Modelling a Time-Dependent Hazard Ratio with Regression Splines

Todd Mackenzie
McGill University, Montreal

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Science.

July 24, 1993

Contents

Abstract

The proportional hazards model proposed by Cox(1972) is by far the most popular method of regressing survival data. This model is attractive because: (i) It has a simple interpretation; the impact of a variable upon survival is a constant and multiplicative effect on the hazard function. (ii) It facilitates the employment of the partial likelihood inference technique so that it requires no assumptions about the baseline distribution of survival times. Many numerical tests as well as graphical approaches have been proposed for assessing the adequacy of the proportional hazards model. However only a few authors have discussed strategies for modelling data for which the hazard ratio varies over time.

In this thesis the topic of survival analysis is overviewed, and methods for assessing the validity of the proportional hazards assumption are reviewed. Finally a method of estimating the hazard ratio as a flexible function of time using the method of regression splines and the AIC model selection criterion is proposed. We report the results of a simulation meant to examine the small sample properties of this technique.

Résumé

Le modèle à hasard proportionnel présenté par Cox (1972) est une méthode de régression trés utilisée dans l'analyse de survie. Ce modèle est intéressant, car (i) il est simple à interpréter, l'impact d'une variable sur la survie ayant un effet constant et multiplicatif sur la fonction de hasard, et (ii) il permet l'utilisation de la technique d'inference par la vraisemblance partielle, ce qui n'oblige aucune supposition sur la distribution de base des temps de survie. De nombreux tests et méthodes graphiques existent pour mesurer l'adéquation du modèle à hasard proportionnel. Toutefois, peu d'ouvrages discutent des stratégies pour la modélisation des données ayant un taux de hasard variable en temps.

Dans cette étude, nous passons en revue les caractérisations essentielles à l'analyse de survie et les méthodes pour mesurer la validité de l'hypothèse d'un modèle à hasard proportionnel. Par la suite, nous présentons un estimateur flexible du taux de hasard, variable en temps, par la méthode des splines de régression et la sélection d'un modèle, par le critère AIC. Enfin, nous rapportons les résultats d'une simulation qui examine les propriétés des petits échantillons.

Acknowledgements

I thank, above everyone else, my wife, Amanda, for being patient as I took time away from childcare to work on this thesis. And maybe now we can put some curtains up in our new apartment. I also thank Delaney and Ambrose for tolerating less dad during the past few months. Thanks also goes to Mum, Margo and Adrienne for helping us out so I could work on my thesis.

I also thank my supervisor, Dr.Michal Abrahamowicz, for his enthusiasm and guidance. I will keep his artful page covering corrections forever.

I would also like to acknowledge the contribution of Dr. David Wolfson who acted as liason with the Department of mathematics and statistics, and Dr. John Esdaile, my employer, whose research originally stimulated the ideas behind this thesis. Furthermore, I would like to thank Roxanne de Berger for her assistance in editing this thesis and for her translation of the abstract into the french language.æ

Introduction

Cox's proportional hazards model is one of the most popular statistical techniques in health care research. Its employment rivals that of t-tests, and 2×2 contingency tables. This is used in regression applications where the dependent variable is a survival or some elapsed time which is subject to censoring. The appeal of this model may be explained by the following properties: (i) it has a relatively simple interpretation, and (ii) it is semi-parametric since it avoids assumptions about the underlying distribution of survival times.

Despite its overwhelming popularity, analysts who employ this model rarely if ever report checking its main assumption, that the hazard functions of any two subjects are proportional. The impact of an independent variable on, or association with, the hazard function is not mitigated by the passage of time. A practical implication of this assumption is that a variable is equally able at predicting short-term and long-term survival. Another way of stating this is that, say, a laboratory test recorded today is not more valuable in determining risk than is a recording from any time in the past. This constancy of the relative risk over time may be approximately true in studies of a suitably short duration but there is no reason to expect it to be

true in general. The widespread failure of analysts to check the assumptions of the model is not due to a lack of graphical or numerical methods for testing the proportionality of hazards. For instance, at least 13 numerical tests have been proposed in the literature.

Alternative modelling strategies that model the relative risk as a function of time would be valuable to health care practitioners. Characteristics that have only a short-term ability to predict events or have a delayed impact on risk may go undetected by the proportional hazards method. In clinical trials it would be useful to determine if the trial treatment is actually effective during the entire follow-up of the study and not just in the short term. In observational studies it would be valuable to know whether an exposure has an immediate effect on risk or if the effect does not occur until some time after the exposure.

Zucker and Karr (1990) propose the modelling of the hazard ratio as a function of time by using smoothing splines. In this thesis the modelling of the hazard ratio by another smoothing method, the method of regression splines is proposed and evaluated in a preliminary simulation study.

In chapter 1 the topic of survival analysis is outlined. Survival analysis, also variously called, failure-time analysis and response-time analysis, is basically the theory and methodology of analyzing data from health care research where the focal measurement, or dependent variable, is an elapsed time. In this chapter, the partial likelihood is presented, and we give heuristic proofs of it is asymptotic properties. The proportional hazards model is introduced

as a method that facilitates the partial likelihood approach. In the final section we draw the link between the proportional hazards model and the log rank test, a test for comparing survival between two samples, and illustrate the assumptions implicitly made about the hazard ratio as function of time by some other popular two sample test statistics.

In chapter 2, we discuss the methods of validating the assumption of proportional hazards. This chapter reviews numerical and graphical tests of assessing the constancy of the hazard ratio, or rather the proportionality of hazards. Particular emphasis is placed on the residuals approach proposed by Schoenfield (1982).

In chapter 3, we discuss methods for modelling the hazard ratio as a function of time. We describe the smoothing spline approach proposed by Zucker and Karr (1990). Then we propose a method that uses regression splines and the model selection criterion proposed by Akaike. We refer to this as the best-AIC regression spline approach. We discuss heuristically its the large sample properties of this technique and make some relevant observations concerning the impact of model selection upon inference.

Finally, in chapter 4, we describe, and report the results of, a simulation meant to examine the sample properties of the best-AIC regression spline approach.

Throughout this manuscript we shall refer to the following data set, reported in Fleming and Harrington (1991) and frequently used to illustrate new survival analytic methods. Between 1974 and 1984 the Mayo Clinic con-

ducted a trial of the effects of the drug D-penicillamine on persons suffering from primary biliary cirrhosis (PBC) of the liver. This is a rare but chronic and fatal disease of the liver whose cause is unknown. The treatment was ultimately deemed ineffective. However, for each subject a number of clinical features and laboratory markers from the time of diagnosis were recorded which makes it possible to examine their possible association with subsequent survival.

We shall refer repeatedly to one of these measurements, prothrombin time. Prothrombin is an agent in blood that is responsible for coagulation. Prothrombin time is the duration of time, usually about 10 seconds in these subjects, required to achieve coagulation in a test tube of the subjects's blood. We shall use this variable to compare the results of different statistical methods discussed in the thesis and to illustrate the ability of the proposed model to provide new insights into the structure of survival data.

A total of 424 persons with PBC were referred to the Mayo clinic during the 10-year period of the study and 312 agreed to participate in the trial. All but 6 of the remaining persons consented to undergo measurements. Of these 418 subjects, 161 were observed to die during the course of the study.

Chapter 1

Analysis of Survival Data

1.1 Follow-up data and censoring

In health statistics one of the most common measurements is an elapsed time. Examples are the time from the diagnosis of cancer until death, or the time from the initiation of some treatment to death or some other event such as heart attack or stroke. In each case there is some well-defined starting point, such as the date of diagnosis or the start of treatment and a well defined endpoint such as death. This type of data is called *follow-up data.* There is broadly speaking two types of studies in which follow-up data is collected. One is the clinical trial, and the other is the observational study.

A (randomized) clinical trial is the evaluation of the efficacy of a drug or more generally any intervention by enrolling a group of subjects, usually subject to some *inclusion criteria*, and administering the treatment to some and not to others (randomly), and by finally comparing the treated with the untreated persons. Often efficacy is defined to be the ability of the treatment

to increase survival. In this case the enrollees would be followed up for some suitable period of time in order to compare the survival between the treated and untreated group.

An observational study is the evaluation of the association between some relevant endpoint, for instance, survival, and the level of, or presence or absence of some exposure or characteristic. For instance, an oncologist could follow up a group of persons newly diagnosed with cancer in order to assess the association between the size of the tumour they present with and their subsequent survival.

It is the nature of the time dimension that we cannot always completely measure survival, or more generally speaking, durations. It usually happens that on the date of the analysis one or more of the subjects in the study are still alive or have not experienced the particular event being examined. On the other hand, perhaps some of these persons could not be followed up for logistical reasons, i.e.; perhaps one of these persons are known to have lived for 5 years after the diagnosis after which they they moved, and the investigators lack knowledge of what happened to them subsequently. Some clinical studies may be predetermined to terminate after a fixed number of deaths or events have been observed leaving the exact survival of the remaining subjects unobserved. In statistics, this type of partial information is referred to as (right) censored data, (Kalbfleisch and Prentice, 1980).

It is very unusual to encounter follow-up data which is complete, that is, uncensored, for every subject. It is not uncommon to encounter studies

in which 50% or more of the follow-ups are censored. It would be wrong to exclude the subjects for which the complete duration is not known. Censored follow-up data provides some information. Furthermore discarding such partial information creates bias in the estimates since a person whose duration is long is more likely to be censored than a person whose duration is short.

The analysis of censored follow-up data is popularly referred to as survival analysis. The term, survival, is used despite the fact that it is just as common to examine non-death endpoints such as heart attacks or strokes.

In this chapter, we shall review some of the major concepts from survival analysis methodology. In section 2, we introduce the survival function, as well as the hazard function, and discuss its estimation. In section 3, we review methods for comparing survival between two. These groups may be determined by the presence or absence of some clinical feature or exposure or may represent the treated and untreated arms of a clinical trial. One method of comparison is the log-rank test. We show how other popular tests can be expressed as 'weighted' log-rank tests. In section 4, we go beyond comparisons between groups and discuss regression methods for quantitating the association between survival and one or more independent variables. In this section, we motivate the idea of the partial likelihood which is the main topic of section 5. In section 5, we also introduce the proportional hazards model and demonstrate how it facilitates use of the partial likelihood. In section 6, the partial likelihood is discussed further. We demonstrate that its use is justified when the number of persons in a study observed to die, or more

generally speaking, experience the examined event, is suitably large. The partial likelihood facilitates the use of *time-dependent* independent variables. These are discussed in section 7. In section 8, we demonstrate the correspondence between the log-rank statistic and the proportional hazards model. In this section also make some interesting observations concerning the relation between other weighted log-rank tests and particular hazard ratio models.

In this chapter, we emphasize the overwhelming popularity and appeal of non-parametric and semi-parametric methods over parametric methods. We discuss how the former became developed only when the methodologists switched their center of attention from time as the *dependent variable*, to the complementary notion of risk.

1.2 Estimation of the survival function

In many statistical settings, it is usually informative to calculate a mean or median as well as a standard deviation or interquartile range. When the variables of interest are survival times, or durations, these summary statistics are also helpful. However it is far more popular to report, instead of this single summary measure, the *survival function*. The survival function, $S(t)$, also referred to as the *survival curve*, is plotted along the time axis and estimates for each time the probabilities that a person would survive that period of time or longer. It T is a random variate from a distribution of survival times, the $S(t) = Pr[T > t]$. The survival function is just the complementary cumulative distribution function.

One approach to estimation of the survival function is to propose a parametric form, $S_\theta(t)$, and proceed to estimate θ and therefore S_θ using an approach such as maximum likelihood. Some possibilities for this parametric form, are the Weibull, $S_{\rho,\kappa}(t) = \exp[-(\rho t)^\kappa]$, and the log-logistic, $S_{\rho,\kappa}(t) = [1 + (\rho t)^\kappa]^{-1}$, as well as the Gamma, and the log-normal, for whom closed forms for the survival function do not exist. See Cox and Oakes (1984) or Kalbfleisch and Prentice (1980) for a thorough discussion of parametric statistical models for survival analysis.

A special case of the Weibull distribution is the exponential distribution which occurs when $\kappa = 1$. The exponential is characterized by the following *memory-less* property. Let T be a random variate from an exponential distribution, and t and u two positive real values then,

$$Pr[T \in [t, t+u) | T \geq t] = Pr[T \in (0, u)]. \tag{1.1}$$

The memory-less property is equivalent to stating that the *hazard function*, $\lambda(t)$, is constant with respect to time, t, where,

$$\lambda(t) = \lim_{\Delta t \to 0} Pr[T \in [t, t+\Delta t) | T \geq t]. \tag{1.2}$$

The hazard function, or just hazard, is also known as the *instantaneous risk.*

The hazard function and the survival function play a central role in survival analysis, just as the density function and the cumulative distribution function play a central role in most other areas of statistics. One can rewrite the right side of equation (1.2) as $-\frac{d}{dt}S(t)/S(t)$ or $f(t)/S(t)$ where $f(t)$ is the

probability density function of survival times. Often it is more meaningful to propose a parametric form for the hazard function instead of the survival function. In this case the survival function can be obtained from the hazard using the formula,

$$S(t) = \exp(-\int_0^t \lambda(t)\,dt). \tag{1.3}$$

The cumulative hazard, $\Lambda(t) = \int_0^t \lambda(t)\,dt = -\log S(t)$, is also a popular quantity in survival analysis.

The hazard function for the Weibull distribution has the parameterization $\kappa\rho(\rho t)^{\kappa-1}$. When κ exceeds 1 this hazard function is strictly increasing whereas the hazard is strictly decreasing when κ is less than 1. A clinical setting where the latter might be true is survival following an operation where patients settle down to a low risk status after living through a high-risk period immediately following the operation.

Having proposed a parametric form for either the survival or the hazard function we could proceed to their estimation by writing the likelihood function. Suppose n persons have been followed up. Typically, survival data is recorded as pairs. The pair (t, δ) consists of a time measurment and a binary variate which is typically 1 if the full duration has been observed and 0 if the duration has been censored. To write the likelihood of observing a pair (T, Δ) it is necessary to recognize the following,

$$\begin{aligned} T &= \min(T^0, C), \\ \Delta &= 1_{T=T^0}. \end{aligned} \tag{1.4}$$

where T^0 is true survival time and C is the censoring time. If we assume that T^0 and C are independent the likelihood of the pair (t, δ) is,

$$L = \begin{cases} dPr(T^0 < t)Pr(C > t), & \text{if } \delta = 1; \\ Pr(T^0 > t)dPr(C < t), & \text{if } \delta = 0. \end{cases} \tag{1.5}$$

This can be rewritten as

$$L = \{dPr(T^0 < t)Pr(C > t)\}^{\delta} \{Pr(T^0 > t)dPr(C < t)\}^{1-\delta} \tag{1.6}$$

which becomes

$$L(\theta) = \{dS_\theta(t)Pr(C > t)\}^{\delta} \{S_\theta(t)dPr(C < t)\}^{1-\delta} \tag{1.7}$$

upon substitution of $S_\theta(t)$ for $Pr(T^0 > t)$. Typically it is assumed that the distribution function of the censoring time, $Pr(C < t)$, carries no information about θ. This is referred to as *noninformative censoring* (Kalbfleisch and Prentice, 1980). Therefore the θ maximizing $L(\theta)$ is the same θ maximizing $dS_\theta(t)^{\delta} S_\theta(t)^{1-\delta} = \lambda_\theta(t)^{\delta} S_\theta(t)dt$. If n persons are followed up yielding the observations $(t_1, \delta_1), \ldots, (t_n, \delta_n)$ and we assume that their survival times and censoring times are **independent** of one another the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \lambda_\theta(t_i)^{\delta} S_\theta(t_i). \tag{1.8}$$

The assumption of the independence of the survival times and censoring times is usually reasonable in clinical trials and observational studies. It would be wrong to assume independence if subjects were censored according

to some characteristics that are related to survival. For instance, if subjects in the PBC study had been withdrawn from the study because their health status either improved or deteriorated substantially the censoring times would not be independent of their survival times.

It is not necessary to assume that T^0 and C are independent to derive the likelihood in (1.8). It is sufficient that,

$$Pr(T^0 \in [t, t+\Delta t) | T^0 \geq t, C \geq t) = Pr(T^0 \in [t, t+\Delta t) | T^0 \geq t). \quad (1.9)$$

This condition is referred to as *weak independence*. Deriving the likelihood in (1.8) using only this condition is more difficult. It involves the partition of the time axis into an infinite number of infinitesimally small intervals. See Kalbfleisch and Prentice (1980) for details.

Choosing a particular parametric model is usually arbitrary although the choice may be guided by *a posteriori* model selection criteria. Employing a non-parametric estimator avoids this arbitrariness. Partly for this reason non-parametric methods are favoured in survival analysis. One example of a non-parametric method is the *Kaplan-Meier* estimator (Kaplan and Meier, 1958). This estimator is usually presented in any article in the medical literature in which survival or durations are being examined.

The Kaplan-Meier estimator of the survival function has an intuitive form when expressed in terms of *risk sets*. The risk set at time t is the set of the subjects known to be alive (or to have not yet incurred the particular endpoint being examined) at time t after the beginning of follow-up. A

subject is part of the risk set at time t if the following 2 conditions are true; (i) the subject survives t units or longer and (ii) the subject is not censored before time t. So for instance all subjects are part of the set at time 0. Usually we are not interested in the continuum of risk sets, but the risk sets evaluated **immediately prior** to each observed death (event). The first risk set consists of all subjects except any subjects censored before the time of the first observed failure. The second risk set consists of all subjects except the subject first observed to fail and any subjects censored before the time of the second observed failure.

The Kaplan-Meier estimator, also known as the product-limit estimator, is calculated as follows. Let $t_1^* < \ldots < t_k^*$ be the (unique) times at which deaths (events) occur and $R_1, ..., R_k$ be the corresponding risk sets. The Kaplan-Meier estimate is given by

$$\hat{S}_{KM}(t) = \prod_{i:t_i^* < t} (1 - 1/|R_i|). \tag{1.10}$$

The Kaplan-Meier puts all its mass at the k death times. Its derivation is based on the following chain-like identity of conditional probabilities

$$Pr[U > u_n] = \prod_{1}^{n} Pr[U > u_i | U > u_{i-1}] \tag{1.11}$$

where $0 = u_0 < u_1 < ... < u_n$. It uses $1 - 1/|R_i|$ as the estimate of $Pr[T > t_i | T > t_i - \epsilon]$ and 1 as the estimate of $Pr[T > t_i - \in | T > t_{i-1}]$. Peterson (1977) showed that the Kaplan-Meier is a consistent estimator of $S(t)$.

Another, very similar, estimator of the survivor function is based upon the estimator given by Nelson (1969). Nelson estimated the cumulative hazard $\Lambda(t) = \int_0^t \lambda(u)\, du$ as

$$\hat{\Lambda}(t) = \sum_{t_i^* < t} 1/|R_i|. \tag{1.12}$$

The estimate of the survivor function is obtained by taking $\hat{S}(t) = \exp[\hat{\Lambda}(t)]$. It is worth noting that the Nelson estimator can be expressed as a first order Taylor expansion of the logarithm of the Kaplan-Meier estimator since

$$\log S_{KM}(t) = \sum_{t_i^* < t} \log(1 - 1/|R_i|) \approx \sum_{t_i^* < t} 1/|R_i| = \hat{\Lambda}(t). \tag{1.13}$$

1.3 Comparison of risks

A health care researcher who wishes to assess if some characteristic or exposure is a possible cause of some disease or adverse outcome would calculate the observed risk of subjects with the characteristic and compare it to the calculated observed risk of subjects without the characteristic. In a clinical trial, the goal is to compare the outcome of subjects receiving a new treatment with subjects receiving a conventional treatment or placebo. In observational studies, we usually refer to characteristics that are associated with an adverse outcome as risk factors for that adverse outcome and we say that the characteristics have *predictive ability.*

To compare the risks of two groups for whom survival data has been observed, one method is to compare the estimates of their respective sur-

vival functions. Figure 1 depicts two Kaplan-Meier estimates superimposed on the same graph. The laboratory marker prothrombin time recorded in the PBC study has been dichotomized into *low* and *high* values using the cut of 11 seconds. This threshold was chosen since it is the closest integer value to the median prothrombin time which is 10.6 seconds. Kaplan-Meier estimates have been plotted for the low (thin line) and high value (thick line) groups. The striking difference between the two estimates suggests that the corresponding survivor functions that they estimate are different. We say that persons with high values are more at risk of death, in other words, prothrombin time has the ability to predict survival.

The superimposition of the two Kaplan-Meier estimates facilitates their comparison. We would like to formalize this comparison procedure by constructing a hypothesis test. If there was no censoring the standard non-parametric approach would be to use the Wilcoxon test statistic (Lehmann, 1975). Let $t_1, \ldots t_n$ be the uniquely valued observed survival times, the Wilcoxon statistic can be expressed as

$$U = \sum_{\substack{i=1..n \\ j=1..n}} I_{a_i \neq a_j} U_{ij}, \tag{1.14}$$

where a_i is 0 or 1 according as the i-th subject does or does not have the characteristic, I is an indicator, and U_{ij}, assuming no ties, is defined as 1 if $T_i > T_j$ and -1 otherwise. Gehan (1967) naturally adapted this formulation to censored data. He defined

$$U_{ij} = \begin{cases} +1, & \text{if } t_i > t_j \text{ and } \delta_j = 1; \\ -1, & \text{if } t_i < t_j \text{ and } \delta_i = 1. \end{cases} \tag{1.15}$$

Mantel (1966) proposed another test which marked a breakthrough in the way survival data is treated. In constructing a test he focused on the idea of risk and not on durations as does Gehan's method. Mantel reformulated the problem in terms of a series of 2 by 2 tables and applied the method of stratified contingency tables that he developed in an earlier paper (Mantel and Haenszel, 1959). The principle underlying this method is to compare each of the subjects who have died with the risk set evaluated just prior to their death. Let O_i be 0 or 1, respectively, (or 1 or 2) depending on whether the i-th death corresponds to a person with, or without the characteristic. O_i is an observation from a hypergeometric distribution. Let $E_i = E[O_i|R_i]$ and $V_i = Var[O_i|R_i]$, be the expected value and variance respectively of O_i given knowledge of the risk set just prior to the death. Under the hypothesis that the characteristic has no predictive ability, O_i is an observation from a hypergeometric distribution, and $E_i = \sum_{j \in R_i} O_j/|R_i|$ and $V_i = \sum_{j \in R_i} O_j * O_j/|R_i| - E_i^2 = E_i - E_i^2$. The Mantel-Haenszel test statistic is

$$\sum_{1}^{k}(O_i - E_i)^2 / \sum_{1}^{k} V_i, \tag{1.16}$$

where k is the total number of deaths. This statistic is non-parametric because the actual times at which the deaths occurred are not used. Their order is used implicitly and in entirety in the construction of the risk sets. If the characteristic has no predictive ability the Mantel-Haenszel test statistic has an approximate normal distribution with null mean and unit variance

which becomes exact asymptotically (Crowley, 1974).

The log-rank test applied to the dichotomized version of the laboratory marker prothrombin time, yields a p-value less than 0.0001 confirming the ability of prothrombin time to predict survival.

The Mantel-Haenszel is usually referred to as the *log-rank* statistic. This name dates to a paper by Peto and Peto (1972), in which they demonstrated that the Mantel-Haenszel is one generalization of Savage's exponential scores test to censored data, which can be interpreted as a sum of logarithm of ranks.

The expression for the log-rank in (1.16) may be generalized by the incorporation of *weights*, $W_1, \ldots, W_k$ as follows,

$$\sum_{i=1}^{k} W_i^2 (O_i - E_i)^2 / \sum_{1}^{k} W_i^2 V_i. \tag{1.17}$$

This is referred to as a *weighted log-rank* statistic. The standard log-rank is recaptured when $W_1 = \cdots = W_k = 1$. When $W_i = |R_i|$, the number of persons at risk just prior to the i-th death, it becomes Gehan's statistic. The Gehan statistic is criticized because the weights $W_i = |R_i|$ depend on the censoring distribution. Indeed, for large n, $|R_i|/n$ is close to $S(t_i^*)S_C(t_i^*)$ where t_i^* are observed failure times and $S_C(t)$ is the probability of not being censored before time t. Prentice (1978) proposed the weighting scheme $W_i = \hat{S}(t_i^*)$ where $\hat{S}$ is any estimator of the combined survival curve for the two groups, for instance, the Kaplan-Meier estimate. Prentice's test, like Gehan's, becomes the Wilcoxon test when there is no censored data.

It is not difficult to generalize the log-rank statistic in (1.16) to ordinal or

continuous data. Earlier we dichotomized the laboratory marker prothrombin time in order to test if subjects with high values had a higher or lower risk of death than subjects with low values but we should be able to use it in its undichotomized form. In the former case each failure corresponds to sampling a 0 or 1 from the set of 0 and 1's indexed by the corresponding risk set. In the latter case each O_i becomes a 1-sample from a discrete distribution whose mass points correspond to the values of the variable, for example prothrombin time, for the subjects in R_i. In section 8 we will demonstrate that this continuous version of the log-rank statistic is actually a score statistic corresponding to the type of regression Cox introduced in 1972.

1.4 Estimating the relative risk

Weighted log-rank statistics are useful for coming to conclusions of the type 'prothrombin time has the ability to predict survival' or 'treated subjects survive longer than untreated subjects'. This may not be enough in some situations. We may want to quantify the difference in survival. One such way would be to estimate the difference in mean or median survival. Another is to estimate the difference or ratio of the risks.

Consider the comparison of subjects with low and high values of prothrombin time. Let $\lambda_0(t)$ and $\lambda_1(t)$ be the respective hazard functions. Suppose the ratio of the hazard functions just before the time of the i-th observed failure, t_i^*, is $\rho_i = \lambda_1(t_i^*-)/\lambda_0(t_i^*-)$. For instance at the time just before the i-th failure a subject with a high prothrombin time (as measured at time 0) is

ρ_i times more likely to fail in the next instant than a subject with a low prothrombin time. The conditional probability that the i-th failure corresponds to a subject with high prothrombin time is

$$\frac{n_{i1}\lambda_1(t_i^*)}{n_{i0}\lambda_0(t_i^*) + n_{i1}\lambda_1(t_i^*)} \tag{1.18}$$

or

$$\frac{n_{i1}\rho_i}{n_{i0} + n_{i1}\rho_i} \tag{1.19}$$

where n_{i0} and n_{i1} are respectively the number of persons with low and high values of prothrombin time at risk just prior to the i-th failure. The conditional probability that the i-th failure corresponds to a subject with high prothrombin time is

$$\frac{n_{i0}}{n_{i0}\rho_i + n_{i1}}. \tag{1.20}$$

By multiplying the each of these probabilities we obtain the following pseudo likelihood

$$\prod_{i=1}^{k} \frac{n_{i0}^{1-O_i}(n_{i1}\rho_i)^{O_i}}{n_{i0} + n_{i1}\rho_i}, \tag{1.21}$$

where O_1 is 1 if the subject failing has high prothrombin time and 0 otherwise. This pseudo-likelihood is not proportional to a probability, and therefore not a true likelihood, since the probabilities we have multiplied do not correspond to independent events. However, it is intuitive that these respective events are not very dependent.

This particular model for the risk ratio's is overparameterized. We have a unique parameter for each failure time. One way of reducing the number of parameters is to assume that $\rho_1 = \rho_2 = ...\rho_k = \rho$ or in effect that at all times during the follow-up, the instantaneous relative risk of death of a subject with high prothrombin is ρ times more likely than the instantaneous relative risk of a subject with low prothrombin. To estimate ρ we could choose that ρ maximizing expression (1.21). Doing so we yields the estimate $\hat{\rho} = 3.3$. The instantaneous risk of subjects with high prothrombin times is on average 3.3 times higher than subjects with low prothrombin times.

In the next section, this method of creating a likelihood is formalized.

1.5 Partial likelihood and the proportional hazards model

The methods we introduce in this section are due to Cox (1972). Cox's contribution was two complementary concepts. The first was to create a likelihood whose terms did not come from strictly independent events. This later became known as the *partial likelihood* (Cox, 1975). The second was the proposal of a model for the way in which covariates affect the hazard function that yields a simple form for the partial likelihood. This is called the *proportional hazards model.*

1.5.1 Partial likelihood

Cox extended Mantel's (1967) idea of conditioning on the risk sets in order to yield test statistics. In this way he obtained the partial likelihood, which may be used to yield maximum likelihood type estimates for the parameters of, for instance, regression models. The appeal of this risk set approach is two-fold: (i) No distributional assumptions about the (baseline) hazard function are necessary, but as we shall show we do parameterize the way in which variables modify the hazard function and (ii) It allows the incorporation of *time-dependent* variables which we shall define later.

Cox associates exactly one term in the partial likelihood with each observed failure. The likelihood term he associates with the first failure is

$$
\begin{aligned}
L_1 &= Pr[T_1^* = t_1^* | T_j \geq t_1^* \; \forall j \in R_1 \text{ and } T_j = t_1^* \text{ for exactly one } j] \\
&= \frac{Pr[T_1^* = t_1^* \text{ and } T_j \geq t_1^* \; \forall j \in R_1 \text{ and } T_j = t_1^* \text{ for exactly one } j \in R_i]}{Pr[T_j \geq t_1^* \; \forall j \in R_1 \text{ and } T_j = t_1^* \text{ for exactly one } j \in R_i]} \\
&= \frac{Pr[T_1^* = t_1^* \text{ and } T_j \geq t_1^* \; \forall j \in R_1]}{Pr[T_j \geq t_1^* \; \forall j \in R_1 \text{ and } T_j = t_1^* \text{ for exactly one } j \in R_i]} \\
&= \frac{Pr[T_j \geq t_1^* \; \forall j \in R_1] Pr[T_1^* = t_1^* | T_1^* \geq t_1^*]}{Pr[T_j \geq t_1^* \; \forall j \in R_1] \sum_{j \in R_1} Pr[T_j = t_1^* | T_j \geq t_1^*]} \\
&= \frac{Pr[T_1^* = t_1^* | T_1^* \geq t_1^*]}{\sum_{j \in R_1} Pr[T_j = t_1^* | T_j \geq t_1^*]} \\
&= \frac{\lambda_1(t_1^*)}{\sum_{j \in R_1} \lambda_j(t_1^*)} \qquad (1.22)
\end{aligned}
$$

where λ_i and S_i are, respectively, the hazard and survival functions of the i-th subject. In the partial likelihood approach, a random variable is *induced* at each observed failure time by conditioning on the corresponding risk set, or rather, on all the information about failures and censorings prior to the time of the i-th failure, which we shall denote by H_i. Let A_i be the discrete random variable which takes the value of the index of the subject having the i-th observed failure, so that

$$Pr[A_i = a_i|R_i] = \begin{cases} \frac{\lambda_{a_i}(t_i^*)}{\sum_{j\in R_i}\lambda_j(t_i^*)} & \text{if } j \in R_i; \\ 0 & \text{otherwise.} \end{cases} \tag{1.23}$$

Cox argued that the event A_i given H_i is rather independent of the event A_j given H_j. For instance, for $i > j$,

$$\begin{aligned} Pr[(A_i = a_i|H_i = h_i)|A_j = a_j] &= Pr[A_i = a_i|H_i = h_i \text{ and } A_j = a_j] \\ &= Pr[A_i = a_i|H_i = h_i], \end{aligned} \tag{1.24}$$

since H_i, the complete history of survival up to time t_i^*, contains the information on the failure, $A_j = a_j$, at time $t_j^* < t_i^*$. Using this notion of independence Cox proposed the likelihood as the product of the k factors, one for each failure time, of the form in (1.22),

$$PL = \prod_1^k \frac{\lambda_i(t_i^*)}{\sum_{j\in R_i}\lambda_j(t_i^*)}, \tag{1.25}$$

1.5.2 The proportional hazards model

Cox's second idea was to propose a form for the way independent variables affect the hazard function that took advantage of the partial likelihood's

structure. Cox proposed the following factorization of the hazard function

$$\lambda(t;x) = \lambda_0(t)e^{\beta x}, \tag{1.26}$$

where $\lambda(t;x)$ is the hazard function of a subject whose value for the independent variable is x. This factorization is referred to as the proportional hazards model. It is semi-parametric: The effect of the covariate x on the hazard is parameterized but the effect of time is not. The first factor on the right in (1.26) is referred to as the *baseline hazard* function. The second factor is the hazard ratio or instantaneous relative risk. When expression (1.26) is substituted into (1.25) the baseline hazard cancels out of the expression yielding

$$\prod_{i=1}^{k} \frac{e^{\beta x_i^*}}{\sum_{j \in R_i} e^{\beta x_j}}. \tag{1.27}$$

In this model, the parameter β measures the effect of the variable x on the risk of failure. When x is a dichotomous variable e^{β} is the ratio of hazards between the two groups. In general, e^{β} is the multiplicative effect on the instantaneous risk of increasing x by 1. Cox proposed that the partial likelihood be maximized with respect to β in order to yield estimates $\hat{\beta}$. The absolute value of $\hat{\beta}$ expresses the strength of the predictive ability of the characterisic x.

Expression (1.27) is independent of the actual failure times. For this reason $\hat{\beta}$ is called a *semi-parametric* estimator. It is not fully non-parametric since the effect of the x on the hazard function is parameterized.

In expression (1.26) the hazard ratio can be replaced by $g(\beta x)$ where g is any positively valued function and x may be a matrix and not just a vector in which case β is a vector. Two more possible generalizations are: (i) the variable may be *time dependent* and (ii) the hazard ratio need not be assumed to be constant. Almost always in clinical applications the hazard ratio is assumed to be $e^{\beta x}$. This is also referred to as the log-linear model. The linear model, $g(\beta x) = \beta x$, is not appealling from a computational point of view since for this model the estimated hazard ratio may be non-positive.

1.5.3 More about the partial likelihood

It is intuitive that the partial likelihood carries information about the hazard ratio, but it may not be completely clear that it carries all of the information about the hazard ratio. Let $(t_{s_1}, \delta_{s_1}), \ldots, (t_{s_n}, \delta_{s_n})$ be the observed follow-up information ordered so that $t_{s_1} < \cdots < t_{s_n}$. Suppose t_{s_i} and t_{s_j} are successive uncensored failure times, so that $\delta_{s_i} = \delta_{s_j} = 1$ and $\delta_{s_{i+1}} = \cdots = \delta_{s_{j+1}} = 0$. The partial likelihood ignores any information carried by the censoring times in the interval (t_{s_i}, t_{s_j}). It seems intuitive, for instance, that knowledge that subject s_{i+1} survived over the interval $(t_{s_i}, t_{s_{i+1}})$ contributes some information about $\lambda(t; x_{i+1})$ and therefore about β. However, since we are making no assumptions about the baseline hazard, this knowledge can contribute little or no information about the hazard ratio. It may be that $\lambda_0(t)$ is zero everywhere except for mass points coinciding with the observed failure times. In that case, the knowledge that subject s_{i+1} survived over

the interval $(t_{s_i}, t_{s_{i+1}})$ carries no information, since subject s_{i+1} cannot fail in that interval.

Kalbfleisch and Prentice (1973) obtained the partial likelihood as the marginal likelihood of the ranks of the follow-up times when the true model is proportional hazards.

The idea that the baseline hazard may be zero everywhere except a finite set of points is troubling. In many cases this may be very unlikely. It is more likely that the baseline hazard is a smooth function. Abrahamowicz and Ciampi (1993) discuss estimation of the hazard ratio when the baseline hazard is assumed be have some minimal level of smoothness and is estimated from the data based on full maximum likelihood density estimation by regression splines (Abrahamowicz, Ciampi and Ramsay, 1992).

The full likelihood, FL, of the follow-up times can be expressed in such a way that the partial likelihood appears as a factor carrying most or all of the information about the impact of a variable upon the baseline hazard (Cox, 1984). Let $H(t)$ be all the information about the failure and censoring times 'up until' time t. The information in $H(t)$ about subject i is

$$\begin{cases} T_i = t_i, \Delta_i = \delta_i, & \text{if } t_i < t; \\ T_i > t_i, & \text{otherwise.} \end{cases} \tag{1.28}$$

The entire information from the sample is then carried in $H(\infty)$. The full likelihood, $FL = Pr[H(\infty) = h(\infty)]$, can be manipulated using the chain rule of probabilities as

$$\prod_{i=1}^{k+1} Pr[H(t_i) = h(t_i) | H(t_{i-1}) = h(t_{i-1})] \tag{1.29}$$

or

$$\prod_{i=1}^{k+1}\{Pr[H(t_i) = h(t_i)|H(t_i - dt) = h(t_i - dt)] \times \quad (1.30)$$
$$Pr[H(t_i - dt) = h(t_i - dt)|H(t_{i-1}) = h(t_{i-1})]\}$$

which can be rewritten as

$$\prod_{i=1}^{k+1} Pr[H(t_i) = h(t_i)|H(t_i - dt) = h(t_i - dt)] \times \quad (1.31)$$
$$\prod_{i=1}^{k+1} Pr[H(t_i - dt) = h(t_i - dt)|H(t_{i-1}) = h(t_{i-1})]$$

where $t_1^* < \cdots < t_k^*$ are the observed failure times and t_0^* and t_{k+1}^* denote 0 and ∞ respectively. The first product in (1.32) is the partial likeliood. The other factor carries information provided by the gaps between successive failures.

1.6 Asymptotics for maximum partial likelihood estimates

Estimates of β that maximize a partial likelihood, called MPLE's have the same important properties of maximum likelihood estimates. In particular, MPLE's are consistent and asymptotically normal subject to some mild regularity conditions. Tsiatis (1981), and Andersen and Gill (1982) have demonstrated the asymptotic theory. The latter two authors reformulate survival analysis in the language of counting processes and martingale theory.

In his original paper, Cox (1972) presented an heuristic proof of the asymptotic normality of MPLE's. For this, it is necessary to employ the

logarithm of the partial likelihood, and its first, $U(\beta)$ and second derivatives $V(\beta)$ with respect to the parameter β. Assume, for simplicity, that β is a scalar. Let U_i and V_i be the contributions to the first and second derivatives of the log-partial likelihood corresponding to the i-th summand, that is, the i-th failure. For instance,

$$U_i(\beta) = \frac{d}{d\beta}\lambda(t_i^*; x_i^*, \beta) - \frac{\sum_{j \in R_i} \frac{d}{d\beta}\lambda_\beta(t_i^*; x_j^*)}{\sum_{j \in R_i} \lambda_\beta(t_i^*; x_j^*)} \tag{1.32}$$

Using the A_i notation defined earlier (see (1.23)

$$U_i(\beta) = \frac{d}{d\beta} Pr_\beta[A_i = a_i | R_i]. \tag{1.33}$$

It is a property of all families of frequency functions (and density functions), g_θ, indexed by a continuous parameter θ for which θ_0 is the true value, that

$$E[\frac{d}{d\theta} g_\theta(X)|_{\theta=\theta_0}] = 0, \tag{1.34}$$

(Casella and Berger, 1990). It follows that

$$E[\frac{d}{d\beta} Pr_\beta(A_i = a_i | R_i)|_{\beta=\beta_0} | R_i] = 0 \tag{1.35}$$

for each i. By the double expectation theorem we can take one more expectation, that is integrate with respect to R_i to yield

$$E[\frac{d}{d\beta} Pr_\beta(A_i = a_i | R_i)]|_{\beta=\beta_0} = 0. \tag{1.36}$$

Therefore

$$E[U_i(\beta_0)] = 0 \tag{1.37}$$

and so $E[U(\beta_0)] = 0$.

Its a further property (Casella and Berger, 1990) of all families of frequency functions, g_θ, that

$$\begin{aligned} E[\frac{d^2}{d\theta^2} g_\theta(X)|_{\theta=\theta_0}] &= -E[\frac{d}{d\theta} g_\theta(X)^2|_{\theta=\theta_0}] \\ &= -Var[\frac{d}{d\theta} g_\theta(X)]. \end{aligned} \tag{1.38}$$

Therefore

$$Var[U_i(\beta_0)|R_i] = -E[V_i(\beta_0)|R_i] \tag{1.39}$$

and again by the double expectation theorem $Var[U_i(\beta_0)] = -E[V_i(\beta_0)]$. To derive $Var[U(\beta_0)]$ we first have to calculate the covariance of U_i and U_j for $i \neq j$. Unlike the corresponding terms in the proof of the asymptotic properties of maximum likelihood estimates (see, for instance, Kendall and Stuart, 1979) these terms are not independent. Expression (1.37) and the double expectation theorem are used in the following derivation of the covariance of U_i and U_j,

$$\begin{aligned} Cov[U_i(\beta_0)U_j(\beta_0)] &= E[U_i(\beta_0)U_j(\beta_0)] - E[U_i(\beta_0)|R_j]E[U_j(\beta_0)] \\ &= E[U_i(\beta_0)U_j(\beta_0)] \\ &= E[E[U_i(\beta_0)U_j(\beta_0)|R_j]] \\ &= E[U_i(\beta_0)E[U_j(\beta_0)|R_j]] \\ &= 0. \end{aligned} \tag{1.40}$$

It follows that

$$Var[U(\beta_0)] = Var[\sum_i U_i(\beta_0)]$$

$$
\begin{aligned}
&= \sum_i Var[U_i(\beta_0)] + 2\sum_{i\neq j} Cov[U_i(\beta_0), U_j(\beta_0)] \\
&= \sum_i -E[V_i(\beta_0)] \\
&= -E[V(\beta_0)]. \qquad (1.41)
\end{aligned}
$$

Cox's (1972) proof proceeds as in the maximum likelihood proof where the first derivative of the log-likelihood is expanded in a Taylor series with remainder around the true parameter, β_0, and evaluated at $\hat{\beta}$,

$$U(\hat{\beta}) = U(\beta_0) + (\hat{\beta} - \beta_0)V(\beta^*) \qquad (1.42)$$

where β^* lies between β_0 and $\hat{\beta}$. Using $U(\hat{\beta}_0) = 0$, this expression can be rewritten as

$$[-V(\beta^*)]^{\frac{1}{2}}(\hat{\beta} - \beta_0) = [-V(\beta^*)]^{-\frac{1}{2}}[U(\beta_0)]. \qquad (1.43)$$

Using the *near* independence of U_i and U_j and the central limit theorem, Cox (1972) concludes that

$$\frac{U(\beta_0)}{Var[U(\beta_0)]^{\frac{1}{2}}} \rightarrow N(0,1). \qquad (1.44)$$

The last expression also yields a score test statistic. For instance to test the hypothesis that the true value of β is 0, compare the statistic $V(0)^{-\frac{1}{2}}U(0)$ to a normal distribution.

Having demonstrated the approximate distribution of a MPLE, we now calculate an estimate and a confidence interval for a measure of the predictive ability of the continuous version of prothrobin time. We shall assume a proportional hazards effect, $\lambda(t;x) = \lambda_0(t)\exp(\beta x)$. We calculate $\hat{\beta} = 0.263$,

with confidence interval $(0.178, 0.349)$. Thus, assuming that hazards are indeed proportional this means that each increase of a second in the time to achieve coagulation is associated with a $e^{\hat{\beta}} = 1.29$ fold increase in instantaneous risk of death at any time during the follow-up. In the next chapter, we shall question this assumption of a proportional hazards effect of prothrombin time.

1.7 Time-dependent variables

The appeal of the partial likelihood as an inference device is two-fold. First, it avoids parameterization of the baseline hazard function, and second, it allows the introduction of *time-dependent covariates* (Cox, 1972). So far we considered variables that are measured at the beginning of follow-up. When we discuss the effect of prothrombin time upon the hazard at time t, $\lambda(t, x)$, we have meant the effect of prothrombin time as measured at the beginning of this subject's followup, $x(0)$. We have not meant and not the effect of of the value of prothrombin time at time t, $\lambda(t, x(t))$, (or the effect of any other function that depends on the values of $x(t)$ for $t > 0$.) However the validity of the partial likelihood is not lost on regression models in which the predictor variable is time varying.

The classic example of the time-dependent variable arises from the Stanford Heart Transplant Study (Turnbull, Brown and Hu, 1974). In this study, the efficacy of heart transplantation was being examined. Subjects became part of the study if they were considered to be candidates for heart trans-

plantation at which time they were put on a waiting list. The time from this decision of a subject's candidacy until the time of the actual transplant depends on the untimely death of another person. This duration is safely assumed to be independent of the particular morbidity of a subject. Some subjects died before ever recieving a heart transplant. Follow-up data consists of the 3-tuple $(T, \delta, \{x(t)\}_{t<T})$ where T is the time from the decision of candidacy until death, $\delta = 1$, or censoring, $\delta = 0$, and $x(t)$ is the dichotomous variable which is 1 if the subject received a heart transplant before time t and 0 otherwise. In order to assess the effect of heart transplantation on the risk of death, one can employ the proportional hazards model, $\lambda(t; \{x(u)\}_{u<t}) = \lambda_0(t)e^{\beta x(t)}$. Estimates of β for which $e^{\hat{\beta}}$ is very low would provide evidence that heart transplantation is effective.

In some instances, a variable changes over the follow-up but its path is completely known at time 0. For instance $x(t) = x(0)g(t)$, where $g(t)$ is some function known at time 0. This type of time-dependent variable is called fixed. We shall employ fixed time-dependent covariates later to model alternatives to the proportional hazards assumption. Another type of time-dependent covariate, which proves useful for uniting some seemingly disparate concepts from survival analysis, is the *evolutionary covariate.* Examples of evolutionay covariates are the number of failure up to time t, $N(t)$, or the number of censorings up until time t, $N^U(t)$, and the Kaplan-Meier estimator, $S_{KM}(t)$, whose value is bases strictly on information acquired before time t. We shall refer to evolutionary covariates in the next section.

1.8 Link between the log rank and proportionality of hazards

Arguments based on the risk set were used to derive both the partial likelihood and the family of weighted log-rank statistics in (1.17). Accordingly it may not be surprising to find that the partial likelihood and log-rank tests are intimately related. Consider the single parameter family of models, $\lambda(t;x) = \lambda_0(t)e^{\beta g(t)x}$. Here the effect of a 1 unit increase in x upon the instantanaous risk of failure at time t is $e^{\beta g(t)}$ for some function g that might be known at time 0 or might be an evolutionary covariate. The logarithm of the partial likelihood, LPL, and its first and second derivatives, U, and V, corresponding to this model are

$$\begin{aligned} LPL(\beta) &= \sum_{i=1}^{k}\{\beta g(t_i^*)x_i^* - \log\sum_{j\in R_i} e^{\beta g(t_i^*)x_j}\}, \\ U(\beta) &= \sum_{i=1}^{k} g(t_i^*)\{x_i^* - \frac{\sum_{j\in R_i} x_j e^{\beta g(t_i^*)x_j}}{\sum_{j\in R_i} e^{\beta g(t_i^*)x_j}}\}, \\ V(\beta) &= -\sum_{i=1}^{k} g(t_i^*)^2\{\frac{\sum_{j\in R_i} x_j^2 e^{\beta g(t_i^*)x_j}}{\sum_{j\in R_i} e^{\beta g(t_i^*)x_j}} - [\frac{\sum_{j\in R_i} x_j e^{\beta g(t_i^*)x_j}}{\sum_{j\in R_i} e^{\beta g(t_i^*)x_j}}]^2\}. \end{aligned}$$

The score statistic corresponding to the hypothesis that $\beta = 0$, is

$$U(\beta) = \sum g(t_i^*)\{x_i^* - \frac{\sum_{j\in R_i} x_j}{|R_i|}\}. \tag{1.45}$$

If x is a dichotomous variable this is simply the weighted log-rank statistic with weights $\{g(t_i^*)\}_{i=1,\ldots,k}$. Furthermore if g is a constant function, the score statistic becomes the original log-rank proposed by Mantel (1966). In employing Mantel's statistic to test the predictive ability of some dichotomous

variable, one is, in a sense, assuming that the predictive ability, if it exists, is constant over the follow-up.

If g is the evolutionary covariate, $|R(t)|$, the number at risk at time t, this becomes Gehan's (1965) test statistic, whereas if $g(t) = S_{KM}(t)$ it is Prentice's (1978) test statistic. If baseline survival is approximately exponentially distributed then $S_{KM}(t) \approx e^{-\theta t}$ for some $\theta > 0$ and so $g(t) \approx e^{\theta t}$. This author notes that in employing Prentice's test one is, in a sense, assuming that predictive ability, if it exists, is exponentially decaying.

Chapter 2

Assessing the Validity of the Proportional Hazards Model

2.1 Introduction

Cox's proportional hazards model is by far the most popular method of regressing survival data. This model is attractive because it facilitates the partial likelihood approach and because it has a relevant and simple interpretation; the effect of a unit increase in a covariate is associated with a uniform multiplicative change in the instantaneous relative risk (hazard). However, the assumption of proportional hazards (referred to later as ph) cannot be expected to be true, or approximately true, in all situations.

There are many instances when the assumption of proportional hazards would seem implausible. For instance, in many clinical trials subjects are followed up even after treatment has been terminated. One would expect that if treatment is effective that the relative risk relating treated to control subjects would change at or some time after the point that treated subjects

terminate treatment. One would also expect the predictive ability of some laboratory markers to decline to null eventually. For example a level of high density lipids measured 10 years ago likely has less predictive value than a current level in predicting the risk of subsequent coronary heart disease.

Suppose a marker is intimately related to the hazard as in $\lambda(t; x(t)) = \lambda_0(t)\exp(\beta x(t))$. For example, consider a cohort of persons suffering from AIDS whose risk of death is largely determined by their ability to fight infection, which in turn is associated with their current level of *helper cells*, a component of the immune system. Rarely would we know $x(t)$ at every moment in time. An analyst might have just one measurement of $x(t)$, the value at the initiation of follow-up, and proceed to use the model $\lambda(t; x(t)) = \lambda_0(t)\exp(\beta x(0))$. The extent to which this latter model is appropiate is proportional to the correlation between $x(t)$ and $x(0)$. If the correlation between $x(t)$ and $x(0)$ decays quickly, then the impact of $x(0)$ upon instantaneous risk is far from proportional. The association between $x(0)$ and the hazard $\lambda(t, x(\cdot))$ will also decay. A more appropiate model would assume that the instantaneous relative risk decays toward unity, i.e.; as in the *exponential decay model* (Gore, 1984)

$$\lambda(t, x(\cdot)) = \lambda_0(t)\exp[e^{-\theta t}x(0)] \tag{2.1}$$

for some $\theta > 0$.

In an 1982 article in Biometrics, Andersen wrote in regard to the proportional hazards assumption, "surprisingly little attention has been paid to the

problem of model checking". In the same article, Andersen proposed a test of the ph assumption. Earlier, graphical assessments of the assumption had been proposed by Kay (1977) and Kalbfleisch & Prentice (1980, Chapter 4) and Cox (1979). Kay (1977) had also proposed a method using the definition of generalized residuals of Cox and Snell (1968) to examine the overll fit of the model. In his original paper, Cox (1972) proposed a numerical test of the proportional hazards assumption and in 1980 Schoenfield proposed a numerical test. Since the time of Andersen's remark, a considerable number of methods have been proposed for assessing the validity of the ph assumption.

We begin this chapter by reviewing some graphical techniques for assessing the proportionality of hazards in section 2. In section 3, we discuss some residual methods for assessing the ph assumption and emphasize the residual approach proposed by Schoenfield (1982). Hypothesis tests, are reviewed in section 4.

2.2 Graphical assessment of the proportionality of hazards

If the impact of a variable, x, (as measured at the start of follow-up) upon the instantaneous relative risk is multiplicative and constant, that is, if hazards are proportional, $\lambda(t;x) = \lambda_0(t)g(\beta x)$, then $S(t;x) = S_0(t)^{g(\beta x)}$. This follows from equation (1.3) which expresses the survival function in terms of the

hazard function. This expression yields the following identity,

$$\log[-\log S(t;x)] = \log[-\log S_0(t)] - \log g(\beta x), \tag{2.2}$$

which motivates one of the graphical approaches for assessing the proportionality of hazards.

Suppose we are measuring the impact of some treatment upon survival in a randomized clinical trial. In this case we can think of x as being a dichotomous variable. Equation (2.2) can be written as $\log[-\log S_T(t)] = \log[-\log S_P(t)] - \log g(\beta)$, where S_T and S_P are, respectively, the survival curves for persons receiving the treatment and the placebo. To check whether hazards are proportional in this setting one could estimate each survival curve, using the Kaplan-Meier estimates for instance, and plot the log[− log] transformation of each. If treatment does truly have a constant or approximately constant effect upon the hazard function then the two plotted curves should be nearly parallel. This approach was one of the methods proposed by Kay (1977). Figure 2 depicts this approach for the dichotomized version of prothrombin time in the PBC study. The thick curve is the $\log - \log S_{KM}(t)$ curve for the subjects with high ($\geq$ 11 seconds) values of prothrombin time. This example illustrates a problem with this approach to ph assessment. The two plotted curves are not quite parallel and it is unclear what conclusion to come to.

For interpretation purposes it is useful to recall that $\log[-\log S(t)]$ is the logarithm of the cumulative (integrated) hazard. Thus figure 2 suggests that

the cumulative hazard of subjects with high prothrobin time is increasing at a greater rate than other subjects, for instance the hazard is greater, for up to 3 years after its measurement but not after this point. This method can be easily generalized to variables with 3 or more levels. An obvious violation of the the proportional hazards assumption are $\log[-\log \hat{S}(t)]$ curves that cross at one more points but are far apart elsewhere. To firm up the notion of *far* the analyst could also plot confidence intervals corresponding to these curves. Confidence intervals for a Kaplan-Meier estimate were demonstrated by Kaplan-Meier (1958). They are based on a standard error calculation usually referred to as Greenwood's formula since it is similar to an estimate proposed within another framework by Greenwood (1926).

When the predictor has 2 levels, an alternative to plotting $\log[-\log \hat{S}_0(t)]$ and $\log[-\log \hat{S}_1(t)]$ versus the time axis is to plot one versus the other. If the predictor has a constant predictive ability then the resulting curve should be nearly linear with slope equal to the true hazard ratio, (Kalbfleisch and Prentice, 1980).

As we have done with the laboratory marker, prothrombin time, continuous measures may be stratified into ordinal variables of 2 or more levels in order to apply this approach. This method does suffer from two disadvantages: (i) The choice of strata is arbitrary ; (ii) Some power is lost in detecting violations or validations to the ph assumption that may occur within strata.

For discrete predictors, a more direct graphical method would be to estimate the hazard function for each level of the predictor. Then the resulting

estimates could be superimposed on the same plot. In order to check proportionality of the hazards, it is preferable to plot the logarithm of each, in which case proportionality of hazards corresponds to nearly parallel curves. Not all estimates of the hazard are ideal for this purpose. For instance the Nelson estimator of the cumulative hazard assigns mass at each of the failure times and 0 everywhere else. Comparing two such estimators by eye is very difficult. It is preferable to assume that that the hazard is *smooth* and use a smoothing approach. Ramlau-Hansen (1983) discusses the use of kernel functions for estimating the hazard rate. Bloxom (1985) uses regression splines to estimate the hazard function while Abrahamowicz, Ciampi and Ramsay (1992) uses regression splines to estimate the density function. O'Sullivan (1988) uses smoothing splines to estimate the logarithm of the hazard.

An even more direct method would be to estimate the hazard ratio itself. A simple way of doing so is to partition the follow-up period and to estimate the relative risk separately in each interval. The method of time-dependent covariate functions can be used to facilitate the estimation. In general, suppose the follow-up is partitioned as $\{[\gamma_{i-1}, \gamma_i)\}_{i=1,...,k}$ using the sequence $0 = \gamma_0 < \gamma_1 < ... < \gamma_k = \infty$. Define the following indicator functions, $I_i(t) = I_{[\gamma_{i-1},\gamma_i)}(t)$ and the following k time-dependent covariates, $xI_i(t)$, $i = 1, ..., k$. One then maximizes the partial likelihood corresponding to the model

$$\lambda(t; x) = \lambda_0(t)\, exp(\sum_{i=0}^{k} \theta_i I_i(t) x) \tag{2.3}$$

to yield estimates $\theta_1, ..., \theta_k$ of the log relative risk in the respective intervals of the partition. For the PBC data the follow-up period was categorized into 5 intervals, $[0,1), [1,2), [2,5), [5,8)$ and $[8,\infty)$. This partition reflects the distribution of the timing of the deaths. The respective number of deaths observed in each interval was 30, 20, 65, 28 and 18. The respective relative risks for prothrobin time with 95 % confidence interval (based on asymptotic theory) are presented in figure 3. Note that that the relative risk axis is log scaled. The dotted line indicates a relative risk of 1. If the confidence interval for any particular point on the time axis does not contain the value 1 then the data suggests that prothrombin time has an impact on the hazard at this point. This figure provides evidence that the predictive ability of prothrombin time is initially high but null 5 years after its measurement. The analyst should reject the assumption of a constant hazard ratio upon seeing these results.

Estimation of the hazard ratio is discussed in more detail in section 5, and in Chapter 3, we discuss the estimation of the hazard ratio by a specific smoothing method, the method of regression splines.

2.3 Residual methods for assessing proportionality of hazards

In his 1977 paper, Kay also suggested that the method of *generalized residuals* can be used to assess the goodness-of-fit of the ph model. Generalized residuals were defined by Cox and Snell (1968). For uncensored data,

$(x_1, Y_1), \ldots, (x_n, Y_n)$, where the Y_i's represent the dependent variable, the i-th generalized residual is defined as $\hat{F}(Y_i; x_i)$ where $\hat{F}(\cdot; x)$ is an estimate of the true cumulative distribution function, $F_0(\cdot; \cdot)$, given the independent variable x. If $\hat{F}(\cdot; \cdot)$ is a good approximation to $F_0(\cdot; \cdot)$ then $\hat{F}(Y_i; x_i)$ is close to $F_0(Y_i; x_i)$. This latter random variable has the uniform distribution, $U(0, 1)$, since,

$$\begin{aligned} Pr[F_0(Y_i; x_i) < u] &= Pr[Y_i < F_0^{-1}(u; x_i)] \\ &= F_0[F_0^{-1}(u; x_i); x_i] \\ &= u. \end{aligned} \tag{2.4}$$

Therefore, the generalized residuals $\hat{F}(Y_1; x_1), \ldots, \hat{F}(Y_n; x_n)$ should resemble an n-sample from $U(0, 1)$. For instance, the empirical distribution function of these residuals should be approximately a straight line connecting $(0, 0)$ and $(1, 1)$.

For a right-censored data set, $(T_1, \Delta_1, x_1), \ldots, (T_n, \Delta_n, x_n)$ whose true survival function is $S(.; .)$ and $C(t) = Pr(\text{subject is not censored before time } t)$, $1 - S(T_i; x_i)$ is a right-censored random variate corresponding to uniformly distributed 'survival' durations and the censoring distribution $C[(1-S)^{-1}(t)]$, where $(1-S)^{-1}$ denotes the inverse function of $1-S$. To apply the method of generalized residuals to the setting in which the proportional hazards model has been applied, we first require an estimate of $S(\cdot; \cdot)$. One appealing estimate of the baseline cumulative hazard function, $\Lambda_0(t) = -\log S(t; 0)$, is

due to Breslow (1972,1974),

$$\hat{\Lambda}(t) = \sum_{t_i^* \leq t} \frac{1}{\sum_{j \in R_i} e^{\hat{\beta} x_j}}. \tag{2.5}$$

This is a natural generalization of the Nelson estimator, of the cumulative hazard (1.12). The Nelson estimator is captured by taking $\beta = 0$. (An analogous generalization of the Kaplan-Meier estimator is, $\hat{S}_0(t) = \prod_{t_i^* \leq t}[1 - \frac{1}{\sum_{j \in R_i} e^{\hat{\beta} x_j}}]$.) To estimate $S_0(t) = S(t;0)$, we take $\hat{S}_0(t) = e^{-\hat{\Lambda}(t)}$. To estimate $S(t;x_i)$ we can take $\hat{S}(t;x_i) = \hat{S}_0(t)^{\exp(\beta x_i)}$. The i-th generalized residual for this censored data set under the proportional hazards model is the pair $(1 - \hat{S}(T_i;x_i), \delta_i)$. If the proportional hazards model is valid, then these n residuals resemble a sample from a censored uniformly distributed sample. Accordingly, an estimate of the survival curve for these transformed times, such as the Kaplan-Meier estimate, should be close to a straight line connecting $(0,1)$ and $(1,0)$ (Kay, 1977).

The method of generalized residuals in this setting is laborious. Furthermore, it is difficult to determine the variance of the survival function estimated for these censored generalized residuals. This makes it difficult for the analyst to determine if this final curve invalidates the ph assumption by not lying close enough to the straight line connecting $(0,1)$ and $(1,0)$.

The generalized residuals approach was based on the durations T_i whereas the partial likelihood is motivated by the concept of the risk set. Schoenfield's (1982) *partial residuals* are based on the risk set interpretation of survival data. He defines one residual for each of the observed failures based on the

discrete distribution that the i-th risk set induces. The expected value of the variable x of the i-th subject observed to fail given the risk set just prior to the time of the i-th failure is

$$\mu_i = \frac{\sum_{j \in R_i} x_j \lambda(t_i^*; x_j)}{\sum_{j \in R_i} \lambda(t_i^*; x_j)}, \tag{2.6}$$

where t_i^* denotes the time of the i-th failure. The i-th partial residual is defined as

$$U_i = x_i^* - \mu_i. \tag{2.7}$$

If proportional hazards are used to model the impact of the variable x upon the hazard the i-th partial residual is

$$U_i = x_i^* - \frac{\sum_{j \in R_i} x_j e^{\beta x_j}}{\sum_{j \in R_i} e^{\beta x_j}}, \tag{2.8}$$

The last expression is equivalent to the i-th summand of the score statistic, (1.32), based on the partial likelihood and Cox's proportional hazards model. As we demonstrated in section 1.6 the quantities u_i have zero mean if the specified model is correct. Furthermore, they are uncorrelated since $E[U_i U_j] = 0$ (see section 1.6). To operationalize these residuals we can substitute $\beta = \hat{\beta}$. In this case, the residuals satisfy the property that they sum to zero since their sum is the derivative of the log-likelihood evaluated at $\hat{\beta}$.

Unlike the residuals we encounter in the usual uncensored continuous response regression model which are identically distributed under conventional assumptions, Schoenfield's partial residuals are not identically distributed.

The partial residuals may be standardized by dividing each by its standard deviation. An estimate of the variance of the i-th residual is

$$\sigma_i^2 = \frac{\sum_{j \in R_i} x_j^2 e^{\beta x_j}}{\sum_{j \in R_i} e^{\beta x_j}} - \left(\frac{\sum_{j \in R_i} x_j e^{\beta x_j}}{\sum_{j \in R_i} e^{\beta x_j}}\right)^2. \tag{2.9}$$

Typically σ_i will not change much as i increases unless the covariate has a very strong effect.

The partial residuals, u_i or u_i/σ_i, may be used to graphically assess suggestive departures from the ph assumption. The analyst can plot them against the timing or rank of the failure associated with the corresponding risk set. Suppose the true model is $\lambda(t;x) = \lambda_0(t)e^{[\beta+g(t)]x}$, for some function g. The function g describes the departure from the proportional hazards model and is centered about zero. Using a Taylor expansion of $E[U_i|R_i]$ about $g(t) = 0$ (i.e.; the case when ph holds,) Schoenfield (1982) shows that

$$E[U_i|R_i] \approx g(t_i^*)Var_0(X_i^*|R_i), \tag{2.10}$$

where $Var_0(X_i^*|R_i)$ is the true variance of X_i^* given the risk set just prior to t_i. The partial residuals will tend to be positive or negative, respectively, depending on whether the true log-relative risk is being underestimated or overestimated. Figure 4 depicts the residuals, u_i, for Cox model estimates of the effect of prothrombin time. There is a definite systematic departure. During early follow-up the residuals tend to be positive. This suggests that subjects with higher levels of prothrombin time are more likely to die in the earlier part of follow-up than the ph model predicts. In other words, the

proportional hazards model underestimates the relative risk function during the first part of follow-up. Conversely, the negatively tending residuals suggest that the relative risk is overestimated later in follow-up. The analyst at this point should reject or at least be cautious about the assumption of a constant hazard ratio.

Partial residuals are most appropiate when the predictor is a continuous variable or at least an ordinal variable with several levels. When proportional hazards do hold and the predictor is binary (0,1) and say level 1 is associated with more risk the quantities μ_t decrease with time since level 1 subjects are being filtered out of the sample at risk. Consequently, as this author has experienced, the eye picks up an increasing trend despite the fact that the average value remains about 0. This same difficulty occurs when the risk of censoring is more likely in one level than another. A useful alternative in this situation is to *smooth* these residuals (Petit and Bin Daud, 1990).

Recent research into diagnostics for survival models using the approach of martingales have yielded the same Schoenfield (1982) residuals. If the specified model is correct, the Schoenfield residuals are simply increments of a martingale process (Fleming and Harrington, 1990). Barlow and Prentice (1988) propose a number of martingale-based residuals for relative risk regression and obtain a generalization of Schoenfield's partial residuals. Therneau, Grambsch and Fleming (1990) propose a score process meant to evaluate the proportionality assumption and note that the increments of the process are the residuals introduced by Schoenfield. Finally Hendersen and Milner

(1991) introduce a general form of residuals for partial likelihood regression a special case of which is the Schoenfield residual.

2.4 Hypothesis testing of the proportionality assumption

There is a wealth of tests of the hypothesis of proportional hazards. The first test was proposed by Cox (1972) in the paper introducing the proportional hazards model. Cox proposed the alternative model that incorporates a time-dependent covariate, $\lambda(t;x) = \lambda_0(t)exp\{[\beta_0 + \beta_1 t]x\}$. He suggests that the proportionality assumption be evaluated by testing the hypothesis $\beta_1 = 0$ using the score statistic for the partial likelihood evaluated at $\beta_0 = \hat{\beta}_0$ and $\beta_1 = 0$. This test is most powerful against the alternative that the hazard ratio is a linear function of time.

Schoenfield (1980) proposed an omnibus goodness-of-fit test of ph. He suggests a partition of the time axis as well as of the range of the covariate and then shows how the expected values and covariance of the number of events in each resulting cell can be calculated. He proceeds to yield a goodness-of-fit statistic which he demonstrates is asymptotically distributed as a chi-square. The partition of the time and covariate axes is arbitrary. Andersen (1982) also proposed a goodness-of-fit test based on an arbitrary partition of the product of the time axis and the covariate range. His method turns out to be computationally simpler than Schoenfield's. Moreau, O'Quigley and Mesbah (1985) propose a test statistic which in the two sample setting is

equivalent to Schoenfield's (1980). O'Quigley and Pessione (1989) propose a general framework in which to test specific alternatives to ph such as linear, quadratic or exponential trends in the hazard ratio. They use rank-invariant scores in order to retain the semi-parametric nature of Cox's model.

A number of tests are implicitly based on the Schoenfield's partial residuals, (2.7). Nagelkerke, Oosting and Hart (1984) propose a test of the ph assumption which assumes that the alternative is a smoothly changing hazard ratio. They use the summands of the score statistic, U_i, defined in (1.32). These are just the partial residuals of Schoenfield (1982), although they make no mention of this. They argue that if the hazard ratio is changing smoothly, the U_i are not uncorrelated, in contrast to the case when the hazard ratio does not change. Indeed, successive values should be positively correlated. They propose the test statistic $\sum_i u_i u_{i-1}$. They use a permutational approach to estimate the mean and find an upper limit for the variance of this statistic. Wei (1984) considered violations to proportionality of hazards in the two-sample setting. Using a stochastic process approach, Wei derived the test statistic $\max_i |\sum_{j \leq i} U_j|$ where the U_i are the summands of the score statistic. Wei does not identify these increments as partial residuals. Harrell and Lee (1986) also use partial residuals. They propose that the partial residuals be correlated with the ranks of the failure times corresponding to these residuals. The null hypothesis is tested using the Fisher z-transform of the correlation coefficient. This test retains the semi-parametric nature of Cox's model and is powerful against alternatives involving monotonic changes of

the relative risk.

Gill and Schumacher (1987) propose a test of ph which is specific to the two-sample setting. They consider ratios of the form

$$\int K(t)d\hat{\Lambda}_2(t)/\int K(t)d\hat{\Lambda}_1(t) \tag{2.11}$$

where Λ_1 and Λ_2 are estimators of the cumulative hazard function in samples 1 and 2 respectively, and $K(t)$ is some arbitrary weighting function. If hazards are proportional then this ratio estimates the constant hazard ratio for any choice of K. If the relative risk is not constant then the expected value of this estimate depends on the choice of K. The authors propose test statistics of the form

$$\int K_1(t)d\hat{\Lambda}_2(t)/\int K_1(t)d\hat{\Lambda}_1(t) - \int K_2(t)d\hat{\Lambda}_2(t)/\int K_2(t)d\hat{\Lambda}_1(t). \tag{2.12}$$

. They discuss the ideal choice of the weight functions K_1 and K_2.

More recently Quantin (1992, personal communication) generalizes Cox's model to $\Lambda(t;x) = [\Lambda_0(t)]^{e^{\gamma x}}e^{\beta x}$. Cox's model is captured by taking $\gamma = 0$. She tests the hypothesis that $\gamma = 0$ using the score statistic evaluated at $(\beta,\gamma) = (\hat{\beta},0)$. The hazard ratio for this model in the two-sample setting is $e^{\beta+\gamma}\Lambda_0(t)^{e^{\gamma}-1}$ which is a monotonous function, since $\Lambda(t)$ is strictly increasing. Thus this test is not powerful against alternatives in which the hazard ratio is U-shape in nature. LLiang, Self and Liu (1990) generalize Cox's model to $\lambda(t;x) = \lambda_0(t)e^{(\beta+\theta I_{(t\leq\gamma)})}$ where γ is a change-point. They propose a test statistic for the hypothesis of no change-point which takes

the supremum of a score statistic over a range of all possible change-point values. Gu (1992) uses a similar change-point model but yields a likelihood ratio test statistic.

There is no lack of test statistics for examining the validity of the proportional hazards assumption. Different tests of ph are powerful against different alternatives although it is usually not clear what is the best test to use. Furthermore even if the assumption of a constant hazard ratio is rejected the analysts is left with the problem of what is the correct model and how to estimate the predictor effects over different portions of the follow-up. There has been but a few papers discussing the modelling of a time-dependent hazard function. In the next chapter, we describe the previous work on modelling of the time varying relative risk and introduce the idea of smooth estimates of the hazard ratio expressed as a function of time.

Chapter 3

Modelling of the Relative Risk by Regression Splines

3.1 Overview

Only a few authors have discussed alternatives to the proportional hazards (ph) model that take advantage of the partial likelihood framework. In section 2, we review alternatives to the ph assumption that retain the partial likelihood framework and introduce the notion of *smoothing* the hazard ratio. In section 3 we discuss the method of smoothing splines for estimating the time-dependent hazard ratio, as proposed by Zucker and Karr (1990). In section 4 we introduce another smoothing technique, regression splines, and describe in general some of their properties. In section 5 we propose a regression spline model for estimating the hazard ratio. We also discuss the large sample properties of the regression spline estimator of the hazard ratio. In section 6 we shall discuss the computation involved in maximizing the partial likelihood when regression splines are employed to model the time-dependent

relative hazard. We introduce the truncated power, B-spline and M-spline bases which span a given family of regression splines. In section 7 we discuss the optimal choice of the level of flexibility of a regression spline. There we propose a method for model identification that uses the Akaike Information Criterion. We refer to this as the best-AIC regression spline approach. In section 8 we discuss the impact of *a posteriori* model selection upon inference. In section 9 we suggest a modification to the best-AIC regression spline approach in order to make it more conservative with respect to rejecting the proportional hazards model.

3.2 Modelling the Time-Dependence of the Hazard Ratio

One way to incorporate the possibility that the relative risk may not necessarily be constant is to extend Cox's proportional hazards model to

$$\lambda(t;x) = \lambda_0(t)e^{\beta(t)x}, \tag{3.1}$$

in which we replace the parameter β by an estimable function of time $\beta(t)$. While the hazards are no longer proportional we have retained a factorization which, like the ph model, facilitates the partial likelihood method since the factor $\lambda_0(t)$ cancels out of the partial likelihood. $\beta(t)$ can be interpreted as the log relative risk at time t.

One of the most simple ways to estimate $\beta(t)$ is to express it as a polynomial function of time, $\lambda(t;x) = \lambda_0(t)\exp[(\beta_0+\beta_1 t+\cdots+\beta_r t^r)x(0)]$. Another

simple way is to partition the time axis and estimate the constant relative risk separately in each interval as in the model (2.3).

Gore, Pocock and Kerr (1984) propose a log relative risk that declines exponentially over time, $\beta(t) = Ae^{bt}$, for some known $b < 0$. It seems intuitive that the predictive ability of a characteristic could often behave this way. This seems to be the case, as we described in chapter 2, section 1, when $x(0)$ is used as a predictor when in fact $\lambda(t; x(t)) = \lambda_0(t)\exp(\beta x(t))$ and the correlation between $x(t)$ and $x(0)$ decays to zero. Gore *et al* were analysing data from a series of breast cancer patients. They were led to the exponential decay model after consideration of the estimates from the stepwise constant relative risk model of (2.3).

Lliang, Self and Liu (1990), as we have previously mentioned, apply a change point model. The change-point model is appropiate if there is reason to believe that the relative risk will change abruptly and at the **same** point on the time axis for every subject. While it is conceivable that for a specific subject the relative risk could undergo a dramatic shift at some point on the time axis it is difficult to imagine any biological mechanism where the relative risk associated with some marker or some treatment would undergo this dramatic shift at the same time for each subject. It makes more sense that the hazard ratio changes smoothly over time.

Zucker and Karr (1990) suggest the use of smoothing splines (Wegman and Wright, 1983) to model the time-dependent relative risk.

3.3 Smoothing splines

One criterion for a function $h(t)$ to be smooth is that it be continuous as well as having a continuous first derivative. Continuity ensures that there is no breakpoints while the continuity of the first derivative ensures that there is no 'corners' in the curve. This criterion is not enough. The functions $\prod_{i=1}^{n}(x - i/n)$ and $\sin(2\pi n)$ defined on $[0, 1]$ are both infinitely differentiable but for the purposes of modelling data cannot be considered smooth since they change value so rapidly. The second derivative is one measure of the local smoothness of a function. As such a popular quantity for measuring the overall smoothness of a function h is $\int_A |h''|^2$, where A is the domain of h. In this case maximal smoothness, attained when $\int_A |h''|^2 = 0$, occurs when h is a linear function. The functional $\int_A |d^2h|^2$ (or more generally $\int_A |Lh|^2$ where L is a differential operator) is central to the topic of smoothing splines (Eubank, 1988).

A smoothing spline is defined to be a solution to the maximization problem,

$$max_{h \in H^r[a,b]} \Phi(h) + \alpha \int_A |Lh|^2 \tag{3.2}$$

where H is a space of r-differentiable functions defined on the interval $[a, b]$, Φ is some functional defined on H, α is a constant and L is an $r - th$ order differential linear operator. Usually $r = 2$ and $L = D^2$. The function h that maximizes expression (3.2) strikes a compromise between the desire to mazimize the log-likelihood and the desire that h be smooth. The constant

α controls the relative degrees of these two desires. As α increases so does the smoothness of the solution to (3.2). Typically, the *smoothing parameter*, α, is chosen using generalized cross-validation (Craven and Wahba, 1979).

Zucker and Karr (1990) use smoothing splines to model the time-dependent relative risk. In this setting $h(t) = \beta(t)$, the hazard ratio expressed as a function of time, $r = 2$, $a = 0$, b is some reasonable upper bound for the follow-up period, $L = D^2$ and most importantly Φ is the log partial likelihood of the observed data assuming model (3.1),

$$\max_{\beta \in H^2[0,b]} LPL(\beta) + \alpha \int_0^b |\beta''(t)|^2 \tag{3.3}$$

The space of functions $H^2[a,b]$ is *huge* but finding this solution is not as daunting as it may initially seem. Zucker and Karr prove that any solution to (3.3) must lie in a finite order linear subspace of $H^2[a,b]$. Let $t_1^*, \ldots, t_k^*$ be the observed failure times. This linear space is the set of all $\beta \in H^2[0,b]$ such that β is a cubic polynomial on any one of the intervals $[0, t_1^*], [t_1^*, t_2^*], \ldots, [t_{k-1}^*, t_k^*], [t_k^*, \infty]$. The fact that β is a member of $H^2[a,b]$ ensures that at the failure times β is continuous as well as having continuous slope and curvature. Zucker and Karr demonstrate that if indeed the model in expression (3.1) is true and that β does indeed have continuous second derivative then the maximizer, $\hat{\beta}(t)$, of expression (3.3) is a consistent estimator and pointwise is asymptotically normal when the sequence of penalty paramters, $\{\alpha_n\}_{n=1,\ldots,\infty}$, is appropiately chosen. Here, the sequence $\{\alpha_n\}_{n=1,\ldots,\infty}$ is indexed by the number of observed failures, n, to indicate the

dependence of the level of smoothing on the available amount of data.

3.4 Regression splines

Regression splines generalize regression by polynomials but are usually considered to be an improvent over polynomials in that fitting is more *localized.* An estimate is called local to the extent that it places relatively more weight on using information from observations close to t than far from t. Localizing an estimate reduces the bias of estimates while increasing the variance of the estimator (Hastie and Tibshirani, 1990).

A family of regression splines are defined by their degree, which is equivalent to the degree of a polynomial, and their *knots.* The knots consist of the exterior knots, 2 points which define the domain of the regression, and interior knots, a set of distinct points that partition the domain. Usually we speak of the *order* of the regression spline which is 1 plus the degree. Between any two adjacent knots the regression spline is a polynomial of the same order. These polynomials join in a smooth fashion. A regression spline of order k has continuous $l-th$ derivative for $l = 0, ..., k-2$. A cubic spline is continuous as well as having continuous slope and curvature at its interior knots. A quadratic spline is continuous as well as having continuous slope at its interior knots.

The family of regression splines of a given order specific to any particular sequence of knots is a linear space. As such there exists a finite basis. Estimation by regression splines is simply accomplished by regressing onto these

basis elements by a method such as maximum likelihood estimation or MPL estimation.

Regression splines have been recently used in survival analysis to (i) estimate the density and hazard functions (Abrahamowicz, Ciampi and Ramsay, 1992) and (ii) model the functional form of the impact of covariates. Durrleman and Simon, 1989 and Sleeper and Harrington, 1990 propose the following generalization of the proportional hazards model,

$$\lambda(t;x) = \lambda_0(t)\exp[g(x)], \tag{3.4}$$

and proceed to estimate g using regression splines.

Regression splines bear a relation to smoothing spline methods. For instance, as Zucker and Karr (1990) demonstrate for the relative risk function, there exists a smoothing spline that maximizes the penalized partial likelihood which is a piecewise cubic polynomial. In other words, there exists a solution which is a cubic regression spline. This particular regression spline has one knot for each observed failure time which is an obvious over-parameterization for a MPLE approach. Typically in regression spline methods no more than 10 knots are ever used.

3.5 Regression spline model for hazard ratio

The set of regression splines, of a given order (1+ degree) and given knot placement, is a finite linear subspace. As such there exists a finite basis to span this set of regression splines. The regression spline model we propose

for the relative hazard can be expressed in the form

$$\lambda(t;x) = \lambda_0(t)exp(\sum_{i=1}^{l} \beta_i g_i(t)x) \tag{3.5}$$

where $\{g_i\}_{i=1,...,l}$ is an l-dimensional basis for a regression spline space. This model is a special case of the generalization of Cox's model to time dependent covariates. The terms $g_i(t)x$ are *fixed* time dependent covariates (see section 1.7). To determine the particular regression spline amongst a space of regression splines of a given order and of given knot positions that best estimates the relative hazard an obvious approach is to maximize the partial likelihood corresponding to this model.

3.5.1 Large sample properties

The proofs of the consistency and asymptotic normality for the partial likelihood, that we discussed in section 1.6, ensure the following: **If** the true relative hazard is an element of the regression spline space over which we are maximizing then the relative hazard at any particular point on the time axis will be estimated without bias and with normal error asymptotically. To match the 'true' curve 'exactly' it is likely that no finite regression space will do. For instance, the true curve corresponds to a regression spline with an infinite number of knots. Thus to establish asymptotic consistency in general we must assume an estimation technique that allows the number of knots to go to infinity. Moreover for any interval (a,b) over which $dPr[T < x] > 0$ for all $x \in (a,b)$ the number of knots falling in the interval must approach

infinity. One such sequence of knots is obtained by spreading them out uniformly amongst the observed failure times. For instance if there are n observed failures, at times $t_{(1)} < \ldots < t_{(n)}$, the knots are chosen to be $\kappa_i = t_{(i \times r_n)}$, $i = 1, \ldots, \lfloor n/r_n \rfloor$, where $\{r_i\}_{i=1,\ldots,\infty}$ is any sequence satisfying $\lim_{n\to\infty} r_n \to \infty$ and $\lim_{n\to\infty} r_n/n \to 0$.

3.6 MPL estimates for the regression spline model

Once a knot sequence has been chosen employing regression splines to smooth the relative hazard can be done at relative ease. For instance one can use the statistical software BMDP module 2L to obtain smoothed estimates with little more code than that necessary to obtain estimates from a proportional hazards model. The Newton-Raphson algorithm can be employed to find the estimates maximizing the partial likelihood of the model in (3.5). The vector of first derivatives is given by,

$$\frac{\partial}{\partial \beta_k} \log PL = \sum_{i=1}^{k} g_k(t_i) x_i^* - \frac{\sum_{j \in R_i} x_j \exp[\beta(t_i^*) x_j]}{\sum_{j \in R_i} \exp[\beta(t_i^*) x_j]}. \tag{3.6}$$

where $\beta(t) = \sum_{i=1}^{l} \beta_i g_i(t)$.

The matrix of second derivatives is given by

$$\frac{\partial^2}{\partial \beta_k \partial \beta_l} \log PL = -\sum g_k(t_i^*) g_l(t_i^*) \{ \frac{S_{2i}(\beta)}{S_{0i}(\beta)} - (\frac{S_{1i}(\beta)}{S_{0i}(\beta)})^2 \} \tag{3.7}$$

where

$$S_{ri}(\beta) = \sum_{j \in R_i} x_j^r \exp[\beta(t_i^*) x_j] \tag{3.8}$$

for $r = 0, 1, 2$, and i indexing the observed failures.

The order of one iteration of this Newton-Raphson is $o(nmq)$ where n is the size of the sample, m is the number of observed failures and q reflects the time necessary to invert the matrix of second derivatives. On the other hand the order of the proportional hazards model with one covariate is n. The factor increase in complexity, m, is due to the fact that the sums, S_{ri}, for $r = 0, 1, 2$, must be recalculated for each failure time. The corresponding sums for the proportional hazards model are

$$S_{ri} = \sum_{j \in R_i} x_j^r \exp(\beta x_j). \tag{3.9}$$

Since β does not vary from one failure time to the next, as it does in the time-dependent model, $S_{r,i+1}$ may be obtained from $S_{r,i}$ sums by deleting the summands corresponding to subjects who are eliminated from the risk set between these failure times.

Estimates of the asymptotic covariance matrix of the the estimate $\hat{\beta}$, denoted by $\hat{\Sigma}(\hat{\beta})$, are yielded by the inverse of the observed information matrix (matrix of second derivatives), evaluated at $\hat{\beta}$. The pointwise estimate of the variance of the log relative risk at time t is then given by

$$\mathbf{g}(t)^T \hat{\Sigma}(\hat{\beta})^{-1} \mathbf{g}(t) \tag{3.10}$$

where $\mathbf{g}(t)$ is the vector of basis functions evaluated at time t.

3.6.1 Regression Splines Bases

The functions $\{g_i\}_{i=1,\cdots,d}$ in (3.5) are a basis for the d-dimensional regression spline basis. A very intuitive and easy to program basis for a space of regression splines is the *truncated power basis*. The set of functions

$$1, t, \ldots, t^r, (t-\kappa_1)_+^r, \ldots, (t-\kappa_m)_+^r \tag{3.11}$$

spans the $r+m$-dimensional regression spline space of degree $r-1$ (order r) with knots placed at $\kappa_1 < \cdots < \kappa_m$. The meaning of the notation $(x)_+^r$ is the following

$$(x)_+^r = \begin{cases} x^r, & \text{if } x \geq 0; \\ 0, & \text{if } x < 0. \end{cases} \tag{3.12}$$

The continuity properties of the regression spline space are then self-evident since

$$\frac{d^i}{dx^i} x_+^r = \begin{cases} r(r-1)\cdots(r-i+1)x^{r-i}, & \text{if } x \geq 0; \\ 0, & \text{if } x < 0. \end{cases} \tag{3.13}$$

which is continuous at 0 for $i < r$. Therefore every element of this basis has continuous i-derivatives for $i < r$.

The truncated power basis is easy to define. For instance one could use it rather easily to obtain regression spline estimates using a statistical package such as BMDP. However, this basis is highly collinear especially due to the fact that every element of this basis is a non-decreasing function (de Boor, 1978). The high level of collinearity decreases the precision with

which the matrix of second derivatives can be inverted. As a result the number of steps necessary for convergence may increase or convergence may not be attained. On the other hand if the basis was orthogonal or 'close' to orthogonal the inversion of the matrix of second derivatives may be changed from a $o(d^2)$ calculation to a $o(d)$ calculation. This becomes important in reducing computing time if the relative hazard function cannot be modelled adequately by a low dimensional regression spline.

An orthogonal basis does exist for any particular regression spline basis but typically the order of any algorithm, such as the Gram-Schmidt process, employed to determine an orthogonal basis is $o(d^2)$ which would negate the one of the motivations for finding an orthogonal basis.

The truncated power basis while highly collinear does have a suggestive property. The property that $(t-\kappa_i)_+^r$ takes the value 0 over the interval $(-\infty, \kappa_i)$ can be taken advantage of to derive a nearly orthogonal basis. A basis, $\{g_i(t)\}_{i=1,\ldots,d}$, for which all elements is zero everywhere except over finite intervals might yield near orthogonality. Suppose g_i is zero on the interval I_i for $i = 1, \ldots, d$ then the second partial derivative, $\frac{\partial^2}{\partial\beta_k \partial\beta_l} \log PL$, as in (3.7) is zero for all k and l such that I_k and I_l do not intersect.

Consider the truncated powers corresponding to the first four knots of a quadratic regression spline space, $(t-\kappa_i)_+^2, i = 1, 2, 3, 4$. Any linear combination of the latter three is zero on $(-\infty, \kappa_1)$. Moreover on the interval

$(\kappa_4, -\infty)$

$$\begin{aligned}\sum_{i=1}^{4} a_i(t-\kappa_i)_+^2 &= \sum_{i=1}^{4} a_i(t-\kappa_i)^2 \\ &= \sum_{i=1}^{4} a_i(t^2 - 2\kappa_i t + \kappa_i^2) \\ &= \sum_{r=0,1,2} (-1)^r \binom{2}{r} C_r t_r \qquad (3.14)\end{aligned}$$

where $C_r = \sum_{i=1}^{4} a_i \kappa_i^{2-r}$. If we could find non-zero $\mathbf{a} = (a_1, a_2, a_3, a_4)^T$ for which each of the three C_rs equalled zero then the function $\sum a_i(t-\kappa_i)_+^2$ would be a regression spline that is non-zero on the interval (κ_1, κ_4), and zero everywhere else. Such a 4-tuple a does exist as one of the many solutions to the matrix-equation

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ \kappa_1 & \kappa_2 & \kappa_3 & \kappa_4 \\ \kappa_1^2 & \kappa_2^2 & \kappa_3^2 & \kappa_4^2 \end{pmatrix} a = 0. \qquad (3.15)$$

For any of the solutions, $\hat{\mathbf{a}}$, to (3.15), the resulting regression spline $\sum \hat{a}_i(t-\kappa_i)_+^2$ has the same approximate shape. This shape is a smooth bump which is either all negative or all positive. It is possible to derive a basis which consists entirely of these type of functions. Curry and Schoenberg (1966) derive such a basis, the *M-spline basis* and de Boor (1978) derives another, the *B-spline basis.* Each generalize the idea of these bases to regression spline bases of arbitrary order. For instance cubic-order B-spline and M-spline bases are defined. Whereas a quadratic-order B-spline or M-spline component is zero everywhere except over an interval (κ_i, κ_{i+3}) an r-th order B-spline or M-spline is zero everywere except an interval (κ_i, κ_{i+r}).

For instance a cubic-order B-spline is positive over 4 consecutive intervals $(\kappa_i, \kappa_{i+1}), (\kappa_{i+1}, \kappa_{i+2}), (\kappa_{i+2}, \kappa_{i+3}), (\kappa_{i+3}, \kappa_{i+4})$. The B-spline and M-spline are each characterized by a different normalization. Each element of a M-spline basis integrates to 1, whereas for B-spline basis elements, $B_1(t), \ldots, B_d(t)$ the sum $\sum_{i=1}^{d} B_i(t)$ is unity for all t. The unit integral constraint for any particular element of the M-splines basis can be incorporated as a final linear equation, $\int_{-\infty}^{+\infty} \sum a_i (t - \kappa_i)_+^2 \, dt = 0$, in expression (3.15).

The B-spline and M-spline bases are nearly orthogonal. For any inner product, specifically the inner product defined by the second derivates of the time-dependent hazard ratio model, in (3.7), the inner product of any two basis elements, b_i and b_j, is 0 when $|i - j| \geq r$. The matrix of second derivatives (3.7) has a banded structure if we employ B-splines or M-splines. This facilitates the inversion of this matrix which is a crucial part of the Newton-Raphson algorithm.

Each element of an M-spline basis can be derived in $o(r^2)$ calculations. Typically M-splines and B-splines are obtained via a set of recursive formulas, (de Boor, 1978, Curry and Schoenberg, 1966, Ramsay, 1988). For instance the M-spline basis can be constructed using the recursive formulas,

$$M_j(t|\kappa, 1) = \frac{I_{[\kappa_j, \kappa_{j+1})}(t)}{\kappa_{j+1} - \kappa_j} \qquad j = 1, \ldots, 2r + m - 1 \tag{3.16}$$

for $j = 1, \ldots, 2r + m - 1$ and

$$M_j(t|\kappa, r) = [\frac{r}{(r-1)(\kappa_{j+r} - \kappa_j)}][(t - \kappa_j) M_j(t|\kappa, r-1) + \tag{3.17}$$

$$(\kappa_{j+r} - t)M_{j+1}(t|\kappa, r-1)] \qquad (3.18)$$

for $k > 1$ and $j = 1, \ldots, 2r + m - 1$, where $\kappa = (\kappa_1, \ldots, \kappa_m)$, the vector of knots, which also incorporates the endpoints L and U of the domain of the regression, and $I_A(t)$ is the indicator function for the set A. The notation $M(t|\kappa, r)$ is used to denote the dependence of the regression spline space on the knots and the order r. A similar recursive formula exists for deriving a B-spline basis.

This recursive formula is $o(r^2)$. It facilitates the calculation of $M(t|\kappa, r)$ at specific points t. A less direct approach would be to calculate each M as a linear combination of 4 elements of the truncated power basis using the 4-tuple **a** which solves the r by r matrix equation defined by (3.15) and the unit integrality constraint together. On the other hand if one is graphing a linear combination of the M-splines over some interval, as for instance the regression spline which maximizes the partial likelihood for the model (3.5), the recursive formula method should be avoided. In this case the recursion has to be performed for every value of t in some fine grid within the interval, for instance for the value of t corresponding to each column of pixels of a graphic device. In this case computing is facilitated by expressing the M-splines in terms of the truncated power basis or polynomials.

M-splines are appealing since they facilitate the imposition of tail constraints such as $\beta(0) = 0$ or $\beta'(0) = 0$. While we have not used any such constraints we have chosen to use the M-spline basis in our calcuations.

3.6.2 Example

The analyses that follow were accomplished using a program written in the language C for a PC-486 by the author. The program was written to yield partial likelihood estimates for the model in (3.5) where the set of functions $\{g_i(t)\}$ are replaced by an M-spline basis. In the example that follow knot selection is automatic. The knots were located along the time axis so that there was the same (or the same less 1) number of failures in between any two adjacent knots.

In section 1.3 we reported that the laboratory marker, prothrombin time, was a statistically significant predictor of deaths based upon the log-rank test (1.16). In sections 2.3 a plot of Schoenfield's partial residuals (2.7) suggested that the predictive ability of this variable is not constant. Now we shall use the regression spline approach to estimate the possible time-dependent nature of the relative risk corresponding to this variable. Given the large number of observed failures, 161, we find it reasonable to use a cubic regression spline with 3 knots. This 7-dimensional regression spline ensures sufficient flexibility to capture the shape of the *true* hazard ratio.

Figure 5 depicts the regression spline estimate (thick line) and corresponding 95% confidence intervals (dashed lines). The horizontal line of unit relative risk is also included. The time axis is measured in years. The relative risk axis is logarithmically scaled. A reasonable, although perhaps liberal, approach to inference is to conclude that $\beta(t) > 0$, that is that prothrombin

time has predictive ability, whenever the lower bound of the confidence interval exceeds 0. In that case we would conclude that prothrombin time has the ability to predict deaths up until about 3 years after its measurement. One lesson from this result is that amongst subjects suffering from PBC a measurement of prothrombin time should be repeated at least every 3 years if its to have any prognostic value.

The estimate in figure 5 is a decreasing function with no local maxima or minima. This shape could be captured by a lower dimensional regression spline, for instance, a quadratic polynomial. In the next section we discuss the choice of the dimension, or in other words, the level of smoothness, of a regression spline.

3.7 Best AIC-Regression Splines

As we have described the smoothing of the relative hazard by regression splines can be done at relative ease. However, as in any modelling application the task is not as automatic as we may initially hope. Before an analyst proceeds with estimation he/she must decide on the number of and position of the knots that are required as well as the order of the regression spline. The choice of these quantities corresponds to the choice of the penalty parameter in the smoothing spline approach.

Regression splines are often criticized for their dependence on the location of knots when they are *a posteriori* (Hastie and Tibshirani, 1990). We shall employ the following *a priori* automatic approach. We shall locate the knots

along the time axis so that the number of failures in between any two adjacent knots is approximately the same. For instance if 3 knots are used they are located at the quartiles of the empirical distribution of survival times.

We shall employ an automated model selection criterion to determine the regression spline order and the **number** of knots. Among these criteria Akaike's Information Criterion, *AIC*, (Akaike, 1974) seems to the most popular (Bloxom, 1985, Sleeper and Harrington, 1990, Abrahamowicz *et al*, 1992).

3.7.1 AIC in the full likelihood setting

In the full likelihood setting the AIC is easily computed. It is simply the log-likelihood minus the degrees of freedom in the model. (Usually this is multiplied by minus 2, but here we shall disregard this constant factor). The AIC rewards models that fit the observed data well but penalizes the degrees of freedom in the model. There is two ways to regard the penalty. The first is that the AIC penalizes complicated models. This is ideal as regression models that an analyst communicates to a clinician, for instance, should be as simple as possible. The second way is that the AIC penalizes spurious fitting which is usually referred to as overfitting.

Akaike's (1974) motivation in deriving the AIC was the maximization of the **expected** log likelihood over a class of models. The expected log likelihood

$$E[\log g(Y;x)] = \int \log g(y;x) f(y;x)\, dy \tag{3.19}$$

is a measure of the closeness of a model $g(\cdot; x)$ to the true model $f(\cdot; x)$. This quantity is maximized over the set of all models when $g = f$. This result follows upon substitution of $v = g(Y; x)/f(Y; x)$ into the inequality

$$-1 + v - \log v \geq 0 \qquad \forall v \in \Re \tag{3.20}$$

and by then taking expected values. The expected log-likelihood plays a dominant role in information theory (Kullback and Leibler, 1951).

The observed log-likelihood of the model $g(\cdot; x)$ is

$$\sum_{1}^{n} \log g(y_i; x_i). \tag{3.21}$$

It is an estimator of $nE[\log g(Y_i; x_i)]$. In the maximum likelihood setting g is parameterized as $g(\cdot; x_i, \theta)$ and we find $\hat{\theta}$ that maximizes the log-likelihood. Let θ^* be the θ that maximizes the expected log-likelihood in this particular parameter space. The maximized log likelihood is a biased estimator of the maximized expected log likelihood. Akaike demonstrates that the difference,

$$\sum \log g(y_i; x_i, \hat{\theta}) - nE[\log g(y; x, \theta^*)] \tag{3.22}$$

is asymptotically distributed as a chi-square with degrees of freedom equalling the dimension of the parameter space. Accordingly, the expected value of the difference is then this dimension. The AIC incorporates this correction. The AIC can be conceptualized as follows. The first component, the observed log-likelihood estimates, with bias, the distance between the true distribution and the best model in the particular parameter space. The second component corrects for this bias.

3.7.2 AIC in the partial likelihood setting

So far we have discussed the AIC with respect to a full likelihood inference device. We must now extend the AIC to the partial likelihood setting. This is accomplished by replacing the log-likelihood by the log-partial likelihood. The component of the AIC that corrects bias remains the same. In this case the AIC estimates the expected value of the partial likelihood. Several authors have used the AIC for model selection based upon the partial likelihood, (Sleeper and Harrington, 1990 and Durrleman and Simon, 1989).

In general we propose that the hazard ratio be estimated in the following fashion; (i) find the MPL estimate of the hazard ratio for the constant and linear relative risk models as well as in all quadratic and cubic regression splines spaces whose dimension does not exceed some reasonable fraction of the observed number of failures. For instance there should not be more than 1 dimension per 10 observed failures (Kalbfleisch and Prentice, 1980). (This fraction should also decrease as the number of events increases if asymptotic properties of the MPLE's are to hold.) (ii) For each estimate calculate the AIC. (iii) Choose the model for which the AIC is optimized.

We shall refer to this method of estimation as the best-AIC regression spline approach.

3.7.3 Example

We used the best-AIC regression spline approach to estimate the hazard ratio corresponding to the variable prothrombin time in the PBC study. The AIC

criterion was maximized using a quadratic regression spline with 0 knots, in other words, a quadratic polynomial. Figure 6 illustrates this estimate. We have included the nominal 95% confidence intervals constructed as in the case where a quadratic polynomial is *a priori* chosen. This estimate much resembles the estimate in figure 5 with the exception that it is smoother.

3.8 Impact of model selection upon inference

After having selected a model based on a criterion such as the AIC an analyst usually constructs confidence intervals. For instance we would chose a regression spline model and proceed to draw the curves corresponding to the 95% confidence intervals for the relative hazard based on the consistency and asymptotic normality of the maximum likelihood estimates. These confidence intervals are constructed under the premise that the parameter space of the true model was known *a priori*, that is, known before the data was sampled. The process of model selection alters the actual coverage of these confidence intervals, (Hurvich and Tsai, 1990). This idea is demonstrated by the fundamental variance partitioning identity,

$$Var(Y) = E[Var(Y|X)] + Var[E(Y|X)] \tag{3.23}$$

where X and Y are any random variables. Suppose we wish to construct a confidence interval for a random variable Y but first choose a model from a family of models that compete to model Y. The process of model selection is

certainly subject to randomness so let X be the particular model we choose. If we had chosen to use a particular model x prior to model selection we would use the variance $Var(Y|X = x)$ to construct a confidence interval for Y. However, this variance underestimates, *on average*, the actual variance by the amount $Var\{E(Y|X)\}$. This latter term will be small if the expected value of Y does not differ much between models or if a particular model is chosen at a rate close to 1 or if a subset of models all of which differ little with regard to $E[Y|X]$ are chosen at a rate close to 1. Furthermore the variance $Var(Y)$ will be well estimated by $Var(Y|X)$ if and only if $Var\{E(Y|X)\}$ is small and the variance of $Var(Y|X)$ as a function of the random variable is small. So the coverage calculated from model X will be reasonable if and only if the the model selection criteria choses with probability approaching 1 a set of models in the support of X for which $E(Y|X)$ and $Var(Y|X)$ vary little. The severity of this problem of underestimating the true variance is addressed partially in section 1.6 in which we propose and perform simulations.

3.9 A modification to the best-AIC regression spline approach

The best-AIC regression spline approach is liberal in its tendency to select models with time-varying hazard ratios and not the more simple proportional hazards model. Consider the comparison of a time-dependent regression spline model of dimension r to the constant relative risk model, which has dimension 1. Suppose the corresponding AIC are A_r and A_1. The probability

that the time-dependent model is selected is

$$\begin{aligned} Pr[A_1 < A_r] &= Pr[\, LPL_1 - 1 < \, LPL_r - 2] \\ &= Pr[-2(LPL_r - LPL_1) > 2(r-1)], \end{aligned} \tag{3.24}$$

where LPL_1 and LPL_r are the maximized log partial likelihoods corresponding to the constant and time-dependent relative risk models, respectively. If the true form of the relative risk is constant, then since the constant relative risk model is nested within the regression spline model then twice the difference of the maximized log partial likelihoods is asymptotically distributed as a chi-square whose degrees of freedom is the difference $r - 1$ (Gill and Andersen, 1980). This is a likelihood ratio test. Therefore

$$Pr[A_1 > A_r] = Pr[\chi^2_{r-1} > 2(r-1)]. \tag{3.25}$$

For instance for $r = 2, 3, 5$ and 10 the respective probabilities of rejecting the constant relative risk model when the relative risk is actually constant are 0.15, 0.13, .09 and .04.

In practise an analyst is certain to reduce the probability of rejecting the hypothesis of either a null effect of the covariate or a proportional hazard effect by appealling to hypothesis tests. Both because of its simple conceptually appealling form and its popularity the analysts would be expected to provide strong evidence that the ph model is invalid. For instance upon finding that say a cubic regression spline with 2 knots has the best AIC an analyst would test if this model is (i) significantly better than the null model

and (ii) significantly better than the proportional hazards model. Since trivially the null model is nested within any space of regression splines and since the proportional hazards model is also nested in any regression spline space a likelihood ratio statistic can be used to test if the best-AIC model fits significantly better than the null and constant hazard ratio models. Of course these tests are only nominal since implicitly by comparing the best-AIC to say the constant model one is performing multiple comparisons but this modification is a reasonable tool for reducing the frequency of type I error. The exact correction of this type I error is not a trivial problem.

When this modification to the best-AIC regression spline approach is applied to the variable prothrombin time we find that the quadratic polynomial for the relative risk of a 1 second increase in prothrombin time fits the observed data *significantly* better than the constant model, P=.0003.

In the simulations that follow we refer to this modification as the modified best-AIC regression spline approach.

Chapter 4

Small Sample Behaviour of Best-AIC Regression Splines

4.1 Overview

The asymptotic properties of the modified best-AIC regression spline, as we discussed in chapter 3 may be of limited value. For practical purposes it is more relevant to examine the behaviour of this approach when there is only a small number of observed failures. A typical approach to determining small sample behaviour is to perform a simulation.

In section 1 we describe the goals of, and the set-up for a simulation we perform. In section 2 we describe the generation of data for this simulation using a novel approach based on the generation of random variates from the risk sets. Section 3 concerns a few other details of the simulation, for instance, the possibility of divergent MPLE's. Finally, in section 4, the results of the simulation are reported.

4.2 Simulation

Through simulation we would like to examine the small sample behaviour of modified best-AIC regression spline estimates with respect to (i) type I error in the sense of rejecting the ph model, (ii) bias, (iii) coverage rates of asymptotically based nominal 95% confidence intervals, (iv) the frequency with which the general shape of the relative risk is captured, be it decreasing or increasing or U-shaped.

We shall generate samples of size 25, 50 and 100 where there is assumed to be 33% censoring. In the simulations that follow we shall refer to the respective sample sizes as $n_f = 17$, $n_f = 33$ and $n_f = 67$, where the notation n_f is meant to signify the number of failures. The variable whose impact upon the hazard function is being assessed is dichotomous taking either value with equal probability. Each sample will be simulated 500 times.

For each sample we consider only smooth regression splines. For instance for $n_f = 67$ we consider all regression splines with cubic or quadratic order and with 0 up to 3 knots, as well as the null, constant hazard ratio and linear hazard ratio models. For smaller sample sizes we consider a smaller array of models. For $n_f = 33$ we consider null, constant, linear, quadratic and 1-knot quadratic regression splines. For $n_f = 17$ we consider only the null, constant, linear and quadratic models. The choice of knot locations is automatic. It is determined by the quantiles of the observed failure times.

We have chosen 4 relative hazard functions from which we shall generate

data. They represent broadly speaking a range of possibilities that would be of interest and intepretable to a clinician who is establishing the predictive ability of some measurable feature or the time of efficacy of a treatment. They are (i) a constant relative hazard, (ii) a linear relative hazard, (iii) an exponentially decaying relative hazard and (Gore, 1984) (iv) a relative hazard initially null that increases and eventually decreases before the end of follow-up. We shall refer to the latter as the *Rise and Fall* model.

4.3 Data generation

There is many ways to generate data. The most popular approach for generating data from an arbitrary distribution F is to generate a random deviate, u, from $U(0,1)$. Then $F^{-1}(u)$ is a random deviate from F. To apply this approach to our problem we would have to specify the baseline hazard function as well as the relative hazard and calculate $F(t) = 1 - exp(-\int_0^t \lambda_0(t)exp\{\beta(t)x\}dt)$. A closed form for the cumulative hazard does exist if the baseline distribution is exponential and the relative hazard is constant or linear but does not exist when the relative hazard takes on the two other forms we are considering for seemingly any choice of the baseline hazard function. To avoid a complex data generation process we propose an alternative approach based on risk sets that also simplifies the interpretion of the simulation.

Assume that events, be they failures or censorings occur at fixed points, $t_1, ..., t_n$. We shall sample randomly from each of the corresponding risk

sets, $R_1, \ldots, R_n$. To incorporate censoring into this process we randomly consider time t_i a censoring time 33% of the time. If it is a censoring time we randomly select a subject from the risk set assuming that each is equiprobable. If it is not a censoring time we randomly select a subject from the risk set and say that the subject has failed using the following method: First we generate a deviate from $U(0,1)$, u. The cumulative discrete distribution induced by this risk set, R_i, assuming that the hazard function is given by $\lambda(t;x) = \lambda_0(t)e^{\beta(t)x}$, is

$$F_i(j) = \frac{\sum_{l \leq j, l \in R_i} e^{\beta(t_i)x_l}}{\sum_{l \in R_i} e^{\beta(t_i)x_l}}. \tag{4.1}$$

defined for each $j \in R_i$. Then the subject failing at time t_i is indexed by the j satisfying $F(j-) < u <= F(j)$. The $i+1$-th risk set consists of the i-th risk set less the subject randomly selected out of it.

4.4 Other details

We shall assume that the t_i's are uniformly spaced over the unit interval. This facilitates summarization of the simulation results. When we report the results we shall frequently refer to properties of $\hat{\beta}(t)$ and $\hat{Var}(\hat{\beta}(t))$ at the quartiles of the empirical distribution of observed failure times, where $t = 0.25, 0.50$ and 0.75, respectively.

For a dichotomous predictor in low sample sizes the probability that no finite MPLE exists is reasonably high. For instance for the proportional hazards model with k observed failures and equal numbers of subjects on

each of the two respective levels of the predictor the probability of no finite MPLE is on the order of $2(\frac{1}{2})^k$ since no finite MPLE would exist if every subject failing has the same level of the predictor. In our simulations the quadratic polynomial has no MPLE when $n_f = 17$ with frequency on the order of 1 in 25. This would happen if amongst the sequence of predictor values corresponding to failing subjects, ordered by failure time, there was a run with 2 or less transitions, since then a quadratic polynomial can be fitted that has zeroes in at or in between these transition points. Although it creates a bias in the simulation results we chose to ignore a particular estimate if it was not finite. So for instance if the quadratic polynomial failed it would not be considered as a choice for best model.

In our report of the results of the simulation we shall stratify the estimates according to whether (i) the null model was chosen, (ii) the constant model was choosen and (iii) a time-dependent relative risk was chosen. This seemed advantageous to us from two viewpoints; (i) It seemed unuseful to report overall results about the pointwise estimated variance and about the pointwise coverage when with a reasonably high frequency the null model would be selected, in which case the corresponding nominal **100%** confidence interval is 0. (ii) The overall results would be too much an artifact of the particular true models we chose. This way we separate the effect of power for detecting time-dependence from that of the bias and precision of time-dependent estimates.

4.5 Results of Simulation

When the true form of the relative risk is constant, $\beta(t) = 1$ for all t, a time-dependent estimate of the relative risk is chosen with frequency 11.6%, in samples with 17 events (on average, i.e.; sample size is 25), 13.8% in samples with 33 events, (say $n_f = 33$), and 19% when $n_f = 66$ (see Table 4.1). The fact that each of these frequencies exceeds 0.05 is to be expected since the likelihood ratio test with nominal type I error of 0.05 compares the best of a set of regression splines to the constant relative risk model. The fact that this emiprical type I error rate is increasing with sample size may be because for each increase in the number of events we consider successively more regression spline models using the AIC criterion.

This disturbing type I error rate when $n_f = 67$ is in part mitigated by the nearly flat shape of the 'significantly' better fitting time-dependent estimators. The mean values of $\hat{\beta}(0.25)$, $\hat{\beta}(0.50)$ and $\hat{\beta}(0.75)$, respectively amongst these time-dependent estimators are 1.4, 1.1 and 1.0. These time-dependent estimators change quickly in the 'tails'. The mean values of $\hat{\beta}(0)$ and $\hat{\beta}(1)$ are 3.3 and 5.1, respectively, with very large standard deviations of 12.4 and 11.6. This erratic behaviour toward the tails is more pronounced in smaller samples. Figure 7 depicts the true relative risk (thick line), the mean estimate(dashed line) and curves representing plus and minus one standard deviation of these estimates(dotted line) when $n_f = 33$ amongst estimates that are time dependent (i.e.; estimates resulting from a rejection of the null

True Hazard Ratio	Sample Size	Null	Constant	Time Dependent
Constant	17	.470	.414	.116
	33	.196	.644	.160
	67	.016	.794	.190
Linear	17	.784	.066	.150
	33	.682	.048	.270
	67	.414	.020	.566
Exponent.	17	.204	.350	.246
	33	.126	.406	.468
	67	.002	.214	.784
Rise/Fall	17	.746	.144	.110
	33	.564	.240	.196
	67	.330	.330	.340

Table 4.1: Frequency with which a particular model was chosen

and constant models).

As we described in section 4 any estimation technique that involves model selection will yield estimates of the variance that underestimate the true variance. In these simulations, the extent of this underestimation **increases** as the sample increases for simulations in which the true relative risk is constant and the estimate chosen is time-dependent. For simulations in which $n_f = 17$ the ratio of the mean estimate of the standard deviation to the empirical standard deviation of the estimates is 0.88, 0.96 and 0.83 at the 25-th, 50-th and 75-th percentiles along the time axis. Table 4.2 reports the coverage rates at each quartile for each sample size. For simulations in which there is 33 events on average the corresponding ratios are 0.58, 0.61 and 0.61. The corresponding ratios for the n_f =67 are 0.6, 0.625 and 0.45. Not surprisingly the coverage rates of the nominal pointwise 95% c.i.s of these relative risk estimates that are time-dependent also decrease as sample size increases. For $n_f = 17$ they range from 0.88 to 0.98 over the time axis. For $n_f = 33$ they range from 0.73 to 0.88 over the time axis. For $n_f = 67$ they range from 0.67 to 0.83 over the time axis. On the other hand amongst relative risk estimates that are constant the coverage remains intact at about 0.97 for all three event sizes.

When the true log relative risk is linear there is substantial inflation amongst the estimates that are time-dependent. Figure 8 depicts for $n_f = 33$ the true log relative risk (thick line), the mean estimate (dashed line) and symettric plus or minus one standard deviation curves for these estimates

True Hazard Ratio	Sample Size	Perentile 0.25	0.50	0.75
Constant	17	.940	.931	.921
	33	.837	.837	.850
	67	.751	.832	.703
Linear	17	.987	.933	.973
	33	.874	.822	.963
	67	.919	.901	.922
Exponent.	17	.000	.992	.976
	33	.991	.944	.927
	67	.929	.903	.926
Rise/Fall	17	.964	.945	.709
	33	.898	.939	.888
	67	.853	.871	.876

Table 4.2: Coverage rate of nominal 95% confidence interval at quartiles

(dotted lines). The line of unit relative risk is also indicated. For n_f = 33 there is also underestimation of the variance. The ratios of the mean estimated standard deviation to the standard deviation of the estimates are .89, .72 and .87, respectively at the quartiles t = 0.25, 0.50 and 0.75. The coverage rate of the nominal 95% c.i. ranges from 0.82 to 0.96 over the time axis. Table 4.2 reports the coverage rates at the quartiles.

In practise an analyst would likely interpret the relative risk as being

significantly different from unity when the nominal 95% c.i. interval for $\beta(t)$ does not contain 0. Figure 9 depicts the frequencies with which (i) the lower limit of the 95% c.i. exceeded 0 (solid line) and (ii) the upper limit of the 95% c.i. is less than 0 (dashed line), when the true log-relative risk is linear amongst those simulations where a time-dependent model is chosen. For instance the empirical estimate of the probability with which an analyst would conclude that $\beta(0.25)$ exceeds 0 is nearly 50%. With frequency of approximately 40% the analyst would conclude that $\beta(0.75)$ is less than 0. This shows limited power but the "significant" conclusions tend to be correct.

The actual estimates at each point in time of the relative risk are not, in practise, as important as the general location (increased risk vs decreased risk) and the shape of the log-relative risk function. For each simulation in which a time-dependent model is chosen we have identified whether the relative risk is (i) generally increasing, (ii) increasing then decreasing (iii) decreasing then increasing and (iv) generally decreasing, based on the estimates of the relative risk at $t = 0.25$, 0.50 and 0.75. When the true log relative risk is linear (and decreasing) 81% ($n_f = 17$), 87% ($n_f = 33$) and 76% ($n_f = 67$) of the time-dependent estimates are generally decreasing (see Table 4.3). The fact that for $n_f = 67$ the correct shape is identified with frequency of only 76% is troubling.

It is not surprising that the low-dimensional regression splines had more difficulty estimating the exponentially decaying version than the linear version of the log-relative risk. For instance when $n_f = 17$ a time-dependent

True Hazard Ratio	Sample Size	Shape			
		Increasing	Increasing Decreasing	Decreasing Increasing	Decreasing
Constant	17	.45	.09	.28	.19
	33	.25	.16	.31	.28
	67	.26	.22	.23	.28
Linear	17	.00	.05	.13	.81
	33	.00	.08	.05	.87
	67	.01	.12	.11	.76
Exponent.	17	.02	.02	.26	.71
	33	.15	.51	.11	.22
	67	.06	.71	.05	.18
Rise/Fall	17	.49	.29	.02	.20
	33	.25	.16	.31	.28
	67	.26	.22	.23	.29

Table 4.3: Frequency with which a particular shape of model was chosen

estimate could either be linear or quadratic. In this case the mean estimate of $\beta(t)$ is a function that is decreasing up to $t = 0.7$ afterwhich it is increasing. The standard deviation of these estimates is large. The true value of $\beta(0.50)$ is 0.67 whereas the standard deviation of the linear and quadratic estimates is 4.5. For $n_f = 17$ the estimates of the variance of $\hat{\beta}(t)$ are conservative. The mean estimated standard deviation typically doubled the empirical estimate of the standard deviation. The mean estimate of the standard deviation of $\hat{\beta}(0.5)$ was 8.8.

Figure 10 depicts the exponentially decaying log relative risk (thick line) with corresponding mean estimates (dashed line) and symettric curves representing plus and minus one standard deviation of the estimates (dotted line) when $n_f = 33$, amongst those simulations where a time-dependent model was chosen over the null or constant models. The true value and mean estimates are better correlated than for $n_f = 17$ but there is systematic bias towards over-estimation of relative risks. There is systematic inflation. Here the variance is slightly under-estimated. The ratios of the empirically estimated standard deviation to the mean estimated standard deviation at the 25-th, 50-th and 75-th percentiles are 0.81, 0.79 and 0.87 respectively. The corresponding coverage rates are 0.99, 0.94 and 0.93 (see Table 4.2). Figure 11 depicts (i) the frequency with which the lower limit of the 95% c.i. exceeded 0 (solid line) and (ii) the frequency with which the upper limit deceeded 0 (dashed line). An analyst using the nominal pointwise 95% c.i. would conclude that the covariate had a short-term predictive ability roughly 75% of

the time.

The corresponding results for the exponentially decaying log-relative risk when $n = 67$ are similar. In general, when $n = 67$, estimates are less biased but the coverage averages only 0.90 between $t = 0.25$ and 0.75.

The general decreasing aspect of the exponentially decaying log-relative risk is captured with frequency 71%($n_f = 17$), 74%($n_f = 33$), and 74%($n_f = 67$) amongst estimates that are time-dependent (see Table 4.3).

Figure 12 depicts the rise and fall log relative risk (thick line) with corresponding mean estimates (dashed line) and symmetric plus and minus one standard deviation of the estimates (dotted line) when $n_f = 33$, amongst the estimates that are time-dependent. The estimates capture the general form of the rise and fall except at the tails and except that they are substantially inflated. There is a second rise in these estimates toward the end of follow-up but the corresponding estimates of the standard deviation are large so that the analyst would rarely be mislead. The mean estimated standard deviation compares favourably to the standard deviation of the estimates. The ratio of the former to the latter at the 25-th, 50-th and 75-th percentiles is 0.79, 1.01 and 1.00. between $t = 0.25$ and 0.75 the coverage averages 0.91 (see Table 4.3). Figure 13 depicts (i) the frequency with which the lower limit of the 95% c.i. exceeded 0 (solid line) and (ii) the frequency with which the upper limit deceeded 0 (dashed line), amongst those simulations where the time-dependent model was chosen. This figure suggests that an analyst would conclude that the predictor had a rise and fall impact on the relative

risk midway through follow-up about 75% of the time The estimate of the relative risk was generally increasing, then decreasing with frequency 51%

For the larger sample size, $n_f = 33$, a generally increasing, then decreasing represented 71% of the time-dependent estimates. However, the coverage was worse, averaging 0.85 between $t = 0.25$ and 0.75 (see Table 1.3). This was due to underestimation of the standard deviation. Between $t = 0$ 25 and 0.75 the ratio of the empirical estimate of $\hat{\beta}(t)$ to the mean value of $\hat{Var}(\hat{\beta}(t))$ was 0.8.

In summary the simulations yield a rather consistent picture, across the variations in the shape of the *true* hazard ratio and sample sizes. Some bias in the estimates of the hazard ratio is underestimable given that the best-AIC regressions splines were often of low dimension. This can be explained to some extent by the problem of insufficient power to detect more complex shapes. More importantly, despite this bias, the best-AIC regression splines often captured the general shape of the hazard ratio, be it increasing, decreasing or U-shaped. We believe these results justify our approach as a useful exploratory tool even in small samples.

As expected the simulations clearly demonstrated that model selection has a large effect on the rate of rejecting the constant hazard ratio and on the validity of MPLE based confidence intervals. Increasing sample size, while simultaneously increasing the array of candidate models, did not eliminate these problems and even aggravated them. Further research is necessary to develop pragmatic strategies to account for the impact of model selection on

MPLE based inference.

Conclusion

Cox's proportional hazards model is conceptually appealling due to its simplicity. However this model is not always appropiate and may mislead the clinical researchers who employ it. It may prevent them from detecting important effects of independent variables that are limited to a portion of the follow-up period such as the short-term. Several tests can be used to test the appropiateness of the constant hazard ratio assumption but very little research has been conducted into the representation of the hazard ratio as a function of time when the assumption of a constant hazard ratio is rejected. In this thesis a regression spline approach has been proprosed for estimating the variable effects of an independent variable on the hazard function over time.

Regression splines can be employed with relatively low computational burden to estimate the hazard ratio as a function of time. Akaike's Information criterion can be used to guide model selection. The combination of these two techniques yields a convenient tool for exploring the time-frame within which variables have predictive ability.

We have examined the small sample properties of the best-AIC regression

spline approach. As we expected this approach is liberal. Further research is required to determine exact hypothesis tests of, for instance, the proportional hazards model versus the alternative of a best-AIC regression spline hazard ratio estimate. As we also expected the 95% confidence intervals based on large sample MPLE theory have low coverage even when the number of observed failures is as large as 67. Further research is required to determine methods with low computational cost for creating more precise confidence intervals.

There is a number of generalizations of our work. First, in our model we have assumed that the effect of an independent variable is log-linear. Our method could be incorporated with the methods of Sleeper and Harrington(1990), and Durrleman and Simon (1988), who model the function form of the impact of the independent variable using regression splines and the AIC. Interesting identifiability problems may arise when both the functional forms for the impact of the variable and for the way the hazard ratio changes over time compete to explain survival. A second obvious generalization of our work is to model the hazard ratio using more than one variable. Other possible directions are to compare the regression spline approach to other non-parametric regression techniques, such as smoothing splines, and kernel smoothing.

Bibliography

[1] Abrhamowicz, M., Ciampi, A. and Ramsay, J.O. (1992) Nonparametric density estimation for censored survival data: Regression spline approach, *Canadian Journal of Statistics*, **2**, 171-175

[2] Akaike, H. (1974) A new look at statistical model identification, *IEEE Transactions on Automatic Control*, **AC-19**, 716-723

[3] Andersen, P.K. (1982) Testing Goodness of Fit of Cox's Regression and Life Model, *Biometrics*, **38**, 67-77

[4] Andersen, P.K. and Gill, R.D. (1982) Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, **10**, 1100-20

[5] Atkinson, (1980) A note on the generalized information criterion for choice of a model. *Biometrika*, **67**, 413-418.

[6] Barlow, W. E. and Prentice, R. L. (1988) Residuals for relative risk regression. *Biometrika*, **75**, 65-75

[7] Bloxom, B. (1985) A constrained spline estimator of a hazard function. *Psychometrika*, **50**, 301-321.

[8] Casella, G. and Berger R. (1990) *Statitical Inference.* Wadsworth, Belmont, California

[9] Cox, D. R. (1972) Regression models and life tables. *J.R.Statis.Soc. B* **34**, 187-220

[10] Cox, D. R. (1975) Partial likelihood, *Biometrika*, **62**, 269-76

[11] Cox, D. R. (1979) A note on the graphical analysis of survival data. *Biometrika*, **66**, 188-190.

[12] Cox, D. R. and Oakes D. (1984) *Analysis of Survival Data.* Chapman and Hall, London

[13] Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, *Numerische Mathematik*, **31**, 377-403

[14] Crowley, J. (1974) Asymptotic normality of a new nonparametric statistic for use in organ transplant studies. *JASA*, **69**, 1006-1011

[15] Curry, H.B. and Schoenberg, I.J. (1966) On Polya frequency functions. IV: The fundamental spline functions and their limits. *J. Analyse Math.*, **17**, 71-107.

[16] deBoor, C. (1978) *A Practical Guide to Splines*, New York: Springer-Verlag

[17] Durrleman, S. and Simon, R. (1989) Flexible Regression Models With Cubic Splines, *Statistics in Medicine* **8**, 551-561

[18] Eubank, R.L. (1988) *Spline Smoothing and Nonparametric Regression* New York: Marcel Dekker.

[19] Fleming, T.R. and Harrington, D.P. (1991) *Counting processes and survival analysis.* Wiley, New York

[20] Gehan, E. A. (1965) A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, **52**, 203-223

[21] Gill, R. and Schumacher, M. (1987) A simple test of the proportional hazards assumption, *Biometrika*, **74**, 289-300

[22] Gore, S.M., Pocock, S.J. and Kerr, G.R. (1984) Regression Models and Non-proportional Hazards in the Analysis of Breast Cancer Survival, *Applied Statistics*, **33**, 176-195

[23] Gu M. (1992) A likelihood ratio test for the fit of proportional hazards model. Report 92-19 Mathematics and Statistics Department, McGill University

[24] Harrington, D.P. and Fleming, T.R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, **69**, 553-66

[25] Hastie, T.J. and Tibshirani, R.J. (1990) *Generalized Additive Models.* Chapman and Hall, New York

[26] Henderson, R. and Milner, A. (1991) On residual plots for relative risk regression, *Biometrika*, **78**, 631-6

[27] Hurvich, C.M. and Tsai, C. (1990) The Impact of Model Selection on Inference in Linear Regression, *The American Statistician*, **44**, 214-217

[28] Kalbfleisch, J.D. and Prentice, R.L. (1973). Marginal likelihoods based on Cox's rgeression and life model. *Biometrika*, **60**, 267-78

[29] Kalbfleisch, J.D. and Prentice, R.L. (1980) *The Statistical Analysis of Failure Time Data.* New York: Wiley

[30] Kaplan, E. L. and Meier, P. (1958) Nonparametric estimation from incomplete observations, *JASA*, **53**, 457-481

[31] Kay, R. (1977) Proportional Hazard Regression Models and the Analysis of Censored Survival Data, *Appl. Statist.*, **26**, 227-237

[32] Kendall, M. and Stuart, A. (1979) *The Advanced Theory of Statistics, Volume II: Inference and Relationship*, 4th edition. New York: Macmillan

[33] Kullback, S. and Leibler, R.A. (1951) On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79-86

[34] Lehmann, E.L. (1975), Nonparametrics: *Statistical Methods Based on Ranks*, Holden-Day, San Fransico

[35] Lliang, Self and Liu (1990)Mantel and Haenszel (1959) Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, **22**, 719-748

[36] Mantel, N. (1966) Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer. Chemother. Rep.*, **50**, 163-170

[37] Moreau T., O'Quigley J. and Mesbah, M. (1985) A Global Goodness-of-fit Statistic for the Proportional Hazards Model, *Appl. Statist.*, **34**, 212-18

[38] Nagelkerke, N. J. D., Oosting, J. and Hart, A.A.M. (1984) A Simple Test for Goodness of Fit of Cox's Proportional Hazards Model, *Biometrics*, **40**, 483-486

[39] Nelson, W. (1969) Hazardplotting for incomplete failure data. *Journal of Quality Technology*, **1**, 27-52

[40] O'Quigley J. and Pessione, F. (1989) Score Tests for Homogeneity of Regression Effects in the Proportional Hazards Model, *Biometrics*, **45**, 135-44

[41] O'Sullivan (1988) Fast Computation of Fully Automated Log-Density and Log-Hazard Estimators, *SIAM J. Sci. St. Stat.*., 363-379

[42] Peterson, A. V., Jr. (1977) Expressing the Kaplan-Meier estimator as a function of empirical subsurvival functions. *JASA*, **72**, 854-858.

[43] Petit A. N. and Bin Daud, I. B. (1990) Investigating Time Dependence in Cox's Proportional Hazards Model, *Appl. Statist.*, **39**, 313-329

[44] Prentice, R.L. (1978) Linear rank tests with right censored data. *Biometrika*, **65**, 167-79

[45] Prentice and Farewell (1986) Relative risk and odds ratio regression. *Annual Review of Public Health*, **7**, 35-58

[46] Prentice and Self (1983) Asymptotic Distribution Theory for Cox-Type Regression Models With General Relative Risk Form, *Annals of Statistics*, **11**, 804-813

[47] Ramsay, J.O., (1988), Monotone Regression Splines in Action, *Statistical Science*, **3**, 425-461

[48] Ramlau-Hansen, (1983) Smoothing counting process intensities by means of kernel functions. *Ann. Statist.*, **11**, 453-466

[49] Schoenfield, D. (1980) Chi-squared goodness-of-fit tests for the proportional hazards regression model, *Biometrika*, **67**, 145-53

[50] Schoenfield, D. (1982) Partial residuals for the proportional hazards regression model, *Biometrika*, **69**, 239-41

[51] Sleeper, L.A. and Harrington, D.P. (1990) Regression Splines in the Cox Model With Application to Covariate Effects in Liver Disease, *JASA*, **85**, 941-949

[52] Stone (1977) An Asymptotic Equivalence of Choice of Model by Cross Validation and Akaike's Criterion, *Journal of Royal Statistical Society, Ser. B*, **39**, 44-47

[53] Therneau, T. M., Grambsch, P. M. and Fleming, T. R. (1990) Martingale-based residuals for survival models, *Biometrika*, **77**, 147-60

[54] Tsiatis, A.A. (1981) A large sample study of Cox's regression model *Ann. of Statist.* **9**, 93-108.

[55] Turnbull, B.W., Brown, B.W. and Hu, M. (1974) Survivorship analysis of heart transplant data. *JASA*, **69**, 74-80

[56] Vuong (1989) Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrika* **57**, 307-333

[57] Wegman and Wright (1983) Splines in statistics. *JASA*, **78**, 351-366

[58] Wei, L. J. (1984) Testing Goodness of Fit for Proportional Hazards Model With Censored Observations, *JASA*, **79**, 649-52

[59] Zucker, D.M. and Karr, A.F. (1990) Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach, *Ann. Statist.*, **18**, 329-353

Figures

Figure 1. Comparison of survival via Kaplan-Meier estimates

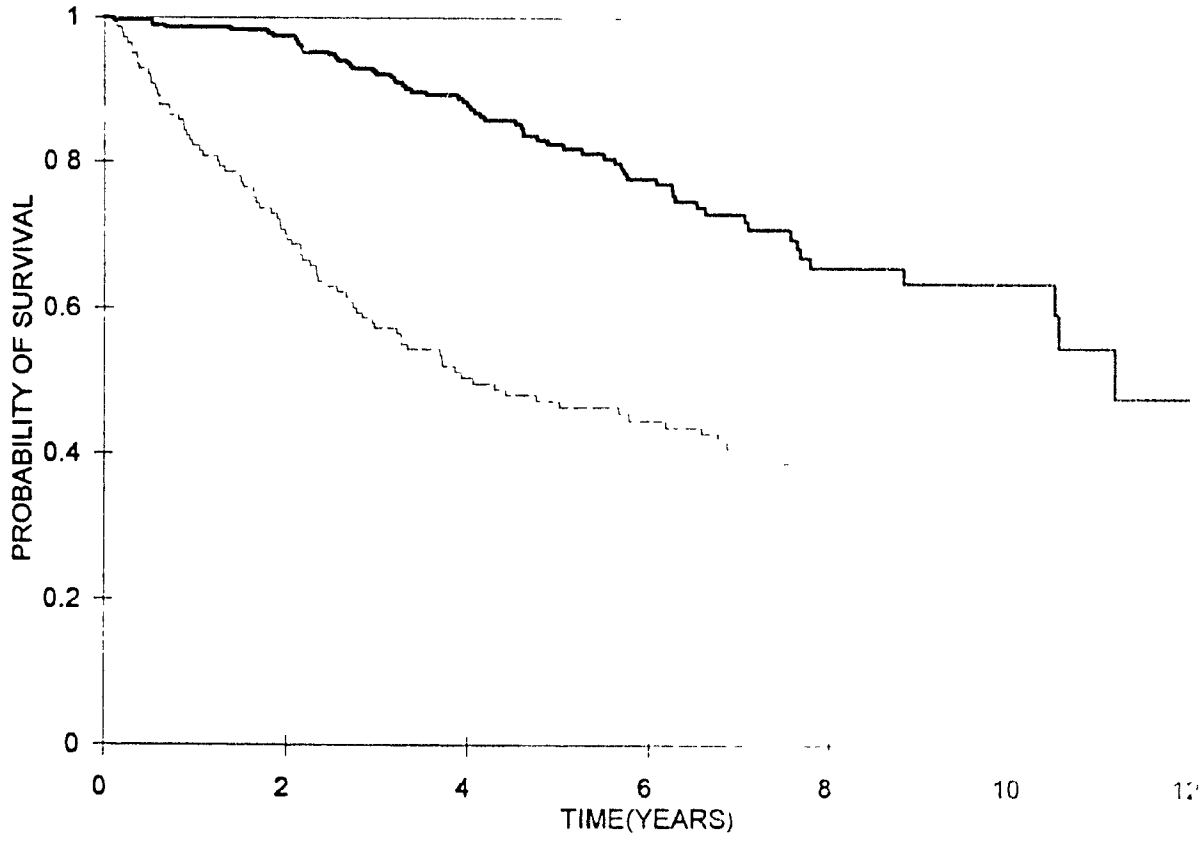


Figure 2. Log(-log(S(t)) curves to assess proportionality of hazards

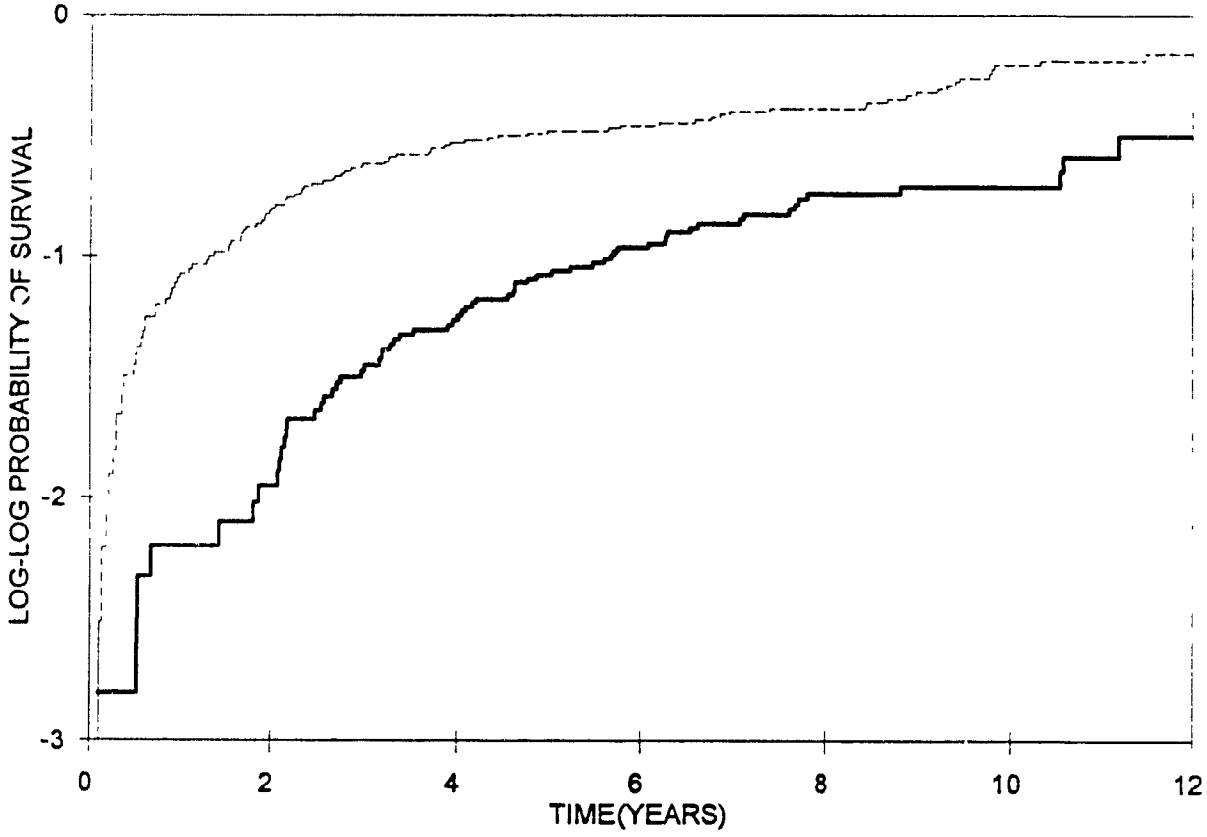


Figure 3. Step function estimate of the hazard ratio

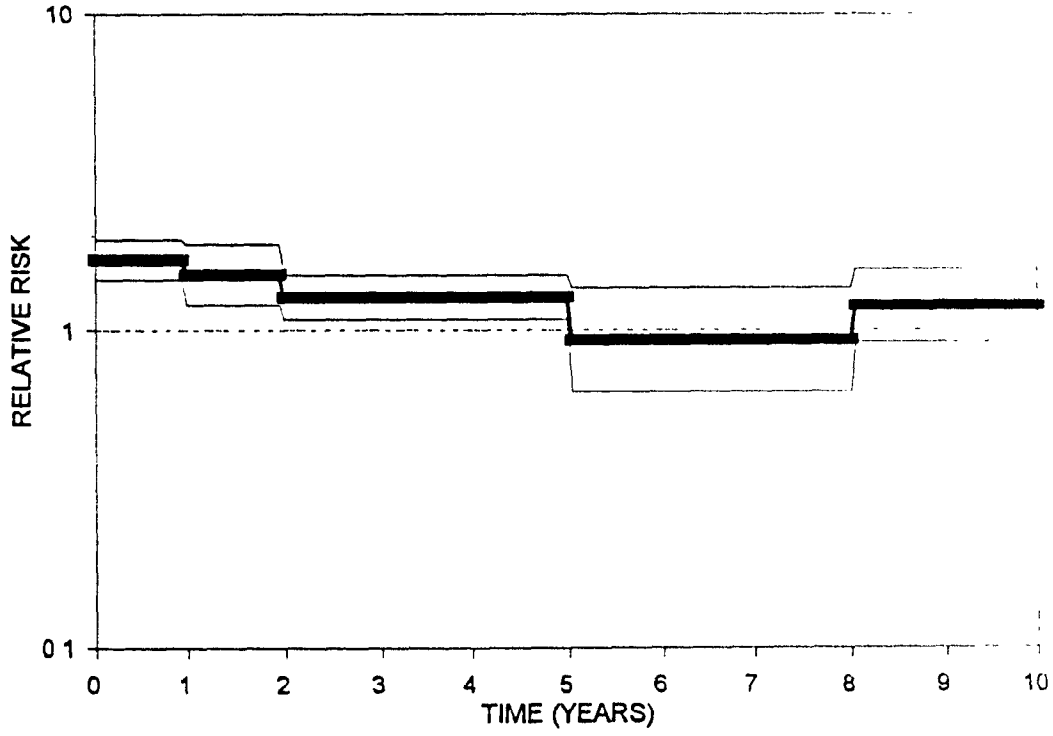


Figure 4. Partial residuals

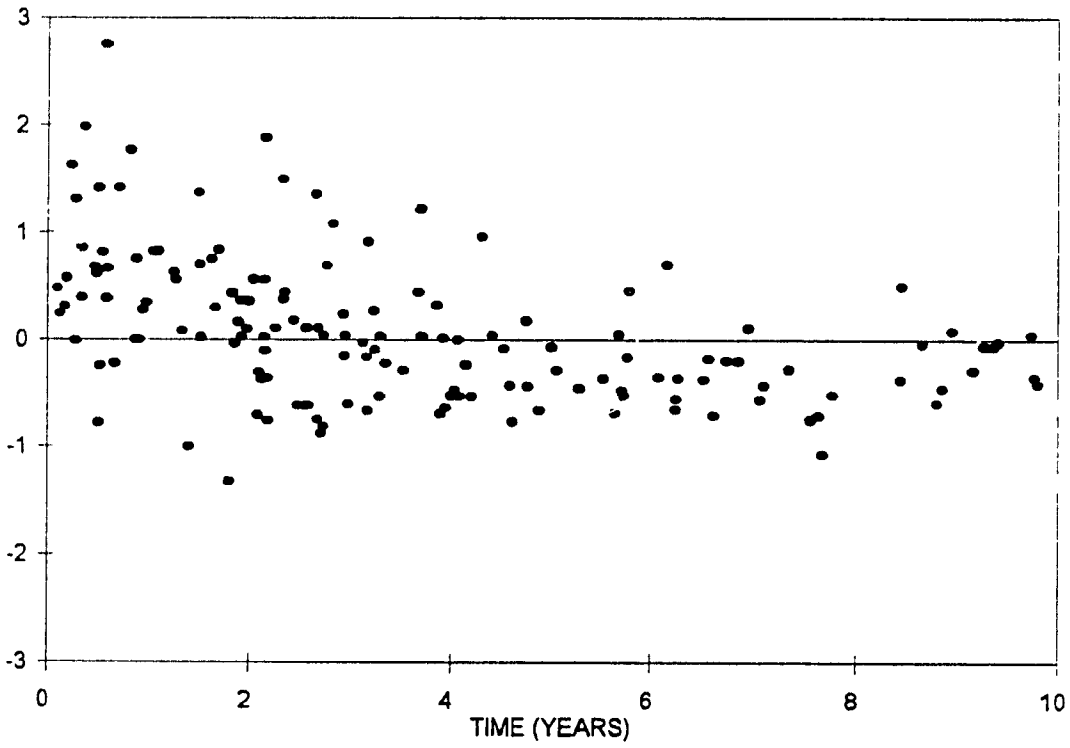


Figure 5. Cubic regression spline with 3 knots

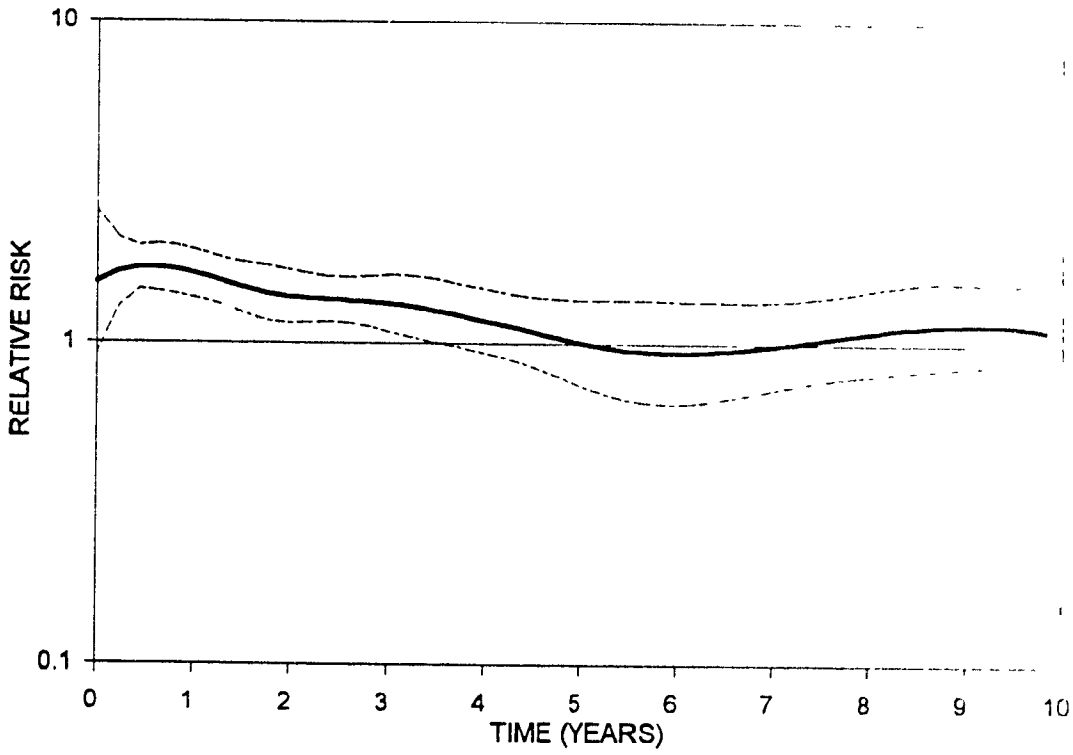


Figure 6. AIC-Best regression spline

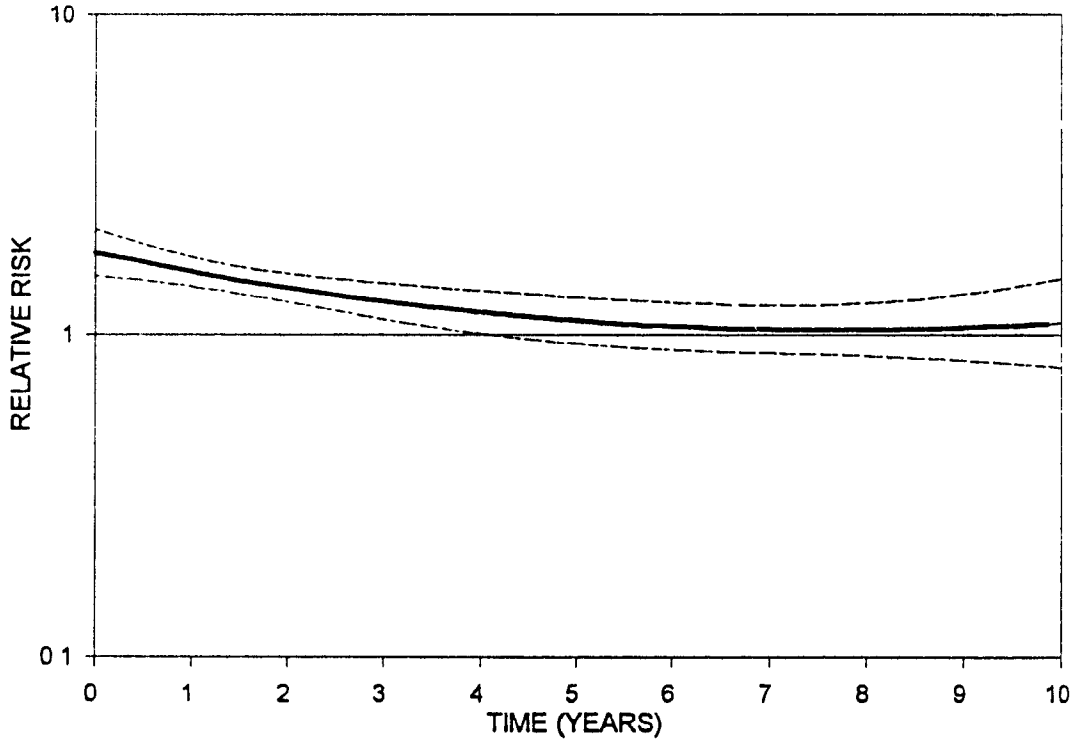


Figure 7

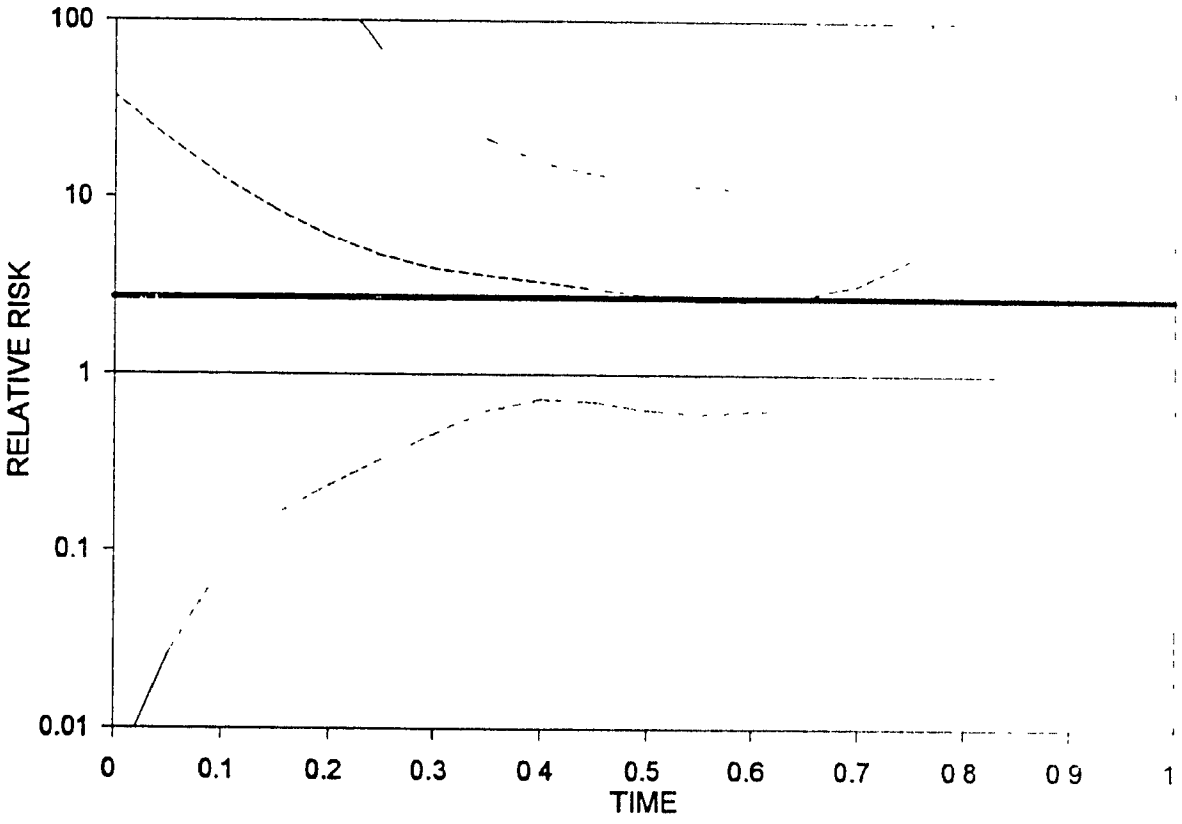
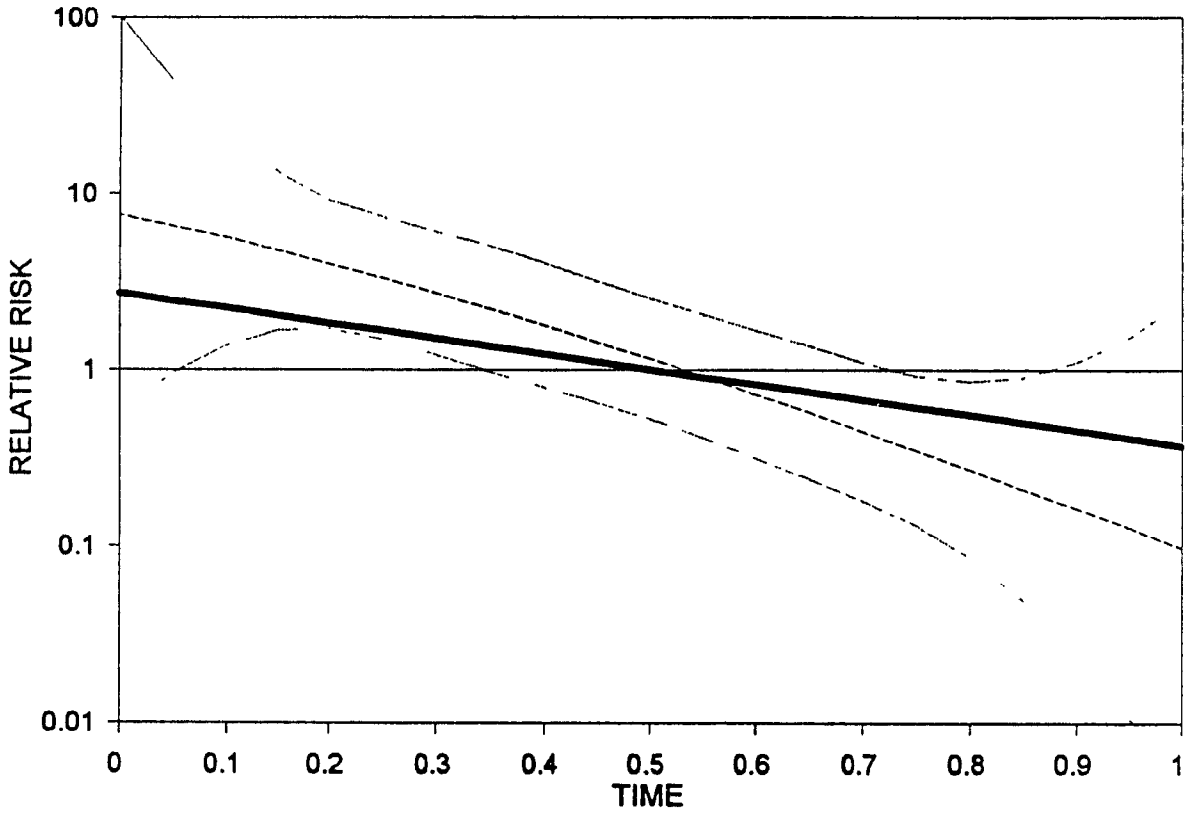


Figure 8



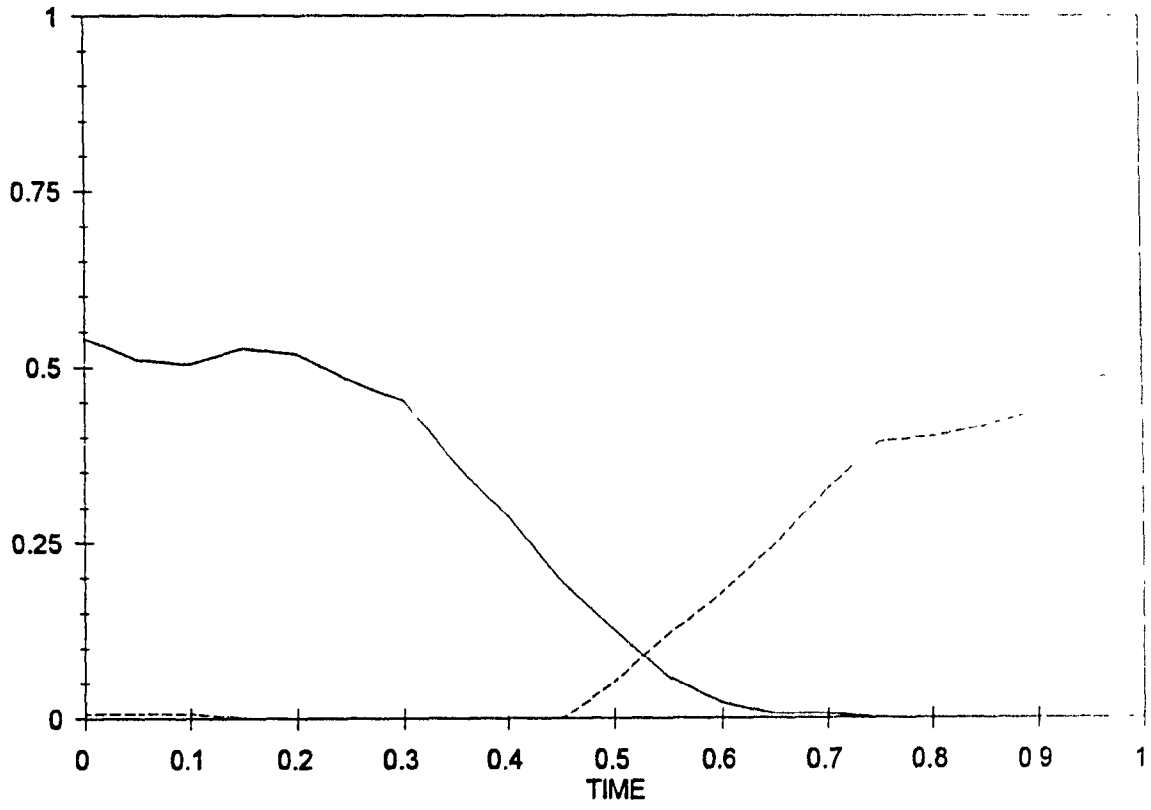

1
0.75
0.5
0.25
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0 9
1
TIME

Figure 10

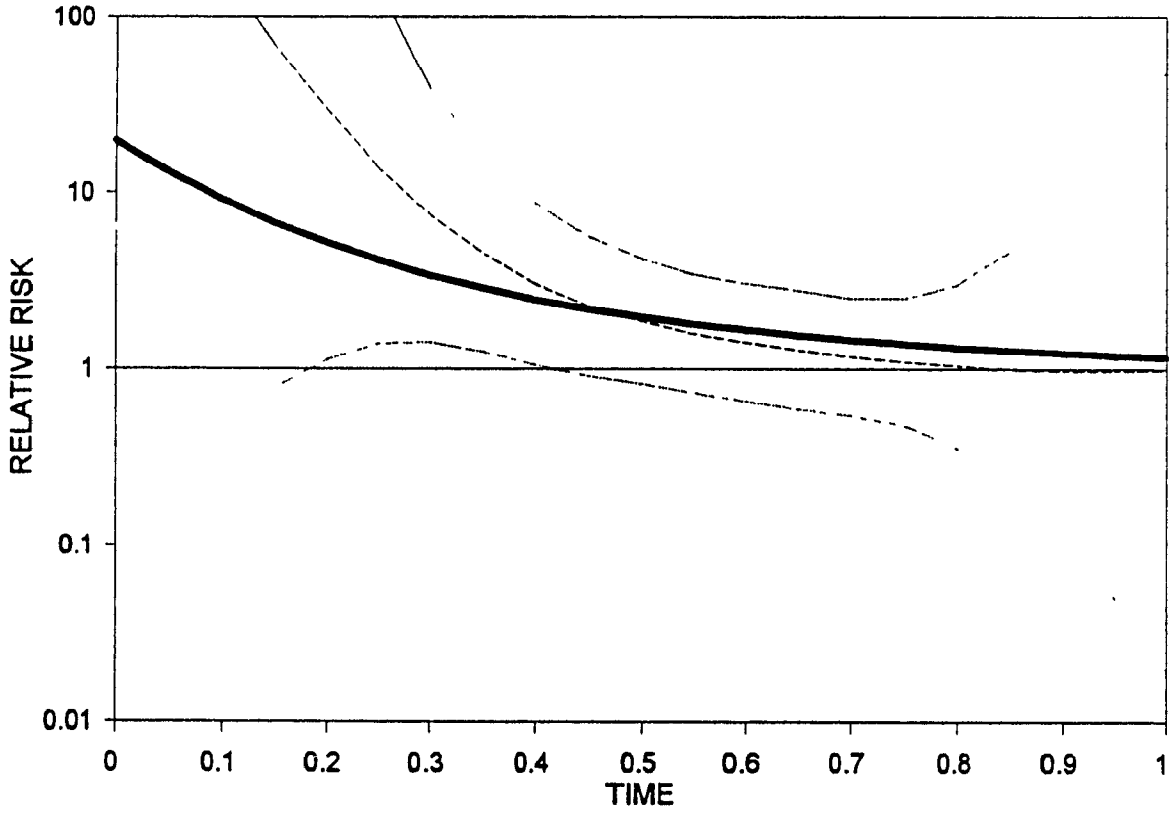


Figure 11

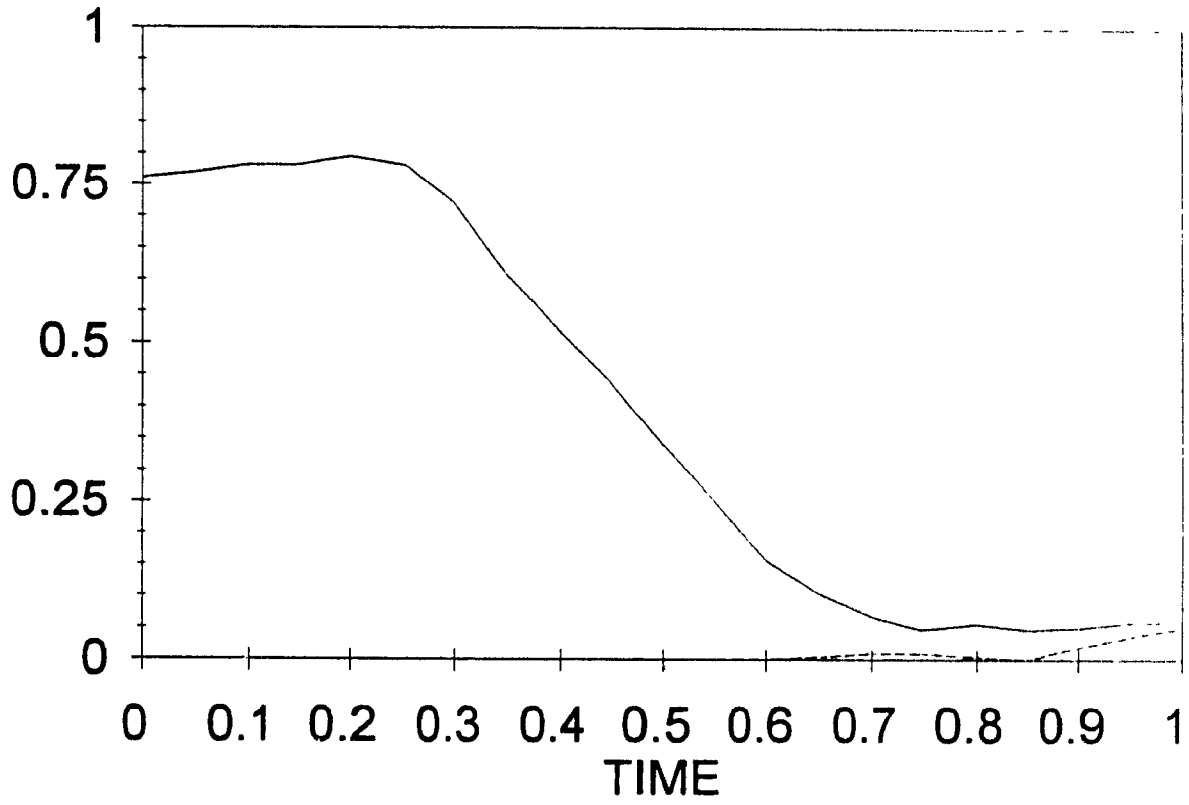


Figure 12

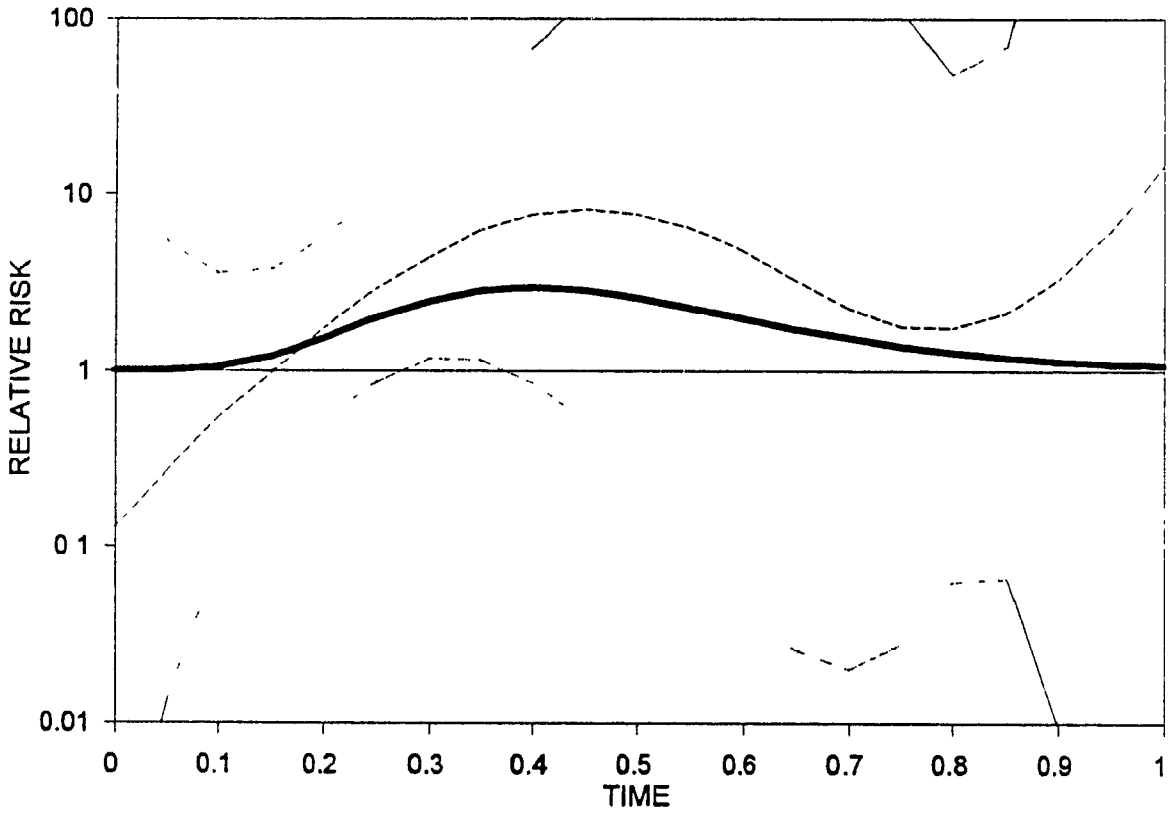


Figure 13