# Nonparametric Methods for the Exploration and Analysis of Survival Data

Eileen McCormick Wright B.Sc.(Hons)

*A Dissertation Submitted to the*

*University of Glasgow*

*for the degree of*

*Doctor of Philosophy*

Department of Statistics

March, 1995.

To my Mum and Dad

## Abstract

Traditional survival analysis methods are primarily those of Kaplan–Meier curves, the log–rank test and Cox's Proportional Hazards model. Only the first of these techniques is routinely used to provide a graphical representation of the data. The idea of a regression curve is used to describe the relationship between survival time and a continuous covariate is rarely considered. This is presumably due to the complexity of estimating a mean when there are censored observations.

Median survival times are often quoted for a set of analysed data and extending this to a median curve across a continuous covariate would provide an intuitive description of the effect of this covariate on survival time. In this thesis, a combination of two nonparametric procedures using kernel estimates provides a doubly–smooth quantile estimator for the $p^{th}$ $(0 \leq p \leq 1)$ quantile of survival time given a covariate.

Similar percentile curves can be derived for both Cox's model and a smooth proportional hazards models. While these allow a more explicit form of the curve to be written down, the doubly–smooth estimator has no assumptions about the baseline hazard rate or the shape of the covariate effect and is therefore more flexible.

Assessing and comparing the fits of each of these approaches can be achieved by the calculation of a form of likelihood statistic. Due to the complexity of the mathematical properties of the nonparametric method, testing procedures are carried out using resampling techniques such as bootstrapping and permutation tests.

One extension of this methodology is to consider the additional effect of a binary covariate on survival time. This is analogous to an analysis of covariance in a Normal regression model and interest lies in how to characterise the behaviour of the curves from each of the two levels. As before, percentile curves can be obtained and appropriate testing procedures applied.

An algorithm based on serum creatinine curves was developed to detect graft deterioration in kidney transplant patients. These diagnoses had previously only been made by the subjective, experienced opinions of physicians, whereas the algorithm provided an explicit rule for detecting these cases. Survival times were also obtained and these data were analysed using standard techniques. Percentile curves were used to provide more information where the interpretation of a covariate effect was difficult.

In the absence of censored data, a different form of nonparametric smoothing was considered to assess the development of children suffering from cerebral palsy. Percentile curves were obtained using cubic splines to describe the growth of children with this condition and to compare them with those of normal children. This not only vindicated the belief that children suffering from cerebral palsy tend to be smaller and lighter than normal children of similar age but also provided standard curves useful in monitoring the development of these children.

# Contents

# List of Tables

# List of Figures

## Acknowledgements

I wish to thank Dr. Adrian W. Bowman for his supervision, insight and patience throughout my Ph.D. I also acknowledge the financial support of the ESPRC who funded me for the duration of this degree.

I would also like to thank Dr. Peter Rowe of the Renal Unit at the Western Infirmary who provided me with a very interesting applied problem, helping me to appreciate the use of statistics in medical applications. I am also grateful to Dr. Scott Zeger at Johns Hopkins University in Baltimore for the valuable experience gained working there during the summer of 1993. I would like to thank Dr. Patrick Royston for his support during the time between my appointment at the Royal Postgraduate Medical School and submission of this thesis.

The members of staff at the Statistics department in Glasgow have provided support during both my postgraduate and undergraduate degrees. In particular I would like to thank my friends Andrew Holmes, Graeme Archer, Catriona Hayes and Rodney Wolff for making the past three years such fun.

Finally, I would like to thank my family. My parents, brother and grandparents have been very supportive throughout my student years and completion of this thesis was aided by their love and confidence in me.

# Chapter 1

# Introduction

## 1.1 Survival Data

Survival data arises when there is interest in the length of time until a particular event occurs. Although there are many areas of application for this type of data, the most common is probably in medicine. Individuals who experience the event are classed as *failures*. However, there will be cases where the failure time is not observed. This may be for reasons independent of the condition under investigation. For example, it may be simply because they have not failed within the period of the study. The end point in these cases are classed as *censored* times. To investigate factors which may affect the observed time to failure, one or more *covariates* for the individuals are often measured and these possible effects can be analysed.

## 1.2 Techniques for Survival Analysis

The most common techniques for analysing survival data can be split into three categories.

1. Nonparametric – The product–limit estimator proposed by Kaplan and Meier (1958) estimates the distribution of the failure times. No covariate information is used in this method.

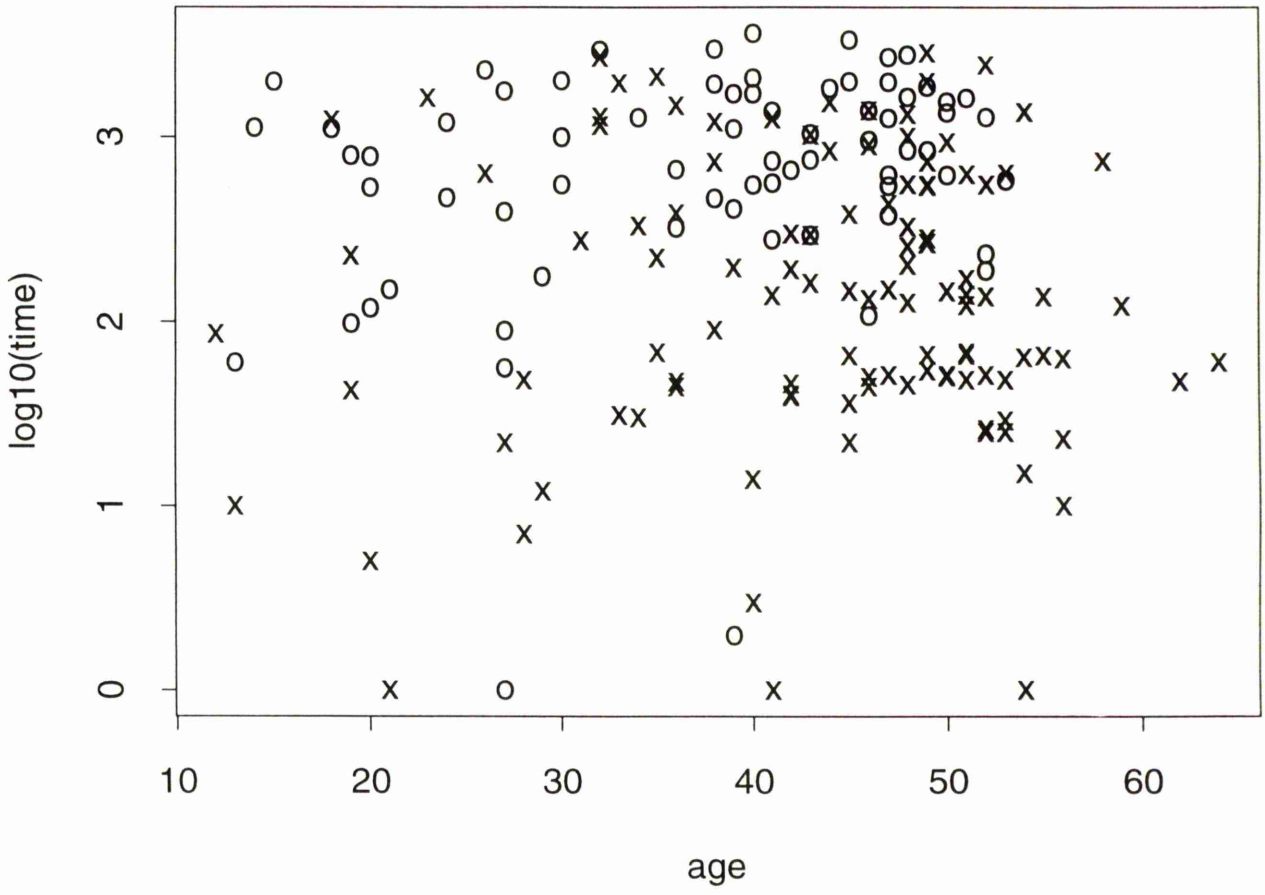
2. Semiparametric – Cox's (1972) Proportional Hazards model proposes a hazard function which is composed of a baseline hazard and a parametric part which assumes that covariate values act multiplicatively on the hazard. Estimation of the parameter involved is usually carried out using partial likelihood. Testing whether this parameter could be zero, provides an assessment of whether the covariate has any effect on survival time.

3. Parametric – One example of this is the Accelerated Failure Time model, as described by Kalbfleisch and Prentice (1980). This also has a hazard function which is expressed as the product of a baseline hazard and a parametric part. However, the covariates act multiplicatively on the time variable under this model, implying that the baseline hazard rate changes (i.e. accelerates or decelerates) unlike that of the Proportional Hazards Model.

**Main Aim** – The main aim of this research is to explore ways of describing the relationship between survival time and a covariate of interest in a nonparametric manner. These descriptions can then be used as graphical methods and as means of carrying out nonparametric inference.

Consider the Stanford Heart Transplant data which have been extensively analysed by survival data techniques. Figure 1.1 is a scatterplot of the logarithm of survival time against age. (A cross denotes a failure time and the circles are censored values) There may be a slight decrease in survival time as age increases but an appropriate line superimposed on the plot to describe this effect would be more informative.

Figure 1.1 : Stanford Heart Transplant Data



To explore and model the data, the most obvious direction to take is to consider the plot of time against covariate and fit a regression curve to the data. For *uncensored* data, there are various techniques to estimate the mean regression line or curve, from the parametric approach of least squares estimation to nonparametric methods of kernel estimation and fitting cubic splines, as well as many others.

This is more difficult when there are *censored* observations, since due to incompleteness of the data, it is not straightforward to calculate a mean. Recent work by Fan and Gijbels (1994) discussed ways of transforming censored times to failures, then by implementing a regression technique, a mean curve could be estimated. However, if we wish to leave the censoring mechanism intact, another approach should be used.

Doksum and Yandell (1982) suggested a nonparametric mean regression curve based on an identifiable integral which represents the expectation of survival time given covariate values. Also discussed in this article is Beran's (1981) product-limit median. The survival time corresponding to the point at which the estimated survivor function is equal to 0.5 is found for the $k^{th}$ nearest neighbours around each covariate value. The collection of these "running medians" provides a simple, but informative, description of the effect of the covariate on survival.

## 1.3 Quantile Curves

For a sample of data, the $p^{th}$ percentile is defined as the point below which $p\%$ of the observations lie. Percentile charts, displaying percentiles as a function of other variables, are particularly popular for describing growth since not only can the median be estimated for a range of covariate values but, the results and ideas also extend to a range of estimable percentiles.

Median survival times are often quoted in the context of survival data to give a rough indication of time to failure for a particular condition. One way to assess the effect of a covariate of interest on survival time would be to analyse how median survival times change for different values of the covariate. Plotting

survival time against covariate, with a superimposed line indicating the median (or more generally, $p^{th}$ quantile) times could be highly informative.

Different methods of estimating percentile survival time will be discussed in the next chapter, particularly those of a nonparametric nature, but a brief summary is given below.

1. Nonparametric

   - The empirical quantile function produces a point estimate. The idea can be extended to estimation over a range of covariate values, see e.g. Beran (1981), Doksum and Yandell (1982)
   - The kernel estimator of Padgett (1986) provides a smoother approach for evaluating a percentile but does not contain any information about the covariate. Alternatively, Gentleman and Crowley (1991) suggest kernel estimates to calculate a Kaplan–Meier estimator weighted by covariate values.
   - Extension of the idea of cubic splines to estimate percentiles is straightforward but it not clear how to deal with censored data.

2. The Proportional Hazards Model

   - Once a proportional hazards model has been fitted, the survivor function can be estimated and used to find the percentile survival time. By estimating over a range of covariate values, Doksum and Yandell (1982) produced a percentile curve based on the model.
   - Tibshirani and Hastie (1987) proposed an extension of Cox's proportional hazards model in which the covariate effect on survival time was assumed to be smooth and estimable by nonparametric regression procedures. Quantiles under the smooth model can be estimated in a similar manner to those produced by Doksum and Yandell (1982). This is effectively a nonparametric approach and could have been listed with those above in 1. However, there is still an assumption of proportionality of the effect of the covariate on the hazard function.

**Proposal : A doubly–smooth quantile estimator**

This estimator smooths across the values of the covariate using a weighted Kaplan–Meier estimator and across the quantile space using a nonparametric quantile estimator. Its basis are the ideas of Padgett (1986) and Gentleman and Crowley (1991) mentioned above. Both procedures come from kernel estimates and the combination of these result in a graphically smooth curve.

This is demonstrated in Figure 1.2 for the Stanford Heart Transplant data, where survival time is now clearly decreasing as age increases which was not as obvious simply from the scatterplot. (This dataset will be discussed further as an example in chapters 3 and 4.)

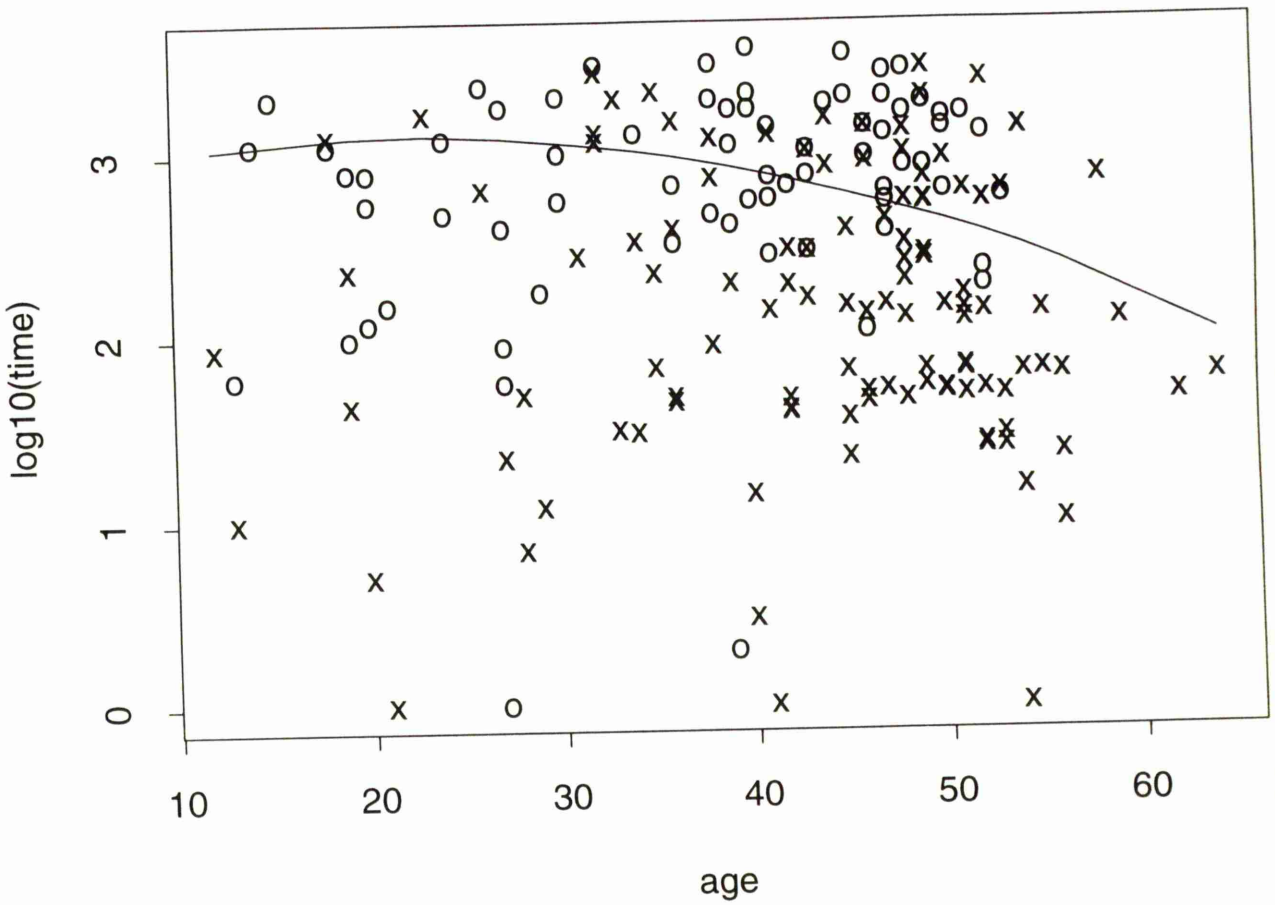
## 1.4 Testing for an effect

A good, clear subjective impression can be gained from the percentile curves but it is also useful to be able to follow it up with a formal testing procedure. We consider a variety of models of interest.

1. The covariate has *no effect* on survival. i.e. the values of the covariate appear to be randomly distributed about the survival times.
2. The effect of the covariate is most aptly described by a curve estimated from nonparametric regression techniques which assumes *neither proportionality* of hazards *nor linearity* of covariate effect.
3. The effect of the covariate is *non–linear*, but still *proportional* with respect to a baseline hazard.
4. The effect of the covariate is *linear*, within a Proportional Hazards Model.

"Linear" here refers to the hazard function in the log scale. To estimate percentile curves from each of the last three types of covariate effect requires the same basic approach, with the main difference being the estimation of the survivor function. In order to summarise the information from the survivor functions across the range of covariates, a form of likelihood statistic is considered.

Figure 1.2 : Median Survival Time



The expression for this is the product of contributions from failure and censored times given their corresponding covariate values. Estimation of the underlying density, or equivalently the gradient of the survivor function, is an appropriate measurement for the failure times, with the censored times contributing the value of the survivor function at that point.

However, this statistic has little informative use if not accompanied by an appropriate testing procedure. Due to the mathematical complexity of the theory behind the proposed nonparametric method, resampling techniques are employed, providing a more formal way of analysing which of the above formulations of the covariate effect provides a suitable description of the data.

## 1.5 Applied Work

Two applied statistical problems have been tackled, both involving nonparametric curves of different types. A brief summary of each of these is given below.

### 1.5.1 Renal Unit

Research with Dr. Peter Rowe of the Renal Unit at Glasgow's Western Infirmary has focussed on the deterioration of a kidney graft in transplant patients. The main aim initially was to construct a way of detecting if, and when, progressive graft dysfunction, a particular type of deterioration, had occurred. This is often routinely decided by physicians on the basis of decreasing serum creatinine levels over time. The aim was to develop an algorithm, using elementary statistical techniques, to mimic the ideas and decisions made by the experienced medical eye. This would provide a more objective and standard procedure for classifying cases where this condition has occurred, as well as making explicit the features of the serum creatinine profile which leads to diagnosis of progressive graft dysfunction.

Using the results from the algorithm, an investigation was carried out into which factors are associated with the occurrence of this condition. A combina-

tion of traditional survival analysis methods and the percentile curves approach revealed that only a very small number of covariates had significant effect, but that these effects were important.

### 1.5.2 Linear Growth Curves

At the Kennedy Krieger Institute in Baltimore, researchers were interested in the growth development of children with cerebral palsy and how they compared to normal children of similar age. The height and weight measurements of children suffering this condition who visit the Institute had been recorded.

A common approach to analysing this kind of data is linear growth percentile curves. Charts for normal children have been produced in this way and provided a means of comparison to children with cerebral palsy. Under the supervision of Dr. Scott Zeger, the percentile curves were estimated using a weighted cubic spline where the weights were used to transform the spline from estimating the mean to estimating the quantile of interest. Since the amount of data used to calculate the percentile curves for the cerebral palsy children was much less than that used to calculate the curves for normal children, it was important to assess the amount of variability in these new curves. This was achieved by plotting the percentile curves from bootstrap samples.

It is hoped that these linear growth curves for children with cerebral palsy will be used in both the Kennedy Krieger Institute and other such centres to help in the understanding of the development of these children.

## 1.6 Summary of Chapters

CHAPTER 2 is a review of quantile estimators and their properties. While it attempts to cover many of the different approaches described in the literature, nonparametric estimators and those relevant to survival data are discussed in most detail.

CHAPTER 3 describes and discusses the doubly–smooth quantile estimator. This is pursued through graphical examples and the calculation of a likelihood statistic. A principal aim of this chapter is to compare this new approach with those based on a commonly used form of model, such as proportional hazards.

CHAPTER 4 considers the use of resampling techniques as the basis of inferential tests. Standard permutation tests are applied and a review of bootstrap methods for survival data is given as a justification for the procedures used to compare different models.

CHAPTER 5 takes all the proposed ideas from the previous two chapters and extends them to the case where, in addition to being interested in a continuous covariate, the effect of a binary covariate may also be important.

CHAPTER 6 gives full details of the analysis carried out to detect a particular form of deterioration in transplanted kidneys and the investigation into factors which influence this condition.

Finally, CHAPTER 7 describes the construction of centile growth curves for children with cerebral palsy using a cubic spline approach.

# Chapter 2

# Review of Quantile Estimators

The following sections review quantile estimators by stepping through a development of ideas, considering at each stage those for the uncensored and censored cases. After stating a few preliminary definitions, basic empirical quantile estimators are discussed and this is followed by an extension to nonparametric smoothing quantiles. To incorporate a covariate, nonparametric quantile regression curves are reviewed. Finally, there is a discussion of how to obtain similar curves under Cox's Proportional Hazards model. This chapter is intended to be a comprehensive review of published work in this field, hence many of the mathematical results are quoted for completeness, without further exploration.

## 2.1 Preliminaries

### 2.1.1 Uncensored Case

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with continuous cumulative distribution function $F$. Suppose we are interested in the $(100 \times p)^{th}$ quantile $(0 \leq p \leq 1)$ of $X$, i.e. where $F(x) = p$. Note that we are effectively considering the inverse of the distribution function and that the quantile function can be written as

$$Q(p) = F^{-1}(p) = \inf\{x : F(x) \geq p\}$$

Let $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ denote the order statistic of the $X_i$'s.

### 2.1.2 Random Censorship Model

Let $X_1^o, \ldots, X_n^o$ denote a sample of failure times for $n$ items or individuals. Since these cannot always be observed, we also have a sequence $U_1, \ldots, U_n$ of censoring variables.

The observed right-censored data are usually denoted by the pairs $(X_i, \Delta_i)$ for $i = 1, \ldots, n$ where

$$X_i = \min\{X_i^o, U_i\} \qquad \Delta_i = \begin{cases} 1 & \text{if } X_i^o \leq U_i \\ 0 & \text{if } X_i^o > U_i \end{cases}$$

So the value $\Delta$ of a corresponding time value $X$ will tell us whether or not that individual was observed to fail.

The most widely used (and probably most realistic) situation is the one in which the *random censorship model* is applicable. Here we assume that the $X_i^o$'s are non-negative, independently and identically distributed random variables with common, unknown distribution function $F_o$. Also that the $U_i$ 's are a random sample from a distribution $G$ and are independent of the $X_i^o$ 's.

Under this model, the distribution function of each $X_i$ is

$$F = 1 - (1 - F_o)(1 - G)$$

Since we are discussing a series of time values, it will often be appropriate to look at these values as an ordered sample. While it should be obvious when the estimators referred to are suitable for survival data, to make the distinction even clearer, $(t_{(i)}, \delta_i)$ will denote the ordered $X_i$'s, with $t_{(i)}$ referring to ordered times and $\delta_i$ referring to the corresponding censoring indicator.

The quantile function when censoring is present is similar to $Q(p)$ above, but depends only on the distribution of the failure times.

$$Q^o(p) = F_o^{-1}(p) = \inf\{t : F_o(t) \geq p\}$$

## 2.2 Empirical Quantiles

### 2.2.1 Uncensored Case

The most common estimator of $F$ is the empirical distribution function, defined by

$$\widehat{F}(X_{(i)}) = \frac{i}{n} \qquad i = 1, \ldots, n$$

The $p^{th}$ quantile can then be estimated by

$$\widehat{Q}(p) = \widehat{F}^{-1}(p) = X_{([np])}$$

where $\widehat{F}$ is the empirical distribution function and [] denotes an integer value.

Another way of formulating this (see David(1970)) is that the probability integral transformation $v = P(x)$ takes the $r^{th}$ order statistic $X_{(r)}$ into the $r^{th}$ order statistic $V_{(r)}$ where $V_1, \ldots, V_n$ is a sample from a rectangular Un$(0, 1)$ distribution.

So, $V_{(r)} = P(X_{(r)})$ or equivalently $X_{(r)} = Q(V_{(r)})$

Expanding $Q(V_{(r)})$ in a Taylor series about $E(V_{(r)}) = \frac{r}{n+1} = p_r$ and $var(V_{(r)}) = \frac{r(n-r+1)}{(n+1)^2(n+2)} = \frac{p_r(1-p_r)}{n+2}$ gives rise to the following properties.

$$E(X_{(r)}) = Q(p_r) + \frac{p_r(1-p_r)}{2(n+2)} Q''(p_r) + O((n+2)^{-2})$$

$$var(X_{(r)}) = \frac{p_r(1-p_r)}{n+2} \left(Q'\right)^2 (p_r) + O((n+2)^{-1})$$

$$cov(X_{(r)}, X_{(s)}) = \frac{p_r(1-p_s)}{n+2} Q'(p_r) Q'(p_s) + O((n+2)^{-1})$$

Note that $Q'(p_r) = \dfrac{1}{f(Q(p_r))}$ where $f$ is the derivative of $F$.

Also, for $0 < p_1 < p_2 < \ldots < p_k < 1$, the joint distribution of $X_{([np_j])}$ $(j = 1, \ldots, k)$ tends to a $k$-dimensional normal distribution with

means $Q(p_1), \ldots, Q(p_k)$ and covariance

$$cov(X_{([np_j])}, X_{([np_l])}) = \frac{p_j(1-p_l)}{nf(Q(p_j))f(Q(p_l))} \qquad j \leq l$$

### 2.2.2 Censored Case

When censoring is present, it is usually more appropriate to discuss the *survivor function* than the distribution function, where

$$S(t) = 1 - F_o(t) = Pr\{\text{survive beyond time t}\}$$

Kaplan and Meier (1958) derived the product-limit (PL) esimator which is defined as follows :

$$\widehat{S}(t) = \begin{cases} 1 & 0 \leq t \leq t_{(1)} \\ \prod_{i=1}^{k-1}(\frac{n-i}{n-i+1})^{\delta_i} & t_{(k-1)} < t \leq t_{(k)} \quad k = 2, ..., n \\ 0 & t_{(n)} < t \end{cases}$$

While the empirical distribution function has 'jumps' of size $\frac{1}{n}$ at each data-point, the PL-estimator has 'jumps' which occur at failure times only, with the size of each being determined by the number of censored values between two consecutive ordered failure times.

Gill(1983) showed that, under very weak conditions on the censoring process, the PL estimator is asymptotically normally distributed. This will be expressed more explicitly in terms of the *PL-process*. Breslow and Crowley(1974) and Aalen(1976) proved that the PL-process

$$\beta_n(t) = \sqrt{n}(\widehat{F}_n(t) - F(t)) \qquad -\infty < t < \infty$$

converges weakly to a mean zero Gaussian process with covariance structure

$$\text{cov}(\beta(s), \beta(t)) = (1 - F_o(s))(1 - F_o(t)) \int_0^s [1 - F(u)]^{-2} d\widetilde{F}(u) \qquad s \leq t$$

where $\widetilde{F}(x) = \int_0^x (1 - G) dF_o$is the subdistribution function. Burke *et al.* (1981) established strong approximations of $\beta_n$ in terms of Gaussian processes.

The *PL-quantile* function is defined as

$$\widehat{Q}_n(p) = \inf\{t : 1 - \widehat{S}(t) \geq p\}$$

which leads to the PL-quantile process

$$\gamma_n(p) = \sqrt{n} f(Q(p))(Q(p) - \widehat{Q}_n(p))$$

where $f$ is the derivative of $F$.

The weak convergence of $\gamma_n$ was proved by Sander (1975). Doss and Gill (1992) derived a similar result by stating and proving a general theorem involving Bahadur representations which can also be applied to quantiles under the proportional hazards model (discussed later). Aly *et al.*(1985) derived strong approximation theorems for the PL-quantile process by a generalised Keifer process.

Cheng(1984) developed an almost sure representation for $\widehat{Q}_n(p)$ which showed that under a set of sufficient conditions, with probability 1,

$$\sup_{0 \leq p \leq p_o} \left| \widehat{Q}_n(p) - Q(p) - \frac{(p - \widehat{F}_n(t))}{F'(Q(p))} \right| = O(n^{-\frac{3}{4}} (\log n)^{\frac{3}{4}})$$

as $n \to \infty$, where $0 \leq p_o \leq 1$ , which led to the corollaries

1. With probability 1

$$\sup_{0 \leq p \leq p_o} |\widehat{Q}_n(p) - Q(p)| = O(n^{-\frac{1}{2}} (\log \log n)^{\frac{1}{2}}) \quad as\ n \to \infty$$

2.

$$\sup_{0 \leq p \leq p_o} |\gamma_n(p) - N(p)| = O(n^{-\frac{1}{3}} (\log n)^{\frac{3}{2}}) \quad as\ n \to \infty$$

where N is a sequence of identically distributed processes with zero mean and covariance function

$$\mathrm{cov}(N(p), N(q)) = (1-p)(1-q) \int_0^1 \frac{dt}{(1-t)^2 [1 - F(F_o^{-1}(t))]} \qquad p \leq q$$

Chung *et al.* (1990) were concerned with estimating confidence bands for the quantile function under the random censorship model. Previously bands had been calculated for the PL-process and then inverted for the quantile function (see e.g. Csörgő and Horváth (1986) ) but these were not asymptotically correct.

The authors of this paper used a more direct approach by proving that the PL-quantile process converges weakly to a Weiner process.

Brookmeyer and Crowley (1982) were also interested in providing a nonparametric aysmptotic confidence interval for median survival time. The approach applied here was to invert a generalisation of the sign test for censored data.

The sign test is also employed by Reid (1981). The facts that $\widehat{F}_n(p)$ is asymptotically normally distributed, and the sign test statistic for censored data is a linear function of $F_n(p)$, imply that the test statistic is also approximately normally distributed.

Slud *et al.* (1984) compare a number of methods proposing estimates for nonparametric confidence bands for median survival time based on right-censored data, including those by Brookmeyer and Crowley (1982) and Reid (1981) mentioned above.

## 2.3 Smooth Quantiles

### 2.3.1 Uncensored Case

The methods in the last section choose the value of the $p^{th}$ quantile to be the data point whose empirical distribution (or PL) estimate is smaller than but, closest to, $p$. The main drawback to sample quantiles is that they experience a substantial lack of efficiency, caused by the variability of individual order statistics.

Reiss (1980) proposed *quasiquantiles* which take the average of two order statistics a distance of $m$ points from the sample quantile of interest,

i.e. for $m \in \{1, 2, \ldots, \min\{(np) - 1, n - (np)\}\}$

$$\tilde{Q}_n(p) = \frac{(X_{([np]-m)} + X_{([np]+m)})}{2}$$

Other linear combinations involving a larger number of order statistics were also suggested. The efficiency of two estimators can be measured by calculating $\lim_{n\to\infty}$ of the ratio of their variances. The relative deficiency of the sample quantile with suitable quasiquantiles tends to infinity for increasing sample sizes.

Linear combinations of this type are simple forms of smoothing and suggests that averaging over the order statistic, by weighting observations according to their proximity to the sample quantile, may be an improvement over the raw sample quantile alone.

Stigler (1974) discussed an estimator of this form, namely

$$B_n = \frac{1}{n}\sum_{i=1}^{n} J\left(\frac{i}{n+1}\right) X_{(i)}$$

where $J$ is a suitably chosen weight function e.g. if $J(u) \equiv 1$ then $B_n$ is the sample mean. Asymptotic formulae for the mean and variance of $B_n$ were derived. Weak convergence to a Gaussian process was also established.

A common technique for smoothing is kernel estimation. When considering distribution functions, it is natural to think of the order statistics referring to the horizontal axis, and the empirical distribution function to the vertical axis. However, when the object of estimation is a quantile, this order can be reversed, with the point of estimation $p$ $(0 \leq p \leq 1)$ referring to the horizontal axis and the quantile scale expressed vertically. Standard nonparametric regression methods can then be applied. This is not a viewpoint which has been adopted in the literature, but it provides a helpful perspective on suggested estimators.

Parzen(1979) suggested an estimator of the form

$$\begin{aligned} T_n(p) &= \int_0^1 \hat{F}^{-1}(u)\frac{1}{h_n}K\left(\frac{u-p}{h_n}\right)du \\ &= \sum_{i=1}^{n} X_{(i)}\frac{1}{h_n}\int_{\frac{i-1}{n}}^{\frac{i}{n}} K\left(\frac{u-p}{h_n}\right)du \end{aligned}$$

where $K$ is a bounded probability density function which is zero outside a finite interval (-c,c) and is symmetric about zero, and $\{h_n\}$ is a "bandwidth" sequence of positive numbers such that $h_n \to 0$ *as* $n \to \infty$. This estimator can be viewed as arising from a nonparametric regression of $X_{(i)}$ on $\frac{i}{n}$ (using a form of smoothing proposed by Gasser and Müller (1979) ).

Yang (1985) approximated this estimator by

$$\hat{T}_n(p) = \frac{1}{n}\sum_{i=1}^{n} X_{(i)}\frac{1}{h}K\left(\frac{\frac{i}{n}-p}{h}\right)$$

Using standard Taylor series expansion arguments, it can be shown that

$$E(\widehat{T}_n) = Q(p) + \frac{1}{2}h^2 Q''(p) \int u^2 K(u)du + O(h^2)$$

$$var(\widehat{T}_n) = \frac{[Q'(p)]^2}{n} p(1-p) - \frac{[Q'(p)]^2}{n} 2h \left\{ \int tK(t)K^*(t)dt \right\} + O(\frac{h}{n})$$

This shows that to first order, the kernel quantile estimator has the same basic properties as the sample quantile estimator.

Yang(1985) also reported the following results :

1. The kernel-quantile process has the property that

$$\sqrt{n}(T_n(p) - Q(p)) = -\sqrt{n}\frac{[F_n(Q(p)) - p]}{f(Q(p))} + O_p(1)$$

2. For $0 < p_1 < p_2 < \ldots < p_k < 1$, the asymptotic distribution of

$$\sqrt{n}[T_n(p_1) - Q(p_1), \ldots, T_n(p_k) - Q(p_k)]$$

is k-dimensional normal with zero mean vector and covariance matrix with elements

$$\frac{p_i(1-p_j)}{f(Q(p_i))f(Q(p_j))} \quad i,j = 1,\ldots,k; i \leq j.$$

Falk(1985) also derived multidimensional normality for this estimator.

3. $\widehat{T}_n(p)$ and $T_n(p)$ are aysmptotically equivalent in mean square, and results 1 and 2 also hold for $\widehat{T}_n(p)$.

Falk (1984) derived the asymptotic mean squared error, and Sheather and Marron (1990) extended this to a higher term.

$$\begin{aligned} MSE(T_n(p)) &= \frac{1}{n}p(1-p)[Q'(p)]^2 - \frac{2h}{n}[Q'(p)]^2 \int_{-\infty}^{\infty} uK(u)K^*(u)du \\ &\quad + \frac{1}{4}h^4[Q''(p)]^2 \left[\int_{-\infty}^{\infty} u^2 K(u)du\right]^2 + O(n^{-1}h) + O(h^4) \end{aligned}$$

where $K^*(x) = \int_{-c}^{y} K(y)dy$

Now define $i(n) = \min\{j \in N : MSE(\widehat{Q}_j(p)) \leq MSE(T_n(p))\}$

Falk(1984) showed that

$$\lim_{n\in N}\left(\frac{i(n)-n}{nh}\right)=2\int\frac{xK(x)K^*(x)dx}{p(1-p)}$$

and so the quantity $\psi(k)=2\int xK(x)K^*(x)dx$ can be regarded as a measure of the asymptotic performance within the class of kernels. The sign determines whether the smooth quantile estimator is better than the sample quantile (i.e. if $\psi(k)<0$) or vice versa. Sheather and Marron (1990) commented that for large samples, the deficiency will be relatively small.

Note that to estimate the variance of $T_n(p)$, and therefore be able to construct confidence intervals, it is necessary to estimate $Q'(p)$( also written as $(F^{-1})'(p)$.) Falk (1986) proposed that since $T_n(p)\rightarrow F^{-1}(p)\ for\ n\in N$, in probability, then the derivative of $T_n$ with respect to $p$ would tend to the derivative of the quantile function in probability also. So the kernel estimator of $(F^{-1})'(p)$ is defined to be

$$T_n'(p)=\int_0^1 F_n^{-1}(x)\frac{1}{h^2}K^+\left(\frac{p-x}{h}\right)dx$$

where $K^+ : R\rightarrow R$ denotes a kernel function. Under fairly general conditions, this is shown to be asymptotically Normal.

Sheather and Marron (1990) derived an expression for an asymptotically optimal smoothing parameter obtained by minimising the asymptotic mean squared error of $T_n$.

For $p\neq 0.5$, this is

$$h_{opt}=\left[\frac{2\int_{-\infty}^{\infty}uK(u)K^*(u)du}{\{\int_{-\infty}^{\infty}u^2K(u)du\}^2}\right]^{\frac{1}{3}}\left[\frac{Q'(p)}{Q''(p)}\right]^{\frac{2}{3}}n^{-\frac{1}{3}}$$

For $p=0.5$ and $F$ symmetric, no single value of $h$ minimises $MSE(T_n(0.5))$. Here, any $h$ satisfying $h=constant*n^{-m}$ $(0<m\leq 0.5)$ will, for larger values of the constant, produce an estimator with smaller mean squared error than $\widehat{Q}(0.5)$.

### 2.3.2 Censored Data Case

Padgett (1986) presented a smooth nonparametric estimator of the quantile function for right-censored data based on kernel estimates.

Extending the sample quantile function to one which was appropriate for censored data required replacing terms involving the empirical distribution function with equivalent ones from the product-limit estimator. The same idea applies here and we adjust the smooth quantile function, $T_n$, accordingly to produce the estimator,

$$\begin{aligned} Q_n(p) &= \frac{1}{h}\int_0^1 \widehat{Q}_n(p)K\left(\frac{u-p}{h}\right)du \\ &= \frac{1}{h}\sum_{i=1}^{n} t_{(i)}\int_{J_{i-1}}^{J_i} K\left(\frac{u-p}{h}\right)du \end{aligned}$$

where $j_i = \widehat{S}(t_{(i)}) - \widehat{S}(t_{(i-1)})$ and $J_i = \sum_{k=1}^{n} j_k$ , the cumulative jump, as defined earlier.

This can be approximated by

$$Q_n^*(p) = \frac{1}{h}\sum_{i=1}^{n} t_{(i)} j_i K\left(\frac{J_i - p}{h}\right)$$

Due to censoring, results similar to those derived by Yang and by Falk are very difficult to obtain. However, under general conditions, Padgett (1986) proved that $Q_n(p)$ is strongly consistent.

Lio *et al.* (1986) showed that the process

$$\sqrt{n}[Q_n(p) - Q^o(p)]$$

converges in distribution to a Gaussian process with zero mean and variance

$$\sigma_p^2 = (1-p)^2 \int_0^{Q^o(p)} \frac{[1-F(u)]^{-2} dF_o^*(u)}{f_o^2(Q^o(p))}$$

where $F_o^*$ is the sub-distribution function of uncensored observations. Lio *et al.* (1986) established an asymptotic uniform mean-squared equivalence of $Q_n$ and $Q_n^*$.

Lio and Padgett(1987) derived an expression for the mean squared error of $Q_n$. However, Ghorai and Rejtő (1990) obtained a slightly different result for this

quantity and were subsequently able to obtain information about the relative deficiency of $Q_n$. The latter authors derived

$$nMSE(Q_n(p)) = \left(\frac{1-p}{f(Q(p))}\right)^2 \int_0^{Q(p)} \frac{(1-G(u))dF_o}{(1-F(u))^2} + O(n^{-\frac{1}{4}} \log n)$$

leading to

$$\left| nMSE(Q_n(p)) - \left[ \left(\frac{1-p}{f(Q(p))}\right)^2 \int_0^t \frac{1-G}{(1-F)^2} dF_o - \frac{2h \int uK(u)K^*(u)du}{f^2(Q(p))(1-G(Q(p)))} \right] \right|$$
$$\leq O(nh^{2m+2}) + O(h^2) + O(n^{\frac{1}{4}} h^{m+1} \log n) + o(n^{-\frac{1}{4}} \log n)$$

Recalling the expression for the mean squared error of the PL-quantile function, the relative deficiency can be assessed, as for the uncensored case, by

$$i(n) - n \quad \text{where} \quad i(n) = \min\{j : MSE(\hat{Q}_j(p)) \leq MSE(Q_n(p))\}$$

and

$$\lim_n \left(\frac{i(n)-n}{nh}\right) = \frac{2 \int xK(x)K^*(x)dx}{f^2(Q(p))(1-G(Q(p)))\sigma^{*2}(p)}$$

where

$$\sigma^{*2} = \left(\frac{1-p}{f(Q(p))}\right)^2 \int_0^{Q(p)} \frac{(1-G(u))dF_o}{(1-F(u))^2}$$

So, if the kernel is such that $\int xK(x)K^*(x)dx > 0$, then $Q_n$ performs better than $\hat{Q}$.

To find the "optimal" bandwidth, Padgett and Thombs(1986) suggested a bootstrapping approach. In order to create a single bootstrap sample, the collection of observations $(t_{(i)}, \delta_i)$ (where $i = 1, \ldots, n$) is sampled with replacement $n$ times. A large number of these samples are drawn and the smooth kernel quantile estimator is calculated for each. The mean squared error of these values can be estimated from standard formulae. From a range of values of $h$, the "optimal" value is chosen to be that which minimises the mean squared error. Bootstrapping will be discussed in more detail in a later chapter.

## 2.4 Incorporating a Regression Effect

### 2.4.1 Quantile Curves

Quantile regression curves provide a helpful graphical representation of how a covariate affects a response of interest. These curves have been studied in a variety of ways. One application is the construction of centile reference charts, which are used in medicine to compare the measurements of individuals with those estimated for the population. Typically, the variables on these charts relate to growth, although there is a much wider area of applications. There are many ways of tackling this problem and what follows is intended as a review of some of the key nonparametric approaches.

Cleveland (1979) proposed fitting a smooth line to a scatterplot using locally weighted regression. Healy *et al.* (1988) simplified and extended this idea to estimating smooth centile curves, making use of polynomial regression and the empirical quantile function.

The LMS method, introduced by Cole (1988) finds appropriate Box–Cox transformations of the explanatory variable across the range of the covariate and summarises the characteristics of the data by 3 curves representing the median, coefficient of variation and skewness. From these, centile curves can easily be obtained. While this method is essentially a parametric approach, the extension proposed by Cole and Green(1992) introduces a nonparametric aspect to the problem. The original method required that the covariate should be grouped but the improvement made by Cole and Green uses maximum penalized likelihood to estimate the curves more directly using natural cubic splines.

Jones, in the discussion of the paper by Cole (1988), proposed that minimising the sum of the absolute deviances and adding on a roughness penalty could produce a spline smoothing regression quantile.

Kernel estimation of centile curves was also discussed by Jones and Hall(1990) and Rossiter(1991). Jones and Hall (1990) regarded this as a nonparametric re-

gression problem. Suppose we have data $(x_i, y_i)$ where the $y_i$'s are assumed to be realisations from the conditional distribution of $Y|X = x$ with associated distribution function $F_x(y)$. The regression quantile $q_\alpha(x)$ is defined by $F_x(q_\alpha(x)) = \alpha$ for $\alpha \in (0, 1)$. Stone (1977) proposed a weighted empirical distribution function, $\widehat{F}_x$, where the weights determine the jump at $y_i$ by the distance between $x$ and $x_i$ (instead of each being equal to $\frac{1}{n}$). Jones and Hall (1990) used kernel weights to produce $\widehat{F}_x$ and solve $\widehat{F}_x(\widehat{q}_\alpha(x)) = \alpha$. This can also be formulated as

$\widehat{q}_\alpha(x)$ is the solution $\theta$ to the equation

$$H_\alpha(\theta) = \sum_{i=1}^{n} W_i(x)\Psi_\alpha(y_i - \theta) = 0$$

where

$$\Psi_\alpha(z) = \alpha I_{(0,\infty)}(z) - (1-\alpha)I_{(-\infty,0)}(z)$$

and

$$W_i(x) = \frac{1}{h_\alpha}\int_{\frac{i-1}{n}}^{\frac{i}{n}} K\left(\frac{x-u}{h_\alpha}\right) du$$

where $I_A(z)$ is the indicator function, $K$ is a kernel function and $h_\alpha$ is the smoothing parameter (as defined earlier).

The mean squared error of this estimator was proved to be the following

$$MSE(\widehat{q}_\alpha(x)) \simeq \frac{h_\alpha^4}{4}\left[\int x^2 K(x)dx\right]^2 \left\{\frac{F_x''(q_\alpha(x))}{f_x(q_\alpha(x))}\right\}^2$$

From this, the asymptotically optimal value of the smoothing parameter is given by

$$h_\alpha^* = \left[\frac{\int K^2(x)dx}{\int x^2 K(x)dx}\frac{\alpha(1-\alpha)}{[F_x''(q(x))]^2}\right]^{\frac{1}{5}} n^{-\frac{1}{5}}$$

Jones and Hall (1990) also derived results of this sort for the more general case of the random design, which replaces the weight function $W_i$, with

$$W_i^*(x) = \frac{1}{nh_\alpha}K\left(\frac{x-x_i}{h_\alpha}\right)$$

Rossiter (1991) used the nonparametric smoothing approach to density estimation, using multidimensional kernels to condition on one or more covariates. The percentiles were calculated by solving the integral of the estimates obtained for the density for $p$ using a Newton-Raphson procedure.

### 2.4.2 Quantile Curves for Censored Data

A report by Doksum and Yandell (1982) explored ways of graphically describing the effect of a covariate on survival time by quantile curves. Semiparametric and parametric methods of achieving this can be based on the Proportional Hazards and Accelerated Failure Time models respectively. The nonparametric approaches discussed in this article are Beran's running product limit median and a running product limit mean.

Consider the data as a set of triples $(t_{(i)}, \delta_i, z_i)$ where $z_i$ is the covariate value for the $i^{th}$ individual ordered by time. Under the random censorship model, it is assumed $X^o$ and $U$ are conditionally independent given $Z$.

Beran(1981) proposed a running product-limit estimator (BRPLE) which is conditional on a covariate. The idea is to choose a set of covariate grid values and, for the time values corresponding to the $k^{th}$ nearest neighbour covariate values, calculate the product-limit estimates. This can be written formally as

$$\widehat{S}(t|z) = \prod_{\{j|t_{(j)}<t\}} [1 - \widehat{\lambda}_{j,k}(z)]$$

where

$$\widehat{\lambda}_{j,k}(z) = \frac{\#[t_{(i)} = t_j, i \in I_k(z)]}{\#[Y_{(i)} \geq t_j, i \in I_k(z)]}$$

and $I_k(z)$ are the indices of the $k^{th}$ nearest neightbours of z.

The conditional median survival can then be defined to be

$$m(z) = \frac{1}{2}[m^+(z) + m^-(z)]$$

$$\text{where} \quad m^+(z) = \inf\{t : S(t|z) < \tfrac{1}{2}\}$$

$$m^-(z) = \sup\{t : S(t|z) > \tfrac{1}{2}\}$$

Plotting $m(z)$ against $z$ provides a nonparametric median regression curve for random right censored data. Doksum and Yandell (1982) established consistency and derived asymptotic distributions and confidence bands for this curve.

Recall that the product–limit estimator for survival times can be written as

$$\widehat{S}(t) = \prod_{i:t_i<t} \left(1 - \frac{1}{n-i+1}\right)$$

Gentleman and Crowley (1991) extended this to the $p^{th}$ quantile and proposed that kernel weights be used instead of $k^{th}$ nearest neighbours to estimate the Kaplan-Meier. This can be written as

$$\widehat{S}_h(t|z) = \prod_{i:t_i<t} \left[1 - \frac{W_i(z)}{\sum_{j\in R_i} W_j(z)}\right]^{\delta_i}$$

where $W_i(z)$ is the weight attached to the $i^{th}$ individual according to its covariate value's proximity to the point of interest and $R_i$ is the risk set (i.e. individuals who have not yet failed or been censored) at time $t_{(i)}$.

$\widehat{S}_h(t|z)$ can also be written as

$$\widehat{S}_h(t|z) = \prod_{i:t_i<t} \left[\frac{\sum_{j=1}^n W_j(z)I(t_j > t) - W_i(z)}{\sum_{j=1}^n W_j(z)I(t_j > t)}\right]^{\delta_i}$$

which in the uncensored case becomes

$$\widehat{S}_h(t|z) = \frac{\sum_{j=1}^n W_j Y_j}{\sum_{j=1}^n W_j} \qquad \text{where } Y_j \sim Bi(1, S(t|z))$$

Using Taylor expansions, it can be shown that

$$E(\widehat{S}_h(t|z)) = S(t|z) + \frac{h^2}{2} S''(t|z) \int u^2 K(u) du$$

$$var(\widehat{S}_h(t|z)) = S(t|z)[1 - S(t|z)] + O(h)$$

Quantiles calculated over a range of covariate values are not necessarily smooth. Gentleman and Crowley (1991) smooth the resulting curve using a kernel smoother.

Doksum and Yandell (1982) defined the running product mean to be

$$\mu_n(z) = \int_0^\infty S_n(t|z) dt$$

However, with right-censored data, the right tail of the distribution function cannot be estimated with much precision and so estimates of the mean will not be accurate.

## 2.5 Proportional Hazards Quantiles

### 2.5.1 The Model

A median curve under Cox's(1972) Proportional Hazards (PH) model was mentioned very briefly as a semi-parametric method in the last section. While all the techniques for quantile estimation mentioned so far have been nonparametric, the main aim of this research is to provide quantile curves for censored data and the proportional hazards model is a widely used tool in survival analysis. This, and the fact that we will use it later in a formal test of covariate effect, makes it appropriate to be discussed in the present context of quantiles.

The proportional hazards model of Cox (1972) specifies that the hazard rate for an individual with covariate $z$ is

$$\lambda(t|z) = \lambda(t)\exp(\beta_o\, z)$$

where $\beta_o$ is an unknown coefficient and $\lambda$ is the underlying baseline hazard rate. The survivor function under this model is written as

$$S_\beta(t|z) = S_o(t)^{\exp(\beta_o z)}$$

where $S_o(t)$ is a baseline survivor function. An estimate for $\beta$ is usually found by maximising the partial likelihood to give $\widehat{\beta}$ and the baseline survivor function can be estimated by

$$\widehat{S}_o(t_{(i)}) = \prod_{j=0}^{i-1}\left[1 - \frac{\delta_i \exp(\widehat{\beta} z_i)}{\sum_{j\in R_i}\exp(\widehat{\beta} z_j)}\right]^{\exp(-\widehat{\beta} z_i)}$$

Tibshirani and Hastie (1987) proposed a more general form of the PH model, namely

$$\lambda(t|z) = \lambda(t)\exp(\ s(z)\ )$$

where $s()$ is simply a smooth curve of unrestricted shape. This will be referred to as the *smooth* proportional hazards model.

One way to estimate the $s(z)$ values, as proposed in the aforementioned paper, is by application of the local likelihood technique. This uses the $k^{th}$ nearest neigh-

bours of a covariate value of interest to estimate local partial likelihoods. These are then maximised to provide the estimates of $s(z)$. This can be formulated as

$$PL_i = \prod_{l \in D \cap N_i} \frac{exp(z_l \beta_i)}{\sum_{j \in R_i \cap N_i} exp(z_j \beta_i)}$$

where $D$ is the set of indexes for the uncensored $t$'s, $R_i$ is the risk set and $N_i$ represents those in the $k^{th}$ nearest neighbours set. The smooth terms are then estimated, using the trapezoidal rule, by

$$\widehat{s}(z_i) = \sum_{j=1}^{i} (z_j - z_{j-1}) \frac{(\widehat{\beta}_j + \widehat{\beta}_{j-1})}{2}$$

To estimate the baseline survivor function, the $\widehat{\beta} z_i$ terms in the formula above for $\widehat{S}_o(t_{(i)})$ in Cox's PH model are simply replaced by the $\widehat{s}(z_i)$ values. This provides a model whose description of the covariate is less restricted than that of the monotonic behaviour in Cox's PH model.

### 2.5.2 Quantile Curves under the PH model

Literature on quantiles under the PH model has focussed on the Cox model. Let $\xi_p(z)$ be the $p^{th}$ quantile of the distribution of the life length of an individual with covariate $z$.

Doss and Gill (1992) showed that

$$\sqrt{n}(\widehat{S}_\beta(\cdot) - S(\cdot), \widehat{\beta} - \beta) \xrightarrow{d} (V(\cdot), W)$$

for some continuous mean zero Gaussian process, V, over a suitable interval and W, a normally distributed random variable with mean zero, where $\widehat{S}_\beta$ is the estimated survivor function under the PH model.

Now the $p^{th}$ quantile survival time for a given covariate value $z^*$ can be found by solving

$$[S_o(t)]^{\exp(\beta z^*)} = 1 - p \qquad \text{for } t$$

which can also be written as

$$\widehat{\xi}_p(z) = \widehat{S}^{-1}((1-p)^{\exp(-\widehat{\beta} z^*)})$$

Doss and Gill (1992) proved the weak convergence of the process

$$\zeta_p(z) = \sqrt{n}(\widehat{\xi}_p(z) - \xi_p(z))$$

Dabrowska and Doksum (1987) also proposed similar estimates of this type, expressed in terms of the integrated hazard, which were discussed by Burr and Doss (1993) along with the above estimator. Although these were shown to have the same first order asymptotics, $\xi_p$ was shown to have "sounder theoretical basis".

Burr and Doss (1993) proposed two approaches to estimating confidence bands for percentile survival time under the PH model. The first method produced what they called "simulated process bands". Noting that the process, $\zeta_p$, converges to a distribution, $V_p$, which is a Gaussian process defined on $K$, the distribution of $\sup |\widehat{V}_p(z)|$ can be simulated and used to find the required critical constants. The second method used the idea of bootstrapping, which will be discussed in more detail in the next chapter.

Choosing a range of values of $z^*$ and plotting the estimated PH quantile will produce a curve which describes the effect of the covariate on survival time under the PH model. Although the mathematical properties of a similar quantile estimation method for the *smooth* PH model case are more difficult to explore, an analogous procedure is obvious and may prove to be a useful comparison to those curves from Cox's model. Because of the structure of the model, PH quantile curves have monotonic behaviour, although for curves under the *smooth* PH model, and for the sample and kernel quantile curves, this need not be the case.

### 2.5.3 Plots of PH quantile curves

To conclude this chapter, the expected shape of the curves under Cox's Proportional Hazards model are discussed.

Consider 3 sets of data, each simulated from a PH model but with values of $\beta$ of -2, 0, 2. Scatter plots of these data are shown in Figures 2.1–3, with

failure times denoted by crosses and censored times by circles. These cases represent increasing hazard, no covariate effect and decreasing hazard respectively, as covariate values increase.

Figure 2.4 depicts the three estimated baseline survivor functions for these data. The curve with the long right hand tail is that with $\beta = -2$. The next plot focusses on the 0–1.5 survival time range and there appears to be very little difference in any of the curves.

The quantile curves are calculated by finding the survival time at which the survivor function for particular covariates is equal to a chosen percentile. The slope of these curves is determined by the estimated value of the regression coefficient, $\beta$. The difference of effect of these values is clearly seen in the last three pictures. Here, in Figures 2.6–8, the median curves are plotted at 30 values across the range of the covariate and superimposed on the scatter plot. It can be seen that, due to the exponential power on the covariate, if the curves were transformed to a logarithmic scale, the quantile curves will be approximately a straight line.

With visual knowledge of what to expect under a PH model, a nonparametric approach is now proposed which has no underlying assumptions about the model and therefore is a much more data–driven technique for obtaining quantile curves.

# Data simulated under PH model

Figure 2.1 : beta=2

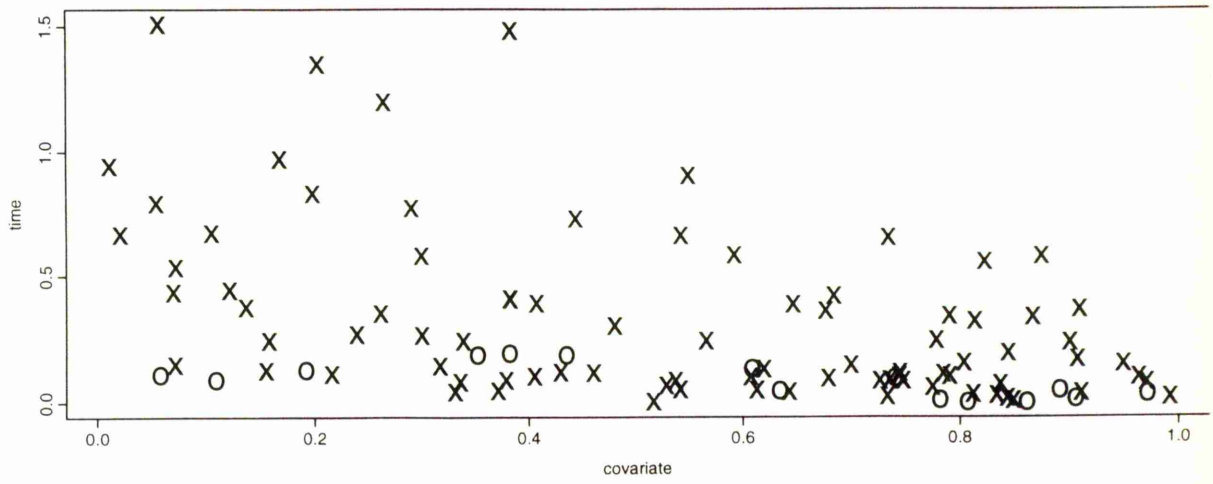



Figure 2.2 : beta=0

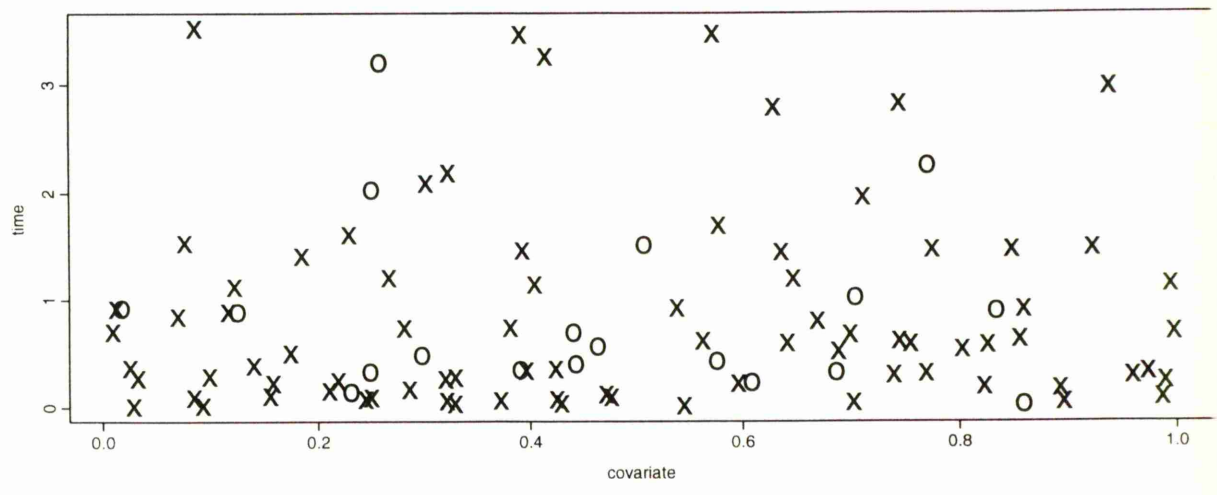
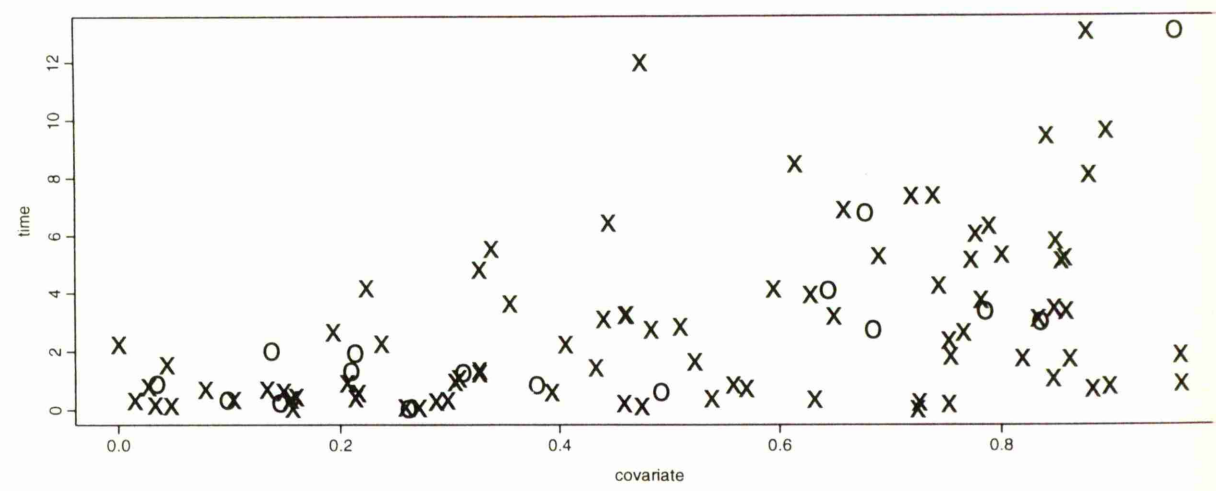



Figure 2.3 : beta=-2



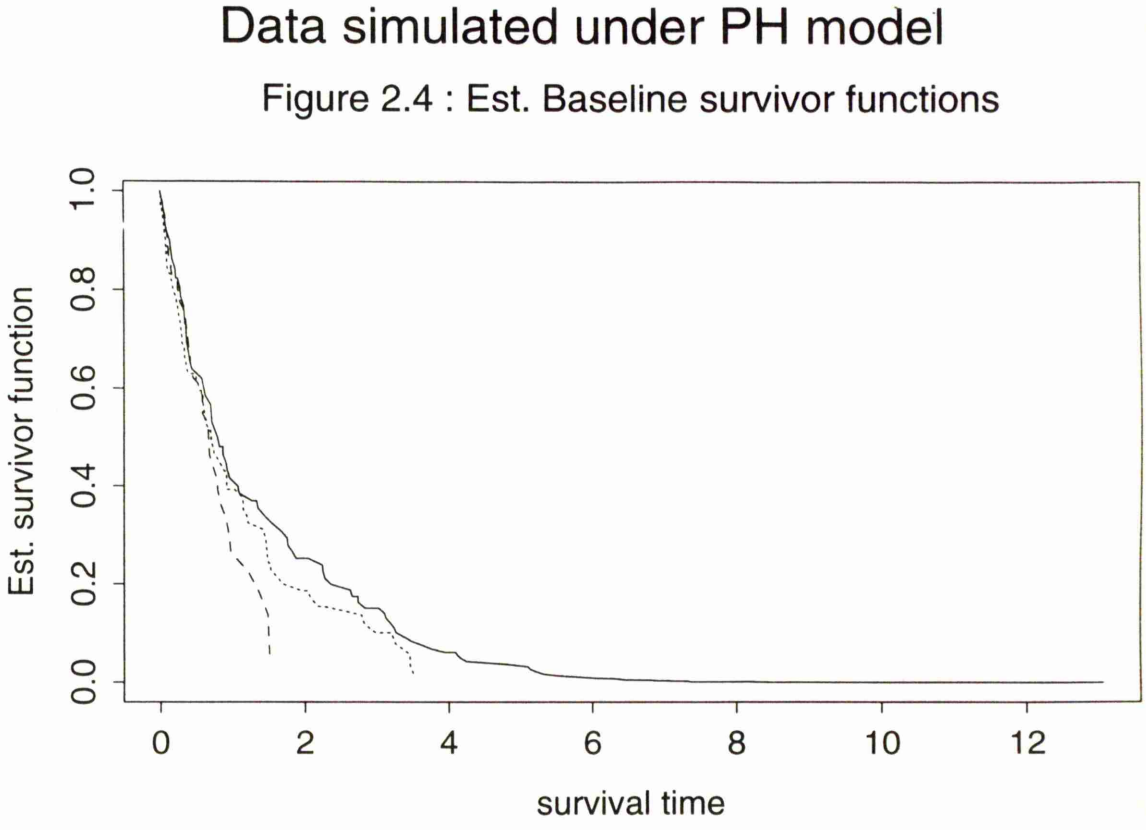



Figure 2.4 : Est. Baseline survivor functions

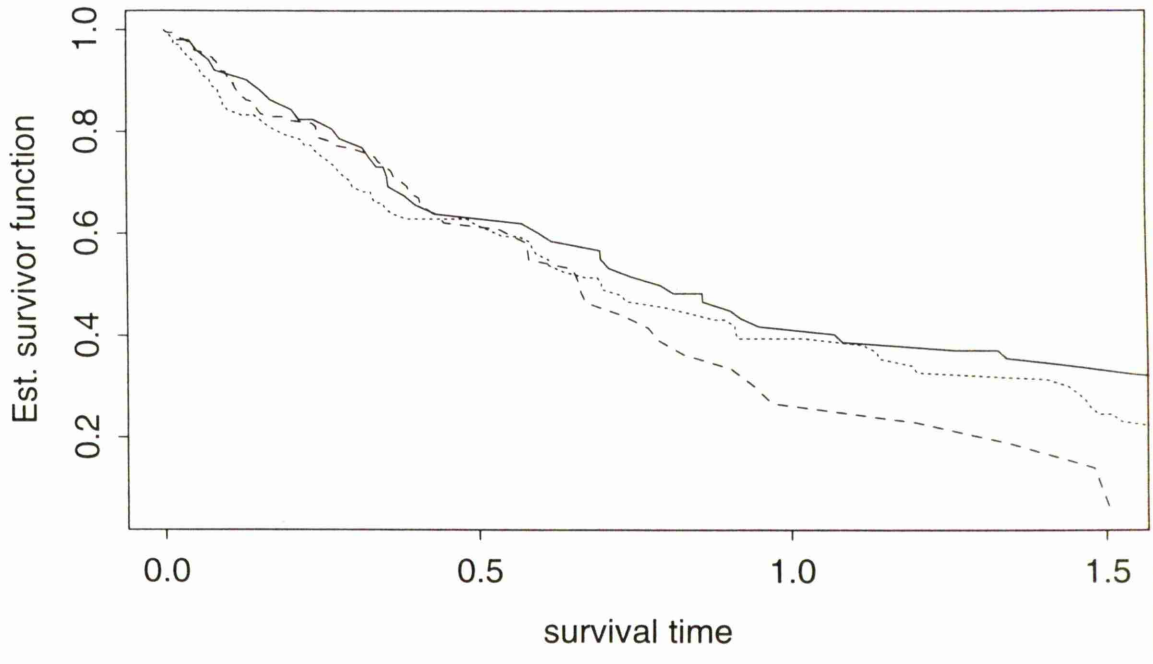



Figure 2.5 : Est. Baseline survivor functions

# Data simulated under PH model

Figure 2.6 : beta=2

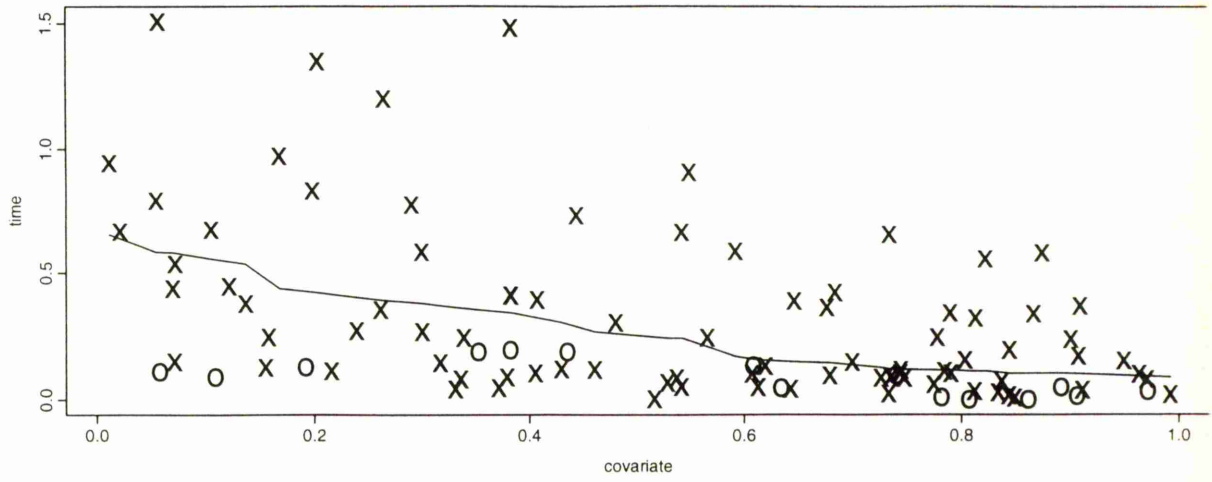



Figure 2.7 : beta=0

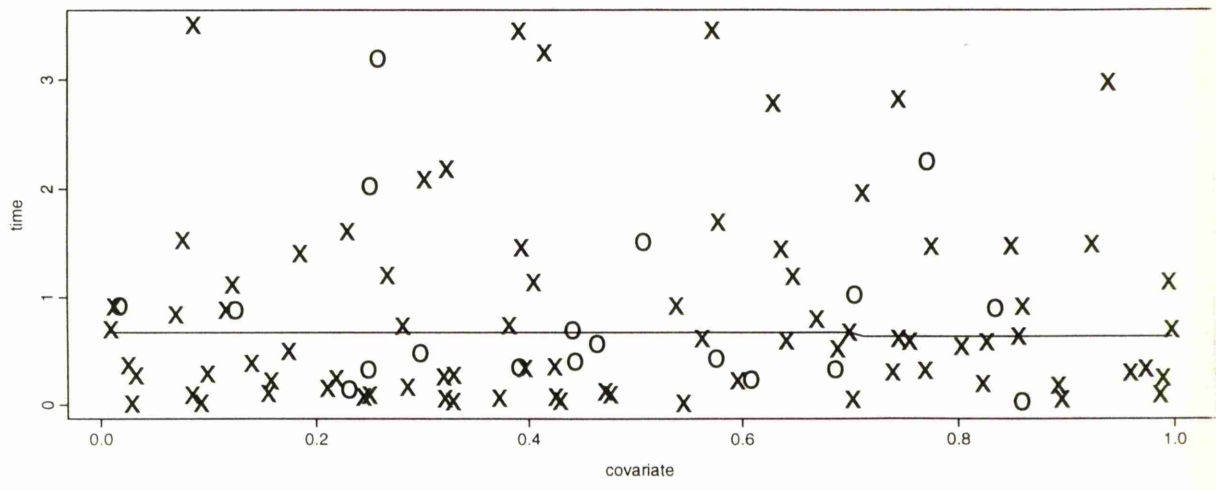
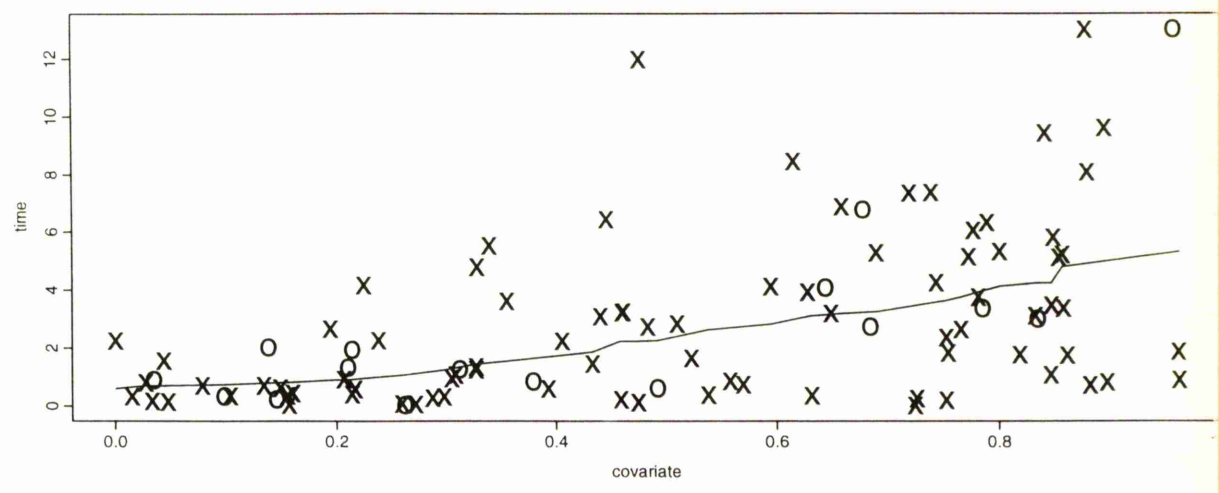



Figure 2.8 : beta=-2



# Chapter 3

# The Doubly–Smooth Quantile Estimator

## 3.1 Introduction

The principal aim of this research is to provide a method for obtaining nonparametric quantile curves for the analysis of survival time in the presence of a continuous covariate. By combining different ideas, described in previous sections, a *doubly–smooth quantile estimator for right–censored data* is proposed. This estimator smooths, using kernel estimates, in two directions – in the quantile space and across the covariate. It is the combination of these procedures which produces a curve which is smooth. Each smoothing component is based on nonparametric regression, applying appropriate weightings of the covariate and order statistic. This simplifies the computation involved for the quantile curves, compared to other existing proposals.

The first step is to construct a local estimate of the distribution of survival times. The data are weighted by the proximity of each individual's covariate values to the point of estimation. The *weighted Kaplan–Meier*, discussed in section 4.2 of the previous chapter, produces a step function, whose size of "jump" is partly determined by the amount of censoring, and which is conditional on the covariate.

The second step comprises of a *smooth quantile function*, similar to that in

section 2.3.2 proposed by Padgett (1986), which provides a direct formula for a fixed quantile (as opposed to solving the survivor function inversely in the manner of more traditional quantile functions). However, in the case of this "doubly–smooth" estimator, information about the covariate is included by using the weighted Kaplan–Meier estimates.

Calculating this estimate at various covariate values produces a curve whose shape will describe the effect of that covariate on survival time. The roughness of the curve will be determined by the smoothing parameters.

Using similar notation to that used before, this can be formulated as follows: Again, denote a sample of right–censored data by the set of triples,

$$\{(t_{(1)}, \delta_1, z_1), (t_{(2)}, \delta_2, z_2), \ldots, (t_{(n)}, \delta_n, z_n)\}$$

where $t_{(i)}$ is the $i^{th}$ ordered survival time
$\delta_i$ denotes the corresponding indicator (1=failure, 0=censored)
$z_i$ is the covariate value for the ith individual
for $i = 1, \ldots, n$

Let $z$ be a point of estimation, then we can represent the data as

$$\left\{\left(t_{(1)}, \delta_1, \frac{1}{h_1} W_1\left(\frac{z - z_1}{h_1}\right)\right), \left(t_{(2)}, \delta_2, \frac{1}{h_1} W_2\left(\frac{z - z_2}{h_1}\right)\right), \ldots, \left(t_{(n)}, \delta_n, \frac{1}{h_1} W_n\left(\frac{z - z_n}{h_1}\right)\right)\right\}$$

where $W$ is a bounded probability density function which is zero outside a finite interval $(-c, c)$ and is symmetric about zero and $h_1$ is a bandwidth sequence of positive numbers such that $h_1 \to 0$ *as* $n \to \infty$.

The weighted Kaplan–Meier is defined to be

$$\hat{S}_{h_1}(t_{(i)}|z) = \prod_{k=1}^{i} \left(\frac{\sum_{j=k+1}^{n} W_j}{\sum_{j=k}^{n} W_j}\right)^{\delta_k}$$

For a set of data containing only failure times, $\delta_i = 1$ for all $i = 1, \ldots n$, and so

$$\hat{\mathrm{S}}_{\mathrm{h}_1}(\mathrm{t}_{(\mathrm{i})}|\mathrm{z}) = \left(\frac{\sum_{j=2}^{n} W_j}{\sum_{j=1}^{n} W_j}\right)\left(\frac{\sum_{j=3}^{n} W_j}{\sum_{j=2}^{n} W_j}\right) \cdots \left(\frac{\sum_{j=i+1}^{n} W_j}{\sum_{j=i}^{n} W_j}\right)$$

$$= \left(\frac{\sum_{j=1}^{n} W_j I(t_j > t_{(i)})}{\sum_{j=1}^{n} W_j}\right)$$

where $I(t_j > t_{(i)}) = 1$ if $t_j > t_{(i)}$ and 0 otherwise. Denote the "jump" at each time point by

$$j_i^* = \widehat{S}_{h_1}(t_{(i)}|z) - \widehat{S}_{h_1}(t_{(i+1)}|z) = \frac{W_i}{\sum_{k=1}^{n} W_k}$$

and the cumulative jump by

$$J_i^* = \sum_{k=1}^{i} j_k^* = 1 - \widehat{S}_{h_1}(t_{(i)}|z)$$

As mentioned in Section 2.4.2, in the case of no censoring, the expectation and variance for the weighted Kaplan–Meier estimator can be derived by standard Taylor series expansions as shown in Appendix A. The covariate here is assumed to be uniformly distributed over the design space.

$$E(\widehat{S}_{h_1}(t|z)) = S(t|z) + \frac{h_1^2}{2} S''(t|z) \int u^2 K(u) du + O(h_1^2)$$

$$var(\widehat{S}_{h_1}(t|z)) = S(t|z)[1 - S(t|z)] + O(h_1)$$

These expressions show both the characteristics of the corresponding properties of the empirical distribution function and standard nonparametric regression estimators. To the first order, these expressions are the same as those for the empirical quantile function. The effect of smoothing is a second order one.

The idea of the doubly–smooth quantile estimator is to take an average of the ordered failure times, weighted by the jumps in $\widehat{S}_{h_1}(t|z)$. This can be written as

$$\widehat{Q}_{h_2}(p) = \frac{1}{h_2} \sum_{i=1}^{n} j_i^* t_{(i)} K(\frac{J_i^* - p}{h_2})$$

where $K$ and $h_2$ are defined similarly to $W$ and $h_1$.

This is a form of smoothing which is appropriate when the "design points" (here the $J_i^*$) are equidistant. A more general form of smoothing ensures the weights sum to one, as suggested for the standard nonparametric case by Nadaraya (1964) and Watson (1964), is

$$\widehat{Q}_h^*(p) = \frac{\sum_{i=1}^{n} j_i^* t_{(i)} K(\frac{J_i^* - p}{h_2})}{\sum_{i=1}^{n} j_i^* K(\frac{J_i^* - p}{h_2})}$$

This form, $\widehat{Q}_h^*$, has better behaviour because it is generally monotonic, unlike $\widehat{Q}_h$. The reason for this is that when there is heavy censoring of large time values the corresponding weights for estimation of large quantiles will be very small. However, $\widehat{Q}_h^*$ will ensure that the sum of the weights at each value of the quantile is 1. So instead of plummetting to zero for p large, the quantile time value will be the same, or slightly higher, than that of the last failure time.

Problems still exist when $p$ is a long distance from the nearest failure time, but this is a situation where the estimation of the quantile would be unwise under any circumstances.

The basic properties of the quantile estimator

$$\widehat{T}_n(p) = \frac{1}{n}\sum_{i=1}^{n} X_{(i)} \frac{1}{h} K\left(\frac{\frac{i}{n} - p}{h}\right)$$

can also be found by standard Taylor series arguments and these are given by

$$E(\widehat{T}_n) = Q(p) + \frac{1}{2}h^2 Q''(p) \int u^2 K(u) du + O(h^2)$$

$$var(\widehat{T}_n) = \frac{[Q'(p)]^2}{n} p(1-p) - \frac{[Q'(p)]^2}{n} 2h \left\{ \int tK(t)K^*(t)dt \right\} + O(\frac{h}{n})$$

as stated in Section 2.3.1 and derived in Appendix A. Note the similarity between the expectations for $\widehat{S}_{h_1}$ and $\widehat{T}_n$. Also, the variance expressions are the same as their empirical estimator equivalents to the first order. When these estimators are jointly used to produce the doubly–smooth effect for quantile curves, $h_2$ has a second–order effect. One way of viewing this issue is to consider that an optimal smoothing parameter for a survivor function is $n^{-\frac{1}{3}}$ (Azzalini, 1981)whereas the optimal rate for nonparametric regression is $n^{-\frac{1}{5}}$. The slower rate of $n^{-\frac{1}{5}}$ will therefore be the dominant effect.

The complex mathematics required to derive formal properties for the doubly–smooth estimator have not been pursued. However, the expressions derived for the properties of the separate smoothing components indicate the behaviour of the processes taking place. In particular, it is worth noting that the crucial step is that of calculating $\widehat{S}_{h_1}$.

## 3.2 Graphical Interpretation of Quantile Curves

All sets of survival data have the common feature that there is interest in observing the length of time until the occurrence of a particular event and that, for various reasons, it may not be experienced by all the individuals during the study. However, beyond this, there are many characteristics which make data-sets different. Several examples will now be used to present some of the possible covariate effects and also to help explain the behaviour of the doubly–smooth estimator. The percentile curves from this estimator will be compared to those calculated under the smooth proportional hazards (PH) model and Cox's PH model. Note the value of $k$ specified for the smooth PH model refers to the proportion of data used to estimate each $s(z)$ value. Finally, in an uncensored example, a comparison will be made with quantiles estimated from cubic splines.

### 3.2.1 Example 1 – Simulated data with no effect

The simplest model of interest is when the survival time is unaffected by a measured covariate. One way to simulate data of this form is to generate survival times and covariate values under independent distributions, then assign a censoring indicator using a binomial distribution. While the choice of parameters for the uniform distributions is relatively unimportant, the probability parameter of the binomial distribution will determine the proportion of censoring in the sample and it will be useful to look at different values.

Note that data of this form are not simulated under the usual random censorship model. Randomly censored data could be generated by, for example, choosing a failure and a censored time for each individual under different uniform distributions and applying the definition in section 2.1.2. The data from the approach used here will result in more of the censored times being smaller since each survival time has equal probability of being classed as censored.

A scatter plot of data representing a sample generated such that $t \sim Un(0.5, 10)$, $z \sim Un(0, 6)$ and $\delta \sim Bi(100, 0.8)$ (chosen to achieve approximately 20% cen-

soring) is shown in Figure 3.1.1. The plot below this, Figure 3.1.2, represents the median doubly–smooth (solid line), proportional hazards (broken line) and smooth PH (dotted line) percentile curves. The smoothing parameters used are $h_1$=5 and $h_2$=0.075 and $k = \frac{1}{4}$. As we would expect, there is little evidence of a covariate effect although the doubly–smooth and smooth PH curves are less straight and appear to be more sensitive to chance patterns in the data. Figures 3.1.3 and 3.1.4 show the $25^{th}$ and $75^{th}$ percentile curves respectively for each of the three methods and convey similar conclusions.

The following plots attempt to explain graphically how the combination of two nonparametric regression procedures produces a doubly–smooth quantile curve. Figures 3.1.5 and 3.1.6 show the Kaplan–Meier curve, calculated using only the survival times and censoring indicator, as a solid line, with the broken lines representing the weighted KM curves for the values 2 and 4 of the covariate respectively. The dotted lines in these plots are the result of the second smoothing procedure, which focusses on the relationship between the survival times and the quantiles (equivalent to 1- the weighted KM estimates). So, when the covariate is equal to 2, Figure 3.1.5 indicates that at, say p=0.25, or equivalently, the estimated survivor function equal to 0.75, the percentile survival time is approximately 5. In other examples, it may be interesting to compare these curves for various covariate values, to enable a closer look at the information summarised in plots such as Figure 3.1.2.

The behaviour of the doubly–smooth estimator under different proportions of censoring is an important issue as most estimators similar to this (see e.g. Padgett (1986) ) are known to perform badly in the presence of a large amount of censoring. It will be discussed later in an inference section but, as an introduction to the problem, the above analysis was repeated with approximately 50% censoring.

Figure 3.1.7 shows the data with the doubly–smooth (solid line) and PH median (broken line) curves. The most notable difference is that the curves are higher than those derived when only 20% of the data was censored. The data

Simulated data with no effect : 20% censoring

Figure 3.1.1 : Scatter plot

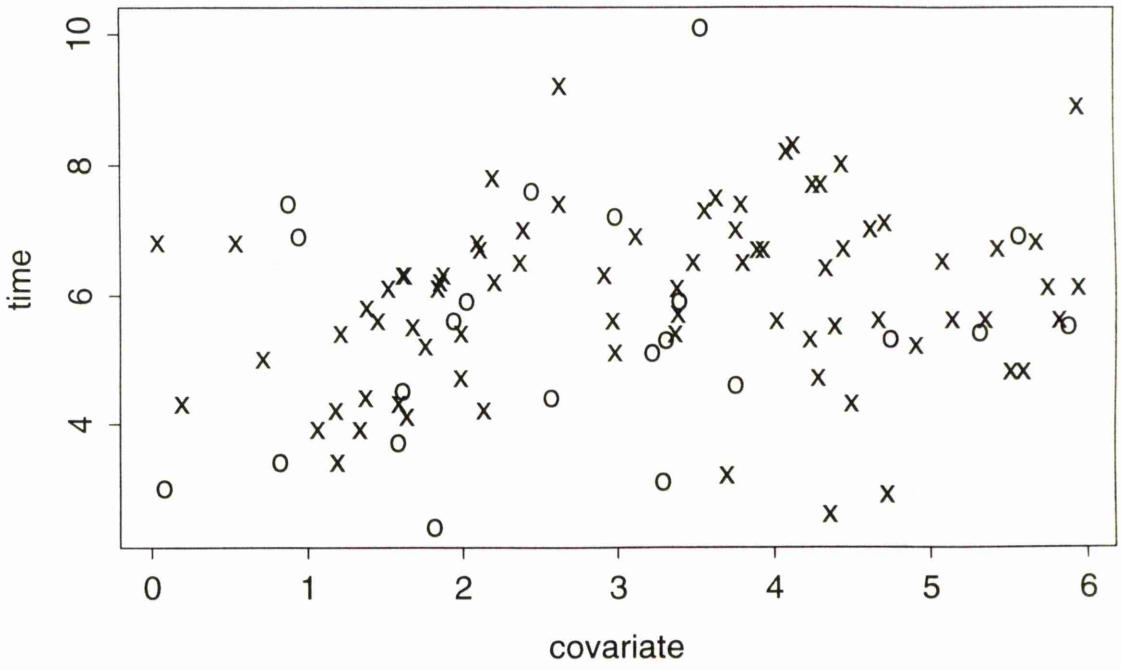
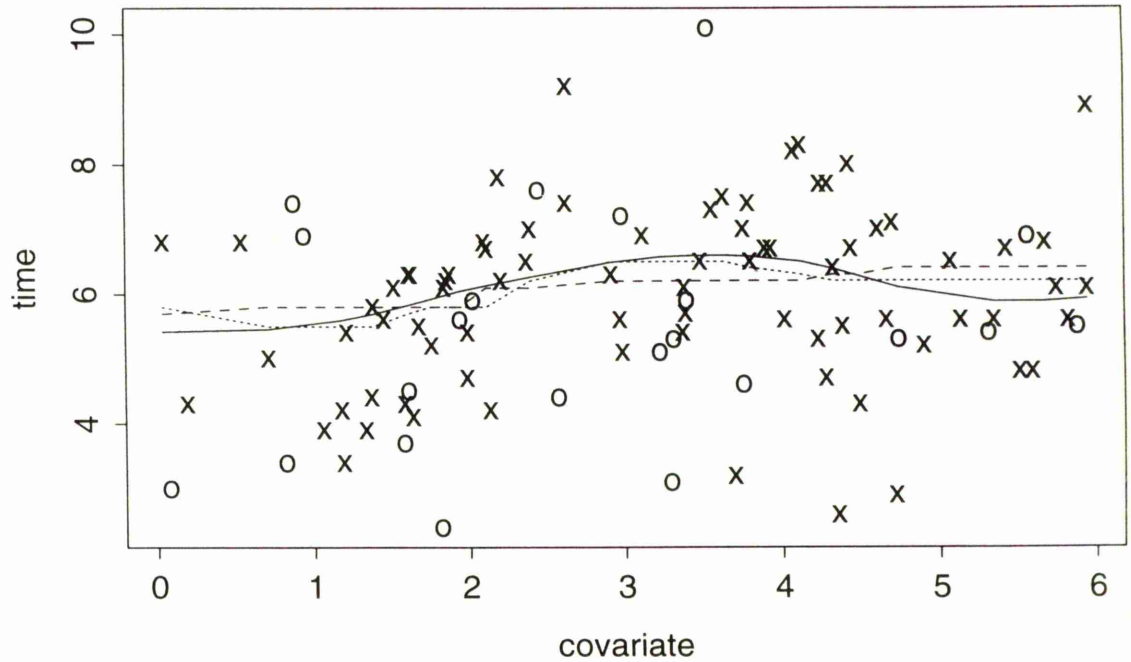



Figure 3.1.2 : 50th percentile curves




h1=0.7 ,h2=0.075

Simulated data with no effect : 20% censoring

Figure 3.1.3 : 25th percentile curves

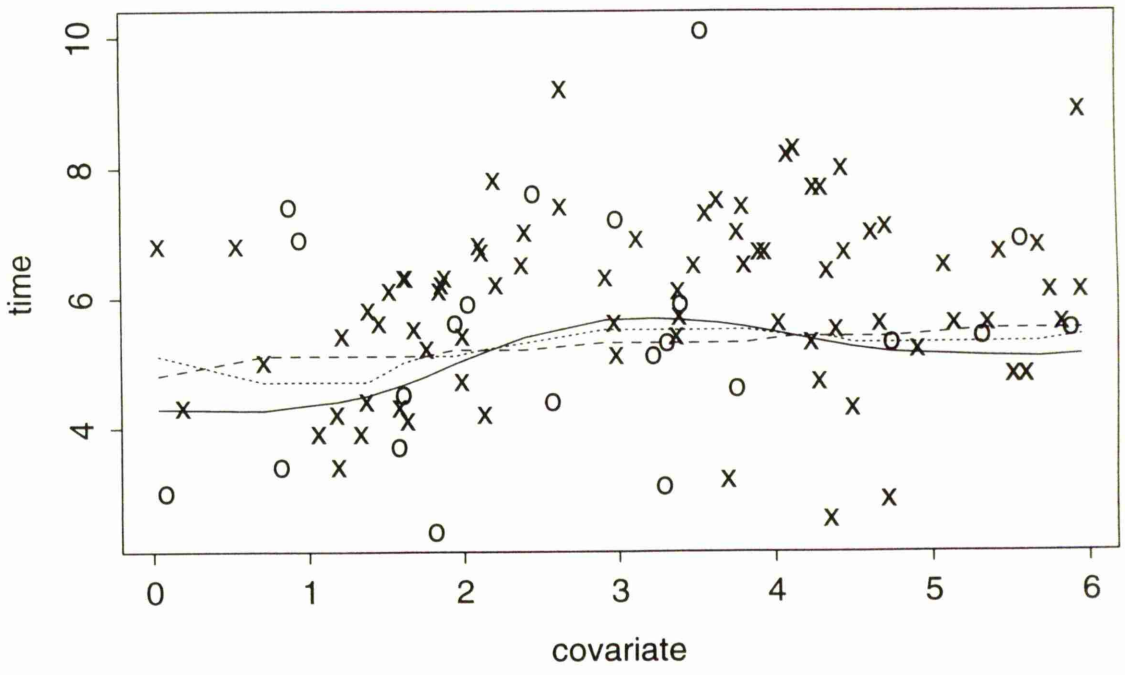
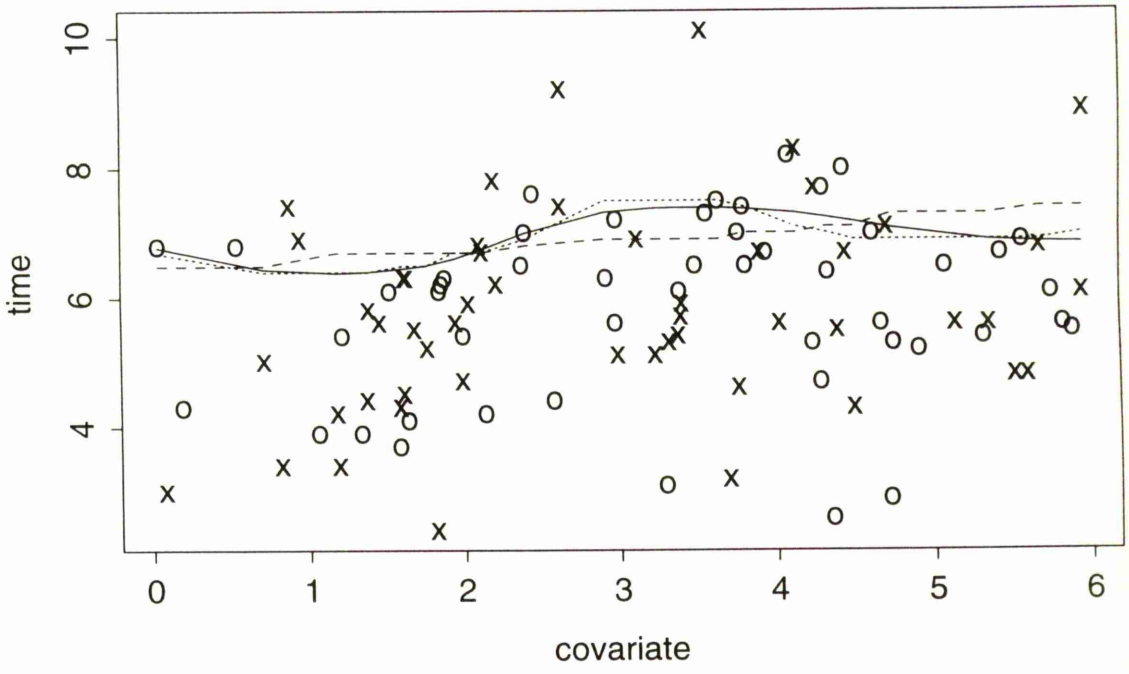



Figure 3.1.4 : 75th percentile curves




h1=0.7 ,h2=0.075

Simulated data with no effect : 20% censoring

Figure 3.1.5 : Survivor curves at covariate=2

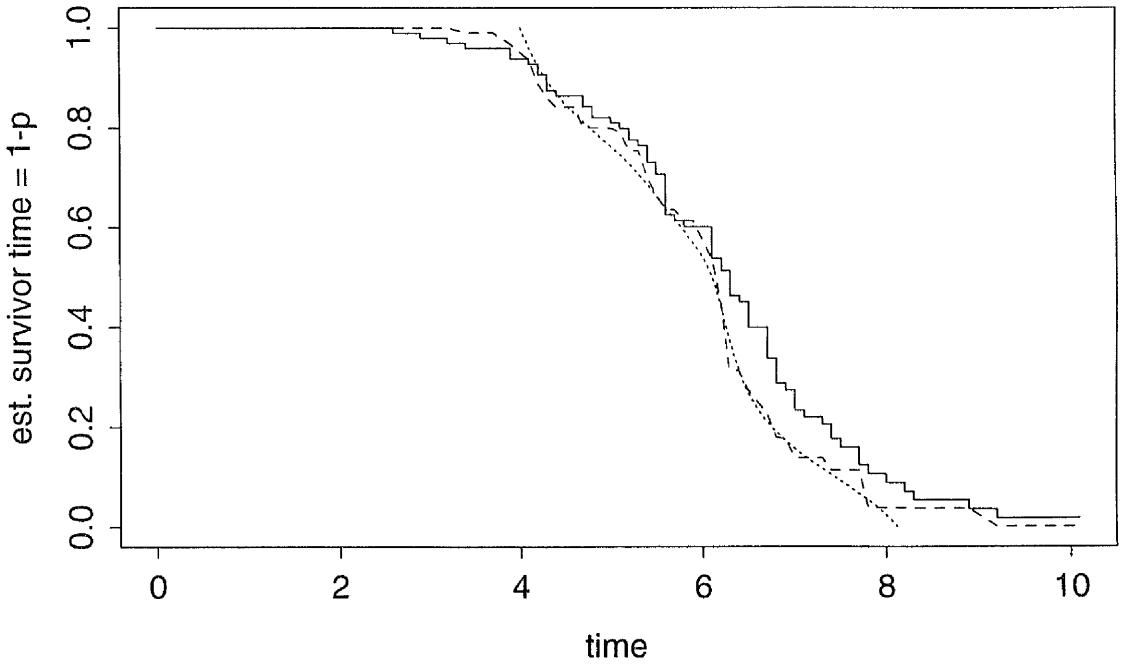
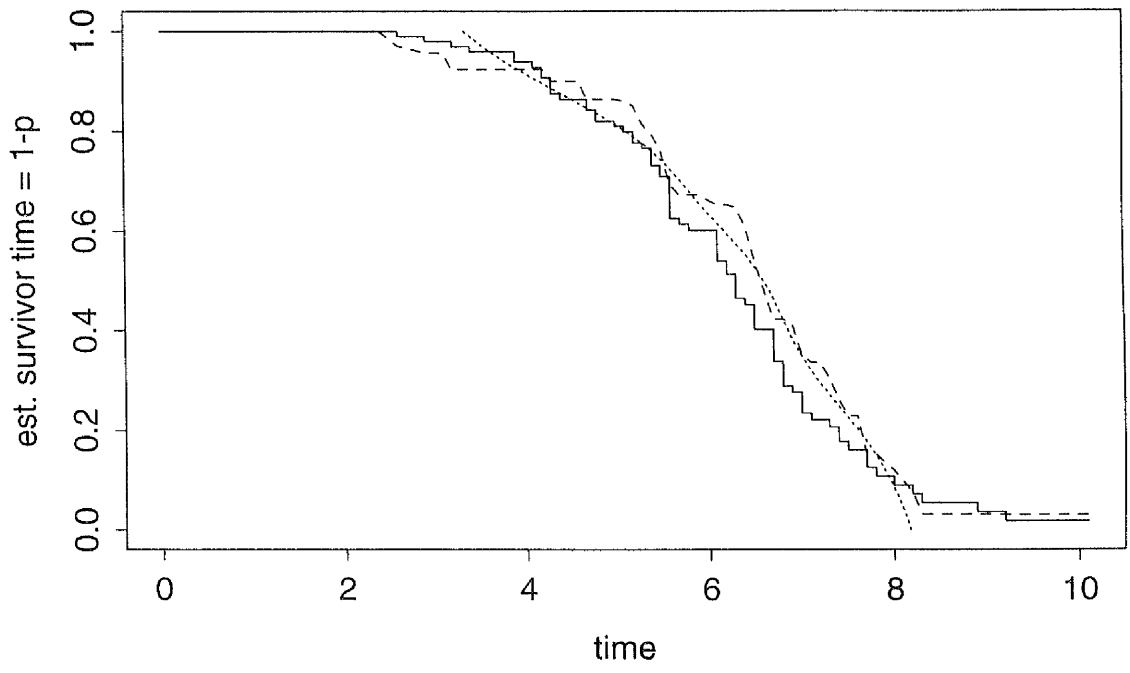



Figure 3.1.6 : Survivor curves at covariate=4




h1=0.7 ,h2=0.075

are not randomly censored by the definition of section 2.1.2, resulting in larger numbers of small censored times. The Kaplan–Meier curve (the solid line in Figure 3.1.8) is more of an "S-shape", with the slope being less steep for times between 2 and 7 but more so for larger survival times. The smoothed survival curves estimated at a covariate value of 4 (referring to the broken and dotted lines in the same figure) have a similar shape and show why the percentile survival times have increased.

When the larger survival times are censored, the weighted Kaplan–Meier curve will not drop to zero. One solution in these cases would be to simply estimate a smaller quantile which will reflect the effect of the covariate for survival times which are mostly failures.

### 3.2.2 Example 2 – Simulated Quadratic Data

A key characteristic of the doubly–smooth quantile estimator is that it is not forced to be *monotonic*, unlike percentile curves derived under the Proportional Hazards model. In Cox's model, the covariate is expressed as $\exp(\beta z)$ where $\beta$ is the regression coefficient and $z$ is the covariate and is therefore linear on a log–scale. The curves from the doubly–smooth approach are much more flexible. There may be situations where young and old people are more susceptible to a certain type of infection, or where low blood pressure is as detrimental to health as high blood pressure. In such cases, as will now be shown, smooth procedures have a certain advantage.

Consider a simulated data set of the form

$$t_i = (z_i - 3)^2 + 5 + \varepsilon_i \quad \text{where } i = 1, \ldots, 100$$

where $z$ is generated under a Un(0,6) distribution and $\varepsilon \sim$ N(0,2.5). The corresponding censoring indicator was generated using a Bi(100,0.8) distribution to produce a set of data which had approximately 20% censoring.

A scatterplot of these data is shown in Figure 3.2.1. Adding the doubly–smooth $50^{th}$ percentile curve ($h_1$=5, $h_2$=0.075) and comparable proportional haz-

simulated data with no effect : 50% censoring

Figure 3.1.7 : 50th percentile curves

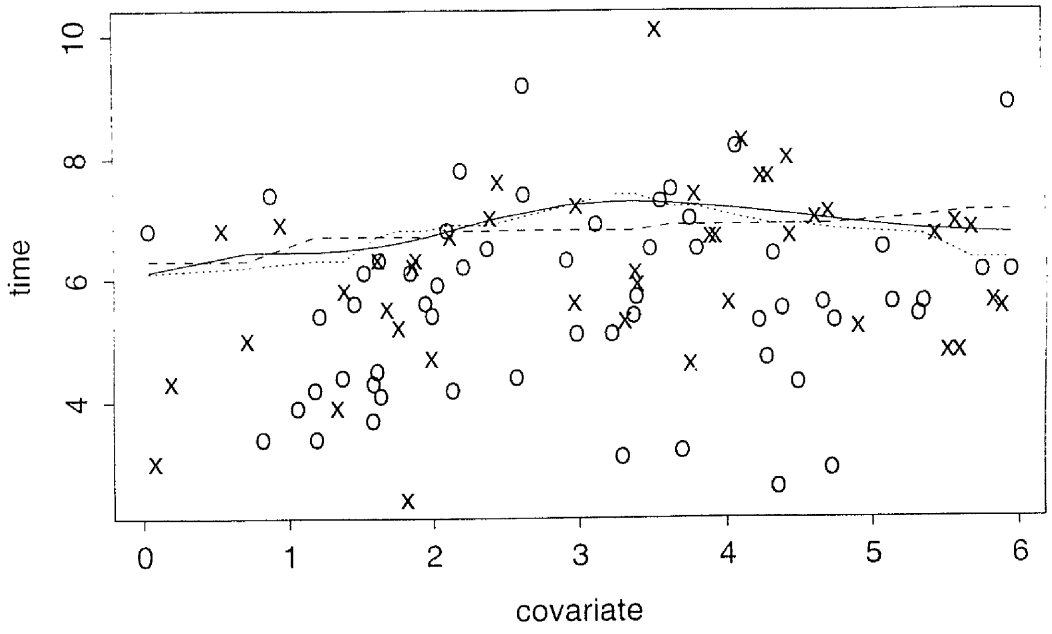
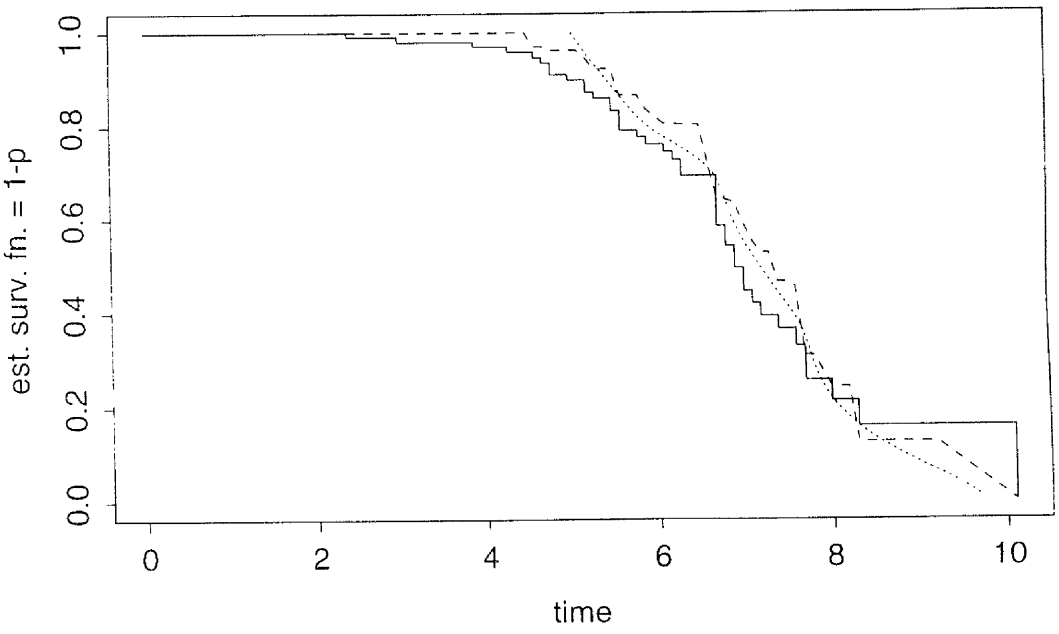



Figure 3.1.8 : Survivor curves for covariate=4




h1=0.75 ; h2=0.075

ards and smooth PH curves (represented as a solid line, a broken line and a dotted line respectively), to this produces the plot in Figure 3.2.2. It is clear that while the proportional hazards assumption of monotonicity is inappropriate for these data, the two smooth approaches present very similar quadratic median curves. The "shape" of the data is very obviously quadratic here, not least because of knowledge of the data's source. However, it may be in other situations that it is not so clear but that a graphical representation might point to its presence.

Again we can look at the behaviour of the estimator for this data more closely. In Figures 3.2.3-4, we have a plot of the Kaplan–Meier curve (solid line) with the weighted Kaplan–Meier curves corresponding to the covariate values of 0.5 and 3 drawn as broken lines. Since these covariate values correspond to extreme survival times, they depict the variability of the different curves across the covariate. Again the curves for the covariate values are looked at separately with the dotted line here representing the second smoothing procedure, which effectively performs a nonparametric regression analysis on $t$ *vs.* $p$. Similar plots for covariate values of 2 and 5 are on the following page and show the different behaviour of the Kaplan–Meier curves at points where the slope of time vs. covariate is more gradual.

### 3.2.3 Example 3 – Data simulated from a Proportional Hazards Model

In the previous two examples, data have been generated in such a way that the scatterplot has a known "shape". In this example, the data are generated under the proportional hazards model. This information then allows us to assess the performance of the doubly–smooth approach by comparing it with the PH curve, also estimated from the data but with the knowledge that the PH assumption is true. The aim here is to examine the performance of the nonparametric approach when the data are generated from a parametric model.

For the definition of Cox's Proportional hazards model, see Cox(1972) or Section 5.1 in the quantile review. Here the data are generated from a proportional

Simulated Quadratic Data : 20% censoring

Figure 3.2.1 : Scatter plot

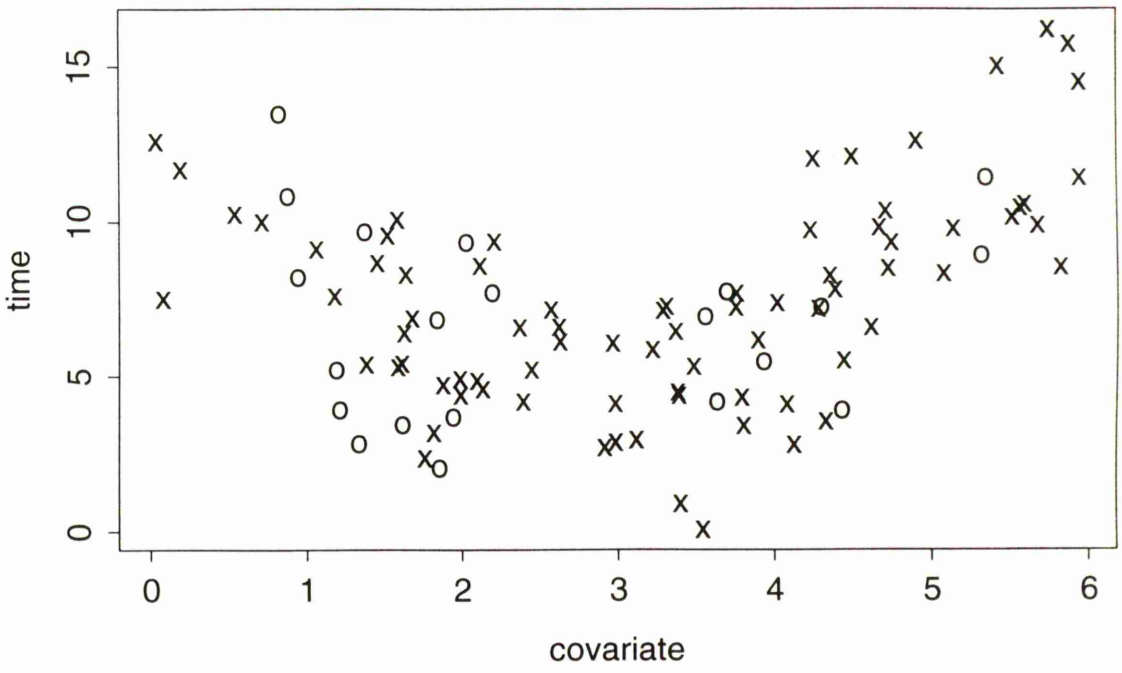
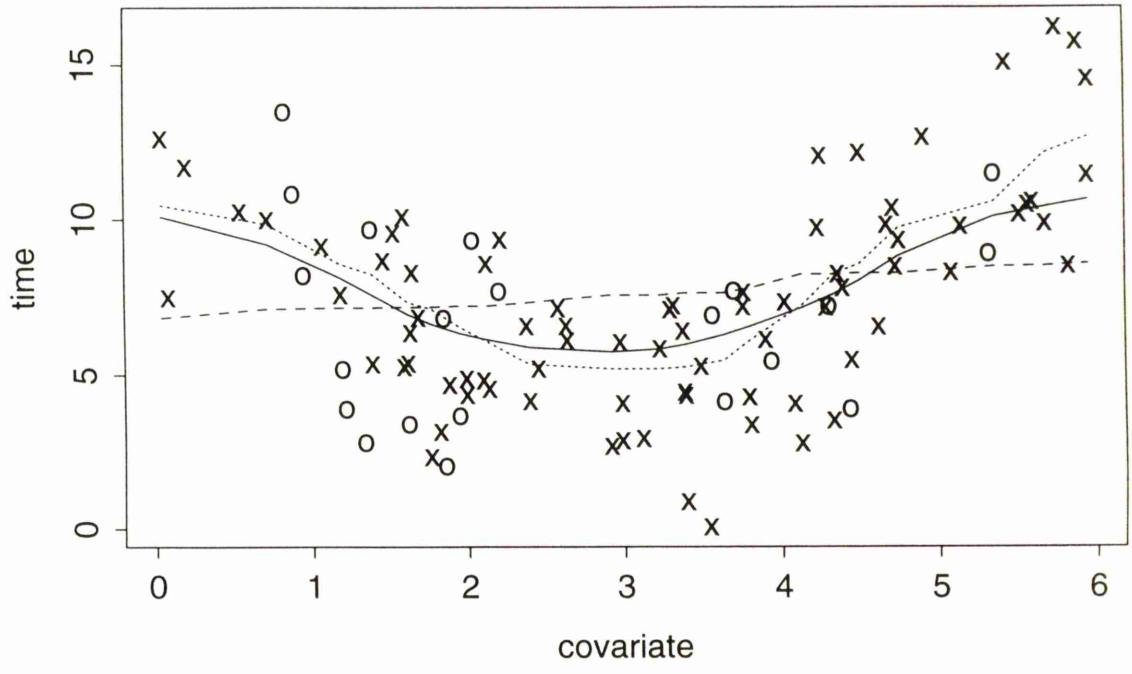



Figure 3.2.2 : 50th percentile curves




h1=0.5, h2=0.075

Simulated Quadratic Data : 20% censoring

Figure 3.2.3 : Survivor curves at covariate=0.5

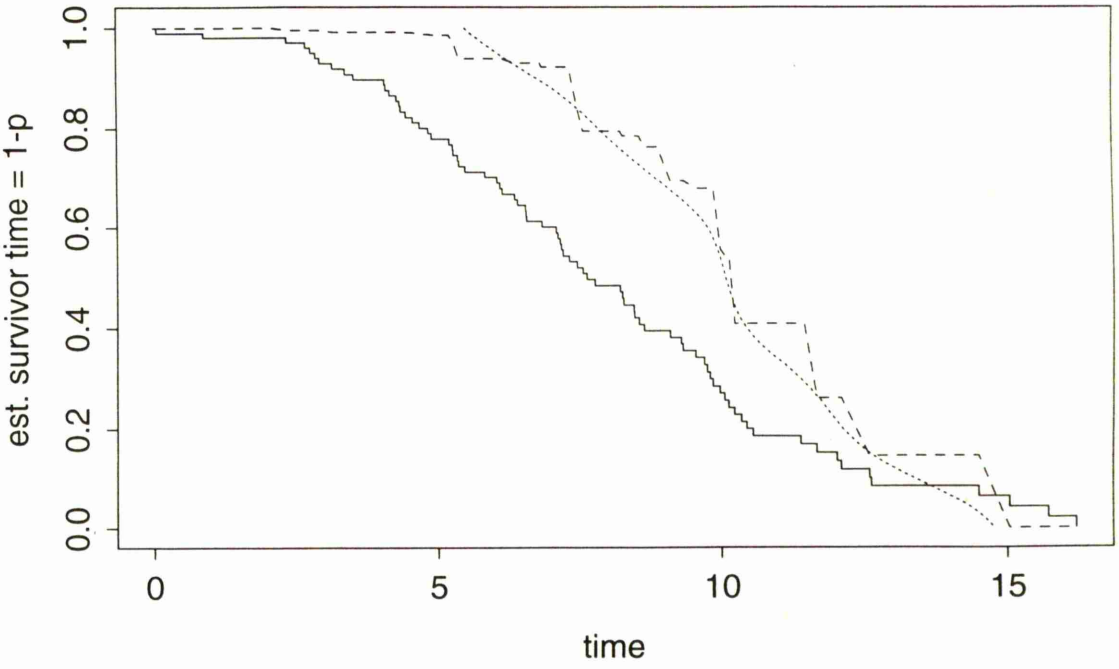
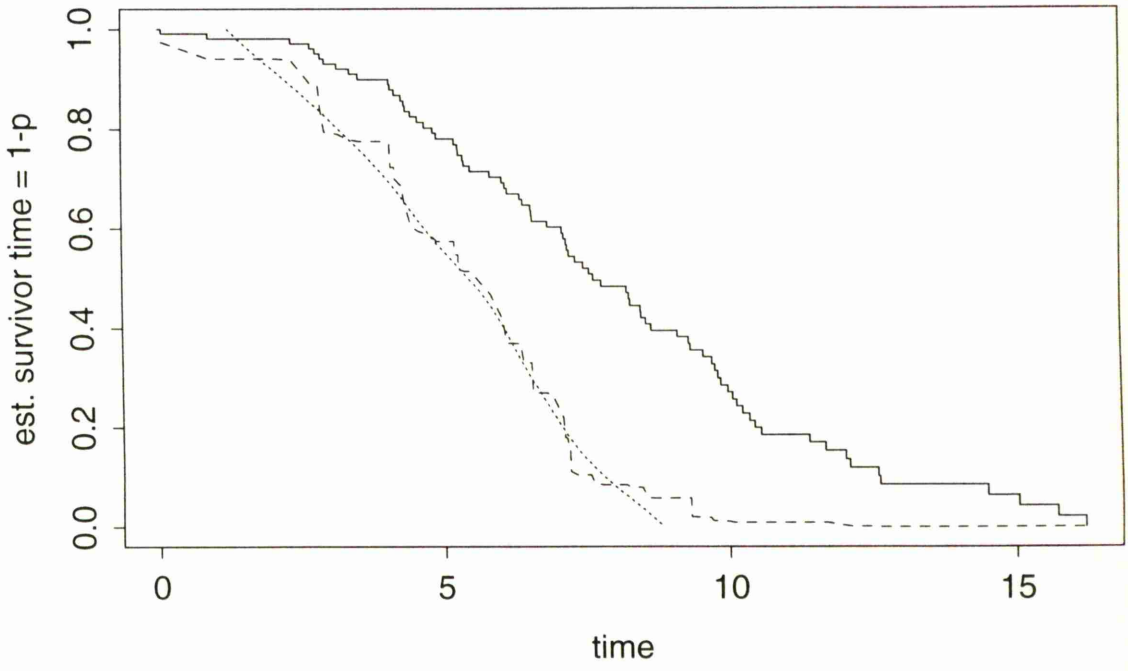



Figure 3.2.4 : Survivor curves at covariate=3




h1=0.5, h2=0.075

Simulated Quadratic Data : 20% censoring

Figure 3.2.5 : Survivor curves at covariate=2

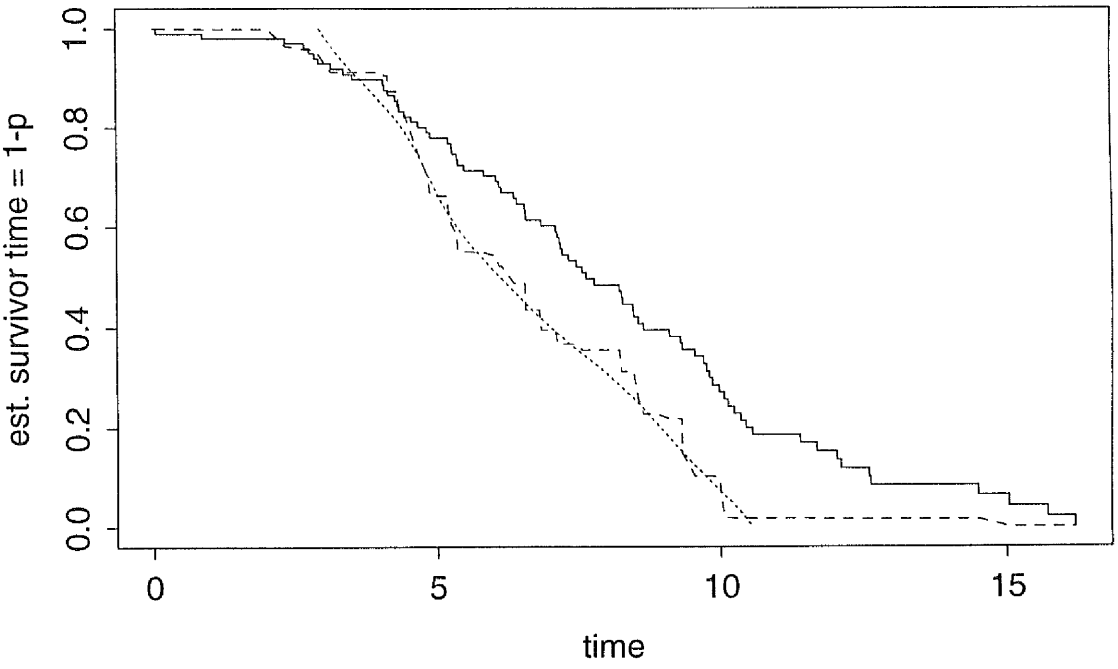
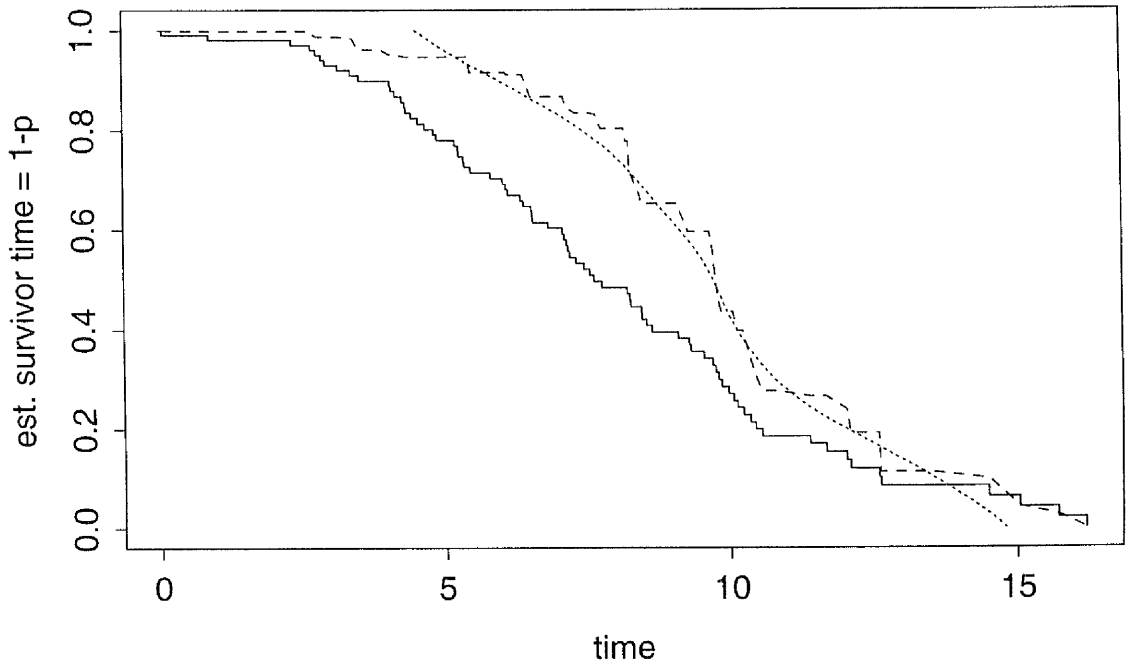



Figure 3.2.6 : Survivor curves at covariate=5




h1=0.5, h2=0.075

hazards model with regression coefficient, $\beta$, equal to -1 and a sample size of 100. Then, for a given covariate $z$ generated under a Un(0,1) distribution, failure times can be simulated from an Ex($e^z$) distribution and censored times from an Ex($2e^z$) (although, in the PH model, censored times should be generated independently of the covariate). A set of data of the form (time,censoring indicator, covariate) is created and the observed amount of censoring here is 22%.

A scatterplot of the data is shown in Figure 3.3.1. As expected from choosing $\beta = -1$, as the values of the covariate increases, the "risk" decreases and so survival time increases but, also, the spread of values increases. As in the other examples, this is followed by a plot with the $50^{th}$ doubly–smooth ($h_1$=0.1, $h_2$=0.075), proportional hazards and smooth PH percentile curves. For the $25^{th}$ and $75^{th}$ percentile curves under the three models, see Figures 3.3.3 and 3.3.4 respectively. The subjective impression of the data is confirmed and the doubly–smooth curves look similar to both those derived from a PH model, except perhaps at the extremes of the covariate on the right where the smooth PH and doubly–smooth agree. The behaviour of the smooth curves is justified here by the relatively big cluster of low survival times in this area.

The three survivor curves, Kaplan–Meier (solid line), weighted Kaplan–Meier (broken line) and regression on p (dotted line), have been plotted for the covariate values of 0.2 and 0.8 and are shown in Figures 3.3.5 and 3.3.6 respectively. By choosing two covariate values so far apart, the increase in percentile time is obvious from the shift of the weighted Kaplan–Meier curves.

This example helps to show that, when the proportional hazards assumption is valid, the doubly–smooth percentile curves present the same basic shape as those derived from the model. Later, a likelihood test statistic calculated from the estimated survivor functions will be used as a formal test of when the Proportional Hazards model is appropriate.

Data simulated from a Proportional Hazards Model

Figure 3.3.1 : Scatter plot

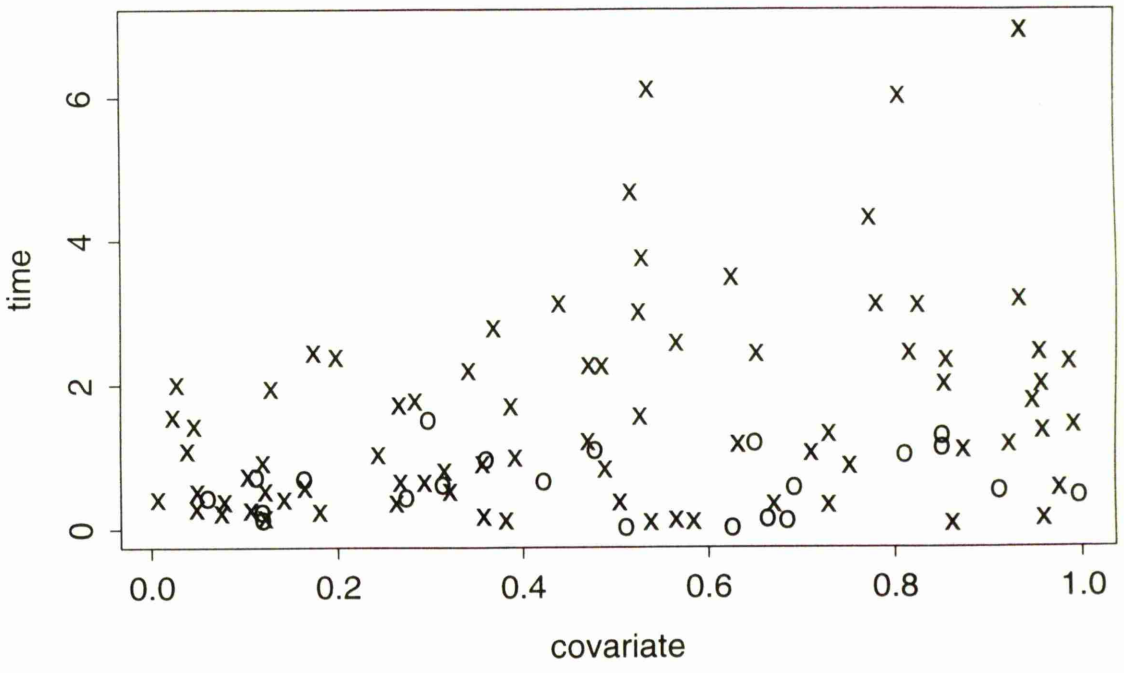
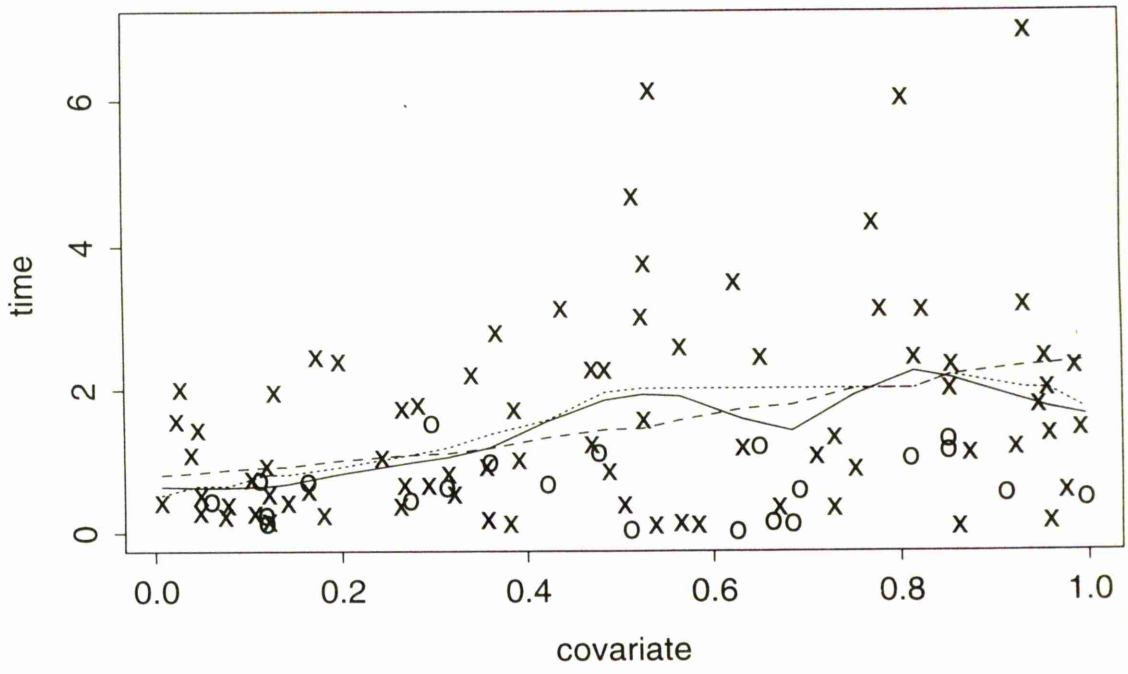



Figure 3.3.2 : 50th percentile curves




h1=0.1 , h1=0.075

Data simulated from a Proportional Hazards Model

Figure 3.3.3 : 25th percentile curves

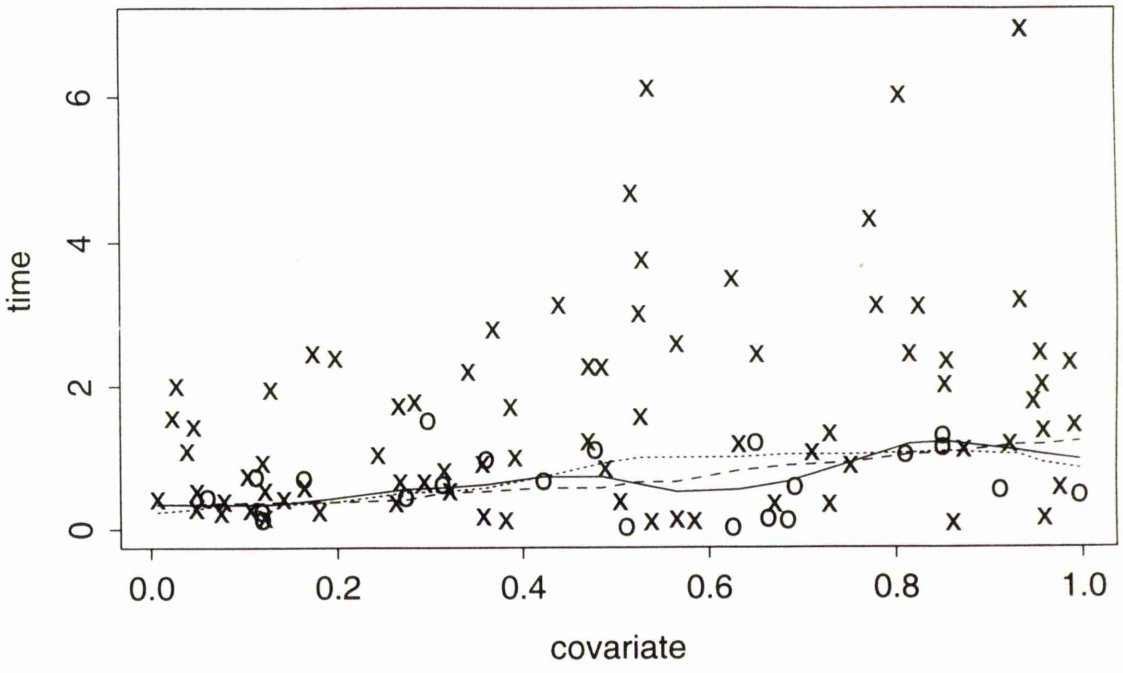
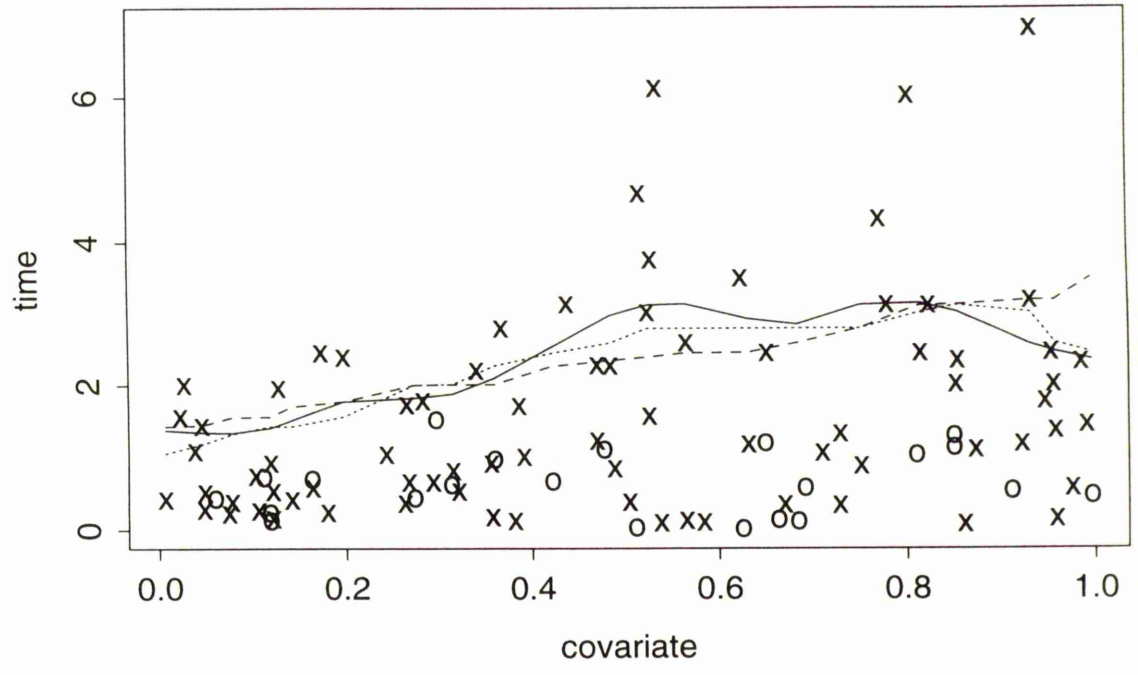



Figure 3.3.4 : 75th percentile curves




h1=0.1 , h1=0.075

Data simulated from a Proportional Hazards Model

Figure 3.3.5 : Survivor curves at covariate=0.2

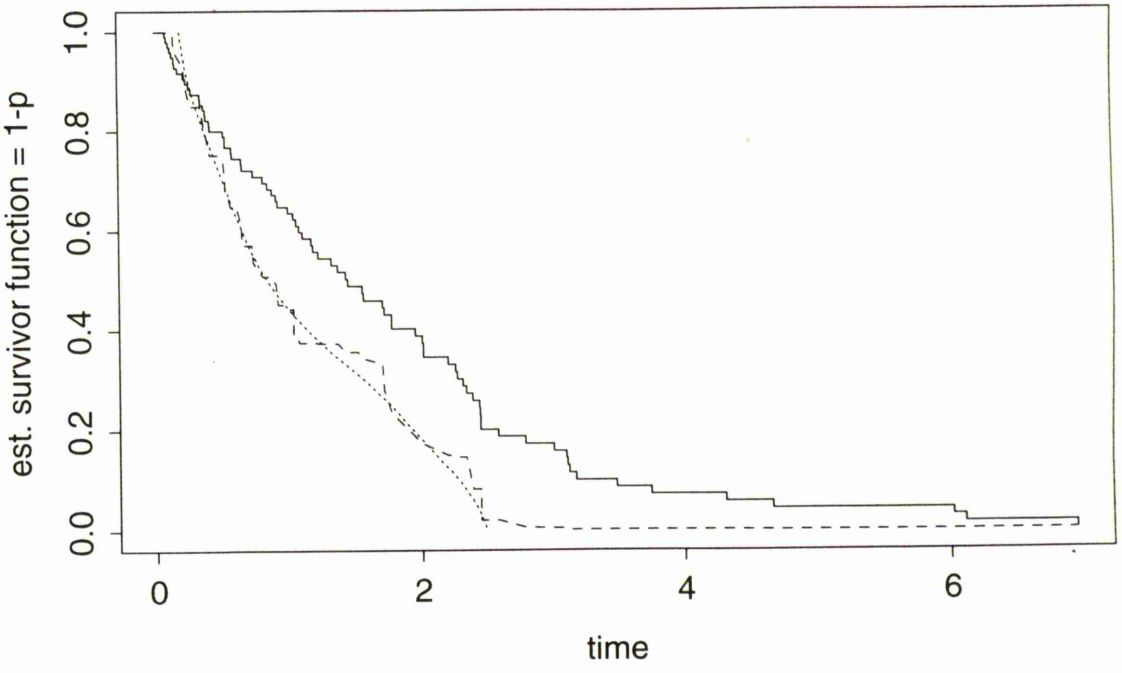
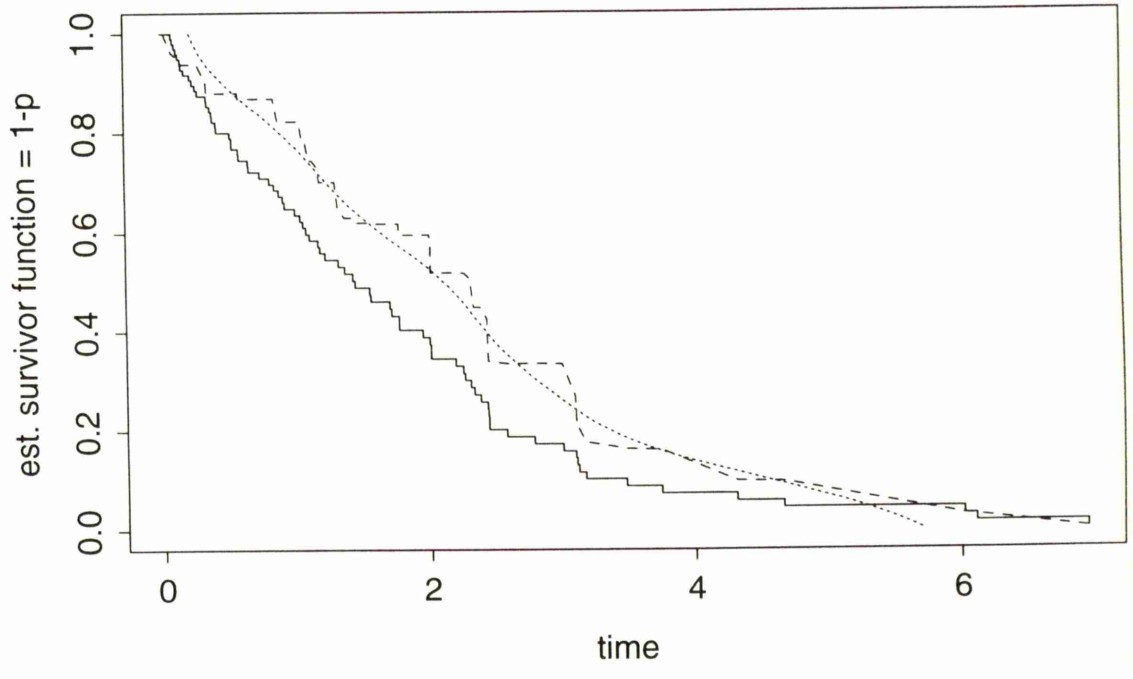



Figure 3.3.6 : Survivor curves at covariate=0.8




h1=0.1 , h1=0.075

### 3.2.4 Example 4 – Non-PH Simulated Data

The final set of simulated data considers a particular example where the proportional hazards model is not applicable. Here the data, plotted in Figure 3.4.1, exhibit decreasing variability of survival time as the covariate increases, converging to a value close to the median time.

The survival times were generated from a Normal distribution with mean 20 and sample size 150. The covariate values are simulations of an Ex(2) distribution and the censoring indicator is chosen from Bi(150,0.5) to attain approximately 20% censoring.

Figure 3.4.2 displays the median percentile curves under each of the 3 different methods. (The line types correspond to the same type of percentile curve as described in the other examples.) It may be worth noting that the median survival time appears to decrease slightly as covariate increases and that there is relatively little difference between any of the curves.

However, Figures 3.4.3 and 3.4.4 considers the $10^{th}$, $90^{th}$ and $3^{rd}$, $97^{th}$ percentiles respectively and highlights the consequence of placing a proportionality constraint on the hazard. While the doubly–smooth estimator is less efficient in its approximation of the underlying true curves, its behaviour more accurately describes the data than the curves produced under the proportional hazards assumption. The key to this issue is that the PH model fails to capture the variance structure across the covariate and so over (under) estimation occurs for large (small) centile values.

Transformations of the covariate may remove some of the heteroscedascity in some examples. However, since this is not routinely performed in PH analysis, it is reasonable to assess the fit of the curves on the raw values.

Non-PH Simulated data : 20% censoring

Figure 3.4.1 : Scatter plot

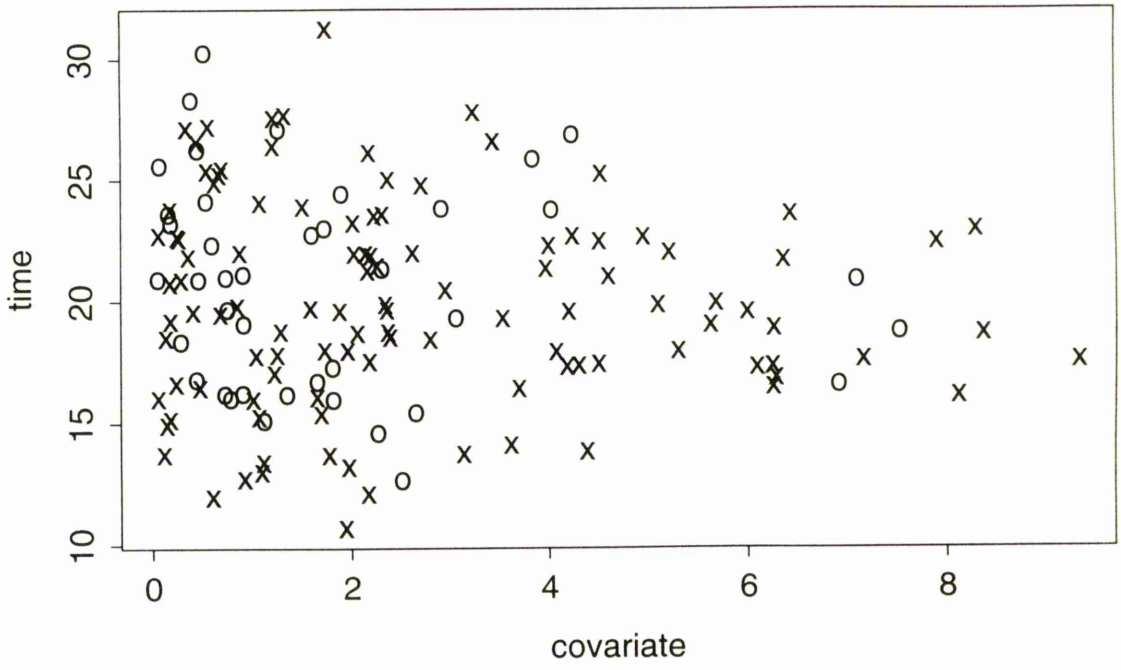
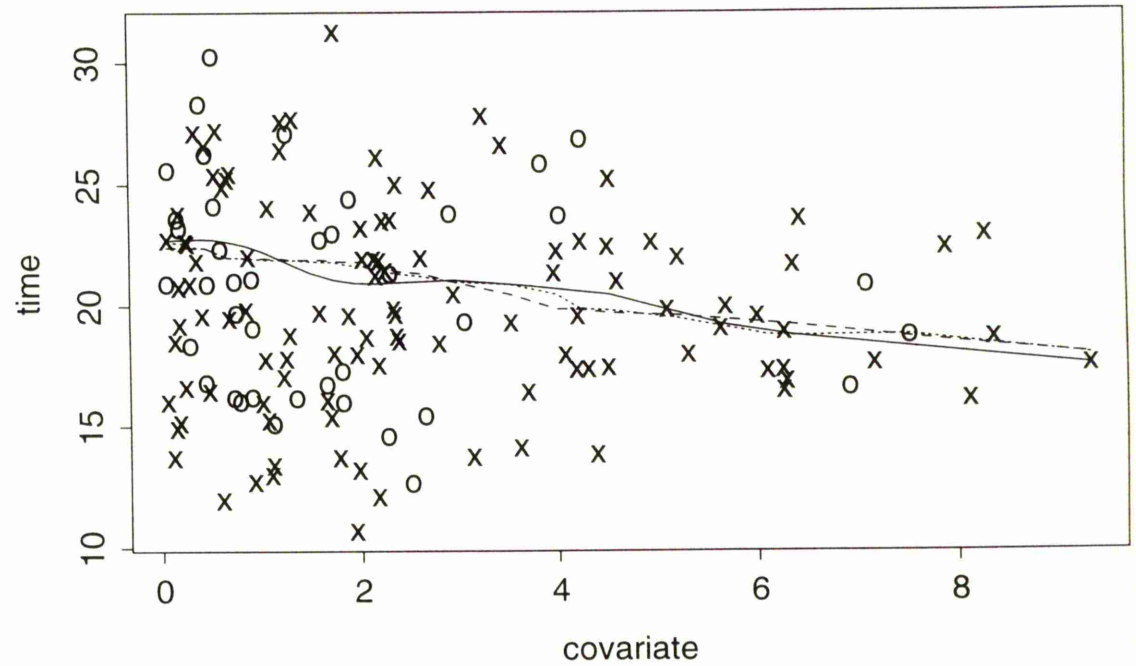



Figure 3.4.2 : 50th percentile curves




h1=0.7 ,h2=0.075

Non-PH Simulated data : 20% censoring

Figure 3.4.3 : 10th and 90th percentile curves

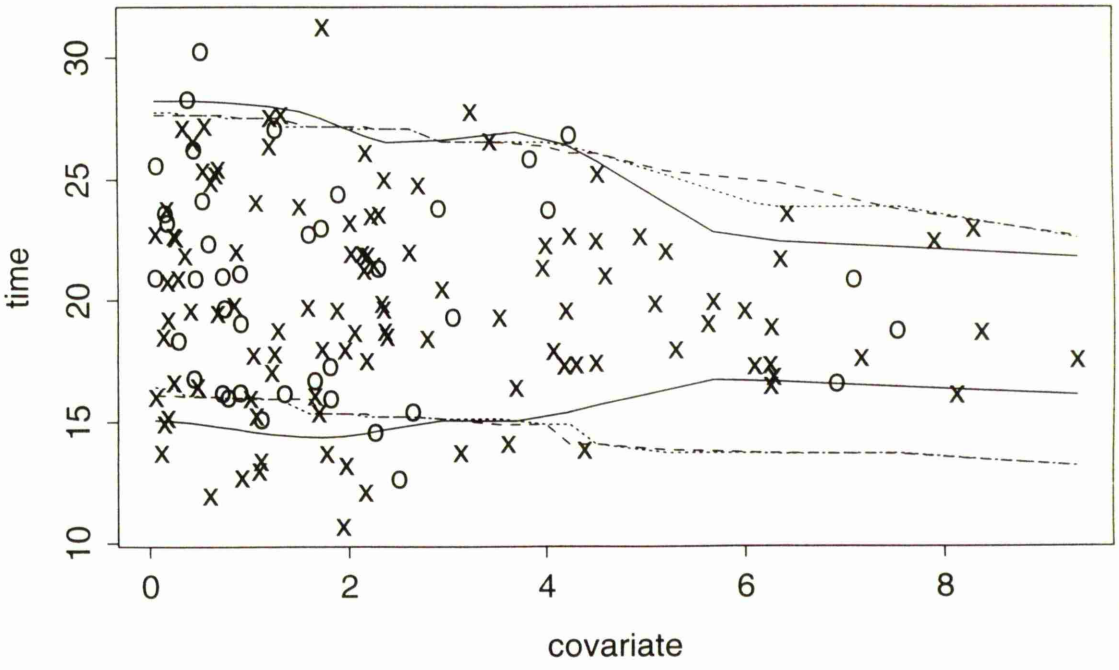
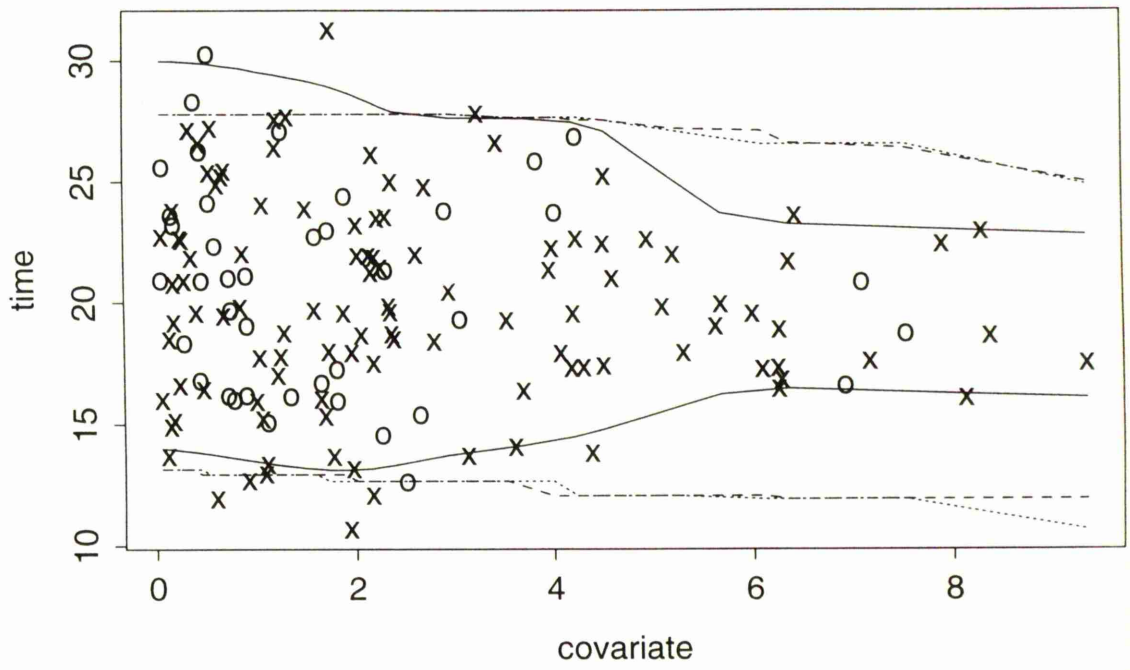



Figure 3.4.4 : 3rd and 97th percentile curves




h1=0.7 ,h2=0.075

### 3.2.5 Example 5 - Stanford Heart Transplant Data

The Stanford Heart Transplant Data is a very commonly used example for survival analysis and was originally presented in a paper by Crowley and Hu (1977). Assessment of the manner in which the time to death following transplantation is influenced by age has been tackled using many different methods. Doksum and Yandell (1987) produced median survival curves based on the Proportional Hazards model similar to those shown already for the simulated data. The nonparametric approach, to which they compared these curves, was based on a "running" median, proposed by Beran (1981). In this example, the number of days to failure is usually transformed by taking the logarithm to the base 10, due to the skewness of the data.

Figures 3.5.1–2 representing a scatterplot, with $40^{th}$ percentile curves from the doubly–smooth ($h_1$=5, $h_2$=0.075), proportional hazards and smooth PH ($k = \frac{1}{5}$) approaches help to describe the age effect on survival time. (Note : we use the 40th percentile as there is a slightly higher proportion, $\approx$ 40%, of censoring in this example.) It is difficult to draw fine conclusions from the collection of these curves as each displays slightly different features. The doubly–smooth curve (solid line) shows a steady decrease in percentile survival time as does the curve under the PH model (broken line). However, before the age of 30 years, the doubly–smooth curve displays a slight increase in failure time as age increases. This may be because the majority of censored values in this region have larger age values than the failure times. The smooth PH quantile curve (dotted line) suggests a constant $40^{th}$ percentile survival time until the age of 40, after which it decreases relatively quickly. It is difficult to assess which of these conclusions is correct. Futher analysis of these models will be carried out in Chapter 4.

The following four plots (Figures 3.5.3–6) show, for age values of 15, 65, 30 and 50, the Kaplan–Meier curve (solid line), the weighted Kaplan–Meier (broken line) at the respective age values with $h_1$=5 and the nonparametric regression of $t$ *vs.* $p$ (dotted line) with $h_2$=0.075.

# Stanford Heart Transplant Data

Figure 3.5.1 : Scatter plot

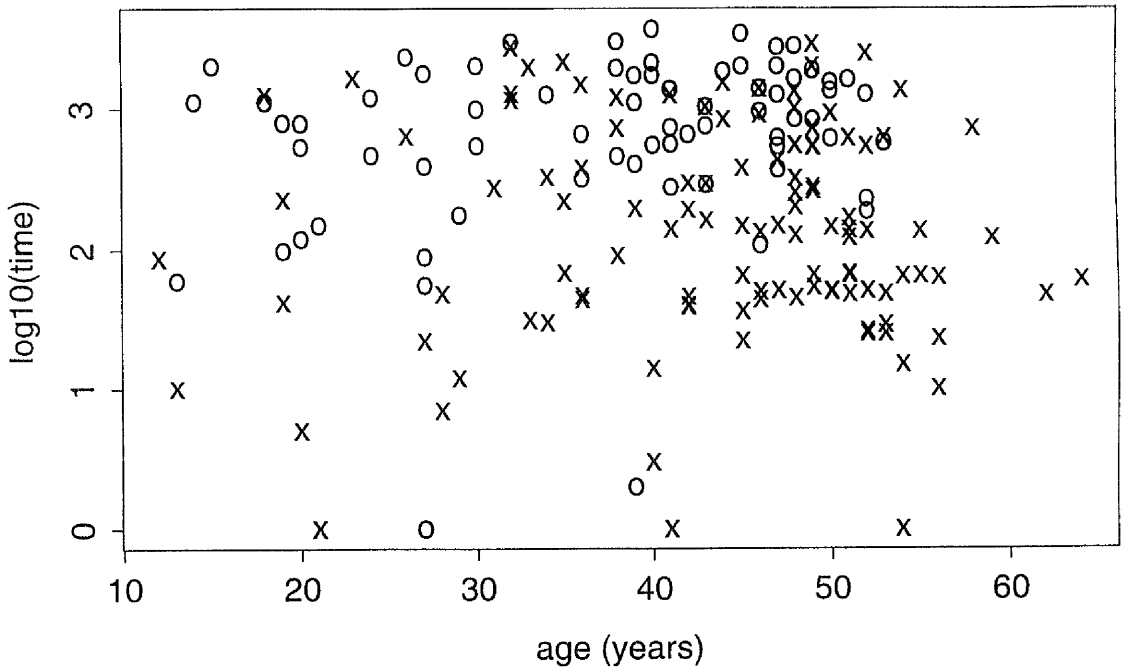
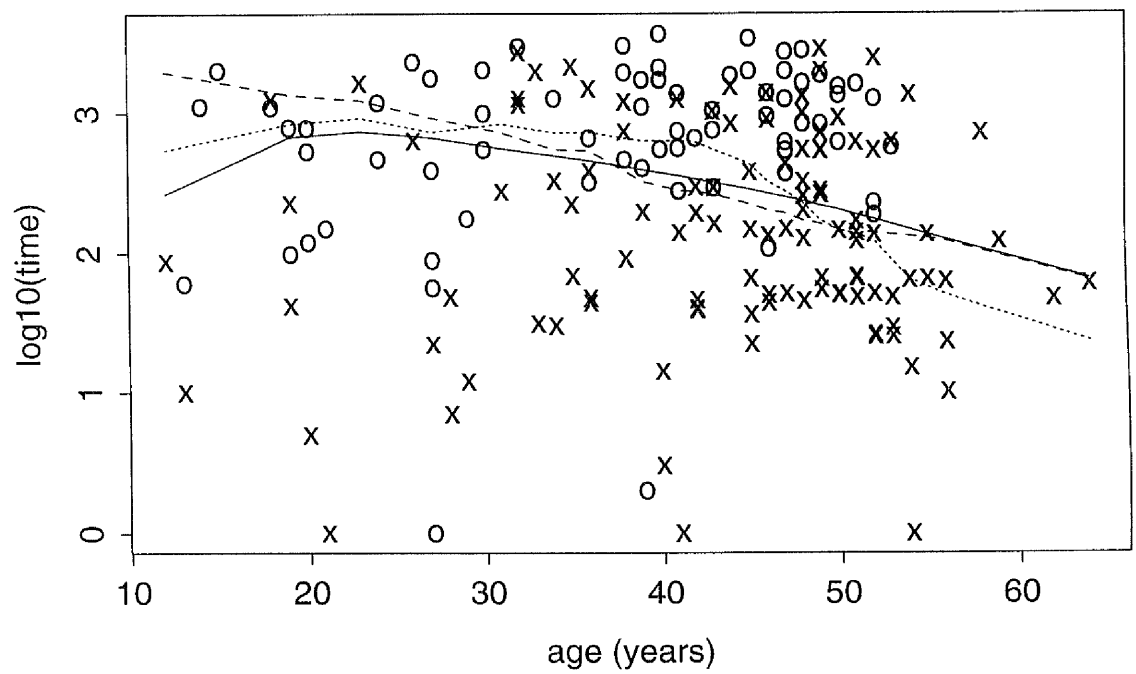



Figure 3.5.2 : 40th percentile curves




h1=5 ; h2=0.075

Above, the behaviour of the percentile curves at age<30 years was discussed. From the plot of the survival curves at 15 years (Figure 3.5.3) we see that the regression of $t$ *vs.* $p$ is perhaps over-smoothing these weighted KM curves. However, $h_2$=0.075 appears to be a suitable smoothing parameter across the other values of the covariate so we will continue using this value but be aware that effects like this may occur, particularly at the edges where there is likely to be less data.

### 3.2.6 Example 6 – Multiple Myeloma data-set

The effect of age on survival time is analysed for a sample of 48 patients in a study on multiple myeloma. Krall, Uthoff and Harley (1975) gives details of the other covariates measured. This is a relatively small data–set with a moderate proportion of censoring (25%).

A subjective impression from the scatter plot of the data, Figure 3.6.1, might suggest that, for most people, survival time will not be influenced by age. However, there appears to be a small group of people for whom survival time is larger than the rest of the sample. The general pattern of the scatterplot is similar to the data simulated under the PH model with a regression coefficient which would be expected to be positive. The effect of using log(time) in this example arises naturally as a suggestion to handle skewness in the survival times (as in the Stanford data).

Figure 3.6.2 shows the scatterplot with survival time given a logarithmic transformation and the doubly–smooth (solid line) median survival curve with $h_1$=3 and $h_2$=0.075, the proportional hazards curve and smooth PH curve (broken line and dotted line respectively). The curves for the $25^{th}$ and $75^{th}$ percentiles are given in Figures 3.6.3 and 3.6.4. While the PH median curve suggests, on the log time scale, that survival time changes very little as age increases, the corresponding doubly–smooth curve proposes an increase in survival time until about 55 years, which then decreases slowly until a much sharper decrease at 70

# Stanford Heart Transplant Data

Figure 3.5.3 : Survivor curves for age=15

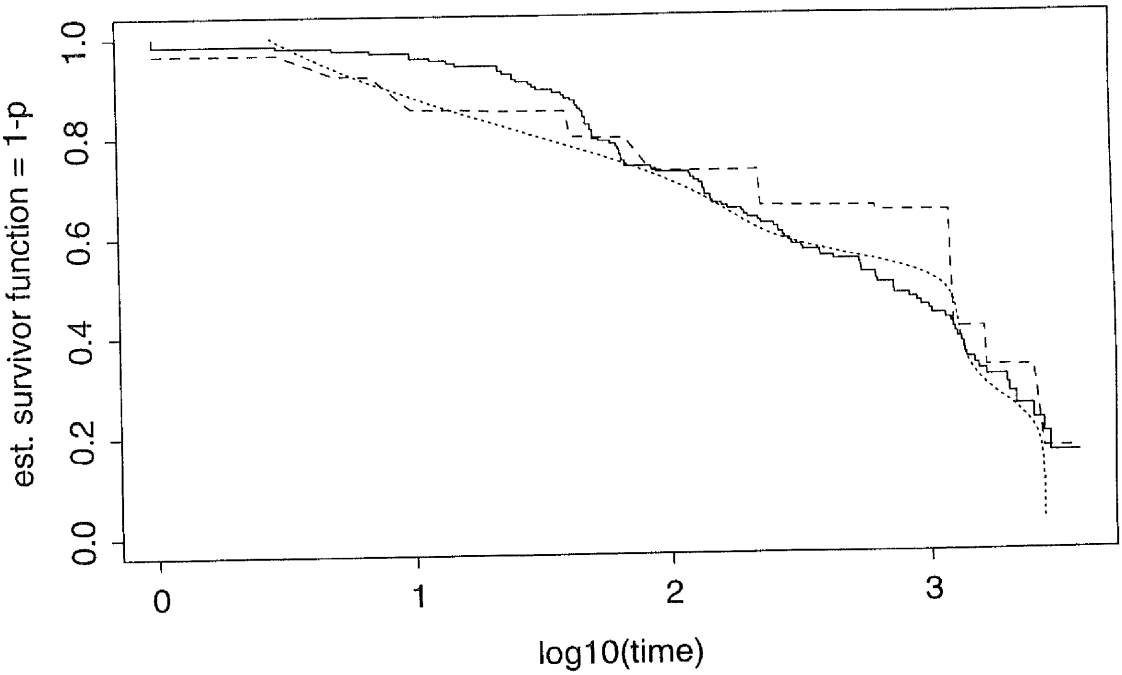
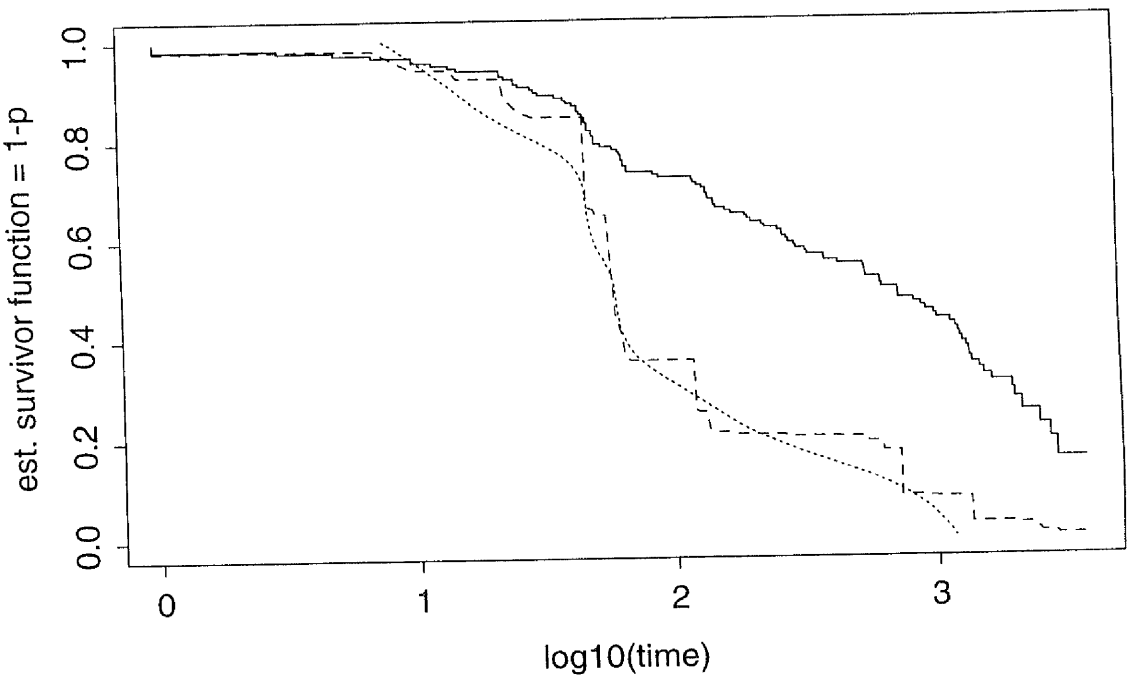



Figure 3.5.4 : Survivor curves for age=65




h1=5 ; h2=0.075

# Stanford Heart Transplant Data

Figure 3.5.5 : Survivor curves for age=30

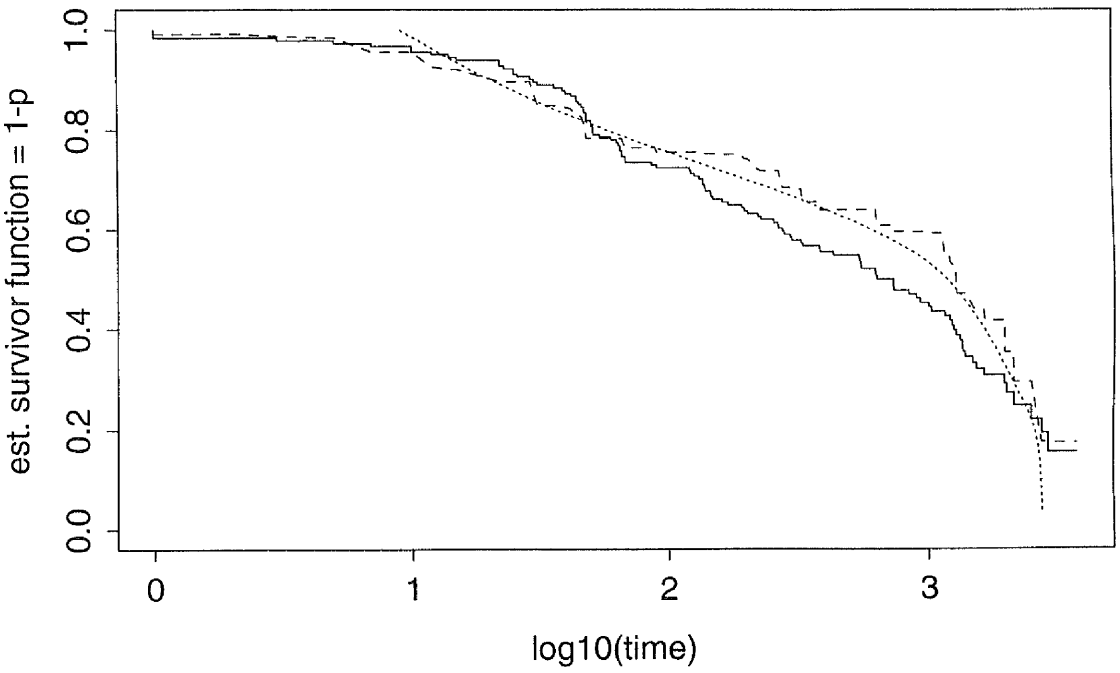
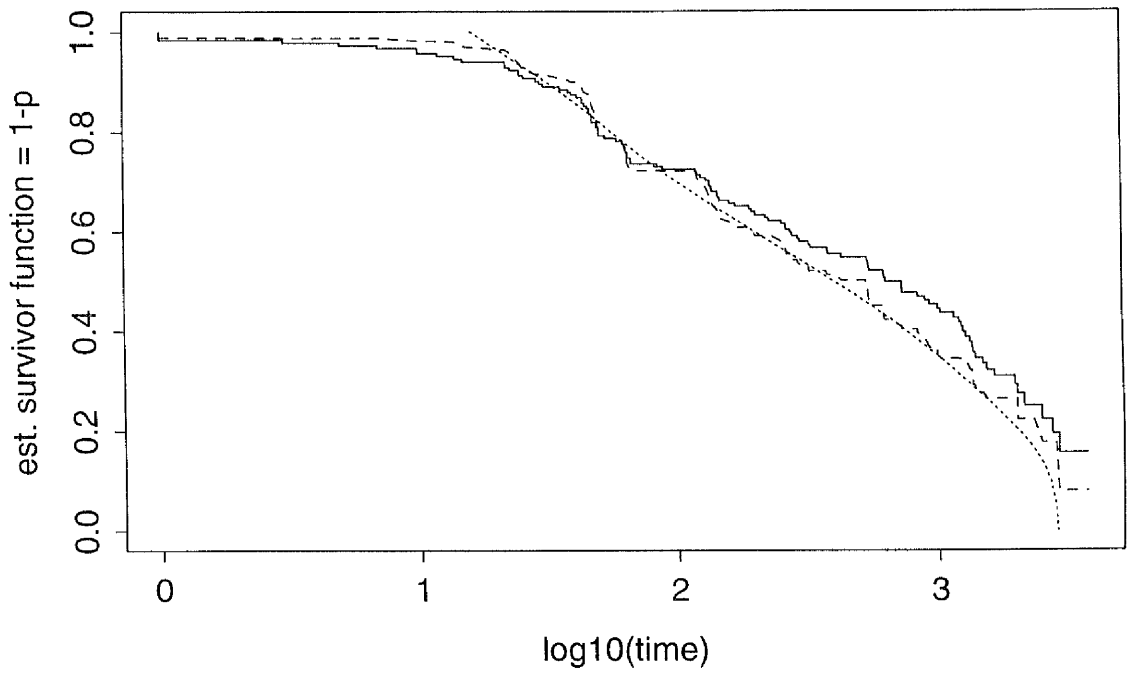



Figure 3.5.6 : Survivor curves for age=50




h1=5 ; h2=0.075

years. The smooth PH curve lies between these two as an intermediate suggestion. However, this effect is perhaps not strong enough to be attributed to more than random variation although it is more pronounced in the $75^{th}$ percentile curves.

Again two covariate values of interest are selected and the weighted KM (broken line) and $t$ *vs.* $p$ regression (dotted line) curves are displayed. Figures 3.6.5 and 3.6.6 make the difference in the higher percentiles between 60 and 70 very obvious.

While the data may be adequately described by a proportional hazards model, the doubly–smooth approach is also providing a plausible relationship between survival time and age which may be worth consideration.

### 3.2.7 Example 7 – Renal Unit Data

Chapter 6 discusses in detail the research and results of work carried out with Dr. Peter Rowe at the Renal Unit of the Western Infirmary in Glasgow. The main aim was to detect a particular type of deterioration sometimes found in transplanted kidneys called progressive graft dysfunction (PGD). Following on from this, there was interest in comparing the time until PGD occurred with survival times from patients who had stable grafts or who had suffered from the more severe event of kidney failure. The latter two groups were treated as censored observations for the analysis and those who suffered PGD as failures. The effect of various covariates was explored. One in particular required careful interpretation and we shall use these data as another example of where the information in a nonparametric percentile curve is valuable.

Yearly blood pressure measurements were extracted from a database of all patients visiting the Renal Unit. Included as a time–dependent covariate in a proportional hazards model, the mean blood pressure (MBP) was found to have a significant effect on time to PGD. The implication of analysing time–dependent covariates is explained in more detail in the section entitled "Factors Affecting the Onset of Progressive Graft Dysfunction" in chapter 6. Briefly, for each year until a

## Multiple myeloma data-set

Figure 3.6.1 : Scatter plot

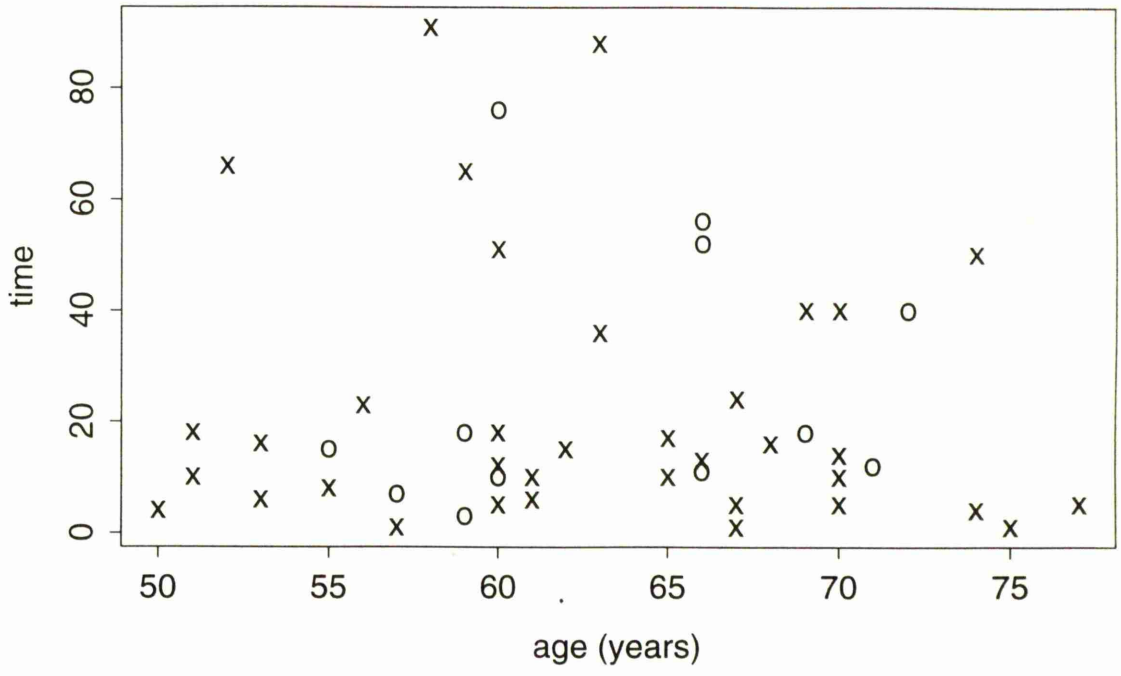
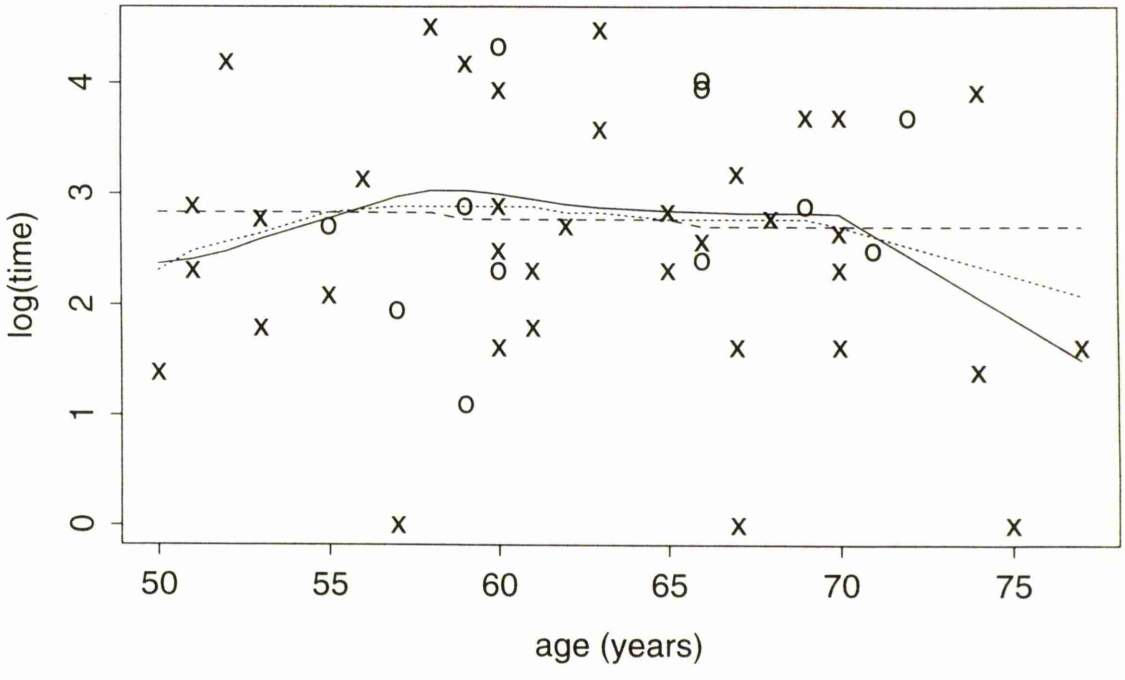



Figure 3.6.2 : 50th percentile curves




h1=3 . h2=0.075

Multiple myeloma data-set

Figure 3.6.3 : 25th percentile curves

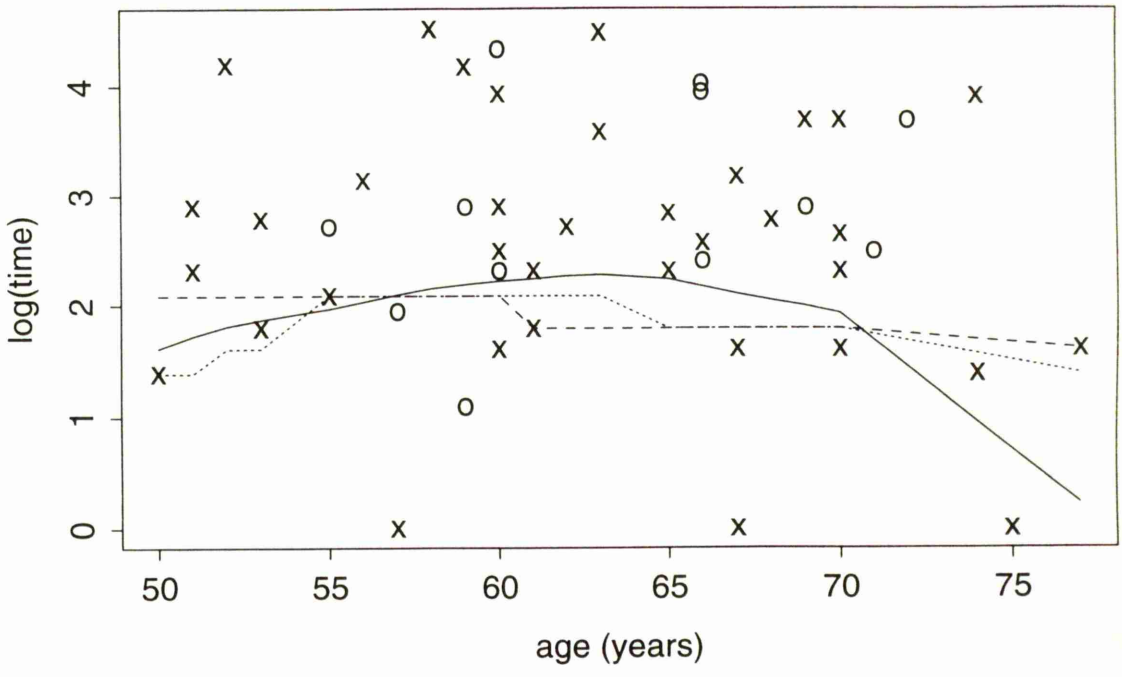
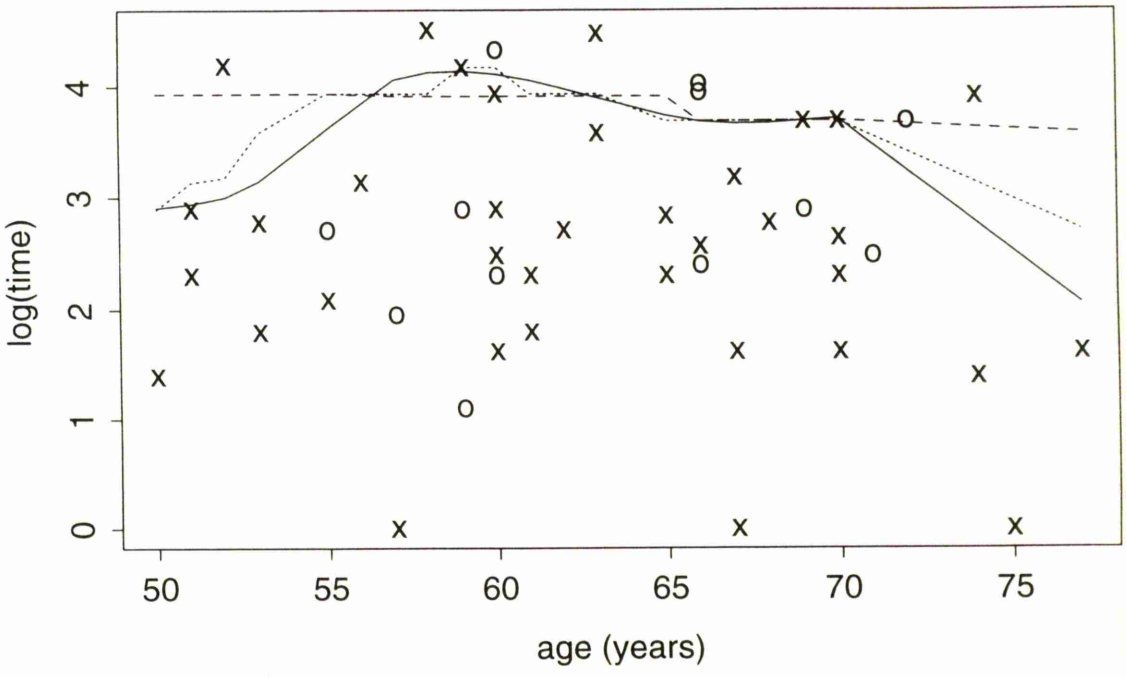



Figure 3.6.4 : 75th percentile curves




h1=3 , h2=0.075

## Multiple myeloma data-set

Figure 3.6.5 : Survivor curves at age=60

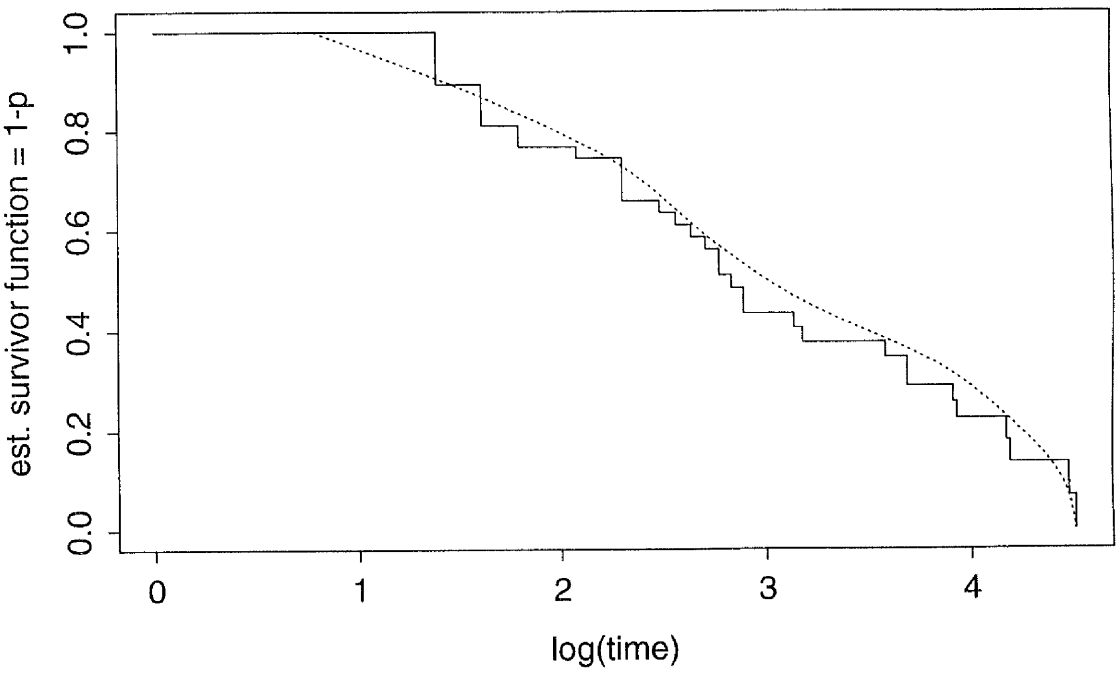
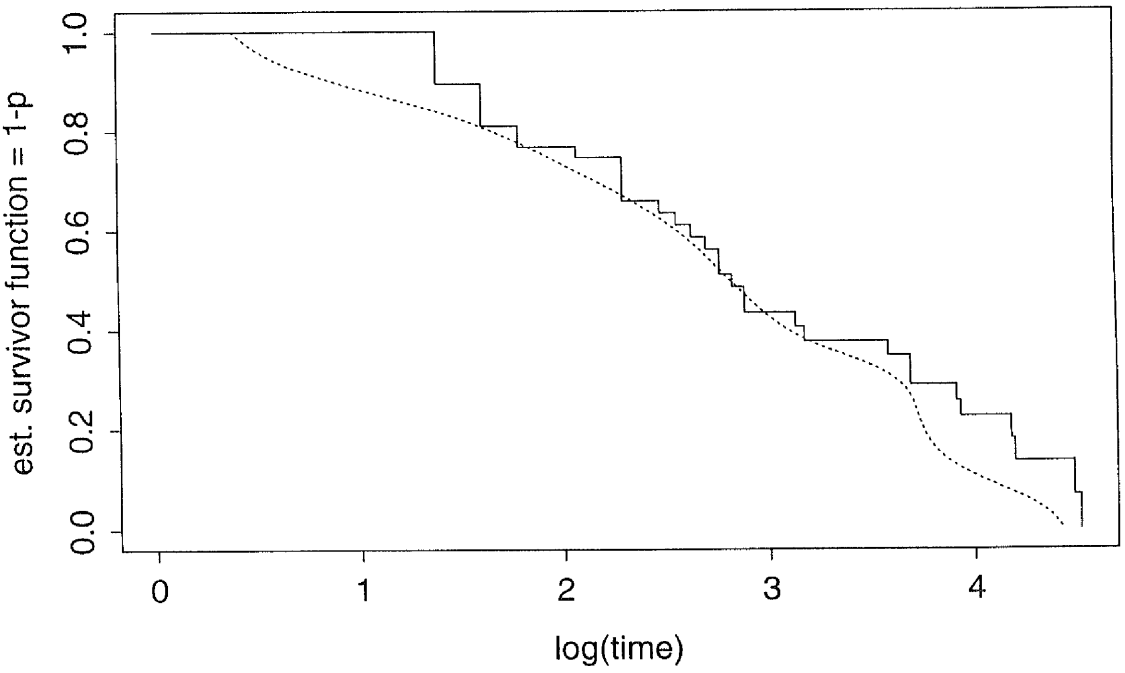



Figure 3.6.6 : Survivor curves at age=70




h1=3 , h2=0.075

patient "fails" (i.e. experiences PGD), they are recorded as being censored at the end of that year with their measured MBP. Each measurement is included in the PH analysis and so most individuals will contribute more than one observation. The scatterplot in Figure 3.7.1 displays the data as it would be entered in the model. In section 6.3, particular attention was paid to the effect of MBP on the occurrence of PGD in the second year following transplanation. It is this subset of the data which is now examined.

A key feature of the data here is the amount of censoring. A large proportion of patients at the Renal Unit don't experience PGD, so this, coupled with the added censored observations for the time–dependent analysis, results in only 5% of the data being regarded as failure times. To take account of this when producing the percentile curves, a much smaller smoothing parameter in the quantile space is required and the $5^{th}$ percentile is calculated. The doubly–smooth ($h_1$=7 , $h_2$=0.02) and proportional hazards curves are shown in Figure 3.7.2 as full and broken lines respectively. The two curves agree in that both show a decrease in survival time as MBP increases. The doubly–smooth curve shows a fairly sharp decrease until around MBP=110, then the curve levels out. The PH curve, however, shows a fairly steep decrease in survival time between 100 and 120.

The survivor functions, and their smooth versions, are plotted for MBP = 100 and 120. Not surprisingly, the curves don't even reach 0.8. It is not sensible for the estimator to approximate smaller percentiles than this, and has been constructed not to evaluate quantiles beyond two bandwidth's distance of the end time point of the weighted Kaplan–Meier.

The percentile curves in Figure 3.7.2 proved useful in aiding the understanding of the mean blood pressure effect although the high level of censoring in the data makes it difficult to identify its exact nature.

# Renal Unit data

Figure 3.7.1 : Scatter plot

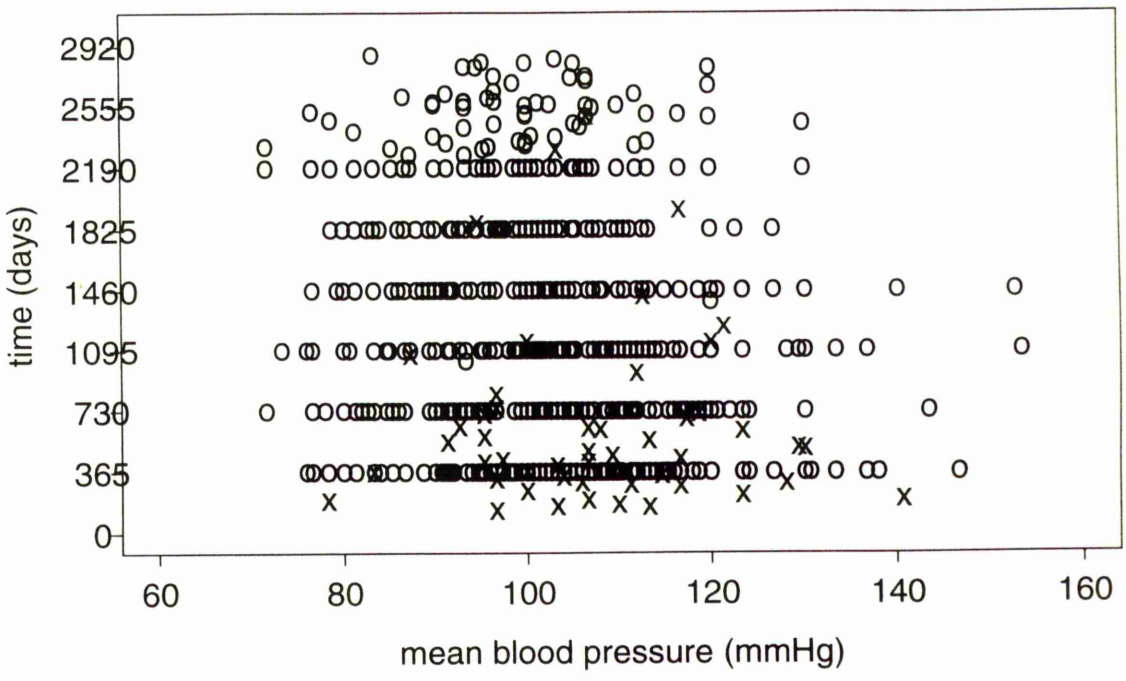
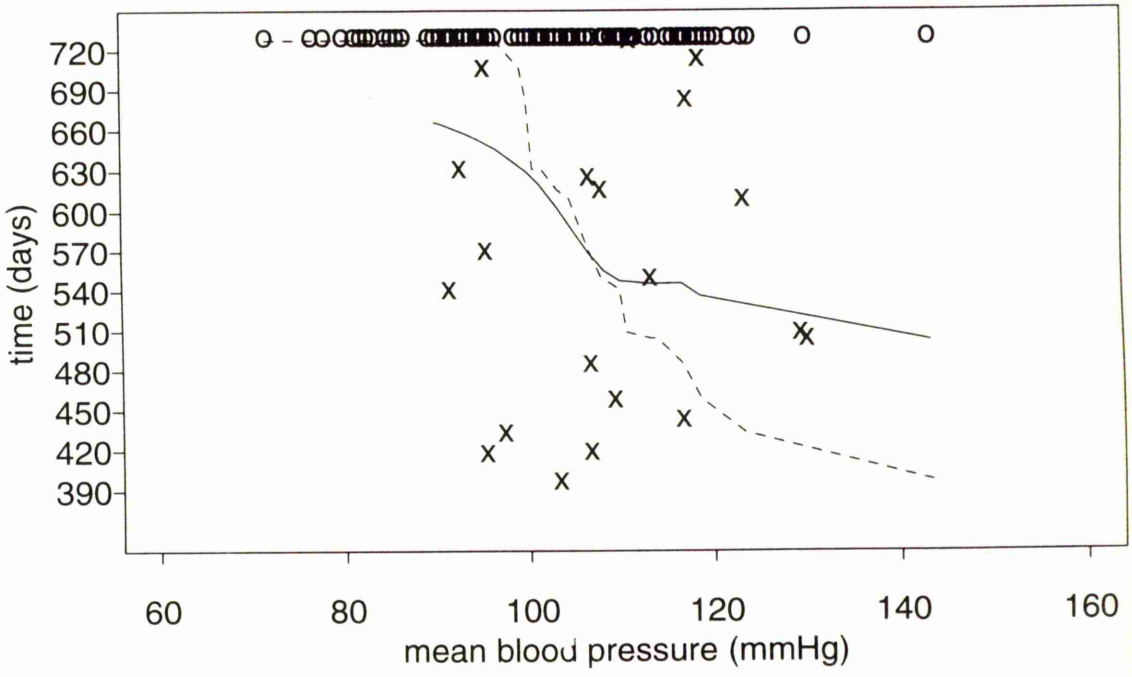



Figure 3.7.2 : 5th percentile curve




h1=7, h2=0.02

# Renal Unit data

Figure 3.7.3 : Survivor curves at covariate=100

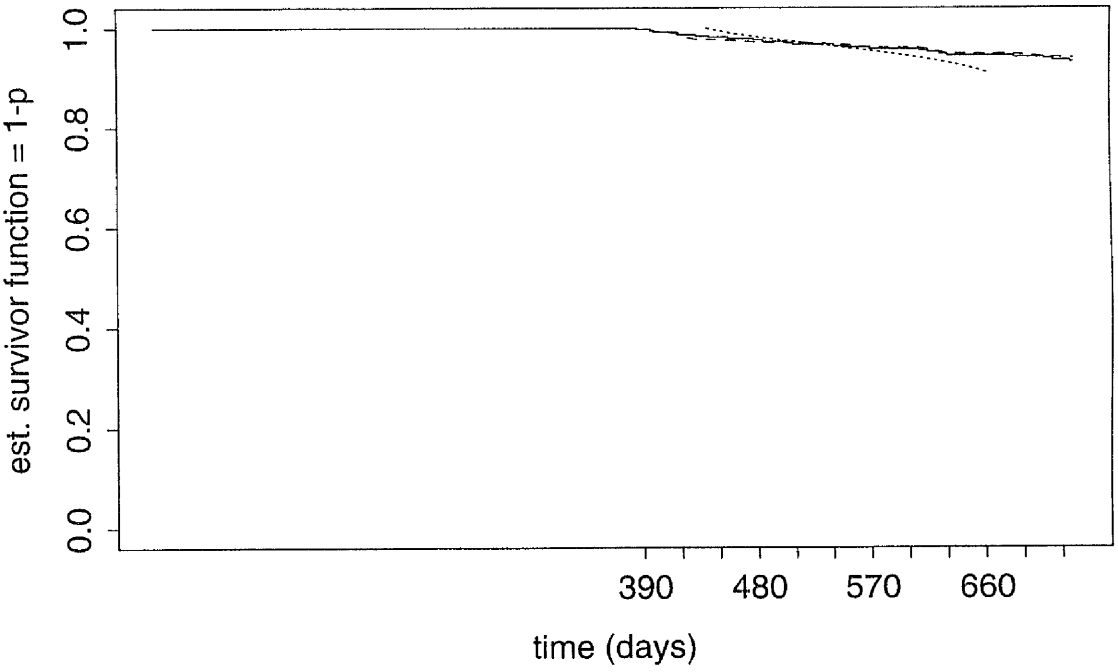
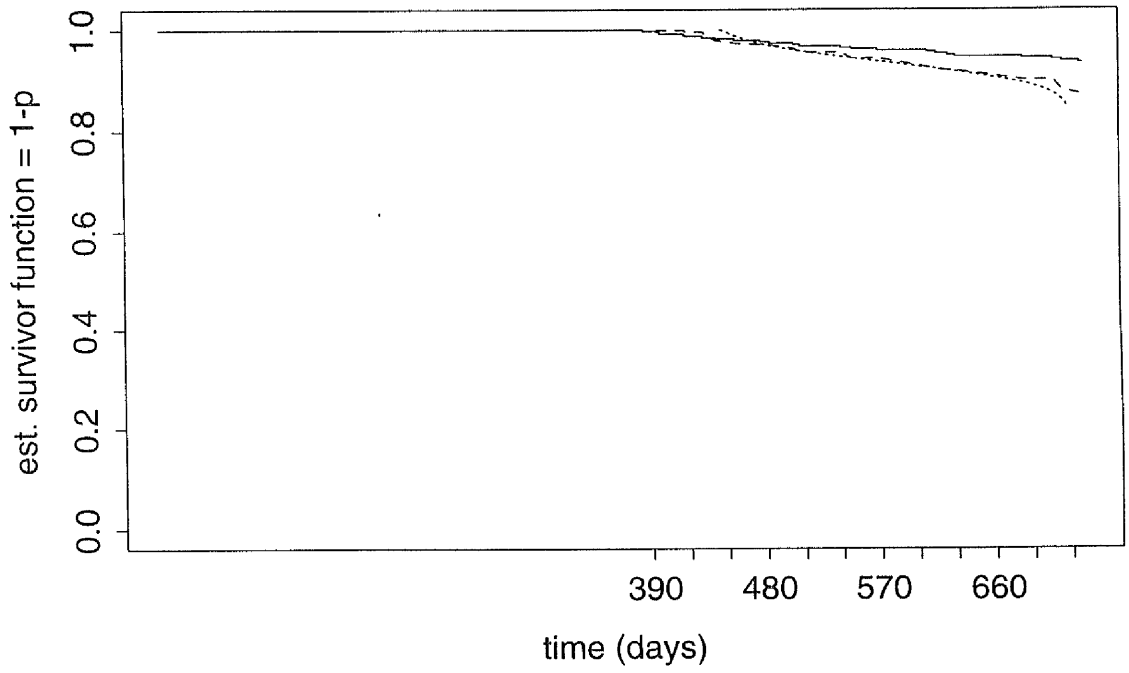



Figure 3.7.4 : Survivor curves at covariate=120




h1=7, h2=0.02

### 3.2.8 Example 8 – Cerebral Palsy Growth Data

While the doubly–smooth quantile estimator has been tailored for censored data, it can be applied in the more usual regression setting by simply regarding all of the responses as "failures". Linear growth charts are one such area of application and were discussed briefly in Chapter 2. At the Kennedy Krieger Institute in Baltimore, data have been collected on the development of children who have cerebral palsy. Interest lies in the growth, particularly the height and weight, of children suffering from this condition and how they compare to normal children of similar age. Full details of the analysis of this data are given in Chapter 7.

Here we consider the relationship between height and age for boys and use it as an example to exhibit the comparison of kernel and spline quantile approaches. (The estimation of quantiles using cubic splines is also discussed in more detail in Chapter 7.)

The linear growth charts for normal children, commonly used as national standards, were contructed by Hamill *et al.* (1977). Measurements on over 20,000 individuals were used to derive curves for a large range of quantile values. For a given variable of interest, for example height, the data were divided into very narrow intervals according to age and the empirical quantile values calculated for each interval. These values were then smoothed using a natural cubic spline with 5 degrees of freedom to produce the necessary curves for each quantile.

A scatterplot of the data is given in Figure 3.8.1. (Note here, the $y$–axis represents the height measurements and the $x$–axis the age in months which correspond to survival time and covariate respectively in previous examples.) While we obviously have less information on cerebral palsy children than in the sample used to produce the percentile curves for normal children, there is still a substantial amount of data. A large number of the height values were recorded on boys under 48 months and the variability of these measurements increases with age. To produce a more even spread of values and remove as much linear trend as possible, the residuals from the least squares estimation fit of height vs. the

square root of age were analysed.

The resulting $10^{th}$, $50^{th}$ and $90^{th}$ percentile curves, transformed back to their original scale, are shown in Figure 3.8.2 as broken lines. To allow a fair comparison, the data were given the same transformation before applying the doubly–smooth estimator, producing the curves represented as full lines in Figure 3.8.2. There is little evidence of differences between the two types of curve, although the doubly–smooth percentile curves are slightly less steep for ages greater than 168 months but this also corresponds to a region where there is less data.

### 3.2.9 Conclusions from Examples

Cox's Proportional Hazards Model assumes that, in the absence of a covariate, there exists a baseline rate of survival. Choosing the effect of the covariate to be linear forces the percentile curves under this model to be monotonic.

The doubly–smooth quantile estimator assumes neither of these and is a much more data–driven approach. In fact, different percentiles from this method can have very different shapes depending on the data.

The smooth PH model provides a combination of these, with the assumption of a baseline hazard rate and the smooth covariate effect removing the monotonicity assumption.

As has been shown in the first seven examples, the percentile curves from these methods are usually very similar. Disagreements tend to occur when there are smaller numbers of data points, especially at the extreme covariate values. Here the heavy influence of the data on the nonparametric approach becomes more obvious. In some instances, the monotonicity of the PH model will be inapproriate and the curves from the doubly–smooth estimator and the smooth PH model will provide a more accurate description of the data.

The smoothing parameter values were chosen subjectively. As discussed in Chapter 2 and at the beginning of Chapter 3, the choice of the parameter in the covariate space ($h_1$) is generally more important to the shape of the resulting

# Cerebral Palsy Growth Data

Figure 3.8.1 : Scatter plot

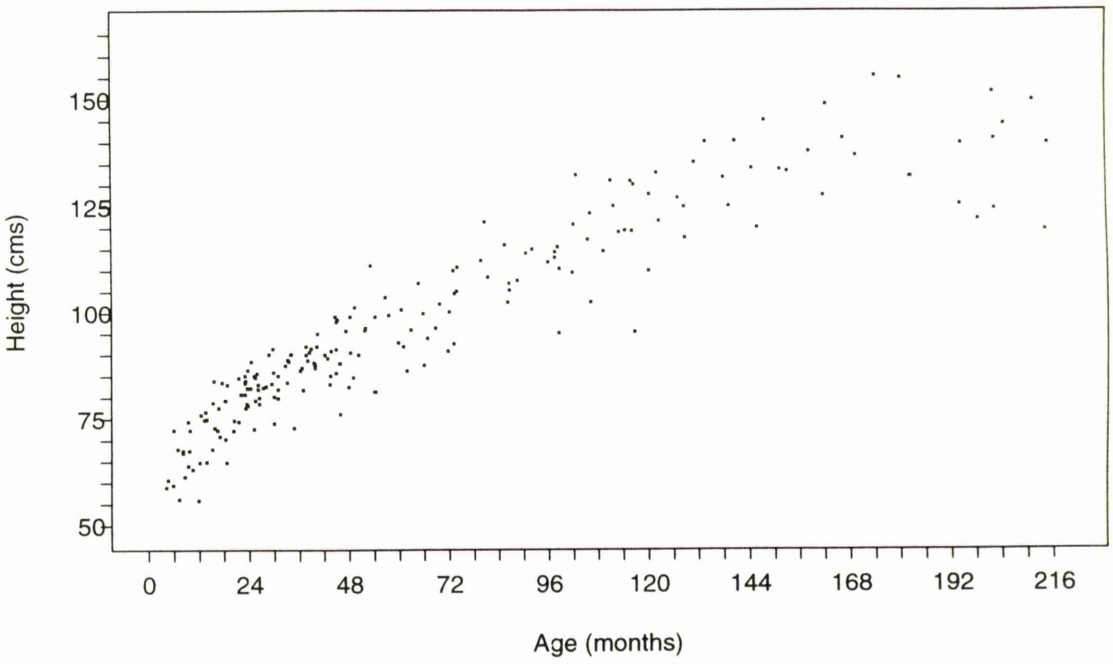
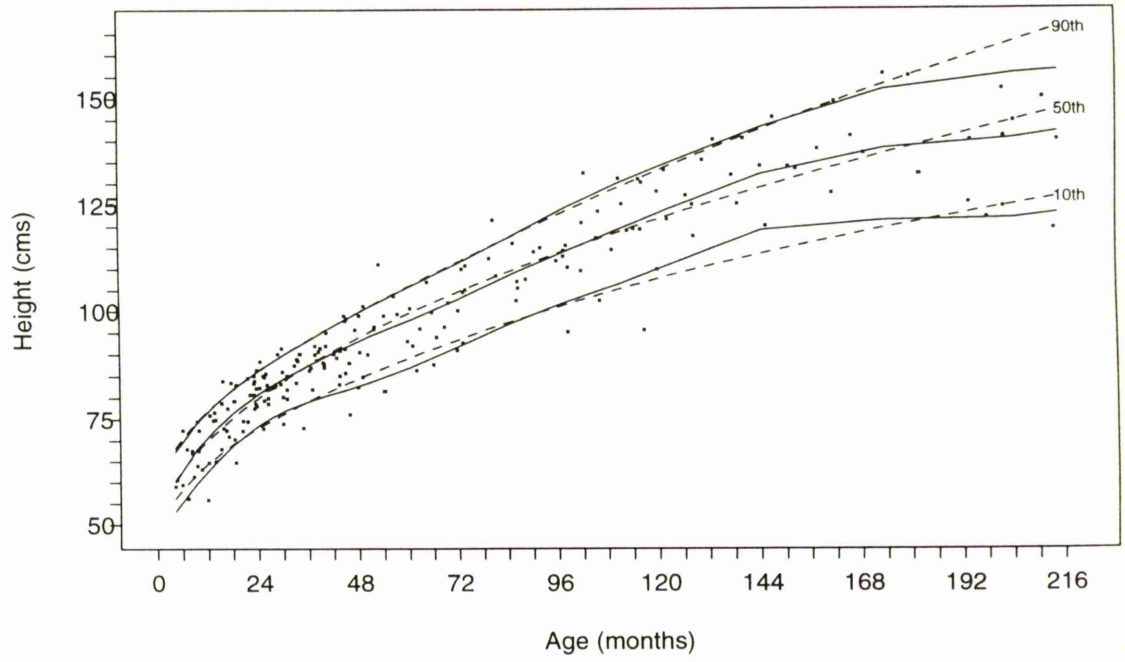



Figure 3.8.2 : 10th, 50th and 90th percentile curves




h1=1, h2=0.075

percentile curve than the parameter in the quantile space ($h_2$). However, in cases where there is a large proportion of censoring, particularly if the higher survival times are censored, then the choice of $h_2$ becomes more important and, in general, we would use a relatively small smoothing parameter. This is primarily because the survivor function may not reach zero and so estimation is over a shorter range than zero to one.

The issue of heavy censoring was also illustrated. In addition to lowering the smoothing parameter $h_2$, smaller percentiles should be plotted, as information for higher values of $p$ can become severely biased.

Finally, two nonparametric approaches were compared in an example where there was no censoring. When calculated under similar circumstances, there was very little difference between the quantile curves estimated from the cubic spline and doubly–smooth approaches.

## 3.3 Choice of smoothing parameter

The choice of smoothing parameter in smoothing problems is a topic on which much literature has been written in a variety of contexts, particularly in the case of kernel density estimation. Many of the basic ideas have been developed and extended to the nonparametric regression setting. Marron (1988) provides a good discussion and comparison of smoothing parameter selection methods, although many more articles have been written on the subject since then.

In some cases, a form of cross–validation procedure is employed, often a least–squares approach as in Bowman (1984) for density estimation and Härdle and Marron (1985) for regression. The criteria for selection purposes is usually the mean square error (MSE), the integrated square error (ISE) or the mean integrated square error (MISE). (Grégoire (1993) lists many references for this.) Various papers discuss the asymptotic rates of convergence for these quantities. However, it is now generally accepted that the best selection methods are those driven by "plug–in" estimators. The choice of smoothing parameter is made using an iterative procedure with a criteria chosen to have good convergence properties. The kernel density estimation case is discussed by Sheather and Jones (1991). Gasser *et al.* (1991) propose the corresponding method appropriate for one form of kernel regression.

Patil (1993) and Grégoire (1993) both recommended using a least–squares cross–validation procedure when choosing a smoothing parameter for the hazard rate estimated using kernel functions. The quantity to measure the amount of error here was the integrated square error. The ideas presented in Tanner and Wong (1984) were very similar, although the proposed selection criteria was a modified likelihood statistic. While this is a different setting to that of quantile estimators, it would appear that the same ideas can be used in the censored data case as in that where no censoring exists.

In the context of kernel quantile estimators, Sheather and Marron (1990) derived an expression for an optimal smoothing parameter based on the min-

imisation of the mean squared error. Padgett and Thombs (1986) discussed the censored data version of this estimator and selected the smoothing parameter by minimising the asyptotic mean squared error for bootstrap samples.

A feature of the doubly–smooth quantile estimator is that two smoothing parameters must be chosen. Boularan *et al.* (1994) discussed a two–stage smoothing procedure for a nonparametric mixed effects model. The smoothing parameters used here were estimated using cross–validation, with MISE as the measure of error, and the second parameter being conditional on the first. While this is an additive model and so the structure is quite different to that of the doubly–smooth quantile estimator, the ideas may still give some clues as to the approach required.

Within the notation of section 3.1 and referring to the discussion of the various examples in section 3.2, importance should be placed on the selection of $h_1$, the parameter which controls the smoothing across the covariate. It is essentially this value which determines the shape of the resulting quantile curve, whereas $h_2$, the smoothing parameter across the quantile space, produces additional smoothing to make the curve more aesthetically pleasing, as well as providing a simple method of calculation of the estimated quantile.

With this in mind, one possible, relatively simple method for choosing the smoothing parameter is now explored. Sometimes data are simulated from a model considered to be true to gain information about a similar, but perhaps less well-defined, approach. In the case of choosing smoothing parameter a PH assumption provides a useful reference model for a set of data. The median under this model can be calculated to provide a set of data points $(z_i, m_i)$ which describe the curve, where $m_i$ are the values of the fitted PH median at the covariate values $z_i$.

Now data are simulated from this fitted PH model and for a grid of values of $h_1$, the doubly–smooth median curve $(\widehat{m}_i)$ is calculated. (The simulation procedure is similar to a bootstrapping method described later in section 4.2.3.) The distance between each curve from the simulated data and the original fitted me-

dian can be measured by $d_i = \sum_i (m_i - \widehat{m}_i)^2$.

These values are then averaged over the number of simulations ($N$), $\overline{d} = \sum d_i/N$, as an estimate of the expected distance between the curves from the two methods,

$$E \int (m(z) - \widehat{m}(z))^2 f(z) dz$$

Then $h_1$ is chosen to be the value which minimises this average.

This procedure was carried out for 100 simulations for each of three data-sets and Figures 3.9.1–3 are plots of $\overline{d}$ vs. $h_1$. (To ease interpretation, a scatterplot smoother using Normal kernel weights was added to each plot.) The results can be summarised as follows.

- Simulated PH data – The criteria $\overline{d}$ is minimised when $h_1$ is about 0.25. However, there would appear to be little difference for values of the smoothing parameter between 0.15 and 0.275.
- The simulated "no effect" data – When $\beta$ for the fitted PH model is equal to, or close to, zero then $\overline{d}$ will not have a unique minimum but will asymptote for large smoothing parameters. For this data, a value of 1.5 appears to be a reasonable choice on the basis of the $\overline{d}$ criteria.
- Stanford Heart Transplant Data – The values of $\overline{d}$ appear to be minimised when $h_1$ is approximately 10, although choosing a value of 6 would be equally acceptable.

This method of choosing $h_1$ does not have as much of a theoretically sound basis as some of the previously mentioned approaches. However, it is a sensible technique and appears to produce reasonable results. Analogous arguments apply in the case of density estimation, where a "rule of thumb" method is to use a smoothing parameter which is optimal for a Normal distribution (Silverman, 1986). This approach seems to work surprisingly well (Bowman, 1985). A practical method of smoothing parameter choice based on the theoretical results given above has yet to be developed.

# Choice of Smoothing Parameter

Figure 3.9.1 : Simulated PH data

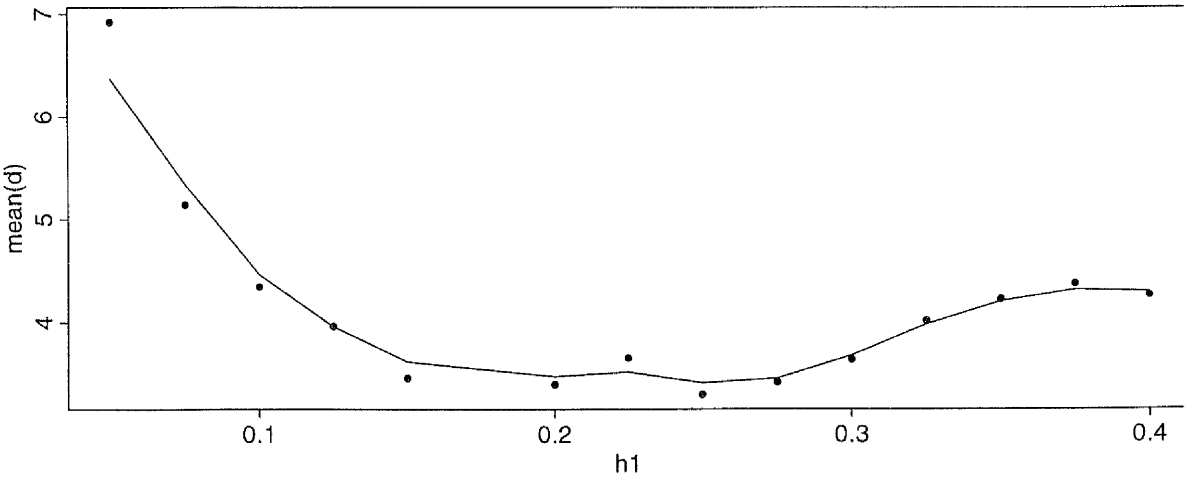



Figure 3.9.2 : Simulated data with no effect

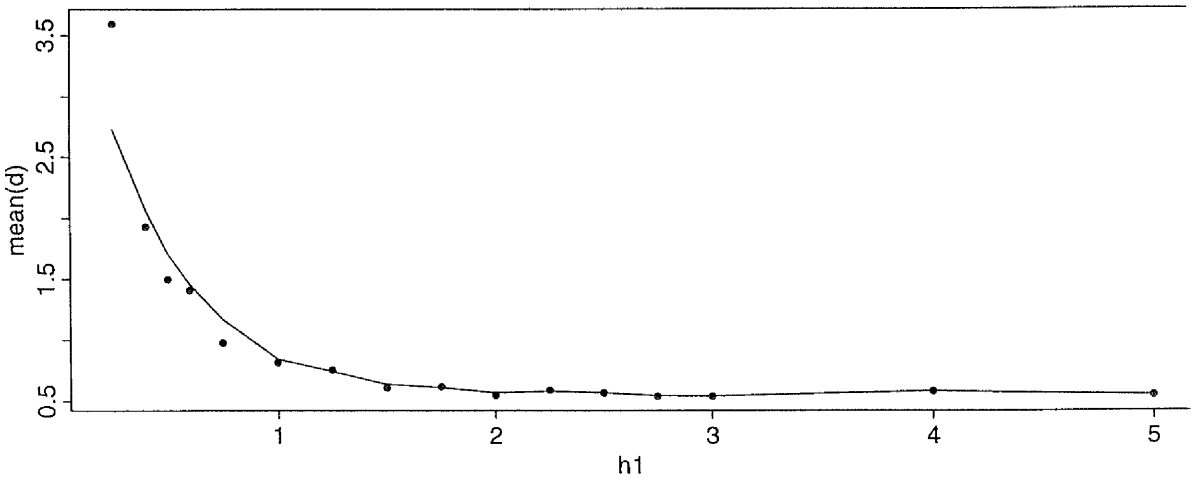
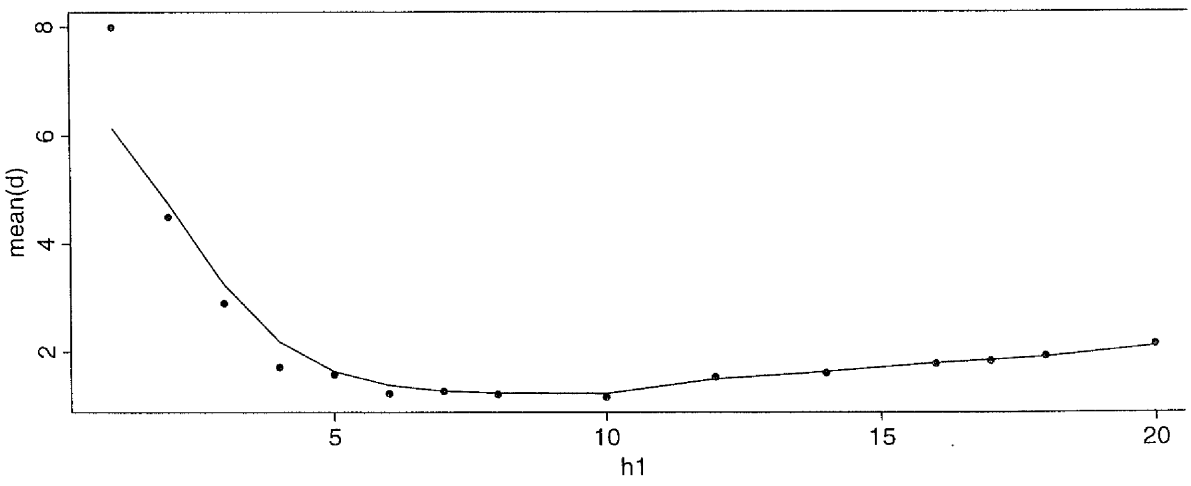



Figure 3.9.3 : Stanford Heart Transplant data



## 3.4 A Likelihood Statistic for Inference

In some situations, the role of the doubly–smooth percentile curve may be to validate the information from the proportional hazards model. Other problems may simply require a subjective impression for the effect of a covariate on survival times with as few constraints as possible. However, it will often be helpful for this to be accompanied by a formal test, based on a quantity which measures the characteristic of interest.

The shape of the centiles curves is determined by the corresponding estimated survivor function. The second smoothing procedure in the doubly–smooth quantile estimator is merely a tool used to ease graphical interpretation. Therefore attention should be focussed on the behaviour of each survivor function across the range of covariate values. The aim is to provide a statistic which will enable distinctions between different survivor functions to be identified.

Likelihood is a fundamental quantity in the field of statistics. Often the aim is to find values of the parameters which maximise the probablility of an event across the range of data, e.g. the partial likelihood under Cox's PH model is maximised to find the "best estimate" of the coefficient, $\beta$. The value of the maximised likelihood is also a basic measure of fit of a model, which can be used in formal testing procedures.

The full likelihood for a set of survival data with a continuous covariate (Kalbfleisch and Prentice, 1980) can be written as

$$\mathcal{L} = \prod_{i=1}^{k} \{ \prod_{l \in D_i} [S(t_{(i)}|z_l) - S(t_{(i)} + 0|z_l)] \prod_{l \in C_i} S(t_l|z_l) \}$$

where $D_i$ is the set of labels for those failing at $t_{(i)}$ and $C_i$ denotes those which are censored in the interval $[t_{(i)}, t_{(i+1)})$.

As survival time increases, survivor functions monotonically decrease over the range [0,1]. Each estimated survivor function can be described by its hazard rate, or by its gradient which is calculated using the actual values of the failure times.

The survivor function uses the ranking order of the times (not the lengths of survival time itself), implying that the gradient is a more appropriate measurement for the failure time. However, since the only information available on censored observations is that their failure time exceeds that of the recorded survival time, the contribution to a statistic should be of a form similar to that in $\mathcal{L}$. There are standard expressions for estimating the hazard rate under the Proportional Hazards model but this is not true for the nonparametric method which uses the weighted Kaplan–Meier. Therefore, it is more appropriate to consider an expression which can be easily extended to any estimated survivor function.

The calculation of a likelihood statistic, $L$, is performed by considering each failure time to contribute the gradient of the esimated survivor function at that time, given its covariate value. The corresponding terms for censored times are the values of the estimated survivor function itself, given its covariate value.

Azzalini (1981) derived an expression for an estimate of the distribution function using kernels which can be extended to the survivor function and written as

$$\tilde{S}(t|z) = \frac{1}{n}\sum_{l\in D} j_l(z) K^*\left(\frac{t-t_l}{h_v}\right)$$

where $h_v$ is the bandwidth and $K^*$ is the kernel density function with $K^*(s) = \int_s^{-\infty} K(y)dy$, in the same notation as chapter 2.

To find an estimate of the gradient from this, we differentiate with respect to $t$ to obtain

$$\frac{d}{dt}\tilde{S}(t|z) = -\frac{1}{n}\sum_{l\in D} j_l(z)\frac{1}{h_v}K\left(\frac{t-t_l}{h_v}\right)$$

The likelihood statistic can now be written as

$$L = \prod_{i=1}^{k}\left\{\prod_{l\in D_i} -\frac{d}{dt}\tilde{S}(t_{(i)}|z_l)\prod_{l\in C_i}\tilde{S}(t_l|z_l)\right\}$$

where $t_{(i)}$ , $i = 1,\ldots,k$ denotes the unique ordered failure times, and $D_i$ and $C_i$ are as before.

This quantity can then be constructed for the survivor functions from each of the three models in exactly the same way. It is common practice to present

the log–likelihood ($ll$), since calculating the likelihood will produce very small values, the mathematical expressions involved are more tractable under the log tranformation, and in parametric cases appropriate distributions can be derived.

While this procedure introduces another smoothing parameter, the effect of different values of $h_v$ on the conclusions will be diminished by using the same value of $h_v$ in models to be compared. This effect is studied below.

The following section presents the results of the examples analysed.

### 3.4.1 Results of Calculations of Likelihood statistic

The aim of the following information is to illustrate the effect of modelling data by each of the three methods and to what extent the statistic changes with different levels of smoothing.

For each of the first seven example data–sets, explored graphically in the last section, the log–likelihood ($ll$) statistic is calculated under each of the three approaches. The smooth PH (SPH) and nonparametric (NP) models require an amount of smoothing to be chosen. Tables of the log–likelihood statistics are given for a wide range of values, from those representing extremely rough to over-smoothed estimates. Also of interest is the difference between pairs of these statistics. So for each of the smooth PH and nonparametric log–likelihoods, the value under the PH model is subtracted as a point estimate description of the difference in fit.

Considering the parameterisation of each of the models and the constraints which these impose, it would be logical to expect, for a given data–set, that the log–likelihood statistic under the PH model would be less than that of the smooth PH model, which in turn would be less (for equivalent amounts of smoothing) than the value under the NP model.

The smoothing parameters of the SPH and NP models control the amount of weight observations near a given observation can have. Therefore, small values of $h1$ and $k$ would result in only points which lay very close to the point of

estimation having some influence on the estimate. Equally, large values of $h1$ and $k$ give more equal weightings to a larger spread of observations at any given point. (In fact when $k$=1, we would expect the value of the SPH ll to be very close to that of PH ll, although not equal to as $s(z_o)$ is constrained to be zero.) With this in mind, as $h1$ and $k$ increase, it is logical to expect the values of NPll and SPHll to decrease.

Let $D_1$ = PH$ll$ - SPH$ll$ and $D_2$ = PH$ll$ - NP $ll$

**Simulated "No Effect" Data – 20% censoring**

For $h_v$=1.0,

PH ll = -175.49

SPH ll = -174.66 for k=0.25

NP ll = -168.36 for h1=0.7

As expected, a slight increase is obtained in the log–likelihood statistics when the smooth version of the PH model is used over the standard PH model. The values of SPHll is very low because the value of $k$ is so small that the local likelihood estimates are based on very small samples, producing instability in the likelihood statistic. The nonparametric approach produces even larger values for the statistic. However, for "sensible" values of the smoothing parameters in Table 3.1, there appears to be little difference between any of the models.

**Simulated "No Effect" Data – 50% censoring**

For $h_v$=1.0

PH ll = -128.75

SPH ll = -128.48 for k=0.25

NP ll = -126.22 for h1=0.7

Increasing the amount of censoring produces larger values of the log–likelihood statistic but the differences, $D_1$ and $D_2$, in Table 3.2 still suggest little discrep-

| $k$ | SPHll | D1 | $h1$ | NPll | D2 |
|---|---|---|---|---|---|
| 0.050 | -212.45 | 36.96 | 0.10 | -139.24 | -36.25 |
| 0.100 | -177.14 | 1.65 | 0.25 | -157.27 | -18.22 |
| 0.125 | -174.05 | -1.44 | 0.50 | -165.69 | -9.80 |
| 0.200 | -174.38 | -1.11 | 0.75 | -168.81 | -6.68 |
| 0.250 | -174.66 | -0.83 | 1.00 | -170.40 | -5.09 |
| 0.333 | -174.88 | -0.61 | 1.50 | -171.99 | -3.50 |
| 0.500 | -174.85 | -0.64 | 2.00 | -172.94 | -2.55 |
| 1.000 | -175.25 | -0.24 | 5.00 | -174.78 | -0.71 |

Table 3.1: Results for Simulated "No Effect" Data – 20% censoring

| $k$ | SPHll | D1 | $h1$ | NPll | D2 |
|---|---|---|---|---|---|
| 0.050 | -140.93 | 12.18 | 0.10 | -98.15 | -30.6 |
| 0.100 | -128.33 | -0.42 | 0.25 | -114.97 | -13.78 |
| 0.125 | -128.28 | -0.47 | 0.50 | -124.10 | -4.65 |
| 0.200 | -128.67 | -0.08 | 0.75 | -126.53 | -2.22 |
| 0.250 | -128.47 | -0.28 | 1.00 | -127.56 | -1.19 |
| 0.333 | -129.01 | 0.26 | 1.50 | -128.52 | -0.23 |
| 0.500 | -129.60 | 0.85 | 2.00 | -129.02 | 0.27 |
| 1.000 | -130.14 | 1.39 | 5.00 | -129.89 | 1.14 |

Table 3.2: Results for Simulated "No Effect" Data – 50% censoring

ancy between the models.

**Simulated "Quadratic" Data**

For $h_v = 1.5$

PH ll = -221.48

SPH ll = -198.32 for k=0.25

NP ll = -194.39 for h1=0.5

The failure of the PH model to detect a relationship between survival time and a covariate which was non–monotonic was examined in the previous section as Example 2. It is important that the corresponding *ll* statistics also reflect this. From Table 3.3, the value of the *ll* statistic under the PH model is very much lower than those observed from the SPH and NP models, except for extremely large smoothing levels, as would be expected. The *ll* statistics from the SPH and NP models appear to be in close agreement, validating the subjective impression from Figure 3.2.2.

**Simulated "PH" Data**

For $h_v = 0.5$

PH ll = -120.63

SPH ll = -118.85 for k=0.25

NP ll = -115.40 for h1=0.1

The results of the statistics calculated using the data simulated from a Proportional Hazards model in Table 3.4 reflect the same conclusions as those for Tables 3.1 and 3.2, that there is little discrepancy between any of the models.

**Non-PH Simulated Data**

For $h_v = 2$

PH ll = -365.52

| $k$ | SPHll | D1 | $h1$ | NPll | D2 |
|---|---|---|---|---|---|
| 0.050 | -196.33 | -25.15 | 0.10 | -164.62 | -56.86 |
| 0.100 | -196.04 | -25.44 | 0.25 | -185.26 | -36.22 |
| 0.125 | -194.46 | -25.02 | 0.50 | -194.39 | -27.09 |
| 0.200 | -198.32 | -23.16 | 0.75 | -200.34 | -21.14 |
| 0.250 | -199.23 | -22.25 | 1.00 | -205.44 | -16.04 |
| 0.333 | -200.96 | -20.52 | 1.50 | -213.13 | -8.35 |
| 0.500 | -208.63 | -12.85 | 2.00 | -217.51 | -3.97 |
| 1.000 | -221.42 | -0.06 | 5.00 | -221.96 | 0.48 |

Table 3.3: Results for Simulated "Quadratic" Data

| $k$ | SPHll | D1 | $h1$ | NPll | D2 |
|---|---|---|---|---|---|
| 0.050 | -304.23 | 183.60 | 0.010 | -76.31 | -44.32 |
| 0.100 | -117.63 | -2.90 | 0.025 | -97.07 | -23.56 |
| 0.125 | -118.01 | -2.62 | 0.050 | -108.91 | -11.72 |
| 0.143 | -117.80 | -2.83 | 0.075 | -113.27 | -7.36 |
| 0.200 | -117.99 | -2.64 | 0.100 | -115.40 | -5.23 |
| 0.250 | -118.85 | -1.78 | 0.125 | -116.64 | -3.99 |
| 0.333 | -119.03 | -1.60 | 0.200 | -118.84 | -1.79 |
| 0.500 | -119.27 | -1.36 | 0.400 | -122.59 | 1.96 |
| 1.000 | -120.64 | 0.01 | 1.000 | -125.63 | 5.00 |

Table 3.4: Results for Simulated "PH" Data

SPH ll = -364.28 for k=0.25

NP ll = -346.08 for h1=0.4

As shown in Table 3.5, the values for the two PH models are very similar for reasonable values of $k$. However, the NPll values appear to be different, confirming the impression from Figures 4.2-4.

**Stanford Heart Transplant Data**

For $h_v = 0.3$

PH ll = -198.35

SPH ll = -189.56 for k=0.25

NP ll = -181.39 for h1=5

The values of $D_1$ and $D_2$ in Table 3.6 show differences in the $ll$ statistics between the smooth models and the PH model. While these values are not as large as those found in Table 3.3, it would appear that the resulting percentile curves under each of the methods may exhibit slightly different characteristics.

**Multiple Myeloma Data**

For $h_v = 5$

PH ll = -102.94

SPH ll = -102.14 for k=0.25

NP ll = -101.99 for h1=3

As shown in Table 3.7, the differences observed in this data–set are relatively small. This may be partially due to the small sample size, although the results reveal the same conclusions as Figures 3.6.1–3.

**Renal Unit Data**

For $h_v$ =30

PH ll = -234.41

| $k$ | SPHll | D1 | $h1$ | NPll | D2 |
|---|---|---|---|---|---|
| 0.050 | -378.95 | 13.43 | 0.10 | -311.50 | -54.02 |
| 0.100 | -362.96 | -2.56 | 0.20 | -332.80 | -32.72 |
| 0.125 | -362.76 | -2.76 | 0.40 | -346.08 | -19.44 |
| 0.200 | -363.57 | -1.95 | 0.60 | -351.96 | -13.56 |
| 0.250 | -364.28 | -1.24 | 0.70 | -353.93 | -11.49 |
| 0.333 | -365.08 | -0.44 | 0.80 | -355.49 | -10.03 |
| 0.500 | -364.87 | -0.65 | 1.00 | -357.72 | -7.80 |
| 1.000 | -364.84 | -0.68 | 2.00 | -362.00 | -3.52 |

Table 3.5: Results for Non-PH Simulated Data

| $k$ | SPHll | D1 | $h1$ | NPll | D2 |
|---|---|---|---|---|---|
| 0.050 | -187.55 | -10.80 | 1 | -141.42 | -56.93 |
| 0.071 | -187.42 | -10.93 | 2 | -161.30 | -37.05 |
| 0.083 | -188.14 | -10.21 | 3 | -171.35 | -27.00 |
| 0.100 | -188.46 | -9.89 | 4 | -177.55 | -20.80 |
| 0.125 | -188.26 | -10.09 | 5 | -181.39 | -16.96 |
| 0.200 | -189.11 | -9.24 | 6 | -183.82 | -14.53 |
| 0.250 | -189.56 | -8.79 | 7 | -185.46 | -12.89 |
| 0.333 | -190.40 | -7.96 | 8 | -186.66 | -11.69 |
| 0.500 | -191.62 | -7.09 | 10 | -188.43 | -9.92 |
| 1.000 | -192.77 | -5.58 | 20 | -193.42 | -4.93 |

Table 3.6: Results for Stanford Heart Transplant Data

| $k$ | SPHll | D1 | $h1$ | NPll | D2 |
|---|---|---|---|---|---|
| 0.125 | -102.30 | -0.64 | 0.5 | -97.14 | -5.80 |
| 0.143 | -102.30 | -0.64 | 1.0 | -99.87 | -3.07 |
| 0.167 | -102.54 | -0.40 | 2.0 | -101.40 | -1.54 |
| 0.200 | -102.40 | -0.54 | 3.0 | -101.99 | -1.04 |
| 0.250 | -102.14 | -0.80 | 4.0 | -102.29 | -0.65 |
| 0.333 | -102.30 | -0.64 | 5.0 | -102.48 | -0.46 |
| 0.500 | -102.65 | -0.29 | 7.0 | -102.71 | -0.23 |
| 1.000 | -103.39 | 0.45 | 10.0 | -102.92 | -0.02 |

Table 3.7: Results for Multiple Myeloma Data

SPH ll = -225.55 for k=0.25

NP ll = -219.00 for h1=5

In both sets of differences in Table 3.8, $D1$ and $D2$, the choice of smoothing parameter has little effect on the values of the likelihood statistic. It would appear that each of the approaches is slightly different to the others although the SPH and NP statistics are closest.

**Remark**

Various values of $h_v$ were considered for each data–set but the results of the log–likelihood ratios were found to alter only very slightly. The results for the data simulated to have "no effect" with 20 % censoring are shown in Table 3.9 ($k = \frac{1}{4}, h_1 = 0.7$). This suggests that the choice of $h_v$ makes little difference to any conclusions provided the same value is used for each of the three statistics. However, a "sensible" subjective value would be advisable.

### 3.4.2 Conclusions

Comparisons of the above log–likelihood statistics support most of the subjective impressions gained from the plots in the previous section.

Within a range of what could be termed "reasonable" smoothing parameters, the effect of a smaller or larger parameter on the test statistics appears to make little difference in most cases. However, choosing a value to represent the degree of smoothing is not trivial and it is important to give various values careful consideration to enable a good subjective choice.

The standard results of 2 × log–likelihood statistic being approximately chi-squared do not hold here since the parameter values have not been chosen to maximise this quantity. Further, the complexity of the theory surrounding the nonparametric method creates difficulties in obtaining an inferential test. An alternative approach is to consider a re–sampling technique such as permutation

tests or bootstrapping. While these can be computationally intensive, they provide a relatively easy way of providing a more objective conclusion. The following chapter discusses the different aims and implementations of these methods.

| $k$ | SPHll | D1 | $h_1$ | NPll | D2 |
|---|---|---|---|---|---|
| 0.050 | -226.18 | -8.23 | 1 | -205.88 | -28.53 |
| 0.100 | -225.70 | -8.71 | 3 | -215.99 | -18.42 |
| 0.125 | -225.65 | -8.76 | 4 | -217.81 | -16.60 |
| 0.200 | -225.47 | -8.94 | 5 | -219.00 | -15.41 |
| 0.250 | -225.55 | -8.86 | 7 | -220.87 | -13.54 |
| 0.333 | -225.58 | -8.83 | 8 | -221.53 | -12.88 |
| 0.500 | -225.59 | -8.82 | 10 | -222.40 | -12.01 |
| 1.000 | -225.61 | -8.80 | 20 | -224.27 | -10.14 |

Table 3.8: Results for Renal Unit Data

| $h_v$ | PHll | SPHll | D1 | NPll | D2 |
|---|---|---|---|---|---|
| 0.10 | -287.11 | -285.73 | -1.38 | -279.69 | -7.42 |
| 0.50 | -176.23 | -174.87 | -1.36 | -163.85 | -12.38 |
| 0.75 | -175.24 | -174.01 | -1.23 | -165.99 | -9.25 |
| 1.00 | -175.49 | -174.66 | -0.83 | -168.36 | -7.13 |
| 1.50 | -178.95 | -178.00 | 0.95 | -173.74 | -5.21 |
| 1.75 | -180.08 | -179.89 | -0.19 | -171.31 | -8.77 |
| 2.00 | -182.65 | -179.84 | -2.81 | -171.85 | -10.80 |

Table 3.9: Comparisons of $h_v$ values

# Chapter 4

# Bootstrapping Methods

## 4.1 Introduction, Aims and Permutation Tests

As discussed in Section 1.4, one of the main aims of this research is to consider which of the four following descriptions is most apt for a set of data, with the corresponding quantile curve method in brackets. These are also formulated below in terms of the hazard function and "linear" here refers to the hazard function on the log scale.

A) The covariate has *no effect* on survival time. $\lambda(t;z)$ does not depend on $z$

B) There is a covariate effect on survival time which is not linear nor can it be modelled under a proportionality of hazards assumption (**doubly–smooth** quantile estimator) $\lambda(t;z)$ varies smoothly with $z$.

C) The covariate has a *non–linear* effect but it can be assumed that a baseline hazard rate exists. ( quantiles estimated under the **smooth proportional hazards model**) $\lambda(t;z) = \lambda_o(t)\exp(s(z))$

D) The covariate effect is *linear* with a baseline hazard rate. (quantiles estimated under the **proportional hazards model**) $\lambda(t;z) = \lambda_o(t)\exp(\beta z)$

The most routinely used method of assessing the applicability of the PH model is to consider a plot of $\log(-\log(\widehat{S}(t|\tilde{z})))$ against survival time. If the model is

appropriate, the lines for the different levels of $\tilde{z}$ will be parallel. If the covariate is continuous, the values should be categoried to provide an interpretable plot.

A more formal score test for $\beta = 0$ is based on derivatives of the partial likelihood under the PH model. Le and Grambsch (1994) provide a test of association for survival time and continuous covariate which does not assume a monotonic relationship. The method, based on ranks and similar to the Siegel–Tukey test, is nonparametric but assumes that the covariate effect is symmetric.

Deriving properties and results which will allow practical inferential procedures with the doubly–smooth quantile estimator is a difficult task. In statistical research, when theoretical aspects of an analysis become mathematically complex, resampling techniques can be useful in investigating properties which would otherwise be extremely difficult to assess.

R.A. Fisher first proposed the idea of the permutation test in the 1930s and, with the advancement of computing technology, this method has become more and more popular. Efron (1981) introduced the resampling technique of bootstrapping, which is also becoming applied to an increasing number of areas of research and is computer intensive. For discussion of both these approaches, Efron and Tibshirani (1993)'s recent publication on the topic provides an excellent introduction.

The basic idea of a permutation test is that the statistic of interest is calculated for resamples where the data have been permuted in a manner which is consistent with the null hypothesis. These values are compared with the statistic value observed for the original data. In the survival data setting considered here, the (time,censoring) pair are permuted while the covariate values are held fixed. The statistic of interest, $\theta$, can then be calculated for each resample of data and compared to the value observed for the original data-set. This comparison can be summarised as an empirical p–value, written as

$$P = \frac{\text{no. of } \theta_i \text{ which are more extreme than that observed}}{N}$$

where $i = 1, \ldots, N$ and $N$ is the number of resamples. This is an estimate of the p–value, allowing an assessment of the presence of a covariate effect on survival

time. Bootstrapping has become a very popular approach for a wide range of applications. The following is intended to be a short introduction to the basic method and a review of bootstrap methods for censored data.

## 4.2 Review of Bootstrap Methods for Censored Data

### 4.2.1 Introduction

Efron (1979) proposed a method, known as *bootstrapping*, which can be used to make inferences about parameters of interest for the population using a large number of resamples of the observed data and standard formulae. The simplicity of the bootstrap makes it easy to estimate quantities such as standard error and bias and to calculate approximate confidence bands in many situations. Because it is a computer-intensive technique, it is extremely useful for cases where theory is very complex. The basic idea is described below.

Suppose we have an independent and identically distributed random sample from an unknown probability distribution function, $F$, so that $X_1, X_2, \ldots, X_n \sim F$. Say these are observed as $X_i = x_i, i = 1, \ldots, n$ and that we are interested in the statistic $\widehat{\theta}(X_1, X_2, \ldots, X_n)$

1. Construct $\widehat{F}$, the empirical distribution function, putting mass $\frac{1}{n}$ at $x_i, i = 1, \ldots, n$.
2. Draw a bootstrap sample $X_1^*, X_2^*, \ldots, X_n^*$ by randomly sampling with replacement from $\widehat{F}$.
3. Calculate $\widehat{\theta}^* = \widehat{\theta}(X_1^*, X_2^*, \ldots, X_n^*)$.
4. Independently repeat steps 2 and 3 a large number, $B$, times to obtain a set of bootstrap replications $\widehat{\theta}^{*1}, \widehat{\theta}^{*2}, \ldots, \widehat{\theta}^{*B}$.

The basic idea of the bootstrap is that the behaviour of $\widehat{\theta}$ about the true $\theta$ will be reflected in the behaviour of $\widehat{\theta}^*$ about $\widehat{\theta}$. In particular, the standard error

of $\widehat{\theta}$ can be estimated using

$$\widehat{\sigma}_{boot} = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} [\widehat{\theta}^{*b} - \widehat{\theta}^{*\bullet}]^2 \right\}^{\frac{1}{2}} \tag{4.1}$$

where $\widehat{\theta}^{*\bullet} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\theta}^{*b}$ and the bias by

$$\widehat{bias}_{boot} = \widehat{\theta}^{*\bullet} - \widehat{\theta} \tag{4.2}$$

The Percentile Method is probably the most straightforward approach for calculating a confidence interval based on the bootstrap. Let the cumulative distribution function for $\widehat{\theta}^*$ be

$$CDF_*(t) \equiv \Pr_*\{\widehat{\theta}^* \leq t\}$$

Under the assumption of no bias, and for $0 < \alpha < 1$, a $1-2\alpha\%$ confidence interval for $\theta$ is $(\widehat{\theta}_{LOW}, \widehat{\theta}_{UP})$ where

$$\widehat{\theta}_{LOW} \equiv CDF_*^{-1}(\alpha) \qquad \widehat{\theta}_{UP} \equiv CDF_*^{-1}(1-\alpha)$$

The scope for applications of this method is very wide and, over the years, there have been many adaptations of the idea and derivations of appropriate asymptotic results.

The following sections describe different ways in which the bootstrap can be used to analyse *censored data.* First we tackle how to sample *(time, censoring)* pairs and discuss the application to quantile estimation. Section 4.2.3 considers how to incorporate covariate information and the assumption of a survival model. These ideas are also extended to the estimation of quantiles.

### 4.2.2 Bootstrapping Survival Times

The obvious complication with bootstrapping survival data is how to deal with the incomplete, or censored, observations. Some researchers propose sampling the failure and censored times separately under appropriate survivor functions while others suggest simply sampling the observations as a pair.

The notation used will be that stated in Sections 1.2 and 2.2 of Chapter 2. For censored data, sometimes it is more convenient to discuss the distributions of variables in terms of the survivor functions, given by

$$S(t) = 1 - F_o(t) \quad ; \quad R(t) = 1 - G(t)$$

where $R$ is the cdf of the censoring variable. $S$ is most commonly estimated using Kaplan–Meier (1958) estimates to give $\widehat{S}$. We can estimate $R$ similarly by simply treating the censored times as failures and vice versa.

Suppose now we observe right–censored data of the form $\{(t_1, \delta_1), (t_2, \delta_2), \ldots, (t_n, \delta_n)\}$ and assume that $t_1 < t_2 < \ldots < t_n$. Efron (1981) proposed a simple extension to the original bootstrapping method, and used this to provide an estimate for the standard error of the Kaplan–Meier curve.

1. Using $\widehat{F}$, put mass $\frac{1}{n}$ on each *pair* of observations $\{(t_i, \delta_i)\}$ for $i = 1, \ldots, n$.
2. Construct bootstrap samples of the form $\{(X_1^*, \Delta_1^*), (X_2^*, \Delta_2^*), \ldots, (X_n^*, \Delta_n^*)\}$
3. Calculate $\widehat{S}^*(t)$, the Kaplan–Meier estimates using the bootstrap sample.
4. Repeat 2 and 3 a large number, $B$, times giving $\widehat{S}_1^*, \widehat{S}_2^*, \ldots, \widehat{S}_B^*$.

The standard error of the survivor function at a particular time $t$ can now be estimated by replacing $\widehat{S}^*(t)$ for $\widehat{\theta}^*$ in equation (1). Greenwood's formula is perhaps the most traditional way of estimating this quantity, but results from the bootstrap method are shown to agree very closely with values from this formula.

Recall the definition of right–censored data in section 2.1.2, where the $X_i^o$ denote a sample of failure times and the $U_i$ are a sequence of censoring variables, with $i = 1, \ldots, n$. The data are then written as the pairs $(X_i, \Delta_i)$ where

$$X_i = \min\{X_i^o, U_i\} \qquad \Delta_i = \begin{cases} 1 & \text{if } X_i^o \leq U_i \\ 0 & \text{if } X_i^o > U_i \end{cases}$$

Efron (1981) discussed sampling the variables $X_i^o$ and $U_i$ separately under their respective estimated Kaplan–Meier curves, using the definition of the random censorship model to obtain bootstrap samples. However, this was shown to

be equivalent to putting mass $\frac{1}{n}$ on each *(time,censoring)* pair where there are no ties between the failure times.

The method proposed by Reid (1981) sampled only from the Kaplan–Meier curves, which implies that samples will consist solely of failure times and be conditional on the amount of censoring in the original sample. The main advantage of this idea is that a bootstrap sample containing *only* censored observations will never occur, which is a possibility under Efron's method.

The asymptotic behaviour of the Kaplan–Meier under Efron and Reid's methods was studied by Akritas (1986). Using the theory of martingales for point processes, it was shown that only Efron's bootstrapping procedure could be used to construct asymptotically correct confidence bands.

Doss and Gill (1992) proposed a bootstrapping scheme which concentrated on the distribution of the censored times. The failure times are generated under the Kaplan–Meier estimates as before. However, when the survival time, $X_i$, is a failure, the only information available on the censored value is that it exceeds that failure time. It is proposed that the censoring variable could be generated from an adjusted Kaplan–Meier curve, calculated using only survival times greater than that value. The method can be written as follows.

1. Sample $X_i^{o^*}$ iid $\sim \widehat{F}^o \quad i = 1, \ldots, n$
2. Independently generate $U_i^*$ as
   - if $\delta_i = 0$, let $U_i^* = X_i$
   - if $\delta_i = 1$, $U_i^*$ is generated from $\widehat{G}_{X_i}$ where $\widehat{G}_{X_i}(t) = \frac{[\widehat{G}(t) - \widehat{G}(X_i)]}{1 - \widehat{G}(X_i)}$ for $t \geq x_i$
3. Form $X_i^* = \min(X_i^{o^*}, U_i^*)$ and $\Delta_i^* = I(X_i^{o^*} \leq U_i^*)$.
4. Steps 1–3 are repeated a large number of times to form bootstrap samples.

This bootstrapping scheme is asymptotically equivalent to Efron's method and can produce confidence bands which are asymptotically correct.

We return now to the subject of estimating quantiles for censored data. The $p^{th}$ percentile survival time ($0 \leq p \leq 1$) can be defined as $\xi_p = inf\{x : F(x) \geq p\}$. Replacing $F(x)$ by $1 - \widehat{S}^o$ provides an estimate of $\xi_p$. Obtaining an estimate of the standard error of the $p^{th}$ percentile lifetime can now be found using one of the bootstrapping schemes mentioned above and applying equation (4.1).

Horváth and Yandell (1987) derived asymptotic results for the bootstrapped Kaplan–Meier process and corresponding quantile process. Efron's sampling method was used although the size of the bootstrap sample was not necessarily assumed to be the same as that of the observed sample. Using Wiener processes, the convergence rates for strong approximations of the bootstrapped Kaplan–Meier and quantile processes were determined.

### 4.2.3 Bootstrapping with a Covariate

Consider now the case where, for each of $n$ individuals, a survival time, a censoring variable (indicating if that person has "failed") and a vector of covariate measurements are observed.

The data can be thought of as a set of triples

$$\{(t_1, \delta_1, \underline{z}_1), (t_2, \delta_2, \underline{z}_2), \ldots, (t_n, \delta_n, \underline{z}_n)\}$$

A common technique for analysing data of this type is to fit Cox(1972)'s Proportional Hazards (PH) model. This model is described in Section 5.1 of Chapter 2. In this situation, bootstrapping could be used to estimate the standard error for $\beta$, the coefficient assigned to the covariate, and also find a confidence interval. Other areas of interest may be assessing the validity of the model and estimating the variance of quantile survival time given a covariate. Methods to investigate these, and other problems, will be discussed. As before, for simplicity, we shall assume we have measured only one covariate value for each individual.

Efron and Tibshirani (1986) discussed a number of examples where bootstrapping was useful, two of which involved survival data and the PH model.

The following resampling scheme was used.

1. Put mass $\frac{1}{n}$ at each triple $(t_i, \delta_i, z_i)$ for $i = 1, \ldots, n$.
2. Generate a bootstrap sample $\{(X_1^*, \Delta_1^*, Z_1^*), (X_2^*, \Delta_2^*, Z_2^*), \ldots, (X_n^*, \Delta_n^*, Z_n^*)\}$ by randomly sampling with replacement.
3. Calculate $\widehat{\beta}^*$ by maximising the partial likelihood of the bootstrap sample.
4. Repeat steps 2 and 3 a large number of time to obtain $\widehat{\beta}_1^*, \widehat{\beta}_2^*, \ldots, \widehat{\beta}_B^*$.

The standard error, and a confidence interval based on the percentile method, for $\widehat{\beta}$ can be estimated as for $\widehat{\theta}$ in the introductory section.

Tibshirani and Hastie (1987) proposed replacing the term $e^{\beta z}$ by a more general form, $e^{s(z)}$, where $s(z)$ is estimated using local likelihood estimation, a nonparametic technique. (See also Section 5.2 of Chapter 2 and Chapter 3 for further discussion of this model.) If step 3 is altered so that it is fitting the general model for each bootstrap sample then we can compare these with the observed sample under the $e^{\beta z}$ form. This is probably most aptly achieved by plotting the relative risk for a range of covariate values. These comparisons will show whether the parametric form of the model is valid.

Chen and George (1985) fitted the PH model to a set of data using a stepwise procedure to select appropriate covariates. To assess the validity of this model, bootstrap samples were drawn using Efron and Tibshirani's method (but sampling the covariate *vector* instead of individual values) and models found for each. The number of covariates and the frequency with which each was chosen were noted and the final model was chosen to be that which was observed more often.

Efron and Gong (1983) discussed the case where the covariate is binary eg. independent treatment and control groups. The bootstrapping method proposed here was that, instead of sampling the set of triples, each group was sampled separately under their respective empirical distribution functions. Then, each $\widehat{\beta}^*$ was calculated under the PH model for a pair of bootstrap samples.

Karrison (1990) chose bootstrap samples under the piecewise exponential (PE) model, which assumes the hazard rate is constant within arbitrarily defined intervals. The failure times are sampled under the survivor function of the PE model for a given value of the covariate. The censored times are sampled from the censored KM estimates. Then the minimum of each pair is chosen as the survival time and the censoring indicator noted. Parameters and quantities of interest under the PE model were then calculated.

Hjört (1985) employed a similar bootstrap scheme for the Proportional Hazards model.

1. Generate $X_i^{o^*} \sim (\widehat{S}_\beta(\cdot))^{\exp(\underline{\widehat{\beta}}^T \underline{z}_i)}$ for $i = 1, \ldots, n$.
2. Generate $U_i^* \sim \widehat{G}$, the KM for censored times for $i = 1, \ldots, n$.
3. Then form $X_i^* = \min(X_i^{o^*}, U_i^*)$ and $\Delta_i^* = I(X_I^{o^*} \leq U_i^*)$ with $Z_i^* = z_i$ for all $i$ to give a bootstrap sample $\{(X_1^*, \Delta_1^*, Z_1^*), (X_2^*, \Delta_2^*, Z_2^*), \ldots, (X_n^*, \Delta_n^*, Z_n^*)\}$.
4. Repeat steps 1–3 a large number of times to form appropriate bootstrap samples.

For each bootstrap sample, the PH model can be fitted and $\beta$ estimated. This method is similar to that mentioned in the previous section by Doss and Gill in that the failure and censored times are resampled separately under different distribution functions. To make the methodology perhaps slightly more "symmetrical" we could sample the censored times from a PH model which treats the censored times as failures and vice versa (as for $\widehat{G}$, the censored KM).

Burr and Doss (1993) proposed that, using this method of bootstrapping within the Proportional Hazards model, a confidence interval for the percentile survival time, given a covariate value, could be found and that these intervals would be asymptotically correct. The $p^{th}$ percentile lifetime given $z$ can be estimated by

$$\widehat{\xi}_p(z) = \widehat{S}_\beta^{-1}((1-p)^{\exp(-\widehat{\beta}z)}) \qquad 0 \leq p \leq 1$$

where $\widehat{S}_\beta$ is the estimated survivor function under the PH model. Bands are constructed using

$$W^* = \sup \frac{\sqrt{n}|\widehat{\xi}_p^*(z) - \widehat{\xi}_p(z)|}{\widehat{\sigma}(p,z)}$$

where $\widehat{\sigma}(p,z)$ is calculated using an asymptotic result about the quantile process. Obtain $W^*$ for each bootstrap sample to produce the sequence $W_1^*, W_2^*, \ldots, W_B^*$. Let $b_\alpha^{(n)}$ be the $(1-\alpha)^{th}$ quantile of the empirical distribution of $W_1^*, W_2^*, \ldots, W_B^*$. Then, the bands

$$\widehat{\xi}_p(z) \pm b_\alpha^{(n)} \frac{\widehat{\sigma}(p,z)}{\sqrt{n}}$$

have asymptotic coverage probability of $1-\alpha$.

### 4.2.4 Summary of Main Methods

| Data | Reference | Method |
|---|---|---|
| (time, cens) | Efron (1981) | Sample pairs under empirical distribution function |
| | Reid (1981) | Sample failure times from KM estimates |
| | Doss and Gill (1992) | Sample failures and censored times separately under respective KM curves |
| (time,cens,covariate) | Efron & Tibshirani (1986) | Sample triples under empirical distribution function |
| | Hjört (1985) | Sample failures under survivor function from PH model and censored times from censored form of KM estimator |

Table 4.1: Table of Main Methods for Bootstrapping Censored Data

## 4.3 Test Procedures

Returning to the aims discussed in the first section of this chapter, investigation of which of the four models best describes the covariate effect for a set of data can now be examined more formally. To recap, the situations of interest from section 4.1 are as follows, where "linear" refers to the hazard function in the log scale.

A) The covariate has *no effect* on survival time.

B) The effect of the covariate is best described by a *smooth* curve which has no assumption of proportionality in the hazard.

C) The covariate has a *non–linear* effect and it can be assumed that a baseline hazard rate exists.

D) The covariate effect is *linear* with a baseline hazard rate.

In Section 4 of Chapter 3, a likelihood statistic was proposed which characterised a set of data by estimating the gradient at failure times, in the form of a likelihood function, and could be calculated under any of the 3 different quantile curve approaches. Using this information and the knowledge of different resampling techniques, a set of progressive testing procedures can be formulated. This can be summarised by the table below.

| Comparison | Approach | |
|---|---|---|
| A vs. B | Means of Inference | A permutation test |
| | Method of Assessment | Compare the value of $NPll$ for the original data, to the $NPll$ values of the resamples and find the empirical p–value for the test of "no covariate effect" |
| B vs. C | Means of Inference | Bootstrapping triples approach of Efron and Tibshirani(1986) |
| | Method of Assessment | The $NPll$ values of the bootstrap samples provide a band representing the variability of the NP approach and can be compared to the observed value of $SPHll$ |
| B vs. D | Means of Inference | Bootstrap under the PH model, similar to Hjört (1985) but using the more symmetrical approach of also sampling the censored times under a PH model. |
| | Method of Assessment | For each bootstrap sample, calculate the difference $D_2$ between $NPll$ and $PHll$ and by comparing this with $D_2$ for the observed data, calculate the empirical p–value |

Table 4.2: Summary of Testing Procedures

This moves into the standard setting of testing which may be expressed in terms of a null and an alternative hypotheses. For example, to test A vs. B,

$\mathcal{H}_0$ : No covariate effect

$\mathcal{H}_1$ : Covariate effect modelled by doubly–smooth quantile estimator

The case of comparing the doubly–smooth estimator with the smooth PH model (B vs. C) is not a formal testing procedure. The implementation here of the SPH model is a local likelihood approach. However, consideration of this method of analysis is moving into the area of generalised additive models (Hastie and Tibshirani, 1990) and this is, in itself, a large topic. To construct a test

would require bootstrapping the data under the SPH model and the results of this would only really be convincing if a GAM was being fitted. Instead an alternative approach is to bootstrap the data as triples (Efron and Tibshirani, 1986) and use the values of $NPll$ or the centile curves from these samples to construct confidence bands. If the value of the likelihood statistic under the SPH model lies outwith this band, it would give some indication that the approaches are reporting different forms of covariate effect.

The case of B vs. D is a test of proportionality of the hazard and the monotonicity of the covariate effect, it is an important comparison of methodologies. This now provides a natural progression of testing the models A,B and D, with the link between B and C considered separately. The following sections give the results of these procedures for the sets of data discussed and analysed in Chapter 3.

## 4.4 Results

### 4.4.1 Testing A vs. B

Plotting the doubly–smooth median curves of permuted data, we would expect to observe a band of curves which are clustered around the median survival time across the range of the covariate.

Figure 4.1 considers the data simulated to have no covariate effect discussed as Example 1 in Chapter 3. The curves for 100 sets of resampled data are plotted in light grey, with the curve for the original data as the bold line. The median curve does not appear to be significantly different from those which have randomly allocated covariate values although the interpretation is arguable. However, when the Stanford Heart Transplant Data is examined by the same procedure, in Figure 4.2, the 40th centile curve is clearly distinct from the curves constructed from permuted data. There is therefore clear evidence that the survival time is decreasing as age increases.

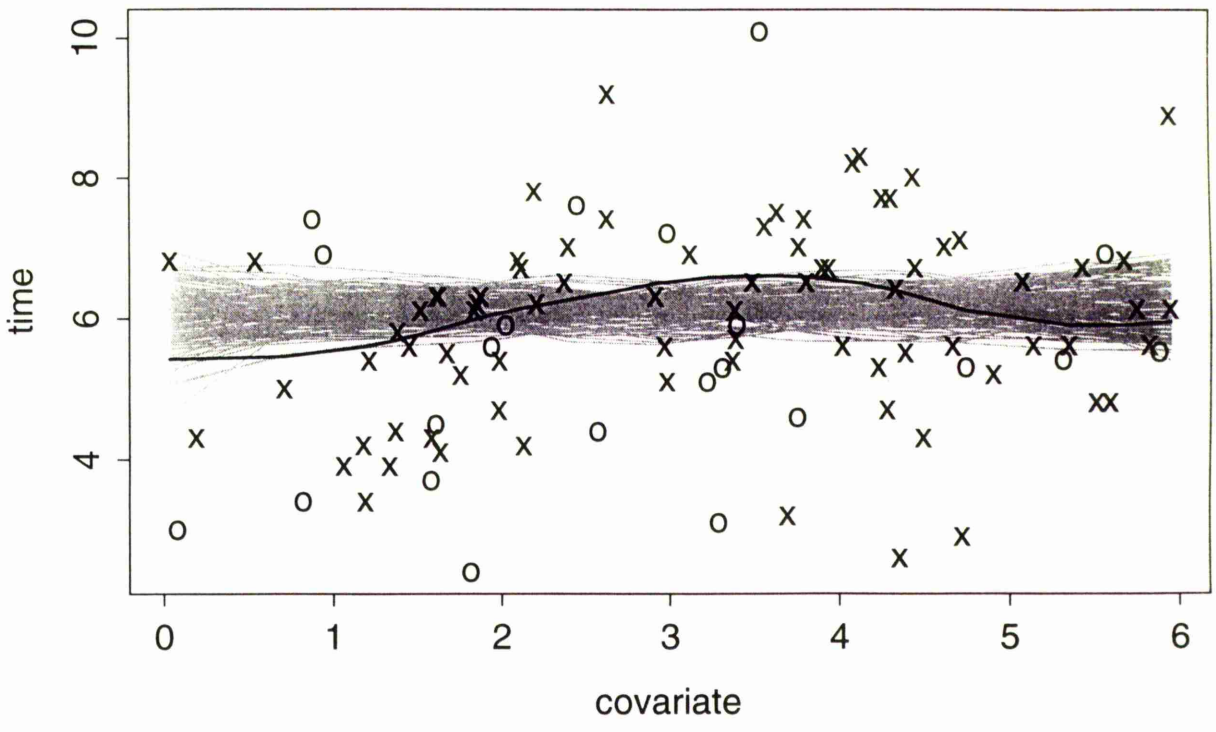



Figure 4.1: No effect data

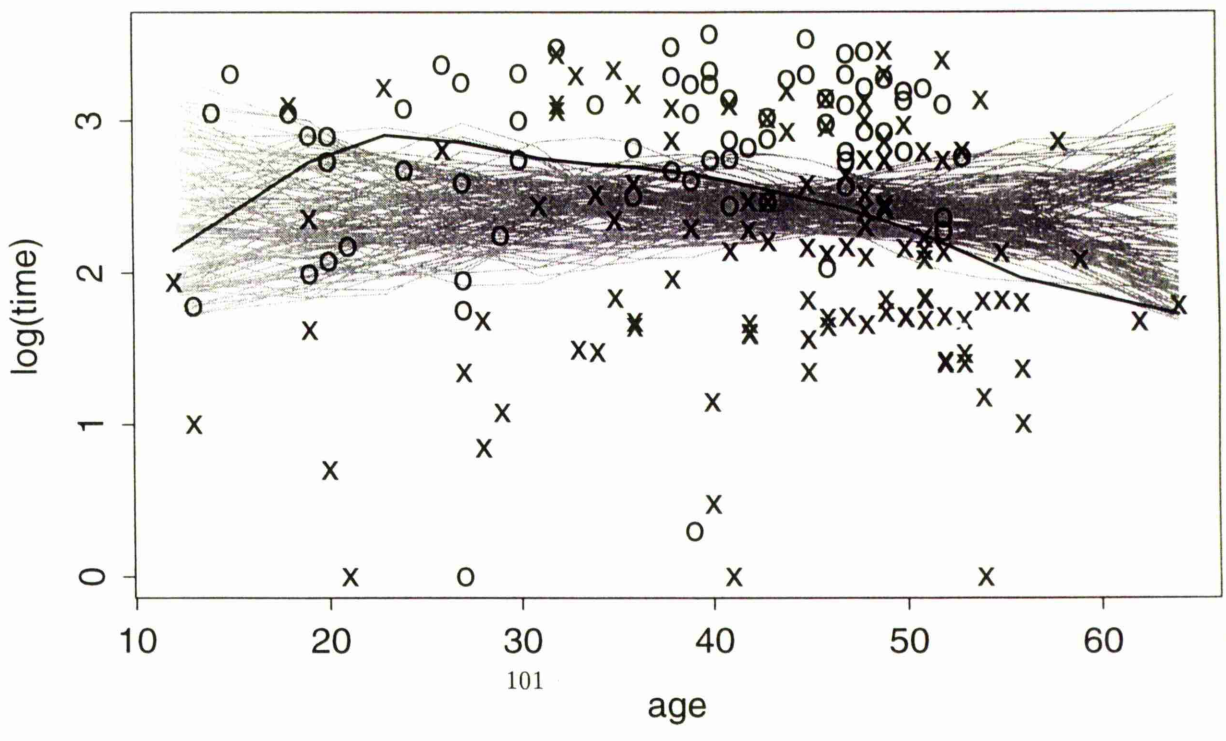



Figure 4.2 : Stanford Heart Transplant Data

To test this informal assessment more formally, the likelihood statistic based on the nonparametric approach is calculated for the 100 sets of permuted data and an empirical p–value is found. This procedure is implemented on each of the six example data-sets discussed in Chapter 3.

| Example No. | Data–set | Emp. p–value |
|---|---|---|
| 1 | No Effect data | 0.27 |
| 2 | Quadratic data | 0.00 |
| 3 | PH data | 0.05 |
| 4 | Non-PH Simulated data | 0.15 |
| 5 | Stanford H.T. data | 0.03 |
| 6 | M. Myeloma data | 0.36 |
| 7 | Renal Unit data | 0.44 |

Table 4.3: Results for Test of A vs. B

The values confirm the conclusions of the subjective impressions from the graphical descriptions in Chapter 3.

The borderline significance of the data simulated from a PH model reflects the slight curvature in the observed quantiles. Typically, low centile curves will stay relatively constant across the covariate since although the range of the survival time increases, the minimum survival time does not.

It is interesting to note that the test concludes no covariate effect for the multiple myeloma data. For higher centiles, there did appear to be a covariate effect, but since this method is effectively testing over the whole range of centiles, effects of this restricted type may not be identified.

Similarly the effect of blood pressure on survival time in the Renal Unit data is not found to be significant. It has proved difficult to produce a definitive analysis of these data as described in detail in Chapter 6, due principally to the extremely high proportion of censoring.

### 4.4.2 Testing B vs. C

Example 4 in Chapter 3, the non–PH simulated data, displayed a case where the percentile curves from a PH model were a poor fit to the data. The variability of the survival times was decreasing across the covariate values, converging to the

median, but the PH model did not adequately capture this structure. However, it is only in the large and small centiles that this is observed. Figure 4.3 displays the doubly–smooth median curves for 100 bootstrap samples. The bold line is the median curve under the smooth PH model. This would suggest that there is little difference in the medians of the two approaches. However, Figure 4.4 illustrates the same data with the lines representing the $10^{th}$ centiles. This clearly shows the difference in the behaviour of the two approaches for the low centile.

In the next chapter, a data-set on leg ulcer patients is discussed. The time of interest was the number of days until the ulcer had healed, and age was thought to be a potentially important covariate. Also available, and thought to be important, is a record of whether the presence of deep vein involvement played a part in the healing process. Here, survival times for those patients who did experience deep vein involvement are analysed to assess the age effect. Figure 4.5 is a scatterplot of the data with the PH (broken line), SPH (dotted line) and doubly–smooth median curves superimposed. There appears to be a clear difference between each of the three methods. The doubly–smooth curves for 100 bootstrap samples are shown with the SPH median curve as a bold line in Figure 4.6. One conclusion from this may be that the proportional hazards assumption is not a particularly good description of the data.

The results of the bootstrap analyses of all the example data–sets are tabulated below with the range of values of $NPll$ for the bootstrap samples written as a confidence band. Checking whether the observed $SPHll$ value is near the centre of this band can be informative when looking for evidence of differences between the two approaches.

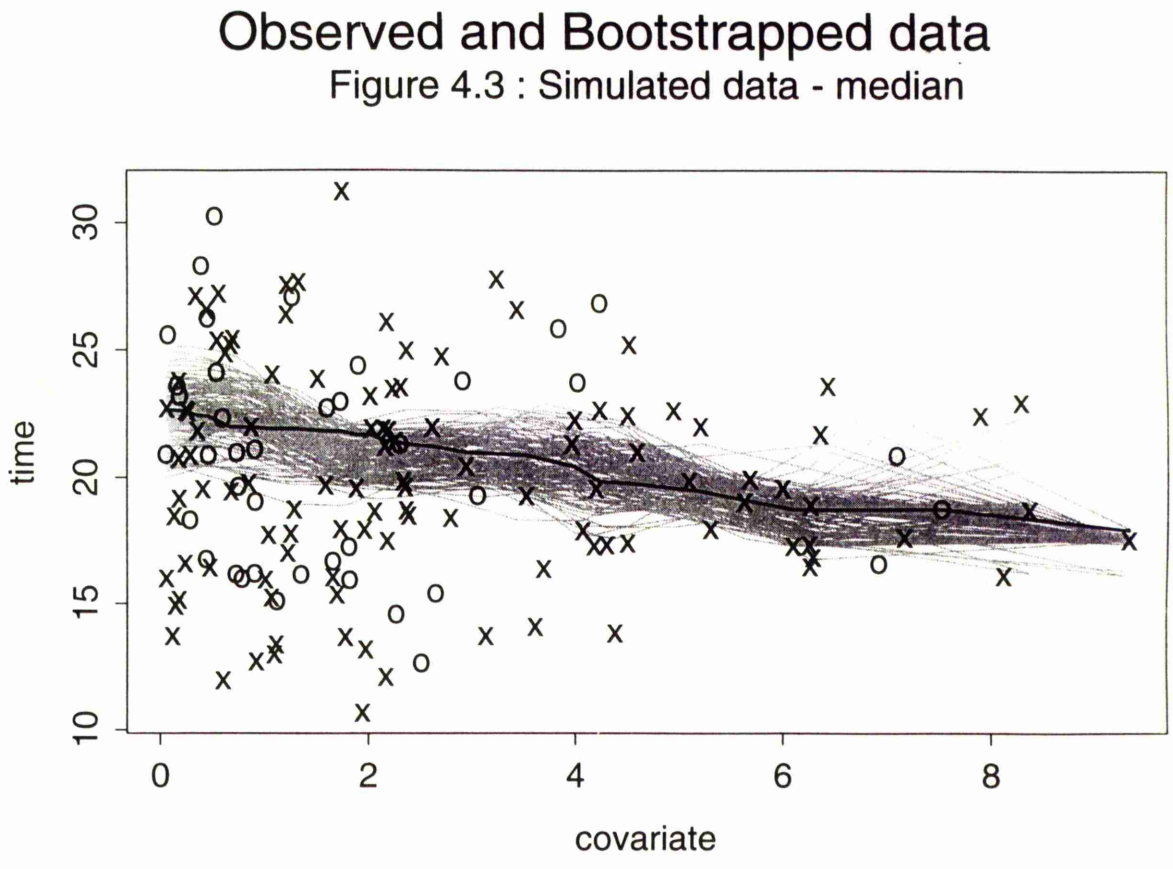



Figure 4.3 : Simulated data - median

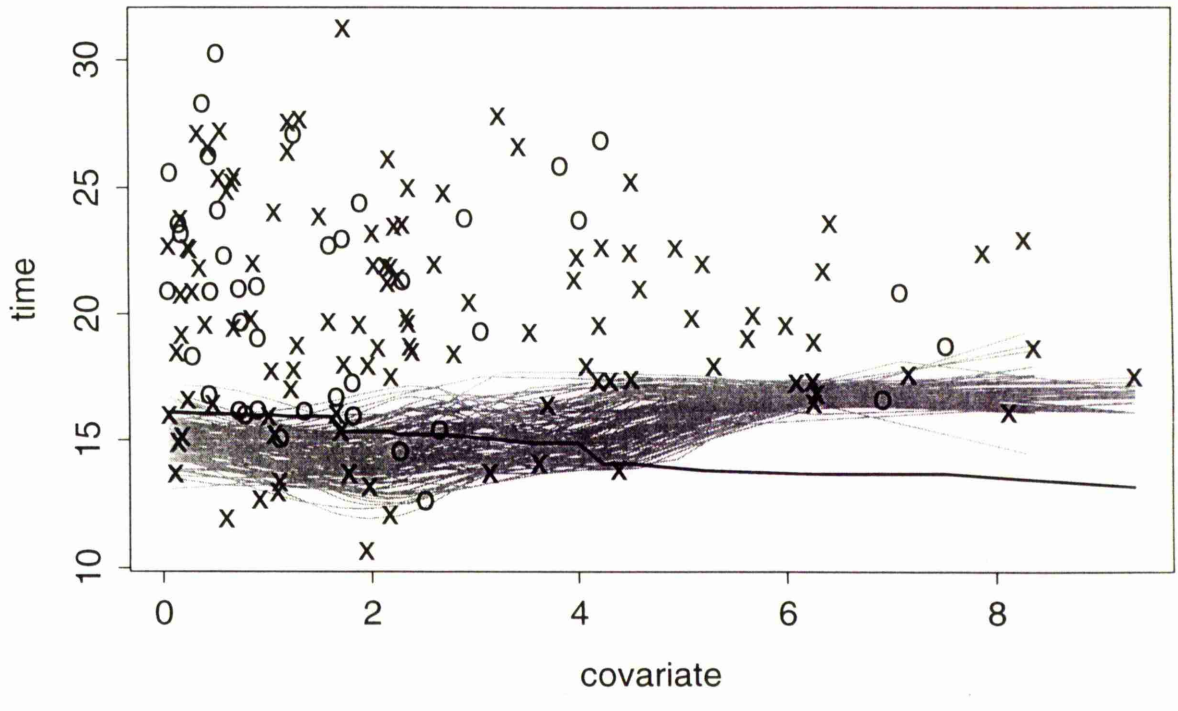



Figure 4.4 : Simulated data - 10th centile

## Observed and Bootstrapped data

Figure 4.5 : Median curves for Leg Ulcer Data

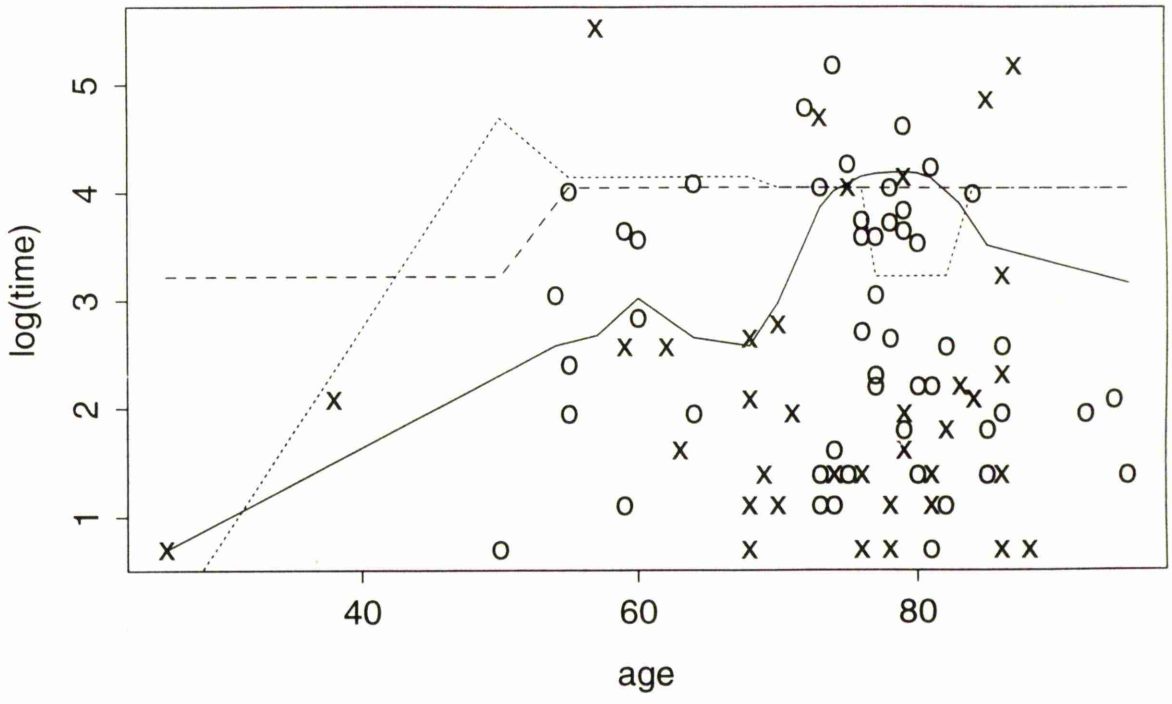
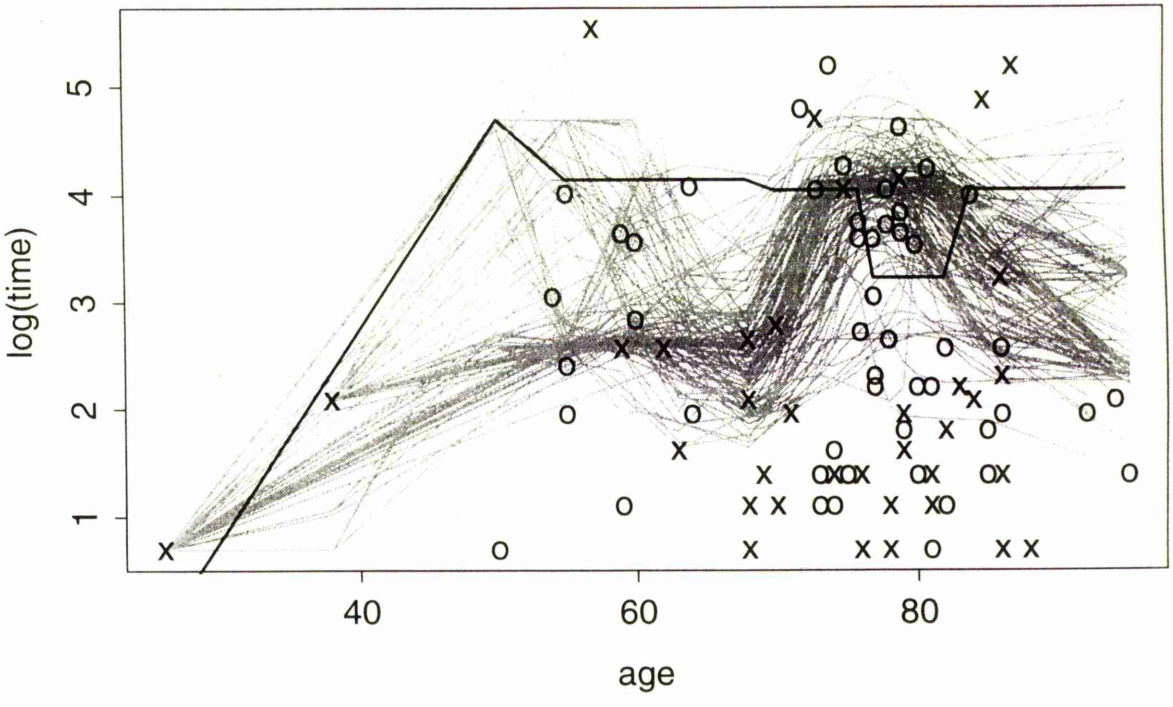



Figure 4.6 : Leg Ulcer Data



| Example No. | Data–set | Confidence Band | Obs. $SPHll$ |
|---|---|---|---|
| 1 | No Effect data | (-197,04,-148.32) | -168.36 |
| 2 | Quadratic data | (-190.46,-154.27) | -172.05 |
| 3 | PH data | (-126.31, -90,35) | -115.39 |
| 4 | Non–PH Simulated data | (-413.85,-297.89) | -364.27 |
| 5 | Stanford H.T. data | (-190.98,-155.21) | -181.39 |
| 6 | M. Myeloma data | (-177.35,-119.69) | -151.07 |
| 7 | Renal Unit data | (-592.44,-199.62) | -225.55 |
| – | Leg Ulcer data | (-116.79, -68.24) | -110.09 |

Table 4.4: Results of Comparing B vs. C

This would appear to suggest that the curves from the doubly–smooth estimator are often very similar to those estimated from the smooth PH model in all of the examples. The only example where the observed $SPHll$ value seemed rather extreme is in the leg ulcer data which was expected from the plot of median curves. For example 4, we know that there are differences in the low and high centiles, but since the likelihood statistic measures the fit of a model across all of the centile values, none are substantially larger as to produce a significant result. To highlight the differences in this example, a test for differences in the individual percentile curves themselves would be necessary, something which may be worth investigating in the future.

### 4.4.3 Testing B vs. D

Four examples are used here to exhibit the lack of differences between the curves under the Proportional Hazards model and the doubly–smooth estimator. In each the band of curves generated from PH resampling are shown as faint lines, with the doubly–smooth quantile curves for the original data as solid, black lines.

For the "No effect" data, it can be seen in Figure 7 that the doubly–smooth median curve moves between the extremes of the PH band of medians but always lies within it. As would be expected, we are unlikely to find differences between the approaches for this type of data.

Figure 4.8 shows similar median curves for the Stanford Heart Transplant data and here the difference in structure for age values less than 20 years is made

obvious. However, for a large part of the age range, there appears to be little difference between these curves.

Figure 4.9 shows the PH bootstrapped 10th centile curves and there is clearly a difference for covariate values greater than 6. But again a formal test based on all centiles may not detect this since these differences will not be so obvious in centiles closer in value to the median.

Finally, Figure 4.10 considers the Leg Ulcer data discussed in the last section (on B vs. C). Although there looked to be differences between the smooth PH model and the doubly–smooth curves, these differences would not appear to exist between the PH and doubly–smooth approaches. This may confirm the reservations expressed over the implementation of the SPH model.

Results of the bootstrap test for each example are given below.

| Example No. | Data–set | Emp. p–value |
|---|---|---|
| 1 | No Effect data | 0.21 |
| 2 | Quadratic data | 0.00 |
| 3 | PH data | 0.46 |
| 4 | Non–PH Simulated data | 0.34 |
| 5 | Stanford H.T. data | 0.24 |
| 6 | M. Myeloma data | 0.17 |
| 7 | Renal Unit data | – |
| – | Leg Ulcer Data | 0.25 |

Table 4.5: Results of Testing B vs. D

An empirical p–value for the Renal Unit data was unobtainable, despite several attempts. The problem was that too many samples consisted solely of censored times, preventing a PH model from being fitted.

These results reflect exactly the conclusions of the subjective impressions made from Figures 4.7-10 and those from Chapter 3. Some differences appear to be observed graphically but do not result in a small p–value. This may be because the covariate effect is only different in one aspect, such as the very low and high centiles of the Non–PH data or at the extremes of the covariate values.

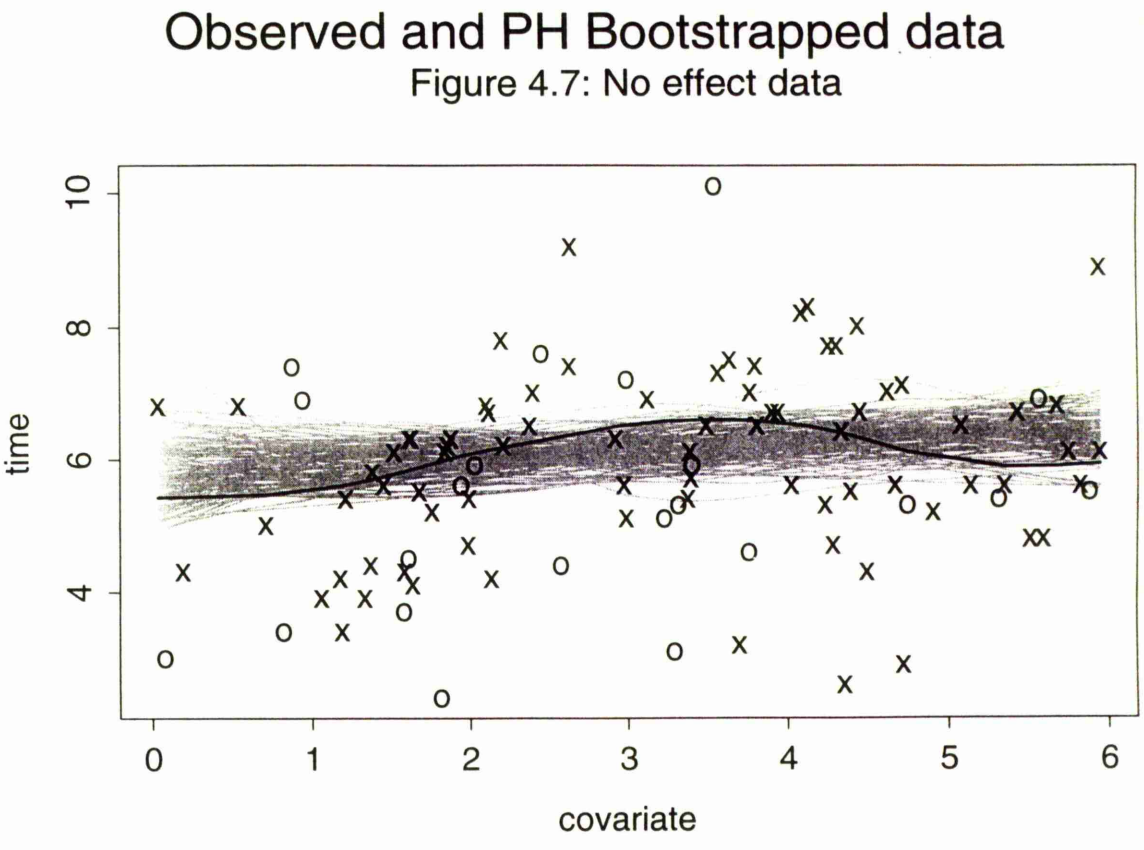



Figure 4.7: No effect data

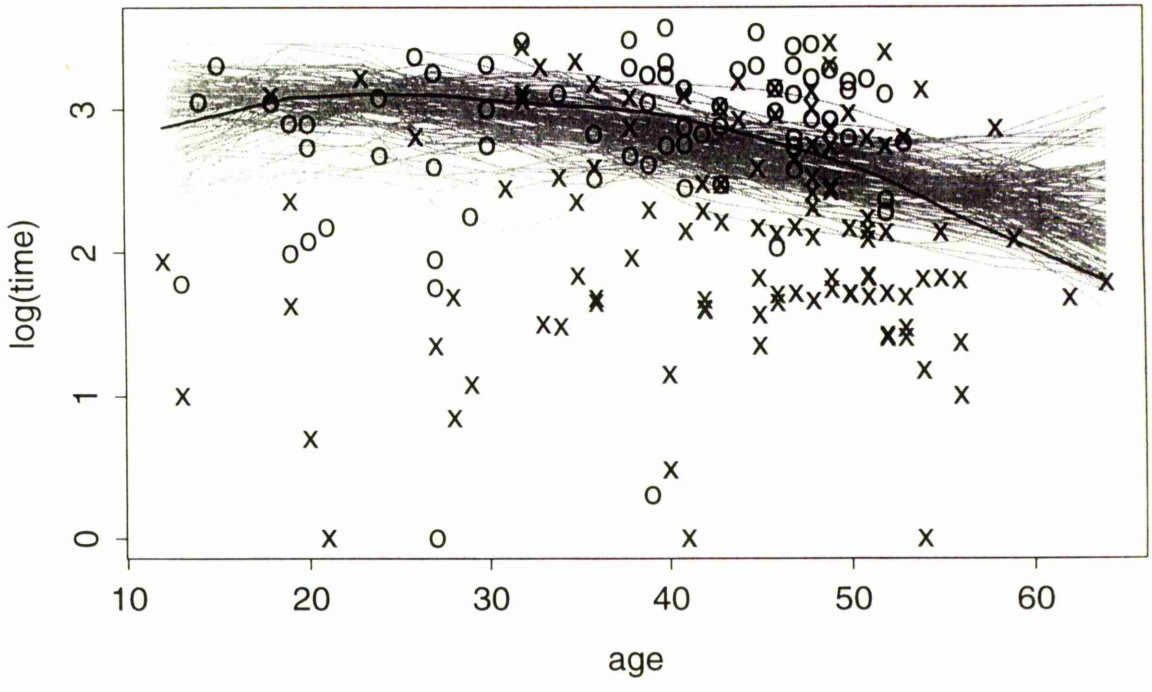



Figure 4.8 : Stanford Heart Transplant Data

Observed and PH Bootstrapped data

Figure 4.9 : 10th centile for Non-PH Simulated data

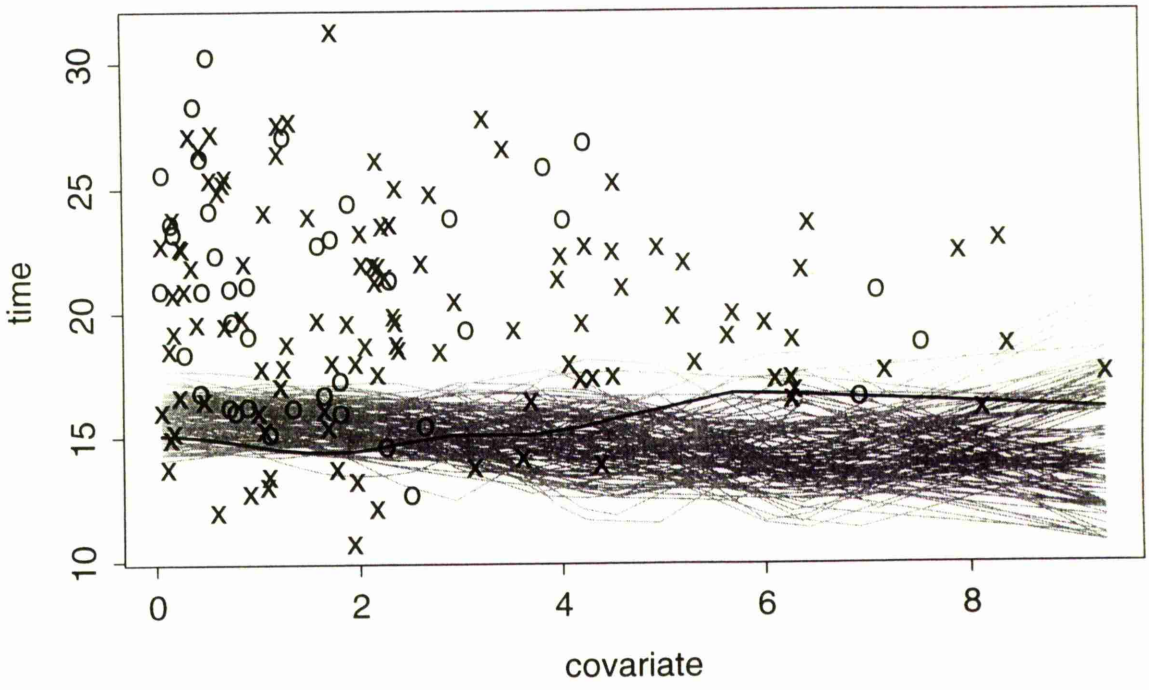
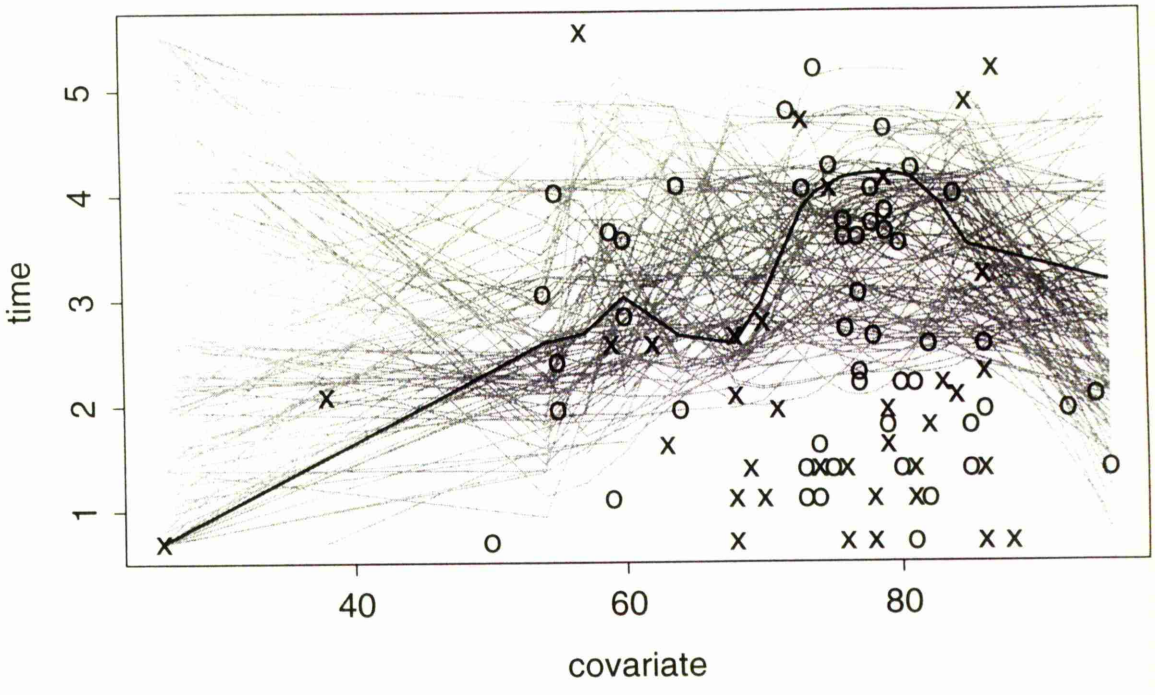



Figure 4.10 : Leg Ulcer data



### 4.4.4 Conclusions

The test statistic, along with the resampling techniques, has proved to be useful in gaining an overall picture of differences between the different quantile curve approaches. In certain cases, where centiles behave differently relative to other centiles, it would perhaps be more appropriate to test only the information from those curves.

## 4.5 A Small Power Study

When proposing a testing procedure, it is important to assess the power and size. Here, there is particular interest in these properties for the test of the nonparametric versus proportional hazards models (B vs. D).

First of all, the probability of wrongly rejecting the hypothesis that the PH model is appropriate was assessed. For each of 100 data–sets simulated under the PH model, the bootstrap test was applied and the empirical p–value noted. These p–values should be approximately uniform and the number of values less than or equal to 0.05 should be very small (hopefully around 5%). The actual observed number was 7, which is within acceptable limits of Binomial variation.

Next, the case where the PH model is inappropriate is considered. The most obvious example to use is the quadratic data and again 100 similarly simulated data–sets were analysed. Here we expect the p–values to be very low, with almost all of them being less than 0.05. In fact, because of the strong non–monotonic covariate effect, none of the empirical p–values were observed to be above 0.00.

While the cases considered here are quite extreme, they provide a reassuring illustration of the performance of the test. It would have been useful to have executed similar exercises for the other examples but each of these has a substantial computational cost.

## 4.6 Summary

The plot of survival time vs. covariate is very useful for obtaining a subjective impression of the effect of that covariate. More information can be extracted from this if percentile curves are superimposed.

Cox's PH model is a familiar and commonly-used one in medical statistics. The theory is well-developed and implemented in many statistics packages. However, it is important to be aware of the assumptions made and the limitations of the model. The smooth PH model, proposed by Efron and Tibshirani (1986), provides an important step to making the original PH model more general, while still retaining the baseline hazard assumption. By introducing the doubly–smooth quantile estimator, the aim was to provide an alternative to the PH model which would be applicable to all sets of data, particularly useful when the assumptions of the model were inappropriate, and to provide a means by which the assumption of proportionality in the SPH model could be checked.

Often a diagram depicting the curves from each of the approaches will allow an informal assessment of whether these ideas agree but this can be tested more formally using a likelihood based statistic and one of the resampling techniques discussed here.

# Chapter 5

# Extension to Analysis of Covariance

## 5.1 Discussion

One possible extension of the work on quantile curves is to consider the case where, as well as investigating the effect of a continuous covariate on survival time, a grouping factor is present. The interest here would be whether the behaviour of the survival time for each level of this factor is different and how this difference can best be characterised. This is analogous to the regression method known as analysis of covariance (ANCOVA). In the simple linear case, the model can be written as

$$Y_{ij} = \alpha_i + \beta_i X_{ij} + \epsilon_{ij} \qquad i = 1, \ldots, m \; ; \; j = 1, \ldots, n_i$$

where $Y$ is the response, $X$ is the explanatory variable and the $\epsilon$'s are the errors, assumed to have an $N(0,1)$ distribution. For each level of the grouping factor, the relationship between the response and the explanatory is expressed in a linear form with slope $\beta_i$ and intercept $\alpha_i$.

For simplicity, the simplest case where the data are grouped by a binary outcome ($m$=2), for example, treatment/placebo or sex, is considered here. There are 3 main situations in which we may be interested. These are where the lines for each outcome $i$ are

1. Different : $\alpha_1 \neq \alpha_2$ and $\beta_1 \neq \beta_2$
2. Parallel: $\alpha_1 \neq \alpha_2$ and $\beta_1 = \beta_2$
3. Same: $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$

Lines representing each of these situations can be fitted using least squares estimation and standard parametric theory is available for testing which of the three descriptions is most plausible for the data.

One alternative to the least squares approach is to consider a generalised additive model (GAM). These models allow more flexible relationships between the response and covariates and so imply the requirement for a smoothing procedure, usually provided by cubic splines. This area has been researched extensively and Hastie and Tibshirani's (1992) book on the topic is an excellent account of this work.

To extend the ideas behind these methods to quantile curve estimation, the problem is best explored by considering the hazard function. While it is not obvious how to produce quantile curves which are parallel for the levels of the binary covariate, under the proportional hazards model, it is possible to constrain the hazard to be parallel in the log scale.

For the smooth PH model case, a form of GAM is assumed. Here a simple extension of the model described earlier is applied but for a fuller description of the more general case see Section 8.3 of Hastie and Tibshirani (1992).

The different models, written in terms of the hazard function, for each of the three approaches, are probably most easily explained by the table below.

| | Different | Parallel | Same |
|---|---|---|---|
| PH | $\lambda_{oi}(t)\exp\{\alpha_i + \beta_i z\}$ | $\lambda_o(t)\exp\{\alpha_i + \beta z\}$ | $\lambda_o(t)\exp\{\beta z\}$ |
| SPH | $\lambda_{oi}(t)\exp\{\alpha_i + s_i(z)\}$ | $\lambda_o(t)\exp\{\alpha_i + s(z)\}$ | $\lambda_o(t)\exp\{s(z)\}$ |
| NP | $\lambda_i(t\|z)$ | - | $\lambda(t\|z)$ |

The models which assume that the curve is the *same* for the 2 treatment groups simply fits a single curve to the entire dataset. When the curves from the

treatments are assumed to have *different* shapes, the data can be split into two groups and analysed separately. The behaviour of the curves where there is an assumption of *parallelism* in the log scale is more difficult to fit and assess.

For the PH model, the value of $\alpha_1$ is chosen to be zero (automatic when the first binary covariate level is coded as zero) and $\alpha_2$ and $\beta$ are chosen by the usual partial likelihood method. For the smooth PH model, the $s(z)$ term controls the behaviour of the continuous covariate (as $\beta$ does in the PH model) with $\alpha_2$ ($\alpha_1$=0) determining the amount of vertical shift. For a given value of $\alpha_2$, $s(z)$ can be found using a similar approach to that described in Section 5.1 of Chapter 2. The value of $\alpha_2$ can be found by a grid–search, with the selection criteria being the partial likelihood expression for which the estimates of $s(z)$ have been substituted.

A model corresponding to parallelism in the NP approach is more difficult to formulate due to the unrestricted nature of the hazard functions in this case. For that reason, parallel curves will be investigated only for the PH and SPH structures.

The likelihood statistic, proposed in Chapter 3, can also be calculated under each of these situations. When considering the quantile curves to be the same, the information from the binary covariate is considered unimportant and the likelihood statistic, $L$, is simply estimated as described in Section 3.4.

$$L = \prod_{i=1}^{k} \left\{ \prod_{l \in D_i} -\frac{d}{dt}\tilde{S}(t_{(i)}|z_l) \prod_{l \in C_i} \tilde{S}(t_l|z_l) \right\}$$

For situations where the quantile curves are proposed to be different, the value of $L$ is calculated for each group separately and these are multiplied to give an overall likelihood statistic for that method. Under the assumption of parallel curves, the same principle applies. The SPH or PH model is fitted, the likelihood is calculated for each group, then multiplied to produce the final value of $L$. Comparisons can be made using bootstrapping and permutation tests as before.

Now, not only do we have to choose between the different survival models, but also, given a model, which of the descriptions discussed above is most appropriate for the data. The structure of the problem has become more complex, although

the two issues of model form and covariate effects can be assessed separately, e.g. it would not be sensible to test the parallel SPH curve against the different NP curves situation.

Analysis procedures similar to those discussed in Chapter 4 can be applied. To compare NP vs. SPH, Efron and Tibshirani (1986)'s bootstrap approach and construction of confidence bands can be used. To test NP vs. PH, bootstrap samples under the corresponding PH model are analysed and result in an empirical p–value to assess the applicability of the PH model.

When choosing a model, the key interest initially is to assess whether the curves are different or if they are the same under a given model. This can be examined by using the permutation argument again. If we randomly allocate the binary covariate, under the assumption that the curves are the same, this should make no difference to the likelihood statistic.

Having established that the curves are different for the two levels, if this is indeed the case, then we may be interested in whether these curves can be assumed to be parallel in the log scale. This is difficult to test since we have no formal way of comparing the SPH and PH models and is a case which shall have to be examined in more detail in the future.

## 5.2 An Application

These data were made available by Caroline Doré of the Royal Postgraduate Medical School in London and were used as part of a clinical trial to investigate factors affecting the healing time of leg ulcers. (For more details of this study, see Skene *et al.* (1992)) Age is known to be an important covariate for this condition and the severity of the ulcers is partly determined by whether or not deep vein involvement (DVI) exists.

Figure 5.1 shows a plot of survival time vs. age. Survival time here is time to healing, with a failure time (denoted by x) here representing the number of

days from the treatment being administered until the leg ulcers have completely healed. (The censored times are, as before, denoted by a o.)

Since there is a large cluster of low survival times and heteroscedasticity across age, the survival times are given a logarithmic transformation. Figure 5.2 shows the data in this form, with the failure and censored times for those who experienced deep vein involvement denoted as solid boxes and solid circles respectively. There may be a case for analysing only observations for which age $> 50$ years due to sparsity of data below this age but, for the purposes of this example, all of the available data will be used.

The median percentile curves are plotted for each of the different approaches in Figures 5.3–5. In each of these plots, the different assumptions of same, parallel and different curves are illustrated. Where two lines are plotted (ie for the "parallel" and "different" cases), the higher one represents individuals who experienced deep vein involvement.

The doubly–smooth curves of Figure 5.3 exhibit quite a lot of curvature but the main point to notice is that when the data are separated by the binary covariate, the curves are quite clearly different. ($h_1 = 5$, $h_2 = 0.075$.)

In Figure 5.5 ($k = \frac{1}{3}$, where $k$ is the proportion of data used in each window of the local estimation procedure), the parallel assumption defines a clear shape to the curves fitted under the smooth PH model, similar to that when the binary covariate is ignored. However, the results are quite different and difficult to interpret when the curves are plotted separately.

The conclusions for the PH model, shown in Figure 5.4, are basically the same although the monotonicity assumption restricts the curve from dipping at 65 years, which was observed in both the nonparametric approaches.

The likelihood statistic for each of the possible situations was calculated, with $h_v$ chosen to be 0.5, and these are given in the following table.

Leg Ulcer data

Figure 5.1 : Survival Time vs. Age

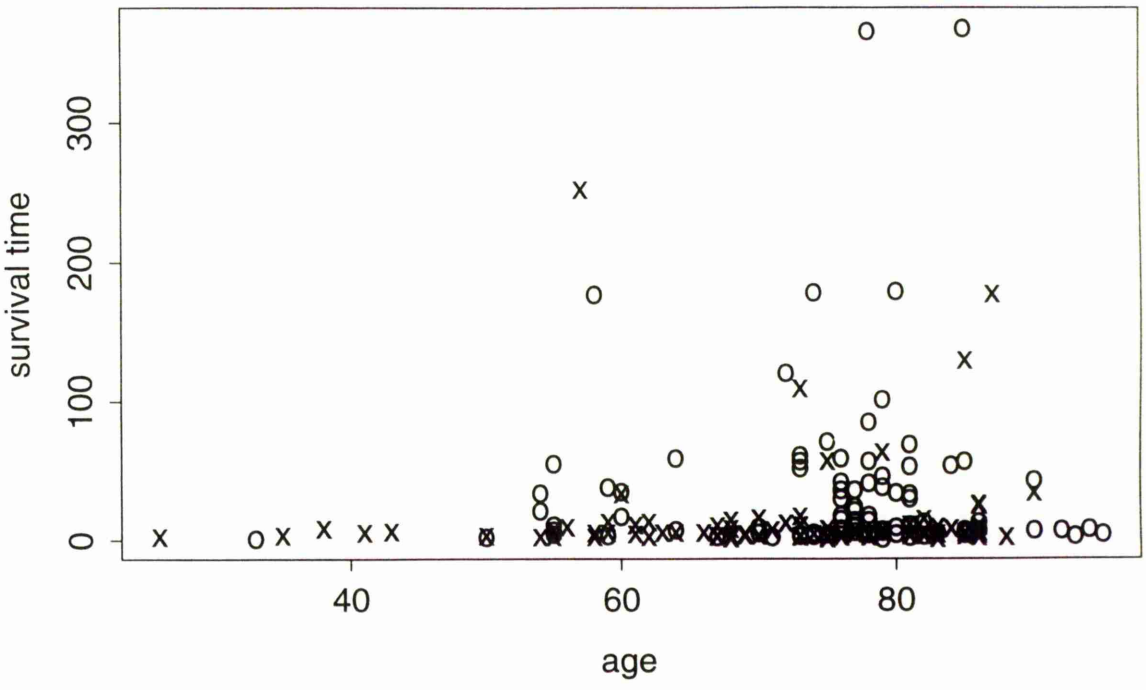
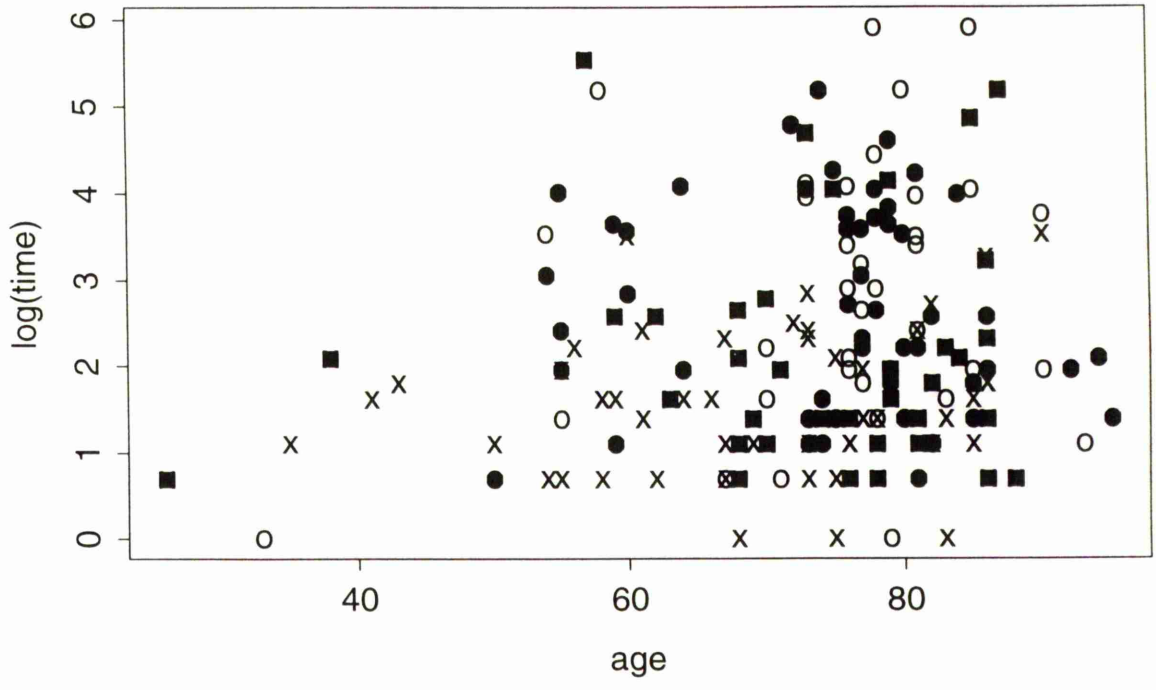



Figure 5.2 : Log(Survival time) vs. age



# Leg Ulcer data

Figure 5.3 : Smooth median curves

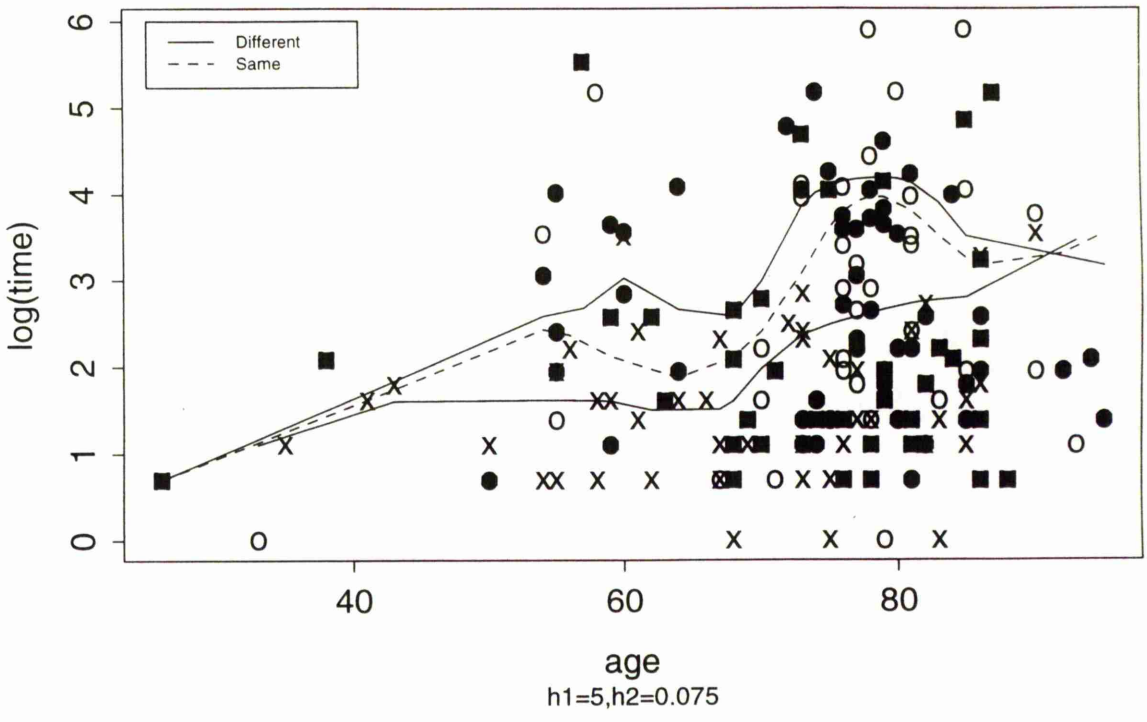
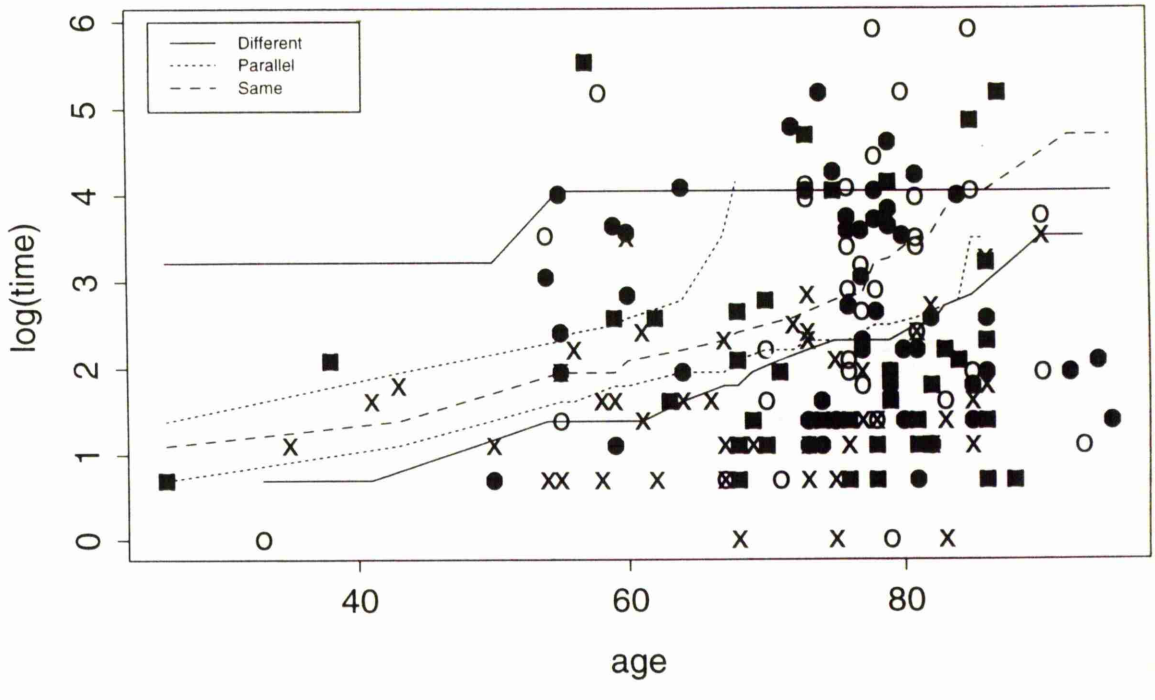



Figure 5.4 : PH median curves



| | Different | Parallel | Same |
|---|---|---|---|
| PH | -208.503 | -222.585 | -225.883 |
| SPH | -196.364 | -216.850 | -223.896 |
| NP | -177.387 | – | -206.701 |

Of course, these are relatively uninformative unless accompanied by an empirical p–value from an appropriate analytical procedure. Let us assume that the age covariate has an effect but it is unclear which model or curve description is most applicable. An analysis of these data is described below.

The first area of interest may be to compare whether, using the likelihood statistic corresponding to the doubly–smooth estimator, the relationship between age and survival time is the same independent of the occurrence of deep vein involvement. A permutation test provided an empirical p–value of 0.00, suggesting that the two groups should be looked at separately as their percentile survival times will be different. (From Figure 5.3, the percentiles for those experiencing deep vein involvement were higher.)

Now consider modelling the covariates under a smooth PH model. A bootstrap band to represent the doubly–smooth approach is found to be of the range (-191.66,-131.22) with the observed $SPHll$ value of -196.36. This would appear to imply that the SPH model is inappropriate here, which is perhaps not a very surprising conclusion in the light of the behaviour of the percentile curves in Figures 5.4 and 5.5 being slightly peculiar. However, to be more confident of this result, the doubly–smooth centile curves for the different levels of binary covariate were plotted with a large smoothing parameter of $h_1$=10. These curves can be seen in Figure 5.6. The curvature between the ages of 60 and 80 years appears to have been "ironed out" and the bootstrap reference band is now (-205.53,-146.92), which contains the observed $SPHll$ value. The shape of the doubly–smooth curves with $h_1 = 5$ suggest that the single values for the regression coefficients in the PH model would not summarise the data in precisely the same way. (The values of the coefficients in the PH model were in fact -0.023 and

-0.48 for age and deep vein involvement respectively and both were statistically significant.) However, as was observed in the test between these two approaches for the data involving deep vein involvement only in section 4.4.3, the likelihood statistics $NPll$ and $PHll$ do not exhibit signs of being significantly different (empirical p–value =0.26). Repeating this test for $h_1$=10, it is not surprising to find that the conclusion does not change and that the empirical p–value increases slightly to 0.28.

### Conclusions

The above was an attempt to explore an example where the methodology described in Chapters 3 and 4 was developed further. Due to the increased variety of models when a binary covariate is involved, relating the different relationships became rather complex. However, plots of time vs. covariate with appropriate percentile curves superimposed provide a good, visual explanation of some of the characteristics of the data–set. The resampling techniques provide a logical procedure for interpreting the covariate effects. This allows some subjective impressions to be strengthened in some cases, while protecting against the over-interpretation of possibly random feaures in other cases.

# Leg Ulcer data

Figure 5.5 : SPH median curves

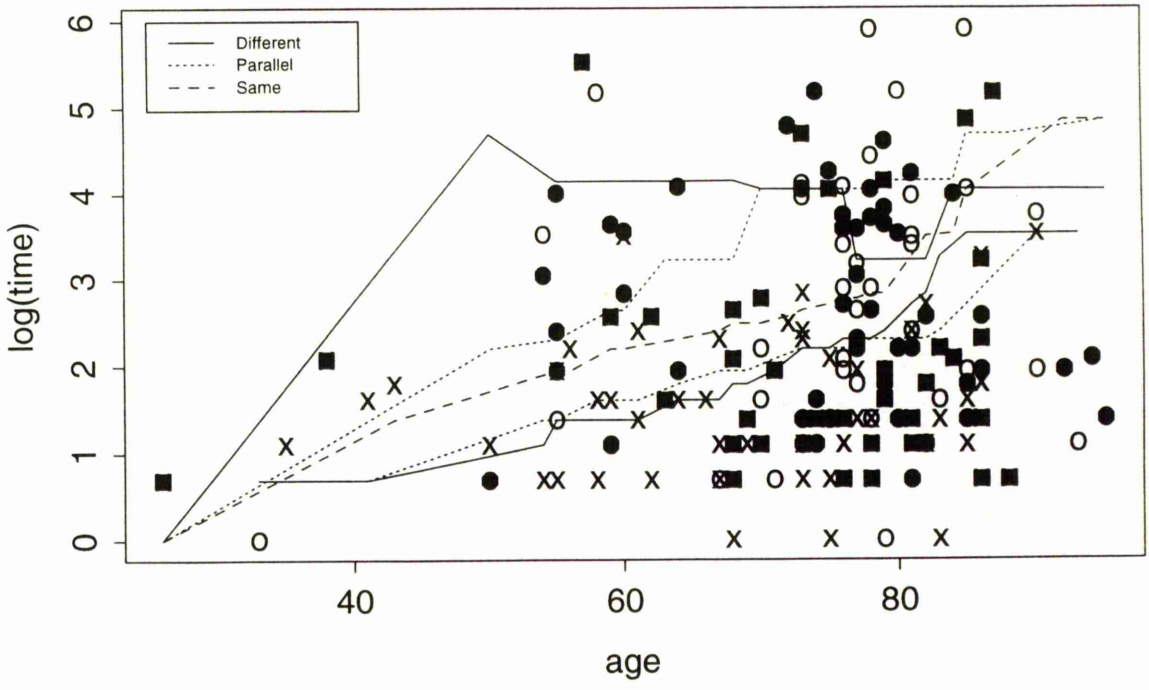
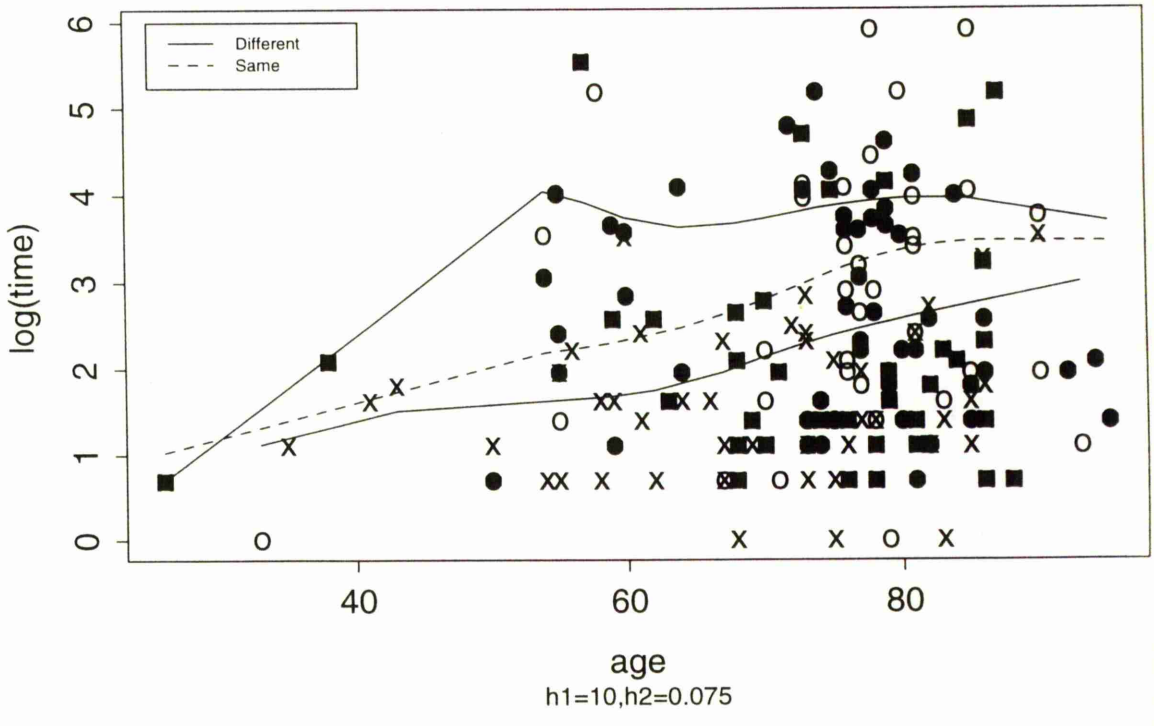



Figure 5.6 : Smooth median curves



# Chapter 6

# The Detection and Investigation of Progressive Graft Dysfunction in Kidney Transplant Patients

## 6.1 Introduction

Advances made in immunosuppresive therapy, most notably with the introduction of the drug cyclosporin in the early 1980s, has lowered the number of cases suffering from acute graft loss in the first year after transplantation. However, the rate of late graft loss, although much smaller initially, has remained virtually constant for the past five to ten years and is thus assuming a greater importance as a cause of transplantation failure. Studies of late graft loss have shown possible important factors to be the race of the patient, the number of previous transplants, the age of the donor and the degree of graft match (Mickey *et al.* (1990) and Gjertson (1991)). Collaborative research with Dr Peter Rowe and Dr Maureen Lafferty of the Renal Unit at the Western Infirmary has focussed on the occurrence of late graft loss in kidney transplant patients.

The main cause of late graft loss is known to be chronic rejection, although it is unclear what causes this condition. One possible explanation is that it is an immunological phenomenon and the infliction of graft injury is by a humoral or cellular process, resulting in graft destruction. Another theory is that a steady deterioration of the renal function occurs, caused by haemodynamic

changes compensating for a reduction in nephron mass. (A transplanted kidney will contribute less than half of the normal complement of functioning nephrons, due to unavoidable damage during the transplantation process.) A third possibility is that injury to the graft could have occurred by chronic cyclosporin nephrotoxicity, where "too large" a dose of the drug has a detrimental effect on the function resulting in renal failure (Salomon, 1992). It could also be that it is a combination of these factors which causes chronic rejection.

We are actually interested in progressive graft dysfunction which, although difficult to define, can be thought of as the condition of the graft during which steady loss of function occurs leading ultimately to failure.

**Objectives of the study**

These are :

1. to investigate what causes progressive graft dysfunction
2. to study which factors are peculiar to patients who have suffered late graft loss.
3. to help to decide upon the best form of treatment for this condition.

Firstly, we must find a method that will routinely detect progressive graft dysfunction. This was achieved by a statistical algorithm. The aim of this algorithm is to determine which grafts have suffered progressive graft dysfunction, and when this occurred.

## 6.2 Detecting Progressive Graft Dysfunction

### 6.2.1 Previous Literature

All the papers published on this topic agree that when progressive graft dysfunction occurs, the negative reciprocals of the increasing body-weight-adjusted

serum creatinine levels generally increase linearly with time. The examination of plots of creatinine over time, and of the negative reciprocal of creatinine over time, are routine tools employed by doctors. Experience has allowed them to determine from this information when a graft has suffered from progressive graft dysfunction.

One reason for examining the negative reciprocal of creatinine levels is that reliance on the non–transformed levels over time alone leads to small changes in the serum creatinine at the lower end of the range being overlooked and their importance underestimated.

A more formal justification is derived from consideration of the glomerular filtration rate (GFR) which forms the basis of kidney excretary function. The definition of GFR is the total volume of plasma water ultrafiltered through nephrons. These values are usually quoted in units of ml/min and corrected for body surface area. Calculation comes from the Fick principle,

$$\text{GFR} = \frac{\text{amount of substance filtered}}{\text{concentration in plasma} \times \text{time}}$$

The substance employed should be freely filtered by glomerulus and neither reabsorbed nor excreted by the remainder of nephron. Creatinine is derived from the turnover of the skeletal muscle protein creatine and just about satisfies these criteria. The GFR can now be written as

$$\text{GFR} = \frac{U_{Cr} \times U_{vol}}{S_{Cr} \times \text{time}} \qquad \text{where} \quad \begin{array}{lcl} U_{Cr} & = & \text{Urine creat. conc.} \\ U_{vol} & = & \text{Urine volume} \\ S_{Cr} & = & \text{Serum (plasma) creat. conc.} \end{array}$$

In a steady state, the creatinine excretion ($U_{Cr} \times U_{vol}$) is constant, providing the relationship

$$\text{GFR} \propto \frac{1}{S_{Cr} \times \text{time}} \qquad \text{or,} \qquad \text{GFR} \propto \frac{1}{S_{Cr}}$$

As a measure of excretary function, these values also give an indication of the condition of the graft.

Most of the formal statistical techniques used to detect the event of chronic rejection are on-line or predictive procedures (as opposed to the retrospective detection we are interested in). They are mainly Bayesian approaches, such as

the Kalman filter(Smith and West, 1983) and the intersection of two lines(Smith and Cook, 1980) although ideas of ROC curves(Smith *et al.* 1983) and CUSUM charts(Stoodley and Mirnia, 1979) have also been explored. The techniques are generally mathematically complex. The aim of this study was to evaluate the effectiveness of simple linear methods in the setting of retrospective data.

The basis of the algorithm is the fact that deterioration (signalling the onset of progressive graft dysfunction) is displayed in a markedly positive slope in a plot of the negative of the reciprocal of serum creatinine levels against time. Some previous work used body-weight-adjusted levels but since the present study is concerned with change and since body weight is unlikely to vary much over the time periods considered, the adjustment was not adopted.

### 6.2.2 Different Classifications

The algorithm must be able to differentiate between the possible ways in which a graft can function and events that can occur. The four main graft types to be recognised are:

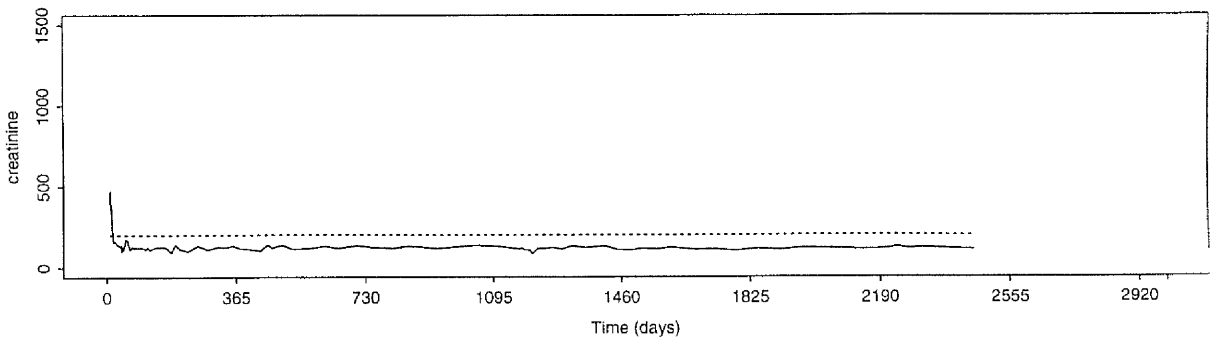
1. stable, good function
2. stable, poor function
3. the occurrence of progressive graft dysfunction (PGD)
4. the occurrence of acute graft loss (AGL)

Typical examples of the plots of interest for each classification are shown in Figures 6.1–4.

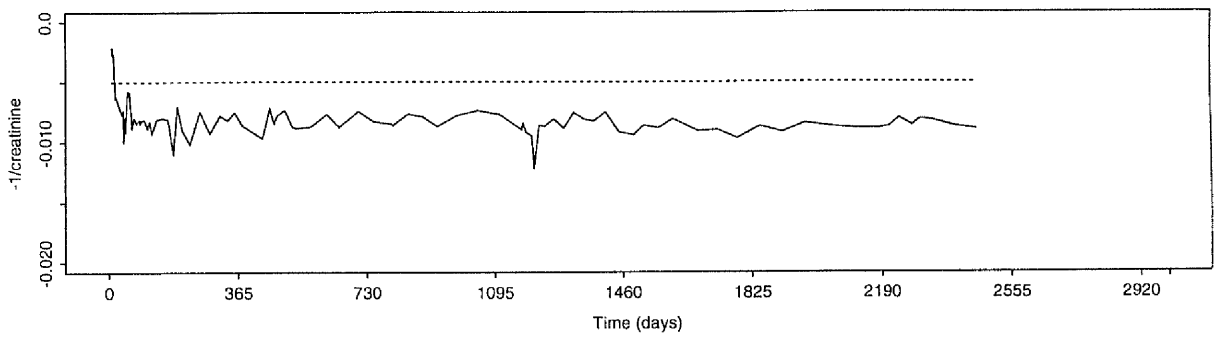
The horizontal dotted line indicates the serum creatinine level of 200$\mu$moles /litre, sometimes used to define a threshold between good and poor functions that are otherwise stable. In each pair, the top figure displays serum creatinine and the bottom figure displays -1/(serum creatinine).

# Example serum creatinine traces

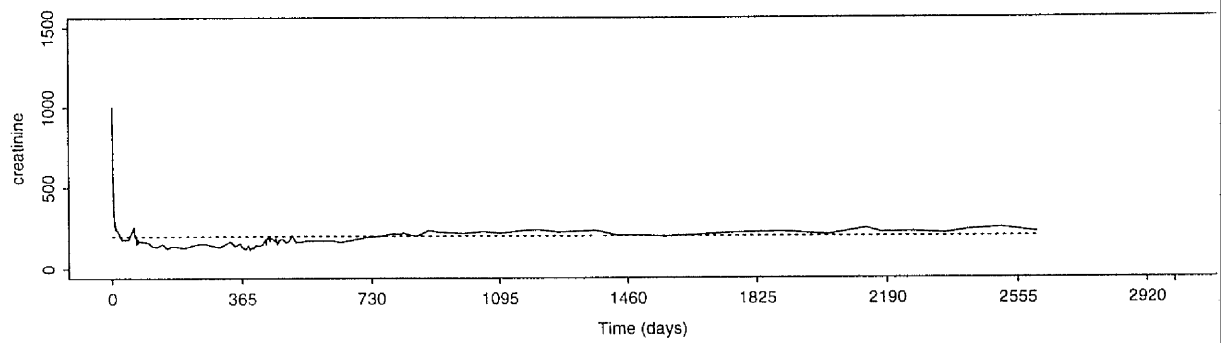
## Figure 6.1 : Hospital number 1



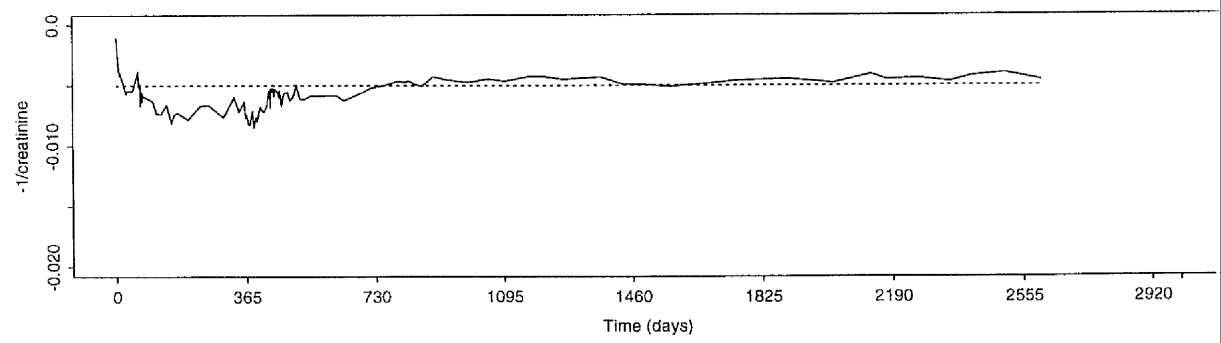

## Hospital number 1




## Figure 6.2 : Hospital number 2

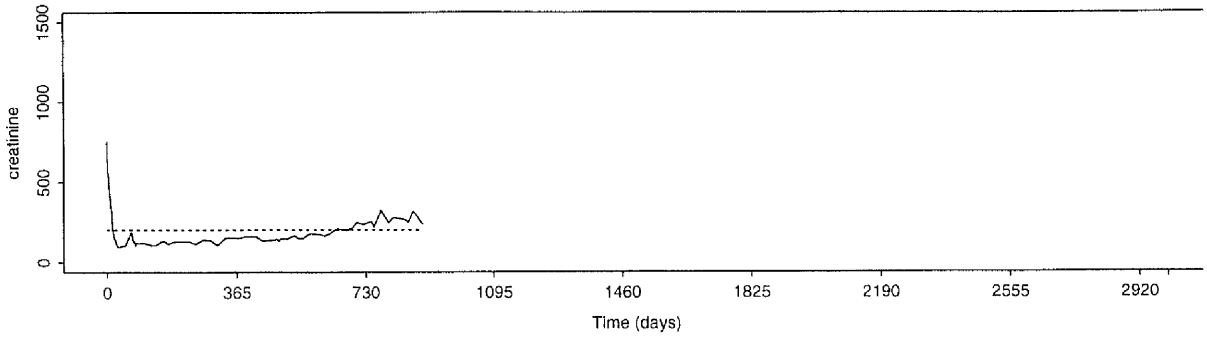



## Hospital number 2



# Example serum creatinine traces

Figure 6.3 : Hospital number 3




Hospital number 3

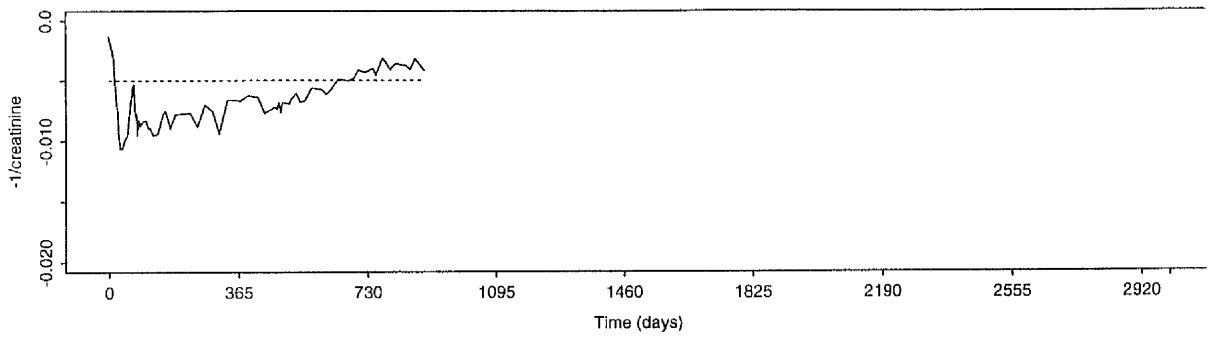
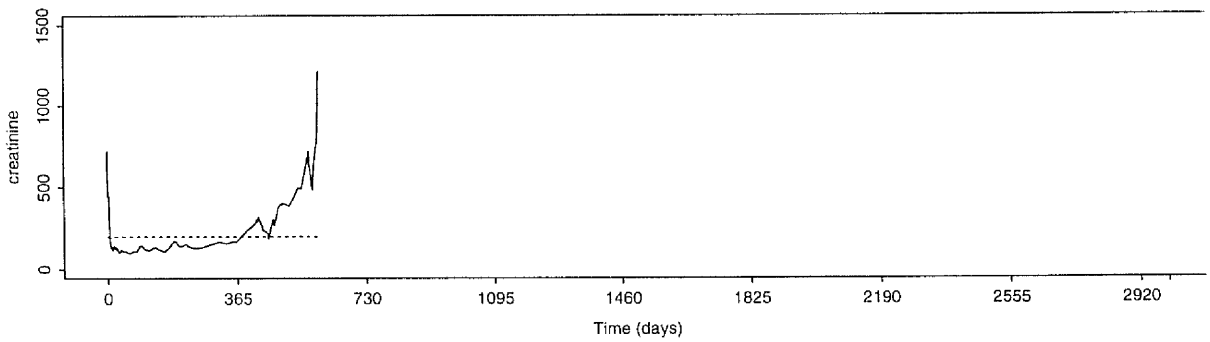
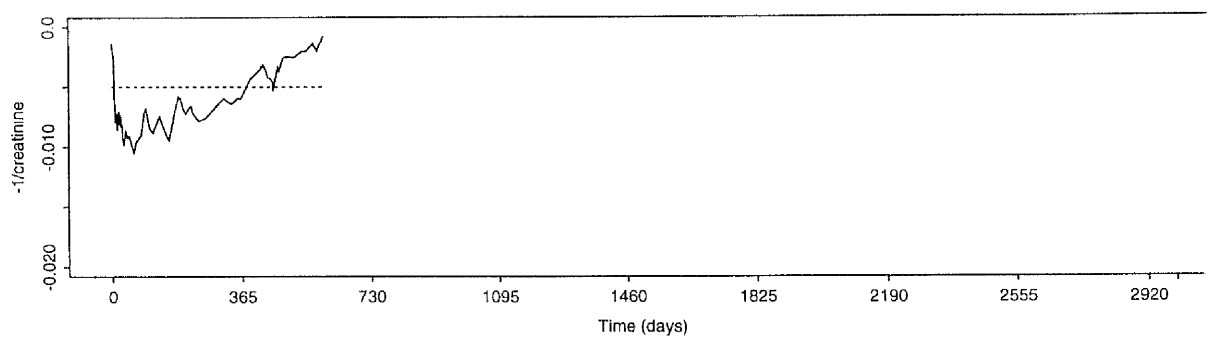



Figure 6.4 : Hospital number 4




Hospital number 4



In all of these traces, serum creatinine levels are initially very high, which is normal after a transplant, but they decrease very quickly in the first few weeks following the operation. Ignoring this typical effect, the key features of each set of plots are :

| | | |
|---|---|---|
| Figs 6.1 a & b | – | a relatively smooth trace with very little increase in creatinine levels. **stable, good** (hospital number – 1 –) |
| Figs 6.2 a & b | – | the creatinine levels are higher than in the previous case but the graft is still exhibiting signs of stability. **stable, poor** (hospital number – 2 –) |
| Figs 6.3 a & b | – | initially fairly smooth with the slope steadily increasing, showing deterioration **PGD** (hospital number – 3 –) |
| Figs 6.4 a & b | – | increasing creatinine levels with a very sharp rise towards the end. **AGL** (hospital number – 4 –) |

### 6.2.3 Detecting Slope Change Over Time

The following method was implemented for each individual graft :

A grid of time values is set up 30 days apart, starting 90 days after the time point corresponding to the lowest creatinine level in the first 90 days (to allow for high creatinine levels after transplantation).

Windows of 180 days are centred round the grid points and for each grid point, the slope of the straight line fitted through the points that lie in that window is evaluated by least squares estimation. Bands of width $\pm 2$ standard errors are then calculated and the slope is deemed to be significantly positive if the lower band is greater than zero.

The data in this problem constitute a time series and so the fact that the errors are correlated should be taken into account. This was investigated by examining sections of data where the process was stationary. In all of these

cases, the autocorrelation was small and so this effect was outweighed by other, more major, problems in the construction of a detection system, as described below.

Another issue to be considered is that of multiple comparisons since, where a large number of tests are carried out, the chances of a spurious result are high. There are two simple ways of overcoming this. One is to use a larger width for the confidence bands; for example $\pm 3$ standard errors might be used. However, it was decided that a more appropriate strategy would be to look for evidence of deterioration in several consecutive windows. This will reduce the risk of a misclassification which could otherwise occur on the basis of one result.

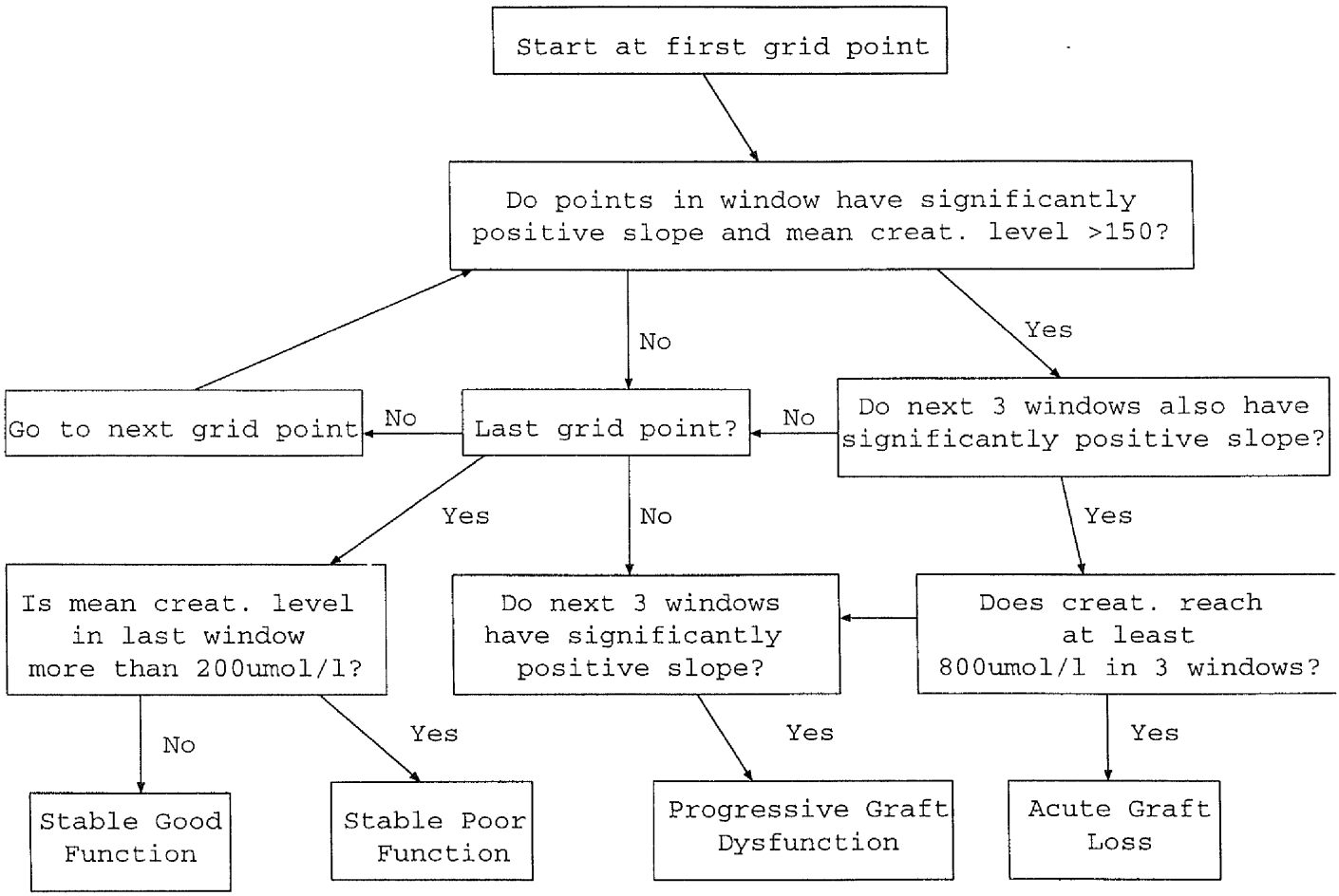
Now that a way of assessing change in slope over time has been established, a number of issues remain to be addressed.

- For how long must the slope be positive for progressive graft dysfunction to be concluded?
- How will acute graft losses be differentiated from progressive graft dysfunction cases?
- When will a stable function be described as poor, as opposed to good?

### 6.2.4 The Algorithm

The initial construction of the algorithm was based on the ideas and recommendations of doctors from the Renal Unit. Amendments and improvements were made on the basis of the true classifications (as decided by Drs Rowe and Lafferty) of a random sample of 100 grafts.

The final version of the algorithm can best be described by the following flow diagram.




### 6.2.5 Classifications

The features of this algorithm which address the outstanding issues identified earlier are as follows.

- For progressive graft dysfunction to be identified there must be a significantly positive slope in seven consecutive windows, corresponding to deterioration over a 1-year period. Where there is insufficient data to construct seven consecutive windows a smaller number may be used, but there must be at least three windows.

- After a fourth consecutive window has resulted in a significantly positive slope, we assess the creatinine levels by looking for a value $>800\mu$moles/litre in those windows. Cases which satisfy these criteria are said to have suffered acute graft loss. This takes into account the high levels and rapid deterioration in acute graft loss.

- The distinction between stable good and poor functions was made by calculating the mean of the serum creatinine levels in the last window. If this value was $\geq 200\mu$moles/litre, then a stable, poor graft was concluded and otherwise stable, good.

### 6.2.6 Time Points

So far classification of grafts into four groups has been discussed but to be able to perform further analysis on the results, it would be useful to have a time corresponding to the event of interest.

For stable functions, we use the last time point as the time of interest.

When a decision of progressive graft dysfunction or acute graft loss has been reached then the time of interest is the value of the grid point corresponding to the first window in which a significantly positive slope has been reached with the appropriate number of consecutive windows also having significantly positive slope.

### 6.2.7 Examples

All this information can now be conveyed conveniently in three plots :

- Time vs. creatinine
- Time vs. -1/creatinine
- Time vs. slope coefficients(with $\pm 2$ standard error bands)

The typical examples of each classification, which were used earlier are shown with the additional graph in Figures 6.5–8. The vertical line denotes when progressive graft dysfunction or acute graft loss has occurred.

### 6.2.8 Results of the Algorithm on the Test Set

As would be expected with this type of analysis, the algorithm does not always deduce the status correctly. However, in the test data available 91 cases were correctly classified and only 9 cases were misclassified. The complete results are displayed in the following table.

| | | Algorithm | | | |
|---|---|---|---|---|---|
| | | stable,good | stable,poor | PGD | AGL |
| True Status | stable,good | 70 | 0 | 2 | 0 |
| | stable,poor | 0 | 4 | 2 | 0 |
| | PGD | 1 | 2 | 15 | 1 |
| | AGL | 0 | 1 | 0 | 2 |

Table 6.1: Classifications of Test Cases

The main area in which we would like to see improvement is in the detection of PGD. However, the four misclassified PGD cases are spread across all the other categories and so it will be difficult to pinpoint the problem, especially since four non–PGD cases of different types have been classified PGD.

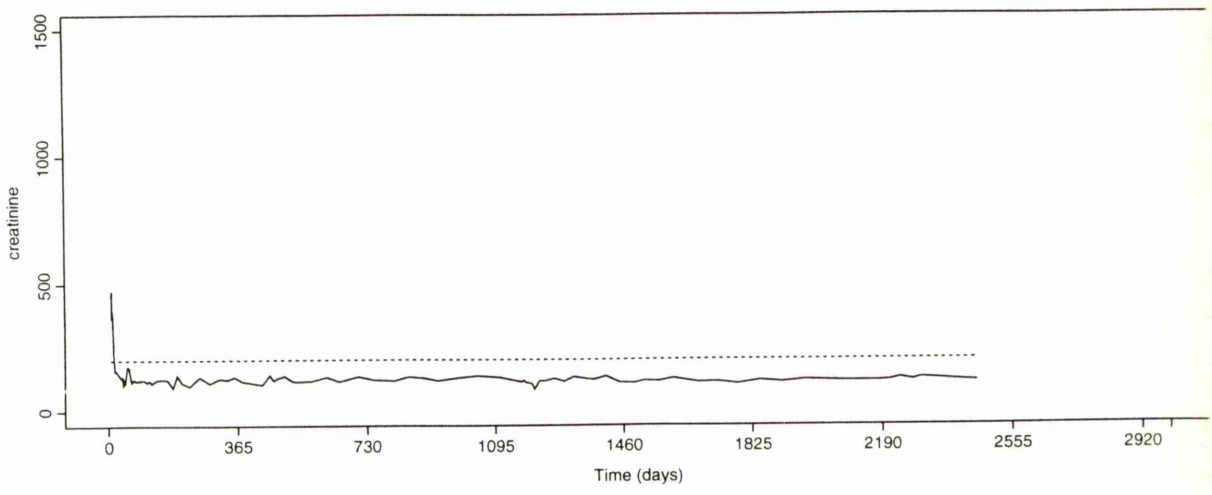
Figures 6.9–12 show four cases where medical opinion has disagreed with the classifications made by the algorithm. These help to highlight some of the difficulties in achieving improved performance.

Looking at the trace of hospital number – 5 –, we see that there is quite a sharp rise in creatinine level which then appears to stabilise, producing an approximate S-shape. The doctors describe this as a stable, poor function. However, the algorithm has detected the rise and finds six subsequent, consecutive windows which have (only just) significantly postive slope.

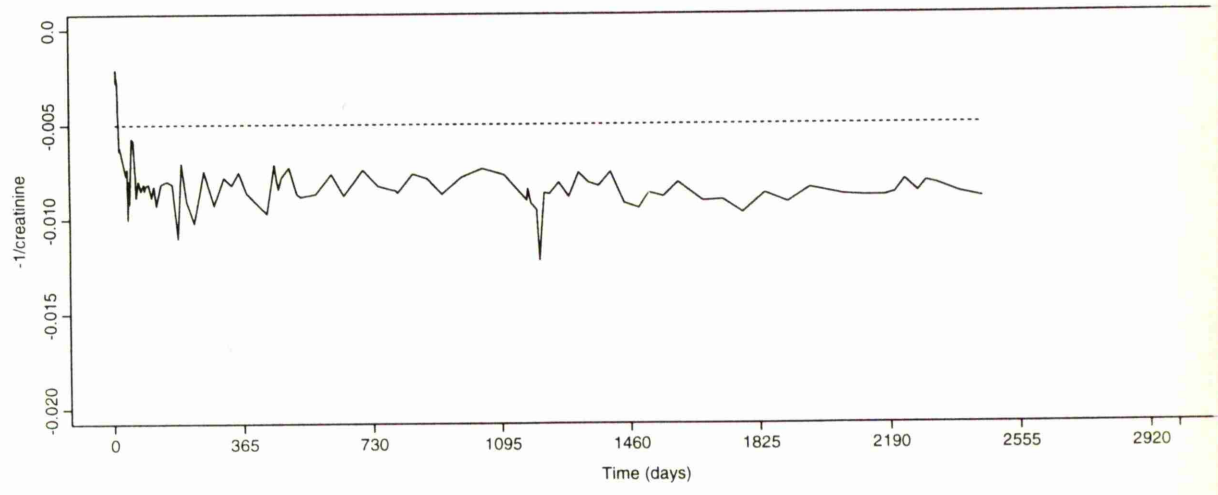
In the case of hospital number – 6 –, the short spikes in the trace mean that

# Example of stable, good trace

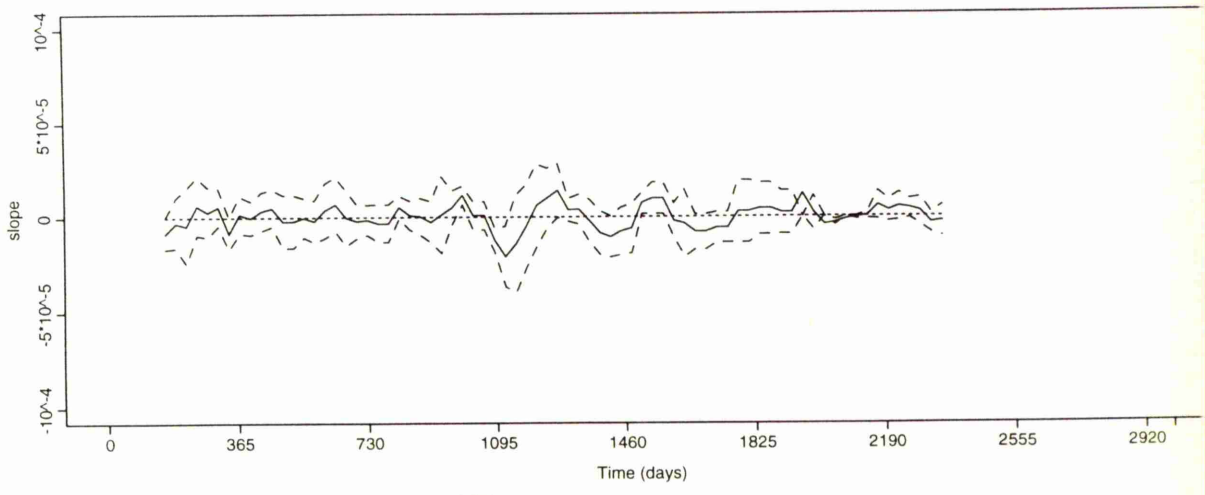
Figure 6.5 : Hospital number 1




Hospital number 1

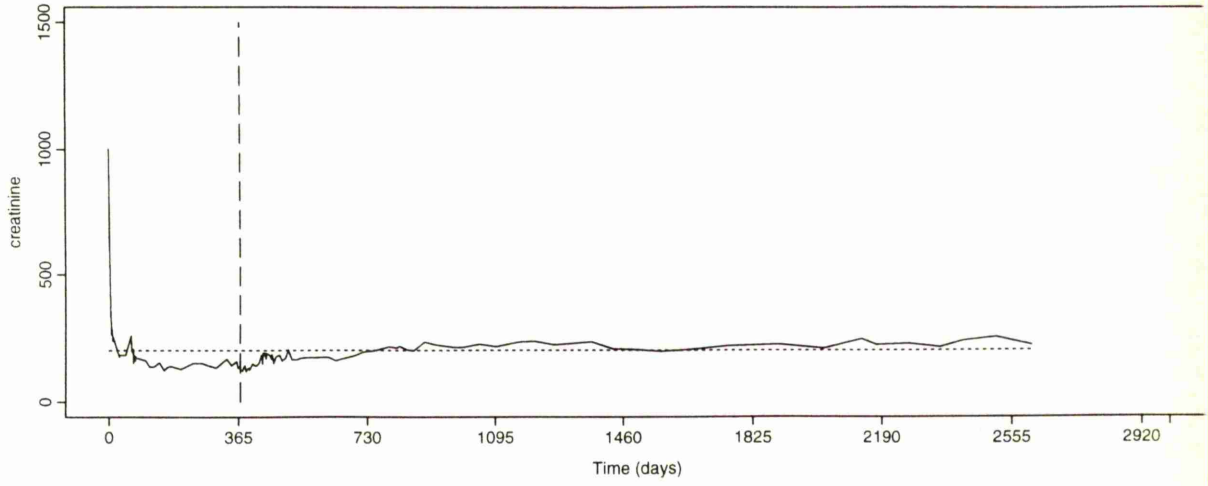



Hospital number 1

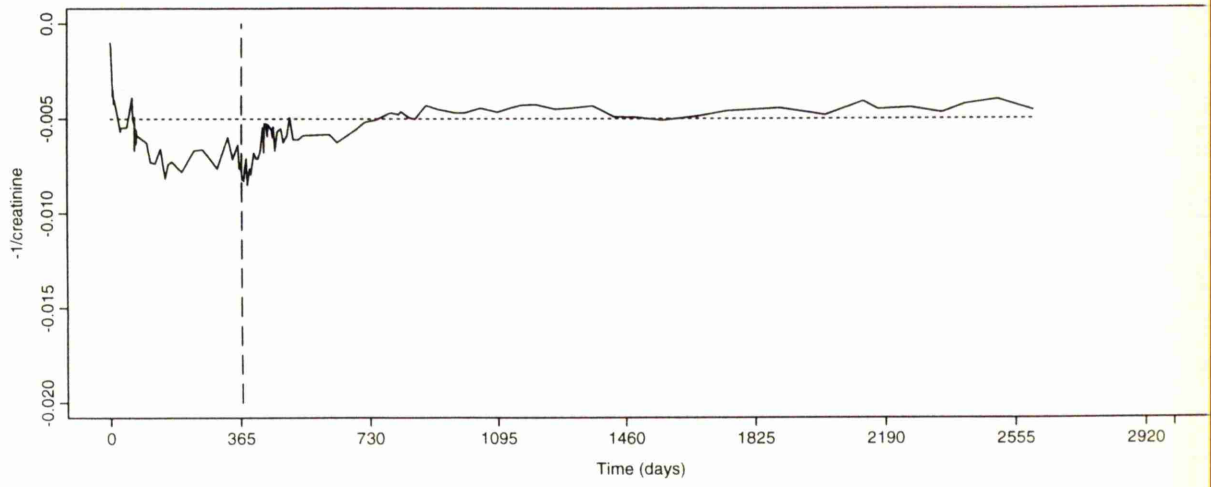


# Example of stable, poor trace

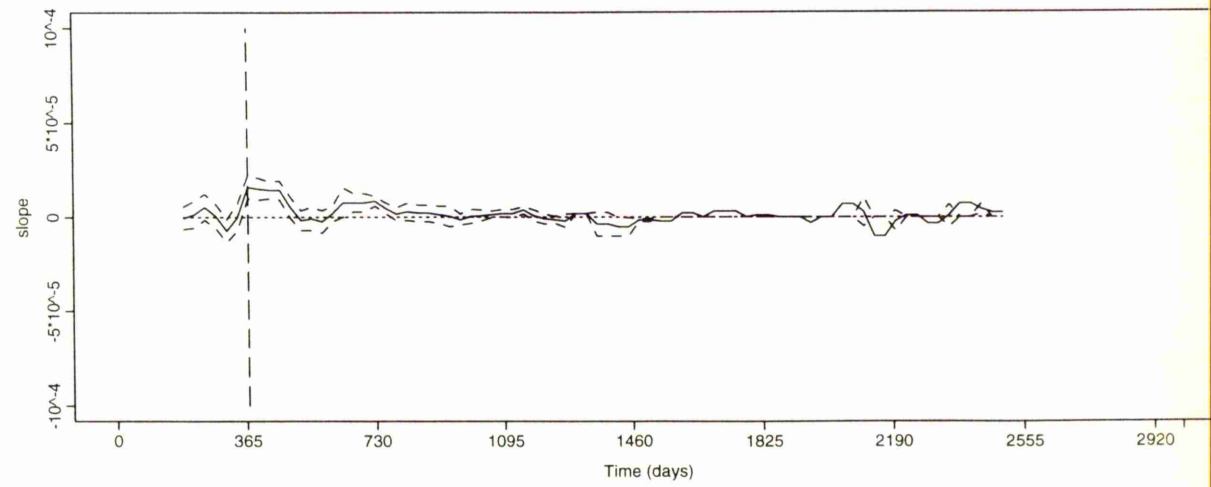
Figure 6.6 : Hospital number 2




Hospital number 2

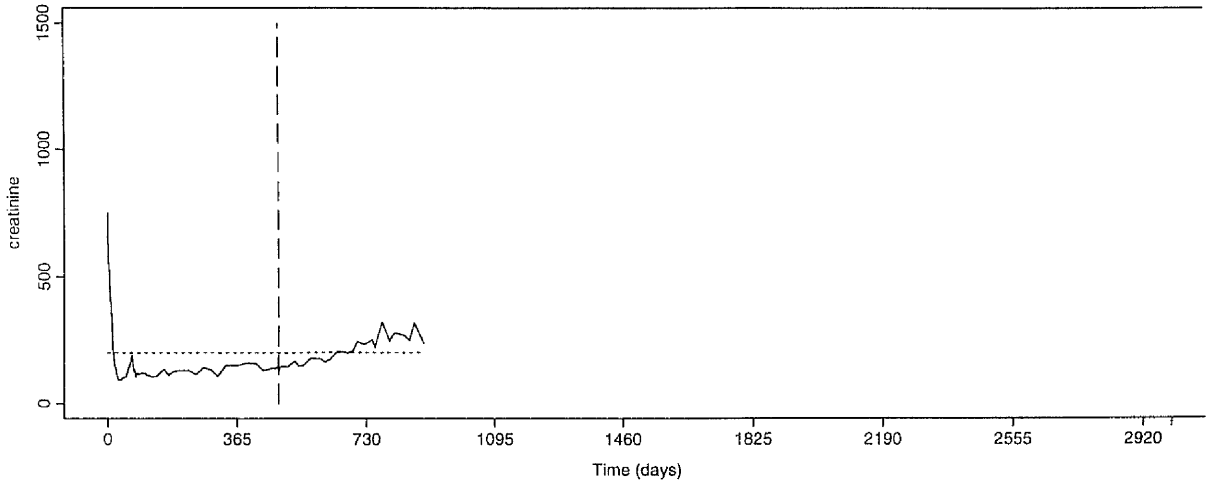



Hospital number 2

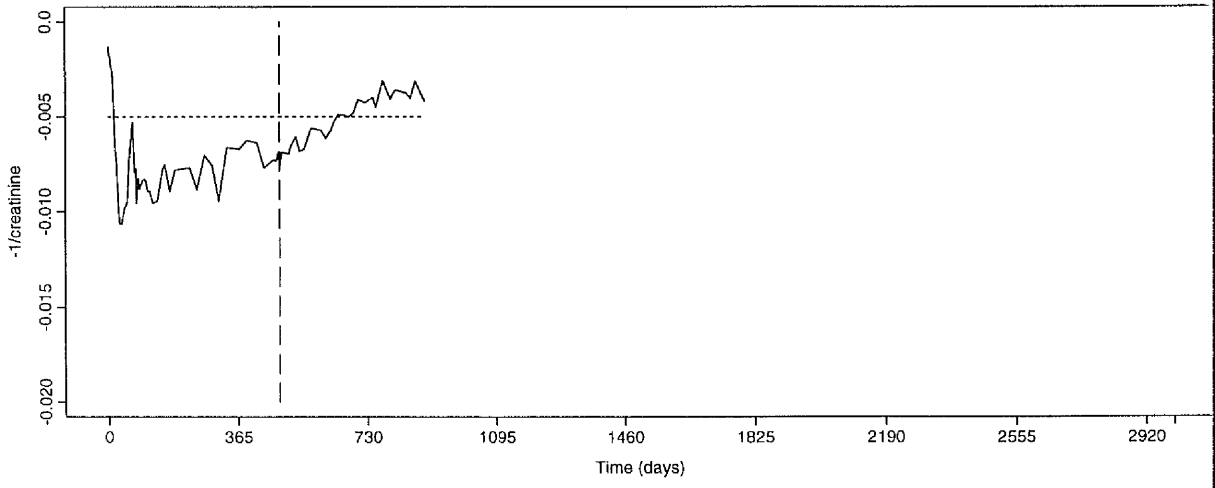


# Example of trace for PGD

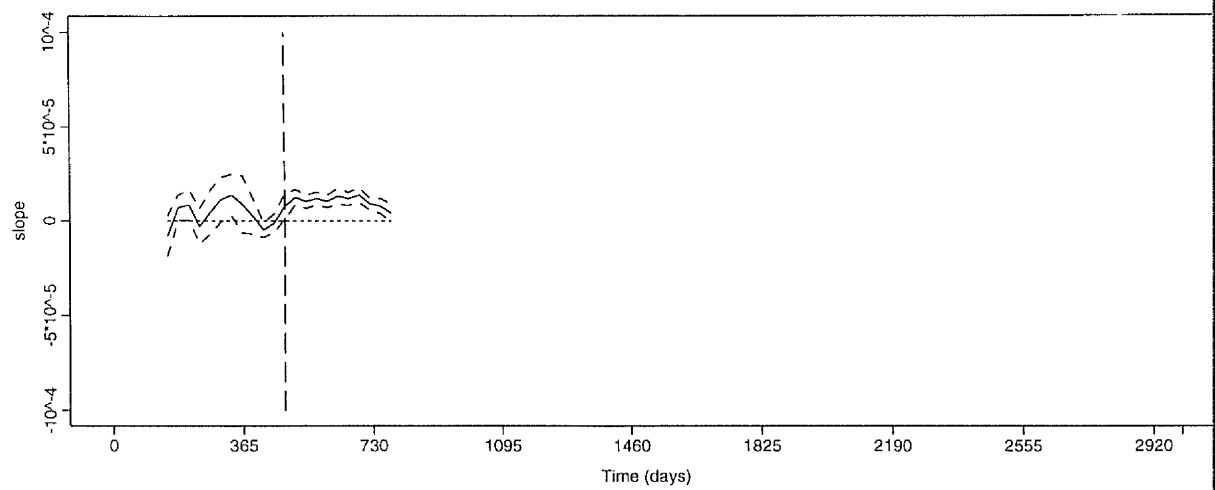
Figure 6.7 : Hospital number 3




Hospital number 3

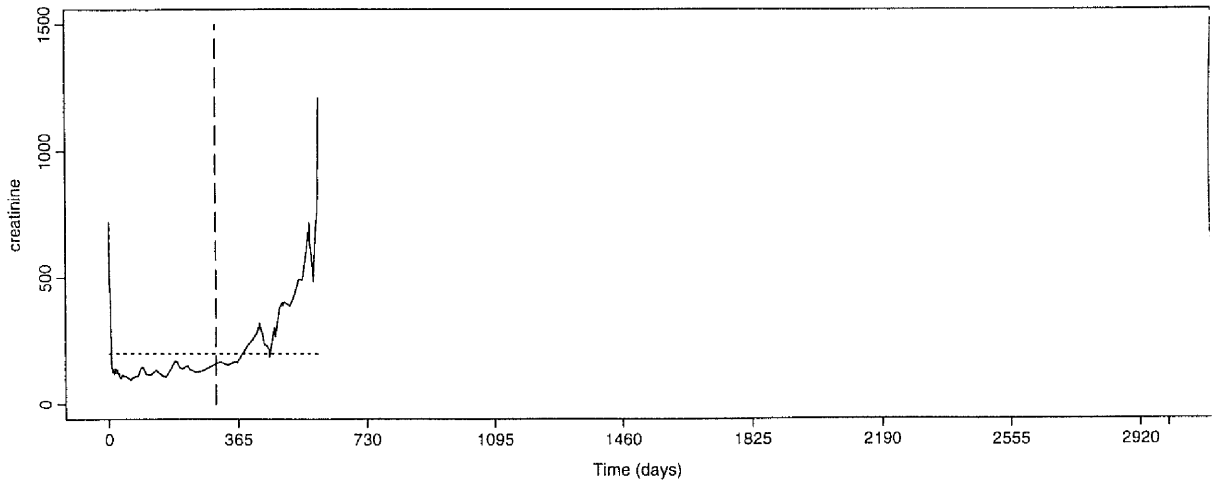



Hospital number 3

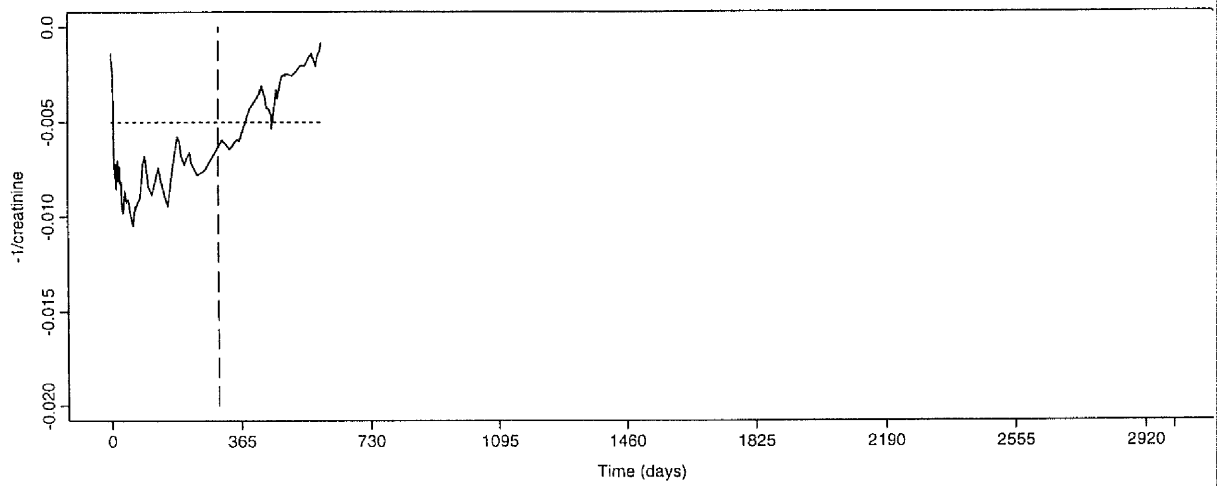


# Example of trace for AGL

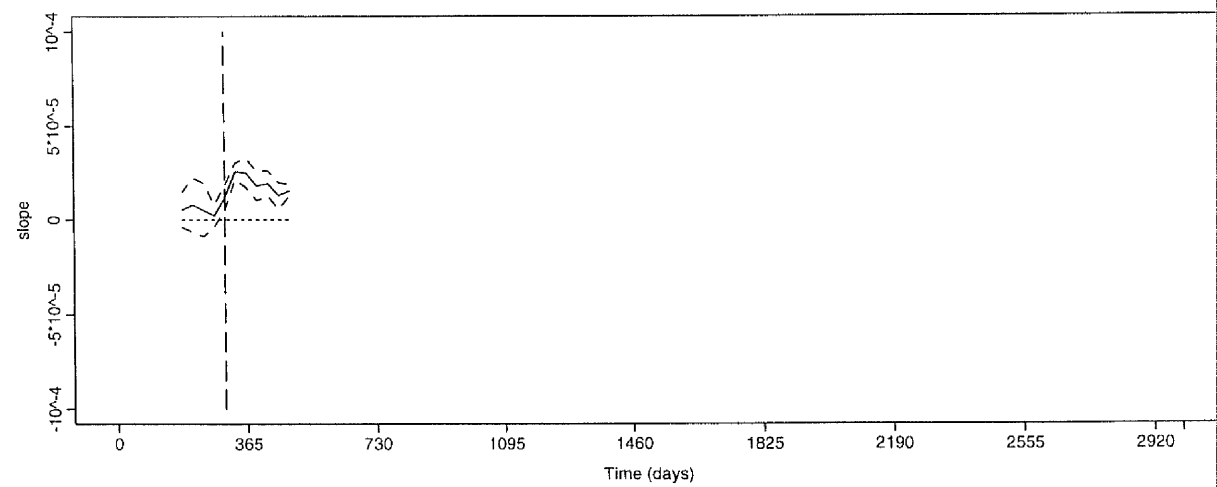
Figure 6.8 : Hospital number 4




Hospital number 4




Hospital number 4



the slope is not positive for a long enough period of time and is being classified as stable, poor instead of its true status PGD.

The trace of hospital number – 7 – is positive just long enough for PGD to be detected instead of being classified as stable, poor but here the values of the lower band are extremely close to zero.

The graft with hospital number – 8 – is said by the medics to be a case of progressive graft dysfunction but we suspect that short spikes and the sparseness of the data are the reasons for the algorithm concluding a stable, good status.

On the whole, the algorithm seems to perform well. It is of interest to apply it to more data and to continue with further analysis. Before doing so, a few notes on problems and possible amendments should be made.

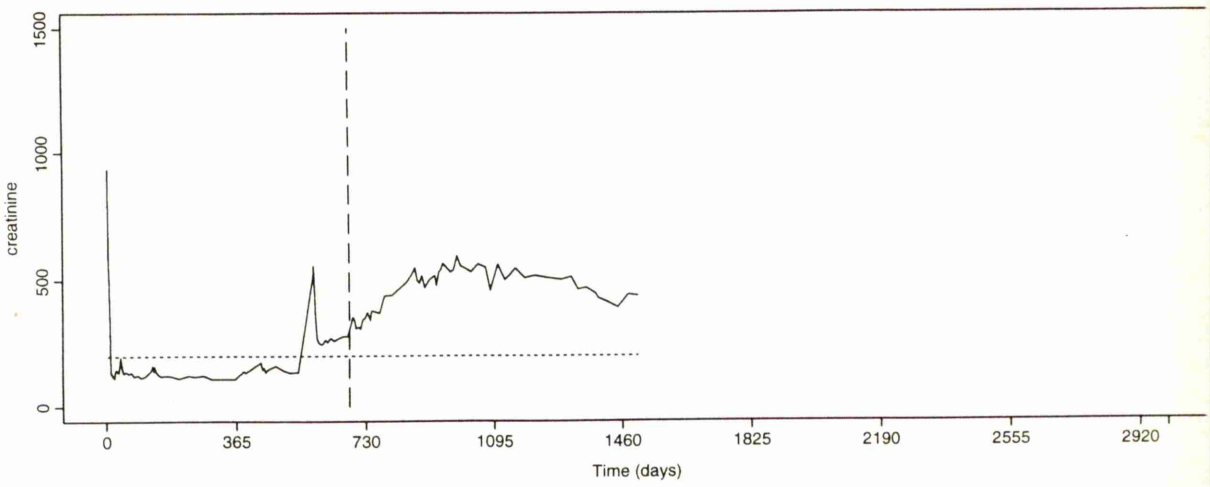
Firstly, the detection of the S-shape curve could be examined by looking at the remaining data after detection of PGD and re-running the algorithm on this. However, when this was tried, there was no further improvement in the success rate already achieved and further adjustments were not attempted.

Some acute graft losses are still going undetected. There may be a need to be more specific about the size of jumps in creatinine levels. However, this would create yet another area requiring subjective tuning of the algorithm.

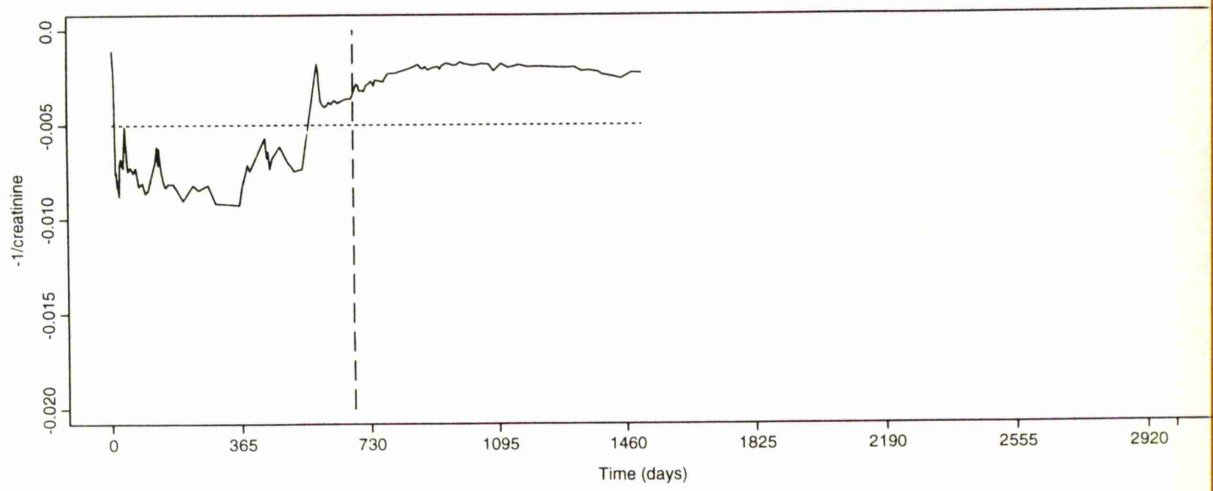
The effect of short spikes in the trace to the slope coefficient could be lessened by using robust regression. This will be more difficult to implement, and could affect those classifications already specified.

# Example of misclassification

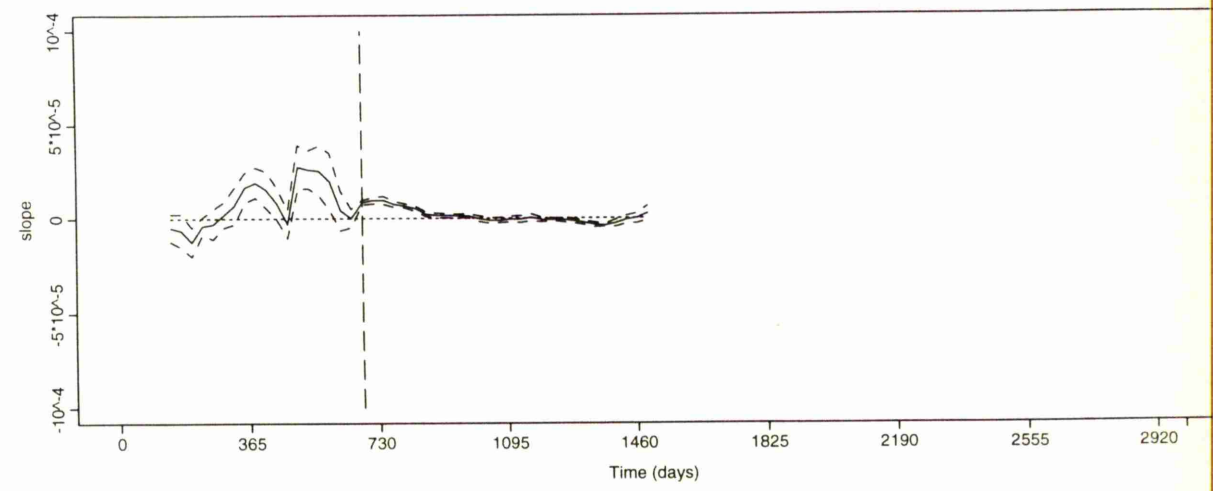
## Figure 6.9 : Hospital number 5




hospital number 5

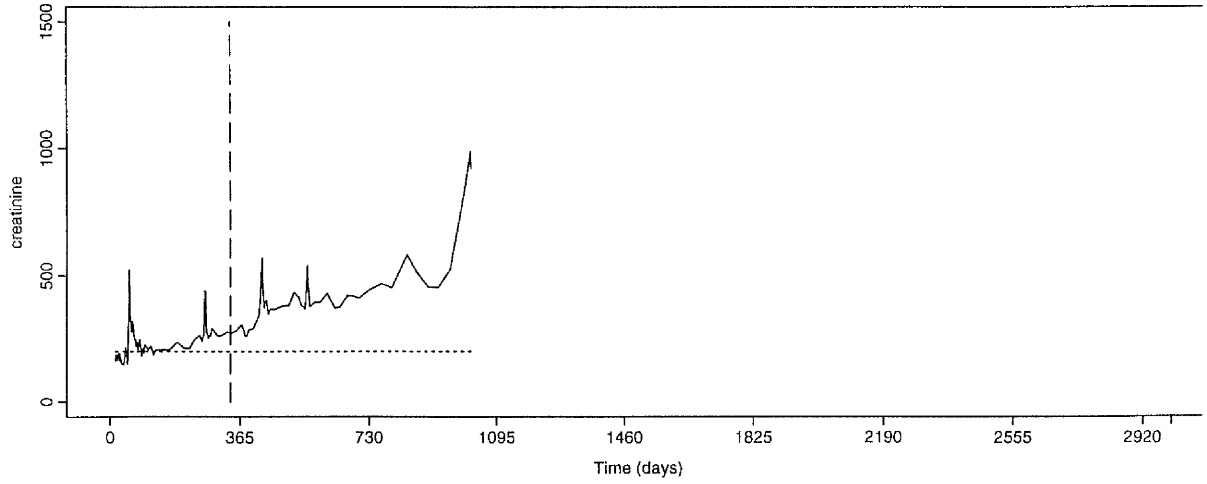



Hospital number 5

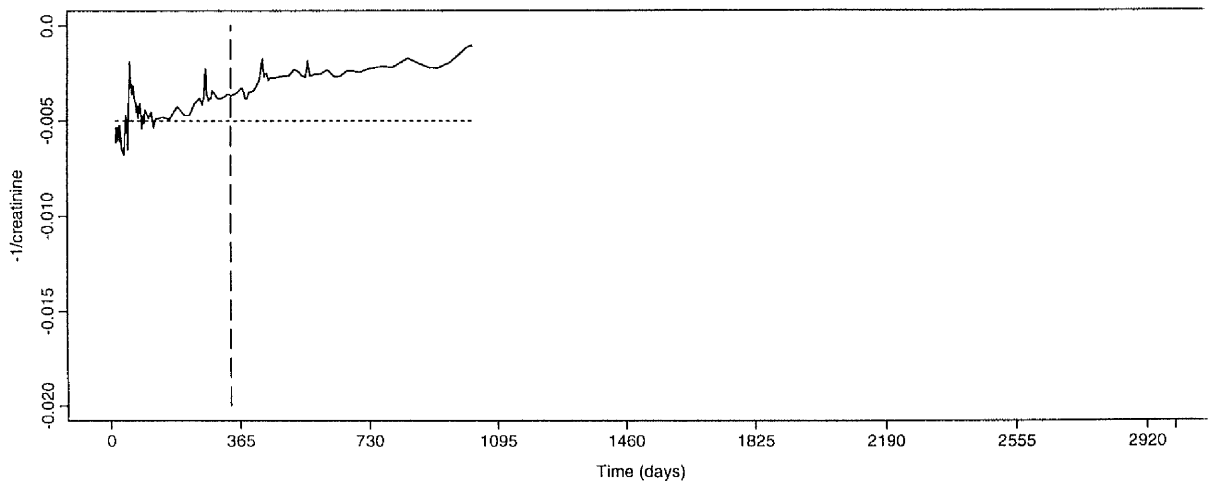


# Example of misclassification

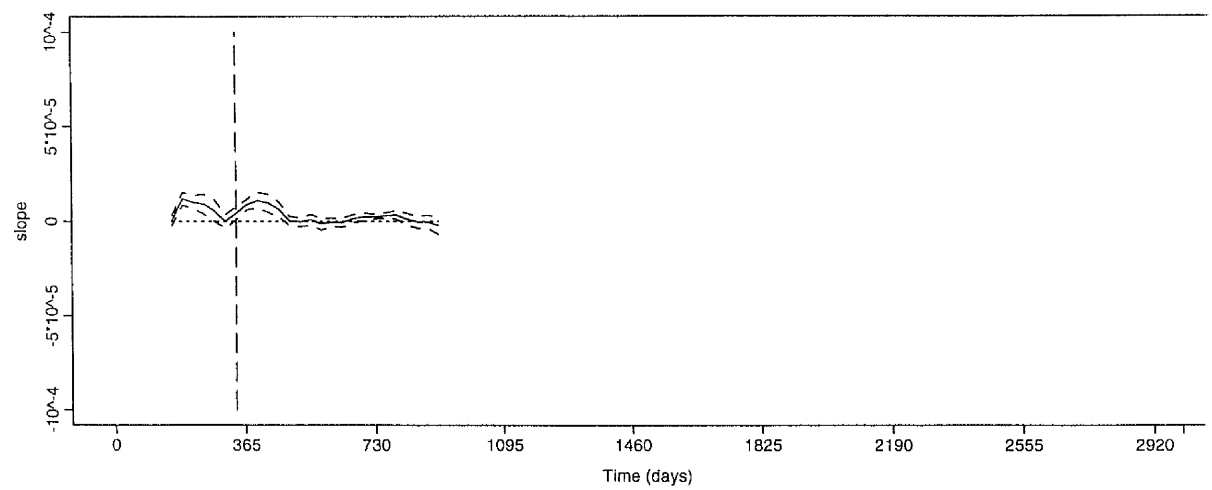
Figure 6.10 : Hospital number 6




Hospital number 6

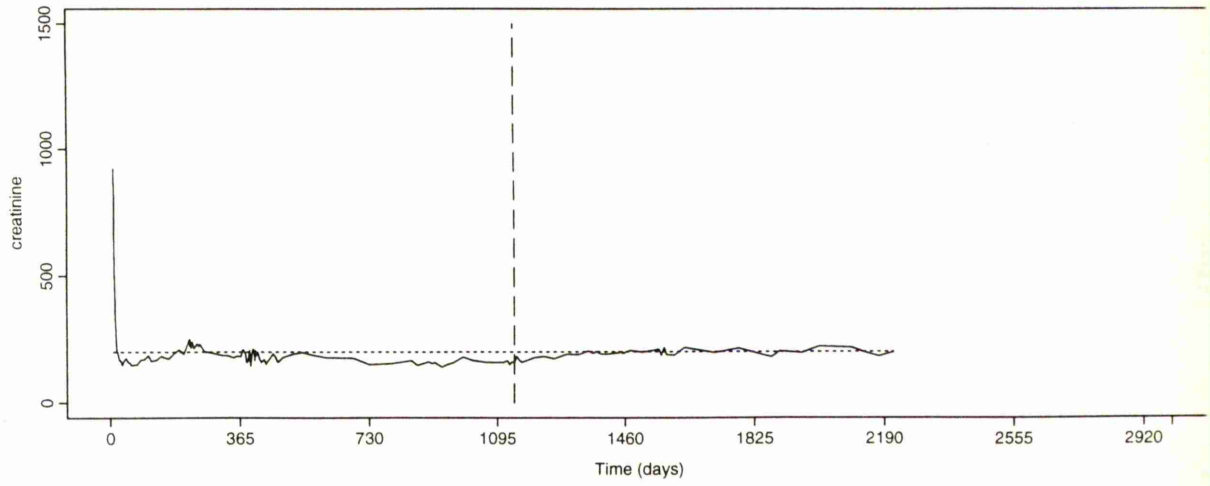



Hospital number 6

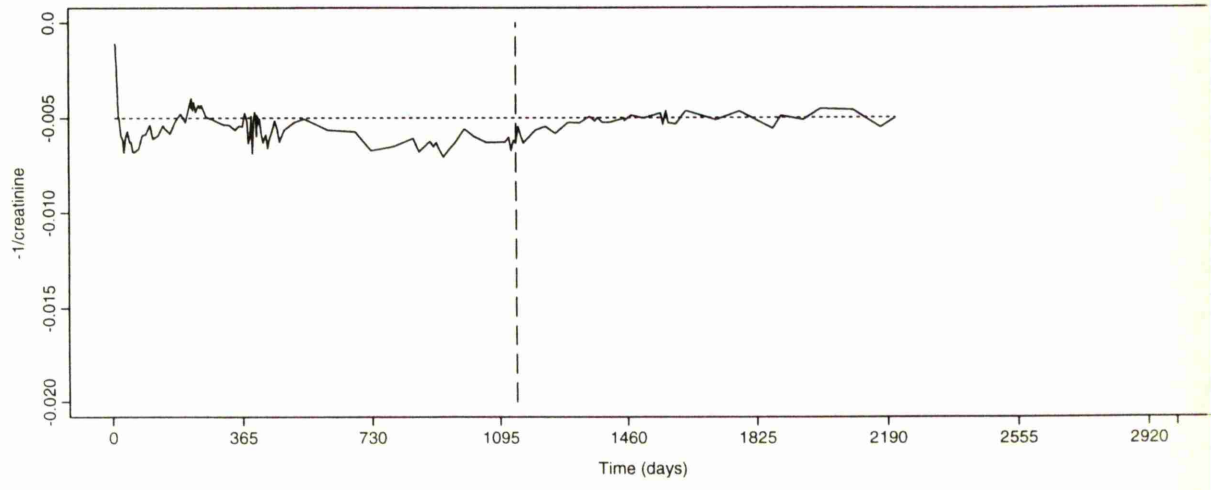


# Example of misclassification

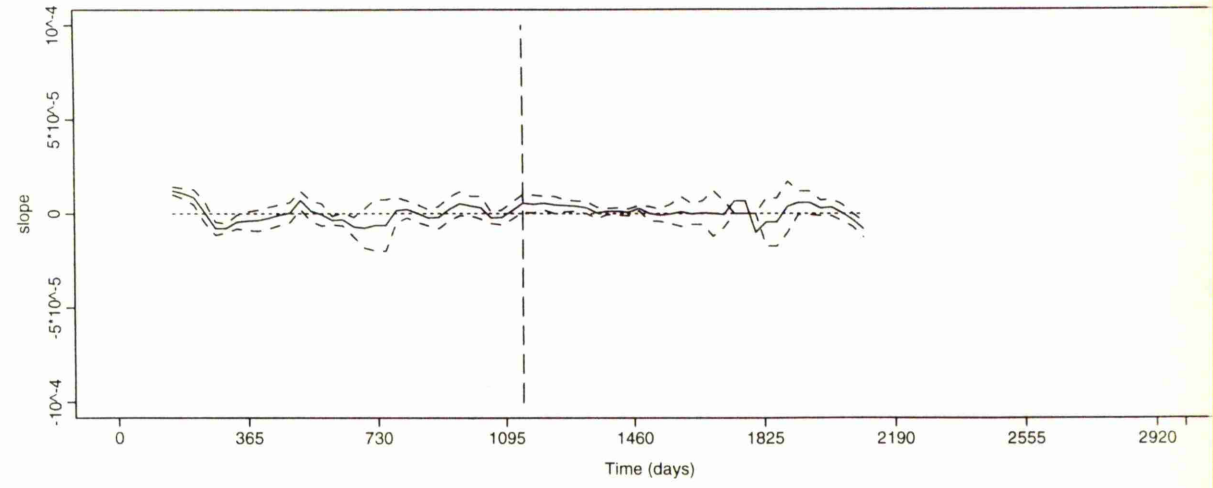
## Figure 6.11 : Hospital number 7




Hospital number 7

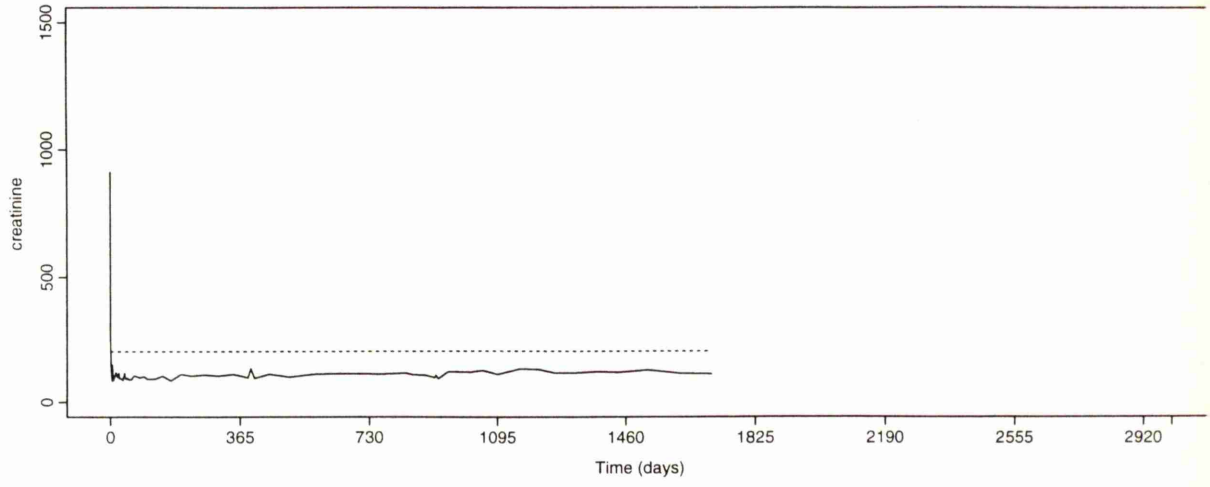



Hospital number 7

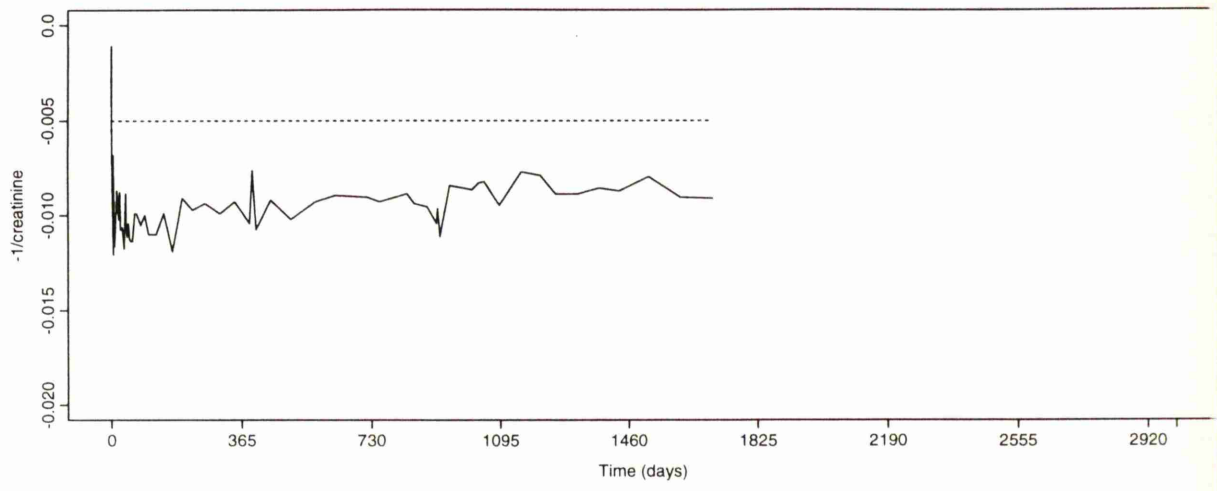


# Example of misclassification

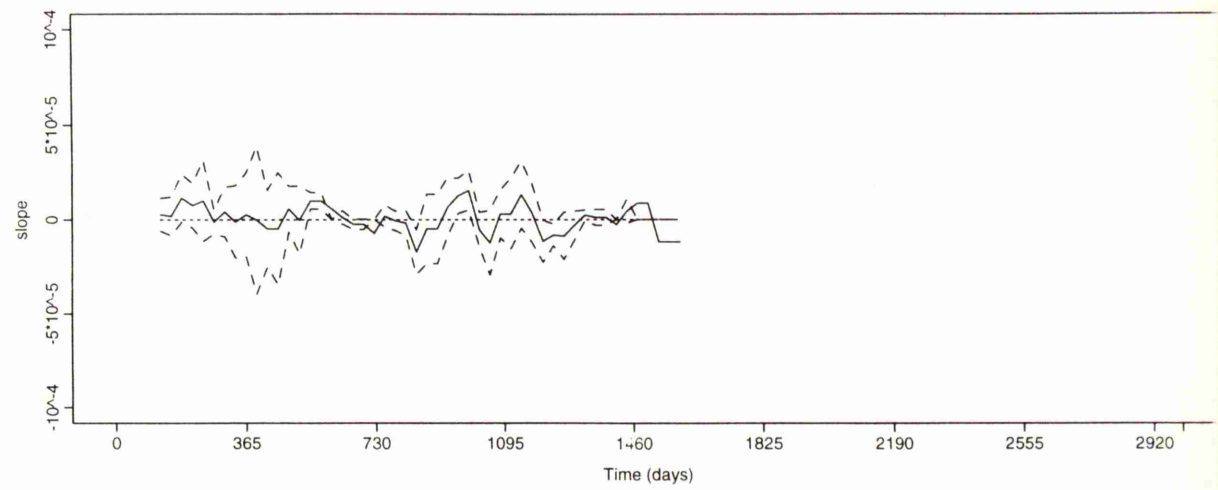
Figure 6.12 : Hospital number 8




Hospital number 8




Hospital number 8



### 6.2.9 Results of the Algorithm

The time of event, status and covariate information were available for 446 individual grafts. The classifications of these were :

| status | number in category |
|---|---|
| stable,good | 306 |
| stable,poor | 49 |
| PGD | 66 |
| AGL | 25 |

Table 6.2: Results of algorithm

The proportions in each category are similar to those observed in the set of test cases. This indicates confidence of a representative test sample and consistency in the algorithm.

A histogram of the times associated with the onset of progressive graft dysfunction (Figure 6.13) shows how the times are distributed, with most of them occurring within two or three years of transplantation.

### 6.2.10 The Experts' Views

The algorithm was tried, tested and improved on the basis of 100 randomly chosen grafts and the views of two experienced medics.

A more rigorous test to assess the performance of the algorithm was carried out using the opinions of three other experienced medics on all 446 of the available grafts. Drs. Douglas Briggs (DB), Brian Junor (BJ) and Stuart Rodger (SR) were given the traces of serum creatinine and -1/serum creatinine for each of the individual grafts (similar to Figures 6.1–4). On the basis of these, each expert independently decided on an appropriate classification from stable good, stable poor, PGD or AGL.

Tables, similar to that in section 6.2.9, could be drawn for each pairing of expert and of expert and algorithm. The question of interest is the extent to

Figure 6.13 : Histogram of PGD Times

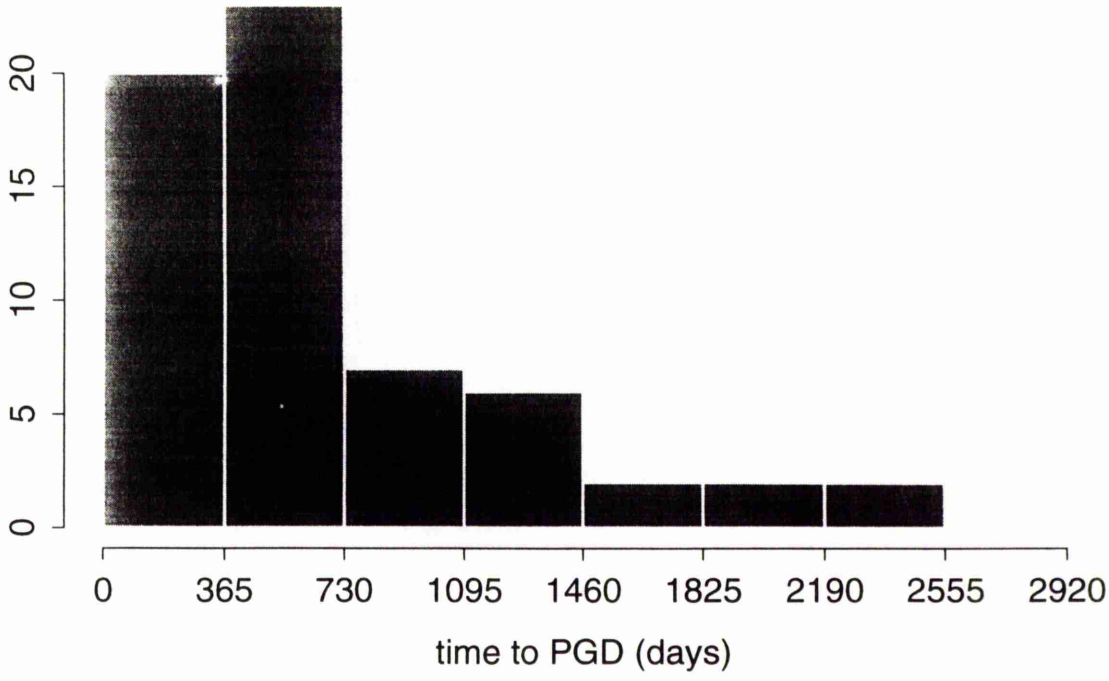
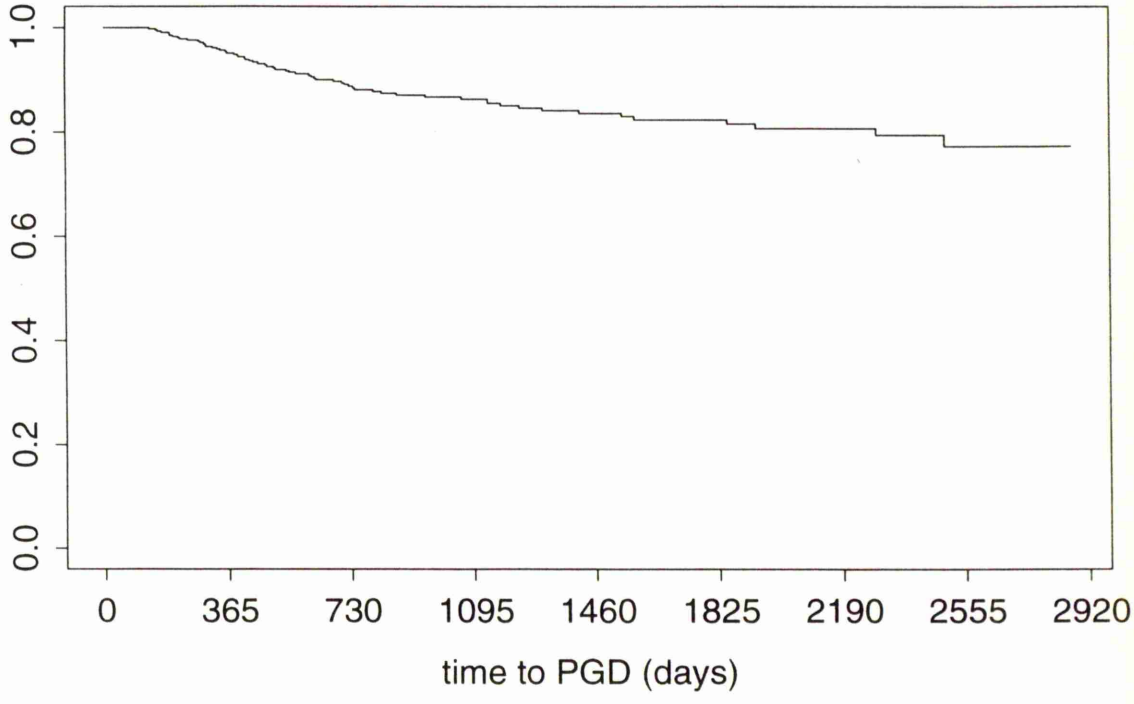



Figure 6.14 : Onset of PGD



which there is agreement in PGD classifications.

The categories stable good, stable poor and AGL are merged to provide a classification which encompasses all cases *except* those diagnosed as PGD. (The motivation for this is clearer when techniques for further analysis are discussed in the next section.) For example consider the comparison of opinions of Dr. Douglas Briggs and the algorithm.

| | | Algorithm | |
|---|---|---|---|
| | | PGD | Other |
| DB | PGD | 57 | 29 |
| | Other | 9 | 347 |

Table 6.3: DB and Algorithm classfications

Note: There were occasions when the experts felt that a particular graft did not fall into any specific classifications. Also the algorithm requires at least one year's data to be able to make a decision. These cases have been disregarded from this analysis.

An obvious quantity to measure agreement would be the proportion of cases correctly classified. For the case of DB and algorithm above, this is 0.914, calculated as 57+347 / 57+29+9+347 . The other combinations of opinion produce the following results.

| | DB | BJ | RR | ALGO |
|---|---|---|---|---|
| DB | 1 | 0.941 | 0.941 | 0.914 |
| BJ | | 1 | 0.947 | 0.881 |
| SR | | | 1 | 0.912 |
| ALGO | | | | 1 |

Table 6.4: Comparisons of Experts' Views

In assessing the performance of the algorithm, it should be borne in mind that the opinions of the experts are themselves not in complete agreement. However, there is slightly greater agreement among the experts than between the experts

and the algorithm. Nevertheless, the algorithm has achieved a good level of performance.

## 6.3 Factors Affecting the Onset of Progressive Graft Dysfunction

To investigate which factors may cause, or have an influence on, the onset of PGD, we grouped together cases in the stable good, poor and AGL categories. This allows us to employ survival analysis techniques since we use our time event of interest (discussed in an earlier section) as "survival time", grafts which have experienced PGD will be said to be "failures" and "censored" observations will be all those which fall into the other three categories. The classifications used are those obtained from the algorithm since these have been shown to be in close agreement with diagnoses by medical experts and the criteria on which they are based are expressed explicitly.

Firstly we found the Kaplan-Meier survivor function estimates for these data and a plot of these can be seen in Figure 6.14. This gives a very basic illustration of the probability of suffering PGD over time. Point estimates for this probability for different years are helpful.

| Year | Estimated cumulative Probability of onset of PGD |
|---|---|
| 1 | 0.05 |
| 2 | 0.12 |
| 3 | 0.15 |
| 4 | 0.17 |
| 5 | 0.19 |

Table 6.5: Survival by Year

This reflects the pattern displayed by the histogram in the previous section, namely that most of the occurrences of the PGD are detected within one or two years of transplantation.

Various covariates which were thought likely to have an effect on the occurrence of PGD, were analysed using Cox's proportional hazards model(Cox,1972). The results of careful model comparisons can be summarised as follows.

A variety of covariates were not found to be related to the occurrence of PGD.

- Neither the **age** of the recipient nor the donor at the time of transplant was found to significantly affect the onset of PGD.
- Neither **sex**, male or female, is significantly more at risk to this condition than the other.
- The **mismatch score**, used to assess the acceptability of the kidney to the patient, and the number of **peak panel level reactive antibodies** were also found to be non-significant in their effect on PGD.
- The number of days between the **date of the kidney first functioning** and the date of transplant does not have a significant effect on the onset of PGD either.
- **Cyclosporin dosage and blood level** and the **number of hypertensive drugs** were included as time-dependent covariates in the model but none of these was found to have a significant effect.
- There are several **immunosuppressive drugs** used to suppress the natural immunity of the body which would otherwise cause the transplanted organ to be damaged by a rejection process, but none of these, nor any combination of them, was found to have a significant effect on the occurrence of PGD. (Perhaps due to small numbers in some groups.)

The following covariates were found to be related to the occurrence of PGD.

- The **type of transplant** (cadaver or living-related donor) was found to have an effect which bordered on significance (p-value = 0.08) when fitted in the Proportional Hazards model.

When investigated further, using the log-rank test statistic to compare the two Kaplan-Meier survivor functions, there was again a borderline effect ($\chi^2 = 3.1$, p-value $= 0.078$). Figure 6.15 shows these two curves and we suspect that the difference would be more statistically significant if more cases were available in the living-related donor group (Only 2 out of 37 grafts experienced PGD in the living-related donor group, as opposed to 60 out of 375 in the cadaver group.)

- The first covariate which was found to be strongly statistically significant was the **number of rejection episodes treated with steroids.** (coeff $=$ 0.06, p-value $< 0.005$)

  Since there were dates available for when each rejection occurred, the number of days to rejection from transplantation was entered as a time-dependent covariate in the Proportional Hazards Model and found to be significant.(coeff $= 0.0007$, p-value $= 0.015$)

  A simple way to analyse this effect is, to use the log-rank test statistic to test for different Kaplan-Meier survival curves.(See Figure 6.16)

Further investigation showed that the principle difference lies in whether or not the graft has suffered a rejection episode, irrespective of how many. ($\chi^2$ $= 16$, p-value $< 0.0001$) Point estimates for the two groups, similar to those used earlier for the basic curve describing onset of PGD, are shown below.

| year | no rejection episodes | at least one rejection episode |
|---|---|---|
| 1 | 0.01 | 0.08 |
| 2 | 0.06 | 0.18 |
| 3 | 0.06 | 0.22 |
| 4 | 0.09 | 0.25 |
| 5 | 0.10 | 0.26 |

Table 6.6: Probability of survival by year and rejection episode

Recalling that the type of transplant was of borderline significance, the effect of the number of rejection episodes treated with steroids was re–analysed on cadaver transplant patients but the results were found to be very similar.

Figure 6.15 : Onset of PGD by transplant type

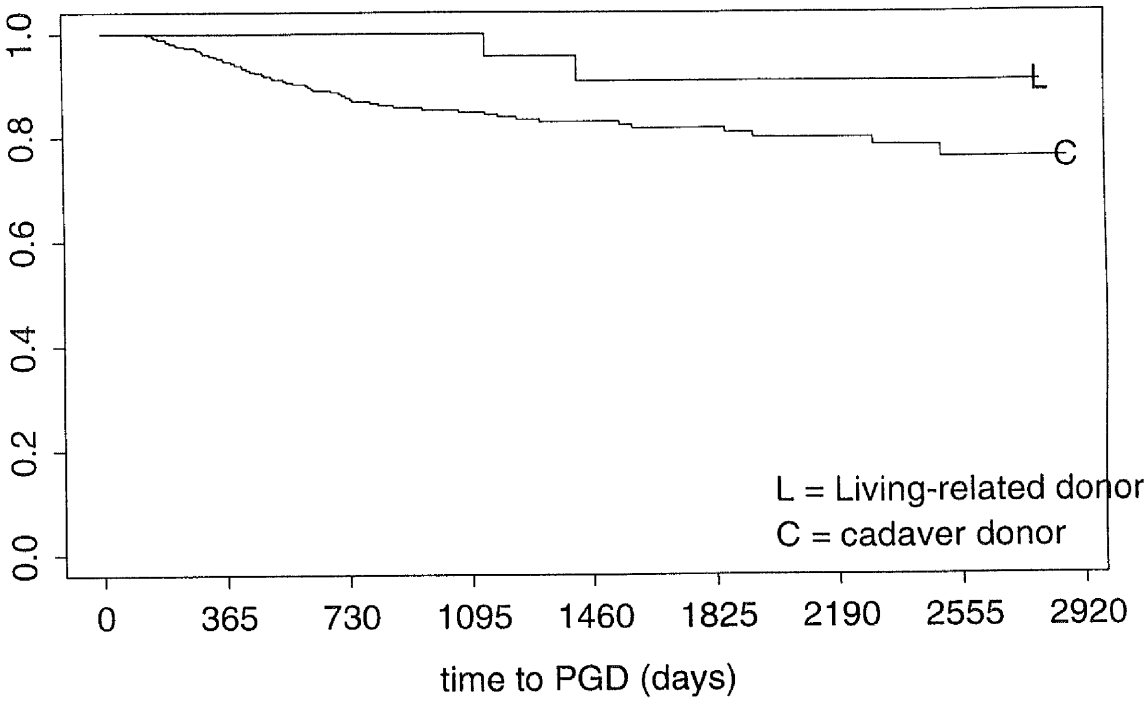
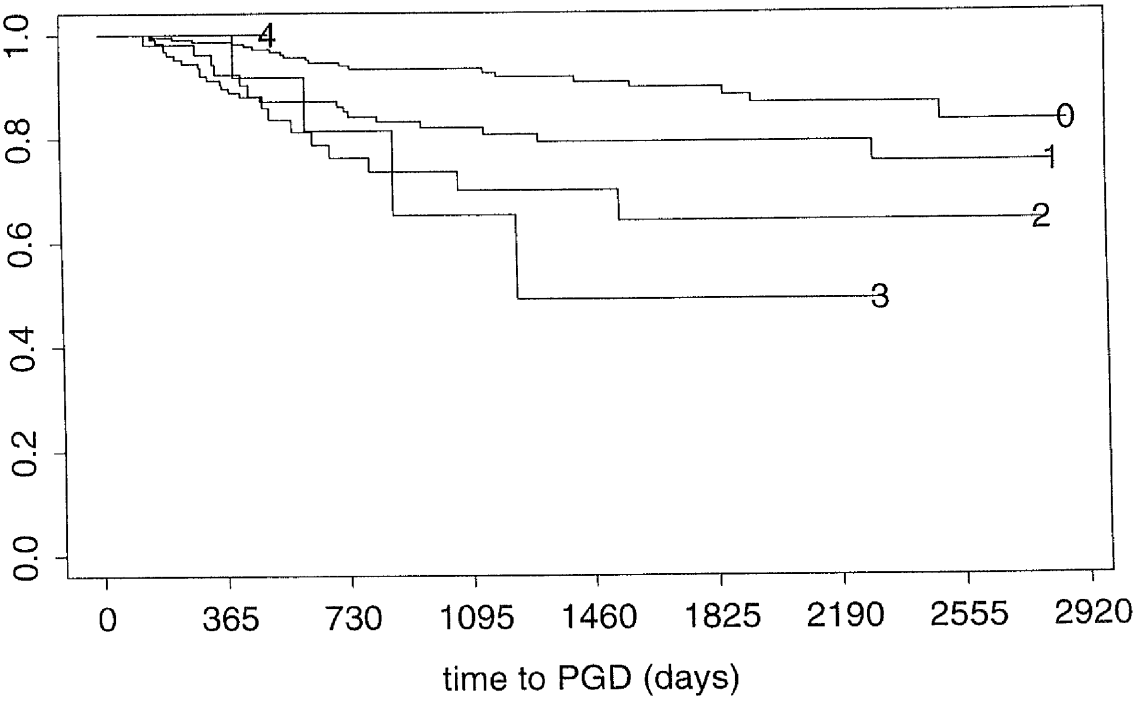



Figure 6.16 : Onset of PGD for different numbers of rejection episodes treated with steroids



- Systolic and diastolic blood pressure, fitted as time-dependent covariates separately, are significantly related to the occurrence of progressive graft dysfunction. (coeff = 0.0284 & 0.0327, p–value = 0.0001 & 0.0199 respectively) These quantities are more traditionally analysed as a mean blood pressure ($\frac{1}{3}$systolic + $\frac{2}{3}$diastolic) and, not surprisingly, this also produces a result which is significant. (coeff = 0.0416, p–value = 0.0002).

This is in apparent contradiction to the non-significant effect of the number of hypertensive drugs and the interpretation of this is not clear.

In this analysis the mean blood pressure (MBP) of individuals was noted annually, for a maximum of six years. The time-dependent proportional hazards model requires a survival time for each observed MBP measurement. So, individuals who had not yet been observed to fail in a particular time interval/year, were classed as being censored at the upper end of that interval. Each of these times is treated as an observation and Cox's PH model is fitted using all these measurements. For example, a patient observed to experience PGD after 1500 days will contribute 5 observations to the analysis.

| Start | End | Status | MBP at year ... |
|---|---|---|---|
| 0 | 365 | 0 | 1 |
| 365 | 730 | 0 | 2 |
| 730 | 1095 | 0 | 3 |
| 1095 | 1460 | 0 | 4 |
| 1460 | 1500 | 1 | 5 |

Figure 6.17 displays the data as it would be analysed using the time-dependent PH model. The high proportion of censored data and sparsity of large failure times may be masking the evidence of an effect. A percentile lifetime curve under the PH model would help to provide a graphical interpretation of how MBP changes with time. It may also be informative to consider a doubly-smooth percentile curve as this does not have the restriction of being forced to be monotonic unlike the PH curve. The smooth and PH $5^{th}$ percentiles across the mean blood pressure are shown in Figure 6.18. The axes have been swapped round to help

the understanding of the key features. Both curves suggest that the time to PGD decreases as MBP increases.

To investigate this further, each year's data, as it is entered in the time–dependent model, is analysed separately.

The results of fitting Cox's PH model to each year were as follows :

| Year | Coeff. | p–value |
|---|---|---|
| 1 | 0.035 | 0.064 |
| 2 | 0.057 | 0.001 |
| 3 | -0.027 | 0.548 |
| 4 | 0.055 | 0.061 |

Table 6.7: Analysis of Mean Blood Pressure by Year

There is too little data to analyse the effect of MBP during years 5 and 6. Also, since only 4 failures occurred in each of the third and fourth years, the results from these time intervals are dubious. The effect of mean blood pressure on progressive graft dysfunction in year 1 is marginally significant but the effect during the second year is larger.

The subjective technique applied earlier to assess the effect of MBP on all the data is employed here for data observed only during the second year. The smooth and PH $5^{th}$ percentile curves are calculated and shown in Figure 6.19. Under the PH model, the curve decreases sharply at $100 mmHg$, with slope gradually becoming less steep at around $120 mmHg$. The smooth curve displays a less dramatic shape although there appears agan to be a change in gradient around 115 or 120 $mmHg$. Categorising the data into 2 groups, survival times with MBP$<115 mmHg$ and MBP$\geq 115 mmHg$, the Kaplan–Meier survival curves can be drawn and compared (see Figure 6.20). Results from this analysis show that there is significant evidence that patients with mean blood pressure $\geq 115 mmHg$ are likely to suffer from PGD earlier than those with lower MBP. ($\chi^2 = 6.4$, p–value $= 0.012$)

Attempts were made to separate the data into three groups to assess whether it is more applicable to consider MBP as low, medium and high but the results implied that the lower 2 groups should be merged. Lack of data would also sug-

Figure 6.17 : Mean BP vs Survival Time

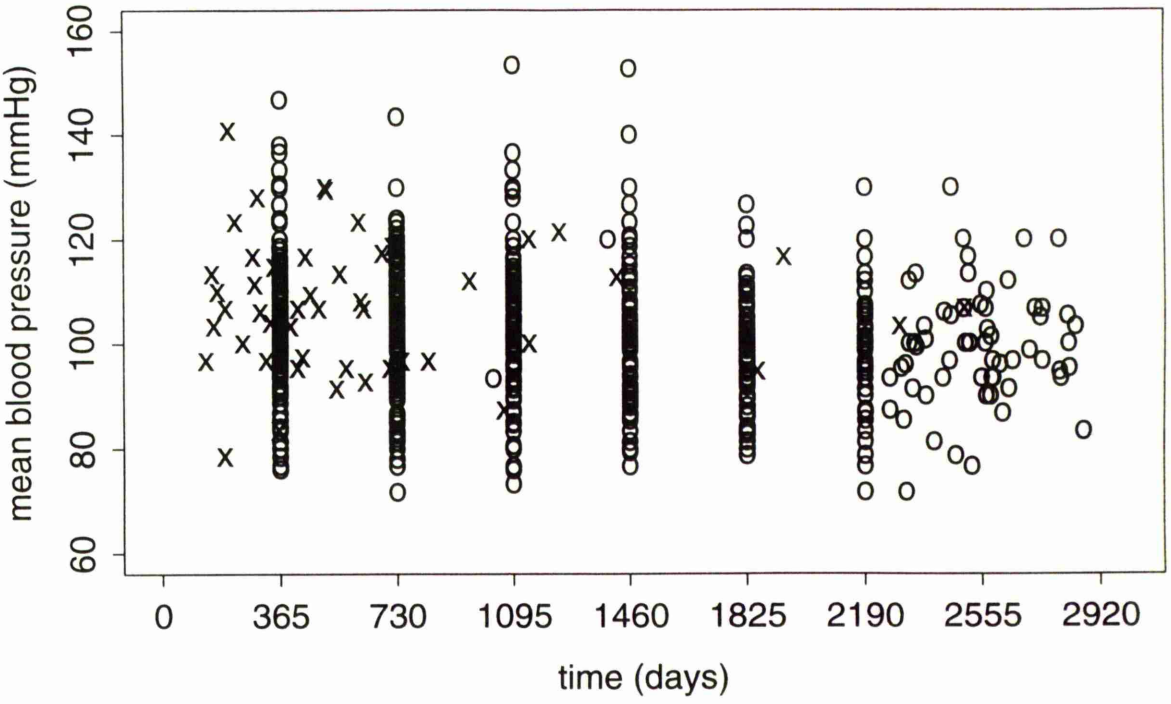
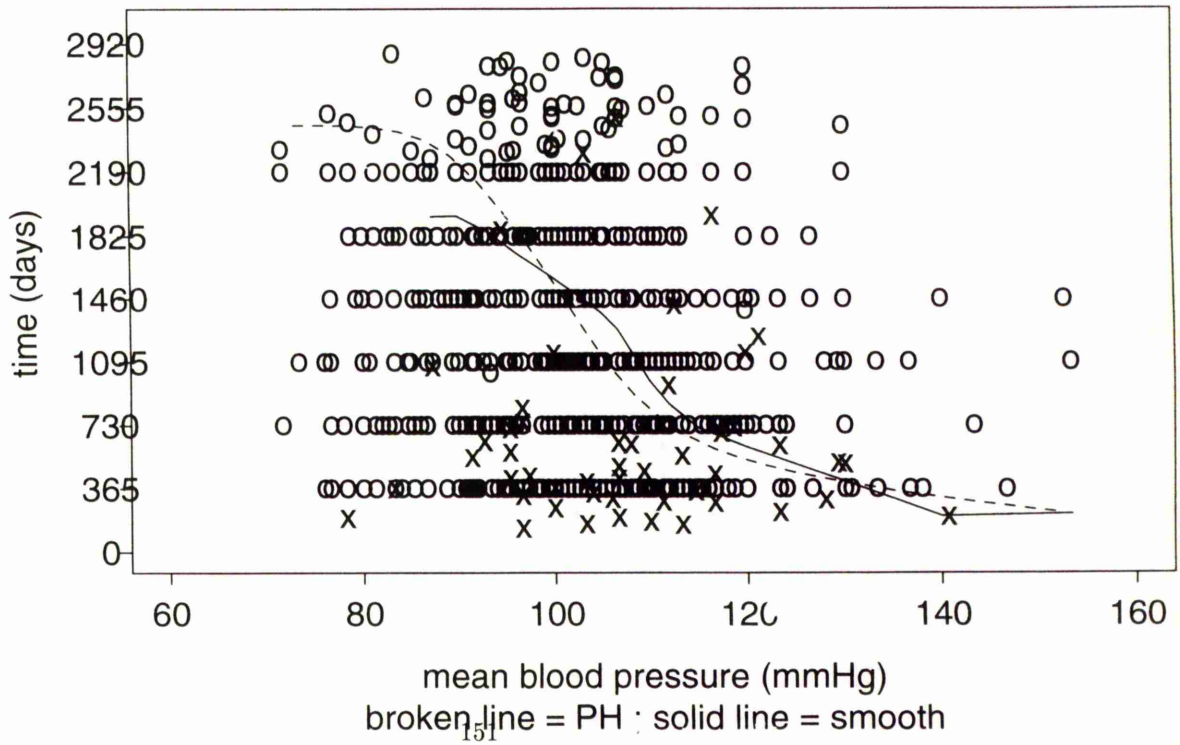



Figure 6.18 : Survival time vs Mean BP
with 5th smooth and PH percentiles



gest keeping only 2 categories as we want a reasonable number of individuals in each.

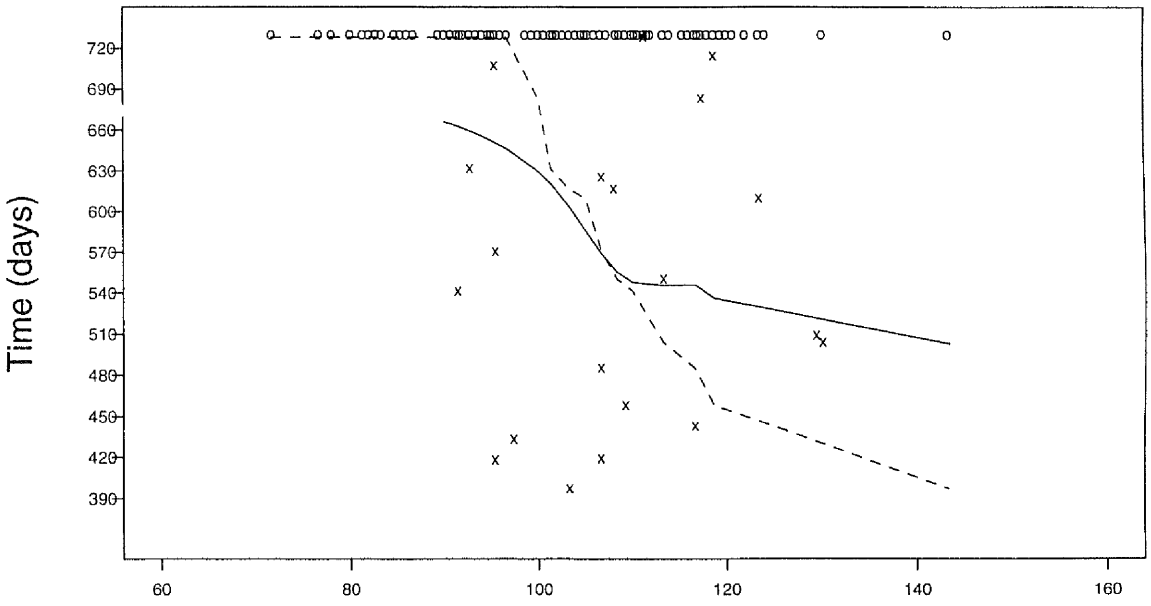
Since the mean blood pressure at year 1 was observed to be a marginally significant effect, it may be worth further investigation. The same procedures were carried out as for the year 2 data but, as would be expected, evidence of an effect was less obvious. In particular, the smooth $5^{th}$ percentile curve, of Figure 6.21, displays very little change over the range of MBP (apart from those areas where it is clear that one particular point is having an effect on the curve). The value at which the PH curve decreases is around 100 $mmHg$ which is consistent with the data analysed in year 2.

The cutpoint $115mmHg$ was used to distinguish between high and low MBP as before but the two Kaplan–Meier curves were not found to be significantly different. ($\chi^2 = 2.2$, p–value $= 0.141$). However the effect of moving the cutpoint to $110mmHg$ produces a result which borders on significance. ($\chi^2 = 3.8$, p–value $= 0.051$). Between 110 and 115, there are a lot of censored values and four failure times, two of which are low. Figure 6.22 shows the 2 sets of KM curves for these different cutpoints. It is very difficult to determine the nature of the effect of blood pressure on survival time due to high censoring proportions and the extent to which even one or two failure times can affect, or change, the conclusions.

This area may be worth further investigation, in particular to find more evidence as to whether blood pressure is a causal or subsequent effect. However, it should be borne in mind that the large amount of missing data and high proportion of censoring casts doubt upon the significance of the results.

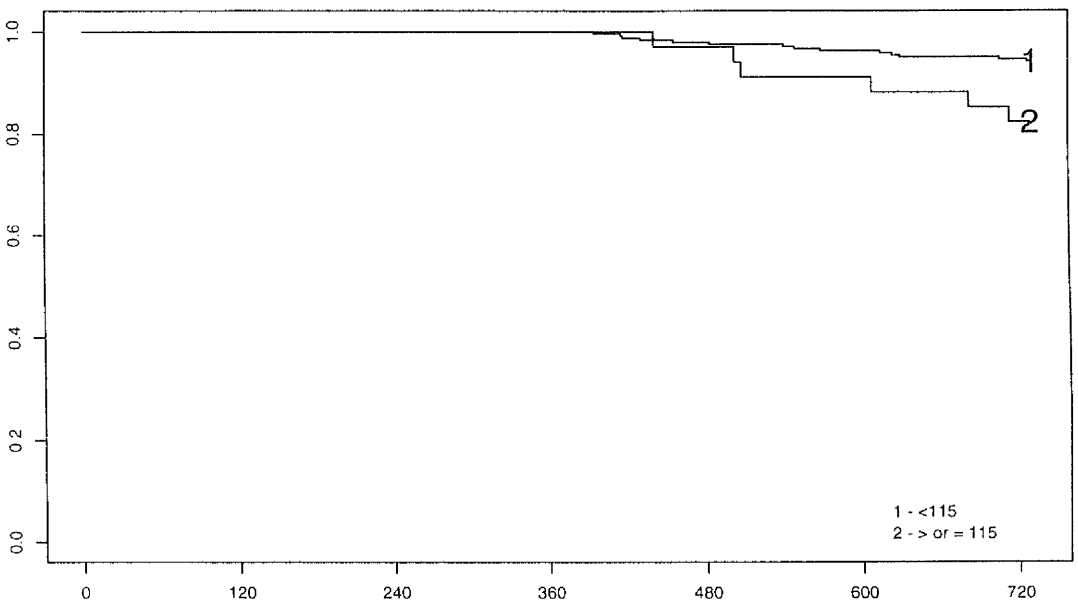
At first, it may seem surprising that only a few covariates were found to have a significant effect on the onset of progressive graft dysfunction. However, since relatively little is know about the occurrence of this type of deterioration perhaps other factors, which have not been considered here, require analysis. These results are still interesting and important and will, hopefully, help future research into this condition.

Figure 6.19 : Survival time vs Mean BP during year 2
with 5th smooth and PH percentiles




Mean Blood Pressure (mmHg)
broken line = PH ; solid line = smooth

Figure 6.20 : Survival curves for low and high MBP




time to PGD (days)

Figure 6.21 : Survival time vs Mean BP during year 1
with 5th smooth and PH percentiles

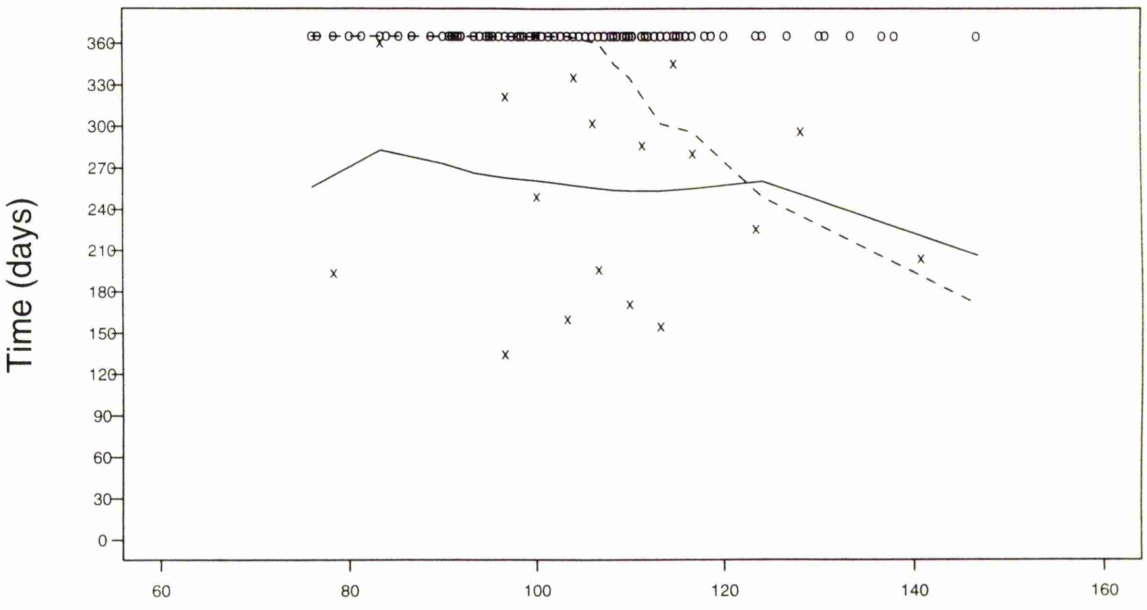
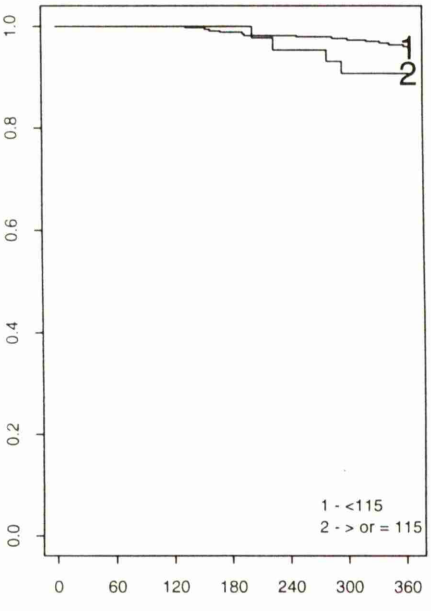
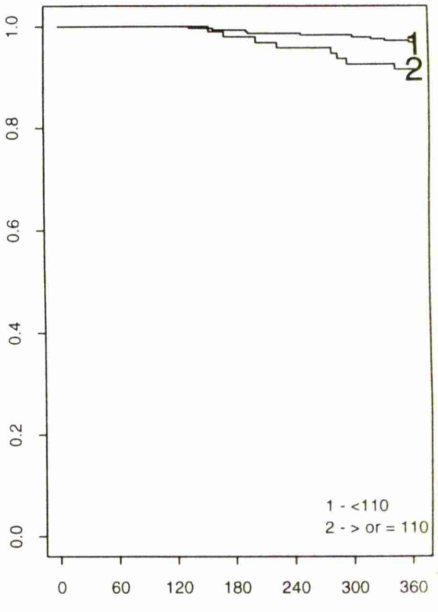



Figure 6.22 : Survival curves
for low and high MBP




Survival curves
for low and high MBP



## 6.4 Summary

Progressive graft dysfunction is a condition suffered by some kidney transplant patients. It is a steady deterioration of the function of renal graft but little is known about what causes it.

An algorithm has been developed to provide an explicit rule for detecting incidents of progressive graft dysfunction. Previously these cases have been diagnosed by the experienced, though subjective, impressions of doctors, who observe the transformed serum creatinine levels measured over time following transplantation. The algorithm also uses these levels as the basis of its diagnoses and a graft can be classified as one of four conditions – experienced progressive graft dysfunction or acute graft loss (a much more severe deterioration) or in a stable, good or poor, state. A time corresponding to the number of days until the event occurred is also given. The proportion of correctly classified grafts by this method has been demonstrated to be high.

Cox's Proportional Hazards Model and log-rank tests were used to analyse the data obtained from the algorithm with appropriate covariate information.

The event of a rejection episode significantly increased the chance of the onset of progressive graft dysfunction, although the risk did not seem to increase if there were subsequent rejection episodes. This may be explained by early acute rejection damaging the graft, rendering it susceptive to later immunological injury, or perhaps further reducing the number of functioning nephrons. Another study by Almond *et al.* also found this to be an important factor.

It also appears likely that patients whose kidney was from a living-related donor are less likely to suffer from progressive graft dysfunction than those who had a cadaver transplant. It is thought that the living–related donor grafts are likely to be better matched and suffer less damage during organ donation and transplantation.

The effect of blood pressure is interesting. It is thought that an increase in blood pressure could cause injury to the graft through haemodynamic mecha-

nisms but as yet it is unclear as to the extent of its significance.

Cyclosporin nephrotoxicity was observed to have an effect on the occurrence of PGD in a study by Salomon (1992) but, in this analysis, the drug dosage did not appear to have a significant effect. At the beginning of this chapter, other causes of late graft loss were listed, such as donor age, but these would not appear to be directly relevant to PGD sufferers and perhaps more distinction between the different causes of rejection is required.

# Chapter 7

# Linear Growth Curves for Children with Cerebral Palsy

## 7.1 Introduction

At the Kennedy Krieger Institute in Baltimore, Jackie Krick and Patti Miller have collected data on a large number of children with cerebral palsy. They are interested in assessing the growth and weight of these children relative to their ages and how this compares to the development of normal children.

Cerebral palsy (CP) is caused by non–progressive abnormality of the immature brain and results in disorders of movement and posture. Other symptoms may include mental retardation, seizures, visual and auditory deficits.

The cause of the brain damage, which results in cerebral palsy, can only be identified in 60% of all cases but is distinguishable from other brain disorders in that it occurs before a child's brain has fully matured, which is usually when a child is about sixteen years old, and its lack of progression.

There are different types of cerebral palsy, corresponding to differently damaged regions of the brain, which in turn determines which limbs are most affected by spasticity. The main types are :-

- quadriplegia - all four limbs affected
- diplegia - legs affected more than arms
- hemiplegia - arms affected more than legs
- paraplegia - legs only affected, no involvement of arms

The National Centre for Health Statistics (NCHS) constructed percentile curves which could be used to assess the physical development of normal children. Data from other previous studies was combined providing measurements on a sample of over 20,000 children. (Care was taken to ensure that this was a representative sample of all children in the United States of America.) The four curves derived from this research were various percentile curves for height, weight and head circumference vs. age, and weight vs. height. Each chart was constructed for males and females individually and for the two age groups 0–36 months and 2–18 years. We want to present similar charts for children with cerebral palsy. Although we obviously do not have as large amount of information as was used to create the NCHS charts, there is enough data to achieve useful results.

It is thought that there are less differences in height and weight between children with cerebral palsy of a similar age, even children of different sex, than we would find in normal children. So, we should assess the similarities of the charts and consider the possibility of creating a "combined cerebral palsy chart" to represent the growth of both sexes.

One of the major aims of the analysis is to compare the rate of growth of normal children with that of the children with cerebral palsy. The latter are thought to be both lighter and smaller than normal children.

Deductions from a visual comparison of the charts may not be enough since, as was mentioned before, the sample of children with cerebral palsy is small relative to that used to construct the NCHS charts. Therefore, we will assess the

sample variability to validate our conclusions.

### Objectives

- Compile growth curves for children with cerebral palsy
- Assess similarity of male and female curves.
- Compare CP and NCHS growth curves.
- Analyse sample variability and differences with NCHS curves.

## 7.2 The Data

We have data on 282 boys and 262 girls. During visits to the Kennedy Krieger Institute, the age, height and weight of each child was recorded. The number of measurements for each child varies widely and there is no set frequency to the visits. In total, we have 1080 measurements on the boys and 1088 measurements on the girls (although there are a few missing data points occurring where the height was not recorded.) Data were not collected on head circumference as it was not thought to be of interest in this research.

While it is tempting to utilise all of the available data, there is a longitudinal aspect to the measurements. If a method appropriate for cross–sectional data is used, there would be a risk of over-, or under-, representing some patterns of growth. For example, very small children may be monitored more closely and would therefore contribute more measurements to the data–set than children of average height. (An informal investigation showed there to be some evidence of this.) Therefore, for each child, only the information recorded for one visit, chosen at random, is used in the analysis.

We must distinguish between the different types ot cerebral palsy from one another as they are thought to develop at different rates, just as their growth rate differs from that of normal children. However, we have an insufficient amount

of data on any of the types of cerebral palsy apart from quadrilplegia so we will consider only children with this condition.

In the NCHS charts, separate curves were derived for two age groups – <36 months and 36 months –18 years. However, we do not have enough information on children under 36 months so charts will be constructed using the whole range of data from 0–18 years.

Figures 7.1–6 show scatterplots of the complete data–set. (Other figures will show only the data analysed.) These are plots of height vs. age, weight vs. age and weight vs. height for males and females individually.

Although a smooth line representing the median growth rate on these plots is informative, it is important to identify a region in which most of the data will occur. This will be illustrated by also considering the 10th and 90th percentile growth curves. The 10th(90th) percentile curve is defined to have 10(90) percent of children's heights below the curve at the corresponding age.

In both the girls' and boys' data, there appears to be a couple of observations which look, almost unreasonably, large. There was assurance however, that these were genuine measurements. These cases help to highlight the need to calculate percentiles since the size of the value will not have as large an effect as it would if we were calculating, say, mean values.

# Scatter plots

Figure 7.1
Boys : Height vs. Age

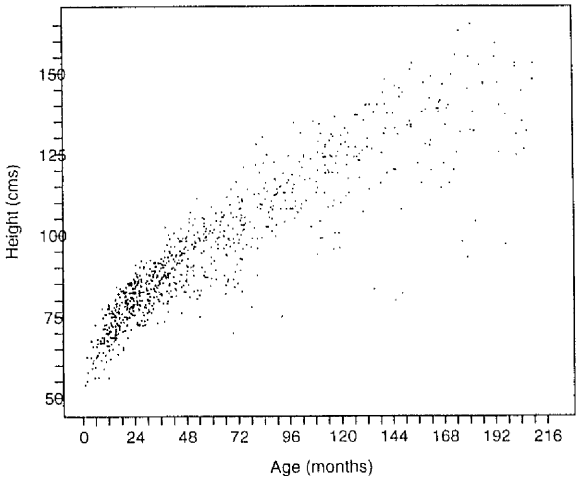



Figure 7.2
Girls : Height vs. Age

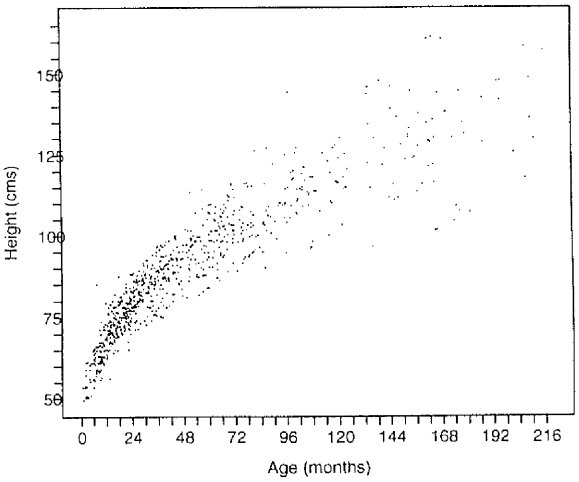



Figure 7.3
Boys : Weight vs. Age

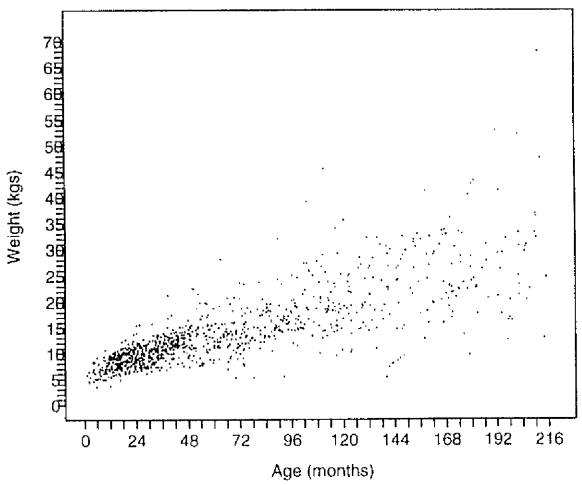



Figure 7.4
Girls : Weight vs. Age

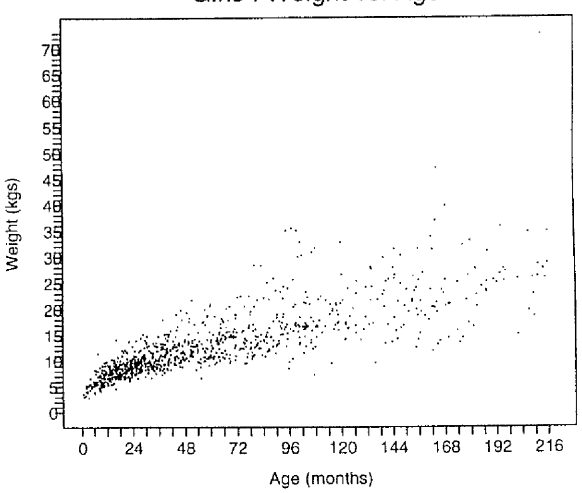



Figure 7.5
Boys : Weight vs. Height

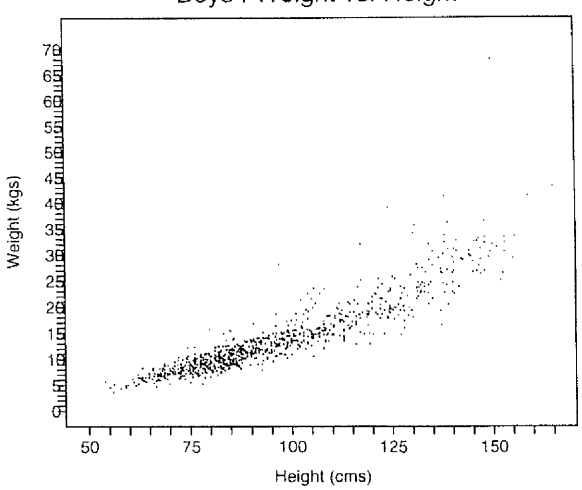
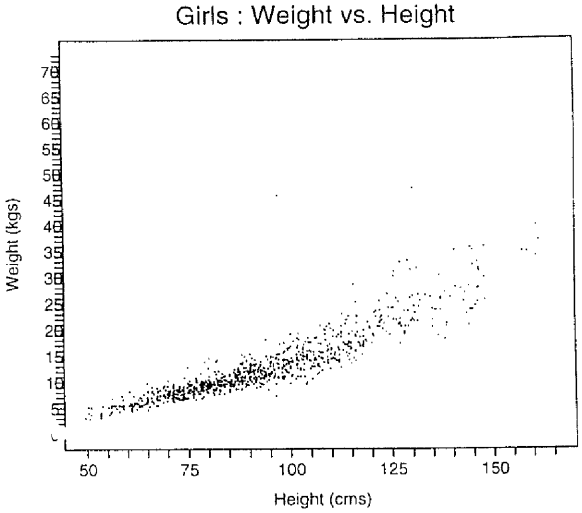



Figure 7.6
Girls : Weight vs. Height



## 7.3 Methodology

The topic of nonparametric percentile curves has been discussed at length already in Chapters 2 and 3, where kernel estimates were used to derive quantile curves for survival data. Here there are no censored observations and this is therefore an opportunity to explore other nonparametric methods.

A common technique for fitting a curve to non-linear data is *smoothing splines*. The rationale behind this is as follows. (See Hastie and Tibshirani (1990).)

Let $\{(X_i, Y_i)\}_{i=1}^n$ be the set of $n$ data points which have been collected, where $Y$ is the response variable and $X$ is the explanatory variable. We wish to characterise the dependence of $Y$ on $X$. Suppose the relationship can be described as

$$Y_i = g_o(X_i) + \epsilon_i$$

where $g_o$ is an unknown mean curve. For any curve, $g$, the residual sum of squares $\sum_{i=1}^n (Y_i - g(X_i))^2$ measures the extent to which this curve is close to the data.

This residual sum of squares can be reduced to zero if we choose any curve which interpolates the data. However, this would obviously produce a curve which is too rough, and we expect our fitted curve to have a reasonable degree of smoothness without too much local variation.

In order to choose a suitable curve, $g$, we can quantify local variation through the integrated second derivative.

$$\int (g''(u))^2 du$$

This measures the roughness of the curve $g$ and can be used as a penalty to ensure sufficient smoothness, by constructing the weighted sum.

$$S_\lambda(g) = \sum_{i=1}^n (Y_i - g(X_i))^2 + \lambda \int (g''(u))^2 du$$

The quantity $\lambda$ denotes a *smoothing parameter*, whose choice determines the trade-off between fidelity to the data and the smoothness of the curve. Choosing

$\lambda$ to be small results in a larger roughness penalty, and hence a less smooth curve, but a smaller residual error. Choosing $\lambda$ to be large gives a smoother curve but larger residual error.

Alternatively, we could select the value of the smoothing parameter by specifying the degrees of freedom for the smooth curve. For the function, $S_\lambda$, the degrees of freedom is defined to be the trace of the matrix $S_\lambda$, or, equivalently, the sum of the eigenvalues of $S_\lambda$. So the choice of degrees of freedom controls the flexibilty of the fitted curve, $g$. This is analogous to degrees of freedom in a linear model. The unique solution to the problem of minimising $S_\lambda(g)$ over the class of all twice differentiable functions is a cubic spline. This produces a good estimator of the true mean function $g_o$.

Hamill *et al.* (1977) derived the NCHS curves by calculating the percentiles from the empirical quantile function for small intervals of data across age. The values for each centile were then smoothed using a cubic spline. Here, less data is available and the number of points within each interval would be insufficient to produce reasonable empirical quantile values so a more direct approach must be used. An estimator of this type was proposed by Jones (1988) and is explained below.

We are interested in producing a spline estimator for percentiles, rather than mean values. In order to construct the median smoothing spline, the sum of the absolute deviances should be used instead of the residual sum of squares. So, we now solve the following

$$\min_{g(x)} \sum_{i=1}^{n} |Y_i - g(X_i)| + \lambda \int (g''(u))^2 du$$

which can also be written as

$$\min_{g(x)} \sum_{i=1}^{n} \frac{1}{|Y_i - g(X_i)|} (Y_i - g(X_i))^2 + \lambda \int (g''(u))^2 du$$

This is now a similar function to the one defined above, but with a weighted residual sum of squares to result in the median spline being evaluated. To generalise to the $p^{th}$ percentile, $0 \leq p \leq 1$, we adapt these weights further and

use

$$\min_{g(x)} \sum_{i=1}^{n} W_i(\underline{g})(Y_i - g(X_i))^2 + \lambda \int (g''(u))^2 du$$

where

$$W_i(\underline{g}) = \begin{cases} \frac{\mathrm{p}}{|\mathrm{Y_i} - \mathrm{g(X_i)}|} & \text{if } Y_i - g(X_i) > 0 \\ \frac{1-\mathrm{p}}{|\mathrm{Y_i} - \mathrm{g(X_i)}|} & \text{if } Y_i - g(X_i) < 0 \\ 0 & \text{if } Y_i - g(X_i) = 0 \end{cases}$$

In applying these techniques, it can be helpful to first remove as much as possible of the linear trend in the data. The quantile spline then has less pattern to detect and so its fit to the data should be improved.

In the cerebral palsy data, transformations of age and weight improved the linearity in the plot of response vs. explanatory in each of the charts. The problem of scarcity of data on older children and the wider variability of height and weight values of these children is also reduced by making these adjustments. The transformations used were to take the square root of age and the natural logarithm of weight values.

Least squares estimation was used to fit a linear regression model to each pair of transformed (where applicable) variables. So the residuals from this analysis and the transformed (again, where applicable) explanatory variable were then used to fit a smoothing spline with 5 degrees of freedom for a specified percentile. It is a straightforward procedure to transform the results from this back to their original scale so that we have our curve of interest.

## 7.4 The Cerebral Palsy Growth Curves

Figures 7.7-12 are the scatterplots of the different pairs of variables for the individual sexes with the 10th, 50th and 90th percentiles superimposed.

As we would have expected, both from looking at the data and by intuition, the rate of growth in the first two or three years following birth is high and eventually slows down as children get older. Also, initially there is little variation in the height and weight of children of similar ages but this variability becomes larger between older children.

As we would expect, the weight vs. height charts show that, for most children, these attributes will vary proportionally implying that small children will tend to be much less heavy than tall children.

As mentioned before, information on older children is sparse and, as a result, we would be wary of using these charts to make any conclusive remarks on children over the age of 120 months. Despite this, the charts could be very useful in determining the size of a child with quadriplegic cerebral palsy relative to other children with this condition.

# CP growth curves

Figure 7.7
Boys : Height vs. Age

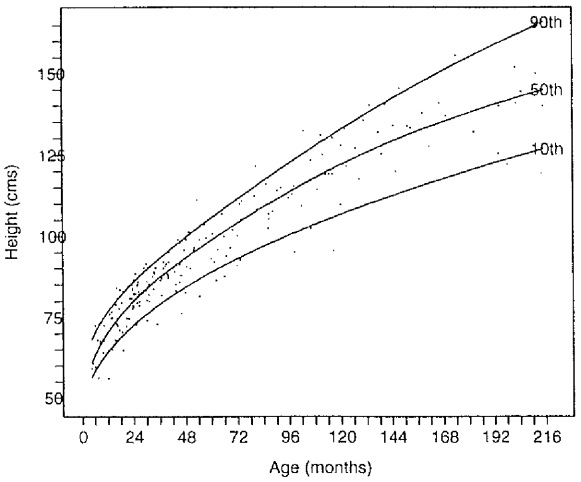



Figure 7.8
Girls : Height vs. Age

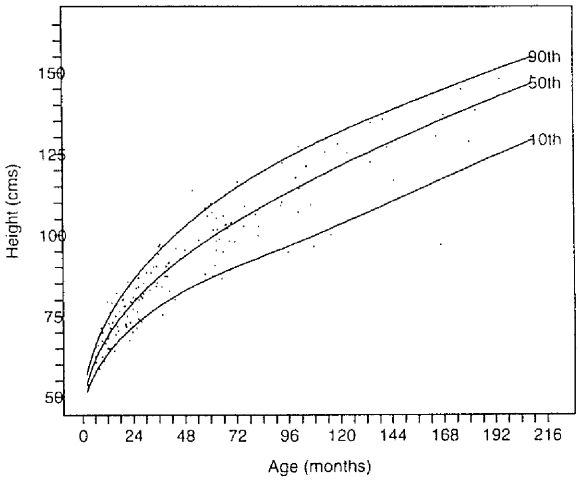



Figure 7.9
Boys : Weight vs. Age

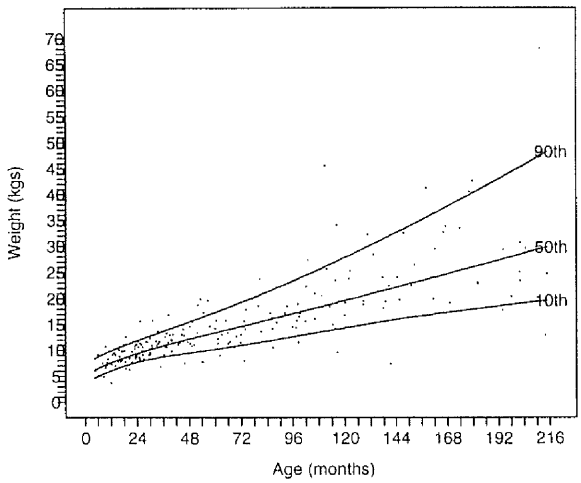



Figure 7.10
Girls : Weight vs. Age

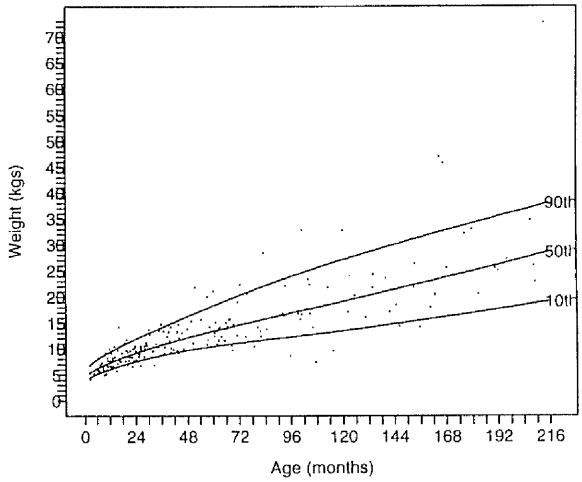



Figure 7.11
Boys : Weight vs. Height

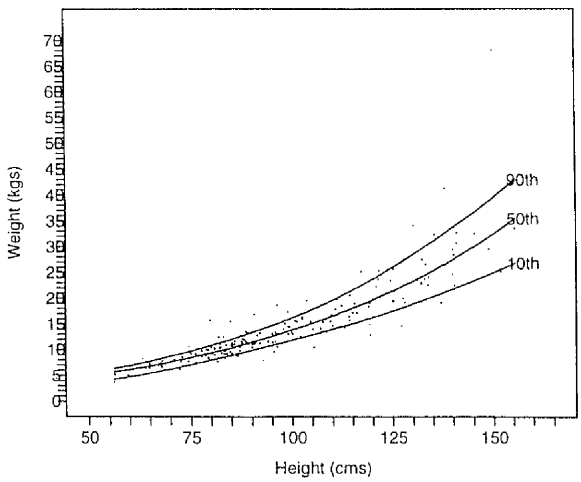
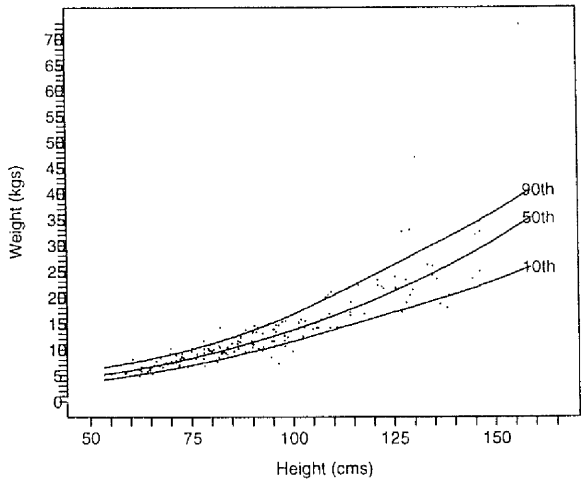



Figure 7.12
Girls : Weight vs. Height



## 7.5 Comparison of Sexes

The possibility of representing both males and females in a single chart was discussed earlier. To assess the plausibility of this, the curves for males and females were plotted together and can be seen in Figures 7.13-15.

There seems to be very little difference between the actual measurements or the rates of growth in height and weight. The largest differences can be seen in children over 120 months. Although it is known that boys are usually bigger than girls as age increases, we should remember about the scarcity of data and our consequent reluctance to make any conclusions about children in this region.

Merging the data for the two sexes results in the curves shown in Figures 7.16-18.

There is an argument that the two sexes are so biologically and physically different that to assess the development of all children together is inappropriate, regardless of the evidence of the similarity between them in Figures 7.13-15. Also, the NCHS charts are constructed for males and females individually and we wish now to compare these with the CP curves. For these two reasons, we shall proceed treating the sexes separately.

## 7.6 Comparison with NCHS Charts

In order to reconstruct the NCHS curves, values for the various charts were taken from the Normalized NCHS/CDC Anthropometric Reference Book.

For the height and weight vs. age curves, the 10th, 50th and 90th percentiles of height and weight were plotted at 3 monthly intervals. This information was only available up to the age of 120 months. However, this was not a problem since we have maintained throughout the analysis that the curves beyond this age were unreliable anyway. The "blip" at 24 months in the height vs. age chart occurs because a distinction is made between how a child is measured. Children

# Comparing Growth of Boys and Girls

Figure 7.13
Boys & Girls : Height vs. Age

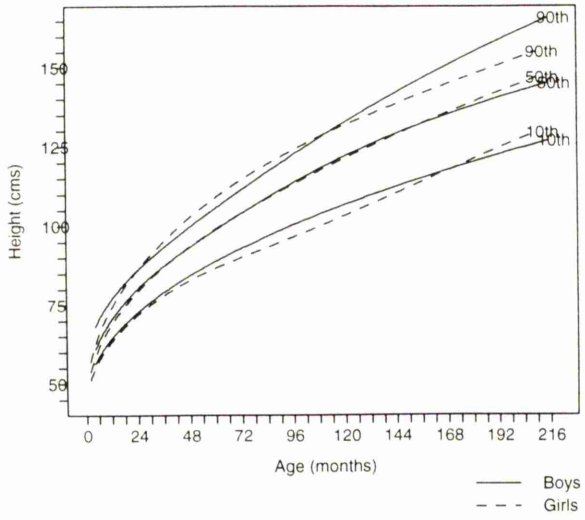



Figure 7,14
Boys & Girls : Weight vs. Age

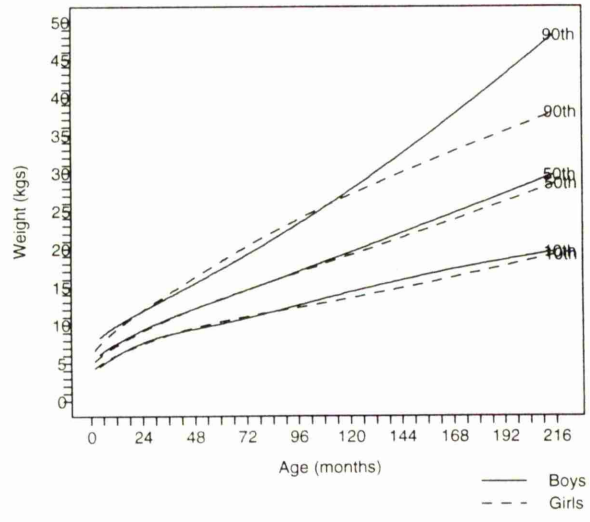



Figure 7.15
Boys & Girls : Weight vs. Height

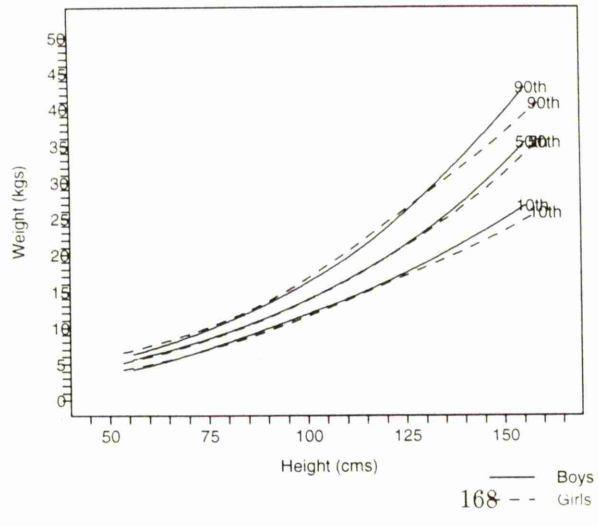



Figure 7.16
Boys & Girls combined : Height vs. Age

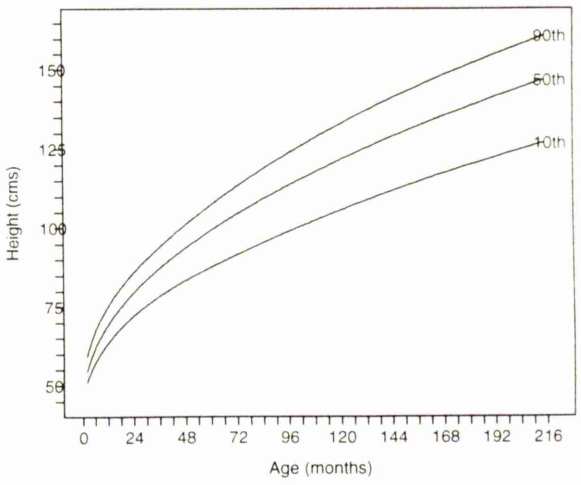



Figure 7.17
Boys & Girls combined : Weight vs. Age

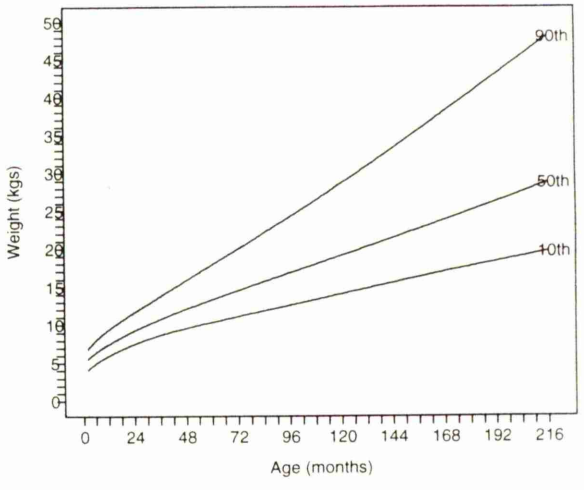
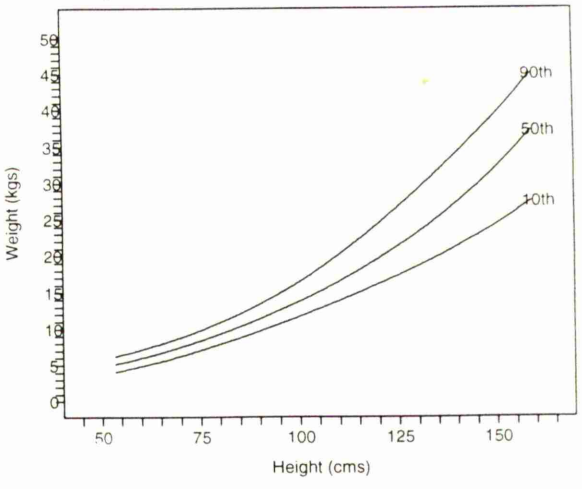



Figure 7.18
Boys & Girls combined : Weight vs. Height



up until the age of 24 months have their *length* measured and children older than this are measured by their *height*. Since children with cerebral palsy are unable to stand up straight, their length is regarded as being equivalent to their height.

For the weight vs. height charts, the 10th, 50th and 90th percentiles of weight were recorded for two values in every 5 centimetres between 50 and 135 centimetres. Again the full range of values available was used and here the height measurements less than 65 centimetres are actually lengths.

Figures 7.19-24 show the charts for the different pairs of variables for both the males and females individually.

In the height vs. age plots, the NCHS 10th percentiles lie between the 50th and 90th percentiles of the CP curves suggesting that the children with cerebral palsy are much smaller than normal children. During the first few months after birth, there is only a small difference between the two types of curve but this increases with time. So it appears that children suffering from cerebral palsy also have a slower rate of development in height than normal children.

We see the same sort of effect in weight vs. age and so children with cerebral palsy also seem to be much lighter and gain weight at a slower rate than normal children.

The differences in the weight vs. height charts are not quite so strong but, since age is not taken into consideration here, this reflects the fact that children with cerebral palsy are only slightly lighter than normal children of similar height.

## 7.7 Assessing Sample Variability

As we discussed in the previous section, from the comparison of NCHS and CP curves, children with cerebral palsy appear to be lighter and smaller than normal children of a similar age. However we must remember that the sample used to construct the CP curves is considerably smaller than that used in the NCHS

# Comparison of NCHS and CP curves

Figure 7.19
Boys : Height vs. Age

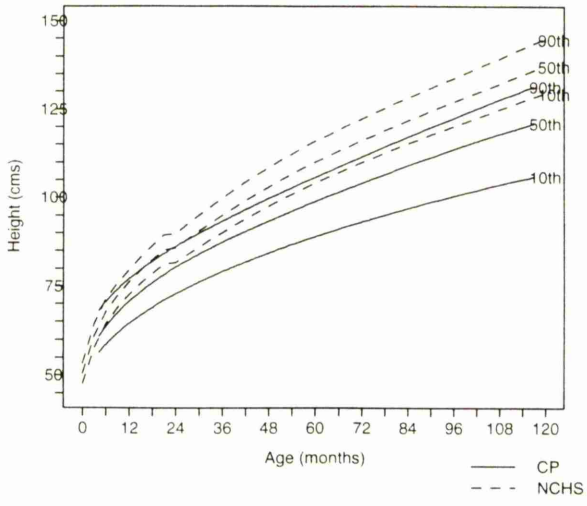



Figure 7.20
Girls : Height vs. Age

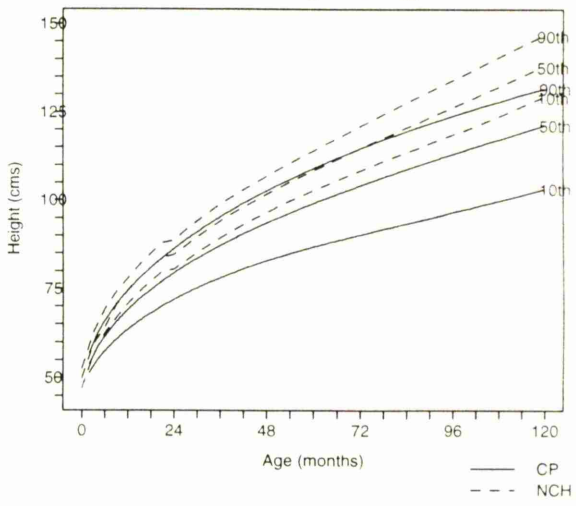



Figure 7.21
Boys : Weight vs. Age

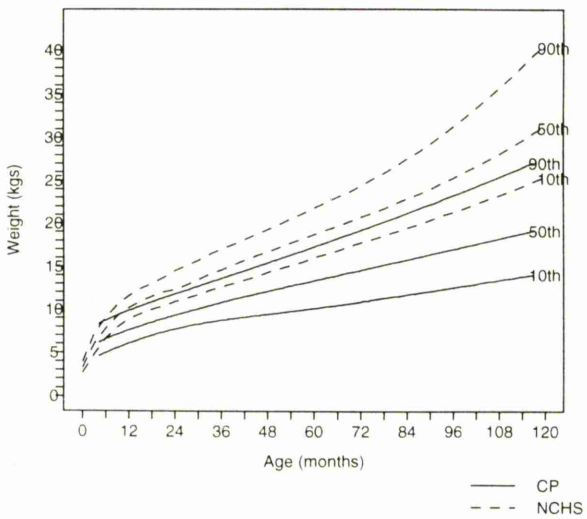



Figure 7.22
Girls : Weight vs. Age

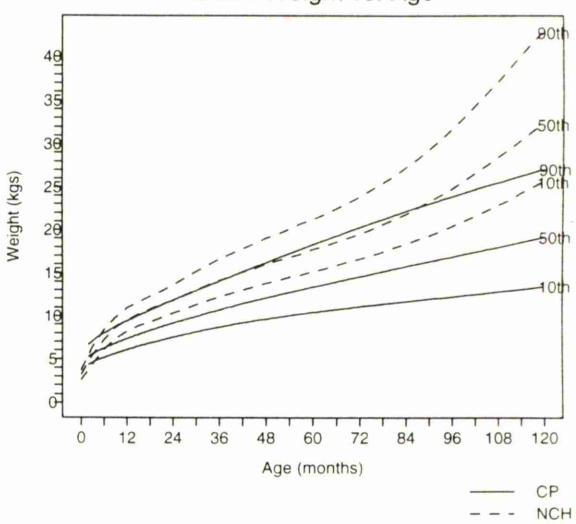



Figure 7.23
Boys : Weight vs. Height

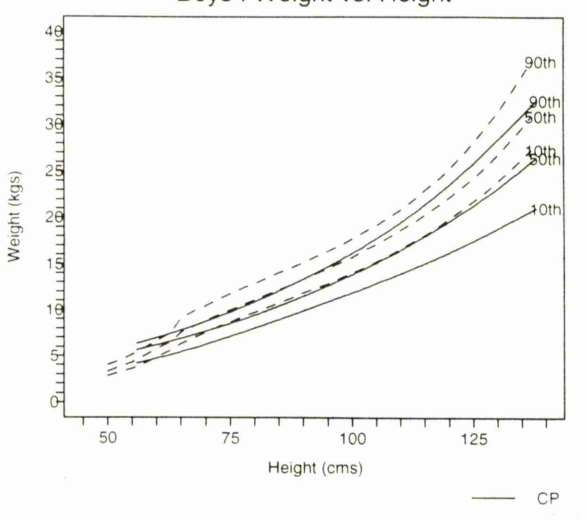
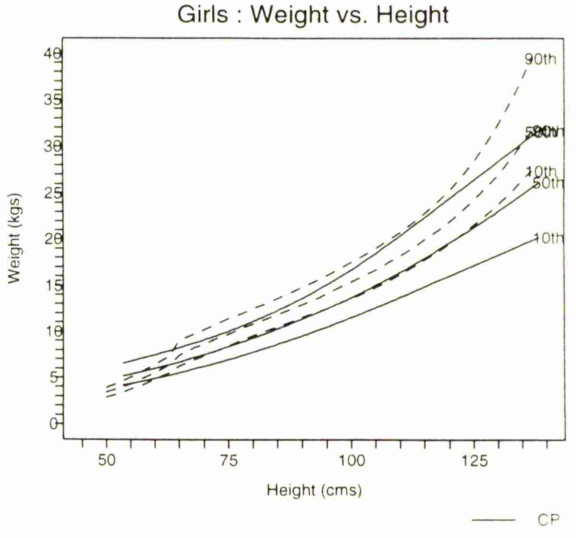



Figure 7.24
Girls : Weight vs. Height



charts. We therefore must assess the sample variability of our CP curves and take this into account when making our comparisons between the two curves.

Estimating the variance of the quantile spline would be mathematically complex but, using bootstrap samples, the variability of the sample can be assessed in a relatively simple way.

As discussed in Chapter 4, the most basic method of bootstrapping is random sampling with replacement from the available dataset under the cumulative distribution function. So, for a dataset containing $n$ observations, a bootstrap sample will be of size $n$ with $0, 1, \ldots, n$ repetitions of each of the original observations.

To assess the variability of a statistic, we draw a large number of bootstrap samples, calculate the statistic of interest for each sample and create a histogram of the values. The variability across the bootstrap replications represents the uncertainty in the statistic calculated for the original sample.

In the CP case, we have resampled children rather than individual observations, so the resampling is from the joint distribution rather than the conditional distribution. The percentile splines were calculated for each of 50 bootstrap samples. Figures 7.25–30 show the NCHS curves as seen before in Figures 7.19-24 but here the shaded areas on the plots indicate the region in which these curves fell, with the boundaries representing the minimum and maximum values at 3 monthly intervals.

As we would expect, the variability in the percentiles varies slightly between the variables and the sexes and there is a lot of overlap in the first few months. The NCHS curves are still above these shaded areas in the comparison of age vs. height and weight, and the comments of the last section still seem viable. This analysis confirms that children with cerebral palsy appear to be shorter and lighter in weight than normal children of the same age.

Figure 7.25
Boys : Height vs. Age

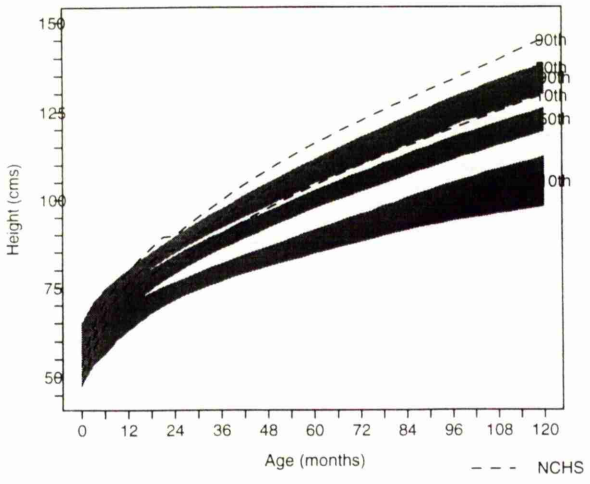



Figure 7.26
Girls : Height vs. Age

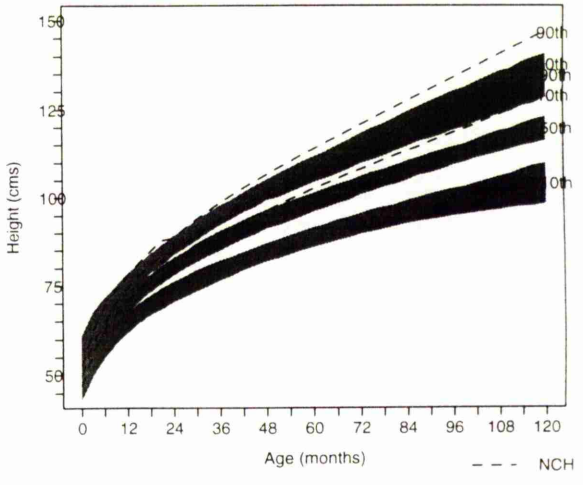



Figure 7.27
Boys : Weight vs. Age

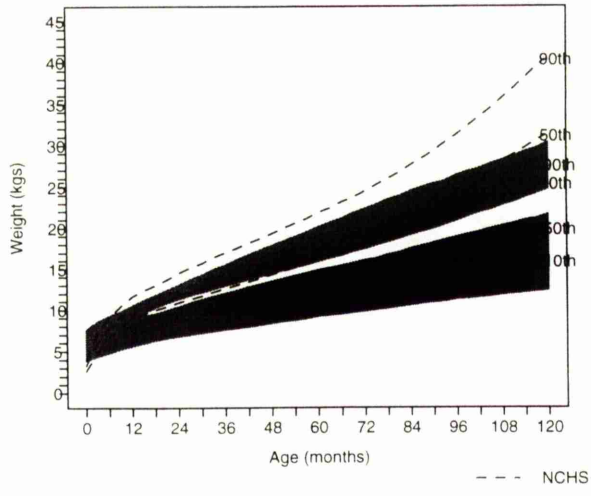



Figure 7.28
Girls : Weight vs. Age

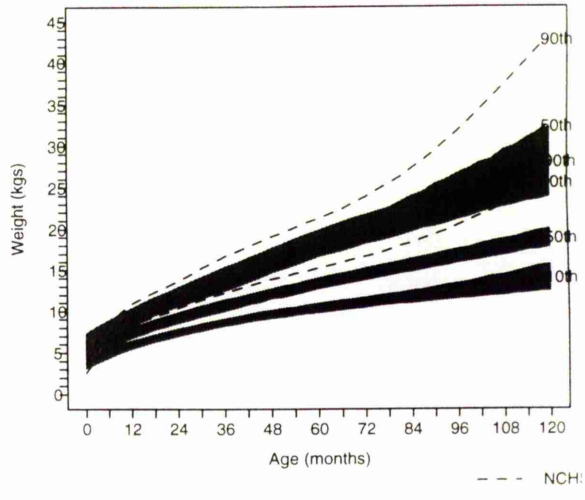



Figure 7.29
Boys : Weight vs. Height

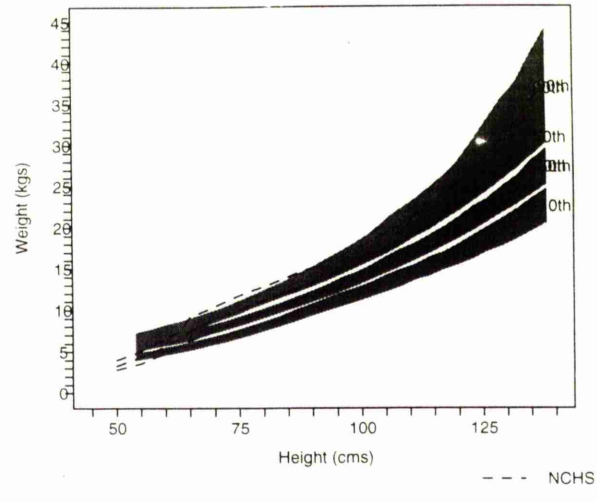
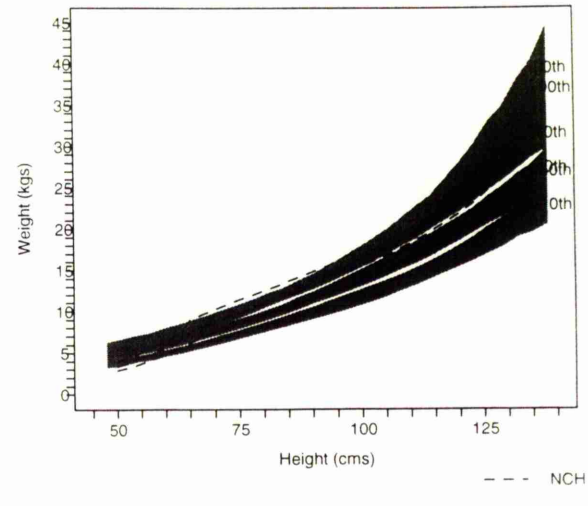



Figure 7.30
Girls : Weight vs. Height



## 7.8 Conclusions

In the previous sections we have built up a picture of what is of interest in assessing the development of children with cerebral palsy.

It is important to know how quadriplegic children compare to normal children and to appreciate the amount of sample variability.

The last six growth charts, Figures 7.25–30, combine all this information. They provide a chart to which a child with cerebral palsy's height and weight could be compared to that of both other children with the same condition and normal children. This comparison can also be quantified in terms of whether the child's measurements lie within the 10th, 50th or 90th percentile curve areas. All of this helps to give parents of children with quadriplegic cerebral palsy some idea of how their child is developing.

# Appendix A

# Taylor Series Arguments

Consider the uncensored case of the weighted Kaplan–Meier estimator $\widehat{S}(t|z)$ and use the results,

$$\text{for } X \sim Bi(n,\theta) \quad E(X) = n\theta \quad \text{var} = n\theta(1-\theta).$$

Recall that, for illustration, the covariate values here are assumed to be uniformly distributed on the design space. The limits of the integration are assumed to correspond to the range of the $z$ values.

$$\begin{aligned}
E(\widehat{S}(t|z)) &= E\left(\frac{\sum_{j=1}^{n} W_j Y_j}{\sum_{j=1}^{n} W_j}\right) \quad \text{where} \quad Y_j \sim Bi(1, S(t|z)) \\
&= \sum W\left(\tfrac{z-z_j}{h}\right) S(t|z_j) \Big/ \sum W\left(\tfrac{z-z_j}{h}\right) \\
&\approx \int W\left(\tfrac{z-v}{h}\right) S(t|v)dy \Big/ \int W\left(\tfrac{z-v}{h}\right) dv \\
&= \int W(u)S(t|z-uh)du \Big/ \int W(u)du \\
&= \int W(u)\{S(t|z) - uhS'(t|z) + \tfrac{(uh)^2}{2}S''(t|z) + O(h^2)\}du \Big/ \int W(u)du \\
&= S(t|z) + \tfrac{h^2}{2}S''(t|z) \int u^2 W(u)du + O(h^2)
\end{aligned}$$

assuming $h$ to be small and $\int_{-\infty}^{\infty} uK(u)du = 0$.

$$
\begin{aligned}
\operatorname{var}(\widehat{S}(t|z)) &= \operatorname{var}\left(\frac{\sum_{j=1}^{n} W_j Y_j}{\sum_{j=1}^{n} W_j}\right) \quad \text{where} \quad Y_j \sim Bi(1, S(t|z)) \\
&= \tfrac{1}{n^2}\textstyle\sum \tfrac{1}{h^2} K^2\left(\tfrac{z-z_j}{h}\right) \operatorname{var}(Y_j) \Big/ \tfrac{1}{n^2}\textstyle\sum \tfrac{1}{h^2} K^2\left(\tfrac{z-z_j}{h}\right) \\
&\approx \tfrac{1}{n}\int \tfrac{1}{h^2} K^2\left(\tfrac{z-v}{h}\right) S(t|v)[1-S(t|v)]dv \Big/ \tfrac{1}{n}\int \tfrac{1}{h^2} K^2\left(\tfrac{z-v}{h}\right) dv \\
&= \tfrac{1}{n}\int \tfrac{1}{h} K^2(u) S(t|z-uh)[1-S(t|z-uh)]du \Big/ \tfrac{1}{n}\int \tfrac{1}{h} K^2(u) du \\
&\approx \int K^2(u)\{S(t|z) - (uh)S'(t|z) + \tfrac{(uh)^2}{2} S''(t|z)\} \\
&\qquad .\{1 - S(t|z) + (uh)S'(t|z) - \tfrac{(uh)^2}{2} S''(t|z)\}du \,/ \int K^2(u)du \\
&= \{\int K^2(u) S(t|z)[1-S(t|z)]du + \int K^2(u) uh S'(t|z)[2S'(t|z) - 1 - uhS'(t|z)]du \\
&\qquad + O(h^2)\} \,/ \int K^2(u)du \\
&= S(t|z)[1-S(t|z)] + O(h)
\end{aligned}
$$

again, assuming $h$ to be small.

Now consider the expectation and variance of the quantile estimator, $\widehat{T}_n$. Recall from sections 2.1.1 and 2.2.1 that

$$Q(p) = F^{-1}(p) = \inf\{x : F(x) \geq p\}$$

with $E(X_{(i)}) = Q(p)$ , $\operatorname{cov}(X_{(i)}, X_{(j)}) = \frac{1}{n}$ , $p_r(1-p_s)(F^{-1})'(p_r)(F^{-1})'(p_s)$ and $p_r = \frac{i}{n}$ , $p_s = \frac{j}{n}$

$$
\begin{aligned}
E(\widehat{T}_h(p)) &= \tfrac{1}{n}\sum_{i=1}^{n} \tfrac{1}{h} K\left(\tfrac{i/n-p}{h}\right) E(X_{(i)}) \\
&= \tfrac{1}{n}\sum_{i=1}^{n} \tfrac{1}{h} K\left(\tfrac{i/n-p}{h}\right) F^{-1}(i/n) \\
&\approx \int_0^1 \tfrac{1}{h} K\left(\tfrac{x-p}{h}\right) F^{-1}(x)dx \\
&= \int_{-p/h}^{(1-p)/h} K(u) F^{-1}(uh+p)du \\
&\approx F^{-1}(p) + \tfrac{h^2}{2}\left(F^{-1}\right)''(p) \int u^2 K(u)du + O(h^2)
\end{aligned}
$$

assuming $h$ to be small.

$$
\begin{aligned}
var(\widehat{T}_h(p)) &= \tfrac{1}{n^2}\textstyle\sum_{i,j}\operatorname{cov}(X_i,X_j)\tfrac{1}{h^2}K\left(\tfrac{i/n-p}{h}\right)K\left(\tfrac{j/n-p}{h}\right)\\
&= \tfrac{1}{n^2}\textstyle\sum_{i,j}\tfrac{1}{n}\tfrac{i}{n}(1-\tfrac{j}{n})(F^{-1})'(\tfrac{i}{n})(F^{-1})'(\tfrac{j}{n})\tfrac{1}{h^2}K\left(\tfrac{i/n-p}{h}\right)K\left(\tfrac{j/n-p}{h}\right)\\
&\approx \int_0^1\int_0^1 \tfrac{x(1-y)}{n}(F^{-1})'(x)(F^{-1})'(y)\tfrac{1}{h^2}K\left(\tfrac{x-p}{h}\right)K\left(\tfrac{y-p}{h}\right)dx\,dy\\
&= 2\int_0^1\left\{\int_0^y \tfrac{x(1-y)}{n}(F^{-1})'(x)(F^{-1})'(y)\tfrac{1}{h^2}K\left(\tfrac{x-p}{h}\right)K\left(\tfrac{y-p}{h}\right)dx\right\}dy\\
&= 2\int_0^1\left\{\int_{-p/h}^{(y-p)/h}\tfrac{1}{n}(uh+p)(1-y)(F^{-1})'(uh+p)(F^{-1})'(y)K(u)\tfrac{1}{h}K\left(\tfrac{y-p}{h}\right)du\right\}dy\\
&\approx 2\int_0^1\int_{-p/h}^{(y-p)/h}\tfrac{1}{n}(uh+p)(1-y)[(F^{-1})'(p)+uh(F^{-1})''(p)](F^{-1})'(y)\\
&\qquad .K(u)\tfrac{1}{h}K\left(\tfrac{y-p}{h}\right)du\,dy\\
&= 2\int_{-p/h}^{(1-p)/h}\int_{-p/h}^{v}\tfrac{1}{n}(uh+p)(1-p-vh)[(F^{-1})'(p)+uh(F^{-1})''(p)]\\
&\qquad .(F^{-1})'(p)+vh(F^{-1})''(p)]K(u)K(v)du\,dv\\
&= 2\textstyle\int\int\tfrac{1}{n}(uh+p)(1-p-vh)\{[(F^{-1})'(p)]^2+uh(F^{-1})'(p)(F^{-1})''(p)\\
&\qquad +vh(F^{-1})'(p)(F^{-1})''(p)\}K(u)K(v)du\,dv\\
&= \tfrac{2}{n}\textstyle\int\int\{p(1-p)+uh(1-p)-vhp\}K(u)K(v)\{[(F^{-1})'(p)]^2+\ldots\}du\,dv\\
&= p(1-p)[(F^{-1})'(p)]^2-\tfrac{1}{n}[(F^{-1})'(p)]^2 2h\textstyle\int uK(u)K^*(u)du+O(\tfrac{h}{n})
\end{aligned}
$$

where $K^*(u)=\int_u^\infty vK(v)\,\mathrm{dv}$ and assuming $h$ small, so that

$$\int_{-p/h}^{(1-p)/h}\int_{-p/h}^{v}K(u)K(v)du\,dv=-\frac{1}{2}$$

and

$$\int_{-p/h}^{(1-p)/h}\int_{-p/h}^{v}uK(u)K(v)du\,dv=-\int_{-p/h}^{(1-p)/h}\int_{-p/h}^{v}vK(u)K(v)du\,dv$$

# References


Aalen, 0. (1976). Nonparametric Inference in Connection with Multiple Decrement Models. *Scandinavian Journal of Statistics*, **3**, 15–27.

Akritas, M.G. (1986). Bootstrapping the Kaplan–Meier Estimator. *Journal of the American Statistical Association*, **81**, 1032–1038.

Almond, P.S., Matas, A., Gillingham, K., Dunn, D.L., Payne, W.D., Gores, P., Gruessner, R. and Najarian, J.S. (1993). Risk–Factors for Chronic Rejection in Renal–Allograft Recipients. *Transplantation*, **55**, 4, 752–757.

Aly, E.E., Csörgő, M. and Horváth, L. (1985). Strong Approximations of the Quantile Process of the Product-Limit Estimator. *Journal of Multivariate Analysis*, **16**, 185–210.

Azzalini, A. (1981). A Note on the Estimation of a Distribution Function and Quantiles by a Kernel Method. *Biometrika*, **68**, 1, 326–328.

Azzalini, A., Bowman, A.W. and Härdle,W. (1989) On the Use of Nonparametric Regression for Model Checking. *Biometrika*, **76**, 1, 1–11.

Beran, R. (1981). Nonparametric Regression with Randomly Censored Survival Data. *Technical Report, University of California, Berkley.*

Boularan, J., Ferré, L. and Vieu, P. (1994). Growth Curves : a Two–Stage Nonparametric approach. *Journal of Statistical Planning and Inference*, **38**, 327–350.

Bowman, A.W. (1984). An Alternative Method of Cross–Validation for the Smoothing of Density Estimates. *Biometrika*, **71**, 2, 353–360.

Bowman, A.W. (1985). A Comparative Study of Some Kernel–Based Nonparametric Density Estimators. *Journal of Statistical Computation and Simulation*, **21**, 313–327.

Breslow, N. and Crowley, J. (1974). A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship. *Annals of Statistics*, **2**, 3, 437–453.

Brookmeyer, R. and Crowley, J. (1982). A Confidence Interval for the Median Survival Time. *Biometrics*, **38**, 29–41.

Burke, M.D., Csörgő, S. and Horváth, L. (1981). Strong Approximations of Some Biometric Estimates under Random Censorship. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **56**, 87–112.

Burr, D. and Doss, H. (1993). Confidence Bands for the Median Survival Time as a Function of the Covariates in the Cox Model. *Journal of the American Statistical Association*, **88**, 1330–1340.

Chen, C.–H. and George, S.L. (1985). The Bootstrap and Identification of Prognostic Factors via Cox's Proportional Hazards Regression Model. *Statistics in Medicine*, **4**, 39–46.

Cheng, K.–F. (1984). On Almost Sure Representation for Quantiles of the Product Limit Estimator with Applications. *Sankhyā, Series A*, **46**, 3, 426–443.

Chung, C.–J.F., Csörgő, M. and Horváth, L. (1990). Confidence Bands for Quantile Function Under Random Censorship. *Annals of the Institute of Statistical Mathematics*, **42**, 1, 21–36.

Cleveland, W.S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Cole, T.J. (1988). Fitting Smoothed Centile Curves to Reference Data - with discussion. *Journal of the Royal Statistical Society - Series A*, **151**, 385–418.

Cole, T.J. and Green, P.J. (1992). Smoothing Reference Centile Curves : The LMS Method and Penalized Likelihood. *Statistics in Medicine*, **11**, 1305–1319.

Cox, D.R. (1972). Regression Models and Life Tables - with discussion. *Journal of the Royal Statistical Society - Series B*, **34**, 187–220.

Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant data. *Journal of the American Statistical Association*, **72**, 27–36.

Csörgő, S. and Horváth, L. (1986). Confidence Bands from Censored Samples. *Canadian Journal of Statistics*, **14**, 2, 131–144.

David, H.A. (1970). *Order Statistics.* Wiley Series in Probability and Mathematical Statistics.

Dabrowska, D.M. and Doksum, K.A. (1987). Estimates and Confidence Intervals for Median and Mean Life in the Proportional Hazards Model. *Biometrika*, **74**, 4, 799–807.

Doksum, K.A. and Yandell, B.S. (1982). Properties of Regression Estimates Based on Censored Survival Data, in *A Festscrift for Erich Lehmann*, eds. P.J. Bickel, K.A. Doksum and J.L. Hodges, Jr., Belmont, CA : Wadsworth International Group, 140–156.

Doss, H. and Gill, R.D. (1992). An Elementary Approach to Weak Convergence for Quantile Processes, with Applications to Censored Survival Data. *Journal of the American Statistical Association*, **87**, 869–877.

Efron, B. (1979). Bootstrap Methods : Another Look at the Jackknife. *Annals of Statistics*, **7**, 1, 1–26.

Efron, B. (1981). Censored Data and the Bootstrap. *Journal of the American Statistical Association*, **76**, 312–319.

Efron, B. and Gong, G. (1983). A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *The American Statistician*, **37**, 1, 36–48.

Efron, B. and Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, **1**, 1, 54–77.

Efron, B. and Tibshirani, R.J. (1993). An Introduction to the Bootstrap. *Chapman and Hall, Monographs on Statistics and Applied Probability*, **57**.

Falk, M. (1984). Asymptotic Normality of the Kernel Quantile Estimators. *Annals of Statistics*, **13**, 1, 428–433.

Falk, M. (1985). Relative Deficiency of Kernel Type Estimators of Quantiles. *Annals of Statistics*, **12**, 1, 261–268.

Falk, M. (1986). On the estimation of the Quantile Density Function. *Statistics and Probability Letters*, **4**, 69–73.

Fan J. and Gijbels I. (1994). Censored Regression : Local Linear Approximations and Their Applications. *Journal of the American Statistical Association*, **89**, 426, 560–570.

Gasser, T. and Müller (1979) Kernel Estimation of Regression Functions *In Lecture Notes in Mathematics, eds. Gasser and Rosenblatt. Stringer–Verlag : New York*, **757**, 23–68.

Gasser, T., Kneip, A. amd Köhler,W. (1991). A Flexible and Fast Method for Automatic Smoothing. *Journal of the American Statistical Society*, **86**, 415, 643–652.

Gentleman, R. and Crowley, J. (1991). Graphical Methods for Censored Data. *Journal of the American Statistical Association*, **86**, 678–683.

Gill, R.D. (1983). Large Sample Behaviour of the Product-Limit estimator on the whole line. *Annals of Statistics*, **11**, 49–58.

Gjertson, D.W. (1991) Survival Trends in Long Term First Cadaver Donor Kidney Transplants. *Terasaki, P.I. Ed. Clinical Transplants 1991. Los Angeles UCLA. Tissue Typing Authority*, 225–235.

Grégoire, G. (1993). Least Squares Cross–Validation for Counting Process Intensities. *Scandinavian Journal of Statistics*, **20**, 343–360.

Hamill, P.V.V., Drizd, T.A., Johnson, C.L., Reed, R.B. and Roche, A.F. (1977) NCHS Growth Curves for Children Birth–18 Years. *Washington D.C. : National Center for Health Statistics, Vital and Health Statistics*, **Series 11**, No. 165.

Härdle, W. and Marron, J.S. (1985) Optimal Bandwidth Selection in Nonparametric Regression Function Estimation. *Annals of Statistics*, **13**, 1465–1481.

Hastie, T.J. and Tibshirani, R.J. (1990). Generalised Additive Models. *Chapman and Hall.*

Healy, M.J.R., Rasbash, J. and Yang, M. (1988). Distribution–free estimation of age-related centiles. *Annals of Human Biology*, **15**, 1, 17–22.

Hjört, N.L. (1985) Bootstrapping Cox's Regression Model. *Technical Report 241, Stanford University, Dept. of Statistics.*

Horváth and Yandell, B.S. (1987). Convergence Rates for the Bootstrapped Product–Limit Process. *Annals of Statistics*, **15**, 3, 1155–1173.

Jones, M.C. In the Discussion to Cole (1988). *Journal of the Royal Statistical Society - Series A*, **151**, 412–413.

Jones, M.C. and Hall, P. (1990). Mean Squared Error Properties of Kernel Estimates of Regression Quantiles. *Statistics and Probability Letters*, **10**, 283–289.

Kalbfleisch, J.D. and Prentice, R.L. (1980) The Statistical Analysis of Failure Time Data. *Wiley Series in Probability and Mathematical Statistics.*

Kaplan, E.L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, **53**, 457–481.

Karrison, T. (1990). Bootstrapping Censored Data with Covariates. *Journal of Statistical Computation and Simulation*, **36**, 195–207.

Knapp, M.S., Smith, A.F.M., Trimble, I.M., Pownall, P. and Gordon, K. (1983). Mathematical and Statistical aids to evaluate data from Renal Patients. *Kidney International*, **24**, 474–486.

Krall, J.M., Uthoff, V.A. and Harley, J.B. (1975). A Step–up Procedure for Selecting Variables Associated with Survival. *Biometrics*, **31**, 49–57.

Le, C.T. and Grambsch, P.M. (1994). Tests of Association Between Survival Time and a Continuous Covariate. *Communications in Statistics – Theory and Methods*, **23**, 4, 1009–1019.

Lio, Y.L., Padgett, W.J. and Yu, K.F. (1986). On the Asymptotic Properties of a Kernel Type Estimator from Censored Samples. *Journal of Statistical Planning and Inference*, **14**, 169–177.

Lio, Y.L. and Padgett, W.J. (1987). On the Mean Squared Error of Nonparametric Quantile Estimators Under Random Right-Censorship. *Communications in Statistics - Theory and Methods*, **16**, 1617–1628.

Marron, J.S. (1988). Automatic Smoothing Parameter Selection : A Survey. *Empirical Economics*, **13**, 187–208.

Mickey, R., Cho, Y.W. and Carnahan, E. (1990) Long–term Graft Survival. *Terasaki, P.I. Ed. Clinical Transplants 1990. Los Angeles UCLA. Tissue Typing Authority*, 385–396.

Nadaraya, E.A. (1964). On Estimating Regression. *Theory of Probability and its Applications*, **9**, 141–142.

Padgett, W.J. (1986). A Kernel-Type Estimator of a Quantile Function from Right–Censored Data. *Journal of the American Statistical Association*, **81**, 215–222.

Padgett, W.J. and Thombs, L.A. (1986). Smooth Nonparametric Quantile Estimation Under Censoring : Simulation and Bootstrap Methods. *Communications in Statistics – Simulation and Computation*, **15**, 1003–1025.

Parzen, E. (1979). Nonparametric Statistical Data Modeling. *Journal of the American Statistical Association*, **74**, 365, 105–131.

Patil, P.N. (1993). Bandwidth Choice for Nonparametric Hazard Rate Estimation. *Journal of Statistical Planning and Inference*, **35**, 15–30.

Reid, N. (1981). Estimating the Median Survival Time. *Biometrika*, **68**, 3, 601–608.

Reiss, R.–D. (1980). Estimation of Quantiles in Certain Nonparametric Models. *Annals of Statistics*, **8**, 1, 81–105.

Rossiter, J.E. (1991). Calculating Centile Curves Using Kernel Density Estimation Methods With Application to Infant Kidney Lengths. *Statistics in Medicine*, **10**, 1693–1701.

Salomon, D.R. (1992). Cyclosporine Nephrotoxicity and Long Term Renal Transplantation. *Transplantation Reviews*, **6**, 10–19.

Sander (1975). The Weak Convergence of Quantiles of the Product Limit Estimator. *Technical Report No. 5*, Department of Statistics, Stanford University, Stanford.

Sheather, S.J. and Jones, M.C. (1991). A Reliable Data–Based Bandwidth Selection Method for Kernel Density Estimation. *Journal of the Royal Statistical Society, Series B*, **53**, 3, 683–690.

Sheather, S.J. and Marron, J.S. (1990). Kernel Quantile Estimators. *Journal of the American Statistical Association*, **85**, 410–416.

Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. *Chapman and Hall : London.*

Skene, A.I., Smith, J.M., Doré, C.J., Charlett,A. and Lewis, J.D. (1992). Venous Leg Ulcers : A Prognostic Index to Predict Time to Healing. *British Medical Journal*, **305**, 1119–1121.

Slud, E.V., Byar, D.P. and Green, S.B. (1984). A Comparison of Reflected Versus Test–Based Confidence Intervals for the Median Survival Time, Based on Censored Data. *Biometrics*, **40**, 587–600.

Smith, A.F.M. and Cook, D.G. (1980). Straight lines with a Change-Point : A Bayesian Analysis of some Renal Transplant Data. *Applied Statistics*, **29**, 2, 180–189.

Smith, A.F.M. and West, M. (1983). Monitoring Renal Transplants : An Application of the Multiprocess Kalman Filter. *Biometrics* **39**, 867–878.

Stigler, S.M. (1974). Linear Functions of Order Statistics with Smooth Weight Functions. *Annals of Statistics*, **2**, 4, 676–693.

Stone, C.J. (1977). Consistent Nonparametric Regression. *Annals of Statistics*, **5**, 4, 595–645.

Stoodley, K.D.C. and Mirnia, M. (1979). The Automatic Detection of Transients, Step Changes and Slope Changes in the Monitoring of Medical Time Series. *The Statistician* **28**, 3, 163–170.

Tanner, M.A. and Wong, W.H. (1984). Data–based Nonparametric Estimation of the Hazard Function with Applications to Model Diagnostics and Exploratory Analysis. *Journal of the American Statistical Association*, **79**, 385, 174–182.

Tibshirani, R. and Hastie, T. (1987). Local Likelihood Estimation. *Journal of the American Statistical Association*, **82**, 559–567.

Watson, G.S. (1964). Smooth Regression Analysis. *Sankhyā, Series A*, **26**, 359–372.

Yang, S.–S. (1985). A Smooth Nonparametric Estimator of a Quantile Function. *Journal of the American Statistical Association*, **80**, 1004–1011.