Estimating the functional form of the effect of a continuous covariate on survival time

Dissertation

zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Dortmund

Dem Fachbereich Statistik
der Universität Dortmund
vorgelegt von

Norbert Holländer

aus

Bockum-Hövel

Freiburg 2001

Gutachter: Prof. Dr. M. Schumacher (Freiburg)
Prof. Dr. W. Urfer
Prof. Dr. S. Schach

Tag der mündlichen Prüfung: 8. Februar 2002

Für die wissenschaftliche Betreuung dieser Arbeit und seine stets motivierenden Anregungen bedanke ich mich bei Prof. Dr. Martin Schumacher ganz herzlich. Mein Dank gilt auch meinen Kolleginnen und Kollegen der Abteilung Medizinische Biometrie und Statistik des Instituts für Medizinische Biometrie und Medizinische Informatik am Universitätsklinikum Freiburg und des Freiburger Zentrums für Datenanalyse und Modellbildung. Besonders danken möchte ich Nicole Augustin für ihre kritischen Kommentare, Thomas Gerds für seine konstruktiv kritischen, manchmal schon zu kritischen Anmerkungen, Regina Gsellinger für ihre Unterstützung bei der Textverarbeitung und unserem lokalen 'S-Plus-Guru' Guido Schwarzer, der es geschafft hat, aus mir einen passablen S-Plus-Programmierer zu machen.

Ich danke auch meiner Frau Christiane Kreienbaum und unseren Töchtern Hannah und Katharina für ihr generell vorhandenes, aber unterschiedlich ausgeprägtes Verständnis für meine häufige häusliche Abwesenheit während sonst unüblicher Bürozeiten.

Freiburg, im Februar 2002

Contents

Zusammenfassung

Bei der statistischen Modellbildung im Rahmen von Regressionsmodellen muss der funktionale Zusammenhang zwischen der Zielgröße und den Einflussfaktoren spezifiziert werden. In der vorliegenden Arbeit werden verschiedene Methoden zur Bestimmung der funktionalen Form des Effekts einer stetigen Kovariable im proportionalen Hazardmodell für Überlebenszeiten untersucht. Dabei werden datenabhängige und datenunabhängige Methoden betrachtet.

Datenunabhängigkeit bedeutet hier, dass die generelle funktionale Form vorgegeben ist. Üblicherweise wird im proportionalen Hazardmodell angenommen, dass die logarithmierte Hazardfunktion linear von einer stetigen Kovariable abhängt. Neben dieser linearen Risikofunktion werden aus der Klasse der datenunabhängigen Methoden die folgenden Methoden betrachtet: eine Risikofunktion mit einem linearen und quadratischen Term, ein an den Rändern linearisierter kubischer Spline mit fest vorgegebenen Knoten (restricted cubic spline) und die Kategorisierung der stetigen Kovariable anhand eines oder mehrerer fester Cutpoints. Die Parameter dieser Funktionen werden aus den Daten geschätzt. Bei den datenabhängigen Methoden wird zusätzlich die funktionale Form aus den Daten geschätzt. Eine in dieser Arbeit untersuchte Methode besteht darin, die stetige Kovariable anhand eines oder mehrerer datenabhängiger Cutpoints zu kategorisieren. Als weitere Methode wird die Modellierung der Risikofunktion durch fractional polynomials untersucht.

Für alle Methoden wird die klassische Annahme einer linearen Risikofunktion als Referenzmodell betrachtet. Hierbei wird untersucht, ob eine komplexere Risikofunktion zu einer signifikanten Verbesserung in der Likelihood führt.

Die Auswirkungen der Modellbildung lassen sich besonders gut bei der Anwendung datenabhängiger Methoden untersuchen. Die Kategorisierung einer stetigen Kovariable anhand datenabhängiger Cutpoints kann dabei als Prototyp für komplexere Strategien der Modellbildung, wie z.B. der Variablenselektion, angesehen werden. Die Selektion eines Cutpoints mittels maximal selektierter Teststatistik führt dazu, dass das relative Risiko zwischen den resultierenden Risikogruppen in der Regel deutlich überschätzt wird. Dies ist darauf zurückzuführen, dass sowohl die Modellbildung als auch die Schätzung des resultierenden Effekts im selben Datensatz vorgenommen wurde. Des Weiteren ist aus der Literatur hinreichend bekannt, dass der P-Wert der maximal selektierten Teststatistik korrigiert werden muss. Korrigierte P-Werte werden daher auch in der vorliegenden Arbeit verwendet. Basierend auf eigenen, früheren Forschungsaktivitäten werden außerdem Shrinkage-Methoden zur Korrektur des geschätzten relativen Risikos verwendet. Die Kategorisierung anhand mehrerer datenabhängiger Cutpoints ist eine Erweiterung des oben beschriebenen Cutpoint-Problems. Die Modellierung der Risikofunktion durch fractional polynomials basiert ebenfalls auf einem intensiven Prozess der Modellbildung: aus einer Klasse von

fractional polynomials wird anhand der Daten die beste Risikofunktion ausgewählt. Wie oben beschrieben wurde, wird auch bei der Anwendung datenunabhängiger Methoden eine einfache Modellselektion verwendet, indem entweder die komplexere Risikofunktion ausgewählt, oder bei nicht signifikanter Verbesserung der Likelihood die lineare Risikofunktion benutzt wird. Da Modellselektion auch bei weniger intensiven Selektionsmethoden dazu führt, dass der wahre Effekt der Einflussgröße häufig zu optimistisch beurteilt wird, ist es wünschenswert stabile Methoden zur Schätzung der Risikofunktion zu finden. Zu diesem Zweck wurde das sogenannte bootstrap aggregating (bagging), das ursprünglich zur Reduzierung des Vorhersagefehlers von Prädiktoren entwickelt wurde, auf das vorliegende Problem übertragen. Die Idee des bagging besteht darin, die Risikofunktion in einer Menge von Bootstrap-Stichproben zu schätzen und die geschätzten Funktionen über alle Bootstrap-Stichproben zu mitteln. In allen Bootstrap-Stichproben wird dabei die gleiche Modellselektionsstrategie angewendet wie in den Originaldaten. Neben der Berechnung der aggregierten Risikofunktion werden die Resultate aus den Bootstrap-Stichproben zudem dazu verwendet, um die Stabilität der Schätzung zu untersuchen.

Das primäre Ziel der vorliegenden Dissertation besteht darin, alle genannten Verfahren zu vergleichen, und die auf bootstrap aggregating basierende Erweiterung der Schätzverfahren zu untersuchen. Zentrale Fragen sind:

- Welche Methoden sind am besten geeignet, um typische Risikofunktionen adäquat zu schätzen ?
- Inwieweit hängen die geschätzten Risikofunktionen von einer mehr oder weniger intensiven Modellbildung ab?
- Ist bagging dazu geeignet, die negativen Auswirkungen der Modellbildung zu korrigieren und zuverlässigere Risikofunktionen zu schätzen?

In Kapitel 2 werden die verwendeten Methoden beschrieben und auf die Daten von zwei Brustkrebsstudien angewendet. Des Weiteren gehe ich hier kurz auf den potentiell zeitabhängigen Effekt der Kovariable ein und stelle weitere Methoden aus der Literatur vor.

Um die einzelnen Methoden beurteilen zu können, wurde eine Simulationsstudie durchgeführt. Neben dem Nullmodell, in dem die stetige Einflussgröße keinen Effekt hat, wurden drei proportionale Hazardmodelle mit verschieden Risikofunktionen generiert: I. ein Cutpointmodell mit einem fest vorgegeben Cutpoint, II. eine lineare Risikofunktion, III. eine Risikofunktion, bei der das Risiko mit zunehmenden Abstand von einem festen Cutpoint linear ansteigt (V-Typ). Da starke Effekte in der Praxis eher selten sind, habe ich dabei einen moderaten Effekt zu Grunde gelegt.

Die Ergebnisse der Simulationsstudie werden in Kapitel 3 beschrieben. Des Weiteren gehe ich in diesem Kapitel darauf ein, wie man die mit verschiedenen Methoden geschätzten Risikofunktionen im proportionalen Hazardmodell vergleichbar macht, und es werden verschiedene Fehlermaße zur Beurteilung der Methoden diskutiert. Neben den quantitativen Fehlermaßen Mean Squared Error und Mean Absolute Error wird in der Arbeit auch ein qualitatives Fehlermaß vorgestellt. Die Auswirkungen der Modellbildung wird anhand einzelner Beispiele dargestellt. Für das Cutpoint-Problem wird dabei auf die Problematik der Schätzung von Konfidenzintervallen nach Modellbildung eingegangen. Es wird gezeigt, wie man durch die Anwendung von Shrinkage-Methoden und Bootstrap-Resampling zu zuverlässigen Ergebnissen gelangt.

Die wichtigsten Resultate der Arbeit lassen sich folgendermaßen zusammenfassen:

- Bei der Analyse der Brustkrebsstudien konnte ein deutlicher nichtlinearer Effekt der stetigen Kovariable Alter auf die progressionsfreie Überlebenszeit nachgewiesen werden. Für die meisten Methoden war die mit bagging geschätzte Risikofunktion fast identisch mit der in den Originaldaten geschätzten Funktion.

- In der Simulation waren die fractional polynomials am besten geeignet, die vorgegebenen Risikofunktionen zu beschreiben.

- Es konnte eine deutliche Überschätzung des vorgegebenen Effekts beobachtet werden, wenn bei der Modellbildung die komplexere Risikofunktion ausgewählt wurde.

- Durch die Anwendung von bagging konnten die Fehler häufig reduziert werden, bagging kann aber nicht in allen Fällen empfohlen werden. Generell kann bagging nur dann funktionieren, wenn sich die in den Bootstrap-Stichproben geschätzten Risikofunktionen unterscheiden. Diese Resultate bestätigen die aus der Literatur bekannten Ergebnisse für Klassifikationsprobleme.

- Die Anwendung von Shrinkage Methoden beim Cutpoint-Problem ermöglicht eine Bias-Reduktion des aufgrund der Modellbildung überschätzten relativen Risikos. Zur Korrektur der Varianzschätzung wurde in dieser Arbeit die Varianz aus den Bootstrap-Stichproben geschätzt. Die simultane Anwendung von Shrinkage und Varianzkorrektur ermöglicht es, korrekte Konfidenzintervalle zu schätzen.

- Das in dieser Arbeit vorgeschlagene qualitative Fehlermaß zur Beurteilung der geschätzten Risikofunktionen ist nur eingeschränkt anwendbar.

- Die geschätzten quantitativen Fehler hängen nur wenig von der Methode ab, mit der die verschiedenen Risikofunktionen vergleichbar gemacht wurden.

1 Introduction

In the analysis of many medical studies the effect of covariates, measured on a continuous scale, on an outcome variable or some transformation of outcome is often assumed to be linear. However, a specific prognostically relevant covariate may exhibit an effect that is markedly nonlinear. Consequently, the assumption of linearity may lead to wrong conclusions. A nonlinear effect can be detected, for example when categorizing the continuous covariate into several categories and estimating the effect with respect to a reference category. However, this approach can be very unstable if the resulting subgroups are too small. Furthermore, even when using several subgroups there is a loss of information. On the other hand, categorization is often preferred due to an easy interpretation of the results. For categorizing a continuous covariate one or several cutpoints need to be determined. This can either be done by using prespecified cutpoints taken from specific medical knowledge or from previous studies, or by selecting the cutpoints from the data. Taking the continuous structure of the covariate into account a nonlinear effect can also be detected by considering a linear and quadratic term or, more generally, by modeling the effect by polynomials or splines. Applying the former methods the general functional form is fixed in advance whereas it is usually estimated within the data when using splines.
In the analysis of survival time data interest centers on the time between a starting point, e.g. diagnosis or start of therapy, and the occurrence of a specific event, e.g. death or progressive disease. As outcome variable we usually consider the hazard function, which is the conditional probability for the occurrence of an event during an infinitesimal small time interval given that the event did not occur before. Modeling the effect of one or several covariates by using the Cox proportional hazards regression model (Cox, 1972) the hazard function depends on this/these covariate(s) and on time. In the Cox model the assumption of a linear effect corresponds to a log-linear dependency, more details are given in section 2.
In this thesis I investigate several methods to estimate the functional form of the effect of one continuous covariate in the framework of the Cox model. In addition, I use bootstrap resampling to extend the chosen methods and to investigate problems of model selection. More details on the contents of this work will be given after illustrating the problem by using an example in oncology. In particular, I consider the effect of the continuous covariate age on the prognosis of breast cancer. After discussing a few examples taken from the literature the problem is motivated by using the data of two breast cancer studies.

1.1 Example: The effect of age on the prognosis of breast cancer

In breast cancer the effect of more than 100 potential prognostic factors were controversially discussed during the last years. Independent from the therapy the strong effect of the number of positive lymph nodes on the prognosis has been proven in the past and, therefore, this covariate can be considered as a standard prognostic factor. Other covariates, e.g. tumor size, tumor grade, the progesterone and estrogen receptor and the patient's age are often assumed to influence the prognosis, too. Especially, the potential prognostic effect of age has been discussed controversially. Table 1.1 shows the results of 5 studies on this topic. Except for the study of Kroman et al. (2000) at least two of the following endpoints were considered: overall survival, disease-free survival and cancer-specific survival. Disease-free survival, which is also referred to as relapse-free or event-free survival, is usually defined as time from diagnosis or treatment start until relapse/progressive disease or death. Patients without any of these event were censored at the last follow up. For cancer specific survival, also called cause- or disease-specific survival, non-disease-related deaths were not considered as event and, therefore, the corresponding deaths were considered as censored observations.

In most of the cited papers the effect of age were investigated by univariate and multivariate analysis. In an univariate analysis age is taken into account as single covariate, whereas the estimated effect of age is adjusted to other covariates when performing a multivariate analysis. Here, I neither comment on the statistical methods used in the cited studies, nor on the different designs and the quality of the selected studies. The results described in table 1.1 show that there seems to be an increased risk for young patients. However, this result was not confirmed in all studies and for all endpoints. Although age is a continuous covariate age subgroups were used in all 5 studies. The categorization of continuous covariates is often preferred because results can be interpreted easily: the estimated parameters in the Cox regression model can be referred to as log relative risks between the subgroups. However, the few examples cited above already show that cutpoints can differ between studies. Therefore, it may be difficult to compare the results. Besides the categorization into 3 age subgroups De la Rochefordiere et al. (1993) also assume a linear, a quadratic and a logarithmic effect of age on the log hazard rate. Out of these risk functions the linear effect led to the best fit. In a recently published paper investigating several pathologic and clinical factors for the prognosis of invasive breast carcinoma Fisher et al. (2001) used a risk function with a linear and a quadratic term for age and found a highly significant effect. To model the potential nonlinear but continuous effect of age and other covariates Sauerbrei and Royston (1999) use fractional polynomials. Using the data of a controlled clinical trial, which will be referred to as GBSG-2 study, they showed that younger women have an increased risk with respect to event-free survival whereas age showed no effect when assuming a (log-)linear relationship. Comparing therapies in the first analysis of the GBSG-2 study age was included as categorized covariate using the subgroups ≤ 45, $45-60$ and > 60 years (Schumacher et al., 1994). In the cited paper there was no significant effect of age with regard to event free survival in the univariate analysis. Therefore, age was not used in the multivariate analysis.

Throughout my thesis I will use the data of the GBSG-2 study, and the data of a second somewhat smaller study, which will be referred to as the Freiburg DNA-study, for illustration. Due to its larger sample size I focus on the GBSG-2 study.

GBSG-2 study

The GBSG-2 study is a prospective, controlled clinical trial on the treatment of node positive breast cancer patients conducted by the German Breast Cancer Study Group (GBSG) (Schumacher et al., 1994). The principal eligibility criterion was a histologically verified primary breast cancer of stage T1a-3aN+M0, i.e. with positive regional lymph nodes but no distant metastases. Primary local treatment was by a modified radical mastectomy with en bloc axillary dissection with at least six identifiable lymph nodes. Patients should not be older than 65 years of age and should present with a Karnofsky index of at least 60. The study was designed as a Comprehensive Cohort Study (Schmoor et al., 1996), i.e. randomized as well as non-randomized patients who fulfilled the entry criteria were included and followed-up according to the study procedures.

The study had a 2x2 factorial design with four adjuvant treatment arms: three vs. six cycles of chemotherapy with and without hormonal treatment. Prognostic factors evaluated in the trial were patient's age, menopausal status, tumor size, estrogen and progesterone receptor, tumor grade according to Bloom and Richardson (1957), histological tumor type and number of involved lymph nodes. Histopathologic classification was reexamined, and grading was performed centrally by one reference pathologist for all cases. Event-free survival (EFS) was defined as time from mastectomy to the first occurrence of either locoregional or distant recurrence, contralateral tumor, secondary tumor or death. During six years 720 patients were recruited, of whom about two thirds were randomized. After a median follow-up time of nearly 5 years, 299 events for EFS and 171 deaths were observed. Event-free survival was about 50% at five years. Complete data of the standard prognostic factors mentioned above were available for 686 (95.3%) patients, who where taken as the basic patient population.

Freiburg DNA study

The database of this study consisted of all patients with primary, previously untreated node positive breast cancer who were operated between 1982 and 1987 in the Department of Gynecology at the University of Freiburg and whose tumor material was available for DNA investigations. Some exclusion criteria (history of malignoma, T_4 and/or M_1 tumors according to the UICC classification system, without adjuvant therapy after primary surgery, older than 80 years etc.) were defined retrospectively. This left 139 patients out of 218 originally investigated for the analysis.

Table 1.1: Results of 5 selected studies investigating the effect of age on the prognosis of breast cancer

study described in the paper of	sample size	Age subgroups	Main results
De la Rochefordiere et al. (1993)	1703	$\leq 3, 34-40, > 40$ continuous: linear, quadratic & logarithmic effect	Younger women had an increased risk for cause specific and disease free survival in univariate and multivariate analyses. Assuming a (log-)linear effect of age led to a better fit as compared to the other two continuous risk functions, a significant linear decrease in risk is obtained for disease free survival only.
Chung et al. (1996)	3722	$< 40, 41-50, 51-60, 61-70, 71-80, > 80$	The worst cancer specific and disease free survival rate were obtained for the oldest ($>$ 80) and the youngest ($<$ 40) patients. Except for the stratification by tumor stage, the authors performed no multivariate analysis
Vanlemmens et al. (1998)	1751	$\leq 3, 34-40, > 40$	The univariate analysis showed a significant decrease in risk with increasing age for overall survival, cancer specific and relapse free survival. In the multivariate analysis a significant prognostic effect of age was found for cancer specific survival only
Ezzat et al. (1998)	710	$< 40, 40-50, > 50$ and $< 30, 30-40, 40-50, 50-60, > 60$	There was no effect of age on relapse free survival and overall survival in univariate and multivariate analyses
Kroman et al. (2000)	10356	$< 35, 35-39, 40-44, 45-49$	An increased risk for younger women with respect to overall survival was exclusively found in patients with low risk disease who did not receive any adjuvant treatment.

Eight patients characteristics were investigated. Besides age, number of positive lymph nodes and size of the primary tumor, the grading score according to Bloom and Richardson (1957) as well as estrogen- and progesterone receptor status were recorded. DNA flow cytometry was used to measure ploidy status of the tumor (using a cutpoint of 1.1 for the DNA index) and S-phase fraction, which is the percentage of tumor cells in the DNA synthesizing phase obtained by cell cycle analysis.
The median follow-up was 83 months. At the time of analysis, 76 events have been observed for event-free survival which was defined as the time from surgery to the first of the following events: occurrence of locoregional recurrence, distant metastases, second malignancy or death. Event-free survival (EFS) was estimated as 50% after five years. Further details of the study can be found elsewhere (Pfisterer et al., 1995).

A first example
I focus on the continuous covariate age and investigate its functional relationship with respect to EFS. The age distribution displayed in figure 1.1 show that the patients are slightly older in the Freiburg DNA-study. The patient's age ranges from 26 to 80 years with a median of 56 years in the Freiburg DNA study, whereas the median age is 53 (range: 21 to 80) in the GBSG-2 study.

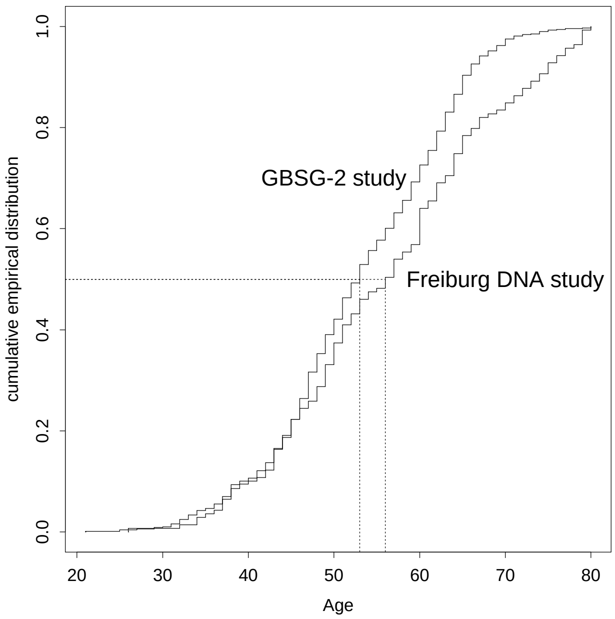


Figure 1.1: Empirical distribution function of age in the GBSG-2 study and the Freiburg DNA study, dotted lines denote the median of age in the two studies

To get an idea of the effect of the continuous covariate age on the hazard of EFS we consider three assumptions with respect to the functional form.

- assuming a linear effect of age
- categorizing the continuous covariate age into two subgroups using 35 and 40 years, respectively, as cutpoints
- allowing a curved functional form, which is modeled by a two term fractional polynomial

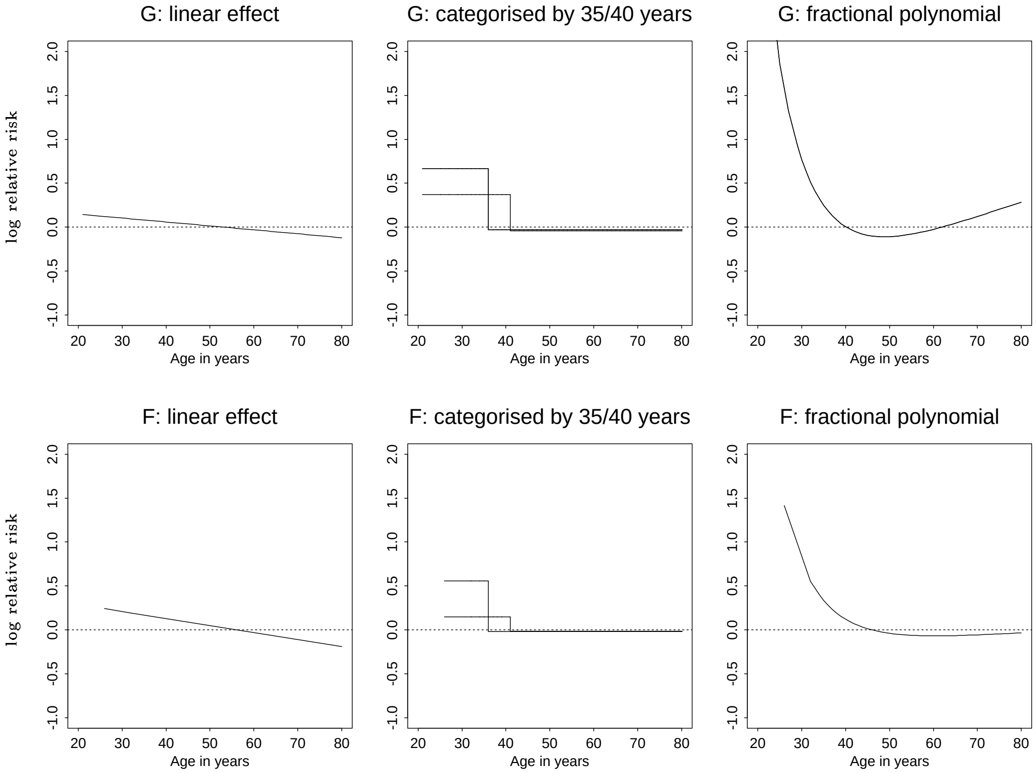


Figure 1.2: Risk function estimates in the GSGG-2 study (G) and the Freiburg DNA study (F) for different methods used to model the effect of age for EFS

The general shape of the first two functions is fixed in advance whereas we use the data in order to find the 'best' fractional polynomial in each study. More details on this model building process, the model and further methods to describe the functional form of a continuous covariate are given in section 2. Figure 1.2 illustrates the log relative risk as a function of age estimated in a univariate Cox proportional hazard model in the GBSG-2 study and the Freiburg DNA study. In the GBSG-2 study the continuous covariate age is significantly (likelihood ratio test with $p < 0.05$) related to EFS when age is categorized into subgroups or the effect of age is described by a fractional polynomial. Assuming a

linear functional form, however, we obtained no significant effect of age. All functions show a decreasing risk with increasing age, where the amount of this decrease depends strongly on the chosen model. Interpreting these results it should be taken into account that there are only a few patients younger than 30 years. Thus, the strong decrease in risk obtained by the estimated two term fractional polynomial should be interpreted carefully. In the Freiburg DNA study none of the functions show a significant effect of age. However, due to the small sample size the estimated variance is very large. Considering the resulting functional form, the results of both studies correspond very well. Using 35 years as a cutpoint the risk difference between the resulting subgroups is nearly the same although there are only 5 patients (3 events) younger than 35 years in the Freiburg DNA study. The fractional polynomial approach shows the strongest decrease in risk with increasing age up to 45 years. However, in the GBSG-2 study we observed again a slight increase for older patients whereas the risk seems to be constant for patients between 45 and 80 years in the Freiburg DNA study. So, should we believe the results of the larger study or is the risk increase for older patients solely caused by the model building process used to find the 'best' two term fractional polynomial? It should be mentioned briefly, that the results with respect to the functional form of the effect of age only change slightly, if we estimate additionally the effects of the other potential prognostic factors in a multivariate proportional hazards model. In my thesis I restrict myself to the univariate situation.
Because of the well known problem that an extensive process of model building is more likely to produce artefact in small studies and/or when the true effect is small (see e.g. Schumacher et al. (1997)), I will use, in contrast to figure 1.2, the Freiburg DNA study only for validation. Thus, the GBSG-2 study is used to determine the functional form (if not fixed in advance) of the effect of age, with respect to EFS and to estimate the corresponding effect. In the Freiburg DNA study we try to validate results by estimating the effect while assuming the functional form obtained in the GBSG-2 study.

1.2 Outline

Considering the effect of age with respect to event-free survival section 1.1 illustrated, that answers to questions like *Is there an effect? How can this effect be described?* may depend strongly on the choice of the risk function. The prespecification of a functional relationship, i.e. as done when assuming a (log-) linear effect may be not suitable to describe the true effect correctly, whereas the data-dependent choice of the functional form can lead to a drastic 'over-fitting'. In order to determine the functional form of a continuous covariate I consider several data-dependent and data-independent methods, where the general functional shape is given for data-independent methods. Thus, the data are used only to estimate the parameter(s) of the risk function. In contrast with data-dependent methods we also try to find the best functional form within the data before estimating the parameters. Besides categorizing the covariate by using fixed and data driven cutpoints, respectively, I consider a linear relationship commonly assumed in regression models. Furthermore, I use a risk function including a linear and a quadratic term, and I estimate the functional form by a fractional polynomial (Royston and Altman, 1994) and by a restricted cubic spline (Harrell, 1997). All methods are extended by adopting a method called 'bootstrap aggregating' that has been proposed by Breiman (1996).

In section 2 I describe the methods in the framework of the Cox proportional hazard model and apply them to the data of the GBSG-2 study and the Freiburg DNA study. In this context it will be investigated how far results can be validated using an independent data set. An overview of all standard methods is given in section 2.8. In section 2.1.10 I comment on potential time-dependency. Some further methods to estimate the functional form of a continuous covariate that were not used in this thesis will be described in section 2.1.11.

In order to investigate the capability of the chosen methods to describe a given functional relationship correctly, I performed a simulation study. In my investigation I focused especially on the following questions:

- Which method is most appropriate to describe typical functions?
- How far does the estimated functional form depend on a more or less extensive process of model building?
- How far is the use of bootstrap resampling helpful to obtain more 'robust' functions and/or to overcome problems of model building?

The design of the simulation study is described in section 3.1. In section 3.2 I discuss two methods to make results comparable, which is a specific problem when analyzing survival time data. After introducing error measures in section 3.3 the results of the null model of no prognostic relevance of the continuous covariate with respect to survival are given in section 3.4. The results of further non-zero risk functions are summarized in sections 3.5 - 3.7. Section 3.8 contains a short comparison of different error measures. Furthermore, I discuss some problems caused by model building and the calculation of confidence intervals after model building by using specific examples from the simulation study. After investigating selected methods again in a further small simulation that is based on the data of the GBSG-2 study, the main results of section 3 are summarized in section 3.9. In section 4 this thesis ends with a final discussion.

2 Methods for estimating the functional form of a continuous covariate

2.1 Standard procedures

A standard tool for analyzing survival time data is the Cox proportional hazards model (Cox, 1972; Andersen et al., 1993). Considering only one continuous covariate X the model is given by

$$\lambda(t|X = x) = \lambda_0(t)\ \exp(h(x, \beta)), \tag{1}$$

where $\lambda(t|\cdot)$ denotes the hazard function of the event-free or overall survival time random variable T and $\lambda_0(t)$ is an unspecified baseline hazard. The effect of the continuous covariate is described by the function $h(x, \beta)$, that will be referred to as risk function. The parameter estimates $\hat{\beta}$ of this risk function are obtained by maximizing the corresponding partial likelihood (Cox, 1972). It should be noted that $h(x, \beta)$ contains no intercept term, because all information not related to the covariate is included in the baseline hazard. Since

$$\frac{\lambda(t|X = x)}{\lambda(t|X = 0)} = \exp(h(x, \beta))$$

the risk function is referred to as log hazard ratio or log relative risk for an individual with $X = x$ having an event (e.g. death) as compared to an individual with $X = 0$. For details on the analysis on survival data I refer to the literature, e.g. the textbooks of Kalbfleisch and Prentice (1980) or Marubini and Valsecchi (1995).

In practical situations, we consider usually several covariates $X_1, \ldots, X_K$ that may be continuous, binary and/or ordinal. However, in this thesis my interest centers around the estimation of the functional form $h(x, \beta)$ of one continuous covariate and, therefore, I restrict to this simple situation. Although the risk function depends on one or more parameters, I use the notation $h(x)$ instead of $h(x, \beta)$ throughout my thesis.

2.1.1 Linear relationship

Modeling the effect of a continuous covariate on survival time the classical assumption is a linear relationship given by

$$h(x) = \beta x. \tag{2}$$

In this situation a log-linear relationship holds between the hazard function and the covariate X. The parameter $\exp(\beta)$, therefore, represents the increase or decrease in risk if X is increased by one unit, e.g. one year when X denotes the age of a patient. However, it is unlikely that this assumption is reasonable under all circumstances. Therefore, it is important to investigate the use of the linear risk function in a nonlinear situation and to compare it to models that are based on nonlinear risk functions.

Besides specifying the functional form correctly, one may also be interested, whether the covariate has an influence on the hazard function and, therefore on survival, at all. In order to test the hypothesis $H_0 : \beta = 0$ one may use the Wald test statistic $\hat{\beta}/S.E.(\hat{\beta})$ which is asymptotically distributed as a standard normal distribution under the null hypothesis. Alternatively one could use the likelihood ratio test comparing the likelihood of the model with covariate against the so called null model omitting all covariates. Throughout this thesis the linear relationship is considered as some kind of basic assumption, i.e. I assume a linear effect (2) if the other methods described in the sequel do not suggest a nonlinear relationship. The test procedure used to select between the linear and the nonlinear risk function depends on the method and will be given when describing the corresponding method. All tests are based on a significance level $\alpha = 0.05$.

2.1.2 Linear and quadratic term

One common approach to investigate potential nonlinearity is to add a quadratic term to the risk function, i.e. considering

$$h(x) = \beta_1 x + \beta_2 x^2. \tag{3}$$

A deviation from linearity corresponds to a significant effect of the quadratic term, i.e. if the hypothesis $H_0 : \beta_2 = 0$ is rejected. In order to test H_0 I use the likelihood ratio test comparing the model based on (3) against that based on the linear risk function (2). Alternatively, one could also use the corresponding Wald test.

2.1.3 Categorization based on prespecified cutpoints

In many applications a continuous covariate (e.g. age) is categorized in two or three categories using one (or two) prespecified cutpoint(s). Using one fixed cutpoint μ_0 the risk function is given by

$$h(x) = \beta x^*, \tag{4}$$

where $x^* = 1_{\{x>\mu_0\}}$. The corresponding proportional hazards model can be written by

$$\lambda(t|X > \mu_0) = \exp(\beta)\lambda(t|X \leq \mu_0). \tag{5}$$

The parameter β is referred to as log-relative risk for observations with $X > \mu_0$ relative to observations with $X \leq \mu_0$.

Assuming three categories, which are defined by two cutpoints μ_{01} and μ_{02} with $\mu_{01} < \mu_{02}$, we consider

$$h(x) = \beta_1 x_1^* + \beta_2 x_2^* \tag{6}$$

with

	$x \leq \mu_{01}$	$\mu_{01} < x \leq \mu_{02}$	$x > \mu_{02}$
x_1^*	0	1	0
x_2^*	0	0	1

The first interval defined by $x \leq \mu_{01}$ serves as reference category, β_1 and β_2 are referred to as log relative risks for observations within the second and the third interval, respectively, relative to observations in the reference category.

This simple interpretation of the results may be seen as the main reason for the categorization of a continuous covariate.

2.1.4 Categorization using data driven cutpoints

Cutpoints are often chosen according to common schemes (e.g. age $>$ 60 years versus age $\leq$ 60 years) or the determination of a cutpoint is based on specific medical knowledge (e.g. progesterone receptor $\geq$ 20 fmol versus progesterone receptor $<$ 20 fmol). However, for a factor where no such prior information is available the data are often used in order to find so called data driven cutpoints.

One cutpoint

Considering only one cutpoint μ the risk function is

$$h(x) = \beta \tilde{x} \tag{7}$$

with $\tilde{x} = 1_{\{x>\mu\}}$,the corresponding proportional hazards cutpoint model is given by

$$\lambda(t \,|\, X > \mu) = \exp(\beta)\lambda(t \,|\, X \leq \mu), \tag{8}$$

which is identical to formula (5) except for the fact, that here the cutpoint μ is unknown and has to be estimated from the data. A popular approach for a data dependent categorization is the so-called minimum P-value method where - within a certain range of the distribution of X, the selection interval - the cutpoint $\hat{\mu}$ is taken such that the P-value for the comparison of observations below and above the cutpoint is a minimum. Here I use the P-value of the logrank-test and consider all covariate values between the 10 and 90 percent quantile as potential cutpoints. It should be mentioned that the cutpoint of 35 years used for the example described in the introduction is outside this selection interval.

Several cutpoints

Applying the minimum P-value method to the whole population we obtain one cutpoint $\hat{\mu}$ and, consequently, two subgroups, namely those with the best separation with respect to patient's survival. With this approach it is implicitly assumed that each resulting subgroup is homogeneous and this assumption may not be adequate. Therefore, it may be sensible to repeat the procedure within each of the two subgroups separately. The gradual selection of subgroups is the basic idea of the method of classification and regression trees (CART).

A comprehensive description of CART can be found in the book of Breiman et al. (1984), an overview including modifications and extensions of CART is given by Zhang et al. (1998). One extension of CART is the application to survival data (Gordon and Olshen, 1985; Le Blanc and Crowley, 1992, 1993; Segal, 1988).

Briefly, the idea of CART is to construct subgroups which are internally as homogeneous as possible with regard to the outcome variable and externally as separated as possible. Usually CART is applied to several covariates, which may be quantitative, ordinal or nominal. Here I restrict to one continuous covariate. We define a minimum number of patients within a subgroup, n_{min} say, and prespecify an upper bound for the P-value of the logrank-statistic, p_{stop}. The tree building algorithm is defined by the following steps:

i) Within a prespecified selection interval - all values of X between the 10% and 90% quantile are considered as potential cutpoints - the minimal P-value of the logrank statistic is computed.

ii) The whole group of patients is split into two subgroups based on the cutpoint $\hat{\mu}$ with the minimal P-value, if the minimal P-value is smaller or equal to p_{stop}.

iii) The partition procedure is stopped if there exists no allowable split, i.e. if the minimum P-value is greater than p_{stop} or because the size of the subgroup is smaller than n_{min}.

iv) For each of the two resulting subgroups the procedure is repeated.

Note that steps i) to iii) correspond to the 'one cutpoint situation' described above.

It should be briefly mentioned that the classical CART procedure consists - as described in the cited references - of a tree building and a tree pruning algorithm. Usually, in the tree building step the partitition process is performed with hardly any stopping rules.

Consequently, the resulting tree has very small final subgroups, the so called final nodes. In the tree pruning step final nodes may again be combined by some amalgation. However, we use only a tree building step and control the size of the tree by the selection interval and n_{min} (Lausen et al., 1994). Furthermore, the number of the final nodes is limited by allowing only one repetition in step iv) leading to maximally 3 cutpoints and, therefore, to maximally 4 final subgroups. However, in principle the procedure can be repeated more than once. The cutpoints obtained by the tree building algorithm are used to categorize the continuous covariate X. Assuming a maximum of 4 final subgroups defined by three cutpoints $\mu_1 < \mu_2 < \mu_3$ the corresponding risk function is given by

$$h(x) = \beta_1 \tilde{x}_1 + \beta_2 \tilde{x}_2 + \beta_3 \tilde{x}_3 \tag{9}$$

with

	$x \leq \mu_1$	$\mu_1 < x \leq \mu_2$	$\mu_2 < x \leq \mu_3$	$x > \mu_3$
$\tilde{x}_1$	0	1	0	0
$\tilde{x}_2$	0	0	1	0
$\tilde{x}_3$	0	0	0	1

As described in 2.1.3 the first interval serves as reference category. Due to its relation to the CART procedure this method will be referred to as CART-based categorization.

2.1.5 Fractional polynomial

In order to model a curved relationship between an outcome variable (in this paper survival time) and a continuous covariate Royston and Altman (1994) propose the use of so-called fractional polynomials in the context of regression models.

A fractional polynomial (FP) is an extension of the ordinary polynomial that offers greater flexibility. A fractional polynomial of degree m is defined by:

$$FP_m(x) = \beta_0 + \sum_{j=1}^{m} \beta_j \, x^{p_j} \tag{10}$$

where m is a positive integer. As with ordinary polynomials one has to select the number of terms. Here I restrict myself to one and two term FP's. In principle the powers p_j, $j = 1, 2$ may take any real value but for pragmatic reasons Royston and Altman (1994) suggest considering only values from the restricted set $S_p = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ where, per definition $x^0 = \log x$.

In the context of the proportional hazards model β_0 can be moved to the baseline hazard and the risk function is given by

$$h(x) = \sum_{j=1}^{m} \beta_j \, x^{p_j} \tag{11}$$

where $m \leq 2$.

For $m = 1$ and $p_1 = 1$ the model reduces to the linear relationship described in 2.1.1 and $m = 2$ with $p_1 = 1$ and $p_1 = 2$ lead to a linear and a quadratic term as described in 2.1.2. However, the best one-term $FP\,(m = 1)$ is found by fitting regression models corresponding to every power in the restricted set S_p, namely x^{-2}, x^{-1}, …, x^3. The power which corresponds to the model with the lowest deviance $D = -2$ log-likelihood is selected. Similarly, to find the best two term FP, regression models are fitted corresponding to every possible pair of powers from S_p.

As done in this thesis it was also postulated by Royston and Altman (1994) to use the linear relationship as basic assumption. Thus, the relationship between X and survival time is only described by a m-degree FP if it is better, i.e. has lower deviance, than the model with the linear risk function. The gain of the FP model is measured by a deviance based test using

$$G = D_{LIN} - D_{FP\,(m)}, \tag{12}$$

where $D_{FP\,(m)}$ is the deviance for the model in question. It should be mentioned that the test is somewhat conservative (Royston and Altman, 1994). Since I consider $m \leq 2$ only I have to choose between the best one and the best two term fractional polynomial, denoted as FP_1 and FP_2, and the linear relationship, here denoted by LIN. Note that G is simply the likelihood ratio test statistic when testing against LIN. However, comparing FP_2 to FP_1 the models are not necessarily nested. In this thesis I will not discuss how far this could be a problem.

Initially the fit of FP_2 is compared to the fit of LIN with the corresponding test statistic G, which is asymptotically distributed as χ^2 with three degrees of freedom when model LIN (the smaller model) is adequate. The degree of freedom is equal to the difference in the number of estimated parameters in the two models. If FP_2 is significantly better than LIN the procedure is continued. Otherwise the procedure is stopped and LIN is accepted as the final model. If the procedure is continued the fit of FP_2 is compared to that of FP_1, the corresponding G has 2 DF. If the fit of FP_2 is significantly better the procedure is stopped and FP_2 is kept as the best model. Otherwise we choose FP_1 as final model.

2.1.6 Restricted cubic spline

The basic concept of splines is to fit piecewise functions rather than using the complete data set at once. Piecewise cubic polynomials were shown to have nice properties with good ability to fit sharply curving shapes. Cubic splines can be made to be smooth at the so called knots, by forcing the first and second derivates of the function to agree at the knots. Stone amd Koo (1985) have found that cubic splines do have a drawback in that they can be poorly behaved in the tails, i.e. before the first and after the last knot. Their restricted cubic spline function, also called natural spline by De Boor (1978), with k knots $a_1 < a_2 < \ldots < a_k$ is given by

$$RCS\,(x) \;=\; \beta_0 \,+\, \beta_1\, x \,+ \beta_2\, x_2 \,+\, \ldots \,+\, \beta_{k-1} x_{k-1}, \tag{13}$$

where $x_1 = x$ and for $j = 1, \ldots, k-2$,

$$x_{j+1} = (x - a_j)_+^3 - (x - a_{k-1})_+^3 \left(\frac{a_k - a_j}{a_k - a_{k-1}}\right) + (x - a_k)_+^3 \left(\frac{a_{k-1} - a_j}{a_k - a_{k-1}}\right) \tag{14}$$

and

$$\begin{aligned} (U)_+ \;&=\; u, \quad u > 0 && (15)\\ &=\; 0, \quad u \le 0. && (16) \end{aligned}$$

It can be shown that x_j is linear in x for $x > a_k$. Besides a better potential fit of the tails (which was also cited as an argument to use fractional polynomials) the restricted cubic spline has the additional advantage that only $k-1$ parameters must be estimated (besides the intercept) as opposed to $k+3$ parameters with the unrestricted cubic spline. Once $\beta_0 \ldots \beta_{k-1}$ are estimated, the restricted cubic spline can be restated in the form

$$RCS\,(x) \;=\; \beta_0 \,+\, \beta_1\, x \,+ \beta_2 (x - a_1)_+^3 \,+\, \beta_3 (x - a_2)_+^3 \,+\, \ldots\; \beta_{k+1} (x - a_k)_+^3 \tag{17}$$

by computing

$$\beta_k \;=\; \frac{1}{(a_k - a_{k-1})} \Big[\beta_2 (a_1 - a_k) + \beta_3 (a_2 - a_k) \;+\; \ldots \,+\, \beta_{k-1} (a_{k-2} - a_k)\Big] \tag{18}$$

$$\beta_{k+1} \;=\; \frac{1}{(a_{k-1} - a_k)} \Big[\beta_2 (a_1 - a_{k-1}) + \beta_3 (a_2 - a_{k-1}) \;+\; \ldots \,+\, \beta_{k-1} (a_{k-2} - a_{k-1})\Big]. \tag{19}$$

Following Stone and Koo (1985) and Harrell (1997) we assume that the location of the knots are specified in advance, i.e. the knot locations are not treated as free parameters to be estimated. Usually no prior knowledge on knot locations is available and, therefore, knots are placed at fixed quantiles of the empirical distribution of X. The following equally spaced quantiles are recommended:

k	quantiles				
3	0.05	0.50	0.95		
4	0.05	0.35	0.65	0.95	
5	0.05	0.275	0.50	0.725	0.95

To ensure that enough points are available in each interval the outer quantiles should be replaced by the 5th smallest and the 5th largest data points, respectively, if the sample size is less than 100. The choice of k should be guided by the sample size. According to Stone (1986) more than 5 knots are seldomly required in a restricted cubic spline model. For many data sets, $k = 4$ offers an adequate fit of the model and is a sufficient compromise between flexibility and loss of precision caused by over-fitting small samples (Harrell, 1997). Fixing the number of knots in advance, restricted cubic splines cannot be expected to perform as flexible as fractional polynomials and, therefore, the comparison of both approaches may not be adequate. Nevertheless, in this thesis I use restricted cubic splines with k=4 knots.

In the framework of the proportional hazards model the risk function is given by

$$h(x) = \beta_1 x + \beta_2 x_2 + \ldots + \beta_{k-1} x_{k-1}, \tag{20}$$

where $x_2, \ldots, x_{k-1}$ are defined as above. Estimates of the coefficients of the spline function are derived with standard techniques allowing statistical inference to be drawn (Harrell et al., 1988). The gain of a restricted cubic spline model as compared to the linear risk function can be investigated by

$$G = D_{LIN} - D_{RCS(k)} \tag{21}$$

where $D_{RCS(k)} = -2$ log likelihood of the fitted k knots restricted cubic spline model and D_{LIN} denotes the corresponding deviance of the proportional hazards model assuming a linear risk function. In contrast to the fractional polynomial test statistic G is always based on nested models and is equal to the likelihood ratio test statistic. Under the null hypothesis , i.e. assuming a linear relationship, G is asymptotically distributed as χ^2 with $k-1$ degrees of freedom. As done for the other risk functions we select the restricted cubic spline model if the test result is significant and the linear risk function otherwise.

2.1.7 A note on bias caused by model building

So far, we considered six approaches in order to determine the functional form of a continuous covariate. Four of these methods are based on risk function that are specified in advance, the data are used to estimate the parameters of the risk function only. Assuming a linear and quadratic relationship (section 2.1.2) and describing the effect by a restricted cubic spline (section 2.1.6) model building is performed in so far, that the more complex model is used only if the likelihood ratio test against the model with a linear risk function was significant. Furthermore, the restricted cubic spline uses data dependent quantiles as knots. A more complex process of model building is involved when categorizing the continuous covariate using data driven cutpoints (2.1.4) and when applying the fractional polynomial approach, where the data is used to find the best functional form. These steps of model building may lead to a considerable amount of over-optimism with respect to the predictive ability of the 'final' regression model. This problem is especially relevant if model building, here the specification of the functional form, and estimation of the resulting effect is performed with the same data set. We have illustrated this phenomenon when categorizing a continuous covariate X by using the minimum P-value in a simulation study (Schumacher et al., 1997).

The selection of one data driven cutpoint is usually based on many tests. In particular, the number of logrank tests used to select $\hat{\mu}$ is equal to the number of different covariate values in the selection interval. Due to the well known problems resulting from multiple testing it is obvious that the minimum P-value method cannot lead to correct results. However, this problem can be solved by using a corrected P-value as proposed in Lausen and Schumacher (1992), which has been developed by taking the minimization process into account. The formula reads

$$p_{cor} = \varphi(z) \left[z - \frac{1}{z}\right] \log\left[\frac{(1-\varepsilon)^2}{\varepsilon^2}\right] + 4\,\frac{\varepsilon(z)}{z} \tag{22}$$

where φ denotes the probability density function and z the $(1 - p_{min}/2)$ - quantile of the standard normal distribution, p_{min} is obtained by the minimum P-value method. The selection interval characterized by the proportion ε of the smallest and largest values of x that are not considered as potential cutpoints. In this thesis I use $\varepsilon = 0.1$. It should be mentioned that other approaches of correcting the minimum P-value could be applied; a comparison of three approaches can be found in a paper by Hilsenbeck and Clark (1996). However, using p_{cor} instead of p_{min} does not solve all problems, because the cutpoint estimate $\hat{\mu}$ and the estimated log-relative risk $\hat{\beta}$ is still obtained by using the same data set. This may lead to a drastic overestimation of the log-relative risk β, if the true value is small or moderate ($\mid \beta \mid \leq 0.7$) (Altman et al., 1994; Schumacher et al., 1997). In order to correct for overestimation it has been proposed (Van Houwelingen and Le Cessie, 1990) to shrink the parameter estimates by a so called shrinkage factor. Considering the cutpoint model the log-relative risk should then be estimated by

$$\hat{\beta}_{cor} = \hat{c} \cdot \hat{\beta} \tag{23}$$

where $\hat{\beta}$ is based on the minimum P-value method and $\hat{c}$ is the estimated shrinkage factor. Values of $\hat{c}$ close to one should indicate a minor degree of overestimation whereas small values of $\hat{c}$ should reflect a substantial overestimation of the log-relative risk. Obviously, with maximum partial likelihood estimation of c in a model

$$\lambda(t \,|\, X > \mu) = \exp{(c\,\hat{\beta})}\lambda(t \,|\, X \leq \mu) \tag{24}$$

using the original data we get $\hat{c} = 1$ since $\hat{\beta}$ is the maximum partial likelihood estimate. In a recent paper we compared several methods for estimation of $\hat{c}$ (Schumacher et al., 1997). Here I use the so called heuristic estimate $\hat{c} = (\hat{\beta}^2 - var(\hat{\beta}))/\hat{\beta}^2$ where $\hat{\beta}$ and $var(\hat{\beta})$ are resulting from the minimum P-value method (Van Houwelingen and Le Cessie, 1990). Besides this heuristic estimator Van Houwelingen and Le Cessie (1990) and Verweij and Van Houwelingen (1993) propose cross-validation calibration. For leave-one-out cross-validation, let $\hat{\beta}_{(-i)}$ denote the estimated regression coefficient obtained when the patient i with covariate value $\tilde{x}_i = 1_{\{x_i > \hat{\mu}\}}$ and observed survival time T_i is removed from the data, where $\hat{\mu}$ denotes the estimated cutpoint in the original data set. Then the 'score' $(\tilde{x}_i - \bar{\tilde{x}})\hat{\beta}_{(-i)}$ can be seen as a 'predictor' for the 'new' observation T_i, where the parameter estimate $\hat{\beta}_{(-i)}$ is independent of $\tilde{x}_i$. In order to assess its predictive potential $(\tilde{x}_i - \bar{\tilde{x}})\hat{\beta}_{(-i)}$ is included as the only covariate for the patient i with survival time T_i in a Cox regression model using all data. The estimated regression coefficient of this covariate obtained by maximizing the corresponding partial likelihood can be used as a shrinkage factor. It should be stressed that the standardization of the binary variable $\tilde{x}_i$ is necessary, because the estimation of the shrinkage factor should be based on all estimated regression coefficients $\hat{\beta}_{(-i)}$ and not only on those where $\tilde{x}_i = 1$. For the categorization using one data driven cutpoint the heuristic estimate performed quite well when compared to cross-validation calibration and other more elaborated resampling approaches (Schumacher et al., 1997).

Categorizing x by using several data-driven cutpoints several parameters has to be estimated (cf. formula (9)). An overall cross-validation calibration shrinkage factor can be obtained by including $\sum_{j=1}^{3} (\tilde{x}_{ij} - \bar{\tilde{x}}_j)\, \hat{\beta}_{(-i)j}$, as the only covariate for patient i with survival time T_i in a Cox regression model using all data (here assuming the maximum of 4 final subgroups). The adjusted 'predictor' is then estimated by $\hat{h}(x) = \hat{c}(\hat{\beta}_1\tilde{x}_1 + \hat{\beta}_2\tilde{x}_2 + \hat{\beta}_3\tilde{x}_3)$. Note, that the same shrinkage factor is applied to all parameter estimates $\hat{\beta}_i$. However, this may not be sensible in practical application and, therefore, we use a slight generalization of the cross-validation approach. Considering all factors $(\tilde{x}_{i1} - \bar{\tilde{x}}_1)\hat{\beta}_{(-i)_1}$, $(\tilde{x}_{i2} - \bar{\tilde{x}}_2)\hat{\beta}_{(-i)_2}$ and $(\tilde{x}_{i3} - \bar{\tilde{x}}_3)\hat{\beta}_{(-i)_3}$ instead of their sum as covariates for the i-th individual in a Cox regression model, we obtain three shrinkage factors $\hat{c}_1, \hat{c}_2$ and $\hat{c}_3$. The resulting estimated risk function is given by

$$\hat{h}(x) = \hat{c}_1\hat{\beta}_1\tilde{x}_1 + \hat{c}_2\hat{\beta}_2\tilde{x}_2 + \hat{c}_3\hat{\beta}_3\tilde{x}_3 \tag{25}$$

Since this approach lead to one shrinkage factor for each estimated parameter $\hat{c}_i$ are referred to as parameterwise shrinkage factors (Sauerbrei et al., 1999). In this thesis I investigate the effect of parameterwise shrinkage factors for the CART based categorization.

It should be briefly mentioned that Van Houwelingen and Le Cessie (1990) propose a simple extension of the heuristic shrinkage factor in multiple linear regression. However, for survival data this approach would be more complicated.

Describing the functional relationship by a fractional polynomial the fitted risk function is also resulting from multiple testing. However, since the number of tests is smaller (when using a restricted set for p_i) as compared to the cutpoint model and prior investigations show that this approach holds the type I error rate (Royston and Altman, 1994) over-optimism of the final regression model may be less extreme. An adaption of the shrinkage procedure to fractional polynomials seems to be rather complicated. Note that both, the parameters β_i and the powers p_i, was estimated in the selected model.

In the simulation study I will investigate for potential over-optimism of all methods. Furthermore, type I error rates will be estimated for all methods and different model selection strategies. Stability of the model selection process and, therefore, of the estimated risk function is investigated by applying the same model building strategy as in the original data set to a set of bootstrap samples. In order to extend the methods described so far the results of the bootstrap samples are aggregated to determine a so called bagged risk function (cf. section 2.2).

Generally, the results obtained by the 'final' regression model should be validated using an independent study - particularly if model building and estimation of the resulting risk function is performed within the same data set. As mentioned earlier the Freiburg DNA study is used to validate the results obtained in the GBSG-2 study.

2.1.8 Overview

All methods described above are summarized in table 2.1 For the CART based categorization and the fractional polynomial approach the table contains the risk function of the highest order only, i.e. a step function describing the relative risk between 4 final subgroups and a fractional polynomial of degree 2, respectively. For each method I selected a label (cf. table 2.1), which will be used in the sequel when describing the results of the application and the simulation study. As described earlier the linear risk function is used as basic assumption or reference model for all approaches.

Table 2.1: Standard procedures to estimate the functional form of a continuous covariate ($1_{\{x>..\}}$ denotes the indicator function)

effect	estimated risk function $\hat{h}(x)$	label
linear[1]	$\hat{\beta}\, x$	**LIN**
linear & quadratic	$\hat{\beta}_1\, x + \hat{\beta}_2\, x^2$	**LINQ**
Categorization based on		
- one prespecified cutpoint μ_0	$\hat{\beta}\, 1_{\{x>\mu_0\}}$	**FIX**
- two prespecified cutpoints $\mu_{01} < \mu_{02}$	$\hat{\beta}_1\, 1_{\{\mu_{01}<x\leq\mu_{02}\}} + \hat{\beta}_2\, 1_{\{x>\mu_{02}\}}$	**FIX2**
Categorization based on		
one data driven cutpoint $\hat{\mu}$		
- using p_{min} without shrinkage	$\hat{\beta} 1_{\{x>\hat{\mu}\}}$	**CUT**
- using p_{min} with shrinkage	$\hat{c}\hat{\beta} 1_{\{x>\hat{\mu}\}}$	**CUTS**
- using p_{cor} without shrinkage	$\hat{\beta} 1_{\{x>\hat{\mu}\}}$	**CUTC**
- using p_{cor} with shrinkage	$\hat{c}\hat{\beta} 1_{\{x>\hat{\mu}\}}$	**CUTCS**
CART based categorization		
- using p_{min} without shrinkage	$\hat{\beta}_1\, 1_{\{\hat{\mu}_1<x\leq\hat{\mu}_2\}} + \hat{\beta}_2\, 1_{\{\hat{\mu}_2<x\leq\hat{\mu}_3\}} + \hat{\beta}_3 1_{\{x>\hat{\mu}_3\}}$	**CART**
- using p_{min} with shrinkage	$\hat{c}_1\hat{\beta}_1 1_{\{\hat{\mu}_1<x\leq\hat{\mu}_2\}} + \hat{c}_2\hat{\beta}_2 1_{\{\hat{\mu}_2<x\leq\hat{\mu}_3\}} + \hat{c}_3\hat{\beta}_3 1_{\{x>\hat{\mu}_3\}}$	**CARTS**
- using p_{cor} without shrinkage	$\hat{\beta}_1\, 1_{\{\hat{\mu}_1<x\leq\hat{\mu}_2\}} + \hat{\beta}_2\, 1_{\{\hat{\mu}_2<x\leq\hat{\mu}_3\}} + \hat{\beta}_3 1_{\{x>\hat{\mu}_3\}}$	**CARTC**
- using p_{cor} with shrinkage	$\hat{c}_1\hat{\beta}_1 1_{\{\hat{\mu}_1<x\leq\hat{\mu}_2\}} + \hat{c}_2\hat{\beta}_2 1_{\{\hat{\mu}_2<x\leq\hat{\mu}_3\}} + \hat{c}_3\hat{\beta}_3 1_{\{x>\hat{\mu}_3\}}$	**CARTCS**
restricted cubic spline	$\hat{\beta}_1 x + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_{n-1} x_{n-1}$	**RCS**
fractional polynomial	$\hat{\beta}_1 x^{\hat{p}_1} + \hat{\beta}_2 x^{\hat{p}_2}$	**FP**

[1] a linear functional relationship is assumed as basic assumption/reference model

2.1.9 Application I: Estimation and Validation of the effect of age in the two breast cancer studies

In this section the procedures summarized in table 2.1 are applied to the data of the GBSG-2 study and the Freiburg DNA study. As mentioned in the introduction the former study is used for a data-driven selection of the functional form (if not prespecified) as well as for estimating the effect of age on event free survival (EFS). In contrast to the first example in the introduction the Freiburg DNA study is only used for validation, i.e. the effect is estimated within the model selected in the GBSG-2 study. The categorization by prespecified cutpoints (FIX2) is based on the two cutpoints 45 years and 60 years. Except for LIN (the linear function is already displayed in figure 1.2) all estimated risk functions obtained in the GBSG-2 study and the Freiburg DNA study are displayed in figure 2.1. Categorizing age by data driven cutpoints I plotted CUTS and CART only. More details on the parameter estimates can be found in appendix A. All functions are standardized such that the mean log-relative risk in the data is zero. I will discuss this point in detail in section 3.2.

In the GBSG-2 study all methods show a decreasing risk with increasing age up to the age of 50 years. However, except for LIN and CUT we observe again a slight increase in risk for older patients. For RCS there is additionally a decrease in risk for patients older than 65 years. However, using only 3 knots the shape of the restricted cubic spline is similar to that obtained by the two term fractional polynomial (not shown, estimates in figure 2.1 are based on 4 knots). Considering the improvement of the log-likelihood as a measure of the model fit we obtained a better fit for the data based methods (CUT, CART, FP). However, the smallest value of the log-likelihood is obtained when using a RCS to describe the effect of age (cf. table A 2.1 in the appendix). With this approach, a data dependent model building is only performed in so far that the knots are quantiles of the data (cf. section 2.1.6). It is obvious from figure 2.1 that nearly all methods used to estimate the functional relationship of the effect of age with respect to EFS in the GBSG-2 study suggest a nonlinear effect. Testing LINQ, FP and RCS against LIN by the corresponding likelihood-ratio test the gain of the more complex function is always significant. These method also show a significant effect of age when testing the model against the null model whereas there is no effect at all for LIN. Since the model building process used with CUT and CART led to a significant cutpoint the corresponding value of the likelihood ratio test comparing these models to the null model is also significant. However, the corrected P-values are not significant (we obtained p_{cor}=0.07 for the first split) and therefore CUTC and CARTC would reduce to LIN with our model building approach. A test of CUT, CART and FIX2 against LIN is formally not allowed, since the underlying models are not nested.

In the Freiburg DNA study none of the fitted functions lead to a significant reduction of the log-likelihood indicating that there may be no influence of age on EFS at all. However, this study is rather small and the variability of the estimated parameters are large. The best improvement in terms of the likelihood was obtained for the RCS (table A 2.1). Again, all methods show a decrease in risk with increasing age (cf figure 2.1). However, this decrease is substantially smaller as in the GBSG-2 study. Furthermore, there is no increase of the log relative risk for older patients as observed with FIX2, CART, FP and LINQ in the GBSG-2 study. For LINQ the quadratic effect can be neglected in the Freiburg DNA study, whereas the functional shape of FP and RCS are similar in both studies. The difference between the two subgroups obtained by CUT is substantially smaller in the Freiburg DNA study. The results obtained by using CART should be interpreted very cautiously in the Freiburg DNA study, because the reference group (patients $\leq$ 32 years) contains only 2 patients and, therefore, the resulting parameter estimates exhibit a very large variability.

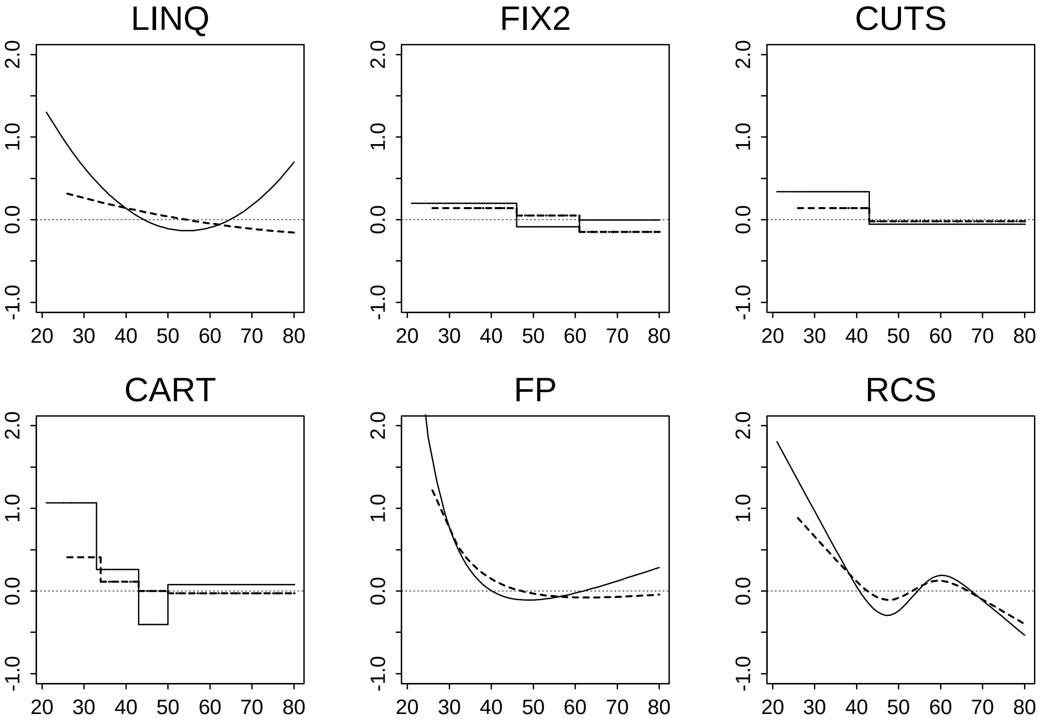


Figure 2.1: Risk function estimates of the effect of age in the GBSG-2 study (solid line) and validation in the Freiburg DNA-study (dashed line)

Recapitulating all results the use of different risk functions may lead to differences with respect to the interpretation of the effect of age on EFS. The results obtained in the GBSG-2 study could only be partially verified by the data of the Freiburg DNA study. The large change in log relative risk observed with CUT and CART in the GBSG-2 study is obviously resulting from the extensive process of model building involved when selecting one or more cutpoints. Furthermore, the huge decrease in risk observed with FP is not very likely, too. And last but not least: Is there an increase in risk for older patients or not? In order to illuminate these questions it is worthwhile to investigate the stability of all risk functions and to search for more robust methods.

2.1.10 Time dependent covariates

All approaches described so far focus on the functional form of the effect of the continuous covariate rather than on the deviation from the proportional hazards (PH) assumption. However, the assumption of a constant effect in time may be wrong. A common example is the treatment effect that decreases with time. In oncological studies patients are usually observed for several years where the effect of the patient's age may also change with time. In this section I investigate the potential time-dependency of age on EFS in the GBSG-2 study. Furthermore, it will be illustrated that the functional form of the effect of a covariate is also related to the PH assumption.

Several graphical methods and tests are available to assess the proportionality of hazards (Marubini and Valsecchi, 1995). I restrict myself to a test proposed in the original paper by Cox (1972) and a simple but effective graphical approach by Grambsch and Therneau (1994).

To investigate the PH assumption Cox proposed to define a time-dependent transform of the continuous covariate by multiplying it by a function $g(t)$ of time, and include it in the classical model:

$$\lambda(t|X=x) = \lambda_0(t)\ \exp(\beta x + \gamma x g(t)) \tag{26}$$

In general, a non-zero value of γ (corresponding to a significant gain as compared to LIN) would indicate a variation in time of the hazard ratio between two individuals with a different value of x. Common choices for $g(t)$ are the identity function $g(t) = t$ and its logarithmic transformation $g(t) = \log(t)$. These functions are often centered around an arbitrary constant to improve interpretability and to avoid instability of the estimates. Investigating the potential time-dependency of age in the GBSG-2 study this constant is chosen as the observed median survival time and the log median survival time, respectively. The resulting models are denoted by LIN & LIN(t) and LIN & LOGLIN(t), respectively. Table 2.2 lists the values of the likelihood ratio test statistic (LRT) against the null model in the GBSG-2 study for several models with and without a time-dependent risk function. For LIN & LIN(t) and LIN & LOGLIN(t) we obtained a significant effect of age. The values of LRT are 8.59 and 8.72, respectively, which is larger than the corresponding value of χ^2 distribution with 2 degrees of freedom $\chi^2_{0.95} = 5.99$. Comparing LIN & LIN(t) to LIN the resulting test statistic with 1 degree of freedom has the value 8.59-0.58=8.01 indicating a strong effect of the time-dependent term LIN(t). Adding a time-dependent term to LIN the improvement in terms of likelihood is as large as with LINQ. However, the value of LRT is substantially smaller than those of FP and RCS. It should be briefly mentioned that the results are similar for the multivariate model including the other prognostic factors as time-independent covariates with a linear risk function. So what to do? Is there a time-dependent effect of age or is the assumption of a linear effect wrong? Last but not least there may be both, a nonlinear and a time-dependent effect. Adding e.g. a linear time-dependent effect to LINQ increases the likelihood ratio test statistic significantly from 8.99 for LINQ to 14.76 for LINQ & LIN(t).

Table 2.2: Investigating time-dependency of the effect of age on the hazard of EFS in the GBSG-2 study

risk function	df	LRT[1]	$\hat{\rho}$	PH-Test
LIN	1	0.58	0.147	7.76
LIN & LIN(t)	2	8.59	–	–
LIN & LOGLIN(t)	2	8.72	–	–
LINQ	2	8.99	-0.073	2.08[2]
LINQ & LIN(t)	3	14.76	–	–
FP	4	17.63	0.007	0.55[2]
RCS	3	21.69	-0.040	0.01[2]

[1] likelihood ratio test statistic against the null model

[2] based on multiple regression coefficients

Considering the same model as described above Grambsch and Therneau (1994) show that smoothed plots of the standardized Schoenfeld residuals can reveal the form of $g(t)$. Restricting to the observed events $j = 1, ..., J$ the Schoenfeld residuals are defined by $\hat{r}_j = x_j - E(x_j \mid R_j)$, which is the difference of the observed covariate for the individual dying at time point j and the expected value of this covariate under the model (R_j denote the risk set at time j). Standardized or scaled Schoenfeld residuals r_j^* can be obtained by dividing $\hat{r}_j$ by its variance estimate. Note that for more than one covariate $\hat{r}_j$ and $\hat{r}_j^*$ are vectors, and that standardization is based on the inverse of the covariance matrix of $\hat{r}_j$. Since the covariance matrix tends to be fairly constant over time Grambsch and Therneau (1994) propose an easy approximation to calculate r_j^*. Furthermore, they show that $\hat{r}_j$ and their approximation have a mean at time t of approximately

$$E(r^*(t)) = \gamma \, g(t) \tag{27}$$

This result suggests that a plot of the scaled Schoenfeld residuals over time including a smoothing line may be used to visually assess whether the coefficient γ is equal to zero and, if not, of what nature the time dependence may be. Grambsch and Therneau derive a generalized least squares estimator for γ (all γ_i, $i = 1, ..., k$, if there are k covariates) and a score test (denoted as PH-test) to test the hypothesis H_0: $\gamma = 0$. Under the null hypothesis the PH-Test statistic has asymptotically a χ^2 distribution with 1 degree of freedom. It should be briefly mentioned that many tests of proportional hazards are closely

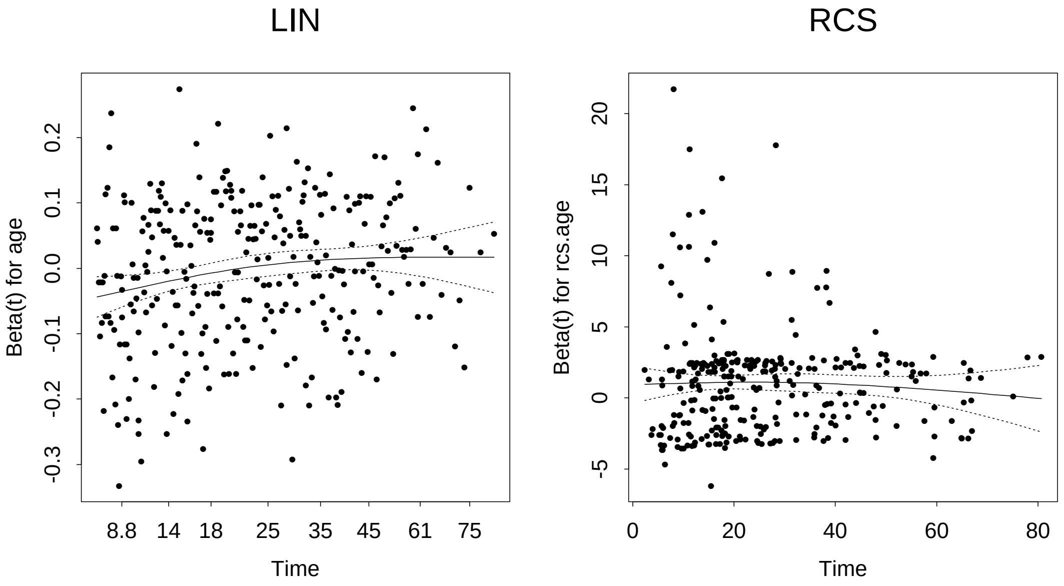


Figure 2.2: Plot of scaled Schoenfeld residuals versus time with smoothing lines for LIN and RCS in the GBSG-2 study

related to this score test. For instance, the score test of Grambsch and Therneau (1994) is equal to that proposed by O'Quigley and Pessione (1989) if g is piecewise constant on non-overlapping time intervals with the intervals and constants chosen in advance. Additionally to the 'correlation with time' score test and the corresponding residual plot time-dependency can also be described by the correlation $\hat{\rho}$ between r_j^* and survival time, here $\hat{\rho}$ is simply Pearson's correlation coefficient. Assuming a linear risk function we obtain a correlation of $\hat{\rho} = 0.147$, the resulting residual plot (figure 2.2) suggests a slight increase of the effect of age with increasing time. Although the PH-test also show a significant time-dependent effect ($\chi^2 = 7.76$, $p = 0.005$, corresponding well to the results obtained above for LIN & LIN(t) and LIN & LOGLIN(t)) this effect seems to be rather small. The smoothing line, which is obtained by fitting the scaled Schoenfeld residuals by a 4 knots restricted cubic spline, increases only slightly with increasing time. A time-dependent effect can not be seen easily without this smoothing line.

Considering multiple regression coefficients for one covariate as done when assuming LINQ, FP or RCS an approximate analysis can be performed in which the covariate is transformed by the estimated risk function (Harrell, 1997) leading again to one term per covariate. Then, the analysis is performed as described above. Doing so it should be taken into account that the score test is based on the false number of degrees of freedom (namely one), and that the risk function used for the transformation was estimated from the data.

As listed in table 2.2 the PH score test does not indicate a time dependent effect of age when using LINQ, FP or RCS as risk function. The value $\hat{\rho}$ is approximately zero for all models and the smoothing line is constant over the time. As an example the residual plot is displayed for RCS in figure 2.2, similar plots would be obtained for LINQ and FP. Although the choice of RCS provides the best fit in terms of the log-likelihood, this risk function is not appropriate for the complete data set. As shown in figure 2.2 there are several observations with extremely large residuals r_j^*.

Analyzing the data of the GBSG-2 study with respect to potential time-dependency of age, we can at least exclude a strong time-dependent effect on EFS. Furthermore, the residual plots have shown that the choice of a specific risk function is also related to the PH assumption. Generally, it would be desirable to investigate both assumptions, the functional form of the risk function and potential time-dependency together. However, in this thesis I focus on the functional form.

2.1.11 Further approaches

In this section I give a short overview on some further data-driven methods for determining the risk function. Some of the methods described in this section are similar to the data driven cutpoint model and the CART based categorization.

Studying also the effect of age on survival in breast cancer Contal and O'Quigley (1999) used a data driven cutpoint to divide the patients population into a high risk and a low risk group. In contrast to the approach described in section 2.1.4 their approach *avoids arbitrarily eliminating potential cutpoints near the extremities'*. They point out that cutpoints near to the boundaries of the covariate's distribution *'may be real or may reflect the presence of outliers'*. To identify the 'optimal' cutpoint and to estimate its significance Contal & O'Quigely construct a process that asymptotically behaves like a Brownian bridge and use classical properties of the Brownian bridge. Doing so Contal & O'Quigely are able to consider all observed values of the continuous covariate as potential cutpoint. Remember, that the cutpoint of 35 years (used as fix cutpoint in section 1.2) was outside the selection interval when using the minimum P-value approach. Since this cutpoint produced a higher risk difference as compared to the categorization by 42 years (obtained by the minimum P-value method), 35 years would probably have been selected when applying the method of Contal & O'Quigely.

Another method to estimate the functional form is the local likelihood approach of Tibshirani and Hastie (1987). Considering only a subset of the data within a given window (around each value of the continuous covariate X) a Cox regression model is fitted using the linear (or another continuous) risk function. Displacing the window step by step from small to large values of X one obtains a set of risk estimates. Then, the trapezoidal rule is used to combine these estimates leading to an estimate of the functional form of the effect of X. Using this local likelihood method to investigate the effect of S-Phase fraction on EFS in the Freiburg DNA study we have shown that the resulting risk function may strongly depend on the size w of the window (Schumacher et al., 1996). Tibshirani and Hastie (1987) propose to try a range of spans (= window sizes) and *'examine the resulting estimate and the value of the global likelihood that it produces'*. The authors propose also an automatic method for selecting w that is based on a form of Akaike's information criterion (AIC) (Akaike, 1973).

As already mentioned in section 2.9 the functional form of a continuous covariate may be visualized by using residual plots. Therneau et al. (1990) suggested the use of martingale based residuals. These residuals are estimated by

$$\hat{M}_i = \delta_i - \hat{\Lambda}_0(t_i)\ \exp(\hat{\beta} X_i), \tag{28}$$

where t_i is the observation time and δ_i the censoring indicator for the i-th individual, $\hat{\Lambda}_0(t_i)$ denotes the cumulative baseline hazard and the risk function is assumed to be linear. The martingale residual can be interpreted as the difference of the number of observed events (0 or 1) and the number of expected events under the model. Suppose now that the covariate x (one covariate x_k out of x if x is a vector) strongly influences survival in a nonlinear fashion, for example with earlier failures occurring at higher values of x. It is likely that in the Cox model without x - and that is the basic assumption of Therneau et al. (1990) - the residuals $\hat{M}_i$ plotted against x would not be centered around zero as expected. They could tend instead to be positive for larger values and negative for smaller values of x. The functional form of the effect of x may then be visualized by a smoothing line. Plotting the martingale-based residuals of the null model against age for the data of the GBSG-2 study the smoothing line – I used lowess (Cleveland, 1979) – suggests a similar functional shape as obtained when fitting a 4 knot RCS to the data (figures 2.1 and 2.3). This illustrates that different concepts may lead to similar results.

As done in the data driven cutpoint model (CUT, CUTS) Le Blanc and Crowley (1995) also consider piecewise constant relative risk functions to describe the effect of one or more continuous covariates on survival. Similar to section 2.1.4 they propose a two step technique for fitting an adaptively chosen step function for a continuous covariate X. In addition to X their model may contain also additional covariates. The models reads

$$\lambda\,(t \mid X, \mathbf{Z}) = \lambda_0(t)\ \exp(\beta_{\mathbf{1}}^{\mathbf{T}}\,\mathbf{Z} + \beta_2\, g(\mathbf{Z}) 1_{\{X \leq \mu\}})\,, \tag{29}$$

where $\mathbf{Z}$ is the vector of additional covariates and $\beta_{\mathbf{1}}$ denotes the corresponding parameter vector. In order to find the approximate maximum likelihood estimates first, a *'good*

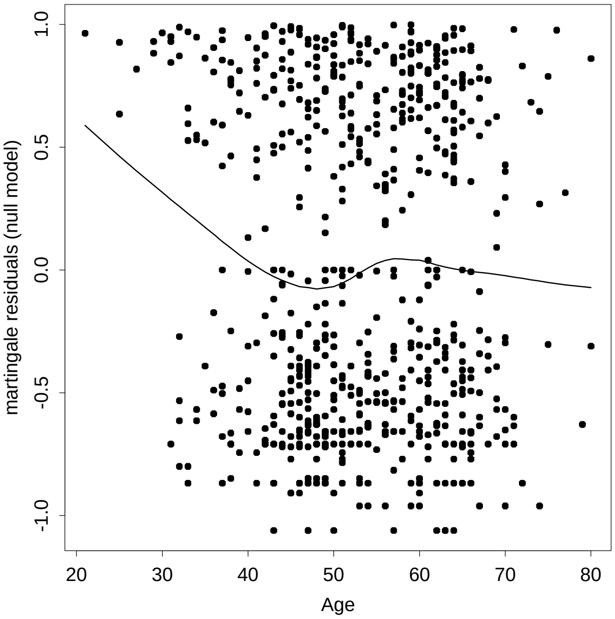


Figure 2.3: Martingale residuals for the null model versus age in the GBSG-2 study including a lowess smoother

value' for the cutpoint is estimated by a weighted least-squares method which allows *'efficient updating'* for different values of μ. To do so Le Blanc and Crowley (1995) proposed a standardized approximate score statistic that is based on the weighted sum of squares including the term $g(\mathbf{Z})1_{\{X \leq \mu\}}$ in a linear model. Considering only one continuous covariate, i.e. for the special case $g(\mathbf{Z}) = 1$ and $\beta_{\mathbf{1}} = 0$, the approximate score statistic is just the logrank test for two groups defined by the cutpoint. After estimating μ by maximization of the score statistic over a selection interval $a \leq \mu \leq b$ the parameters $(\hat{\beta}_{\mathbf{1}}^{\mathbf{T}}, \hat{\beta}_2(\hat{\mu})$ are obtained by maximizing the corresponding partial likelihood. Tail probabilities of the score statistic and corrected P-values for testing H_0:$\beta_2 = 0$ for an adaptively chosen step function are obtained by a permutation approach. As outlined by Le Blanc and Crowley (1995) for the special case $g(\mathbf{Z}) = 1$ the weighted least squares approach has connection to the residual analysis idea proposed by Therneau et al. (1990) that I described above. Developing models with step functions may often involve repeated use of the step function algorithm, since a model may contain more than one step function. To find an appropriate final model Le Blanc and Crowley (1995) propose a modified Akaike information criterion penalizing – in addition to the number of parameters – the model building process.

In a further paper Le Blanc and Crowley (1999) use adaptive regression splines to explore the effect of continuous covariates. With this approach the risk function consists of piecewise linear terms. Knot positions are selected adaptively, additional constraints are used to avoid placing knots to close to the extremes of the data. Model selection is based on backward elimination using a modified version of AIC, which is similar to that described in their step function paper. Although adaptive regression splines may suggest departures from the linear risk functions Le Blanc and Crowley (1999) point out that their *'modeling technique does not test whether the piecewise term is better than the linear term, the technique simply finds the best piecewise linear term'.*

In a recently published paper Xu and Adak (2001) use a tree based approach to approximate the time-varying regression effect of several prognostic factors in breast cancer as piecewise constants. A fast algorithm that relies on maximized modified score statistics is designed for recursive segmentation of the time axis. Following segmentation, i.e. the tree growing step, some of the segments are then recombined by using a pruning algorithm similar to that of the classical CART procedure (Breiman et al., 1984). Based on the finally selected change-points (=cutpoints) step functions of the time-varying effect of the prognostic factors are estimated in the proportional hazards model. Similar to CUT/CART that I used to estimate a step function for the effect of a time-constant continuous covariate the approach of Xu and Adak is also based on an extensive process of model building. In order to correct for over-optimism due to change-point optimization the authors propose a bootstrap procedure for the final segmentation of the time axis. With this procedure the model selection process is repeated in a set of bootstrap samples. An average of the results over all bootstrap samples are used together with a penalty term for the final determination of change-points. Alternatively to the bootstrap procedure the final model can also be selected by using a measure of the explained variation (Xu and Adak, 2001).

There are lot of further approaches to determine the functional form of the effect of time varying or time-constant covariates. However, it is beyond the scope of my thesis to describe all of them.

2.2 Extensions based on bootstrap resampling

The categorization of a continuous covariate or the specification of its functional relationship with respect to the outcome variable is an important step of model building. However, building a statistical model and estimating the resulting risk function within the same data set may lead to a considerable amount of over-optimism with respect to the predictive ability of the 'final' model. The bootstrap is often used to investigate the influence of data-dependent decisions in model building-strategies (Sauerbrei, 1998).
In order to reduce the variability due to model building Breiman (1996) proposed the concept of bootstrap aggregating (bagging). Considering a series of samples $L_1, \dots, L_r$ and building a predictor in each of these samples he suggested to use an aggregated predictor. The exact method of aggregation depends on the response type. For the regression case the arithmetic mean is used, while for the classification case, e.g. using CART in order to predict a class, a simple voting procedure is involved, so that the aggregated predictor assigns that class to an element which was predicted most often in the r single predictors. Breiman shows that, in the case of independent samples $L_1, \dots, L_r$, the aggregated predictor is always at least as good as a single predictor. However, since the luxury of independent samples is available in simulations only, Breiman suggested to use bootstrap samples instead.
In this thesis bagging is adapted to the current situation. I consider B bootstrap samples $S_1, \dots, S_B$ of the complete patient's vector (i.e. the survival time variable, the censoring indicator and the covariate X) sampled with replacement out of the original data (Efron and Tibshirani, 1993). In each of the B bootstrap samples the functional form of the covariate X is determined with the methods described in section 2.1. Doing so, the same model selection process as in the original data is applied in each bootstrap sample. Using LIN and FIX2 there is no model selection involved and, therefore, the general functional form is the same in all bootstrap samples. Consequently, the aggregated function that will be referred to as bagged risk function h_{bagg} is obviously of the same type as that of the corresponding standard procedure. For the other approaches the form of the fitted function may differ between bootstrap samples (e.g. FP_1 or FP_2) and, therefore, it may be difficult to determine a specific form for the resulting bagged function. However, it is sufficient to aggregate the bootstrap results for fixed values of the continuous covariate. I use all observations of the original data, more details will be given below. Formally, a bagging estimate of the functional form of the continuous covariate X is given by

$$h_{bagg}(x) = \frac{\sum\limits_{b=1}^{B} h_b(x)}{B}, \tag{30}$$

where $h_b(x)$ is the risk function obtained in the b-th bootstrap sample ($b = 1, \dots, B$).
Besides the determination of a bagged function, that may be more stable than one single

function, we explore the results of bootstrap resampling in order to get further insight into the stability of the estimated risk functions obtained by the underlying standard procedure. In particular, the results of all bootstrap samples are illustrated graphically. Furthermore, we count the frequencies of linear and higher order risk functions.

a) Linear relationship (LIN)

Since the assumption of a linear relationship is considered as basic assumption (significant and non-significant effects are both considered as linear) the bagging estimator

$$h_{bagg}(x) = \overline{\beta}^B x = \frac{\sum\limits_{b=1}^{B} \beta_b x}{B} \tag{31}$$

is also linear in x.

b) Linear and quadratic term (LINQ)

For this approach the functional form of the b-th bootstrap sample is assumed to be linear ($h_b(x) = \beta_1 x$) if the hypothesis $H_0 : \beta_2 = 0$ cannot be rejected and the effect is assumed to be linear and quadratic ($h_b(x) = \beta_1 x + \beta_2 x^2$) if $H_0 : \beta_2 = 0$ is rejected. Due to the fact that the functional form differs between bootstrap samples h_{bagg} is determined for all observations $x_1, \ldots, x_n$ in the original data. The functional form of h_{bagg} can be obtained by using $\beta_2 = 0$ for all bootstrap samples with a linear relationship. Let l be the number of bootstrap samples with a linear risk function and q the number of bootstrap samples where the risk function contains a linear and a quadratic term ($B = l + q$) with corresponding sets $M(l)$ and $M(q)$. Then h_{bagg} can be written by

$$\begin{aligned} h_{bagg}(x) &= \frac{1}{B}\sum_{b=1}^{B} \beta_{1b}\, x + \frac{1}{B}\left(\sum_{b \in M(q)} \beta_{2b}\, x^2 + \sum_{b \in M(l)} 0\, x^2\right) \\ &= \overline{\beta}_1^B\, x + \overline{\beta}_2^B\, x^2, \end{aligned}$$

i.e. h_{bagg} consists of a linear and a quadratic term.

c) Categorization based on prespecified cutpoints (FIX)

This approach uses the same categorization in each bootstrap sample and, therefore, h_{bagg} is given by $h_{bagg}(x) = \overline{\beta}^B \cdot x^*$ with $x^* = 1_{\{x > \mu_0\}}$ when using one cutpoint μ_0 and $h_{bagg}(x) = \overline{\beta}_1^B x_1^* + \overline{\beta}_2^B x_2^*$ (x_1^*, x_2^* as defined in section 2.1.3) when using two cutpoints.

d) Categorization using data driven cutpoints (CUT, CART)

In contrast to c) the cutpoint(s) may differ between the bootstrap samples and, consequently, the functional form of h_{bagg} cannot be given explicitly. As described in b) h_{bagg} is determined for all observed values of X. If there is at least one significant cutpoint ($p_{min} \leq 0.05$) the functional relationship between X and the survival time variable T

is described by a step function describing a change in relative risk at each cutpoint (cf. 2.1.4). Consequently the bagged function is also a step function with changes in risk at each cutpoint obtained in any bootstrap sample. For all bootstrap samples with no significant cutpoint ($p_{min} > 0.05$) we assume a linear relationship and correct the bagged step function by the mean linear effect: Let k be the number of bootstrap samples with at least one significant cutpoint with corresponding set $M(k)$. Based on $M(k)$ I calculate h^k_{bagg} by aggregating k single step functions. The bagged risk function for all bootstrap samples is then obtained by

$$h_{bagg}(x) = \frac{k}{B} h^k_{bagg}(x) + \frac{l}{B}\overline{\beta}^l x, \tag{32}$$

where $\overline{\beta}^l$ is the mean linear effect of the $l = (B - k)$ bootstrap samples with no significant cutpoint.

e) Fractional polynomials (FP)

The functional form of the b-th bootstrap sample is either described by a two term fractional polynomial (FP_2) or a one term fractional polynomial (FP_1) or it is assumed to be linear (cf. section 2.1.5). Furthermore, the power(s) of the one or two term fractional polynomial may differ between bootstrap samples. Therefore, the functional form of h_{bagg} would be more complex as the linear and quadratic function obtained when including a linear and a (potential) quadratic term only. However, in principle h_{bagg} could be described as fractional polynomial by setting non selected terms to zero as done with LINQ.

f) Restricted cubic spline (RCS)

A restricted cubic spline is only selected if the corresponding likelihood ratio test show a significant improvement as compared to the assumption of a linear risk function (cf. section 2.1.6). It should be noted that the knots can differ slightly between the bootstrap samples. As for the fractional polynomials h_{bagg} is calculated for all observed values of X in the original data set.

2.2.1 Application II: Estimation of bagged risk functions for the effect of age in the GBSG-2 study

In order to estimate the bagging estimates of the functional form of the effect of age I drew 100 bootstrap samples of the complete patients vector (EFS, the censoring indicator, age) out of the original data of the GBSG-2 study. In each of these bootstrap samples the functional form and/or the corresponding parameters were estimated by using the standard methods. As in the original data in each bootstrap sample the functions are standardized such that the mean log relative risk of all patients is zero. Furthermore, the model building process described in sections 2.1.2 - 2.1.6. is applied in each bootstrap sample.

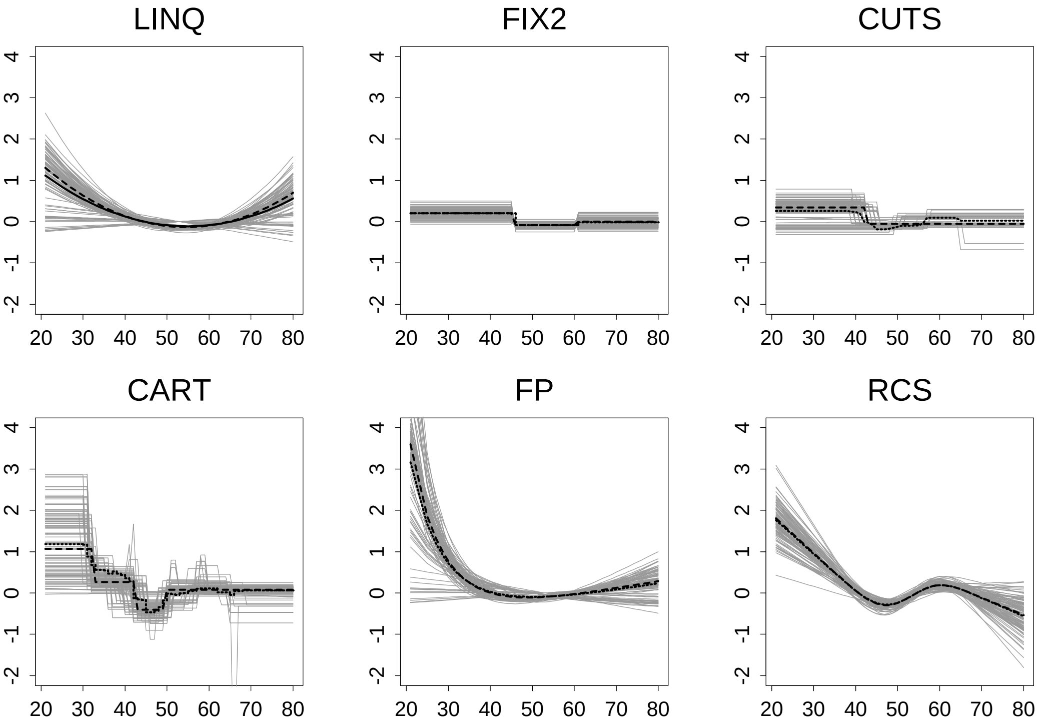


Figure 2.4: Estimated risk functions in 100 bootstrap samples of the GBSG-2 study and comparison of the estimated bagged risk function to the risk function obtained in the original data, the black dashed line describes the risk function obtained in the original data,h_{bagg} is given by the black solid line

For LINQ the quadratic effect was significant in 77 bootstrap samples, whereas the model reduced to a linear risk function in 23%. A restricted cubic spline was selected in all 100 bootstrap samples. For FP a linear risk function was selected in 12, a fractional polynomial of degree 1 in 16, and a two-term fractional polynomial in 72 bootstrap samples. None of these FPs had the same powers as obtained in the original data of the GBSG-2 study ($\hat{p}_1 = -0.5$ and $\hat{p}_2 = -2$). However, the functional form obtained in most bootstrap samples is similar to that estimated in the original data. Using CUT/CUTS or CART the continuous covariate age was categorized in 97 bootstrap samples. In 3 bootstrap samples there was no significant cutpoint and, therefore, I assumed a linear risk function. Taking the corrected instead of the minimum P-value the rate of categorized risk functions reduced to 63 percent. The estimated standardized risk functions for LINQ, FIX2, CUTS, CART, FP and RCS of all bootstrap samples are displayed in figure 2.4. As in the original data for each method most functions show a decrease in risk with increasing age up to 50 years. Some of the methods show again also a slight risk increase for older patients. It is obvious from figure 2.4 that the variability of the estimated risk functions between bootstrap samples is – except for FIX2 – more relevant in the tails of the age distribution. These results should be taken into account when considering the bagged risk function (solid black lines in figure 2.4). Comparing $\hat{h}_{\mathrm{bagg}}$ to the corresponding risk function obtained in the original data of the GBSG-2 study (dashed black line in figure 2.4) there is no difference for LIN (not shown) and FIX2. This is not astonishing, because the functional shape is completely prespecified with both approaches and, therefore, is the same in all bootstrap samples. Consequently, the variability between bootstrap samples is only small. Averaging over a set of random bootstrap samples led us to the same result as in the original data. This explanation also holds for RCS. In all bootstrap samples the restricted cubic spline produced a better fit than the linear risk function. Due to the small variation with respect to the knots the functional shape is similar in each bootstrap sample. Therefore, $\hat{h}_{\mathrm{bagg}}$ differs only slightly from the RCS obtained in the original data. For LINQ and FP the bagged risk function is less extreme in the tails of the age distribution resulting from the fact that a linear risk function was selected in some bootstrap samples. However the difference as compared to the 'original' function is only small. The largest difference can be observed for CUT/CUTS and CART, respectively. However the functional shape of the bagged risk function is still similar to that obtained in the original data, but $\hat{h}_{\mathrm{bagg}}$ is more smooth.

The results described above agree with the Breimans basic result that bagging may work if the underlying model building process is unstable (Breiman, 1996). In the current situation instability occurs when the model selection process produce different risk functions in different bootstrap samples. In the GBSG-2 study the nonlinear effect of age on EFS seems to be relatively strong. Therefore, the models building process led to similar results in most bootstrap samples.

2.3 Summary

In section 2.1 I introduced different methods to estimate the risk function for the effect of a continuous covariate in the Cox proportional hazards model. Section 2.2 describes how to calculate aggregated risk functions based on bootstrap samples. To do so it is necessary to repeat the whole model building process in each bootstrap sample. This allows also to investigate for stability of the risk function estimated in the original data.

Applying the proposed methods to estimate the effect of age on event-free survival of breast cancer patients I obtained the following results:

- All methods showed a decrease in risk with increasing age up to 45-50 years. For patients older than 50 years the risk seems to be rather constant. Thus, the effect of age cannot be described correctly by a linear risk function

- The results obtained in the GBSG-2 study could be partially verified in the Freiburg DNA study. However, especially the decrease in risk is substantially smaller in the latter study. Although the Freiburg DNA study is very small and a comparison of result may not be fair, there seems to be a certain amount of over-optimism in the GBSG-2 study

- In most bootstrap samples the estimated risk function is similar to that obtained by the corresponding methods in the original data of the GBSG-2 study. More complex risk functions reduced to the linear effect in a few bootstrap samples only. The highest rate (23%) of linear risk functions is obtained for LINQ, whereas a restricted cubic spline was selected in all 100 bootstrap samples. These results indicate a strong nonlinear effect of age on event-free survival

- Except for the categorization by data-driven cutpoints there is hardly any difference between the bagged risk function and that obtained in the original data. This may be caused by the strong nonlinear effect of age and, therefore, the stability of the risk functions in the bootstrap samples

3 Simulation Study

The application to the data of the two breast cancer studies has shown, that the choice of different risk functions may lead to different results. However, the true functional relationship between age and EFS is of course unknown and, therefore, the stability of the functional form in the bootstrap samples (observed for example for RCS) cannot guarantee that the chosen method is generally adequate. Therefore, I investigate the capability of all methods to provide good estimates of the true functional form of a continuous covariate by a simulation study.

The design of this simulation is presented in section 3.1, and in section 3.2 it will be discussed how to make different risk functions comparable. Measures for assessing the fit of the estimated functions are described in section 3.3. Some notes on software and the concept of programming will be given in appendix B. The results are described in sections 3.4 - 3.7. After discussing further topics, e.g. the estimation of confidence intervals after model building, in section 3.8 a short summary is given in section 3.9.

3.1 Design

A continuous covariate X taken as uniformly distributed on the interval $[1, 2]$ is considered. This interval was selected to avoid problems that may be caused by the logarithmic term of the fractional polynomial risk function for values close to 0. The survival time random variable T is taken from an exponential distribution by using the transformation $T = -(1/\lambda_z)\, log(U)$, where U is uniformly distributed on $[0, 1]$ and λ_z is chosen according to the underlying functional relationship. For simplicity, only uncensored survival times are considered. The simulation is performed using $n = 100$ observations, $R = 1000$ replications and $B = 100$ bootstrap samples in each replication for the procedures, which are based on bootstrap resampling. To avoid too small subgroups when using the CART based categorization we set $n_{min} = 20$ (cf. section 2.1.4).

As given models we consider four situations:

1. The null model with given risk function $h(x) = 0$ for all values x

$$0 \qquad \lambda(t|X = x) = \lambda_0(t)\, exp(0) = \lambda_0(t). \tag{33}$$

2. A proportional hazards cutpoint model

$$\mathrm{I} \qquad \lambda(t\,|X = x) = \lambda_0(t)\, exp(\beta \cdot 1_{\{x>\mu\}}) \tag{34}$$

with $\mu = 1.5$, $\beta = 0.5$ and $\lambda_0(t) = 1$.

3. A linear relationship, in the sequel also referred to as linear model

$$\mathrm{II} \qquad \lambda(t\,|X = x) = \lambda_0(t)\, exp(\beta x) \tag{35}$$

with $\beta = 0.5$.

4. A V-type risk function, i.e. a model where the risk increases linearly to the distance from a given cutpoint μ:

$$\text{III} \qquad \lambda(t\,|X = x) = \lambda_0(t)\, exp(2\beta\,|x - \mu\,|), \tag{36}$$

where $\beta = 0.5$ and $\mu = 1.5$. This model will also be referred to as V-type model.

All models are displayed in figure 3.1. In models I–III it is assumed that the change in risk β is rather moderate. This assumption corresponds well to practical situations, where a large effect of a (new) prognostic factor cannot be expected.

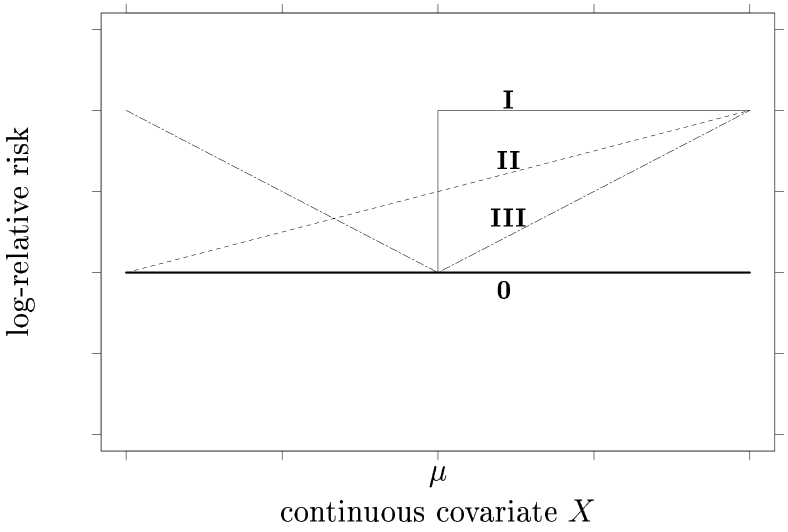


Figure 3.1: Risk function used in the simulation study

3.2 Making results comparable

Using e.g. two different risk functions, say h_1 and h_2, to estimate the true functional relationship g of the effect of a (continuous) covariate X both functions cannot be compared directly, because they refer to two different baseline hazards $\lambda_0^{h_1}(t)$ and $\lambda_0^{h_2}(t)$ in the underlying Cox regression model (cf. formula (1)). This is illustrated by using one selected simulated data set generated for the V-type model (Figure 3.2 a). For instance, the shape of the estimated fractional polynomial is similar to that obtained by LINQ. However the difference with respect to the corresponding baseline hazards lead to a shift in location. To investigate the fit with respect to g and to compare all methods with each other we have to make the results comparable. The estimated risk functions of the two breast cancer studies were standardized such as the mean log-relative risk is zero (cf. figure 2.1). Formally, we considered

$$\hat{h}_{c_1}(x_i) = \hat{h}(x_i) - \frac{1}{n}\sum_{i=1}^{n}\hat{h}(x_i) \tag{37}$$

for all observed values $x_i \quad (i = 1, \ldots, n)$ instead of the fitted values $\hat{h}(x_i)$ leading to the standardization $\sum_{i=1}^{n}\hat{h}_{c_1}(x_i) = 0$. The standardized functional forms obtained in the selected simulated data set is shown in figure 3.2 b.

As mentioned above, lack of comparability is resulting from the difference with respect to the baseline hazard. Since $\lambda_0(t)$ contains the intercept term we calculate

$$\hat{h}_{c_2}(x_i) = log(\hat{\lambda}_0(t^*)) + \hat{h}(x_i) \tag{38}$$

using the Breslow estimator of $\lambda_0(t^*)$ for a fixed time point t^*, which is obtained by

$$\hat{\lambda}_0(t^*) = \frac{d_j}{(t_{(j)} - t_{(j-1)})\sum\limits_{i \in R_j} exp(\hat{h}(x_i))}, \tag{39}$$

where $t_{(j)} - t_{(j-1)}$ is the time interval between two consecutive failure times and $t^* \in (t_{(j-1)}, t_{(j)}]$. The term d_j is the number of observed deaths at $t_{(j)}$ and R_j denotes the risk set at time $t_{(j)}$ (Marubini and Valsecchi, 1995). In contrast to the standardization to zero mean of the log relative risk $\hat{h}_{c_2}$ depends on time via $\lambda_0(t^*)$, where t^* has to be chosen out of the time range between the first and the last observed failure time. Considering two functions h_1 and h_2 (one of them may be the true function g) the difference between the corresponding estimated logarithmic baseline hazards

$$\begin{aligned}
D_{\hat{\lambda}_0(t)} &= log(\hat{\lambda}_0^{\hat{h}_1}(t)) - log(\hat{\lambda}_0^{\hat{h}_2}(t)) \\
&= log(d_j) - log(t_{(j)} - t_{(j-1)}) - log\sum_{R_j} exp(\hat{h}_1(x)) \\
&\quad -log(d_j) + log(t_{(j)} - t_{(j-1)}) + log\sum_{R_j} exp(\hat{h}_2(x)) \\
&= log\sum_{R_j} exp(\hat{h}_2(x)) - log\sum_{R_j} exp(\hat{h}_1(x))
\end{aligned}$$

depends on t via the risk set R_j.

In order to investigate the effect of time figure 3.2 c shows the estimated logarithmic baseline hazard for the linear risk function in the simulated data set. Furthermore we consider the difference $D_{\hat{\lambda}_0(t)}$ with respect to the linear / linear & quadratic risk function and the linear / fractional polynomial risk function, respectively. Although $log(\hat{\lambda}_0(t))$ depends substantially on t the difference between the three functions seems to be the same for all values of t. This phenomenon is also observed when considering further functions, the restriction to three functions is only for better illustration. These results show that we may choose any value of t as fixed time point t^*. However, the value of $D_{\hat{\lambda}_0(t)}$ cannot expected to remain constant over time if the risk set R_j gets small and/or $h_1(x)$ and

$h_2(x)$ differ substantially for individual observations $X = x_i$. The *standardized* functions obtained by using the median survival time as fixed value t^* to calculate the corresponding functions $\hat{h}_{c_2}$ are displayed in figure 3.2 d. The fact that $D_{\hat{\lambda}_0(t)}$ remains constant in time in the simulated data set may be caused by the design of the simulation study: the survival time was generated from an exponential distribution, which is characterized by a constant hazard rate. However, in the GBSG-2 study, where the distribution of survival time may differ from an exponential distribution, we observed similar results (not shown) indicating that $\hat{h}_{c_2}$ can also be used in practical situations.

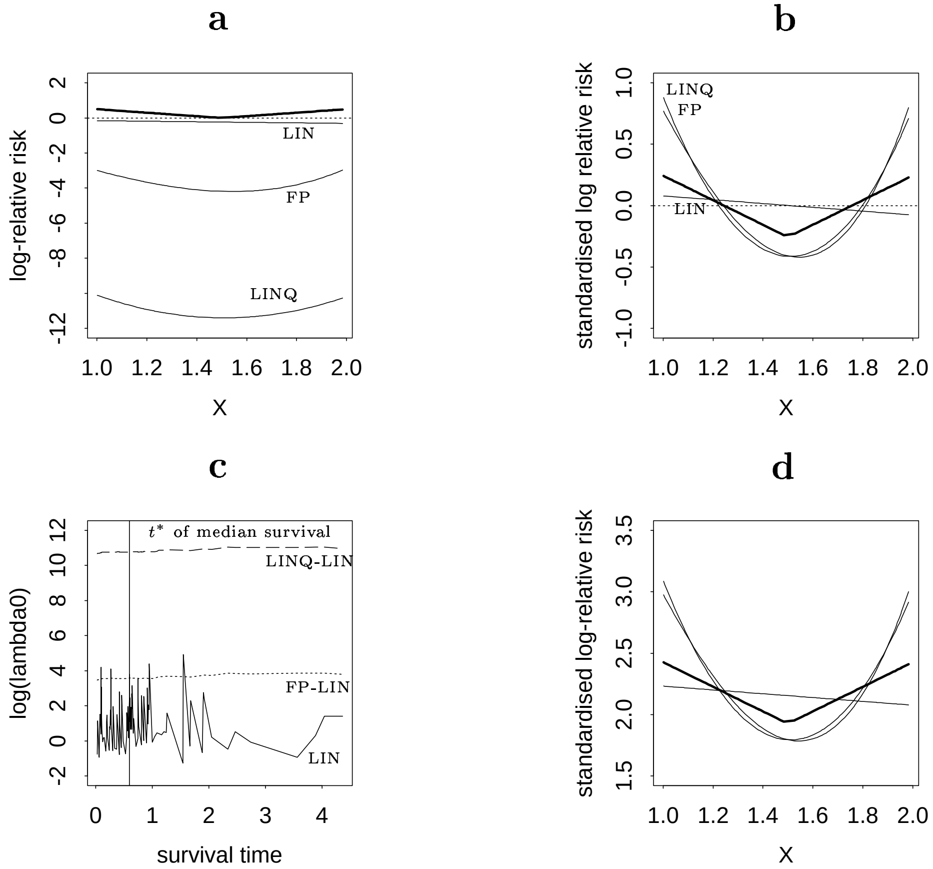


Figure 3.2: Results obtained in the selected simulated data set: **a:** without any *standardization* **b:** standardization to zero mean log relative risk **c:** $\log \hat{\lambda}_0(t)$ and $D_{\hat{\lambda}_0(t)}$ versus survival time for LIN, LINQ-LIN and FP-LIN, respectively **d:** making the functions comparable by adding $\log \hat{\lambda}_0(t^*)$ with t^* corresponding to median survival time

Due to the results of this section it can be expected that both methods for making risk functions comparable would lead to similar results. However, using the standardization to a zero mean log relative risk one should be aware that this approach neglects potential time dependency and censoring. Censoring is taken into account when estimating the baseline hazard whereas time dependency is not investigated when choosing one fix value $t*$.

3.3 Assessment of the fit

Several risk functions have been proposed in order to estimate the true functional relationship g of the effect of a continuous covariate. Here I discuss a few measures for comparing the estimated risk function $\hat{h}(x)$ with the true function $g(x)$.

Quantitative Error

A commonly used measure is the estimated mean squared error, which is - in our design - given by

$$\widehat{MSE} = \int_1^2 (\hat{h}_c(x) - g_c(x))^2 \, dF_n(x) = \frac{1}{n}\sum_{i=1}^{n}(\hat{h}_c(x_i) - g_c(x_i))^2,$$

where g is given by the simulation design (e.g. $g(x) = \beta \cdot 1_{\{x>\mu\}}$ for model I) and F_n denotes the empirical distribution function. Note, that we have to use the *standardized* risk functions $\hat{h}_c$ and g_c, respectively, instead of $\hat{h}$ and g respectively.

In addition to the estimated MSE we will calculate the mean absolute error by

$$\widehat{MAE} = \int_1^2 |\hat{h}_c(x) - g_c(x)| \, dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} |\, \hat{h}_c(x_i) - g_c(x_i)|,$$

a measure that puts the same weight on each observation.

Since the distribution F of X is known in the simulation study, we can also calculate

$$MSE = \int_1^2 (\hat{h}_c(x) - g_c(x))^2 \, dF(x)$$

and

$$MAE = \int_1^2 |\hat{h}_c(x) - g_c(x)| \, dF(x).$$

However, the calculation of MSE and MAE is not straight forward when h is estimated by bootstrap aggregating, because the form of the resulting function $\hat{h}_{bagg}$ cannot be determined explicitly for all methods. Due to the Glivenko-Cantelli theorem F_n converges to F for $n \to \infty$ and, therefore $\widehat{MSE} \approx MSE$ and for $\widehat{MAE} \approx MAE$ for large samples.

Qualitative Error

If the true effect of the continuous covariate is rather moderate as in our simulation study a small quantitative error does not guarantee that the corresponding $\hat{h}$ describes the true functional relationship adequately. Considering the standardized functions obtained in the selected simulated data set (cf. figure 3.2) the quantitative error (e.g. MSE) of the linear risk function can expected to be smaller than the error of the fitted fractional polynomial, although the latter describes the change in risk with increasing values of X better. To take account for this problem we propose a measure of the qualitative error by comparing the change in risk of $\hat{h}_c$ and the true function g_c (or between two function $\hat{h}_1$ and $\hat{h}_2$) for consecutive values of x. For two consecutive values $x_{(j)}$ and $x_{(j+1)}$ we consider

$$Err(j) = \begin{cases} & \text{if} \quad \hat{h}_c(x_{(j)}) > \hat{h}_c(x_{(j+1)}) \wedge g_c(x_{(j)}) > g_c(x_{(j+1)}) \\ 0 & \text{or} \quad \hat{h}_c(x_{(j)}) < \hat{h}_c(x_{(j+1)}) \wedge g_c(x_{(j)}) < g_c(x_{(j+1)}) \\ & \text{or} \quad \hat{h}_c(x_{(j)}) = \hat{h}_c(x_{(j+1)}) \wedge g_c(x_{(j)}) = g_c(x_{(j+1)}) \\ & \\ & \text{if} \quad \hat{h}_c(x_{(j)}) = \hat{h}_c(x_{(j+1)}) \wedge g_c(x_{(j)}) < g_c(x_{(j+1)}) \\ 0.5 & \text{or} \quad \hat{h}_c(x_{(j)}) = \hat{h}_c(x_{(j+1)}) \wedge g_c(x_{(j)}) > g_c(x_{(j+1)}) \\ & \text{or} \quad \hat{h}_c(x_{(j)}) < \hat{h}_c(x_{(j+1)}) \wedge g_c(x_{(j)}) = g_c(x_{(j+1)}) \\ & \text{or} \quad \hat{h}_c(x_{(j)}) > \hat{h}_c(x_{(j+1)}) \wedge g_c(x_{(j)}) = g_c(x_{(j+1)}) \\ & \\ 1 & \text{if} \quad \hat{h}_c(x_{(j)}) > \hat{h}_c(x_{(j+1)}) \wedge g_c(x_{(j)}) < g_c(x_{(j+1)}) \\ & \text{or} \quad \hat{h}_c(x_{(j)}) < \hat{h}_c(x_{(j+1)}) \wedge g_c(x_{(j)}) > g_c(x_{(j+1)}) \end{cases} \quad (40)$$

If both functions $\hat{h}_c$ and g_c are continuous $Err(j)$ can also be written in a shorter form by using the derivatives $\hat{h}'_c$ and g'_c. The qualitative error that will be denoted by Err_{qual} is defined by

$$Err_{qual} = \frac{1}{n-1} \sum_{j=1}^{n-1} Err(j). \quad (41)$$

Obviously, the values of Err_{qual} lies between 0 and 1, where $Err_{qual} = 0$ if the qualitative change in risk (e.g. increase / decrease) is the same for $\hat{h}_c$ and g_c for all observed pairs $(x_{(j)}, x_{(j+1)})$, $j = 1, \ldots, n-1$.

Illustration based on the selected simulated data set for the V-type model

Quantitative and qualitative errors were calculated for the simulated data set used in section 3.2, the results are listed in table 3.1. The $\widehat{MSE}$ and the $\widehat{MAE}$ were calculated for both procedures proposed to make results comparable, whereas MSE and MAE are based on $\hat{h}_{c_2}(x)$ and $g_{c_2}(x)$ using the median survival time as fixed time point t^*. The values of MSE and MAE indicate, that the standardization to a zero mean log-relative risk and the addition of $log(\hat{\lambda}_0(t^*))$ may lead to similar results in the simulation study. Except for the data driven cutpoint model the values of the MSE and the MAE, respectively, are always slightly larger than the corresponding values of the $\widehat{MSE}$ and $\widehat{MAE}$, respectively.

Comparing the quantitative errors between the different methods used to estimate the functional relationship, LIN turned out to be the best method, whereas the results differ only slightly between the other approaches. However, the linear approach suggests a slight

decrease of the log relative risk with increasing values of x although there is an increase in risk with increasing values of x for $x > 1.5$. In contrast the fitted curve obtained by LINQ and FP describe the given functional form adequately (cf. figure 3.2). The qualitative error Err_{qual} for both methods is close to 0 showing that this measure is sensible in the current situation. For LIN we obtained $Err_{qual} = 0.475$, a value that seems also to be sensible, since there is a decrease in risk for increasing values of x up to $x = 1.5$.
However, there are still problems for categorized functions as obtained by CUT. A change in risk can only be obtained for a few pairs $(x_{(j)}, x_{(j+1)})$ and, therefore, Err_{qual} is always close to 0.5 - even if the categorized function is a good approximation to the given function. However, in this situation, there is a small quantitative error. Consequently the qualitative error cannot be considered separately. This is also true for continuous functions as e.g. a fractional polynomial. Note, that the change in risk is substantially overestimated by the fitted fractional polynomial. In spite of the value $Err_{qual} = 0.030$ this overestimation may lead to wrong interpretations. In order to combine the information of a quantitative and the qualitative error, the latter one may be used as kind of penalty term for the first measure. This may be done by considering

$$\widehat{MSE}_{qual} = \widehat{MSE} + w \cdot Err_{qual}, \tag{42}$$

where w is a weight that should take the value of $\widehat{MSE}$ into account.

Table 3.1: Quantitative and qualitative errors with respect to the given function (model III) in the selected simulated data set

	$\widehat{MSE}^{1)}$	$MSE^{2)}$	$\widehat{MAE}^{1)}$	$MAE^{2)}$	Err_{qual}	$\widehat{MSE}_{qual}^{3)}$
LIN	0.0196/0.0204	0.0233	0.1184/0.1175	0.1285	0.525	0.0402
LINQ	0.0518/0.0524	0.0575	0.1901/0.1845	0.1946	0.010	0.0528
CUT	0.0575/0.0579	0.0579	0.1984/0.2030	0.1987	0.495	0.1144
FP	0.0522/0.0527	0.0588	0.1907/0.1858	0.1982	0.030	0.0553

1) first value correspond to the standardization to zero mean, second value obtained, when adding $log(\hat{\lambda}_0(t^*))$ for t^* =median survival time

2) based on the addition of $log(\hat{\lambda}_0(t^*))$ for $t^* =$ median survival time

3) based on the standardization to zero mean, $w = 2 \cdot \widehat{MSE}$

We consider $w = 2 \cdot \widehat{MSE}$ leading to values of $\widehat{MSE}_{qual}$ between $\widehat{MSE}$ (if $Err_{qual} = 0$) and $3 \cdot \widehat{MSE}$ (if $Err_{qual} = 1$), the results are also displayed in table 3.1. Due to the very small quantitative error the linear risk function is still the best choice in the simulated data set. However, considering $\widehat{MSE}_{qual}$ the advantage of the linear risk function is not as obvious as for $\widehat{MSE}$. Analogous to $\widehat{MSE}_{qual}$ one may also calculate $\widehat{MAE}_{qual}$.

Concluding remarks

Although the estimated risk functions of the selected simulated data set clearly showed the need of a qualitative error, the proposed measure Err_{qual} seems to be sensible for continuous and monotone functions only. For the null model (0) and the cutpoint model (I) the qualitative error is not sensible, because there is no difference in risk between all (0) and all but one (I), respectively, consecutive values $x_{(j)}$ and $x_{(j+1)}, \quad j = 1, \ldots, n-1$. Therefore, I focus on quantitative errors in order to assess the results of the simulation study. For better illustration I will focus on the $\widehat{MAE}$. This measure may also be preferred to the $\widehat{MSE}$, because the advantage of the wrong linear risk function is less extreme for $\widehat{MAE}$ in the selected simulated data set. Results with respect to the more common $\widehat{MSE}$ are given in appendix A. In section 3.8.1 I compare $\widehat{MAE}$, $\widehat{MSE}$ and $\widehat{MSE}_{qual}$ for the V-type model.

3.4 The null model

This section summarizes the results for the null model of no prognostic relevance of the continuous covariate X on survival time. Section 3.4.1 describes the results obtained by using the standard procedures. Furthermore, I illustrate the effect of using P-value correction and shrinkage methods in the data driven cutpoint model and for the CART based categorization. Type-I error rates will be given for different model selection strategies. Section 3.4.2 describes the differences obtained when using the two approaches to make results comparable. The bagged risk functions and its error estimates as compared to those of the corresponding standard procedures are considered in section 3.4.3. Additionally I give some results of the model selection process in the bootstrap samples.

3.4.1 Using standard procedures

To illustrate the fit of the estimated risk function to the given function $h = 0$ figures 3.3 and 3.4 show the results of the first 100 replications. In each replication the estimated risk function is standardized to a zero mean log relative risk. A comparison to the results obtained when adding the logarithm of the estimated baseline hazard is given in section 3.4.2.

The visual inspection of the estimated risk functions clearly show that the best fit is obtained with LIN. Note, that the null model is also linear with $\beta = 0$. In most replications more complex risk functions reduce to the linear risk function because of the underlying model building process (cf. section 2.1.7). A risk function containing a linear and a quadratic term is only chosen in 5%, and the restricted cubic spline is preferred in 6.3% of all replications (cf. table A 3.1 in appendix A). A fractional polynomial is selected in 10 out of 1000 replications (1%). Thus, FP seems to respond to a strong curvature in the data only, whereas other approaches may be more sensible to slight deviations from linearity. Note, that the variability in the data generation process may also produce deviations from the null model). As shown in figure 3.3 the fitted curve for the log relative risk function obtained by LINQ, RCS and FP are similar in some replications. In the present simulation we obtained with CUT 43.2% significant minimum P-values of the logrank test and, therefore, 432 cutpoint models. This result confirm prior findings and theoretical results that the minimum P-value approach produce type I error rates of about 40% when selecting the cutpoint out of all potential values between the 10% and 90% quantile of the empirical distribution of the continuous covariate (Lausen and Schumacher, 1992; Schumacher et al., 1997).

Due to its construction a CART based categorization is also obtained in 432 replications. However, figure 3.4 shows that the deviation from the null model is larger than that of CUT.

As described in section 2.1.7 corrected instead of minimum P-values should be used in the

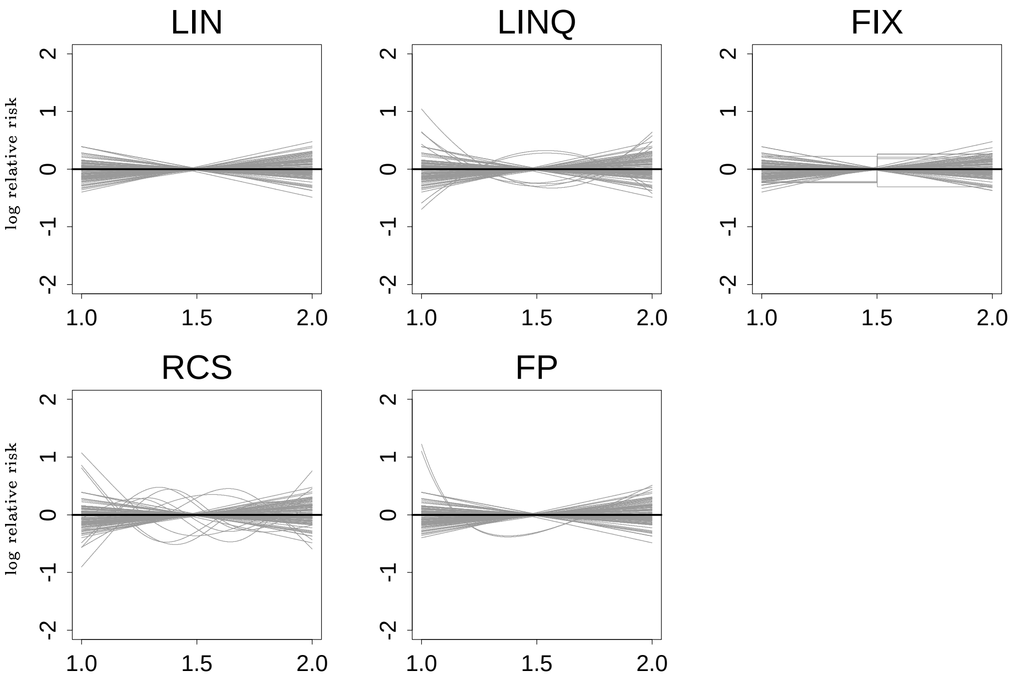


Figure 3.3: Estimated risk functions obtained by using the standard procedures, results of the first 100 replications in the simulated null model (thick line), standardization to zero mean log relative risk

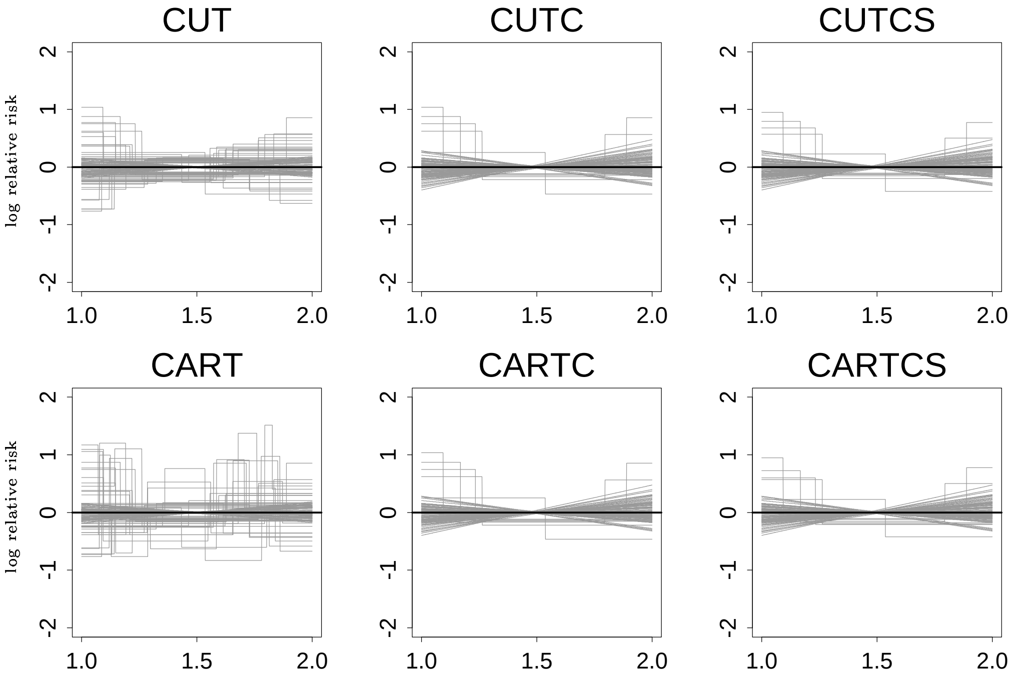


Figure 3.4: Estimated categorized risk functions based on data-driven cutpoints and effect of P-value correction and shrinkage, results of the first 100 replications in the simulated null model (thick line), standardization to zero mean log relative risk

data driven cutpoint model and for the CART based categorization. Using the P-value correction according to formula 22 there are 67 cutpoint models (6.7%) with $p_{corr} < 0.05$ left. Since the first split of the CART based categorization corresponds to that of the cutpoint model the resulting piecewise linear risk function also obtained in 67 replications. A slight correction of the overestimated log relative risks in the cutpoint model and for the CART based categorization is obtained when applying the shrinkage procedures introduced in section 2.1.7. However, as illustrated for the first 100 replications the corrected risk functions are still far away from the given null model (figure 3.4). The heuristic estimates of the shrinkage factor used in the data driven cutpoint model is positive with $\hat{c} < 1$ in all replications with a significant cutpoint model. However, at least one of the parameterwise shrinkage factors that we obtained when using the modified form of cross-validation calibration (cf. section 2.1.7) was negative in several replications and, therefore, there may be no shrinkage effect. The results are shown in table 3.2.

Table 3.2: Estimated parameterwise shrinkage factors $\hat{c}_i$ used to correct for overestimation in the CART based risk function

		number of selected cutpoints								
	0 = linear	1			2			3		
		value of $\hat{c}_i$								
		≤ 0	$\epsilon(0,1]$	> 1	≤ 0	$\epsilon(0,1]$	> 1	≤ 0	$\epsilon(0,1]$	> 1
$\hat{c}_1$	—	—	193	—	1	163	18	1	40	16
$\hat{c}_2$	—	—	—	—	55	127	—	10	32	15
$\hat{c}_3$	—	—	—	—	—	—	—	27	26	4
no. of replications	568		193			182			57	

In 27 out of 57 replications, where the model building process leads to a risk function with 3 cutpoints, the estimated shrinkage factor $\hat{c}_3$ is negative. Negative and/or values of $\hat{c}_i$ $(i = 1, 2, 3)$ that are several times larger than 1 may be caused by the small size of the patients subgroups and small values of β_i $(i = 1, 2, 3)$. Note, that we allowed a maximum of 4 subgroups for n = 100 observations in our simulation. Using $\hat{c}_i = 0$ $(i = 1, 2, 3)$ for negative estimates of the shrinkage factors we adopt the proposal of Van Houwelingen and Le Cessie (1990) to the modified cross-validation calibration procedure.

In the case of only one cutpoint, where the CART based approach is identical to the data driven cutpoint model all shrinkage factors $\hat{c}_1$ are between 0 and 1. Furthermore, there was hardly any difference between the values of the cross-validation shrinkage factor $\hat{c}_1$ and the corresponding heuristic estimate (not shown). These results agree with those of a simulation that we performed a few years ago for the data driven cutpoint model (Schumacher et al., 1997).

Quantitative errors

For each method figure 3.5 show the boxplot of the distribution of the $\widehat{MAE}$ based on all replications. Results in terms of empirical quantiles of these distributions and the corresponding distributions of the $\widehat{MSE}$ are given in tables A 3.1 and A 3.2 in appendix A. Naturally, the error estimates correspond to the results described above. Except for the categorization by data-driven cutpoints the distribution of $\widehat{MAE}/\widehat{MSE}$ is similar for all methods, since the more complex models reduced to the linear risk function in most replications. The error reduction obtained when using CUTC instead of CUT and CARTC instead of CART is resulting from the higher rate of linear risk functions (cf. figures 3.4 and 3.6). The P-value correction reduce the number of cutpoint models and CART based risk function. Due to the smaller error of the linear risk function a reduction of the $\widehat{MAE}$ is obtained in nearly all replications with $p_{min} \leq 0.05$ and $p_{corr} > 0.05$. For those replications with a significant corrected P-value ($p_{corr} \leq 0.05$) the error is reduced by applying shrinkage methods. However, especially for the CART based categorization the results should be interpreted carefully, because so far there is no theoretical justification to use parameterwise shrinkage factors. Furthermore, cutting negative shrinkage factors at zero and values larger than 1 to 1 would force the error of the shrinked risk function to be smaller than the error of the unshrinked estimated risk function. In the present simulation negative values of $\hat{c}_2$ and $\hat{c}_3$ are observed more often than values that are extremely larger than 1 and, therefore, we set negative values to zero. Comparing the categorization by data-driven cutpoints to LIN the former methods lead to risk functions with larger errors in nearly all replications with a significant cutpoint (figure 3.6, table A 3.1). This could have been expected, because the functional form of the null model is correctly specified when assuming a linear effect.

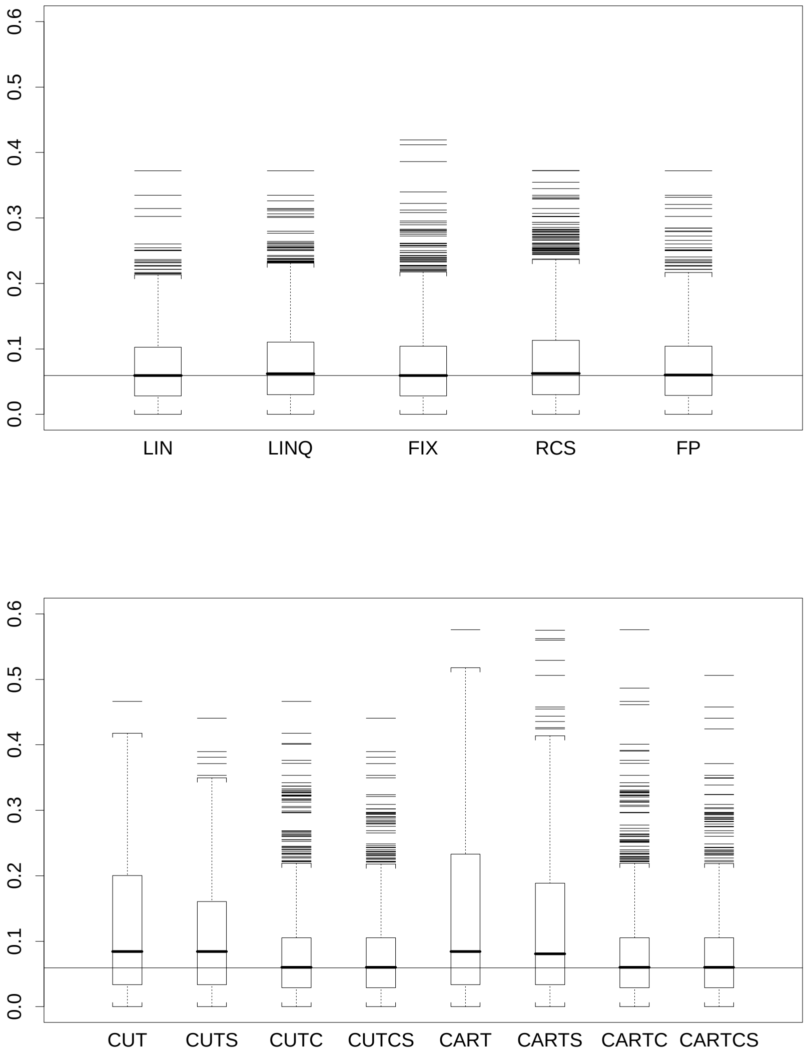


Figure 3.5: Distribution of the $\widehat{MAE}$ for all standard procedures in the simulated null model based on all 1000 replications, horizontal line denotes the median $\widehat{MAE}$ of LIN

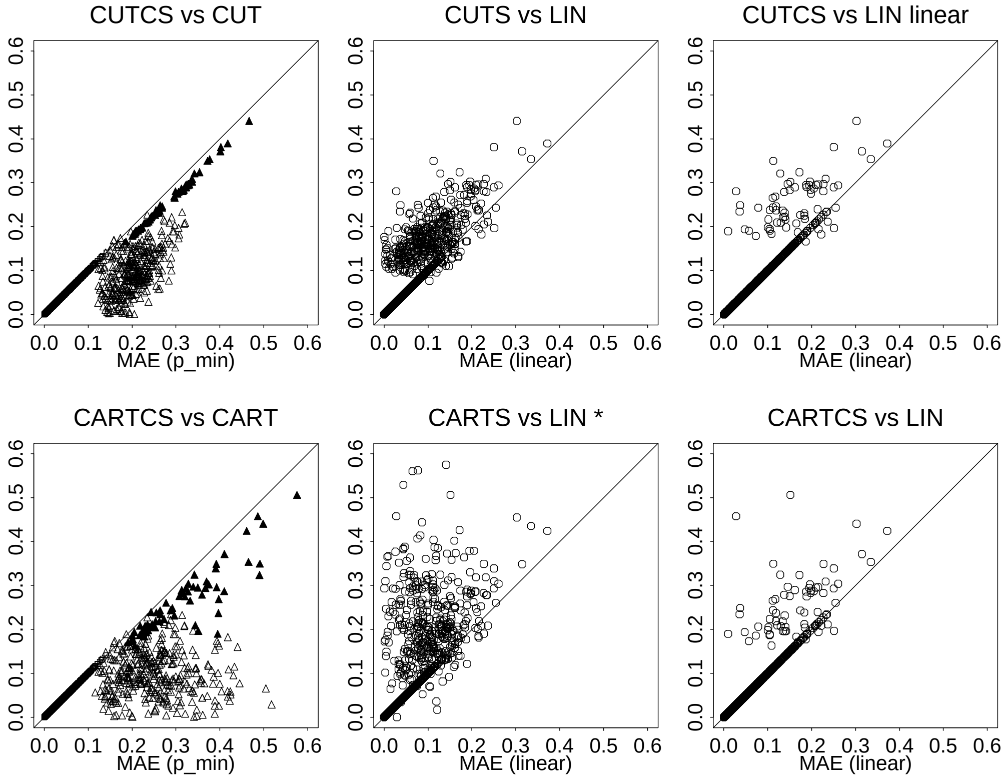


Figure 3.6: Effect of P-value correction and shrinkage on the $\widehat{MAE}$ in the data driven cutpoint model and the CART based categorization (CUTCS vs CUT, CARTCS vs CART), comparison of CUTS, CUTCS, CARTS and CARTCS to LIN, points on the diagonal denotes the $\widehat{MAE}$ obtained by LIN, triangles ($\triangle$) correspond to the $\widehat{MAE}$ of replications with a linear risk function after P-value correction, filled triangles describe replications with a categorized risk function before and after P-value correction. (* for CARTS the $\widehat{MAE}$ was larger than 0.6 in 3 replication, the axis of the plot CARTS vs LIN has been cut off at 0.6 for better illustration).

Type I error rates

Under the simulated null model the rate of selected cutpoint models can be taken as type I error rate for CUT. However, according to our model selection procedure, which will be denoted by M in this section, the linear risk function is used to describe the effect of X if there is no significant cutpoint. Therefore, the type I error rate for CUT (CART) based on M is the rate of significant cutpoint models (CART based categorizations) plus the rate of significant linear risk functions in the case of non-significant cutpoints. Using M for all other risk functions the type I error rate is also based on the higher order model (if selected) and the linear risk function if the higher order function is not selected.
To estimate type I error rates I used the likelihood ratio test testing the model with the selected estimated risk function against the null, i.e. the model with risk function $h = 0$. This was done for all methods except for the categorization by data driven cutpoints (CUT, CART, CUTC and CARTC). Here, the logrank test is used, because this test is implemented in the C programme used to select cutpoints.
In addition to strategy M two further model selection strategies were considered:

H: Using the model of the highest order as e.g. a two term FP in all replications

B: Using the best model measured by the smallest P-value of the corresponding likelihood ratio test against the null model.

M: Using the higher order model if the likelihood ratio against the linear model is significant, assuming a linear effect otherwise

Estimated type I error rates denoted as $\hat{p}_{err}$ are taken as the rate of replications with a significant test result (p<0.05). Using B instead of M it is neglected whether the higher order model is *better* than the linear risk function
Since only one parameter has to be estimated for CUT and FIX, these risk functions are of the same order as LIN. However, the categorized risk function will be considered as higher order model for strategy H. Thus, the type I error rate for CUT is simply the rate of significant cutpoint models when using H. Since FIX and LIN are not nested, FIX is tested against the null model for strategy M. Therefore, M leads to the same results as B. Additionally to $\hat{p}_{err}$ the corresponding 95% confidence intervals where calculated by

$$\left[\quad \hat{p}_{err} \pm u_{1-\alpha/2}\sqrt{\frac{\hat{p}_{err}(1-\hat{p}_{err})}{R}} \quad \right],$$

where $u_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

Table 3.3: Estimated type I error rates with 95% confidence intervals based on different model selection strategies

risk function	model selection strategy	$\hat{p}_{err}$	95% confidence interval
LIN	$H = M = B$	0.054	[0.040;0.068]
LINQ	H	0.049	[0.036;0.062]
	M	0.074	[0.058;0.090]
	B	0.075	[0.059;0.091]
FIX[1]	H	0.056	[0.042;0.070]
	$M = B$	0.076	[0.060;0.092]
CUT[1]/CART	H	0.432[2]	[0.401;0.463]
	$M = B$	0.432[2]	[0.401;0.463]
CUTC[1]/CARTC	H	0.067	[0.052;0.082]
	$M = B$	0.091	[0.073;0.109]
RCS	H	0.052	[0.038;0.066]
	M	0.084	[0.067;0.101]
	B	0.086	[0.069;0.103]
FP	H	0.014	[0.007;0.021]
	M	0.057	[0.043;0.071]
	B	0.057	[0.043;0.071]

[1] categorization in all replications are considered as higher order model

[2] values are slightly smaller when using the likelihood ratio test instead of the logrank test

Table 3.3 summarizes the estimated type I error rates. Using LIN the estimated value $\hat{p}_{err}$ corresponds well to the given test level $\alpha = 0.05$. However, the given α is also correctly estimated when using LINQ or RCS in all replications (selection strategy H). As described earlier the type I error rate is inflated for CUT and CART. P-value correction led us to a estimated type I error rate of $\hat{p}_{err} = 0.067$, which is at least close to the given level. These results disagree with that of Lausen and Schumacher (1992, 1996). In the cited papers the test based on corrected P-values tends to be conservative. Lausen and Schumacher used the standardized rank statistic with logrank scores whereas my calculation is based on the *common* logrank test. However, both test should be identical without censored observations. The differences with respect to the estimated type I error rate cannot be explained so far. Considering a fractional polynomial of degree 2 in all replications (FP based on strategy H) the true error rate is underestimated. This finding corresponds well to remarks of Royston and Altman (1994) who stated that the fractional polynomial approach tends to be conservative.

Generally, the model selection strategy M results in higher estimated type I error rates than strategy H. For FP based on strategy M $\hat{p}_{err}$ corresponds well to the given level whereas we observed a slight overestimation of the true level for LIN and RCS. Using the P-values correction, i.e. CUTC and CARTC strategy M produces larger type I errors than strategy H whereas there is no significant linear risk function when using the minimum P-value (CUT and CART).

It is not astonishing that there is hardly any difference between strategy M and B, because strategy M also tries to find the "best" model. Generally, the rate of significant models obtained by strategy B is at least as large as that based on strategy M. For instance, testing RCS and LIN against the null $(h(x) = 0)$ we obtained

method	LIN			
RCS	LRT[1]	$p > 0.05$	$p \leq 0.05$	Σ
	$p > 0.05$	914	34	948
	$p \leq 0.05$	32	20	52
	Σ	946	54	1000

[1] likelihood ratio test against the null model $(h(x) = 0)$

Using strategy B a significant effect is obtained in 34+32+20 replications leading to $\hat{p}_{err} = 86/1000 = 0.086$. In the first step of strategy M the higher order restricted cubic spline is tested against the linear risk function:

- In 28 out of the 914 non significant replications the restricted cubic spline was better than the linear risk function.
- All 34 replications, which were significant against the null for LIN and not significant for RCS, cannot be improved by a restricted cubic spline.
- In 30 out of 32 replications, which were significant against the null for RCS and not significant for LIN, the restricted cubic spline was better than the linear risk function.
- In 5 out of the 20 replications, where both methods showed a significant effect, the restricted cubic spline is better than the linear risk function

All in all, for strategy M a significant effect was found in $34 + 30 + 20 = 84$ replications, 2 less than for strategy B.

In appendix A the rate of nonlinear risk functions is given for all methods, cf. e.g. table A 3.1. Considering the results described above I obtained $28 + 0 + 30 + 5 = 63$ nonlinear risk functions for RCS.

Although the model selection strategy M tends to overestimate the type I error rate we restrict to this model building process in order to estimate the given functional relationship. Thus, the bagging estimator is also based on strategy M. Doing so, I produce variability between bootstrap samples whereas one would expect only slight differences with respect to the functional shape of the estimated risk function when using strategy H. Remember, that the bagged risk function was identical to the function obtained in the GBSG-2 study for FIX2 and RCS due to the lack of variability between bootstrap samples.

3.4.2 Comparing the strategies to make results comparable with respect to the error estimation

For the simulated data set that has been used in section 3.2 for illustration the two strategies proposed to make results comparable lead to similar results: The values of $\widehat{MSE}$ and $\widehat{MAE}$ obtained when adding $\log \hat{\lambda}_0(t^*)$ using t^* with $\hat{S}(t^*) = 0.5$ differ only slightly from those based on the standardization to a zero mean log relative risk. Furthermore, the dependence of the log baseline hazard on time (i.e. the question which value t^* should be chosen) seemed to be not very important, because the difference $D_{\hat{\lambda}_0}$ between two risk functions was constant over time.

In this section I consider several values of the fixed time point t^* and investigate the effect on the estimation of the MSE and the MAE. The resulting estimates for all replications in the null model are compared to the corresponding error estimates obtained after standardization to a zero mean log relative risk. The $\widehat{MSE}$ and $\widehat{MAE}$ of the latter approach are taken as a baseline (100%). The ratios $\widehat{MSE}(\text{adding } \lambda_0)/\widehat{MSE}(\text{zero mean})$ and $\widehat{MAE}(\text{adding } \lambda_0)/\widehat{MAE}(\text{zero mean})$, respectively, are then used (in terms of %) to compare results. It should be mentioned that the former ratio is always larger than the baseline, because $\widehat{MSE}$ is minimal if $\sum_{i=1}^{n} h(x_i) = \sum_{i=1}^{n} g(x_i)$. This condition must not hold for the $\widehat{MAE}$ and, therefore, we observed deviation in both directions.

For LIN, LINQ and CUT the distributions of the error ratios using values t^* with $\hat{S}(t^*) = 0.9, 0.75, 0.5, 0.25$ and 0.1 are displayed in figure 3.7. For better illustration the whiskers of the boxplots end at the maximal and minimal value, respectively. The results for LIN indicate hardly any difference between $\hat{S}(t^*) = 0.9, 0.75, 0.5$ and as compared to the standardization to zero mean log relative risk. However, especially for t^* with $\hat{S}(t^*) = 0.1$ there is a strong deviation from the baseline and the variability of the distribution of the ratios is very large. These results may be caused by the fact that the risk set used to estimate $\lambda_0(t^*)$ is rather small for $\hat{S}(t^*) = 0.1$. Therefore, we have to deal with a large variability. Furthermore, the influence of single observations can be extremely strong.

According to the our model selection procedure LINQ and CUT, respectively, reduces to LIN, if there is no significant quadratic effect and cutpoint, respectively. Therefore, I illustrate the distribution of error ratios for replications with a nonlinear risk function only. Similar to LIN the difference between $\hat{S}(t^*) = 0.9, 0.75, 0.5$ is rather small, whereas the variability increases substantially for $\hat{S}(t^*) = 0.25$ and $\hat{S}(t^*) = 0.1$. Considering $\hat{S}(t^*) \geq 0.5$ the deviation of the corresponding error estimates is less than 10% as compared to the standardization to a zero mean risk for LINQ, whereas we observed a larger difference for CUT. The results for the other risk functions are listed in the appendix (table A 3.4). For FIX, RCS and FP there is again hardly any difference between the chosen values t^* and as compared to the baseline. The results for CART are similar to those obtained for CUT: The variability of the error estimates as well as the deviation from the baseline is larger.

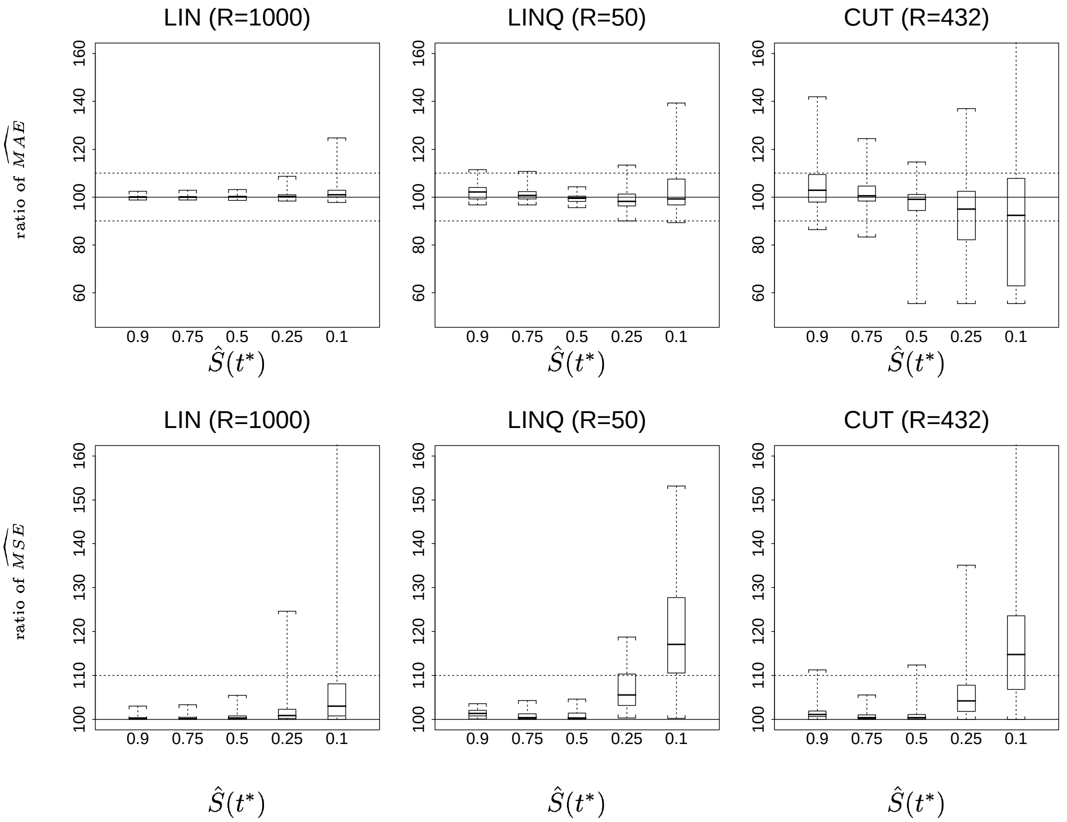


Figure 3.7: Comparing error estimates based on the addition of $log\hat{\lambda}_0(t^*)$ using different values of t^* to those obtained when standardizing to a zero mean log relative risk: Distribution of the ratio $\widehat{MAE}$(adding λ_0)/$\widehat{MAE}$(zero mean) in % (top) for LIN, LINQ for replications with quadratic term and CUT for replications with significant cutpoint, corresponding distributions for $\widehat{MSE}$ at the bottom

Concluding remarks

Adding the logarithm of the underlying estimated baseline hazard to each risk function in order to make different functions comparable t^* should be chosen such that the risk set is large enough to calculate $\log\hat{\lambda}_0(t^*)$. Even if this is the case results may depend on the choice of this fix time point for specific risk functions. Therefore, and due to the fact that the simulation is restricted to the situation of no censoring, I will mainly use the standardization to a zero mean log relative risk to make results comparable. Generally, it would be worthwhile to investigate the dependency of the standardization on the choice of t^*. In this context it should be taken into account that $\log \hat{\lambda}_o(t^*)$ also depends on the chosen risk function. Thus, if the true risk function is completely misspecified (as done when selecting several data-driven cutpoints in the null model) $\log \hat{\lambda}_0(t^*)$ is also misspecified.

3.4.3 Application of the bootstrap

For each replication 100 bootstrap samples were generated in order to estimate the bagged risk functions. As mentioned above the model selection strategy M is used in each bootstrap sample to select between the higher order and the linear risk function. Investigating the capability of the bagged risk function to estimate the given risk function (i.e. the null model) more adequately than the underlying standard procedure I compare $\widehat{MAE}$ and $\widehat{MSE}$, respectively, of $\hat{h}_{bagg}$ to the corresponding error estimates obtained when using $\hat{h}$. Restricting again to the first 100 replications the estimated values of $\widehat{MAE}$ are illustrated for LINQ, FIX, RCS, CUTS, CART and FP in figure 3.8. Dots are used for replications with nonlinear risk functions in the original data. Points below the diagonal line indicate an error reduction of the bagged risk function as compared to the corresponding standard procedure.

The results for all replications and all approaches are summarized in table A 3.3 in the appendix: Error ratios (bagged versus original) are given for $\widehat{MAE}$ and $\widehat{MSE}$. These ratios were also calculated separately for replications with a linear and a nonlinear risk function in the original data set. Additionally to the distribution of the error ratios table A 3.3 show the number of replications with a smaller estimated $\widehat{MAE}$ ($\widehat{MSE}$) for the bagged risk function. As observed in the GBSG-2 study there is hardly any difference between the bagged risk function and the estimated function in the original data for LIN. This is not astonishing since there is no variability between bootstrap samples. For all other approaches – except for CART – bagging lead to an error reduction in replications with a nonlinear risk function in the original data set. This is especially true for CUT (CUTS, CUTC, CUTCS), where the error estimates are smaller in (nearly) all cases. For these approaches we also observed the highest error reduction: the median $\widehat{MAE}$ is about 3 times smaller as compared to the corresponding value obtained in the original data. The ratio for $\widehat{MSE}$ range between 6% for CUTCS and 18% for CUT. Note, that the ratio would be equal to 100% if errors are identical. For the other approaches the rate of replications with a better bagged risk function range between 75 and 98%.

For the CART based categorization bagging cannot be recommended at all – even in replications with huge errors in the original data set the bagged risk function may be worse. Note, that the axes for CART used in figure 3.8 has been cut at 0.4. In 4 out of the first 100 replications the $\widehat{MAE}$ was larger than 0.4, 3 of these replications showed a smaller

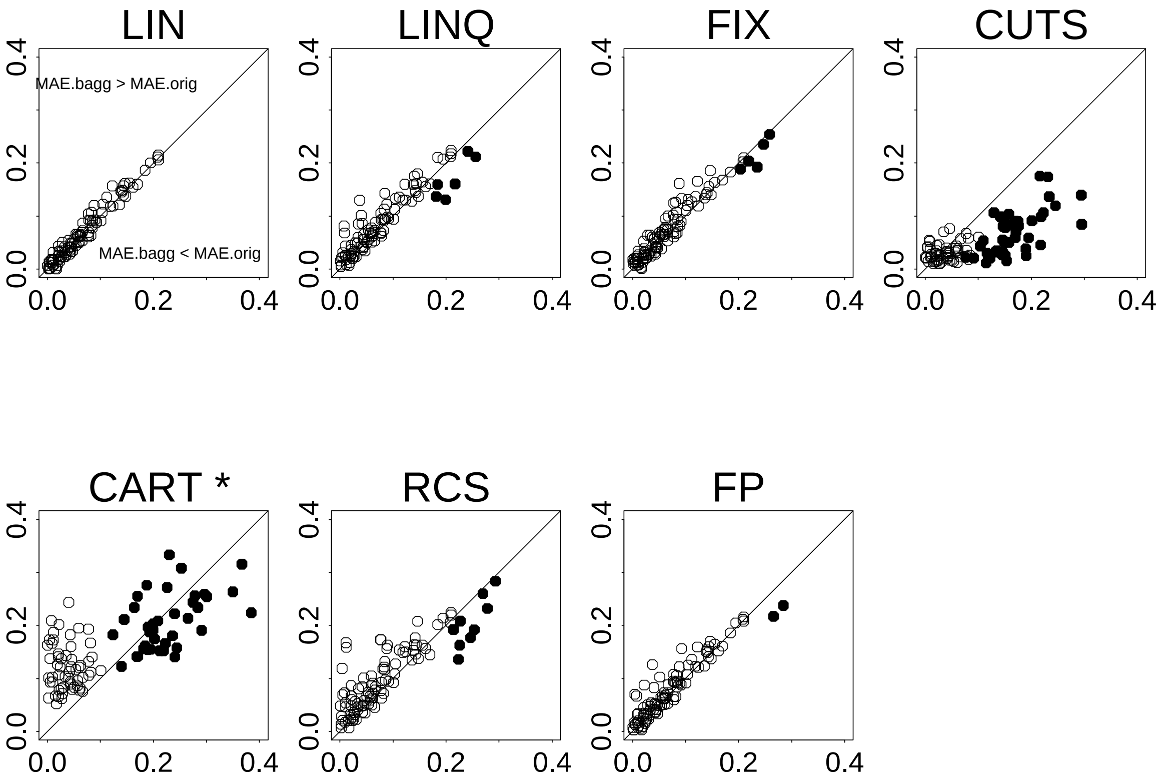


Figure 3.8: $\widehat{MAE}$ of the bagged risk function as compared to the the $\widehat{MAE}$ obtained in the original data set for the first 100 replications of the simulated null model, dots denote replications with a nonlinear risk function in the original data (* $\widehat{MAE}$ larger than 0.4 in 4 replications, axes cut off at 0.4)

$\widehat{MAE}$ of the bagged risk function. As described earlier the application of parameterwise shrinkage factors seemed to be not sensible to correct for over-optimism of the categorized risk function, because the estimates $\hat{c}_i$ were negative or larger than 1 in many replications. This is probably resulting from the fact that the CART based categorization may produce very small subgroups. In many bootstrap samples parameterwise shrinkage factors could not be calculated at all due to missing convergence of the log-likelihood. Therefore, in the sequel I consider CART only.

As figure 3.8 shows bagging increases the error, if the risk function is linear in the underlying replication. Considering RCS for example these findings could be explained easily: as shown in section 3.4.1 the linear risk function turned out to be the best method to estimate the null model. Given the data of one replication in which RCS reduced to the linear risk function it can be expected that the model selection strategy M lead to

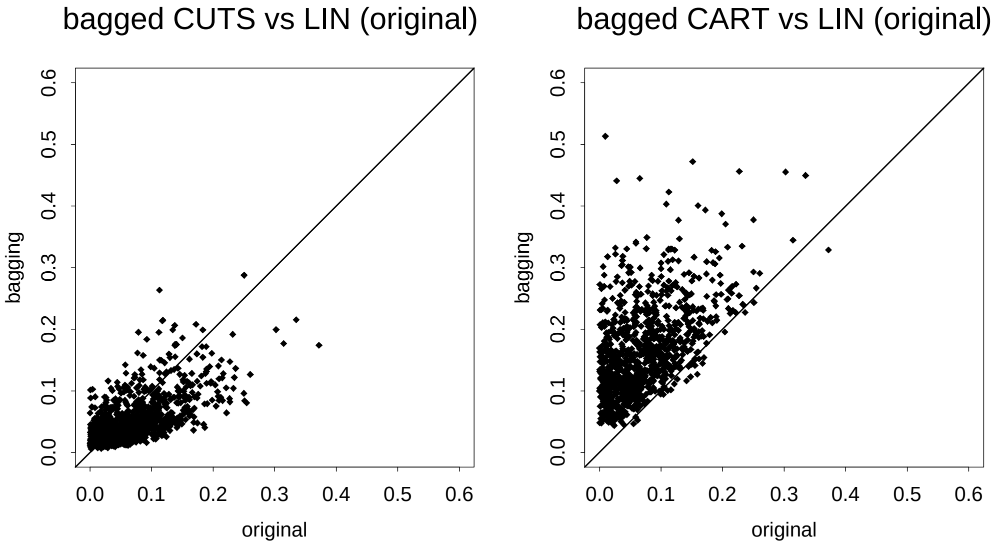


Figure 3.9: Comparing $\widehat{MAE}$ of the bagged risk function for CUTS and CART to the $\widehat{MAE}$ obtained for LIN in the original data set (all replications)

a linear risk function in most of the bootstrap samples, too. However, there may be a few bootstrap samples with a restricted cubic spline selected. The resulting functions of these bootstrap samples may be far away from the linear risk function and the null model. Thus, the influence of a few bootstrap samples on the estimation of the bagged risk function could be very strong and, therefore, may lead to larger errors. The same argument may be used for the other approaches. One exception to this rule is CUT (CUTS, CUTC, CUTCS), where the bagged risk functions produces smaller errors in 51 to 84 per cent of the corresponding replications. Comparing the distribution of the $\widehat{MAE}$ and $\widehat{MSE}$ for the bagged risk functions to those obtained in the original data set it is obvious that bagging reduces errors mainly in replications with larger errors. For CUTS this can also be seen in figure 3.8: For all replications (liner or cutpoint model in the original data) with an $\widehat{MAE}$ larger than 0.05 points are below the diagonal line. In contrast, bagging is worse for CART in nearly all replications with a linear risk function. This may be explained by the fact that CART led to the most extreme results (cf. section 3.4.1) and, therefore, to the largest errors.

As described earlier the error obtained when categorizing X by one or several data driven cutpoint(s) in the original data is larger than LIN in nearly all replications. This is not astonishing, because the assumption of a linear risk function is the best we can do in the null model. In contrast the risk function is misspecified if X is categorized. Comparing the $\widehat{MAE}$ and $\widehat{MSE}$ of the bagged risk function based on CUT (CUTS, CUTC, CUTCS) to the corresponding errors of LIN in the original data we obtained smaller errors for the former in more than 70 per cent of all replications (cf. table A 3.3). Therefore, bagging can produce better estimates of the true risk function even if this function is misspecified in a part of the bootstrap samples. However, as described above bagging does not work at all with CART. Figure 3.9 compares the the values of the $\widehat{MAE}$ for CUTS and CART, respectively, to the $\widehat{MAE}$ of LIN in the corresponding original data set.

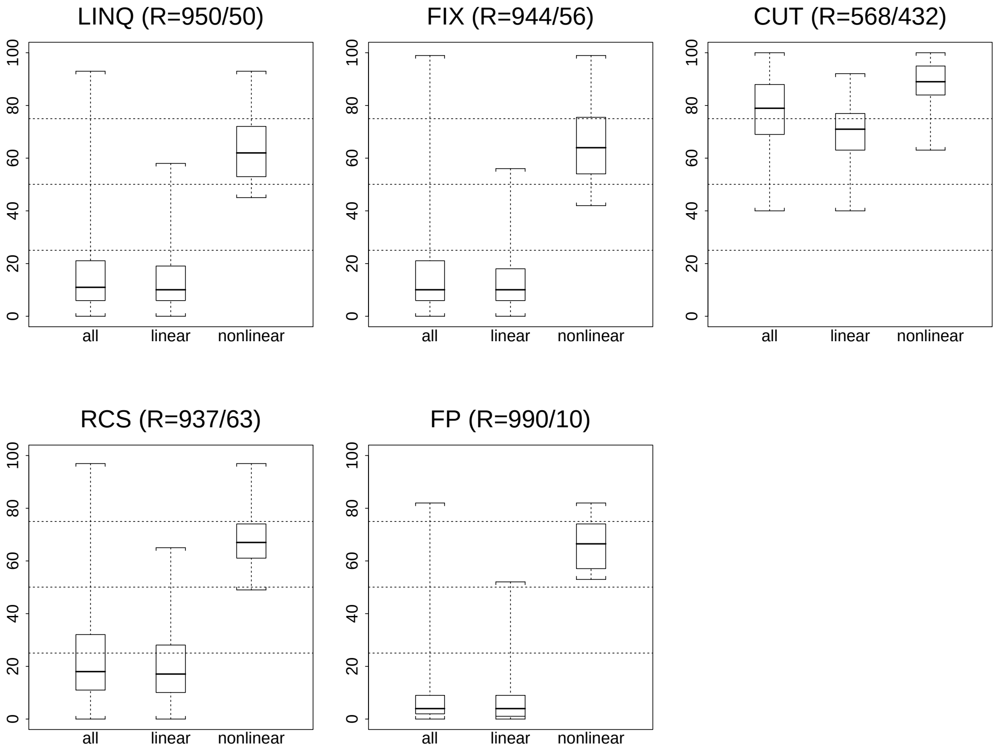


Figure 3.10: Distribution of the number of nonlinear risk functions selected in 100 bootstrap samples for all replications and separated by linear/nonlinear risk function in the original data set, R denotes the corresponding number of replications with a linear/nonlinear risk function

Selection frequencies in the bootstrap samples

As in the GBSG-2 study the stability of the estimated risk functions can be investigated by analyzing the results of the model selection procedure M in the bootstrap samples. For each replication I counted the number of bootstrap samples with a higher order model and the number of bootstrap samples, where the risk function reduced to the linear effect. Categorizing X by data-driven cutpoints it is not distinguished between one, two or three cutpoints. Therefore, CUT and CART lead to the same results. Similar, one and two term

fractional polynomials are considered together as higher order model when using FP. The empirical distributions of the count of higher order models are displayed in figure 3.10. Besides considering all replication, these distributions are also shown for replications with a linear and nonlinear risk function, respectively, in the original data. The corresponding number of replications are displayed in the heading of the plots (cf. also table A 3.1).
If the higher order risk function was selected in the original data the number of bootstrap samples with a nonlinear risk function is large, too. This is resulting from the fact that bootstrap samples are not independent of the underlying original data set. A strong quadratic effect found by LINQ in the original data set, for example, will carry through to the bootstrap samples. Therefore, the number of bootstrap samples showing also a significant quadratic effect can expected to be high. For instance, for the R=50 replications with a quadratic term in the original data, a quadratic term is usually chosen in more than 50% of the corresponding bootstrap samples, too. The median rate of bootstrap samples with a quadratic term is about 63%. In contrast, the rate of bootstrap samples with a linear risk is high if the corresponding higher order risk function reduced to the linear one in the original data set. As shown in table A 3.1 in 100 replications FP reduced to the linear risk function in all 100 bootstrap samples. In contrast, the rate of bootstrap samples with a linear risk function is smaller than 50 per cent for the 10 replications with a fractional polynomial selected in the original data set (cf. figure 3.10). Due to its construction CUT/CART shows the highest rate of bootstrap samples with a nonlinear risk function. Furthermore, this rate is also relatively high in replications with no significant cutpoint (i.e. a linear risk function) in the original data.

3.5 The cutpoint model

In this section I present the results obtained when assuming a cutpoint model with given cutpoint $\mu = 1.5$ and log relative risk $\beta = 0.5$. Using the given cutpoint to categorize X and estimating the resulting log relative risk β is the best we can do in the current situation. Therefore, this approach that is equal to FIX without model selection can be considered, besides LIN, also as some kind of reference for the assessment of the other risk functions. Comparability of all risk functions is obtained by standardization to a zero mean log relative risk. Some results obtaned when adding the estimated logarithmic baseline hazard are summarized in appendix A.

3.5.1 Using standard procedures

Risk function estimates based on the standard procedures of the first 100 replications are displayed in figures 3.11 and 3.12. The error estimates based on all replications and results with respect to model selection are given in appendix A. As shown in figure 3.11 the linear risk function seems to be a good approximation to the given cutpoint model. The given increase in risk was recognized in nearly all simulated data sets, the estimated linear effect is positive in 982 (98.2%) replications. Consequently, the values of $\widehat{MAE}$ and $\widehat{MSE}$ are rather small for LIN (figure 3.13 and tables A 3.5, A 3.6). The good fit of the linear risk function may be the reason that more complex continuous risk functions reduced to LIN in most replications: The rate of replications with a linear risk function is 94.4% for LINQ, 88.8% for RCS and 98.9% for FP (cf. table A 3.5). Error estimates for replications with a nonlinear effect are larger than those based on a linear effect. For instance, the median $\widehat{MAE}$ of RCS is 0.128, whereas this value is increased to 0.222 when considering the 112 replications with a selected restricted cubic spline only. The fact that nonlinear risk functions tend to larger errors is also obvious from figure 3.11. Categorization based on the given cutpoint (FIX) leads to a significant effect in 67.2% of all replications, otherwise we used the linear risk function. As mentioned above we also consider FIX without model selection, i.e. the categorization by μ in all replications (FIXALL). Using FIXALL the resulting estimated log relative risks center around the given value $\beta = 0.5$, whereas the estimates $\hat{\beta}$ tends to be larger than 0.5 for the cutpoint models selected with FIX (cf. figure 3.11). Smaller values of $\hat{\beta}$ are often not significant and, therefore, FIX reduced to the linear risk function. Using the minimum P-value approach we obtained a significant cutpoint in 908 of 1000 replications. There are several replications, in which the selected cutpoint is close to $\mu = 0.5$ and the estimated log relative risk obtained by CUT are not far away from the given value $\beta = 0.5$. However, the rate of estimated risk functions which ar far away from the given cutpoint model is also high. As for the null model the deviation from the given model is more extreme for CART. Using the corrected instead of the minimum P-value the rate of significant

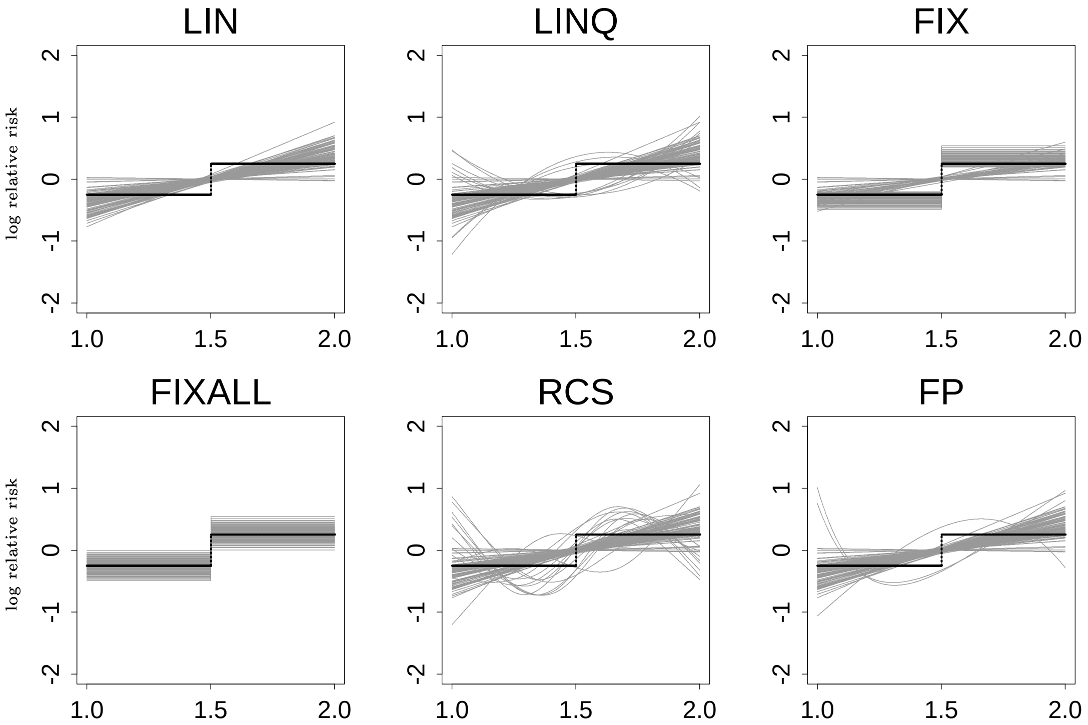


Figure 3.11: Estimated risk functions obtained by using the standard procedures, results of the first 100 replications in the simulated cutpoint model (thick line), standardization to zero mean log relative risk

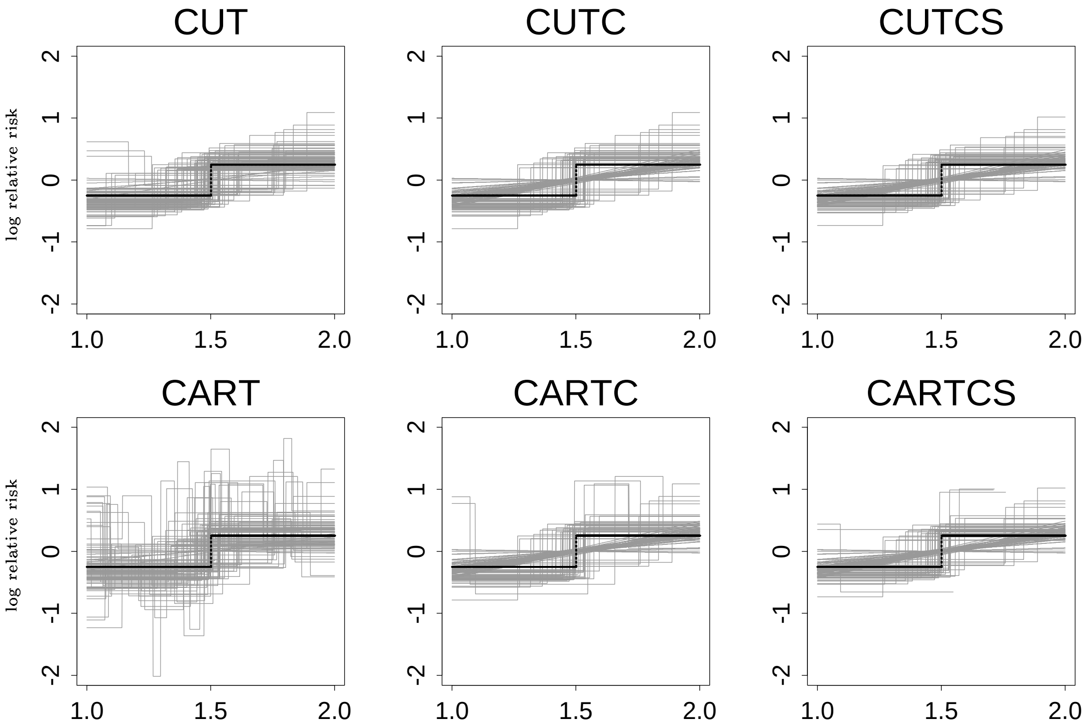


Figure 3.12: Estimated categorized risk functions based on data-driven cutpoints and effect of P-value correction and shrinkage, results of the first 100 replications in the simulated cutpoint model (thick line), standardization to zero mean log relative risk

cutpoint models and CART based categorizations, respectively, reduced to 44.9%. Due to the higher rate as well as due to the good fit of linear risk functions CUTC/CARTC have smaller errors than CUT/CART. A further reduction of the error estimates can be obtained by shrinkage. However, it can easily be seen that the data driven categorization leads to larger errors as compared to the methods based on continuous risk functions. As in the null model the highest error is observed for CART. A slight reduction of the $\widehat{MAE}$ and $\widehat{MSE}$ can be obtained by applying shrinkage methods. However, many of the categorized risk functions are still far away from the given model (cf. figure 3.12). Furthermore, for several replications I obtained parameterwise shrinkage factors larger than 1 and, therefore, an increase in the error of the shrinked risk functions (CARTS, CARTCS) as compared to CART and CARTC, respectively. As in the null model, this problem is relevant only if the number of selected cutpoints is at least 2. In some of these replications I obtained also values smaller than 0. All in all the results with respect to the estimated parameterwise shrinkage factors are similar to those described in 3.4.1 for the null model. Selecting only 1 cutpoint all values of the cross-validation based shrinkage factor are close to the heuristic estimate of the shrinkage factor and both are between 0 and 1.

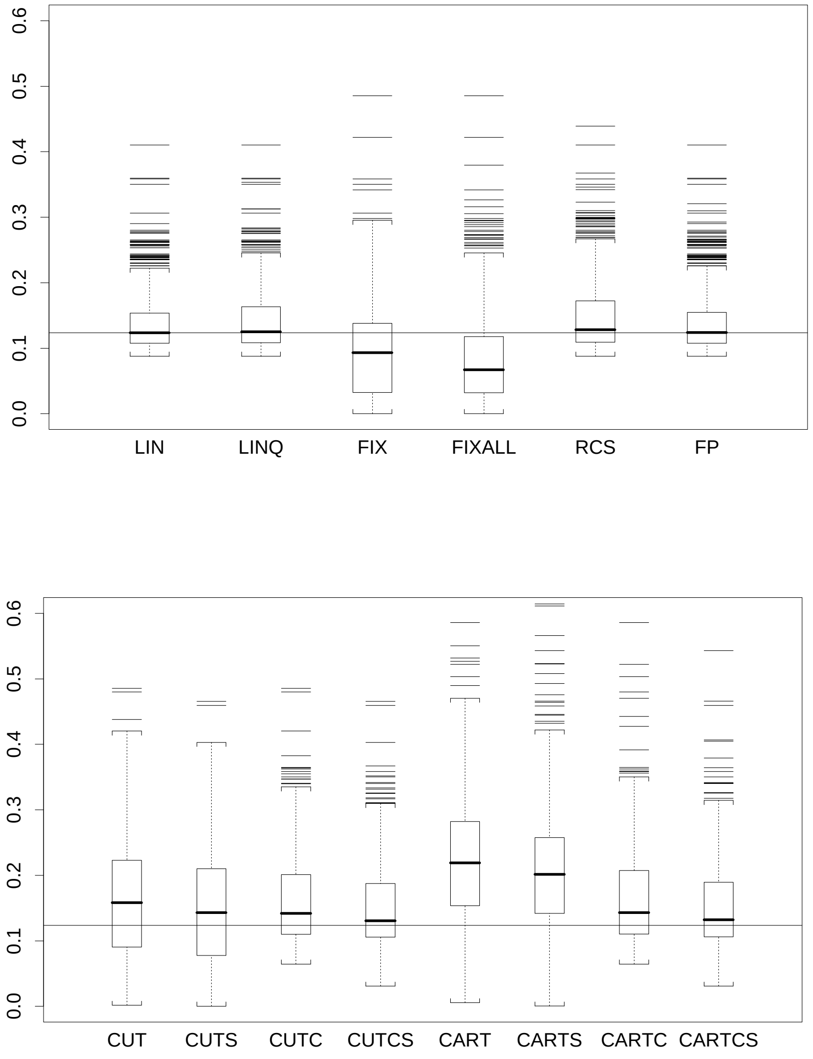


Figure 3.13: Distribution of the $\widehat{MAE}$ for all standard procedures in the simulated cutpoint model based on all 1000 replications, horizontal line denotes the median $\widehat{MAE}$ of LIN

3.5.2 Application of the bootstrap

To investigate the effect of bagging error estimates obtained for the bagged risk functions are compared to those of the underlying standard procedures. Figure 3.14 compares the $\widehat{MAE}$ of the bagged risk function to those obtained for h in the original data set for the first 100 replications. More details on the distribution of the $\widehat{MAE}$ and $\widehat{MSE}$, respectively, as well as error ratios of bagged versus original risk functions are given in appendix A. In contrast to the null model I considered bootstrap results of the of the first 100 replications only. Except for a higher accuracy it cannot be accepted that results would change when using bootstrap samples of all 1000 replications.
Since there is no model selection strategy involved with LIN and, therefore, there is no variability between bootstrap samples we obtained nearly the same error estimates for the bagged risk function as compared to the linear risk function estimated in the original data set. Slight differences can be observed for LINQ, RCS and FP. For these approaches bagging leads to an error reduction in replications with a nonlinear risk function in the original data set. From figure 3.11 it can easily be seen that the deviation from the given cutpoint model is usually higher for nonlinear risk functions leading to larger error estimates. This is especially true if the estimated risk function shows a strong curvature. Although nonlinearity in these replications often carry through to the corresponding bootstrap samples, the resulting bagged risk function is also based on bootstrap samples, in which the more complex risk function reduced to LIN. Thus, the curvature of $\hat{h}_{bagg}$ can expected to be less extreme than that of the underlying function $\hat{h}$. Consequently, the deviation from the given cutpoint model tends to be smaller and, therefore, the estimated error is smaller, too. If LINQ, RCS and FP reduced to the linear risk function in the original data set, bagging can either lead to a reduction or to an increase of the error. Using FIX bagging leads to smaller error estimates mainly for those replications with a linear risk function in the original data set. The largest differences between the error estimates of $\hat{h}_{bagg}$ and $\hat{h}$ is obtained for CUT (CUTS) and CART. In contrast to the null model, where bagging leads to a reduction of the error for replications with a cutpoint model in the original data, we observed both, replications with a large increase as well as replications with a large decrease of the $\widehat{MAE}$. As mentioned above CUT (CUTS) the estimated data driven cutpoint model corresponds well to the given cutpoint model in several replications (cf. also section 3.8.2). For these replications the fit of the bagged risk function cannot be expected to be better than that obtained in the original data. In contrast there are also many replications where the estimated data driven cutpoint model differ substantially from the given model. Here, bagging has a good chance to improve the fit. Similar results are obtained for CART although the values of $\widehat{MAE}$ are generally larger.

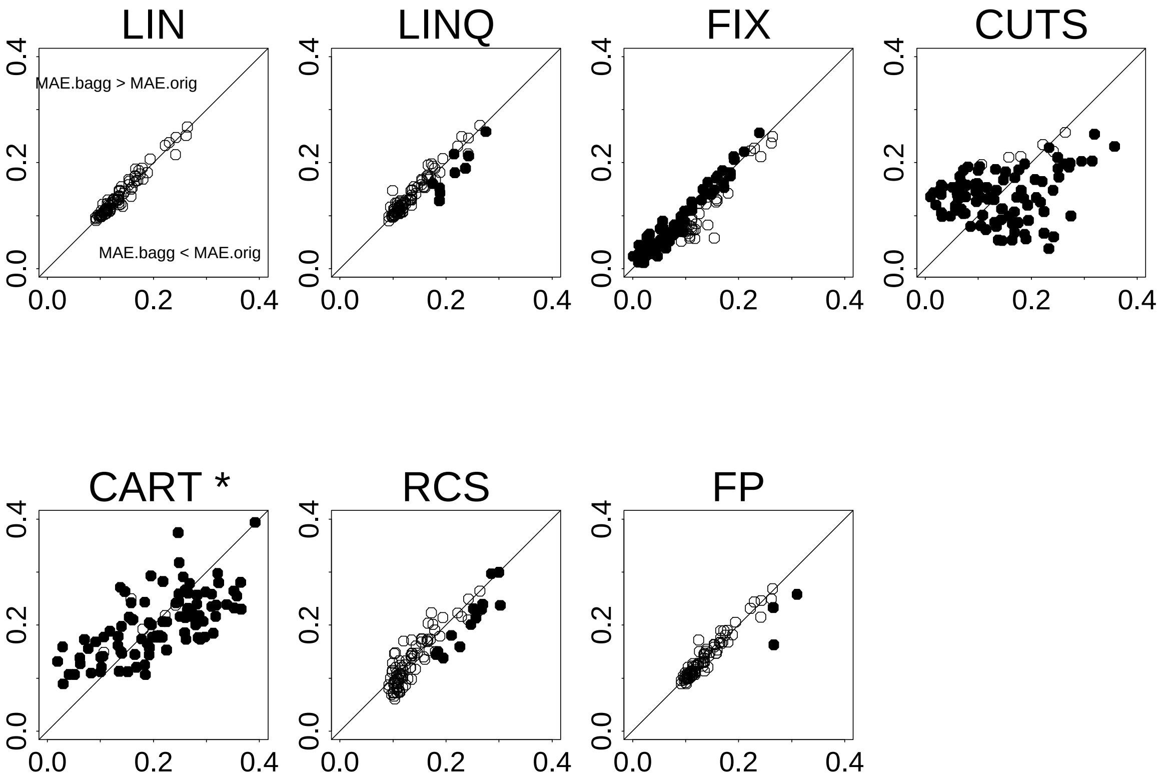


Figure 3.14: $\widehat{MAE}$ of the bagged risk function as compared to the the $\widehat{MAE}$ obtained in the original data set for the first 100 replications of the simulated cutpoint model, dots denote replications with a nonlinear risk function in the original data ($\star$ $\widehat{MAE}$ larger than 0.4 in 2 replications, axes cut off at 0.4)

Selection frequencies in the bootstrap samples

Reporting on the results of the model selection strategy M in the bootstrap samples I restrict myself again to CUT when considering data driven cutpoints and I neglect the order of the fractional polynomial if this nonlinear risk function is chosen by FP. Analogous to figure 3.10 describing the results obtained in the null model figure 3.15 shows the distributions of the number of nonlinear risk function in the bootstrap samples. Generally, the results obtained in the original data sets also carry through to bootstrap samples. For LINQ and FP the results are similar to that obtained in the null model: the number of nonlinear risk function is large for replications with a nonlinear risk function in the underlying original data and it is small if the risk function reduced to the linear effect in the original data set. As compared to the null model the rate of significant data driven cutpoint models is higher and, therefore, the number of bootstrap samples with at least one significant data driven cutpoint tends to be larger, too. In 27 out of the the first 100 replications CUT selected a cutpoint model in all 100 bootstrap samples. For FIX the number of nonlinear risk functions in the bootstrap samples is also higher as in the null model: the median number is about 30 (null model: 10) for the $R = 32$ replications with a linear risk functions in the original data and about 85 (null model: 65) for the $R = 68$ replications with a categorized risk function in the original data. For RCS the selection frequencies in the bootstrap samples of the cutpoint model are similar to that obtained for the bootstrap samples in the null model.

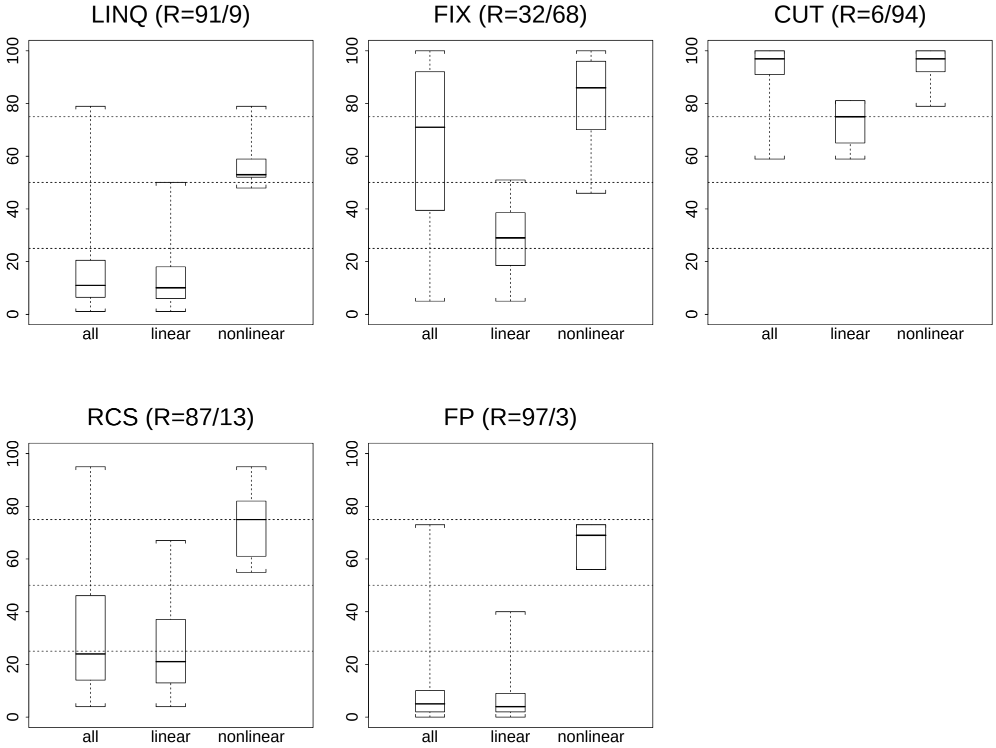


Figure 3.15: Distribution of the number of nonlinear risk functions selected in the bootstrap samples for the first 100 replications in the simulated cutpoint model, results separated by linear/nonlinear risk function in the original data, R denotes the corresponding number of replications with a linear/nonlinear risk function.

3.6 The linear effect model

This section summarizes the results obtained when assuming the *classical* standard assumption of a linear risk function with $\beta = 0.5$. Thus, the given functional form is correctly specified by LIN. All risk functions are made comparable by standardization to a zero mean log relative risk. A short comparison to the results based on the addition of the logarithmic baseline hazard is given in appendix A.

3.6.1 Using standard procedures

The estimated risk functions obtained in the first 100 replications are shown in figures 3.16 and 3.17. The distribution of the $\widehat{MAE}$ for all approaches is displayed in figure 3.18, empirical percentiles with respect to both error estimates, the $\widehat{MAE}$ and the $\widehat{MSE}$, are given in tables A 3.9 and A 3.10 in the appendix. Obviously, the best fit is obtained when using LIN, which is equal to the given risk function in the current situation. LINQ and RCS, respectively, reduced to the linear risk function in 93.9% and 94.1%, respectbely, of all replications. There is hardly any difference between FP and LIN, because FP reduced to the linear risk function in 99.1% (cf. table A 3.9). Using FIX a significant categorized risk function is obtained in 215 out of 1000 replications. As shown in figure 3.16 the categorized risk function is a good approximation of the given linear risk function, the corresponding values of $\hat{\beta}$ range from 0.41 to 0.91. This is also true for FIXALL, where $\hat{\beta}$ is larger than 0 for more than 80% of all replications. All in all the distribution of the $\widehat{MAE}$ is nearly equal for LIN, LINQ, FIX, RCS and FP, which is caused by the high rate of linear risk functions. Larger errors and a worse fit is obtained for all approaches based on data driven cutpoints. At least one significant data driven cutpoint is obtained in 68.9% of all replications. Using corrected P-values this rate reduced to 21.9%. The resulting categorized risk functions seems to be a good approximation to the given linear risk function especially for CUTCS: In 218 of the 219 replications with a categorized risk function the estimated log relative risk is larger than 0 describing the given increase in risk with increasing values of X correctly. The corresponding number for CUT is 663 out of 689. Only one significant cutpoint was found in 97 out of 219 (367 out of 689) replications. Thus, CARTCS is equal to CUTCS in these replications. Although the increase in risk is often specified correctly by the approaches based on data driven cutpoints, the $\widehat{MAE}$ obtained for CUTCS and CARTCS is still larger than that obtained by all approaches based on continuous risk functions.

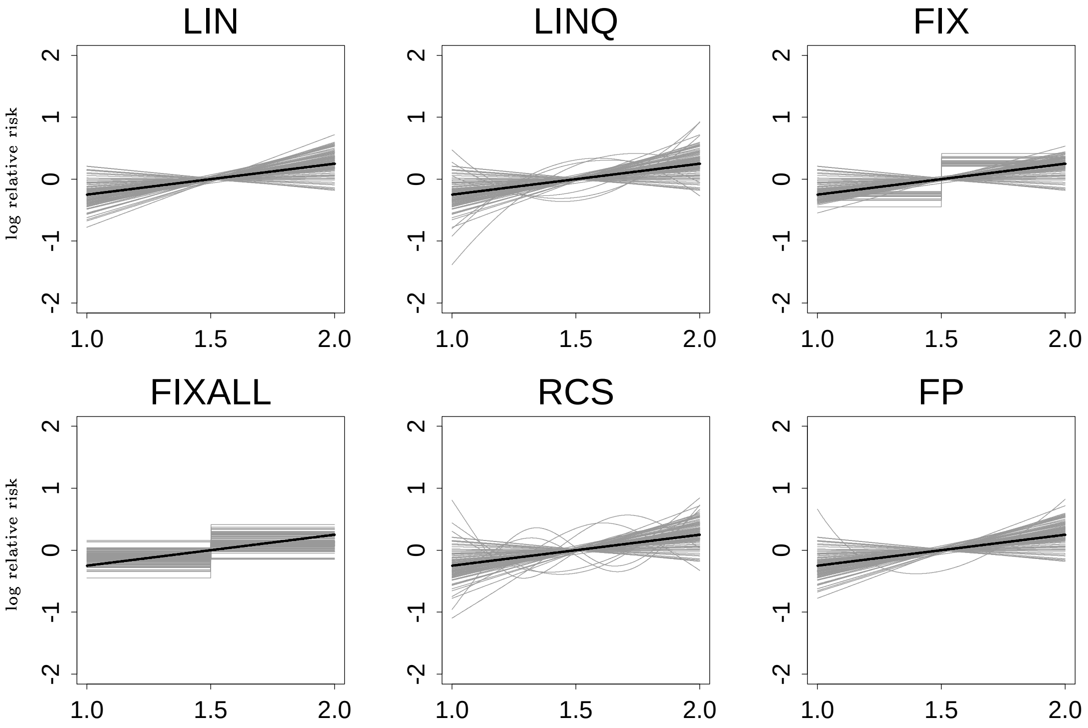


Figure 3.16: Estimated risk functions obtained by using the standard procedures, results of the first 100 replications in the simulated linear model (thick line), standardization to zero mean log relative risk

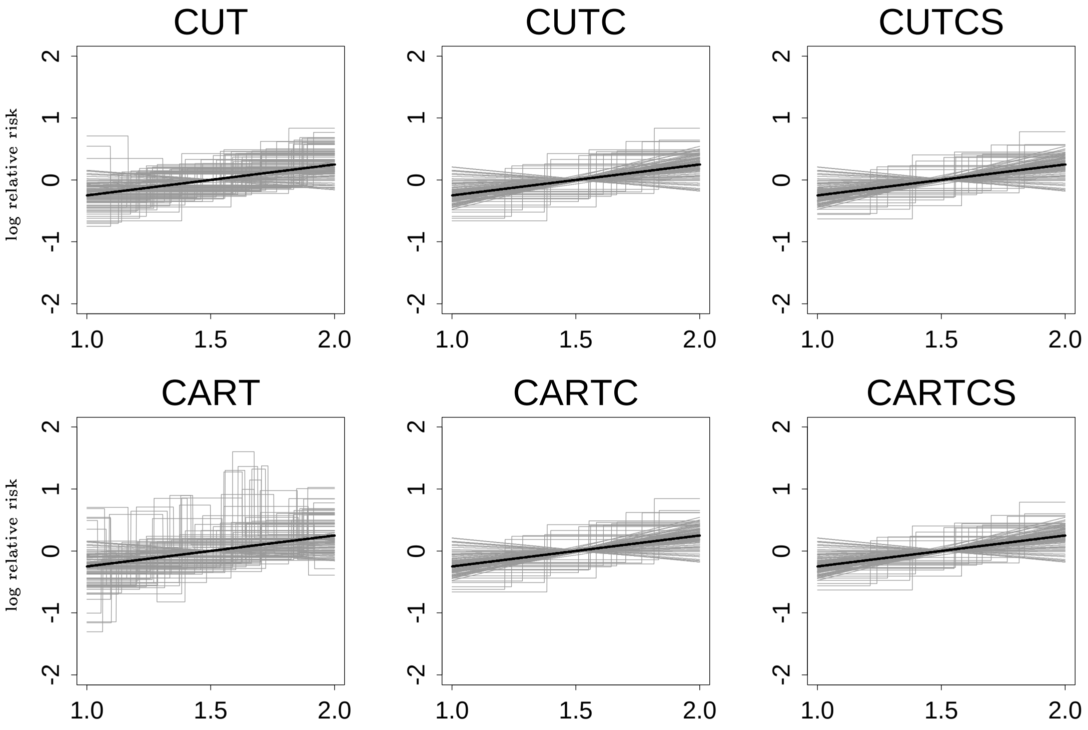


Figure 3.17: Estimated categorized risk functions based on data-driven cutpoints and effect of P-value correction and shrinkage, results of the first 100 replications in the simulated linear model (thick line), standardization to zero mean log relative risk

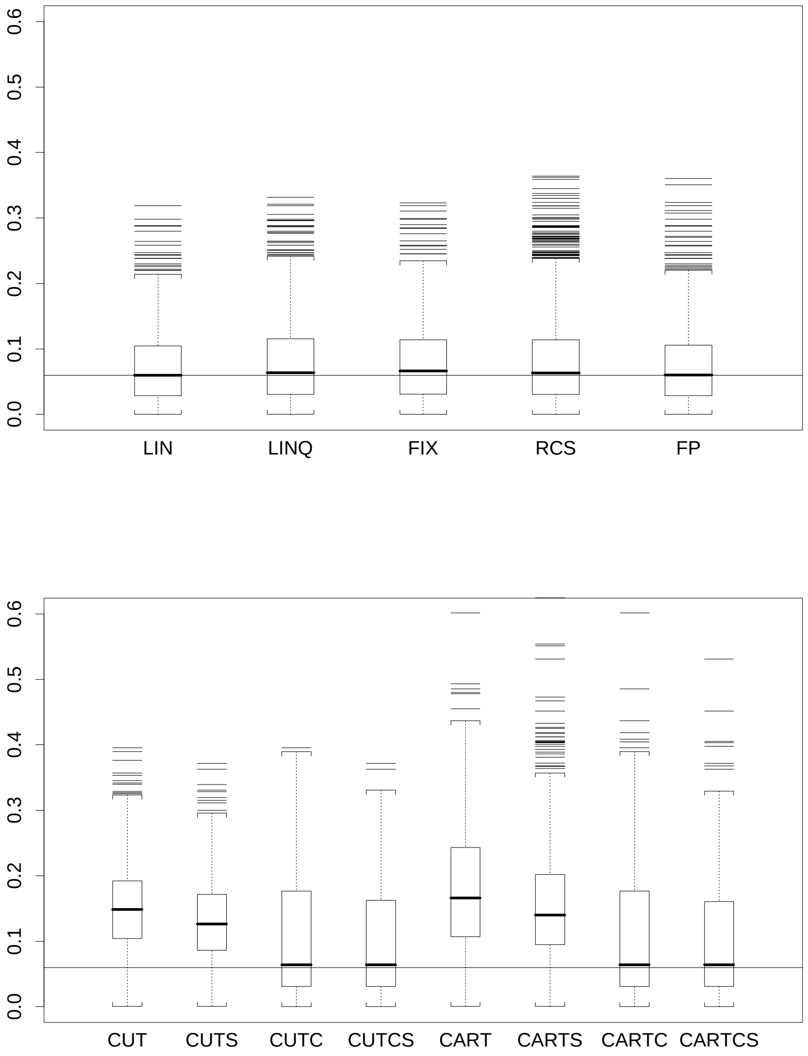


Figure 3.18: Distribution of the $\widehat{MAE}$ for all standard procedures in the simulated linear model based on all 1000 replications, horizontal line denotes the median $\widehat{MAE}$ of LIN

3.6.2 Application of the bootstrap

As for the cutpoint model I used bootstrap samples of the first 100 replications only. Comparing the $\widehat{MAE}$ of the bagged risk function to the corresponding error estimates obtained in the original data there is hardly any difference for LIN and FIX. Especially for LIN this could have been expected in advance, because bagging is based on the same, namely the linear risk functions only. For LINQ, RCS and FP bagging leads to a small error reduction for those replications with a nonlinear risk function in the original data set. However, if the more complex risk function reduced to the linear effect in the original data, bagging may lead to a substantial increase of the error. Similar to the given cutpoint model described in section 3.5 the deviation from the given model, here a linear risk function, is usually higher if the selected risk function is nonlinear. Thus, to explain the effect of bagging we can use the same arguments as for the given cutpoint model: The curvature of the bagged risk function is less extreme as compared to the underlying nonlinear risk function obtained in the original data set. Otherwise, if a linear risk function is chosen in the original data bagging may produce curvature and, consequently, a stronger deviation from the given linear risk function. The largest difference between the risk function obtained in the original data and the corresponding bagged risk function is obtained for CUT(S) and CART. If a significant cutpoint was selected in the original data bagging leads to an error reduction in nearly all replications for CUTS (94%) and in more than 70% for CART. Especially for CUTS the improvement may be substantially (cf. figure 3.19, table A 3.11). If CUTS reduced to the linear risk function in the original data set, we observed both replications with an increase as well as replications with a decrease of the error estimates. For CART the error of the bagged risk function is substantially larger as compared to that of the linear risk function obtained in the original simulated data set in nearly all replications.

Selection frequencies in the bootstrap samples

Figure 3.20 shows the distribution of the number of nonlinear risk functions obtained for the bootstrap samples of the first 100 replications. For LINQ and FP the numbers of bootstrap samples with a nonlinear risk function is similar to that observed in the null model. Note, that the risk function reduced to the linear effect in most of the original data sets (cf. table A 3.9). Therefore, the number of nonlinear risk functions is also small in the corresponding bootstrap samples. If a higher order model was selected in the original data, a nonlinear risk function is also chosen in most of the corresponding bootstrap samples. For FIX, CUT and RCS the rate of nonlinear risk functions obtained in the original data sets was higher as compared to the null model. These results also carry through to the bootstrap samples. However, except for this shift towards higher rates, the distribution of the number of nonlinear risk function in the 100 bootstrap samples is similar to the

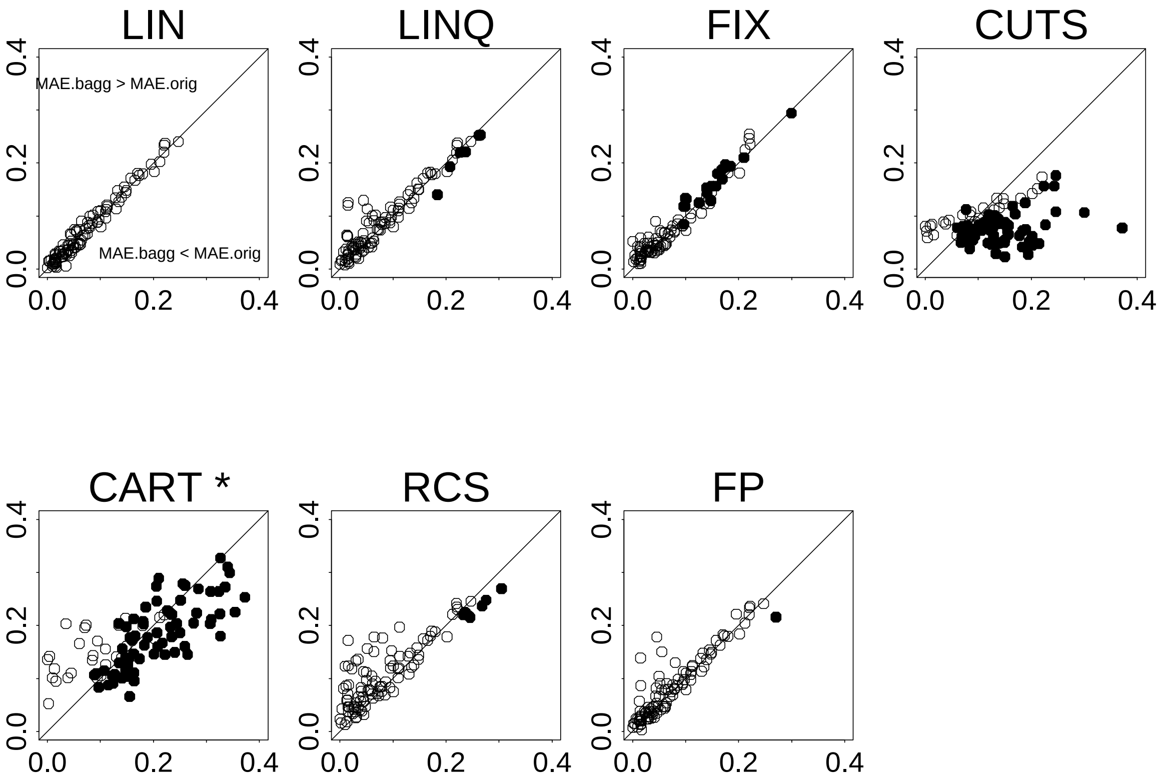


Figure 3.19: $\widehat{MAE}$ of the bagged risk function as compared to the the $\widehat{MAE}$ obtained in the original data set for the first 100 replications of the simulated linear model, dots denote replications with a nonlinear risk function in the original data (* $\widehat{MAE}$ larger than 0.4 in 1 replication, axes cut off at 0.4)

corresponding selection frequencies obtained in the null model (cf. figure 3.10).

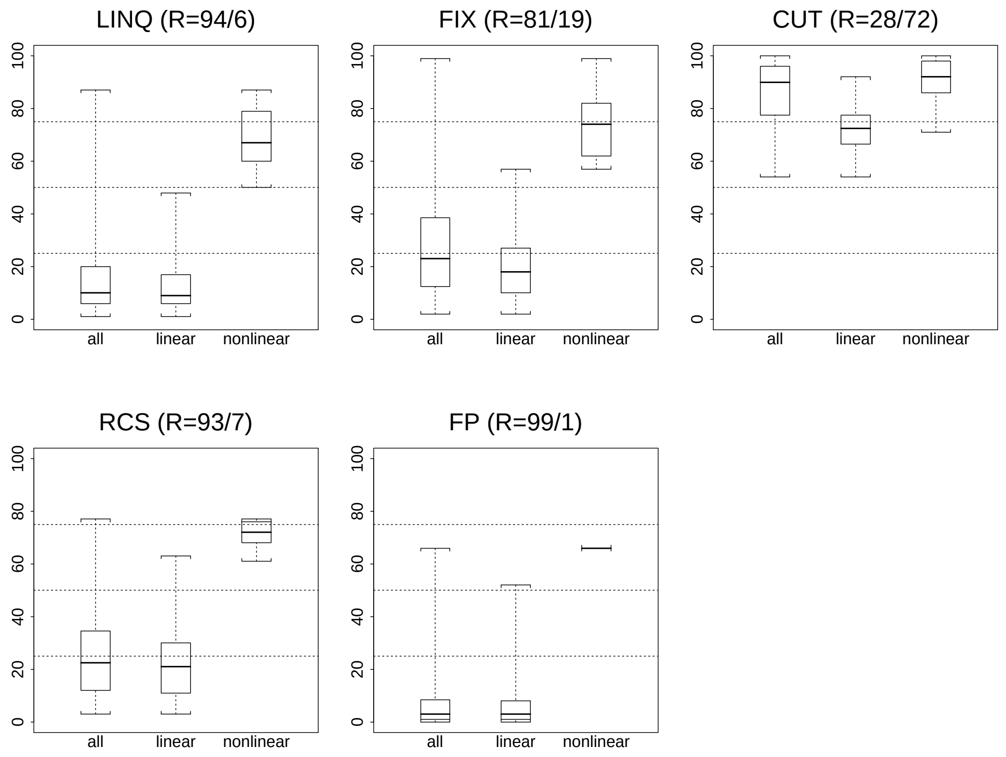


Figure 3.20: Distribution of the number of nonlinear risk functions selected in the bootstrap samples for the first 100 replications in the simulated linear model, results separated by linear/nonlinear risk function in the original data, R denotes the corresponding number of replications with a linear/nonlinear risk function.

3.7 The V-type model

The V-type model describes the situation of a linear increase in risk with increasing distance from a given value μ. Using the transformation of the simulated V-type model, i.e. $y = 2 \mid x - \mu \mid$ and estimating the log relative risk β in the model $\lambda(t \,|Y = y) = \lambda_0(t)\, exp(\beta y)$ is the best one can do in the current situation. Doing so for all replications this method is denoted by TRANSALL. Using the transformation only if the resulting effect estimate is significant (likelihood ratio test) as compared to the null model and taking the linear risk function otherwise, this method will be referred to as TRANS. All risk functions are standardized to a zero mean log relative risk. Appendix A lists some results obtained when adding the estimated logarithmic baseline hazard.

3.7.1 Using standard procedures

As done for the other models of the simulation study the standardized risk functions obtained in the first 100 replications are illustrated graphically (figure 3.21 and 3.22). Error estimates are listed in tables A 3.13 and A 3.14 in the appendix, the distribution of the $\widehat{MAE}$ is displayed in figure 3.23. Obviously, the best fit is obtained when using TRANSALL whereas the given risk function cannot be described adequately by LIN. Consequently, the error estimates obtained for TRANS are slightly larger as compared to TRANSALL. Considering only those 280 replications with a nonlinear risk function selected by TRANS the given effect seems to be overestimated (cf. figure 3.21). However, in spite of this over-optimism the errors are still smaller as compared to LIN (cf. Tables A 3.13 and A 3.14). Replications with small effect estimates $\hat{\beta}$ of the transformed covariate reduced to the linear risk function. The same phenomenon was observed with FIX and FIXALL in the cutpoint model (section 3.5). Although the linear risk function cannot fit the given V-type model correctly, the rate of replications with a linear risk function is rather high for LINQ, RCS and FP (69.8%, 75.9% and 87.4%). If the more complex risk function is chosen the functional shape is usually described adequately. This is especially true for LINQ and FP. However, again the change in risk seems to be overestimated by the nonlinear risk function. Due to its restricted functional shape neither CUT (CUTC, CUTS) nor FIX can describe the given V-type risk function correctly. FIX reduced to the linear risk function in nearly all replications (93.9%), whereas the extensive process of model building leads to a significant data driven cutpoint model in 68.4% of all replications.

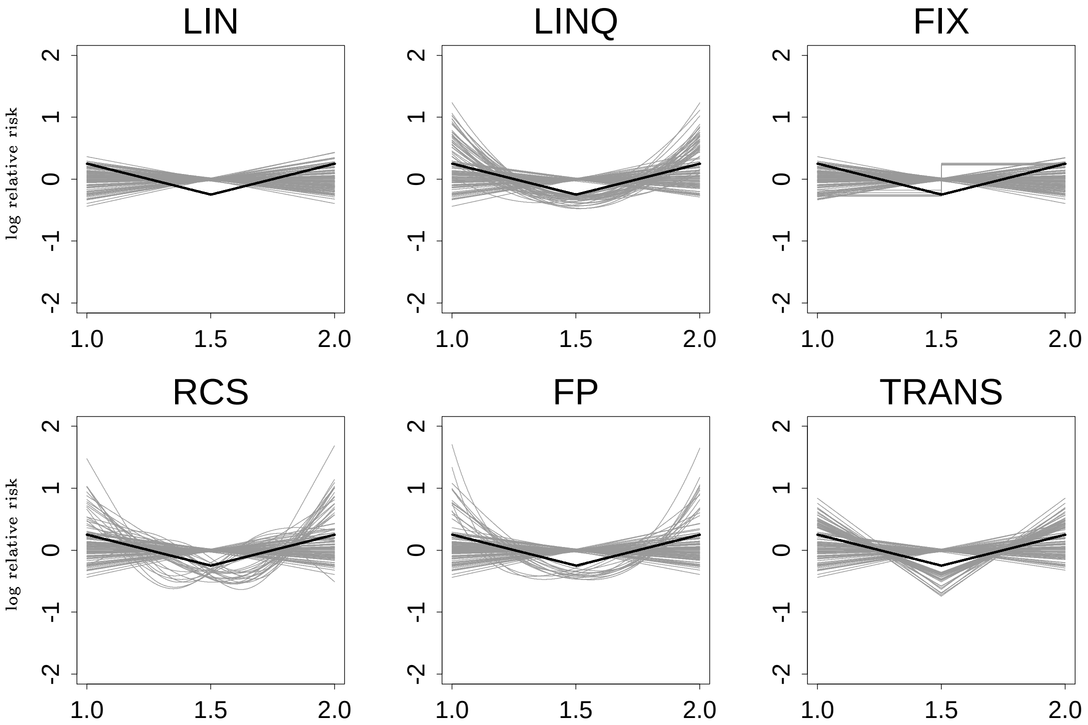


Figure 3.21: Estimated risk functions obtained by using the standard procedures, results of the first 100 replications in the simulated V-type model (thick line), standardization to zero mean log relative risk

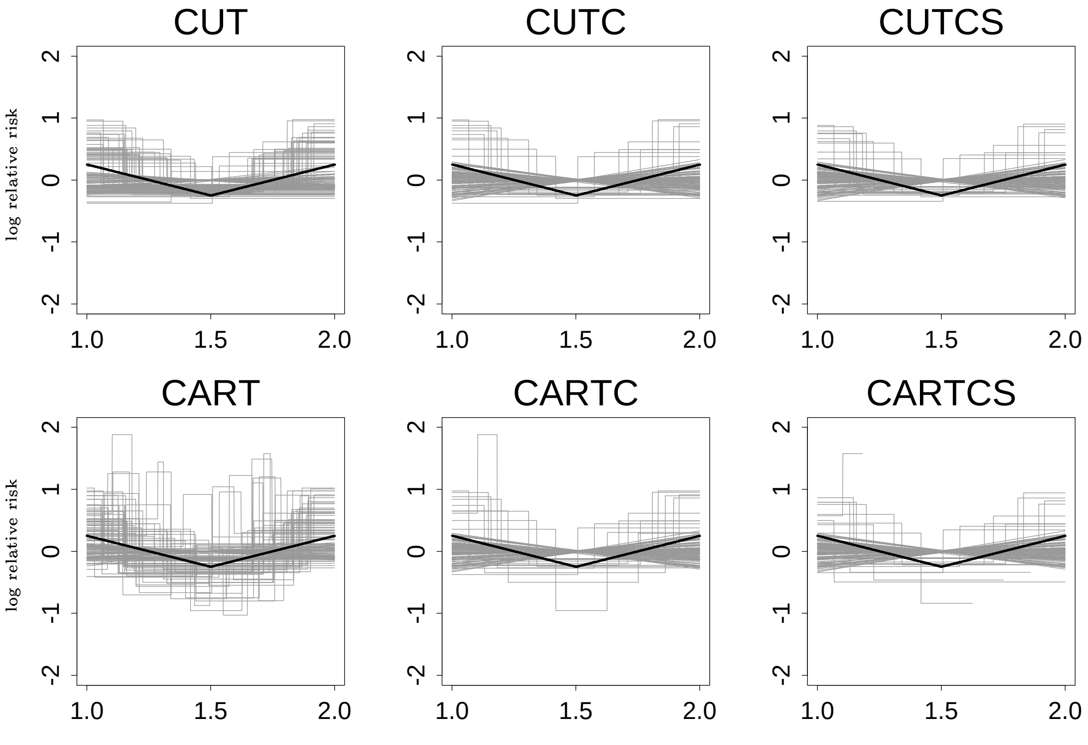
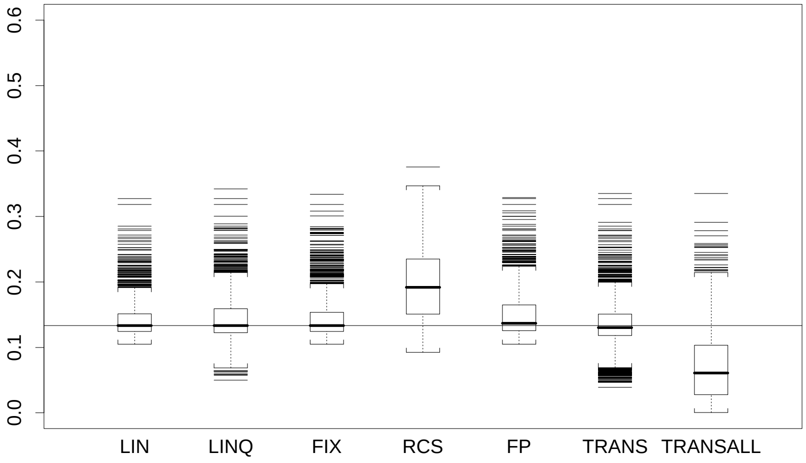


Figure 3.22: Estimated categorized risk functions based on data-driven cutpoints and effect of P-value correction and shrinkage, results of the first 100 replications in the simulated V-type model (thick line), standardization to zero mean log relative risk



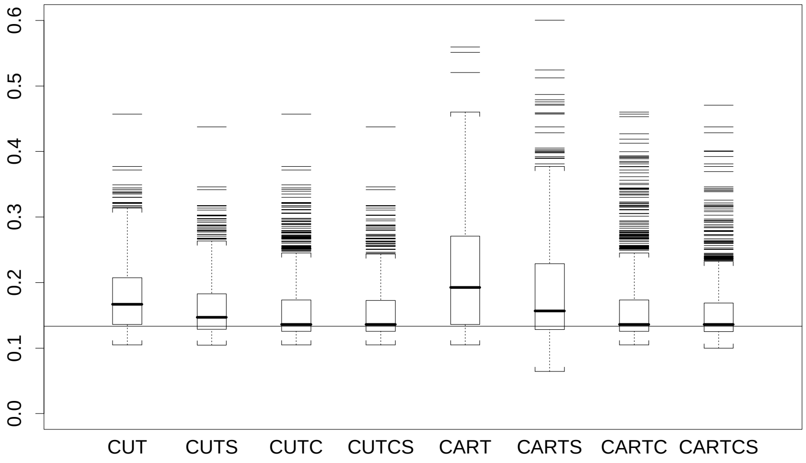


Figure 3.23: Distribution of the $\widehat{MAE}$ for all standard procedures in the simulated V-type model based on all 1000 replications, horizontal line denotes the median $\widehat{MAE}$ of LIN

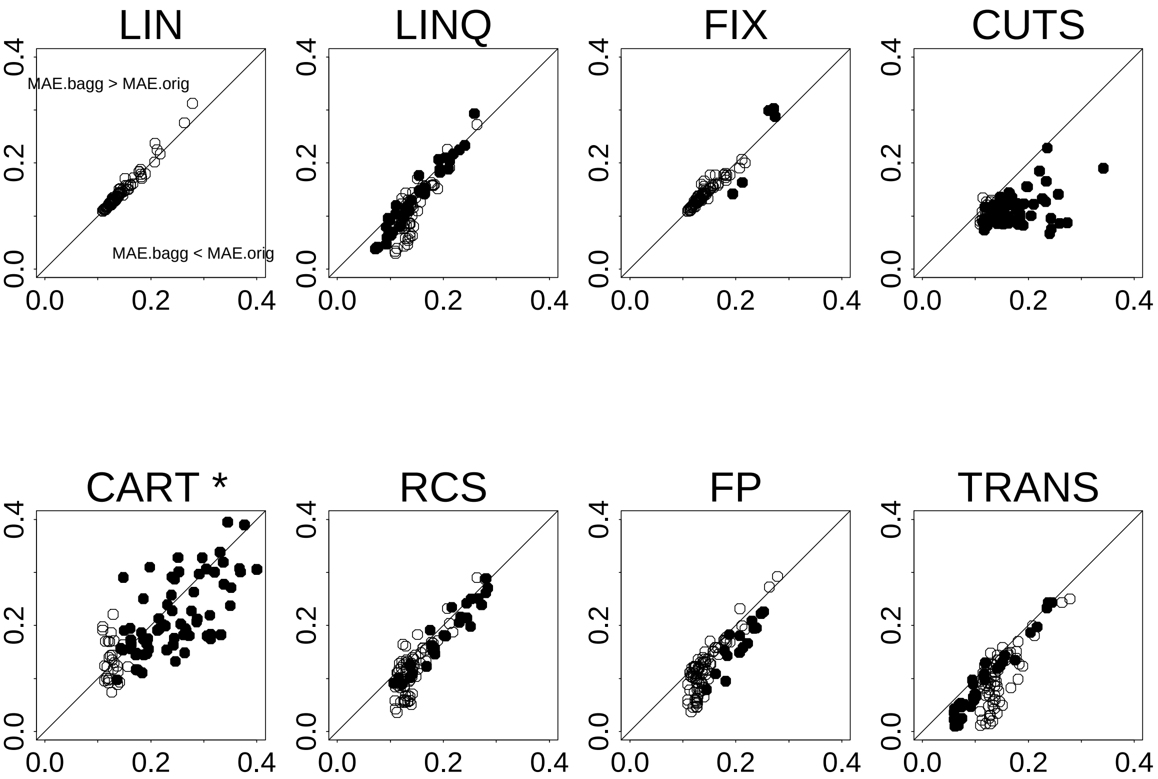


Figure 3.24: $\widehat{MAE}$ of the bagged risk function as compared to the the $\widehat{MAE}$ obtained in the original data set for the first 100 replications of the simulated V-type model, dots denote replications with a nonlinear risk function in the original data (* $\widehat{MAE}$ larger than 0.4 in 2 replications, axes cut off at 0.4)

3.7.2 Application of the bootstrap

The plot of the $\widehat{MAE}$ of the bagged risk function versus the corresponding error estimates obtained in the original data set for the first 100 replications are displayed in figure 3.24. More results can be found in appendix A.

Similar to the previous models there is hardly any effect of bagging for LIN and FIX. For FIX the linear risk function was selected in nearly all replications. This is also true for the corresponding bootstrap samples and, therefore, the bagged risk function is similar to the linear risk function obtained in the original data sets. For LINQ, CUTS, RCS, FP and TRANS the deviation of the bagged risk function from the given V-type risk function is smaller in most replications as compared to the deviation of the underlying risk function obtained in the original data set. In contrast to the simulated models 0, I and II the improvement is also obtained in the majority of replications with a linear risk function in the original data set. For instance, the $\widehat{MAE}$ of $\hat{h}_{\text{bagg}}$ is smaller in 62 out of the 74 (83.8%) replications in which RCS reduced to the linear risk function. The corresponding numbers were 122 out of 937 (13%) in the null model, 49 out of 87 (56.3%) in the cutpoint model and 18 out of 93 (19.4%) in the linear model. These results can be explained by the incapability of the linear risk function to specify the V-type risk function adequately. Since $\hat{h}_{\text{bagg}}$ uses also nonlinear risk functions bagging had a good chance to improve the fit. Considering CUTS neither the categorized nor the linear risk function are able to fit the given function correctly. Thus, bagging can be expected to work for CUTS, too. Due to the large errors produced by CART we observed both, replications with smaller and replications with higher errors of the bagged risk function.

Selection frequencies in the bootstrap samples

Assuming the V-type risk function LINQ, RCS, and FP selects a nonlinear risk function more often as compared to model 0, I and II. Again, these results observed in the original data sets also carry through to the bootstrap samples. Since the linear risk function cannot describe the given functional form correctly, the number of bootstrap samples with a nonlinear risk functions tends also to be higher (as compared to the other models) for those replications, where the more complex function reduced to the linear risk function in the original data sets. This can especially be seen for FP when comparing the distribution displayed in figure 3.25 to the corresponding distributions obtained for model I and II: The 75% quantile of the distribution of the number of nonlinear risk functions selected in the bootstrap samples (of the $R = 85$ replications, in which FP reduced to LIN in the original data) is about 30 for the V-type model, whereas the corresponding quantiles were less than 10 for the simulated cutpoint model (figure 3.15) and the simulated linear model (figure 3.20). However, as for all other models higher selection frequencies are of course obtained for replications with a nonlinear risk function selected in the original data set. For CUT the selection frequencies are similar to that obtained when assuming a linear effect (model II) and for FIX the results corresponds to that of the null model. Note, that the number of significant categorized risk functions obtained by FIX is nearly equal in the null model and the V-type model.

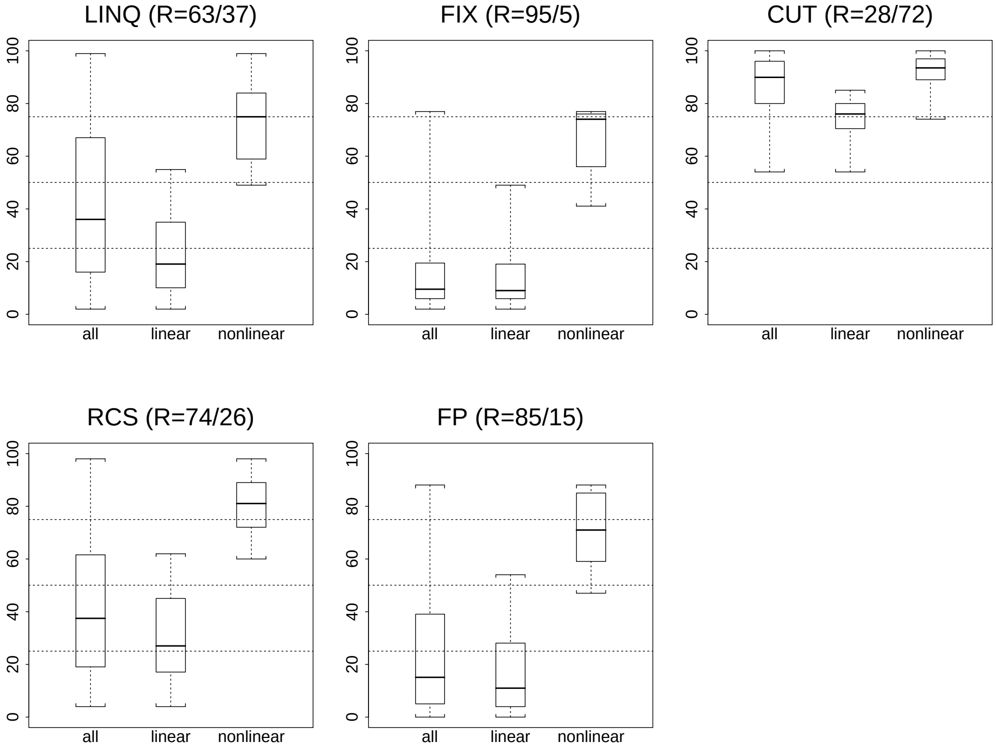


Figure 3.25: Distribution of the number of nonlinear risk functions selected in the bootstrap samples for the first 100 replications in the simulated V-type model, results separated by linear/nonlinear risk function in the original data, R denotes the corresponding number of replications with a linear/nonlinear risk function.

3.8 Further topics

3.8.1 Comparing qualitative and quantitative errors

In order to investigate the capability of all standard procedures to estimate the given risk function adequately I focused on graphical methods and the estimated mean absolute error $\widehat{MAE}$. One reason to prefer $\widehat{MAE}$ was that the difference between an observation $X = x_1$ with a large deviation from the the given function at x_1 and an observation $X = x_2$ with a small deviation from the given function at x_2 is less extreme as compared to the more common measure $\widehat{MSE}$. Furthermore, the $\widehat{MAE}$ was chosen for better illustration of the results. Using one specific simulated data set I discussed the need of an qualitative error. However, my proposal defined in section 3.3 showed that it is difficult to define such an error adequately, especially if the risk function or its estimate is constant for all or nearly all values of X. Therefore, I did not use this measure for the assessment of the standard procedures and the corresponding bagged risk functions.

In this section I compare the qualitative errors obtained for LINQ, RCS, FP and CUT to those obtained for LIN in the linear model (model II) and for the V-type risk function (model III). Qualitative errors of the bagged risk functions were not calculated. Furthermore, I will investigate, how far results differ when using $\widehat{MAE}$, $\widehat{MSE}$ and $\widehat{MSE}_{\text{qual}}$, respectively, for the assessment of the estimated risk functions.

Qualitative errors as defined in section 3.3 for the more complex models are plotted against those obtained for LIN: figure 3.26 show the results obtained for the linear effect assumption (model II) and figure 3.27 display the corresponding results for the V-type risk function. Obviously, if LINQ, RCS, FP or CUT reduced to the linear risk function the value of $\widehat{Err}_{\text{qual}}$ is equal to that obtained for LIN.

Let us first consider the results obtained for the simulated linear model (figure 3.26), where the given linear effect was $\beta = 0.5$. If the estimated risk function is also linear $\widehat{Err}_{\text{qual}}$ is either equal to its minimum 0 (if $\hat{\beta}$ is positive) or equal to its maximum 1 (if $\hat{\beta} < 0$). In the first case the qualitative error of the more complex risk function can only be larger whereas it must be smaller in the latter situation. The curvature produced by the nonlinear risk functions of LINQ, RCS and FP usually leads to both, an increase as well as a decrease in risk between two consecutive values $x_{(j)}$ and $x_{(j+1)}$. The value of $\widehat{Err}_{\text{qual}}$ decreases with an increasing rate of pairs $(x_{(j)}, x_{(j+1)})$ with $\hat{h}_c(x_{(j)}) < \hat{h}_c(x_{(j+1)})$. Therefore, depending on the rate of those pairs the value of $\widehat{Err}_{\text{qual}}$ centers around 0.5. For the categorized risk function obtained by CUT this value is nearly equal to 0.5 according to the definition of the qualitative error.

For the V-type risk function the value of $\widehat{Err}_{\text{qual}}$ obtained for CUT is, per definition, also nearly equal to 0.5 (figure 3.27). As mentioned earlier the V-type risk function cannot be estimated adequately by LIN, because this approach models either an increase or a

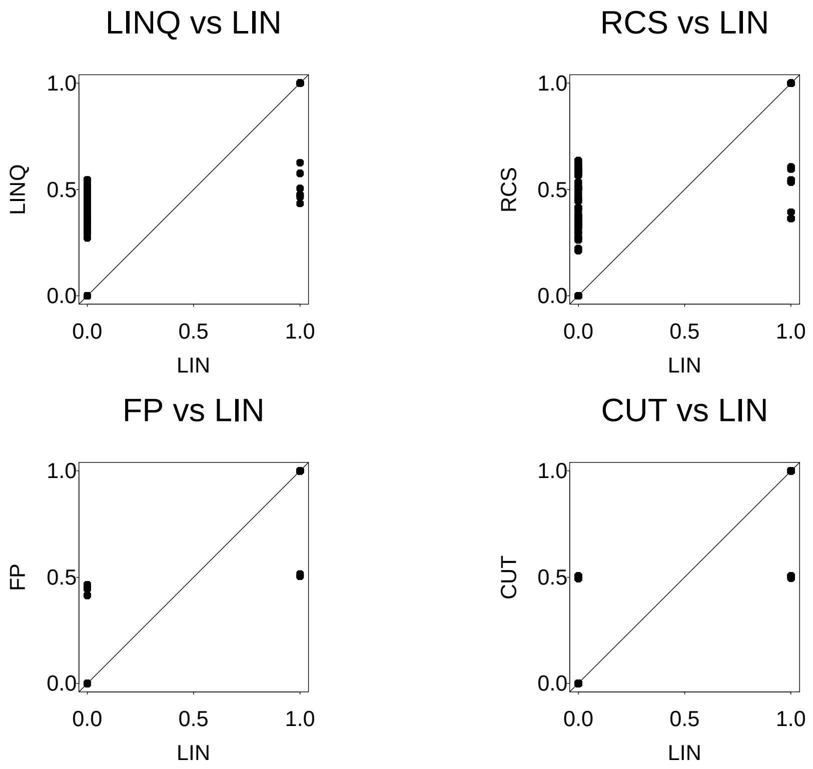


Figure 3.26: Qualitative errors for the simulated linear model, $\widehat{Err}_{\text{qual}}$ is equal for both methods on the diagonal line

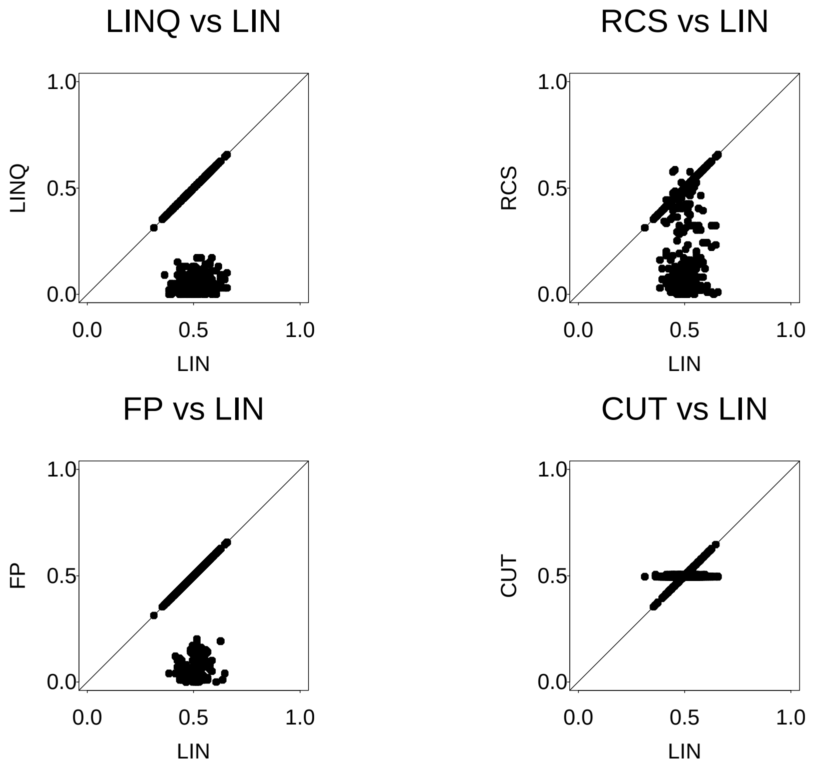


Figure 3.27: Qualitative errors for the simulated V-type model, $\widehat{Err}_{\text{qual}}$ is equal for both methods on the diagonal line

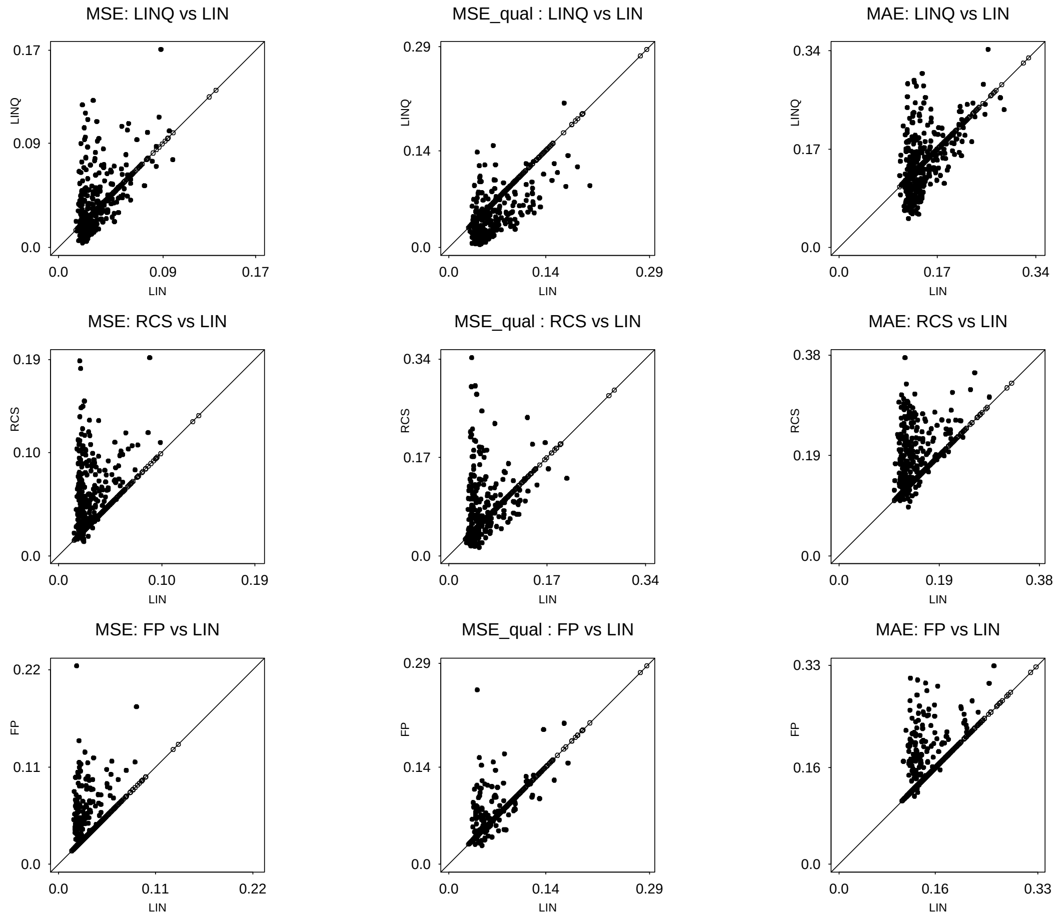


Figure 3.28: Comparing $\widehat{MSE}$, $\widehat{MSE}_{\text{qual}}$ and $\widehat{MAE}$ for the simulated V-type model

decrease in risk. Therefore, the qualitative error obtained for the estimated linear risk function centers also around 0.5. In contrast, the given functional form may be specified correctly by the nonlinear risk functions obtained by LINQ, RCS and FP. For LINQ and FP we obtained an improvement with respect to the qualitative error as compared to LIN in all replications with a nonlinear risk function. For RCS there are a few replications with a slightly larger value $\widehat{Err}_{\text{qual}}$ than that of the corresponding estimated linear risk function.

The fact that the qualitative error of LINQ, RCS and/or FP is close to 0 in many replications indicate that these approaches are capable to model the decrease as well as the increase in risk of the V-type risk function correctly. However, it was shown in the last sections that the difference in risk may be overestimated substantially. Therefore, the qualitative error should not be used without considering a quantitative measure of the deviation between estimated and given risk function.

In order to combine the quantitative and the qualitative error a so called qualitative MSE that uses the qualitative error as some kind of penalty term was defined in section 3.3. Here I use again the weight $w = 2\widehat{MSE}$ to calculate this error. For the V-type model figure 3.28 plots the values of $\widehat{MSE}_{\text{qual}}$ obtained for LINQ, RCS and FP, respectively, against those calculated for LIN. Furthermore, these results are compared to those based on $\widehat{MSE}$ and $\widehat{MAE}$. Generally, the number of replications with a smaller error of the nonlinear risk function as compared to LIN is higher when using $\widehat{MSE}_{\text{qual}}$ instead of $\widehat{MSE}$, the corresponding numbers are listed in table 3.4. For instance, LINQ is better than LIN in 161 out of 302 replications with a nonlinear risk function when using $\widehat{MSE}$ for the assessment, whereas the number is 251 when using $\widehat{MSE}_{\text{qual}}$. Using the latter error measure the correct functional shape obtained by LINQ is rewarded or – in other words – $\widehat{MSE}_{\text{qual}}$ penalizes the wrong functional form of LIN. However, interpreting these results it should be taken into mind that the choice of w to calculate $\widehat{MSE}_{\text{qual}}$ is somewhat arbitrary. Taking $\widehat{MAE}$ for the assessment the results are similar to those based on $\widehat{MSE}$. However, the number of nonlinear risk functions with a smaller error than LIN is slightly larger as compared to $\widehat{MSE}$ indicating that the small quantitative deviation of the linear risk function from the give V-type risk function has a smaller influence when using $\widehat{MAE}$.

Table 3.4: Comparing the fit obtained by LINQ, RCS and FP to that obtained by LIN in the V-type model by using different error measures

		no. of replications with a smaller error than LIN		
method	no. replications with linear/nonlinear risk function	$\widehat{MSE}$	$\widehat{MSE}_{\text{qual}}$	$\widehat{MAE}$
LINQ	698/302	161	251	175
RCS	759/241	18	105	32
FP	874/126	1	46	5

Concluding remarks

Although it may be sensible to define a qualitative error in addition to a quantitative error for the assessment of the estimated risk functions $\widehat{Err}_{\text{qual}}$ seems to be useful for the V-type model and monotone risk functions only. Furthermore, the examples described in this section support the use of $\widehat{MAE}$ rather than $\widehat{MSE}$ in the current situation.

3.8.2 Visualizing the effect of bagging

In order to illustrate the capability of bagging to improve the estimation of a risk function on the one side, and to show the negative effect of bagging when the risk function estimate is quite good in the original data I consider two examples. Each of these examples is based on one selected simulated data set and the corresponding 100 bootstrap samples only, the results are displayed in figure 3.29 and error estimates are listed in table 3.5.

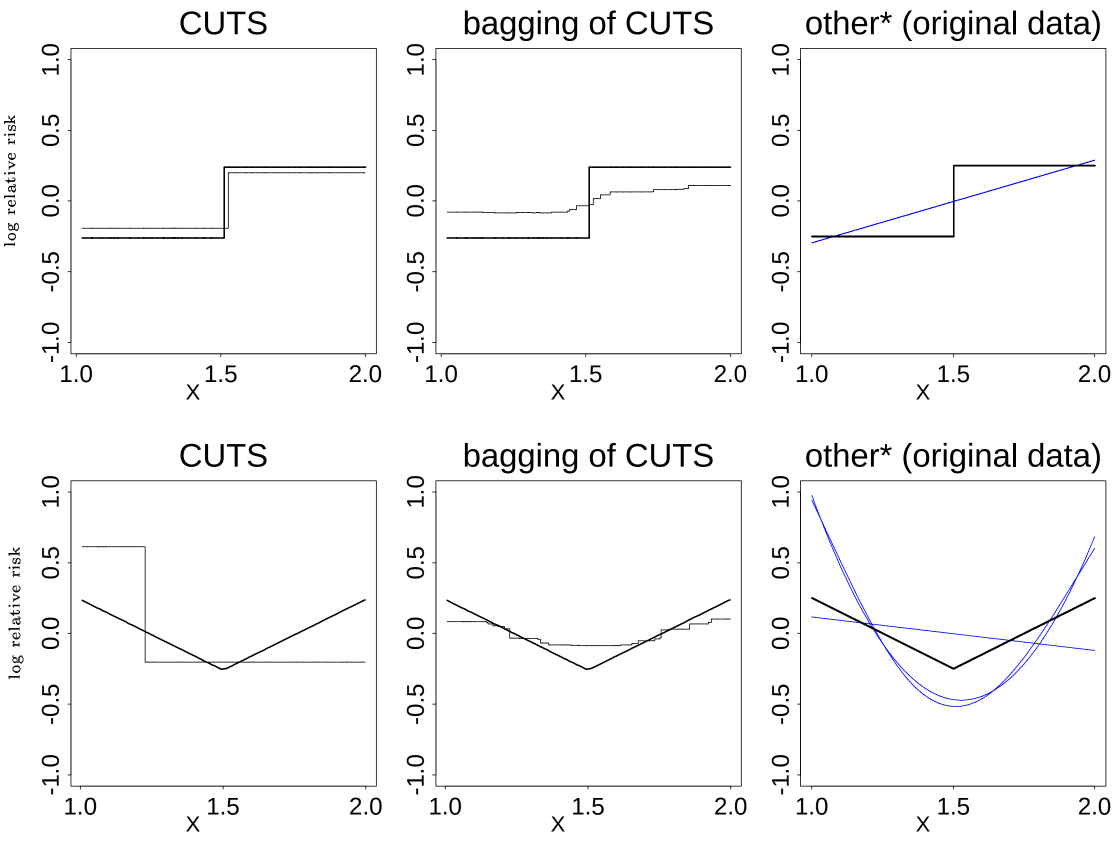


Figure 3.29: The effect of bagging on the estimation of the risk function: Examples based on one simulated data set (replication) for the cutpoint model and the V-type model, (* LIN, RCS and FP)

For the selected data set from the simulated cutpoint model the categorized risk function obtained by CUTS corresponds well to the given cutpoint model. The estimated data driven cutpoint as well as the estimated log relative risk between the resulting risk groups are close to the given value and, therefore, the resulting $\widehat{MAE}$ is very small. Using RCS or FP the risk function reduced to the linear effect for both approaches, the estimated linear effect seems to be a good approximation to the given cutpoint model in the current situation. The bagged risk function based on CUTS is more smooth and the change in risk is less extreme as compared to the risk functions obtained in the original data. Consequently, the resulting difference to the given cutpoint model measured by the $\widehat{MAE}$ gets larger. Similar results were obtained in other replications, remember that (in contrast to models 0, II and III) bagging does not necessarily lead to an decrease of the error for CUTS in the given cutpoint model (cf. figure 3.14). For LIN and FP the error estimates are similar to that obtained in the original data resulting from the fact that the bagged risk function is the average over 100 linear risk functions. Note that for the selected simulated data set FP reduced to LIN in all 100 bootstrap samples. Consequently, the bagged risk function based on FP is equal to that of LIN. RCS reduced to the linear risk function in 14 out of the 100 bootstrap samples.

The second example is based on one simulated data set from the V-type model. Neither by CUTS nor by LIN the given risk function can be specified correctly. However, the $\widehat{MAE}$ of LIN is rather moderate indicating that a restriction to this quantitative measure may lead to a false interpretation. Although the given functional form is described quite well by RCS and FP, the resulting error estimates are large due to the huge change in risk. Bagging leads only to a slight error reduction for FP and LIN, whereas the error of the bagged risk function based on RCS is substantially smaller than that estimated in the original data. The smallest $\widehat{MAE}$ is obtained for the bagged risk function of CUTS, which is a good approximation to the given V-type model in the current situation. Thus, bagging may lead to a substantial improvement, even if the underlying method used to estimate the risk functions in the bootstrap samples is misspecified.

Table 3.5: The effect of bagging on the $\widehat{MAE}$ for the two examples described in figure 3.29

method	cutpoint model		V-type model	
	original data	bagging	original data	bagging
LIN	0.113	0.123	0.139	0.137
CUTS	**0.028**	**0.174**	**0.262**	**0.067**
RCS	0.113[1]	0.110	0.244	0.075
FP	0.113[1]	0.123[2]	0.230	0.209

[1] reduced to the linear risk function

[2] FP reduced to the linear risk function in all 100 bootstrap samples

3.8.3 The influence of model building on the estimation of parameters

In section 2.1.7 I have already discussed the problem of bias caused by model selection. In order to correct for overestimation of the log relative risk in the data driven cutpoint model we used shrinkage procedures. Considering a given cutpoint model in previous simulations (Schumacher et al., 1997) we have shown that the selection of a data-driven cutpoint (CUT) led to overestimation of the resulting log relative risk β, if β is small or, as in the current simulation study, of moderate size. However, overestimation caused by the model building process may also be relevant to other models and when applying less intensive selection strategies, e.g. the selection between a higher order risk function and the linear risk function. Throughout the simulation the higher order risk function was only taken if it led to an improvement in terms of log-likelihood as compared to the linear risk function. Except for the simulated null model it was not investigated whether there was an effect at all, i.e. if the model including the risk function is significantly better than the model with a null risk $h(x) = 0$ for all values x. For the simulated null model we tested the selected model against the null in order to estimate type I error rates. In this section we investigate the effect of the continuous covariate for the other simulated models (cutpoint, linear, V-type). As for the simulated null model we considered the model selection strategies H, M and B (cf. section 3.4).

Applying the *best* method for each simulated model, namely the estimation of the risk function by using the given functional form, figure 3.30 plots the P-Values of the likelihood ratio test testing the model with the *best* risk function against the model with a null risk versus the estimated parameters $\hat{\beta}$ of the corresponding risk function for all 1000 replications. A significant log relative risk is obtained, if the P-value of the likelihood ratio test is smaller than 0.05. In figure 3.30 a significant log relative risk $\hat{\beta}$ corresponds to a value below the horizontal line. Averaging $\hat{\beta}$ over all replications the resulting mean $\bar{\hat{\beta}}$ corresponds well to the given value β (vertical line) in all models. This would not be true when restricting to replications with a significant effect only (cf. table 3.6). For instance, FIXALL led to $\bar{\hat{\beta}} = 0.507$ when using all replications, whereas the given value $\beta = 0.5$ is overestimated when averaging over the 672 (67.2%) replications with a significant effect only ($\bar{\hat{\beta}} = 0.618$). For the linear and the V-type model the model selection process (the rate of significant replications in table 3.6 is based on strategy H) would led to average parameter estimates, which are nearly twice as large as the given β.

The rate of significant replications obtained in the cutpoint model (67.2 %) corresponds well to power calculations: using the simulated median survival times of the two subgroups defined by the given cutpoint μ and assuming subgroups with 50 patients I obtained a power between 65% and 70% (depending on the assumption for the accrual and follow up times). However, using the given risk functions (LIN and TRANS) in the simulated linear and V-type model, respectively, a significant effect is obtained in 27.9 and 28.0%,

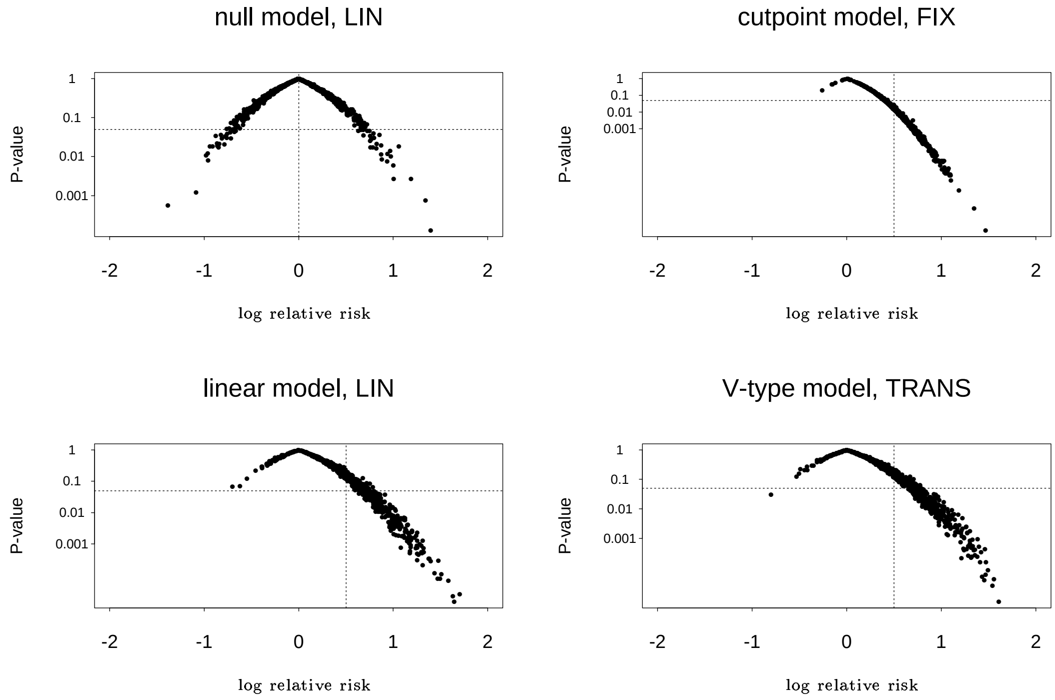


Figure 3.30: P-values of the likelihood ratio test against the null ($h(x) = 0$) versus parameter estimates $\hat{\beta}$ based on the given risk function, vertical lines denote the given effect β, horizontal line denote P=0.05

respectively, of all replications only. Thus, even if the model is correctly specified, it seems to be difficult to detect a moderate effect of the covariate on survival in these models.

Table 3.6: The effect of model selection when estimating the risk function by using the prespecified function: Estimated mean log hazard ratio $\bar{\hat{\beta}}$ obtained in all replications to that based on replications with a significant effect only

model	given β	method	all (R=1000)	based on significant $\hat{\beta}$ only	
			$\bar{\hat{\beta}}$	No. of replications	$\bar{\hat{\beta}}$
null (0)	0	LIN	0.005	54[1]	0.136[2]
cutpoint (I)	0.5	FIX(ALL)	0.507	672	0.618
linear (II)	0.5	LIN	0.503	279	0.955
V-type (III)	0.5	TRANS(ALL)	0.504	280	0.951

[1] corresponding to the estimated type I error $\hat{p}_{err} = 0.054$ (cf. section 3.4.1)

[2] model selection lead to large positive and negative values of $\hat{\beta}$ (cf. figure 3.30)

Table 3.7: Rate of replications with a significant effect obtained for all risk functions in the cutpoint model, the linear model and the V-type model

method	cutpoint model			linear model			V-type model		
	strategy			strategy			strategy		
	H	M	B	H	M	B	H	M	B
LIN	53.1	–	–	27.9	–	–	4.0	–	–
LINQ	41.6	55.0	55.5	22.7	30.8	32.0	22.1	22.3	23.2
FIX[1]	67.2	71.8	71.8	21.5	31.0	31.0	6.1	7.8	7.8
CUT/CART[1]	90.8	90.8	90.8	68.9	68.9	68.9	68.4	68.4	
CUTC/CARTC[1]	44.9	57.5	57.5	21.9	31.6	31.6	18.3	18.3	
RCS	45.0	58.0	60.0	21.1	31.2	33.2	20.4	21.5	22.0
FP	23.6	53.5	53.5	9.5	28.6	28.6	10.1	13.0	13.1
TRANS[1]	–	–	–	–	–		28.0	31.3	31.3

[1] not nested to LIN, strategy M is equal to B, because the risk functions are not tested against LIN.

Table 3.7 summarizes the results obtained for all standard procedures in the cutpoint model, the linear model and the V-type model. The highest rate of replications with a significant effect is obtained in the simulated cutpoint model. This may be caused by the clear separation of risk groups making it probably easier to detect the increase in risk. For instance, LIN led to 53.1% significant models. As described in section 3.5 the linear risk function was a good approximation of the given cutpoint model. In contrast, for the simulated linear model the rate of replications with a significant linear effect is 27.9% only. In the V-type model LIN showed a significant effect in 4% of all replications only. This is not astonishing, since the given V-type risk function cannot be described appropriately by a linear function.

For the simulated cutpoint model and the simulated linear model we obtained smaller rates for LINQ, RCS and FP as compared to LIN when using strategy H. However, the rate is larger than LIN when using strategies M and B. For instance with strategy H we obtained a significant restricted cubic spline in 45% of all replications in the simulated cutpoint model. This rate is increased to 56% and 60% respectively, when using M and B, respectively. Generally, a nonlinear risk function should be used only if it is better than the linear risk function. Doing so, the chance to detect an effect of a continuous covariate is increased as compared to LIN. Investigating the effect of age on event-free survival in the GBSG-2 study (cf. section 2.1.9), there was no effect at all when assuming a linear risk function but a highly significant effect when using a nonlinear risk function, e.g. RCS.

However, as shown in the simulation study a significant nonlinear risk function is no guaranty for a good fit. Simulating the V-type model for example, a significant data-driven cutpoint is obtained in 68.4% of all replications, although the V-type model cannot be described correctly by a cutpoint model. In contrast, a good approximation of the V-type shape can be obtained when using LINQ or FP. For these methods the rate of replications with a significant effect is increased from 4% for LIN to 22.1% for LINQ and 10.1% for FP, respectively (table 3.7). A further increase of this rate is obtained when using strategies M or B, e.g. 13% for FP when using M. However, since this increase is based on significant linear risk functions only there is no improvement with respect to the fit of the given functional shape. Note that all significant fractional polynomials have been used to calculate the rate 10.1% for strategy H.

3.8.4 A note on the estimation of confidence intervals

In the simulation study I focus especially on the capability of the chosen methods to estimate the given function correctly. Categorizing X by using one or more data driven cutpoints I have already addressed the problem of bias caused by model building. In order to correct for overestimation of the log relative risk I used shrinkage methods. Furthermore, corrected P-values should be used to get reliable test results (cf. section 2.1.7). I have also mentioned that selected cutpoints may differ between studies and, therefore, it is difficult to compare results. Figure 3.31 shows that both, a large variability of selected cutpoints as well as overestimation of the resulting log relative risk, is more relevant in the null model than in the cutpoint model. Applying shrinkage methods we can at least correct for overestimation. Notice, that figure 3.31 uses the same results as figure 3.4 and 3.12.

However, so far I have not investigated the validity of the estimated variances and confidence intervals of the selected model. As illustrated above variable selection may lead to overestimation of effects in the selected model. This phenomenon is well known in the multivariate situation, where specific variables are selected and included into a *final* regression model. Besides overestimation of the effect the corresponding variance estimates and, consequently the resulting confidence intervals may be too small (Miller, 1990).

Using the selection of a data driven cutpoint referred to as CUT as a *prototype* for other, more complicated problems of model selection I illustrate the (potential) impact of the model building process on the estimation of confidence intervals in the null model. After categorizing X by using the selected cutpoint $\hat{\mu}$ the estimated log relative risk $\hat{\beta}$ between the resulting subgroups is obtained by maximization of the corresponding partial likelihood (cf. section 2). A confidence interval for β is then calculated as

$$\hat{\beta} \pm z_{1-\alpha/2} \sqrt{\widehat{\text{var}}_{\text{mod}}\,(\hat{\beta})}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution and $\widehat{\text{var}}_{\text{mod}}\,(\hat{\beta})$ is the model-based estimated variance of $\hat{\beta}$ that is also derived from the proportional hazards cutpoint model based on $\hat{\mu}$. Figure 3.32 A shows the estimated log relative risk with corresponding 95% confidence intervals obtained in the first 100 replications of the null model. The samples are ordered according to the value of $\hat{\beta}$ from smallest to largest. As mentioned previously there are no values of $\hat{\beta}$ close to 0 resulting from the optimization process of the minimum P-value approach. 40 of the 100 model based confidence intervals do not contain the given log relative 0. These results correspond well to the inflated type I error obtained for CUT ($\hat{\alpha} = 0.432$, cf. section 3.4). Overestimation can be reduced by considering the shrinked estimates $\hat{c}\hat{\beta}$ instead of $\hat{\beta}$. In this thesis I used a heuristic shrinkage factor $\hat{c} = (\hat{\beta}^2 - \widehat{\text{var}}_{\text{mod}}\,(\hat{\beta})\,)/\hat{\beta}^2$. Figure 3.32 B and figure 3.31 N2 show that the shrinked estimated log relative risk is closer to 0. However, we still obtain 14 out of

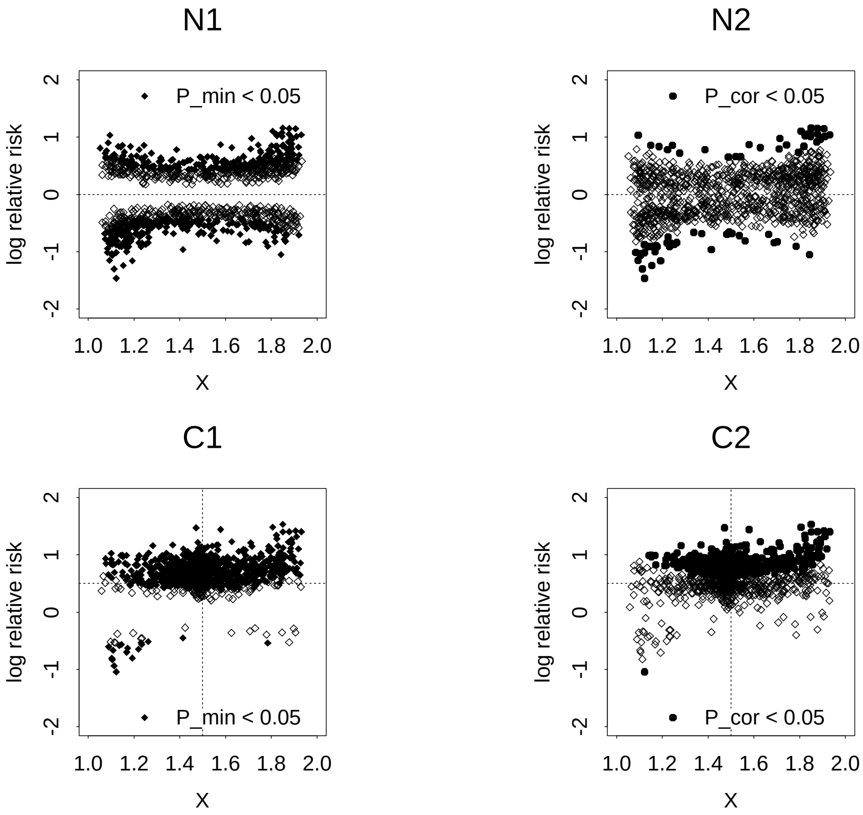


Figure 3.31: Estimated log relative risk versus selected data driven cutpoint obtained for CUT/CUTS in the null model (N1/N2) and the cutpoint model (C1/C2)

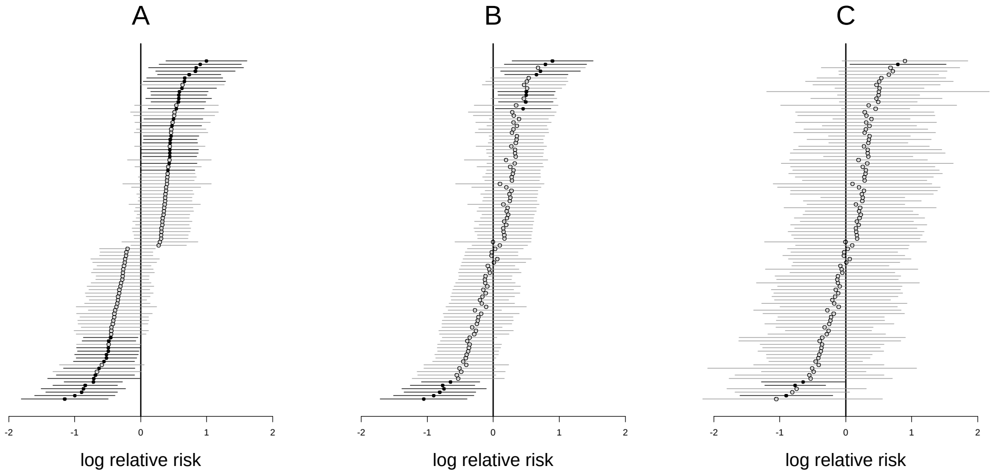


Figure 3.32: Estimated log relative risk versus selected data driven cutpoint obtained for CUT/CUTCS in the null model, black lines denote confidence intervals that do not cover the null, all samples are ordered according to the value of $\hat{\beta}$ from smallest to largest

100 significant confidence intervals when calculating

$$\hat{c}\hat{\beta} \pm z_{1-\alpha/2} \sqrt{\widehat{\mathrm{var}}_{\mathrm{mod}}\,(\hat{\beta})} \tag{43}$$

The reason is that the variance of the estimated log relative is still model based, i.e. is derived from a proportional cutpoint model where a fixed and predefined cutpoint is assumed. Thus, the variance estimate and the resulting confidence intervals tends to be too small in the current situation.

In order to take the additional variability of both the estimated cutpoint $\hat{\mu}$ and the estimated shrinkage factor into account, the results of the bootstrap samples can be used for variance estimation. For each replication we simply take the empirical variance of the shrinked log relative risk over the corresponding bootstrap samples given by

$$\widehat{\mathrm{var}}_{\mathrm{boot}}\,(\hat{c}\hat{\beta}) = \frac{1}{B-1} \sum_{j=1}^{B} \left(\hat{c}_j\hat{\beta}_j - \overline{c\beta}_{\mathrm{boot}}\right)^2$$

where $\overline{c\beta}_{\mathrm{boot}}$ denotes the average of the shrinked estimated log relative risks over the B bootstrap samples. Replacing $\widehat{\mathrm{var}}_{\mathrm{mod}}$ in formula (43) by $\widehat{\mathrm{var}}_{\mathrm{boot}}$ the coverage of the resulting confidence intervals corresponds well to the results obtained by P-value correction. For the first 100 replications in the null model 4 out of 100 confidence intervals do not contain the value 0 (figure 3.32 C). Using corrected P-values we obtained an estimated type I error of $\hat{p}_{err} = 0.067$ (cf. section 3.4.1).

3.8.5 Simulating the null hypothesis from the data

Estimating the effect of age on event-free survival (EFS) in the GBSG-2 study nearly all risk functions showed a strong decrease in risk with increasing age up to 50 years. Furthermore, there were only slight differences between the bagged risk functions and those estimated in the original data. Using e.g. RCS a restricted cubic spline was selected in all 100 bootstrap samples and for FP the risk function reduced to the linear risk function in 12 out of 100 bootstrap samples only. These results suggest a strong nonlinear effect of age on EFS.

In order to connect real data to the simulation study I artificially generated independence between age and EFS. Independence was obtained by a random re-allocation of the observed values of patient's age to the observed EFS, which is equivalent to the null hypothesis of no prognostic relevance of age with respect to EFS. This problem was repeated 100 times and in each repetition the risk function was estimated by LIN, LINQ, FIX2, CUTS, RCS and FP. Since the results of CART was not very promising in the simulation study I did not use this approach in this section. The estimated risk function that were standardized to a zero mean log relative risk in all 100 repetitions are displayed in figure 3.33. All in all the random re-allocation example confirms the results that I obtained for the null model. The best fit is obtained for LIN. Using model selection strategy M , i.e. using the higher order model if it leads to a significant improvement as compared to LIN, a nonlinear risk function was selected in 3, 4 and 1, respectively, repetitions when using LINQ, RCS and FP, respectively. Furthermore, all three models hold the type I error. For FIX2 a categorized risk function was selected in 7 out of 100 repetitions. CUT selects a significant data driven cutpoint in 30 repetitions indicating again the inflation of the type I error rate of the minimum P-value approach.

To investigate the effect of bagging in the current situation I consider an example based on one selected data set only: I use the data from one repetition of the 100 random reallocations, in which LINQ, RCS and FP selected the higher order risk function. From this selected data set 100 bootstrap samples were generated. Risk functions were estimated in each of these bootstrap samples using again the model selection strategy M. The resulting risk functions for LINQ and RCS are shown in figure 3.34. In 20 bootstrap samples LINQ reduced to the linear risk function, for RCS model selection yields to 29 bootstrap samples with a linear risk function. Considering the results of all bootstrap samples nonlinearity as obtained in the selected data set cannot be assumed for RCS and is at least questionable when interpreting the risk functions based on LINQ. Especially for RCS the risk functions show more heterogeneity between bootstrap samples than those based on the original data set of the GBSG-2 study (cf. figure 2.4). This may be caused by the fact that the risk difference is smaller in the selected data set of the random re-allocation experiment whereas the effect of age was strong in the original data.

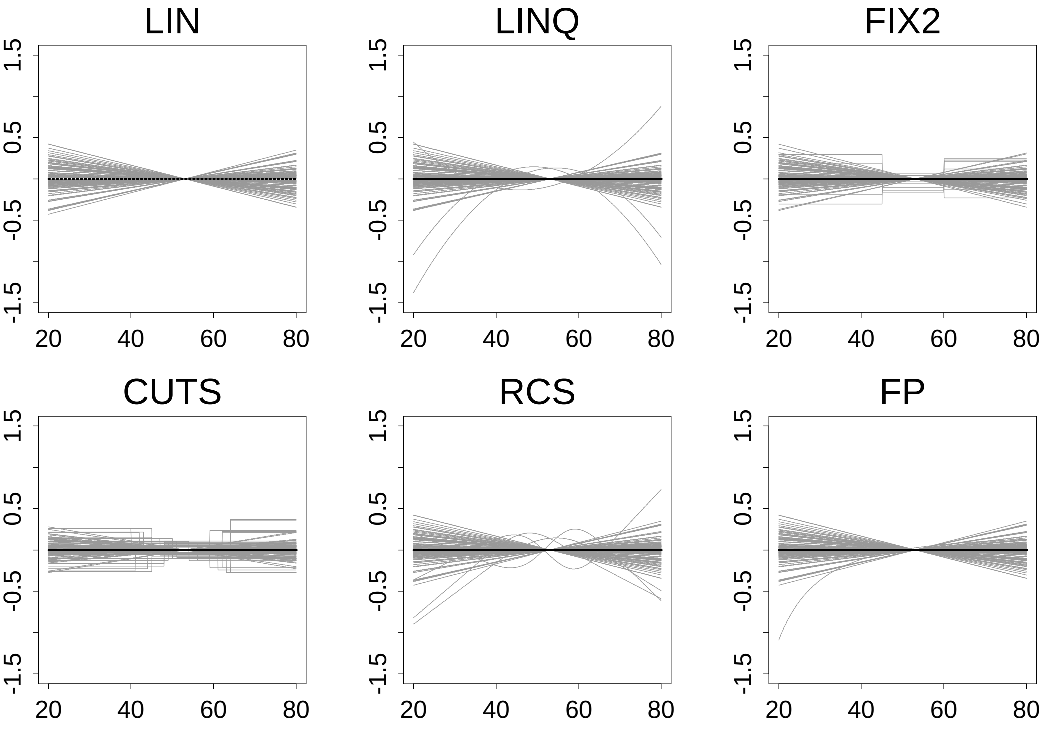


Figure 3.33: Estimated risk functions for in 100 random reallocations of the data of the GBSG-2 study

However, although the curvature of the bagged risk function (solid lines in figure 3.34) is less extreme as compared to that estimated in the selected repetition (dashed lines in figure 3.34), and therefore $\hat{h}_{\text{bagg}}$ has a smaller deviation from the given null risk, the functional shape is still similar. In general the variability of the risk functions should be taken into account when interpreting $\hat{h}_{\text{bagg}}$. This is of course also true for $\hat{h}$. Pointwise confidence bounds for $\hat{h}$ and/or $\hat{h}_{\text{bagg}}$ may be constructed based on the risk functions obtained in the original data set. Since the bagged risk function is based on the model selection strategy M the corresponding confidence bounds should be based on all (e.g. the linear and the restricted cubic splines for RCS) risk functions estimated in the bootstrap samples.

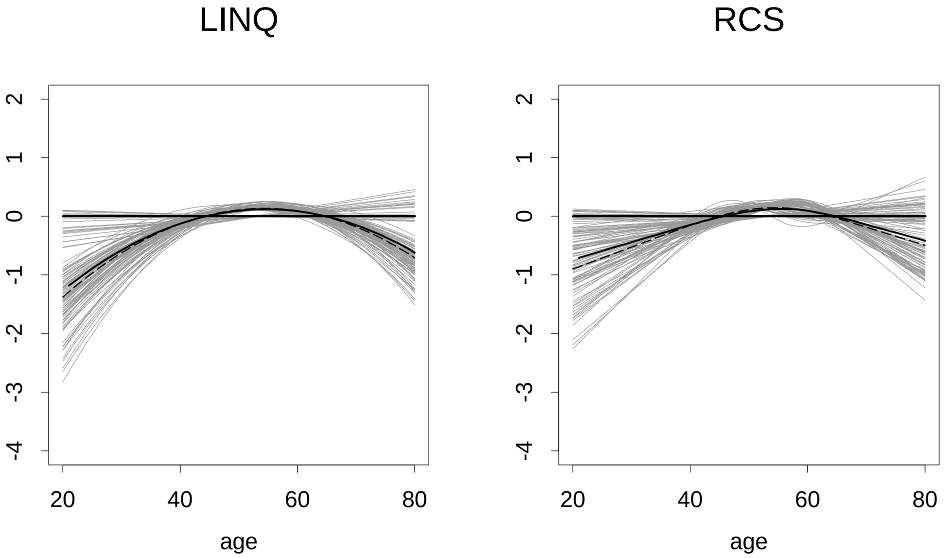


Figure 3.34: Estimated risk functions in 100 bootstrap samples of the selected data set in the random reallocation example, the black dashed line describes the risk function obtained selected data set, h_{bagg} is given by the black solid line

3.9 Summary of the simulation study

In order to investigate the capability of the standard procedures and the corresponding bagged risk functions to describe a given functional form adequately I performed a simulation study considering four different situations: the null model (0), the cutpoint model (I), a linear (II) and a V-type (III) risk function. Assessment of all procedures is based on the deviation between the estimated and the given risk function, where the latter is based on the given parameter β. The classical assumption of a linear risk function was taken as a reference. Furthermore, I used also the given model and estimated the underlying parameter in each data set. Using a linear risk function, for example, is the best we can do in the null model, whereas the categorization by using the given cutpoint (denoted as FIXALL) is the best choice for the given cutpoint model. The main results of the simulation study can be summarized as follows:

- The best fit and, therefore, the smallest errors are obtained when estimating the risk function by using the given functional form in all replications

- The linear risk function (LIN), which is the best for the null model and model II performs also quite good in the cutpoint model. However, this risk function is not suitable for model III.

- Estimated type I error rates $\hat{p}_{err}$ depend on the model selection strategy. Using the model of the highest order, the given type I error is held by LIN, LINQ FIX, RCS and FP whereas the categorization based on data driven cutpoints (CUT, CART) lead to a drastic inflation of the type I error. Using corrected P-values (CUTC, CARTC) the estimate $\hat{p}_{err}$ is close to the given value. Selecting the higher order model if it leads to a significant improvement in terms of likelihood and using the linear risk function otherwise (strategy M) the given α is obtained for FP only.

- For models I, II and III the power to detect the given effect was rather small. However the log relative risk was only moderate in the simulation study

- Using the model selection strategy M LINQ, RCS and FP reduce often to the linear risk function, even for model III. The smallest rate of nonlinear risk functions is obtained for FP (e.g 1.1% for model I). Due to the underlying optimization process I obtained the highest rate of nonlinear risk function for CUT and CART corresponding to the large type I error rate when considering the null model

- Besides TRANS the shape of the V-type risk function given by model III is described best by the nonlinear risk functions obtained by LINQ and FP. However, the change in risk is overestimated as a result of the model selection process and, therefore, quantitative errors are large.

- The qualitative error proposed in section 3.3 is sensible for continuous and monotone risk functions only. Thus, assessment of results was mainly based on quantitative errors

- Bagging is capable to reduce errors, but cannot be recommended in all situations. For models 0, I and II the error of the bagged risk function is similar or even larger in replications with a linear risk function. An improvement is mainly obtained in replications with a nonlinear risk functions in the original data set. This is caused by the good fit of the linear risk function, which is the best choice for models 0 and II and a good approximation for model I. In contrast, for model III bagging also leads to an improvement in replications, in which the higher order model reduced to the linear risk function in the original data set. Note, that the V-type model cannot be estimated correctly by LIN

- Generally bagging can work only, if the selected risk function differ between bootstrap samples. Note that there was no effect at all for LIN, whereas the highest error reduction is obtained for CUT/CUTS.

- Comparing the two strategies to make risk functions comparable we obtained rather small differences between error estimates (cf. also appendix A). However, larger differences may be observed when the risk set used to estimate the baseline hazard gets too small and/or the risk function shows an extremely nonlinear change in in risk with increasing values of the continuous covariates of X. The effect of censoring was not investigated so far

- Using the selection of a data driven cutpoint as a prototype of model building it was shown that the log relative risk between the resulting subgroups is overestimated and that confidence intervals calculated with the model based variances are too small. These deficiencies, that are caused by the extensive process of model building when selecting data driven cutpoints, can be corrected by applying shrinkage methods and by using the results of the bootstrap samples to estimate the variance. The coverage of the corrected confidence intervals corresponds well to the results obtained by using corrected P-values.

- The application of shrinkage methods for CUT and CART led to an error reduction in nearly all replications. However, the parameterwise shrinkage factors used for CART were negative or took values that were substantially larger than 1 in several replications. Therefore, there was no shrinkage effect at all in these replication. In many bootstrap samples parameterwise shrinkage factors could not be calculated at all due to missing convergence.

- Generally, the results obtained when simulating the null hypothesis from the data confirms the results of the null model.

4 Discussion

In this thesis I investigated several methods to estimate the functional form of the effect of a continuous covariate on survival time. To determine the so called risk function of a continuous covariate in the proportional hazards model, data-dependent as well as data-independent methods were considered. The former approaches are often based on an extensive process of model building, whereas the general functional form is prespecified when using the latter. Out of the class of data dependent methods I investigated the categorization of the continuous covariate by one or more data-driven cutpoints, and the use of fractional polynomials. For the categorization by data-driven cutpoints corrected P-values and shrinkage methods were used to correct for over-optimism of the selected model. A linear risk function, assuming a linear and a quadratic effect, the categorization by prespecified cutpoints, and the fit of a restricted cubic spline with fixed knots were taken as data independent methods. For each method the classical assumption of a linear risk function was taken as some kind of basic assumption, i.e. the more complex nonlinear risk function (e.g. a fractional polynomial) was used only if it improved the fit as compared to the linear effect. In so far, there was also one step of model building involved for the data independent methods. All methods were extended by adopting an approach called bootstrap aggregating (bagging) that has been proposed by Breiman (1996) for improvement of the error rate of predictors. Bootstrap samples were generated from the original data. In each bootstrap sample the risk function was estimated by using the same methods as in the original data. A so called bagged risk function was estimated as average over the corresponding risk functions of all bootstrap samples. Furthermore, the results of the bootstrap samples were used to investigate the stability of the model selection process.

Applying the methods to estimate the effect of age on event-free survival of breast cancer patients we observed a clear nonlinear effect: Up to 50 years there was a decrease in risk with increasing age, whereas the risk seemed to be rather constant for patients older than 50 years. However, the change in the observed risk differed between methods and the results could only partially be verified in the second, smaller Freiburg DNA study. Except for the categorization by data driven cutpoints there was hardly any difference between the bagged risk functions and the corresponding risk functions estimated in the original data. For most methods the estimated risk functions are similar in nearly all bootstrap samples confirming a strong nonlinear effect of age.

To assess the capability of all methods to describe a given functional form correctly I performed a simulation study considering 4 different models. Except for the null model assuming no effect at all the change in risk was assumed to be rather moderate. This assumption corresponds well to practical situations, since a large effect of a (new) prognostic factor on survival cannot be expected in most circumstances. Naturally, the best results are obtained when using the given assumption to estimate the risk function. However, analyzing real data the given functional form is of course unknown. Taking all models into account the method based on fractional polynomials performed best: If the given risk function is linear FP reduced to LIN in nearly all replications, whereas this rate was lower for the other methods. Although the functional shape of the V-type model is described correctly by the nonlinear risk functions of FP and LINQ the change in risk is rather overestimated. For the simulated cutpoint model the more complex risk functions of LINQ, FP and RCS reduced to LIN in nearly all replications indicating that it is rather difficult to fit a cutpoint model by continuous functions. However, the linear risk function serves as good approximation to the cutpoint model. The worst fit of all simulated models is observed for CART.

The application of bootstrap aggregating can improve the fit, but cannot be recommended in all situations. Bagging was efficient only if the estimated risk function is nonlinear and/or the given risk function is nonlinear. However, as illustrated in section 3.8.2 for CUTS bagging may also increase the error if the risk function estimate in the original data corresponds well to the given function. All in all the positive effect of bagging observed in the simulation study is not as large as we have expected in advance. Generally, the bagged risk function should be considered together with the results of all bootstrap samples. Doing so, we get insight into the variability between bootstrap samples and, therefore, also into the stability of the risk function estimated in the original data.

The fact, that bagging works only if the selected functional form differ between bootstrap samples agree with the statement that the *'vital element'* [of bagging] *'is the instability of the prediction method'* (Breiman, 1996). In a forthcoming paper Bühlmann and Yu (2000) formalize the notation of instability. Furthermore, they derive theoretical results to explain a variance reduction effect of bagging for so called *'hard decision problems'* like classification and regression trees or variable selection in regression models.

The impact of model building on the estimation of risk functions was especially illustrated in sections 3.8.3 and 3.8.4. For instance, it was demonstrated that parameters of risk functions tend to be overestimated when considering the selected nonlinear risk function only. Methods used to correct for errors caused by model building were investigated in the data-driven cutpoint model only. Although the application of shrinkage and bootstrap methods lead to more reliable results, the selected data-driven cutpoint may differ substantially between studies. Therefore, this approach should be applied only if the selected cutpoints are similar in most bootstrap samples and if the risk function do not disagree to other continuous risk functions. In any case results should be based on corrected P-values and shrinked estimates. Unfortunately, there are still publications in the medical literature using data-driven cutpoints without any correction, although the dangers of using this approach were discussed extensively in the statistical and medical literature (Lausen and Schumacher, 1992; Altman et al., 1994; Lausen and Schumacher, 1996; Schumacher et al., 1997; Altman, 1998; Holländer and Schumacher, 2001). Just to give one example from the huge literature on prognostic factors in breast cancer, in a recent paper by Linderholm et al. (2000) the minimum P-value approach was used to find a data-driven cutpoint for vascular endothetical growth factor (VEGF). As main result the authors found a significant difference of the resulting subgroups with respect to overall survival.

What did we learn on the estimation of a risk function?

It is not possible to fix definitive rules for future data analysis. None of the methods can be recommended in all situations. This would also be true for other approaches not investigated in this thesis. Out of the methods considered here the CART based categorization should not be used. Generally, one should not focus on one specific method and neglect all the other. Bootstrap methods and other resampling methods can help to overcome some problems caused by model building, at least it can be used to investigate for stability of the model selection process. Based on the results of the simulation and the application to the data I would recommend the following strategy:

– Investigate always for potential nonlinearity of the risk function, use several method to do so

– If all methods reduce to the linear risk function when applying an appropriate model selection strategy, assume a linear effect

– If one or more methods suggest a strong nonlinear effect use the same model building process as in the original data to estimate the risk functions in a set of bootstrap samples

a) If the risk functions corresponds well to that obtained in the original data in nearly all bootstrap samples, and results are similar for different methods, there is a strong evidence for a nonlinear effect. The bagged risk function would be nearly equal to the original estimate and, therefore, bagging is not necessary

b) If the risk functions reduce to the linear effect in several bootstrap samples the nonlinear effect is not as strong as seen in the original data. The bagged risk function may be used to correct for over-optimism, but should be considered together with the results of all bootstrap samples

– Consider always the distribution of the underlying covariate when interpreting results

– Investigate also for potential time-dependency of the covariate

– Validate all results by using one or more independent data sets

In practical situations we should of course not restrict to univariate analyses. In a recent paper Sauerbrei and Royston (1999) proposed a model selection strategy for multivariable prognostic and diagnostic models. Their approach is based on the transformation of the covariates by using fractional polynomials and incorporates also medical knowledge into the analysis.

For the assessment of all approaches I considered the deviation between the estimated and the given risk function. On the one hand the risk function of each covariate can be interpreted itself. On the other hand the answer to the question *Is there an effect of a specific covariate on survival?* can also depend on the functional form. Remember, that in the GBSG-2 study age showed no effect on event-free survival when assuming a linear effect. In contrast the effect of age was highly significant when using a restricted cubic spline. This question is not limited to the analysis of survival data, but has also relevance for other situations.

However, analyzing survival data it should be taken into account that the effect of covariates can change over time. Therefore, it would be worthwhile to model both, the functional form and the effect over time, by using one risk function. As I did in my work all approaches cited in section 2.1.11 referred to one of these problems only.

Furthermore, I have not investigated how far the use of appropriate risk functions can improve the prediction. It should be noted that survival of an individual patient cannot adequately be predicted (Henderson, 1995). Actually, there seems to be no sound and commonly agreed statistical methodology to assess the accuracy of predictions derived from a survival model. An outline on some recent developments is given in Schumacher et al. (2001), more details can be found elsewhere Graf et al. (1999).

All in all there remain a lot of questions calling for further research. The results of this thesis can be useful for some additional investigations.

5 Appendix A: Tables of results

The main results obtained in the breast cancer studies and the simulation study were illustrated graphically in sections 2 and 3. This appendix lists additional results: Table A 2.1 describes the estimated risk functions in the GBSG-2 study and the Freiburg DNA study displayed in figure 2.1. Tables A 3.1 - A 3.16 describe results of the simulation study. I have already referred to all tables when interpreting the results in the text.

For each simulated model of the simulation study the following tables are given:

R1 : Estimated MAE obtained in the original data set and for the bagged risk functions
R2 : Estimated MSE obtained in the original data set and for the bagged risk functions
R3 : The effect of bagging on the estimated MAE and MSE
R4: The comparison of the procedures used to make results comparable

Furthermore, all tables contain information on the results of model selection. A detailed description is given in the heading of each table.

Overview:

simulated model	R1	R2	R3	R4
null model (model 0)	table A 3.1	table A 3.2	table A 3.3	table A 3.4
cutpoint model (model I)	table A 3.5	table A 3.6	table A 3.7	table A 3.8
linear effect (model II)	table A 3.9	table A 3.10	table A 3.11	table A 3.12
V-type effect (model III)	table A 3.13	table A 3.14	table A 3.15	table A 3.16

Table A 2.1: Estimated functions of the effect of age with respect to event-free survival time: Results obtained by using the standard procedures in the GBSG-2 study and validation in the Freiburg DNA study

functional	GBSG-2 study			Freiburg DNA study	
relationship	DF[1]	LRT[2]	estimated risk function	LRT[2]	estimated risk function
LIN	1	0.58	$-0.0048 \cdot \text{age}$	0.71	$-0.0080 \cdot \text{age}$
LINQ	2	9.00	$-0.1395 \cdot \text{age} + 0.0013 \cdot \text{age}^2$	0.73	$-0.0185 \cdot \text{age} + 0.0001 \cdot \text{age}^2$
FIX2	2	3.83	$-0.2866 \cdot 1_{\{45 < \text{age} \le 60\}}$ $-0.2032 \cdot 1_{\{\text{age} > 60\}}$	1.01	$-0.0891 \cdot 1_{\{45 < \text{age} \le 60\}}$ $-0.2890 \cdot 1_{\{\text{age} > 60\}}$
CUT CUTS	1	7.79	$-0.4441 \cdot 1_{\{\text{age} > 42\}}$ $-0.3942 \cdot 1_{\{\text{age} > 42\}}$	0.19	$-0.159 \cdot 1_{\{\text{age} > 42\}}$
CART	3	23.40	$-0.8022 \cdot 1_{\{32 < \text{age} \le 42\}}$ $-1.4697 \cdot 1_{\{42 < \text{age} \le 49\}}$ $-0.9879 \cdot 1_{\{\text{age} > 49\}}$	0.19	$-0.299 \cdot 1_{\{32 < \text{age} \le 42\}}$ $-0.409 \cdot 1_{\{42 < \text{age} \le 49\}}$ $-0.437 \cdot 1_{\{\text{age} > 49\}}$
FP	4	17.62	$1.527 \cdot (\text{age} / 50)^{-0.5}$ $-6.38237 \cdot (\text{age}/50)^{-2}$	1.38	$2.18 \cdot (\text{age} / 50)^{-0.5}$ $-2.207 \cdot (\text{age}/50)^{-2}$
RCS	3	21.69	$-0.09315 * \text{age}$ $+0.00024 * \max(\text{age} -6,30)^3$ $-0.00076 * \max(\text{age} - 48, 0)^3$ $+0.00075 * \max(\text{age} - 58, 0)^3$ $-0.00023 * \max(\text{age} - 68, 0)^3$	2.17	$-0.05582 * age$ $+0.00014 * \max(age -6,30)^3$ $-0.00044 * \max(age - 48, 0)^3$ $+0.00044 * \max(age - 58, 0)^3$ $-0.00014 * \max(age - 68, 0)^3$

[1] degrees of freedom = number of parameters

[2] value of likelihood-ratio test statistic testing the model against the null model

Table A 3.1: Empirical distribution in terms of quantiles of the $\widehat{MAE}$ for the standard and bagged risk functions in the simulated **null model** (standardization to zero mean log relative risk) and results of model selection

original data

risk function	Empirical quantile based on **all replications (R=1000)**									no. replications with linear/nonlinear risk function	Empirical quantile based on **replications with nonlinear risk function only**								
	min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.		min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.
LIN	0	0.0053	0.0118	0.0284	0.0594	0.1027	0.1462	0.1744	0.372	1000/0									
LINQ	0	0.0054	0.0121	0.0301	0.0619	0.1103	0.1694	0.2124	0.372	950/50	0.1682	0.176	0.1797	0.1944	0.2207	0.2549	0.2819	0.3088	0.3263
FIX	0	0.0053	0.0118	0.0284	0.0594	0.1042	0.1536	0.2134	0.4191	944/56	0.2021	0.2048	0.2068	0.2197	0.2407	0.2782	0.3103	0.3514	0.4191
CUT	0	0.0062	0.0133	0.0338	0.0844	0.2003	0.2449	0.274	0.4664	568/432	0.1227	0.14	0.1535	0.1775	0.2086	0.2408	0.2888	0.3198	0.4664
CUTS	0	0.0062	0.0133	0.0338	0.0842	0.1604	0.209	0.2436	0.4405	568/432	0.0765	0.1093	0.1183	0.1416	0.1683	0.2055	0.2525	0.2853	0.4405
CUTC	0	0.0053	0.012	0.0289	0.0601	0.1051	0.1617	0.2346	0.4664	933/67	0.1856	0.2068	0.2127	0.2355	0.2657	0.3233	0.3466	0.3938	0.4664
CUTCS	0	0.0053	0.012	0.0289	0.0601	0.1051	0.1617	0.2179	0.4405	933/67	0.166	0.1844	0.1905	0.2103	0.2433	0.2941	0.3221	0.3661	0.4405
CART	0	0.0062	0.0133	0.0338	0.0844	0.2331	0.313	0.3602	0.5758	568/432	0.1161	0.1472	0.1606	0.1979	0.249	0.3081	0.3738	0.4064	0.5758
CARTS	0	0.006	0.0133	0.0337	0.0808	0.1885	0.2821	0.3229	1.0939	568/432	0	0.0975	0.1168	0.1537	0.2019	0.2698	0.3299	0.3834	1.0939
CARTC	0	0.0053	0.012	0.0289	0.0601	0.1051	0.1617	0.2328	0.5758	933/67	0.1912	0.2033	0.2162	0.2294	0.2637	0.327	0.3822	0.4432	0.5758
CARTCS	0	0.0053	0.012	0.0289	0.0601	0.1051	0.1617	0.2096	0.5061	933/67	0.163	0.1788	0.1879	0.2045	0.2429	0.2947	0.3492	0.4085	0.5061
RCS	0	0.0053	0.0121	0.0302	0.0626	0.1131	0.1887	0.2467	0.3724	937/63	0.2055	0.2148	0.2239	0.2369	0.2581	0.2804	0.3063	0.3318	0.3724
FP	0	0.0053	0.012	0.029	0.0601	0.1043	0.1514	0.1866	0.372	990/10	0.2406	0.2455	0.2505	0.2674	0.2797	0.2844	0.3219	0.3268	0.3317

bagged risk functions

risk function	Empirical quantile based on **all replications (R=1000)**									no. h linear/ nonlinear	no. **linear** in **all** bootstrap samples	Empirical quantile based on **replications with nonlinear risk function only**								
	min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.			min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.
LIN	0.0001	0.0063	0.0131	0.0297	0.0612	0.1054	0.1488	0.1843	0.3658	1000/0	1000									
LINQ	0.0039	0.0161	0.0229	0.0453	0.0776	0.124	0.1663	0.2063	0.3658	950/50	5	0.1075	0.1296	0.1323	0.1489	0.1925	0.2257	0.2649	0.3016	0.3072
FIX	0.001	0.0085	0.015	0.0325	0.0668	0.1179	0.1649	0.2039	0.4308	944/56	1	0.1462	0.171	0.179	0.2057	0.2388	0.2744	0.2964	0.3545	0.4308
CUT	0.0063	0.0161	0.02	0.0302	0.0489	0.0807	0.116	0.145	0.3026	568/432	0	0.0133	0.0283	0.0337	0.0507	0.0841	0.1145	0.15	0.1855	0.3026
CUTS	0.0057	0.0135	0.0168	0.0261	0.0422	0.0712	0.1036	0.1292	0.2876	568/432	0	0.0113	0.0245	0.0292	0.0443	0.0738	0.101	0.1363	0.1718	0.2876
CUTC	0.0012	0.0048	0.0061	0.0101	0.0179	0.0357	0.0695	0.0944	0.2887	933/67	0	0.0129	0.0196	0.0224	0.0401	0.0793	0.1463	0.1831	0.2073	0.2887
CUTCS	0.0012	0.0042	0.0057	0.009	0.0153	0.0294	0.0591	0.0821	0.271	933/67	0	0.0084	0.012	0.0137	0.0283	0.0634	0.1326	0.1654	0.1862	0.271
CART	0.0437	0.0656	0.0825	0.1128	0.1548	0.2081	0.2632	0.3013	0.5132	568/432	0	0.1032	0.1405	0.1503	0.1716	0.2108	0.2563	0.308	0.3326	0.5132
RCS	0.003	0.0231	0.0341	0.0567	0.0906	0.138	0.1916	0.2241	0.3756	937/63	1	0.1361	0.1655	0.1734	0.1943	0.2118	0.2516	0.2963	0.3185	0.3666
FP	0.0008	0.0119	0.018	0.0366	0.0691	0.1141	0.1595	0.1963	0.3658	990/10	100	0.1916	0.2031	0.2145	0.2231	0.2329	0.2762	0.281	0.288	0.295

Table A 3.2: Empirical distribution in terms of quantiles of the $\widehat{MSE}$ for the standard and bagged risk functions in the simulated **null model** (standardisation to zero mean log relative risk) and results of model selection

original data

risk function	Empirical quantile based on **all replications (R=1000)**									no. replications with linear/nonlinear risk function	Empirical quantile based on **replications with nonlinear risk function only**								
	min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.		min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.
LIN	0	0	0.0002	0.0011	0.0048	0.0141	0.029	0.0407	0.1741	1000/0									
LINQ	0	0	0.0002	0.0012	0.0052	0.0165	0.0396	0.0612	0.1741	950/50	0.0386	0.0435	0.0469	0.053	0.0678	0.0942	0.1272	0.1427	0.1486
FIX	0	0	0.0002	0.0011	0.0048	0.0143	0.0311	0.0494	0.1757	944/56	0.0409	0.0421	0.0435	0.0488	0.0583	0.0782	0.0986	0.1243	0.1757
CUT	0	0.0001	0.0002	0.0015	0.0097	0.0555	0.0853	0.1071	0.2248	568/432	0.0285	0.0413	0.0428	0.0483	0.0585	0.0826	0.1094	0.1289	0.2248
CUTS	0	0.0001	0.0002	0.0015	0.0097	0.0356	0.0646	0.0859	0.2006	568/432	0.0109	0.023	0.0245	0.029	0.0392	0.0613	0.0886	0.1059	0.2006
CUTC	0	0	0.0002	0.0011	0.0049	0.0148	0.0359	0.1071	0.2248	933/67	0.0944	0.0972	0.0997	0.1071	0.1155	0.1391	0.1828	0.1944	0.2248
CUTCS	0	0	0.0002	0.0011	0.0049	0.0148	0.0359	0.0859	0.2006	933/67	0.0743	0.077	0.0798	0.0859	0.095	0.1162	0.1598	0.1723	0.2006
CART	0	0.0001	0.0002	0.0015	0.0097	0.0996	0.1655	0.2056	0.4468	568/432	0.0285	0.044	0.0488	0.0675	0.1125	0.1615	0.2118	0.2568	0.4468
CARTS	0	0	0.0002	0.0015	0.0093	0.0639	0.1228	0.1673	1.2464	568/432	0	0.0189	0.0256	0.0411	0.0761	0.1175	0.178	0.2435	1.2464
CARTC	0	0	0.0002	0.0011	0.0049	0.0148	0.0359	0.1071	0.4468	933/67	0.0944	0.0983	0.1001	0.1072	0.1161	0.1394	0.1963	0.2643	0.4468
CARTCS	0	0	0.0002	0.0011	0.0049	0.0148	0.0359	0.0803	0.3491	933/67	0.0587	0.0754	0.0771	0.0805	0.0947	0.1148	0.1721	0.1953	0.3491
RCS	0	0	0.0002	0.0012	0.0052	0.0173	0.0471	0.0845	0.1931	937/63	0.0577	0.0681	0.0699	0.0816	0.0953	0.1141	0.1447	0.1697	0.1931
FP	0	0	0.0002	0.0011	0.0049	0.0143	0.0303	0.0469	0.1741	990/10	0.0817	0.0894	0.0971	0.1006	0.1134	0.1197	0.1423	0.1476	0.1529

bagged risk functions

risk function	Empirical quantile based on **all replications (R=1000)**									no. h linear/ nonlinear	no. **linear** in **all** bootstrap samples	Empirical quantile based on **replications with nonlinear risk function only**								
	min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.			min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.
LIN	0	0.0001	0.0002	0.0011	0.005	0.0148	0.0296	0.0446	0.1684	1000/0	1000									
LINQ	0	0.0004	0.0008	0.0028	0.0083	0.0214	0.0388	0.0583	0.1684	950/50	5	0.017	0.0248	0.0259	0.0323	0.051	0.0749	0.1142	0.128	0.1425
FIX	0	0.0001	0.0003	0.0013	0.0053	0.0158	0.0297	0.0442	0.1856	944/56	1	0.0227	0.0295	0.0324	0.0438	0.0584	0.0759	0.0885	0.1265	0.1856
CUT	0.0001	0.0004	0.0006	0.0013	0.0036	0.0088	0.0191	0.0297	0.1274	568/432	0	0.0003	0.0012	0.0017	0.0039	0.0089	0.018	0.0344	0.0486	0.1274
CUTS	0	0.0003	0.0004	0.001	0.0027	0.0067	0.0153	0.0242	0.1151	568/432	0	0.0002	0.0009	0.0013	0.003	0.0071	0.014	0.0281	0.042	0.1151
CUTC	0	0	0.0001	0.0002	0.0005	0.002	0.0067	0.0137	0.116	933/67	0	0.0004	0.0009	0.0012	0.003	0.0079	0.038	0.053	0.0667	0.116
CUTCS	0	0	0.0001	0.0001	0.0004	0.0014	0.0049	0.0112	0.1022	933/67	0	0.0001	0.0003	0.0005	0.0013	0.0051	0.0312	0.0437	0.0559	0.1022
CART	0.003	0.0071	0.0108	0.0194	0.0361	0.0652	0.1009	0.129	0.342	568/432	0	0.023	0.0312	0.0355	0.0468	0.0675	0.0961	0.1308	0.1538	0.342
RCS	0	0.0008	0.0019	0.0048	0.0119	0.0265	0.0501	0.0689	0.207	937/63	1	0.0255	0.0385	0.041	0.0524	0.0672	0.0882	0.1283	0.1616	0.207
FP	0	0.0002	0.0005	0.0019	0.0067	0.0179	0.0354	0.0515	0.1684	990/10	100	0.0481	0.057	0.0659	0.0733	0.0813	0.1118	0.1173	0.1184	0.1194

Table A 3.3: The effect of bagging on the estimated error in the simulated **null model**:
Comparing the $\widehat{MAE}/\widehat{MSE}$ obtained in the original data to the corresponding errors of the bagged risk function

risk	linear	nonlinear	all (R=1000)					linear					nonlinear				
function			number (percent)		ratio bagging/original			number (percent)		ratio bagging/original			number (percent)		ratio bagging/original		
			of smaller $\widehat{MAE}$ for		quantiles			of smaller $\widehat{MAE}$ for		quantiles			of smaller $\widehat{MAE}$ for		quantiles		
			$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75	$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75	$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75
LIN	1000	0	441(44.1)	559(55.9)	91	102	115	( )	( )				( )	( )			
LINQ	950	50	285(28.5)	715(71.5)	98	110	143	238(25.1)	712(74.9)	100	112	147	47 (94)	3 (6)	76	86	93
FIX	944	56	355(35.5)	645(64.5)	94	108	131	318(33.7)	626(66.3)	94	110	133	3 7 (66.1)	19 (33.9)	92	98	101
CUT	568	432	744(74.4)	256(25.6)	38	58	101	312(54.9)	256(45.1)	59	91	181	43 2(100)	0 (0)	25	39	52
CUTS	568	432	787(78.7)	213(21.3)	37	57	89	356(62.7)	212(37.3)	51	78	156	43 1(99.8)	1 (0.2)	27	43	57
CUTC	933	67	829(82.9)	171(17.1)	18	34	67	762(81.7)	171(18.3)	18	34	70	67 (100)	0 (0)	18	30	52
CUTCS	933	67	851(85.1)	149(14.9)	16	28	61	784(84)	149(16)	16	28	64	67 (100)	0 (0)	13	28	53
CART	568	432	335(33.5)	665(66.5)	87	146	347	10 (1.8)	558(98.2)	176	299	635	325 (75.2)	107(24.8)	73	84	100
RCS	937	63	242(24.2)	758(75.8)	101	121	185	180(19.2)	757(80.8)	104	125	198	62 (98.4)	1 (1.6)	79	84	91
FP	990	10	323(32.3)	677(67.7)	96	107	132	315(31.8)	675(68.2)	96	107	132	8 (80)	2 (20)	82	85	90

risk	linear	nonlinear	all (R=1000)					linear					linear				
function			number (percent)		ratio bagging/original			number (percent)		ratio bagging/original			number (percent)		ratio bagging/original		
			of smaller $\widehat{MSE}$ for		quantiles			of smaller $\widehat{MSE}$ for		quantiles			of smaller $\widehat{MSE}$ for		quantiles		
			$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75	$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75	$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75
LIN	1000	0	441(44.1)	559(55.9)	82	103	132	( )	( )				( )	( )			
LINQ	950	50	266(26.6)	734(73.4)	98	126	223	218(22.9)	732(77.1)	103	130	240	48 (96)	2 (4)	59	75	87
FIX	944	56	453(45.3)	547(54.7)	81	104	148	417(44.2)	527(55.8)	81	105	152	3 6 (64.3)	20(35.7)	84	97	103
CUT	568	432	722(72.2)	278(27.8)	15	36	118	290(51.1)	278(48.9)	38	93	411	43 2(100)	0 (0)	6	15	29
CUTS	568	432	770(77)	230(23)	15	35	86	338(59.5)	230(40.5)	27	67	299	432 (100)	0 (0)	8	18	35
CUTC	933	67	819(81.9)	181(18.1)	4	12	49	752(80.6)	181(19.4)	4	13	57	67 (100)	0 (0)	2	7	26
CUTCS	933	67	837(83.7)	163(16.3)	3	9	40	770(82.5)	163(17.5)	3	9	43	67 (100)	0 (0)	1	6	27
CART	568	432	373(37.3)	627(62.7)	76	223	1490	6 (1.1)	562(98.9)	366	1015	4811	367 (85)	65(15)	47	69	89
RCS	937	63	184(18.4)	816(81.6)	108	154	392	122(13)	815(87)	113	167	420	62 (98.4)	1 (1.6)	61	73	87
FP	990	10	307(30.7)	693(69.3)	94	117	183	298(30.1)	692(69.9)	95	118	185	9 (90)	1 (10)	67	75	80

Table A 3.4: Empirical quantiles of the ratios 100^{*} ($\widehat{MSE}_{\log(\lambda_0(t^{*}))}/\widehat{MSE}_{\text{zero mean}}$) and 100^{*} ($\widehat{MAE}_{\log(\lambda_0(t^{*}))}/\widehat{MAE}_{\text{zero mean}}$) for the simulated **null model**

risk function		Empirical quantiles of the ratio for $\widehat{\textbf{MSE}}$ (R=1000)							number of	Empirical quantiles (**only nonlinear replications**)						
	$\hat{S}(t*)$	min.	0.05	0.25	0.5	0.75	0.95	max.	replications	min.	0.05	0.25	0.5	0.75	0.95	max.
LIN	0.50	100	100	100.1	100.3	100.7	102.1	105.5	0							
	0.75	100	100	100	100.1	100.4	101.2	103.3	0							
	0.90	100	100	100	100.1	100.3	100.9	103	0							
LINQ	0.50	100	100	100.1	100.3	100.7	102.3	105.5	50	100	100	100.1	100.3	101.3	103.3	104.6
	0.75	100	100	100	100.1	100.4	101.2	104.2	50	100	100	100.2	100.4	101.2	102.5	104.2
	0.90	100	100	100	100.1	100.4	101.3	103.6	50	100.1	100.2	100.8	101.3	102.1	102.9	103.6
FIX	0.5	100	100	100.1	100.3	100.7	102.1	105.5	56	100	100	100.1	100.3	100.8	102	103.6
	0.75	100	100	100	100.1	100.4	101.2	103.3	56	100	100	100.1	100.4	100.9	101.9	102.6
	0.90	100	100	100	100.1	100.4	101.1	103.1	56	100	100.2	100.5	101.2	101.6	102.6	103.1
CUT	0.50	100	100	100.1	100.3	100.8	102.6	112.4	432	100	100	100.1	100.4	101.1	103.4	112.4
	0.75	100	100	100.1	100.2	100.6	101.7	105.5	432	100	100	100.1	100.4	101	102.3	105.5
	0.90	100	100	100	100.3	100.9	102.7	111.3	432	100	100.2	100.6	101.1	101.9	103.4	111.3
CART	0.50	100	100	100.1	100.4	101	103.7	113.9	432	100	100	100.2	100.6	101.8	104.9	113.9
	0.75	100	100	100.1	100.2	100.7	102.3	105.8	432	100	100	100.2	100.6	101.5	103.3	105.8
	0.90	100	100	100.1	100.3	101.7	105.4	118.4	432	100	100.3	101	102	103.7	106.9	118.4
RCS	0.50	100	100	100.1	100.3	100.7	102.1	105.5	63	100	100	100.1	100.4	100.9	102.6	104.1
	0.75	100	100	100	100.1	100.4	101.2	103.3	63	100	100	100.2	100.5	101.2	102.5	104.1
	0.90	100	100	100	100.1	100.4	101	104.7	63	100.2	100.5	101.1	101.6	102.4	103.5	106.9
FP	0.50	100	100	100.1	100.3	100.7	102.1	105.5	10	100	100	100.2	100.8	101.2	104.3	104.5
	0.75	100	100	100	100.1	100.4	101.2	103.3	10	100	100	100	100.3	100.6	101.8	102.3
	0.90	100	100	100	100.1	100.4	101	104.7	10	100.3	100.4	100.8	101.5	103	104.2	104.7

risk function		Empirical quantiles of the ratio for $\widehat{\textbf{MAE}}$ (R=1000)							number of	Empirical quantiles (**only nonlinear replications**)						
	$\hat{S}(t*)$	min.	0.05	0.25	0.5	0.75	0.95	max.	replications	min.	0.05	0.25	0.5	0.75	0.95	max.
LIN	0.50	98.6	99.5	99.9	100.1	100.4	101.2	103.1	0							
	0.75	98.8	99.6	99.9	100	100.2	100.8	102.9	0							
	0.90	98.9	99.6	99.9	100	100.2	100.7	102.4	0							
LINQ	0.50	95.6	99.4	99.9	100.1	100.4	101.3	104.4	50	95.6	97	98.3	99.5	100.3	103.3	104.4
	0.75	96.8	99.5	99.9	100	100.3	101	110.7	50	96.8	98.1	99.2	100.7	102.3	104.4	110.7
	0.90	96.8	99.5	99.9	100	100.2	101.1	111.5	50	96.8	97	99.2	102.2	103.9	106.3	111.5
FIX	0.5	96.8	99.5	99.9	100.1	100.3	101.2	103.1	56	96.8	98.9	99.7	100	100.3	100.9	101.2
	0.75	98.1	99.6	99.9	100	100.2	100.8	102.9	56	98.1	98.8	99.7	100	100.3	101.7	102
	0.90	97.7	99.5	99.9	100	100.2	100.7	102.5	56	97.7	98.3	99.4	100.1	100.9	102.1	102.5
CUT	0.50	55.6	88.9	99.4	100	100.4	103.7	114.7	432	55.6	81.6	94.5	99	101.1	106.9	114.7
	0.75	83.3	95.2	99.8	100.1	100.6	108.6	124.4	432	83.3	92.8	98.4	100.6	104.7	112.5	124.4
	0.90	86.3	93.6	99.9	100.1	101.8	114.5	141.9	432	86.3	90.6	97.9	102.9	109.4	119.9	141.9
CART	0.50	69.6	88.3	99.5	100	100.5	105.5	120.1	432	69.6	82.2	94	99	101.6	109.3	120.1
	0.75	85.1	95.1	99.8	100.1	100.8	111	134.6	432	85.1	92.6	98	101	106.1	115.3	134.6
	0.90	82.5	94	99.9	100.1	103.3	118.8	151.7	432	82.5	90.3	98.3	105.3	112.7	124.9	151.7
RCS	0.50	95.8	99.5	99.9	100.1	100.4	101.3	103.1	63	97.2	98	99.6	100	101	102.5	104.2
	0.75	98.8	99.6	99.9	100	100.2	100.8	104.4	63	97.5	98.6	99.8	100.2	101.3	102.6	109
	0.90	98.2	99.6	99.9	100	100.2	100.7	107.8	63	97	98.4	99.6	100.5	102.1	105.1	114.3
FP	0.50	95.8	99.5	99.9	100.1	100.4	101.3	103.1	10	95.8	97	98.7	99.1	99.6	101.3	102.3
	0.75	98.8	99.6	99.9	100	100.2	100.8	104.4	10	98.9	99	99.8	100.4	101.2	103.5	104.4
	0.90	98.2	99.6	99.9	100	100.2	100.7	107.8	10	98.2	98.3	101	102.1	103.8	106.9	107.8

Table A 3.5: Empirical distribution in terms of quantiles of the $\widehat{MAE}$ for the standard and bagged risk functions in the simulated **cutpoint model** (standardisation to zero mean log relative risk) and results of model selection

original data

risk function	Empirical quantile based on **all replications (R=1000)**									no. replications with linear/nonlinear risk function	Empirical quantile based on **replications with nonlinear risk function only**								
	min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.		min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.
LIN	0.0877	0.0978	0.1014	0.1078	0.1235	0.1535	0.1883	0.2165	0.4102	1000/0									
FIXALL	0.0003	0.0066	0.0129	0.0323	0.0673	0.1178	0.1673	0.2107	0.4856	0/1000									
LINQ	0.0877	0.0978	0.1016	0.1085	0.1252	0.1633	0.2045	0.2311	0.4102	944/56	0.1724	0.1839	0.1865	0.1946	0.2156	0.2425	0.2768	0.2909	0.3534
FIX	0.0003	0.0066	0.0129	0.0324	0.0932	0.1381	0.1901	0.2255	0.4856	328/672	0.0003	0.0048	0.009	0.0218	0.0464	0.0981	0.1603	0.209	0.4856
CUT	0.0016	0.029	0.042	0.0905	0.1582	0.223	0.2716	0.3013	0.4856	92/908	0.0016	0.0271	0.0396	0.0841	0.1513	0.2248	0.2736	0.3028	0.4856
CUTS	0.0003	0.0293	0.0453	0.0777	0.1429	0.2101	0.2585	0.286	0.4657	92/908	0.0003	0.0287	0.0418	0.073	0.1313	0.2079	0.2599	0.2876	0.4657
CUTC	0.0646	0.0947	0.1002	0.1099	0.1417	0.201	0.2585	0.2876	0.4856	551/449	0.0646	0.093	0.104	0.1372	0.1907	0.2482	0.2887	0.3145	0.4856
CUTCS	0.0307	0.0751	0.0933	0.1056	0.1306	0.1873	0.247	0.274	0.4657	551/449	0.0307	0.0597	0.0719	0.1084	0.1697	0.2257	0.2736	0.2952	0.4657
CART	0.0055	0.057	0.0959	0.154	0.219	0.2819	0.3381	0.372	0.5858	92/908	0.0055	0.0541	0.0917	0.1517	0.2264	0.2883	0.3406	0.3757	0.5858
CARTS	0.0004	0.0531	0.0845	0.1418	0.2016	0.2577	0.3145	0.3638	1.0129	92/908	0.0004	0.0497	0.0749	0.135	0.2039	0.261	0.3199	0.3671	1.0129
CARTC	0.0646	0.095	0.101	0.1102	0.1432	0.2074	0.2651	0.3007	0.5858	551/449	0.0646	0.0947	0.1076	0.1401	0.1985	0.2538	0.3013	0.324	0.5858
CARTCS	0.0307	0.0764	0.0939	0.1061	0.1323	0.1893	0.2558	0.2799	0.9335	551/449	0.0307	0.0597	0.0735	0.11	0.1752	0.2355	0.2829	0.3072	0.9335
RCS	0.0877	0.0981	0.102	0.109	0.1282	0.1724	0.219	0.2542	0.4393	888/112	0.1408	0.1657	0.1807	0.1961	0.2221	0.2575	0.2981	0.3086	0.4393
FP	0.0877	0.0978	0.1014	0.1078	0.1241	0.1549	0.192	0.2218	0.4102	989/11	0.1582	0.2	0.2418	0.2591	0.2665	0.2819	0.3099	0.3154	0.3208

bagged risk functions (first 100 replications only)

risk function	Empirical quantile based on **all replications (R=1000)**									no. h linear/ nonlinear	no. **linear** in **all** bootstrap samples	Empirical quantile based on **replications with nonlinear risk function only**								
	min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.			min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.
LIN	0.0911	0.0984	0.1007	0.1058	0.1199	0.1469	0.1816	0.2159	0.2672	100/0	100									
LINQ	0.0905	0.1	0.1021	0.1081	0.1248	0.1593	0.1957	0.2172	0.2701	91/9	0	0.1283	0.134	0.1398	0.1524	0.1809	0.2126	0.2245	0.2416	0.2586
FIX	0.0113	0.0189	0.0276	0.0466	0.076	0.1301	0.1813	0.2209	0.2563	32/68	0	0.0113	0.0134	0.0236	0.0344	0.0646	0.1272	0.1752	0.1985	0.2563
CUT	0.041	0.0566	0.0668	0.0984	0.132	0.1659	0.1973	0.2172	0.26	6/94	0	0.041	0.0548	0.0659	0.0931	0.1305	0.1562	0.1914	0.1992	0.2531
CUTS	0.0379	0.06	0.0791	0.1042	0.1412	0.1732	0.2026	0.2216	0.257	6/94	0	0.0379	0.0588	0.0757	0.1016	0.1358	0.1642	0.1924	0.2027	0.2537
CUTC	0.0462	0.0642	0.0783	0.119	0.1684	0.215	0.2415	0.2452	0.2601	52/48	0	0.0462	0.0543	0.0637	0.0892	0.1185	0.1416	0.1697	0.1905	0.2048
CUTCS	0.0342	0.0744	0.0976	0.1336	0.1807	0.2235	0.2467	0.2495	0.2567	52/48	0	0.0342	0.0655	0.0728	0.0997	0.1323	0.1577	0.1823	0.1959	0.2158
CART	0.0892	0.1118	0.1245	0.1611	0.2053	0.2426	0.2786	0.2931	0.4734	6/94	0	0.0892	0.1111	0.1221	0.1596	0.204	0.2431	0.2795	0.2945	0.4734
RCS	0.0607	0.0732	0.0803	0.0977	0.1305	0.1721	0.2229	0.2374	0.2994	87/13	0	0.138	0.142	0.1456	0.1592	0.2136	0.2373	0.2854	0.298	0.2994
FP	0.0894	0.0981	0.1002	0.1064	0.1209	0.1556	0.1897	0.2336	0.2685	97/3	9	0.1629	0.1699	0.177	0.198	0.233	0.2455	0.2529	0.2554	0.2579

Table A 3.6: Empirical distribution in terms of quantiles of the $\widehat{MSE}$ for the standard and bagged risk functions in the simulated **cutpoint model** (standardisation to zero mean log relative risk) and results of model selection

original data

risk function	Empirical quantile based on **all replications (R=1000)**									no. replications with linear/nonlinear risk function	Empirical quantile based on **replications with nonlinear risk function only**								
	min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.		min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.
LIN	0.0121	0.0143	0.0149	0.0166	0.0203	0.0292	0.0444	0.0566	0.2431	1000/0									
FIXALL	0	0	0.0002	0.0011	0.0046	0.0141	0.0285	0.0444	0.2359	0/1000									
LINQ	0.0121	0.0143	0.0151	0.0167	0.021	0.0321	0.056	0.0717	0.2431	944/56	0.0524	0.0587	0.0607	0.0671	0.0769	0.1028	0.1203	0.1426	0.1825
FIX	0	0	0.0002	0.0011	0.0093	0.0221	0.0369	0.0515	0.2359	328/672	0	0	0.0001	0.0005	0.0022	0.0097	0.0259	0.0438	0.2359
CUT	0	0.0033	0.0066	0.0176	0.0378	0.0686	0.1014	0.134	0.2849	92/908	0	0.003	0.0059	0.0161	0.0383	0.0715	0.1045	0.1368	0.2849
CUTS	0	0.0033	0.0059	0.0136	0.0323	0.0588	0.0874	0.1146	0.2654	92/908	0	0.003	0.0055	0.0121	0.0316	0.0599	0.0898	0.1199	0.2654
CUTC	0.0048	0.0138	0.0149	0.0174	0.0262	0.0558	0.0955	0.1172	0.2849	551/449	0.0048	0.0108	0.0167	0.0292	0.0566	0.0901	0.1194	0.1453	0.2849
CUTCS	0.0012	0.0111	0.014	0.0167	0.024	0.0473	0.0823	0.1027	0.2654	551/449	0.0012	0.0057	0.0104	0.0208	0.0456	0.0774	0.1046	0.1295	0.2654
CART	0	0.0102	0.0199	0.0419	0.0815	0.1406	0.1963	0.2361	0.4686	92/908	0	0.0093	0.0195	0.0509	0.0911	0.1455	0.2018	0.2403	0.4686
CARTS	0	0.0074	0.0158	0.0345	0.0625	0.1086	0.1698	0.2249	1.5306	92/908	0	0.007	0.0143	0.0371	0.0673	0.1141	0.1772	0.2303	1.5306
CARTC	0.0048	0.0139	0.015	0.0175	0.0267	0.0591	0.1055	0.1385	0.4218	551/449	0.0048	0.0113	0.0177	0.0311	0.0609	0.1	0.1451	0.1814	0.4218
CARTCS	0.0012	0.0117	0.0142	0.0168	0.0245	0.0502	0.0894	0.1215	1.5306	551/449	0.0012	0.0066	0.0112	0.0229	0.0499	0.0841	0.1223	0.1586	1.5306
RCS	0.0121	0.0144	0.0151	0.0169	0.0218	0.036	0.0649	0.0859	0.2431	888/112	0.0323	0.0396	0.0462	0.0592	0.0764	0.0961	0.1181	0.1597	0.2324
FP	0.0121	0.0143	0.015	0.0166	0.0204	0.0298	0.0461	0.061	0.2431	989/11	0.0416	0.0695	0.0975	0.1007	0.1152	0.1295	0.1423	0.156	0.1697

bagged risk functions (first 100 replications only)

risk function	Empirical quantile based on **all replications (R=1000)**									no. h linear/ nonlinear	no. **linear** in **all** bootstrap samples	Empirical quantile based on **replications with nonlinear risk function only**								
	min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.			min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.
LIN	0.012	0.014	0.0147	0.0165	0.0192	0.0288	0.0461	0.0546	0.0903	100/0	100									
LINQ	0.0119	0.0148	0.0159	0.0178	0.0215	0.0364	0.054	0.0718	0.091	91/9	0	0.034	0.0342	0.0344	0.042	0.0499	0.0763	0.0806	0.0858	0.091
FIX	0.0002	0.0005	0.001	0.0029	0.0067	0.0183	0.0333	0.0489	0.0664	32/68	0	0.0002	0.0002	0.0006	0.0016	0.0048	0.0164	0.0308	0.0396	0.0664
CUT	0.0039	0.0073	0.0089	0.0127	0.0228	0.0332	0.0438	0.0543	0.0687	6/94	0	0.0039	0.0073	0.0086	0.0126	0.0215	0.0305	0.039	0.0517	0.0665
CUTS	0.0049	0.0083	0.0096	0.0147	0.0246	0.0358	0.0459	0.0545	0.0668	6/94	0	0.0049	0.0081	0.0094	0.0138	0.0236	0.0313	0.0409	0.05	0.0653
CUTC	0.0053	0.0085	0.0102	0.0183	0.0326	0.0493	0.0593	0.0619	0.0683	52/48	0	0.0053	0.0073	0.0085	0.0117	0.0177	0.0244	0.0314	0.0471	0.0655
CUTCS	0.0049	0.01	0.0131	0.0221	0.0364	0.0512	0.0612	0.0639	0.0663	52/48	0	0.0049	0.009	0.0097	0.0137	0.021	0.0272	0.0364	0.0481	0.0605
CART	0.0116	0.0206	0.0224	0.0422	0.0632	0.0944	0.1182	0.1329	0.2932	6/94	0	0.0116	0.0203	0.0219	0.0419	0.0632	0.0976	0.1223	0.1337	0.2932
RCS	0.0072	0.0105	0.0111	0.0138	0.0227	0.0412	0.066	0.0746	0.1277	87/13	0	0.026	0.0288	0.0323	0.0433	0.06	0.0692	0.1075	0.1189	0.1277
FP	0.0117	0.0142	0.015	0.0167	0.02	0.0331	0.0488	0.068	0.0958	97/3	9	0.0439	0.0486	0.0532	0.0672	0.0906	0.0932	0.0948	0.0953	0.0958

Table A 3.7: The effect of bagging on the estimated error in the simulated **cutpoint model**:
Comparing the $\widehat{MAE}/\widehat{MSE}$ obtained in the original data to the corresponding errors of the bagged risk function

risk	linear	nonlinear	all (R=1000)					linear					nonlinear				
function			number (percent)		ratio bagging/original			number (percent)		ratio bagging/original			number (percent)		ratio bagging/original		
			of smaller $\widehat{MAE}$ for		quantiles			of smaller $\widehat{MAE}$ for		quantiles			of smaller $\widehat{MAE}$ for		quantiles		
			$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75	$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75	$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75
LIN	100	0	49(49)	51(51)	97	100	103	( )	( )				( )	( )			
LINQ	91	9	42(42)	58(58)	97	101	106	34(37.4)	57(62.6)	98	102	107	8 (88.9)	1 (11.1)	80	83	92
FIX	32	68	55(55)	45(45)	80	97	111	29(90.6)	3 (9.4)	64	79	89	26(38.2)	42(61.8)	94	106	135
CUT	6	94	58(58)	42(42)	57	90	136	2 (33.3)	4 (66.7)	100	109	126	56(59.6)	38(40.4)	56	84	138
CUTS	6	94	48(48)	52(52)	65	103	186	2 (33.3)	4 (66.7)	99	111	129	46(48.9)	48(51.1)	65	101	192
CUTC	52	48	39(39)	61(61)	67	130	181	3 (5.8)	49(94.2)	146	178	200	36(75)	12(25)	43	66	94
CUTCS	52	48	34(34)	66(66)	83	153	199	3 (5.8)	49(94.2)	154	186	208	31(64.6)	17(35.4)	6 0	81	137
CART	6	94	57(57)	43(43)	77	94	133	3 (50)	3 (50)	98	103	132	54(57.4)	40(42.6)	76	91	132
RCS	87	13	61(61)	39(39)	82	92	104	49(56.3)	38(43.7)	84	95	106	12(92.3)	1 (7.7)	80	83	89
FP	97	3	51(51)	49(49)	95	100	103	48(49.5)	49(50.5)	96	100	103	3 (100)	0 (0)	72	83	86

risk	linear	nonlinear	all (R=1000)					linear					linear				
function			number (percent)		ratio bagging/original			number (percent)		ratio bagging/original			number (percent)		ratio bagging/original		
			of smaller $\widehat{MSE}$ for		quantiles			of smaller $\widehat{MSE}$ for		quantiles			of smaller $\widehat{MSE}$ for		quantiles		
			$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75	$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75	$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75
LIN	100	0	51(51)	49(49)	96	100	105	( )	( )				( )	( )			
LINQ	91	9	38(38)	62(62)	97	103	116	30(33)	61(67)	99	105	118	8 (88.9)	1 (11.1)	57	7 4	81
FIX	32	68	55(55)	45(45)	60	95	127	31(96.9)	1 (3.1)	40	56	78	24(35.3)	44(64.7)	90	115	204
CUT	6	94	71(71)	29(29)	34	60	114	2 (33.3)	4 (66.7)	101	118	145	69(73.4)	25(26.6)	32	57	103
CUTS	6	94	60(60)	40(40)	46	86	148	2 (33.3)	4 (66.7)	100	122	152	58(61.7)	36(38.3)	43	78	146
CUTC	52	48	43(43)	57(57)	42	136	244	3 (5.8)	49(94.2)	187	239	277	40(83.3)	8 (16.7)	23	41	70
CUTCS	52	48	37(37)	63(63)	61	171	274	3 (5.8)	49(94.2)	198	261	305	34(70.8)	14(29.2)	3 3	54	127
CART	6	94	58(58)	42(42)	56	89	155	1 (16.7)	5 (83.3)	107	134	183	57(60.6)	37(39.4)	55	81	153
RCS	87	13	60(60)	40(40)	72	91	112	48(55.2)	39(44.8)	74	94	116	12(92.3)	1 (7.7)	64	68	78
FP	97	3	52(52)	48(48)	95	99	110	49(50.5)	48(49.5)	95	100	110	3 (100)	0 (0)	56	67	74

Table A 3.8: Empirical quantiles of the ratios 100* ($\widehat{MSE}_{\log(\lambda_0(t^*))}/\widehat{MSE}_{\text{zero mean}}$) and 100* ($\widehat{MAE}_{\log(\lambda_0(t^*))}/\widehat{MAE}_{\text{zero mean}}$) for the simulated **cutpoint model**

risk function		Empirical quantiles of the ratio for $\widehat{\textbf{MSE}}$ (R=1000)							number of	Empirical quantiles (**only nonlinear replications**)						
	$\hat{S}(t*)$	min.	0.05	0.25	0.5	0.75	0.95	max.	replications	min.	0.05	0.25	0.5	0.75	0.95	max.
LIN	0.50	100	100	100.1	100.5	101.2	103.5	109.5	0							
	0.75	100	100	100.2	100.6	101.6	103.5	107.9	0							
	0.90	100	100	100.4	101	102	103.9	114.3	0							
LINQ	0.50	100	100	100.1	100.5	101.2	103.4	109.5	56	100	100	100.1	100.3	100.9	102.4	104.3
	0.75	100	100	100.1	100.6	101.6	103.4	107.9	56	100	100	100.1	100.3	101.3	102.4	105
	0.90	100	100	100.3	101	102	104	114.3	56	100	100	100.1	100.8	102	105.8	112.2
FIX	0.50	100	100	100.1	100.6	101.6	103.3	109.1	672	100	100	100.1	100.6	101.6	103.3	107.5
	0.75	100	100.2	101.3	102.8	104.4	107.1	112	672	100.2	101.1	102.4	103.8	105.1	107.7	112
	0.90	100	100.6	102.2	104.7	106.5	109.3	116.3	672	102.1	103.3	104.7	105.9	107.2	109.7	116.3
CUT	0.50	100	100	100.1	100.5	101.2	103.2	110	908	100	100	100.1	100.4	101.2	103.2	110
	0.75	100	100	100.2	100.8	102.4	105.5	111.1	908	100	100	100.1	100.7	102.3	105.6	111.1
	0.90	100	100	100.3	101.4	103.8	107.8	115.1	908	100	100	100.2	101.2	104.1	107.9	115.1
CART	0.50	100	100	100.1	100.6	101.8	105.5	117.2	908	100	100	100.1	100.6	101.8	105.7	117.2
	0.75	100	100	100.2	100.8	102.2	104.8	110.8	908	100	100	100.2	100.8	102.2	104.9	110.8
	0.90	100	100	100.6	102.1	104.6	109.9	137	908	100	100	100.6	102.2	105	110.2	137
RCS	0.50	100	100	100.1	100.5	101.2	103.5	109.5	112	100	100	100	100.3	100.8	102	102.8
	0.75	100	100	100.1	100.6	101.6	103.5	107.9	112	100	100	100.2	100.9	102.1	105.3	108.4
	0.90	100	100	100.4	101	102	104	114.3	112	100	100.1	100.7	102.3	104	107.7	113.1
FP	0.50	100	100	100.1	100.5	101.2	103.5	109.5	11	100	100	100.3	100.8	101.7	102.5	102.8
	0.75	100	100	100.1	100.6	101.6	103.5	107.9	11	100	100	100.1	100.4	101	102.2	102.2
	0.90	100	100	100.4	101	102	104	114.3	11	100.1	100.2	100.9	102	103.3	105.2	106.1

risk function		Empirical quantiles of the ratio for $\widehat{\textbf{MAE}}$ (R=1000)							number of	Empirical quantiles (**only nonlinear replications**)						
	$\hat{S}(t*)$	min.	0.05	0.25	0.5	0.75	0.95	max.	replications	min.	0.05	0.25	0.5	0.75	0.95	max.
LIN	0.50	95.4	99.3	99.9	100.1	100.6	102.1	105.7	0							
	0.75	95	99.3	99.9	100.1	100.7	102.1	104.8	0							
	0.90	96.5	99	99.9	100.2	100.9	102.5	107.8	0							
LINQ	0.50	86.5	99.1	99.9	100.1	100.6	102.1	107.2	56	86.5	96.6	98.3	99.9	100.6	102.2	107.2
	0.75	92.6	99.2	99.9	100.1	100.7	102.4	111.7	56	92.6	97.4	100	100.5	102.6	105.7	111.7
	0.90	91.1	98.9	99.9	100.2	100.9	103	128.9	56	91.1	97.3	100	100.7	103.8	111.4	128.9
FIX	0.50	94	98.4	99.6	100	100.5	101.8	107.2	672	94	98.2	99.5	100	100.3	101.4	107.2
	0.75	93	97.3	99	100	101	102.9	108.9	672	93	96.9	98.6	100	101	103.1	108.9
	0.90	92.7	96.6	98.8	100.1	101.3	103.4	109.1	672	92.7	96.2	98.2	100	101.4	103.8	109.1
CUT	0.50	57.2	93.1	98.9	100	101.6	112.2	152.6	908	57.2	92.4	98.8	100	101.8	113.1	152.6
	0.75	57.4	93.8	98.4	100	101.9	108.2	124.3	908	57.4	93.6	98.1	100	102	109	124.3
	0.90	74.1	94.2	97.9	100	102.4	109.6	132.1	908	74.1	93.8	97.6	100	102.6	109.7	132.1
CART	0.50	54.3	85.4	97.1	99.9	101.4	109.2	152.5	908	54.3	84.8	96.6	99.8	101.6	109.5	152.5
	0.75	57.4	92.7	98.5	100.6	103.7	113.5	134.3	908	57.4	92.4	98.2	100.7	104.3	114.5	134.3
	0.90	75.8	92.1	98.4	101.4	107	129.4	180.9	908	75.8	91.8	98.2	102	108.4	130.1	180.9
RCS	0.50	95.2	99.2	99.9	100.1	100.5	102.1	105.7	112	95.4	98.7	99.8	100.1	100.5	102.5	108.7
	0.75	95	99.2	99.9	100.1	100.7	102.1	104.8	112	95.9	98.9	99.8	100.3	101	104.2	106
	0.90	96.5	98.9	99.9	100.2	100.9	102.7	107.8	112	97.1	98.8	99.9	100.6	101.9	104.6	114.2
FP	0.50	95.2	99.2	99.9	100.1	100.5	102.1	105.7	11	95.2	95.8	97	98.6	100	101.1	101.7
	0.75	95	99.2	99.9	100.1	100.7	102.1	104.8	11	97.3	98.1	99.5	100.9	102.1	103.4	104.2
	0.90	96.5	98.9	99.9	100.2	100.9	102.7	107.8	11	97.7	98	101.5	103.4	105.2	106.6	106.9

Table A 3.9: Empirical distribution in terms of quantiles of the $\widehat{MAE}$ for the standard and bagged risk functions in the simulated **linear model** (standardisation to zero mean log relative risk) and results of model selection

original data

risk function	Empirical quantile based on **all replications (R=1000)**									no. replications with linear/nonlinear risk function	Empirical quantile based on **replications with nonlinear risk function only**								
	min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.		min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.
LIN	0.0001	0.0054	0.0121	0.0286	0.0597	0.1043	0.1546	0.1869	0.3188	1000/0									
LINQ	0.0001	0.006	0.0134	0.0306	0.0636	0.1153	0.186	0.2132	0.3314	939/61	0.1708	0.1791	0.1831	0.1968	0.2159	0.2409	0.2779	0.2964	0.3314
FIX	0.0001	0.0059	0.013	0.0311	0.0664	0.1137	0.1684	0.1923	0.323	785/215	0.077	0.0865	0.0913	0.1035	0.1365	0.1746	0.2063	0.2452	0.323
CUT	0.0004	0.035	0.0525	0.1042	0.1486	0.1922	0.238	0.2747	0.3957	311/689	0.0887	0.1072	0.1179	0.1415	0.167	0.2125	0.2579	0.2904	0.3957
CUTS	0.0004	0.035	0.0525	0.0865	0.1262	0.1716	0.2146	0.2474	0.3716	311/689	0.0606	0.077	0.0871	0.1164	0.1427	0.1883	0.2308	0.2633	0.3716
CUTC	0.0001	0.0059	0.0126	0.0309	0.0639	0.1766	0.2287	0.2669	0.3957	781/219	0.1764	0.1904	0.1949	0.2085	0.2259	0.2587	0.2983	0.3116	0.3957
CUTCS	0.0001	0.0059	0.0126	0.0309	0.0639	0.1626	0.2086	0.2433	0.3716	781/219	0.1503	0.1643	0.1712	0.1844	0.1976	0.2361	0.2736	0.2859	0.3716
CART	0.0004	0.035	0.0525	0.1068	0.166	0.243	0.3076	0.3537	0.6016	311/689	0.0887	0.1191	0.1346	0.157	0.2076	0.27	0.3281	0.3695	0.6016
CARTS	0.0004	0.035	0.0525	0.0947	0.1398	0.2017	0.2668	0.3288	1.145	311/689	0.0617	0.0855	0.1018	0.1286	0.1691	0.2279	0.2949	0.3568	1.145
CARTC	0.0001	0.0059	0.0126	0.0309	0.0639	0.1766	0.2337	0.272	0.6016	781/219	0.1764	0.1889	0.196	0.2089	0.2281	0.2694	0.3052	0.3459	0.6016
CARTCS	0.0001	0.0059	0.0126	0.0309	0.0639	0.1604	0.2116	0.2509	0.6275	781/219	0.1325	0.1606	0.166	0.1837	0.1985	0.2429	0.2807	0.318	0.6275
RCS	0.0001	0.0059	0.0126	0.0306	0.0634	0.1137	0.1891	0.2431	0.3639	941/59	0.2019	0.2151	0.2291	0.2385	0.268	0.291	0.3308	0.3464	0.3639
FP	0.0001	0.0057	0.0121	0.0287	0.0601	0.1055	0.1624	0.1913	0.3603	991/9	0.2236	0.2371	0.2506	0.2704	0.3074	0.3236	0.3527	0.3565	0.3603

bagged risk functions (first 100 replications only)

risk function	Empirical quantile based on **all replications (R=1000)**									no. h linear/ nonlinear	no. **linear** in **all** bootstrap samples	Empirical quantile based on **replications with nonlinear risk function only**								
	min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.			min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.
LIN	0.0025	0.009	0.0135	0.0272	0.0513	0.1083	0.1722	0.1978	0.2405	100/0	100									
LINQ	0.008	0.0157	0.0209	0.0406	0.0775	0.1259	0.1847	0.2218	0.2529	94/6	0	0.1402	0.1534	0.1666	0.1997	0.2206	0.2444	0.2525	0.2527	0.2529
FIX	0.0103	0.0176	0.0238	0.0352	0.0593	0.1237	0.1823	0.2109	0.294	81/19	0	0.0846	0.1148	0.1183	0.1311	0.1565	0.1898	0.1997	0.2185	0.294
CUT	0.026	0.0453	0.0471	0.058	0.0792	0.0979	0.1266	0.1457	0.1883	28/72	0	0.026	0.0376	0.0467	0.0541	0.0753	0.0889	0.1085	0.1306	0.1883
CUTS	0.0234	0.0416	0.0485	0.0618	0.0809	0.0952	0.1236	0.1429	0.1764	28/72	0	0.0234	0.0398	0.0467	0.0532	0.0742	0.0874	0.1063	0.1216	0.1764
CUTC	0.0298	0.05	0.0549	0.0796	0.1046	0.1185	0.1263	0.1312	0.1787	92/18	0	0.0314	0.0373	0.0423	0.0506	0.0634	0.0808	0.1171	0.1302	0.1787
CUTCS	0.0387	0.0495	0.0566	0.085	0.1076	0.1206	0.1268	0.1351	0.166	92/18	0	0.0387	0.0424	0.0436	0.0498	0.0617	0.088	0.1078	0.1274	0.166
CART	0.0527	0.0948	0.1014	0.1282	0.1714	0.2138	0.2645	0.2793	0.4029	28/72	0	0.066	0.0935	0.1053	0.1353	0.1803	0.2254	0.2736	0.2936	0.4029
RCS	0.0129	0.0276	0.0385	0.0594	0.0943	0.1511	0.2184	0.2351	0.2694	93/7	0	0.2151	0.216	0.2169	0.2193	0.2247	0.2423	0.2565	0.2629	0.2694
FP	0.0034	0.0151	0.0204	0.0341	0.0718	0.122	0.1797	0.2159	0.2411	99/1	14	0.2157	0.2157	0.2157	0.2157	0.2157	0.2157	0.2157	0.2157	0.2157

Table A 3.10: Empirical distribution in terms of quantiles of the $\widehat{MSE}$ for the standard and bagged risk functions in the simulated **linear model** (standardisation to zero mean log relative risk) and results of model selection

original data

risk function	Empirical quantile based on **all replications (R=1000)**									no. replications with linear/nonlinear risk function	Empirical quantile based on **replications with nonlinear risk function only**								
	min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.		min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.
LIN	0	0	0.0002	0.0011	0.0047	0.0146	0.0325	0.0458	0.126	1000/0									
LINQ	0	0	0.0002	0.0012	0.0054	0.0178	0.0465	0.062	0.1595	939/61	0.0419	0.0462	0.0467	0.0532	0.062	0.0834	0.1186	0.1322	0.1595
FIX	0	0	0.0002	0.0013	0.0059	0.0173	0.0357	0.0459	0.126	785/215	0.0091	0.0115	0.0124	0.0154	0.0237	0.0359	0.0482	0.0655	0.1091
CUT	0	0.0016	0.0037	0.0147	0.0305	0.0538	0.0846	0.1125	0.222	311/689	0.0115	0.0158	0.0184	0.0284	0.0415	0.066	0.101	0.1234	0.222
CUTS	0	0.0016	0.0037	0.0103	0.0218	0.0422	0.0676	0.0919	0.201	311/689	0.0059	0.0087	0.0104	0.0178	0.029	0.0515	0.0807	0.1034	0.201
CUTC	0	0	0.0002	0.0013	0.0054	0.0404	0.0748	0.0983	0.222	781/219	0.0402	0.0477	0.0528	0.0609	0.0724	0.095	0.1248	0.1375	0.222
CUTCS	0	0	0.0002	0.0013	0.0054	0.0347	0.0608	0.084	0.201	781/219	0.0287	0.035	0.0396	0.0473	0.0566	0.0782	0.1073	0.1182	0.201
CART	0	0.0016	0.0037	0.0157	0.0434	0.1069	0.1654	0.2057	0.5778	311/689	0.0115	0.0191	0.0252	0.0388	0.0745	0.1338	0.1847	0.2214	0.5778
CARTS	0	0.0016	0.0037	0.0123	0.0285	0.069	0.1281	0.1728	1.3282	311/689	0.006	0.0106	0.0141	0.023	0.0507	0.0967	0.1557	0.1941	1.3282
CARTC	0	0	0.0002	0.0013	0.0054	0.0404	0.0786	0.1075	0.5778	781/219	0.0402	0.0484	0.0526	0.0626	0.0744	0.1048	0.1394	0.1815	0.5778
CARTCS	0	0	0.0002	0.0013	0.0054	0.0343	0.0636	0.09	0.6327	781/219	0.024	0.0346	0.0375	0.0459	0.0585	0.0873	0.1191	0.1516	0.6327
RCS	0	0	0.0002	0.0012	0.0054	0.0174	0.0481	0.0811	0.1856	941/ 59	0.056	0.071	0.0727	0.0803	0.0972	0.1236	0.1477	0.1615	0.1856
FP	0	0	0.0002	0.0011	0.0048	0.0148	0.035	0.0484	0.1794	991/9	0.0741	0.085	0.0959	0.1032	0.1325	0.1521	0.1734	0.1764	0.1794

bagged risk functions (first 100 replications only)

risk function	Empirical quantile based on **all replications (R=1000)**									no. *h* linear/ nonlinear	no. **linear** in **all** bootstrap samples	Empirical quantile based on **replications with nonlinear risk function only**								
	min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.			min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.
LIN	0	0.0001	0.0002	0.001	0.0035	0.015	0.0393	0.05	0.0751	100/0	100									
LINQ	0.0001	0.0003	0.0006	0.0023	0.0088	0.0219	0.0471	0.0701	0.1048	94/6	0	0.0268	0.0334	0.0401	0.055	0.0653	0.0843	0.0968	0.1008	0.1048
FIX	0.0002	0.0004	0.0008	0.0016	0.0044	0.0181	0.0402	0.0488	0.0912	81/19	0	0.01	0.0145	0.0156	0.0184	0.026	0.0382	0.0451	0.0525	0.0912
CUT	0.001	0.0031	0.0034	0.005	0.0083	0.0126	0.0241	0.0311	0.0621	28/72	0	0.001	0.0022	0.0032	0.0043	0.0077	0.0103	0.0168	0.0279	0.0621
CUTS	0.0008	0.0024	0.0034	0.0054	0.009	0.0123	0.0222	0.0293	0.0535	28/72	0	0.0008	0.0023	0.0033	0.0043	0.0078	0.01	0.0148	0.0233	0.0535
CUTC	0.0013	0.0032	0.0046	0.0089	0.0146	0.0184	0.0216	0.0231	0.0549	82/18	0	0.0014	0.002	0.0024	0.0034	0.0055	0.0092	0.0184	0.027	0.0549
CUTCS	0.0024	0.0033	0.0046	0.0095	0.0157	0.019	0.0218	0.0238	0.0463	82/18	0	0.0025	0.0027	0.0029	0.0033	0.0054	0.0107	0.015	0.0245	0.0463
CART	0.0049	0.0141	0.0164	0.0257	0.043	0.0715	0.0963	0.1134	0.2462	28/72	0	0.0077	0.015	0.0169	0.0289	0.0508	0.0815	0.1073	0.1236	0.2462
RCS	0.0002	0.001	0.0024	0.0051	0.0149	0.0349	0.0583	0.0707	0.11	93/7	0	0.0578	0.0591	0.0603	0.0643	0.0704	0.0777	0.0914	0.1007	0.11
FP	0	0.0003	0.0007	0.0015	0.0074	0.0201	0.0442	0.0648	0.0772	99/1	14	0.0677	0.0677	0.0677	0.0677	0.0677	0.0677	0.0677	0.0677	0.0677

Table A 3.1: The effect of bagging on the estimated error in the simulated **linear model**:

Comparing the $\widehat{MAE}/\widehat{MSE}$ obtained in the original data to the corresponding errors of the bagged risk function

risk	linear	nonlinear	all (R=1000)					linear					nonlinear				
function			number (percent) of smaller $\widehat{MAE}$ for		ratio bagging/original quantiles			number (percent) of smaller $\widehat{MAE}$ for		ratio bagging/original quantiles			number (percent) of smaller $\widehat{MAE}$ for		ratio bagging/original quantiles		
			$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75	$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75	$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75
LIN	100	0	51(51)	49(49)	88	100	116	( )	( )				( )	( )			
LINQ	94	6	34(34)	66(66)	95	108	153	28(29.8)	66(70.2)	97	109	156	6 (100)	0 (0)	93	94	96
FIX	81	19	38(38)	62(62)	91	108	131	33(40.7)	48(59.3)	90	108	140	5 (26.3)	14(73.7)	100	105	113
CUT	28	72	85(85)	15(15)	38	58	82	14(50)	14(50)	83	100	215	71(98.6)	1 (1.4)	34	47	60
CUTS	28	72	82(82)	18(18)	43	67	89	14(50)	14(50)	85	102	229	68(94.4)	4 (5.6)	37	61	70
CUTC	82	18	36(36)	64(64)	67	157	365	18(22)	64(78)	117	226	440	18(100)	0 (0)	21	29	35
CUTCS	82	18	36(36)	64(64)	69	160	390	18(22)	64(78)	118	234	465	18(100)	0 (0)	23	29	40
CART	28	72	54(54)	46(46)	79	96	119	2 (7.1)	26(92.9)	108	162	353	52(72.2)	20(27.8)	74	87	102
RCS	93	7	25(25)	75(75)	100	126	210	18(19.4)	75(80.6)	105	134	225	7 (100)	0 (0)	88	90	92
FP	99	1	38(38)	62(62)	93	106	142	37(37.4)	62(62.6)	94	107	142	1 (100)	0 (0)	80	80	80

risk	linear	nonlinear	all (R=1000)					linear					linear				
function			number (percent) of smaller $\widehat{MSE}$ for		ratio bagging/original quantiles			number (percent) of smaller $\widehat{MSE}$ for		ratio bagging/original quantiles			number (percent) of smaller $\widehat{MSE}$ for		ratio bagging/original quantiles		
			$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75	$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75	$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75
LIN	100	0	51(51)	49(49)	77	100	135	( )	( )				( )	( )			
LINQ	94	6	32(32)	68(68)	93	118	244	26(27.7)	68(72.3)	96	122	271	6 (100)	0 (0)	87	90	91
FIX	81	19	45(45)	55(55)	84	108	163	36(44.4)	45(55.6)	83	110	207	9 (47.4)	10(52.6)	94	105	113
CUT	28	72	84(84)	16(16)	15	30	69	12(42.9)	16(57.1)	75	124	514	72(100)	0 (0)	11	20	32
CUTS	28	72	80(80)	20(20)	19	43	90	12(42.9)	16(57.1)	76	123	557	68(94.4)	4 (5.6)	14	32	46
CUTC	82	18	36(36)	64(64)	44	256	1439	18(22)	64(78)	143	535	1906	18(100)	0 (0)	5	9	13
CUTCS	82	18	36(36)	64(64)	47	262	1511	18(22)	64(78)	146	552	2121	18(100)	0 (0)	6	10	16
CART	28	72	52(52)	48(48)	56	95	151	1 (3.6)	27(96.4)	135	323	1544	51(70.8)	21(29.2)	49	78	111
RCS	93	7	16(16)	84(84)	109	165	511	9 (9.7)	84(90.3)	117	202	583	7 (100)	0 (0)	79	81	83
FP	99	1	37(37)	63(63)	89	113	212	36(36.4)	63(63.6)	90	115	212	1 (100)	0 (0)	64	64	64

Table A 3.12: Empirical quantiles of the ratios 100* ($\widehat{MSE}_{\log(\lambda_0(t^*))}/\widehat{MSE}_{\text{zero mean}}$) and 100* ($\widehat{MAE}_{\log(\lambda_0(t^*))}/\widehat{MAE}_{\text{zero mean}}$) for the simulated **linear model**

risk function		Empirical quantiles of the ratio for $\widehat{\textbf{MSE}}$ (R=1000)							number of	Empirical quantiles (**only nonlinear replications**)						
	$\hat{S}(t*)$	min.	0.05	0.25	0.5	0.75	0.95	max.	replications	min.	0.05	0.25	0.5	0.75	0.95	max.
LIN	0.50	100	100	100.1	100.4	101.1	102.7	106.6	0							
	0.75	100	100	100.5	101	102	103.7	108.2	0							
	0.90	100	100.2	100.8	101.6	102.7	104.8	109.3	0							
LINQ	0.50	100	100	100.1	100.4	101.1	102.8	106.6	61	100	100	100.1	100.4	100.9	103.2	104.8
	0.75	100	100	100.4	101	101.9	103.6	108.2	61	100	100	100.1	100.2	100.8	102	102.7
	0.90	100	100.2	100.8	101.6	102.6	104.8	109.3	61	100	100	100.4	101	101.8	104.4	108.7
FIX	0.50	100	100	100.1	100.4	101.1	102.8	108	215	100	100	100.1	100.3	101	103.1	108
	0.75	100	100	100.4	100.9	101.8	103.4	108.2	215	100	100	100.4	101.1	102	103.7	107.6
	0.90	100	100.2	100.8	101.4	102.3	103.9	107.3	215	100.1	100.4	101	101.6	102.7	104.5	107.3
CUT	0.50	100	100	100.1	100.3	101	102.8	118.8	689	100	100	100.1	100.3	100.9	102.8	118.8
	0.75	100	100	100.2	100.6	101.5	103.3	108.7	689	100	100	100.2	100.6	101.6	103.4	108.7
	0.90	100	100.1	100.6	101.2	102.2	104.9	113	689	100	100.1	100.7	101.5	102.8	105.7	113
CART	0.50	100	100	100.1	100.5	101.4	104.2	118.8	689	100	100	100.1	100.5	101.5	104.9	118.8
	0.75	100	100	100.2	100.7	101.6	103.5	108.2	689	100	100	100.2	100.7	101.7	103.8	108.1
	0.90	100	100.1	100.7	101.4	102.9	107	119.8	689	100	100.2	101	102	103.9	108.5	119.8
RCS	0.50	100	100	100.1	100.4	101.1	102.8	106.6	59	100	100	100.1	100.4	101.3	104.4	107.1
	0.75	100	100	100.4	101	101.9	103.7	108.2	59	100	100	100.3	100.6	101.1	102.2	103.8
	0.90	100	100.2	100.9	101.6	102.7	104.8	109.3	59	100	100.2	100.9	101.6	102.5	105.1	106.8
FP	0.50	100	100	100.1	100.4	101.1	102.8	106.6	9	100	100	100.2	100.9	103.3	105.8	106.5
	0.75	100	100	100.4	101	101.9	103.7	108.2	9	100	100	100.1	100.3	100.4	100.9	101
	0.90	100	100.2	100.9	101.6	102.7	104.8	109.3	9	100.1	100.2	100.7	101.3	101.9	102.2	102.4

risk function		Empirical quantiles of the ratio for $\widehat{\textbf{MAE}}$ (R=1000)							number of	Empirical quantiles (**only nonlinear replications**)						
	$\hat{S}(t*)$	min.	0.05	0.25	0.5	0.75	0.95	max.	replications	min.	0.05	0.25	0.5	0.75	0.95	max.
LIN	0.50	98.6	99.6	99.9	100.1	100.5	101.4	104.2	0							
	0.75	98.2	99.3	99.9	100.3	100.8	102	105	0							
	0.90	98.1	99.3	100	100.4	101.1	102.5	104.8	0							
LINQ	0.50	96.2	99.4	99.9	100.1	100.5	101.6	107.9	61	96.2	97.1	98.7	99.7	100.7	102	107.9
	0.75	95.6	99.2	99.9	100.3	100.8	102.1	105	61	95.6	98.1	99.1	99.7	100.7	103.1	104.6
	0.90	95.2	99.2	99.9	100.4	101.2	102.7	112.1	61	95.2	97.1	98.7	100.2	101.7	105	112.1
FIX	0.50	97.9	99.4	99.9	100.1	100.5	101.5	104.2	215	97.9	99.1	99.9	100.1	100.5	101.8	104.1
	0.75	96.3	99.1	99.9	100.2	100.7	102	105.3	215	96.3	98.4	99.7	100.1	100.7	102	105.3
	0.90	94.7	99	99.9	100.4	100.9	102.2	105.4	215	94.7	98.1	99.5	100.2	100.9	102.4	105.4
CUT	0.50	87.2	96.7	99.5	100	100.5	102.1	114.3	689	87.2	96	98.9	99.8	100.5	102.2	114.3
	0.75	90.2	96.7	99.6	100.1	101	105.4	115	689	90.2	96	99.1	100	101.5	106.3	115
	0.90	88.5	96.1	99.5	100.3	101.6	109.1	135.2	689	88.5	95.2	99	100.5	103	111.2	135.2
CART	0.50	75	93.6	98.9	100	100.4	102.2	114.1	689	75	91.6	97.9	99.7	100.4	102.6	114.1
	0.75	86.8	96.4	99.7	100.2	101.8	107.9	126.5	689	86.8	95.9	99.3	100.4	103	109.8	126.5
	0.90	85.8	95.9	99.8	100.6	104.3	116.3	165.4	689	85.8	94.7	99.5	101.7	107.1	119.9	165.4
RCS	0.50	96.9	99.5	99.9	100.1	100.5	101.4	106.6	59	96.5	98.6	99.7	100	100.9	103.2	106.9
	0.75	98.2	99.3	99.9	100.3	100.8	102	105	59	97.9	98.4	99.6	100	100.6	102.2	105.5
	0.90	97.2	99.3	100	100.4	101.1	102.5	104.8	59	96.6	97.8	99.6	100.3	101.4	103.2	111.2
FP	0.50	96.9	99.5	99.9	100.1	100.5	101.4	106.6	9	96.9	97	97.7	98.9	100.8	105.1	106.6
	0.75	98.2	99.3	99.9	100.3	100.8	102	105	9	98.8	98.8	98.9	99.4	100.4	102.8	103
	0.90	97.2	99.3	100	100.4	101.1	102.5	104.8	9	97.2	97.4	98.1	100.8	102.6	104.4	104.4

Table A 3.13: Empirical distribution in terms of quantiles of the $\widehat{MAE}$ for the standard and bagged risk functions in the simulated **V-type model** (standardisation to zero mean log relative risk) and results of model selection

original data

risk function	Empirical quantile based on **all replications (R=1000)**									no. replications with linear/nonlinear risk function	Empirical quantile based on **replications with nonlinear risk function only**								
	min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.		min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.
LIN	0.1049	0.115	0.118	0.1243	0.1334	0.1513	0.1856	0.2113	0.3273	1000/0									
TRANSALL	0.0004	0.0055	0.0112	0.0279	0.0611	0.1033	0.1531	0.1863	0.335	0/1000									
LINQ	0.0501	0.0883	0.1088	0.1223	0.1334	0.1591	0.1957	0.2255	0.342	698/302	0.0501	0.072	0.0804	0.1004	0.1346	0.1799	0.2233	0.2468	0.342
FIX	0.1049	0.115	0.118	0.1243	0.1335	0.1534	0.1979	0.2239	0.3338	939/61	0.1906	0.1978	0.2093	0.2163	0.2335	0.2623	0.2807	0.2847	0.3338
TRANS	0.0391	0.067	0.0829	0.118	0.1303	0.1507	0.1893	0.2149	0.335	720/280	0.0391	0.0554	0.0591	0.0721	0.0998	0.1431	0.1909	0.2168	0.335
CUT	0.105	0.118	0.1236	0.136	0.167	0.2074	0.2516	0.2771	0.4569	316/684	0.1302	0.1444	0.1501	0.1655	0.1894	0.2237	0.27	0.2902	0.4569
CUTS	0.1045	0.1165	0.1198	0.1288	0.147	0.1828	0.2245	0.2486	0.4375	316/684	0.1045	0.1197	0.1275	0.1435	0.1665	0.1996	0.2381	0.2631	0.4375
CUTC	0.1049	0.1157	0.1188	0.1254	0.136	0.1733	0.2311	0.2672	0.4569	828/172	0.1824	0.1954	0.1985	0.2149	0.235	0.2706	0.3063	0.3258	0.4569
CUTCS	0.1049	0.1157	0.1188	0.1254	0.136	0.1727	0.2156	0.2404	0.4375	828/172	0.1694	0.1774	0.1813	0.1964	0.2165	0.243	0.2788	0.2996	0.4375
CART	0.105	0.118	0.1236	0.1363	0.1924	0.2709	0.3296	0.3547	0.5595	316/684	0.1308	0.1518	0.1628	0.1892	0.2423	0.2967	0.3442	0.3732	0.5595
CARTS	0.0646	0.1122	0.1179	0.1284	0.1568	0.2288	0.2929	0.3387	0.9063	316/684	0.0646	0.1126	0.1256	0.1516	0.2008	0.2572	0.316	0.3628	0.9063
CARTC	0.1049	0.1157	0.1188	0.1254	0.136	0.1733	0.239	0.2906	0.4602	828/172	0.1824	0.1956	0.1997	0.2173	0.2516	0.3109	0.3714	0.3918	0.4602
CARTCS	0.1	0.1155	0.1188	0.1253	0.1359	0.1685	0.2195	0.2606	0.9063	828/172	0.1	0.1682	0.1761	0.1904	0.2238	0.2632	0.3212	0.379	0.9063
RCS	0.0926	0.1153	0.1191	0.1265	0.139	0.1774	0.223	0.2558	0.3756	759/241	0.0926	0.121	0.1298	0.1511	0.1918	0.2351	0.2681	0.288	0.3756
FP	0.1049	0.1151	0.1187	0.1254	0.1369	0.165	0.2077	0.232	0.329	874/126	0.1126	0.1429	0.1495	0.1701	0.1971	0.229	0.2598	0.2861	0.329

bagged risk functions (first 100 replications only)

risk function	Empirical quantile based on **all replications (R=1000)**									no. *h* linear/ nonlinear	no. **linear** in **all** bootstrap samples	Empirical quantile based on **replications with nonlinear risk function only**								
	min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.			min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.
LIN	0.1091	0.1124	0.1166	0.1251	0.1331	0.1489	0.1788	0.2027	0.3124	100/0	100									
LINQ	0.0301	0.0414	0.0571	0.0804	0.1095	0.149	0.189	0.2176	0.2932	63/37	0	0.0381	0.0413	0.0547	0.0803	0.1199	0.1823	0.2129	0.2265	0.2932
FIX	0.1092	0.1119	0.1168	0.1263	0.1361	0.1528	0.1775	0.191	0.3027	95/5	0	0.1419	0.1462	0.1505	0.1634	0.2875	0.2988	0.3011	0.3019	0.3027
CUT	0.0624	0.0811	0.0847	0.092	0.1036	0.126	0.1433	0.1626	0.2385	28/72	0	0.0624	0.0792	0.0842	0.092	0.1072	0.1304	0.15	0.1699	0.2385
CUTS	0.0671	0.0842	0.0863	0.0942	0.1039	0.1216	0.1355	0.1552	0.2284	28/72	0	0.0671	0.0821	0.0857	0.0917	0.1071	0.1226	0.1421	0.16	0.2284
CUTC	0.0808	0.175	0.1962	0.2187	0.2342	0.2482	0.2562	0.266	0.3099	81/19	0	0.0808	0.1375	0.1512	0.1944	0.235	0.2627	0.2934	0.3073	0.3099
CUTCS	0.1009	0.182	0.1996	0.2203	0.2327	0.2467	0.2538	0.2595	0.2959	81/19	0	0.1009	0.141	0.1492	0.1927	0.2283	0.256	0.283	0.2951	0.2959
CART	0.0741	0.0967	0.1123	0.1443	0.1804	0.2376	0.3066	0.3286	0.3948	28/72	0	0.0975	0.1165	0.1445	0.1641	0.1942	0.2879	0.3182	0.3389	0.3948
RCS	0.0361	0.0561	0.0587	0.0943	0.1255	0.1626	0.232	0.2515	0.29	74/26	0	0.0892	0.0935	0.1012	0.1305	0.1862	0.2377	0.2562	0.2681	0.2885
FP	0.037	0.0526	0.0595	0.0905	0.1183	0.1521	0.1838	0.2092	0.2925	85/15	4	0.079	0.0902	0.1006	0.1457	0.1663	0.1948	0.2166	0.2231	0.2258
TRANS	0.0098	0.0198	0.0249	0.0526	0.0947	0.1303	0.1594	0.2008	0.2501	68/32	0	0.0118	0.0206	0.0306	0.0657	0.0961	0.129	0.1498	0.1768	0.2501

Table A 3.14: Empirical distribution in terms of quantiles of the $\widehat{MAE}$ for the standard and bagged risk functions in the simulated **V-type model** (standardisation to zero mean log relative risk) and results of model selection

original data

risk function	Empirical quantile based on **all replications (R=1000)**									no. replications with linear/nonlinear risk function	Empirical quantile based on **replications with nonlinear risk function only**								
	min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.		min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.
LIN	0.0149	0.0186	0.0197	0.0216	0.0258	0.0349	0.0494	0.0616	0.1356	1000/0									
TRANSALL	0	0	0.0002	0.0011	0.0049	0.0146	0.0309	0.0468	0.1461	0/1000									
LINQ	0.0041	0.0132	0.0176	0.0212	0.0263	0.0386	0.0563	0.0711	0.1708	698/302	0.0041	0.0082	0.0106	0.0161	0.0291	0.0485	0.0735	0.0932	0.1708
FIX	0.0149	0.0186	0.0197	0.0216	0.0258	0.0357	0.055	0.0694	0.1372	939/61	0.0547	0.0589	0.0608	0.0656	0.0751	0.0887	0.0986	0.1022	0.1372
TRANS	0.0021	0.0059	0.0091	0.0197	0.0239	0.0336	0.0497	0.0632	0.1461	720/280	0.0021	0.0041	0.0047	0.0069	0.0134	0.0274	0.049	0.064	0.1461
CUT	0.015	0.0195	0.0212	0.0263	0.045	0.0706	0.0983	0.1178	0.2714	316/684	0.0259	0.0322	0.035	0.0439	0.0589	0.0831	0.1088	0.1259	0.2714
CUTS	0.015	0.0191	0.0206	0.0234	0.033	0.0541	0.0799	0.097	0.2498	316/684	0.0156	0.0213	0.0239	0.0308	0.0435	0.0663	0.0896	0.1045	0.2498
CUTC	0.0149	0.0188	0.0199	0.0219	0.0268	0.0433	0.0939	0.1139	0.2714	828/172	0.0612	0.0708	0.0752	0.0838	0.0966	0.1177	0.1417	0.1535	0.2714
CUTCS	0.0149	0.0188	0.0199	0.0219	0.0268	0.0433	0.0773	0.0953	0.2498	828/172	0.0467	0.0559	0.0589	0.0678	0.0796	0.099	0.122	0.1317	0.2498
CART	0.015	0.0195	0.0212	0.0263	0.058	0.1161	0.1847	0.2219	0.5361	316/684	0.0259	0.0358	0.0415	0.0569	0.0925	0.1453	0.2104	0.2397	0.5361
CARTS	0.0061	0.0179	0.0195	0.0228	0.0373	0.0846	0.145	0.1875	0.8366	316/684	0.0061	0.0177	0.0219	0.0348	0.0651	0.1095	0.1692	0.2104	0.8366
CARTC	0.0149	0.0188	0.0199	0.0219	0.0268	0.0433	0.0964	0.135	0.2837	828/172	0.0612	0.0716	0.0755	0.088	0.1019	0.1426	0.1772	0.2117	0.2837
CARTCS	0.0142	0.0187	0.0199	0.0219	0.0267	0.0424	0.0756	0.1001	0.8366	828/172	0.0142	0.0473	0.0523	0.0655	0.0775	0.1074	0.1516	0.1833	0.8366
RCS	0.0138	0.0188	0.0202	0.0224	0.0291	0.0463	0.072	0.0919	0.1918	759/241	0.0138	0.022	0.0264	0.0374	0.0569	0.0831	0.1067	0.1189	0.1918
FP	0.0149	0.0187	0.0199	0.022	0.0276	0.0417	0.0634	0.0815	0.2245	874/126	0.0248	0.0316	0.0378	0.0482	0.0648	0.0863	0.1057	0.1154	0.2245

bagged risk functions (first 100 replications only)

risk function	Empirical quantile based on **all replications (R=1000)**									no. *h* linear/ nonlinear	no. **linear** in **all** bootstrap samples	Empirical quantile based on **replications with nonlinear risk function only**								
	min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.			min.	0.05	0.1	0.25	0.5	0.75	0.9	0.95	max.
LIN	0.017	0.0181	0.019	0.0214	0.0258	0.034	0.0454	0.0563	0.1186	100/0	100									
LINQ	0.0014	0.0026	0.0041	0.0095	0.0167	0.0314	0.0489	0.0669	0.1256	63/37	0	0.0024	0.0027	0.0055	0.0112	0.0206	0.0439	0.0676	0.0757	0.1256
FIX	0.017	0.018	0.0191	0.0221	0.0265	0.0347	0.0449	0.0513	0.1117	95/5	0	0.0312	0.0329	0.0345	0.0395	0.1025	0.1091	0.1106	0.1111	0.1117
CUT	0.0056	0.0093	0.0103	0.0122	0.0155	0.0222	0.0293	0.0419	0.1056	28/72	0	0.0056	0.0087	0.0098	0.0122	0.0159	0.0253	0.0311	0.0473	0.1056
CUTS	0.0064	0.0099	0.0109	0.0125	0.0151	0.021	0.0262	0.0366	0.0951	28/72	0	0.0064	0.0094	0.0103	0.0123	0.0164	0.0224	0.028	0.0409	0.0951
CUTC	0.0084	0.0372	0.0442	0.0514	0.0575	0.063	0.0674	0.0752	0.1001	81/19	0	0.0084	0.0282	0.0323	0.0476	0.0627	0.0808	0.0926	0.0997	0.1001
CUTCS	0.0117	0.0387	0.0449	0.0511	0.057	0.0616	0.0656	0.0701	0.0906	81/19	0	0.0117	0.03	0.0326	0.0469	0.0582	0.0747	0.0857	0.0905	0.0906
CART	0.009	0.0145	0.0189	0.0317	0.0494	0.0878	0.1428	0.1487	0.2081	28/72	0	0.0159	0.0237	0.0308	0.0414	0.0581	0.1102	0.1468	0.1701	0.2081
RCS	0.0021	0.0046	0.0057	0.0138	0.023	0.0381	0.0692	0.0905	0.1293	74/26	0	0.0131	0.0139	0.0149	0.0243	0.0515	0.0741	0.099	0.1115	0.1293
FP	0.0024	0.0042	0.0053	0.0126	0.0196	0.0341	0.0497	0.0636	0.1153	85/15	4	0.0108	0.0133	0.0169	0.0302	0.0495	0.0591	0.0667	0.0714	0.0775
TRANS	0.0001	0.0006	0.001	0.0036	0.0121	0.0245	0.0396	0.0543	0.0814	68/32	0	0.0001	0.0004	0.0008	0.0032	0.0067	0.0222	0.0525	0.0755	0.0814

Table A 3.15: The effect of bagging on the estimated error in the simulated **V-type model**:
Comparing the $\widehat{MAE}/\widehat{MSE}$ obtained in the original data to the corresponding errors of the bagged risk function

risk	linear	nonlinear	(R=100)					linear					nonlinear				
function			number (percent)		ratio bagging/original			number (percent)		ratio bagging/original			number (percent)		ratio bagging/original		
			of smaller $\widehat{MAE}$ for		quantiles			of smaller $\widehat{MAE}$ for		quantiles			of smaller $\widehat{MAE}$ for		quantiles		
			$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75	$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75	$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75
LIN	100	0	43(43)	57(57)	99	100	102	( )	( )				( )	( )			
LINQ	63	37	87(87)	13(13)	66	85	94	56(88.9)	7 (11.1)	63	80	92	31(8 3.8)	6 (16.2)	78	90	97
FIX	95	5	46(46)	54(54)	98	100	102	44(46.3)	51(53.7)	99	100	102	2 (40)	3 (60)	77	105	112
CUT	28	72	98(98)	2 (2)	56	66	77	26(92.9)	2 (7.1)	76	79	83	72(100)	0 (0)	50	60	68
CUTS	28	72	98(98)	2 (2)	62	74	81	26(92.9)	2 (7.1)	76	81	85	72(100)	0 (0)	58	69	76
CUTC	81	19	9 (9)	91(91)	133	167	191	1 (1.2)	80(98.8)	158	176	195	8 (42.1)	11(57.9)	80	105	113
CUTCS	81	19	7 (7)	93(93)	135	167	190	1 (1.2)	80(98.8)	159	174	195	6 (31.6)	13(68.4)	88	111	121
CART	28	72	63(63)	37(37)	76	90	108	13(46.4)	15(53.6)	82	105	133	50(69.4)	22(30.6)	72	86	102
RCS	74	26	85(85)	15(15)	74	87	96	62(83.8)	12(16.2)	69	85	97	23(8 8.5)	3 (11.5)	85	89	94
FP	85	15	90(90)	10(10)	67	86	95	75(88.2)	10(11.8)	66	88	96	15(10 0)	0 (0)	73	82	88
TRANS	68	32	91(91)	9 (9)	53	73	91	62(91.2)	6 (8.8)	48	73	90	29(90. 6)	3 (9.4)	62	73	92

risk	linear	nonlinear	all (R=1000)					linear					linear				
function			number (percent)		ratio bagging/original			number (percent)		ratio bagging/original			number (percent)		ratio bagging/original		
			of smaller $\widehat{MSE}$ for		quantiles			of smaller $\widehat{MSE}$ for		quantiles			of smaller $\widehat{MSE}$ for		quantiles		
			$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75	$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75	$\hat{\mathbf{h}}_{\mathbf{bagg}}$	$\hat{\mathbf{h}}$	0.25	0.5	0.75
LIN	100	0	40(40)	60(60)	98	101	104	( )	( )				( )	( )			
LINQ	63	37	89(89)	11(11)	43	71	87	57(90.5)	6 (9.5)	39	62	82	32(86 .5)	5 (13.5)	63	83	90
FIX	95	5	43(43)	57(57)	96	101	104	41(43.2)	54(56.8)	96	101	103	2 (40)	3 (60)	65	108	120
CUT	28	72	98(98)	2 (2)	28	41	60	26(92.9)	2 (7.1)	61	63	72	72(100)	0 (0)	25	33	45
CUTS	28	72	98(98)	2 (2)	35	51	66	26(92.9)	2 (7.1)	63	66	73	72(100)	0 (0)	34	46	55
CUTC	81	19	22(22)	78(78)	126	206	267	3 (3.7)	78(96.3)	178	235	274	19(100)	0 (0)	55	69	76
CUTCS	81	19	19(19)	81(81)	129	206	264	3 (3.7)	78(96.3)	179	229	271	16(84.2)	3 (15.8)	65	77	88
CART	28	72	63(63)	37(37)	58	77	118	10(35.7)	18(64.3)	77	123	189	53(73.6)	19(26.4)	54	69	102
RCS	74	26	82(82)	18(18)	56	75	90	59(79.7)	15(20.3)	48	71	91	23(8 8.5)	3 (11.5)	67	78	90
FP	85	15	89(89)	11(11)	47	74	87	74(87.1)	11(12.9)	47	75	90	15(10 0)	0 (0)	55	68	79
TRANS	68	32	93(93)	7 (7)	28	53	83	63(92.6)	5 (7.4)	20	53	81	30(93. 8)	2 (6.2)	40	53	84

Table A 3.16: Empirical quantiles of the ratios 100* ($\widehat{MSE}_{\log(\lambda_0(t^*))}/\widehat{MSE}_{\text{zero mean}}$) and 100* ($\widehat{MAE}_{\log(\lambda_0(t^*))}/\widehat{MAE}_{\text{zero mean}}$) for the simulated **V-type model**

risk function		Empirical quantiles of the ratio for $\widehat{\textbf{MSE}}$ (R=1000)							number of	Empirical quantiles (**only nonlinear replications**)						
	$\hat{S}(t*)$	min.	0.05	0.25	0.5	0.75	0.95	max.	replications	min.	0.05	0.25	0.5	0.75	0.95	max.
LIN	0.50	100	100	100.1	100.3	101.2	103.4	107.1	0							
	0.75	100	100	100	100.2	100.5	101.4	103.5	0							
	0.90	100	100	100	100.2	100.4	101	102.3	0							
LINQ	0.50	100	100	100.1	100.3	101	102.7	109	302	100	100	100.1	100.3	101.1	102.8	109
	0.75	100	100	100.1	100.3	101	103.5	108.4	302	100	100	100.4	101.2	102.7	105.5	108.4
	0.90	100	100	100.1	100.4	101.6	105.3	111.4	302	100	100.8	102	102.9	104.4	107.6	111.4
FIX	0.50	100	100	100.1	100.4	101.2	103.4	107.1	61	100	100	100	100.2	100.7	102.3	105.2
	0.75	100	100	100	100.2	100.5	101.4	103.5	61	100	100	100.1	100.2	100.5	102.2	102.9
	0.90	100	100	100	100.2	100.4	101.1	103	61	100	100	100.1	100.2	100.7	101.5	103
CUT	0.50	100	100	100.1	100.4	101.2	103.4	111.9	684	100	100	100.1	100.4	101.2	103.6	111.9
	0.75	100	100	100.1	100.4	101	102.6	111.5	684	100	100	100.1	100.4	101.1	102.8	111.5
	0.90	100	100	100.2	100.6	101.5	103.9	111.5	684	100	100	100.3	101	102	104.4	111.5
CART	0.50	100	100	100.1	100.5	101.5	104.7	115.8	684	100	100	100.1	100.6	101.7	105.1	115.8
	0.75	100	100	100.1	100.5	101.4	103.6	113.4	684	100	100	100.2	100.8	101.8	104.3	113.4
	0.90	100	100	100.3	101.3	103.3	107.7	114.7	684	100	100.1	101	102.4	104.4	108.7	114.7
RCS	0.50	100	100	100.1	100.3	101	103.1	109.3	241	100	100	100.1	100.4	101.1	102.7	107.1
	0.75	100	100	100	100.2	100.7	102.5	109.3	241	100	100	100.6	101.2	102.6	105	111.4
	0.90	100	100	100.1	100.2	100.6	105.3	113	241	100.1	100.9	101.8	103.2	104.7	108.2	111.5
FP	0.50	100	100	100.1	100.3	101	103.1	109.3	126	100	100	100.1	100.4	101.2	105.7	109.3
	0.75	100	100	100	100.2	100.7	102.5	109.3	126	100	100.1	100.7	101.5	103.3	106	109.3
	0.90	100	100	100.1	100.2	100.6	105.3	113	126	100.1	101.4	103.2	104.7	106.7	109.2	113

risk function		Empirical quantiles of the ratio for $\widehat{\textbf{MAE}}$ (R=1000)							number of	Empirical quantiles (**only nonlinear replications**)						
	$\hat{S}(t*)$	min.	0.05	0.25	0.5	0.75	0.95	max.	replications	min.	0.05	0.25	0.5	0.75	0.95	max.
LIN	0.50	92.8	97.9	99.7	100.1	100.9	103.1	114.2	0							
	0.75	93.8	98.1	99.8	100.1	100.5	102.1	105	0							
	0.90	95	98.5	99.9	100.1	100.6	101.4	103	0							
LINQ	0.50	89.6	96.8	99.5	100.1	101.2	104.4	114.2	302	89.6	95.2	98	99.8	101.6	105.5	110.7
	0.75	93.8	98.6	100	100.5	102.3	107.7	114.6	302	98	100.1	102.2	104	106.6	109.5	114.6
	0.90	95	98.8	100	100.6	104.3	109.9	115.1	302	99.8	102.9	105.1	106.8	108.9	111.6	115.1
FIX	0.50	92.8	98	99.8	100.1	100.8	102.9	114.2	61	98.2	99.4	99.9	100	100.3	101.1	101.2
	0.75	94.7	98.5	99.9	100.1	100.5	102	105	61	97.2	99.3	99.9	100.1	100.3	101.1	103.8
	0.90	95.3	98.8	99.9	100.1	100.5	101.4	103.5	61	97.5	99.3	99.9	100	100.3	100.9	103.5
CUT	0.50	89.9	96.3	99.4	100	100.6	103	114.2	684	89.9	95.4	98.8	99.8	100.5	103.3	109
	0.75	95.9	99.3	100	100.4	101.7	106	129.2	684	95.9	99	100	100.7	102.6	107	129.2
	0.90	94.9	99.6	100.1	100.8	103	110.7	128.7	684	94.9	99.6	100.3	101.8	104.9	112	128.7
CART	0.50	84.4	93.8	98.9	99.9	100.5	102.9	114.2	684	84.4	92.6	97.9	99.6	100.4	103.2	111
	0.75	91	97.8	99.9	100.4	102.1	107.9	123.5	684	91	97.1	99.8	100.8	103.1	110.3	123.5
	0.90	89	97.6	100	100.8	105	116.1	138.6	684	89	96.7	100.1	102.7	107.7	117.9	138.6
RCS	0.50	83.8	97.2	99.7	100.1	100.9	103.2	114.2	241	87.3	95	99.2	100	100.6	102.1	108.4
	0.75	93.8	98.3	99.9	100.2	101.1	105.4	115.3	241	96.1	98.5	100	101.4	103.5	109.1	114
	0.90	95	98.6	99.9	100.3	100.9	109.3	129.5	241	95.9	97.8	100.2	102.8	106.3	115.9	127.1
FP	0.50	83.8	97.2	99.7	100.1	100.9	103.2	114.2	126	83.8	90.4	97	99.6	100.6	102.7	107.7
	0.75	93.8	98.3	99.9	100.2	101.1	105.4	115.3	126	95.8	99.8	102	104	107.5	110.8	115.3
	0.90	95	98.6	99.9	100.3	100.9	109.3	129.5	126	98.7	101.2	105.1	107.9	111.3	119.1	129.5

6 Appendix B: Software and concept of programming

In this appendix I comment briefly on the software used in data generation and estimation of risk functions. Most computations have been performed using S-Plus, Version 3.4 (Statistical Sciences, 1993) on Sparc workstations under UNIX. To estimate the parameters of the restricted cubic spline I used the procedure *rcspline* from Frank Harrell, which is part of the *Hmisc* library (available from the S-library via *http://lib.stat.cmu.edu/S/*). Fractional polynomials were calculated by using the procedure *fp* of Stata 5.0/Stata 7.0 (StataCorp, 1997, 2001) . Meanwhile fractional polynomials can also be estimated in S-Plus, the corresponding programs by Gareth Ambler are also available in the S-library *http://lib.stat.cmu.edu/S/fracpoly*). Data driven cutpoints for CUT and CART as well as the corresponding minimum and corrected P-values were calculated by using a program written in C. This program was developped years ago at our institute. It was extended by my colleague Willi Sauerbrei with technical support by Martin Nehring and offers several possibilities to build and validate classification and regression trees (not only for survival data). In order to use this interactive programm in my simulation study it was necessary to write a connection to S-Plus. Thanks to Martin Nehring who made this connection possible. Running loops, the data set of each iteration has to be read into an ASCII file, which can then be used by the C program. Besides the fact that simulation is rather slow (e.g. several days when using 1000 simulated data sets with 100 bootstrap samples of each data set, i.e. 100.000 data sets at all) enormous memory is needed. Therefore, it was necessary to run the simulation programm in small steps using only up to 1000 data sets in each step. Besides these difficulties the main advantage of the C programe (with respect to my purpose) is its possibility to control the tree building algorithm by determining e.g. the selection interval for cutpoints or the depth of the tree (cf section 2.1.4). Furthermore, not only minimal P-values but also corrected P-values are calculated for each split. Alternativeley, to our C program there are a few programs available to caculate classification and regression trees for survival data in the S-library. To transfer the results of our C program into S-Plus the connection of both programms was constructed similar to that of *tssa* by Mark Segal *http://lib.stat.cmu.edu/S/tssa*). However, *tssa* do not need an extra ASCII file but can use the S-Plus data directly. In contrast to our C program *tssa* is based on the classical tree building and tree pruning steps (cf. section 2.1.4) and it do not calculate corrected P-values.

To generate bootstrap samples I used the library *http://lib.stat.cmu.edu/S/bootstrap.funs*), that contains all functions described in the textbook of Efron and Tibshirani (1993).

The connection between the different programs of the simulation study is illustrated in figure A 3.1. To avoid problems caused by the lack of memory I used two steps for estimating risk functions and making the estimated risk functions commparable. For the analysis of the breast cancer studies the concept is similar to figure A 3.1, instead of the simulated data sets I used the data of the study. Error estmation is of course not possible, because the true functional form of age is unknown.

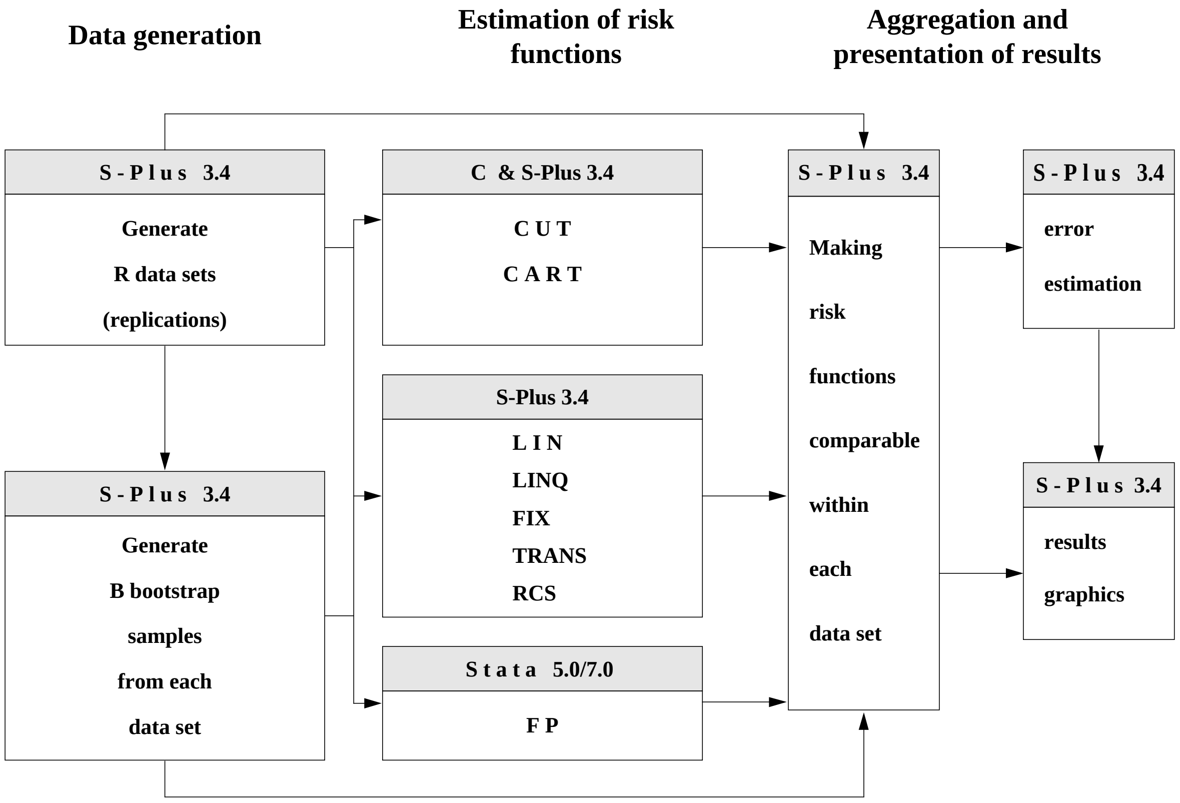


Figure A 3.1: Flow chart of programs of the simulation study

References

Akaike, H. (1973). Information theory and an extension of the entropy maximization principle. In Petrov, B. and Csak, F., editors, *2nd International Symposium on Information Theory*, pages 267–281. Akademia, Kiado.

Altman, D. (1998). Suboptimal analysis using 'optimal' cutpoints. letter to the editor. *British Journal of Cancer*, 78:556–557.

Altman, D., Lausen, B., Sauerbrei, W., and Schumacher, M. (1994). Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. commentary. *Journal of the National Cancer Institute*, 86:829–835.

Andersen, P., Borgan, O., Gill, R., and Keiding, N. (1993). *Statitical Methods Based on Counting Processes*. Springer-Verlag, New York.

Bloom, H. J. G. and Richardson, W. W. (1957). Histological grading and prognosis in breast cancer. *British Journal of Cancer*, 2:359–377.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 26:123–140.

Breiman, L., Friedman, J. H., Olsen, R. J., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Monterey.

Bühlmann, P. and Yu, B. (2000). Explaining bagging. Technical report, Seminar für Statistik, ETH, Zürich.

Chung, M., Chang, H., Bland, K., and Wanebo, H. (1996). Younger women with breast carcinoma have a poorer prognosis than older women. *Cancer*, 77:97–103.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836.

Contal, C. and O'Quigley, J. (1999). An application of changepoint methods in studying the effect of age on survival in breast cancer. *Computational Statistics and Data Analysis*, 30:253–270.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34:187–220.

De Boor, C. (1978). *A practical guide to splines*. Springer, New York.

De la Rochefordiere, A., Asselain, B., Campana, F., Scholl, S., Fenton, J., Viloq, J., Durand, J.-C., Pouillart, P., Magdelenat, H., and Fourquet, A. (1993). Age as parognostic factor in premenopausal breast carcinoma. *The Lancet*, 341:1039–1043.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap.* Chapman and Hall, New York.

Ezzat, A., Raja, M., Zwaan, F., Brigden, M., Rostom, A., and Bazarbashi, S. (1998). The lack of age as a significant prognostic factor in non-metastatic breast cancer. 1:23–27.

Fisher, E., Anderson, S., Tan-Chiu, E., Fisher, B., Eaton, L., and Wolmark, N. (2001). Fifteen-year prognostic discriminants for invasive breast carcinoma. *Cancer (Supplement)*, 91:1679–1687.

Gordon, L. and Olshen, R. (1985). Tree-structured survival analysis. *CanTreatmRep*, 69:1065–1069.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18:2592–2545.

Grambsch, P. and Therneau, T. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81:515–526.

Harrell, F., Lee, K., and Pollock, B. (1988). Regression models in clinical studies: determining relationships between predictors and response. *Journal of the National Cancer Institute*, 87:1198–1202.

Harrell, F. E. (1997). *Predicting Ouutcomes: Applied Survival Analysis and Logistic Regression (Technical Report).* University of Virginia, USA, Virginia.

Henderson, R. (1995). Problems and prediction in survival analysis. *Statistics in Medicine*, 14:143–152.

Hilsenbeck, S. G. and Clark, G. M. (1996). Practical p-value adjustment for optimally selected cutpoints. *Statistics in Medicine*, 15:103–112.

Holländer, N. and Schumacher, M. (2001). On the problem of using 'optimal' cutpoints in the assessment of quantitative prognostic factors. *Onkologie*, 24:194–199.

Kalbfleisch, J. and Prentice, R. (1980). *The statistical analysis of failure time data.* Wiley, New York.

Kroman, N., Jensen, M.-B., Wohlfahrt, J., Mouridsen, H., Andersen, P., and Melbye, M. (2000). Factors influencing the effect of age on prognosis in breast cancer: population study. *British Medical Journal*, 320:474–479.

Lausen, B., Sauerbrei, W., and Schumacher, M. (1994). Classification and regression trees (cart) used for the exploration of prognostic factors measured on different scales. In

Dirschedl, P. and Ostermann, R., editors, *Computational Statistics*, pages 1483–1496. Physica-Verlag, Heidelberg.

Lausen, B. and Schumacher, M. (1992). Maximally selected rank statistics. *Biomtrics*, 48:73–85.

Lausen, B. and Schumacher, M. (1996). Evaluating the effect of optimized cutoff values in the assessment of prognostic factors. *Computational Statistics and Data Analysis*, 21:307–326.

Le Blanc, M. and Crowley, J. (1992). Relative risk regression trees for censored survival data. *Biometrics*, 77:411–425.

Le Blanc, M. and Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association*, 88:457–467.

Le Blanc, M. and Crowley, J. (1995). Step-function covariate effects in the proportional-hazards model. *The Canadian Journal of Statistics*, 23:109–129.

Le Blanc, M. and Crowley, J. (1999). Adaptive regression splines in the cox model. *Biometrics*, 55:204–213.

Linderholm, B., Grankvist, K., Wilking, N., Johannson, M., Tavelin, B., and Henriksson, R. (2000). Correlation of vascular endothetical growth factor content with recurrences, survival, and first relapse side in primary node-positive breast carcinoma after adjuvant treatment. *Journal of Clinical Oncology*, 18:1423–1431.

Marubini, E. and Valsecchi, M. G. (1995). *Analysing survival data from clinical trials and observational studies.* Wiley, New York.

Miller, A. J. (1990). *Subset selection in regression.* Chapman and Hall, London.

O'Quigley, J. and Pessione, F. (1989). Score tests for homogenity of regression effects in the proportional hazards model. *Biometrics*, 45:135–144.

Pfisterer, J., Kommoss, F., Sauerbrei, W., Menzel, D., Kiechle, M., Giese, E., Hilgarth, M., and Pfleiderer, A. (1995). DNA flow cytometry in node positive breast cancer: prognostic value and correlation to morphological and clinical factors. *Analytical and Quantitative Cytology and Histology*, 17:406–412.

Royston, P. and Altman, D. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Applied Statistics*, 43:429–467.

Sauerbrei, W. (1998). Bootstraping in survival analysis. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics*, pages 433–436. Wiley, Chichester.

Sauerbrei, W. and Royston, P. (1999). Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society, Series A*, 162:71–94.

Sauerbrei, W., Royston, P., Bojar, H., Schmoor, C., and Schumacher, M. f. t. G. B. C. S. G. G. (1999). Modelling the effects of standard prognostic factors in node-positive breast cancer. *British Journal of Cancer*, 79:1752–1760.

Schmoor, C., Olschewski, M., and Schumacher, M. (1996). Randomized and non-randomized patients in clinical trials: experiences with comprehensive cohort studies. *Statistics in Medicine*, 15:263–271.

Schumacher, M., Bastert, G., Bojar, H., Hübner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Neumann, R., and Rauschecker, H. f. t. B. C. S. G. (1994). Randomized 2x2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology*, 12:2086–2093.

Schumacher, M., Holländer, N., and Sauerbrei, W. (1996). Reduction of bias caused by model building. In *ASA Proceedings of the Statistical Computing Section*, pages 1–7. American Statistical Association, Alexandria.

Schumacher, M., Holländer, N., and Sauerbrei, W. (1997). Resampling and cross-validation techniques: a tool to reduce bias caused by model building. *Statistics in Medicine*, 16:2813–2827.

Schumacher, M., Holländer, N., Schwarzer, G., and Sauerbrei, W. (2001). Prognostic factor studies. In Crowley, J., editor, *Handbook of Statistical Methods in Clinical Oncology*, pages 321–378. Dekker, New York.

Segal, M. (1988). Regression trees for censored data. *Biometrics*, 44:35–47.

StataCorp (1997). *Stata statistical software: release 5.0.* College Station, Texas: Stata Corporation.

StataCorp (2001). *Stata statistical software: release 7.0.* College Station, Stata Corporation, Texas.

Statistical Sciences (1993). *S-PLUS programmer's manual, Version 3.2.* StatSci, a division of MathSoft Inc., Seattle.

Stone, C. (1986). Commnet: Generalized additive models. *Statistical Science*, 1:312–314.

Stone, C. J. and Koo, C. Y. (1985). Additive splines in statistics. In *Proceedings of the Statistical Computing Section ASA*, pages 45–48.

Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990). Martingale based residuals for survival models. *Biometrika*, 48:147–460.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82:559–567.

Van Houwelingen, H. and Le Cessie, S. (1990). Predictive value of statistical models. *Statistics in Medicine*, 9:1303–1325.

Vanlemmens, L., Hebbar, M., Peyrat, J., and Bonneterre, J. (1998). Age as a prognostic factor in breast cancer. *Anticancer Research*, 18:1891–1896.

Verweij, P. and Van Houwelingen, H. (1993). Cross-validation in survival analysis. *Statistics in Medicine*, 12:2305–2314.

Xu, R. and Adak, S. (2001). Survival analysis with time-varying relative risks: a tree-based approach. *Methods of Information in Medicine*, 40:141–147.

Zhang, H., Crowley, J., Sox, H., and Olshen, R. (1998). Tree-structured statistical methods. pages 4561–4573. Wiley, Chichester.