  and with first row consisting of ones, and let  $v_i$  be the  $i$ th column of  $V$  and  $v_{ji}$  be the  $j$ th element of  $v_i$ , then the log-likelihood function is given by

$$\ell = \sum_{i=1}^n f(p_i, \lambda_i, \theta_i, x_i, y_i) + \sum_{i=n+1}^N g(p_i, \lambda_i, \theta_i, t_i) \quad (3.20)$$

where the first sum is over the uncensored cases and the second sum is over the censored cases and

$$f(p_i, \lambda_i, \theta_i, x_i, y_i) = \log p_i + \log \lambda_i - \lambda_i x_i + \log \theta_i - \theta_i y_i$$

$$g(p_i, \lambda_i, \theta_i, t_i) = \log \left\{ 1 - p_i + p_i \frac{\theta_i e^{-\lambda_i t_i} - \lambda_i e^{-\theta_i t_i}}{\theta_i - \lambda_i} \right\}$$

$$\log \left( \frac{p_i}{1 - p_i} \right) = c^T v_i = \sum_{j=1}^5 c_j v_{ji}$$

$$\log \lambda_i = a^T v_i = \sum_{j=1}^5 a_j v_{ji}$$

$$\log \theta_i = b^T v_i = \sum_{j=1}^5 b_j v_{ji}.$$

To obtain standard errors we calculate the second derivatives of  $\ell$ , defined by equation (3.20), implicitly, and then estimate the inverse of the information matrix in the usual way. The information matrix is given by

$$\begin{pmatrix} -\partial^2 \ell / \partial c^2 & -\partial^2 \ell / \partial c \partial a & -\partial^2 \ell / \partial c \partial b \\ -\partial^2 \ell / \partial a \partial c & -\partial^2 \ell / \partial a^2 & -\partial^2 \ell / \partial a \partial b \\ -\partial^2 \ell / \partial b \partial c & -\partial^2 \ell / \partial b \partial a & -\partial^2 \ell / \partial b^2 \end{pmatrix}$$

which is a  $15 \times 15$  symmetric matrix. Each element of this information matrix is a  $5 \times 5$  matrix itself which can be determined as follows. For brevity, only the calculations corresponding to the first element,  $-\partial^2 \ell / \partial c^2$ , are presented here.

Using the chain rule for the derivatives of the functions , we get

$$\partial^2 \ell / \partial c_j \partial c_k = \sum_u \phi_1 + \sum_c \phi_2 \quad (3.21)$$

where

$$\phi_1 = (\partial^2 f / \partial p_i^2) (\partial p_i / \partial c_k) (\partial p_i / \partial c_j) + (\partial f / \partial p_i) (\partial^2 p_i / \partial c_j \partial c_k)$$

$$\phi_2 = (\partial^2 g / \partial p_i^2) (\partial p_i / \partial c_k) (\partial p_i / \partial c_j) + (\partial g / \partial p_i) (\partial^2 p_i / \partial c_j \partial c_k)$$

$$\partial p_i / \partial c_j = p_i (1 - p_i) v_{ji}$$

$$\partial^2 p_i / \partial c_j \partial c_k = p_i (1 - p_i) (1 - 2p_i) v_{ji} v_{ki}$$

or, in vector notation, we have

$$\partial p_i / \partial c = p_i (1 - p_i) v_i$$

$$\partial^2 p_i / \partial c^2 = p_i(1 - p_i)(1 - 2p_i)v_i v_i^T$$

and

$$\partial^2 g / \partial p_i^2 = -(\partial g / \partial p_i)^2$$

$$\partial f_i / \partial p_i = 1/p_i, \quad \partial^2 f / \partial p_i^2 = -1/p_i^2.$$

Therefore, we can write equation (3.21) as

$$-\partial^2 \ell / \partial c_j \partial c_k = \sum_{\mathbf{u}} \psi_1 + \sum_c \psi_2 \quad (3.22)$$

where

$$\psi_1 = p_i(1 - p_i)v_{ji}v_{ki}$$

$$\psi_2 = \{p_i(1 - p_i)(\partial g / \partial p_i) + 2p_i - 1\}(\partial g / \partial p_i)p_i(1 - p_i)v_{ji}v_{ki}$$

and

$$\partial g / \partial p_i = f_{1i} / f_{2i}$$

in which

$$f_{1i} = \lambda_i(1 - e^{-\theta_i t_i}) - \theta_i(1 - e^{-\lambda_i t_i})$$

$$f_{2i} = (\theta_i - \lambda_i)(1 - p_i) + p_i(\theta_i e^{-\lambda_i t_i} - \lambda_i e^{-\theta_i t_i}).$$

Note that equation (3.22) gives the  $(j, k)$ th element of  $(-\partial^2 \ell / \partial c^2)$ . Now, in the  $\sum_{\mathbf{u}}$  part of this equation let

$$d_{1i} = p_i(1 - p_i)$$

and let  $D_1$  be the diagonal matrix with  $i$ th main diagonal element  $d_{1i}$ . For the  $\sum_c$  part, let

$$d_{2i} = \{p_i(1 - p_i)(\partial g / \partial p_i) + 2p_i - 1\}(\partial g / \partial p_i)p_i(1 - p_i)$$

and  $D_2$  be the diagonal matrix with  $i$ th main diagonal element  $d_{2i}$ .

Then equation (3.22) can be written as

$$\begin{aligned}
 -\partial^2 \ell / \partial c_j \partial c_k &= \sum_{i=1}^n d_{1i} v_{ji} v_{ki} + \sum_{i=n+1}^N d_{2i} v_{ji} v_{ki} \\
 &= (j, k\text{th element of } V_1 D_1 V_1^T) + (j, k\text{th element of } V_2 D_2 V_2^T) \\
 &= j, k\text{th element of } (-\partial^2 \ell / \partial c^2)
 \end{aligned}$$

where  $V_1$  is the covariates matrix corresponding to the uncensored cases with dimension  $5 \times 486$  and  $V_2$  is the covariates matrix corresponding to the censored cases with dimension  $5 \times 693$ .

Therefore, it follows that

$$-\partial^2 \ell / \partial c^2 = V_1 D_1 V_1^T + V_2 D_2 V_2^T. \quad (3.23)$$

The other elements of the information matrix can be obtained by a similar procedure. Indeed, a rather long calculation shows that

$$-\partial^2 \ell / \partial c \partial a = V_2 D_3 V_2^T \quad (3.24)$$

$$-\partial^2 \ell / \partial c \partial b = V_2 D_4 V_2^T \quad (3.25)$$

$$-\partial^2 \ell / \partial a^2 = V_1 D_5 V_1^T + V_2 D_6 V_2^T \quad (3.26)$$

$$-\partial^2 \ell / \partial a \partial b = V_2 D_7 V_2^T \quad (3.27)$$

$$-\partial^2 \ell / \partial b^2 = V_1 D_8 V_1^T + V_2 D_9 V_2^T \quad (3.28)$$

where

$D_m =$  the diagonal matrix with  $i$ th main diagonal element  $d_{mi}$

for  $m = 3, \dots, 9$  and  $d_{mi}$ 's are given by

$$d_{3i} = \left( \frac{f_{1i}f_{3i} - f_{2i}f_{4i}}{f_{2i}^2} \right) p_i (1 - p_i) \lambda_i$$

$$d_{4i} = \left( \frac{f_{1i}f_{5i} - f_{2i}f_{6i}}{f_{2i}^2} \right) p_i (1 - p_i) \theta_i$$

$$d_{5i} = \lambda_i x_i$$

$$d_{6i} = \lambda_i^2 r_1 - \lambda_i \left( \frac{f_{3i}}{f_{2i}} + \frac{1}{\theta_i - \lambda_i} \right)$$

$$d_{7i} = \left\{ \frac{f_{3i}f_{5i} - f_{2i}f_{8i}}{f_{2i}^2} + \frac{1}{(\theta_i - \lambda_i)^2} \right\} \lambda_i \theta_i$$

$$d_{8i} = \theta_i y_i$$

$$d_{9i} = \theta_i^2 r_2 + \theta_i \left( \frac{1}{\theta_i - \lambda_i} - \frac{f_{3i}}{f_{2i}} \right)$$

where

$$r_1 = \frac{f_{3i}^2 - f_{2i}f_{7i}}{f_{2i}^2} - \frac{1}{(\theta_i - \lambda_i)^2}$$

$$r_2 = \frac{f_{2i}f_{9i} + f_{5i}^2}{f_{2i}^2} - \frac{1}{(\theta_i - \lambda_i)^2}$$

Here, the functions  $f_{1i}, \dots, f_{9i}$  are defined by

$$f_{1i}(p_i, \lambda_i, \theta_i, t_i) = \lambda_i(1 - e^{-\theta_i t_i}) - \theta_i(1 - e^{-\lambda_i t_i})$$

$$f_{2i}(p_i, \lambda_i, \theta_i, t_i) = (\theta_i - \lambda_i)(1 - p_i) + p_i(\theta_i e^{-\lambda_i t_i} - \lambda_i e^{-\theta_i t_i})$$

$$f_{3i}(p_i, \lambda_i, \theta_i, t_i) = p_i(1 - \theta_i t_i e^{-\lambda_i t_i} - e^{-\theta_i t_i}) - 1$$

$$f_{4i}(p_i, \lambda_i, \theta_i, t_i) = 1 - \theta_i t_i e^{-\lambda_i t_i} - e^{-\theta_i t_i}$$

$$f_{5i}(p_i, \lambda_i, \theta_i, t_i) = p_i(\lambda_i t_i e^{-\theta_i t_i} + e^{-\lambda_i t_i} - 1) + 1$$

$$f_{6i}(p_i, \lambda_i, \theta_i, t_i) = \lambda_i t_i e^{-\theta_i t_i} + e^{-\lambda_i t_i} - 1$$

$$f_{7i}(p_i, \lambda_i, \theta_i, t_i) = p_i \theta_i t_i^2 e^{-\lambda_i t_i}$$

Table 3.2: Standard errors for  $p$ -part of full model

covariate	parameter	standard error	standardized estimate
<i>age</i>	$c_2$	0.016	-3.944
<i>ac</i>	$c_3$	0.078	0.784
<i>pre</i>	$c_4$	0.019	4.090
<i>jc</i>	$c_5$	0.086	0.915

$$f_{8i}(p_i, \lambda_i, \theta_i, t_i) = p_i t_i (e^{-\theta_i t_i} - e^{-\lambda_i t_i})$$

$$f_{9i}(p_i, \lambda_i, \theta_i, t_i) = p_i \lambda_i t_i^2 e^{-\theta_i t_i}.$$

Analogous to what we did in subsection 3.10.1, using the estimated information matrix and the variance-covariance matrix, the standard errors of the parameter estimates and consequently the standardized estimates of the parameters can be obtained. The summaries of these analyses corresponding to  $p$ -part,  $\lambda$ -part and  $\theta$ -part of the full model (3.13), are illustrated in Tables 3.2, 3.3, and 3.4 respectively.

In the  $p$ -part of the model the standardized estimates corresponding to the covariates *age* and *pre* are -3.944 and 4.090 respectively, which are large numbers, implying that in this part of the model these two covariates are statistically significant. In the  $\lambda$ -part of the model, the covariates *ac* and *jc* are moderately statistically significant because their corresponding standardized estimates are 1.899 and 2.729 respectively. In the  $\theta$ -part of the model none

Table 3.3: Standard errors for  $\lambda$ -part of full model

covariate	parameter	standard error	standardized estimate
<i>age</i>	$a_2$	0.016	-1.288
<i>ac</i>	$a_3$	0.045	1.899
<i>pre</i>	$a_4$	0.008	0.624
<i>jc</i>	$a_5$	0.045	2.729

Table 3.4: Standard errors for  $\theta$ -part of full model

covariate	parameter	standard error	standardized estimate
<i>age</i>	$b_2$	0.009	-1.689
<i>ac</i>	$b_3$	0.037	-0.714
<i>pre</i>	$b_4$	0.006	1.062
<i>jc</i>	$b_5$	0.037	-1.014

of the covariates is statistically significant because their corresponding standardized estimates are small, so if an individual covariate is deleted from this part of the model, then the resulting model would not be worse as compared to the original model. Note that over the whole model all the covariates seem to be statistically significant, two of them in the  $p$ -part and two others in the  $\lambda$ -part of the model.

### 3.10.3 The Restricted Model

Let  $\ell^*(A, B, a, b)$  and  $\ell(c, a, b)$  be the log-likelihood functions of the restricted model (3.17) and the full model (3.13) respectively. Then, we have

$$\ell^*(A, B, a, b) = \ell(c, a, b)$$

where

$$c = Ae + Ba,$$

for scalars  $A, B$  and  $e$  is the unit vector  $(1, 0, 0, 0, 0)^T$ . Here the vector  $e$  is to permit different constant terms for the two linear scores  $\log \lambda$  and  $\log \theta$ . The information matrix for the restricted model (3.17) can be represented by

$$\begin{pmatrix} M_1 & M_2 & M_3 \\ M_4 & M_5 & M_6 \\ M_7 & M_8 & M_9 \end{pmatrix}$$

where

$$M_1 = \begin{pmatrix} -\partial^2 \ell^* / \partial A^2 & -\partial^2 \ell^* / \partial A \partial B \\ -\partial^2 \ell^* / \partial B \partial A & -\partial^2 \ell^* / \partial B^2 \end{pmatrix}, \quad M_2 = \begin{pmatrix} -\partial^2 \ell^* / \partial A \partial a \\ -\partial^2 \ell^* / \partial B \partial a \end{pmatrix}$$

$$M_3 = \begin{pmatrix} -\partial^2 \ell^* / \partial A \partial b \\ -\partial^2 \ell^* / \partial B \partial b \end{pmatrix}, \quad M_4 = \begin{pmatrix} -\partial^2 \ell^* / \partial a \partial A & -\partial^2 \ell^* / \partial a \partial B \end{pmatrix}$$

$$M_5 = -\partial^2 \ell^* / \partial a^2, \quad M_6 = -\partial^2 \ell^* / \partial a \partial b$$

$$M_7 = \begin{pmatrix} -\partial^2 \ell^* / \partial b \partial A & -\partial^2 \ell^* / \partial b \partial B \end{pmatrix}$$

$$M_8 = -\partial^2 \ell^* / \partial b \partial a, \quad M_9 = -\partial^2 \ell^* / \partial b^2.$$

After some calculations the entries of the matrices  $M_1, \dots, M_9$  are given by

$$(-\partial^2 \ell^* / \partial A^2) = e^T (-\partial^2 \ell / \partial c^2) e$$

$$(-\partial^2 \ell^* / \partial B^2) = a^T (-\partial^2 \ell / \partial c^2) a$$

$$(-\partial^2 \ell^* / \partial A \partial B) = (-\partial^2 \ell^* / \partial B \partial A) = e^T (-\partial^2 \ell / \partial c^2) a$$

$$(-\partial^2 \ell^* / \partial A \partial a) = e^T (-\partial^2 \ell / \partial c \partial a)$$

$$(-\partial^2 \ell^* / \partial B \partial a) = a^T (-\partial^2 \ell / \partial c \partial a)$$

$$(-\partial^2 \ell^* / \partial A \partial b) = e^T (-\partial^2 \ell / \partial c \partial b)$$

$$(-\partial^2 \ell^* / \partial B \partial b) = a^T (-\partial^2 \ell / \partial c \partial b)$$

$$(-\partial^2 \ell^* / \partial a \partial A) = (-\partial^2 \ell^* / \partial A \partial a)^T$$

$$(-\partial^2 \ell^* / \partial a \partial B) = (-\partial^2 \ell^* / \partial B \partial a)^T$$

$$(-\partial^2 \ell^* / \partial a^2) = (-\partial^2 \ell / \partial c \partial a) B + (-\partial^2 \ell / \partial a^2)$$

$$(-\partial^2 \ell^* / \partial a \partial b) = (-\partial^2 \ell / \partial c \partial b) B + (-\partial^2 \ell / \partial a \partial b)$$

$$(-\partial^2 \ell^* / \partial b \partial A) = (-\partial^2 \ell^* / \partial A \partial b)^T$$

$$(-\partial^2 \ell^* / \partial b \partial B) = (-\partial^2 \ell^* / \partial B \partial b)^T$$

Table 3.5: Standard errors under restricted model

covariate	parameter	standard error	standardized estimate
<i>age</i>	$a_2$	0.006	-3.082
<i>ac</i>	$a_3$	0.028	1.464
<i>pre</i>	$a_4$	0.006	3.092
<i>jc</i>	$a_5$	0.033	1.747
<i>age</i>	$b_2$	0.009	-1.703
<i>ac</i>	$b_3$	0.037	-0.816
<i>pre</i>	$b_4$	0.006	1.196
<i>jc</i>	$b_5$	0.037	-1.111

$$(-\partial^2 \ell^* / \partial b \partial a) = (-\partial^2 \ell / \partial b \partial a)$$

$$(-\partial^2 \ell^* / \partial b^2) = (-\partial^2 \ell / \partial b^2).$$

Now using equations (3.23)–(3.28), the matrices  $M_1, \dots, M_9$  will be obtained. Having these matrices, we can get the estimated information matrix and the corresponding variance-covariance matrix. Using these information, the standard errors of the parameter estimates and consequently the standardized estimates of the parameters will be available. Table 3.5 sets out the summary of this analysis for the restricted model (3.17) with at least three years of follow-up. On the basis of the magnitudes of the standardized estimates we observe that in the  $\lambda$ -part of the restricted model the covariates *age* and *pre*

are statistically significant, but the covariates  $ac$  and  $jc$  are not. Note that in the  $\lambda$ -part of the full model (3.13) the covariates  $age$  and  $pre$  are not statistically significant but the covariate  $ac$  and  $jc$  are moderately significant. Also, from Tables 3.3 and 3.5 we see that the restricted model has smaller standard errors of the coefficients of  $\lambda$  as compared with the full model. In the  $\theta$ -part of the restricted model none of the covariates is statistically significant, implying that if an individual covariate is discarded from this part of the model, then the subsequent model will still have the same goodness of fit as the original model. But, if more than one covariate, for instance two covariates, are deleted from this part of the model, then we can not anticipate what will happen without considering the value of the deviance created by this alteration of the covariates. Here, the standardized estimate of the parameter  $B$  is 10.328 which is a large number, indicating that if we are looking for a good model of this type, then the  $p$ -part of the model must be a function of the covariates rather than being a constant. In other words, if we constrain the restricted model to have  $B=0$ , that is  $p$  equals constant, then the subsequent model will not fit the data well, as we have already mentioned in the subsection 3.6.1.

### 3.11 Comparison of the Score Coefficients

Table 3.6 sets out the score coefficients for the actual and imputed data, discussed in section 3.5, under the restricted model (3.17). Note that the parameters  $a_2, \dots, a_5$  corresponds to the  $\lambda$ -part and  $b_2, \dots, b_5$  are for the  $\theta$ -part

Table 3.6: Score coefficients for actual and imputed data

covariate	parameter	actual data	imputed data
<i>age</i>	$a_2$	-0.021	-0.020
<i>ac</i>	$a_3$	0.041	0.041
<i>pre</i>	$a_4$	0.018	0.018
<i>jc</i>	$a_5$	0.058	0.058
<i>age</i>	$b_2$	-0.016	-0.015
<i>ac</i>	$b_3$	-0.028	-0.030
<i>pre</i>	$b_4$	0.006	0.008
<i>jc</i>	$b_5$	-0.040	-0.041

of the model. From this table we see that in the  $\lambda$ -part there is in fact no difference between the corresponding score coefficients but in the  $\theta$ -part there is slightly difference, and so the imputed data makes only a little difference to the fitted models.

# Chapter 4

## Choice of Follow-Up and Fairer Assessment of Risk Scores

### 4.1 Introduction

The models that we have fitted to the data so far are based upon a follow up time of at least 3 years. Now, suppose that the data had only been collected for  $t$  years, for instance,  $t = 1$  year. Then an interesting and imperative question that might be asked at this stage of our statistical analysis is: could we still estimate the indicated models for a new follow up time less than that of the original one and get similar reoffending risk score? If so, then the main advantage of the new scheme will be in ease of data collection by using the most recent data possible and the cost of follow-up will be decreased considerably as well, which is of great financial interest. Fortunately, the answer to the proposed question is positive as shown in the following sections of the present

chapter.

In the data, with at least 3 years of follow-up, we are concerned with the assessment of the risk of reoffending within a *variable* time period  $T$ , as the period of follow-up will vary from one prisoner to another. Now suppose that we are interested in estimating the risk of reoffending within a *fixed* period of time,  $t_0$  say, after release which we assume to be less than 3 years. From now on, for a given fixed follow-up time  $t_0$ , throughout the calculations we pick up those parts of the sample corresponding to the new follow-up time  $t_0$ .

In general, no statistical model will ever perfectly fit a given set of data and there will be some discrepancies, large or small, between the data and their model. How to assess these discrepancies, and how to come to a conclusion about the adequacy of the proposed models with new follow-up times will form the central themes of this chapter.

In the following sections under this new approach we shall consider the model (3.10) and the restricted model (3.17) with 1, 2 and at least 3 years of follow-up (1 yr F/U, 2 yr F/U, full F/U). In both cases, after finding the maximum likelihood estimates of the parameters, we apply several statistical checks for the goodness of fit of the models. As before an informative procedure in this regard is based upon the diagnostic plots, which play an important role in the model validation and will be used repeatedly in later sections.

Moreover, in order to make a fairer assessment of the performance of the risk scores, the split-sample (the training sample and the validation sample) method is discussed and applied to the data.

Table 4.1: Parameter estimates of marginal model

estimate	F/U=1 yr	F/U=2 yr	full F/U
$\hat{p}$	0.990	0.490	0.470
$\hat{\lambda}$	0.001	0.002	0.002
$\hat{\theta}$	0.004	0.006	0.007

We have also carried out some simulation experiments to check the sensitivity of the diagnostic plots to model misspecification.

## 4.2 Estimation of the Marginal Model

The maximum likelihood estimates of the parameters  $p$ ,  $\lambda$  and  $\theta$  under the model (3.10) with 1, 2, and at least 3 years of follow-up are given in Table 4.1. For the model with 1 year of follow-up, the estimate of the split proportion  $p$  is almost equal to 1 and this model tends to be similar to the model (3.9) which does not fit the data well. In other words, considering this aspect of the analysis, parameter estimation, for 1 year of follow-up the model (3.10) does not appear to predict the future data well. But, for 2 and at least 3 years of follow-up the corresponding parameter estimates are very similar under this model.

### 4.3 Diagnostics for $X$ in Marginal Model

The diagnostic plots corresponding to 1, 2 and at least 3 years of follow-up under the model (3.10) are illustrated in Figures 4.1, 4.2 and 4.3 respectively, which are all reasonably linear. Therefore, this aspect of the analysis, i.e. the diagnostic check, suggests that the model with 1, 2 and at least 3 years of follow-up fits the data well. However, in order to reach a definite conclusion about the adequacy of the model, we shall consider further analyses of the data under this model.

Fig. 4.1 Diagnostic Plot for X (Marginal Model:1 yr F/U)

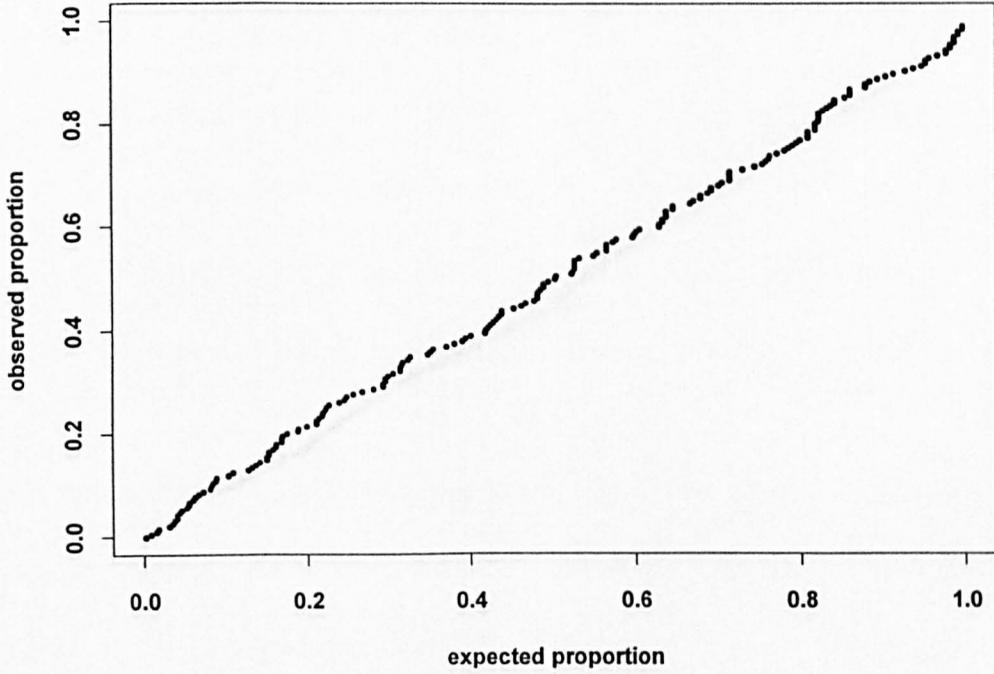


Fig. 4.2 Diagnostic Plot for X (Marginal Model:2 yr F/U)

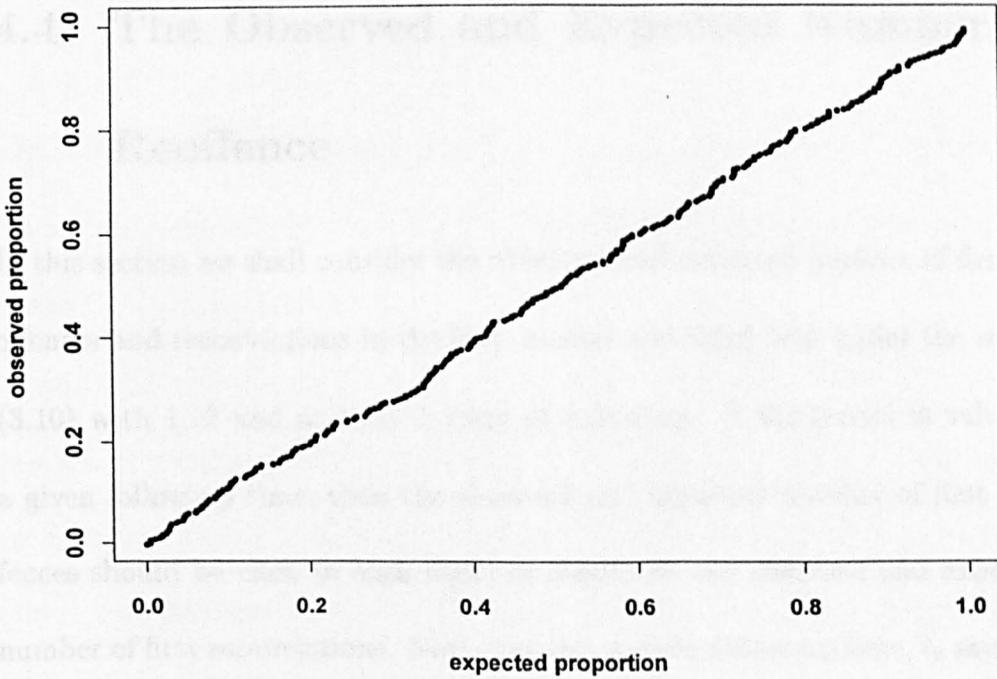
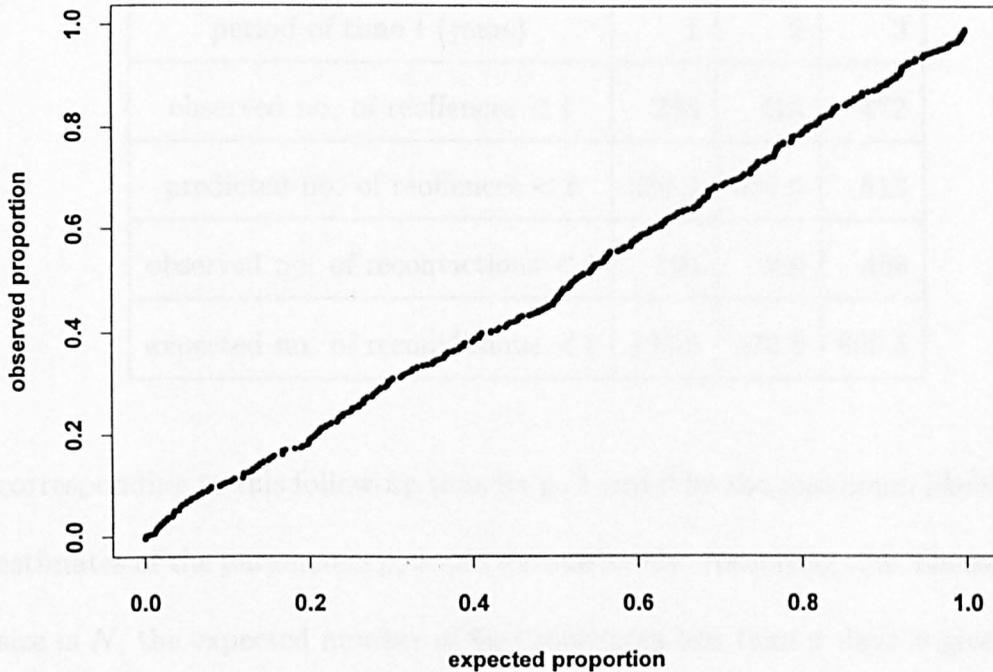


Fig. 4.3 Diagnostic Plot for X (Marginal Model:full F/U)



## 4.4 The Observed and Expected Number of Reoffence

In this section we shall consider the observed and expected number of first reoffences and reconvictions in the first, second and third year under the model (3.10) with 1, 2 and at least 3 years of follow-up. If the model is valid for a given follow-up time, then the observed and expected number of first reoffences should be close to each other as should be the observed and expected number of first reconvictions. Now, consider a given follow-up time,  $t_0$  say, and

Table 4.2: Expected no. of reoffence and reconviction, F/U=1 yr

period of time $t$ (years)	1	2	3
observed no. of reoffences $< t$	288	413	472
predicted no. of reoffences $< t$	380.7	638.5	813
observed no. of reconvictions $< t$	194	368	458
expected no. of reconvictions $< t$	193.9	470.9	690.5

corresponding to this follow-up time let  $\hat{p}$ ,  $\hat{\lambda}$  and  $\hat{\theta}$  be the maximum likelihood estimates of the parameters  $p$ ,  $\lambda$  and  $\theta$  respectively. Assuming that the sample size is  $N$ , the expected number of first reoffences less than  $x$  days is given by

$$N\hat{p}P(X < x) = N\hat{p}(1 - e^{-\hat{\lambda}x})$$

and the expected number of first reconvictions less than  $z$  days is given by

$$N\hat{p}P(Z < z) = N\hat{p}\left(1 - \frac{\hat{\theta}e^{-\hat{\lambda}z} - \hat{\lambda}e^{-\hat{\theta}z}}{\hat{\theta} - \hat{\lambda}}\right).$$

The calculations for the observed and expected number of first reoffences and reconvictions under the model (3.10) with different follow-up times are given in the Tables 4.2, 4.3 and 4.4.

In Table 4.2, for each given period of time the observed and expected number of reoffences are quite different from each other. Also, except in first year, the observed and predicted number of reconvictions are not close to each other. So, this aspect of the analysis implies that the model (3.10) with 1 year of follow-up does not extrapolate the data well. On the other hand,

Table 4.3: Expected no. of reoffence and reconviction, F/U=2 yr

period of time $t$ (years)	1	2	3
observed no. of reoffences $< t$	288	413	472
predicted no. of reoffences $< t$	287.1	431	503.2
observed no. of reconvictions $< t$	194	368	458
expected no. of reconvictions $< t$	184.5	368.3	470.5

Table 4.4: Expected no. of reoffence and reconviction, full F/U

period of time $t$ (years)	1	2	3
observed no. of reoffences $< t$	288	413	472
predicted no. of reoffences $< t$	278.8	417.9	487.3
observed no. of reconvictions $< t$	194	368	458
expected no. of reconvictions $< t$	187.4	364.3	459.8

considering Tables 4.3 and 4.4 we see that for a given period of time the observed and expected number of reoffences as well as those of convictions are very close to each other. Therefore, this aspect of the analysis suggests that for 2 and at least 3 years of follow-up the model fits the data well.

## 4.5 Kaplan-Meier and Fitted Survival Plots

In this section we shall introduce an alternative procedure for checking the goodness of fit of the model (3.10) with 1, 2 and at least 3 years of follow-up. Let  $F_X$  and  $F_Z$  be the probability distribution functions of  $X$  and  $Z$  respectively, then

$$F_X(\lambda, x) = P(X \leq x) = 1 - e^{-\lambda x}$$

and

$$F_Z(\lambda, \theta, z) = P(Z \leq z) = 1 - \frac{\theta e^{-\lambda z} - \lambda e^{-\theta z}}{\theta - \lambda}. \quad (4.1)$$

Using the split population model given by  $p$ ,  $\lambda$  and  $\theta$ , the predicted survival function of  $Z$  is given by

$$S(z) = 1 - p + p P(Z > z) = 1 - p F_Z(\lambda, \theta, z).$$

The estimated survival function of the actual reoffence time  $X$  is defined by

$$S(x) = 1 - p + p P(X > x) = 1 - p F_X(\lambda, x) = 1 - p P(X \leq x). \quad (4.2)$$

However, because of the delayed censoring, an offence may in fact have occurred by time  $x$  but not be observed as the case has not yet come to court.

So the fitted survival function of the observed values of  $X$  is given by

$$S^*(x) = 1 - p P(X \leq x, X + Y \leq t)$$

$t$  being the time to follow-up, or

$$S^*(x) = 1 - p \{1 - F_2(\lambda, \theta, \lambda x, t)\} F_Z(\lambda, \theta, t)$$

where  $F_2$  and  $F_Z$  are defined by equations (3.7) and (4.1) respectively. Using equations (3.7) and (4.1), this is equal to

$$S^*(x) = \begin{cases} 1 - p + p e^{-\lambda x} + p \lambda e^{-\theta t} \frac{1 - e^{-(\lambda - \theta)x}}{\lambda - \theta} & \text{if } x < t \\ 1 - p F_Z(\lambda, \theta, t) & \text{if } x \geq t. \end{cases} \quad (4.3)$$

With the more usual kind of censoring, the survival model (4.2) can be illustrated by comparing the corresponding predicted survival curve  $S(x)$  with a Kaplan-Meier survival plot (the observed survival plot) calculated directly from the data. With delayed censoring, the Kaplan-Meier survival plot has to be compared not with the fitted survival function  $S(x)$  but with the estimated survival function  $S^*$ . If the quantities in the right hand side of equation (4.3) differ from case to case, either through dependence on covariates or through differing follow-up times, equation (4.3) defines the case-specific survival curves  $S_i^*(x)$ ,  $i = 1, \dots, N$ ,  $N$  being the total sample size. Then the Kaplan-Meier survival curve for the observed reoffence times must be compared with the sample average

$$\bar{S}^*(x) = \frac{1}{N} \sum_{i=1}^N S_i^*(x). \quad (4.4)$$

This is most usefully done for interesting subsets of the data, for example the subsets defined by ranges of the risk score  $R_{sc} = a^T v$ . This analysis has been

considered in section 4.7. The Kaplan-Meier survival curves for  $X$  and  $Z$  can be easily done in the package S-PLUS by using the function `Surv.fit`.

In order to compare the extent of the survival curves to flatten off at larger values of  $X$ , caused by under-reporting of the later reoffences, we make an initial exploration of the survival curves by ignoring the covariates and treating the data as an homogenous sample. Figure 4.4 shows the Kaplan-Meier survival plot of the reoffence times (broken curve), and superimposes (full curve) the theoretical curve  $\bar{S}^*(x)$  from equation (4.4). This gives a good fit. Also shown (dotted curve) is the estimated survival curve  $S(x)$  from equation (4.2). Each survival plot has used the data over the full follow-up period 3 to 4 years. As expected the two curves  $S(x)$  and  $\bar{S}^*(x)$  are very similar for the smaller reoffence times but  $\bar{S}^*(x)$  correctly reflects the levelling off towards the right hand side of the graph caused by the under-reporting of late reoffences.

With delayed censoring the predicted survival curves for  $X$  under the marginal model (3.10) with 1, 2 and at least 3 years of follow-up together with the Kaplan-Meier survival curve of  $X$  corresponding to the original follow-up, at least 3 years, are shown in Figure 4.5 from which it is readily apparent that the predicted survival curves corresponding to 2 and at least 3 years of follow-up are very nearly along the Kaplan-Meier survival curve. So, for 2 and at least 3 years of follow-up the model fits the data well. For 1 year of follow-up the model gives a better fit to that part of the data corresponding to those values of  $X$  up to about 2 months. But, for other values of  $X$ , the predicted

Table 4.5: Committed, Convicted and Censored no. of reoffences

period of time $t$	1 yr	2 yrs	3 yrs	end of follow-up
no. of reoffences $< t$	288	413	472	486
no. of reconvictions $< t$	194	368	458	486
no. of censored reoffences	94	45	14	0

survival curve deviates substantially from the Kaplan-Meier survival curve. Therefore, the model with 1 year of follow-up does not predict the future data well.

Similarly, the Kaplan-Meier survival curve and the fitted survival curves for  $Z$  are shown in Figure 4.6. Analogous to the argument just done for  $X$  we can say that for  $Z$  the model with 2 and at least 3 years of follow-up fits the data well. But, for 1 year of follow-up the model gives the better fit to that part of the data corresponding to  $Z$  up to about 1 year. Of course, it should be pointed out that when the follow-up time is 1 year the actual number of reoffences in first year is 288 from which only 194 cases will be observed, the remaining 94 cases being censored. In fact these 94 censored cases will be added to the number of reoffences between the first and second year for further investigation. Table 4.5 sets out this sort of calculations.

Fig. 4.4 Observed and Fitted Survival Curves without covariates (Marginal Model)

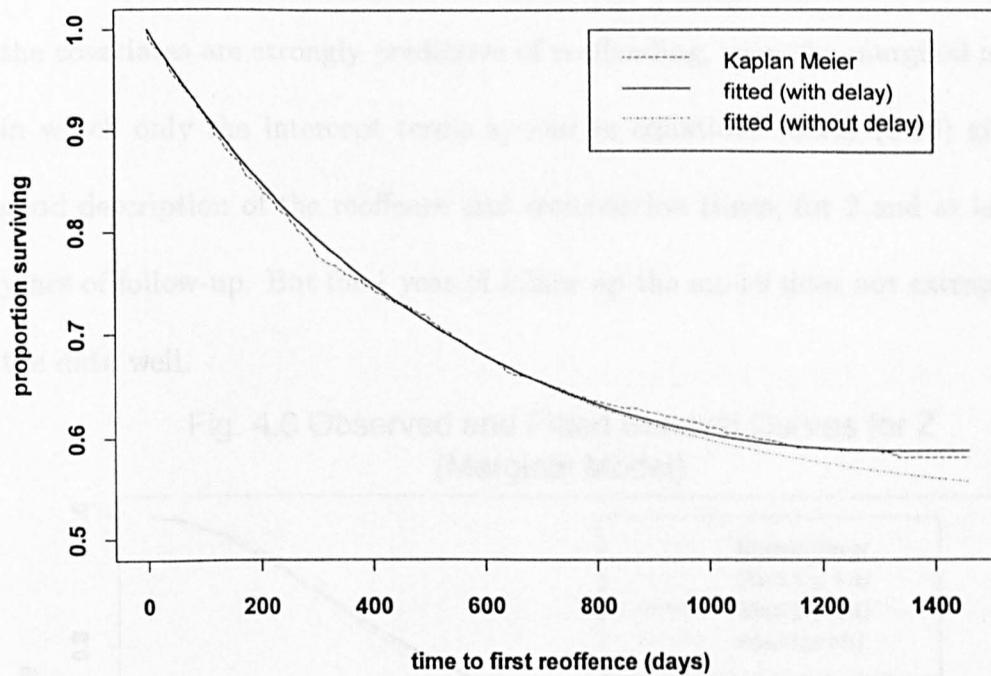
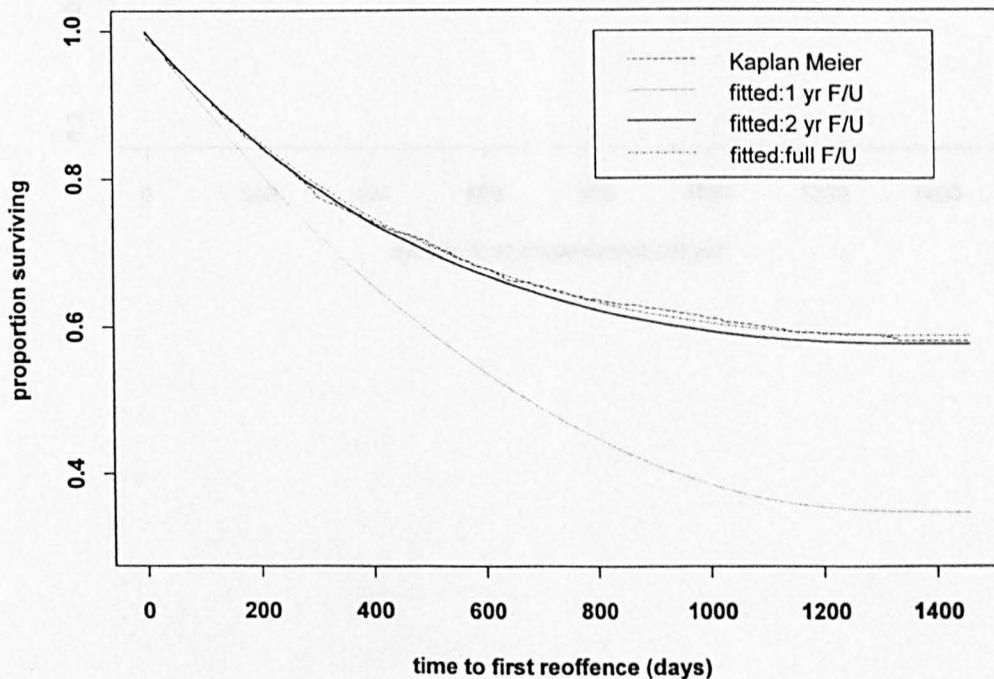
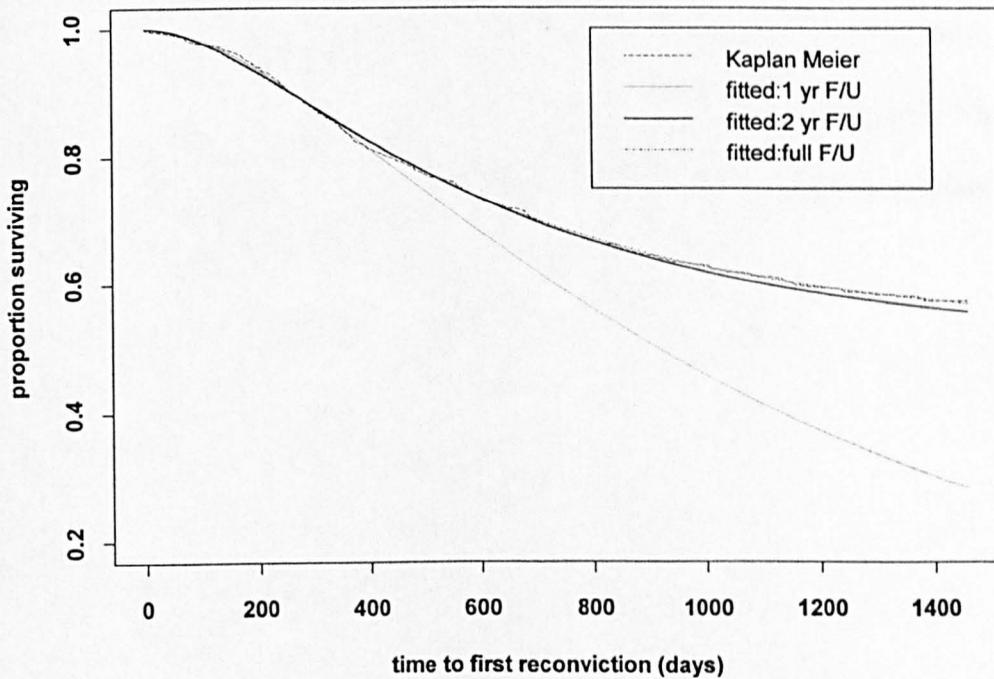


Fig. 4.5 Observed and Fitted Survival curves for X (Marginal Model)



From the various approaches developed so far we can now sum up the results of fitting the marginal model (3.10) to the data as follows: although the covariates are strongly predictive of reoffending, even the marginal model in which only the intercept terms appear in equations (3.14)–(3.16) gives a good description of the reoffence and reconviction times, for 2 and at least 3 years of follow-up. But for 1 year of follow up the model does not extrapolate the data well.

Fig. 4.6 Observed and Fitted Survival Curves for Z  
(Marginal Model)



## 4.6 Diagnostics for $X$ in Restricted Model

As we have already mentioned in section 3.3, in order to check the goodness of fit of the model we need to plot

$$\frac{n+1-j}{n+1} \text{ against } \frac{1}{n} \sum_{i=1}^n F_2(\lambda_i, \theta_i, u_{(j)}, t_i)$$

that is, the observed proportion of  $U$ 's against the expected proportion of  $U$ 's where  $U = \lambda X$ . If the model fits the data well, then the plot should resemble a straight line. Under the restricted model (3.17) with 1, 2 and at least 3 years of follow up, these diagnostic plots are pictured in Figures 4.7, 4.8 and 4.9 respectively, which are all reasonably linear. Therefore, this aspect of the analysis implies that for 1, 2 and at least 3 years of follow up the model fits the data well.

Fig. 4.7 Diagnostic Plot for X (Restricted Model:1 yr F/U)

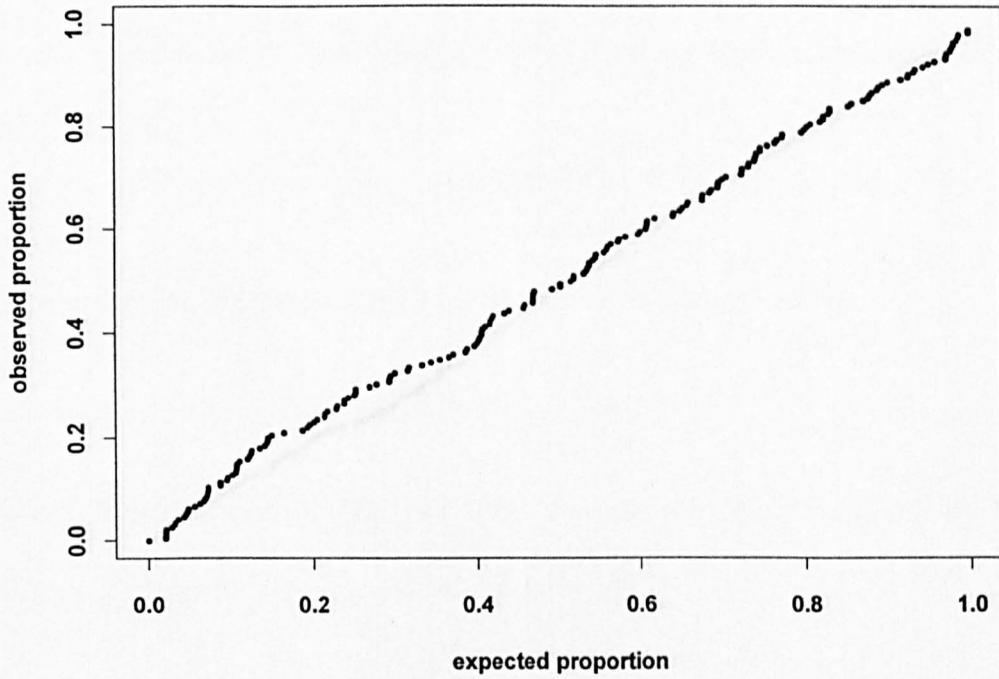


Fig. 4.8 Diagnostic Plot for X (Restricted Model:2 yr F/U)

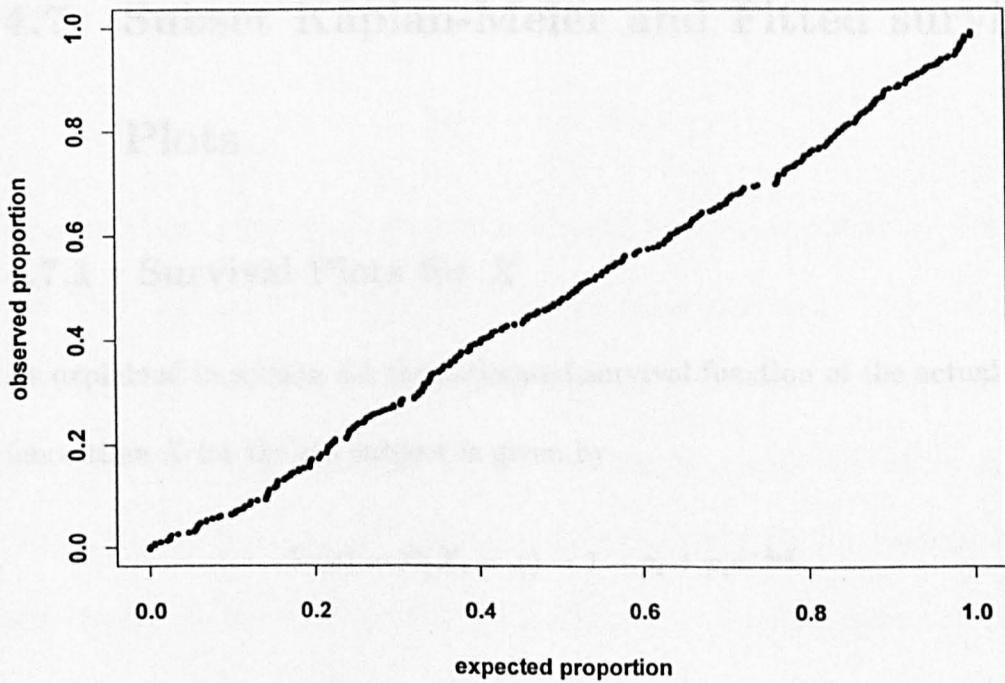
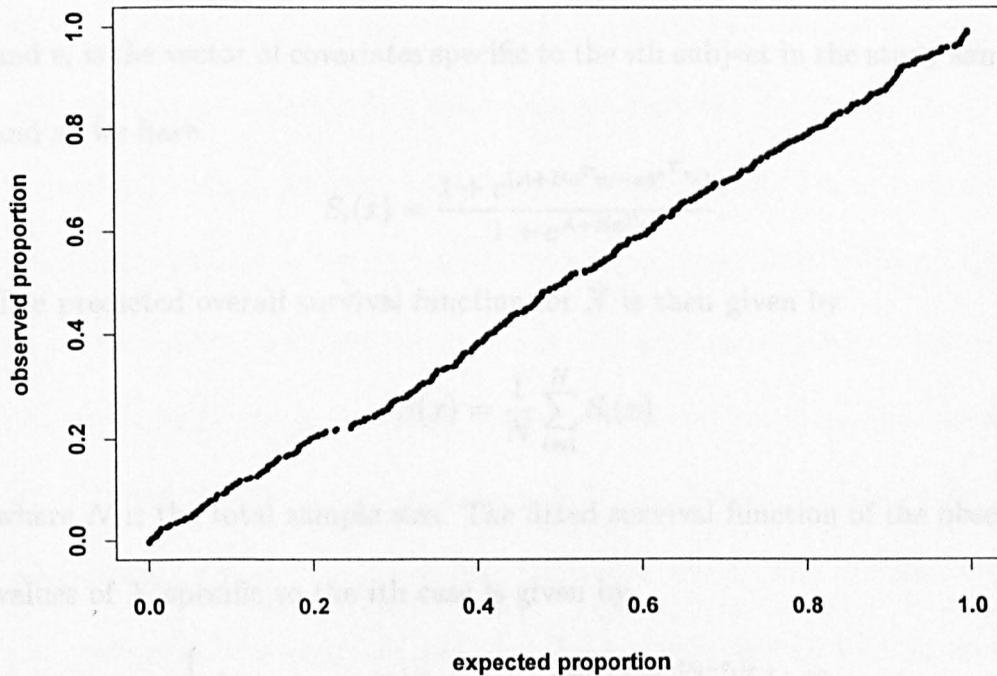


Fig. 4.9 Diagnostic Plot for X (Restricted Model:full F/U)



## 4.7 Subset Kaplan-Meier and Fitted survival Plots

### 4.7.1 Survival Plots for $X$

As explained in section 4.5 the estimated survival function of the actual reference time  $X$  for the  $i$ th subject is given by

$$S_i(x) = P(X_i > x) = 1 - p_i + p_i e^{-\lambda_i x}$$

where

$$p_i = \frac{e^{A+Ba^T v_i}}{1 + e^{A+Ba^T v_i}}, \quad \lambda_i = e^{a^T v_i}$$

and  $v_i$  is the vector of covariates specific to the  $i$ th subject in the study sample,

and so we have

$$S_i(x) = \frac{1 + e^{(A+Ba^T v_i - x e^{a^T v_i})}}{1 + e^{A+Ba^T v_i}}.$$

The predicted overall survival function for  $X$  is then given by

$$S(x) = \frac{1}{N} \sum_{i=1}^N S_i(x)$$

where  $N$  is the total sample size. The fitted survival function of the observed values of  $X$  specific to the  $i$ th case is given by

$$S_i^*(x) = \begin{cases} 1 - p_i + p_i e^{-\lambda_i x} + p_i \lambda_i e^{-\theta_i t_i} \left\{ \frac{1 - e^{-(\lambda_i - \theta_i)x}}{\lambda_i - \theta_i} \right\} & \text{if } x < t_i \\ 1 - p_i F_Z(\lambda_i, \theta_i, t_i) & \text{if } x \geq t_i \end{cases} \quad (4.5)$$

where  $F_Z(\lambda_i, \theta_i, t_i)$  is defined by equation (4.1). The estimated sample average survival function is therefore given by

$$\bar{S}^*(x) = \frac{1}{N} \sum_{i=1}^N S_i^*(x). \quad (4.6)$$

After estimating the parameters of the model for a given follow up time, we plot  $\bar{S}^*(x)$  against  $x$  and then compare the behaviour of this survival curve with the corresponding Kaplan-Meier survival plot calculated directly from the data. The approach used in the analysis is to split the whole sample reoffending score  $\log \lambda = a^T v$  into separate groups and for each subgroup we plot the Kaplan-Meier survival curve and the corresponding estimated sample average survival curve  $\bar{S}^*(x)$  for different follow up times under the restricted

model. By this method the risk of reoffending between different subgroups can be examined and compared carefully. For a given follow up time if the Kaplan-Meier plot and the fitted survival curve corresponding to each subgroup are very close to each other, then the model predicts the data well. Here, three groups of cases are selected, those with the lowest 20% of scores, those with the middle 20% of scores and those with the highest 20% of scores. These represent low risk, medium risk and high risk cases respectively. Figures 4.10, 4.11 and 4.12 illustrate the subset Kaplan-Meier and the fitted survival curves for  $X$  under the restricted model with 1, 2 and at least 3 years of follow up respectively. Figure 4.10 shows that the model with 1 year of follow up does not extrapolate the data well and from Figures 4.11 and 4.12 we observe that for 2 and at least 3 years of follow up the model predicts the data well.

Fig. 4.10 Observed and Fitted Survival Curves by Risk Group for X (Restricted Model:1 yr F/U)

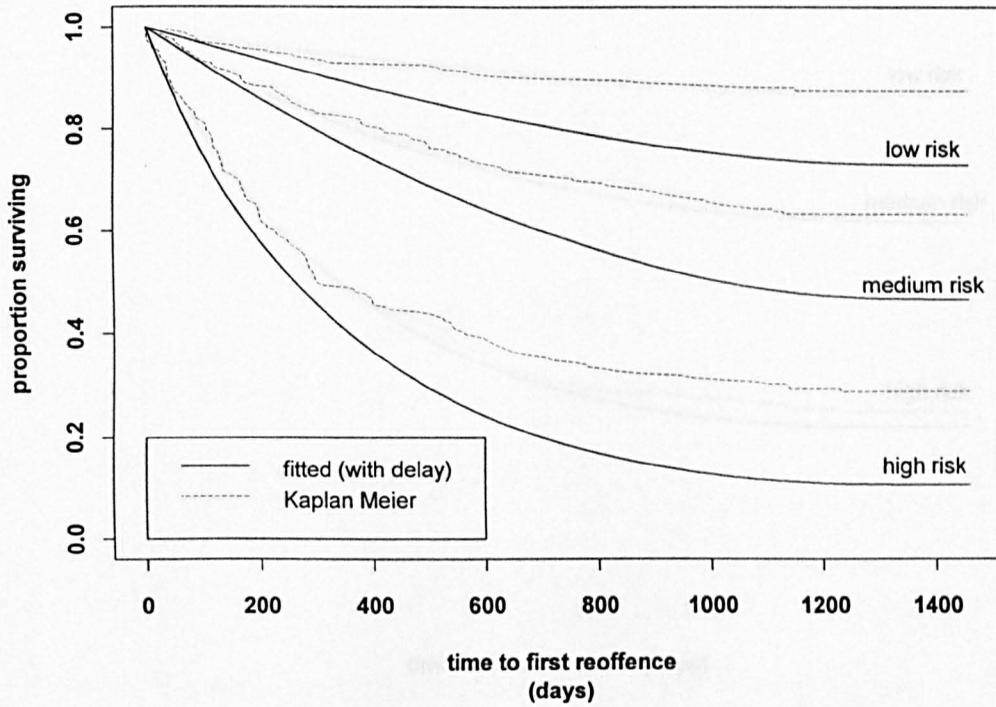


Fig. 4.11 Observed and Fitted Survival Curves by Risk Group for X (Restricted Model:2 yr F/U)

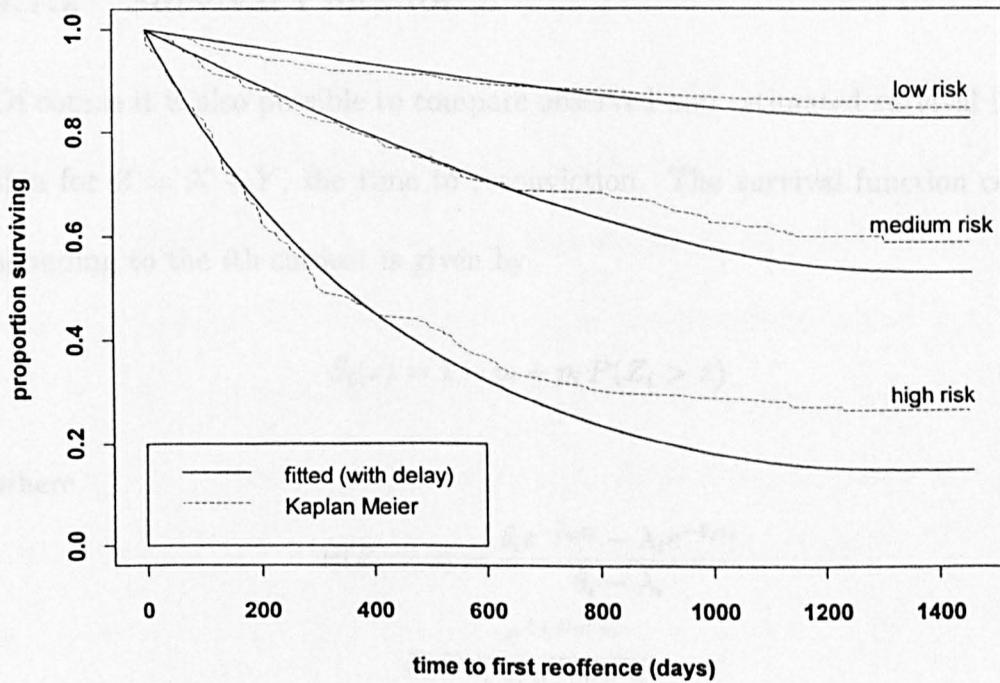
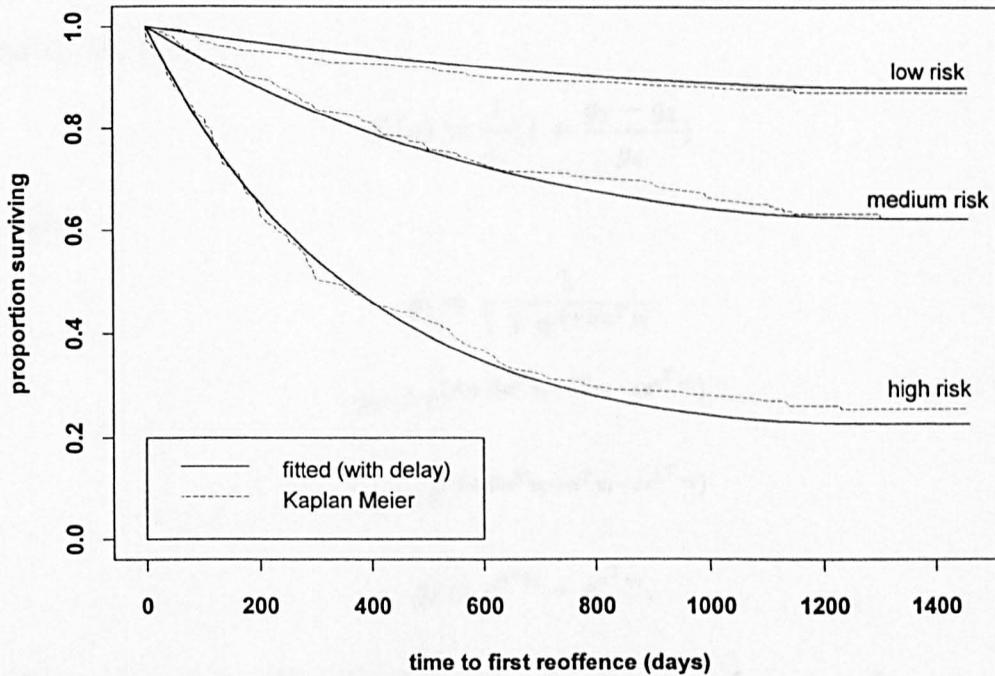


Fig. 4.12 Observed and Fitted Survival Curves by Risk Group for X (Restricted Model:full F/U)



#### 4.7.2 Survival Plots for $Z$

Of course it is also possible to compare observed and estimated survival function for  $Z = X + Y$ , the time to reconviction. The survival function corresponding to the  $i$ th subject is given by

$$S_i(z) = 1 - p_i + p_i P(Z_i > z) \quad (4.7)$$

where

$$P(Z_i > z) = \frac{\theta_i e^{-\lambda_i z_i} - \lambda_i e^{-\theta_i z_i}}{\theta_i - \lambda_i}$$

$$p_i = \frac{e^{A + Ba^T v_i}}{1 + e^{A + Ba^T v_i}}$$

$$\lambda_i = e^{a^T v_i}, \quad \theta_i = e^{b^T v_i}$$

and  $v_i$  is the vector of covariates corresponding to the  $i$ th case. Then it can be shown that

$$S_i(z) = \frac{1}{g_1} \left( 1 + \frac{g_2 - g_3}{g_4} \right)$$

where

$$g_1 = \frac{1}{1 + e^{A + Ba^T v_i}}$$

$$g_2 = e^{(A + Ba^T v_i + b^T v_i - ze^{a^T v_i})}$$

$$g_3 = e^{(A + Ba^T v_i + a^T v_i - ze^{b^T v_i})}$$

$$g_4 = e^{b^T v_i} - e^{a^T v_i}.$$

The predicted overall survival function for  $Z$  is therefore given by

$$S(z) = \frac{1}{N} \sum_{i=1}^N S_i(z) \quad (4.8)$$

where  $N$  is the sample size. Applying exactly the same procedure used in subsection 4.7.1, the Kaplan-Meier and fitted overall survival curves for  $Z$  under the restricted model with 2 and at least 3 years of follow up are depicted in Figures 4.13 and 4.14 respectively. In both diagnostic plots the Kaplan-Meier survival curve and the fitted survival curve corresponding to each subgroup are very nearly along each other, implying that for 2 and at least 3 years of follow up the model predicts the data well. Also, it can be shown that for 1 year of follow up the model does not extrapolate the data well.

Fig. 4.13 Observed and Fitted Survival Curves by Risk Group for Z (Restricted Model:2 yr F/U)

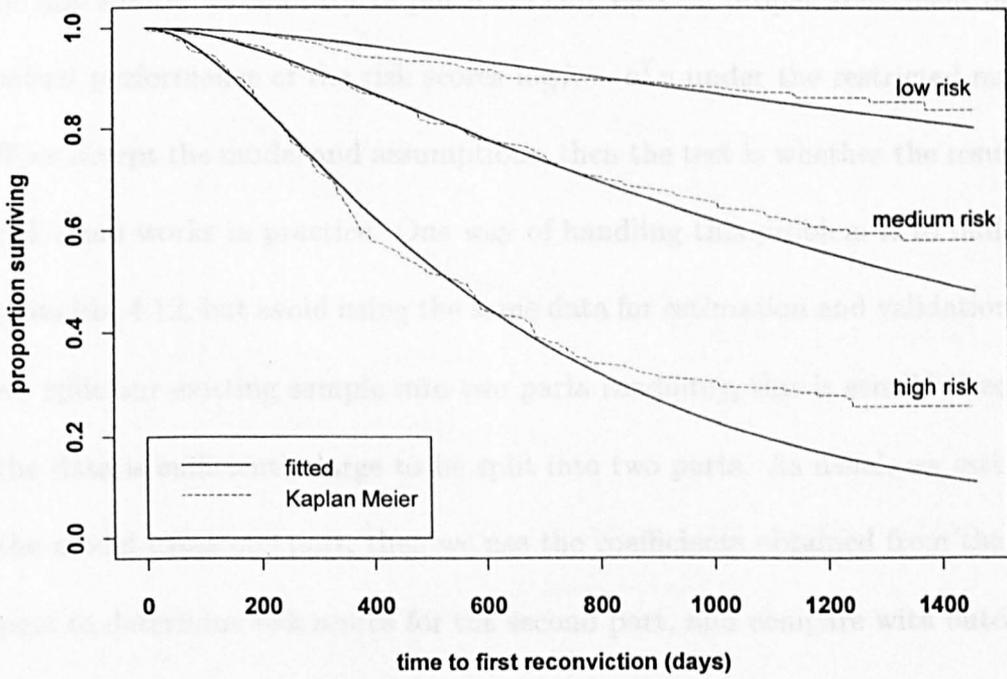
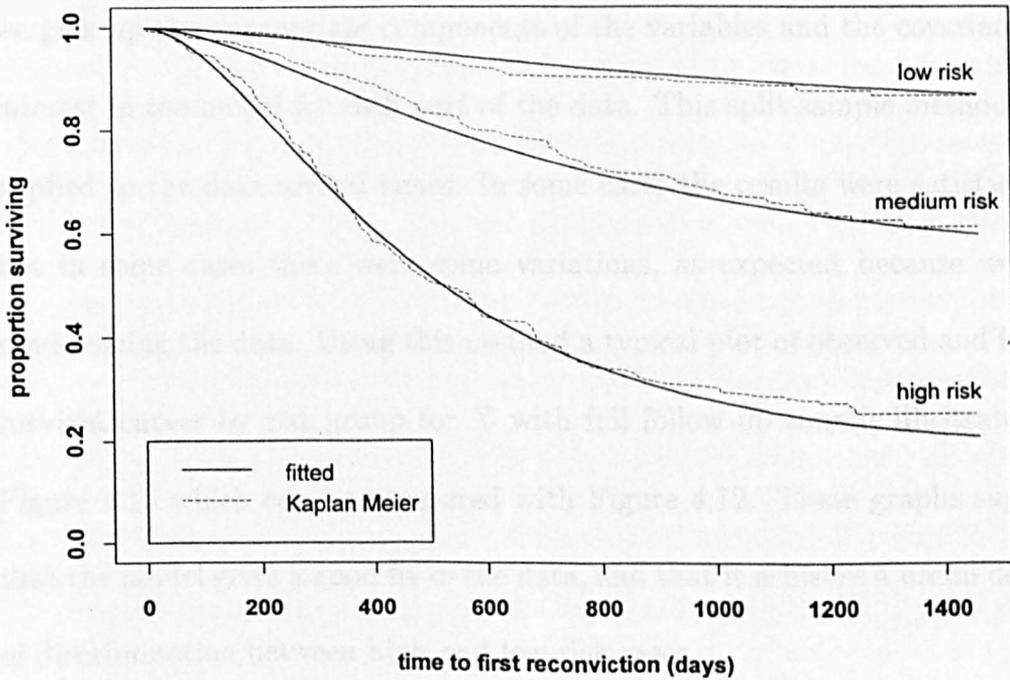


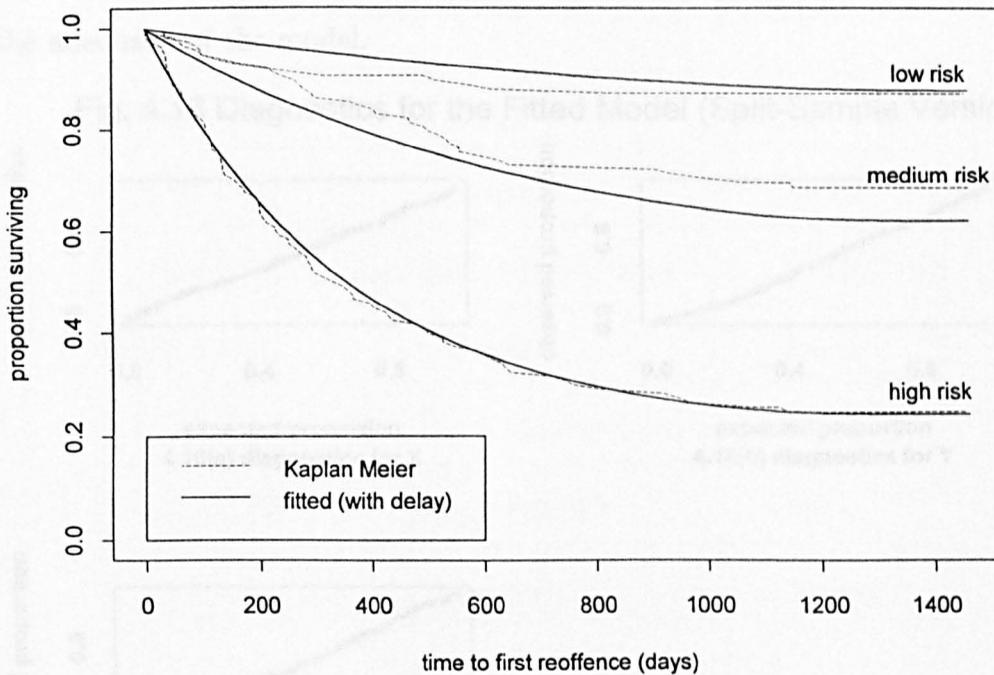
Fig. 4.14 Observed and Fitted Survival Curves by Risk Group for Z (Restricted Model:full F/U)



## 4.8 Split-Sample and Simulation Methods

In this section we shall try to put more emphasis on proper assessment of the actual performance of the risk scores  $\log \lambda = a^T v$  under the restricted model. If we accept the model and assumptions, then the test is whether the resulting risk score works in practice. One way of handling this problem is to build up plots like 4.12, but avoid using the same data for estimation and validation. So we split our existing sample into two parts randomly, this is sensible because the data is sufficiently large to be split into two parts. As usual, we estimate the model using one part, then we use the coefficients obtained from the first part to determine risk scores for the second part, and compare with outcome. This gives a fairer assessment of the performance of the scores. Note that in order to get the split data, we generate a random sample of indices of size  $N$ ,  $N$  being the total sample size, and then corresponding to these random indices we pick up the appropriate components of the variables and the covariates of interest in the model for each part of the data. This split-sample method was applied to the data several times. In some cases the results were satisfactory but in some cases there were some variations, as expected because we are randomizing the data. Using this method a typical plot of observed and fitted survival curves by risk group for  $X$  with full follow up time is illustrated in Figure 4.15 which can be compared with Figure 4.12. These graphs suggest that the model gives a good fit to the data, and that it achieves a useful degree of discrimination between high and low risk cases.

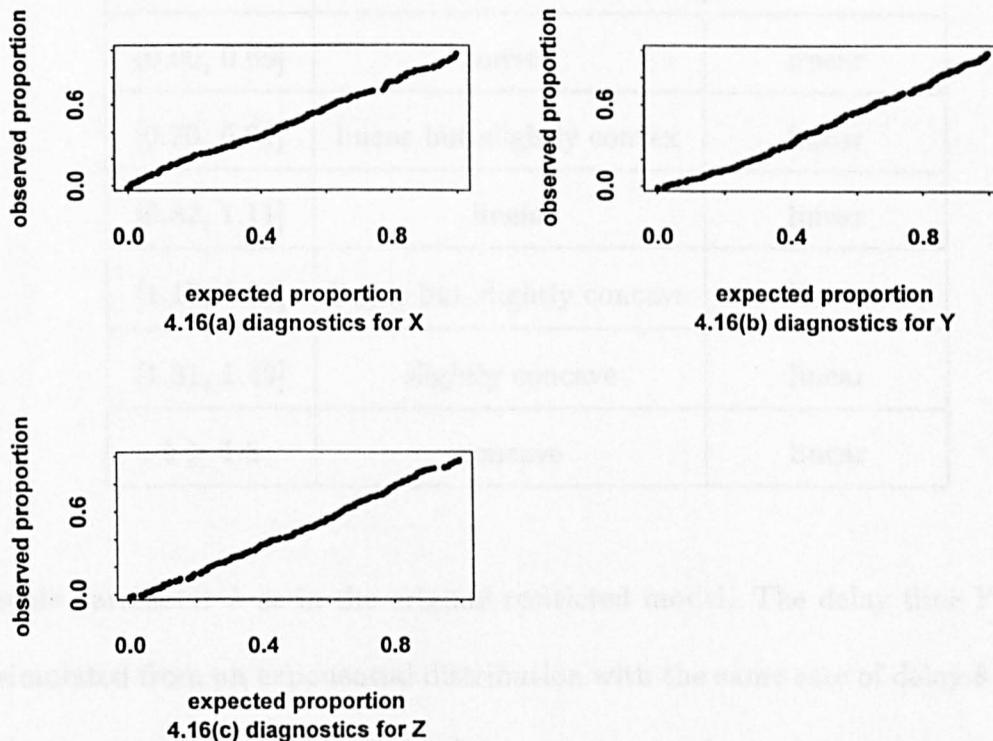
Fig. 4.15 Split-Sample Version of Figure 4.12



If we consider the new score—the score under the restricted model with 2 years of follow up, then by analogy with Figures 4.12 and 4.15, we can again examine the survival curves for the three risk groups defined by the top 20%, the middle 20% and the lowest 20% of the values of the new score, and compare with the curves  $\bar{S}^*(x)$  calculated from the model fitted to the 2-years data. The plot is very similar to Figure 4.12: the new version of model gives a good fit to the data up to 2 years as expected, but also gives a good prediction of the data which were subsequently collected during the third and fourth years of follow up. The split-sample method was also applied to the order-based diagnostics in the analysis and in all cases the results were satisfactory. Figure 4.16 shows a randomly chosen diagnostic plots for the fitted model which are similar to

those for the full analysis shown in Figures 4.9, 3.6 and 3.11, again indicating the adequacy of the model.

Fig. 4.16 Diagnostics for the Fitted Model (Split-Sample Version)



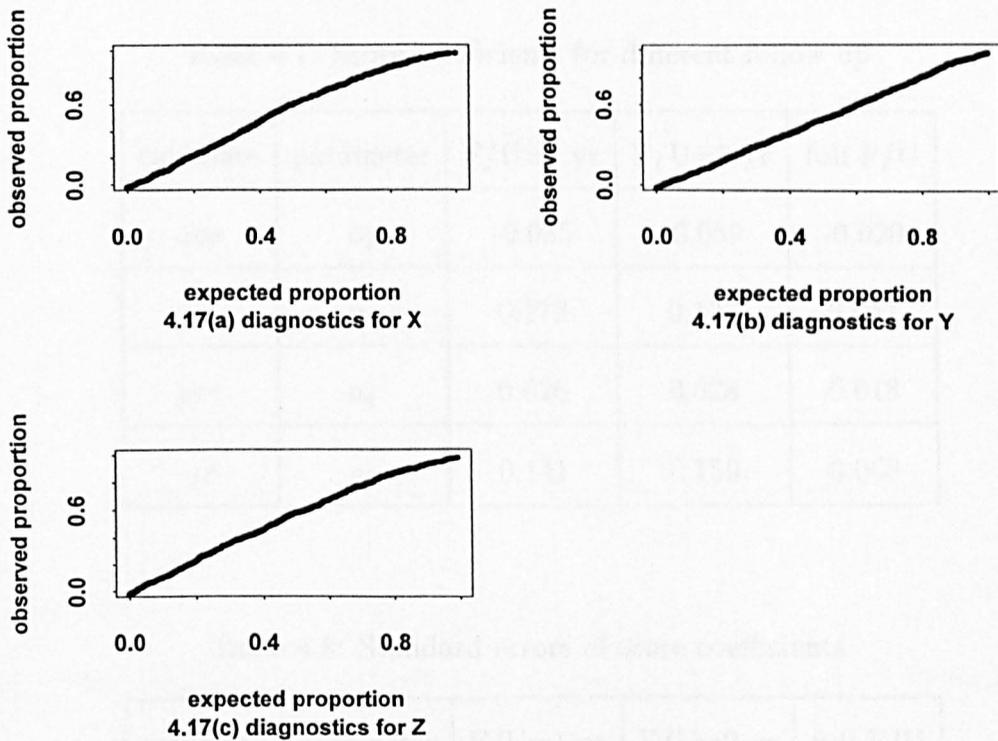
A more powerful approach for thorough testing of the suggested order-based diagnostics is based on simulation method in which we simulate a data set of size  $N$ ,  $N$  being the total sample size, from an alternative model,  $M$  say. Then we refit our original proposed model (the restricted model) with these simulated data and check the subsequent diagnostics for linearity as before. We do these simulation experiments to check the sensitivity of these diagnostic plots to model misspecification. Here, keeping the covariates and follow up time  $T$  fixed, we have simulated the reoffence time  $X$  from Weibull and gamma distributions with different shape parameters but with the same

Table 4.6: Pattern of Diagnostics with Simulation

interval for $b$	diagnostic of $X$	diagnostic of $Y$
(0.00, 0.69]	convex	linear
[0.70, 0.81]	linear but slightly convex	linear
[0.82, 1.11]	linear	linear
[1.12, 1.30]	linear but slightly concave	linear
[1.31, 1.49]	slightly concave	linear
$b \geq 1.5$	concave	linear

scale parameter  $\lambda$  as in the original restricted model. The delay time  $Y$  was simulated from an exponential distribution with the same rate of delay  $\theta$  as in the restricted model. Note that the reason for simulating  $Y$  from exponential model and  $X$  from an alternative Weibull or gamma model is to observe possible changes in the diagnostics of  $X$ , although it is possible to simulate both  $X$  and  $Y$  from the same Weibull or gamma model. Similarly, this procedure was repeated vice versa to simulate  $X$  from exponential model and  $Y$  from Weibull and gamma models. The diagnostics were then examined for straightness. For the shape parameter  $b$  in different ranges, the outcome of the corresponding order-based diagnostics are summarized in Table 4.6, which implies that under this simulation the diagnostic plots are sensitive to shape parameter  $b$ . Some typical illustrations for the shape parameter  $b=1.3$  are shown in Figure 4.17.

Fig. 4.17 Diagnostic Plots for the Fitted Model (Simulated Data)



Compared with the results obtained from the restricted model, the simulation results presented here indicate that the diagnostics appear to be not much affected by the change in shape parameter  $b$ , for  $b \in [0.82, 1.20]$ . Thus, for this range of shape parameter, the diagnostics are not able to detect model misspecification, but outside this range the diagnostics do show clear non-linearity. This indicates the level of sensitivity of the diagnostic plots for model misspecification.

## 4.9 Standardized Estimates

Using the log-likelihood function (3.18) we can find the maximum likelihood estimates of the parameters under the restricted model (3.17). The standard

Table 4.7: Score coefficients for different follow up

covariate	parameter	F/U=1 yr	F/U=2 yr	full F/U
<i>age</i>	$a_2$	-0.035	-0.059	-0.020
<i>ac</i>	$a_3$	0.172	0.137	0.041
<i>pre</i>	$a_4$	0.026	0.028	0.018
<i>jc</i>	$a_5$	0.141	0.159	0.058

Table 4.8: Standard errors of score coefficients

covariate	parameter	F/U=1 yr	F/U=2 yr	full F/U
<i>age</i>	$a_2$	0.027	0.011	0.006
<i>ac</i>	$a_3$	0.102	0.042	0.028
<i>pre</i>	$a_4$	0.020	0.008	0.006
<i>jc</i>	$a_5$	0.085	0.044	0.033

errors of the parameter estimates and the standardized estimates of the parameters are found by using the procedures developed in subsection 3.10.3. Under the restricted model with 1, 2 and at least 3 years of follow up the results of these analyses corresponding to the score coefficients  $a_2, a_3, a_4, a_5$  for the reoffending score  $\log \lambda = a^T v$  are summarized in Tables 4.7, 4.8 and 4.9.

As mentioned in subsection 3.10.2, the standardized estimates of the parameters can be used as a measure of importance of the covariates in a model.

Table 4.9: Standardized estimates

covariate	parameter	F/U=1 yr	F/U=2 yr	full F/U
<i>age</i>	$a_2$	-1.309	-5.446	-3.082
<i>ac</i>	$a_3$	1.688	3.287	1.464
<i>pre</i>	$a_4$	1.313	3.340	3.092
<i>jc</i>	$a_5$	1.665	3.640	1.747

From Table 4.9 we see that:

- (i) in the model with 1 year of follow up, none of the covariates is statistically significant because their corresponding standardized estimates are small, implying that we can remove any covariate individually from this  $\lambda$ -part of the model without affecting the results substantially. However, when more than one covariate is deleted from the model we must consider the estimated value of the deviance  $\Delta D$  created by this alteration of the covariates and find out if this value is significant as compared to the degrees of freedom of  $\Delta D$ .
- (ii) in the model with 2 years of follow up all the covariates are statistically significant because their corresponding standardized estimates are large. In this case the covariates are retained in the model.
- (iii) in the model with at least 3 years of follow up the covariates *age* and *pre* are statistically significant but the covariates *ac* and *jc* are not.

According to Table 4.7 the score coefficients are different for 2 and at least 3 years of follow-up. However, the covariates are strongly correlated and we

go on to show that the scores themselves are in fact very similar.

## 4.10 Correlation Matrix

The correlation matrix of the reoffending scores  $\log \lambda = a^T v$ ,  $v$  being the vector of covariates, under the restricted model with 1, 2 and at least 3 years of follow up is given by

$$\begin{pmatrix} 1.00 & 0.96 & 0.98 \\ 0.96 & 1.00 & 0.99 \\ 0.98 & 0.99 & 1.00 \end{pmatrix}$$

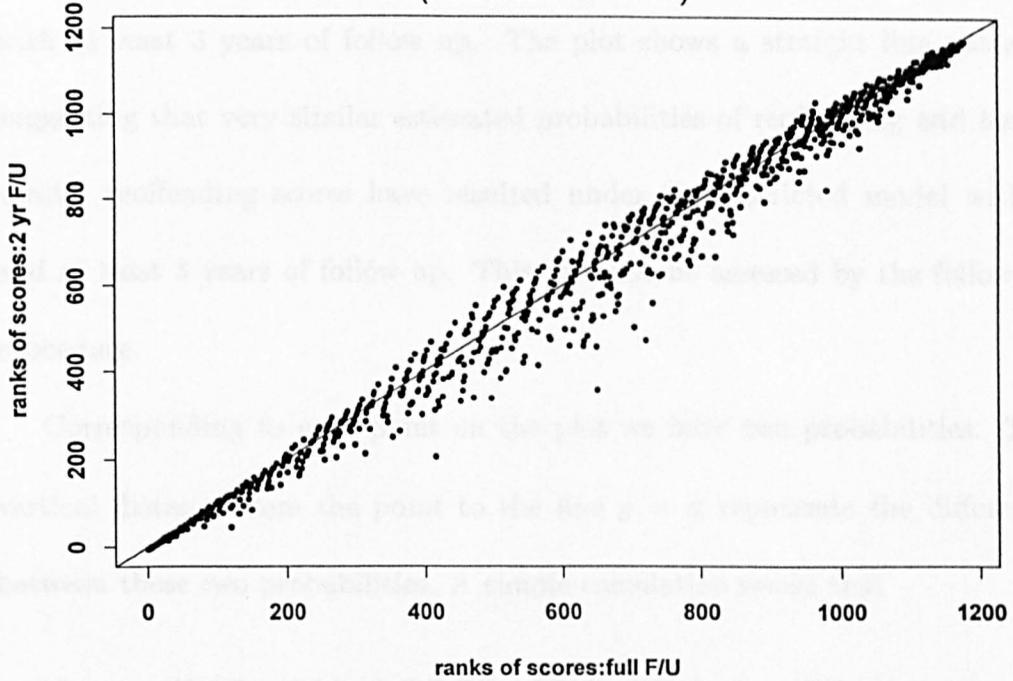
from which we observe that the scores are highly correlated with each other, particularly when the follow-up times are 2 and at least 3 years. In fact, as shown in section 4.11, the scores are remarkably similar.

An alternative method of comparing the reoffending scores,  $\log \lambda = a^T v$ , corresponding to 2 and at least 3 years of follow-up is to look at the plot of the ranks of these two set of scores. If the scores are similar, the plot should resemble a straight line. Figure 4.18 illustrates this plot, substantiating the similarity of these two set of scores corresponding to 2 and at least 3 years of follow-up under the restricted model.

## 4.11 Probability of $X$ Within a Time Period $t$

In section 4.4 we discussed the expected number of first reoffences in the first, second and third year under the marginal model (3.10) with different follow up

Fig. 4.18 Plot of Ranks of Reoffending Scores  
(Restricted Model)



times. An idea similar to this and in fact an alternative method of testing the similarity of the reoffending scores,  $\log \lambda = a^T v$ , under the restricted model is to compare, for different follow-up times, the overall probability of first reoffences occurring within a time period  $t$ ,  $t$  being the length of potential parole, which is not fixed in general but varies from one subject to another. This probability is given by

$$P(X < t) = p(1 - e^{-\lambda t}),$$

$p$  being the split population proportion, or

$$P(X < t) = \frac{e^{A+Ba^T v}}{1 + e^{A+Ba^T v}} (1 - e^{-te^{a^T v}}) \quad (4.9)$$

where the probability depends on  $v$ , the vector of covariates or risk factors.

Under the restricted model (3.17) and for a time period  $t = 2$  years, Figure 4.19

illustrates the plot of  $P(X < t)$  with 2 years of follow up against  $P(X < t)$  with at least 3 years of follow up. The plot shows a straight line pattern, suggesting that very similar estimated probabilities of reoffending and hence similar reoffending scores have resulted under the restricted model with 2 and at least 3 years of follow up. This can also be assessed by the following procedure.

Corresponding to each point on the plot we have two probabilities. The vertical distance from the point to the line  $y = x$  represents the difference between these two probabilities. A simple calculation shows that

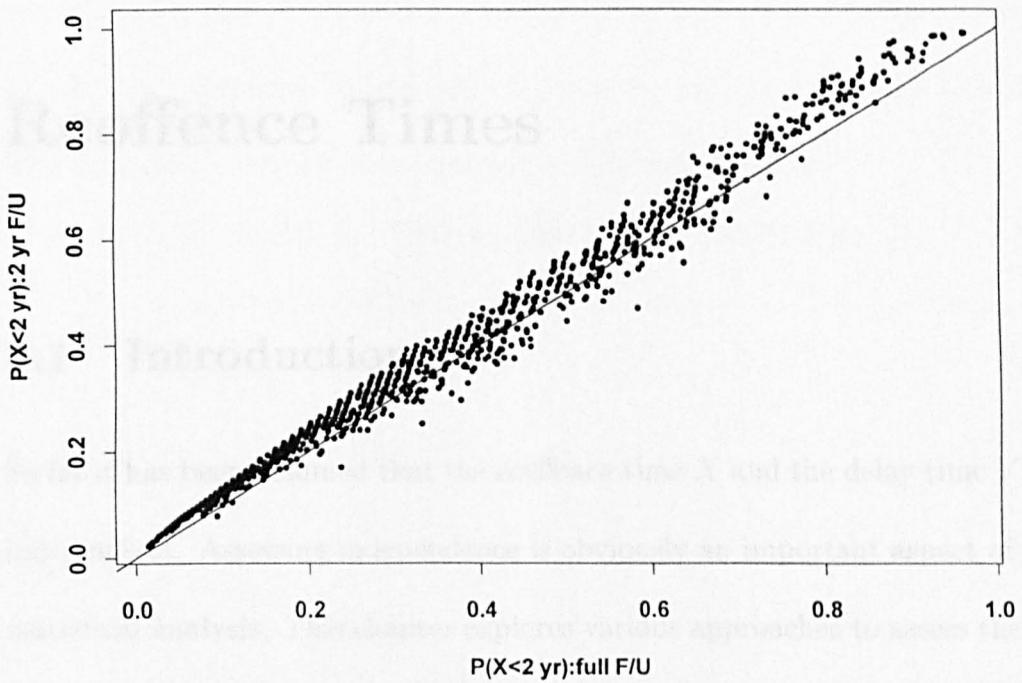
$$\#\{\text{cases} : |[P(X < 2\text{yr}) : \text{full F/U}] - [P(X < 2\text{yr}) : 2 \text{ yr F/U}]\} > 0.1\} = 9$$

which is a small number as compared with the total sample size 1179, and where  $\#A$  is the cardinality of a set  $A$ . This means that if, for example, the scores are used to estimate the probability of a reoffence occurring within 2 years, these probabilities differ by more than 0.1 in only 9 of the 1179 cases in the sample.

From the various statistical checks developed so far we can now sum up the results of fitting the restricted model to the data as follows: for 2 and at least 3 years of follow up the model gives a good fit to the data but for 1 year of follow up the model does not extrapolate the data well. Using the model suggested, risk scores for reoffending can be fitted using more up-to-date data and with lower data collection costs.

# Chapter 5

Fig. 4.19 Comparing  $P(X < 2 \text{ yr})$  for Restricted Model



# Chapter 5

## Independence of Delay and Reoffence Times

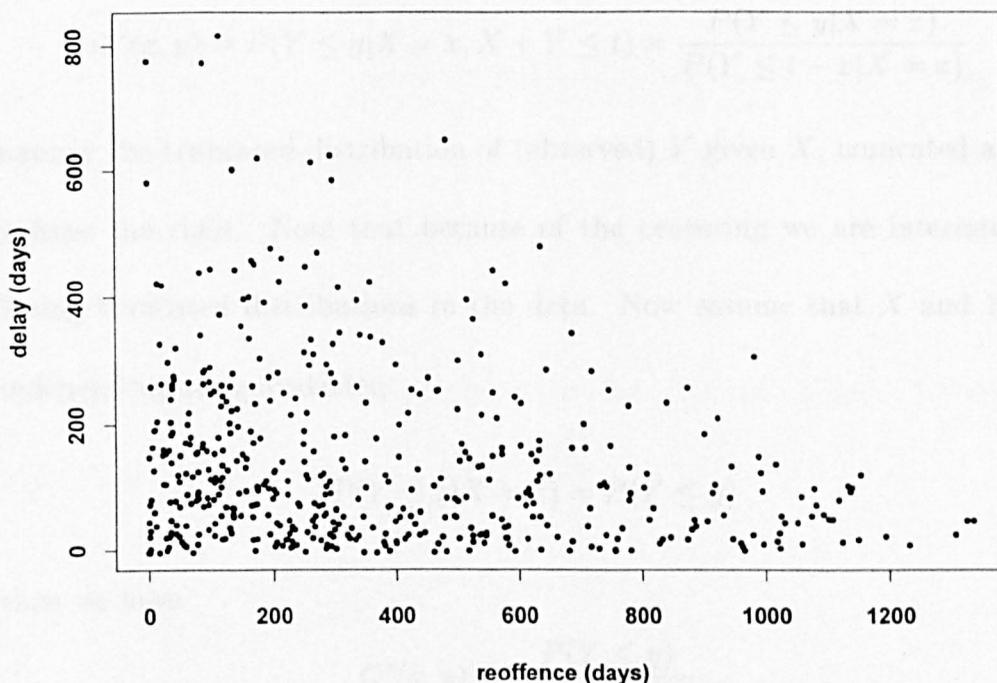
### 5.1 Introduction

So far it has been assumed that the reoffence time  $X$  and the delay time  $Y$  are independent. Assessing independence is obviously an important aspect of any statistical analysis. This chapter explores various approaches to assess the dependence of the delay and reoffence times by studying truncated distributions fitted to these data, through parametric, semi-parametric and nonparametric models without covariates. We have already checked the goodness of fit of the exponential distribution for  $X$  and  $Y$ . Also we have checked the validity of the distribution of the reconviction time  $Z = X + Y$  in the DCA model. If the distributional assumptions about  $X$  and  $Y$  seem reasonable in the light of the data, then assessing the distribution of  $Z$  is essentially assessing the

independence assumption on which the formula (4.1) for  $F_Z(\lambda, \theta, z)$  is based.

Examining the joint distribution of  $X$  and  $Y$  is rather difficult because of the censoring, as the sum of the reoffence and the delay times is equal to the reconviction time and also an observation is censored if its reconviction time exceeds its follow-up time. However, we can look at the scatterplot of the data. Figure 5.1 plots the observed delay times against the times to the observed reoffences. The censoring removes any point to the top right of this plot where the sum of the reoffence and the delay times is bigger than the time to follow-up. However, in the light of this censoring the scatterplot does not show any clear evidence of an association between  $X$  and  $Y$ .

Fig. 5.1 Plot of Delay versus Reoffence



Another possibility is to assess the conditional distribution of (observed)

$Y$  given  $X$  which is induced by the censoring. This analysis has been carried out in the following sections.

## 5.2 Analysis of Truncated Data Without Covariates

### 5.2.1 Parametric Analysis

#### Exponential Distribution

Let  $G^*$  denote the conditional probability distribution function of  $Y$  given  $X$  under the condition  $X + Y \leq t$ ,  $t$  being the time to follow up. Then

$$G^*(x, y) = P(Y \leq y | X = x, X + Y \leq t) = \frac{P(Y \leq y | X = x)}{P(Y \leq t - x | X = x)}$$

namely the truncated distribution of (observed)  $Y$  given  $X$ , truncated at  $t - x$  from the right. Note that because of the censoring we are interested in fitting truncated distributions to the data. Now assume that  $X$  and  $Y$  are independent or equivalently,

$$P(Y \leq y | X = x) = P(Y \leq y)$$

then we have

$$G^*(x, y) = \frac{P(Y \leq y)}{P(Y \leq t - x)}.$$

If  $Y$  has exponential distribution with mean  $1/\theta$ , then

$$G^*(x, y, \theta, t) = \frac{1 - e^{-\theta y}}{1 - e^{-\theta(t-x)}}. \quad (5.1)$$

Note that here we are not using any distributional assumption about  $X$ . The truncated density function of  $Y$  is now given by

$$f_{Y|X=x, X+Y \leq t}(x, y, \theta, t) = \frac{\theta e^{-\theta y}}{1 - e^{-\theta(t-x)}}$$

this gives

$$E(Y|X = x, X + Y \leq t) = \frac{1}{\theta} - \frac{t - x}{e^{\theta(t-x)} - 1} \quad (5.2)$$

and

$$E(Y^2|X = x, X + Y \leq t) = \frac{2}{\theta^2} - \frac{2(t-x)}{\theta [e^{\theta(t-x)} - 1]} - \frac{(t-x)^2}{[e^{\theta(t-x)} - 1]}.$$

Thus the conditional variance of  $Y$  is given by

$$\text{Var}(Y|X = x, X + Y \leq t) = \frac{1}{\theta^2} - \left[ \frac{t-x}{e^{\theta(t-x)} - 1} \right]^2 e^{\theta(t-x)}. \quad (5.3)$$

In the limit as  $t \rightarrow \infty$ , equation (5.2) tends to  $1/\theta$ , the mean of the marginal distribution of  $Y$ , and (5.3) tends to  $1/\theta^2$ , the variance of the marginal distribution of  $Y$ , as expected. Also as  $x \rightarrow t$ , equations (5.2) and (5.3) both tend to 0.

The conditional log-likelihood function for observed values of  $Y$  based on this truncated distribution is now given by

$$\ell = \sum_{i=1}^n \{ \log \theta - \theta y_i - \log(1 - e^{-\theta(t_i - x_i)}) \} \quad (5.4)$$

where  $n$  is the number of observed cases. The standard error of the maximum likelihood estimate of  $\theta$  is defined by

$$\text{se}(\hat{\theta}) = \left( -\frac{\partial^2 \ell}{\partial \theta^2} \Big|_{\theta = \hat{\theta}} \right)^{-1/2}$$

where

$$\partial^2 \ell / \partial \theta^2 = \sum_{i=1}^n \left\{ -\frac{1}{\theta^2} + (t_i - x_i)^2 \frac{e^{-\theta(t_i - x_i)}}{(1 - e^{-\theta(t_i - x_i)})^2} \right\}.$$

The estimated values of  $\theta$  and  $\text{se}(\hat{\theta})$  for the criminological data are

$$\hat{\theta} = 0.0066, \text{ se}(\hat{\theta}) = 0.00033$$

and the maximum value of the log-likelihood function is  $\ell(\hat{\theta}) = -2875.207$ .

Now we can assess the independence assumption of  $X$  and  $Y$ . This can be done by plotting the estimated residuals

$$\{y - E(Y|X = x, X + Y \leq t)\} \text{ against } x, \quad (5.5)$$

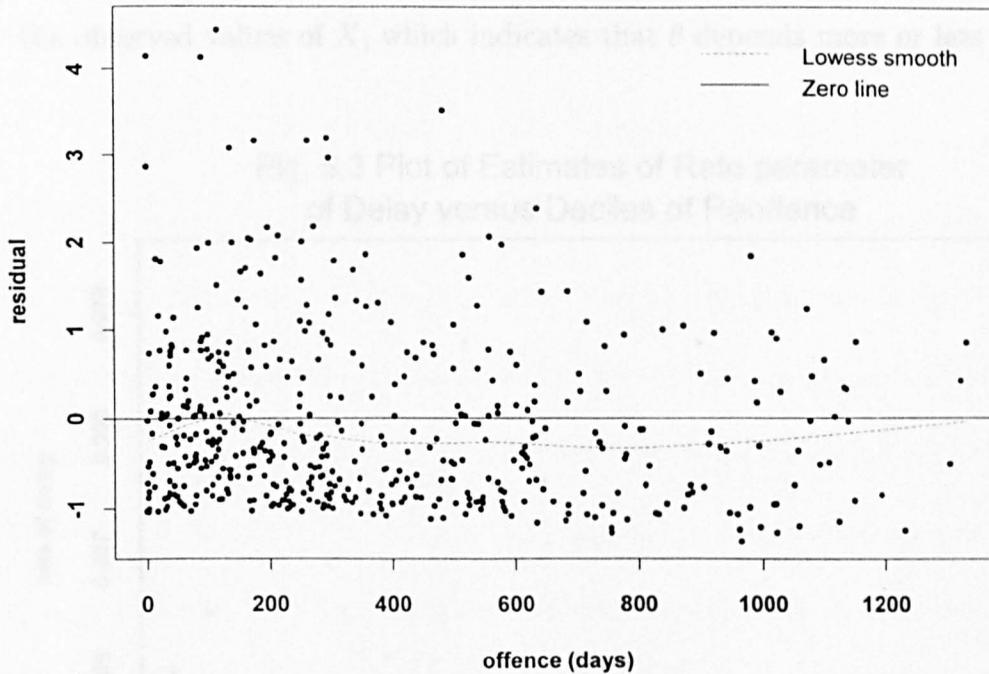
where the expectation here is given by equation (5.2). The conditional variance of the residual in expression (5.5) is given by equation (5.3) which depends on  $x$ , thus an alternative approach for checking the independence assumption between  $X$  and  $Y$  is to plot the estimated standardized residuals

$$\frac{\{y - E(Y|X = x, X + Y \leq t)\}}{\sqrt{\text{Var}(Y|X = x, X + Y \leq t)}}, \text{ against } x. \quad (5.6)$$

This plot is shown in Figure 5.2. The nonparametric regression line LOWESS (Cleveland, 1979) shown on the plot suggests a small decrease in the mean of  $Y$  as  $X$  increases and hence there is some dependence between the reoffence and the delay times but not very much (correlation -0.10, just significant at the nominal 5% level).

We shall now go on to consider an alternative possibility for assessing the independence assumption between  $X$  and  $Y$ . We can find the maximum

Fig. 5.2 Plot of Standardized Residuals of Delay versus Reoffence



likelihood estimates of  $\theta$  for different ranges of  $X$  to determine whether  $\theta$  is a function of  $X$ . If the estimated values of  $\theta$  are considerably different from each other, corresponding to different ranges of  $X$ , then this will suggest that  $\theta$  depends on  $X$ , and hence  $Y$  and  $X$  are correlated with each other, contrary to the independence assumption between  $X$  and  $Y$ .

Applying this method to the deciles of the observed values of  $X$ , the corresponding maximum likelihood estimates and standard errors of  $\theta$  (with standard errors in brackets) are

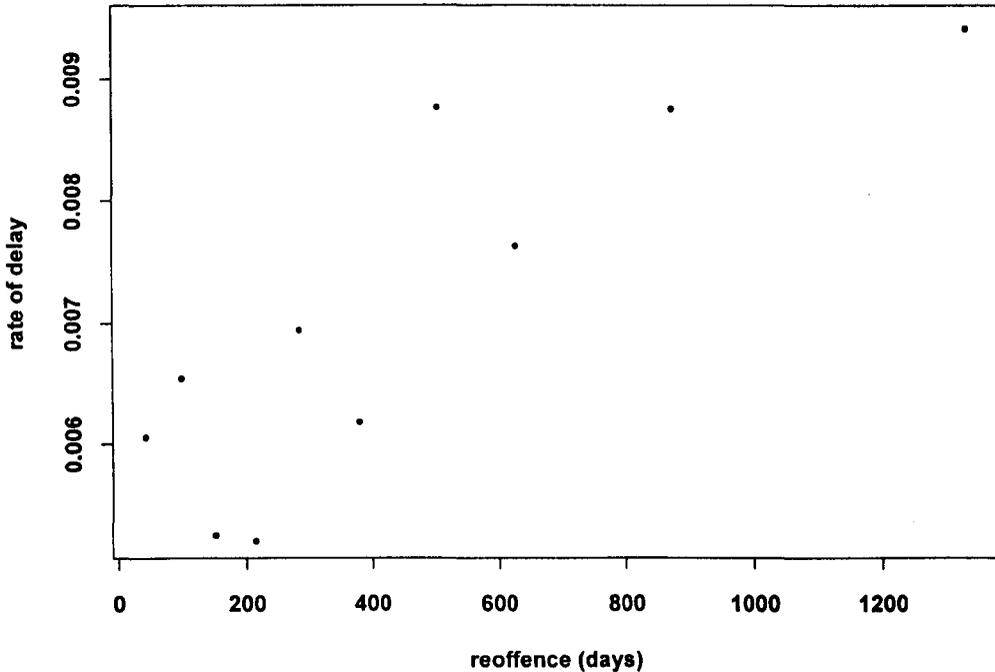
0.00607 (0.00088), 0.00656 (0.00094), 0.00526 (0.00081), 0.00522 (0.00079),

0.00696 (0.00102), 0.00619 (0.00093), 0.00878 (0.00129), 0.00765 (0.00118),

0.00877 (0.00142), 0.00944 (0.00216).

Figure 5.3 shows the plot of these estimated values of  $\theta$  against the deciles of the observed values of  $X$ , which indicates that  $\theta$  depends more or less on  $X$ .

Fig. 5.3 Plot of Estimates of Rate parameter of Delay versus Deciles of Reoffence



Also the plot suggests that we can fit a log-linear model such as

$$\log \theta = a + b(x - \bar{x}) \quad (5.7)$$

to the data, where  $a$  and  $b$  are scalars and  $\bar{x}$  is the mean of  $x$  over the sample. This model allows us to test the dependence of  $\theta$  upon the values of  $X$ , by testing the null hypothesis that  $b = 0$ . This can be done with a log-likelihood ratio test. Using equation (5.4) the conditional log-likelihood of  $Y$  under model (5.7) is given by

$$\ell = \sum_{i=1}^n (\ell_{1i} - \ell_{2i})$$

where

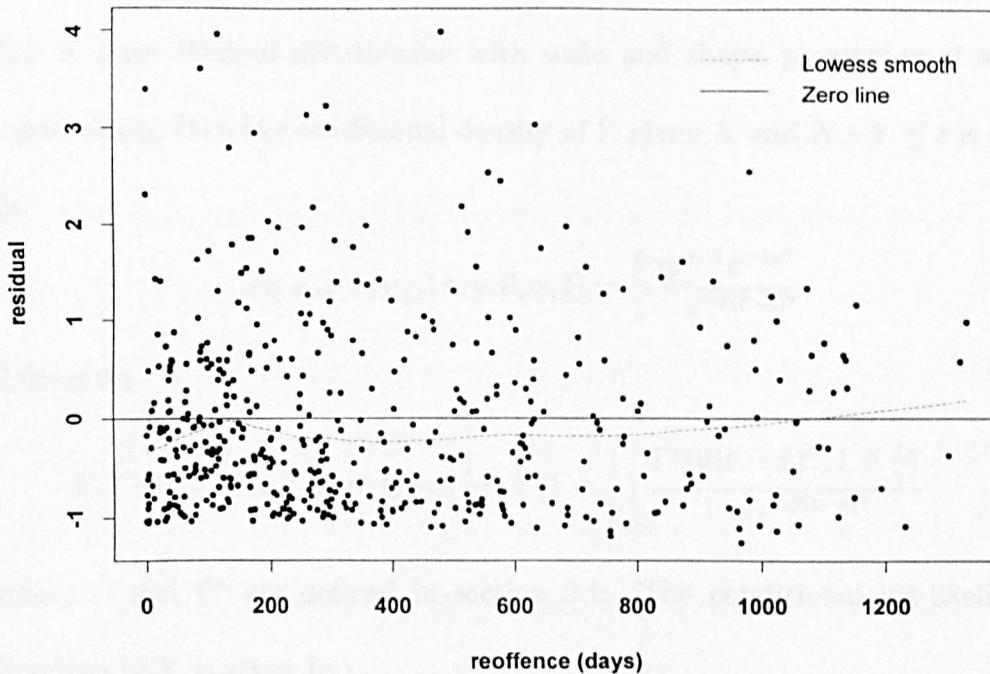
$$\ell_{1i} = a + b(x_i - \bar{x}) - y_i \exp\{a + b(x_i - \bar{x})\}$$

$$\ell_{2i} = \log\{1 - \exp\{-(t_i - x_i) \exp\{a + b(x_i - \bar{x})\}\}\}.$$

Fitting the model (5.7) to the data, the estimates of  $a$  and  $b$  are  $\hat{a} = -4.96$ ,  $\hat{b} = 0.0006$  and the corresponding maximum value of  $\ell$  is  $-2870.676$ . Now using the log-likelihood ratio test for testing the significance of  $b$  we obtain  $\Delta D = 9.062$ , which is statistically significant when compared with the  $\chi_1^2$  distribution. Accepting the model (5.7) and the exponential assumption for  $Y$ , we still need to test whether this model fits the data well in practice. To do this we repeat the plot 5.2 under the model (5.7). This is illustrated in Figure 5.4 and as before the nonparametric regression line LOWESS shown on the plot suggests that there is some dependence between the delay and the reoffence times. We also examined models similar to model (5.7) with quadratic, cubic and quartic terms. But these models did not make any improvement as compared to the fitted model (5.7).

Assuming exponential distribution for  $Y$ , all the models described here indicate that there is some dependence between  $X$  and  $Y$ . However, these models do not seem to fit the data well, as Figure 5.4 shows that, for a few points in the middle of the plot, there is some dispersion between the the nonparametric estimate of the regression line, the LOWESS line, and the zero residual line  $y = 0$ . Of course, all these analyses have been affected by the imputation of the delay times, described in section 3.6, a rather arbitrary

Fig. 5.4 Plot of Standardized Residuals of Delay versus Reoffence (model:[5.4])



aspect of this particular data set. This is a reason to study the nonparametric approach for our analysis.

Since Fig. 5.4 suggests that there may be some doubts about the validity of the conditional exponential distribution of  $Y$  given  $X = x$ , it is also worth exploring whether any more complicated parametric models might also be useful in explaining the dependence of  $Y$  on  $X$ .

In the following sections we shall do similar parametric analyses of truncated data by using Weibull, gamma and mixed exponential distributions for  $Y$ , rather than single exponential distribution.

## Weibull Distribution

Let  $Y$  have Weibull distribution with scale and shape parameters  $\theta$  and  $\gamma$  respectively, then the conditional density of  $Y$  given  $X$  and  $X + Y \leq t$  is given by

$$f_{Y|X=x, X+Y \leq t}(x, y, \theta, \gamma, t) = \frac{\theta \gamma y^{\gamma-1} e^{-\theta y^\gamma}}{1 - e^{-\theta(t-x)^\gamma}}.$$

This gives

$$E(Y|X = x, X + Y \leq t) = \left(\frac{1}{\theta}\right)^\gamma \Gamma\left(1 + \frac{1}{\gamma}\right) \frac{\Gamma^*(\theta(t-x)^\gamma, 1 + \frac{1}{\gamma})}{1 - e^{-\theta(t-x)^\gamma}} \quad (5.8)$$

where  $\Gamma$  and  $\Gamma^*$  are defined in section 2.1. The conditional log-likelihood function of  $Y$  is given by

$$\ell = \sum_{i=1}^n \{\log \theta + \log \gamma + (\gamma - 1) \log y_i - \theta y_i^\gamma - \log(1 - e^{-\theta(t_i - x_i)^\gamma})\}.$$

The following models were fitted to the truncated data:

$$\log \theta = a, \log \gamma = b \quad (5.9)$$

$$\log \theta = a_1 + a_2(x - \bar{x}), \log \gamma = a_3 \quad (5.10)$$

$$\log \theta = a_1 + a_2(x - \bar{x}), \log \gamma = a_3 + a_4(x - \bar{x}). \quad (5.11)$$

The estimates of the parameters and the value of the log-likelihood function for these models were found to be respectively,

$$\text{model (5.9) : } \hat{\theta} = 0.0068, \hat{\gamma} = 0.992, \ell(\hat{\theta}, \hat{\gamma}) = -2875.187,$$

$$\text{model (5.10) : } \hat{a}_1 = -5.08395, \hat{a}_2 = 0.00062, \hat{a}_3 = 0.02245,$$

$$\hat{\gamma} = 1.023, \ell(\hat{\theta}, \hat{\gamma}) = -2870.505$$

model (5.11) :  $\hat{a}_1 = -5.08392, \hat{a}_2 = 0.00049, \hat{a}_3 = 0.02294,$

$$\hat{a}_4 = 0.000026, \ell(\hat{\theta}, \hat{\gamma}) = -2870.485.$$

Note that in models (5.9) and (5.10), the shape parameter  $\gamma$  is nearly equal to 1, suggesting that the exponential distribution for  $Y$  seems to be adequate. We shall now apply the log-likelihood ratio test as follows. Using the models (5.7) and (5.10),  $\Delta D = 0.342$ , which is not statistically significant as compared with the  $\chi_1^2$  distribution. Again this suggests that the exponential assumption for  $Y$  is sensible. On the other hand using the models (5.9) and (5.10),  $\Delta D = 9.364$  and using the models (5.9) and (5.11),  $\Delta D = 9.404$ . In each case the value of  $\Delta D$  is statistically significant when compared with the  $\chi_1^2$  and  $\chi_2^2$  distributions. So we reject the null model (5.9) in favour of models (5.10) or (5.11). But using the models (5.10) and (5.11),  $\Delta D = 0.04$  which is not significant as compared with the  $\chi_1^2$  distribution. Thus there is no evidence to reject the model (5.10) when compared with the model (5.11). Note that as in model (5.7), the model (5.10) also indicates that  $\theta$  depends slightly on  $x$ , which could be seen graphically as well. This was done by plotting the estimated residuals in expression (5.5) against  $x$  under the model (5.10), where the expectation here was given by equation (5.8). Again the nonparametric regression line LOWESS showed that there is some dependence between  $X$  and  $Y$ .

## Gamma Distribution

If  $Y$  has gamma distribution with scale and shape parameters  $\theta$  and  $\gamma$ , then the conditional density of  $Y$  given  $X$  and  $X + Y \leq t$  is given by

$$f_{Y|X=x, X+Y \leq t}(x, y, \theta, \gamma, t) = \frac{\theta^\gamma y^{\gamma-1} e^{-\theta y}}{\Gamma(\gamma) \Gamma^*(\theta(t-x), \gamma)},$$

this gives

$$E(Y|X = x, X + Y \leq t) = \frac{\Gamma(1 + \gamma) \Gamma^*(\theta(t-x), 1 + \gamma)}{\theta \Gamma(\gamma) \Gamma^*(\theta(t-x), \gamma)}.$$

The conditional log-likelihood of  $Y$  is now given by

$$\ell = \sum_{i=1}^n \{\gamma \log \theta + (\gamma - 1) \log y_i - \theta y_i - \log \Gamma(\gamma) - \log(\Gamma^*(\theta(t-x), \gamma))\}.$$

Here the models (5.9), (5.10) and (5.11) were fitted to the truncated data and similar results as before were obtained, suggesting again that the exponential distribution for  $Y$  in the analysis of truncated data seems to be adequate. Also similar residual plots were examined which gave almost the same results as before for specifying the dependence of  $X$  and  $Y$ . For a good background discussion of lifetime data analysis, and more applications of graphical approaches using residuals, see the recent book by J. I. Ansell and M. J. Phillips (1994), chapters 1–3. There are also useful detailed discussions on choice of statistical models and dependency analysis in the same book, chapters 6–9.

## Mixed Exponential Distribution

Let  $Y$  be a mixed exponential random variable with density function

$$f_Y(y, \pi, \theta_1, \theta_2) = \pi \theta_1 e^{-\theta_1 y} + (1 - \pi) \theta_2 e^{-\theta_2 y} \quad (5.12)$$

and probability distribution function

$$F_Y(y) = P(Y \leq y) = \pi (1 - e^{-\theta_1 y}) + (1 - \pi) (1 - e^{-\theta_2 y})$$

where  $0 \leq \pi, \theta_1, \theta_2 \leq 1$ . Suppose that no distributional assumption is made about  $X$ , but  $X$  and  $Y$  are independent. Then the conditional density of  $Y$  given  $X$  and  $X + Y \leq t$  is equal to

$$P(Y|X = x, X + Y \leq t) = \frac{\pi\theta_1 e^{-\theta_1 y} + (1 - \pi)\theta_2 e^{-\theta_2 y}}{\pi(1 - e^{-\theta_1(t-x)}) + (1 - \pi)(1 - e^{-\theta_2(t-x)})}$$

and

$$E(Y|X = x, X + Y \leq t) = \frac{f_1(x, t, \pi, \theta_1) + f_2(x, t, \pi, \theta_2)}{f_3(x, t, \pi, \theta_1, \theta_2)}$$

where

$$\begin{aligned} f_1(x, t, \pi, \theta_1) &= \frac{\pi}{\theta_1} (1 - e^{-\theta_1(t-x)}) - \pi(t-x)e^{-\theta_1(t-x)} \\ f_2(x, t, \pi, \theta_2) &= \frac{(1-\pi)}{\theta_2} (1 - e^{-\theta_2(t-x)}) - (1-\pi)(t-x)e^{-\theta_2(t-x)} \\ f_3(x, t, \pi, \theta_1, \theta_2) &= \pi(1 - e^{-\theta_1(t-x)}) + (1-\pi)(1 - e^{-\theta_2(t-x)}). \end{aligned}$$

Note that in the limit as  $t \rightarrow \infty$ ,

$$E(Y|X = x, X + Y \leq t) \rightarrow \frac{\pi}{\theta_1} + \frac{1-\pi}{\theta_2} = E(Y).$$

Also, when  $\theta_1 = \theta_2 = \theta$ ,

$$E(Y|X = x, X + Y \leq t) = \frac{1}{\theta} - \frac{(t-x)}{e^{\theta(t-x)} - 1},$$

as expected. The conditional log-likelihood function of  $Y$  is given by

$$\ell = \sum_{i=1}^n \log \left\{ \frac{\pi\theta_1 e^{-\theta_1 y_i} + (1-\pi)\theta_2 e^{-\theta_2 y_i}}{\pi(1 - e^{-\theta_1(t_i-x_i)}) + (1-\pi)(1 - e^{-\theta_2(t_i-x_i)})} \right\}$$

Here, the estimated values of the parameters are  $\hat{\theta}_1=0.0065$   $\hat{\theta}_2=0.6219$ ,  $\hat{\pi}=0.99$ , and the corresponding maximum value of the log-likelihood function  $\ell$ , is -2874.651. The proximity of  $\hat{\pi}$  to 1 suggests that the consideration of single exponential distribution for  $Y$  is probably sensible. In fact it can be shown that this result is true not only for the truncated data but also for all the data, including both uncensored and censored cases. To do this we proceed as follows.

Suppose that  $Y$  is a mixed exponential random variable with density function defined by equation (5.12), and  $X$  has a single exponential distribution with density

$$f_X(x, \lambda) = \lambda e^{-\lambda x}, \quad x, \lambda > 0.$$

If  $X$  and  $Y$  are independent, then it can be shown that the density function of  $Z = X + Y$  is given by

$$f_Z(z, \beta) = f_{Z_1}(z, \beta_1) + f_{Z_2}(z, \beta_2)$$

where

$$\beta = (\pi, \lambda, \theta_1, \theta_2), \quad \beta_1 = (\pi, \lambda, \theta_1), \quad \beta_2 = (\pi, \lambda, \theta_2)$$

and

$$f_{Z_1}(z, \beta_1) = \frac{\pi \lambda \theta_1}{\theta_1 - \lambda} (e^{-\lambda z} - e^{-\theta_1 z})$$

$$f_{Z_2}(z, \beta_2) = \frac{(1 - \pi) \lambda \theta_2}{\theta_2 - \lambda} (e^{-\lambda z} - e^{-\theta_2 z}).$$

This gives

$$P(Z > t) = S_{Z_1}(t, \beta_1) + S_{Z_2}(t, \beta_2)$$

where  $t$  is the time to follow-up and

$$S_{Z1}(t, \beta_1) = \frac{\pi(\theta_1 e^{-\lambda t} - \lambda e^{-\theta_1 t})}{\theta_1 - \lambda}$$

$$S_{Z2}(t, \beta_2) = \frac{(1 - \pi)(\theta_2 e^{-\lambda t} - \lambda e^{-\theta_2 t})}{\theta_2 - \lambda}.$$

Now using the split population model given by  $p, \lambda, \pi, \theta_1, \theta_2$ , the log-likelihood function for the full data is given by

$$\ell = \sum_{i=1}^n \ell_{1i} + \sum_{i=n+1}^N \ell_{2i}$$

where the first sum is over the uncensored cases and the second sum is over the censored cases and

$$\ell_{1i} = \log p + \log \lambda - \lambda x_i + \log\{f_Y(y_i, \pi, \theta_1, \theta_2)\}$$

$$\ell_{2i} = \log\{1 - p + p (S_{Z1}(t_i, \beta_1) + S_{Z2}(t_i, \beta_2))\}.$$

Note that in order to get the maximum likelihood estimates of  $p, \lambda, \pi, \theta_1, \theta_2$ , it is more convenient to use the following parameterizations

$$\log\left(\frac{p}{1-p}\right) = a_1, \log \lambda = a_2, \log\left(\frac{\pi}{1-\pi}\right) = a_3, \log \theta_1 = a_4, \log \theta_2 = a_5.$$

Fitting the split population model to the data, we get  $\hat{p}=0.47$ ,  $\hat{\lambda}=0.002$ ,  $\hat{\pi}=0.98$ ,  $\hat{\theta}_1= 0.0066$ ,  $\hat{\theta}_2=0.049$ . Again the proximity of  $\hat{\pi}$  to 1 suggests that the single exponential distribution for  $Y$  is reasonable.

## 5.2.2 Semi-parametric Analysis

Doubts about exponential assumption under parametric analysis, discussed in subsection 5.2.1, suggests that we need a nonparametric approach for the

estimation of the truncated data. To start with, for simplicity, suppose that  $Y$  is a discrete random variable with values  $0, 1, \dots, k$  and assume that

$$P(Y = j) = p_j, \quad j = 0, 1, \dots, k$$

$$P(Y \leq i) = \sum_{j=0}^i p_j = q_i.$$

If  $Y$  is truncated at  $i$  from the right, then

$$P(Y = j|Y \leq i) = \frac{p_j}{q_i}, \quad i = 0, 1, \dots, k, \quad j \leq i. \quad (5.13)$$

Suppose the data consist of the pairs  $(i, j), j \leq i$  with multiplicity  $n_{ij}$ , that is  $n_{ij}$  is the number of times that we observe the values of  $i$  and  $j$  simultaneously in all possible combinations of  $i$  and  $j$  and  $\sum_{i,j} n_{ij}$  is equal to the sample size. A natural and more convenient way of looking at these observed values of  $n_{ij}$  is to arrange them into a matrix configuration. To do this let  $N$  be a  $(k + 1) \times (k + 1)$  lower triangular matrix with

$$n_{ij} = (i + 1, j + 1)\text{th entry of } N, \quad i = 0, \dots, k, \quad j \leq i.$$

Now on the basis of the conditional sample  $(i, j), j \leq i$ , the problem considered is to find the maximum likelihood estimates of  $p_0, p_1, \dots, p_k$  (the unconditional probabilities of  $Y$  at  $0, 1, \dots, k$ ) or equivalently to find the maximum likelihood estimates of  $q_0, q_1, q_2, \dots, q_k$  and hence the probability distribution function of  $Y$ . The likelihood function is given by

$$L = \prod_{i=0}^k \prod_{j=0}^i \left( \frac{p_j}{q_i} \right)^{n_{ij}} \quad (5.14)$$

and the conditional log-likelihood function is

$$\ell = \sum_{i=0}^k \sum_{j=0}^i n_{ij} (\log p_j - \log q_i). \quad (5.15)$$

Now define

$$n_i = \sum_{j=0}^i n_{ij}, \quad m_j = \sum_{i=j}^k n_{ij} \quad (5.16)$$

$$\alpha_j = \sum_{i=j}^k n_i - \sum_{i=j+1}^k m_i, \quad (5.17)$$

then it can be shown that

$$\frac{\partial \ell}{\partial p_j} = \frac{m_j}{p_j} - \sum_{i=j}^{k-1} \left( \frac{n_i}{q_i} \right), \quad j = 0, 1, \dots, k-1$$

$$\frac{\partial \ell}{\partial p_k} = \frac{n_{kk}}{p_k}.$$

Note that in terms of rows and columns of the counting matrix  $N$ ,  $n_i$  is sum of the entries of  $(i+1)$ th row and  $m_j$  is sum of the entries of  $(j+1)$ th column.

In order to maximize  $\ell$ , we use the Lagrange multiplier method, with

$$g(p_0, p_1, \dots, p_k) = \sum_{j=0}^k p_j = 1$$

and we need to solve

$$\frac{\partial \ell}{\partial p_j} = c \frac{\partial g}{\partial p_j}, \quad j = 0, 1, \dots, k.$$

But  $\frac{\partial g}{\partial p_j} = 1$  for all  $j$ , thus

$$\frac{\partial \ell}{\partial p_j} = \frac{\partial \ell}{\partial p_{j+1}}, \quad j = 0, 1, \dots, k-1. \quad (5.18)$$

After some calculations it can be shown that  $c = n_k$ , the sum of the  $(k+1)$ th row of matrix  $N$  and

$$p_j = \frac{m_j}{\alpha_j} q_j, \quad j = 0, 1, \dots, k. \quad (5.19)$$

Equation (5.19) can most easily be justified by mathematical induction as follows.

By definition

$$\alpha_0 = m_0, \quad q_0 = p_0 \quad (5.20)$$

and we can write

$$\alpha_{r+1} = \alpha_r + m_{r+1} - n_r \quad (5.21)$$

and in particular

$$m_0 - n_0 = \alpha_1 - m_1. \quad (5.22)$$

Now using equation (5.18) with  $j = 0$  and then substituting from equations (5.20) and (5.22) we get

$$p_0 = \left( \frac{\alpha_1}{m_1} - 1 \right) p_1. \quad (5.23)$$

Also by definition

$$q_1 = p_0 + p_1. \quad (5.24)$$

Combining equations (5.23) and (5.24) yields

$$p_1 = \frac{m_1}{\alpha_1} q_1.$$

Thus, equation (5.19) is true for  $j = 1$ . Now assume that equation (5.19) is true for  $j = r$ , that is

$$p_r = \frac{m_r}{\alpha_r} q_r. \quad (5.25)$$

To prove equation (5.19) for  $j = r + 1$ , we proceed as follows. Using equation (5.18) for  $j = r$  and then substituting from equation (5.25), we get

$$p_r = \frac{(\alpha_r - n_r) m_r}{\alpha_r m_{r+1}} p_{r+1}. \quad (5.26)$$

Again by definition

$$q_{r+1} = q_r + p_{r+1}. \quad (5.27)$$

Now combining equations (5.21), (5.25), (5.26) and (5.27) together we have

$$p_{r+1} = \frac{m_{r+1}}{\alpha_{r+1}} q_{r+1}.$$

Therefore, by mathematical induction, equation (5.19) is true for all  $j$ .

We have by definition

$$q_{j-1} = q_j - p_j. \quad (5.28)$$

Thus, using equations (5.19) and (5.28), the maximum likelihood estimate of  $q_j$  satisfies

$$\hat{q}_{j-1} = \hat{q}_j \left(1 - \frac{m_j}{\alpha_j}\right), \quad j = 0, 1, \dots, k \quad (5.29)$$

which can be solved by iteration. Once equation (5.29) is solved for all  $j$ , we can estimate the distribution of  $Y$ . Note that  $q_k = 1$  might be used as an initial value for this iteration process. But when  $q_k = 1$  is considered as an initial value for iteration, we are actually assuming that

$$q_{k+1} = q_{k+2} = \dots = 1 \text{ or } p_{k+1} = p_{k+2} = \dots = 0. \quad (5.30)$$

This may be a disturbing property of the estimators, and may lead to unreasonable estimates. The theory developed here works only for discrete random variables with finite ranges. Thus, the value of  $q_k=1$  is all right if  $Y$  really is a discrete random variable with a finite range. But in the applications we are using this theory for a data set which may not have a definite upper bound. So in practice we need to take  $k$  sufficiently large so that  $P(Y > k) =$

$1 - q_k$  is very small, hence  $q_k$  tends to 1. Another possibility is to choose a suitable value of  $k$  and estimate  $q_k$  from a parametric model.

### Estimating Distribution of Delay

To illustrate an application of model (5.29), we shall now consider our criminological truncated data. Let  $(y_1, w_1), \dots, (y_n, w_n)$  be a random sample from  $Y$  and  $W = T - X$ ,  $T$  being the time to follow-up, with  $y_i \leq w_i, i = 1, \dots, n$ , and assume that  $Y$  and  $X$  are independent. Then

$$P(Y = y_j | X = x_i, X + Y \leq t_i) = P(Y = y_j | Y \leq w_i) \quad (5.31)$$

where  $w_i = t_i - x_i$  is the truncation point. If we compare equations (5.31) and (5.13), then corresponding to  $i$ , in the left hand side of equation (5.13), there is now  $w_i = t_i - x_i$  in equation (5.31) and corresponding to  $j$  we have now  $y_j$  in equation (5.31). In order to apply model (5.29) to the truncated data we proceed as follows:

(i) we integerize the original observed values of  $Y$  in the sample by grouping the data into class intervals. For instance, we may define

$$y' = \text{round}((y - 1)/2). \quad (5.32)$$

Under transformations such as expression (5.32), the observed values of  $Y$  in the sample will be converted into the form  $0, 1, \dots, k$ , where  $k$  is the maximum observed value of  $Y$  after the integerizing process. Similarly, we integerize the observed values of  $W$  in the sample and denote this set by  $w'$ . Of course some of the values of  $w'$  might be bigger than  $k$ . We replace by  $k$  those elements

of  $w'$  which are bigger than  $k$ , and leave the same those which are not and denote this set by  $w''$ . Note that the choice of  $w''$  is sensible. To see this let  $i > k$ , then if  $Y \leq i$  we have  $Y \leq k$ , because  $P(Y > k) = 0$ . Thus  $P(Y = j|Y \leq i) = P(Y = j|Y \leq k)$ , which implies that defining  $w''$  as before is reasonable.

(ii) Using the values of  $y'$  and  $w''$  obtained in part (i), we generate the counting matrix  $N$  in the usual way with

$$n_{ij} = \#\{w'' = i, y' = j\}, \quad i = 0, \dots, k, \quad j \leq i,$$

where  $\#A$  denotes the cardinality of a set  $A$ . Here the values of  $n_i$  and  $m_j$ , with the same definitions as in equation (5.16), can be written as

$$n_i = \#\{w'' = i\}, \quad m_j = \#\{y' = j\}.$$

We now evaluate  $\alpha_j$  by equation (5.17) and hence estimate  $\hat{q}_j$  from equation (5.29) provided a proper initial value is chosen for  $q_k$ . As mentioned before the value of  $q_k = 1$  may lead to unreasonable estimates. To overcome the problem of  $q_k = 1$ , we can choose

$$\hat{q}_k = 1 - e^{-\hat{\theta}k},$$

$\hat{\theta}$  is the maximum likelihood estimate of  $\theta$ , obtained from the observed values of  $Y$  after the integerizing process and under the truncated distribution of  $Y$  defined by equation (5.1), which we found to be  $\hat{\theta}=0.0133$ .

Once we have obtained the maximum likelihood estimates of  $q_0, \dots, q_k$ , we can examine the unconditional distribution of  $Y$ . This can be done by plotting

$\hat{q}_j$  against  $j$ , for  $j = 0, \dots, k$ , or equivalently plotting  $\log(1 - \hat{q}_j)$  versus  $j$  as a diagnostic plot for checking the unconditional distribution of  $Y$ . If distribution of  $Y$  is exponential then the plot should be approximately a straight line.

For the criminological data, the transformation (5.32) gives class intervals  $[0, 2]$ ,  $[2, 4]$ ,  $\dots$ , and the value of  $k$  here is 410. The diagnostic plot is illustrated in Figure 5.5 which is fairly close to a straight line, suggesting that the unconditional distribution of  $Y$  is exponential. Of course, by considering the conditional distribution of (observed)  $X$  given  $Y$  which is induced by the censoring, we can do similar analysis for  $X$  and obtain the formula (5.29) but now for  $X$ , and consequently examine the unconditional distribution of  $X$ . The diagnostic plot for checking the unconditional distribution of  $X$  is pictured in Figure 5.6. This is close to a straight line, again implying that the exponential distribution for  $X$  is sensible.

Fig. 5.5 Marginal Diagnostics for Delay Time

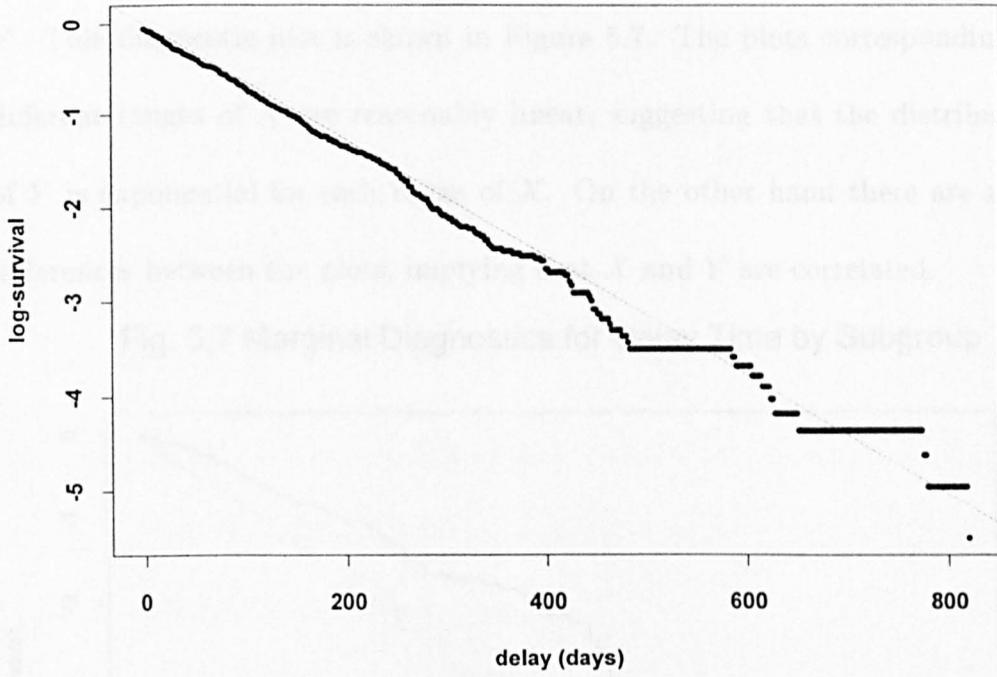
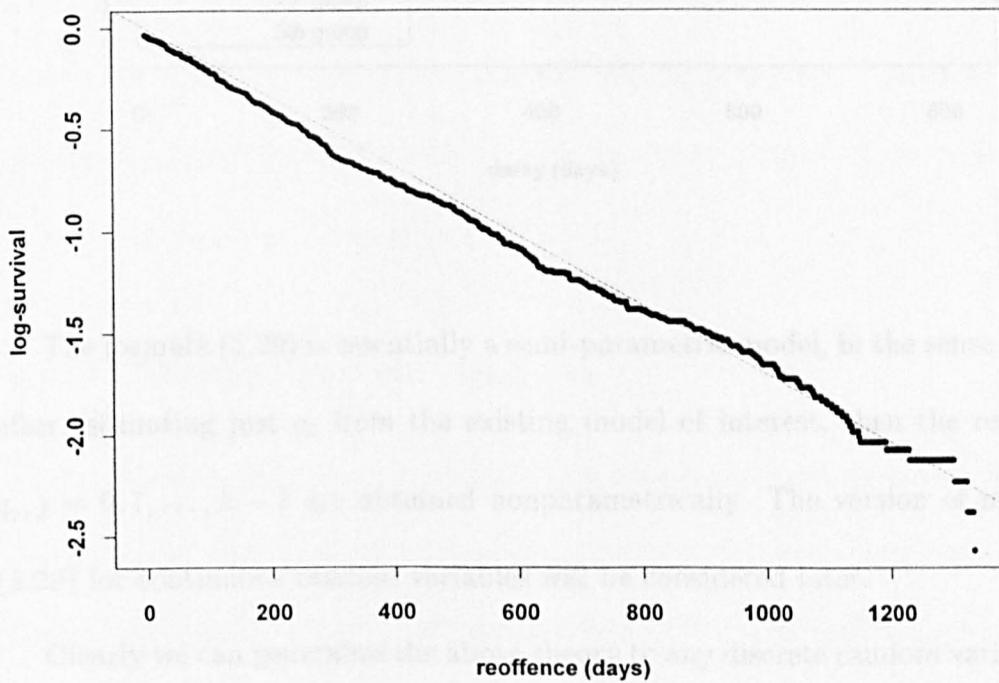
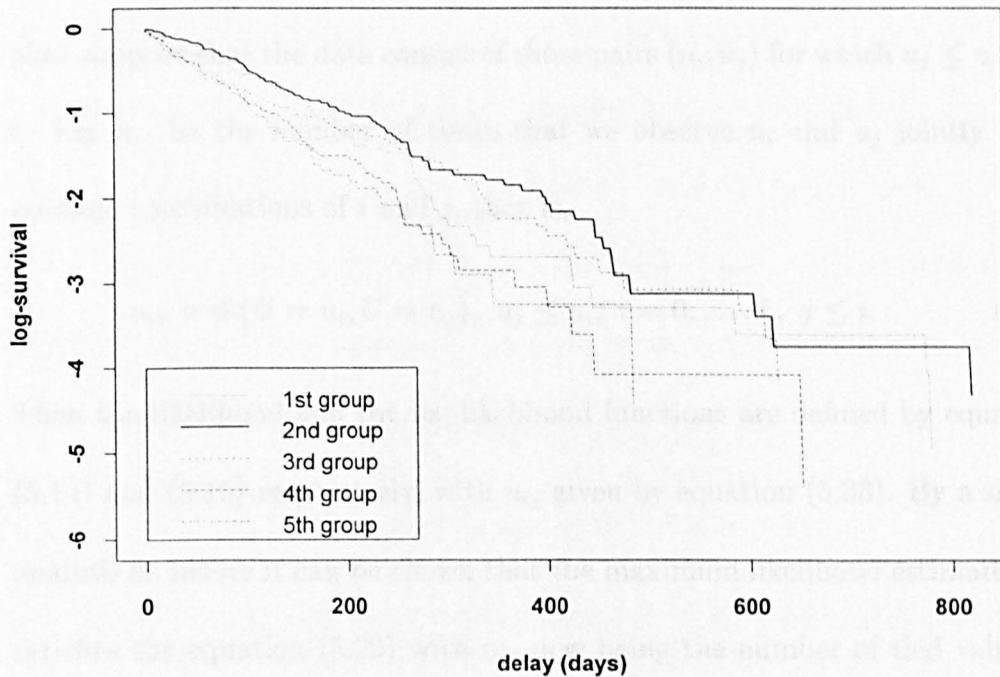


Fig. 5.6 Marginal Diagnostics for Reoffence Time



We shall now go on to repeat the diagnostic plot in Figure 5.5 for different ranges of  $X$ , in order to specify the problem of independence between  $X$  and  $Y$ . This diagnostic plot is shown in Figure 5.7. The plots corresponding to different ranges of  $X$  are reasonably linear, suggesting that the distribution of  $Y$  is exponential for each range of  $X$ . On the other hand there are some differences between the plots, implying that  $X$  and  $Y$  are correlated.

Fig. 5.7 Marginal Diagnostics for Delay Time by Subgroup



The formula (5.29) is essentially a semi-parametric model, in the sense that after estimating just  $q_k$  from the existing model of interest, then the rest of  $q_j, j = 0, 1, \dots, k - 1$  are obtained nonparametrically. The version of model (5.29) for continuous random variables will be considered later.

Clearly we can generalize the above theory to any discrete random variable,

not just a variable taking integer values. Suppose that  $U$  is a discrete random variable which assumes distinct values  $u_0 < u_1 < \dots < u_k$  and let  $m_j$  be the number of tied values at  $u = u_j$   $j = 0, \dots, k$ ,  $\sum_{j=0}^k m_j = n$ . Let

$$P(U = u_j) = p_j, \quad P(U \leq u_i) = \sum_{j=0}^i p_j = q_i,$$

for  $i = 0, \dots, k$ ,  $j \leq i$ . Then

$$P(U = u_j | U \leq u_i) = \frac{p_j}{q_i}.$$

Now suppose that the data consist of those pairs  $(u_i, u_j)$  for which  $u_j \leq u_i$ ,  $j \leq i$ . Let  $n_{ij}$  be the number of times that we observe  $u_i$  and  $u_j$  jointly in all possible combinations of  $i$  and  $j$ , that is,

$$n_{ij} = \#\{U = u_i, U = u_j\}, \quad u_j \leq u_i, \quad i = 0, \dots, k, \quad j \leq i. \quad (5.33)$$

Then the likelihood and the log-likelihood functions are defined by equations (5.14) and (5.15) respectively, with  $n_{ij}$  given by equation (5.33). By a similar analysis as before it can be shown that the maximum likelihood estimate of  $q_j$  satisfies the equation (5.29) with  $m_j$  now being the number of tied values at  $u = u_j$  and  $\alpha_j$  given by equation (5.17).

### 5.2.3 Nonparametric Analysis

As mentioned before, when model (5.29) is used for estimating a distribution function we need to generate the counting matrix  $N$ . This can be done easily as long as the dimension of  $N$  is not too big. But as the number of observations is increased or when we are dealing with continuous data the matrix  $N$  becomes

very large and it requires too much computer memory. This suggests that we need to generalize model (5.29) and obtain a formula which is independent of using the counting matrix  $N$ . This can be done if we are able to express  $\alpha_j$  in equation (5.29) in terms of empirical distributions of the relevant random variables involved in the truncated data analysis. The basic theory for doing this has been described by Woodroffe, M. (1985). Other relevant references in this area include Wang, M.C., Jewell, N.P. and Tsai, W.Y. (1986), Keiding, N. and Gill, R.D. (1990) and van der Lann, M.J. (1995).

Following Woodroffe, let  $G$  and  $F$  denote unconditional distribution functions of two independent random variables  $Y$  and  $W$  respectively, where  $G$  and  $F$  are completely unknown. Also let  $G^*$  and  $F^*$  denote the marginal conditional distribution functions of  $Y$  and  $W$  given  $Y \leq W$  respectively. Thus,

$$G^*(u) = P(Y \leq u | Y \leq W),$$

$$F^*(u) = P(W \leq u | Y \leq W), \quad 0 \leq u < \infty.$$

Let  $(y_1, w_1), \dots, (y_n, w_n)$  be a random sample of size  $n$  from  $(Y, W)$  for which  $y_i \leq w_i$ ,  $i = 1, \dots, n$ . So we are sampling from the joint conditional distribution of  $((Y, W) | Y \leq W)$  instead of the distribution of  $(Y, W)$  itself. The problem considered is the nonparametric estimation of  $G$ . Of course we can estimate  $F$  nonparametrically as well.

To describe the estimator of  $G$ , let  $G_n^*$  and  $F_n^*$  denote the empirical distributions of  $y_1, \dots, y_n$  and  $w_1, \dots, w_n$  respectively, then

$$G_n^*(u) = \left(\frac{1}{n}\right) \#\{i \leq n : y_i \leq u\} = \frac{1}{n} \sum_{i=1}^n I(y_i \leq u),$$

$$F_n^*(u) = \left(\frac{1}{n}\right) \#\{i \leq n : w_i \leq u\} = \frac{1}{n} \sum_{i=1}^n I(w_i \leq u), \quad 0 \leq u < \infty,$$

where as before  $\#A$  denotes the cardinality of a set  $A$  and  $I$  is the usual indicator function. Thus,  $G_n^*$  and  $F_n^*$  estimate the conditional distribution functions  $G^*$  and  $F^*$  respectively. To construct an estimator of  $G$  from  $G_n^*$  and  $F_n^*$ , we proceed as follows. Let

$$\alpha_n(u) = \sum_{i=1}^n I(y_i \leq u, w_i > u), \quad 0 \leq u < \infty,$$

which can be written as

$$\alpha_n(u) = \sum_{i=1}^n I(y_i \leq u) - \sum_{i=1}^n I(w_i \leq u).$$

Thus

$$\alpha_n(u) = n D_n(u),$$

where

$$D_n(u) = G_n^*(u) - F_n^*(u). \quad (5.34)$$

Let  $m(y_j)$  be the number of tied values at  $y = y_j$ ,

$$m(y_j) = \sum_{i=1}^n I(y_i = y_j), \quad 1 \leq j \leq n, \quad (5.35)$$

and

$$D^*(u) = G^*(u) - F^*(u), \quad 0 \leq u < \infty. \quad (5.36)$$

Also let

$$q_{j-1} = P(Y \leq y_{j-1}) = G(y_{j-1})$$

and

$$q_j = P(Y \leq y_j) = G(y_j).$$

Now assume that  $\hat{G}_n$ , the maximum likelihood estimator of  $G$ , exists. Then we have

$$\hat{q}_{j-1} = \hat{G}_n(y_{j-1}), \quad \hat{q}_j = \hat{G}_n(y_j),$$

where  $\hat{q}_{j-1}$  and  $\hat{q}_j$  are the maximum likelihood estimates of  $q_{j-1}$  and  $q_j$  respectively. On the other hand we know that if  $\hat{G}_n$  exists, then it satisfies equation (5.29). Thus,

$$\hat{G}_n(y_{j-1}) = \hat{G}_n(y_j) \left\{ 1 - \frac{m(y_j)}{\alpha_n(y_j)} \right\},$$

or

$$\frac{d\hat{G}_n(y_j)}{\hat{G}_n(y_j)} = \frac{m(y_j)}{nD_n(y_j)},$$

where

$$d\hat{G}_n(y_j) = \hat{G}_n(y_j) - \hat{G}_n(y_{j-1}),$$

and  $D_n(y_j)$ ,  $m(y_j)$  are defined by equations (5.34) and (5.35) respectively. Now define the function  $\phi$  by

$$\phi(v) = - \int_v^\infty \frac{dG(u)}{G(u)}, \quad 0 \leq v < \infty. \quad (5.37)$$

Since  $\hat{G}_n$  (MLE of  $G$ ) estimates  $G$ , then equation (5.37) suggests estimating  $\phi$  by

$$\hat{\phi}_n(v) = - \int_v^\infty \frac{d\hat{G}_n(y)}{\hat{G}_n(y)}, \quad 0 \leq v < \infty.$$

Thus,

$$d\hat{\phi}_n(v) = \frac{d\hat{G}_n(v)}{\hat{G}_n(v)}, \quad (5.38)$$

which gives

$$d\hat{\phi}_n(y_j) = \frac{d\hat{G}_n(y_j)}{\hat{G}_n(y_j)} = \frac{m(y_j)}{nD_n(y_j)}, \quad 1 \leq j \leq n.$$

On the other hand, we have always

$$\hat{G}_n(v) = 1 - \int_v^\infty d\hat{G}_n(y). \quad (5.39)$$

Substituting from equation (5.38) into equation (5.39) gives

$$\hat{G}_n(v) = 1 + \int_v^\infty \hat{G}_n(y) d(-\hat{\phi}_n(y)), \quad (5.40)$$

where

$$d(-\hat{\phi}_n(y)) = -d\hat{\phi}_n(y).$$

Since  $\hat{G}_n$  satisfies the integral equation (5.40), then by product-integration or product-limit method, see Andersen *et al.* (1993), it follows that  $\hat{G}_n$  is the product integral of  $-\hat{\phi}_n$  over the interval  $(v, \infty)$ . Thus,

$$\hat{G}_n(v) = \mathcal{P}_v^\infty \{1 + d(-\hat{\phi}_n(y))\} = \mathcal{P}_v^\infty \{1 - d\hat{\phi}_n(y)\},$$

or

$$\hat{G}_n(v) = \prod_{j:y_j > v} \{1 - d\hat{\phi}_n(y_j)\} = \prod_{j:y_j > v} \left\{1 - \frac{m(y_j)}{nD_n(y_j)}\right\}, \quad (5.41)$$

where the product is taken over distinct values of  $y_1, y_2, \dots, y_n$  and  $m(y_j)$  is the number of tied values at  $y = y_j$ , and if there is no  $y_j > v$  the product is taken to be 1. Note that for continuous data where there are no ties,  $m(y_j) = 1$  for all  $j$ . The function  $\hat{G}_n$ , defined by equation (5.41) is the nonparametric maximum likelihood estimator of the distribution function  $G$  under random truncation. The estimator is the analogue of the product-limit estimator of Kaplan-Meier for randomly censored data.

The estimator  $\hat{G}_n$  is a step function and

$$\hat{G}_n(v) = 1, \text{ if } v > \max\{y_j\}, \quad j = 1, \dots, n.$$

Since we are estimating  $G$  by  $\hat{G}_n$ , then

$$G(v) = 1, \text{ for } v > \max\{y_j\}, j = 1, \dots, n,$$

implying that

$$P(Y > \max\{y_j\}) = 0, j = 1, \dots, n.$$

This is similar to the problem of  $q_k=1$ , described in the semi-parametric analysis of the truncated data in subsection 5.2.2. Clearly this is not sensible if the censoring is very heavy and may lead to unreasonable estimates for those values of  $v > \max\{y_j\}, j = 1, \dots, n$ .

Another nonparametric estimator of  $G$  can be constructed as follows. If the function  $\phi$  is defined by equation (5.37), then it can be shown that

$$\phi(v) = - \int_v^\infty \frac{dG^*(u)}{D^*(u)}, 0 \leq v < \infty, \quad (5.42)$$

where  $D^*(u)$  is given by equation (5.36). Since  $G_n^*$  estimates  $G^*$ , then equation (5.42) suggests estimating the function  $\phi$  by

$$\tilde{\phi}_n(v) = - \int_v^\infty \frac{dG_n^*(y)}{D_n(y)} = - \sum_{j:y_j > v} \frac{m(y_j)}{nD_n(y_j)}, 0 \leq v < \infty, \quad (5.43)$$

where the summation is taken over distinct values of  $y_1, \dots, y_n$ . Note that because  $G_n^*$  is a step function it follows that  $\tilde{\phi}_n$  is also a step function with discontinuities at  $y_1, \dots, y_n$ . Solving the equation (5.37) for  $G$  we have

$$G(v) = e^{\phi(v)}. \quad (5.44)$$

Thus, by substituting from equation (5.43) into equation (5.44), a natural estimator of  $G$  is given by

$$\tilde{G}_n(v) = e^{\tilde{\phi}_n(v)} = e^{-\sum_{j:y_j > v} \left\{ \frac{m(y_j)}{nD_n(y_j)} \right\}}, 0 \leq v < \infty. \quad (5.45)$$

The two estimators  $\hat{G}_n$  and  $\tilde{G}_n$  are in fact asymptotically equivalent. For if  $C_j$  is small

$$e^{-C_j} \approx (1 - C_j),$$

and so

$$\prod_j (1 - C_j) \approx e^{-\sum_j C_j}. \quad (5.46)$$

Now in equation (5.46) let

$$C_j = \frac{m(y_j)}{nD_n(y_j)},$$

and observe that as the sample size  $n$  increases then  $C_j$  becomes small. Therefore,

$$\prod_{j:y_j > v} \left\{ 1 - \frac{m(y_j)}{nD_n(y_j)} \right\} \approx e^{-\sum_{j:y_j > v} \left\{ \frac{m(y_j)}{nD_n(y_j)} \right\}},$$

that is, for large  $n$ ,

$$\hat{G}_n(v) \approx \tilde{G}_n(v), \quad 0 \leq v < \infty.$$

To prove that

$$\int_v^\infty \frac{dG^*(u)}{D^*(u)} = \int_v^\infty \frac{dG(u)}{G(u)},$$

let

$$\alpha = P(Y \leq W),$$

then

$$\alpha = \int_0^\infty \int_0^w dG(y) dF(w) = \int_0^\infty G(w) dF(w),$$

or

$$\alpha = \int_0^\infty \int_y^\infty dF(w) dG(y) = \int_0^\infty (1 - F(y)) dG(y).$$

Also by definition we have

$$G^*(y) = P(Y \leq y | Y \leq W) = \alpha^{-1} P(Y \leq y, Y \leq W),$$

or

$$G^*(y) = \alpha^{-1} \int_0^y \int_v^\infty dF(u) dG(v) = \alpha^{-1} \int_0^y (1 - F(v)) dG(v). \quad (5.47)$$

Similarly,

$$F^*(w) = P(W \leq w | Y \leq W) = \alpha^{-1} P(W \leq w, Y \leq W),$$

or

$$F^*(w) = \alpha^{-1} \int_0^w \int_0^u dG(v) dF(u) = \alpha^{-1} \int_0^w G(u) dF(u). \quad (5.48)$$

Now substituting from equations (5.47) and (5.48) into equation(5.36) we get

$$D^*(u) = G^*(u) - F^*(u) = \alpha^{-1} \int_0^u d[G(v) (1 - F(v))],$$

or

$$D^*(u) = \alpha^{-1} G(u) (1 - F(u)).$$

Therefore,

$$\int_v^\infty \frac{dG^*(u)}{D^*(u)} = \int_v^\infty \frac{\alpha^{-1} dG(u) (1 - F(u))}{\alpha^{-1} G(u) (1 - F(u))} = \int_v^\infty \frac{dG(u)}{G(u)}.$$

Of course, similar derivations are possible for the estimation of  $F$ . If we consider the cumulative hazard function  $\psi$  which is defined by

$$\psi(v) = \int_0^v \frac{dF(u)}{1 - F(u)}, \quad (5.49)$$

then it can be shown that

$$\psi(v) = \int_0^v \frac{dF^*(u)}{D^*(u)}. \quad (5.50)$$

Using equations (5.49), (5.50) and after details of algebra, we get

$$1 - \hat{F}_n(v) = \prod_{j:w_j \leq v} \left\{ 1 - \frac{m(w_j)}{nD_n(w_j)} \right\}, \quad (5.51)$$

and

$$1 - \tilde{F}_n(v) = e^{-\sum_{j:w_j \leq v} \left\{ \frac{m(w_j)}{nD_n(w_j)} \right\}}, \quad (5.52)$$

where  $m(w_j)$  is the number of tied values at  $w = w_j$ .

We now turn to some diagnostic plots. By applying model (5.41) or model (5.45) to our data we can estimate the unconditional distribution functions of both the delay  $Y$  and the reoffence time  $X$ . In each case this can be done by plotting the estimated values of  $\hat{G}_n(v)$  or  $\tilde{G}_n(v)$  against the values of  $v$ . Alternatively, as a diagnostic plot for checking the unconditional distributions of  $X$  and  $Y$ , we can also plot  $\log\{1 - \hat{G}_n(v)\}$  or  $\log\{1 - \tilde{G}_n(v)\}$  against  $v$  and examine the straightness of the outcomes. If the distributions of  $X$  and  $Y$  are exponential, then for each case the corresponding diagnostic plot should resemble a straight line. The diagnostic plot,  $\hat{G}_n$  version, relating to the delay  $Y$  is illustrated in Figure 5.8, which is reasonably linear, substantiating that the unconditional distribution of  $Y$  is exponential. The plot corresponding to  $\tilde{G}_n$  version is the same as that of  $\hat{G}_n$ , as expected since  $\hat{G}_n$  and  $\tilde{G}_n$  are asymptotically equivalent. Similarly, the diagnostic plot corresponding to the reoffence time  $X$  is shown in Figure 5.9, which is close to a straight line, suggesting that the unconditional distribution of  $X$  is also exponential.

Fig. 5.8 Nonparametric Marginal Diagnostics for Delay Time

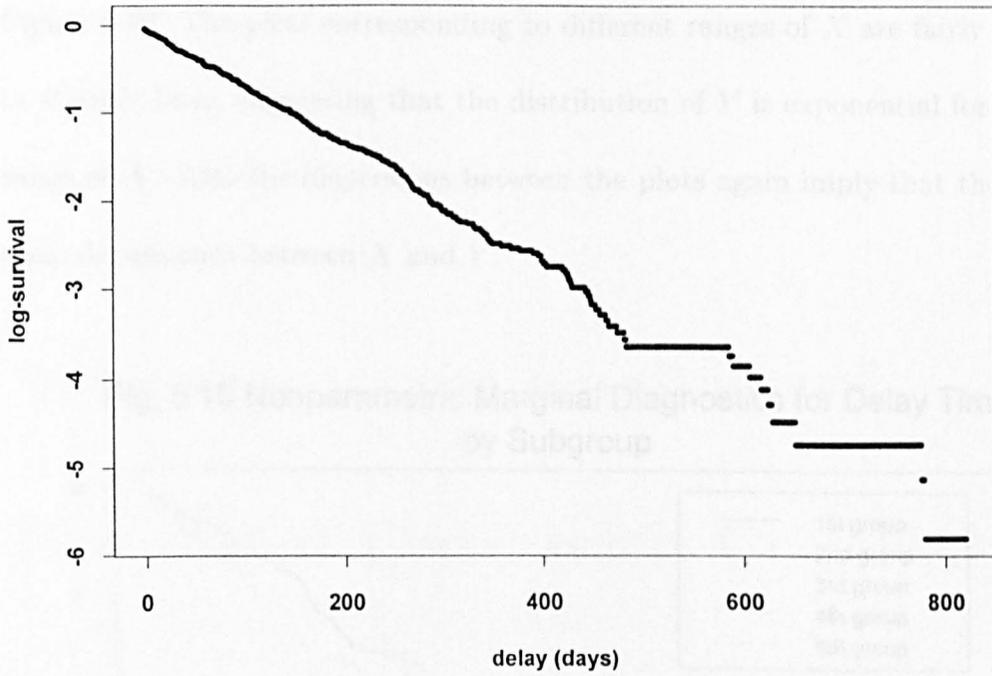
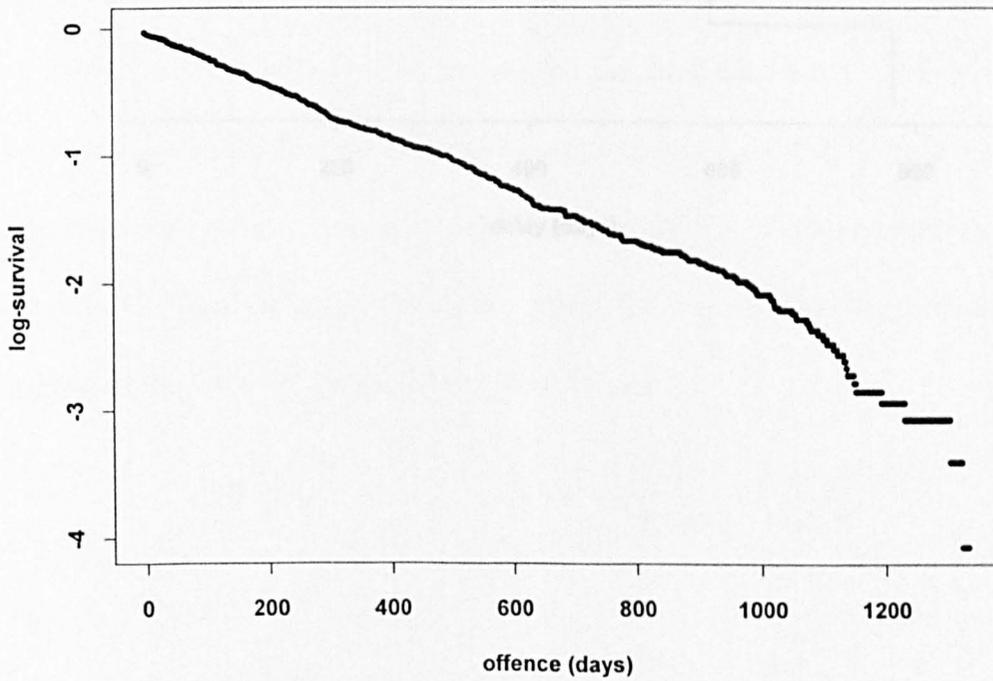
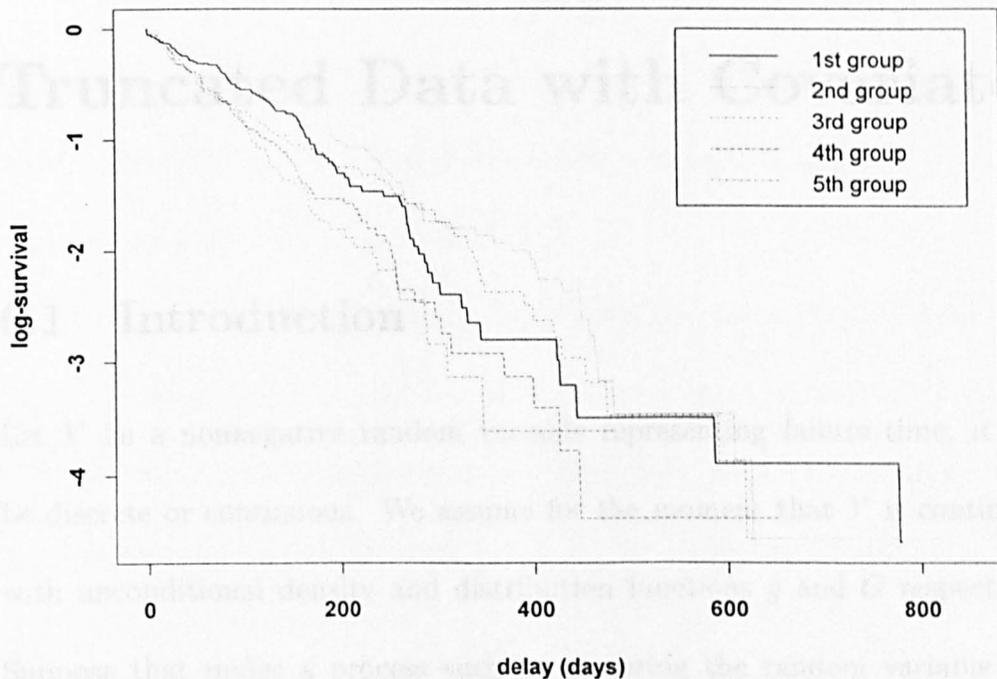


Fig. 5.9 Nonparametric Marginal Diagnostics for offence Time



In order to study the independence between  $X$  and  $Y$  we repeat plot 5.8, with  $\hat{G}_n$  version, for different ranges of  $X$ . This diagnostic plot is depicted in Figure 5.10. The plots corresponding to different ranges of  $X$  are fairly close to straight lines, suggesting that the distribution of  $Y$  is exponential for each range of  $X$ . Also the dispersions between the plots again imply that there is some dependence between  $X$  and  $Y$ .

Fig. 5.10 Nonparametric Marginal Diagnostics for Delay Time by Subgroup



# Chapter 6

## Nonparametric Analysis of Truncated Data with Covariates

### 6.1 Introduction

Let  $Y$  be a nonnegative random variable representing failure time, it may be discrete or continuous. We assume for the moment that  $Y$  is continuous with unconditional density and distribution functions  $g$  and  $G$  respectively. Suppose that under a process such as censoring the random variable  $Y$  is truncated at time  $W$  from the right. Then the conditional distribution and density functions of  $Y$  given  $W$  and  $Y \leq W$  are

$$P(Y \leq y | W = w, Y \leq W) = \frac{G(y)}{\int_0^w g(y) dy} = \frac{G(y)}{G(w)},$$

and

$$P(Y | W = w, Y \leq W) = \frac{dG(y)}{\int_0^w g(y) dy} = \frac{g(y)}{G(w)},$$

respectively. By analogy with ordinary hazard function

$$h_*(y) = \frac{g(y)}{P(Y > y)}, \quad (6.1)$$

we define the *backward hazard function*  $h$  to be

$$h(y) = \frac{g(y)}{P(Y < y)}, \quad (6.2)$$

and if, corresponding to  $y$ , there is available a vector of covariates  $v$ , we define  $h(y; v)$  to be

$$h(y; v) = \frac{g(y; v)}{P(Y < y; v)}. \quad (6.3)$$

We call this ‘backward hazard function’, since in a sense, we are ‘running time backwards’, as  $P(Y > y)$  in the ordinary hazard function (6.1) is being replaced by  $P(Y < y)$  in equation (6.2).

Assume that under the truncation process the data observed from  $Y$  are  $y_1, \dots, y_n$  where each  $y_i$  has its own truncation time  $w_i$  with  $y_i \leq w_i$ ,  $i = 1, \dots, n$ . Note that when we define equation (6.3), we are conditioning on the truncation times and so we treat  $w_1, \dots, w_n$  as if they were fixed.

## 6.2 Backward Regression Model

Suppose now that corresponding to each  $y_i$  there is available a vector of covariates  $v_i$ ,  $v_i^T = (v_{1i}, v_{2i}, \dots, v_{pi})$ ,  $i = 1, \dots, n$ , where  $v_{ki}$ ,  $k = 1, \dots, p$ , is the  $i$ th element of covariate  $k$ . Note that this vector of covariates can consist of the covariates studied earlier, but can also include  $x$  or  $y$ , hence giving a new way of assessing the independence of  $X$  and  $Y$ . If we denote by  $V$  the matrix

of covariates, then  $v_i$ , the vector of covariates specific to the  $i$ th subject, is the  $i$ th column of  $V$  and  $v_{ki}$  is the  $(k, i)$ th entry of  $V$ . On the basis of the truncated data the main problem considered is that of assessing the relation between the unconditional distribution of  $Y$  and the vector of covariates  $v$ . We do this by assuming a proportional hazards model in which the backward hazard is assumed to be

$$h(y; v) = h_0(y)e^{\beta^T v}, \quad (6.4)$$

where  $\beta$  is a  $1 \times p$  vector of unknown parameters and  $h_0(y)$  is an arbitrary and unknown function of  $y$  giving the value of  $h(y; v)$  for an individual under the set of conditions  $v = \underline{0}$ . In other words,  $h_0(y)$  is an unspecified function of time common to all individuals. The vector of regression coefficients  $\beta$  is of particular interest, and the function  $h_0(y)$  can be considered as a nuisance parameter. In the functional form this model is similar to the Cox proportional hazards model (Cox, 1972) in which we are running time forwards.

The set of individuals at risk at time  $y_i$  is called the risk set at time  $y_i$  and is denoted by  $R(y_i)$ ; here this set consists of those individuals whose failure time is at most  $y_i$  and whose truncation time is at least  $y_i$ . Thus

$$R(y_i) = \{j \leq n : y_j \leq y_i, w_j \geq y_i\},$$

which we call the *backward risk set*. Note that in the Cox proportional hazards model, the risk set at time  $y_i$  consists of those individuals whose failure time or censoring time is at least  $y_i$ .

## 6.3 Partial Likelihood

The idea of partial likelihood was introduced by Cox (1975) as a technique for making inferences in the presence of many nuisance parameters. The methods allow reduction in the dimensionality of certain problems.

Following Cox, now consider a particular case with vector of covariates  $v_i$  and with failure at time  $y_i$ . Conditioning on the risk set  $R(y_i)$ , the probability that the failure on this individual is as observed is

$$\frac{h(y_i; v_i)}{\sum_{j \in R(y_i)} h(y_i; v_j)} = \frac{h_0(y_i) e^{\beta^T v_i}}{\sum_{j \in R(y_i)} h_0(y_i) e^{\beta^T v_j}} = \frac{e^{\beta^T v_i}}{\sum_{j \in R(y_i)} e^{\beta^T v_j}},$$

for  $i = 1, \dots, n$ . Each failure contributes a factor of this kind to the likelihood function. Thus the required partial likelihood function is

$$L(\beta) = \prod_{i=1}^n \left\{ \frac{e^{\beta^T v_i}}{\sum_{j \in R(y_i)} e^{\beta^T v_j}} \right\},$$

which is independent of  $h_0(y)$  and the corresponding partial log-likelihood function is given by

$$\ell(\beta) = \sum_{i=1}^n \beta^T v_i - \sum_{i=1}^n \log \left\{ \sum_{j \in R(y_i)} e^{\beta^T v_j} \right\}. \quad (6.5)$$

Direct calculation from equation (6.5) gives for  $k, m = 1, \dots, p$

$$\frac{\partial \ell(\beta)}{\partial \beta_k} = \sum_{i=1}^n \{v_{ki} - A_{ki}(\beta)\},$$

where

$$A_{ki}(\beta) = \frac{\sum_{j \in R(y_i)} v_{kj} e^{\beta^T v_j}}{\sum_{j \in R(y_i)} e^{\beta^T v_j}},$$

and

$$-\frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_m} = \sum_{i=1}^n B_{kmi}(\beta), \quad (6.6)$$

where

$$B_{kmi}(\beta) = \frac{\sum_{j \in R(y_i)} v_{kj} v_{mj} e^{\beta^T v_j}}{\sum_{j \in R(y_i)} e^{\beta^T v_j}} - A_{ki}(\beta) A_{mi}(\beta).$$

Maximum likelihood estimate of  $\beta$ , denoted by  $\hat{\beta}$ , can be obtained by direct numerical maximization of  $\ell(\beta)$  in the usual way. After finding  $\hat{\beta}$  if we evaluate equation (6.6) at  $\hat{\beta}_k$  and  $\hat{\beta}_m$ , then an estimate for  $(k, m)$ th element of the information matrix will be obtained. So the variance-covariance matrix of the estimator  $\hat{\beta}$ , which is the inverse of the information matrix, will be available and hence the standard errors of  $\hat{\beta}$  can be obtained. Of course, this assumes that the partial likelihood function satisfies the usual properties. Cox (1975) indicated that the usual asymptotic properties for maximum likelihood estimators hold for estimators obtained from maximization of partial likelihoods. For the more usual partial likelihood this is discussed in detail in the literature on the Cox model. Relevant references in this area include Andersen, P. K. and Gill, R. D. (1982), Cox, D. R. (1972), Cox, D. R. (1975), Liu, P. Y. and Crowley, J. (1978), Oakes, D. (1981) and Tsiatis, A. A. (1981). Also significance tests about subsets of parameters can be derived by comparison of the maximum log-likelihoods achieved, that is, by using the log-likelihood ratio test. If the distribution of  $Y$  does not depend on the vector of covariates  $v$ , then  $\beta = \underline{0}$ . Thus we can test the null hypothesis  $H_0 : \beta = \underline{0}$  against the alternative hypothesis  $H_1 : \beta \neq \underline{0}$ .

## 6.4 Estimating Distribution of Failure Time

Once we have obtained the maximum likelihood estimate of  $\beta$ , we can consider the estimation of the distribution function of  $Y$  under the backward proportional hazards model (6.4) as follows. We know that

$$dG(y; v) = g(y; v) dy,$$

or by using equation (6.3) we have

$$\frac{dG(y; v)}{G(y; v)} = h(y; v) dy. \quad (6.7)$$

Integrating both sides of equation (6.7) from  $y$  to  $+\infty$ , gives

$$\int_y^\infty \frac{dG(u; v)}{G(u; v)} = \int_y^\infty h(u; v) du. \quad (6.8)$$

Solving equation (6.8) for  $G$  we get

$$G(y; v) = e^{-\int_y^\infty h(u; v) du}, \quad (6.9)$$

or by substituting from model (6.4) into equation (6.9) we obtain

$$G(y; v) = e^{-c\beta^T v \int_y^\infty h_0(u) du}, \quad (6.10)$$

and

$$g(y; v) = h_0(y) e^{\beta^T v} e^{-e^{\beta^T v} \int_y^\infty h_0(u) du}. \quad (6.11)$$

Thus we can estimate the distribution function  $G$  from equation (6.10), provided an estimate of  $h_0$  is available. One way of estimating  $h_0$  is to carry out a separate maximum likelihood estimation procedure. To do this first we consider the case where there are no ties among the observed values of  $Y$ .

Suppose that the observed ordered values of  $Y$  are

$$y_{(1)} < y_{(2)} < \cdots < y_{(n)}.$$

There are several possibilities for estimating  $h_0$ , but we assume that  $h_0$  is constant between the distinct observed values of  $Y$ . Thus we define

$$h_0(y) = \begin{cases} h_1 & \text{if } y_{(0)} < y \leq y_{(1)} \\ h_2 & \text{if } y_{(1)} \leq y \leq y_{(2)} \\ h_i & \text{if } y_{(i-1)} < y \leq y_{(i)}, \quad i = 3, \dots, n \\ 0 & \text{if } y > y_{(n)} \end{cases} \quad (6.12)$$

where we have defined  $y_{(0)} = 0$ . Note that in this model the term  $h_1$  is unidentifiable. So we need to put a constraint, arbitrarily we assume  $h_1 = h_2$ .

Then

$$\int_{y_{(i)}}^{\infty} h_0(u) du = \sum_{j=i+1}^n h_j \{y_{(j)} - y_{(j-1)}\}. \quad (6.13)$$

Let  $v_{(i)}$  be the vector of covariates corresponding to  $y_{(i)}$ , then substituting from equation (6.13) into equation (6.11) gives

$$g(y_{(i)}; v_{(i)}) = h_i e^{\beta^T v_{(i)}} e^{-e^{\beta^T v_{(i)}} \sum_{j=i+1}^n h_j \{y_{(j)} - y_{(j-1)}\}},$$

for  $i = 1, \dots, n-1$ . Therefore, the likelihood function associated with the estimation of  $h_0 = (h_1, \dots, h_n)$  is

$$L(h_0) = \prod_{i=1}^{n-1} g(y_{(i)}; v_{(i)}),$$

and the required log-likelihood function is

$$\ell(h_0) = \sum_{i=1}^{n-1} \left\{ \log h_i + \beta^T v_{(i)} - e^{\beta^T v_{(i)}} \sum_{j=i+1}^n h_j [y_{(j)} - y_{(j-1)}] \right\}. \quad (6.14)$$

Using equation (6.14), the maximum likelihood estimates of  $h_1, \dots, h_n$  are found to be

$$\hat{h}_1 = \hat{h}_2 = \frac{2}{\{y_{(2)} - y_{(1)}\} e^{\beta T v_{(1)}}},$$

and

$$\hat{h}_i = \frac{1}{\{y_{(i)} - y_{(i-1)}\} \sum_{j=1}^{i-1} e^{\beta T v_{(j)}}},$$

for  $i = 3, \dots, n$ . After finding the estimate of  $h_0$  we have

$$\int_y^\infty h_0(u) du = h_i \{y_{(i)} - y\} + \sum_{j=i+1}^n h_j \{y_{(j)} - y_{(j-1)}\}, \quad (6.15)$$

for  $i = 1, \dots, n - 1$  and  $y_{(i-1)} < y \leq y_{(i)}$ . Using equations (6.15) and (6.10), we get

$$G(u; v) = e^{-e^{\beta T v} \{h_i [y_{(i)} - u] + \sum_{j=i+1}^n h_j [y_{(j)} - y_{(j-1)}]\}}, \quad (6.16)$$

for  $y_{(i-1)} < u \leq y_{(i)}$ ,  $i = 1, \dots, n - 1$ , and

$$G(u; v) = \begin{cases} 0 & \text{if } u \leq y_{(0)} = 0 \\ 1 & \text{if } u > y_{(n)}. \end{cases} \quad (6.17)$$

## 6.5 Analysis in Discrete Time

Now consider the case where there are ties among the observed values of  $Y$ .

Denote the distinct observed values of  $Y$  by

$$y_{(1)} < y_{(2)} < \dots < y_{(k)}.$$

Let  $m_{(i)}$  be the number of tied values at  $y_{(i)}$ , the frequency of  $y_{(i)}$ , so that

$\sum_{i=1}^k m_{(i)} = n$ ,  $n$  being the sample size. (In the continuous case  $k = n$ ,  $m_{(i)} = 1$

for all  $i$ ). Define  $h_0$  as before,

$$h_0(y) = \begin{cases} h_1 & \text{if } y_{(0)} < y \leq y_{(1)} \\ h_2 & \text{if } y_{(1)} \leq y \leq y_{(2)} \\ h_i & \text{if } y_{(i-1)} < y \leq y_{(i)}, \quad i = 3, \dots, k \\ 0 & \text{if } y > y_{(k)}, \end{cases}$$

where again  $y_{(0)} = 0$ ,  $h_1 = h_2$ . Also we have

$$\int_{y_{(i)}}^{\infty} h_0(u) du = \sum_{j=i+1}^k h_j \{y_{(j)} - y_{(j-1)}\}. \quad (6.18)$$

Let  $v_{(ij)}$  be the vector of covariates for  $j$ th observation from the  $i$ th tied group,  $i = 1, \dots, k$ ,  $j = 1, \dots, m_{(i)}$ . Now substituting from equation (6.18) into equation (6.11) gives

$$g(y_{(i)}; v_{(ij)}) = h_i e^{\beta^T v_{(ij)}} e^{-e^{\beta^T v_{(ij)}} \sum_{\ell=i+1}^k h_\ell \{y_{(\ell)} - y_{(\ell-1)}\}},$$

for  $i = 1, \dots, k-1$ ,  $j = 1, \dots, m_{(i)}$ . So in this case the full likelihood function for estimation of  $h_0 = (h_1, \dots, h_k)$  is

$$L(h_0) = \prod_{i=1}^{k-1} \prod_{j=1}^{m_{(i)}} g(y_{(i)}; v_{(ij)}),$$

and the required full log-likelihood function is

$$\ell(h_0) = \sum_{i=1}^{k-1} \sum_{j=1}^{m_{(i)}} \{ \log h_i + \beta^T v_{(ij)} - e^{\beta^T v_{(ij)}} \sum_{\ell=i+1}^k h_\ell \{y_{(\ell)} - y_{(\ell-1)}\} \}. \quad (6.19)$$

Using equation (6.19), the maximum likelihood estimates of  $h_1, \dots, h_k$  can then be shown to satisfy

$$\hat{h}_1 = \hat{h}_2 = \frac{m_{(1)} + m_{(2)}}{\{y_{(2)} - y_{(1)}\} \sum_{r=1}^{m_{(1)}} e^{\beta^T v_{(1r)}}}, \quad (6.20)$$

and

$$\hat{h}_i = \frac{m_{(i)}}{\{y_{(i)} - y_{(i-1)}\} \sum_{j=1}^{i-1} \sum_{r=1}^{m_{(j)}} e^{\beta^T v_{(jr)}}, \quad (6.21)$$

for  $i = 3, \dots, k$ . Once we have obtained the maximum likelihood estimates of  $h_0 = (h_1, \dots, h_k)$ , equations (6.15) to (6.17) follow as before.

Note that in equation (6.16),  $u$  is the argument of the distribution function  $G$  and  $v$  is the vector of covariates of an arbitrary case or individual.

## 6.6 Diagnostic Plots

In order to plot the estimated values of  $G(u; v)$  versus the distinct observed values  $y_{(1)} < y_{(2)} < \dots < y_{(k)}$ , let  $u = y_{(i)}$  in equation (6.16) and for clarity of the notation denote  $v$  by  $v_\ell$ , so we have

$$G(y_{(i)}; v_\ell) = e^{-e^{\beta^T v_\ell} \{\sum_{j=i+1}^k h_j \{y_{(j)} - y_{(j-1)}\}\}}, \quad (6.22)$$

for  $i = 1, \dots, k - 1$ ,  $\ell = 1, \dots, n$ ,  $n$  being the sample size; and  $v_\ell$  is now the vector of covariates specific to the  $\ell$ th individual in the sample, or we can say that  $v_\ell$  is the  $\ell$ th column of the covariates matrix  $V$ . First we consider the model (6.4) for the more general case where covariates are included in the model. Using equation (6.22), for each fixed value of  $i$  we get  $n$  values of  $G(y_{(i)}; v_\ell)$  as  $\ell$  changes from 1 to  $n$ . Then

$$G^*(i) = \frac{1}{n} \sum_{\ell=1}^n G(y_{(i)}; v_\ell), \quad i = 1, \dots, k - 1,$$

is an estimate of the overall proportion of the individuals  $\ell = 1$  to  $n$  whom we would expect to have value of  $y$  less than  $y_{(i)}$ . Thus we plot

$$\{G^*(1), \dots, G^*(k-1)\} \text{ against } \{y_{(1)}, \dots, y_{(k-1)}\}, \quad (6.23)$$

to get an overall estimate of the marginal distribution function  $G$  (after averaging over the covariates). Also if we define the function  $S^*(i)$  by

$$S^*(i) = 1 - G^*(i), \quad i = 1, \dots, k-1,$$

then we can plot

$$\{S^*(1), \dots, S^*(k-1)\} \text{ against } \{y_{(1)}, \dots, y_{(k-1)}\}, \quad (6.24)$$

to get an overall estimate for the marginal survival function of  $Y$ . Alternatively, as a diagnostic plot for checking the marginal distribution of  $Y$ , we can plot

$$\{\log S^*(1), \dots, \log S^*(k-1)\} \text{ against } \{y_{(1)}, \dots, y_{(k-1)}\}, \quad (6.25)$$

to obtain an overall log-survival plot.

The plots given by expressions (6.23), (6.24) and (6.25) can also be repeated for different subsets of the individuals in the sample defined by different ranges of the risk score  $R_{sc} = \beta^T v$  under the model (6.4). We did ‘subset plots’ like these before—subset Kaplan-Meier and fitted survival plots, described in section 4.7.

We shall now consider a special case of the previous procedure. Suppose that there is only one covariate  $v$ ,  $v$  now being a scalar. Then for this covariate

we have

$$G(y_{(i)}; v) = e^{-e^{\beta T} v \sum_{j=i+1}^k h_j \{y_{(j)} - y_{(j-1)}\}} \quad (6.26)$$

Now consider the minimum, median and maximum of  $v$  over the sample (over the  $n$  cases). Then for each of these three extreme values of  $v$  we can plot  $G(y_{(i)}; v)$  against  $y_{(i)}$ ,  $i = 1, \dots, k - 1$ . If we superimpose these three plots, then we can check the extent to which the distribution of  $Y$  depends on this covariate. If there is little dispersion among the superimposed plots, then we can say that the distribution of  $Y$  does not seem to depend on this covariate. Of course, in each case we can do a similar analysis for the survival and log-survival curves of  $Y$ .

Finally, we consider another special case of the model (6.4), in which there is no covariate, that is, the model

$$h(y, \underline{Q}) = h_0(y). \quad (6.27)$$

For this model we have

$$G(y_{(i)}) = e^{-\sum_{j=i+1}^k h_j \{y_{(j)} - y_{(j-1)}\}}, \quad (6.28)$$

for  $i = 1, \dots, k - 1$ . Analogous to previous cases, after estimating  $h_1, \dots, h_k$ , we plot separately  $G(y_{(i)})$ ,  $1 - G(y_{(i)})$  and  $\log\{1 - G(y_{(i)})\}$  against  $y_{(i)}$ ,  $i = 1, \dots, k - 1$ , to get the cumulative distribution, survival and log-survival curves of  $Y$  respectively.

## 6.7 Application of the Regression Model

To illustrate some of the results outlined in previous section, we consider our criminological data in which  $X$  is the time from release to first reoffence, and  $Y$  is the time from this reoffence to conviction. We know that  $Y$  and  $X$  are truncated from the right at  $T - X$  and  $T - Y$  respectively,  $T$  being the time to follow-up. As mentioned before, because of the censoring we are interested in fitting truncated distributions to these data. Using the truncated data, first we consider the analysis about  $Y$  under the model (6.4) for the following cases. (1) with  $v = (x, age, ac, pre, jc)$  as the vector of covariates, and  $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$  as the the regression parameters, the maximum likelihood estimates and standard errors of these parameters (with standard errors in brackets) are

$$\hat{\beta}_1 = -0.00023 (0.00015), \hat{\beta}_2 = 0.00798 (0.00760)$$

$$\hat{\beta}_3 = 0.01699 (0.02964), \hat{\beta}_4 = -0.00627 (0.00635)$$

$$\hat{\beta}_5 = 0.04291 (0.03467).$$

Also the maximum value of the log-likelihood achieved is  $\ell(\hat{\beta}) = -2509.018$ . Here, in order to assess the possibility of a relationship between the unconditional distribution of  $Y$  and  $x$ , we consider  $x$  as a covariate. The maximum likelihood estimates of  $h_1, \dots, h_k$  can now be obtained by substituting the estimated value of  $\beta$ , the vector of regression parameters, into equations (6.20) and (6.21). Now using expression (6.25), Figure 6.1 shows the subset log-survival plots of  $Y$  for different ranges of the risk score  $R_{sc} = \hat{\beta}^T v$ . The plots

are fairly close to straight lines, indicating that the exponential distribution for  $Y$  is adequate. Also there is little dispersion among the plots, suggesting that the distribution of  $Y$  depends on the covariates, but apparently not too much.

(2) with  $v = (age, ac, pre, jc)$ , and  $\beta_1, \beta_2, \beta_3, \beta_4$  as regression parameters, we get

$$\hat{\beta}_1 = 0.00745 (0.00763), \hat{\beta}_2 = 0.02136 (0.02948)$$

$$\hat{\beta}_3 = -0.00630 (0.00632), \hat{\beta}_4 = 0.04781 (0.03457)$$

and  $\ell(\hat{\beta}) = -2510.171$ . The estimates of  $h_1, \dots, h_k$  can be obtained as before. Again using (6.25), the subset log-survival plots of  $Y$ , for different ranges of the risk score  $R_{sc} = \hat{\beta}^T v$ , are depicted in Figure 6.2, suggesting similar results as in case (1). Using the maximum log-likelihoods obtained in parts (1) and (2), the estimate of the log-likelihood ratio statistic is  $\Delta D = 2.306$  which is not significant when compared with the  $\chi_1^2$  distribution, so the hypothesis that the distribution of  $Y$  does not depend on covariate  $x$  is not rejected. We would expect this, as the Figures 6.1 and 6.2 are almost the same. Interestingly, the highly significant result in subsection 5.2.1 is no longer obtained in the current nonparametric model.

(3) with  $v = x$  as the only covariate, we obtain  $\hat{\beta} = -0.00024 (0.00015)$ ,  $\ell(\hat{\beta}) = -2510.260$ ,  $\ell(0) = -2511.586$  and  $\Delta D = 2.652$ , which is not significant when compared with the  $\chi_1^2$  distribution. Thus there is no clear evidence to reject the hypothesis that the distribution of  $Y$  does not depend on  $X$ . The estimates

of  $h_1, \dots, h_k$  can be obtained in the usual way. Now using equation (6.26), the subset log-survival curves of  $Y$  for some extreme values of  $X$  are shown in Figure 6.3, suggesting again that the distribution of  $Y$  is exponential for different values of  $X$ , and there is some dependence between  $X$  and  $Y$ .

(4) with no covariate, the model (6.4) reduces to model (6.27). The estimates of  $h_1, \dots, h_k$  are obtained as before. Now using equation (6.28), the diagnostic plot for checking the distribution of  $Y$  is shown in Figure 6.4. This is fairly close to a straight line, substantiating that the exponential distribution for  $Y$  is reasonable. This gives another confirmation of the exponential distribution, hence the results obtained under the nonparametric analysis of truncated data without covariates in subsection 5.2.3

Fig. 6.1 Marginal Diagnostics for Y by Risk Group (Model:[6.4])

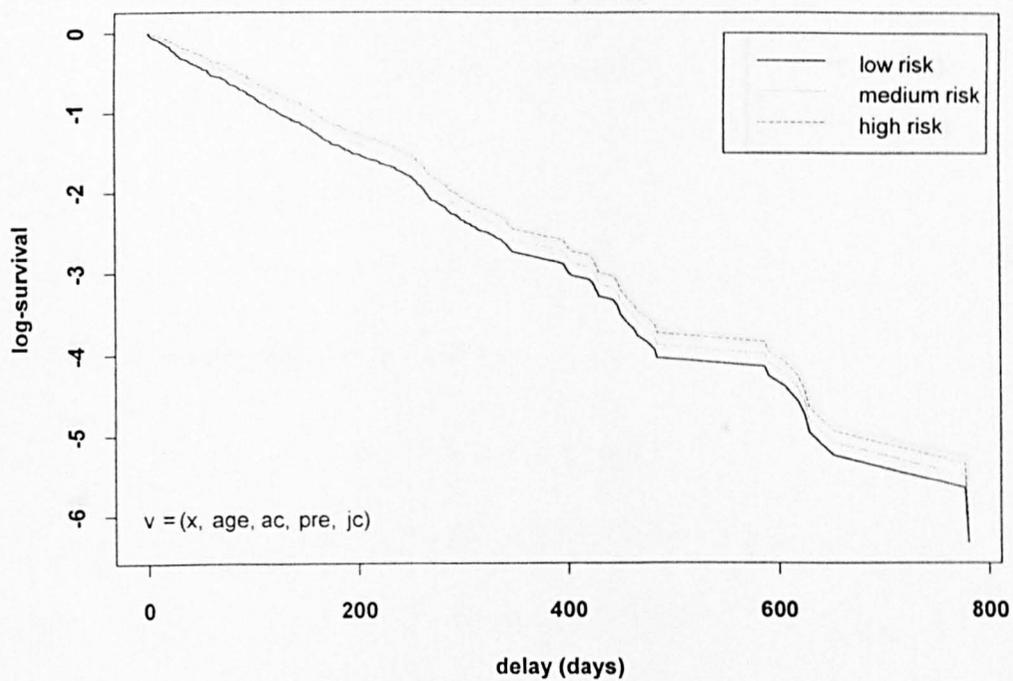


Fig. 6.2 Marginal Diagnostics for Y by Risk Group (Model:[6.4])

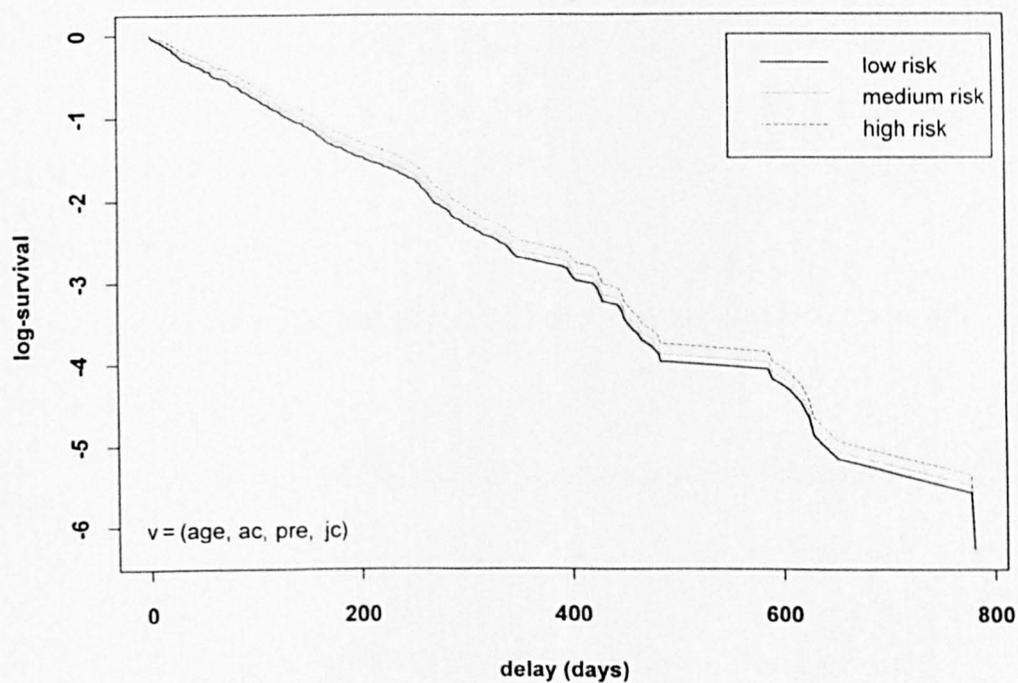


Fig. 6.3 Marginal Diagnostics for Y with Extreme values of X  
(Model:[6.4])

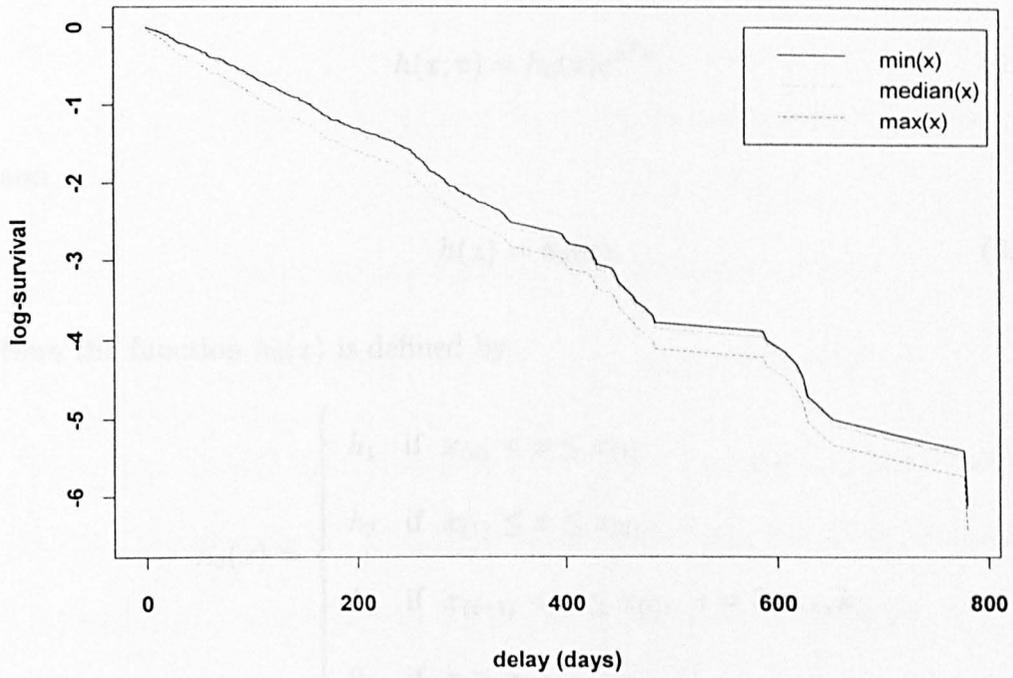
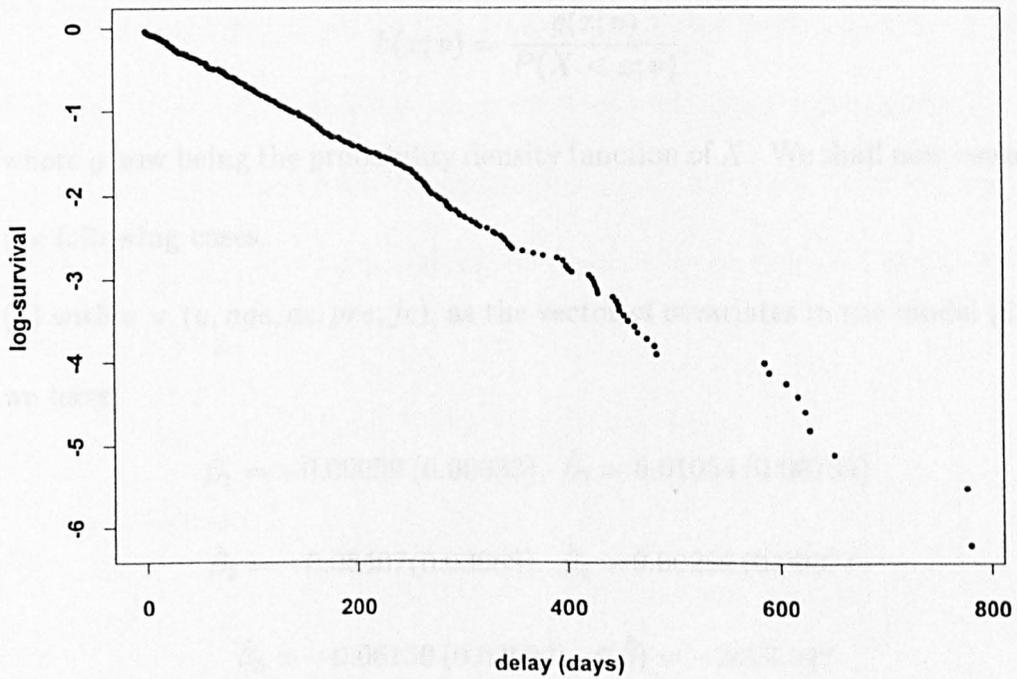


Fig. 6.4 Marginal Diagnostics for Y without Covariates  
(Model:[6.27])



With similar procedures as before, we now briefly turn to the analysis of the reoffence time  $X$  under the models

$$h(x; v) = h_0(x)e^{\beta^T v}, \quad (6.29)$$

and

$$h(x) = h_0(x). \quad (6.30)$$

Here the function  $h_0(x)$  is defined by

$$h_0(x) = \begin{cases} h_1 & \text{if } x_{(0)} < x \leq x_{(1)} \\ h_2 & \text{if } x_{(1)} \leq x \leq x_{(2)} \\ h_i & \text{if } x_{(i-1)} < x \leq x_{(i)}, \quad i = 3, \dots, k \\ 0 & \text{if } x > x_{(k)}, \end{cases}$$

where we have defined  $x_{(0)} = 0$  and  $h_1 = h_2$  as before. Also by analogy with function (6.3) we define

$$h(x; v) = \frac{g(x; v)}{P(X < x; v)},$$

where  $g$  now being the probability density function of  $X$ . We shall now consider the following cases.

(1) with  $v = (y, \text{age}, \text{ac}, \text{pre}, \text{jc})$ , as the vector of covariates in the model (6.29)

we have

$$\hat{\beta}_1 = -0.00059 (0.00033), \quad \hat{\beta}_2 = 0.01054 (0.00735)$$

$$\hat{\beta}_3 = -0.08407(0.03068), \quad \hat{\beta}_4 = 0.00269 (0.00674)$$

$$\hat{\beta}_5 = -0.06159 (0.03592), \quad \ell(\hat{\beta}) = -2493.547.$$

Note that we are interested in  $y$  as a covariate here because we want to study the possibility of a relationship between the unconditional distribution of  $X$  and  $y$ .

(2) with  $v = (age, ac, pre, jc)$ , and  $\beta_1, \beta_2, \beta_3, \beta_4$  as regression parameters, we get

$$\hat{\beta}_1 = 0.00924 (0.00734), \hat{\beta}_2 = -0.08436 (0.03086)$$

$$\hat{\beta}_3 = 0.00308 (0.00673), \hat{\beta}_4 = -0.06474 (0.03599)$$

and  $\ell(\hat{\beta}) = -2495.239$ . Using the two maximum log-likelihoods achieved,  $\Delta D = 3.384$  which is not significant as compared with the  $\chi_1^2$  distribution.

(3) with  $v = y$  as the only covariate, we obtain  $\hat{\beta} = -0.00057 (0.00035)$ ,  $\ell(\hat{\beta}) = -2503.378$  and  $\ell(0) = -2504.977$ . Here,  $\Delta D = 3.198$  which is again not significant.

(4) with no covariate, we consider the diagnostic plots for checking the distribution of  $X$  under the model (6.30).

The analogues of plots 6.1–6.4 for  $X$  are shown in Figures 6.5–6.8 respectively. Although these plots are close to straight lines and suggesting an exponential distribution for the reoffence time  $X$ , but there are some deviations from linearity at the right tails of the plots. This corresponds to the fact that we are assuming

$$h_0(x) = 0, \text{ for } x > x_{(k)},$$

or equivalently

$$g(x; v) = 0, \text{ for } x > x_{(k)},$$

a problem similar to that of  $q_k = 1$ , discussed in subsection 5.2.2. This is the

problem of our backward regression model (6.29) for large values of  $X$ . Clearly this assumption is not reasonable for the reoffence time  $X$ , as the censoring is very heavy in this case and the large reoffence times are being censored. Note that this assumption is sensible for the delay time  $Y$ , because in this case the censoring is not heavy and only very unusually large delay times are likely to be censored.

Fig. 6.5 Marginal Diagnostics for X by Risk Group (Model:[6.29])

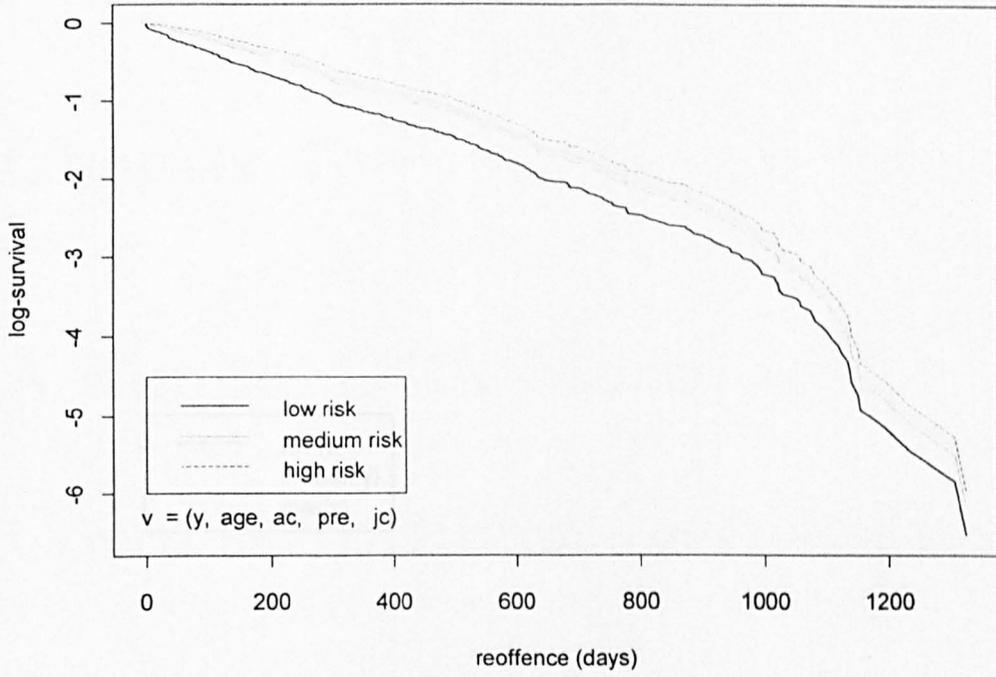


Fig. 6.6 Marginal Diagnostics for X by Risk Group (Model:[6.29])

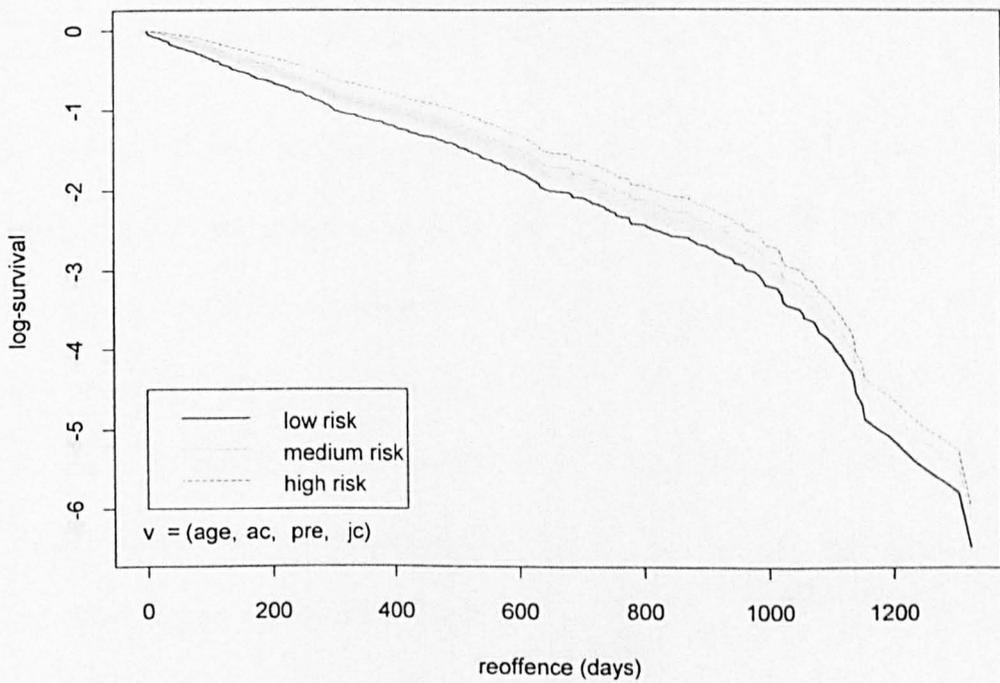


Fig. 6.7 Marginal Diagnostics for X with Extreme values of Y  
(Model:[6.29])

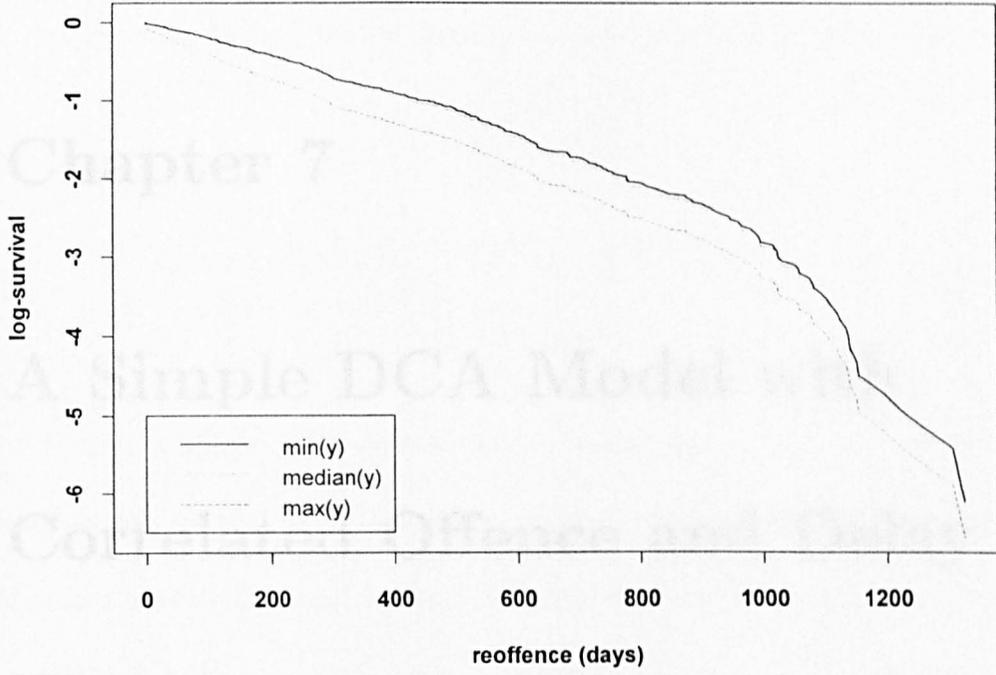
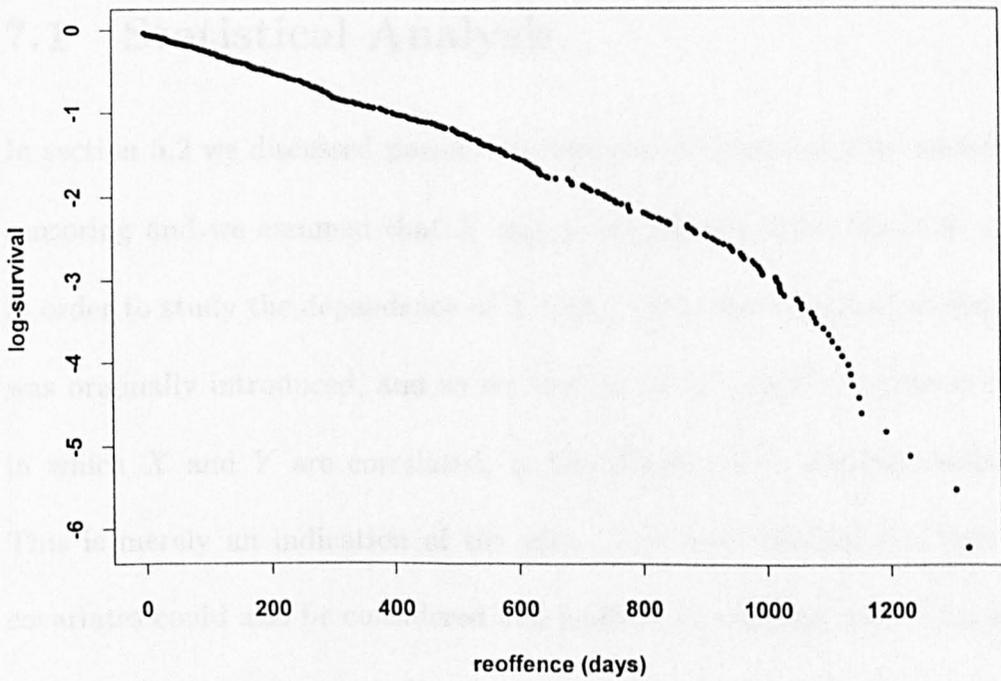


Fig. 6.8 Marginal Diagnostics for X without Covariates  
(Model:[6.30])



# Chapter 7

## A Simple DCA Model with Correlated Offence and Delay Times

### 7.1 Statistical Analysis

In section 5.2 we discussed parametric analysis of truncated data induced by censoring and we assumed that  $X$  and  $Y$  are independent. However, it was in order to study the dependence of  $X$  and  $Y$ , that the truncated model (5.7) was originally introduced, and so we now go on to consider a general model in which  $X$  and  $Y$  are correlated, in the simple case excluding covariates. This is merely an indication of the idea. The more general case including covariates could also be considered and handled in a similar way. This model extends to all the data including both uncensored and censored observations.

We now consider the conditional distribution of  $Y$  given  $X = x$ , and use the notations  $S_{Y|x}$  and  $f_{Y|x}$  for the conditional survival and density functions of  $Y$  respectively. Let  $f_X$  be the density function of  $X$  and suppose that

$$\begin{aligned} f_X(x) &= \lambda e^{-\lambda x}, \\ f_{Y|x}(y) &= \theta^* e^{-\theta^* y}, \quad \theta^* = \theta e^{bx}, \\ S_{Y|x}(y) &= e^{-\theta^* y}. \end{aligned} \tag{7.1}$$

Note that in model (7.1) dependency is represented by the parameter  $b$ . In next section we will test to see if this parameter is significantly different from zero and hence there would be an indication of dependence between  $X$  and  $Y$ .

Following the same argument as in section 3.1 and using the split population model given by  $p$ ,  $\lambda$  and  $\theta^*$ , the likelihood function for estimating the parameters is given by

$$L = \prod_{i=1}^n \{p \lambda e^{-\lambda x_i} \theta_i^* e^{-\theta_i^* y_i}\} \prod_{i=n+1}^N \{1 - p + p P(X + Y > t_i)\},$$

where  $t_i$  is the time to follow-up (or censoring time) for the  $i$ th case,  $n$  is the number of uncensored observations,  $N$  is the total sample size and  $N - n$  is the number of censored observations. Thus the log-likelihood function is given by

$$\ell = \ell_1 + \ell_2 \tag{7.2}$$

where

$$\ell_1 = \sum_{i=1}^n \{\log p + \log \lambda - \lambda x_i + \log \theta + b x_i - y_i \theta e^{bx_i}\}$$

$$\ell_2 = \sum_{i=n+1}^N \log\{1 - p + pP(X + Y > t_i)\}.$$

To evaluate  $P(X + Y > t_i)$  we proceed as follows.

$$P(X + Y > t_i) = E_X\{P(X + Y > t_i|X)\} = E_X\{P(Y > t_i - X|X)\}.$$

Now

$$P(Y > t_i - X|X) = \begin{cases} e^{-\theta(t_i - X)} & \text{if } X < t_i \\ 1 & \text{if } X > t_i, \end{cases}$$

so

$$P(X + Y > t_i) = 1 \times P(X > t_i) + \int_0^{t_i} e^{-\theta(t_i - x)} \lambda e^{-\lambda x} dx$$

or

$$P(X + Y > t_i) = e^{-\lambda t_i} + \lambda \int_0^{t_i} e^{-\theta e^{bx}(t_i - x)} e^{-\lambda x} dx.$$

For small  $b$ ,

$$e^{bx} \approx 1 + bx,$$

thus

$$P(X + Y > t_i) \approx e^{-\lambda t_i} + \lambda \int_0^{t_i} e^{-\theta(1+bx)(t_i - x)} e^{-\lambda x} dx$$

which can be written as

$$P(X + Y > t_i) \approx e^{-\lambda t_i} + \lambda e^{-\theta t_i} \int_0^{t_i} e^{d_i x} e^{b\theta x^2} dx,$$

where

$$d_i = \theta - b\theta t_i - \lambda. \tag{7.3}$$

Again for small  $b$  we have

$$e^{b\theta x^2} \approx 1 + b\theta x^2,$$

thus

$$P(X + Y > t_i) \approx e^{-\lambda t_i} + \lambda e^{-\theta t_i} \left\{ \int_0^{t_i} e^{d_i x} dx + b \theta \int_0^{t_i} x^2 e^{d_i x} dx \right\}.$$

Hence we get, on simplifying the relevant expression,

$$P(X + Y > t_i) \approx e^{-\lambda t_i} + \lambda e^{-\theta t_i} g_i \quad (7.4)$$

where

$$g_i = \left\{ \frac{e^{d_i t_i} - 1}{d_i} + b \theta \frac{e^{d_i t_i} [(d_i t_i - 1)^2 + 1] - 2}{d_i^3} \right\}.$$

Note that when  $b=0$ , we have

$$P(X + Y > t_i) = \frac{\theta e^{-\lambda t_i} - \lambda e^{-\theta t_i}}{\theta - \lambda}$$

which agrees with equation (3.2), as expected. Substituting from equation (7.4) into equation (7.2), the maximum likelihood estimates of  $p$ ,  $\lambda$ ,  $\theta$  and  $b$  can then be obtained by direct numerical maximization of  $\ell$ . For the criminological data this gives

$$\hat{p} = 0.442, \hat{\lambda} = 0.00216, \hat{\theta} = 0.00527, \hat{b} = 0.000815.$$

Note that the estimated value of  $b$  is very small which is consistent with the assumption about  $b$  in the analysis. However, to see if  $b$  is significantly different from zero, we need to know the standard error of  $\hat{b}$ .

## 7.2 Standard Errors

To determine the significance of the parameters obtained in section 7.1, we need to estimate the standard errors of  $\hat{p}$ ,  $\hat{\lambda}$ ,  $\hat{\theta}$  and  $\hat{b}$  from the corresponding

variance-covariance matrix. To do this, we proceed as follows.

The function  $g_i$  in equation (7.4) can be expressed as

$$g_i = g_1(d_i) + b\theta g_2(d_i) \quad (7.5)$$

where  $d_i$ , defined by equation (7.3), satisfies

$$\frac{\partial d_i}{\partial \lambda} = -1, \quad \frac{\partial d_i}{\partial \theta} = 1 - b t_i, \quad \frac{\partial d_i}{\partial b} = -\theta t_i \quad (7.6)$$

and

$$g_1(d_i) = \frac{e^{d_i t_i} - 1}{d_i} \quad (7.7)$$

$$g_2(d_i) = \frac{e^{d_i t_i} \{(d_i t_i - 1)^2 + 1\} - 2}{d_i^3}. \quad (7.8)$$

Let

$$g_{11} = \frac{\partial g_1}{\partial d_i}, \quad g_{21} = \frac{\partial g_2}{\partial d_i}, \quad g_{12} = \frac{\partial^2 g_1}{\partial d_i^2}, \quad g_{22} = \frac{\partial^2 g_2}{\partial d_i^2} \quad (7.9)$$

then differentiating equations (7.7) and (7.8), we have

$$g_{11} = \frac{(d_i t_i - 1) e^{d_i t_i} + 1}{d_i^2} \quad (7.10)$$

$$g_{12} = \frac{\{(d_i t_i - 1)^2 + 1\} e^{d_i t_i} - 2}{d_i^3} = g_2 \quad (7.11)$$

$$g_{21} = \frac{\{(d_i t_i)^3 - 3[(d_i t_i - 1)^2 + 1]\} e^{d_i t_i} + 6}{d_i^4} \quad (7.12)$$

$$g_{22} = \frac{\{(d_i t_i)^2 [(d_i t_i - 2)^2 + 5] - 24(d_i t_i - 1)\} e^{d_i t_i} - 24}{d_i^5}. \quad (7.13)$$

Let

$$k_1 = g_{11} + b\theta g_{21} \quad (7.14)$$

then from equation (7.5), by using chain rule for derivatives of functions, and on substituting from equations (7.6) and (7.14), we get

$$\frac{\partial g_i}{\partial \lambda} = -k_1 \quad (7.15)$$

$$\frac{\partial g_i}{\partial \theta} = b g_2 + k_1 (1 - b t_i) \quad (7.16)$$

$$\frac{\partial g_i}{\partial b} = \theta (g_2 - k_1 t_i). \quad (7.17)$$

For brevity of calculations we also use the following conventional notations.

Let

$$P(t_i) = P(X + Y > t_i), \quad Q(t_i) = 1 - p + p P(t_i) \quad (7.18)$$

$$P_1(t_i) = \frac{\partial P(t_i)}{\partial \lambda}, \quad P_2(t_i) = \frac{\partial P(t_i)}{\partial \theta}, \quad P_3(t_i) = \frac{\partial P(t_i)}{\partial b} \quad (7.19)$$

$$P_{11}(t_i) = \frac{\partial^2 P(t_i)}{\partial \lambda^2}, \quad P_{22}(t_i) = \frac{\partial^2 P(t_i)}{\partial \theta^2}, \quad P_{33}(t_i) = \frac{\partial^2 P(t_i)}{\partial b^2} \quad (7.20)$$

$$P_{12}(t_i) = \frac{\partial^2 P(t_i)}{\partial \lambda \partial \theta}, \quad P_{13}(t_i) = \frac{\partial^2 P(t_i)}{\partial \lambda \partial b}, \quad P_{23}(t_i) = \frac{\partial^2 P(t_i)}{\partial \theta \partial b}. \quad (7.21)$$

Differentiating equation (7.4) with respect to  $\lambda$ ,  $\theta$  and  $b$ , we have

$$P_1(t_i) = -t_i e^{-\lambda t_i} + e^{-\theta t_i} \left\{ g_i + \lambda \left( \frac{\partial g_i}{\partial \lambda} \right) \right\} \quad (7.22)$$

$$P_2(t_i) = \lambda e^{-\theta t_i} \left\{ -t_i g_i + \left( \frac{\partial g_i}{\partial \theta} \right) \right\} \quad (7.23)$$

$$P_3(t_i) = \lambda e^{-\theta t_i} \left( \frac{\partial g_i}{\partial b} \right). \quad (7.24)$$

Substituting from equations (7.15)–(7.17) into equations (7.22)–(7.24) respectively, we get

$$P_1(t_i) = -t_i e^{-\lambda t_i} + e^{-\theta t_i} (g_i - k_1 \lambda) \quad (7.25)$$

$$P_2(t_i) = \lambda e^{-\theta t_i} \{ -t_i g_i + b g_2 + k_1 (1 - b t_i) \} \quad (7.26)$$

$$P_3(t_i) = \lambda \theta e^{-\theta t_i} (g_2 - k_1 t_i). \quad (7.27)$$

Now, using equation (7.2), the first and second order partial derivatives of the log-likelihood function  $\ell$  with respect to  $p$ ,  $\lambda$ ,  $\theta$  and  $b$  can be expressed in terms

of the quantities in equations (7.18)–(7.21) as follows.

$$\begin{aligned} \frac{\partial \ell}{\partial p} &= \frac{n}{p} + \sum_{i=n+1}^N \left\{ \frac{P(t_i) - 1}{Q(t_i)} \right\} \\ \frac{\partial^2 \ell}{\partial p^2} &= - \left\{ \frac{n}{p^2} + \sum_{i=n+1}^N \left( \frac{P(t_i) - 1}{Q(t_i)} \right)^2 \right\} \\ \frac{\partial^2 \ell}{\partial p \partial \lambda} &= \sum_{i=n+1}^N \left\{ \frac{P_1(t_i)}{Q^2(t_i)} \right\} \\ \\ \frac{\partial^2 \ell}{\partial p \partial \theta} &= \sum_{i=n+1}^N \left\{ \frac{P_2(t_i)}{Q^2(t_i)} \right\} \\ \frac{\partial^2 \ell}{\partial p \partial b} &= \sum_{i=n+1}^N \left\{ \frac{P_3(t_i)}{Q^2(t_i)} \right\} \\ \\ \frac{\partial \ell}{\partial \lambda} &= \frac{n}{\lambda} - \sum_{i=1}^n x_i + p \sum_{i=n+1}^N \left\{ \frac{P_1(t_i)}{Q(t_i)} \right\} \\ \frac{\partial^2 \ell}{\partial \lambda^2} &= - \frac{n}{\lambda^2} + p \sum_{i=n+1}^N \left\{ \frac{P_{11}(t_i) Q(t_i) - p P_1^2(t_i)}{Q^2(t_i)} \right\} \\ \\ \frac{\partial^2 \ell}{\partial \lambda \partial \theta} &= p \sum_{i=n+1}^N \left\{ \frac{P_{12}(t_i) Q(t_i) - p P_1(t_i) P_2(t_i)}{Q^2(t_i)} \right\} \\ \\ \frac{\partial^2 \ell}{\partial \lambda \partial b} &= p \sum_{i=n+1}^N \left\{ \frac{P_{13}(t_i) Q(t_i) - p P_1(t_i) P_3(t_i)}{Q^2(t_i)} \right\} \\ \\ \frac{\partial \ell}{\partial \theta} &= \frac{n}{\theta} - \sum_{i=1}^n y_i e^{b x_i} + p \sum_{i=n+1}^N \left\{ \frac{P_2(t_i)}{Q(t_i)} \right\} \\ \\ \frac{\partial^2 \ell}{\partial \theta^2} &= - \frac{n}{\theta^2} + p \sum_{i=n+1}^N \left\{ \frac{P_{22}(t_i) Q(t_i) - p P_2^2(t_i)}{Q^2(t_i)} \right\} \\ \\ \frac{\partial^2 \ell}{\partial \theta \partial b} &= p \sum_{i=n+1}^N \left\{ \frac{P_{23}(t_i) Q(t_i) - p P_2(t_i) P_3(t_i)}{Q^2(t_i)} \right\} \end{aligned}$$

$$\frac{\partial \ell}{\partial b} = \sum_{i=1}^n \{x_i (1 - \theta y_i e^{b x_i})\} + p \sum_{i=n+1}^N \left\{ \frac{P_3(t_i)}{Q(t_i)} \right\}$$

$$\frac{\partial^2 \ell}{\partial b^2} = -\theta \sum_{i=1}^n (x_i^2 y_i e^{b x_i}) + p \sum_{i=n+1}^N \left\{ \frac{P_{33}(t_i) Q(t_i) - p P_3^2(t_i)}{Q^2(t_i)} \right\}.$$

To get the estimates of second order partial derivatives of the log-likelihood function  $\ell$ , first we need to estimate the quantities listed in equations (7.18)–(7.21). Having the observed sample values and the maximum likelihood estimates of the parameters  $p$ ,  $\lambda$ ,  $\theta$  and  $b$ , we can estimate  $P(t_i)$  for each  $t_i$ ,  $i = n + 1, \dots, N$ , from equations (7.3) and (7.4). Using equation (7.18), the estimate of  $Q(t_i)$  is then available. Also, using equations (7.25)–(7.27), the estimates of  $P_1(t_i)$ ,  $P_2(t_i)$ ,  $P_3(t_i)$  can be obtained. To get the estimates of the second order quantities in equations (7.20) and (7.21), we still need some further calculations as follows. Let

$$k_2 = g_2 + b \theta g_{22}$$

then after some calculations, the equations for obtaining the estimates of  $P_{11}(t_i)$ ,  $P_{12}(t_i)$ ,  $P_{13}(t_i)$ ,  $P_{22}(t_i)$ ,  $P_{23}(t_i)$  and  $P_{33}(t_i)$  are given by

$$P_{11}(t_i) = t_i^2 e^{-\lambda t_i} + e^{-\theta t_i} (\lambda k_2 - 2k_1)$$

$$P_{12}(t_i) = e^{-\theta t_i} \{f_1(t_i) - f_2(t_i)\}$$

where

$$f_1(t_i) = b g_2 - t_i g_i + k_1 \{1 + (\lambda - b) t_i\}$$

$$f_2(t_i) = \lambda \{b g_{21} + k_2 (1 - b t_i)\}$$

$$P_{13}(t_i) = \theta e^{-\theta t_i} \{g_2 + t_i (\lambda k_2 - k_1) - \lambda g_{21}\}$$

$$P_{22}(t_i) = \lambda e^{-\theta t_i} \{f_3(t_i) + f_4(t_i)\}$$

where

$$f_3(t_i) = t_i (t_i g_i - 2 b g_2)$$

$$f_4(t_i) = (1 - b t_i) \{k_2(1 - b t_i) + 2 (b g_{21} - k_1 t_i)\}$$

$$P_{23}(t_i) = \lambda e^{-\theta t_i} \{f_5(t_i) + f_6(t_i)\}$$

where

$$f_5(t_i) = (1 - \theta t_i) (g_2 - k_1 t_i)$$

$$f_6(t_i) = \theta \{(1 - 2 b t_i) g_{21} - k_2 t_i (1 - b t_i)\}$$

and

$$P_{33}(t_i) = \lambda \theta^2 t_i e^{-\theta t_i} (k_2 t_i - 2 g_{21}).$$

Hence, on substituting the sample values and replacing the parameters,  $p$ ,  $\lambda$ ,  $\theta$  and  $b$ , by their maximum likelihood estimates,  $\hat{p}$ ,  $\hat{\lambda}$ ,  $\hat{\theta}$  and  $\hat{b}$ , the estimate of information matrix is given by symmetric matrix

$$I = \begin{pmatrix} -\frac{\partial^2 \ell}{\partial p^2} & -\frac{\partial^2 \ell}{\partial p \partial \lambda} & -\frac{\partial^2 \ell}{\partial p \partial \theta} & -\frac{\partial^2 \ell}{\partial p \partial b} \\ -\frac{\partial^2 \ell}{\partial \lambda \partial p} & -\frac{\partial^2 \ell}{\partial \lambda^2} & -\frac{\partial^2 \ell}{\partial \lambda \partial \theta} & -\frac{\partial^2 \ell}{\partial \lambda \partial b} \\ -\frac{\partial^2 \ell}{\partial \theta \partial p} & -\frac{\partial^2 \ell}{\partial \theta \partial \lambda} & -\frac{\partial^2 \ell}{\partial \theta^2} & -\frac{\partial^2 \ell}{\partial \theta \partial b} \\ -\frac{\partial^2 \ell}{\partial b \partial p} & -\frac{\partial^2 \ell}{\partial b \partial \lambda} & -\frac{\partial^2 \ell}{\partial b \partial \theta} & -\frac{\partial^2 \ell}{\partial b^2} \end{pmatrix}.$$

Thus, by inverting this matrix, and taking the square roots of the entries in the main diagonal, the standard errors of  $\hat{p}$ ,  $\hat{\lambda}$ ,  $\hat{\theta}$  and  $\hat{b}$  are found to be

$$\text{se}(\hat{p}) = 0.016, \quad \text{se}(\hat{\lambda}) = 0.00014,$$

$$\text{se}(\hat{\theta}) = 0.00025, \quad \text{se}(\hat{b}) = 0.000111.$$

These are to be compared with the corresponding parameter estimates obtained in section 7.1. It follows that  $p$ ,  $\lambda$  and  $\theta$  are quite accurately estimated from this model, and that  $\frac{\hat{b}}{\text{se}(\hat{b})} = 7.34$ . Therefore, under model (7.1),  $b$  is significantly different from zero, suggesting that there is clear dependence between  $X$  and  $Y$ . Also we are interested in the hypothesis  $p=1$ , so we can calculate and comment on  $\frac{1-\hat{p}}{\text{se}(\hat{p})}$  as evidence that the split population model is needed. The estimated value of this quantity is equal to 34.87, which is extremely large, indicating that the split population model is definitely necessary.

# Chapter 8

## Delayed Censoring Modification to the Cox Model

### 8.1 Introduction

As we know truncated data are those that can be observed only in certain ranges, that is, in models dealing with truncated data the observations are limited to their ranges because of some stochastic mechanism such as censoring. In Chapters 5 and 6 we discussed some parametric and nonparametric analyses of truncated data induced by censoring. In the present Chapter we develop a semi-parametric model for all the data, including both uncensored and censored observations. This will be done by applying the delayed censoring analysis (DCA) model to the Cox proportional hazards regression model (Cox, 1972). This modified Cox model under delayed censoring is then compared with the parametric restricted model discussed earlier in Chapter 3, and

for criminological data very similar results under the two models are obtained. In the general form under this model, the hazard function of observed failure time is expressed in terms of the hazard function of actual failure time multiplied by a weight function, which can be estimated from one of the parametric models developed in the earlier chapters of this thesis. We refer to this model as the ‘weighted hazards model’. This model is then extended to a more general case in which the delay and reoffence times are allowed to be correlated. We refer to this as the ‘generalized weighted hazards model’.

## 8.2 Statistical Theory for Weighted Hazards Model

Let  $X$  and  $Y$  be offence and delay times respectively and assume that they are independent. Denote hazard, density and survival functions of  $X$  by  $h_X$ ,  $f_X$  and  $S_X$  respectively, and let  $S_Y$  be the survival function of  $Y$ . Define the events  $A$  and  $B$  as follows:

$$A = \{\text{observed offence in } (x, x + dx)|t, v\} \quad (8.1)$$

$$B = \{\text{no observed offence in } [0, x]|t, v\} \quad (8.2)$$

where  $t$  is the time of censoring (the time to follow-up) and  $v$  is the vector of covariates specific to a particular observation in the study sample. Recall that an observation is uncensored if and only if  $X + Y < t$ . Note that here,

$A \cap B = A$  and let

$$h^*(x, t, v) = \frac{1}{dx} P(A|B) = \frac{1}{dx} \frac{P(A)}{P(B)} \quad (8.3)$$

be the hazard function for observed offence time  $x$ . Then we will show that

$$h^*(x, t, v) = h_X(x, v) w(x, t, v) \quad (8.4)$$

where

$$h_X(x, v) = \frac{f_X(x, v)}{S_X(x, v)} \quad (8.5)$$

is the hazard function for the actual offence time, and

$$w(x, t, v) = \frac{1 - S_Y(t - x, v)}{1 + \frac{\int_0^x f_X(u, v) S_Y(t - u, v) du}{S_X(x, v)}}. \quad (8.6)$$

We refer to equation (8.6) as the weight function,  $0 \leq w(x, t, v) \leq 1$ , and so we refer to equation (8.4) as ‘weighted hazards model’. Alternatively, we can write

$$w(x, t, v) = \frac{S_X(x, v) \{1 - S_Y(t - x, v)\}}{S_X(x, v) + \int_0^x f_X(u, v) S_Y(t - u, v) du}. \quad (8.7)$$

To prove formula (8.4), we proceed as follows.

(i)  $t > x$ . For this case we have

$$P(A) = P(x < X < x + dx, X + Y < t|t, v)$$

or

$$P(A) = \int_x^{x+dx} f_X(u, v) du \int_0^{t-x} f_Y(y, v) dy.$$

This gives

$$P(A) = [F_X(x + dx, v) - F_X(x, v)] [1 - S_Y(t - x, v)]$$

and so

$$\frac{1}{dx}P(A) = f_X(x, v) [1 - S_Y(t - x, v)]. \quad (8.8)$$

Also we have

$$P(B) = 1 - P(\{\text{observed offence in } [0, x]|t, v\})$$

or

$$P(B) = 1 - P(X \leq x, X + Y < t|t, v) \quad (8.9)$$

which can be written as

$$P(B) = 1 - \int_0^x f_X(u, v) du \int_0^{t-u} f_Y(y, v) dy$$

or

$$P(B) = 1 - \int_0^x f_X(u, v) du + \int_0^x f_X(u, v) S_Y(t - u, v) du$$

and so

$$P(B) = S_X(x, v) + \int_0^x f_X(u, v) S_Y(t - u, v) du. \quad (8.10)$$

Substituting from equations (8.8) and (8.10) into formula (8.3), yields

$$h^*(x, t, v) = \frac{f_X(x, v) [1 - S_Y(t - x, v)]}{S_X(x, v) + \int_0^x f_X(u, v) S_Y(t - u, v) du}. \quad (8.11)$$

If now from equation (8.5) we substitute  $f_X(x, v)$  into equation (8.11), then

we get

$$h^*(x, t, v) = h_X(x, v) \frac{S_X(x, v) [1 - S_Y(t - x, v)]}{S_X(x, v) + \int_0^x f_X(u, v) S_Y(t - u, v) du}.$$

Hence, we have

$$h^*(x, t, v) = h_X(x, v) w(x, t, v).$$

(ii)  $t < x$ . For this case equation (8.10) can be written as

$$P(B) = S_X(x, v) + \int_0^t f_X(u, v) S_Y(t - u, v) du + J$$

where

$$J = \int_t^x f_X(u, v) S_Y(t - u, v) du.$$

Now for  $t < u < x$ , we have  $t - u < 0$  and so

$$S_Y(t - u, v) = P(Y > t - u | t, v) = 1.$$

Thus, in this case we get

$$P(B) = S_X(x, v) + \int_0^t f_X(u, v) S_Y(t - u, v) du + \int_t^x f_X(u, v) du.$$

Therefore, equations (8.4) and (8.6) are defined for all  $x$ .

**Remark:**

(i) If  $t < x$ , then  $S_Y(t - x, v) = 1$  and  $w(x, t, v) = 0$ , implying that  $h^*(x, t, v) = 0$ .

(ii) If  $t$  is very much bigger than  $x$  or  $t - x$  is a very large positive number, in other words the censoring time  $t$  tends to infinity, then  $S_Y(t - x, v)$  becomes very small and  $w(x, t, v) \approx 1$ , and so  $h^*(x, t, v) \approx h_X(x, v)$ .

(iii) For  $t$  close to  $x$  ( $t > x$ ), we have  $0 < S_Y(t - x, v) < 1$ , and hence  $0 < w(x, t, v) < 1$ .

Suppose now that there is available a sample of size  $N$ , including both uncensored and censored observations. On the basis of the whole data and under the weighted hazards model (8.4), the main problem considered is that

of assessing the relation between the unconditional distribution of  $X$  and the vector of covariates  $v$ , for both the actual and observed reoffence times. For the observed reoffence time we will then compare the observed survival curve with the Kaplan Meier survival curve calculated directly from the data. To do this, first we consider the Cox proportional hazards regression model

$$h_X(x, v) = h_0(x) e^{\beta^T v} \quad (8.12)$$

where  $\beta$  is a  $1 \times p$  vector of unknown parameters and  $h_0(x)$  is an unspecified function of time  $x$  common to all observations. Then the weighted hazards model (8.4) becomes

$$h^*(x, t, v) = h_0(x) e^{\beta^T v} w(x, t, v). \quad (8.13)$$

Note that in the special case in which the weight function  $w(x, t, v)$  is equal to 1, the weighted hazards model (8.13) reduces to the ordinary Cox model (8.12). To derive the partial likelihood function (Cox, 1975) for this model, first we need to estimate the weight function  $w(x, t, v)$ . Note that for the sample of  $N$  observations  $(x_i, t_i, v_i)$ ,  $i = 1, 2, \dots, N$ , the estimated values of the weights can be arranged as a matrix  $W$  with  $(i, j)$ th entry given by

$$W_{ij} = w(x_i, t_j, v_j), \text{ for } i, j = 1, \dots, N.$$

### 8.3 Estimation of the Weights

First, we consider estimating the weights from the parametric model considered in section 3.6 (the marginal model). So let  $X$  and  $Y$  have exponential

distribution with mean  $\frac{1}{\lambda(v)}$  and  $\frac{1}{\theta(v)}$  respectively. Then under the split population model, defined in section 2.2, given by  $p(v)$ ,  $\lambda(v)$  and  $\theta(v)$ , where  $p(v)$  is the split proportion parameter, we have

$$f_Y(y, v) = \theta(v) e^{-\theta(v)y}$$

$$f_X(x, v) = p(v) \lambda(v) e^{-\lambda(v)x}.$$

Note that  $f_X$  is an improper density and does not integrate to 1. Thus, on using the split population model, we have

$$S_Y(t - x, v) = e^{-\theta(v)(t-x)}$$

$$S_X(x, v) = 1 - p(v) + p(v) e^{-\lambda(v)x}$$

and

$$\int_0^x f_X(u, v) S_Y(t - u, v) du = \frac{p(v) \lambda(v) e^{-\theta(v)t}}{\theta(v) - \lambda(v)} \{e^{[\theta(v) - \lambda(v)]x} - 1\}.$$

Substituting from these equations into equation (8.6), we can then estimate the weight function  $w(x, t, v)$  for each case, provided the estimates of  $p(v)$ ,  $\lambda(v)$  and  $\theta(v)$  are available. In the following sections (in the applications), for simplicity we treat the case where  $p$ ,  $\lambda$  and  $\theta$  are constants and hence the weight function  $w(x, t, v)$  does not include the covariates, the general case can be handled in a similar way.

For the criminological data and under the delayed censoring models discussed in Chapter 3 (the marginal model), the maximum likelihood estimates of the parameters  $p$ ,  $\lambda$  and  $\theta$  were found to be  $\hat{p}=0.472$ ,  $\hat{\lambda}=0.00191$ ,  $\hat{\theta}=0.00668$ . Having these estimates, the estimates of the weights will then be available.

## 8.4 Estimating the Regression Parameters

Once we have obtained the estimates of the weights  $w(x, t, v)$ , we can consider the estimation of the regression parameters  $\beta$  under the weighted hazards model (8.13). To do this we proceed as follows, by analogy with the development of the partial likelihood for the Cox model. Let  $\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}$  be the set of ordered observed offence times and  $\{t_{(1)}, t_{(2)}, \dots, t_{(N-n)}\}$  be the set of ordered censoring times corresponding to the censored cases,  $n$  and  $N - n$  are the number of observed and censored cases respectively and  $N$  is the total sample size. We merge these two sets of times together to get the merged order statistics

$$\tau_1 \leq \tau_2 \leq \dots \leq \tau_{N-1} \leq \tau_N. \quad (8.14)$$

The set of individuals at risk at time  $\tau_i$  is called the risk set at time  $\tau_i$  and is denoted by  $R(\tau_i)$ ; this set is the number of people (out of  $N$ ) who have not yet failed (committed a crime) or been censored before time  $\tau_i$ . Let  $U$  be the set of uncensored cases and denote by  $C$  the set of censored cases. Then

$$R(\tau_i) = \{j \leq N \mid \text{if } j \in U \text{ then } \tau_j \geq \tau_i \text{ or if } j \in C \text{ then } c_j \geq \tau_i\} \quad (8.15)$$

for  $i = 1, 2, \dots, N$ , where  $c_j$ ,  $j = 1, 2, \dots, N$ , is the censoring time corresponding to the  $j$ th observation in the sample according to the ordering system for the merged order statistics  $\tau_1 \leq \tau_2 \leq \dots \leq \tau_{N-1} \leq \tau_N$ . Now define

$$\delta_i = \begin{cases} 1 & \text{if } \tau_i \text{ corresponds to an observed case} \\ 0 & \text{if } \tau_i \text{ corresponds to a censored case} \end{cases} \quad (8.16)$$

for  $i = 1, 2, \dots, N$ . Let  $\nu_i$  be the vector of covariates corresponding to  $\tau_i$ ,  $i = 1, 2, \dots, N$ . If  $\tau_i$  has  $\delta_i = 1$ , then

$$R(\tau_i) = \{i, i + 1, \dots, N\}. \quad (8.17)$$

Let  $h^*(\tau_i, c_j, \nu_j)$  and  $w(\tau_i, c_j, \nu_j)$  be the values of  $h^*(x, t, v)$  and  $w(x, t, v)$  respectively for the  $j$ th case at time  $\tau_i$ ,  $i = 1, \dots, N$ ,  $j \in R(\tau_i)$ . Then, as in the Cox model, the partial log-likelihood for estimating the parameters  $\beta = (\beta_1, \beta_2, \dots, \beta_p)$  is given by

$$\ell(\beta) = \sum_{i=1}^N \delta_i \log \left\{ \frac{h^*(\tau_i, c_i, \nu_i)}{\sum_{j=i}^N h^*(\tau_i, c_j, \nu_j)} \right\}. \quad (8.18)$$

Substituting from the weighted hazards model (8.13) into equation (8.18), we have

$$\ell(\beta) = \sum_{i=1}^N \delta_i \log \left\{ \frac{h_0(\tau_i) e^{\beta^T \nu_i} w(\tau_i, c_i, \nu_i)}{\sum_{j=i}^N h_0(\tau_i) e^{\beta^T \nu_j} w(\tau_i, c_j, \nu_j)} \right\}$$

or

$$\ell(\beta) = \sum_{i=1}^N \delta_i \left\{ \beta^T \nu_i + \log w(\tau_i, c_i, \nu_i) - \log \sum_{j=i}^N e^{\beta^T \nu_j} w(\tau_i, c_j, \nu_j) \right\}. \quad (8.19)$$

Note that we only need  $W_{ij}$  when  $\delta_i = 1$  and  $j \geq i$ , that is, only the upper triangular entries of the weight matrix  $W$  for the observed cases. Now assume that the estimated weights obtained in the previous section are known, then differentiating equation (8.19) with respect to  $\beta$  yields

$$\frac{\partial \ell(\beta)}{\partial \beta_k} = \sum_{i=1}^N \delta_i (\nu_{ki} - A_{ki}(\beta)), \quad k = 1, 2, \dots, p \quad (8.20)$$

where  $\nu_{ki}$  is the  $k$ th element of  $\nu_i = (\nu_{1i}, \nu_{2i}, \dots, \nu_{pi})^T$ , for  $i = 1, 2, \dots, N$ ,  $k = 1, 2, \dots, p$  and

$$A_{ki}(\beta) = \frac{\sum_{j=i}^N \nu_{kj} e^{\beta^T \nu_j} w(\tau_i, c_j, \nu_j)}{\sum_{j=i}^N e^{\beta^T \nu_j} w(\tau_i, c_j, \nu_j)}. \quad (8.21)$$

The maximum likelihood estimator for  $\beta$  can be obtained as a solution of the equations

$$\frac{\partial \ell(\beta)}{\partial \beta_k} = 0, \quad k = 1, 2, \dots, p.$$

These equations are non-linear in  $\beta_k$ , because  $A_{ki}(\beta)$  is a non-linear function of  $\beta_1, \dots, \beta_p$ , and thus we have to solve them by an iterative procedure. Differentiating equations (8.20) again, we get

$$-\frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_m} = \sum_{i=1}^N \delta_i B_{kmi}(\beta), \quad k, m = 1, 2, \dots, p \quad (8.22)$$

where

$$B_{kmi}(\beta) = \frac{\sum_{j=i}^N \nu_{kj} \nu_{mj} e^{\beta^T \nu_j} w(\tau_i, c_j, \nu_j)}{\sum_{j=i}^N e^{\beta^T \nu_j} w(\tau_i, c_j, \nu_j)} - A_{ki}(\beta) A_{mi}(\beta). \quad (8.23)$$

The information matrix from the partial likelihood is

$$I(\beta) = E\left(-\frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_m}\right).$$

An estimate for the  $(k, m)$ th element of this matrix can be obtained if we evaluate equation (8.22) at  $\hat{\beta}$ . The asymptotic variance-covariance matrix of the estimator  $\hat{\beta}$ , which is the inverse of the information matrix, will then be available and thus the standard errors of  $\hat{\beta}$  can be obtained. These can be used to conduct tests of significance about the covariates in the model. Of course, as mentioned in Chapter 6, this assumes that the partial likelihood satisfies the usual properties—Cox (1975). Also significant tests about subsets of parameters can be achieved by using the common log-likelihood ratio test.

For the criminological data, with  $\nu = (\text{age}, \text{ac}, \text{pre}, \text{jc})$  as the vector of covariates and  $\beta_1, \beta_2, \beta_3, \beta_4$  as the regression parameters, the maximum like-

Table 8.1: Parameter estimates (Cox Model)

coef	se(coef)	z-value
-0.0633	0.00825	-7.68
0.1329	0.03091	4.30
0.0295	0.00621	4.75
0.1304	0.03631	3.59

likelihood estimates and standard errors of these parameters (with asymptotic standard errors in parenthesis) are

$$\hat{\beta}_1 = -0.0635 (0.00824), \hat{\beta}_2 = 0.1369 (0.03072)$$

$$\hat{\beta}_3 = 0.0289 (0.00620), \hat{\beta}_4 = 0.1306 (0.03631).$$

Also the corresponding standardized estimates (the z-values) are  $z_1=-7.70$ ,  $z_2=4.45$ ,  $z_3=4.66$ ,  $z_4=3.59$  which are all statistically significant. The maximum value of the log-likelihood achieved is  $\ell(\beta)=-3159.242$ . These estimated values under the weighted hazards model (8.13) are then compared with those of the Cox proportional hazards regression model (8.12) which are given in Table 8.1. As we observe, very similar results under the two models are obtained, suggesting the validity of the weighted hazards model (8.13).

## 8.5 Generalized Weighted Hazards Model

In section 8.2 we assumed that  $X$  and  $Y$  are independent. However, the method extends to a more general model in which  $X$  and  $Y$  are correlated. We now consider the conditional distribution of  $Y$  given  $X = x$ , and use the notation  $S_{Y|x}$  for the conditional survival function of  $Y$ . Using the same argument as before, we now have the more general weight function

$$w(x, t, v) = \frac{1 - S_{Y|x}(t - x, v)}{1 + \frac{\int_0^x f_X(u, v) S_{Y|u}(t - u, v) du}{S_X(x, v)}}. \quad (8.24)$$

As an example, following section 7.1, suppose that

$$f_X(x, v) = p \lambda e^{-\lambda x}$$

$$S_X(x, v) = 1 - p + p e^{-\lambda x} \quad (8.25)$$

$$f_{Y|x}(y, v) = \theta^* e^{-\theta^* y}, \quad \theta^* = \theta e^{bx} \quad (8.26)$$

$$S_{Y|x}(t - x, v) = 1 - F_{Y|x}(t - x, v) = e^{-\theta^* (t - x)}. \quad (8.27)$$

Note that for  $b = 0$  we have  $\theta^* = \theta$ , which corresponds to the analysis in section 8.3. The parameters  $p, \lambda, \theta, b$  and hence  $\theta^*$  can in general be functions of vector of covariates  $v$ , but here for brevity of calculations we have dropped the argument  $v$  from these parameters.

Following the numerical results in section 7.1, we will assume that the parameter  $b$  is small. Now to evaluate the integral term in equation (8.24) we proceed as follows. Let

$$H(x, t, v) = \int_0^x f_X(u, v) S_{Y|u}(t - u, v) du,$$

then

$$H(x, t, v) = p \lambda \int_0^x e^{-\lambda u} e^{-\theta e^{bu}(t-u)} du,$$

but for small  $b$ ,

$$e^{bu} \approx 1 + bu,$$

so

$$H(x, t, v) \approx p \lambda \int_0^x e^{-\lambda u} e^{-\theta(1+bu)(t-u)} du,$$

which can be written as

$$H(x, t, v) \approx p \lambda e^{-\theta t} \int_0^x e^{du} e^{b\theta u^2} du,$$

where

$$d = \theta - b\theta t - \lambda.$$

Again for small  $b$  we have

$$e^{b\theta u^2} \approx 1 + b\theta u^2,$$

thus

$$H(x, t, v) \approx p \lambda e^{-\theta t} \left\{ \int_0^x e^{du} du + b\theta \int_0^x u^2 e^{du} du \right\}.$$

Hence we get, on simplifying the relevant expression,

$$H(x, t, v) \approx p \lambda e^{-\theta t} \left\{ \frac{e^{dx} - 1}{d} + b\theta \frac{e^{dx} [(dx - 1)^2 + 1] - 2}{d^3} \right\}. \quad (8.28)$$

Note that in the special case when  $b=0$ , we have

$$H(x, t, v) = p \lambda e^{-\theta t} \left\{ \frac{e^{x(\theta-\lambda)} - 1}{\theta - \lambda} \right\},$$

which agrees with simple result obtained in section 8.3, as expected.

Therefore,

$$w(x, t, v) = \frac{1 - S_{Y|x}(t - x, v)}{1 + \frac{H(x, t, v)}{S_X(x, v)}} \quad (8.29)$$

where  $S_X(x, v)$ ,  $S_{Y|x}(t - x, v)$  and  $H(x, t, v)$  are given by equations (8.25), (8.27) and (8.28) respectively.

In order to estimate the weights from equation (8.29), first we need to estimate the parameters  $p$ ,  $\lambda$ ,  $\theta$  and  $b$  under the assumption that the distribution of  $Y$  conditioning on  $x$  is given by equation (8.27). Using the method developed in section 7.1, the maximum likelihood estimates of these parameters are  $\hat{p}=0.442$ ,  $\hat{\lambda}=0.00216$ ,  $\hat{\theta}=0.00527$ ,  $\hat{b}=0.000815$ . Note that these estimated values of  $p$ ,  $\lambda$ ,  $\theta$  are very close to those obtained in section 8.3, and the value of  $b$  is very small which is consistent with the assumption about  $b$  in the analysis. Also following the numerical results in section 7.2, we have  $\frac{\hat{b}}{\text{se}(\hat{b})}=7.34$ , indicating that  $X$  and  $Y$  are correlated. Having the estimates of the parameters  $p$ ,  $\lambda$ ,  $\theta$  and  $b$ , the estimates of the weights will then be available and hence we can estimate the regression parameters  $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)$  as in section 8.4.

Here, the maximum likelihood estimates and standard errors of the regression parameters together with the corresponding  $z$ -values are summarized in Table 8.2. These results are very similar to those obtained in section 8.4, and hence again confirming the adequacy of the weighted hazards model (8.13).

We shall now compare the risk scores under the weighted hazards model (8.13) and the parametric restricted model developed in Chapter 3. For a given model, if  $V$  is the matrix of covariates and  $\beta$  is the vector of regression

Table 8.2: Parameter estimates (Weighted Model)

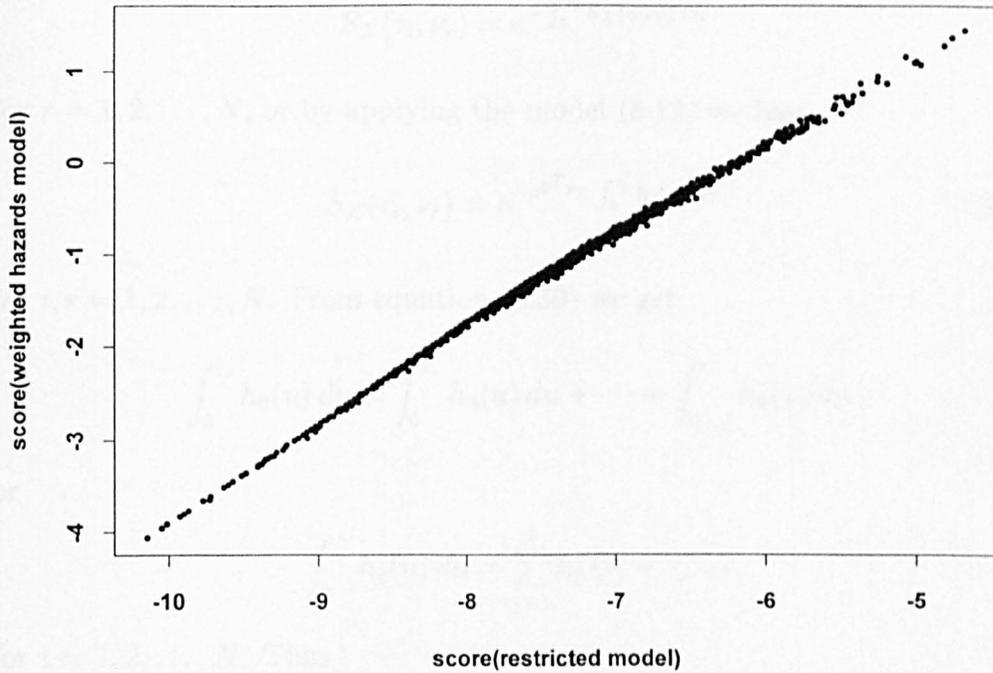
coef	se(coef)	z-value
-0.0634	0.00824	-7.69
0.1366	0.03072	4.44
0.0290	0.00620	4.68
0.1308	0.03630	3.60

parameters, then we define the risk score to be  $R_{sc} = \beta^T V$ . For the model (8.13) with estimated regression parameters given by Table 8.2, and for the restricted model with two years of follow up together with the relevant estimated regression parameters, the two sets of risk scores are illustrated in Fig. 8.1. The plot is very nearly linear, indicating that the two scores are very similar and highly correlated with each other.

## 8.6 Survival Curves by Risk Group

Once we have obtained the maximum likelihood estimate of the regression parameter  $\beta$ , we can consider the estimation of the survival curves of both the observed and actual reoffence time  $X$  under the weighted hazards model (8.13), and hence we can get the analogue of survival curves for risk groups as discussed in Chapter 4. To do this, first we need to find an estimate for the baseline hazard function  $h_0$  in model (8.13). One way of estimating  $h_0$ ,

Fig. 8.1 Comparison of Risk Scores for Two Models



as explained in section 6.4, is to carry out a separate maximum likelihood estimation procedure.

Let

$$\tau_1 \leq \tau_2 \leq \dots \leq \tau_{N-1} \leq \tau_N$$

be the merged order statistics as in section 8.4, and assume that

$$h_0(t) = \begin{cases} h_i & \text{if } \tau_{i-1} < t \leq \tau_i \\ 0 & \text{if } t > \tau_N \end{cases} \quad (8.30)$$

for  $i = 1, 2, \dots, N$ , and  $\tau_0=0$ . Then the model (8.12) can be written as

$$h_X(\tau_i, \nu_i) = h_i e^{\beta^T \nu_i}, \quad (8.31)$$

where  $\nu_i$  is the vector of covariates corresponding to  $\tau_i$ ,  $i = 1, 2, \dots, N$ . We

know that

$$S_X(\tau_i, \nu_r) = e^{-\int_0^{\tau_i} h_X(u, \nu_r) du}$$

for  $r = 1, 2, \dots, N$ , or by applying the model (8.12) we have

$$S_X(\tau_i, \nu_r) = e^{-e^{\beta T} \nu_r \int_0^{\tau_i} h_0(u) du} \quad (8.32)$$

for  $i, r = 1, 2, \dots, N$ . From equation (8.30) we get

$$\int_0^{\tau_i} h_0(u) du = \int_0^{\tau_1} h_0(u) du + \dots + \int_{\tau_{i-1}}^{\tau_i} h_0(u) du,$$

or

$$\int_0^{\tau_i} h_0(u) du = \sum_{j=1}^i h_j (\tau_j - \tau_{j-1}),$$

for  $i = 1, 2, \dots, N$ . Thus

$$S_X(\tau_i, \nu_r) = e^{-e^{\beta T} \nu_r \sum_{j=1}^i h_j (\tau_j - \tau_{j-1})}, \quad (8.33)$$

for  $i, r = 1, 2, \dots, N$ . Using equations (8.5) and (8.8), we have

$$\frac{1}{dx} P(A) = h_X(x, v) S_X(x, v) \{1 - S_Y(t - x, v)\}, \quad (8.34)$$

where the event  $A$  is defined by expression (8.1). Also from equations (8.7)

and (8.10) we get

$$P(B) = S_X(x, v) \frac{1 - S_Y(t - x, v)}{w(x, t, v)}, \quad (8.35)$$

where the event  $B$  is given by expression (8.2). Now using equations (8.34)

and (8.35) with the merged order statistics  $\tau_1 \leq \tau_2 \leq \dots \leq \tau_{N-1} \leq \tau_N$

and the other relevant quantities, the log-likelihood function for estimating

$h_i$ ,  $i = 1, \dots, N$ , is given by

$$\ell(h) = \sum_{i=1}^N \delta_i g_1 + \sum_{i=1}^N (1 - \delta_i) g_2, \quad (8.36)$$

where  $\delta_i$  is defined by equation (8.16) and

$$h = (h_1, h_2, \dots, h_N),$$

$$g_1 = \log h_X(\tau_i, \nu_i) + \log S_X(\tau_i, \nu_i) + \log\{1 - S_Y(c_i - \tau_i, \nu_i)\},$$

$$g_2 = \log S_X(\tau_i, \nu_i) + \log\{1 - S_Y(c_i - \tau_i, \nu_i)\} - \log w(\tau_i, c_i, \nu_i).$$

Alternatively, we can write

$$\ell(h) = \ell_1(h) + \ell_2 - \ell_3, \tag{8.37}$$

where

$$\ell_1(h) = \sum_{i=1}^N \delta_i \log h_X(\tau_i, \nu_i) + \sum_{i=1}^N \log S_X(\tau_i, \nu_i), \tag{8.38}$$

as in the ordinary Cox model, and

$$\ell_2 = \sum_{i=1}^N \log\{1 - S_Y(c_i - \tau_i, \nu_i)\},$$

which does not depend on  $h_X(\tau_i, \nu_i)$ , and

$$\ell_3 = \sum_{i=1}^N (1 - \delta_i) \log w(\tau_i, c_i, \nu_i),$$

which is known if we assume that the weights are known. Thus the approach developed here suggests that  $h_1, h_2, \dots, h_N$  are the same as in the ordinary Cox model. From equations (8.31) and (8.33) we have

$$\log h_X(\tau_i, \nu_i) = \log h_i + \beta^T \nu_i,$$

$$\log S_X(\tau_i, \nu_i) = -e^{\beta^T \nu_i} \sum_{j=1}^i h_j (\tau_j - \tau_{j-1}),$$

and hence

$$\ell_1(h) = \sum_{i=1}^N \delta_i \log h_i + \delta_i \beta^T \nu_i - \sum_{i=1}^N \{e^{\beta^T \nu_i} \sum_{j=1}^i h_j (\tau_j - \tau_{j-1})\}. \quad (8.39)$$

We have also

$$\frac{\partial \ell(h)}{\partial h_j} = \frac{\partial \ell_1(h)}{\partial h_j}, \quad j = 1, 2, \dots, N.$$

Differentiating equation (8.39) with respect to  $h_j$  we get, on simplifying the relevant expressions,

$$\frac{\partial \ell_1(h)}{\partial h_j} = \frac{\delta_j}{h_j} - (\tau_j - \tau_{j-1}) \sum_{k=j}^N e^{\beta^T \nu_k}, \quad (8.40)$$

for  $j = 1, 2, \dots, N$ . Thus, using equation (8.40), the maximum likelihood estimates of  $h_1, h_2, \dots, h_N$  are given by

$$\hat{h}_j = \frac{\delta_j}{(\tau_j - \tau_{j-1}) \sum_{k=j}^N e^{\beta^T \nu_k}}, \quad (8.41)$$

for  $j = 1, 2, \dots, N$ . Therefore, using equations (8.33) and (8.41), the fitted survival function of actual offence time  $X$  is given by

$$S_X(\tau_i, \nu_r) = e^{-e^{\beta^T \nu_r} \sum_{j=1}^i \left\{ \frac{\delta_j}{\sum_{k=j}^N e^{\beta^T \nu_k}} \right\}}, \quad (8.42)$$

for  $i, r = 1, 2, \dots, N$ .

Under the weighted hazards model when  $X$  and  $Y$  are assumed to be correlated, the new version of equation (8.35) is given by

$$P(B) = S_X^*(x, t, v) = S_X(x, v) \frac{1 - S_{Y|x}(t - x, v)}{w(x, t, v)} \quad (8.43)$$

where  $w(x, t, v)$  is defined by equation (8.29). Substituting from equation (8.29) into equation (8.43) yields

$$S_X^*(x, t, v) = S_X(x, v) + H(x, t, v) \quad (8.44)$$

where  $H(x, t, v)$  is given by equation (8.28). In terms of the merged order statistics defined by expression (8.14) we have now

$$S_X^*(\tau_i, c_r, \nu_r) = S_X(\tau_i, \nu_r) + H(\tau_i, c_r, \nu_r) \quad (8.45)$$

for  $i = 1, 2, \dots, N$ , where  $c_r, r = 1, 2, \dots, N$  is defined as in expression (8.15). Note that equation (8.45) defines the case-specific fitted survival function  $S_X^*(\tau_i, c_r, \nu_r)$ , for the observed offence times. So the Kaplan Meier (observed) survival curve for the offence times must be compared with the sample average (averaging over the covariates)

$$\bar{S}_X^*(\tau_i) = \frac{1}{N} \sum_{r=1}^N S_X^*(\tau_i, c_r, \nu_r) = \bar{S}_X(\tau_i) + \bar{H}(\tau_i) \quad (8.46)$$

where

$$\begin{aligned} \bar{S}_X(\tau_i) &= \frac{1}{N} \sum_{r=1}^N S_X(\tau_i, \nu_r), \\ \bar{H}(\tau_i) &= \frac{1}{N} \sum_{r=1}^N H(\tau_i, c_r, \nu_r) \end{aligned}$$

for  $i = 1, 2, \dots, N$ . If we plot  $\bar{S}_X^*(\tau_i)$  against  $\tau_i, i = 1, 2, \dots, N$ , then we will get an overall fitted survival curve for observed offence time  $X$ , and then we can compare it with the corresponding Kaplan Meier survival curve. However, as explained in section 4.7, this approach is most usefully done for interesting subsets of the data, e.g. the subsets defined by ranges of the risk score  $R_{sc} = \beta^T V$ .

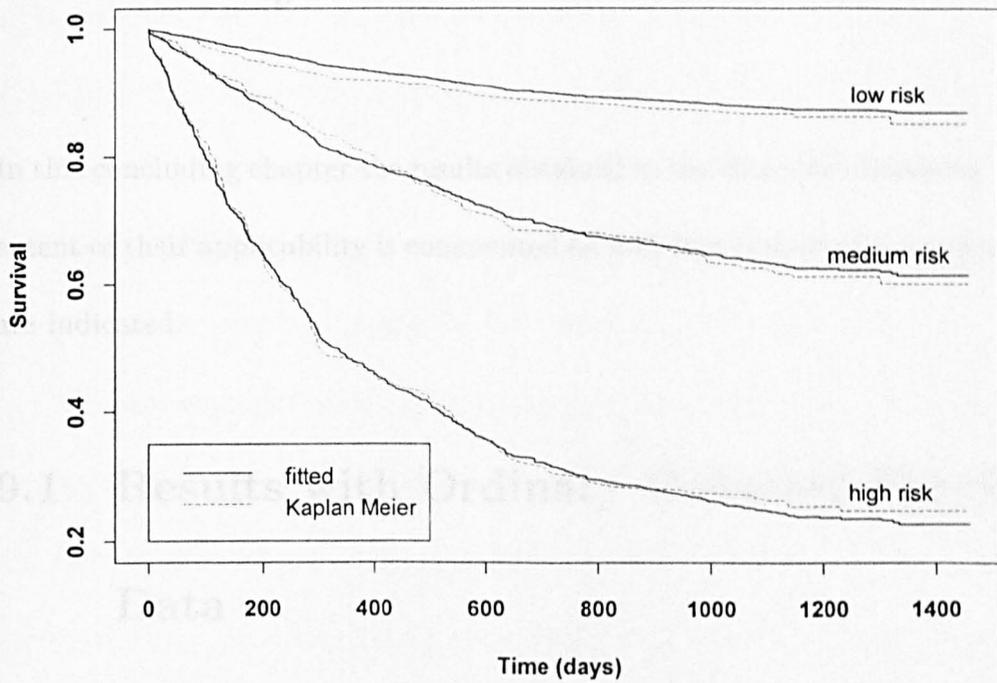
For the criminological data, fitted survival curves by risk group for the observed offence times with estimated risk scores  $R_{sc}$  under the weighted hazards model (8.13), using the weight function (8.29), are illustrated in Fig. 8.2,

which ranks the subjects in the data on the basis of  $R_{sc}$  and extracts those with the top 20% of values of  $R_{sc}$  (high risk cases), those with the middle 20% of values of  $R_{sc}$  (average risk group) and those with the lowest 20% of values of  $R_{sc}$  (low risk cases). Kaplan Meier (observed) survival curves for the reoffence times in each risk group are then compared with  $\bar{S}_X^*(\tau_i)$  averaging over the appropriate group of cases. This is very similar to Fig. 4.12 and gives an even slightly better fit to the data, as we might expect from the fact that the weighted hazards model is now a semi-parametric model.

The analysis could be extended by allowing the weights also to depend on the covariates. The fact that the results were similar for the different weight functions considered in this Chapter, suggests that more complicated weight functions are unlikely to be necessary.

Iterating calculations by reestimating the weights using the current model could also be considered.

Fig. 8.2 Observed and Fitted Survival Curves by Risk Group



# Chapter 9

## Summary and Conclusions

In this concluding chapter the results obtained in the thesis are discussed. The extent of their applicability is commented on and directions for future research are indicated.

### 9.1 Results with Ordinary Censored Survival

#### Data

For an initial exploration of survival distributions, we have considered the data as ordinary censored survival data by ignoring delayed censoring. We have explored the basic parametric survival models, exponential, Weibull and gamma with and without the split population model.

The results include:

- (1) Without split population model the exponential distribution is not a suitable model assumption for the reoffence times but with split population model

the exponential distribution gives a good fit to the reoffence times, suggesting that the split population model is necessary in studying the reoffence data.

(2) With and without split population model, the Weibull and gamma distributions give an almost equivalent good fit to the reoffence times. The fact that with and without split population model, the goodness of fits of each of these models (Weibull or gamma) are almost the same, is confirmed by noting that under the split population model the estimated value of  $p$  (the proportion of population who will reoffend) is close to 1.

(3) Under different prediction models for reoffence, models (1)–(5), discussed in Chapter 2, the *Reoffending Prediction Risk Scores* are highly correlated with each other. The risk scores are almost the same, implying that each of these models can be used as a model for deriving reoffending prediction risk scores.

We have explored more complicated parametric survival models, Weibull and gamma, but we have chosen the exponential model for the full analysis of the data in the thesis, as the exponential model is simple and interpretable in terms of  $\lambda$  (the rate parameter) and  $p$  (the proportion of population who will reoffend).

Most studies in the past have studied the time to reconviction rather than the time to reoffence. For completeness we have examined which if any of these models could also be used for examining the reconviction times. This is simpler in the sense that the problem of delayed censoring does not arise—data on reconviction times take the form of standard censored survival data.

Fitting these models to reconviction data, the results include:

(4) With and without split population model, the Weibull and gamma distributions give an almost equivalent good fit to the data. However, even with split population model, diagnostics shows that the exponential distribution would not be an appropriate model assumption for reconvictions, also as noted by earlier researchers in this area.

(5) Under different prediction models for reconviction, the risk scores are highly correlated with each other and they are almost the same, suggesting that each of these models can be used as a model for deriving reconviction prediction risk scores.

(6) Comparison of the relative importance of covariates shows that the reconviction and reoffending risk scores are almost the same under different models considered for the offence and conviction times. Consequently, the statistical risk scores obtained in the prediction models for reconviction can be used for reoffence as well.

## 9.2 Results with Delayed Censoring Analysis

With delayed censoring the results include:

(1) Diagnostics suggests that without a split population the exponential distribution is not a suitable model assumption for reoffence time, but with a split population the exponential model gives a good fit to the reoffence time. This result is consistent with the case in which the data were considered as ordinary censored survival data, suggesting that with delayed censoring the split

population model is also necessary for the full analysis of the reoffence data. This reflects the usefulness and applicability of the split population model.

(2) Using the split population model we have developed 3 parametric exponential mixture models which give good fit to the data. These include:

(a) the marginal model—model (3.10)

(b) the full model—model (3.13)

(c) the restricted model—model (3.17).

Comparison of the marginal, restricted and full models by hypothesis testing indicates that although the marginal model gives a good fit to the data (reoffence times), each of the restricted and full models provides a significantly better description of the data than the marginal model, suggesting that the covariates are strongly predictive of reoffending.

The restricted model has fewer parameters than the full model, which is difficult to fit to the data. Also analysis shows that the restricted model has smaller standard errors of the coefficients of  $\lambda$  (the rate parameter). Thus, we have reason for preferring the simpler restricted model (3.17). This is the main parametric model discussed in the thesis. We now go on to discuss the applicability of this model.

Longitudinal data is difficult and expensive to collect, more importantly, the longer the follow-up the earlier the sample has to be and so the less relevant is the study to current judicial and social conditions. Also there might be temporal changes in the incidence of crime and in the criminal justice system itself. Thus, there is a strong incentive to use the most recent data possible,

and hence the incentive to use as short a length of follow-up as possible. An advantage of a model which makes proper allowance for delayed censoring is that it can be fitted to data with a length of follow-up that is substantially shorter than the 3–4 years actually used. In Chapter 4, this has been illustrated by refitting the restricted model to the data that would have been obtained if all the subjects in the sample had been followed up for only 2 years. That is we have taken the time to follow-up  $T = 730$  and considered all cases with  $T > 730$  as if they were censored. This leaves 368 uncensored observations compared with the  $n = 486$  uncensored cases in the full data. The analysis shows that for 2-years data the restricted model still gives a good fit. Therefore, using the model suggested, risk scores for reoffending can be constructed using more up-to-date data and with lower data collection costs.

### 9.3 Results on Independence Analysis

Identification of independence is obviously an important aspect of any statistical analysis which should be justified. In our statistical analysis of censored survival data it has been assumed that the delay and reoffence times are independent exponentially distributed random variables. In order to assess whether the data collected conform to these assumptions, we have developed and applied a number of models to the data.

Examining the joint distribution of the delay time and the reoffence time is rather difficult because of the censoring— the sum of the two must always

be less than the time to follow-up. A simple way of examining the relationship it to plot the delay versus reoffence. However, in the light of this censoring the scatterplots of the data did not show any clear evidence of an association between these two quantities.

Another possibility of assessing the dependence of the delay and reoffence times is to study truncated distributions fitted to these data (truncation induced by censoring), through parametric, semi-parametric and nonparametric approaches. In this section we will discuss the results under these approaches.

In parametric approaches without covariates the following methods have been explored.

(a) Assuming  $Y$  (the time to delay) has an exponential distribution, with mean  $1/\theta$ , independent of  $X$  (the time to reoffence), we have derived the conditional mean and variance of the observed value of  $Y$  given  $X = x$ , from which we have plotted the standardized residuals of  $Y$  versus  $X$ . The plot suggests a small decrease in the mean of  $Y$  as  $X$  increases and hence there is some dependence between  $X$  and  $Y$  but not very much (correlation  $-0.01$ , just significant at the nominal 5% level).

(b) We have reestimated the same model in part (a) for different ranges of  $X$ , and hence plotted the estimates of  $\theta$  versus deciles of  $X$ . The plot indicates that  $\theta$  depends on  $X$ , and hence there is some dependence between  $X$  and  $Y$ . This plot suggests that we could fit a log-linear model

$$\text{model (5.7) : } \log \theta = a + b(x - \bar{x})$$

to the data. A significance test (log-likelihood ratio test) shows that  $b$ , the dependence parameter in the model, is significantly different from zero, suggesting dependence between  $X$  and  $Y$ . We have also repeated the standardized residuals plot using model (5.7); again the plot suggests that there is some dependence between  $X$  and  $Y$ .

(c) We have also examined models similar to the model (5.7) with quadratic, cubic and quartic terms. All these model gave an almost equivalent fit to the data and did not make any improvement as compared to the simple fitted model (5.7).

(d) The analyses in parts (a)—(c) suggest that there may be some doubts about the validity of the conditional exponential distribution of  $Y$  given  $X = x$ . And so we have explored more complicated parametric models, such as Weibull, gamma and mixed exponential to see if they might also be useful in explaining the dependence of  $Y$  on  $X$ . These models provide almost equivalent results as before for specifying the dependence of  $X$  and  $Y$ , suggesting that the simple exponential distribution for  $Y$  in the analysis of truncated data seems to be adequate. Of course it should be noted that all these analyses have been affected by the imputation of delay times, a rather arbitrary aspect of this particular data set.

It should be emphasized that the simple imputation given in Chapter 3 was merely an indication of the idea. Better approximations may also be possible and the investigation of other more complicated approaches is clearly of great interest for future research.

Doubts about the exponential assumption in the parametric analyses, suggests that we need a semi-parametric or nonparametric approach for the estimation of the truncated data.

In a semi-parametric approach without covariates we have developed the following model

$$\text{model (5.29) : } \hat{q}_{j-1} = \hat{q}_j \left(1 - \frac{m_j}{\alpha_j}\right), \quad j = 1, \dots, k$$

for discrete random variables with finite ranges, where  $k$  is the maximum observed value of the relevant random variable under consideration and  $q_j$  is the probability that the random variable takes the values less than or equal to  $j$ . See subsection 5.2.2 for details of the notations. Under suitable transformations this model can also be applied to estimate the marginal (unconditional) distribution of a continuous random variable.

As mentioned in subsection 5.2.2, the choice of  $q_k=1$  as an initial value for the iteration process may lead to unreasonable estimates. Great care is therefore needed in assigning initial values to  $q_k$ . Another possibility is to choose a suitable value of  $k$  and estimate  $q_k$  from a parametric model. Note that the value of  $q_k=1$  is all right if the random variable of interest really is discrete with a finite range.

Another difficulty associated with this model is that, when the model is used for estimating a distribution function we need to generate a counting matrix. This is easily possible provided that the dimension of this matrix is not too big. However, as the number of observations is increased or when we

are dealing with continuous data, this matrix becomes very large and requires too much computer memory. Thus, to overcome these difficulties, we have generalized this model and developed nonparametric models in which there is no need for the counting matrix of this kind.

For the criminological data this model gives a good fit to both the delay and reoffence times, and again suggesting exponential distributions for these two quantities. However, the marginal diagnostics by subgroups suggests that the delay and reoffence are correlated.

In the nonparametric approach without covariates we have developed models for estimating a distribution function with truncated data. These models are based on the empirical distributions of the truncated observations.

To introduce these models, let  $G$  and  $F$  denote unconditional distribution functions of two independent random variables  $Y$  and  $W$  respectively, where  $G$  and  $F$  are completely unknown. Suppose that  $Y$  is truncated at  $W$  from the right under some stochastic process such as censoring.

Let  $(y_1, w_1), \dots, (y_n, w_n)$  be a random sample of size  $n$  from  $(Y, W)$  for which  $y_i \leq w_i, i = 1, 2, \dots, n$ . So we are sampling from the joint conditional distribution of  $((Y, W)|Y \leq W)$  instead of the distribution of  $(Y, W)$  itself. Let  $G_n^*$  and  $F_n^*$  denote the empirical distributions of  $y_1, y_2, \dots, y_n$  and  $w_1, w_2, \dots, w_n$  respectively.

The main problem considered is to find nonparametric estimator of  $G$  from  $G_n^*$  and  $F_n^*$ .

For this problem we have developed the following asymptotically equivalent

models:

$$\text{model (5.41) : } \hat{G}_n(v) = \prod_{j:y_j > v} \left\{ 1 - \frac{m(y_j)}{nD_n(y_j)} \right\}$$

and

$$\text{model (5.45) : } \tilde{G}_n(v) = e^{-\sum_{j:y_j > v} \left\{ \frac{m(y_j)}{nD_n(y_j)} \right\}}$$

for  $0 \leq v < \infty$ , where the product and the summation are taken over distinct values of  $y_1, \dots, y_n$ , the observed values of the truncated random variable  $Y$ ,  $m(y_j)$  is the number of tied values at  $y = y_j$  and

$$D_n(u) = G_n^*(u) - F_n^*(u), \quad 0 \leq u < \infty.$$

The function  $\hat{G}_n$  defined in model (5.41) is the nonparametric maximum likelihood estimator of  $G$ . The estimator is the analogue of the product-limit estimator of Kaplan-Meier for randomly censored data.

The estimator  $\hat{G}_n$  is a step function and

$$\hat{G}_n(v) = 1, \text{ if } v > \max\{y_j\}, j = 1, \dots, n.$$

Since we are estimating  $G$  by  $\hat{G}_n$ , then

$$G(v) = 1, \text{ for } v > \max\{y_j\}, j = 1, \dots, n,$$

implying that

$$P(Y > \max\{y_j\}) = 0, j = 1, \dots, n.$$

This is similar to the problem of  $q_k=1$ , discussed in the semi-parametric model. Clearly this is not sensible if the censoring is very heavy and may lead to unreasonable estimates for those values of  $v > \max\{y_j\} = 0, j = 1, \dots, n$ .

Recall that in the criminological data if  $Y$  is the time to delay and  $X$  is the time to reoffence then  $Y$  and  $X$  are truncated at  $T - X$  and  $T - Y$  respectively from the right, where  $T$  is the time to follow-up (censoring time). So we can apply these models to the data.

For the criminological data the model,  $\hat{G}_n$  version, gives a good fit to both the delay and reoffence times, suggesting again exponential distributions for the delay and reoffence times. However, the marginal diagnostics by subgroups indicates that again there is dependence between the two quantities. The model (5.45) also gives a similar fit to the data, as expected, since  $\tilde{G}_n$  and  $\hat{G}_n$  are asymptotically equivalent.

Of course, similar derivations are possible for the estimation of  $F$ . In this regard we have developed the following two asymptotically equivalent models:

$$\text{model (5.51) : } 1 - \hat{F}_n(v) = \prod_{j:w_j \leq v} \left\{ 1 - \frac{m(w_j)}{nD_n(w_j)} \right\}$$

and

$$\text{model (5.52) : } 1 - \tilde{F}_n(v) = e^{-\sum_{j:w_j \leq v} \left\{ \frac{m(w_j)}{nD_n(w_j)} \right\}}$$

for  $0 \leq v < \infty$ , where  $m(w_j)$  is the number of tied values at  $w = w_j$ .

In the nonparametric approach incorporating the covariates we have developed a 'backward regression model' which is similar to the Cox proportional hazards model (Cox, 1972). This model has the form

$$h(x; v) = h_0(x) e^{\beta^T v}$$

in which we define the *backward hazard function*  $h$  to be

$$h(x; v) = \frac{g(x; v)}{P(X < x; v)}$$

We call this ‘backward hazard function’, since in a sense, we are ‘running time backwards’, as  $P(X > x)$  in the ordinary hazard function

$$h_*(x) = \frac{g(x)}{P(X > x)}$$

is being replaced by  $P(X < x)$  in ‘backward hazard function’. Here,  $g$  is the unconditional probability density function of  $X$ ,  $v$  is a vector of covariates and  $\beta$  is a vector of unknown parameters (regression parameters). In this model used for analysing truncated data, the main problem is that of assessing the relation between the unconditional distribution of  $X$  and the vector of covariates  $v$ . Note that in this model  $v$  can consist of the covariates actually being used, but can also include  $x$  or  $y$ , hence giving a new way of assessing the independence of  $X$ , the time to reoffence, and  $Y$ , the time to delay.

To estimate  $G$ , the unconditional distribution function of  $X$ , we define

$$h_0(x) = \begin{cases} h_1 & \text{if } x_{(0)} < x \leq x_{(1)} \\ h_2 & \text{if } x_{(1)} \leq x \leq x_{(2)} \\ h_i & \text{if } x_{(i-1)} < x \leq x_{(i)}, \quad i = 3, \dots, k \\ 0 & \text{if } x > x_{(k)}, \end{cases}$$

where we have defined  $x_{(0)} = 0$ . Note that in this model the term  $h_1$  is unidentifiable. So we need to put a constraint; arbitrarily we assume  $h_1 = h_2$ .

For the criminological data, marginal diagnostics suggest that this model gives a good fit to both the delay and reoffence times. This gives another

confirmation of the exponential distributions for the delay and reoffence times. However, examining different marginal diagnostics for the reoffence time  $X$ , indicate that there are some deviations from linearity at the right tails of the plots. This corresponds to the fact that we are assuming

$$h_0(x) = 0, \quad \text{for } x > x_{(k)}$$

or equivalently

$$g(x; v) = 0, \quad \text{for } x > x_{(k)},$$

a problem similar to that of  $q_k = 1$ , discussed in semi-parametric approach. This is the difficulty of our 'backward regression model' for large values of  $X$ . Clearly this assumption is not reasonable for the reoffence time  $X$ , as the censoring is very heavy in this case and the large reoffence times are being censored. Note that this assumption is sensible for the delay time  $Y$ , because in this case the censoring is not heavy and only very unusually large delay times are likely to be censored.

Comparison of the models by hypothesis testing, using log-likelihood ratio tests, shows that the highly significant result (dependence of  $X$  and  $Y$ ) in the parametric approach is no longer obtained in the current nonparametric model.

## 9.4 Results from Further Delayed Censoring Analysis (DCA)

In order to study the dependence of  $X$  and  $Y$  for all the data including both observed and censored cases, we have also developed a simple parametric DCA model in which the delay and reoffence times are allowed to be correlated, for the simple case excluding covariates. This model is given by

$$\text{model (7.1) : } f_{Y|x}(y) = \theta^* e^{-\theta^* y}, \quad \theta^* = \theta e^{bx}$$

$$f_X(x) = \lambda e^{-\lambda x}$$

$$S_{Y|x}(y) = e^{-\theta^* y},$$

where  $S_{Y|x}$  and  $f_{Y|x}$  denote the conditional survival and density functions of  $Y$  given  $X = x$  respectively and  $f_X$  is the density function of  $X$ . Note that in this model, dependency is represented by the parameter  $b$ . In models of this type it will then be possible to test if the parameter  $b$  is significantly different from zero and hence there will be an indication of dependence between  $X$  and  $Y$ .

For the criminological data, using split population model given by  $p$ ,  $\lambda$  and  $\theta$ , the analysis suggests that: the parameters

$p$  (the proportion of population who will reoffend)

$\lambda$  (the reoffending rate parameter)

$\theta$  (the rate parameter for delay)

are quite accurately estimated for this model because the estimated standard errors of these parameters are very small as compared with the maximum likelihood estimates of the parameters. Under this model there is clear dependence between the delay and reoffence times, as the regression coefficient  $b$  (the dependence parameter) in the model is significantly different from zero. Also under this model the estimated value of the quantity  $\frac{1-\hat{p}}{\text{se}(\hat{p})}$  is extremely large, indicating that the split population model is definitely necessary.

Note that this simple DCA model is merely an indication of the idea. The extension to the case of covariates could also be considered and handled in a similar way. As an alternative general parametric model for the analysis of the data, this can be another possibility for future research.

So far we have discussed some of the results under a number of truncated models in which the data are limited and can be observed in certain ranges because of the censoring. To allow for the censored observations as well, we have proposed a delayed censoring modification to the Cox model (Cox, 1972) and developed a general semi-parametric model for all the data including both observed and censored observations, firstly for the case in which the delay and reoffence times are assumed to be independent. This model is then extended to a more general model in which the delay and reoffence times are allowed to be correlated.

To represent these models let  $X$  and  $Y$  be the offence and delay times respectively and first assume that they are independent. Denote hazard, density and survival functions of  $X$  by  $h_X$ ,  $f_X$  and  $S_X$  respectively, and let  $S_Y$  be the

survival function of  $Y$ . Let  $t$  be the time of censoring (the time to follow-up) and  $v$  be the vector of covariates specific to each observation in the sample. Let  $h^*$  be the hazard function for observed offence time. Then the model is given by

$$\text{model (8.4) : } h^*(x, t, v) = h_X(x, v) w(x, t, v).$$

In this model we refer to  $w(x, t, v)$  as the weight function,  $0 \leq w(x, t, v) \leq 1$ , and so we refer to the model as 'weighted hazards model'.

If we put some Cox proportional hazards assumptions (Cox, 1972) on  $h_X$ , then the model is given by

$$\text{model (8.13) : } h^*(x, t, v) = h_0(x) e^{\beta^T v} w(x, t, v)$$

where  $\beta$  is a  $1 \times p$  vector of unknown parameters and  $h_0(x)$  is an unspecified function of time  $x$  common to all observations. In the special case where the weight function  $w(x, t, v)$  is equal to 1, the weighted hazards model reduces to the ordinary Cox model.

When  $X$  and  $Y$  are independent the weight function is given by

$$w(x, t, v) = \frac{S_X(x, v) \{1 - S_Y(t - x, v)\}}{S_X(x, v) + \int_0^x f_X(u, v) S_Y(t - u, v) du}$$

which can be estimated from one of the parametric models developed in the thesis.

When  $X$  and  $Y$  are allowed to be correlated, we have the more general weight function

$$w(x, t, v) = \frac{S_X(x, v) \{1 - S_{Y|x}(t - x, v)\}}{S_X(x, v) + \int_0^x f_X(u, v) S_{Y|u}(t - u, v) du}$$

where  $S_{Y|x}$  is now the conditional survival function of  $Y$  given  $X = x$ . In this case we refer to the model (8.4) or (8.13) as ‘generalized weighted hazards model’.

On the basis of the whole data and under the weighted hazards models discussed, the main problem is that of assessing the relation between the unconditional distribution of  $X$  and the vector of covariates  $v$ , for both the actual and observed reoffence times. For the observed reoffence time we can then compare the observed survival curve with the Kaplan Meier survival curve calculated directly from the data.

Once we have obtained the estimates of the weights  $w(x, t, v)$ , we can estimate the regression parameters  $\beta$  under the weighted hazards models, by analogy with the development of the partial likelihood for the Cox model (Cox, 1975). See section 8.4 for details. Significant tests about subsets of parameters can be achieved by using the common log-likelihood ratio test.

Once we have obtained the estimates of the regression parameter  $\beta$ , we can estimate the survival curves of both the observed and actual reoffence time  $X$  under the weighted hazards models. The estimated survival function of actual reoffence time  $X$  at  $X = \tau_i$  and covariate  $\nu_r$  is given by

$$S_X(\tau_i, \nu_r) = \exp\{-\exp(\beta^T \nu_r) \sum_{j=1}^i \left\{ \frac{\delta_j}{\sum_{k=j}^N \exp(\beta^T \nu_k)} \right\}\}$$

for  $i, r = 1, 2, \dots, N$ ,  $N$  being the total sample size, and the fitted survival function for the observed reoffence time is given by

$$S_X^*(\tau_i, c_r, \nu_r) = S_X(\tau_i, \nu_r) \frac{1 - S_{Y|x}(c_r - \tau_i, \nu_r)}{w(\tau_i, c_r, \nu_r)}.$$

Again, see section 8.6 for details of the notations. These equations define the case-specific fitted survival curves. So the Kaplan Meier (observed) survival curve for the offence times must be compared with the sample average (averaging over the covariates)

$$\bar{S}_X^*(\tau_i) = \frac{1}{N} \sum_{r=1}^N S_X^*(\tau_i, c_r, \nu_r).$$

We have estimated the weight functions, but without covariates, and applied these weighted hazards models to the criminological data. The results of the analysis are:

(1) Under different weight functions considered, very similar results for parameter estimates and standard errors of these regression parameters are obtained, suggesting the validity of these weighted hazards models.

(2) Plot of reoffending risk scores ( $R_{sc} = \beta^T V$ ,  $V$  the matrix of covariates) under the generalized weighted hazards model and the parametric restricted model developed in Chapter 3, suggests that the two risk scores are highly correlated with each other and very similar.

(3) We have compared the survival curves by risk groups under the restricted and the generalized weighted hazards model and very similar results are obtained. The generalized weighted hazards model gives an even slightly better fit to the reoffence data, as we might expect from the fact that the generalized weighted hazards model is now a semi-parametric model.

Note that the analysis could be extended by allowing the weights also to depend on the covariates. The fact that the results were similar for the different

weight functions considered in the analysis, suggests that more complicated weight functions are unlikely to be necessary. This can be another possibility for future research.

Iterating calculations by reestimating the weights using the current model could also be considered. Again, this is another area of future research.

## 9.5 Further Applications

In this thesis we have proposed the 'delayed censoring' problem. On the basis of this new approach we have developed a number of new statistical models for the analysis of censored survival data. Application of these models to the criminological data indicates the usefulness and applicability of the models for analyzing the recidivism data. However, these models can be widely used in other applications of statistics.

In a typical medical application, consider the death registration problem in a cancer register. For an individual in the data let  $X$  be the time from onset of cancer to death (the actual time of the death) and  $Z$  be the time to notification of the death. Then there is often several weeks of delay,  $Y$ , in the notification of the death time. Here, censoring may be caused by the patient moving to another area, or to the end of a particular study period. So we have a delayed censoring problem and we can apply the delayed censoring models developed in the thesis.

In an engineering application, suppose that  $X$  is the time from the installa-

tion of a piece of equipment to its first failure, and let  $Z$  be the corresponding time of notification of the failure time. Then there may in fact be a delay,  $Y$ , in the notification of the failure time. Here, censoring could be withdrawal of the equipment for some other reason. Again, this is an example of a delayed censoring problem which can be analyzed using delayed censoring models.

In general these delayed censoring models can be used in any situation where there is a delay in investigating failure times.

# Bibliography

Allison, P. D. (1984) Event history analysis. *Quant. Applic. Socl Sci.*, **46**.

Beverly Hills: Sage Publications.

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993) *Statistical Models Based on Counting Processes*. New York: Springer Verlag.

Andersen, P. K. and Gill, R. D. (1982) Cox's Regression Model for Counting Processes: A Large Sample Study. *Ann. Statist.*, **10**, 1100–1120.

Ansell, J. I. and Phillips, M. J. (1994) *Practical Methods for reliability Data Analysis*. New York: Oxford University Press Inc.

Barclay, G. C., Tavares, C. and Prout, A. (1995) *Information on the Criminal Justice System in England and Wales*. London: Home Office Research and Statistics Department.

Barlow, W. E. and Prentice, R. L. (1988) Residuals for relative risk regression. *Biometrika*, **75**, 65–74.

Barton, R. R. and Turnbull, B. W. (1979) Failure rate regression models for evaluation of recidivism data. *Eval. Q.*, **3**, 629–641.

——(1981) A failure rate regression model for the study of recidivism. In

- Models in Qualitative Criminology* (ed. A. J. Fox). New York: Academic Press.
- Carlisle Committee (1988) *The Parole System in England and Wales: Report of the Review Committee*. London: Her Majesty's Stationery Office.
- Cleveland, W. S. (1979) Robust locally weighted regression and Smoothing scatterplots. *J. Am. Statist. Ass.*, **74**, 829-836.
- Copas, J. B., Marshall, P. and Tarling, R. (1996) *Predicting Reoffending for Discretionary Conditional Release*. London: Her Majesty's Stationery Office.
- Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc. B*, **34**, 187-220.
- Cox, D. R. (1975) Partial likelihood. *Biometrika*, **62**, 269-276.
- Farewell, V. T. (1982) The use of mixture models for the analysis of Survival data with long-term survivors. *Biometrics*, **38**, 1041-1046.
- Harris, C. M. and Moitra, S. (1978) Improved statistical techniques for the measurement of recidivism. *J. Res. Crime and Delinq.*, **15**, 194-213.
- Kaplan, E. L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. *J. Am. Statist. Assoc.*, **53**, 457-481.
- Keiding, N. and Gill, R. D. (1990) Random truncation models and Markov processes. *Ann. Statist.*, **18**, 582-602.
- Liu, P. Y. and Crowley, J. (1978) "Large Sample Theory of the MLE based on Cox's Regression Model for Survival Data." *Tech. Rep. NO. 1*, Wisconsin Clinical Cancer Center.

- Mallor, R. A. and Zhou, S. (1992) Estimating the proportion of immunes in a censored sample. *Biometrika*, **79**, 731–739.
- (1996) *Survival Analysis with Long Term Survivors*. New York: Wiley.
- Maltz, M. D. (1984) *Recidivism*. Orlando: Academic Press.
- Nuttall, C. P. (1977) *Parole in England and Wales*. London: Her Majesty's Stationery Office.
- Oakes, D. (1981) Survival Times: Aspects of Partial Likelihood. *Int. Statist. Rev.*, **49**, 235–264.
- Schmidt, P. and Witte, A. D. (1988) *Predicting Recidivism using Survival Models*. New York: Springer.
- Stollmark, S. and Harris, C. M. (1974) Failure rate analysis applied to recidivism data. *Ops Res*, **23**, 1192–1205.
- Tarling, R. (1993) *Analysing Offending*. London: Her Majesty's Stationery Office.
- Tsiatis, A. A. (1981) A large sample study of Cox's regression model. *Ann. Statist.*, **9**, 93–108.
- van der Lann, M. J. (1995) An identity for the nonparametric maximum likelihood estimator in missing data and biased sampling models. *Bernoulli* **1(4)**, 335–341.
- Wang, M. C., Jewell, N. P. and Tsai, W. Y. (1986) Asymptotic properties of the product limit estimate under random truncation. *Ann. Statist.*, **14**, 1597–1605.
- Ward, D. (1987) *The Validity of the Reconviction Prediction Score*. London:

Her Majesty's Stationery Office.

Woodroffe, M. (1985) Estimating a distribution function with truncated data.

*Ann. Statist.*, **13**, 163–177.

—Correction: *Ann. Statist.*, **15** (1987), 883.

This thesis has been prepared using the guidelines on thesis presentation in the *Guide to Examinations for Higher Degrees by Research*, published by the University of Warwick, Graduate School, September 1996.