HUILING CAO

A Comparison Between the Additive and Multiplicative Risk Models

Mémoire présenté
à la Faculté des études supérieures de l'Université Laval
dans le cadre du programme de maîtrise en statistique
pour l'obtention du grade de Maître ès sciences (M.Sc.)

FACULTÉ DES SCIENCES ET DE GÉNIE
UNIVERSITÉ LAVAL
QUÉBEC

Septembre 2005

Résumé

L'analyse des durées de vie étudie et modélise le temps avant que les événements ne se produisent. Elle se concentre surtout sur la distribution des temps d'événements; les données de survie sont en fait des données sur les temps d'événements, comme par exemple le temps avant l'apparition d'une tumeur ou le retour d'une maladie. Les modèles de régression pour les données de durées de vie sont traditionnellement basés sur le modèle des risques proportionnels de Cox, le principal cheval de bataille pour les analyses de régression en présence de données censurées. L'effet des covariables y est supposé multiplicatif sur une fonction de risque de base arbitraire, ce qui rend la modélisation des effets variants dans le temps difficile. Deuxièmement, si des covariables sont enlevées du modèle ou mesurées avec erreur, l'hypothèse de proportionnalité des risques peut s'avérer fausse. Ces failles du modèle de Cox ont généré de l'intérêt pour d'autres modèles. Un de ces modèles est le modèle des risques additifs d'Aalen (1989). Ce modèle suppose que les covariables agissent de façon additive sur une fonction de risque de base arbitraire. Les coefficients de régression peuvent également y être fonction du temps, ce qui permet de supposer que l'effet des covariables varie dans le temps. Le modèle d'Aalen n'est tout-de-même pas communément utilisé. Une explication possible est que le modèle ne peut être ajusté simplement aux données à partir de logiciels communs comme SAS ou S-Plus. Dans ce mémoire, nous utilisons une macro SAS pour ajuster le modèle aux risques additifs. Le but de ce mémoire est de comparer les modèles aux risques proportionnels et additifs du point de vue théorique, par des exemples d'application et par une étude de simulation. En plus d'énumérer les avantages et inconvénients des modèles, nous donnons des instructions sur le choix d'un bon modèle à ajuster dans une situation donnée.

Thierry Duchesne
Directeur de recherche

Huiling Cao
Étudiante

Abstract

Survival analysis examines and models the time it takes for events to occur. It focuses on the distribution of survival times. Survival data is time-to-event data, such as time to death, appearance of a tumor, or recurrence of a disease. Regression models for survival data have traditionally been based on the proportional hazards model, proposed by Cox, that has become the workhorse of regression analysis for censored data. The effect of the covariates on survival is to act multiplicatively on some unknown baseline hazard rate, which makes it difficult to model covariate effects that change over time. Secondly, if covariates are deleted from a model or measured with a different level of precision, the proportional hazards assumption is no longer valid. These weaknesses in the Cox model have generated interest in alternative models. One such alternative model is Aalen's (1989) additive model. This model assumes that covariates act in an additive manner on an unknown baseline hazard rate. The unknown risk coefficients are allowed to be functions of time, so that the effect of a covariate may vary over time. Aalen's additive model is not yet widely used. One reason for this is that the model is not available in any commonly used computer package, such as SAS or S-plus. In this thesis we use a SAS macro that performs the additive hazards regression. The aim of this thesis is to compare the proportional and additive hazards models through theory, application and simulation. We also highlight their respective advantages and disadvantages, and give guidelines as to which model to choose to fit given survival data.

Acknowledgements

I would first and foremost like to express my deep gratitude to my supervisor, professor Thierry Duchesne, Department of Mathematics and Statistics of Laval University for all the great help and guidance he has provided during the work on this thesis. He has not only assisted me with his excellent advice, but he also was always making himself available whenever I needed. Working with him has been one of the most enjoyable parts of my learning experience in this department. Merci beaucoup, Thierry.

My thanks also go to professor of Medical College of Wisconsin, Christian Boudreau, for sending me a copy of Howell, A. (1996)'s thesis. This thesis gave me much inspiration, so that I could smoothly proceed with my thesis. Thanks for your help.

I must also thank Mr.Gaetan Daigle, professional statistician at the Department of Mathematics and Statistics of Laval University. He wrote a program to do all the simulations in SAS. So thanks again.

Finally, special thanks go to my family, it has been great during this period of time. My husband Chunsheng and my daughter Simo have given me all the love and support I could possibly ask for.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Survival analysis is a statistical method designed to study the amount of time an experimental unit survives, or the study of time between entry into observation and a subsequent event. Originally, the event of interest was death and the analysis consisted of following the subjects until death.

The term “survival data” has been used in a broad sense for data involving time to the occurrence of a certain event. This event may be death, the appearance of a tumor, the development of some disease, recurrence of a disease, conception, cessation of smoking, and so forth. In the past, application of the statistical methods for survival data analysis have been extended beyond biomedical and reliability research to other fields, such as the social sciences and business. For example, we could look at the duration of a first marriage (sociology), the length of subscription to a newspaper or a magazine (marketing), and so on. The study of survival data was previously focused on predicting the probability of survival or mean lifetime, and comparing the survival distributions of experimental animals or of human patients under different conditions. In recent years, the identification of risk and/or prognostic factors related to survival, and the development of disease have become important applications of survival analysis.

A key characteristic that distinguishes survival analysis from other areas in statistics is that survival data are usually censored. The defining feature of censored data is that the event time of interest is not fully observed on all subjects under study. Censored data arise in a broad range of application areas (finance, industrial life testing, biomedical studies...). A major complication in analyzing such data is right censoring, where the event of interest is known to occur only after a certain time point. Other types of censoring, such as interval censoring or truncation often occur and present challenges along with complicated data structures in the analysis of failure time data.

1.1 Basic survival analysis definitions

In this chapter, we consider the basic parameters used in modeling survival data.

Let X denote the survival time. In fact, survival times measure the time to certain events such as failure, death, divorce. X is a nonnegative random variable from a homogeneous population. The distribution of X is usually described or characterized by three functions, namely

1. the survival function;
2. the hazard rate function or risk function;
3. the probability density (or probability mass) function.

The three functions are mathematically equivalent – if one of them is given, the other two can be derived.

Next, we briefly describe these three equivalent functions that can characterize the distribution of X and discuss the relationship among the three functions (Lee, 1992).

1. Probability density function.

   This function, denoted by $f(x)$, is defined as the limit of the probability that an individual fails in the short interval x to $x + \Delta x$ per unit width Δx, or simply the probability of failure in a small interval per unit time:

   $$\begin{aligned} f(x) &= \lim_{\Delta x \to 0} \frac{P\{\text{an individual dying in the interval } (x, x + \Delta x)\}}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{P[X \in (x, x + \Delta x)]}{\Delta x}. \end{aligned}$$

   The density function has the following two properties:

   (1) $f(x)$ is a nonnegative function, $f(x) \geq 0, \forall x \geq 0$.

   (2) The area between the density curve and the x axis is equal to 1, i.e,
   $$\int_0^\infty f(x) \mathrm{d}x = 1.$$

2. Survival function.

This function, denoted by $S(x)$, is defined as the probability that an individual survives longer than x:

$$\begin{aligned} S(x) &= P(\text{an individual survives longer than } x) \\ &= P(X > x) \\ &= \int_x^\infty f(t)\mathrm{d}t. \end{aligned}$$

From the definition of the cumulative distribution function, $F(x) = P(X \leq x)$, we have

$$\begin{aligned} S(x) &= 1 - P(\text{an individual fails before time } x) \\ &= 1 - F(x), \end{aligned}$$

where $S(x) = 1$ for $x = 0$ and $S(x) = 0$ for $x = \infty$, that is, the probability of surviving at least to time zero is 1 and that of surviving to infinite time is zero. If X is a continuous random variable, then $S(x)$ is a continuous, strictly decreasing function.

The survival curve describes the relationship between the probability of survival and time. Many types of survival curves are observed in practice, but the important point to note is that they all have the same basic properties: they are monotone, non increasing functions equal to one at time zero and zero as the time approaches infinity.

3. Hazard function.

This function, denoted by $h(x)$, is defined as the probability of failure in a very small time interval, assuming that the individual has survived to the beginning of the interval, or as the limit of the probability that an individual fails in a very short interval x to $x + \Delta x$ per unit time, given that the individual has survived to time x:

$$\begin{aligned} h(x) &= \lim_{\Delta x \to 0} \frac{P\{\text{an individual of age } x \text{ fails in the time interval } (x, x+\Delta x)|\text{alive at x}\}}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{P\{x \leq X < x + \Delta x | X \geq x\}}{\Delta x}. \end{aligned}$$

If X is a continuous random variable, then

$$h(x) = \frac{f(x)}{S(x)} = \frac{f(x)}{1 - F(x)} = -\frac{d}{\mathrm{d}x}\ln[S(x)].$$

The cumulative hazard function, $H(x)$, is defined as

$$H(x) = \int_0^x h(u)\mathrm{d}u = -\ln[S(x)].$$

Thus at $x = 0$, $S(x) = 1$, $H(x) = 0$ and at $x = \infty$, $S(x) = 0$ and $H(x) = \infty$.

The cumulative hazard function can take any value between zero and infinity. There are many general shapes for the hazard rate, the only restriction on $h(x)$ being that it is nonnegative, i.e. $h(x) \geq 0$. From many practical examples, we may get that the hazard rate for the occurrence of a particular event is increasing, decreasing, constant, bathtub-shaped, hump-shaped, and so on.

In fact, the hazard function is usually more informative about the underlying mechanism of failure than the survival function. For this reason, consideration of the hazard function may be the dominant method for summarizing survival data.

3. Relationship among the three functions.

   The three functions defined above are mathematically equivalent. Given any one of them, the other two can be derived:

$$
\begin{aligned}
S(x) &= \int_x^\infty f(t)\mathrm{d}t = \exp\left[-\int_0^x h(u)\mathrm{d}u\right] = \exp[-H(x)];\\
f(x) &= -\frac{d}{\mathrm{d}x}S(x) = h(x)S(x) = h(x)e^{-\int_0^x h(u)\mathrm{d}u};\\
h(x) &= -\frac{d}{\mathrm{d}x}\ln[S(x)] = \frac{f(x)}{S(x)} = \frac{f(x)}{\int_x^\infty f(u)\mathrm{d}u}.
\end{aligned}
$$

1.2 Censoring and truncation

The analysis of survival data is complicated by censoring and/or truncation.

- Censored data arise when an individual's life length is not known exactly, but only known to occur in a certain interval of time. One possible type of censoring is right censoring, where all that is known is that the individual is still alive at a given time. Here, the period of observation expires, or an individual is removed from the study, before the event occurs– for example, some individual may still be alive at the end of a clinical trial, or may drop out of the study for various reasons other than death, prior to its termination. In this case, we only have a lower bound for the value of X for this individual. One type of right censoring that is very common is type I censoring, where the event is observed only if it occurs prior to some prespecified time, e.g, at the closing of a study. A second type of right censoring is type II censoring, in which the study continues until the failure of the first r individuals, where r is some predetermined integer.

Left censoring is when all that is known is that the individual experienced the event of interest prior to the start of the study, i.e., only an upper bound for the value of X is available.

For interval censoring, the only information available is that the event occurs within some interval of time.

- A second feature of many survival studies, sometimes confused with censoring, is truncation. Truncation is a condition which screens certain subjects so that the investigator will not be aware of their existence, i.e., only individuals whose survival time X meets some condition are observed. Types of truncations are (1) left truncation, where only individuals who survive a certain time before the study starts are included; (2) right truncation, where only individuals who have experienced the event by a certain time are included in the study.

- Most methods used in survival analysis are designed to handle right censored data.

1.3 Likelihood construction for censored and truncated data

- Likelihood construction for Type I right censoring

In this section, we construct the likelihood for fixed right-censored (Type I) data and begin with a review of the notation.

Type I right censoring definition:

The survival variables X_1,X_2,X_3....X_n are right-censored by fixed constants c_1, c_2, c_3, $\ldots$, c_n, if the observed sample consists of the ordered pairs $(z_i, \delta_i), i = 1, 2, 3 \cdots n$, where, for each i,

$$z_i = \min\{X_i, c_i\}$$
$$\delta_i = \begin{cases} 1 & \text{if } X_i \leq c_i \text{ (uncensored)} \\ 0 & \text{if } X_i > c_i \text{ (censored)} \end{cases},$$

where c_i is the fixed censoring time and δ_i is the censoring indicator for X_i.

Our assumptions here are that X_1,X_2,X_3....X_n are independent of the c_1, c_2, c_3, $\cdots$, c_n, and identically distributed (iid) from a continuous distribution with p.d.f

$f(c|\theta)$, where θ belongs to some parameter space. Here, θ could be either a real-valued or vector-valued parameter. Our aim is to determine the joint density of the observed data $(z_i, \delta_i), i = 1, 2, 3 \cdots n$, so that θ may be estimated by the method of maximum likelilood.

Theorem 1 (Lee, 1992, Chapter 1)

Under type I right-censoring with fixed censoring times, the joint likelihood $L(\theta)$ of the observed data $(z_i, \delta_i), i = 1, 2, 3 \cdots$, is given by

$$L(\theta) = c \prod_{i=1}^{n} f(z_i)^{\delta_i} S(z_i)^{1-\delta_i}, \qquad \textit{where } c \textit{ is a constant.} \tag{1.1}$$

Equation (1.1) may be generalized further to accommodate other types of censoring such as interval-censoring, left-censoring, $\cdots$. We can generalize this equation so that it has explicit terms for observed deaths, right-censoring, left-censoring, interval-censoring. To construct a likelihood of this general form where censoring times are independent of lifetimes, we take a product of terms of the following form:

$$\begin{aligned} f(y) &: \text{ for observed death at } y; \\ S(t) &: \text{ for right-censoring at } t; \\ 1 - S(l) &: \text{ for left-censoring at } l; \\ S(t_1) - S(t_2) &: \text{ for interval-censoring in } [t_1, t_2). \end{aligned}$$

Therefore, equation (1.1) may be generalized to

$$L(\theta) = c \prod_{D} f(y) \prod_{R} S(t) \prod_{L} [1 - S(l)] \prod_{I} [S(t_1) - S(t_2)], \tag{1.2}$$

where

$$\begin{aligned} D &: \text{the set of observed deaths;} \\ R &: \text{the set of right-censored observations;} \\ L &: \text{ the set of left-censored observations ;} \\ I &: \text{the set of interval-censored observations.} \end{aligned}$$

- Likelihood construction under truncation

Definition:

Suppose a population of an unknown number N of possible observations. We say that the dataset Y_1,Y_2,Y_3....Y_n of $n < N$ elements is left-truncated by the fixed truncation constants t_1,t_2,t_3....t_n if the observed sample consists of ordered pairs

(y_i, t_i), where $y_i \geq t_i$, so that y_i is left-truncated by t_i for each $i = 1, 2, \cdots n$. Right-truncation is defined similarly, with the observed sample consisting of ordered pairs (y_i, t_i) where $y_i \leq t_i$. In other words, elements can be observed only if their value falls in a specific interval.

Likelihood construction:

Under left-truncation, the observed data are ordered pairs $(y_i, t_i), i = 1, 2, \cdots n$ with $y_i \geq t_i$. Assuming that the times of death are independent of the truncation constants, we may write the conditional density as

$$f(y|y > t) = \frac{f(y)}{S(t)}.$$

Each ordered pair in the data makes such a contribution to the likelihood, so that

$$L(\theta) = c \prod_{i=1}^{n} \frac{f(y_i)}{S(t_i)}.$$

Similar ideas apply for right-truncation. In this case, the observed data are ordered pairs $(y_i, t_i), i = 1, 2, \cdots n$,with $y_i \leq t_i$. In this case, conditional densities are of the form

$$f(y|y < t) = \frac{f(y)}{1 - S(t)}.$$

Each ordered pair in the data makes such a contribution to the likelihood, so that

$$L(\theta) = c \prod_{i=1}^{n} \frac{f(y_i)}{1 - S(t_i)}.$$

For example:

We use the maximum likelihood estimation (MLE) theory above to estimate parameters. The SAS software is used to implement this method. We use the data listed in Dataset 1.1 of the Appendix for 100 subjects:

TIME: The follow-up time is the number of months between the entry date and the end date.

AGE: The age of the subject at the start of follow-up (in years).

CENSOR: Vital status at the end of the study (1=Death due to AIDS, 0=Lost to follow-up or alive).

For the moment, we only focus on how to use MLE theory to estimate parameters (Hosmer and Lemeshow, 1999); we do not care about the model fit to the data yet.

If AGE is the only explanatory variable, that will be labeled z, then a model for lifetime as a function of z may be expressed as follows:

$$X = e^{\beta_0+\beta_1 z} \times \varepsilon,$$

where

- X denotes survival time;
- ε follows the exponential distribution with parameter equal to one.

Note that this model is not linear in its parameters. However, it may be 'linearized' by taking the natural log. This yields the following model:

$$Y = \beta_0 + \beta_1 z + \theta,$$

where

$$\begin{cases} Y = \ln(X) \\ \theta = \ln(\varepsilon). \end{cases}$$

Here $\theta \sim G(0,1)$, meaning θ follows a extreme value distribution. The density function of the $G(0,1)$ is $f(\theta) = e^{[\theta - \exp(\theta)]}$, $\theta \in R$, and the survival function is $S(\theta) = e^{-\exp(\theta)}$.

Now we use MLE with an adaptation for censored data to estimate β_0 and β_1. Here, two variables are used to characterize a subject's time, the actual observed time, X, and a censoring indicator variable δ (δ=1 or 0). We denote the density function $f(x,\beta,z)$, the cumulative distribution function $F(x,\beta,z)$, and the survival function $S(x,\beta,z) = 1 - F(x,\beta,z)$. Under the assumption of independent observations, the full likelihood function is

$$L(\beta) = \prod_{i=1}^{n}\{[f(x_i,\beta,z_i)]^{\delta_i} \times [S(x_i,\beta,z_i)]^{1-\delta_i}\}, \qquad \text{where } \delta_i = 0 \text{ or } 1.$$

To obtain the maximized likelihood with respect to the parameter of interest, β, we maximize the log-likelihood function

$$l(\beta) = \sum_{i=1}^{n}\{\delta_i \ln[f(x_i,\beta,z_i)] + (1-\delta_i)\ln[s(x_i,\beta,z_i)]\}. \tag{1.3}$$

Because $y - (\beta_0 + \beta_1 z) = \theta$ and $\theta \sim G(0,1)$, we have $f(\theta) = e^{[\theta - \exp(\theta)]}$, $S(\theta) = e^{-\exp[\theta]}$, so that

$$S(y,\beta,z) = e^{-\exp[y-(\beta_0+\beta_1 z)]}, \tag{1.4}$$

$$f(y, \beta, z) = e^{\{y-(\beta_0+\beta_1 z)-\exp[y-(\beta_0+\beta_1 z)]\}}. \tag{1.5}$$

Substituting the expressions (1.4) and (1.5) into (1.3) yields the following log-likelihood:

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \delta_i \ln(e^{\{y_i-(\beta_0+\beta_1 z_i)-\exp[y_i-(\beta_0+\beta_1 z_i)]\}}) + (1-\delta_i)\ln(e^{-\exp[y_i-(\beta_0+\beta_1 z_i)]}) \\ &= \sum_{i=1}^{n} \delta_i[y_i-(\beta_0+\beta_1 z_i)] - e^{[y_i-(\beta_0+\beta_1 z_i)]}. \end{aligned} \tag{1.6}$$

In order to obtain the MLE of β, we must take the derivatives of the log-likelihood in (1.6) with respect to β_0 and β_1. The two score equations obtained to be solved are

$$\sum_{i=1}^{n} (\delta_i - e^{[y_i-(\beta_0+\beta_1 z_i)]}) = 0 \tag{1.7}$$

$$\sum_{i=1}^{n} z_i(\delta_i - e^{[y_i-(\beta_0+\beta_1 z_i)]}) = 0. \tag{1.8}$$

The equations (1.7) and (1.8) are nonlinear in β_0 and β_1 and must be solved using an iterative method. Many software packages can solve these equations numerically with Newton-Raphson or Fisher scoring algorithms (Kalbfleisch and Prentice (2002)) .

Next we discuss how to get estimates of the standard error of the estimated parameters in the column labeled “Std.err.” of Table 1.1. The negative of the second derivative of the log likelihood in (1.6) is called the observed information, and we will denote it as

$$I(\beta) = -\frac{\partial^2 l(\beta)}{\partial \beta_0 \partial \beta_1}. \tag{1.9}$$

The estimator of the variance of the estimated coefficient is the inverse of (1.9) evaluated at $\hat{\beta}$, i.e.,

$$\hat{Var}(\hat{\beta}) = I(\hat{\beta})^{-1}. \tag{1.10}$$

The estimator of the standard error, denoted $\hat{SE}(\hat{\beta}) = \sqrt{\hat{Var}(\hat{\beta})}$, is the positive square root of the variance estimator in (1.10).

We begin by presenting different tests to assess the significance of the coefficient: the partial likelihood ratio test and the Wald test.

The partial likelihood ratio test, denoted G, is calculated as twice the difference between the log partial likelihood of the model containing the covariate and the log partial likelihood for the model not containing the covariate. Specifically,

$$G = 2\{l(\hat{\beta}) - l(0)\}$$

Under the null hypothesis that the coefficient is equal to zero, this statistic will follow a chi-square distribution with 1 degree-of-freedom. This distribution can be used to obtain p-values to test the significance of the coefficient. Another test for significance of the coefficient can be computed from the ratio of the estimated coefficient to its estimated standard error. This ratio is commonly referred to as a Wald statistic. It follows a standard normal distribution under the null hypothesis that the coefficient is equal to zero. The equation for the Wald statistic is

$$z = \frac{\hat{\beta}}{\hat{SE}(\hat{\beta})}.$$

The endpoints of a 100(1-α) percent confidence interval for the coefficient are

$$\hat{\beta} \pm z_{1-\alpha/2}\hat{SE}(\hat{\beta}).$$

Some statistical packages (SAS) report the square of the Wald statistic, which follows a chi-square distribution with one degree-of-freedom.

Using SAS (SAS program is given in the Appendix), we get the results of Table 1.1 (Dataset 1.1).

Table 1.1: Estimators, standard errors, confidence intervals, Chi-Square, pr > ChiSq

variable	coeff.	std.err	Chi-Square	pr > ChiSq	95%	conf.int
AGE	-0.0941	0.0160	34.68	< .0001	-0.1254	-0.0628
constant	5.8607	0.5918	98.08	< .0001	4.7008	7.0206

The output in Table 1.1 shows that the maximum likelihood estimates of the two parameters are $\hat{\beta}_0 = 5.8607$, $\hat{SE}(\hat{\beta}_0) = 0.5918$, 95% confidence interval is (4.7008, 7.0206) and the estimate of variable AGE, $\hat{\beta}_1 = -0.0941$, $\hat{SE}(\hat{\beta}_1) = 0.0160$, 95% confidence interval is (-0.1254, -0.0628).

Chapter 2

Cox proportional hazards model

2.1 Proportional hazards regression model

The most common approach to model covariate effects on survival is the Cox proportional hazards model (Cox, 1972), which can handle censored and/or truncated observations. The Cox proportional hazards model has been probably the most important piece of work in the statistical analysis of survival data. We will look at it closely in this chapter.

The data, based on a sample of size n, consists of $(t_j, \delta_j, z_j), j = 1, 2...n$, where t_j is the time on study for the jth individual, δ_j is the event indicator ($\delta_j = 1$ if the event has occurred and $\delta_j = 0$ if the lifetime is right-censored) and z_j is the vector of covariates or risk factors for the jth individual (z_j may be also a function of time) that may affect the distribution of X, the time to event.

Let $h(t \mid z)$ be the hazard rate in the subpopulation with covariate value(s) z. The Cox proportional hazards regression model relates covariates to the hazard function as follows:

$$h(t|z) = h_0(t)c(\beta' z),$$

where $h_0(t)c(0)$ is the hazard function for the subpopulation with covariate value $z = 0$ and it is called the baseline hazard function, $\beta = (\beta_1, \beta_2, ...\beta_p)$ is a parameter vector of regression coefficients, $\beta' z = \sum_{i=1}^{p} \beta_k z_k$, and $c(\cdot)$ is a fixed, known scalar function. This is a semi-parametric model in which the baseline hazard, $h_0(t)$, is estimated non parametrically, while the covariate effect is constrained by the parametric representation

$c(\beta' z)$. Most commonly, an exponential form is used for c(·):

$$c(\beta' z) = \exp(\beta' z) = \exp(\sum_{k=1}^{p} \beta_k z_k) = e^{\sum_{k=1}^{p} \beta_k z_k},$$

which assures that the hazard is non-negative and assumes that covariate effects on the hazard are multiplicative. In this case, we have

$$h(t|z) = h_0(t)c(\beta' z) = h_0(t)\exp(\beta' z) = h_0(t)\exp(\sum_{k=1}^{p} \beta_k z_k).$$

The Cox model is often called a proportional hazards model because, if we look at two individuals with covariate values z_1 and z_2, the ratio of their hazard functions at time t is

$$\frac{h(t|z_1)}{h(t|z_2)} = \frac{h_0(t)\exp(\beta' z_1)}{h_0(t)\exp(\beta' z_2)} = \exp[(\beta'(z_1 - z_2)],$$

which is a constant (does not vary over time), that is, the ratio does not depend on t and the hazard rates are proportional, hence a proportional hazards model.

Recall that the hazard function at t given covariate z is $h(t|z) = h_0(t)e^{(\beta' z)}$. The cumulative hazard function, p.d.f. and survival functions given z can respectively be derived as follows:

$$\begin{aligned}
H(t|z) &= \int_0^t h(s|z)ds = \int_0^t h_0(s)e^{\beta' z}ds = H_0(t)e^{\beta' z},\\
S(t|z) &= \exp(-H(t|z)) = \exp(-H_0(t)e^{\beta' z}),\\
f(t|z) &= h_0(t)e^{\beta' z}\exp(-H_0(t)e^{\beta' z}).
\end{aligned}$$

2.2 The Cox proportional hazards partial likelihood

The contribution of Cox was to show how to efficiently estimate the parameters β when the functional form of $h_0(t)$ is unknown. The Cox methodology can be extended to more complicated models. For example, time dependent covariates $z(t)$ can be included in the model so that the characteristics of individuals are allowed to change through time.

2.2.1 Full likelihood

The full likelihood for n observations of (T_j, δ_j, z_j) under the proportional hazards model can be written in the usual fashion (Kalbfleisch and Prentice, 2002, Section 3.5):

$$L = \prod_{j=1}^{n} f_{x|z_j}(T_j)^{\delta_j} S_{x|z_j}(T_j)^{1-\delta_j} = \prod_{j=1}^{n} h_{x|z_j}(T_j)^{\delta_j} S_{x|z_j}(T_j),$$

where $f_{x|z}(t) = h_{x|z}(t)S_{x|z}(t)$. This likelihood could be maximized simultaneously in terms of the parameters β and $h_0(t)$. The Cox methodology uses a partial likelihood to yield estimates of β that are consistent and efficient regardless of the form of $h_0(t)$.

2.2.2 Cox partial likelihood

A sample of n subjects yields data with D distinct failure times, $t_1 < t_2 < ...t_D$, and $n - D$ censored times. Note that $t_{(i)}$ is an ordered event time, while T_j is the follow-up time for subject j. The set of indices of subjects at risk (alive and on study) at time t_i^- is denoted by $R_i = R(t_i)$. The covariate for the subject who has an event at time t_i is denoted by $Z_{(i)}$, to be distinguished from the covariate Z_j for subject j. [1] This allows the notation for all the subjects at risk at time t_i:

$$\{j \in R_i\} = \{j \mid T_j \geq t_i\}$$

Cox proposed the following ‘partial likelihood’ for the parameter β:

$$L(\beta) = \prod_{i=1}^{D} \frac{e^{(\beta' z_{(i)})}}{\sum_{j \in R_i} e^{(\beta' z_j)}},$$

where (i) denotes the subscript of the subject who dies at time $t_{(i)}$.

Remark 1

The Cox partial likelihood is a product over death times of

P_r(*subject (i) dies* | *one subject dies at time* $t_{(i)}$ *among subjects* $j\in R_i$ *at risk*)=

$$\frac{h(t_i|z_{(i)})}{\sum_{j\in R_i} h(t_i|z_j)} = \frac{e^{(\beta' z_{(i)})}}{\sum_{j\in R_i} e^{(\beta' z_j)}}.$$

[1] Here we assume no tied failure times to simplify the exposition. In the case of tied failure times, the notation and likelihood can be adjusted (Klein & Moeschberger, 2003, Section 8.4).

The log partial likelihood is

$$l(\beta) = \ln L(\beta) = \sum_{i=1}^{D} \left(\beta' z_{(i)} - \ln \left(\sum_{j \in R_i} e^{\beta' z_j} \right) \right).$$

The score functions are the first partial derivatives

$$U_h(\beta) = \frac{\partial \ln L}{\partial \beta_h} = \sum_{i=1}^{D} \left(z_{(i)h} - \frac{\sum_{j \in R_i} z_{jh} e^{\beta' z_j}}{\sum_{j \in R_i} e^{\beta' z_j}} \right),$$

for $h = 1, 2 \ldots p$. The maximum likelihood estimates satisfy

$$U_h(\hat{\beta}) = 0, \qquad h = 1, 2, ..., p.$$

The information matrix is the negative of the matrix of second derivatives of the log partial likelihood:

$$I_{gh}(\beta) = \frac{-\partial^2 \ln L}{\partial \beta_g \partial \beta_h} = \sum_{i=1}^{D} \left[\frac{\sum_{j \in R_i} z_{jg} z_{jh} e^{\beta' z_j}}{\sum_{j \in R_i} e^{\beta' z_j}} - \frac{(\sum_{j \in R_i} z_{jg} e^{\beta' z_j})(\sum_{j \in R_i} z_{jh} e^{\beta' z_j})}{(\sum_{j \in R_i} e^{\beta' z_j})^2} \right].$$

This matrix for $g = 1, 2 \ldots p$ and $h = 1, 2 \ldots p$ is a sum over $i = 1, 2 \ldots D$ of weighted covariance matrices for the z vector in the populations at risk at the time t_i.

2.3 Asymptotic distribution

For the majority of inference procedures (hypothesis tests, confidence intervals, forecasts) in survival analysis, we use approximations based either on the asymptotic distribution of $U(\hat{\beta})$, or on the asymptotic distribution of $\hat{\beta}$ (Duchesne, 2003, Section 0.2.3). Let β_0 be the true value of the parameter that we seek to estimate. In practice, we suppose that the sample size is sufficiently large to assume that

$$U(\beta_0) \approx N_q(0, I(\beta_0)), \tag{2.1}$$

where N_q indicates a multivariate normal distribution of dimension q, and

$$I(\beta_0) = -E \left[\frac{\partial}{\partial \beta'} U(\beta) \bigg|_{\beta = \beta_0} \right], \tag{2.2}$$

The asymptotic distribution of $\hat{\beta}$ is

$$\hat{\beta} \approx N_q(\beta_0, I^{-1}(\beta_0)), \tag{2.3}$$

or, equivalently,

$$\frac{\hat{\beta} - \beta_0}{\sqrt{I^{-1}(\beta_0)}} \sim N_q(0, 1).$$

Since the value of β_0 is not known in practice, formula (2.2) can sometimes be difficult to be evaluated. In practice, we use the fact that β_0 is well estimated by $\hat{\beta}$ and that $I(\beta_0)$ is well estimated by $\mathcal{I}(\hat{\beta})$ in calculations, where $\mathcal{I}(\beta)$ is the observed information matrix:

$$\mathcal{I}(\hat{\beta}) = -\left.\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta'}\right|_{\beta=\hat{\beta}}.$$

2.4 Hypothesis tests

The standard asymptotic likelihood inference tests, the Wald, score, and likelihood ratio tests, are also valid under the Cox partial likelihood to test hypotheses about β. Wald statistics are based on the asymptotic normality of the estimated regression coefficients according to formula (2.3). Likelihood ratio statistics are based on the log likelihood ratio for two nested models. Score statistics are based on the asymptotic normal distribution of the score function according to formula (2.1).

We are often interested in a hypothesis about a subset of the covariates. Generally, we partition the covariate vector $\beta = ({\beta_1}', {\beta_2}')'$ where β_1 is the $q \times 1$ subvector of coefficients of interest and β_2 is the $(p-q) \times 1$ vector of other covariate coefficients. Correspondingly, we get

$$\mathcal{I}(\beta) = (\mathcal{I}_{gh}(\beta))_{(p\times q)} = \left(-\frac{\partial^2 logL}{\partial\beta_g\partial\beta_h}\right)_{(p\times q)} = \begin{pmatrix} \mathcal{I}_{11}(\beta) & \mathcal{I}_{12}(\beta) \\ \mathcal{I}_{21}(\beta) & \mathcal{I}_{22}(\beta) \end{pmatrix}$$

and we let

$$\mathcal{I}^{-1}(\beta) = \begin{pmatrix} \mathcal{I}^{11}(\beta) & \mathcal{I}^{12}(\beta) \\ \mathcal{I}^{21}(\beta) & \mathcal{I}^{22}(\beta) \end{pmatrix} \text{be the partition of its inverse.} \tag{2.4}$$

Let $b_{(p\times 1)} = (b_1^{'}, b_2^{'})'$ be the partitioned maximum partial likelihood estimate for β. Consider tests about β_1 of the form $H_0 : \beta_1 = \beta_{01}$.

The estimator of the variance of the estimated coefficient is (2.4) evaluated at $\hat{\beta}$ and is $\hat{Var}(\hat{\beta}) = \mathcal{I}(\hat{\beta})^{-1}$. The estimator of the standard error, denoted $\hat{SE}(\hat{\beta}) = \sqrt{\hat{Var}(\hat{\beta})}$.

The Wald test statistic: ${X_w}^2 = (b_1 - \beta_{01})'[\mathcal{I}^{11}(b)]^{-1}(b_1 - \beta_{01})$.

Note that this statistic depends upon the entire vector b in the inverse information calculation.

The likelihood ratio test statistic: ${X_{LR}}^2 = 2\{l(b) - l[\beta_{01}, b_2(\beta_{01})]\}$, where $b_2(\beta_{01})$ is the maximum partial likelihood estimate of β_2 with β_1 fixed at β_{01}, $l(b)$ is the log partial likelihood function.

The score test statistic : ${X_{sc}}^2 = U_1[\beta_{01}, b_2(\beta_{01})]'[\mathcal{I}^{11}(\beta_{01}, b_2(\beta_{01}))]U_1[\beta_{01}, b_2(\beta_{01})]$, where $U_1[\beta_{01}, b_2(\beta_{01})]$ is the $(q \times 1)$ subvector of the score statistic of first partial derivatives of the log partial likelihood function.

Asymptotically, all three of these statistics have an approximate chi-square distribution with q degrees of freedom when the null hypothesis is true.

2.5 Time-dependent covariates

We can model the hazard function for an individual as a function of covariates whose values were fixed. These are explanatory variables recorded at the start of the study whose values are fixed throughout the course of the study. As is typical in many survival studies, individuals are monitored during the study, and other explanatory variables are recorded whose values may change during the course of the study. Such variables that change over time are called time-dependent variables (Klein and Moeschberger, 2003, Section 9.2).

Let X denote the time to some event and $z(t) = [z_1(t), z_2(t), \cdots, z_p(t)]'$ denote a set of covariates or risk factors at time t. Here the $z_k(t)$'s may be time-dependent covariates, whose value changes over time or they may be constant (or fixed) values known at time 0.

The basic model due to Cox (1972) is as with z replaced by $z(t)$, and for the commonly used model,

$$\begin{aligned} h[t \mid z(s), 0 \le s \le t] &= \lim_{\varepsilon \to 0} \frac{P\{t \le X < t + \varepsilon | X \ge t, z(s), 0 \le s \le t\}}{\varepsilon} \\ &= h_0(t) e^{\beta' z(t)} \\ &= h_0(t) \exp\left[\sum_{k=1}^{p} \beta_k z_k(t)\right]. \end{aligned} \tag{2.5}$$

In this section, our data, based on a sample of size n, consist of the triple

$$(T_j, \delta_j, \{z_j(t), 0 \le t \le T_j\}) \qquad j = 1, 2, \cdots, n,$$

where

- T_j: the time on study for the jth patient
- δ_j: the event indicator for the jth patient ($\delta_j = 1$ if event has occurred, 0 if the lifetime is right-censored).
- $z_j(t) = [z_{j1}(t), z_{j2}(t), \cdots, z_{jp}(t)]'$: the vector of covariates for the jth individual.

We assume that censoring is non informative: given $z_j(t)$, the event and censoring times for the jth patient are independent.

If the event times are distinct and $t_1 < t_2 < \cdots < t_D$ denotes the ordered event times, we define

$z_{(i)}(t_i)$: the covariate associated with the individual whose failure time is t_i;

$R(t_i)$: the risk set at time t_i.

The partial likelihood is given by

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp[\sum_{h=1}^{p} \beta_h z_{(i)h}(t_i)]}{\sum_{j \in R(t_i)} \exp[\sum_{h=1}^{p} \beta_h z_{jh}(t_i)]},$$

based on the hazard formulation (2.2). Note that in order to evaluate the likelihood contribution at t_i, we need to know the values of the covariates at that time for everyone in $R(t_i)$. Estimation and testing may proceed with the appropriate alterations of z to $z(t)$ in the inference procedures described in sections 2.2 to 2.4.

Example:

In the book by Klein and Moeschberger (2003, Section 9.2), a study of acute leukemia patients being given a bone marrow transplant is presented. Bone marrow transplants are a standard treatment for acute leukemia. Prediction for recovery may depend on risk factors known at the time of transplantation, such as patient and/or donor age and sex, the stage of initial disease, the time from diagnosis to transplantation, etc. The final prediction may change as the patient's post transplantation history develops with the occurrence of events at random times during the recovery process, such as development of acute or chronic graft-versus-host disease (GVHD), return of the platelet count to normal levels, return of granulocytes to normal levels, or development of infections. Transplantation can be considered a failure when a patient's leukemia returns (relapse) or when he or she dies while in remission (treatment related death). The variable time t_2 denotes the disease-free-survival time and the event indicator for disease-free-survival is d_3.

There are three risk groups: acute lymphoblastic leukemia (ALL), low-risk acute myeloctic (AML low-risk), and high-risk acute myeloctic leukemia (AML high-risk). We define two binary covariates (AMLL = 1 if AML low-risk, AMLH = 1 if AML high-risk) for the factor of interest. There are many other fixed factors.

In addition to the covariates fixed at the time of transplant, there are three intermediate events that occur during the transplant recovery process that may be related to the disease-free survival time of a patient. These are the development of acute graft-versus-host disease (aGVHD), the development of chronic graft-versus-host disease (cGVHD) and the return of the patients platelet count to a self-sustaining level (platelet recovery). The timing of these events, if they occur, is random. In this example, we shall examine their relationship to the disease-survival time and see how the effects of the fixed covariates change when these intermediate events occur. Each of these time-dependent variables may be coded as an indicator variable whose value changes from 0 to 1 at the time of the occurrence of the intermediate event. We define the covariates as follows:

$$Z_A(t) = \begin{cases} 0 & \text{if} \quad t < \text{time at which acute graft-versus-host disease occurs,} \\ 1 & \text{if} \quad t >= \text{time at which acute graft-versus-host disease occurs.} \end{cases}$$

$$Z_P(t) = \begin{cases} 0 & \text{if} \quad t < \text{time at which the platelet recovered,} \\ 1 & \text{if} \quad t >= \text{time at which the platelet recovered.} \end{cases}$$

$$Z_C(t) = \begin{cases} 0 & \text{if} \quad t < \text{time at which chronic graft-versus-host disease occurs,} \\ 1 & \text{if} \quad t >= \text{time at which chronic graft-versus-host disease occurs.} \end{cases}$$

Local tests may be performed to assess the significance for each time-dependent covariate in a model that already has covariates for the two risk groups included (AMLL and AMLH). We fit a separate Cox model for each of the three intermediate events which includes the disease factor AMLL and AMLH, and form three Cox models: Model 1 (AMLL, AMLH, $Z_A(t)$), Model 2 (AMLL, AMLH, $Z_C(t)$), Model 3 (AMLL, AMLH, $Z_P(t)$). We use SAS (SAS program is given in Appendix 2.1) to get results from tables 2.1-2.3 below.

Table 2.1: Parameter Estimate, Standard Errors, Chi-Square, pr > ChiSq, Hazard Ratio for model 1

variable	coeff.	std.err	Chi-Square	pr > ChiSq	Hazard Ratio
AMLL	-0.55164	0.28799	3.6690	0.0554	0.576
AMLH	0.43381	0.27222	2.5396	0.1110	1.543
Z_A	0.31836	0.28514	1.2466	0.2642	1.375

Table 2.2: Parameter Estimate, Standard Errors, Chi-Square, pr > ChiSq, Hazard Ratio for model 2

variable	coeff.	std.err	Chi-Square	pr > ChiSq	Hazard Ratio
AMLL	-0.62251	0.29622	4.4163	0.0356	0.537
AMLH	0.36567	0.26850	1.8548	0.1732	1.441
Z_C	-0.19478	0.28757	0.4587	0.4982	0.823

Table 2.3: Parameter Estimate, Standard Errors, Chi-Square, pr > ChiSq, Hazard Ratio for model 3

variable	coeff.	std.err	Chi-Square	pr > ChiSq	Hazard Ratio
AMLL	-0.49624	0.28924	2.9435	0.0862	0.609
AMLH	0.38134	0.26761	2.0306	0.1542	1.464
Z_P	-1.12986	0.32800	11.8658	0.0006	0.323

Here, we see that only the return to a sustaining level of the platelets has a significant impact on disease-free survival. The negative value of coefficients suggests that a patient who has the intermediate event ($Z_C(t) = 1$ or $Z_P(t) = 1$) has a better chance of survival than a patient who, at that time, has yet to have these event.

2.6 Regression diagnostics

In this section, we will discuss methods to check the various assumptions of a proportional hazards (PH) model.

2.6.1 Cox-Snell residuals for assessing the overall fit of a proportional hazards model

Suppose X follows the Cox model with covariate z and the regression coefficient β is known. Then the survival probability of X is

$$S(X) = \exp(-H_0(X)e^{\beta' z}).$$

Note that $S(X) = 1 - F(X) \sim U(0,1)$, and $H(X) = -\ln S(X) = -\ln(1 - F(X))$, so $-\ln S(X) \sim Exp(1)$. Thus, $H_0(X)e^{\beta' z} \sim Exp(1)$.

In practice, of course, we don't know H_0 and β, but they can be estimated. If the estimates of the β's from the postulated model are b = $(b_1, ...b_p)'$, then, the Cox-Snell residuals are defined as

$$r_j = \hat{H}_0(T_j)\exp(\hat{\beta}' z_j), \qquad j = 1, 2, \cdots n.$$

Here, $\hat{H}_0(t)$ is Breslow's estimator of the baseline hazard rate (Klein and Moeschberger, 2003, Section 8.8). Let $t_1 < t_2 < \cdots$ denote the distinct death times. Define

$$W(t_i; b) = \sum_{j \in R(t_i)} \exp(b' z_j).$$

The estimator of the cumulative baseline hazard rate $H_0(t) = \int_0^t h_0(u)du$ is given by

$$\hat{H}_0(t) = \sum_{t_i \leq t} \frac{1}{W(t_i; b)},$$

which is a step function with jumps at the observed death times.

If the final PH model is correct and the $\hat{\beta}_j$'s are close to the true values of the β_j's, the r_j's should resemble a censored sample from a unit exponential distribution. To check whether the r_j's resemble a censored sample from a unit exponential, the plot of $\hat{H}_r(r_j)$ against r_j, where $\hat{H}_r$ is the Nelson-Aalen estimator, should roughly be a 45° line through the origin. Ideally, the plot of $\hat{H}_r(r_j)$ against r_j should include a confidence band so that significance can be addressed. Unfortunately, the r_j are not exactly a censored sample from a distribution. So this plot is generally used only as a rough diagnostic.

2.6.2 Martingale residuals for identifying the best functional form of a covariate

As before, X follows the Cox model with covariate z and the regression coefficient β is known. We observe (t_j, δ_j, z_j), for j=1,2...n. Under the assumption that the proportional hazards model is true, the process

$$M_j(t) = N_j(t) - \int_0^t Y_j(s) e^{z_j(s)\beta} dH_0(s)$$

is a martingale for $j = 1, 2...n$, so we expect $\mathrm{E}[M_i(t)] = 0, t \geq 0$. Define the martingale residuals

$$M_j = N_j(\infty) - \int_0^\infty Y_j(s) e^{z_j(s)\beta} dH_0(s), \qquad j = 1, 2...n$$

In practice, to compute the martingale residuals, β and H_0 will be replaced with $\hat{\beta}$ and $\hat{H}_0$. We thus get

$$\hat{M_j} = N_j(\infty) - \int_0^\infty Y_j(s) e^{z_j(s)\hat{\beta}} d\hat{H_0}(s) = \delta_j - r_j,$$

where r_j is the Cox-Snell residual.

Now we shall use these martingale residuals to examine the best functional form for a given covariate using an assumed Cox model for the remaining covariates. Suppose that the covariate vector z is partitioned into a vector z^*, for which we know the proper functional form of the Cox model, and a single covariate z_1 for which we are unsure of what functional form to use. Let $f(z_1)$ be the best function for z_1 to explain its effect on survival. Then

$$H(t \mid z^*, z_1) = H_0(t)\exp(\beta^* z^*)\exp[f(z_1)].$$

To find f, we fit a Cox model to the data based on z^* and compute the martingale residuals, $\hat{M_j}$, j=1,2...n. These residuals are plotted against the values of z_1. If the plot is linear, no transformation of z_1 is needed. If there appears to be a threshold, then, a discretized version of the covariate is indicated. If the plot is not linear nor threshold, then we should use a transform such as log, squared root or power of the variable z_1.

2.6.3 Schoenfeld residuals to examine the fit and detect outlying covariate values.

The kth Schoenfeld residuals defined for the kth subject on the explanatory variable $x^{(j)}$, $j = 1, 2, \cdots, p$, is given by

$$r_{s_{jk}} = \delta_k\{x_k^{(j)} - a_k^{(j)}\}, \qquad k = 1, \ldots n, \qquad j = 1, \ldots p,$$

where

- δ_k is the kth subject's censoring indicator,
- $x_k^{(j)}$ is the value of the jth explanatory variable for the kth individual in the study,
- $a_k^{(j)} = \frac{\sum_{m\in R(y_k)} \exp(x'_m\hat{\beta})x_m^{(j)}}{\sum_{m\in R(y_k)} \exp(x'_m\hat{\beta})}$ and $R(y_k)$ is the risk set at time y_k,
- the dimension of $r_{s_{jk}}$ is $p \times n$.

If the assumption of proportional hazards holds, large Schoenfeld residuals are not expected to appear at late failure times. It means if PH assumption is satisfied, a plot of these residuals against ordered death times should look like a tied down random walk. Starting at 0 at time 0 and ending at 0 at time τ.

2.6.4 Scaled Schoenfeld residuals to test proportional hazard assumption.

The S-plus default for checking the proportional hazards assumption is a formal test and plot of scaled Schoenfeld residuals. The Schoenfeld residual is the difference between the covariate at the failure time and the expected value of the covariate at this time. As an alternative to proportional hazards, Therneau and Grambsch (2000) consider time-varying coefficients $\beta(t) = \beta + \theta * g(t)$, for some smooth function g. The function

$g(t)$ used here is the S-plus default, $g(t) = 1 - S(t)$. Given $g(t)$, they develop a score test for $H_0 : \theta = 0$ based on a generalized least squares estimator for θ.

Defining *scaled Schoenfeld residuals* by the product of the inverse of the estimated variance-covariance matrix of the kth Schoenfeld residual and the kth Schoenfeld residual, they show the kth scaled Schoenfeld residual has approximately mean $\theta g(t_k)$ and the kth Schoenfeld residual has an easily computable variance-covariance matrix. Motivated by these results, they also develop a graphical method. They show by Monte Carlo simulation studies that a smoothed scatter plot of $\hat{\beta}(t_k)$, the kth scaled Schoenfeld residual plus $\hat{\beta}$, versus t_k reveals the functional form of $\beta(t)$. Under H_0, we expect to see a constant function over time. Both of these can be easily done with the S functions *cox.zph* and *plot*. When the proportional hazards assumption holds, a relatively straight horizontal line is expected.

2.6.5 Dfbetas to assess influence of each observation

To investigate influence of observation j on the regression coefficient estimates, one can estimate the difference in β with and without observation j, i.e., $\hat{\beta} - \widehat{\beta_{(j)}}$, where $\widehat{\beta_{(j)}}$ is the estimate of β without observation j. If $\hat{\beta} - \widehat{\beta_{(j)}}$ is close to zero, the jth observation has little influence on the estimate, whereas large deviations suggest a large influence.

2.7 Example 1: CNS Lymphoma

The CNS lymphoma data is listed in Appendix Dataset 2.1. In this example we check the adequacy of the PH model (Tableman and Kim, 2004). The data result from an observational clinical study. It contained 7 variables for each patient:

NUMBER: patient number;

GROUP: 1=prior radiation; 0=no prior radiation with respect to 1st blood brain-barrier disruption (BBBD) procedure to deliver chemotherapy;

SEX: 1=female; 0=male;

AGE: at time of 1st BBBD, recorded in years;

STATUS: 1=dead; 0=alive;

DEATHTIME: time from 1st BBBBD to death in year;

SCORE: Karnofsky performance score before 1st BBBD, numerical value 0-100.

The full model for DEATHTIME in this situation is

$$h(t|z) = h_0(t)\exp(\beta' z) = h_0(t)\exp(\beta_1\text{SCORE} + \beta_2\text{GROUP} + \beta_3\text{SEX} + \beta_4\text{AGE}).$$

Next, we go through model diagnostics to confirm whether or not the Cox model does fit the data.

1. Cox-Snell residuals (Splus program given in Appendix 2.2).

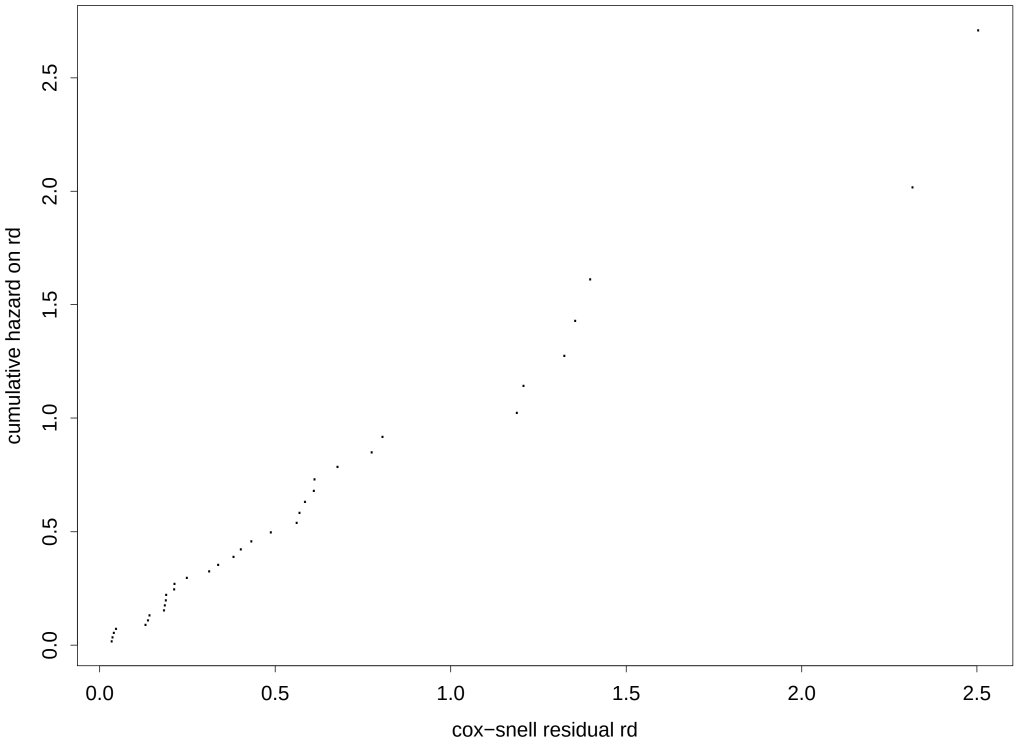

Figure 2.1: Cox-Snell residuals to asses model fit

Result:

We see from the Cox-Snell residual plot (Figure 2.1) that the model gives a reasonable fit to the data. Overall the residuals fall on a straight line with an intercept zero and a slope one. Further, there are no large departures from the straight line and no large variation at the right-hand tail.

2. The martingale residual plot to check functional form of the covariate SCORE follows (the Splus program is given in Appendix 2.3).

   Suppose that the covariate vector z is partitioned into a vector z^* (GROUP, SEX, AGE), for which we know the functional form, and a single continuous covariate z_1 (SCORE) for which we are unsure of what functional form to use. Let $f(z_1)$ denote the best function for z_1 to explain its effect on survival. Then the model fitted to obtain the martingale residuals is

$$H(t \mid z^*, z_1) = H_0(t)\exp(\beta^* z^*)\exp[f(z_1)].$$

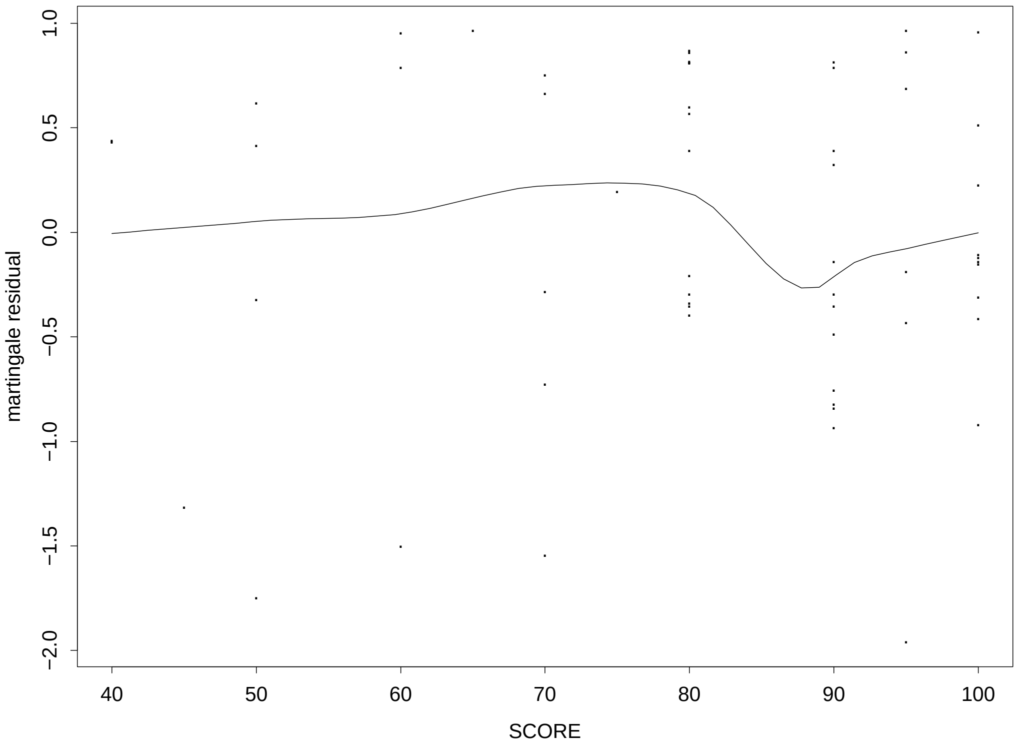

Figure 2.2: Martingale residuals to check functional form of the continuous variable SCORE

Result:

In the plot of the martingale residuals, there appears to be a bump for SCORE between 80 and 90. The lines before and after the bump nearly coincide. Therefore, a linear form seems appropriate for SCORE, but from Figure 2.2, we see that there appears to be a discrete time point where the slope changes. A dichotomized transformation of the variable SCORE may be indicated. Later we will consider this option.

3. Schoenfeld residuals to examine the fit and detect outlying covariate values (Splus program given in Appendix 2.4).

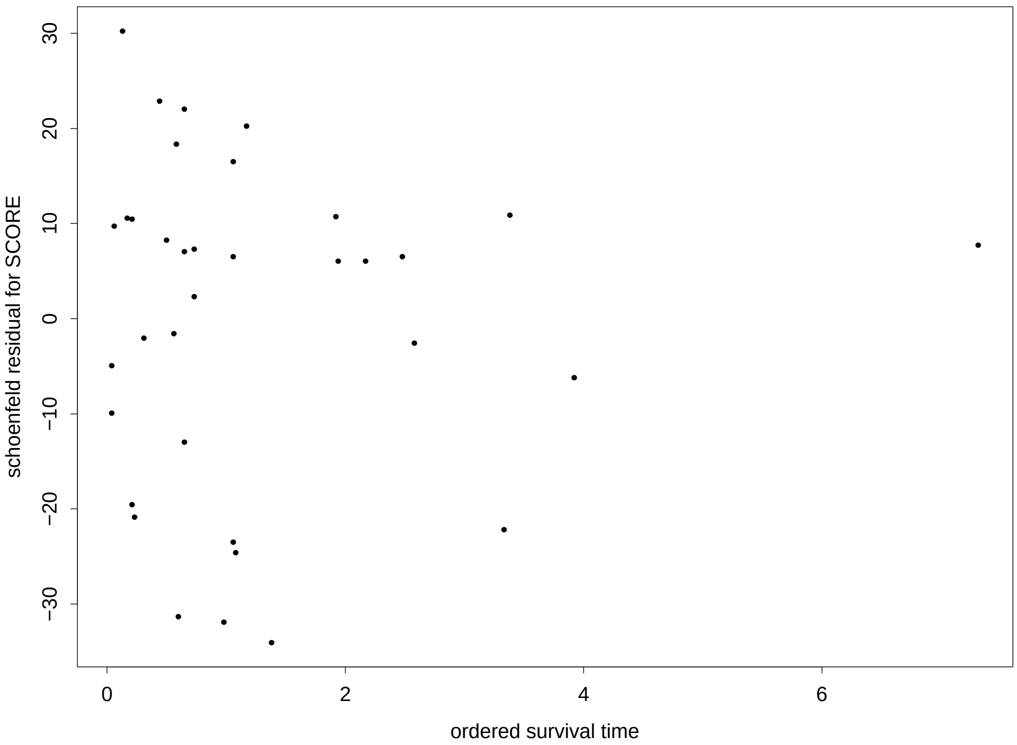

Figure 2.3: Schoenfeld residuals for SCORE against ordered survival time

Result:

From the Schoenfeld residuals plot (Figure 2.3), we find that the subjects with the large absolute valued Schoenfeld residuals for SCORE have very early failure times. Thus, these residuals do not cause specific concern. Therefore, the PH assumption seems to be appropriate.

4. The scaled Schoenfeld residuals and the Grambsch and Therneau's test for time-varying coefficients to assess PH assumption (Splus program given in Appendix 2.5).

The cox.zph tests for the proportional hazards assumption are obtained for each covariate, along with a global test for the model as a whole.

Result:

	rcho	chisq	p
SCORE	0.0218	0.0126	0.911
GROUP	0.1446	0.7782	0.378
SEX	0.1999	1.5761	0.209
AGE	-0.0486	0.0944	0.759
GLOBAL	NA	2.8853	0.577

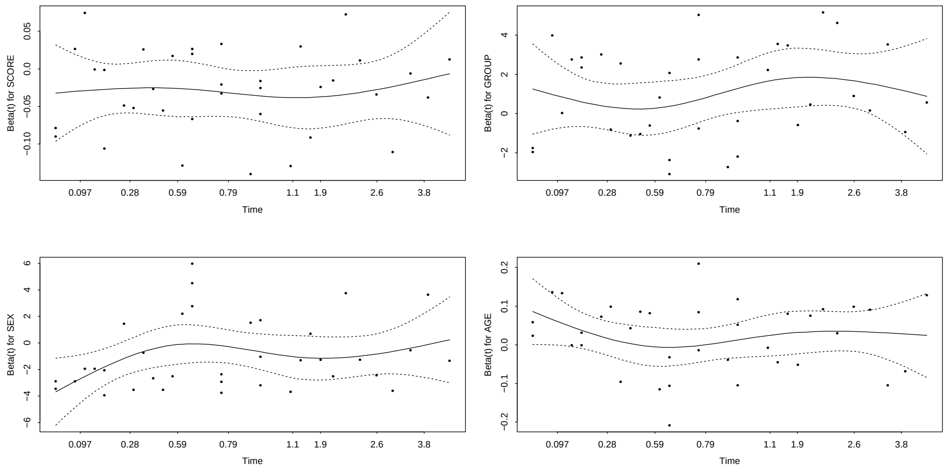

Figure 2.4: Plots of scaled Schoenfeld residuals against ordered time for each covariate in a model fit to the CNS lymphoma data.

The results from the test for constancy of the coefficients based on scaled Schoenfeld residuals indicate the PH assumption is satisfied by all four covariates in the model with all p-values being at least 0.209. Figure 2.4, also supports that the PH assumption is satisfied for all the covariates in the model.

5. The dfbetas to assess influence of each observation (Splus program given in Appendix 2.6).

   Result:

   The plot of the dfbetas, Figure 2.5, shows that the change in the regression coefficients are less than 0.4. Therefore, we conclude that there are no influential subjects.

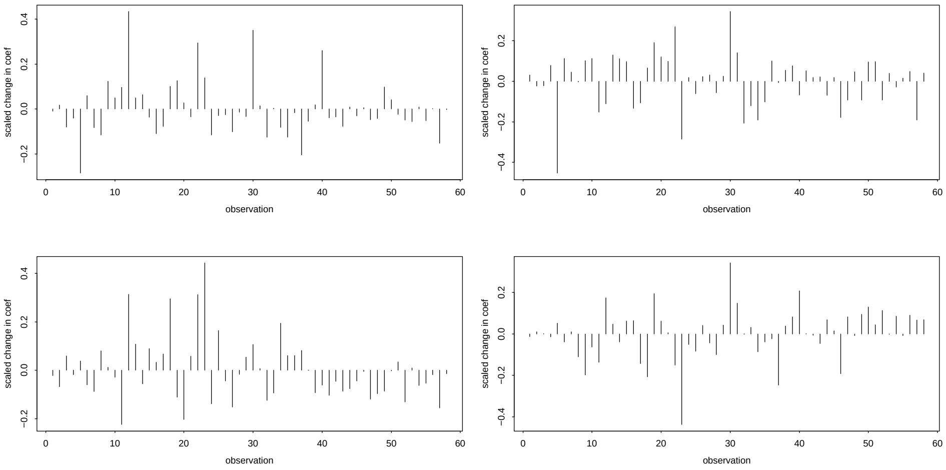

Figure 2.5: The dfbetas to detect influential observations

Conclusion:

Through the model diagnostics, we find that the model considered fits the data very well.

2.8 Example 2: Inmate data

The file name.txt (data listed in Appendix Dataset 2.2) contains information on 432 inmates who were released from state prisons in the early 1970s. The aim of this study was to determine the efficacy of financial aid to released inmates as a means of reducing recividism (Allison, 1995). Half the inmates were randomly assigned to receive financial aid. They were followed for one year after their release and were interviewed monthly during that period. The dataset used here contains the following 9 variables: week, arrest, fin, age, race, wexp, mar, paro, prio:

- week is the week of first arrest, week has a value of 52 if not arrested;
- arrest has a value of 1 if arrested, otherwise arrest has a value of 0;

- fin has a value of 1 if the inmate received financial aid after release, otherwise, fin has a value of 0. fin is randomly assigned, with equal numbers in each category;
- age is the age in years at the time of release;
- race has a value of 1 if the inmate is black, otherwise race has a value of 0;
- wexp has a value of 1 if the inmate has full-time work experience before incarceration, otherwise wexp has a value of 0;
- mar has a value of 1 if the inmate was married at the time of release, otherwise mar has a value of 0;
- paro has a value of 1 if released on parole, otherwise paro has a value of 0;
- prio is the number of convictions prior to current incarceration.

Now we study a Cox regression of time to rearrest with the constant time covariates specified as follows (Splus program is given in Appendix 2.7). The fitted model is summarized as follows:

	coef	exp(coef)	se(coef)	z	p	lower .95	upper .95
fin	-0.3794	0.684	0.1914	-1.983	0.0470	0.470	0.996
age	-0.0574	0.944	0.0220	-2.611	0.0090	0.904	0.986
race	0.3139	1.369	0.3080	1.019	0.3100	0.748	2.503
wexp	-0.1498	0.861	0.2122	-0.706	0.4800	0.568	1.305
mar	-0.4337	0.648	0.3819	-1.136	0.2600	0.307	1.370
paro	-0.0849	0.919	0.1958	-0.434	0.6600	0.626	1.348
prio	0.0915	1.096	0.0286	3.195	0.0014	1.036	1.159

Likelihood ratio test = 33.3 on 7 df, p=2.36e-05

Wald test = 32.1 on 7 df, p=3.86e-05

score (logrank) test = 33.5 on 7 df, p=2.11e-05

From the results, we see that:

1. The covariates age and prio have highly statistically significant coefficients, while the coefficient for fin is marginally significant(p-value = 0.047). When we do a backward elimination, we get the same result. So thereafter, we study the model that only includes the three significant variables age, prio and fin.

2. The exponentiated coefficients are interpretable as multiplicative effects on the hazard. For example, holding the other covariates constant, an additional year of age reduces the weekly hazard of rearrest by a factor of $e^{-0.0574} = 0.0944$ on average, that is, by 5.6 percent. Similarly, each prior increases the hazard by a factor of $e^{0.0915} = 1.096$, or 9.6 percent.

3. The likelihood-ratio, Wald and score tests are asymptotically equivalent tests of the null hypothesis that all of the β's are zero. In this example, the test statistics are in close agreement, and the hypothesis is soundly rejected.

- Model diagnostics

  As is the case for a linear or generalized linear model, it is desirable to determine whether a fitted Cox regression model adequately describes the data. We will briefly consider three kinds of diagnostics: checking proportional hazards, influential data, and nonlinearity (Therneau and Grambsch, 2000, Section 6 and 7). All of these diagnostics use different types of the residuals.

  1. Checking the proportional hazards assumption (Splus program is given in Appendix 2.8).

     Test and graphical diagnostics for the proportional hazards assumption may be based on the scaled Schoenfeld residuals. More conveniently, the cox.zph function, details are in 2.6.4 or (Therneau and Foundation, 1999) calculates tests of the proportional hazards assumption for each covariate by correlating the corresponding set of scaled Schoenfeld residuals with ordered time. There is strong evidence for nonproportionality as shown by the large global test statistic.

     Now we eliminate the covariates whose coefficients were not statistically significant.

	coef	exp(coef)	se(coef)	z	p
fin	-0.3469	0.707	0.1902	-1.82	0.06800
age	-0.0671	0.935	0.0209	-3.22	0.00130
prio	0.0969	1.102	0.0273	3.56	0.00038

Likelihood ratio test = 29.1 on 3 df, p=2.19e-06 n=432

Note:

The coefficient for financial aid is the focus of the study. It now has a two-sided p-value of 0.0608, so there is still marginal evidence for the effect of this covariate on the time of rearrest. The negative coefficient (-0.3469) is what we expected, as financial aid is supposed to reduce the risk of recidivism.

The cox.zph tests for the proportional hazards assumption are obtained for each covariate, along with a global test for the model as a whole.

	rho	chisq	p
fin	-0.00657	0.00507	0.9432
age	-0.20976	6.54118	0.0105
prio	-0.08003	0.77263	0.3794
GLOBAL	NA	7.12999	0.0679

Therefore, there is strong evidence of non-proportional hazard for age, while the global test is not quite statistically significant. One way of accommodating non-proportional hazards is to build interactions between covariates and time into the Cox regression model; such interactions are themselves time-dependent covariates, here we don't study this case. Next we plot graphs of the scaled Schoenfeld residuals (Figure 2.6) against ordered time.

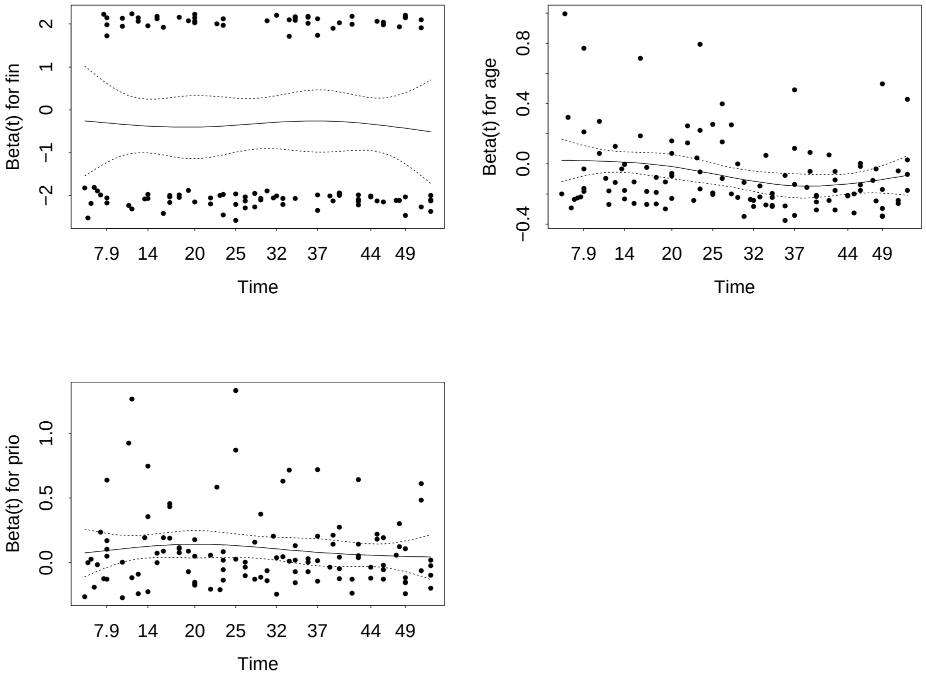

Figure 2.6: Plots of scaled Shoenfeld residual against ordered time

Interpretation of these graphs is greatly facilitated by smoothing, for which cox.zph uses a smoothing spline, shown on each graph by a solid line. The broken lines represent ±2-standard-error envelopes around the fit. Systematic departures from a horizontal line are indicative of non-proportional hazards. The assumption of proportional hazards appears to be supported for the covariates fin and prio, but there appears to be a little trend in the plot for age, with the age effect declining with time. This effect was detected in the test reported above.

2. Influential observations, for the model regressing time to rearrest on financial aid, age and prior. (Splus program given in Appendix 2.9)

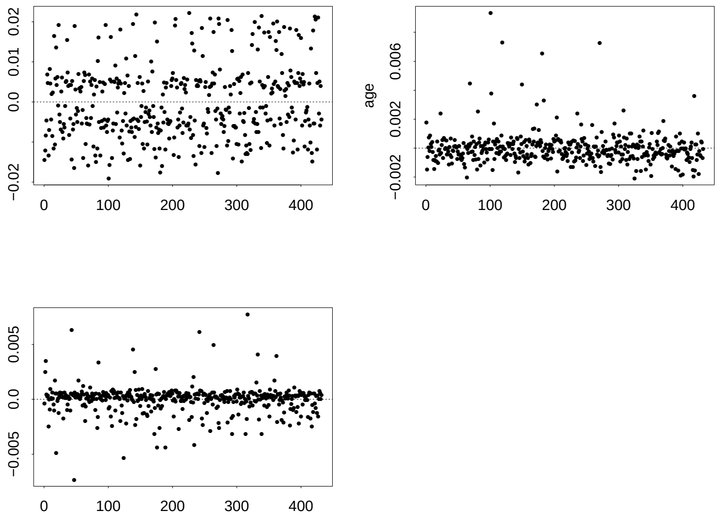

Figure 2.7: Schoenfeld residuals for fin, age, prio against ordered survival time

Comparing the magnitudes of the largest dfbeta values to the regression coefficients in Figure 2.7 suggests that none of the observations is terribly influential individually (even though some of the dfbeta values for age are large compared with the others).

3. Nonlinearity, that is, an incorrectly specified functional form in the parametric part of the model. The martingale residuals may be plotted against covariates to detect nonlinearity. (Splus program is given in Appendix 2.10)

For the regression of time to rearrest on financial aid, age, and number of prior arrests, let us examine the plots of martingale residuals against the last two of these covariates: Nonlinearity is not an issue for financial aid, because this covariate is dichotomous.

The resulting residual plots appear in Figure 2.8. The smooths in Figure 2.8 are produced by local linear regression (using the lowess function). There is no evidence of nonlinearity here.

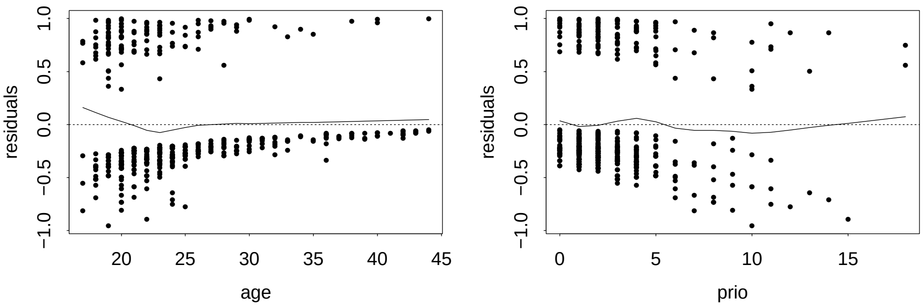

Figure 2.8: Martingale residuals to check the functional form of the continuous variables age and prio.

Chapter 3

Additive hazards regression models

One of the main purposes in survival analysis is to investigate the effects of risk factors on disease occurrence or death. For this purpose, two models are predominantly considered, the proportional hazards model and the additive risk model. The log-linear (or accelerated life) model is also widely used (Klein & Moeschberger, 2003, Section 12). The proportional hazards model assumes multiplicative effects of risk factors on the hazard function while the additive risk model assumes that the hazard function associated with a set of covariates is the sum of a baseline hazard function and a regression function of covariates. The proportional hazards model has been more popular than the additive risk model. As we showed in detail in the previous chapter, the multiplicative model is the major framework for regression analysis of survival data, it is extremely useful in practice since the estimated coefficients themselves or simple functions of them can be used to provide estimates of hazard ratios. In addition, statistical software is readily available and easy to use to fit models, check model assumptions and assess model fit. However, the additive risk model is useful when risk difference, rather than relative risk, is of main interest. Moreover, the additive risk model allows covariate effects to vary with time.

The proportional hazards model has been studied by many authors since Cox (1972), and the additive risk model has been considered by Aalen (1980, 1989). In this thesis, we mainly study Aalen's nonparametric additive hazards model, who discusses issues of estimation, testing and assessment of model fit. His model is fully additive and nonparametric and values of the regression coefficients are allowed to vary over time. We present his model in more detail than the other additive models as it provides the opportunity to not only fit an additive model, but the results of the fit can be used to provide graphical descriptions that supplement fits of other models, such as the proportional hazards model. Even though there are many advantages in using the additive hazards model, it is not widely used. One reason for this is that the model is not available in any commonly used computer packages, such as SAS, S-PLUS. Presented here is a SAS macro that fits the additive hazards regression.

In Chapter 2, we discussed the proportional hazards model, where the estimation of the risk coefficients was based on the partial likelihood. In this model, these risk coefficients were unknown constants whose value did not change over time. In this chapter, we present an alternative model based on assuming that the covariates act in an additive manner on an unknown baseline hazard rate. The unknown risk coefficients in this model are allowed to be functions of time so that the effect of a covariate may vary over time. As opposed to the proportional hazards model where likelihood based estimation techniques are used, estimators of the risk coefficients are based on a least-squares technique.

3.1 Description of Aalen's additive regression model

A number of individuals are observed over time to see if a specified event occurs. The individuals are assumed to be independent and any events happening to the individuals are also assumed to be independent between individuals. The lifetime we observe may be right-censored.

In this section, our data, based on a sample of size n, consist of the triple

$$[T_j, \delta_j, [z_j(t), 0 \leq t \leq T_j]], \qquad j = 1, 2, \cdots, n,$$

where

- T_j: the time on study for the jth patient;
- δ_j: the event indicator for the jth patient ($\delta_j = 1$ if event has occurred, 0 if the lifetime is right-censored);
- $z_j(t) = [z_{j1}(t), z_{j2}(t), \cdots, z_{jp}(t)]$ is a p-vector of, possibly, time-dependent covariates.

For the jth individual we define

$$Y_j(t) = \begin{cases} 1 \text{ if individual } j \text{ is under observation (at risk) at time } t^- \\ 0 \text{ if individual } j \text{ is not under observation (not at risk) at time } t^- \end{cases}$$

For right-censored data, $Y_j(t)$ is 1 if $t \leq T_j$.

For the jth individual, the conditional hazard rate at time t, given $z_j(t)$, can be modeled by the following linear model:

$$h[t|z_j(t)] = \beta_0(t) + \sum_{k=1}^{p} \beta_k(t) z_{jk}(t).$$

The hazard at any time is thus a sum of a baseline hazard, $\beta_0(t)$, and a linear combination of the covariate values, $z_j(t)$. The coefficients $\beta_k(t)$, $k = 1, 2, \cdots, p$, are unknown regression functions to be estimated. These functions measure the influence of the respective covariates. Because regression functions may vary with time, their analysis may reveal changes in the influence of the covariates over time, which is one of the main advantages of the additive model. This model is non-parametric in the sense that no assumption is made about the functional forms of the regression functions.

Estimation of the risk coefficients is based on a least-squares technique (Huffler and McKeague, 1991). This differs from estimation in the proportional hazards model which is based on a partial or conditional likelihood. In fact, direct estimation of the $\beta_k(t)$ is difficult. It is much easier to estimate the cumulative regression functions $B_k(t)$ (see equation (3.1) below) than the regression functions themselves. The column vector $B(t)$, with elements $B_k(t)$, $k = 1, 2, \cdots, p$, will be estimated, where

$$B_k(t) = \int_0^t \beta_k(s) ds \qquad k = 0, 1, 2, \cdots, p. \tag{3.1}$$

To obtain the estimates, we first compute the $n \times (p + 1)$ matrix $X(t)$ which is defined as follows: for the ith row of $X(t)$, we set $X_i(t) = Y_i(t)(1, Z_j(t))$. That is, if the ith individual is a member of the risk set at time t (event has not happened and the individual is not censored), then the ith row of $X(t)$ is the vector $X_i(t) = (1, z_{j1}(t), \cdots, z_{jp}(t))$. If the ith individual is not in the risk set at time t, i.e, the event of interest has already occurred or the individual has been censored, then the corresponding row of $X(t)$ contains only zeros.

Let $T_1 < T_2 < \cdots$ be the ordered observed times when at least one event occurs. The least-squares estimate of the vector $B(t) = (B_0(t), B_1(t), \cdots B_p(t))'$, given in Klein and Moeschberger (2003, section 10.2), is

$$\hat{B}(t) = \sum_{T_i \leq t} [X^{'}(T_i)X(T_i)]^{-1} X^{'}(T_i) I(T_i) = \sum_{T_i \leq t} V(T_i) I(T_i), \tag{3.2}$$

where $V(T_i) = [X'(T_i)X(T_i)]^{-1}X'(T_i)$ and $I(T_i)$ is the $n \times 1$ vector with ith element equal to 1 if subject i experiences an event at time T_i and 0 otherwise.

It should be noted that the estimator $\hat{B}(t)$ is well defined as long as $X(t)$ has full rank and, therefore, $X'X$ is invertible. Therefore, estimates are restricted to the time interval where X is not singular. Also, the estimates of the baseline hazard rate are not constrained to be non-negative.

From equation (3.1), we know that the estimators $\hat{B}_k(t)$ estimate the integral of the regression function $\beta_k(t)$. A crude estimate of $\beta_k(t)$ is given by the slope of our estimate of $B_k(t)$. Better estimates of $\beta_k(t)$ can be obtained by using a kernel-smoothing technique, which we do not pursue here.

The main focus of the analysis in the additive risk model is on cumulative regression plots; the cumulative regression functions are plotted against time and give a description of how the covariates influence the survival over time. The slope of the plot of the cumulative regression function against time gives information on whether the particular covariate has a constant or time dependent effect. Positive slopes occur during periods when increasing covariate values are associated with increases in the hazard function. Negative slopes occur during periods when increasing covariate values are associated with decreases in the hazard function. The cumulative regression function will have roughly zero slope during periods when the covariate has no effect on the hazard.

The following estimator of the covariance matrix of $\hat{B}(t)$ is used:

$$\begin{aligned} COV = \hat{\text{var}}(\hat{B}(t)) &= \sum_{T_i \leq t} [X'(T_i)X(T_i)]^{-1}X'(T_i)I^D(T_i)X(T_i)\{[X'(T_i)X(T_i)]^{-1}\}' \\ &= \sum_{T_i \leq t} V(T_i)I^D(T_i)V(T_i)', \end{aligned}$$

where $I^D(T_i)$ is the diagonal matrix with diagonal elements equal to $I(T_i)$. Confidence intervals for $B(t)$ can be constructed in the usual fashion:

$$\hat{B}_j(t) \pm Z_{1-\alpha/2}[\hat{\text{var}}(\hat{B}_j(t))]^{1/2}.$$

3.2 Hypothesis tests

Hypothesis tests can be done with the additive hazards model. One primary question may be whether a specific covariate has any influence on the distribution of lifetimes.

We discuss testing the hypothesis of no regression effect for one or more covariates. This corresponds to testing the following null hypothesis for some $j = 1, 2, \cdots, p$:

$$\begin{aligned} &H_{0j} : \beta_j(t) = 0, \qquad \text{for all } t \\ &H_{1j} : \text{at least one of the } \beta_j(t)'s \qquad \text{is not 0 for some } t \end{aligned} \tag{3.3}$$

Index (j) corresponds to the jth covariate in the analysis. Testing this hypothesis can only be done in the range where $X(t)$ has full rank. A test statistic for H_{0j} is given by the jth element U_j of the vector

$$U = \sum_{T_i} W(T_i)[X^{'}(T_i)X(T_i)]^{-1}X^{'}(T_i)I(T_i) = \sum_{T_i} W(T_i)V(T_i)I(T_i),$$

where $V(T_i) = [X^{'}(T_i)X(T_i)]^{-1}X^{'}(T_i)$ and $W(t)$ is a $(p+1) \times (p+1)$ diagonal matrix of weight functions. Any weight function can be used in the calculation of the test statistics. Here, we follow the suggestion of Aalen and use the weight matrix

$$W(t) = \{\text{diag}[[X^{'}(t)X(t)]^{-1}]\}^{-1}.$$

The test statistic obtained is simply a weighted sum of the cumulative regression function estimator for all event times, because $V(T_i)I(T_i)$ is recognized as the cumulative regression function estimator $\hat{B}(t)$. The covariance matrix of U is estimated by the formula

$$\begin{aligned} V &= \sum_{T_i} W(T_i)[X^{'}(T_i)X(T_i)]^{-1}X^{'}(T_i)I^D(T_i)X(T_i)\{[X^{'}(T_i)X(T_i)]^{-1}\}^{'}W(T_i) \\ &= \sum_{T_i} W(T_i)V(T_i)I^D(T_i)V(T_i)^{'}W(T_i). \end{aligned}$$

To test an individual H_{0j}, the test statistic $U_jV_{jj}^{-1/2}$ can be used. It has an asymptotic standard normal distribution under the null hypothesis. The global test statistic for testing simultaneously H_{0j}, for all j=1,2,...q, with $q \leq p$, is obtained by constructing the q-vector $U_q^* = (U_1, U_2, \cdots, U_q)^{'}$ and the $q \times q$ matrix $V_q^* = ((V_{ge}), g = 1, 2, \cdots, q, e = 1, 2, \cdots, q)$. The test statistic is the quadratic form

$$U_q^{'*}V_q^{-1*}U_q^*,$$

which has an asymptotic chi-square distribution with q degrees of freedom if the null hypothesis is true.

It is also possible to generalize testing to contrasts, or linear combinations of the β's. Let C be a $r \times (p+1)$ matrix of r contrasts. The hypothesis tested will be

$$\begin{aligned} H_{0j} &: C\beta(t) = 0, && \text{for all } t \\ H_{1j} &: C\beta(t) \neq 0, && \text{for some } t. \end{aligned}$$

The formulas for U, K, and V change slightly:

$$\begin{aligned} W_c(t) &= \{\text{diag}[C(X^{'}(t)X(t))^{-1}C^{'}]\}^{-1} \\ U_c^* &= \sum_{T_i} W_c(T_i)C[X^{'}(T_i)X(T_i)]^{-1}X^{'}(T_i)I(T_i) \\ &= \sum_{T_i} W_c(T_i)CV(T_i)I(T_i) \\ V_c^* &= \sum_{T_i} W_c(T_i)C[X^{'}(T_i)X(T_i)]^{-1}X^{'}(T_i)I^D(T_i)X(T_i)\{[X^{'}(T_i)X(T_i)]^{-1}\}^{'}C^{'}W_c(T_i) \\ &= \sum_{T_i} W_c(T_i)CV(T_i)I^D(T_i)V(T_i)^{'}C^{'}W_c(T_i). \end{aligned}$$

The test statistic for H_0 is $U_c^{'*}V_c^{-1*}U_c^*$, which has a limiting chi-square distribution with r degrees of freedom if the null hypothesis is true.

3.3 Assessing the fit of the additive model

One major question when applying the additive regression model, as well as other statistical models, is whether it actually fits the data. Plotting methods for judging goodness of fit, similar to those suggested for the Cox model, have been proposed. There is a number of methods for checking the model fit for the Cox model, and several of them can be extended to the additive model. Two such extensions are discussed below: one is the Arjas plot, which simply compares the observed and expected number of events as a function of time, for various subgroups of covariate values, and the other is martingale residuals.

For the additive model at time t, when the covariate $z(t)$ is time independent, the estimated cumulative hazard rate, given in Klein and Moeschberger (1997, section 11.7) is estimated by

$$\hat{H}[t|z] = \hat{B}_0(t) + \sum_{k=1}^{p} \hat{B}_k(t)z_k,$$

where $\hat{B}_k(t)$, $k = 0, 1, \cdots p$, are the least squares estimators given by (3.2).

As before, let $N_j(t)$ have a value 1 at time t if individual j has been observed to experience the event of interest before or at t and 0 if the individual has yet to experience the event of interest (until the event of interest has occurred); if the individual is censored, $N_j(t)$ will stay at 0.

The martingale residual for the jth individual at time t is given by the difference between $N_j(t)$ (the observed number of deaths) and $\hat{H}[t|z_j(t)]$ (the expected number of deaths under the additive model):

$$\hat{M}_j(t) = N_j(t) - \hat{H}[t|z_j(t)], \qquad j = 1, 2, \cdots, n.$$

These residuals, which are defined for $t \leq \tau$, (τ is the maximal value of t for which the matrix $X(t)$ is a nonsingular matrix), are martingales and, at any event time, the sum of these residuals over all n observations is zero.

To assess model fit, we pick groups of individuals who might be expected to show deviation from the proposed model. Suppose there are q such groups. The first plot is the Arjas plot. Here, we plot the sum of $N_j(t)$ over the gth group against the values of $\hat{H}[t|z_j(t)]$ summed over this group. A point is generated for each group at each event time, and the points are connected. Here, we are plotting the observed number of deaths in a group against the expected number of deaths in a group. If the model holds, this plot should look like a 45° line through the origin for each group.

In the second plot, we graph the martingale residuals (Kim and Lee, 1996). The advantage of looking at the martingale residuals is that it gives a picture of how accumulated hazard compares to events occurred over time. The idea is to compare the martingale residual for a subgroup (for example the gth group) within a data set with different covariate values, to see if the model is valid for all subgroups. The martingale residual at time t for a given group is the sum of the martingale residuals at time t over the members of the group. These sums are then plotted against time. If the model holds, the plotted curves should be close to zero. To determine if the martingale residual process is too far from zero for a model to be acceptable, we need to compute an estimate of the variance of the martingale residual process. Let Q be the $n \times q$ matrix which has as its jth row a 1 in the column of the group the jth observation belongs to and 0 in other columns. Let $M(t)$ be the vector $[\hat{M}_1(t), \cdots, \hat{M}_n(t)]'$. The Q-vector of martingale residuals summed over groups is given by

$$M_{res}(t) = Q'M'.$$

At an event time t_i, let D_i be the $n \times n$ matrix of all zeros except for the diagonal elements corresponding to individuals who die at time t_i, where the diagonal element has the value 1. Let X_i be the $n \times (p+1)$ matrix whose jth row is zero if the jth individual is not at risk at time t_i and has the value $(1, z_1(t_i), \cdots, z_p(t_i))$ if individual j is at risk. Finally, let I be the $n \times n$ identity matrix. Then the covariance matrix for $M_{res}(t)$ is

$$\text{Cov}[M_{res}(t)] = \sum_{t_i \le t} Q^{'}[I - X_i(X_i^{'}X_i)^{-1}X_i^{'}]D_i[I - X_i(X_i^{'}X_i)^{-1}X_i^{'}]^{'}Q. \tag{3.4}$$

Note that the covariance matrix is singular when each individual are in one of the q groups. Confidence intervals for $M_{res}(t)$ can be constructed in the usual fashion:

$$\text{M}_{res}(t) \pm \text{Z}_{1-\alpha/2}(\text{Cov}[M_{res}(t)])^{1/2}. \tag{3.5}$$

A plot of $M_{res}(t)$ against time for various groups with 95% pointwise confidence intervals constructed using (3.5) is used to assess model fit.

We can use both types of plots to assess the fit of the additive model. The Arjas plot gives a clearer indication of lack of model fit than the martingale residual plot, but the martingale residual plot, which explicitly involves time, gives a clearer indication where problems may be arising from in the fit of the model.

3.4 An illustration

In order to illustrate the use of the Aalen additive model, we fit it to some of the data from the UIS study data set (data listed in Appendix Dataset 3.1) which contains 628 records. The variables represented in the dataset are as follows:

TIME: Time to Return to Drug Use (Measured from Admission);

CENSOR: Returned to Drug Use (1 = Returned to Drug Use, 0 = Otherwise);

AGE: Age at Enrollment years;

BECKTOTA: Beck DepressionScore (0.000 - 54.000);

NDRUGTX: Number of Prior Drug Treatment (0 - 40);

IVHX: History of IV Drug Use (1 = Never, 2 = Previous, 3 = Recent);

RACE: Subject's Race (0 = White, 1 = Non-White);

TREAT: Treatment Randomization (0 = Short, 1 = long);

SITE: Treatment Site (0 = A, 1 = B).

Where we fit a model for TIME containing seven main effect. We know that the main focus of the analysis in the additive risk model is on cumulative regression plots (Howell, 1996). The cumulative regression functions are plotted against time and give a description of how the covariates influence the survival over time. The slope of the plot of the cumulative regression function against time gives information on whether the particular covariate has a constant or time dependent effect. Now we describe what the plots of the cumulative regression coefficients are expected to look like under different types of covariate effects. If a regression coefficient is constant over time, it follows that the plot of the estimated cumulative regression coefficient should look like a straight line through the origin, with slope equal to the value of the coefficient. Deviation from a straight line in any time interval in the plot provides empirical evidence for a time-varying effect in the covariate.

- Now we study cumulative regression functions plots for this data set.

  Figures 3.1-3.8 contain eight separate plots, one for the baseline cumulative hazard model, and one for each term in the fitted model. Each of the eight subfigures contains the plot of an estimated cumulative regression coefficient, along with its upper and lower 95 percent pointwise confidence limits.

  Figure 3.1 presents the graph of the estimated baseline cumulative hazard function. We note that the function increases sharply in a nearly linear fashion over the first 400 days, suggesting that the hazard for the baseline subject described above is approximately constant. There is little or no further increase beyond 400 days.

  Figure 3.2 presents the graph for AGE. The estimated cumulative regression coefficient decreases nearly linearly over the entire 600-day interval. There is a slight upwards bump in the plot between 300 and 400 days, but the plot continues to decrease linearly after 400 days. Overall, the plot suggests that there is a decrease in the hazard rate with increasing age that remains in effect over the entire time period.

  Figure 3.3 presents the graph for BECKTOTA (Beck Depression Score). The plot is nearly linear with a positive slope for the first 200 days, at which point it decreases slowly toward zero. We also note that after about 200 days, a horizontal line is contained within the band for the lower 95 percent confidence limit. This plot suggests that increasing values of the BECKTOTA initially increase the hazard rate and then have no effect.

Figure 3.4 presents the graph for NDRUGTX (Number of Prior Drug Treatments). The plot is nearly linear, with a positive slope over the entire 500 days. This plot suggests that the number of treatments increases the hazard over the entire time period.

Figure 3.5 presents the graph for IVHX (History of IV Drug Use). The plot is nearly linear with a slight positive slope over the first 400 days. However, in this time period the zero line is contained within the lower 95 percent confidence bands, which suggests that the covariate may not provide a significant additive increase to the hazard rate during the first 400 days of follow-up.

Figure 3.6 presents the graph for RACE (0 = White, 1 = Non-White). The plot decreases linearly for the first 100 days of follow-up, so it appears that, in this time interval, non-white race is associated with a constant and significant decrease in the hazard rate. After about 150 days, it appears that the covariate no longer has any effect, as the upper confidence band contains a horizontal line. This plot suggests that non-white race has only an early effect on the hazard rate.

Figure 3.7 presents the graph for TREAT (0 = short, 1 = long). Examining the plot we see that during the first 150 days, the covariate has no significant effect. This conclusion is based on the observation that the confidence bands contain a horizontal line in this interval. For the next days the plot decreases sharply and nearly linearly, and the confidence bands no longer include the zero line. This suggests that assignment to the long treatment provides a significant decrease in the hazard rate, starting after 150 days.

Figure 3.8 presents the graph for SITE. The plot shows no consistent trend in any time interval, and a horizontal line is contained within the 95 percent confidence bands. Thus, there appears to be no significant increase or decrease in the hazard rate associated with SITE.

All the cumulative regression function plots above and outputs below were obtained with a SAS macro program given in Appendix 3.1.

Output results:

```
                    Additive hazards model


                575 observation used in analysis
```

```
Estimates are restricted to the time interval 0 to 654

                Global Test
        Chi-Square    d.f    p-value
         41.9600       7     0.0000
```

Table 3.1: Effect, Chi-Square, d.f, p-value

Effect	Chi-Square	d.f	p-value
AGE	13.6903	1	0.0002
BECKTOTA	2.3413	1	0.1260
NDRUGTX	7.7907	1	0.0053
IVHX	7.6689	1	0.0056
RACE	3.9500	1	0.0469
TREATE	6.6554	1	0.0099
SITE	0.5295	1	0.4668

From the results of Table 3.1 above, we see that variables BECKTOTA and SITE have no effect and after we do a backward elimination, we get similar results. We also drew similar conclusions from the cumulative regression function plots.

- Model diagnostics

  We shall use residual theory to examine the fit of the additive model to the UIS dataset. To assess model fit, we shall focus on the AGE covariate, the only continuous covariate with a significant effect. We divide subjects into two groups: those with age less than 32.4 and those with age greater than or equal to 32.4. Figure 3.9 shows the Arjas plot for the two groups. We see that both curves follow a 45° line very well, and there is no indication of incorrect modeling of the age covariate. Figure 3.10 shows the martingale residual process plot for the less than 32.4 age group and 95% pointwise confidence limits. We note that the confidence intervals contain zero, so, again, there is no evidence of lack of model fit. The conclusion is the same for the greater than 32.4 age group. Similar plots, where the observations are grouped by other covariate, show very good fit of the model.

  Since programs to make Arjas and martingale residual plots to examine the fit of the additive model are not directly available in SAS, we use a MATLAB program to plot them. The program is given in the Appendix 3.2.

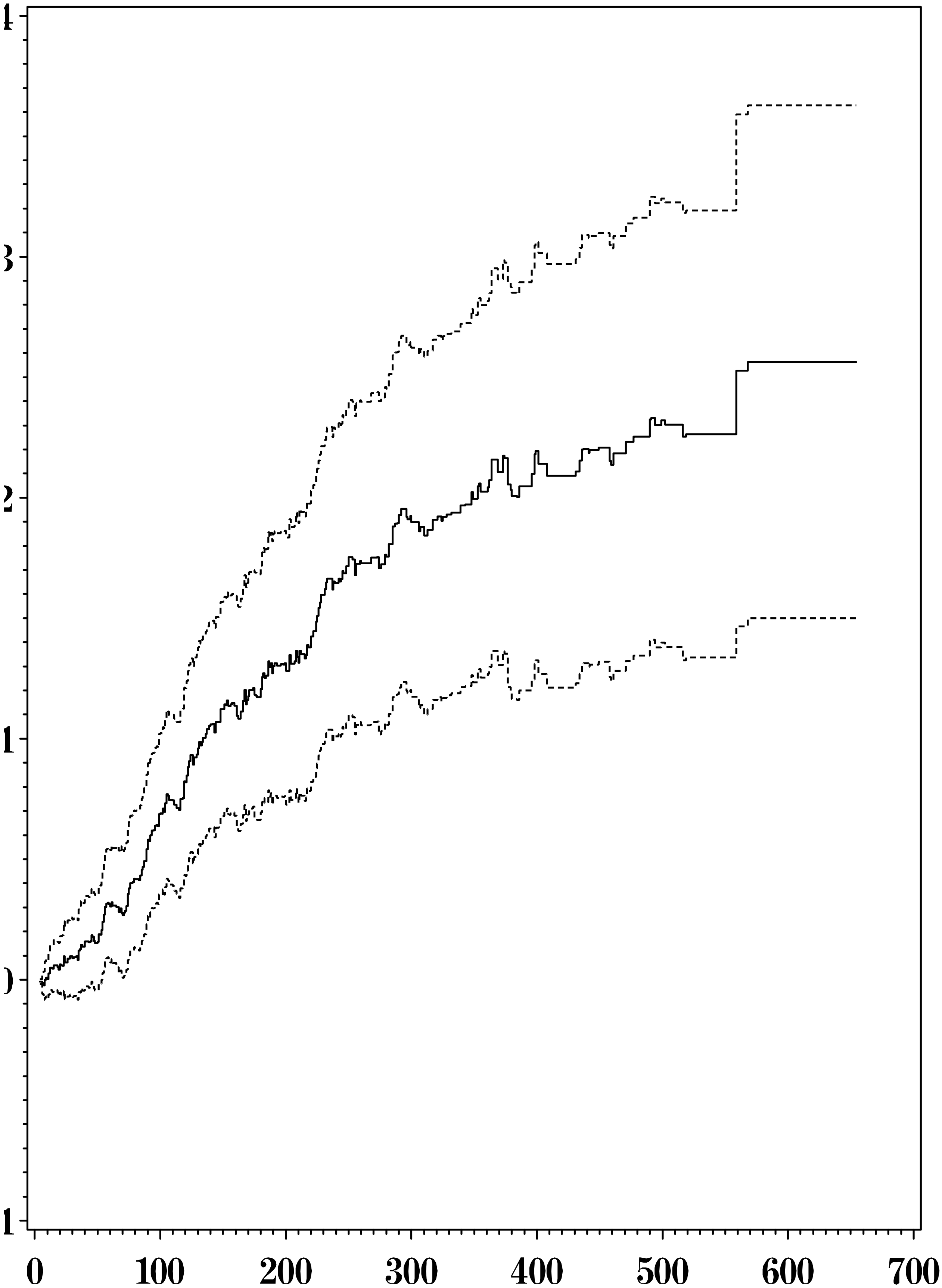

Figure 3.1: Baseline cumulative hazard function

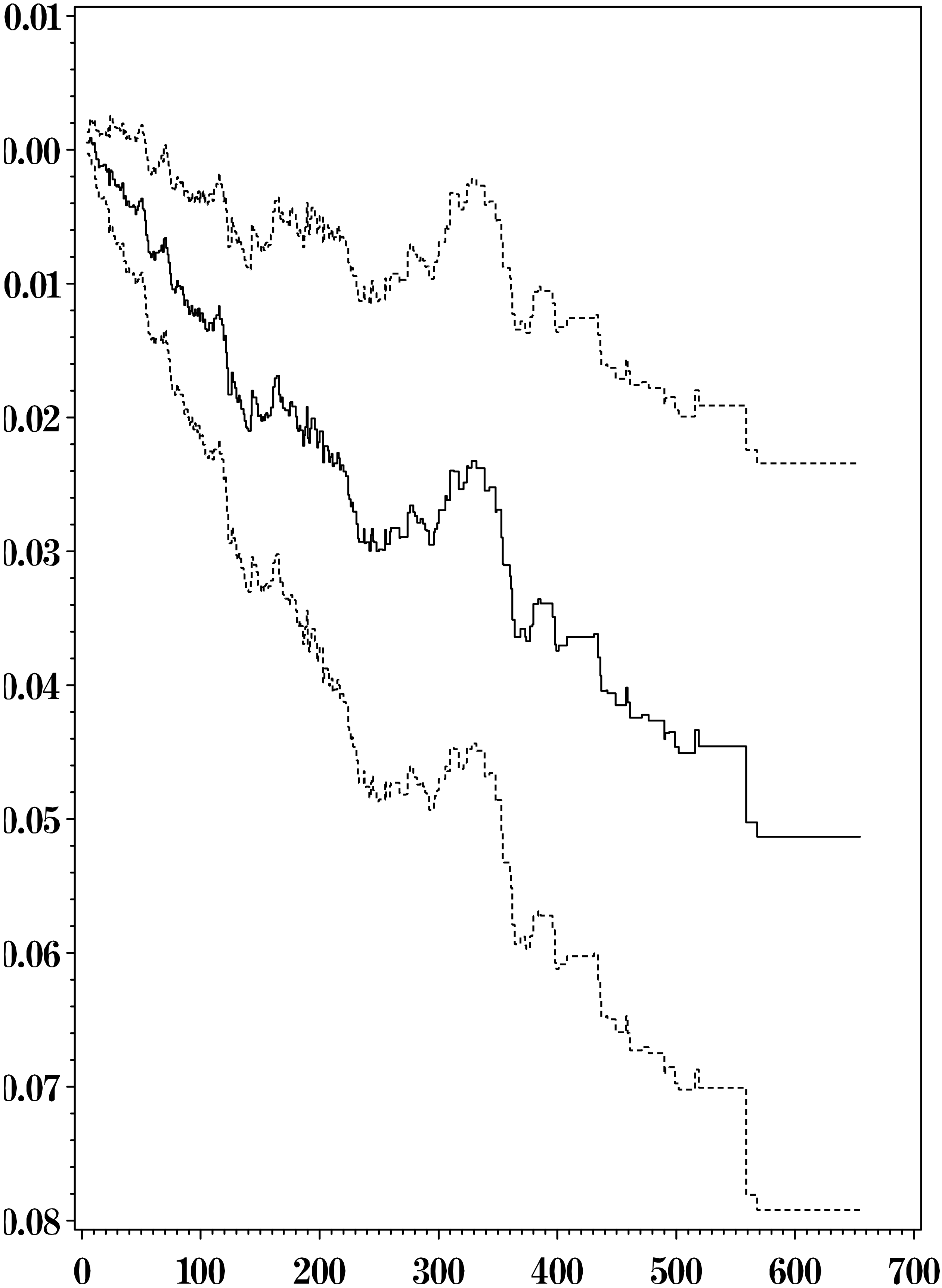

Figure 3.2: Cumulative regression coefficient for AGE

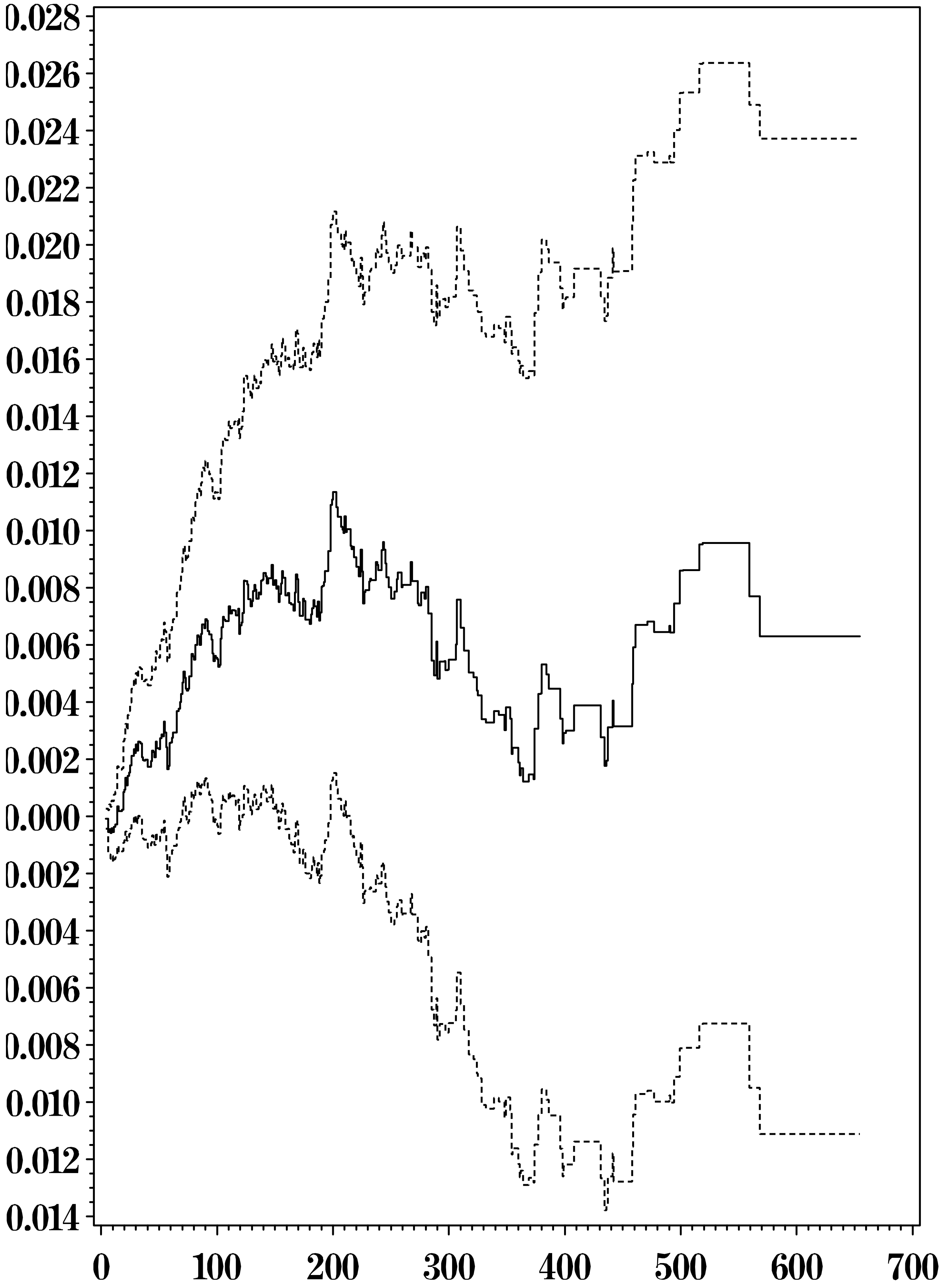

Figure 3.3: Cumulative regression coefficient for BECKTOTA

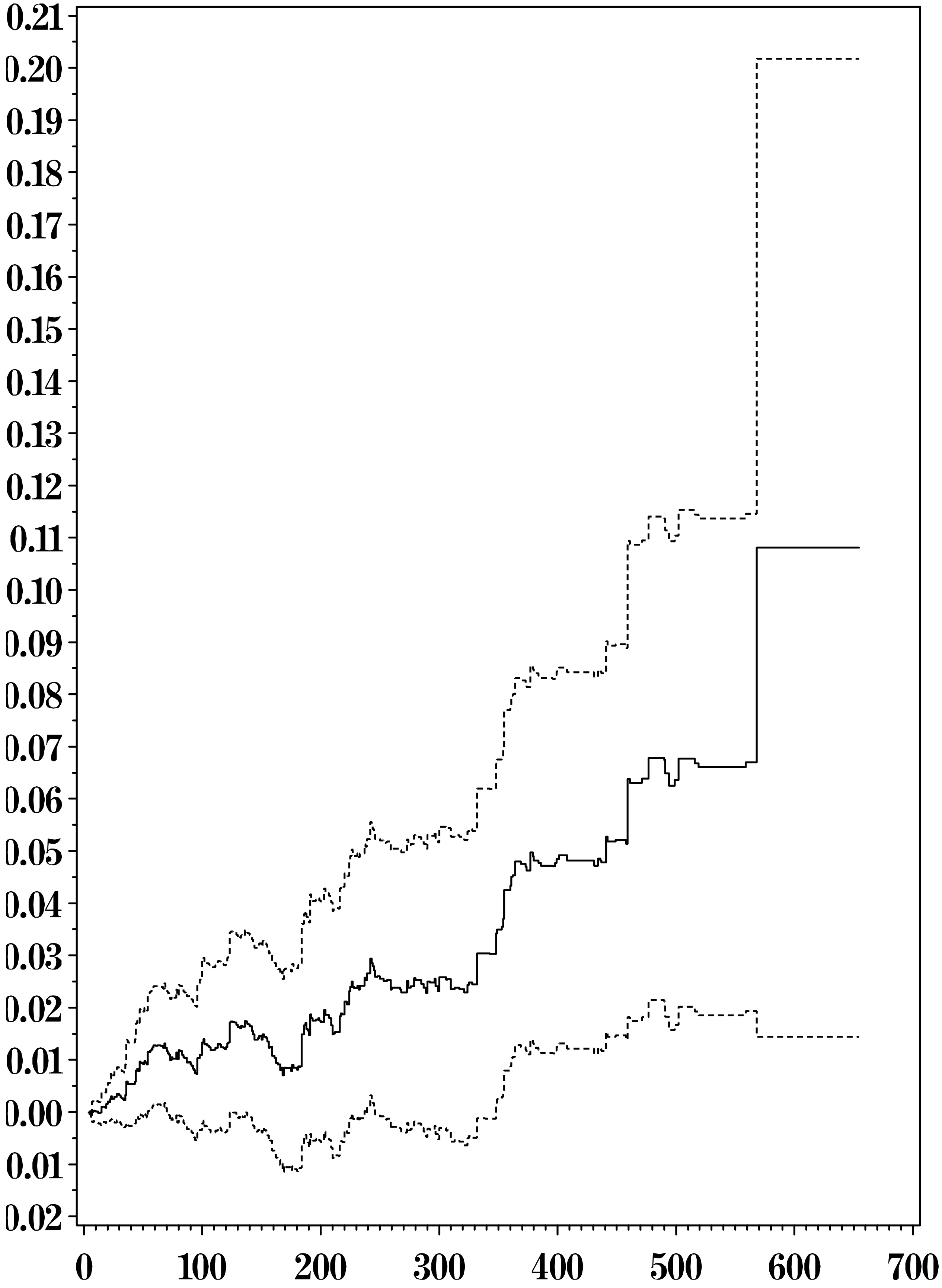

Figure 3.4: Cumulative regression coefficient for NDRUGTX

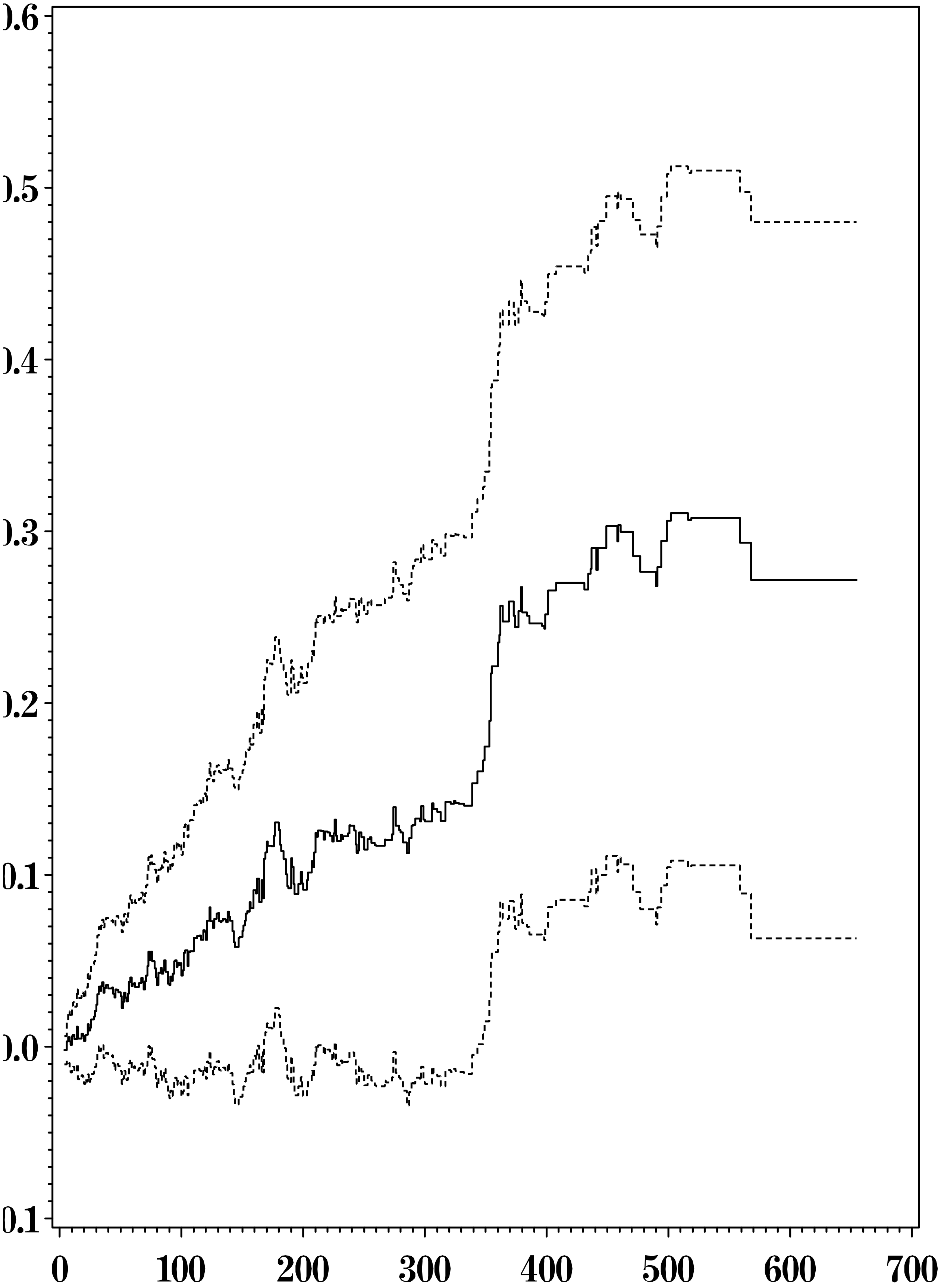

Figure 3.5: Cumulative regression coefficient for IVHX

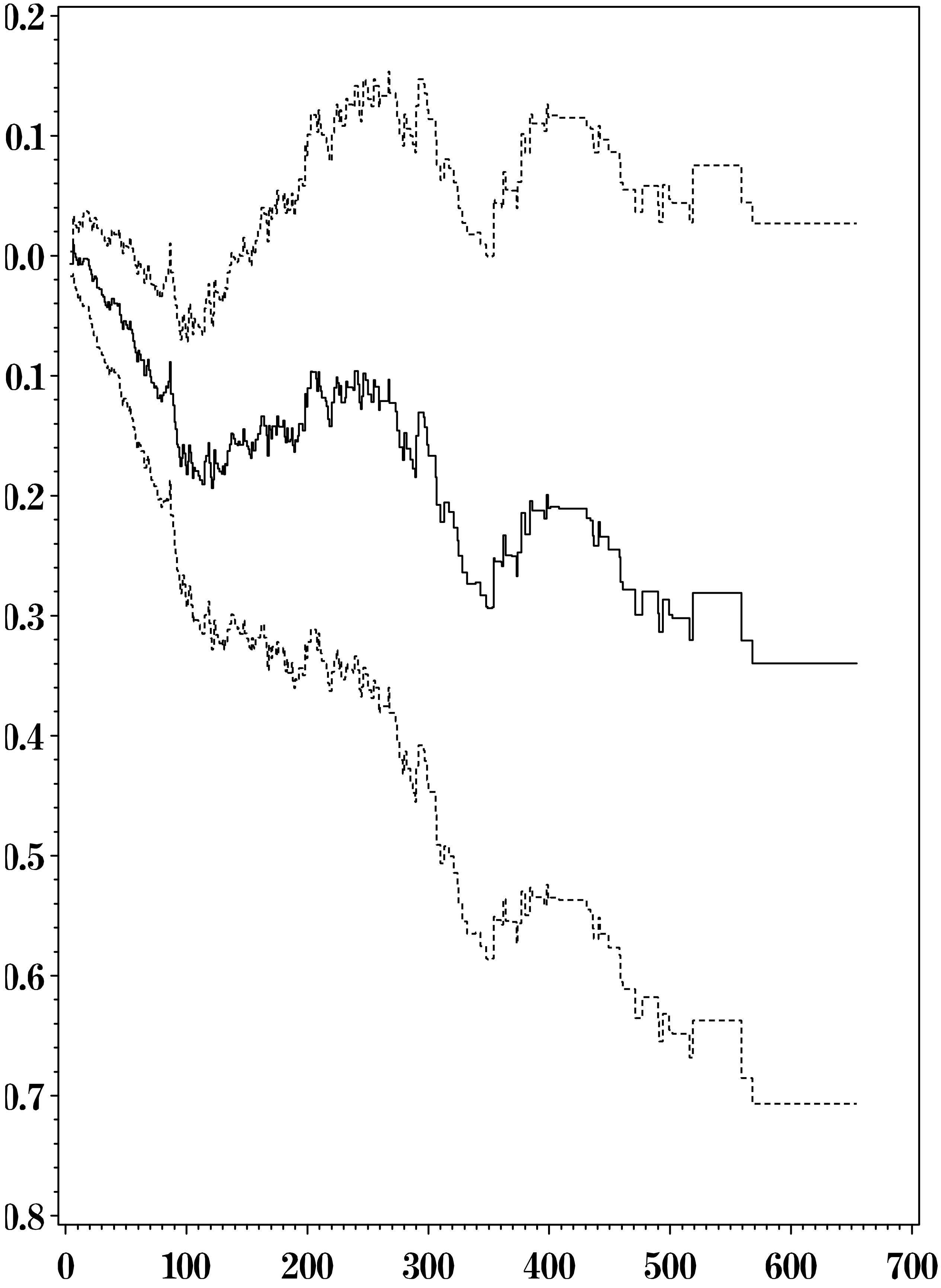

Figure 3.6: Cumulative regression coefficient for RACE

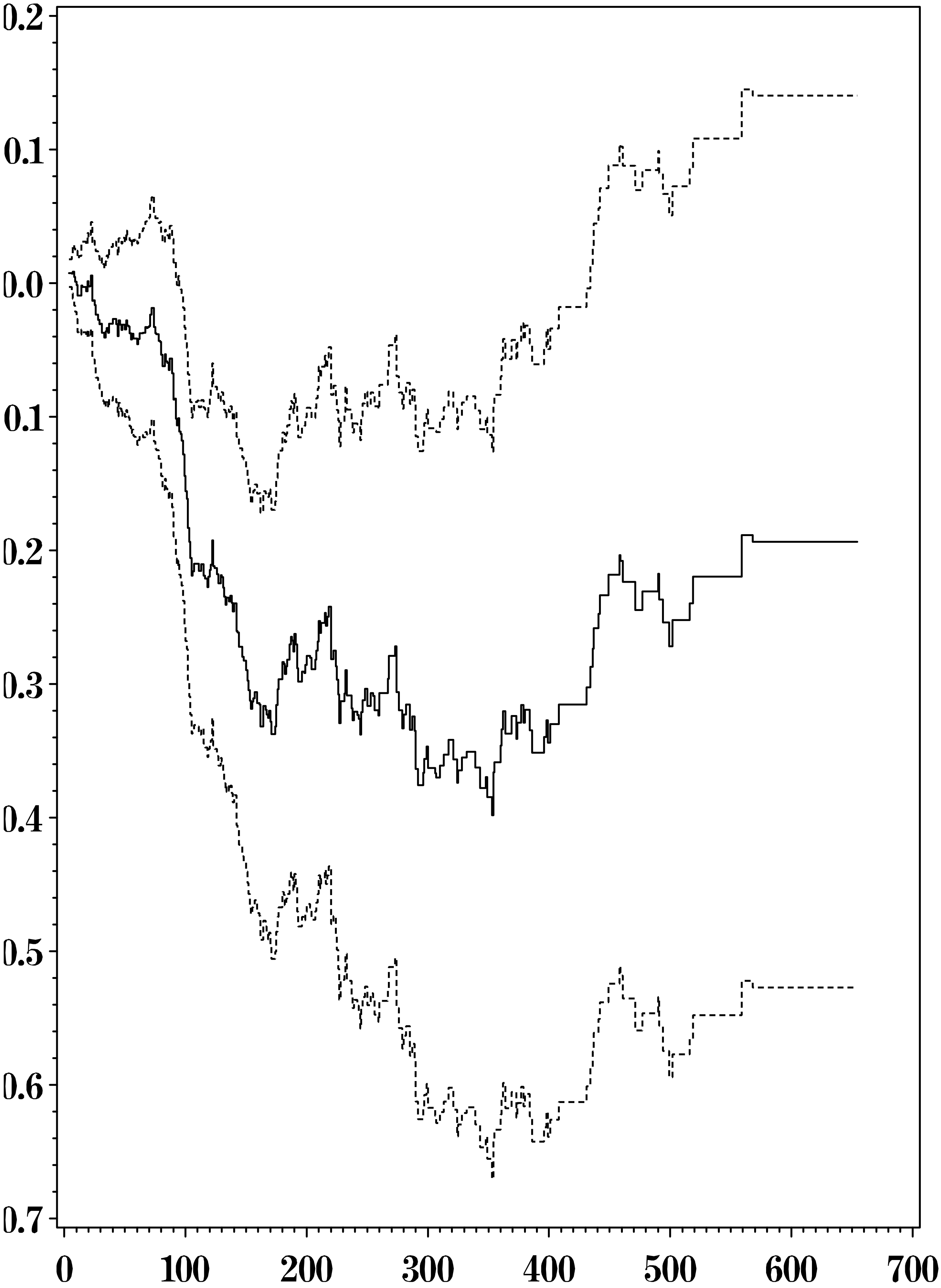

Figure 3.7: Cumulative regression coefficient for TREATE

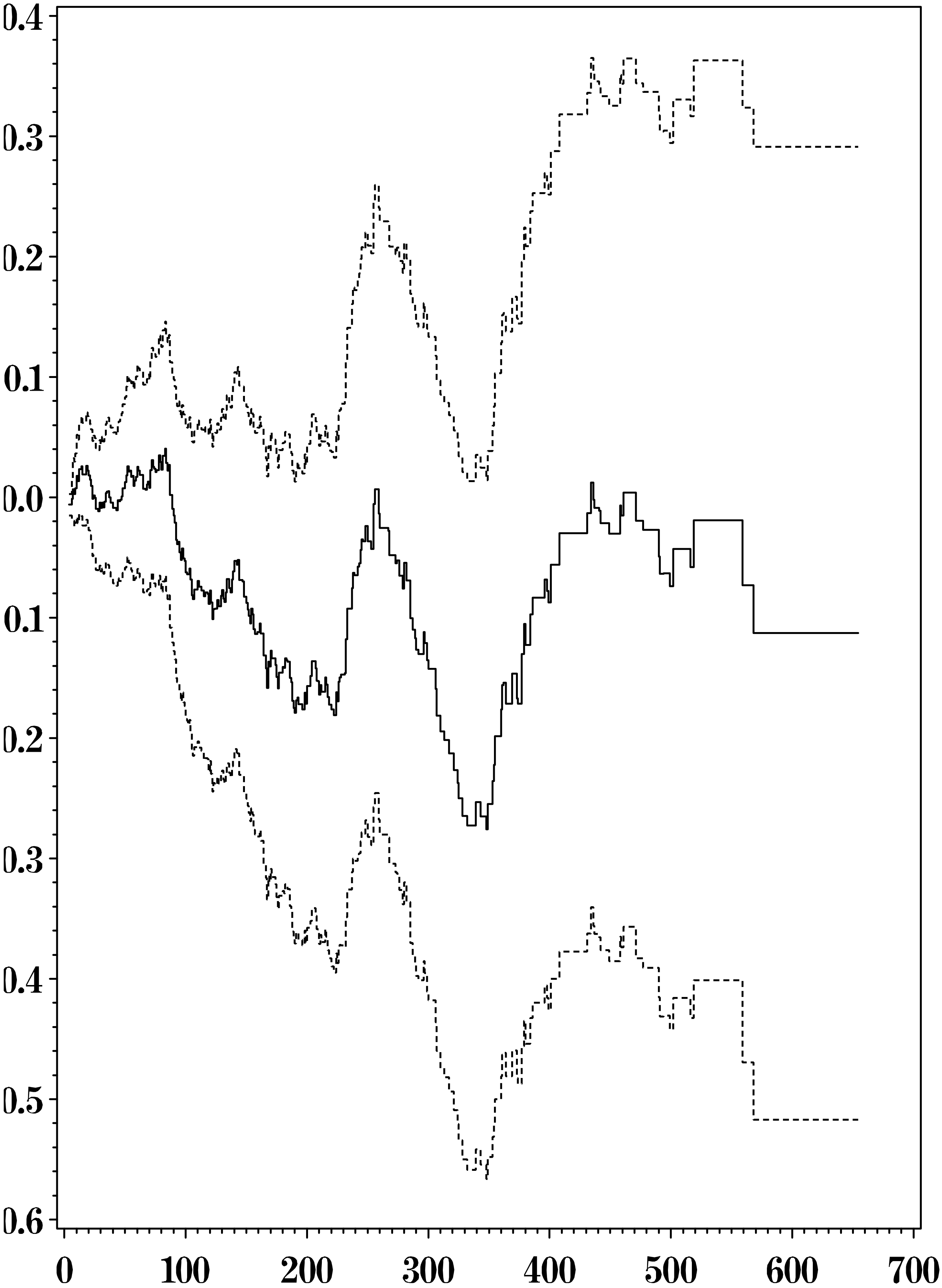

Figure 3.8: Cumulative regression coefficient for SITE

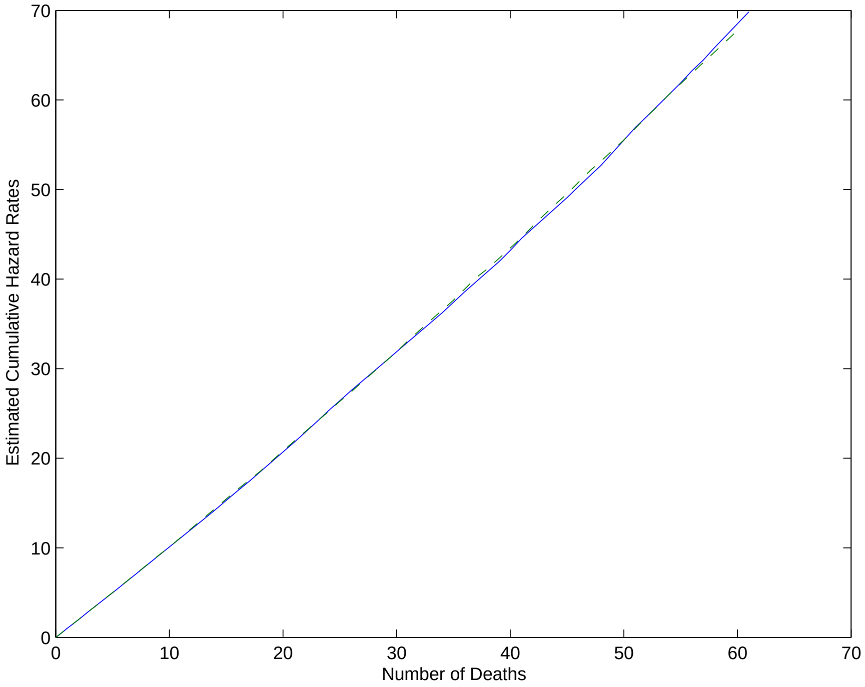

Figure 3.9: Arjas plot to check the adequacy of the additive hazard model. Age less than 32.4 (blue), age greater than 32.4 (green)

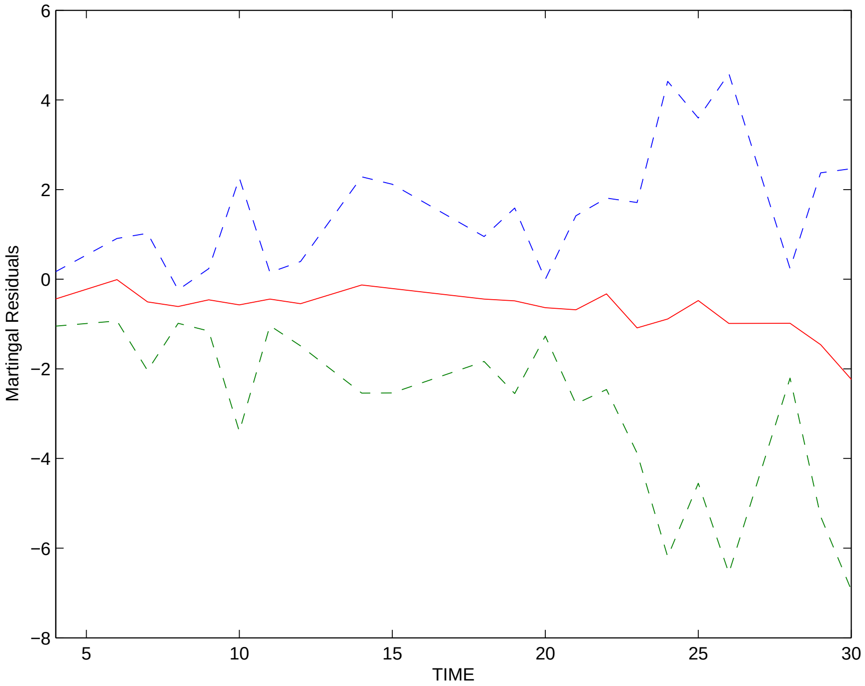

Figure 3.10: Martingale residual process plot and 95% pointwise confidence limits for the under 32.4 age group.

Chapter 4

A comparison between the multiplicative and additive risk models

The primary objective of this chapter is to investigate and compare the use of the Cox proportional hazards model and Aalen's additive model in the analysis of survival data through application and simulation (Torner, 2004).

In sections 4.2 to 4.4, survival data from a study of 90 male laryngeal cancer patients is investigated using the Cox proportional hazards model. The model is optimized by examining different aspects and use of appropriate residual plots. After optimizing the Cox model, the same data is used to fit an additive model. Plots of the martingale residual process and Arjas plots are used to check model fit and optimize model options. The information gained from fitting the two models is similar in some respects, but also quite different in others. Both procedures result in the same covariates selected to remain in the model. The Cox model yields easily interpreted estimates of the covariate effects, but the assumption of proportional hazards is necessary to make these estimates valid. The additive model and plots of the cumulative regression functions give an appealing understanding of how the hazard profile evolves with time.

In section 4.5, the power and robustness of tests of significant covariate effects based on the Cox and additive hazards model are compared through simulation.

4.1 Introduction

Laryngeal cancer is the most common malignant disease for males in northern Europe and North America. The data used is from Kardaun (1983) who reported data on 90 males diagnosed with cancer of the larynx during the period 1970-1978 at a Dutch hospital. Times recorded are the intervals (in years) between first treatment and either death or the end of the study (January 2, 1983). Also recorded are the patient's age at the time of diagnosis, the year of diagnosis, and the stage of the patient's cancer. The four stages of disease in the study were based on the T.N.M (primary tumor (T), nodal involvement (N), and distant metastasis (M) grading) classification used by the American Joint Committee for Cancer Staging (1972). The four groups are Stage 1, $T_1N_0M_0$ with 33 patients; Stage 2, $T_2N_0M_0$ with 17 patients; Stage 3, $T_3N_0M_0$ and $T_xN_1M_0$, with 27 patients; x=1,2, or 3; and Stage 4, all other TNM combinations with 13 patients. These stages are used internationally to classify the disease for treatment decisions and prognosis. Since this classification system is truly international it is convenient to compare different studies and regimes used in different countries. The four groups are labeled Stage 1 through Stage 4, which is ordering the stages from least serious to most serious. Now we describe this data set. The variables represented in the dataset are as follows:

stage: Stage of disease (1=stage 1, 2=stage 2, 3=stage 3, 4=stage 4)

time: Time to death or on-study time, months

age: Age at diagnosis of larynx cancer

dyear: Year of diagnosis of larynx cancer

death: Death indicator (0=alive, 1=dead)

This present work aims to investigate the use of two statistical models to model the survival of patients. The work is divided in the following parts:

- To build a traditional Cox proportional hazards model to describe which factors influence the survival of these patients.
- To fit an additive regression model to the same data with the same purpose.
- To compare the results derived using the Cox proportional hazards model with the results from fitting the additive model.

Data from a clinical study have kindly been made available by Klein and Moeschberger (2003).

4.2 Cox Proportional Hazard Regression Model

In Chapter 2, we have already studied in detail the Cox proportional hazards model.

4.2.1 Cox model with several covariates

A dataset of 90 males diagnosed with cancer of the larynx was described in Section 4.1. Here we code the variable "stage of disease" in preparation for performing a proportional hazards regression test. Since stage has four levels, we adopt the usual indicator variable coding methodology and construct the indicator variables as follows:

$$\begin{aligned} \text{stage2} &= 1, \text{if the patient is in stage 2, 0 otherwise ;} \\ \text{stage3} &= 1, \text{if the patient is in stage 3, 0 otherwise ;} \\ \text{stage4} &= 1, \text{if the patient is in stage 4, 0 otherwise .} \end{aligned}$$

For a patient with Stage 1 cancer, we have stage2 = stage3 = stage4 =0.

The full model for this situation is

$$h(t|z) = h_0(t)\text{exp}(\beta' z) = h_0(t)\text{exp}(\beta_1\text{age} + \beta_2\text{stage2} + \beta_3\text{stage3} + \beta_4\text{stage4}).$$

The initial fitting of this Cox proportional hazards model yielded the results of Table 4.1.

Table 4.1: Results of preliminary fitting of Cox model

Variable	DF	Parameter	Standard	Chi-Square	pr > ChiSq	Hazard
		Estimate	Error			Ratio
AGE	1	0.01890	0.01425	1.7589	0.1848	1.019
stage2	1	0.13842	0.46232	0.0896	0.7646	1.148
stage3	1	0.63815	0.35609	3.2116	0.0731	1.893
stage4	1	1.69333	0.42218	16.0876	< .0001	5.438

The estimates of the parameters are obtained as

$$b_1 = 0.0189 \qquad b_2 = 0.13842 \qquad b_3 = 0.63815 \qquad b_4 = 1.69333.$$

From Table 4.1, we find the variable AGE was not significant (p=0.1848), but variable AGE has often been an important predictor in larynx cancer, so it will be kept in the model with the other variables. Through model checking in section 4.2.2, we find that the model considered fits the data very well.

The plot of the survival probabilities for the stage1, stage2, stage3, stage4 in Figure 4.1 at the end of this chapter shows that the curves are not equal but diverge. This is also consistent with the general clinical perception of stage1-stage4, stage4 is a more advanced stage of the disease and the progress of the disease is inevitable. For stage1, the prognosis is more uncertain and some patients may have somewhat longer life expectancy.

From Table 4.2, we see at six years the estimated survival probabilities for a 65-year-old are 0.64235 for a stage1 patient, 0.60150 for a stage2 patient, 0.43263 for a stage3 patient, and 0.09010 for a stage4 patient. At 6 years, 95% confidence intervals for the survival function, based on the log transformation, are (0.46668, 0.77332), (0.33486, 0.78963), (0.24430, 0.60767), (0.00887, 0.29349), for stage1, stage2, stage3, stage4, respectively. (SAS programme is given in the Appendix 4.1)

Table 4.2: The estimated survival probabilities and 95% confidence limits

age	stage2	stage3	stage4	time	survival	slower	supper
65	0	0	0	6	0.64235	0.46668	0.77332
65	0	1	0	6	0.43263	0.24430	0.60767
65	0	0	1	6	0.09010	0.00887	0.29349
65	1	0	0	6	0.60150	0.33486	0.78963

4.2.2 Model checking

This initial model has been fitted without considering the best functional form of the continuous variables (age) and without questioning the underlying assumption of proportional hazard. The fit of this preliminary model was therefore investigated by examining the following residual plots.

- The overall fit of the Cox model was investigated by Cox-Snell residuals, as described in section 2.6 (SAS program given in the Appendix 4.2).

  Figure 4.2 is a plot of the residuals versus the estimated cumulative hazard of the residuals. If the Cox model fits the data, the plot should follow the 45° line. The plot suggests that this model does not fit too badly.

- The functional form of the continuous variable AGE was investigated by examining martingale residuals.

  We shall examine the problem of determining the functional form to be used for a given covariate to best explain its effect on survival through a Cox proportional hazards model. The best functional form may be a transform of the covariate, such as lnZ, Z^2, or $ZlnZ$, or it may be a discretized version of the covariate. In fact, it is common practice in many medical studies to discretize continuous covariates, and the residuals presented here are useful for determining cut points for the covariates.

  The functional form of the covariate AGE needs to be checked. We have chosen to use martingale residuals to try to determine the correct functional form of the AGE covariate. In Section 2.6, we have studied in detail its theory and its formula. If the martingale plot is linear, no transformation is needed. Including the untransformed covariate in the model together with the other covariates yields an appropriate regression coefficient. If, however, there appears to be a discrete time point where the slope changes, a dichotomized transformation of the covariate may be indicated.

  Looking at a martingale residual plot (Figure 4.3) for our data, for the AGE covariate, the smoothed curve is roughly linear at all time interval, but from this plot, we find at time 65 and at time 75, there might be a little bit of change. So we might think that the AGE covariate may be coded as an indicator variable. To verify, the indicator variable NEWAGE is thus defined as follows:

$$\text{NEWAGE} = \begin{cases} 0 & \text{if} \quad \text{AGE} < \Theta \\ 1 & \text{if} \quad \text{AGE} \geq \Theta \ . \end{cases}$$

  The cut-off value Θ is chosen from the values of AGE in the dataset. A profile likelihood may be plotted for each AGE value in the data set and the Θ value yielding the highest value of the log-likelihood is chosen (Klein and Moeschberger, 2003, Section 8.4). Here we found (from Table 4.3) that each AGE value yielded the same log-likelihood value (-195.906), so we decide that the original covariate AGE in the model does not need to be transformed (SAS program is shown in the Appendix 4.3).

- The proportional hazards assumption was investigated by examining scaled Schoenfeld residuals

  The proportional hazards assumption was examined for the variables STAGE and AGE. The S-plus default for checking the proportional hazards assumption

Table 4.3: Log Partial Likelihood as a Function of Θ at the Failure Times

Value of AGE	Log Partial Likelihood
41	-195.906
43	-195.906
45	-195.906
47	-195.906
48	-195.906
.	-195.906
.	-195.906
70	-195.906
71	-195.906
72	-195.906

is a formal test and plot of scaled Schoenfeld residuals. The Schoenfeld residual is the difference between the covariate at the failure time and the expected value of the covariate at this time. The results of the test of the proportional hazards assumption is presented in Table 4.4. The scaled Schoenfeld residuals for the the variables are plotted in Figure 4.4, together with a smooth. When the proportional hazards assumption holds, a relatively straight horizontal line is expected. The results from Table 4.4 and Figure 4.4 based on scaled Schoenfeld residuals indicate the PH assumption is satisfied by all variables in the model.

Table 4.4: Test of proportional hazards assumption

covariate	rho	chisq	p
stage	-0.285	4.23	0.0897
age	0.133	1.14	0.2848
GLOBAL	NA	5.21	0.0740

Overall, the residuals seem reasonable and no subjects will be considered for exclusion from the analysis.

4.3 Additive Hazards Regression Model

In the Cox model the covariates are assumed to act multiplicatively on a baseline hazard. The baseline hazard is the hazard for individuals with covariate values equal to zero and this hazard is a function of time. The model is semi-parametric in the sense that constant proportional hazard throughout the study is assumed. In some cases this assumption of constant proportional hazards may not always be valid.

An alternative to the Cox model, which does not assume constant proportional hazards, is an additive model, proposed by Aalen (1989). In this model the covariates are modeled as additive risks to a baseline hazard and the regression coefficients are allowed to vary freely over time.

4.3.1 Fitting of the additive model

The data described above, with covariate formation for stage2, stage3, stage4 and age, was fitted as an additive model. The results are presented graphically below with time on the x-axis and cumulative regression functions on the y-axis. The dotted lines indicate 95 percent pointwise confidence intervals.

Here, the estimated cumulative baseline hazard $\hat{B}_0(t)$ (Figure 4.5) is an estimate of the cumulative hazard rate of a stage 1 patient aged 64.11. $\hat{B}_1(t)$, $\hat{B}_2(t)$, $\hat{B}_3(t)$ (Figure 4.6 - 4.8) show the cumulative excess risk due to stage 2, 3 or 4 patients of a given age compared to stage 1 patients with a similar age. Here, it appears there is little excess risk due to being a stage 2 patient, whereas stage 3 and stage 4 have an elevated risk in the first two years following diagnosis where the slopes of the two cumulative hazards are nonzero. Figure 4.9 shows the excess risk due to age.

All the cumulative regression function plots above and outputs below were obtained with a SAS macro programme given in the Appendix 4.4. These results are summarized below.

Output results:

```
                    Additive hazards model


      Estimates are restricted to the time interval 0 to 4.30


                          Global Test
```

Chi-Square	d.f	p-value
10.9613	4	0.0270

Table 4.5: Effect, Chi-Square, d.f, p-value

Effect	Chi-Square	d.f	p-value
stage2	0.1456	1	0.7027
stage3	3.0062	1	0.0829
stage4	8.4655	1	0.0036
age	0.2333	1	0.6291

This table (Table 4.5) suggests that, adjusted for age, there is little difference between the survival rates of Stage 2 or Stage 3 patients compared to Stage 1 patients (p-values > 0.05), but that Stage 4 patients have a significantly different survival (p-value of 0.0036).

4.3.2 Model Checking of Additive Model

Now, we use martingale residual plots and Arjas plots to check the initial fitting of the additive model.

In our data set there is only one covariate with continuous values, AGE. Fitting the Cox model it was shown that this covariate should not be transformed. Below we investigate if and how well the untransformed AGE covariate fits in the additive model. The concept behind Arjas plots is to plot the expected number of failures against the actual number of failures in sub-groups with different covariate values. The sub-groups chosen were the same sub-groups of AGE values as in the martingale residual plot. These are the sub-groups relevant from a clinical perspective and it also gives an opportunity to compare the model information from these two residual plots. An Arjas plot is not a true residual plot, but deviations from the 45° slope will give essentially the same information. We shall use these techniques to examine the initial fit of the additive model to data on laryngeal cancer.

To assess model fit, we shall focus on the age covariate. We divide subjects into two groups: those with ages less than 64.11 and those with ages greater than or equal to 64.11. Figure 4.10 shows the Arjas plot for the two groups. We see that both curves follow a 45° line quite well, and there is no indication of incorrect modeling of the AGE covariate. The martingale residual plot was plotted for subgroups of Age in our data set. Figure 4.11 shows the martingale residual process plot for the less than 64.11 age group and 95% point wise confidence limits. We note that the confidence intervals all contain zero, so, again, there is no evidence of lack of model fit.

4.4 Concluding remarks of example

The data set examined in this comparative study is fairly small, 90 patients. The comparison made here is very informal in nature and information gained from fitting the Cox model has been used in optimizing choices for the additive model as well. The Cox model and Aalen's additive model give similar results with regard to covariates selected to remain in the model. Even though we find that the variable AGE was not significant (p=0.1848) under the Cox model and (p=0.6291) under the additive model, the variable AGE is a very important predictor in larynx cancer, so it was kept in the model with other variables.

Now, we summarize common characters between the Cox model and Aalen's additive model.

1. These two models can both be used to investigate the effects of risk factors on time to event. They can handle censored and/or truncated observations.

2. Using both models, we got almost the same initial p-value. Table 4.6 shows this comparison of p-values for the Cox model and the additive model.

Table 4.6: Comparison of p-values for covariates under the Cox model and the additive model

Covariate	Cox p-value	Additive p-value
AGE	0.1848	0.6291
stage2	0.7646	0.7027
stage3	0.0731	0.0829
stage4	$< .0001$	0.0036

3. They all use residual plots (martingale residual plot and Arjas plot) to check initial fitting.

Next, we summarize different characters between the Cox model and Aalen's additive model.

1. Even though these two models all can be used to investigate the effects of risk factors on disease occurrence or death, the Cox model has been more popular than the additive risk model. As we showed in detail in Chapter 2, the Cox model is established as the major framework for regression analysis of survival data, it is extremely useful in practice since the estimated coefficients themselves or simple functions of them can be used to provide estimates of hazard ratios.

2. For the Cox model, the effect of the covariates on survival is to act multiplicatively on some unknown baseline hazard rate, which makes it difficult to model covariate effects that change over time. An alternate model is Aalen's additive model. This model assumes that the covariates act in an additive manner on an unknown baseline hazard rate. The unknown risk coefficients are allowed to be functions of time so that the effect of a covariate may vary over time. It is an advantage of using the additive hazards model.

3. For the Cox model, estimation of the risk coefficient was based on the partial likelihood. For the additive model, estimators of the risk coefficients are based on a least-squares technique. The estimates of the baseline hazard rate are not constrained to be nonnegative by this least-squares estimation procedure. In fact, if continuous covariates are not centered at their mean values, the estimator of $\beta_0(t)$ may be negative.

4. Aalen (1989) shows that if a covariate is independent of all the other covariates in the model, then, the regression model with this covariate eliminated is the same as the regression model with this variable included. Note that this is not true for the Cox proportional hazards model.

5. The test procedure for the effect of covariates is similar but not equivalent. The null hypothesis for the additive model:

$$H_0 : \beta_j(t) = 0 \qquad \text{for all } t$$

corresponds to the null hypothesis in the Cox model:

$$H_0 : \beta_j = 0.$$

The alternative hypotheses, in favor of which the null hypothesis may be rejected, are, however, quite dissimilar for the two models. The alternative for the additive model states

$$H_1 : \beta_j(t) \neq 0 \qquad \text{for some } t,$$

which is a weaker alternative compared to the alternative hypothesis in the Cox model,

$$H_1 : \beta_j \neq 0,$$

which is valid for all t. This would imply that to reject the null hypothesis in the Cox model, we would require that the best overall estimate of β, which is valid for all time points, is different from 0. However, one must remember that the test statistic in the additive model is designed as a weighted combination of all $\beta(t)$, which means that the null hypothesis may not be so easily rejected even if there are significant deviations from the null hypothesis at a few time points.

6. In practice, we found that it is not appropriate to use Aalen's additive hazards model for all datasets, because when we estimate cumulative regression functions $B(t)$, they are restricted to the time interval where X (X has been defined in Chapter 3) is of full rank, that means $X'X$ is invertible. Sometimes we found that X is not of full rank, which was not a problem with the Cox model.

7. In addition, for the Cox model, statistical software (SAS, SPLUS, R) is readily available and easy to use to fit models, check model assumptions and assess model fit. For Aalen's additive model, the model is not available in any commonly used computer packages, such as SAS, SPSS, SPLUS. In Chapter 3, we use a SAS macro that performs the additive hazards regression. This macro calculates the parameter estimates and respective standard deviations and confidence intervals. Line plots can be printed of each parameter estimate versus time to view how the covariate effects may change over time. But SAS macro is not publicly available to print the residual plots suggested by Aalen (1989) to check the validity of the model, we had to write MATLAB programs to create these plots.

An overall conclusion is that the two models give different pieces of information and should not be viewed as alternatives to each other, but as complementary methods that may be used together to give a fuller and more comprehensive understanding of data.

4.5 A comparison by simulation

In this section, we want to assess robustness and power of tests based on the Cox model and additive hazards model. We do so by generating random datasets. Here, the Monte Carlo power of each test will be calculated to find which test is more powerful.

Here is a quick description of the models that will be used in the simulation, with a very quick explanation of how to simulate data from these models. We generate uncensored random samples of the form $(t_i, z_i), i = 1, 2...n$, where z is a binary covariate and t_i, given z_i, is simulated from three different models. The first model is a proportional hazards model, the second model is an accelerated failure time model and the third model is an additive hazards model. All models give the same log logistic distribution at $z = 0$, but different distributions at $z = 1$ (see Figure 4.12). The parameters under $z = 1$ have been chosen such that the probability of survival to 10 years is the same in each model.

Model 1: Cox model with $\beta = 0.6$ and a log logistic baseline hazard with parameters $\lambda = 0.2$ and $\gamma = 4$, i.e.,

$$h(x|z) = \frac{0.2 \cdot 4 \cdot (0.2x)^{4-1}}{1 + (0.2x)^4} e^{0.6z}.$$

Model 2: Log linear model

$$\ln X = \beta_0 + \beta_1 z + \sigma W,$$

where $\beta_0 = -\ln 0.2$, $\beta_1 = -0.568245$, $\sigma = 1/4$ and W is logistic(0,1).

Model 3: Additive hazards model

$$H(x|z) = B_0(x) + B_1(x)z,$$

where $B_0(x) = ln(1 + (0.2x)^4)$, $B_1(x) = ln(1 + (0.15874x)^6) - ln(1 + (0.2x)^4)$.

Model 0: Any of models 1 to 3 with $z = 0$, i.e.,

$$h(x|z = 0) = \frac{0.2 \cdot 4 \cdot (0.2x)^{4-1}}{1 + (0.2x)^4}.$$

We performed a series of simulation studies to assess the power and robustness of the test of no effect of z with each of the proportional and additive hazards models. Sample sizes of 50 and 200 were considered. At first, we simulated datasets from all 3 models with R and SAS, with 50 observations (25 with $z = 0$ and 25 with $z = 1$) and fitted the Cox and the additive hazards model. We tested for no covariate effect using the score test under the Cox model and the test of (3.3) under the additive hazards model. We have repeated this procedure with samples of 200 observations (100 with $z = 0$ and 100 with $z = 1$). For each model and each sample size, 1000 samples were simulated.

For each sample simulated, the test statistics under both models were saved and each simulation was classified as a "success" or a "failure", depending on whether the corresponding p-value was less than or greater than 0.05, respectively. We know that for both models, the test statistic follows a chi-square with 1 degree of freedom under H_0. The 95th percentile of that distribution being 3.841, we reject H_0 when the test statistic is $>= 3.841$, and we do not reject H_0 when the test statistic is < 3.841.

The simulation results are summarized in Table 4.7 (the programs are given in the Appendix).

Table 4.7: Proportions of rejections of H_0: no covariate effect

N=1000		Cox model	Additive hazards model
Model 0:	n=50	0.053	0.045
	n=200	0.048	0.044
Model 1:	n=50	0.556	0.515
	n=200	0.993	0.99
Model 2:	n=50	0.971	0.959
	n=200	1	1
Model 3:	n=50	0.28	0.308
	n=200	0.707	0.939

From Table 4.7, we find that for model 0, both the Cox and additive hazard models accept H_0 in a proportion very close to the nominal 0.05 level.

For Model 1, both models reject H_0 often, as expected, but we find that the test statistic based on the Cox model has slightly greater power than the test statistic based on the additive hazards model. This had to be expected since Model 1 is a Cox model. However, we must admit that the additive hazards model is surprisingly powerful here.

For Model 2, both models reject H_0 often. Though Model 2 is neither a Cox nor an additive hazards model, both tests seem to be roughly as powerful in this case.

For Model 3, both models reject H_0 as often, but we find that the test statistic based on the additive hazards model has greater power than the test statistic based on Cox model (especially when n=200). This had to be expected since Model 3 is an additive hazards model.

From this small study, we can see that when testing for the effect of a binary covariate on survival, both tests exhibit good robustness properties. The additive hazards model proved to be surprisingly powerful under non additive alternatives.

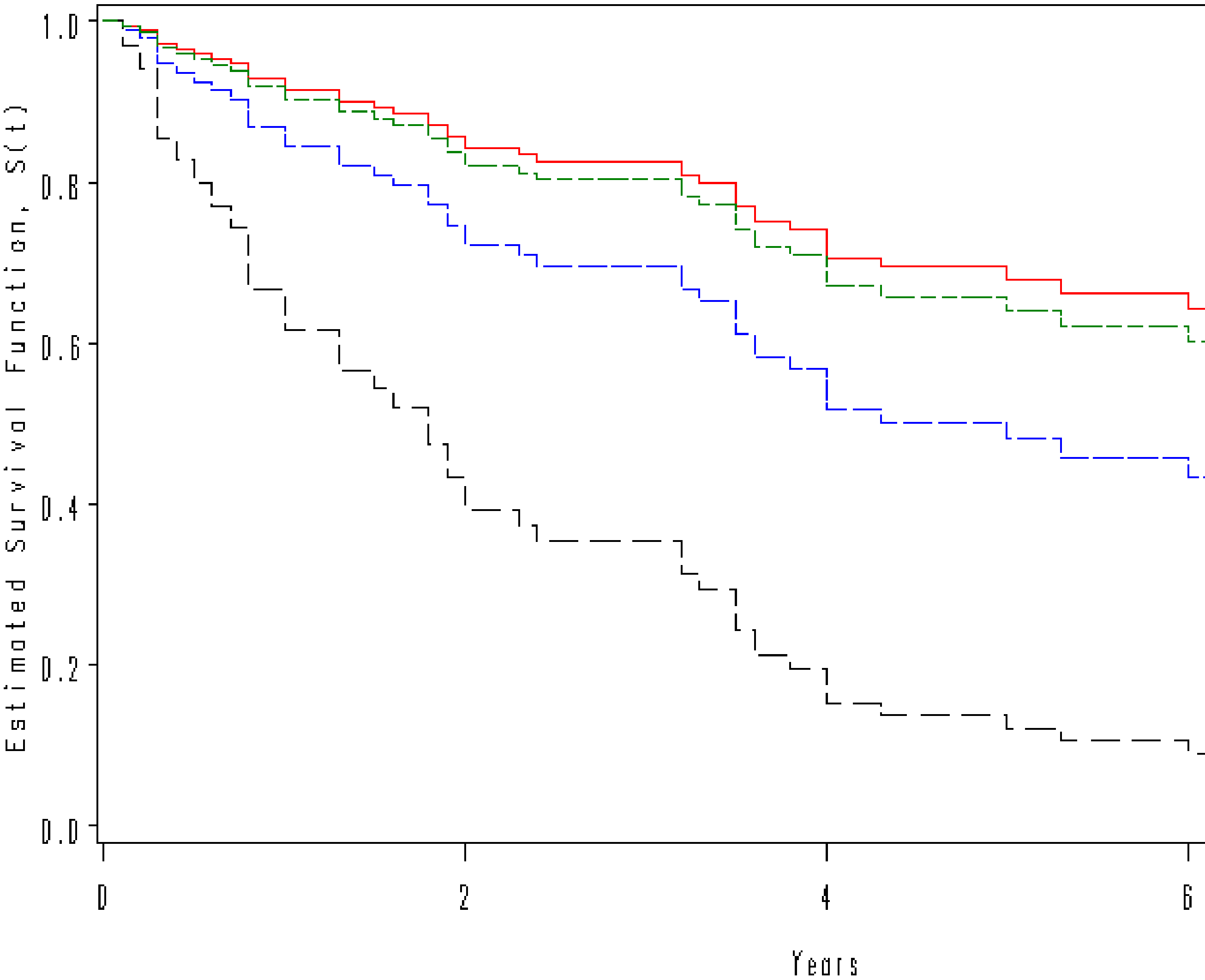

Figure 4.1: Estimated survival functions for a 65 year old larynx cancer patient. Stage1 cancer (red) stage2 (green) stage3 (blue) stage4 (black).

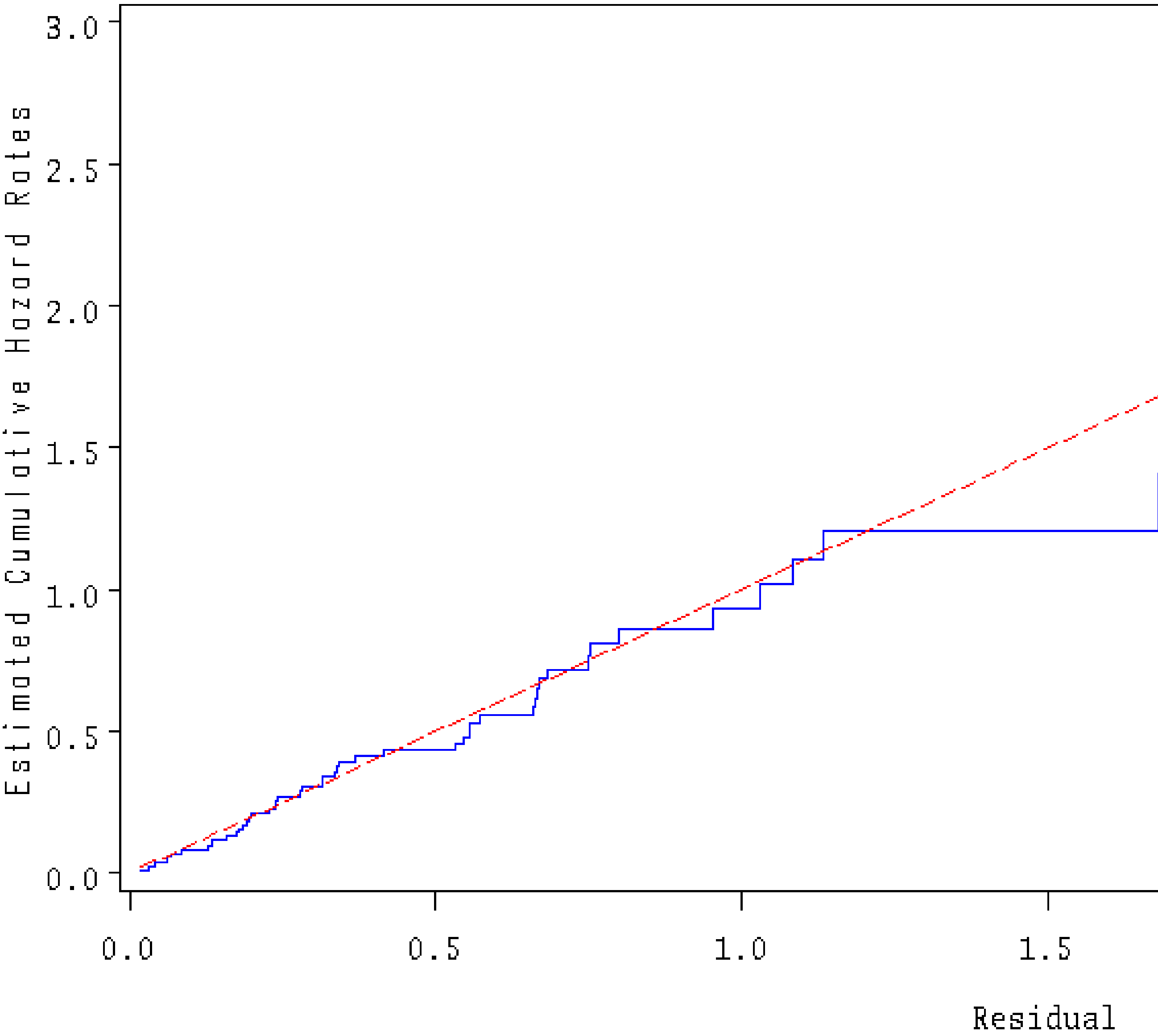

Figure 4.2: Cox-Snell residual plot

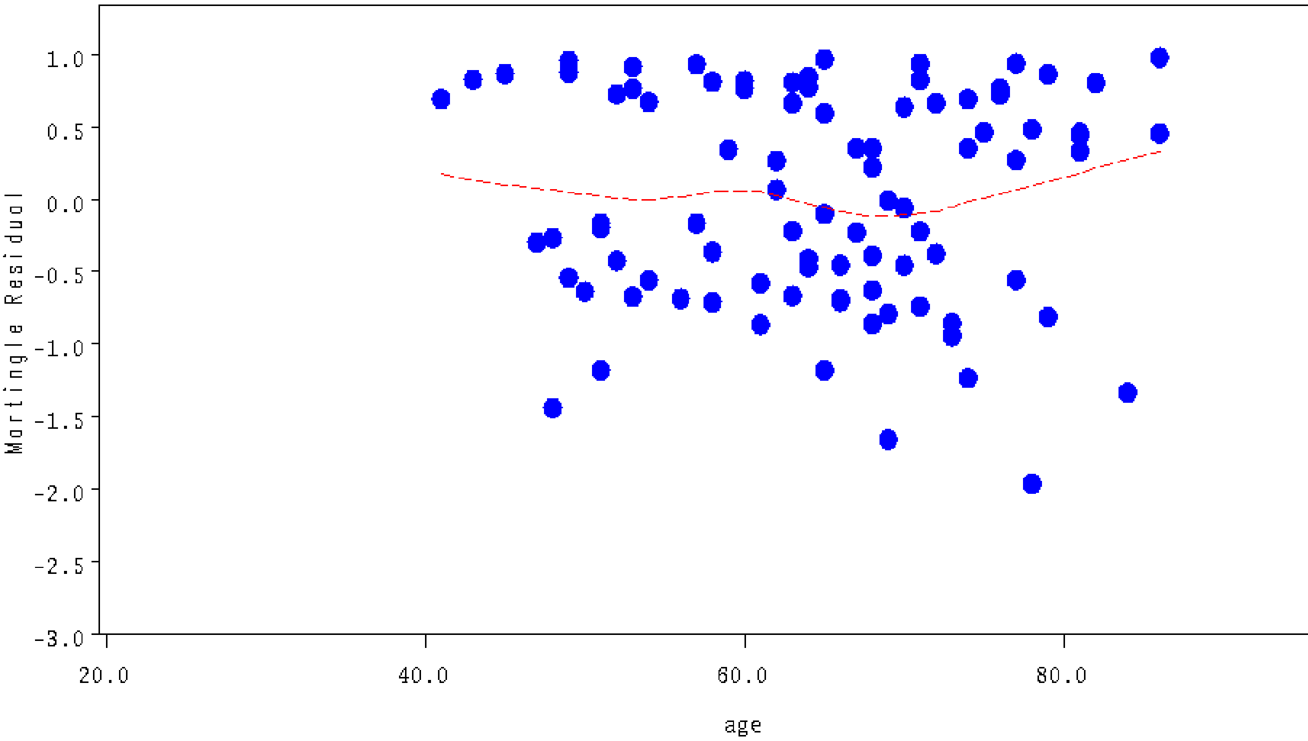

Figure 4.3: Martingale residuals plot

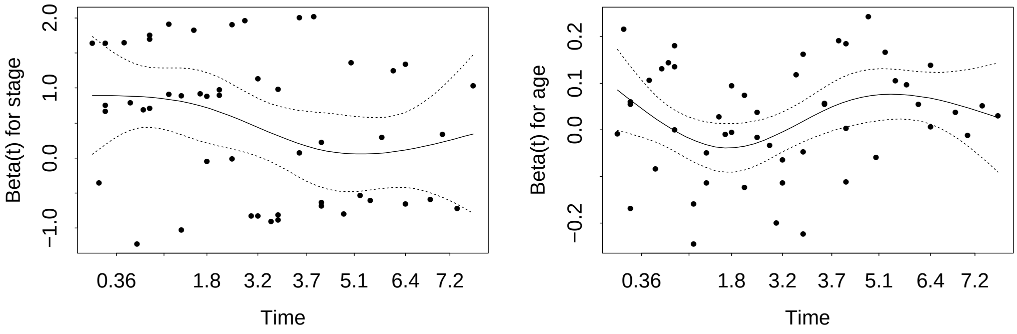

Figure 4.4: Scaled Schoenfeld Residuals

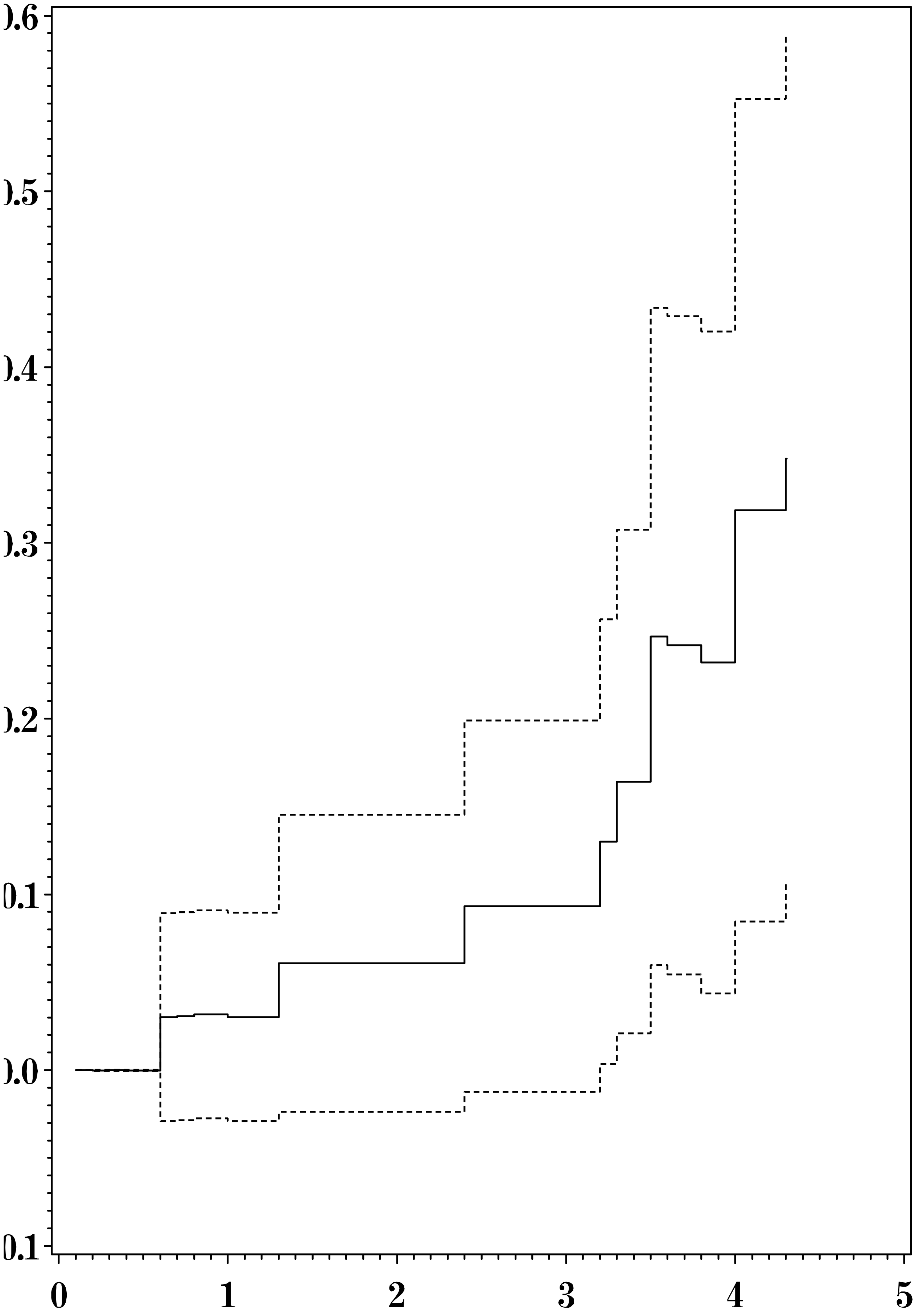

Figure 4.5: Baseline cumulative hazard function

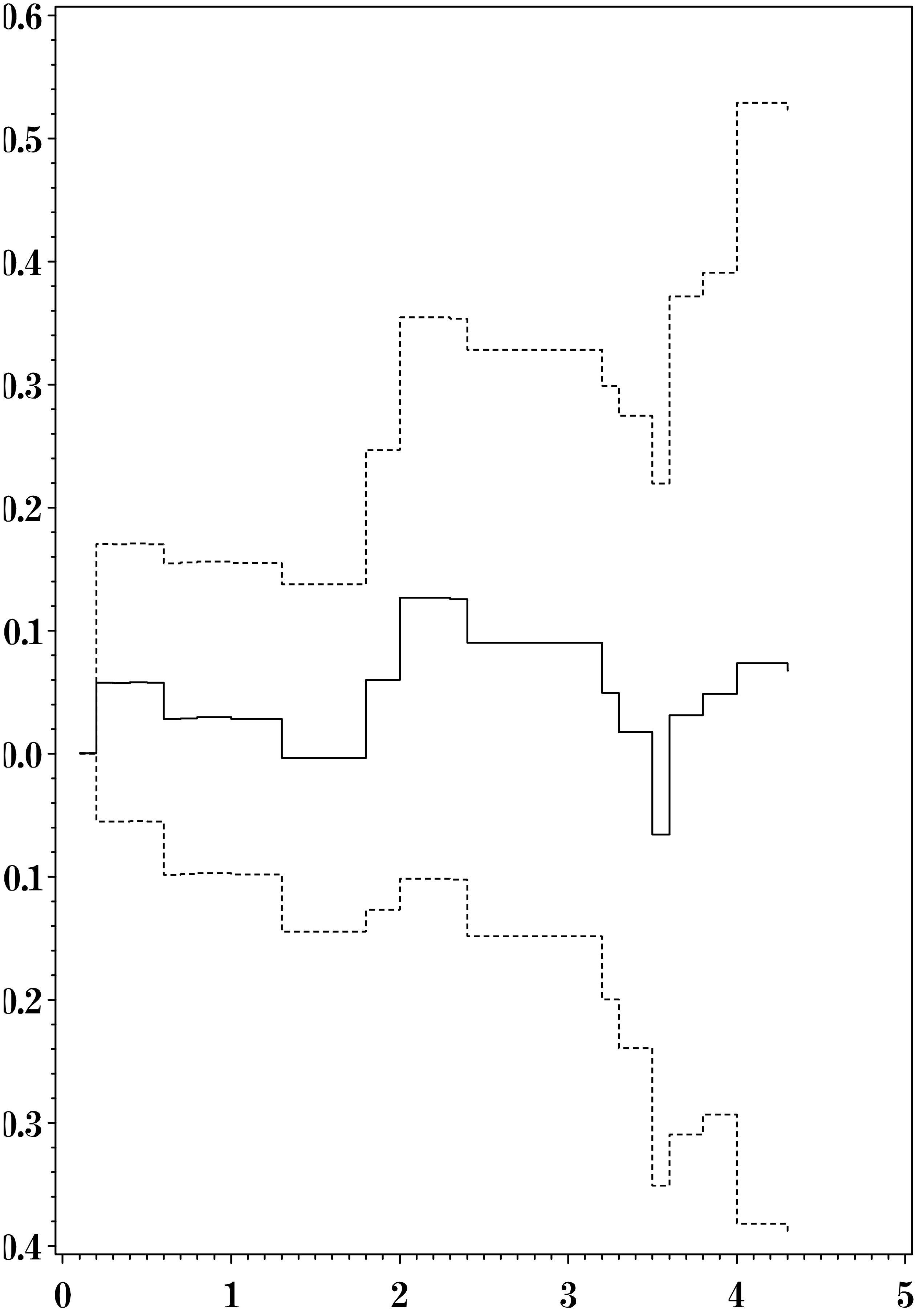

Figure 4.6: Cumulative regression coefficient for stage2

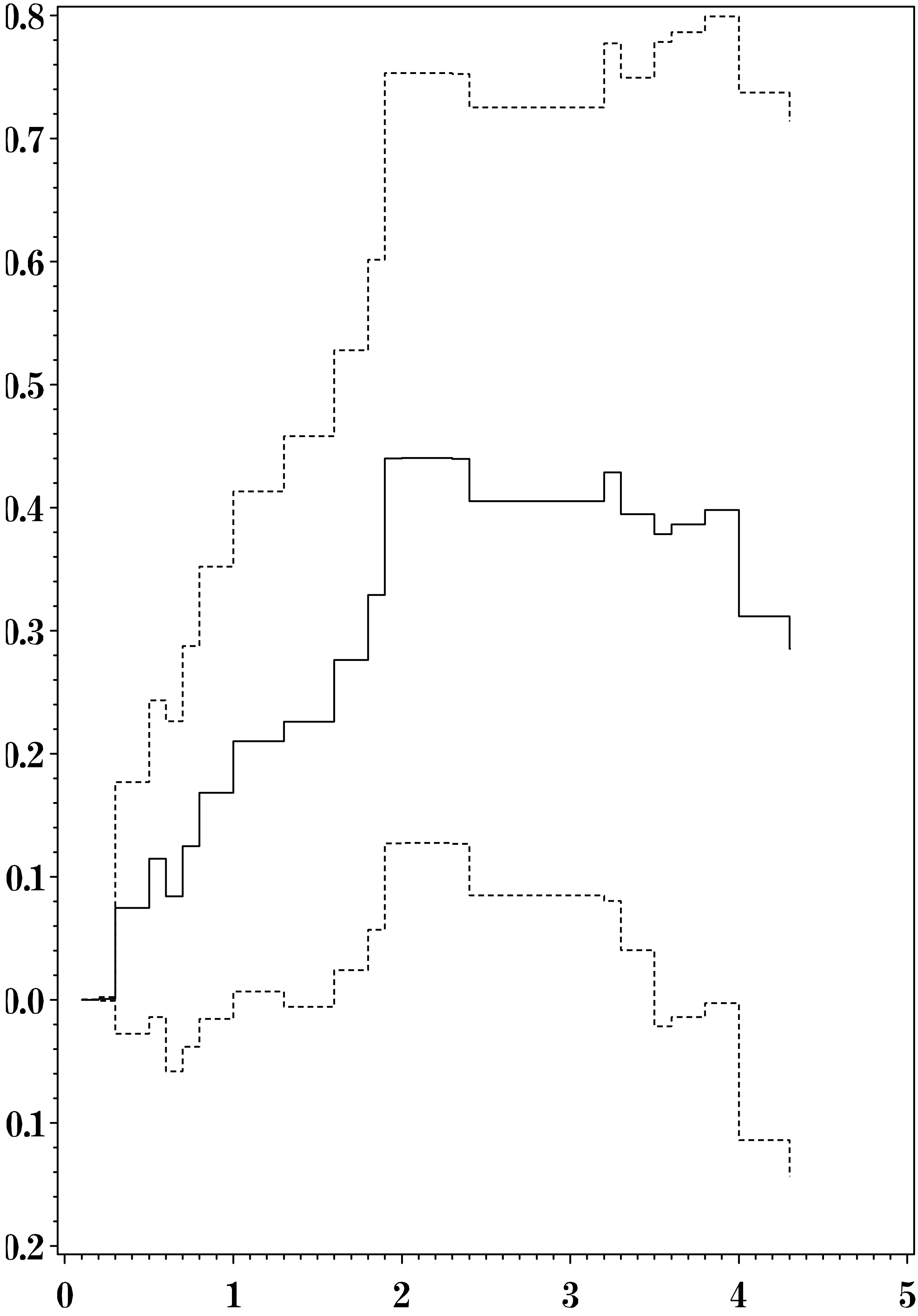

Figure 4.7: Cumulative regression coefficient for stage3

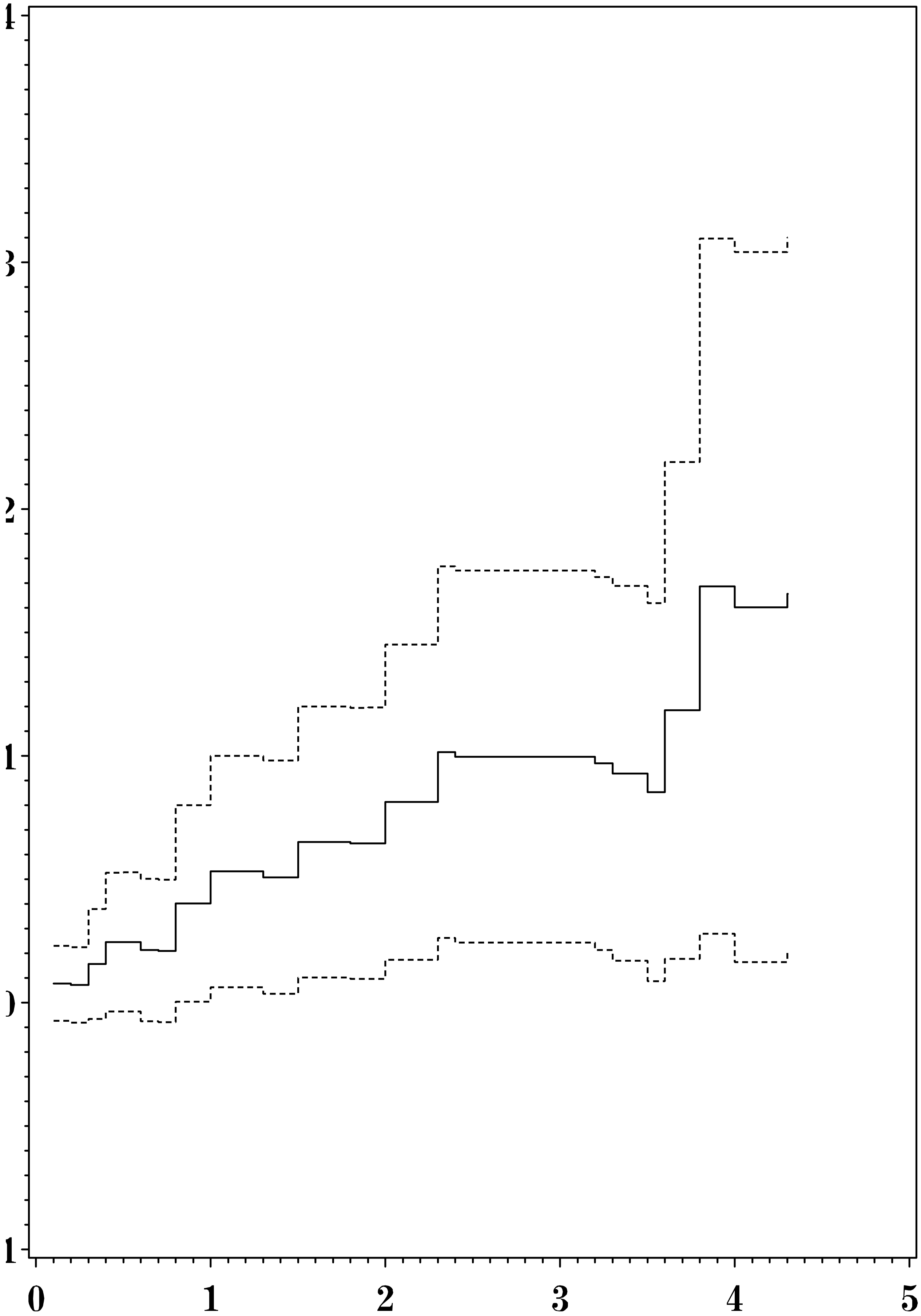

Figure 4.8: Cumulative regression coefficient for stage4

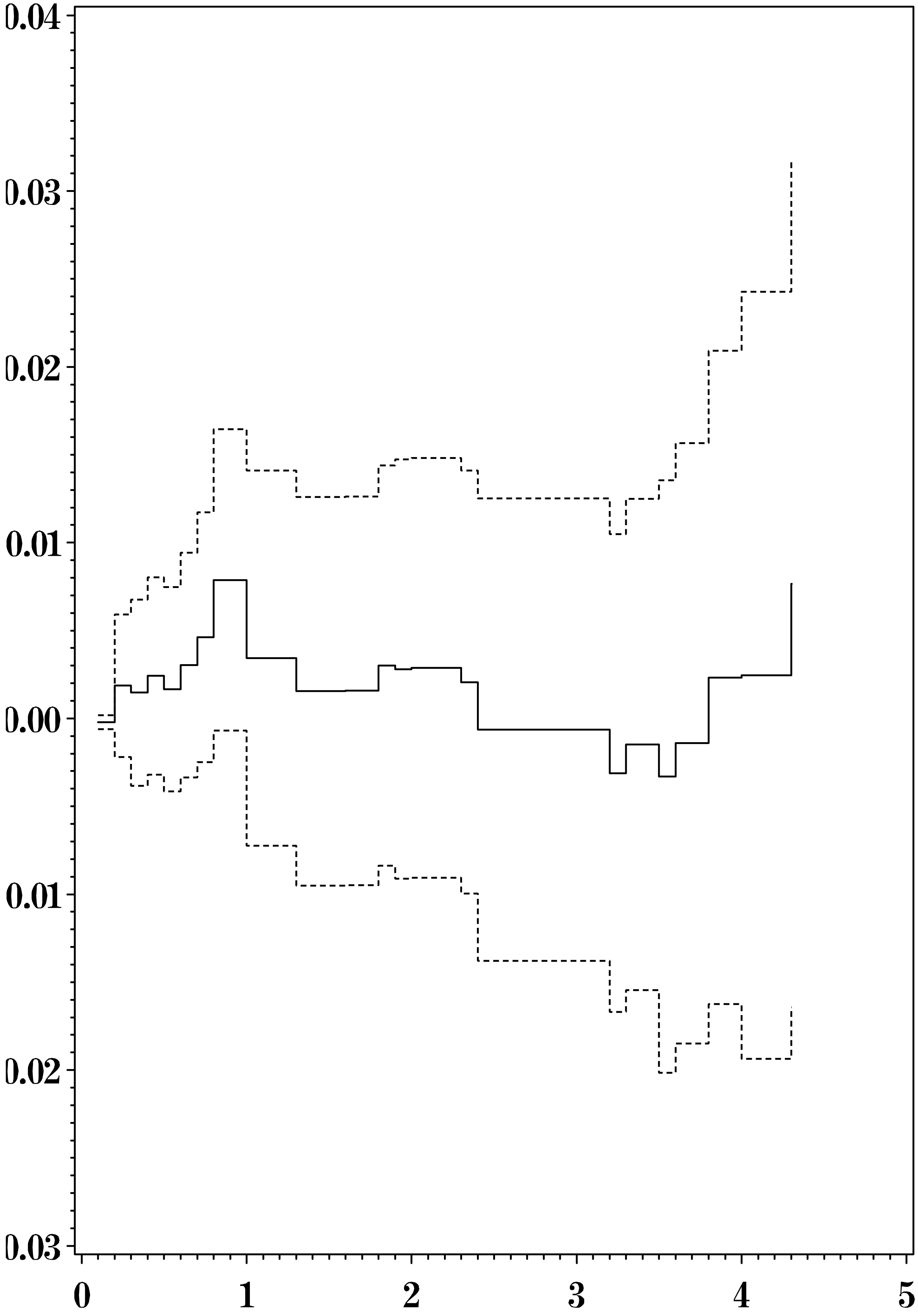

Figure 4.9: Cumulative regression coefficient for age

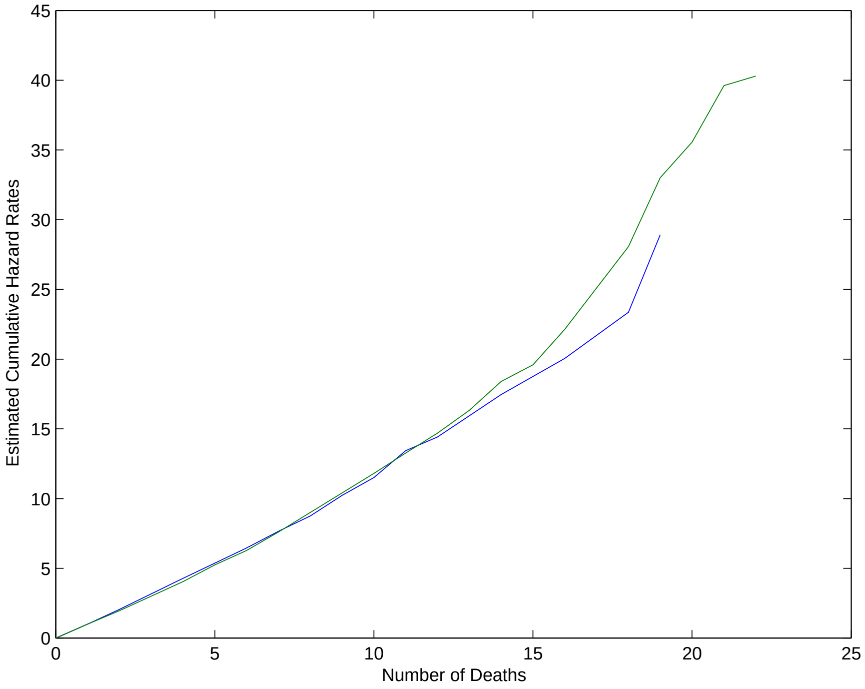

Figure 4.10: Arjas plot to check the adequacy of the additive hazard model. Age less than 64.11 (blue), age greater than 64.11 (green)

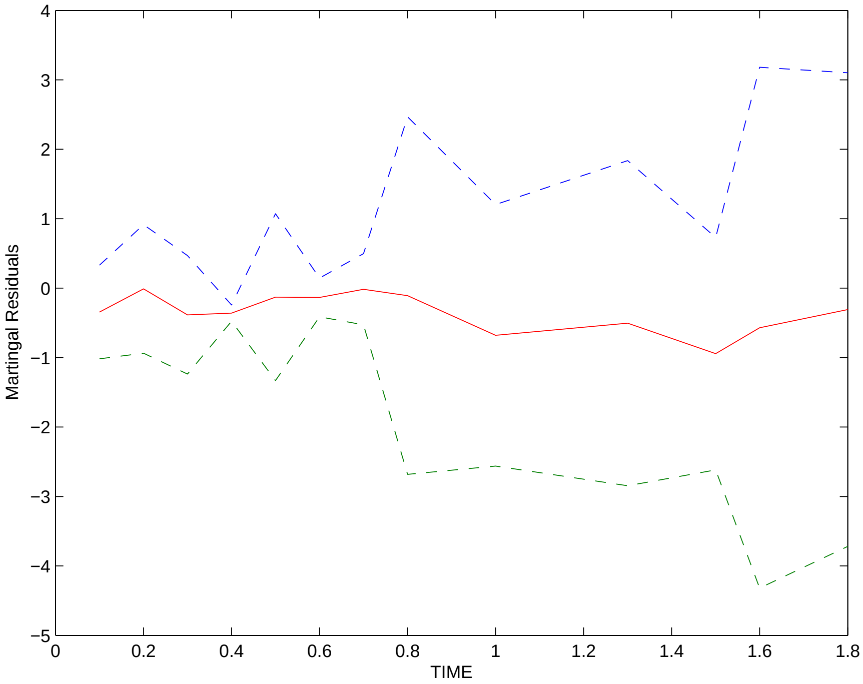

Figure 4.11: Martingale residual process plot and 95% pointwise confidence limits for the under 64.11 age group.

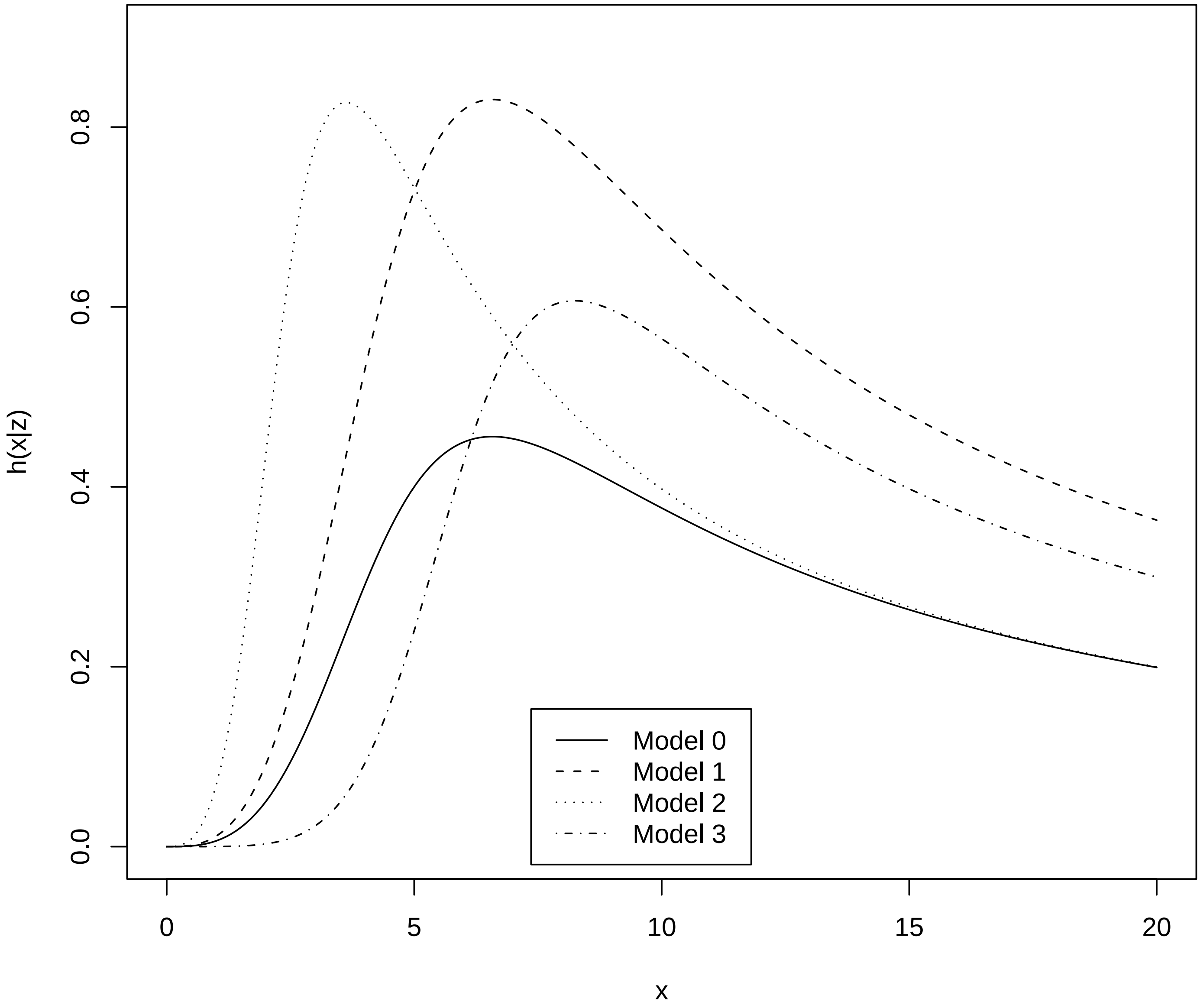

Figure 4.12: Hazards function for all models ($z = 0$ or $z = 1$)

Chapter 5

Conclusion

A problem frequently faced by applied statisticians is the analysis of time to event data. Survival analysis is just another name for time to event analysis. It is being extensively used in clinical trials, biological and epidemiological studies, engineering, finance and social sciences. The analysis of survival experiments is complicated by issues of censoring. In this thesis, we focused on right censored data since this type of data is most frequently encountered in applications.

In Chapter 1, we briefly introduced basic concepts and terminology of survival analysis and how to construct likelihood functions for censored and truncated data. These formed the theoretical basis for following chapters.

In Chapter 2, we presented a detailed discussion of the Cox proportional hazards model through theory and application. It included the Cox model partial likelihood construction, hypothesis tests, and discussions of a variety of residual plots one can make to check the fit of the Cox model.

In Chapter 3, we presented a detailed discussion of Aalen's additive hazards regression model, which may be the model of choice in situations where the proportional hazards model is not available. It included estimation of the cumulative regression functions as well as standard deviations and confidence intervals. Testing the hypothesis of no regression effect for one or more covariates can be done as well as tests of contrasts, and we showed how to assess model fit by using Arjas plots and martingale residual plots.

Chapter 4 was the main element in this thesis. Our interest was to investigate and compare the use of the these two models, and draw some conclusions from this analysis.

Survival data from a study of 90 male laryngeal cancer patients was analyzed using these two models. We presented a detailed discussion of their common and different characters in Section 4.4. We briefly restate their mainly different characters.

1. The Cox model assumes multiplicative effects of risk factors on the hazard function. Covariate effects do not change over time. However, Aalen's additive model assumes that covariates act in an additive manner, risk coefficients are allowed to be functions of time so that the effect of a covariate may vary over time.

2. For the Cox model, statistical software is available and easy to use to fit models, check model assumptions and assess model fit. For Aalen's additive model, standard procedures is not available in any commonly used computer package, such as SAS or S-plus. In this thesis, we used a SAS marco that performs the additive hazards regression. It provides graphical summaries of the covariate effects and tests the hypothesis of no covariate effect. We had to write MATLAB code to make residual plots to assess the fit of the additive hazards model.

3. For the Cox model, if covariates are deleted from a model, regression coefficients for other covariates may change. However, for the additive hazards model, the regression model with this covariate eliminated is the same as the regression model with this covariate included.

4. In practice, we found that it is not appropriate to use Aalen's additive hazards model for all datasets, but it is not a problem with the Cox model.

Finally, we used Monte Carlo simulation to study the power and robustness of tests based on the Cox and additive hazards models. For a Cox model, we find that the test statistic based on Cox model has slightly greater power than the test statistic based on the additive hazards model. For an additive hazards model, we find that the test statistic based on the additive hazards model has slightly greater power than the test statistic based on Cox model. These had to be expected.

References

Aalen, O.O. (1980). *A model for nonparametric regression analysis of counting processes*, Lecture Notes In Statistics 2: 1-25.

Aalen, O.O. (1989). *A Linear Regression Model for the Analysis of Life Times*, Statistic in Medicine, 8: 907-925.

Allison, P.D. (1995). *Survival Analysis Using the SAS System*, SAS Institute, Cary, NC.

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

Duchesne, Thierry (2003). *Note de cours: Analyse des durées de vie.*

Hosmer, D.W. and Lemeshow, S. (1999). *Applied Survival Analysis (Regression Modeling of Time to Event Data)*, John Wiley & Sons, Inc., New York.

Howell, A. (1996). *A SAS Macro for the Additive Regression Hazards Model*, Master's Thesis, Medical College of Wisconsin, Milwaukee, Wisconsin.

Huffler, F.W. and McKeague, I.W. (1991). *Weighted least squares estimation for Aalen's additive risk model*, Journal of the American Statistical Association, 86: 114-129.

Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data (Second Edition)*, John Wiley & Sons, Inc., Hoboken, New Jersey.

Kim, J.H. and Lee, S.Y. (1996). *A goodness-of-fit test based on martingale residuals for the additive risk model*, The Korean Journal of Applied Statistics, 9: 75-89.

Klein, J.P. and Moeschberger, M.L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*, Springer-Verlag, New York.

Klein, J.P. and Moeschberger, M.L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data (Second Edition)*, Springer-Verlag, New York.

Lee, E. T. (1992). *Statistical methods for Survival Data Analysis*, John Wiley & Sons, Inc., New York.

Tableman, M. and Kim, J. S. (2004). *Survival Analysis Using S (Analysis of Time-to-Event Data)*, CRC Press, New York.

Therneau, T.M. and Foundation, M. (1999). *A Package for Survival Analysis in S*, Technical Report.

Therneau, T.M. and Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*, Springer-Verlag, New York.

Torner, A. (2004). *Proportional hazards and additive regression analysis of survival for severe breast cancer*, Stockholm University.

Appendix A

Programs

Chapter 1:

SAS program shown below:

```
data exemple;

infile “/mat/usr/hcao/thesis/hmohiv.dat” firstobs=2;

input g TIME AGE CENSOR;

run;

proc lifereg data=exemple;

model TIME*CENSOR(0) = AGE;

run;
```

Chapter 2:

SAS program 2.1 shown below:

```
data exemple;

infile “/mat/usr/hcao/thesis2/bone.dat” firstobs=2;
```

```
input g t1 t2 d1 d2 d3 tA dA tC dC tP dP z1 z2 z3 z4 z5 z6 z7 z8 z9 z10;

if g=2 then AMLL=1; Else AMLL=0;

if g=3 then AMLH=1; Else AMLH=0;

run;

proc phreg data=exemple;

model t2*d3(0) = AMLL AMLH zA;

if ( t2 >= tA and dA=1) then zA=1; else zA=0;

run;

proc phreg data=exemple;

model t2*d3(0) = AMLL AMLH zC;

if ( t2 >= tC and dC=1) then zC=1; else zC=0;

run;

proc phreg data=exemple;

model t2*d3(0) = AMLL AMLH zP;

if ( t2 >= tP and dP=1) then zP=1; else zP=0;

run;
```

Splus program 2.2 shown below:

```
lym < − read.table(‘/mat/usr/hcao/thesis2/true1.dat’, header=T,

col.names=c(‘NUMBER’,‘GROUP’,‘SEX’,‘AGE’,‘STATUS’,‘DEATHTIME’,

‘SCORE’))
```

```
lym[1:5,1:7]

NUMBER  GROUP  SEX  AGE  STATUS  DEATHTIME  SCORE
   1      1     0    41    1        0.73      75
   2      1     1    61    1        1.06      50
   3      1     0    43    1        2.48      90
   4      1     0    18    1        3.38     100
   5      1     0    37    0        8.81      95

attach(lym)

lym2 < − coxph(Surv(DEATHTIME,STATUS)∼ SCORE+GROUP+SEX+AGE)

rd < − abs(STATUS - lym2$residuals)

km.rd < − survfit(Surv(rd,STATUS) ∼ 1)

summary.km.rd < − summary(km.rd)

rcd < − summary.km.rd$time

surv.rc < − summary.km.rd$surv

plot(rcd, -log(surv.rc), type=‘p’, pch=‘.’, xlab=‘cox-snell residual rd’, ylab=‘cumulative hazard on rd’)
```

Splus program 2.3 shown below:

```
lym2 < − coxph(Surv(DEATHTIME,STATUS)∼ SCORE+GROUP+SEX+AGE)

scatter.smooth(lym$SCORE,resid(lym2),type=‘P’,pch=‘.’, xlab=‘SCORE’,

ylab=‘martingale residual’)
```

Splus program 2.4 shown below:

```
detail < − coxph.detail(lym2)

time < − detail$y[,2]
```

```
status < − detail$y[,3]

sch < − resid(lym2, type=‘schoenfeld’)

plot(time[status==1], sch[,1], xlab=‘ordered survival time’, ylab=‘schoenfeld residuals for SCORE’)
```

Splus program 2.5 shown below:

```
lym2 < − coxph(Surv(DEATHTIME,STATUS)∼ SCORE+GROUP+SEX+AGE)

PH.test < − cox.zph(lym2)

PH.test

par(mfrow = c(3,2))

plot(PH.test)
```

Splus program 2.6 shown below:

```
par(mfrow = c(3,2))

bresid < − resid(lym2, type=‘dfbetas’)

index < − seq(1:58)

plot (index,bresid[,1],type=‘h’,ylab=‘scaled change in coef’,xlab=‘observation’)

plot (index,bresid[,2],type=‘h’,ylab=‘scaled change in coef’,xlab=‘observation’)

plot (index,bresid[,3],type=‘h’,ylab=‘scaled change in coef’,xlab=‘observation’)

plot (index,bresid[,4],type=‘h’,ylab=‘scaled change in coef’,xlab=‘observation’)
```

Splus program 2.7 shown below:

```
name < − read.table(‘/mat/usr/hcao/thesis2/name.txt’,

header=T, col.names=c(‘week’,‘arrest’,‘fin’,‘age’,‘race’,‘wexp’,‘mar’,‘paro’,‘prio’))
```

```
name[1:5,1:9]

name2 < − coxph(Surv(week,arrest)∼ fin + age + race + wexp + mar + paro + prio, data=name)

summary(name2)
```

Splus program 2.8 shown below:

```
name4 < − coxph(Surv(week,arrest) ∼ fin + age + prio, data=name)

name4

cox.zph(name4)

par(mfrow = c(2,2))

plot(cox.zph(name4))
```

Splus program 2.9 shown below:

```
dfbeta < − residuals(name4, type = ‘dfbeta’)

par(mfrow = c(2,2))

for (j in 1:3) { plot(dfbeta[, j], ylab = names(coef(name4))[j])

    abline(h=0,lty=2)}
```

Splus program 2.10 shown below:

```
par(mfrow = c(2,2))

res < − residuals(name4, type = ‘martingale’)

x < − as.matrix(name[, c(‘age’,‘prio’)])

par(mfrow = c(2,2))

for (j in 1:2) { plot(X[,j], res, xlab = c(‘age’,‘prio’)[j], ylab = ‘residuals’)
abline(h=0,lty=2) lines(lowess(X[,j], res, iter = 0))}
```

Chapter 3:

SAS macro program 3.1 shown below:

```
data exemple;

infile “/mat/usr/hcao/thesis2/uisnew.txt” firstobs=2;

input TIME CENSOR AGE BECKTOTA NDRUGTX IVHX RACE TREATE SITE;

run;

%include ‘/mat/usr/hcao/thesis2/additive.sas’;

proc sort;

by TIME;

run;

proc iml;

option = { n, n, y, n, n };

effects ={‘AGE’, ‘BECKTOTA’, ‘NDRUGTX’, ‘IVHX’, ‘RACE’, ‘TREATE’, ‘SITE’};

timenuit = {‘Years’};

% additive (exemple, 0.05, timenuit, effects, option, dummy1, beta, dummy2);

quit;

filename gsasfile ‘graph1.ps’;

goptions reset=global gunit=pct htext=2.5 ftext=centb

gsfmode=replace device=ps gaccess=gsasfile
```

```
gprolog=‘25210D0A’xgepilog=‘04’x;

symboll c=black v=none i=step1j;

symbol2 c=black v=none i=step1j l=3;

symbol3 c=black v=none i=step1j l=3;

axis1 label=(‘Baseline cumulative hazard function’);

axis2 label=(a=90 ‘cumulative regression coefficient’);

proc gplot data=beta;

footnote ‘Figure 1.1’;

plot col2*col1 col18*col1 col26*col1 /overlay haxis=asix1 vaxis=axis2;

run;
```

Note:

*col1 = time;

*col2 - col9 = estimate of AGE, BECKTOTA, NDRUGTX, IVHX, RACE, TREATE, SITE over time;

*col10 - col7 = standard deviation of variables above over time;

*col18 -col25 = lower confidence limit of variables above over time;

*col26 -col33 = upper confidence limit of variables above over time;

The graph resulting from the program above is shown in Figure 3.1. It estimates the baseline cumulative regression function. The parameter estimates for AGE, BECKTOTA, NDRUGTX, IVHX, RACE, TREATE, SITE were also graphed (see Figures 3.2 - 3.8). The SAS code for these graphs is similar to program above.

MATLAB program 3.2 shown below:

Arjas plot:

- main program: data1

```
clear;
dataAge = sortrows(data111, 3);
data1Age = dataAge(1:292, :);
data2Age = dataAge(293:575, :);
data11 = sortrows(data1Age, 1);
data12 = sortrows(data2Age, 1);
NH = Aalen1(data11);
N = NH(1, :);
H = NH(2, :);
NH = Aalen1(data12);
N1 = NH(1, :);
H1 = NH(2, :);
plot(N, H, ‘-’, N1, H1, ‘–’);
```

- function program: Aalen1

```
function NH = Aalen1(lary11)
ti = lary11(:, 1);
sigma = lary11(:, 2);
z = lary11(:, 3:9);
b = zeros(size(z,2) + 1, 1);
flag = 1;
for i = 1:60;
N(i) = CalcuN(ti(i), ti, sigma);
if i == 1
x = xti(ti(i), ti, z);
b = inv(x’ * x) * x’ * Iti(ti(i), ti, sigma);
else
if ti(i)  = ti(i-1)
x = xti(ti(i), ti, z);
```

```
b = b + inv(x' * x) * x' * Iti(ti(i), ti, sigma);
end;
end;
h = b(1) + z * b(2: length(b));
H(i) = sum(h);
end;
N1 = [0 N];
H1 = [0 H];
NH = [N1;H1];
```

- function program: CalcuN

```
function N = CalcuN(t, ti, sigma)
N1 = t ≥ ti & sigma;
N = sum(N1);
```

- function program: xti

```
function matrixx = xti(t, ti, z)
[n1 n2] = size(z);
matrixx = [ ];
for i = 1: n1;
if t <= ti(i)
matrixx = [matrixx; [1 z(i,:)]];
else
matrixx = [matrixx; zeros(1, n2+1)];
end;
end;
```

- function program: Iti

```
function allayx = Iti(t, ti, sigma)
la = length(sigma);
for i = 1: la;
if t == ti(i) & sigma(i) == 1
allayx(i) = 1;
```

```
else
allayx(i) = 0;
end;
end;
allayx = allayx’;
```

Martingale residual plot:

- main program: datacov

```
clear;
dataForCov1 = sortrows(data111, 1);
ti = dataForCov1(:, 1);
for i = 1: 60;
M = AalenCov(dataForCov1, i);
MM = M(1, :);
MM1(i) = MM;
CovMres = AalenCovMres(dataForCov1, i);
CC = CovMres(1, 1);
MT1(i) = MM + 1.96 * sqrt(CC);
MT2(i) = MM - 1.96 * sqrt(CC);
TT(i) = ti(i);
end;
plot(TT, MT1, ‘–’, TT, MT2, ‘–’, TT, MM1);
```

- function program: AalenCov

```
function Mres = AalenCov(lary11,t)
ti = lary11(:, 1);
sigma = lary11(:, 2);
z = lary11(:, 3:9);
Q = FunctionQ(lary11,32);
b = zeros(size(z,2)+1,1);
flag = 1;
```

```
N = ti(t) >= ti & sigma;
N1 = sum(ti(t) >= ti);
for i = 1:N1;
if i == 1
x = xti(ti(i), ti, z);
b = inv(x' * x) * x' * Iti(ti(i), ti, sigma);
else if ti(i)  = ti(i-1)
x = xti(ti(i),ti,z);
b = b + inv(x' * x) * x' * Iti(ti(i),ti,sigma);
end;
end;
end;
h = b(1) + z * b(2: length(b));
M = N - h;
Mres = Q' * M;
```

- function program: FunctionQ

```
function Q1 = FunctionQ(SetA,ValueAge)
[n1 n2] = size(SetA);
Q1 = [ ];
AgeSet = SetA(:,3);
for i = 1: n1;
if AgeSet(i) <= ValueAge
Q1 = [Q1; [1 0] ];
else
Q1 = [Q1; [0 1] ];
end;
end;
```

- function program: AalenCovMres

```
function CovMres = AalenCovMres(lary11,t)
[n1 n2] = size(lary11);
ti = lary11(:,1);
```

```
sigma = lary11(:,2);
z = lary11(:,3:9);
Q = FunctionQ(lary11,32);
flag = 1;
N1 = sum(ti(t) >= ti);
for i = 1:N1;
if i == 1
x = xti(ti(i), ti, z);
y = eye(n1) - x * inv(x’ * x) * x’;
CovMres = Q’ * y * FunctionD(lary11,t) * y’ * Q;
else
if ti(i)  = ti(i-1)
x = xti(ti(i), ti, z);
y = eye(n1) - x * inv(x’ * x) * x’;
CovMres = CovMres + Q’ * y * FunctionD(lary11,t) * y’ * Q;
end;
end;
end;
```

- function program: FunctionD

```
function D = FunctionD(lary11,t)
ti = lary11(:,1);
sigma = lary11(:,2);
z = lary11(:,3:9);
[n1 n2] = size(lary11);
D = zeros(n1);
for i = 1:n1;
if ti(t) == ti(i) & sigma(i) D(i,i) = 1; end;
end;
```

Chapter 4:

SAS program 4.1 shown below:

```
data larynx;
infile “/mat/usr/hcao/thesis2/lary.dat” firstobs=2;
input stage time age dyear death;
stage2=0; if stage=2 then stage2=1;
stage3=0; if stage=3 then stage3=1;
stage4=0; if stage=4 then stage4=1;
RUN;
data age65;
input age stage2 stage3 stage4;
cards;
65 0 0 0
65 0 1 0
65 0 0 1
65 1 0 0
; run;
proc phreg data = larynx;
model time*death(0) = age stage2 stage3 stage4;
baseline out = surv65 survival = survival lower = slower upper = supper
covariates = age65 /method = ch nomean cltype = loglog ;
```

```
run;

proc print data = surv65 noobs;

where time = 6;

run;

proc sort data= surv65;

by time;

run;

proc transpose data = surv65 out = fig8.2 (drop=.name. .label.) prefix = s;

by time;

var survival;

run;

symbol1 i = stepjl l = 1 c=red ;

symbol2 i = stepjl l = 4 c=blue ;

symbol3 i = stepjl l = 21 c=black ;

symbol4 i = stepjl l = 29 c=green ;

axis1 order = (0 to 1 by .2) label=(a= 90 ‘Estimated Survival Function, S(t)’) minor = none;

axis2 order = (0 to 8 by 2) label = (‘Years’) minor = none;

proc gplot data = fig8.2;

plot s1*time s2*time s3*time s4*time /overlay vaxis = axis1 haxis = axis2;

run;
```

```
quit;
```

SAS program 4.2 shown below:

```
data larynx;
infile “/mat/usr/hcao/thesis2/lary.dat” firstobs=2;
input stage time age dyear death;
stage2=0; if stage=2 then stage2=1;
stage3=0; if stage=3 then stage3=1;
stage4=0; if stage=4 then stage4=1;
run;
proc phreg data = larynx;
model time*death(0) = age stage2 stage3 stage4;
output out = figure11.1 LOGSURV = h;
run;
data figure11.1a;
set figure11.1;
h = -h;
cons = 1;
run;
proc phreg data = figure11.1a ;
model h*death(0) = cons;
```

```
output out = figure11.1b logsurv = ls /method = ch;

run;

data figure11.1c;

set figure11.1b;

haz = - ls;

run;

proc sort data = figure11.1c;

by h;

run;

axis1 order = (0 to 3 by .5) minor = none;

axis2 order = (0 to 3 by .5) minor = none label = ( a=90);

symbol1 i = stepjl c= blue;

symbol2 i = join c = red l = 3;

proc gplot data = figure11.1c;

plot haz*h =1 h*h =2 /overlay haxis=axis1 vaxis= axis2;

label haz = “Estimated Cumulative Hazard Rates”;

label h = “Residual”;

run;

quit;
```

SAS program 4.3 shown below:

1. Martingale residual plot

```
data larynx;

infile “/mat/usr/hcao/thesis2/lary.dat” firstobs=2;

input stage time age dyear death;

stage2=0; if stage=2 then stage2=1;

stage3=0; if stage=3 then stage3=1;

stage4=0; if stage=4 then stage4=1;

run;

proc phreg data = larynx;

model time*death(0) = age stage2 stage3 stage4;

output out = figure11.6 RESMART = mgale ;

run;

proc loess data=figure11.6;

ods output OutputStatistics=figure11.6a;

model mgale = age / smooth=0.6 direct;

run;

quit;

proc sort data = figure11.6a;

by age;

run;
```

```
axis1 order = (-3.0 to 1 by .5) offset = (0, 2) label= (a = 90) minor = none;

axis2 order = (20 to 100 by 20) minor = none;

symbol1 i = none v = dot h = 1.5 c = blue;

symbol2 i = join v = none c = red;

proc gplot data = figure11.6a;

format depvar f4.1;

format age f4.1;

plot depvar* age = 1 pred* age = 2 /haxis = axis2 vaxis = axis1 overlay;

label depvar = “Martingle Residual”;

label age = “age”;

run;

quit;
```

2. Log Partial Likelihood

```
data larynx;

infile “/mat/usr/hcao/thesis2/lary.dat” firstobs=2;

input stage time age dyear death;

run;

proc sql noprint;

select distinct age into :event.time separated by ’ ’

from larynx;
```

```
quit;
ods listing close;
proc phreg data= & data;
model &time*&censor(0) = &var z2;
if &time ¿ &dep then z2 = &var;
else z2 = 0;
ods output FitStatistics = .temp&k;
run;
ods output close;
ods listing;
data whole;
set &whole;
if Criterion = ”-2 LOG L” ;
run;
data whole;
set whole;
e.time = scan(“&delta.list”, .n., ’ ’);
logp = - withcovariates/2;
run;
proc print data = whole noobs;
```

```
var e.time logp;

run;
```

SAS program 4.4 shown below:

```
data exemple;

infile “/mat/usr/hcao/thesis2/larytry3.dat” firstobs=2;

input time death stage2 stage3 stage4 age;

age=age-64.11;

RUN;

proc sort;

by time;

run;

proc iml;

option={ n, n, y, n, n};

effects={‘stage2’,‘stage3’,‘stage4’,‘age’};

timeunit={‘Years’};

quit;

filename gsasfile ‘graph10.eps’;

goptions reset=global gunit=pct htext=2.5 ftext=centb

gsfmode=replace device=ps gaccess=gsasfile

gprolog=‘25210D0A’xgepilog=‘04’x;
```

```
symbol1 c=black v=none i=steplj;

symbol2 c=black v=none i=steplj l=2;

symbol3 c=black v=none i=steplj l=2;

axis1 label=('Baseline cumulative hazard function');

axis2 label=(a=90 'Cumulative regression coefficient');

proc gplot data=beta;

plot col2*col1 col12*col1 col17*col1/overlay haxis=axis1 vaxis=axis2;

run;

* col1 = time;

* col2 - col6 = estimate of A0-A4 over time;

* col7 - col11 = standard deviation of A0-A4 over time;

* col12 - col16 = lower confidence limit of A0-A4 over time;

* col17 - col21 = upper confidence limit of A0-A4 over time;
```

Chapter 5:

1. The R code is as follows:

For model 0:(N=1000, n=50)

```
Simulate < −function ()

{

u < − runif(50)

z < − c(rep(0,25),rep(1,25))
```

fail < − rep(1,50)

x < − ((1-u)$^{-1}$ - 1)$^{1/4}$/0.2

test.stat < − coxph(Surv(x,fail)∼ *z*)\$score

return(test.stat)

}

model1.cox < − replicate(1000,Simulate())

sum(model1.cox >= 3.841)/1000

For model 1:(N=1000, n=50)

Simulate < −function ()

{

u < − runif(50)

z < − c(rep(0,25),rep(1,25))

fail < − rep(1,50)

x < − ((1-u)$^{-exp(-0.6*z)}$-1)$^{1/4}$/0.2

test.stat < − coxph(Surv(x,fail)∼ *z*)\$score

return(test.stat)

}

model1.cox < − replicate(1000,Simulate())

sum(model1.cox >= 3.841)/1000

For model 2:(N=1000, n=50)

Simulate $< -$function ()

{

u $< -$ runif(50)

z $< -$ c(rep(0,25),rep(1,25))

fail $< -$ rep(1,50)

x $< -$ exp(-log(0.2)-0.568245*z+(1/4)*rlogis(50))

test.stat $< -$ coxph(Surv(x,fail)$\sim z$)\$score

return(test.stat)

}

For model 3:(N=1000, n=50)

Simulate $< -$function ()

{

u $< -$ runif(50)

z $< -$ c(rep(0,25),rep(1,25))

fail $< -$ rep(1,50)

x1 $< -$ ((1-u[1:25])$^{-1}$-1)$^{1/4}$/0.2

x2 $< -$ ((1-u[26:50])$^{-1}$-1)$^{1/6}$/0.15874

x $< -$ c(x1,x2)

test.stat $< -$ coxph(Surv(x,fail)$\sim z$)\$score

return(test.stat)

```
}
```

2. The SAS code is as follows:

For model 0:(N=1000, n=200)

```
data exemple;

infile “/mat/usr/hcao/thesis2/larytry3.dat” firstobs=2;

input time death stage2 stage3 stage4 age;

age=age-64.11;

run;

%include “/mat/usr/hcao/thesis2/additive.sas”;

proc sort data=exemple; by time;

proc iml;

option={ n, n, y, n, n };

effects={‘stage2’,‘stage3’,‘stage4’,‘age’};

timeunit=’Years’;

%additive(exemple, 0.05, timeunit, effects, option, dummy1, beta, dummy2);

create stat from gltstat;

append from gltstat;

run; quit;

proc print data=stat; run;

%macro analyse(nsimul=,samplesize=);
```

```
options nonotes; ods listing close;

proc datasets; delete final; run; quit;

% do i=1 % to & nsimul;

% put Simulation # &i;

data simul;

do i=1 to & samplesize;

u = ranuni(0);

if i <= (&samplesize/2) then zz=0; else zz=1;

time = ((1-u)^{-1}-1)^{1/4}/0.2;

death = 1;

z = zz;

keep time death z;

output;

end;

run;

proc sort data=simul; by time;

proc iml;

option={ n, n, y, n, n };

effects={'z'};

timeunit={'Years'};
```

```
%additive(simul, 0.05, timeunit, effects, option, dummy1, beta, dummy2);

create stat from gltstat;

append from gltstat;

quit;

proc append data=stat base=final;

run; quit;

%end;

options notes; ods listing;

%mend;

%analyse(nsimul=1000,samplesize=200);

proc print data=final;

run;
```

For model 1:(N=1000, n=200)

```
%macro analyse(nsimul=,samplesize=);

options nonotes; ods listing close;

proc datasets; delete final; run; quit;

% do i=1 % to & nsimul;

% put Simulation # &i;

data simul;

do i=1 to & samplesize;
```

```
u = ranuni(0);

if i <= (&samplesize/2) then zz=0; else zz=1;
```

time = $((1\text{-u})^{-exp(-0.6*zz)}\text{-}1)^{1/4}/0.2$;

```
death = 1;

z = zz;

keep time death z;

output;

end;

run;
```

For model 2:(N=1000, n=200)

```
%macro analyse(nsimul=,samplesize=);

options nonotes; ods listing close;

proc datasets; delete final; run; quit;

% do i=1 % to & nsimul;

% put Simulation # &i;

data simul;

do i=1 to & samplesize;

u = ranuni(0);

if i <= (&samplesize/2) then zz=0; else zz=1;

time = exp(-log(0.2)-0.568245*zz+(1/4)*log(u/(1-u)));
```

death = 1;

z = zz;

keep time death z;

output;

end;

run;

For model 3:(N=1000, n=200)

%macro analyse(nsimul=,samplesize=);

options nonotes; ods listing close;

proc datasets; delete final; run; quit;

% do i=1 % to & nsimul;

% put Simulation # &i;

data simul;

do i=1 to & samplesize;

u = ranuni(0);

if i <= (&samplesize/2) then zz=0; else zz=1;

if i <=(&samplesize/2) then time = $((1\text{-u})^{-1}\text{-}1)^{1/4}/0.2$;

else time = $((1\text{-u})^{-1}\text{-}1)^{1/6}/0.15874$;

death = 1;

z = zz;

```
keep time death z;

output;

end;

run;
```

Appendix B

Datasets

Chapter 1:

dataset 1.1:

g	TIME	AGE	CENSOR
1	5	46	1
2	6	35	0
3	8	30	1
4	3	30	1
5	22	36	1
6	1	32	0
7	7	36	1
8	9	31	1
9	3	48	1
10	12	47	1
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.

Chapter 2:

dataset 2.1:

NUMBER	GROUP	SEX	AGE	STATUS	DEATHTIME	SCORE
1	1	0	41	1	0.73	75
2	1	1	61	1	1.06	50
3	1	0	43	1	2.48	90
4	1	0	18	1	3.38	100
5	1	0	37	0	8.81	95
6	1	0	42	1	0.21	80
7	1	0	53	1	0.21	50
8	1	1	41	1	0.6	40
9	1	0	25	1	0.44	95
10	1	0	30	1	1.17	100
11	1	1	67	1	1.38	45
12	1	0	45	1	1.08	50
13	1	1	59	1	2.17	90
14	1	0	42	1	0.17	80
15	1	1	66	1	0.23	50
14	1	0	42	1	0.17	80
15	1	1	66	1	0.23	50
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

dataset 2.2:

week	arrest	fin	age	race	wexp	mar	paro	prio
20	1	0	27	1	0	0	1	3
17	1	0	18	1	0	0	1	8
25	1	0	19	0	1	0	1	13
52	0	1	23	1	1	1	1	1
52	0	0	19	0	1	0	1	3
52	0	0	24	1	1	0	0	2
23	1	0	25	1	1	1	1	0
52	0	1	21	1	1	0	1	4
52	0	0	22	1	0	0	0	6
52	0	0	20	1	1	0	0	0
52	0	1	26	1	0	0	1	3
52	0	0	40	1	1	0	0	2
37	1	0	17	1	1	0	1	5
52	0	0	37	1	1	0	0	2
25	1	0	20	1	0	0	1	3
46	1	1	22	1	1	0	1	2
28	1	0	19	1	0	0	0	7
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

Chapter 3:

dataset 3.1:

TIME	CENSOR	AGE	BECKTOTA	NDRUGTX	IVHX	RACE	TREAT	SITE
188	1	39	9	1	3	0	1	0
26	1	33	34	8	2	0	1	0
207	1	33	10	3	3	0	1	0
144	1	32	20	1	3	0	0	0
551	0	24	5	5	1	1	1	0
32	1	30	32.55	1	3	0	1	0
459	1	39	19	34	3	0	1	0
22	1	27	10	2	3	0	1	0
210	1	40	29	3	3	0	1	0
184	1	36	25	7	3	0	1	0
5	1	35	.	12	.	1	1	0
212	1	38	18.9	8	3	0	1	0
87	1	29	16	1	1	0	1	0
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.