May 2015

Semiparametric Estimation of the Survival Function in the Presence of Covariates

Madlen Gebauer
University of Wisconsin-Milwaukee

Semiparametric Estimation of the Survival Function in the Presence of Covariates

by

Madlen Gebauer

A Thesis Submitted in
Partial Fulfillment of the
Requirements for the Degree of

Master of Science
in
Mathematics

at

The University of Wisconsin-Milwaukee
May 2015

ABSTRACT

Semiparametric Estimation of the Survival Function in the Presence of Covariates

by

Madlen Gebauer

The University of Wisconsin-Milwaukee, 2015
Under the Supervision of Professor Jugal K. Ghorai

The main interest of survival analysis is to estimate the distribution function of the survival time based on observations of a random sample. In this thesis, a semiparametric estimator is used not only to estimate the survival probability, but also to consider the influence of explanatory variables within the estimation. Therefore, the weighted maximum likelihood estimator of the conditional survival function is derived and a corresponding pointwise likelihood ratio confidence band is developed. Subsequently, the established estimator is compared to a similar estimator which was proposed by Iglesias-Pérez and de Uña-Álvarez [8]. Since the idea of this paper arose in cooperation with an automotive company, the focus is on the application of this model in context of the automotive industry.

A method to select covariates which seem to have the most impact on the failure behavior is derived, using the proposed estimate. Furthermore, the strength of the impact is identified and a profile of the effect is established.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor, Professor Jugal K. Ghorai, for his continuous support of my academic study and research. My earnest thanks also goes to Thomas Köttermann who encouraged my research on behalf of an international leading automotive corporation.

Moreover, I would like to thank Professor Jay H. Beder and Professor Richard H. Stockbridge for their comments, remarks, and engagements in addition to being part of the committee.

Introduction

By analyzing survival data we are interested in the time that passes until a certain event occurs. In contrast to classical statistics, these studies are more complicated due to fragmented data and the fact that a variety of additional characteristics influence the time to failure. There are many data sets that display these features, thus, the application of survival analysis is vital. The methods are usually applied in the medical sector, but are also implemented in the fields of engineering, economics and sociology. The explanatory variables have a decisive influence on the results of the analyses. For example, time independent quantities such as gender, age, or medication, as well as time dependent quantities like blood pressure or location can affect the survival time of a subject.
Often, we are not able to observe the whole information of the failure development. A study might be terminated before all of the subjects reach completion due to time or cost considerations. These diverse influences of explanatory variables and censored data on the survival behavior must be considered within the application of statistical models.

One of the most popular nonparametric estimators of survival functions is the Kaplan-Meier estimator. The semiparametric Cox proportional hazards model additionally takes the influence of external information into account and provides the opportunity to compare these effects through the relationship of their corresponding regression coefficients. However, the assumption of proportional hazards in this model is a major restriction, and therefore may not provide accurate results for many practical situations. Thus, Dikta [5] proposed an alternative semiparametric

estimator for the survival function where weaker assumptions are required. This thesis explores an extension of Dikta's semiparametric estimator which considers covariates to estimate the survival function.

The main purpose of this thesis is to assess the effect of several covariates on the survival time. More precisely, which of the additional factors seem to have the largest impact on the failure of an object.

This concern arose within an automotive environment where additional information on failure behavior is available but is not actually used. Covariates such as average speed, engine performance or torque are potential reasons for car failure. Thus, the impact of each explanatory variable on failure and the use of this information in order to improve the particular vehicle parts, relative to the expenses, is of high interest.

A method to analyze this issue is developed in this paper. First, the semiparametric likelihood estimator of the conditional survival function is characterized and pointwise confidence bands based on the likelihood ratio test statistic are provided. Next, all covariates except one are fixed in order to observe their particular effects on the survival probability and a conclusion is drawn on how strong the impact of each variable is on failure.

Finally, the theoretical results are illustrated in an applied framework using an authentic data set provided by an automotive company. The basic steps were to reduce the numbers of covariates before estimating the conditional probability that an automobile does not fail until a certain time point, taking into account the additional information. Therefore, the established semiparametric estimator is used and likelihood ratio confidence bands are provided. Lastly, this estimator is compared to the semiparametric estimator proposed by Iglesias-Pérez and Uña-Álvarez [8].

This thesis is divided into four Chapters. The first Chapter introduces the reader to the data set and preliminaries of survival analysis. The semiparametric estimator of the conditional survival function in the presence of covariates is derived in Chapter 2. Moreover, the maximum likelihood property is shown and a confidence band is

developed before the established method is applied to a practical example and the effects of several covariates are assessed in Chapter 3. A summary and future plans in Chapter 4 conclude this thesis.

Chapter 1

Preliminaries

In this chapter, the fundamentals and relations in survival analysis that are necessary to follow this paper are explained.

1.1 Motivation

First, let us analyze the data set we are working with which was provided from an automotive company.
The whole data set includes around $17,500$ cars. Within this fleet, events on about 500 cars were observed from which we therefore have complete information. For the remaining $17,000$ cars only fragmentary data is provided due to no observed failure before the study ended. Any events after the endpoint of the data collection are unknown. The data set deals with three different types of engines, thus, one particular type is considered to begin with and the remaining two sets are used for validation.
To begin, we investigate a car fleet which consists of around $8,900$ cars where 265 failures are observed. For each car, observations of the mileage, an indicator if an event occurred for that specific car, and 12 variables that contain additional information such as average fuel consumption, average speed, engine performance or torque are observed.
In this study, the preliminary variable of interest is the mileage of a car which is in this case understood as the survival time. Since 12 variables is a lot compared to the

265 observed failures, and additionally some of the variables are highly correlated, such as average speed and torque, we have to select those variables which are actually influencing the failure time of a car.

First of all, the most important methods of descriptive analysis on the data set are applied to obtain an initial impression and feeling for better understanding. Since we focus on the observed survival time, the sample $(z_1, \ldots, z_n)$ is considered to be realizations of the statistical model $(Z_1, \ldots, Z_n)$, where $Z_1, \ldots, Z_n$ are independent and identically distributed positive random variables with distribution function H. This sample has a mean of $39,910$, a median of $29,740$ and a standard deviation of $35,313$. A histogram of the data is found in Figure 1.1.

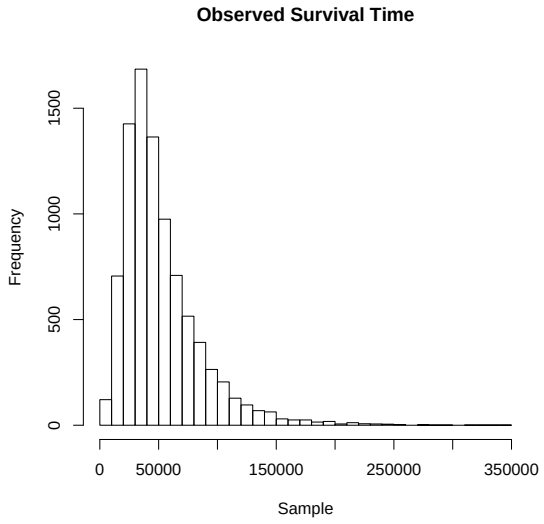

Figure 1.1: Histogram of the observed survival time.

We observe an unimodal distribution with positive skew and a very flat right tail.

In conventional statistics we would now determine the empirical distribution function in order to estimate the actual distribution function. Since the examined data set is censored, we need to apply modified methods which are introduced in the following sections.

1.2 Basic Quantities

In survival analysis, the most important variable is the survival time that describes the time Y until a certain event occurs.
We understand Y to be a nonnegative random variable which can be interpreted as the time until an event occurs, or the time until decease. Note that this variable does not have to be a time-based quantity. It can also represent the mileage until the engine breaks down which is the case in this study.
Our goal is to analyze Y, thus, we are interested in the *survival function* $S(t)$ of Y:

$$S(t) = \mathbb{P}[Y > t] = 1 - F(t), \qquad (t \geq 0) \tag{1.1}$$

where F denotes the distribution function of Y.
Besides the survival function, the *cumulative hazard function* $\Lambda : [0, \infty] \to [0, \infty]$ corresponding to F with

$$\Lambda(t) = -\int_0^t \frac{1}{S(v-)} dS(v) = \int_0^t \frac{1}{S(v-)} dF(v) \tag{1.2}$$

is another term for describing the survival time. This result can be derived from the idea of the corresponding density function, the hazard rate, which is explored in the following:
One is interested in the chance that an event occurs within the next time step Δt, given that one survived t units of time:

$$\frac{\mathbb{P}[t < Y \leq t + \triangle t | Y \geq t]}{\triangle t} = \frac{\mathbb{P}[t < Y \leq t + \triangle t]}{\triangle t \, \mathbb{P}[Y \geq t]}$$

$$= -\frac{S(t + \triangle t) - S(t)}{\triangle t} \cdot \frac{1}{S(t-)} \underset{\triangle t \to 0}{\longrightarrow} -\frac{dS(t)}{S(t-)}.$$

If F is an absolute continuous distribution function, then Λ has a density function λ which we call the *hazard rate* corresponding to F and denote it by

$$\lambda(t) = \frac{f(t)}{S(t)}, \qquad (t \geq 0).$$

Here, we define $\frac{0}{0} = 0$.
Therefore,

$$\Lambda(t) = \int_0^t \frac{f(v)}{S(v)} dv, \qquad (t \geq 0),$$

if F is absolute continuous. Note that $\Lambda(0) = 0$ and Λ is monotone increasing but not necessarily bounded and therefore not a distribution function.

Remark 1.2.1. *In the continuous case, one can observe the following relation between the survival function S and the cumulative hazard function Λ, for all $t \geq 0$:*

$$S(t) = e^{-\Lambda(t)} \tag{1.3}$$

$$\Lambda(t) = -\ln(S(t)) \tag{1.4}$$

1.3 Censoring

When events are narrowed down to a particular time frame, we are talking about *censored* data. The chosen censoring times may vary from subject to subject. Censored data occurs in real-life studies where the study might be terminated before all of the participating subjects realized their events due to time or cost considerations. There are three different types of censoring: right censoring, left censoring and interval censoring. We focus on right censoring, since the introduced data set only contains observations from the registration of a car until warranty time expires. Figure 1.2 provides an example for six cars. The square and the subsequent mileage C_i denote that the car's life length was not completely observed, since no event occurred within this certain period of time. We do know that car i did not experience an event until time C_i but we were not able to observe what happened after this time point. However, the arrow denotes that we observe an event for car i at time Y_i in the given time horizon. This means that cars 2 and 4 failed within the warranty time whereas events for cars 1, 3, 5 and 6 have not been observed.
Although, this is an exaggerated representation of failures within a car fleet, it provides a helpful illustration of the term censoring.

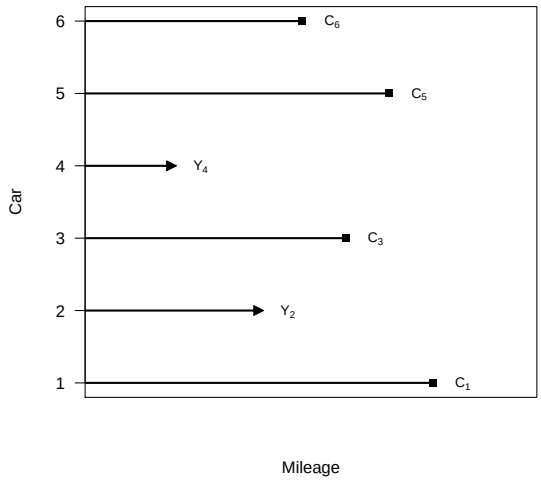

Figure 1.2: Right censoring.

1.4 Estimating the Survival Function

First, let us introduce some notation:
Let $Y_1, \ldots, Y_n$ be n independent and identical distributed positive random variables defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with unknown continuous distribution function F. We consider the random censorship model where these quantities are censored on the right and we only observe the random variables $Z_1, \ldots, Z_n$ with distribution function H, where $Z_i = \min\{Y_i, C_i\}$ denotes the minimum of the survival time Y_i and the censoring time C_i with distribution function G. The censoring indicator $\delta_i = \mathbb{1}_{\{Y_i \leq C_i\}}$ indicates whether Z_i is censored ($\delta_i = 0$) or not ($\delta_i = 1$). Furthermore, let X_i be a p-dimensional covariate vector. We assume the censoring times to be independent of the survival times.

1.4.1 The Kaplan-Meier Estimator

The Kaplan-Meier estimator proposed by Kaplan and Meier [9] is a consistent non-parametric estimator of the survival function, in case of independent censoring. It is a modification of the empirical distribution function in case of censored data.

Moreover, if data is not censored, the estimator reduces to the empirical distribution function.

The *Kaplan-Meier estimator* is denoted by

$$\hat{S}_{KM}(t) = \begin{cases} 1, & 0 \leq t < t_1 \\ \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right), & \text{otherwise} \end{cases}, \tag{1.5}$$

where d_j denotes the number of events at time t_j and n_j the number of individuals at risk at time t_j-.
In the following, we assume that there is at the most one failure at each time t_j and therefore $d_j = 1$. Rewriting the number of individuals at risk shortly before t_j yields

$$n_j = n - R_j + 1,$$

where R_j denotes the rank of Z_j within the Z-sample. This is a consequence of considering the number of objects for which an event occurred at time t_j- to be equal to $R_j - 1$. Since the observation time t_j is related to the occurrence of events Z_j, we can substitute the t_j's by the Z_j's regarding the product boundaries.
Finally, the Kaplan-Meier estimator can be rewritten as

$$\hat{S}_{KM}(t) = \begin{cases} 1, & \text{if } 0 \leq t < Z_{1:n} \\ \prod_{j:Z_j \leq t} \left(\frac{n - R_j}{n - R_j + 1}\right)^{\delta_j}, & \text{otherwise} \end{cases}, \tag{1.6}$$

where $Z_{1:n}$ denotes the first order statistic of the Z-sample.
Note that the Kaplan-Meier estimator is a monotonic non-increasing step function with $\hat{S}_{KM}(0) = 1$, but not necessarily equal to zero for $\hat{S}_{KM}(t)$ as $t \to \infty$.

1.4.2 The Nelson-Aalen Estimator

Because of the relationship of the survival function and the cumulative hazard function which was pointed out in Remark 1.2.1, we can also estimate the latter one in order to achieve information about the survival function:

First define $H^1(x) := \mathbb{P}[\delta = 1, Z \leq x]$ which can be rewritten as:

$$H^1(x) = \int_0^x \mathbb{P}[\delta = 1|Z = z]dH(z). \tag{1.7}$$

Using equation(1.2) and applying (1.7) leads to

$$\Lambda(t) = \int_0^t \frac{1}{1 - F(x-)}dF(x) = \int_0^t \frac{1}{1 - H(x-)}dH^1(x). \tag{1.8}$$

A consistent estimator for the cumulative hazard function is given by the Nelson-Aalen estimator [1].
Let $H_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Z_i \leq x\}}$ denote the empirical distribution function of the Z-sample with $H_n(x-) = \lim_{z \uparrow x} H_n(z)$ and $\bar{H}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Z_i \leq x\}}\delta_i$.
Then, the *Nelson-Aalen Estimator* is defined by

$$\hat{\Lambda}_{NA}(t) = \int_0^t \frac{1}{1 - H_n(v-)}d\bar{H}_n(v) = \sum_{i:Z_i \leq t} \frac{\delta_i}{n - R_i + 1}. \tag{1.9}$$

Using Remark 1.2.1 and the approximation $\exp(-x) \approx 1 - x$ for small x, one can derive

$$e^{-\hat{\Lambda}_{NA}(t)} \approx \hat{S}_{KM}(t) \tag{1.10}$$

considering

$$\begin{aligned} e^{-\hat{\Lambda}_{NA}(t)} &= \exp\left\{\sum_{i:Z_i \leq t} -\frac{\delta_i}{n - R_i + 1}\right\} = \prod_{i:Z_i \leq t} \exp\left\{-\frac{\delta_i}{n - R_i + 1}\right\} \\ &= \prod_{i:Z_i \leq t} \exp\left\{-\frac{1}{n - R_i + 1}\right\}^{\delta_i} \approx \prod_{i:Z_i \leq t} \left(1 - \frac{1}{n - R_i + 1}\right)^{\delta_i} = \hat{S}_{KM}(t). \end{aligned}$$

1.4.3 The Semiparametric Estimator

In order to introduce the semiparametric estimator which was proposed by Dikta [5], the integrand of H^1 is defined by $m(x) := \mathbb{P}[\delta = 1|Z = x] = \mathbb{E}[\delta|Z = x]$. That

is the conditional expectation of the censoring indicator δ, given $Z = x$.
Using equation (1.2) and applying the above result (1.8) leads to

$$\Lambda(t) = \int_0^t \frac{1}{1 - H(x-)} dH^1(x) = \int_0^t \frac{m(x)}{1 - H(x-)} dH(x). \tag{1.11}$$

Furthermore, since F is a continuous distribution function we can write:

$$\Lambda(t) = \int_0^t \frac{m(x)}{1 - H(x)} dH(x) = \int_0^t \frac{m(x)h(x)}{1 - H(x)} dx.$$

Based on equation (1.11), a variety of estimators can be derived simply by specifying the function m in an appropriate way. Dikta [4] used this possibility and estimated m parametrically:
Therefore, it is assumed that $m(x;\theta)$ belongs to a parametric family $\{m(x;\theta) : \theta \in \Theta\}$, where the continuous function $m(\cdot;\cdot)$ is known. Here, $\theta = (\theta_1, \ldots, \theta_k) \in \Theta$ is the k-dimensional parameter vector and $\Theta \subset \mathbb{R}^k$ denotes the parameter space. The unknown parameter vector θ is estimated using the sample data $(z_1, \ldots, z_n)$. The resulting point estimator is then denoted by $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^k$.

Estimating θ yields a semiparametric estimator of the cumulative hazard function (1.11)

$$\hat{\Lambda}_{SP}(t) = \int_0^t \frac{m(x;\hat{\theta})}{1 - H_n(x-)} dH_n(x) = \sum_{i:Z_{i:n} \le t} \left(\frac{m(Z_{i:n};\hat{\theta})}{n - i + 1} \right), \tag{1.12}$$

where $Z_{1:n} \le \ldots \le Z_{n:n}$ denote the order statistics of the Z-values.

Applying the same steps we used in order to obtain the approximation in (1.10) to the semiparametric estimator for the cumulative hazard function (1.12) finally leads to the semeiparametric estimator which has been proposed by Dikta [5]:

$$\hat{S}_{SP}(t) = \prod_{i:Z_{i:n} \le t} \left(1 - \frac{m(Z_{i:n};\hat{\theta})}{n - i + 1} \right). \tag{1.13}$$

Note that this estimator assigns mass to every observation, whereas in case of the Kaplan-Meier estimator (1.6) only uncensored data points are considered. Basically, the censoring indicator δ_i is substituted by the function $m(Z_{i:n}, \hat{\theta})$ for every i.

Interesting properties of the introduced estimator, such as consistency and asymptotic normality, the central limit theorem for the process $\sqrt{n}(\hat{\Lambda}_{SP}(t) - \Lambda(t))$ and weak convergence of the process $\sqrt{n}((1 - \hat{S}_{SP}) - F)$ to a centred Gaussian process were shown by Dikta [4]. The importance of this semiparametric estimator is due to its superiority to the Kaplan-Meier estimator with respect to asymptotic variance. Moreover, an extended version of the estimator can simply be applied when using covariates, as we see in the following section.

1.5 Estimating the Survival Function in the Presence of Covariates

Since our data set contains additional information which might have an impact on the occurrence of certain events, we want to incorporate these covariates in our estimations. Hence, a modification of the semiparametric estimator which was introduced in Section 1.4.3 is considered to estimate the survival probability in the presence of covariates. According to Külheim, Dikta and Ghorai [11] the covariates can simply be included in the function m from the semiparametric estimator (1.13) in order to derive the desired estimator

$$\hat{S}_{CV}(t) = \prod_{i:Z_{i:n} \leq t} \left(1 - \frac{m(Z_{i:n}, X_{[i:n]}; \hat{\theta})}{n - i + 1} \right), \tag{1.14}$$

where $Z_{1:n} \leq \ldots \leq Z_{n:n}$ denote the order statistics of the Z-values and $X_{[i:n]}$ is the concomitant of the i-th order statistic. In other words, if $Z_{i:n} = Z_j$ then $X_{[i:n]} = X_j$. Moreover,

$$m(z, x; \theta) = \mathbb{P}[\delta = 1 | Z = z, X = x] = \mathbb{E}[\delta | Z = z, X = X]$$

and $m(z, x; \hat{\theta})$ denotes the estimated function m which is derived in the following. First, we want to state the following remark:

Remark 1.5.1. *(ref.[11] Remark 1.1)*
The choice of the parametric form of $m(z, x; \theta)$ is very important because $\hat{S}_{CV}$ con-

verges to the survival function provided the assumed form of $m(x, z; \theta)$ *is the correct one.*

Since $m(z, x; \theta) = \mathbb{P}[\delta = 1 | Z = z, X = x] = \mathbb{E}[\delta | Z = z, X = x]$ it is reasonable to fit the multiple logistic regression model in order to find an estimator for the unknown parameter θ of $m(z, x; \theta)$.

1.5.1 The Multiple Logistic Regression Model[1]

Regression methods are a popular technique for data analysis which are concerned with describing the relationship between an outcome variable and a set of explanatory variables.
In our case we want to describe the relationship between the censoring indicator δ, i.e. the ability of observing an event, and the survival time with corresponding covariates. Since the outcome of δ is either 0, in case the data is censored, or 1 if we observed the actual survival time, the linear regression model $m(z, x; \theta) = \beta_0 + \beta_1' x + \beta_2 z$, where $\theta = (\beta_1, \beta_2, \beta_3)$, seems inappropriate due to the domain of m, which would not necessarily be bounded between 0 and 1.
Therefore, we apply the logit transformation on $m(z, x; \theta) = \mathbb{E}[\delta | X = x, Z = z]$:

$$\pi(z, x; \theta) = \ln\left(\frac{m(z, x; \theta)}{1 - m(z, x; \theta)}\right) = \beta_0 + \beta_1' x + \beta_2 z. \tag{1.15}$$

Solving for $m(x, z; \theta)$ leads to

$$m(z, x; \theta) = \frac{\pi(z, x; \theta)}{1 + \pi(z, x; \theta)} = \frac{\exp(\beta_0 + \beta_1' x + \beta_2 z)}{1 + \exp(\beta_0 + \beta_1' x + \beta_2 z)} \tag{1.16}$$

for arbitrary parameter $\beta_0 \in \mathbb{R}$, $\beta_1 \in \mathbb{R}^p$, $\beta_2 \in \mathbb{R}$ and $\theta = (\beta_1, \beta_2, \beta_3)$.

To estimate these regression coefficients, we use the maximum likelihood approach. In order to construct the likelihood function we consider the following:
δ is either 1 or 0 and $m(z, x; \theta)$ describes the vector of conditional probabilities $\mathbb{P}[\delta = 1 | X = x, Z = z]$. The probability of the complementary event is then

[1] This Section is based on the concepts in [7].

denoted by $\mathbb{P}[\delta = 0|X = x, Z = z] = 1 - m(z, x; \theta)$.
Consider now the n independent observations (Z_i, δ_i, X_i).
Since the likelihood function represents the probability of the observed data as a function of the unknown parameter, the contribution if $\delta_i = 1$ is $m(Z_i, X_i; \theta)$. For triples with $\delta_i = 0$, the contribution to the likelihood function is $1 - m(Z_i, X_i; \theta)$.
As the observations are independent, the likelihood function equals

$$L(\theta) = \prod_{i=1}^{n} m(Z_i, X_i; \theta)^{\delta_i} (1 - m(Z_i, X_i; \theta))^{1-\delta_i}. \tag{1.17}$$

In order to maximize L, we consider the log likelihood function $\ell(\theta) := \ln(L(\theta))$:

$$\ell(\theta) = \sum_{i=1}^{n} \delta_i \ln m(Z_i, X_i; \theta) + (1 - \delta_i) \ln (1 - m(Z_i, X_i; \theta)). \tag{1.18}$$

The values $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ which maximize the log likelihood function make up the maximum likelihood estimator. Then, $\hat{m}(z, x; \hat{\theta})$ with $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ is the estimator of the conditional probability that the data is censored, provided that x and z are known.

There are various possibilities testing for the significance of the logistic model. Selected methods, for instance the Wald test, are discussed within the context of application in Chapter 3.

Chapter 2

Estimating the Conditional Survival Function

At this point, an estimator for the survival function is achieved where additional information from covariates is incorporated. In order to predict probabilities of survival for arbitrary values of covariates, the conditional survival function

$$S(z|x) = \mathbb{P}[Z > z|X_1 = x_1, \ldots, X_p = x_p].$$

is considered.

To determine this quantity we propose an estimator $\tilde{S}(z|x)$. In order to give an idea how we came up with the proposal of the weighted semiparametric maximum likelihood estimator, we first derive the general nonparametric maximum likelihood estimator for the conditional survival function. Then we introduce a modification of this estimator, which was suggested by Inglesias-Pérez and Uña-Álvarez [8], before we show the maximum likelihood property of our proposed estimator.

Finally, a pointwise Likelihood Ratio confidence band is derived in a similar way it was proposed in Li and Van Keilegom [12] and we conclude this chapter illustrating properties and modifications of $\tilde{S}$.

2.1 The Semiparametric Maximum Likelihood Estimator for the Conditional Survival Function

First, the nonparametric estimator for the conditional survival function which forms the basis for the later derivation is motivated.

To begin with, the nonparametric kernel estimate of the conditional mean $\mu(x) = \mathbb{E}[Z|X = x]$ is explained by means of the following simple example:

Example 2.1.1 (Motivating the nonparametric kernel estimator)**.**
Consider n randomly selected cars of the same type with similar engines.
Suppose $Z_1, Z_2, \ldots, Z_n$ denote the survival time of the engine and let X be the average fuel consumption in liter per 100 kilometer.
Furthermore, assume that every car of the sample consumes 8 liter per 100 kilometer. Based on this setup, a nonparametric estimator of $\mu(8) = \mathbb{E}[Z|X = 8]$ is obtained by

$$\hat{\mu}(8) = \frac{1}{n}\sum_{i=1}^{n} Z_i = \sum_{i=1}^{n} W_i Z_i,$$

where $W_i := \frac{1}{n}$.
In other words, each Z_i contributes to $\hat{\mu}(8)$ by the same amount.
Since not every car consumes exactly the same amount of fuel, we need to adjust our estimator.
Therefore, we now assume that the cars do not have the same average fuel consumption. Furthermore, suppose only the first car in the sample has an average fuel consumption of 8 liter per 100 kilometer. If only data from cars which hit exactly 8 liter per 100 kilometer is used, the estimator looks like the following: $\hat{\mu}(8) = Z_1$. However, an estimated sample mean based on only one observation can not be very reliable.
Another approach is to assign more weight to the observations Z_i which have similar average fuel consumption and less weight than those where the average fuel consumption deviates more.

Thus, let $K(v)$ be a nonnegative kernel function and h_n denotes the bandwidth. Then, we define the weights by

$$W_i(x; h_n) = \frac{K(\frac{x-X_i}{h_n})}{\sum_{j=1}^{n} K(\frac{x-X_j}{h_n})}$$

and hence, a more appropriate nonparametric estimator of the conditional sample mean $\mu(x)$ based on all observations $Z_1, \ldots, Z_n$ is

$$\hat{\mu}(x) = \sum_{i=1}^{n} W_i(x; h_n) Z_i.$$

The same principle can be applied to derive a nonparametric estimator of the conditional survival function $S(z|x)$ based on the observations (Z_i, X_i), when not all covariates $X_i = x$.

Let $Z_{1:n} < Z_{2:n} < \cdots < Z_{n:n}$ denote the ordered values of $Z_1, \ldots, Z_n$ and $Z_{n+1:n} := \infty$. If the value of the covariate is the same for every observed survival time Z_i, that is $X_i = x$ for all $i = 1, \ldots, n$, then the conditional survival function $S(z|x)$ can be estimated by the sample proportion

$$\hat{S}(z|x) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{Z_i > z\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{Z_i > 0\}}}$$

$$= \left(\frac{\sum_{i=1}^{n} \mathbb{1}_{\{Z_i > Z_{1:n}\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{Z_i > 0\}}}\right) \left(\frac{\sum_{i=1}^{n} \mathbb{1}_{\{Z_i > Z_{2:n}\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{Z_i > Z_{1:n}\}}}\right) \cdots \left(\frac{\sum_{i=1}^{n} \mathbb{1}_{\{Z_i > Z_{r-1:n}\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{Z_i > Z_{r-2:n}\}}}\right) \left(\frac{\sum_{i=1}^{n} \mathbb{1}_{\{Z_i > Z_{r:n}\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{Z_i > Z_{r-1:n}\}}}\right)$$

$$= \left(1 - \frac{\sum_{i=1}^{n} \mathbb{1}_{\{Z_i = Z_{1:n}\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{Z_i > 0\}}}\right) \left(1 - \frac{\sum_{i=1}^{n} \mathbb{1}_{\{Z_i = Z_{2:n}\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{Z_i > Z_{1:n}\}}}\right) \cdots$$

$$\left(1 - \frac{\sum_{i=1}^{n} \mathbb{1}_{\{Z_i = Z_{r-1:n}\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{Z_i > Z_{r-2:n}\}}}\right) \left(1 - \frac{\sum_{i=1}^{n} \mathbb{1}_{\{Z_i = Z_{r:n}\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{Z_i > Z_{r-1:n}\}}}\right) \tag{2.1}$$

for $Z_{r:n} \leq z < Z_{r+1:n}$.

From equation (2.1) it is obvious that each Z_i contributes equally to the estimation of $S(z|x)$.

A more reasonable situation is that the covariates differ in their values. That means that the X_i's are not all equal to x. Therefore, we weight the terms $\mathbb{1}_{\{Z_i=Z_{j:n}\}}$ by $W_i(x) := W_i(x; h_n)$ for $i, j = 1, \ldots, n$. Then $\hat{S}(z|x)$ can be expressed by

$$\begin{aligned}\hat{S}(z|x) = &\left(1 - \frac{\sum_{i=1}^n W_i(x)\mathbb{1}_{\{Z_i=Z_{1:n}\}}}{\sum_{i=1}^n W_i(x)\mathbb{1}_{\{Z_i>0\}}}\right)\left(1 - \frac{\sum_{i=1}^n W_i(x)\mathbb{1}_{\{Z_i=Z_{2:n}\}}}{\sum_{i=1}^n W_i(x)\mathbb{1}_{\{Z_i>Z_{1:n}\}}}\right)\cdots \\ &\left(1 - \frac{\sum_{i=1}^n W_i(x)\mathbb{1}_{\{Z_i=Z_{r-1:n}\}}}{\sum_{i=1}^n W_i(x)\mathbb{1}_{\{Z_i>Z_{r-2:n}\}}}\right)\left(1 - \frac{\sum_{i=1}^n W_i(x)\mathbb{1}_{\{Z_i=Z_{r:n}\}}}{\sum_{i=1}^n W_i(x)\mathbb{1}_{\{Z_i>Z_{r-1:n}\}}}\right) \\ = &\left(1 - \frac{\sum_{i=1}^n W_i(x)\mathbb{1}_{\{Z_i=Z_{1:n}\}}}{\sum_{i=1}^n W_i(x)\mathbb{1}_{\{Z_i\geq Z_{1:n}\}}}\right)\left(1 - \frac{\sum_{i=1}^n W_i(x)\mathbb{1}_{\{Z_i=Z_{2:n}\}}}{\sum_{i=1}^n W_i(x)\mathbb{1}_{\{Z_i\geq Z_{2:n}\}}}\right)\cdots \\ &\left(1 - \frac{\sum_{i=1}^n W_i(x)\mathbb{1}_{\{Z_i=Z_{r-1:n}\}}}{\sum_{i=1}^n W_i(x)\mathbb{1}_{\{Z_i\geq Z_{r-1:n}\}}}\right)\left(1 - \frac{\sum_{i=1}^n W_i(x)\mathbb{1}_{\{Z_i=Z_{r:n}\}}}{\sum_{i=1}^n W_i(x)\mathbb{1}_{\{Z_i\geq Z_{r:n}\}}}\right).\end{aligned} \tag{2.2}$$

Note that the last factor in (2.2) equals zero for $z \geq Z_{n:n}$.
Since at most one term in the numerator of (2.2) is nonzero, $\hat{S}(z|x)$ can be rewritten as

$$\hat{S}(z|x) = \prod_{i:Z_i\leq z}\left(1 - \frac{W_i(x)}{\sum_{j=1}^n W_j(x)\mathbb{1}_{\{Z_j\geq Z_i\}}}\right). \tag{2.3}$$

As we consider p covariates in our model, we choose $K(\cdot)$ to be a multidimensional nonnegative kernel function and determine the weights as shown in the example above by $W_i(x) = \dfrac{K(\frac{x-X_i}{h_n})}{\sum_{j=1}^n K(\frac{x-X_j}{h_n})}$ for $i = 1, \ldots, n$, where h_n represents the vector of the corresponding bandwidth.

2.1.1 Estimator of the Conditional Survival Function by Virtue of Kaplan and Meier

The basis of nonparametric estimation from incomplete data was first introduced by Kaplan and Meier [9]. Since we pick up this idea to derive the weighted semi-parametric conditional estimator, we first give an overview of the derivation of the

general nonparametric maximum likelihood estimate.
The aim is to locally maximize the weighted likelihood function $L(S(\cdot|x)|(Z_i, \delta_i, X_i))$.

In the following derivation we allow ties and therefore introduce some additional notations in order to avoid any confusion:
Let $0 < T_1 < T_2 < \cdots < T_k < T_{k+1} := \infty$ represent k distinct event times among the Z-sample and let c_j denote the number of censored observations in the interval $[T_j, T_{j+1})$. Further, let n_j represent the number of uncensored observations at T_j. Then, the total number of observations adds up to $n = c_0 + c_1 + \cdots + c_k + n_1 + \cdots + n_k$. An illustration of the notation is given in Figure 2.1.

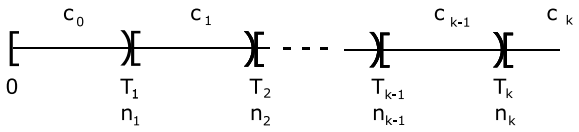

Figure 2.1: Distinct event times for the interval $[0, T_k]$ including the number of observations.

Moreover, let $C_j^{(i)}$ be the j-th censored observation in the interval $[T_i, T_{i+1})$ for $i = 0, \ldots, k$ and $j = 1, \ldots, c_i$ and $X_j^{(i)}$ denote the corresponding value for X. For the weights which are associated with the censored observations $(C_j^{(i)}, X_j^{(i)})$ we use the notation $W(x; C_j^{(i)}, X_j^{(i)}, h_n)$ where $i = 0, 1, \ldots, k$ and $j = 1, \ldots, c_i$ and similarly for the weights corresponding to the uncensored data $(T_i, X_j^{(i)})$ we use $W(x; T_i, X_j^{(i)}, h_n)$ for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$.

We first analyze the influence of the uncensored observations in order to derive the associated likelihood function.
The weighted contribution of the j-th uncensored observation at time T_i is

$$(S(T_i - |x) - S(T_i|x))^{W(x;T_i,X_j^{(i)},h_n)}.$$

Therefore, the total contribution of all uncensored observations at time T_i is

$$(S(T_i - |x) - S(T_i|x))^{\sum_{j=1}^{n_i} W(x;T_i,X_j^{(i)},h_n)}$$

and the total contribution from all uncensored observations at time points $T_1, \ldots, T_k$ is finally

$$\prod_{i=1}^{k} (S(T_i - |x) - S(T_i|x))^{\sum_{j=1}^{n_i} W(x;T_i,X_j^{(i)},h_n)} \tag{2.4}$$

and note that $S(0|x) = 1$.

Let us now consider the influence of the censored observations:
The contribution of the censored observations $(C_j^{(i)}, X_j^{(i)})$ in the time interval $[T_i, T_{i+1})$ is given by

$$\left(S(C_j^{(i)}|x)\right)^{W(x;C_j^{(i)},X_j^{(i)},h_n)}.$$

Then, the total input of all censored observations in $[T_i, T_{i+1})$ is

$$\prod_{j=1}^{c_i} \left(S(C_j^{(i)}|x)\right)^{W(x;C_j^{(i)},X_j^{(i)},h_n)}$$

and hence the total contribution of all censored observations can be expressed by

$$\prod_{i=0}^{k} \prod_{j=1}^{c_i} \left(S(C_j^{(i)}|x)\right)^{W(x;C_j^{(i)},X_j^{(i)},h_n)}. \tag{2.5}$$

Finally, the weighted likelihood function based on all censored and uncensored observations is given by the composition of equations (2.4) and (2.5)

$$L(S(\cdot|x)|(Z_i, \delta_i, X_i)) = \prod_{i=1}^{k} (S(T_i - |x) - S(T_i|x))^{\sum_{j=1}^{n_i} W(x;T_i,X_j^{(i)},h_n)} \times$$

$$\prod_{i=0}^{k} \prod_{j=1}^{c_i} \left(S(C_j^{(i)}|x)\right)^{W(x;C_j^{(i)},X_j^{(i)},h_n)}. \tag{2.6}$$

It was shown by Kaplan and Meier [9] that the likelihood function is maximized at a survival function that is a step function with jumps at the uncensored observations. Therefore, we only consider those $S(\cdot|x)$ which are constant on the interval $[T_i, T_{i+1})$

where $i = 0, \ldots, k$ in order to maximize the weighted likelihood function in (2.6). Choosing an $S(\cdot|x)$ which is constant on $[T_i, T_{i+1})$ can be done by setting

$$S(C_j^{(i)}|x)^{W(x;C_j^{(i)},X_j^{(i)},h_n)} = S(T_i|x)^{W(x;C_j^{(i)},X_j^{(i)},h_n)}.$$

Therefore, the likelihood function (2.6) reduces to

$$L(S(\cdot|x)|(Z_i,\delta_i,X_i)) = \prod_{i=1}^{k}(S(T_i - |x) - S(T_i|x))^{\sum_{j=1}^{n_i} W(x;T_i,X_j^{(i)},h_n)} \times$$

$$\prod_{i=0}^{k}\big(S(T_i|x)\big)^{\sum_{j=1}^{c_i} W(x;C_j^{(i)},X_j^{(i)},h_n)}. \tag{2.7}$$

Define $p_0 = 1$ and

$$p_i := \frac{S(T_i|x)}{S(T_i - |x)} = \frac{S(T_i|x)}{S(T_{i-1}|x)}$$

for $i = 1, \ldots, k$.

Further, let $T_0 = 0$ and $S(0|x) = 1$. Observe that for every fixed $1 \leq i \leq k$ the unknown conditional survival function which is to be estimated equals

$$S(T_i|x) = \left(\frac{S(T_i|x)}{S(T_{i-1}|x)}\right)\left(\frac{S(T_{i-1}|x)}{S(T_{i-2}|x)}\right)\cdots\left(\frac{S(T_2|x)}{S(T_1|x)}\right)\left(\frac{S(T_1|x)}{S(T_0|x)}\right)$$

$$= p_i p_{i-1} \cdots p_1. \tag{2.8}$$

Using these results leads to the following simplification for the censored contribution in equation (2.5):

$$\prod_{i=0}^{k}\big(S(T_i|x)\big)^{\sum_{j=1}^{c_i} W(x;C_j^{(i)},X_j^{(i)},h_n)} = \prod_{i=0}^{k}\big(p_1 p_2 \ldots p_i\big)^{\sum_{j=1}^{c_i} W(x;C_j^{(i)},X_j^{(i)},h_n)}$$

$$= p_1^{\sum_{i=1}^{k}\sum_{j=1}^{c_i} W(x;C_j^{(i)},X_j^{(i)},h_n)} \times p_2^{\sum_{i=2}^{k}\sum_{j=1}^{c_i} W(x;C_j^{(i)},X_j^{(i)},h_n)}$$

$$\times p_{k-1}^{\sum_{i=k-1}^{k}\sum_{j=1}^{c_i} W(x;C_j^{(i)},X_j^{(i)},h_n)} \times p_k^{\sum_{j=1}^{c_k} W(x;C_j^{(k)},X_j^{(k)},h_n)}$$

$$= \prod_{l=1}^{k} p_l^{\sum_{i=l}^{k}\sum_{j=1}^{c_i} W(x;C_j^{(i)},X_j^{(i)},h_n)}. \tag{2.9}$$

In order to rewrite the uncensored contribution in equation (2.4) we consider

$$\begin{aligned} S(T_i - |x) - S(T_i|x) &= S(T_{i-1}|x) - S(T_i|x) \\ &= S(T_{i-1}|x)\left(1 - \frac{S(T_i|x)}{S(T_{i-1}|x)}\right) = p_1 p_2 \cdots p_{i-1}(1-p_i) \end{aligned}$$

for $2 \leq i \leq k$, where the last equality holds because of equation (2.8) and $S(T_1 - |x) - S(T_1|x) = 1 - p_1$.

Substituting this result in (2.4) we obtain

$$\prod_{i=1}^{k}(S(T_i - |x) - S(T_i|x))^{\sum_{j=1}^{n_i} W(x;T_i,X_j^{(i)},h_n)}$$

$$= \prod_{i=1}^{k}((p_1 p_2 \ldots p_{i-1})(1-p_i))^{\sum_{j=1}^{n_i} W(x;T_i,X_j^{(i)},h_n)}$$

$$= \left(\prod_{i=1}^{k}(1-p_i)^{\sum_{j=1}^{n_i} W(x;T_i,X_j^{(i)},h_n)}\right)\left(\prod_{i=2}^{k}(p_1 p_2 \ldots p_{i-1})^{\sum_{j=1}^{n_i} W(x;T_i,X_j^{(i)},h_n)}\right)$$

$$= \left(\prod_{i=1}^{k}(1-p_i)^{\sum_{j=1}^{n_i} W(x;T_i,X_j^{(i)},h_n)}\right)\left(p_1^{\sum_{i=2}^{k}\sum_{j=1}^{n_i} W(x;T_i,X_j^{(i)},h_n)} p_2^{\sum_{i=3}^{k}\sum_{j=1}^{n_i} W(x;T_i,X_j^{(i)},h_n)} \times \right.$$

$$\left. \cdots \times p_{k-2}^{\sum_{i=k-1}^{k}\sum_{j=1}^{n_i} W(x;T_i,X_j^{(i)},h_n)} p_{k-1}^{\sum_{j=1}^{n_k} W(x;T_k,X_j^{(k)},h_n)}\right)$$

$$= \left(\prod_{i=1}^{k}(1-p_i)^{\sum_{j=1}^{n_i} W(x;T_i,X_j^{(i)},h_n)}\right)\left(\prod_{l=1}^{k-1} p_l^{\sum_{i=l+1}^{k}\sum_{j=1}^{n_i} W(x;T_i,X_j^{(i)},h_n)}\right) \tag{2.10}$$

Replacing these two quantities from (2.9) and (2.10) in equation (2.7) yields the

weighted likelihood function

$$L(S(\cdot|x)|(Z_i,\delta_i,X_i)) = \left(\prod_{l=1}^{k} p_l^{\sum_{i=l}^{k}\sum_{j=1}^{c_i} W(x;C_j^{(i)},X_j^{(i)},h_n)+\sum_{i=l+1}^{k}\sum_{j=1}^{n_i} W(x;T_i,X_j^{(i)},h_n)}\right)$$

$$\times \left(\prod_{i=1}^{k}(1-p_i)^{\sum_{j=1}^{n_i} W(x;T_i,X_j^{(i)},h_n)}\right) \tag{2.11}$$

Then, the maximum likelihood estimates of $p_1,\ldots,p_k$ are obtained by solving the equations

$$\frac{\partial \ln L(S(\cdot|x)|(Z_i,\delta_i,X_i))}{\partial p_l} = 0,$$

for $l = 1,\ldots,k$.

We obtain

$$\hat{p}_l =$$

$$\left(1 - \frac{\sum_{j=1}^{n_l} W(x;T_l,X_j^{(l)},h_n)}{\sum_{i=l}^{k}\sum_{j=1}^{c_i} W(x;C_j^{(i)},X_j^{(i)},h_n)+\sum_{l+1}^{k}\sum_{j=1}^{n_i} W(x;T_i,X_j^{(i)},h_n)+\sum_{j=1}^{n_l} W(x;T_l,X_j^{(l)},h_n)}\right) \tag{2.12}$$

and for $l = k$ this expression reduces to

$$\hat{p}_k = \left(1 - \frac{\sum_{j=1}^{n_k} W(X;T_k,X_j^{(k)},h_n)}{\sum_{j=1}^{c_k} W(x;C_j^{(k)},X_j^{(k)},h_n)+\sum_{j=1}^{n_k} W(x;T_k,X_j^{(k)},h_n)}\right).$$

Hence, from equation (2.8), it follows that the nonparametric maximum likelihood estimator of $S(z|x)$ based on the data $\{(Z_i,\delta_i,X_i), i = 1,\ldots,n\}$ is given by

$$\hat{S}(z|x) = \begin{cases} 1, & \text{if } 0 \le z < T_1 \\ \prod_{j=1}^{i}\hat{p}_j, & \text{if } T_i \le z < T_{i+1}, \end{cases} \tag{2.13}$$

for $1 \le i \le k$.

Note that if there is no censored observation after or at time T_k, the last time point where a failure is observed, then

$$\hat{p}_k = 1 - \frac{\sum_{j=1}^{n_k} W(x;T_k,X_j^{(k)},h_n)}{\sum_{j=1}^{n_k} W(x;T_k,X_j^{(k)},h_n)} = 0.$$

In this case, $\hat{S}(z|x)$ is a proper survival function, since $\hat{S}(z|x) = 0$ for all $z \in [T_k, \infty)$. However, if there is at least one censored observation which is greater than or equal to the largest uncensored observation T_k, then $c_k > 0$ and $\hat{p}_k > 0$ and therefore, $\hat{S}(z|x) = \hat{p_1}\hat{p_2}\ldots\hat{p_k} > 0$ for all $z \in [T_k, \infty)$. Hence, $\hat{S}(z|x)$ is not a proper survival function.

Note that in the absence of ties among the uncensored observations, that is if all failures are distinct, and weight is only given to uncensored data, $\hat{S}(z|x)$ reduces to an estimator which was proposed by Dabrowska [3]

$$\begin{aligned}\hat{S}(z|x) &= \prod_{i:Z_i \le z;\delta_i=1} \left(1 - \frac{W_i(x;h_n)}{\sum_{j=1}^n \mathbb{1}_{\{Z_j \ge Z_i\}} W_j(x;h_n)}\right) \\ &= \prod_{i:Z_i \le z} \left(1 - \frac{\delta_i W_i(x;h_n)}{\sum_{j=1}^n \mathbb{1}_{\{Z_j \ge Z_i\}} W_j(x;h_n)}\right) \\ &= \prod_{i:Z_i \le z} \left(1 - \frac{W_i(x;h_n)}{\sum_{j=1}^n \mathbb{1}_{\{Z_j \ge Z_i\}} W_j(x;h_n)}\right)^{\delta_i} \qquad (2.14)\end{aligned}$$

Comparing this Kernel Conditional Kaplan-Meier Estimate to the estimator in (2.3), weight is given only to those observations which are uncensored.
As a consequence we observe great variance in estimation in the presence of moderate to heavy censoring which motivates a search for alternative estimators.
Inspired by the results of Dikta [4] which were introduced in Section 1.4.3 and Dabrowska [3], Inglesias-Pérez and de Uña-Álvarez [8] derived the following semiparametric estimator for the conditional cumulative hazard function

$$\hat{\Lambda}(z|x) = \sum_{i:Z_i \le z} \frac{m(Z_i, x; \hat{\theta}) W_i(x;h_n)}{\sum_{j=1}^n \mathbb{1}_{\{Z_j \ge Z_i\}} W_j(x;h_n)}, \qquad (2.15)$$

where the function m appears because of the considered semiparametric conditional censorship model which was already introduced in Section 1.4.3. The estimator for the conditional survival function is then given by:

$$\hat{S}(z|x) = \prod_{i:Z_i \leq z} \left(1 - \frac{m(Z_i, x; \hat{\theta}) W_i(x; h_n)}{\sum_{j=1}^{n} \mathbb{1}_{\{Z_j \geq Z_i\}} W_j(x; h_n)}\right). \tag{2.16}$$

The same idea that was observed in Dikta's derivation of the semiparametric unconditional estimator in Section 1.4.3 is applied: Censoring indicators δ_i are substituted by $m(Z_i, x; \hat{\theta})$. Note that this estimator reduces to Dikta's estimator (1.13) in the absence of covariates.

Iglesias-Pérez and de Uña-Álvarez [8] established an asymptotic representation and showed asymptotic normality of the estimator in (2.16).

In contrast to the estimator in (2.14), this estimator has jumps not only on all uncensored observations, but on all observations; however, it fails to be a proper survival function, in the sense that for $t \to \infty$, $S(t|x) \nrightarrow 0$ in general, where x is fixed. Furthermore, when the last observation in the study is very volatile such that $W_j(x; h_n) = 0$ for these observations, the denominator eventually equals zero and therefore, the estimator results in an undefined term. Let us consider the following example: Let $K\big(\frac{x - X_{[n:n]}}{h_n}\big) = 0$. It follows that $W_{[n:n]}(x; h_n) = \dfrac{K\big(\frac{x - X_{[n:n]}}{h_n}\big)}{\sum_{j=1}^{n} K\big(\frac{x - X_j}{h_n}\big)} = 0$ and thus, $\sum_{i=1}^{n-1} W_{[i:n]}(x; h_n) = 1$. In the denominator of (2.16) one obtains $\sum_{j=1}^{n} \mathbb{1}_{\{Z_{j:n} \geq Z_{n:n}\}} W_{[j:n]}(x; h_n) = W_{[n:n]}(x; h_n) = 0$ for the very last multiplicand. Note that $Z_{1:n} \leq \cdots \leq Z_{n:n}$ are the order statistics of $Z_1, \ldots Z_n$ and $X_{[i:n]}, W_{[i:n]}$ are the concomitants of the i-th order statistic. That is, if $Z_{i:n} = Z_j$ then $X_{[i:n]} = X_j$ and $W_{[i:n]}(x; h_n) = \dfrac{K\big(\frac{x - X_{[i:n]}}{h_n}\big)}{\sum_{l=1}^{n} K\big(\frac{x - X_l}{h_n}\big)}$.

Because of these drawbacks, we propose the following estimator $\tilde{S}(z|x)$ which fulfills all the required properties (see Section 2.3) and where problems regarding dividing by zero are avoided due to modification regarding standardization and weights.

Let us first introduce the modified cumulative conditional hazard rate

$$\tilde{\Lambda}(z|x) = \sum_{i:Z_i \leq z} \frac{\tilde{m}(Z_i, x; \hat{\theta}) W_i(x; h_n)}{\sum_{j=1}^{n} \mathbb{1}_{\{Z_j \geq Z_i\}} \tilde{m}(Z_j, x; \hat{\theta}) W_j(x; h_n)} \tag{2.17}$$

and the corresponding estimator for the conditional survival function is given by

$$\tilde{S}(z|x) = \begin{cases} 1, & \text{if } 0 \leq z < Z_{1:n} \\ \prod_{i:Z_i \leq z} \left(1 - \frac{\tilde{m}(Z_i, x; \hat{\theta}) W_i(x; h_n)}{\sum_{j=1}^{n} \mathbb{1}_{\{Z_j \geq Z_i\}} \tilde{m}(Z_j, x; \hat{\theta}) W_j(x; h_n)}\right), & \text{if } Z_{i:n} \leq z < Z_{i+1:n} \end{cases} \tag{2.18}$$

which can be derived in the same way, as it was shown in the unconditional framework for equation (1.10).

Define the expression $\tilde{m}(Z_i, x; \hat{\theta}) := \delta_i m(Z_i, x; \hat{\theta}) + (1 - \delta_i)(1 - m(Z_i, x; \hat{\theta}))$ such that we distinguish between censored and uncensored data regarding the weights, but nevertheless weight all of the observations.
This is meaningful, since the contribution for the likelihood function for uncensored data is $m(Z_i, x : \hat{\theta})$ and for censored observations $1 - m(Z_i, x : \hat{\theta})$ respectively, as it was elaborated in Section 1.5.1.

In the following section it is shown that this newly proposed estimator $\tilde{S}(z|x)$ is the semiparametric maximum likelihood estimator.

2.1.2 The Maximum Likelihood Property

Below, the maximum likelihood estimator for the weighted likelihood function is derived by considering similar steps as in Section 2.1.1.
However, there is a significant difference between the mass which is given to the observed data. As stated above, in this approach every observation, regardless of being censored or uncensored, contributes to the weighted semiparametric likelihood function.
In the following elaboration, we do not allow ties and therefore, let $Z_1 < \cdots < Z_n < Z_{n+1} := \infty$ denote the ordered values of the Z-sample.

These thoughts result in the semiparametric weighted likelihood function

$$L(S(\cdot|x)|(Z_i,\delta_i,X_i)) = \prod_{i=1}^{n} (f(Z_i|x))^{\tilde{m}(Z_i,x;\hat{\theta})W_i(x;h_n)}$$

$$= \prod_{i=1}^{n} (S(Z_i - |x) - S(Z_i|x))^{(\delta_i m(Z_i,x;\hat{\theta})+(1-\delta_i)(1-m(Z_i,x;\hat{\theta}))W_i(x;h_n)}. \tag{2.19}$$

Recall that in order to maximize L only those functions $S(\cdot|x)$ which are constant on the intervals $[Z_i, Z_{i+1})$ for $i = 0, 1, \ldots, n$ need to be considered.

From equation (2.8), we obtain that for every fixed $1 \leq i \leq n$ the unknown conditional survival function, which is to be estimated, equals

$$S(Z_i|x) = p_i p_{i-1} \cdots p_1, \tag{2.20}$$

and $Z_0 = 0$ and $S(0|x) = 1$.
Using this result leads to

$$S(Z_i - |x) - S(Z_i|x) = p_1 p_2 \cdots p_{i-1}(1 - p_i)$$

for $2 \leq i \leq n$ and $S(Z_1 - |x) - S(Z_1|x) = 1 - p_1$.

And finally, replacing these two quantities in (2.19) yields

$$L(S(\cdot|x)|(Z_i,\delta_i,X_i)) = \prod_{i=1}^{n} ((p_1 p_2 \cdots p_{i-1})(1 - p_i))^{\tilde{m}(Z_i,x;\hat{\theta})W_i(x;h_n)}$$

$$= \left(\prod_{i=1}^{n} (1 - p_i)^{\tilde{m}(Z_i,x;\hat{\theta})W_i(x;h_n)} \right) \cdot \left(\prod_{i=2}^{n} (p_1 p_2 \cdots p_{i-1})^{\tilde{m}(Z_i,x;\hat{\theta})W_i(x;h_n)} \right)$$

$$= \left(\prod_{i=1}^{n} (1 - p_i)^{\tilde{m}(Z_i,x;\hat{\theta})W_i(x;h_n)} \right) \cdot \left(\prod_{i=1}^{n-1} p_i^{\sum_{j=i+1}^{n} \tilde{m}(Z_j,x;\hat{\theta})W_j(x;h_n)} \right). \tag{2.21}$$

In order to obtain the maximum likelihood estimates $\hat{p}_1, \ldots, \hat{p}_n$ for $p_1, \ldots, p_n$, we solve the equations

$$\frac{\partial \ln L(S(\cdot|x)|(Z_i,\delta_i,X_i))}{\partial p_l} = 0$$

for $l = 1, \ldots, n$ which results in

$$\hat{p}_l = \left(1 - \frac{(\delta_i m(Z_i, x; \hat{\theta}) + (1-\delta_i)(1 - m(Z_i, x; \hat{\theta}))W_i(x; h_n)}{\sum_{j=i}^{n}(\delta_j m(Z_j, x; \hat{\theta}) + (1-\delta_j)(1 - m(Z_j, x; \hat{\theta}))W_j(x; h_n)}\right).$$

Note that this quantity reduces to $\hat{p}_n = 0$ for $l = n$.

Replacing the p_i's in equation (2.20) by those estimates, the semiparametric maximum likelihood estimator of $S(z|x)$ based on the data $\{(Z_i, \delta_i, X_i), i = 1, \ldots, n\}$ is given by

$$\tilde{S}(z|x) = \begin{cases} 1, & \text{if } 0 \leq z < Z_{1:n} \\ \prod_{i:Z_i \leq z}\left(1 - \frac{\tilde{m}(Z_i, x; \hat{\theta})W_i(x; h_n)}{\sum_{j=1}^{n} \mathbb{1}_{\{Z_j \geq Z_i\}}\tilde{m}(Z_j, x; \hat{\theta})W_j(x; h_n)}\right), & \text{if } Z_{i:n} \leq z < Z_{i+1:n} \end{cases} \tag{2.22}$$

and $2 \leq i \leq n$.

2.2 Likelihood Ratio Confidence Bands for the Newly Proposed Estimator

In order to get an idea how accurately the estimator $\tilde{S}(z|x)$ works, we derive a confidence interval for the conditional survival function $S(z|x)$ for a fixed time point z. Pasting together an appropriate set of confidence intervals for different values of z leads to the pointwise confidence band.

Confidence intervals and confidence bands can be obtained based on normal approximations, see Li and Van Keilegom [12]. However, this method has some drawbacks. First, these confidence intervals do not provide values exclusively from the interval $[0, 1]$, in general. Second, for small sample sizes, normal confidence intervals are somewhat misleading.

A nonparametric Likelihood Ratio method was proposed by Thomas and Grunkemeier [16] in the absence of covariates. Based on a Likelihood Ratio test statistic R, all values of p, for which the null hypothesis $H_0 : S(a) = p$ is not rejected, are

included in the confidence interval. Using this method, confidence intervals with values in $[0, 1]$ exclusively are obtained as well as better small sample performance than confidence intervals based on the normal approximations do achieve.
Li and Van Keilegom [12] extended this method in order to derive confidence bands for the nonparametric estimator from equation (2.14)

$$\hat{S}(z|x) = \prod_{i:Z_i \leq z} \left(1 - \frac{\delta_i W_i(x; h_n)}{\sum_{j=1}^{n} \mathbb{1}_{\{Z_j \geq Z_i\}} W_j(x; h_n)}\right)$$

based on right censored data.
In the following, we derive confidence bands for the proposed estimator $\tilde{S}(z|x)$ from equation (2.18) applying the Likelihood Ratio method in a similar way.

The Likelihood Ratio confidence interval is based on a pivotal quantity derived from the Likelihood Ratio test statistic which is used for testing the null hypothesis

$$H_0 : S(t|x) = p.$$

Let $\mathscr{S}$ be the class of all survival functions supported on $(0, \infty)$ and

$$\mathscr{S}_0 = \{S(\cdot|x) | S(t|x) = p, S(\cdot|x) \in \mathscr{S}\}$$

denote the class of survival functions such that $S(t|x) = p$.
Furthermore, let $L(S(\cdot|x)|(Z_i, \delta_i, X_i))$ be the weighted likelihood function based on the given sample. Then, the Likelihood Ratio test statistic is represented by

$$R(p, t|x) = \frac{\sup_{S(\cdot|x) \in \mathscr{S}_0} L(S(\cdot|x)|(Z_i, \delta_i, X_i))}{\sup_{S(\cdot|x) \in \mathscr{S}} L(S(\cdot|x)|(Z_i, \delta_i, X_i))}. \tag{2.23}$$

In order to obtain $\sup_{S(\cdot|x) \in \mathscr{S}_0} L(S(\cdot|x)|(Z_i, \delta_i, X_i))$ we need to maximize the likelihood function in (2.21) under the constraint $S(t|x) = p$. Therefore, let D(t) denote the number of all events that occurred in the interval $(0, t]$ ($D(\infty) = D$). We assume that there is at least one failure in the interval $(0, t]$, such that $D(t) \geq 1$. In Section 2.1.2 we discussed that the maximum of the likelihood function is obtained at some $S(\cdot|x)$ which is constant on each interval $[Z_i, Z_{i+1})$. Hence, without loss of generality we assume that $S(t|x) = S(Z_{D(t)}|x)$.

From equation (2.20) it follows that the null hypothesis can be rewritten in the following way:

$$H_0 : S(t|x) = p \equiv H_0 : \prod_{j=1}^{D(t)} p_j = p.$$

The constrained maximum likelihood estimates can be calculated by applying the Lagrange multiplier method to the constrained weighted logarithmic likelihood function

$$\ln L_{\mathscr{S}_0}(p_1, \ldots, p_n, \lambda) = \sum_{i=1}^{n} (d_i \ln p_i + e_i \ln(1 - p_i)) + \lambda \left(\sum_{i=1}^{D(t)} \ln p_i - \ln p \right),$$

where $e_i = \tilde{m}(Z_i, x; \hat{\theta}) W_i(x; h_n) = \delta_i m(Z_i, x; \hat{\theta}) + (1 - \delta_i)(1 - m(Z_i, x; \hat{\theta})) W_i(x; h_n)$ for $i = 1, \ldots, n$ and $d_i = \sum_{j=i+1}^{n} e_j$, for $i = 1, \ldots, n-1$ and $d_n = 0$.

Solving $\left(\dfrac{\partial \ln L_{\mathscr{S}_0}(p_1, \ldots, p_n, \lambda)}{\partial p_i} \right) = 0$ yields the restricted maximum likelihood estimates

$$\tilde{p_i} = \begin{cases} 1 - \dfrac{e_i}{d_i + e_i + \lambda} & \text{for } i = 1, \ldots, D(t) \\ 1 - \dfrac{e_i}{d_i + e_i} & \text{for } i = D(t) + 1, \ldots, n. \end{cases}$$

Obviously, $\tilde{p_i} = \hat{p_i}$, for $i = D(t) + 1, \ldots, n$, where the $\hat{p_i}$'s denote the maximum likelihood estimates of the usual weighted maximum likelihood function, see Section 2.1.2.

Using equation (2.21) it follows that

$$\sup_{S(\cdot|x) \in \mathscr{S}_0} L(S(\cdot|x)|(Z_i, \delta_i, X_i)) = \sup_{p_1, \ldots, p_n : \prod_{i=1}^{D(t)} p_i = p} L(p_1, \ldots, p_n | (Z_i, \delta_i, X_i))$$

$$= \left(\prod_{i=1}^{D(t)} \tilde{p_i}^{d_i} (1 - \tilde{p_i})^{e_i} \right) \left(\prod_{i=D(t)+1}^{n} \hat{p_i}^{d_i} (1 - \hat{p_i})^{e_i} \right). \qquad (2.24)$$

After this preliminary work, the Likelihood Ratio test statistic can be represented by:

$$R(p, t, \lambda | x) = \frac{\sup_{S(\cdot|x) \in \mathscr{S}_0} L(S(\cdot|x)|(Z_i, \delta_i, X_i))}{\sup_{S(\cdot|x) \in \mathscr{S}} L(S(\cdot|x)|(Z_i, \delta_i, X_i))}$$

$$= \frac{\left(\prod_{i=1}^{D(t)} \tilde{p_i}^{d_i}(1-\tilde{p_i})^{e_i}\right)\left(\prod_{i=D(t)+1}^{n} \hat{p_i}^{d_i}(1-\hat{p_i})^{e_i}\right)}{\prod_{i=1}^{n} \hat{p_i}^{d_i}(1-\hat{p_i})^{e_i}}$$

$$= \prod_{i=1}^{D(t)} \left(\frac{\tilde{p_i}}{\hat{p_i}}\right)^{d_i} \left(\frac{1-\tilde{p_i}}{1-\hat{p_i}}\right)^{e_i}. \tag{2.25}$$

In order to determine the confidence intervals, we investigate this test statistic. More precisely, we are interested in the behavior of R by varying values of λ. For simplicity, we investigate the logarithmic Likelihood Ratio test statistic

$$\ln R(p,t,\lambda|x) = \sum_{i=1}^{D(t)} d_i \ln\left(\frac{\tilde{p_i}}{\hat{p_i}}\right) + e_i \ln\left(\frac{1-\tilde{p_i}}{1-\hat{p_i}}\right)$$

$$= \sum_{i=1}^{D(t)} d_i \ln\left(\left(\frac{d_i+e_i}{d_i}\right)\left(\frac{d_i+\lambda}{d_i+e_i+\lambda}\right)\right) + \sum_{i=1}^{D(t)} e_i \ln\left(\frac{d_i+e_i}{d_i+e_i+\lambda}\right).$$

In order to observe the manner of this function when changing λ, we derive the partial derivative

$$\frac{\partial \ln R(p,t,\lambda|x)}{\partial \lambda} = \sum_{i=1}^{D(t)} -\frac{e_i\lambda}{(d_i+\lambda)(e_i+d_i+\lambda)}.$$

Note that the partial derivative equals zero, if $\lambda = 0$.

Furthermore, the second derivative of the logarithmic test statistic is less than zero for $\lambda = 0$, since

$$\frac{\partial^2 \ln R(p,t,\lambda|x)}{\partial \lambda^2} = -\sum_{i=1}^{n} \frac{e_i(d_i^2+e_id_i-\lambda^2)}{(d_i+\lambda)^2(d_i+e_i+\lambda)^2}$$

and $e_i > 0$ for all $i = 1,\ldots,n$.

Therefore, $\ln R$ and hence R has a maximum at $\lambda = 0$. Obviously, the function $-2nh_n\dfrac{\hat{f}_X(x)}{\int K^2(u)du}\ln R(p,t,\lambda|x)$ has a minimum at $\lambda = 0$, where f_X denotes the density of the covariates X and K the nonnegative kernel function as introduced before. We claim that similar to Li and Van Keilegom [12] it can be shown that

$$-2nh_n\frac{\hat{f}_X(x)}{\int K^2(u)du}\ln R(p,t,\lambda|x)$$

is asymptotically chi-squared distributed with one degree of freedom.

Then, the asymptotic Likelihood Ratio confidence set of level α for the parameter λ is given by

$$\{\lambda : -2nh_n \frac{\hat{f}_X(x)}{\int K^2(u)du} \ln R(p,t,\lambda|x) \leq \chi^2_{1,1-\alpha}\}$$

for $\alpha \in (0,1)$, where $\chi^2_{1,1-\alpha}$ denotes the $(1-\alpha)$ chi-squared quantile with one degree of freedom.
In Appendix A.1 we proved that there is a $\lambda_L < 0$ and $\lambda_U > 0$ such that

$$\mathbb{P}[\{\lambda(p)|\lambda_L \leq \lambda(p) \leq \lambda_U\}] = 1-\alpha,$$

where λ_L and λ_U are obtained by solving

$$-2nh_n \frac{\hat{f}_X(x)}{\int K^2(u)du} \ln R(p,t,\lambda|x) \leq \chi^2_{1,1-\alpha},$$

if $0 < \tilde{S}(t|x) < 1$.
The interval boundaries of the closed interval $[\lambda_L, \lambda_U]$ are $0 = \lambda_L < \lambda_U$, if $\tilde{S}(t|x) = 0$.
In order to obtain a confidence interval $[p_L, p_U]$ for p, we recall the function

$$\prod_{i=1}^{D(t)} \tilde{p}_i(\lambda) = \prod_{i=1}^{D(t)} \frac{d_i + \lambda}{d_i + e_i + \lambda} = p.$$

As shown in Appendix A.2, this function is increasing in λ and therefore, the corresponding interval for $S(t|x) = p$ is also a closed interval of the form $[p_L, p_U]$ with $0 < p_L < p_U < 1$, if $\tilde{S}(t|x) \in (0,1)$, and $0 = p_L < p_U < 1$, if $\tilde{S}(t|x) = 0$. Hence, the piecewise constant interval boundaries p_L and p_U are given by

$$p_L = \prod_{i=1}^{D(t)} \frac{d_i + \lambda_L}{d_i + e_i + \lambda_L} \quad \text{and} \quad p_U = \prod_{i=1}^{D(t)} \frac{d_i + \lambda_U}{d_i + e_i + \lambda_U}.$$

To achieve pointwise confidence bands, we calculate the confidence interval which was established above for appropriate times t.

2.3 Properties and Modifications of the Established Estimator

2.3.1 Properties

Note that the estimator which was introduced in (2.18) represents a proper survival function with $\tilde{S}(0|x) = 1$ and $\tilde{S}(z|x) = 0$ for any $z \geq Z_{n:n}$.

$$\begin{aligned}
\tilde{S}(t|x) &= \prod_{i=1}^{n} \left(1 - \frac{\mathbb{1}_{\{Z_i \leq t\}} \tilde{m}(Z_i, x; \hat{\theta}) W_i(x; h_n)}{\sum_{j=1}^{n} \mathbb{1}_{\{Z_j \geq Z_i\}} \tilde{m}(Z_j, x; \hat{\theta}) W_j(x; h_n)} \right) \\
&\underset{t \to \infty}{\longrightarrow} \prod_{i=1}^{n} \left(1 - \frac{\tilde{m}(Z_i, x; \hat{\theta}) W_i(x; h_n)}{\sum_{j=1}^{n} \mathbb{1}_{\{Z_j \geq Z_i\}} \tilde{m}(Z_j, x; \hat{\theta}) W_j(x; h_n)} \right) \\
&= \prod_{i: Z_i < Z_{n:n}} \left(1 - \frac{\tilde{m}(Z_i, x; \hat{\theta}) W_i(x; h_n)}{\sum_{j=1}^{n} \mathbb{1}_{\{Z_j \geq Z_i\}} \tilde{m}(Z_j, x; \hat{\theta}) W_j(x; h_n)} \right) \\
&\quad \times \left(1 - \frac{\tilde{m}(Z_{n:n}, x; \hat{\theta}) W_{[n:n]}(x; h_n)}{\sum_{j=1}^{n} \mathbb{1}_{\{Z_j \geq Z_{n:n}\}} \tilde{m}(Z_j, x; \hat{\theta}) W_j(x; h_n)} \right) \\
&= 0
\end{aligned}$$

Furthermore, $\tilde{S}$ is monotonically decreasing in the range from 0 to 1.
The estimator has jumps at all observed data points and maximizes a locally weighted likelihood function among all survival functions which have jumps at the observed data points. Moreover, Likelihood Ratio confidence intervals with lower and upper bounds that have values between zero and one exclusively can be constructed. And last but not least, the newly proposed estimator incorporates covariates through the terms $\tilde{m}_i$ and W_i respectively. In order to obtain appropriate weights W_i, the distance to the "neighbours" of each covariate is taken into account.

2.3.2 Modifications

In the following paragraph, a modification of $\tilde{S}$ which arose while developing this estimator is addressed.

Having a closer look at the weights W_i one might observe that it is reasonable to weight the terms in the likelihood function (2.21) by $\frac{1}{W_i}$ instead of W_i. Note that both the values of probabilities being weighted and the weights themselves range from zero to one. Since observations which include covariates closer to x should gain more influence than observations whose covariates differ greatly from x on the estimation, the former should be weighted by the bigger weight which is $\frac{1}{W_i}$ for the considered range of numbers. Since the reciprocal of the weights does not sum up to one, we standardize the values in the following way:

$$B_i(x; h_n) := \frac{\frac{1}{W_i(x; h_n)}}{\sum_{j=1}^{n} \frac{1}{W_j(x; h_n)}}$$

Applications of this modification and comparisons to the derived estimator are provided in the following chapter.

Chapter 3

Application - Analyzing Data from an Automotive Company

In this chapter, the semiparametric estimators are applied and illustrated on the data set from an automotive environment which was introduced in Chapter 1. Furthermore, a corresponding Likelihood Ratio confidence band is derived and modifications of the estimator as well as comparisons to the other illustrated estimators are demonstrated.

Recall that the data set consists of $8,879$ cars. For each of those cars there are observations of mileage, which is considered as the survival time, censoring times and explanatory variables such as average fuel consumption or average speed: (Z_i, δ_i, X_i) All of the considered covariates are one dimensional, but some of them contain information which is divided in several classes, for instance engine revolutions. This covariate is separated in four different classes. In this example the current state of the vehicle is measured with a frequency. We know how much time a vehicle spent in each of the specific classes: $0 - 500$ rotations, $500 - 1500$ rotations, $1500 - 3000$ rotations and the number of rotations that exceed 3000. For simplicity, these classes are combined. In the example above we then end up with two remaining classes: engine revolutions between 0 and 1500 rotations and number of rotations higher than 1500. Only one of these two classes has to be considered in the following evaluation, since it contains already all needed information.

3.1 The Semiparametric Estimator in Comparison to the Kaplan-Meier Estimator

Since the examined data set is censored, we first apply the nonparametric Kaplan-Meier Estimator (1.6) to the observations (Z_i, δ_i). Figure 3.1 displays a plot of the resulting estimated survival curve.

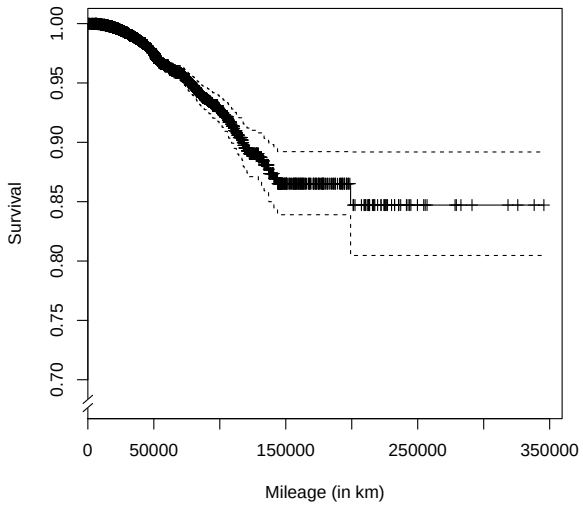

Figure 3.1: Estimated survival function using the Kaplan-Meier estimator $\hat{S}_{KM}$.

Moreover, Dikta's semiparametric estimator (1.13) is applied to the data. A comparison of the Kaplan-Meier Estimator and this semiparametric estimator is illustrated in Figure 3.2.

One observes that in the beginning, these two estimators are similar, whereas with increasing mileage they diverge. This behavior is reasonable, since the Kaplan-Meier estimator gives mass only to uncensored data, whereas with Dikta's semiparametric estimator, all observations are considered and reasonably weighted. Nevertheless, both estimators do not show the properties we expect from a survival function. More precisely, the functions equal one at mileage zero, but lack the property of being

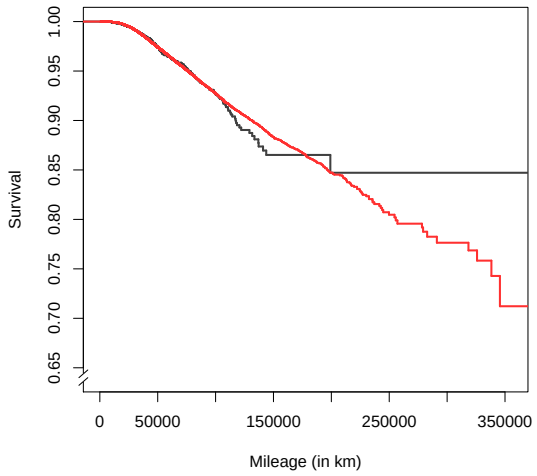

Figure 3.2: Estimated survival function: Dikta's semiparametric estimator $\hat{S}_{SP}$ (—) and the Kaplan-Meier estimator $\hat{S}_{KM}$ (—) in comparison.

equal to zero for a mileage greater or equal than the last observation $Z_{n:n}$.

Besides the observed survival time and the censoring indicator, the data set contains further information in form of covariates. The application of Dikta's semiparametric estimator in the presence of covariates (1.14) is illustrated in Figure 3.3.
Noticeable is the stagnation already at $250,000$ km which is reasonable due to the influence of additional information which decreases the survival probability. Also, more jumps in less time are plausible by considering all observations instead of just the uncensored ones.

3.2 Variable Selection

Indeed, we wish to include all available information that has an impact on the occurrence of an event in our model in order to estimate the reliability as closely as possible. On the other hand, including much additional information leads to a

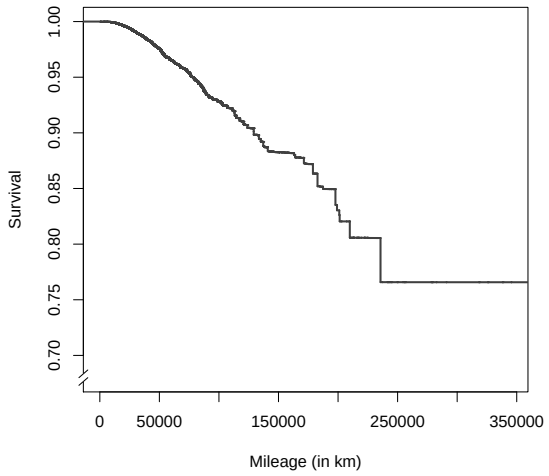

Figure 3.3: Semiparametric estimator $\hat{S}_{CV}$ in the presence of covariates.

higher variance in the estimation. To find a way achieving the greatest possible compromise out of those two conflicting objectives, we have to select some of the information which has an essential effect on the occurrence of events. In the following, we find an appropriate subset of regressors for the model by variable selection.

The additional information in the data set consists of more than 10 different categories and some of them are highly correlated. Including correlated regressors in the model is unnecessary, since they have the same effect on reliability. Therefore, all but one variable of this related additional information can be neglected.

First, let us have a look at the correlation between the covariates. The correlation matrix can be found in Table B.2, Appendix B. An illustration of the correlation of the data is given in the associated correlogram in Figure 3.4, where pies and shades indicate the strength of the correlation respectively and the color blue indicates positive correlation, whereas red indicates the negative correlated variables .
We observe, X_1 and X_4 are highly negatively correlated with X_7 and X_9, respectively. Moreover, X_6 with X_7, X_8 and X_9 and especially X_7 with X_9 show some

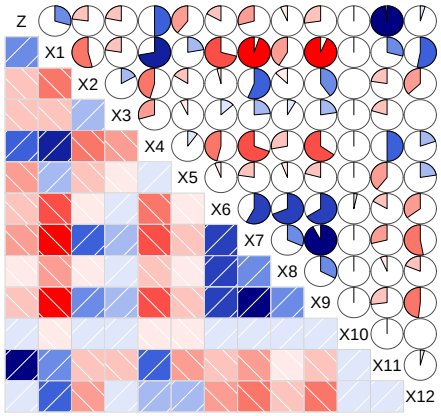

Figure 3.4: Correlation of the covariates.

positive correlation.
Another important observation is the correlation of X_{11} with Z. Due to this high correlation of 0.98 we can already state that X_{11} is definitely one of the variables which seem to have a crucial effect on the occurrence of events.
But let us consult some alternative sources to obtain additional information before drawing the final conclusion.

Considering the Wald test performing the logistic regression for the full model, the covariates X_4, X_5, X_{11} and X_{12} are highly significant, whereas X_1, X_6, X_8, X_9 and X_{10} can be neglected.
Next, we want to compute the All Possible Subset Method.

3.2.1 All Possible Subset Method[1]

Let p denote the number of covariates $x_1, \ldots, x_p$ containing the additional information and $x_i \in \mathbb{R}^n$, where $n > p$. That is, for each car i, there exist observations on

[1]This section is based on the concepts in Montgomery, Peck and Vinning [13].

each of the covariates.
Then, the *full model* is defined by

$$z_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon, \tag{3.1}$$

where $i = 1, \ldots, n$.

In order to find the "best" subset of variables, it is natural to compare different combinations of regressors and select the most significant one. Therefore, we delete little by little regressors from the full model.
Let r denote the number of regressors that are deleted from equation (3.1). Rewriting the model with only $p - r$ covariates leads to the following equation:

$$z_i = \beta_0 + \sum_{j=1}^{p-r} \tilde{\beta}_j \tilde{x}_{ij} + \epsilon, \tag{3.2}$$

for $i = 1, \ldots, n$ and where $\tilde{x}$ represents a vector containing all the remaining $p - r$ regressors and $\tilde{\beta}$ the corresponding coefficients.

In the All Possible Subset Method, all possible regression equations are considered, starting with one regressor, two regressors, until concluding with the full model. For the selection of the final model one consults criteria as the coefficient of determination R^2 or the quadratic mean squared error.
We assume the intercept term β_0 to be an inherent component of each model such that there are p candidate regressors to choose. Note that this model becomes more and more complex by increasing the number of candidates, since there are $2^p - 1$ linear equations to be estimated and investigated.
Many numerical computations are proposed by now. For an example, see Furnival and Wilson [6]. If this model becomes too complex due to a high amount of possible candidates, there are several alternative models such as Stepwise Regression Models available.
For our purpose, we apply the All Possible Subset Method. An extract of the top

three R^2-values for every subset size is displayed in Table B.1 in Appendix B. Let us have a look at the corresponding plot of the R^2-values in Figure 3.5.

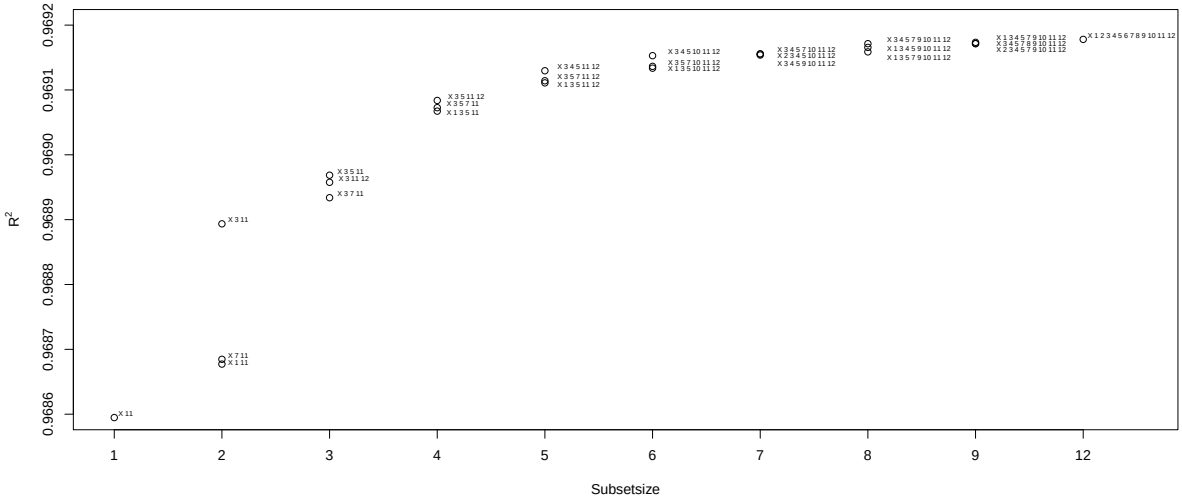

Figure 3.5: Selected R^2-values of the All Possible Subset Method.

It is natural that increasing the variables in the model leads to an increase in the value R^2. In order to find the "best" subset we have to consider the biggest subset for which the increase of the R^2-value is significant and keeping another covariate is worthwhile. For instance, consider the variable X_{11}. When the model only contains X_{11}, the corresponding R^2-value is 0.9685. Additionally including the covariate X_3 in the model leads to a R^2-value of 0.9687.

The value for X_{11} is very high compared to the other subset size one values. Thus, it should be enough to just keep this one covariate in the subset model.

Notice that this case is more of an extraordinary nature. Moreover, we are interested in illustrating our whole developed model by considering various covariates which seems to be the more usual case.

Taking some subjective opinions as well as the regression analysis into account, we

decide to include the covariates X_3, X_5 and X_{11} in the final model.
In the following we continue our application on this subset which is denoted by $\tilde{X} = (X_3, X_5, X_{11})$.

3.3 Thinned Data Set

After the main components of the data set are selected, we finally get to the crucial issue which of the remaining covariates of the sample $\tilde{X}$ seems to have the most influence in the occurrence of an event.
To give a short outline of this section, we first estimate the function m. Thereafter, Dikta's semiparametric estimator for the thinned model is applied which gives an impression about the marginal survival function. In order to obtain an estimator for the conditional survival function the newly proposed estimator $\tilde{S}(z|x)$ of the survival time for every arbitrary value of z and $(x_1, \ldots, x_p)$ is provided. Subsequently, all the contributing variables except one are fixed in order to investigate the effect of this specific quantity on the survival behavior. In conclusion, a characterization of the particular effect is presented.

3.3.1 Application of the Semiparametric Estimator

In order to determine the semiparametric estimator $\hat{S}_{CV}$ which was introduced in Section 1.5, we estimate $m(x, z; \theta)$.
Using the multiple regression model (see Section 1.5.1) we obtain the following maximum likelihood estimates

$$\hat{\beta}_0 = -2.395$$

$$\hat{\beta}_1 = (2.946,\ -3.235,\ 0.0004962)$$

$$\hat{\beta}_2 = -0.0002706, \tag{3.3}$$

where $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$.
Note that all p-values associated with the Wald test statistic for the null hypothesis

$H_0 : \beta_k = 0$ of the regressors $\tilde{X}$ are highly significant. That means, by analyzing the conditional probability of uncensoring, the covariates X_3, X_5 and X_{11} should be kept in the model.
Using those maximum likelihood estimators leads to an estimator of the conditional probability $\mathbb{P}[\delta = 1|Z = z, X_3 = x_3, X_5 = x_5, X_{11} = x_{11}]$:

$$m(z, x; \hat{\theta}) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1' x + \hat{\beta}_2 z)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1' x + \hat{\beta}_2 z)}.$$

Figure 3.6 represents the function m for different values of x. The function behaves as expected. First note, it appears logical that m, the probability of being not censored, decreases with the mileage. Second, it is reasonable that a higher amount of engine starts, cold starts and mileage per year is associated with a lower proportion of uncensoring which leads to poor survival conditions.

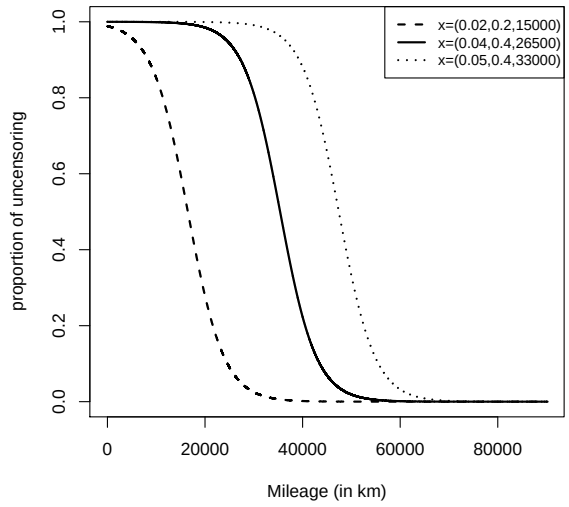

Figure 3.6: Logistic estimation curves of the proportion of no censoring for different values of x.

Inserting the estimates $\hat{\theta}$ from equation (3.3) in equation (1.14) yields the semipara-

metric estimator in the presence of covariates

$$\hat{S}_{CV}(z) = \begin{cases} 1, & \text{if } 0 < z < Z_{1:n} \\ \prod_{i=1}^{n} \left(1 - \frac{m(Z_{i:n}, \tilde{X}_{[i:n]}; \hat{\theta})}{n-i+1}\right)^{\mathbb{1}_{\{Z_{i:n} \leq z\}}}, & \text{otherwise.} \end{cases}$$

The plot of this estimate applied to the data example is shown in Figure 3.7.

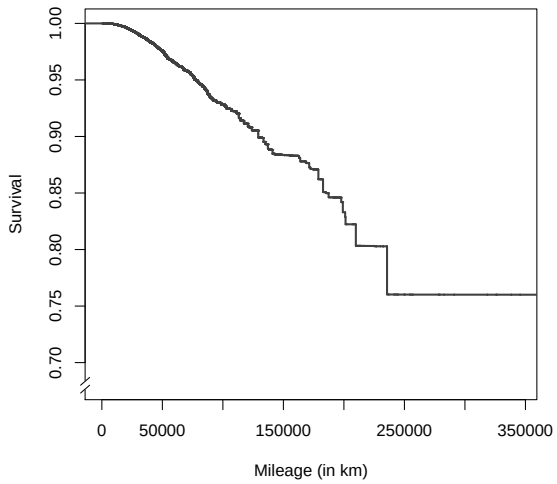

Figure 3.7: The semiparametric estimator in the presence of covariates for the thinned model.

So far, the analysis represents the marginal survival function and gives us some idea of the data set. In order to gain deeper knowledge and more influence on the selection of covariate values, the conditional survival function

$$\mathbb{P}[Z > z | X_3 = x_3, X_5 = x_5, X_{11} = x_{11}]$$

is investigated in the following.

3.3.2 Application of the Estimator of the Conditional Survival Function

In order to estimate the conditional survival function we apply the estimator $\tilde{S}$ which was proposed in equation (2.18) in Chapter 2.

Using the maximum likelihood estimator for $\hat{\theta}$ from equation (3.3), the weights W_i must yet be determined.

For the multidimensional kernel function $K(u)$ we use a multiplicative truncated Cauchy kernel

$$K(u) = K_1(u_1) \cdot K_2(u_2) \cdot K_3(u_3).$$

Then, $W_i(x; h_n) = \dfrac{K(\frac{x-X_i}{h_n})}{\sum_{j=1}^{n} K(\frac{x-X_j}{h_n})}$ for $i = 1, \ldots, n$.

An outstanding issue is to determine the bandwidth $h_n = (h_{1n}, h_{2n}, h_{3n})$. To get some idea about the behavior of the kernel function for different values of h_n, kernel functions for several bandwidths are illustrated in Figure 3.8.

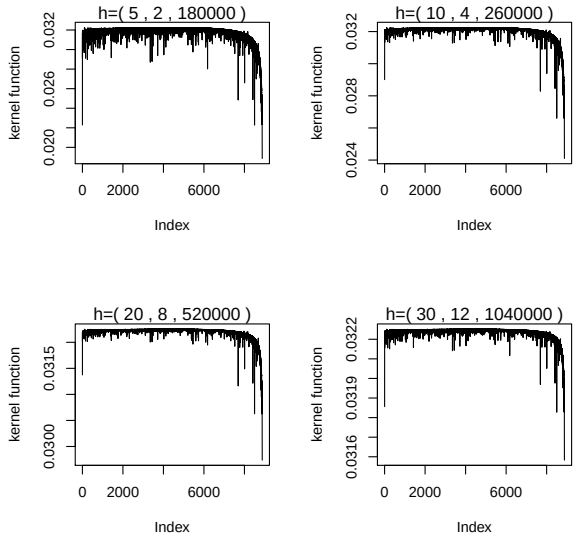

Figure 3.8: Kernel function for a variety of bandwidths.

As expected, the kernel function has less fluctuation for higher values of h_n.

Note that there are several ways to determine the optimal bandwidth. For our purpose, we choose $h_n = (10, 4, 260000)$ to be the bandwidth we are working with in the following.

Eventually, the estimated conditional survival function from Section 2.1 can be computed by

$$\tilde{S}(z|x) = \begin{cases} 1, & \text{if } 0 < z < Z_{1:n} \\ \prod_{i:Z_i \leq z} \left(1 - \frac{\tilde{m}(Z_i, x; \hat{\theta}) W_i(x; h_n)}{\sum_{j=1}^{n} \mathbb{1}_{\{Z_j \geq Z_i\}} \tilde{m}(Z_j, x; \hat{\theta}) W_j(x; h_n)}\right), & \text{otherwise} \end{cases}$$

and is shown in Figure 3.9.

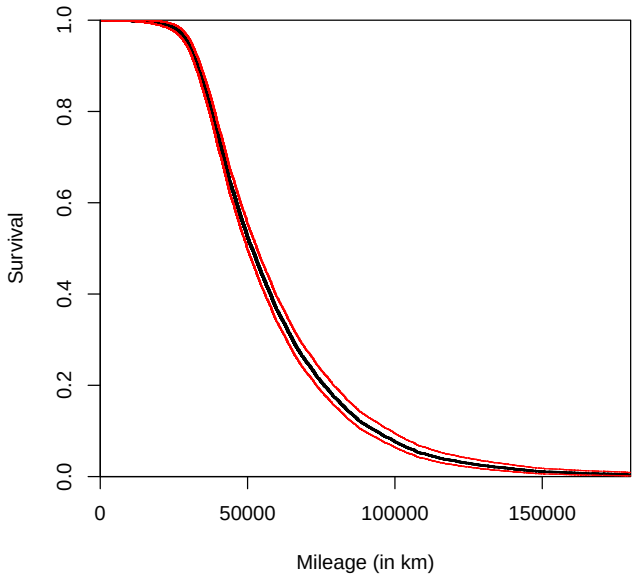

Figure 3.9: Estimated conditional survival function $\tilde{S}(z|x)$ (—) with pointwise confidence band (—).

The corresponding confidence intervals are computed by solving the non-linear equation

$$-4\pi \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \ln R(p, t, \lambda|x) = \chi^2_{1,1-0.05}$$

for λ numerically. As introduced in Section 2.2 the resulting two roots λ_L and λ_U

have to be inserted in the equation respectively

$$p_l = \prod_{i=1}^{D(t)} \frac{d_i + \lambda_l}{d_i + e_i + \lambda_l}, \qquad l \in \{L, U\}$$

in order to achieve boundaries $[p_L, p_U]$ for the Likelihood Ratio confidence interval. Repeating this procedure for various values of t leads to the Likelihood Ratio confidence band which is displayed in Figure 3.9.

3.3.3 Most Influential Covariate

Let us analyze the particular effect on the survival probability of each single covariate by considering some illustrations. For a better comparison all covariates are standardized. Figure 3.10 displays plots of the estimated survival probabilities $\tilde{S}(z|x)$ for different values of x. Based on the standardized vector $x = (0, 0, 0)$ (—) one of these three entries is varied, and the remaining two covariate values are handled as constants. For instance, the first picture shows estimated survival functions for the values $x = (-1, 0, 0)$ (—), $x = (0, 0, 0)$ (—), $x = (5, 0, 0)$ (—) and $x = (8, 0, 0)$ (—) demonstrating the influence of covariate X_3 on the survival function. The remaining two pictures display a similar variation in x for the covariates X_5 and X_{11} respectively. Note that the influence of X_{11} causes the most fluctuation in all of the three pictures.

For a better comparison, all x-values which were varied by the same amount for each considered covariate are shown in Figure 3.11. In the first picture the estimated survival function for the covariate vectors $x = (-1, 0, 0)$ (—), $x = (0, -1, 0)$ (—) and $x = (0, 0, -1)$ (—) , that is all orange functions (—) from Figure 3.10, are illustrated and compared to the standardized version with $x = (0, 0, 0)$ (—).

Obviously, the most fluctuation from the standardized black curve is observed for covariate X_{11} in picture three.

In order to provide a more precise evidence, the area between the standardized curve $\tilde{S}(z|(0, 0, 0))$ and the survival function with a varied covariate vector $\tilde{x}$

$$\int |\tilde{S}(z|\tilde{x}) - \tilde{S}(z|(0, 0, 0))| dz$$

is illustrated in Table 3.1.

Covariate vector x	L^1 norm
(0,0,0)	0
(-1,0,0)	9.828
(0,-1,0)	29.994
(0,0,-1)	237.493
(5,0,0)	53.097
(0,5,0)	117.795
(0,0,5)	11074.210
(8,0,0)	88.121
(0,8,0)	165.251
(0,0,8)	5414.472

Table 3.1: Area between the standardized survival function and the survival probability considering diverse covariates over the time horizon $[0, Z_{n:n}]$.

As expected, the influence of covariate X_{11} in these calculations is outstanding.

Another interesting point of view is given by the plots of survival functions in terms of covariates for a fixed time point z. The corresponding plot of $\tilde{S}(50,000|x)$ where x varies from -1 to 9 for each covariate respectively is found in Figure 3.12. It characterizes the effect of X_3, X_5 and X_{11} on the survival function for a fixed time point.

As assumed from considering the above illustrations, the covariate X_{11} is the most fluctuating one considering the plots and additionally shows the largest L^1 norm for every considered value.

Obviously, the more deviation in the survival probability the more influential the covariate. In this case the variation of the survival probability is the largest for covariate X_{11}. The corresponding effect is shown in Figure 3.12.

Note that this result is consistent with the outcomes from the All Possible Subset method in Section 3.2.1.

This result is also reasonable from a practical point of view. The third picture in Figure 3.10 shows the estimated survival function for different values of X_{11} where

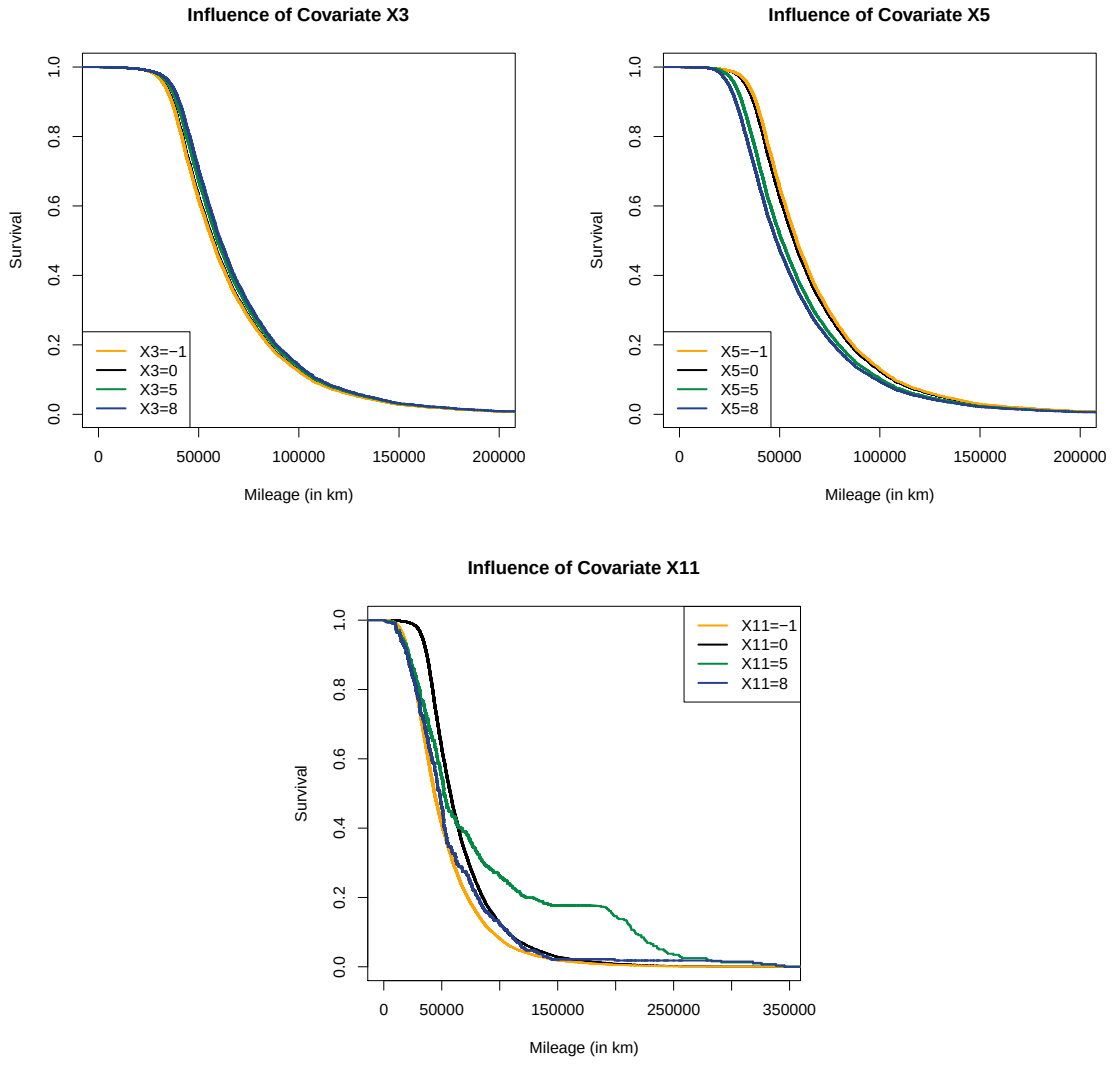

Figure 3.10: Effect of one particular covariate on the estimated survival function.

the values of $X_{11} = -1$, $X_{11} = 5$ and $X_{11} = 8$ lead to lower survival probabilities than for the case $X_{11} = 0$ if mileage is less than $60,000$ km. After this particular time point, the probabilities which belong to $X_{11} = 5$ exceed the probabilities that are countered along the standardized ones. With increasing mileage, the remaining two functions exceed the standardized probabilities as well.

Keeping in mind that the covariate X_{11} represents the influence of mileage per year,

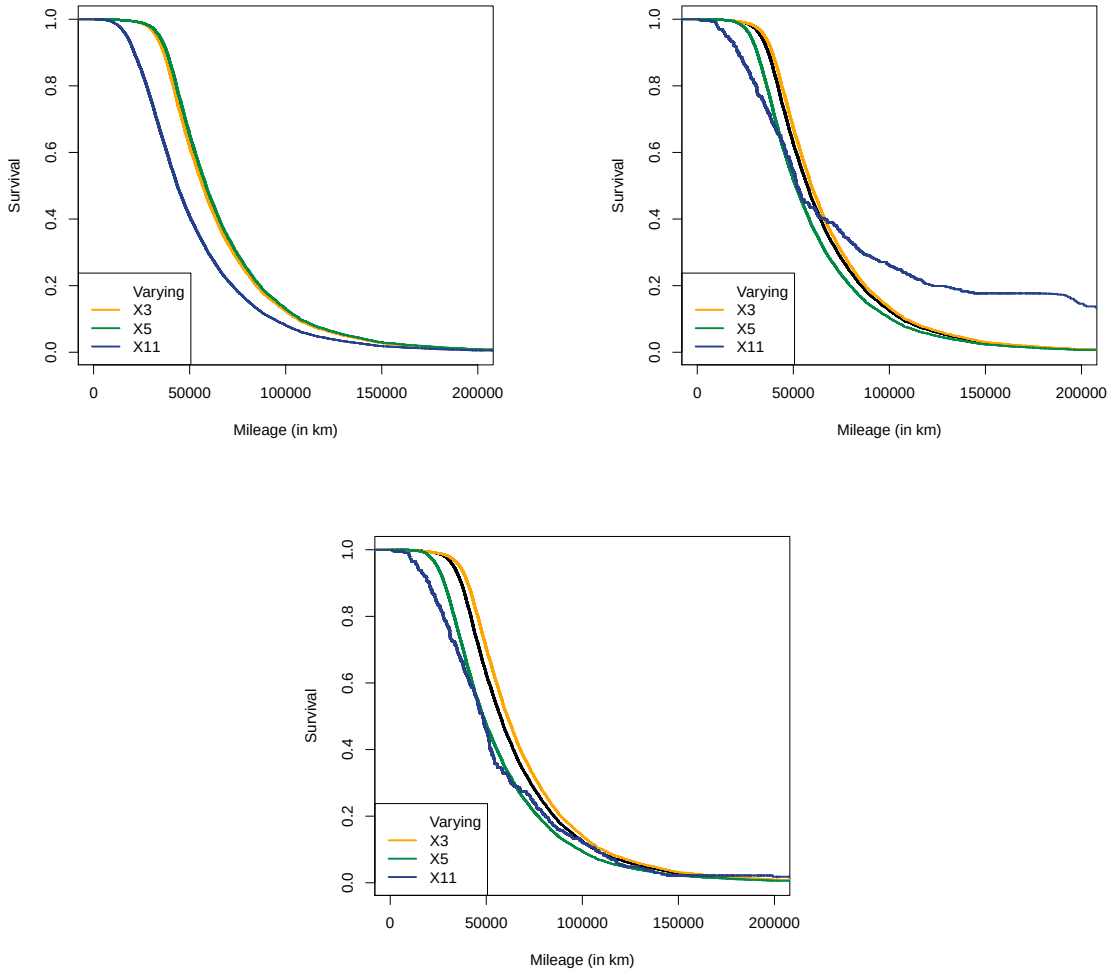

Figure 3.11: Variation of covariate values.

the observed effect is reasonable: An automobile with lower mileage per year tends to have a higher probability of failure than a vehicle with higher mileage. Considering a value for the variable X_{11} which is close to the maximal observed value on the other hand, yields to higher probability of failure as well. A vehicle with high mileage per year corresponds to a driving style on highways, whereas lower mileage is associated with city drivers. Note that driving in cities burdens certain engine components heavier, than rides on highways. Furthermore, it is obvious that almost maximal

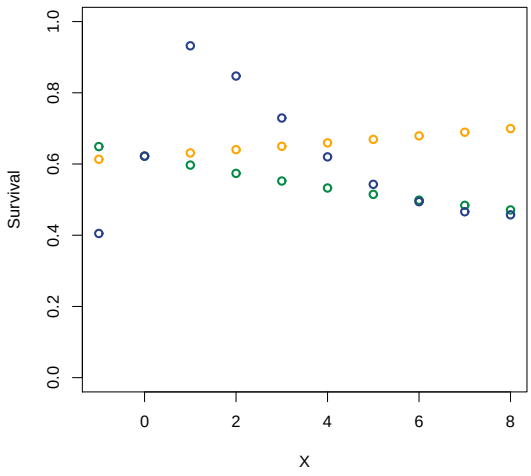

Figure 3.12: Survival probabilities for a fixed time point but different covariate values for X_3 (—), X_5 (—) and X_{11} (—).

burden leads to early failure.

3.3.4 Comparison and Modifications

Modifying the weights W_i through B_i as mentioned in Section 2.3.2 yields the step functions in Figure 3.13. A slight change in the estimation can merely be identified but does not really affect the estimation as a whole.

Comparing the derived estimator $\tilde{S}(z|x)$ to the estimator which was proposed by Iglesias-Pérez and Uña-Álvarez [8], it is conspicuous that $\tilde{S}$ shows a higher survival probability within the first $50,000$ km, whereas the function tends to zero and finally gets close to zero at a mileage of around $200,000$ km. The compared estimator stays constant on a survival probability of about 0.75 from $60,000$ km which is not reasonable for high values of z. The covariate vector was set at $x = (0.03, 0.35, 21800)$ which indicates a certain manner of driving that matches with the behavior of the estimated survival function $\tilde{S}$.

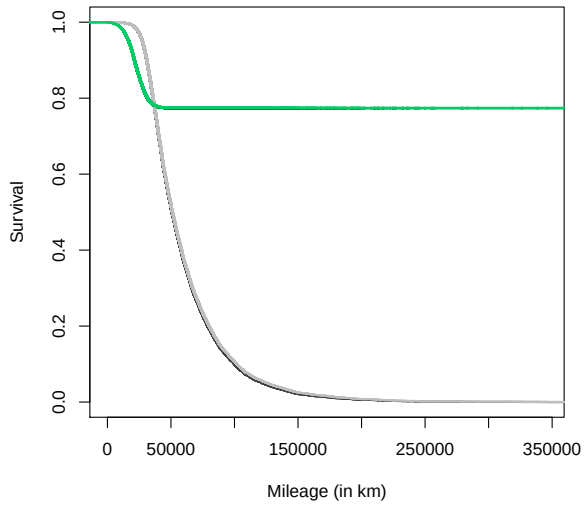

Figure 3.13: Comparison of the estimated survival functions $\hat{S}$ with usual (—) and modified weights (—) and $\tilde{S}$ with usual weights (—) and modified weights (—).

Chapter 4

Conclusion

The goal of this thesis was to select the covariate which seems to have the most impact on survival probability. A method for estimating the conditional survival function was developed, and thereafter the potential factors were varied in order to observe the corresponding change in the survival function.

We started using the conditional semiparametric estimator which was proposed by Iglesia-Pérez and de Uña-Álvarez. During the development of the method, drawbacks of this estimator, such as the lack of being a proper survival function or being undefined for certain weights, were uncovered. Thus, a different estimator was derived where mass is assigned to all observations but is distinguished between censored or uncensored values. We proved that the new estimator not only represents the likelihood estimator but also performs like a proper survival function.

In the application on a real data set from an automotive company the estimator performed as would be expected for a corresponding manner of driving. Also the selected covariate is a reasonably assumption to have the greatest impact on failure out of all considered covariates. For practical applications, another benefit, in contrast to the ordinary applied Kaplan-Meier estimator, is the dependence on covariates and survival time. The derived estimator can easily be applied to an automobile for which a certain driving style is known and whose information is incorporated into the covariates.

An interesting topic for future research would be to investigate the efficiency of

this new estimator by comparing the coverage region to different semiparametric estimators. Furthermore, Dikta, Dabrowska and Iglesia-Pérez and de Uña-Álvarez showed the strong law as well as the central limit theorem for their estimators. It would be interesting to prove similar results for the new estimator. The proof of asymptotic normality will finally lead to the proof of the asymptotically chi-squared distribution of the likelihood ratio confidence intervals. Additionally, some specifics such as selecting the optimal bandwidths h_n could be considered by applying a procedure based on bootstrap similar to that shown in Gang Li and Van Keilegom [12].

We accomplished our goal of developing a method which is imperative for estimating the conditional survival function and thus, selecting proper influential covariates. An application of this model will improve the reliability of the considered objects and consequently is not only image enhancing, but also contributes to cost reduction.

Bibliography

[1] Odd O. Aalen, Ørnulf Borgan, Håkon K. Gjessing, *Survival and Event History Analysis: A Process Point of View*, Springer New York, 2008

[2] David R. Cox, E. J. Snell, *The Choice of Variables in Observational Studies*, Journal of the Royal Statistical Society, 23, 51-59, 1974

[3] Dorota M. Dabrowska, *Uniform Consistency of the Kernel Conditional Kaplan-Meier Estimate*, The Annals of Statistics, 17, 1157-1167, 1989

[4] Gerhard Dikta, *On semiparametric random censorship models* , Journal of Statistical Planning and Inference, 66, 253-279,1998

[5] Gerhard Dikta, *The strong law under semiparametric random censorship models* , Journal of Statistical Planning and Inference, 83, 1-10, 2000

[6] George M. Furnival, Robert W. M. Wilson Jr., *Regression by leaps and bounds*, Technometrics, 16, 499-511, 1974

[7] David W. Hosmer, Stanley Lemeshow, *Applied logistic regression*, Wiley, New York, 2000

[8] Maria C. Iglesias-Pérez, Jacobo de Uña-Álvarez, *Nonparametric estimation of the conditional distribution function in a semiparametric censorship model*, Journal of Statistical Planning and Inference, 138, 3044-3058, 2008

[9] Edward L. Kaplan, Paul Meier, *Nonparametric Estimation from Incomplete Observations*, Journal of the American Statistical Association, 53, 457-481, 1958

[10] John P. Klein, Melvin L. Moeschberger, *Survival Analysis*, Springer, 1998

[11] René Külheim, Gerhard Dikta, Jugal Ghorai, *The strong law under semiparametric random censorship models with covariables*, 2013

[12] Gang Li, Ingrid Van Keilegom, *Likelihood Ratio Confidence Bands in Non-parametric Regression with Censored Data*, Scandinavian Journal of Statistics, 29, 547-562, 2002

[13] Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vinning, *Introduction to linear regression analysis*, Wiley, 2006

[14] Evgeny Spodarev, *Stochastik I*, Lecture Notes, 2015, URL `http://www.uni-ulm.de/fileadmin/website_uni_ulm/mawi.inst.110/mitarbeiter/spodarev/publications/scripts/Stochastik1_01.pdf`

[15] Ulrich Stadtmüller, *Survival Analysis*, Lecture Notes, 2012

[16] David R. Thomas, Gary L. Grunkemeier, *Confidence Interval Estimation of Survival Probabilities for Censored Data*, Journal of the American Statistical Association, 70, 865-871, 1975

Appendix A

Mathematical Results

A.1 Interval Property of the LR Confidence Sets

First, we have to assure that $\tilde{p}_i = 1 - \frac{e_i}{d_i+e_i+\lambda}$ and $\tilde{p}_i = 1 - \frac{e_i}{d_i+e_i}$ are nonnegative. This is the case for $\lambda > -d_{D(t)}$. Considering the partial derivative of $\ln R(p, t, \cdot|x)$ from Section 2.2 leads to

$$\left(\frac{\partial \ln R}{\partial \lambda}\right)_{\lambda=\lambda_L} > \left(\frac{\partial \ln R}{\partial \lambda}\right)_{\lambda=0} = 0 > \left(\frac{\partial \ln R}{\partial \lambda}\right)_{\lambda=\lambda_U},$$

where $-d_{D(t)} < \lambda_L < 0 < \lambda_U$.

Note that $D(t) = 0$ and $d_n = 0$ corresponds to $\tilde{S}(t|x) = 0$ and therefore $\lambda_L = 0$ in the case where the estimated survival function equals zero. Furthermore, $\ln R \longrightarrow -\infty$ for $\lambda \to \infty$ and $\lambda \to -d_{D(t)}$.

A.2 Function p in dependency of λ

We consider the first derivative of $\ln p$:

$$\frac{\partial \ln p}{\partial \lambda} = \frac{\partial}{\partial \lambda} \ln\left(\prod_{i=1}^{D(t)} \tilde{p}_i(\lambda)\right) = \frac{\partial}{\partial \lambda}\left(\sum_{i=1}^{D(t)} \ln(d_i + \lambda) - \ln(d_i + e_i + \lambda)\right)$$

$$= \sum_{i=1}^{D(t)} \frac{1}{d_i + \lambda} - \frac{1}{d_i + e_i + \lambda} > 0,$$

since $e_i > 0$ for $i = 1, \ldots, n$ and $-d_{D(t)} < \lambda$. Therefore, p is increasing in λ.

Appendix B

Some Informative Results

Size	Regressors	R^2
1	X_{11}	0.968595
1	X_4	0.249402
1	X_5	0.153269
2	X_3, X_{11}	0.968894
2	X_7, X_{11}	0.968685
2	X_1, X_{11}	0.968677
3	X_3, X_5, X_{11}	0.968969
3	X_3, X_{11}, X_{12}	0.968958
3	X_3, X_7, X_{11}	0.968934
4	X_3, X_5, X_{11}, X_{12}	0.969084
4	X_1, X_3, X_4, X_{11}	0.969073
4	X_3, X_4, X_{10}, X_{11}	0.969067
5	X_3, X_4, X_5, X_{11}, X_{12}	0.969130
5	X_3, X_5, X_7, X_{11}, X_{12}	0.969114
5	X_1, X_3, X_5, X_{11}, X_{12}	0.969111
6	X_3, X_4, X_5, X_{10}, X_{11}, X_{12}	0.969153
6	X_3, X_5, X_7, X_{10}, X_{11}, X_{12}	0.969136
6	X_1, X_3, X_5, X_{10}, X_{11}, X_{12}	0.969134
7	X_3, X_4, X_5, X_7, X_{10}, X_{11}, X_{12}	0.969156
7	X_2, X_3, X_4, X_5, X_{10}, X_{11}, X_{12}	0.969156
7	X_3, X_4, X_5, X_9, X_{10}, X_{11}, X_{12}	0.969154
8	X_3, X_4, X_5, X_7, X_9, X_{10}, X_{11}, X_{12}	0.969171
8	X_1, X_3, X_4, X_5, X_9, X_{10}, X_{11}, X_{12}	0.969166
8	X_1, X_3, X_5, X_7, X_9, X_{10}, X_{11}, X_{12}	0.969159
9	X_1, X_3, X_4, X_5, X_7, X_9, X_{10}, X_{11}, X_{12}	0.969173
9	X_3, X_4, X_5, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}	0.969171
9	X_2, X_3, X_4, X_5, X_7, X_9, X_{10}, X_{11}, X_{12}	0.969171
Full model	X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}	0.969178

Table B.1: Relevant results of the All Possible Subset Method.

	Z	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
Z	1.00	0.29	-0.23	-0.22	0.50	-0.39	-0.17	-0.29	-0.07	-0.27	0.02	0.98	0.05
X1	0.29	1.00	-0.54	-0.23	0.72	0.22	-0.71	-0.95	-0.41	-0.94	-0.02	0.29	0.53
X2	-0.23	-0.54	1.00	0.17	-0.46	-0.16	-0.04	0.56	-0.13	0.40	0.01	-0.23	-0.37
X3	-0.22	-0.23	0.17	1.00	-0.30	-0.08	0.13	0.24	0.08	0.23	0.02	-0.21	0.01
X4	0.50	0.72	-0.46	-0.30	1.00	0.10	-0.46	-0.69	-0.28	-0.65	-0.02	0.50	0.20
X5	-0.39	0.22	-0.16	-0.08	0.10	1.00	-0.06	-0.23	-0.06	-0.21	-0.02	-0.39	0.23
X6	-0.17	-0.71	-0.04	0.13	-0.46	-0.06	1.00	0.58	0.69	0.66	0.03	-0.17	-0.35
X7	-0.29	-0.95	0.56	0.24	-0.69	-0.23	0.58	1.00	0.31	0.92	0.01	-0.29	-0.53
X8	-0.07	-0.41	-0.13	0.08	-0.28	-0.06	0.69	0.31	1.00	0.33	0.02	-0.07	-0.19
X9	-0.27	-0.94	0.40	0.23	-0.65	-0.21	0.66	0.92	0.33	1.00	0.02	-0.27	-0.48
X10	0.02	-0.02	0.01	0.02	-0.02	-0.02	0.03	0.01	0.02	0.02	1.00	0.01	0.00
X11	0.98	0.29	-0.23	-0.21	0.50	-0.39	-0.17	-0.29	-0.07	-0.27	0.01	1.00	0.04
X12	0.05	0.53	-0.37	0.01	0.20	0.23	-0.35	-0.53	-0.19	-0.48	0.00	0.04	1.00

Table B.2: Correlation matrix of potential influential variables.

Appendix C

Programm Code R

```
## Estimator - Conditional Survival Function with Likelihood ...
    Ratio Confidence Band
#============================================================

#CB=1: Calculate Pointwise Confidence Band
S.tilde <- function(Z,Δ,X,CB){
  # sort dataset
  sortdata<-data.frame(Z,Δ,X)

  sortdata<-sortdata[order(sortdata$Z), ]
  Z.tilde <- sortdata$Z
  Δ.tilde <- sortdata$Δ
  X.tilde <- sortdata[ ,3:dim(sortdata)[2]]

  n <- length(Z.tilde)
  R <- rank(Z.tilde)


  ## Semi Parametric estimator (considering covariates)
  #=====================================================
  ## Estimation of m(Z_i,t) using logistic regression
  # Generalized linearregression (MULTIVARIATE) -> binomial ...
     response variable Δ
  logit.out <- glm(Δ ~ Z + X[1] + X[2] + X[3], data = sortdata, ...
     family = "binomial")
  logitsummary <- summary(logit.out)

  beta_0 <- logitsummary$coef[1,1]
  beta_1 <- logitsummary$coef[3:dim(logitsummary$coef)[1],1]
  beta_2 <- logitsummary$coef[2,1]
```

```
28
29
30    ## Conditional Survival Function (Estimator)
31    #=========================================
32    ## Estimation of S(y|x)
33
34    #general m
35    m.fun <- function(z,x){
36      return(exp(beta_0+beta_1%*%x+beta_2*z)/...
37          (1+exp(beta_0+beta_1%*%x+beta_2*z)))
38    }
39
40    # Multivariate truncated Cauchy kernel function
41    K.multi <- function(x){
42      K<-numeric(0)
43      for (i in 1:length(x)){
44        K[i] <- 1/(pi*(1+x[i]^2))
45      }
46      return(prod(K))
47    }
48
49    boundary<-length(Z.tilde[Z.tilde<=y])
50
51    h <- c(10,4,260000)     # Bandwidth
52    kern <- numeric(0)
53    for(j in 1:n){
54      kern[j] <-K.multi((x-X.tilde[j,])/h)
55    }
56
57    ## Weights
58    B <- numeric(0)
59    for(i in 1:n){
60      B[i] <- kern[i]/sum(kern)   # Weights
61    }
62
63    ## Estimator
64  m.ges<-m.fun(Z.tilde,x)
65  m.mod<-Δ.tilde*m.ges+(1-Δ.tilde)*(1-m.ges)
66  ## modification in m
67  S_tilde<-numeric(0)
68  S_tilde[1]<-1
69  for (i in 1:n){
70    S_tilde[i+1] <- ...
        S_tilde[i]*(1-(1*(Z.tilde[i]<=y)*m.mod[i]*B[i])/ ...
71                  (sum(1*(Z.tilde>=Z.tilde[i])*B*m.mod)))
72  }
73
74    cat("S_tilde(",y,"|x)= ", S_tilde[boundary+1])
```

```
75
76    if(CB==1){
77      ## Pointwise Confidence Band
78      #============================
79      ## Significance niveau
80      alpha<-0.05
81
82      coef <- sum(kern)*(2*pi)
83
84      ## Bisection method
85      CI_lambda_bisec <- function(a,b){
86        fa <- f(a)
87        fb <- f(b)
88        eps<- 1e-6
89        i<-1
90        temp <-numeric(0)
91        while(abs(a-b)>2*eps){
92          m <- (a+b)/2
93          fm <- f(m)
94          if(fm==0){
95            return(m)
96          }else{
97            if(fa*fm<0){
98              b=m
99              fb=fm
100           }else{
101             a=m
102             fa=fm
103           }
104         }
105         temp[i]<-m
106         i<-i+1
107       }
108       return(temp)
109     }
110
111     ## Calculation of e_i
112     e <- m.fun(Z.tilde,x)*B
113
114     ## Calculation of d_i
115     d <- numeric(0)
116     for(i in 1:(n-1)){
117       d[i] <- sum(e[(i+1):n])
118     }
119     d[n] <- 0
120
121     ## Confidence Interval
122     pL<-numeric(0)
```

```
123     pU<-numeric(0)
124     S.est<-numeric(0)
125
126     i<-1
127     for(D in 1:boundary){
128
129       f <-function(x){-2*coef*(sum(d[1:D]*log((d[1:D]+e[1:D])/...
130         d[1:D]*(d[1:D]+x)/(d[1:D]+e[1:D]+x)))+sum(e[1:D]*...
131         log((d[1:D]+e[1:D])/(d[1:D]+e[1:D]+x))))-qchisq(1-alpha,1)}
132
133       lambda_temp <- CI_lambda_bisec(0,1000)
134       lambdaU <- lambda_temp[length(lambda_temp)]
135
136       lambda_temp <- CI_lambda_bisec(-min(d[0:D])+10^(-6),0)
137       lambdaL <- lambda_temp[length(lambda_temp)]
138
139       ## Calculation of CI
140       pL[i]<-prod((d[1:D]+lambdaL)/(d[1:D]+e[1:D]+lambdaL))
141       pU[i]<-prod((d[1:D]+lambdaU)/(d[1:D]+e[1:D]+lambdaU))
142       i<-i+1
143     }
144
145     par(mfrow=c(1,1))
146     plot(stepfun(Z.tilde[1:boundary],S_tilde[1:(boundary+1)]),...
147         main="Conditional Survival Function",xlab="Mileage (in ...
                km)",...
148         ylab="Estimator",lwd=2,xlim=c(0,40000))
149     lines(stepfun(Z.tilde[1:(boundary-2)],pL[1:(boundary-1)]),...
150         lty=2,col="red",lwd=1)
151     lines(stepfun(Z.tilde[1:(boundary-2)],pU[1:(boundary-1)]),...
152          lty=2,col="red",lwd=1)
153
154   }
155 }
```