DIPLOMARBEIT

Nonparametric hazard rate estimation for relative survival models

Ausgeführt am Institut für
Medizinische Statistik
der Medizinischen Universität Wien

unter der Anleitung von
Ao. Univ. Prof. Dr. Werner Brannath

eingereicht an der Technischen Universität Wien
Fakultät für Mathematik

von
Sophie Frantal
e0225826
Urselbrunnengasse 17/4/41, 1100 Wien

Wien, 01.07.2010 _____

Contents

1 Acknowledgment

I have been working for the Medical University of Vienna, Institute of Statistics already for more than three years. During the last year I had a very interesting project. I was working on cardiac surgery data on a comparatively old patient population and therefore was using relative survival models. During a period with a lot of work, times of desperation due to not finding a mistake and finally writing a high-level paper I found a very interesting statistical topic for my diploma thesis: nonparametric hazard rate estimation in relative survival models.

I want to thank Ao. Univ. Prof. Dr. Werner Brannath for supervising my work. Through the time we worked together on my diploma thesis we had a lot of ideas how to improve my work. Unfortunately only a few of them could be implemented in time. Most of the open issues are mentioned in the Section Discussion on page 43. Now Werner moves to Germany to meet a new challenge and our joint work will stop at least for the moment. Perhaps we will find a way to translate our open ideas together in the future. For now, Werner, I wish you and your family a great time and good luck in Germany!

I want to thank Ao. Univ. Prof. Dr. Laßnigg and Univ. Prof. Dr. Hiesmayr for allowing me to use their high quality data. Working together with Andrea Laßnigg on her study was a real pleasure. She has an admirable personality which I appreciate very much. Not only does she regard for results, but she also attaches particular value on the human approach of one another.

I want to thank my boss, Univ. Prof. Dr. Peter Bauer, for forcing me and Werner to come to an end in writing my diploma thesis during the time when we had so many ideas. He asked me every week if my work is going on appropriately and towards the end of my work, if I had finished it yet. Working in this section with him as my boss is a privilege for me. He is an amazingly intelligent person and remembers almost everything he heard or read once. He gives us the space to work independently which is very beneficial. In September he will go on pension and I hope that our next boss will have some of his outstanding qualities and skills.

Last but not least I want to thank my mother, Gertrude Frantal, for her assistance in questions related to the English language. She is not only my mother, but also my friend. She gave me moral support in times of reluctance and indifference. She helped me with many intensive discussions.

2 Introduction

Survival analysis has received considerable attention in the broad field of statistics. A selection of different books and papers from different authors is used for this work to summarize existing theory. Usual survival theory can be found in various books, like Andersen et al. [1993], Kalbfleisch and Prentice [2002], and Vittinghoff et al. [2005]. All information about survival functions, hazard rates and estimation methods are taken from literature like these.

The theory of kernel estimation is basically adapted from Wand and Chris [1995]. The book gives a full overview over kernel smoothing. The motivation and ideas of kernel smoothing are introduced by nonparametric regression. However, it is shown that kernel smoothing can be applied to many other important curve estimation problems, such as estimating hazard rate functions. A chapter focused on nonparametric kernel estimates of the hazard function can also be found in Haerdle et al. [2007]. The attention there is drawn to the choice of the bandwidth. A special iterative method for its estimation is proposed. Some of the basic properties of kernel estimates are also discussed in Andersen et al. [1993]. Even if the book concerns the more general topic of counting processes and gives nonparametric as well as parametric estimation techniques there are some interesting results on kernel smoothing.

The three mainly used papers for smoothing hazard rates are Mueller and Wang [1990], Mueller and Wang [1994] and Hess et al. [1999]. Mueller and Wang [1990] consider nonparametric estimation of hazard functions and their derivatives under random censorship, based on kernel smoothing of the Nelson estimate. Data-based local bandwidth choice is proposed. In particular an asymptotically efficient method derived from pilot estimates of the hazard function and of the local mean square error is recommended. Based on the results of Mueller and Wang [1990], Mueller and Wang [1994] specialize on solutions of the practically relevant problems of boundary effects near the endpoints of the support of the hazard rate. A new class of boundary kernels and a data-adaptive varying bandwidth selection procedure are proposed to solve these problems. The statistical properties of these estimates are compared through computer simulations by Hess et al. [1999]. Further, information about hazard smoothing can also be found in Andersen et al. [1993], Wells [1994], Karunamuni and Alberts [2004], Wang [2005] and Haerdle et al. [2007].

Andersen et al. [1985] introduce a Cox-type regression model for the relative mortality, hence, a model for the ratio between the mortality in a cohort and that in a reference population. They declare that it is possible to include in survival analysis also changing trends in the mortality in the reference population which is particularly relevant in long-term follow-up studies where there may be considerable changes. They discuss estimation procedures

and outline large-sample properties of the estimates. A couple of years later Cao et al. [2005] propose a kernel estimate for the relative hazard rate for the case where two groups are compared which are subject to left truncation and right censoring. Two populations are compared by means of the relative hazard rate of the lifetimes of the first population with respect to the lifetimes of the second population. Furthermore relative survival analysis is intelligibly achieved by Maja Pohar and Janez Stare. In Pohar and Stare [2006] and Pohar and Stare [2007] they review different techniques to model relative survival models and describe there R-package *relsurv* which provides functions for easy and flexible fitting of all the commonly used relative survival regression models.

The approach of this work is the estimation of relative hazard rates under random censoring using kernel methods. At first a summary of the most important functions and methods for usual and relative survival models are given, to impose smooth estimates of relative hazard rates including boundary correction and bandwidth choice, based on previous results, in a second step. The paper is structured as follows: In Section 3 the underlying cardiac surgery data are demonstrated. Section 4 gives a review of the theory of survival models. Survival function and its estimation, hazard rate, Log-rank test and Cox-Regression are explained. Smoothing hazard rates for usual survival models including bandwidth choice and correction for boundary effects, basically based on Mueller and Wang [1994], is the topic of Section 5. Section 6 reviews relative survival models basically based on Pohar and Stare [2006]. Relative survival function, relative hazard models, in particular the Andersen multiplicative model, and the according regression analysis are introduced. In Section 7 we then apply the methods of smoothing hazard rates to relative survival models. A summary of the results is given in Section 8. As all calculations and estimations are done using R, the utilized code is embodied in the Appendix.

3 The case study

The underlying data are taken from the prospectively collected data-base from the general hospital of Vienna. Data from n=4374 cardiac surgery patients, operated between 1st January, 1997 and 31st December, 2001, are collected. Follow-up time ranged from 4 to 9 years (median 6.5, IQR 5.3-7.7), with the end of the follow-up period on 31st December, 2005. After the follow-up time data were combined with the hospital central database, which contains information from the Federal Austrian Statistical Office (Statistik Austria) on any patient's death in Austria. The life tables were obtained online from Statistik Austria ($http://www.statistik.at/web_de/statistiken/bevoelkerung/demographische\_masszahlen/sterbetafeln/index.html$).

The 4374 patients included only adults (>18 years). 2854 (65%) were male and 1520 (35%) were female patients. The mean age at surgery was 70.4 ± 12.5 years. The patients were scheduled for cardiac surgery with cardiopulmonary bypass (CPB), coronary artery bypass grafting without CBP and thoracic aortic surgery with CPB. The following interventions were excluded: transplant surgery, scheduled insertion of a cardiac assist device, operation on the descending aorta, thromboendarterectomy of the pulmonary arteries and surgery of congenital heart disease.

The recorded parameters should represent established, validated risk indicators for mortality after cardiac and thoracic aortic surgery according to the EuroSCORE ($http://www.euroscore.org$). Patients demographics and preoperative morbidity variables such as age, sex, weight, congestive heart failure, diabetes, asthma bronchiale (and others) are included. Additionally, information about the therapy with diuretics and/or angiotensin converting enzyme inhibitors were collected prospectively at the time of premedication. As procedural parameters surgery parameters, pre- and postoperative complications and treatment were recorded. In total 28 variables (1 reference group) were considered as potential confounders.

The primary outcome was mortality (survival time) after the operation. 1086 patients died during the observed study time period. 321 of these died within the first 30 days, further 312 patients died in the following two years after the surgical intervention.

Further information, especially on the medical approach, and an extension of data will be published by Lassnigg et al. [2010]. Here is a short extract of the relevant data, all examples are based on this data set:

	PatId	Age	Weight	SexCode	OPDate	OPCABGplus	OPCABGOff	OPValv	OPTAA	Revision48	Revisionspaeter	CardialeDekomp	Asthma	COPD	Diabetes	ANiereninsufchron	MACE	MDiuretika
1	1	76	51	2	14608	1	0	0	0	0	1	1	0	0	0	0	1	1
2	2	40	78	1	14515	0	0	1	0	0	0	0	0	1	0	0	0	0
3	3	73	65	1	14544	0	0	1	0	0	1	0	0	0	0	0	0	0
4	4	46	80	1	14611	0	0	1	0	0	1	0	0	0	0	0	0	0
5	5	64	125	1	15283	0	1	0	0	0	1	1	0	0	0	0	1	0
6	6	50	83	2	14025	0	0	1	0	1	0	0	0	0	0	0	1	1
7	7	50	60	1	13894	1	0	0	0	0	0	0	0	0	0	0	0	1
8	8	61	83	1	15267	0	0	1	0	0	1	0	0	1	0	1	1	0
9	9	84	70	1	14799	0	1	0	0	0	1	0	0	0	0	0	0	0
10	10	71	70	1	14788	0	0	1	0	0	0	0	0	0	0	0	0	0

	UrgentOP	HLMIABP	Ery	TK	Death	Death_SurvivalDays	HF	EF1neu	EF2neu	EKG_VHFL	Heartpump	Infarct	CaroPAVK
1	0	0	2	0	0	2194	0	1	0	0	0	0	0
2	0	0	0	0	1	786	0	0	0	0	0	0	0
3	0	0	4	0	1	1620	0	1	0	0	0	0	0
4	1	1	2	0	0	2191	1	1	1	1	0	0	0
5	0	0	0	0	0	1519	0	0	0	0	0	1	0
6	0	0	1	0	0	2777	0	0	0	0	0	0	0
7	0	0	0	0	0	2908	0	0	0	0	0	0	0
8	0	0	0	0	0	1535	0	1	0	0	0	0	0
9	1	0	1	0	1	44	0	0	0	0	0	0	0
10	0	0	2	0	0	2014	0	1	0	0	0	0	0

Table 1: Medical data example

4 Survival Models

Survival Analysis can be found in many different fields of science. Actually it is not only a matter of survival time, but a matter of time to event. Time to event is defined as the period between a fixed starting point and the occurrence of a predefined endpoint. Such an event could be the failure of a machine, the replication of an action, the change of living conditions and many others. The most typical and perhaps the most frequent event type is death, which also is the interesting event in our example.

This chapter follows textbooks like Andersen et al. [1993], Kalbfleisch and Prentice [2002], Vittinghoff et al. [2005] and articles like Fox [2002], Collett [2005].

4.1 Survival Function

Survival analysis is based on the so called survival function S. Let T be a positive, continuously distributed life time variable. Then the survival probability is defined as the probability that T exceeds the current time t which leads to the survival function:

$$S(t) = P(T > t) \tag{1}$$

Usually $S(0) = 1$, i.e. at the beginning of a study all observed patients are alive. In exceptional cases it could be less than 1 which allows for the possibility of immediate death. The survival function has to be non-increasing, $S(t) \geq S(\tilde{t})$ if $t \leq \tilde{t}$.

The survival function is related to the distribution function of the event times F and the appropriate continuous density function f. The distribution function is defined as:

$$F(t) = 1 - S(t) \tag{2}$$

$F(t)$ is the probability that the time of death T is earlier than or equal to the current time t. The density function is the derivative of the distribution function:

$$f(t) = F'(t) \tag{3}$$

Here $f(t)$ gives the rate of death per time. The connection between survival function and density function is specified by:

$$S(t) = \int_t^\infty f(u)\, du \tag{4}$$

The true data generating survival function is usually not known. However, the survival function can be estimated from data using information of the observed patients. If T is

observed for all patients, this can be done by the ratio of patients alive at each time point t:

$$\widehat{S}(t) = \frac{\text{Number of patients with survival time} > t}{\text{Total number of patients in the study}} \tag{5}$$

Often not all survival times in a study are observed. Every study is temporary and normally not all observed patients die till the end of the study. Patients may also drop out of the study before study end because of other reasons than death (e.g.: patients request, move,...) and hence are lost to follow-up. The data of patients that survive the end of the study or drop-out before study end are called (right-)censored. In this case it is only known that the patient survived until the end of his follow-up period, i.e. his survival time is at least as long as his follow-up time. The data of patients for whom the event (death) has been observed during the study are called uncensored. Patients who are alive and not censored at time t are called at risk at time t.

Assume that $\tilde{T}_1, \tilde{T}_2, ..., \tilde{T}_n$ are independent, identically distributed uncensored lifetimes and $C_1, C_2, ..., C_n$ are independent, identically distributed censoring times, both with absolutely continuous distribution functions. Because either the life time or the censoring time (depending on what occurs first) is observed in the study, the actual observations are specified by the pair (T_i, δ_i) and $T_i = \min(\tilde{T}_i, C_i)$ gives the follow-up time of a patient. $\delta_i = I_{\{\tilde{T}_i \leq C_i\}}$ is the indicator function giving the censoring status which is one if the patient dies and 0 if the patient is censored. Ordered time data are marked by round brackets around the subscript, i.e. $T_{(1)} \leq T_{(2)} \leq ... \leq T_{(n)}$ and $\delta_{(i)}$ is the censoring indicator variable of $T_{(i)}$. Hence, (i) indicates the patient(s) with the i-th largest survival time $t_{(i)}$. This set of assumptions is called the random censorship model.

In our example the data are right censored. The only reason for censoring was the end of the study, all other patients had a complete follow-up.

4.2 Kaplan-Meier estimate

When the survival times are censored then (5) is not appropriate for an estimation of the survival function. However, estimation can be achieved by using the Kaplan-Meier estimate. The Kaplan-Meier estimate tries to take all available information into account, also the partial information that a censored patient has survived up to his censoring time. Therefore, the most important advantage (and difference to the simple estimate above) is that it can handle censored data.

Given the random censorship model the Kaplan-Meier estimate is defined as:

$$\widehat{S}(t) = \prod_{i=1, t_{(i)} \leq t}^{n} (1 - \frac{\delta_{(i)}}{r_i}) \tag{6}$$

where r_i is the number of patients at risk just before $t_{(i)}$. In the case of no ties $r_i = n - i + 1$. Note that $\frac{\delta_{(i)}}{r_i}$ is the proportion of deaths at $t_{(i)}$ among those patients who are at risk. We further define $\widehat{S}(t) = 1$ for all $t < t_{(i)}$. The precision of the survival estimate decreases with increasing t, because of the decreasing number of patients at risk.

Often Kaplan-Meier estimates are presented as plots. The plot is a step function where horizontal sections show time intervals with no event and vertical sections represent one ore more event(s), depending on the hight of the step. The censored times are sometimes marked by small vertical tick-marks at their censoring times. Using Kaplan-Meier plots it is also possible to split the data in K groups (e.g.: different sex gives two groups) and plot one step function for each group.

In Figure 1 it can be seen that in our medical data example many patients died within the first days after the operation. After this time the death rate seem to be similar over the whole remaining time period. Remember, the only reason for censoring was the end of the study. As the last patient was included into the study at the end of 2001, the shortest follow up period of the censored patients was four years. So all censoring times lay between four and nine years. Comparing sex, in our study, female patients seem to have worse survival than male patients. See the right plot of Figure 1.

4.3 Hazard rate

Another important tool in survival analysis is the hazard rate λ. It is the probability that given a patient is alive at time t, he dies in an arbitrarily small time interval after t:

$$\lambda(t) = \lim_{\Delta \to 0} \frac{1}{\Delta} P(t < \tilde{T} \leq t + \Delta t | \tilde{T} \geq t) \tag{7}$$

By definition the hazard rate is non-negative ($\lambda \geq 0$). It can be non-monotonic or discontinuous. The hazard rate characterizes the risk to die at a specific time point. If it is constant over time this means that the risk to die is the same during the whole observed time interval. If it increases over time this means an increase in the risk to die and if it decreases a decrease in the risk to die.

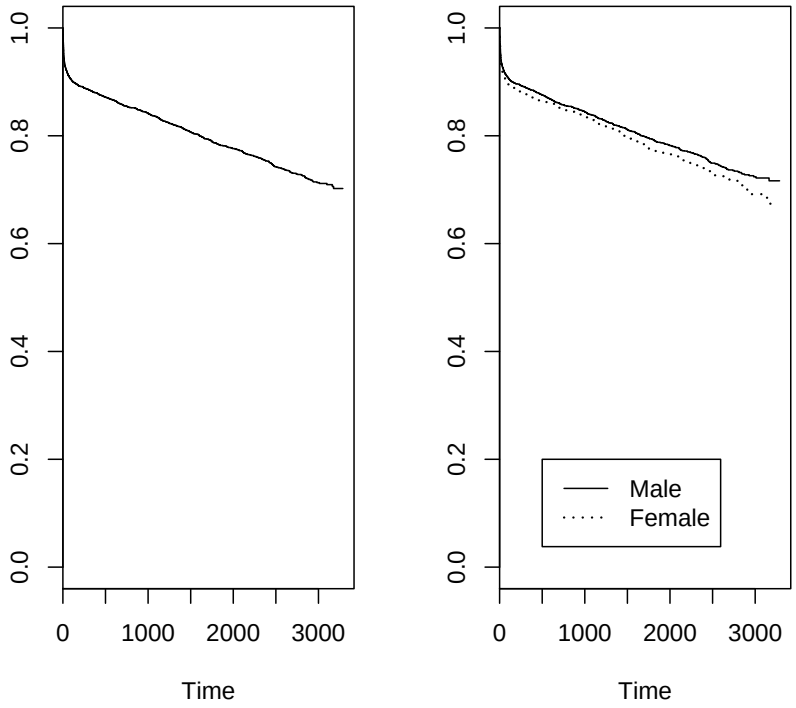

Figure 1: Kaplan-Meier plots

The relation between the hazard rate and the survival function (and therefore also the distribution function and the density function) can be explained by introducing the cumulative hazard rate Λ:

$$\Lambda(t) = \int_0^t \lambda(u)\, du \tag{8}$$

Hence, Λ is the "accumulation" of the hazard over time. Often the cumulative hazard rate is preferred over the hazard rate because it is easier to estimate. The mentioned relation is as follows:

$$S(t) = \exp^{-\Lambda(t)} \tag{9}$$

This implies that the cumulative hazard rate determines the survival function.

Beside the Kaplan-Meier estimate for the survival function also estimates for the hazard rate and the cumulative hazard rate exist. One estimate for the hazard rate which is rarely

used because it is too variable looks as follows:

$$\widehat{\lambda}(t) = \frac{1}{t_{(i+1)} - t_{(i)}} \frac{\delta_{(i)}}{r_i} \tag{10}$$

An example of a non-parametric estimate for the cumulative hazard rate is the so called Nelson-Aalen estimate which is defined by:

$$\widehat{\Lambda}(t) = \sum_{i=1, t_{(i)} \leq t}^{n} \frac{\delta_{(i)}}{r_i} \tag{11}$$

Remember $\delta_{(i)}$ is the number of deaths at time $t_{(i)}$ and r_i is the number of patients at risk just before $t_{(i)}$. When n grows the estimate will get closer to the true cumulative hazard rate. The Nelson-Aalen estimate will be important in Section 4, when smoothing the hazard rate.

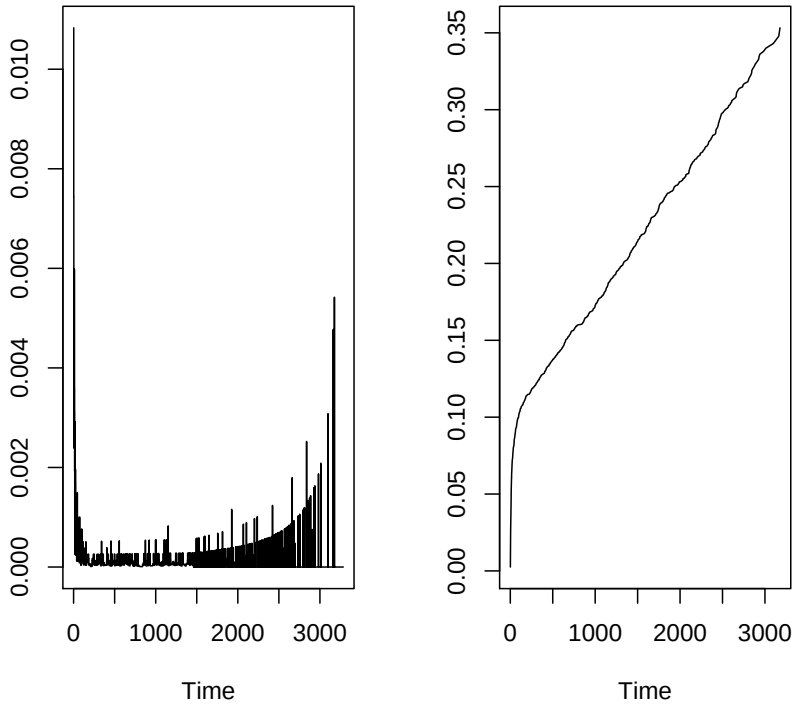

Figure 2: Hazard and cumulative hazard rate

Figure 2 shows the hazard and the cumulative hazard rate in our example. In the left plot it can be seen that, consistently with the rapid decrease of the Kaplan-Meier estimate,

the hazard rate is high at the beginning of the time interval. After the first days the hazard stays at a low level until it grows up again at the end of the time interval. This increase at later time points occurs because censoring is not considered here. The right plot shows the cumulative hazard estimated using the Nelson-Aalen estimate. According to the hazard rate it grows at the beginning and has a constant slope afterwards. Censoring is considered in the Nelson-Aalen estimate. Due to relation (9) on page 11 the plot looks similar to the turned over Kaplan-Meier plot.

4.4 Log-rank test

The Log-rank test is a nonparametric test to compare groups (e.g. two groups like different sex) in terms of their survival times. More precisely the survival functions $S_g(t)$ of the G different, unpaired groups are compared. The two most important features of the Log-rank test are that it can handle with censoring and that it can work with skewed distributed data. Often survival data are not normally distributed, but left-skewed. This could occur, for example, if many patients die shortly after the surgery and only those patients who survive the first critical time, survive longer. Differences in uncensored data could also be calculated by other nonparametric test (e.g. Mann-Whitney-U test for two groups).

The null hypothesis H_0 of the Log-rank test is that the distribution of the survival time is equal in all groups. The related Log-rank test statistic compares the observed and expected number of events under the null hypothesis in each group and consequently looks as follows:

$$LR = \sum_{g=1}^{G} \frac{[O_g - E_g]^2}{E_g} \tag{12}$$

Here $O_g = \sum_{i=1}^{n} \delta_{g(i)}$ is the observed and $E_g = \sum_{i=1}^{n} e_{g(i)}$ is the expected number of events in group g under H_0. Hence, $\delta_{g(i)}$ and $e_{g(i)}$ are the observed respectively the expected number of events in group g at time $t_{(i)}$. The latter is calculated by:

$$e_{g(i)} = r_{gi} \frac{\delta_{(i)}}{r_i} \tag{13}$$

Here r_{gi} is the number of patients at risk in group g right before time $t_{(i)}$. Again, r_i is the number of patients at risk right before time $t_{(i)}$ over all groups and $\delta_{(i)}$ is the number of observed events at time $t_{(i)}$ over all groups. The Log-rank test statistic is approximately χ^2-distributed with $G - 1$ degrees of freedom.

Comparing different sexes in our example (see Kaplan-Meier plot on page 10) using the

Log-rank test no statistically significant difference can be found (p=0.125).

4.5 Cox-Regression

To analyze the influence of several (including metric) covariables (e.g. sex and age) on survival the Log-rank test cannot be used. The applicable method in such cases is the so called Cox proportional-hazards model, or simply Cox model. The Cox-regression can, like the Log-rank test, also handle censored data. As the name implies it is based on the hazard rate, conditional on the covariables:

$$\lambda(t, x_1, ..., x_F) = \lim_{s \to 0} \frac{1}{s} P(t \leq T < t + st | T > t, x_1, ..., x_F) \tag{14}$$

where x_1,...,x_F are F covariables. The proportional-hazards assumption of this model is that the covariables influence the hazard in a time independent manner. This means that a variable increases or decreases the hazard by the same factor at each time point. Based on this assumption the hazard rate has the following form:

$$\lambda(t, x_1, ..., x_F) = \lambda_0(t) f(x_1, ..., x_F) \tag{15}$$

where $\lambda_0(t)$ is the so called baseline hazard rate and $f(x_1, ..., x_F)$ is a positive regression function which is independent of time. So the hazard consists of the baseline hazard multiplied by a time constant factor. Note, that $\lambda_0(t)$ is the hazard of a patient with $f(x_1, ..., x_F) = 1$.

Normally the regression function is assumed to be of the form $f(x_1, ..., x_F) = \exp^{\beta_1 x_1 + ... + \beta_F x_F}$. Hence, the hazard function is defined as follows:

$$\lambda(t, x_1, ..., x_F) = \lambda_0(t) \exp^{\beta' x} \tag{16}$$

With regression coefficients $\beta = (\beta_1, ..., \beta_F)$ and covariables $x = (x_1, ..., x_F)$. Both, the baseline hazard and the regression coefficients β are unknown and must be estimated from the data. To estimate the regression coefficients it is not necessary to know the baseline hazard. The regression coefficients can be estimated by maximizing the so called logarithmized partial likelihood. Provided that only one event per time point occurs this partial likelihood looks as follows:

$$PL(\beta) = \log(\prod_{i=1}^{n} \frac{\exp^{\beta' x_{(i)}}}{\sum_{m \in M_i} \exp^{\beta' x_m}}) \tag{17}$$

Where $x_{(i)}$ is the vector of covariables for a patient who has an event at time $t_{(i)}$ and M_i is

the set of patients at risk right before time $t_{(i)}$. To get the maximum of the logarithmized partial likelihood the partial derivatives with respect to β has to be equated with zero. This affords the same result as solving the following system of equations:

$$\sum_{i=1}^{n} \left[x_{(i)} - \frac{\sum_{m \in M_i} \exp^{\beta' x_m} x_m}{\sum_{m \in M_i} \exp^{\beta' x_m}} \right] = 0 \tag{18}$$

The solution is a vector $\widehat{\beta}$ which contains the maximum-likelihood-estimates. Of course it is possible that more events per time point appear and not only one. In this case different approximation methods exist. Efron's approximation for example is often used as it is numerically simple and very precise in its approximation. The partial likelihood of the Efron approximation is given by:

$$PLE(\beta) = \log(\prod_{i=1}^{n} \frac{\exp^{\beta' x_{(i)}}}{\prod_{j=1}^{\delta_{(i)}} \left[\sum_{m \in M_i} \exp^{\beta' x_m} - \frac{j-1}{\delta_{(i)}} \sum_{m \in O_{(i)}} \exp^{\beta' x_m} \right]}) \tag{19}$$

Once more $\delta_{(i)}$ is the number of events at time $t_{(i)}$ and $O_{(i)}$ is the set of patients with events at time $t_{(i)}$.

The name Cox proportional-hazards model develops from comparing two observations m and m'. The fitting linear predictors are $\eta_m = \beta_1 x_{m1} + ... + \beta_M x_{mF}$ and $\eta_{m'} = \beta_1 x_{m'1} + ... + \beta_M x_{m'F}$. One can easily see that:

$$\frac{\lambda_m(t)}{\lambda_{m'}(t)} = \frac{\lambda_0(t) \exp^{\eta_m}}{\lambda_0(t) \exp^{\eta_{m'}}} = \frac{\exp^{\eta_m}}{\exp^{\eta_{m'}}} \tag{20}$$

Hence, the hazard ratio of the two observations does not depend on the time t.

In our example the Cox-Regression, with all 27 interesting risk factors included (see page 6), shows that patient-related as well as surgery-related and therapy variables are important. Risk factors like age, congestive heart failure, diabetes mellitus, COPD, extracardiac arteriopathy, LVEF<50%, atrial fibrillation and chronic renal failure increase mortality significantly ($p<0.05$). Weight and sex (i.e. being a male patient) decrease the risk significantly. Sex appears significant only in the multivariate Cox regression analysis, but not in the univariate Log-rank test. This is possible because in the multivariate model the influence of sex is affected by all other covariables.

5 Smoothing Survival Models

Like the Cox-Regression is used to estimate the regression coefficients, i.e. the influence of each risk factor on the hazard, there exist non-parametric estimates for the baseline hazard $\lambda_0(t)$ and therefore the overall hazard rate $\lambda(t)$ (see Wells [1994]), as the latter is the baseline hazard multiplied by constant time values. Remember that the hazard rate is a non-negative function that can (and usually will) change from one time point to another. Like densities, hazard rates are difficult to estimate. Hence, often cumulative hazard rates $\Lambda(t)$ are used as they are smoother and therefore easier to interpret, especially for non-statisticians.

Remember the random censorship model (see page 9) where we have n independent and identically distributed uncensored life times $\tilde{T}_1, \tilde{T}_2, ..., \tilde{T}_n$ and $C_1, C_2, ..., C_n$ as n independent, identically distributed censoring times. Both are assumed to have absolutely continuous distribution functions. The actual observations are specified by the pair (T_i, δ_i) where $T_i = \min(\tilde{T}_i, C_i)$ gives the follow-up time of patient i and $\delta_i = I_{\{\tilde{T}_i \leq C_i\}}$ is the indicator function giving the censoring status. As before, ordered time data are marked by round brackets around the subscript, e.g. $(T_{(i)}, \delta_{(i)})$ is an ordered sample with $T_{(1)} \leq T_{(2)} \leq ... \leq T_{(n)}$ the ordered follow-up times and $\delta_{(i)}$ is the corresponding censoring indicator variable of $T_{(i)}$.

The estimate for the cumulative hazard rate used here is the already mentioned Nelson-Aalen estimate. Remember this non-parametric estimate is defined as:

$$\widehat{\Lambda}(t) = \sum_{i=1, t_{(i)} \leq t}^{n} \frac{\delta_{(i)}}{r_i} \tag{21}$$

Here r_i is the number of patients at risk just before $t_{(i)}$. The increments of this estimate provide information about the hazard rate at the i-th death time (Hess et al. [1999]). An estimate of the hazard rate can generally be obtained by smoothing the increments of the Nelson-Aalen estimate (see Wang [2005]). The most common non-parametric smoothing method, which will be also used here, is the so called kernel method.

5.1 Kernel Smoothing

Kernel smoothing is a statistical technique to estimate functions in a non-parametric way. Principally it is used for density or regression function estimation. A kernel estimate is defined via a kernel function. A kernel function is a bounded weighting function. In most cases kernel functions of order 2 are used and the kernel estimation theory discussed here is only for this special case. Information about higher order kernels can be found in textbooks like Wand and Chris [1995], Haerdle et al. [2007] or Andersen et al. [1993]. Symmetric,

non-negative second order kernel functions are real valued functions $K(u)$ that satisfy:

$$\int_{-\infty}^{\infty} K(u)\,du = 1 \tag{22}$$

$$\int_{-\infty}^{\infty} K^2(u)\,du < \infty \tag{23}$$

$$\int_{-\infty}^{\infty} uK(u)\,du = 0 \tag{24}$$

$$\int_{-\infty}^{\infty} u^2K(u)\,du < \infty \text{ and } \neq 0 \tag{25}$$

Furthermore, the support of a kernel function has to be bounded. For simplicity the support is often defined as $[-1, 1]$. Outside the support the kernel function is 0. A kernel function can be interpreted as a local weighting function, leading to a local weight of the data on the observed time interval (Hess et al. [1999]). There exist many different kernel functions. Some important second order kernels are:

- Rectangle kernel

$$K(u) = \frac{1}{2} I_{\{|u|\leq 1\}} \tag{26}$$

- Epanechnikov kernel

$$K(u) = \frac{3}{4}(1 - u^2) I_{\{|u|\leq 1\}} \tag{27}$$

- Biquadratic kernel

$$K(u) = \frac{15}{16}(1 - u^2)^2 I_{\{|u|\leq 1\}} \tag{28}$$

- Triquadratic kernel

$$K(u) = \frac{35}{32}(1 - u^2)^3 I_{\{|u|\leq 1\}} \tag{29}$$

Although the kernels are quite different (see Figure 3) the resulting estimates are usually quite similar.

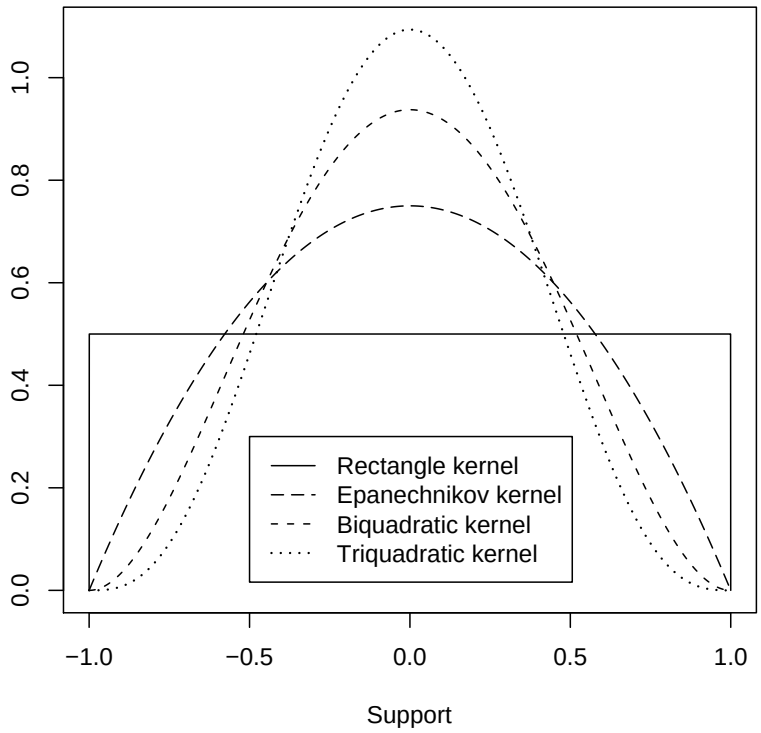

Figure 3: Second order kernels

5.2 Kernel hazard estimate

A general kernel hazard estimate is a convolution of a kernel function $K(u)$ and the Nelson-Aalen estimate $\widehat{\Lambda}(t)$ and looks as follows:

$$\widehat{\lambda}(t) = \frac{1}{b}\int K(\frac{t-s}{b})\, d\widehat{\Lambda}(s) = \frac{1}{b}\sum_{i=1}^{n} K(\frac{t-t_{(i)}}{b})\frac{\delta_{(i)}}{r_i} \tag{30}$$

In the above expression b is a positive, fixed number, called bandwidth of the kernel estimate and should be out of a non-random bandwidth-sequence b_n satisfying $\lim_{n\to\infty} b_n(n) = 0$ and $\lim_{n\to\infty} nb_n(n) = \infty$. The bandwidth influences both, bias and variance of the kernel hazard estimate (Haerdle et al. [2007] and Wang [2005]) and is therefore responsible for a balance in the trade-off between the two factors. If the bandwidth gets larger, the estimate becomes smoother and the variance gets lower, but the bias increases. Hence, the choice of the bandwidth is very important as it is responsible for the degree of smoothing (for details see the following chapter).

The bandwidth of a kernel estimate can be fix i.e. the same on the whole time interval

in which case it is called global bandwidth b. The bandwidth can also change at different time points in which case it is called local bandwidth $b(t)$. Although fixed-bandwidth kernels are often used due to their simplicity, the problem of undesirable effects if the data are not evenly distributed is serious (see Hess et al. [1999]).

Especially near the endpoints of the data unsatisfactory results, so called boundary effects, may occur because of bias problems. Such complications can emerge when the support of the kernel exceeds the available range of the data (Hess et al. [1999]). Usual unmodified kernel estimates may lead to meaningless estimates near the boundaries of the observation interval (see Mueller and Wang [1994]).

There are different possibilities to choose an optimal bandwidth and to handle boundary effects. Some of them will be the topic of the following sections.

5.3 Bandwidth Choice

As mentioned, one distinguishes between local and global bandwidths. In both cases an optimal bandwidth should smooth the data in an appropriate way. So the bandwidth should be large when data are sparse and small when data are dense. If the bandwidth is too small the variance is high and it is possible that random structures of the underlying data that are not part of the true hazard rate may be seen in the estimate. If the bandwidth is too large the bias is high and important structures of the hazard function may be smoothed away.

5.3.1 Optimal global bandwidth

The optimal global bandwidth is usually defined to be the one that minimizes the integrated mean square error:

$$IMSE(b) = \int_0^\infty var(t,b) + bias^2(t,b), dt \tag{31}$$

The IMSE provides a good compromise between variance and bias. As the optimal global bandwidth is based on variance and bias it depends on unknown quantities like the hazard and the empirical survival function and is therefore not directly available in practice. To estimate it cross-validation is often used (see references in Haerdle et al. [2007]). Alternatively, the global optimal bandwidth can be consistently estimated (under mild regularity conditions) by minimizing an asymptotic approximation of the following estimate of the

integrated mean square error:

$$\widehat{IMSE}(b) = \int_0^\infty \widehat{var}(t,b) + \widehat{bias}^2(t,b), dt \tag{32}$$

Bias and variance can be estimated as in Mueller and Wang [1994] in the following way:

$$\widehat{var}(t,b) = \frac{1}{nb} \int K^2(u) \frac{\widehat{\lambda}(t-bu)}{\bar{L}_n(t-bu)} \, du \tag{33}$$

$$\widehat{bias}(t,b) = \int \widehat{\lambda}(t-bu)K(u) \, du - \widehat{\lambda}(t) \tag{34}$$

In the first expression $\bar{L}_n(t) = 1 - L_n(t)$ where $L_n(t) = \frac{1}{n+1} \sum_{i=1}^n I_{\{t_i \leq t, \delta_i = 1\}}$ is the modified empirical survival function of the uncensored observations and $\widehat{\lambda}(t)$ is a pilot estimate of $\lambda(t)$. It can be shown that the asymptotic distribution of the estimated hazard rate is the same for the true optimal global bandwidth and the estimated optimal global bandwidth that minimizes (32) (see Mueller and Wang [1994]). Therefore, minimizing (32) provides an adequate method to estimate the bandwidth. An explicit formula for this optimal bandwidth of the asymptotic approximation can be found, for example, in Andersen et al. [1993] (page 240).

In our example the Epanechnikov-kernel is used because it is very popular and it has certain optimality properties. For example the efficiency of the Epanechnikov-kernel is 1, while other kernels, like the rectangle, biquadratic or triquadratic kernel have efficiency lower than 1 (see Wand and Chris [1995]). The global-bandwidth Epanechnikov-kernel-smoothed Nelson-Aalen estimate for the hazard rate is defined as:

$$\widehat{\lambda}(t) = \frac{1}{b} \sum_{i=1}^n \frac{3}{4} \left[1 - (\frac{t - t_{(i)}}{b})^2 \right] I_{\{|\frac{t-t_{(i)}}{b}| \leq 1\}} \frac{\delta_{(i)}}{r_i} \tag{35}$$

where b is chosen to minimize (32).

Figure 4 shows, based on our medical data example, the Epanechnikov-kernel-smoothed Nelson-Aalen estimate of the hazard rate with an estimated optimal global bandwidth of 103.83. It is based on an initial bandwidth of 101.3020, calculated as recommended in Mueller and Wang [1994] by $b_0 = \frac{R}{8n_u^{\frac{1}{5}}}$ where n_u is the number of uncensored observations and R is the maximal observation time, which is specified here as the time at which at least ten patients are at risk. The hazard rate is then estimated on a predefined time grid using a smoothed version of the optimal global bandwidth. For details see Mueller and Wang [1994].

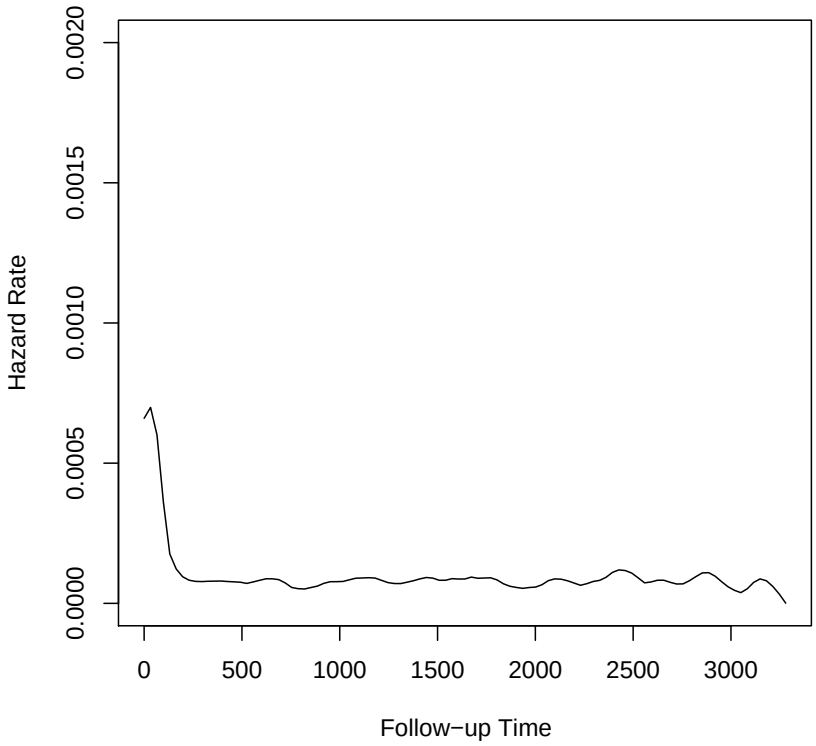

Figure 4: Using optimal global bandwidth

5.3.2 Optimal local bandwidth

The main problem of a global bandwidth is, as mentioned previously, the impossibility of adaption of unevenness in the distribution of the data. A fixed-bandwidth estimate tends to oversmooth in regions with many observations and undersmooth in regions with few observations. There exist different possibilities to deal with this difficulty. One possibility is to use the nearest-neighbor approach, where the information of the k nearest points (neighbors) is included when estimating the hazard rate at a specified time point t. Another possibility is to work with locally optimal bandwidths as discussed below.

The optimal local bandwidth at time point t is usually defined to be the one that minimizes the mean square error:

$$MSE(t, b(t)) = var(t, b(t)) + bias^2(t, b(t)) \tag{36}$$

Like the optimal global bandwidth it can not be calculated exactly since necessary parameters

are unknown. The local optimal bandwidth at time point t can be estimated by minimizing an asymptotic approximation of the following estimate of the local mean square error:

$$\widehat{MSE}(t, b(t)) = \widehat{var}(t, b(t)) + \widehat{bias}^2(t, b(t)) \tag{37}$$

Here variance and bias are estimated as in (33) and (34) on page 20 with b replaced by $b(t)$. Again the asymptotic distribution of the estimated hazard rate is the same using the true and the estimated optimal local bandwidth (see Mueller and Wang [1994]). The IMSE will usually be less for optimal local bandwidths than for an optimal global bandwidth, as for optimal local bandwidths the MSE is minimized at each time point and not only overall. The variance depends on the number of available observations at a considered time point. Therefore local bandwidth estimates permits larger bandwidths at points of larger variance, for example at late time points where few observations are left. Generally local estimates lead to increasing bandwidths near the endpoints (cf. Hess et al. [1999]).

In practice not all time points in the observed time interval are considered to estimate local optimal bandwidths. A grid of time points is chosen at which (37) is minimized. This grid should not be too dense to prevent overfitting. The hazard rate is then estimated on a predefined grid of time points using a smoothed version of the local optimal bandwidths there. For details see Mueller and Wang [1994].

The Epanechnikov-kernel-smoothed Nelson-Aalen estimate for the hazard rate with optimal local bandwidths can be written as:

$$\widehat{\lambda}(t, b(t)) = \frac{1}{b(t)} \sum_{i=1}^{n} \frac{3}{4} \left[1 - (\frac{t - t_{(i)}}{b(t)})^2 \right] I_{\{|\frac{t-t_{(i)}}{b(t)}| \leq 1\}} \frac{\delta_{(i)}}{r_i} \tag{38}$$

The bandwidth $b(t)$ belongs on the actual time point t and minimizes (37).

Figure 5 gives the local-bandwidths Epanechnikov-kernel-smoothed Nelson-Aalen estimate for our medical data example. An initial bandwidth of 101.3020 was chosen, which is the same as for the estimation with a global bandwidth. The estimated optimal local bandwidths change from 20 near the left endpoint up to 2026 over the remaining time interval. These optimal bandwidths are then smoothed at each point of the estimation grid. The time period is again cut when the last ten patients are at risk.

When searching for an optimal bandwidth a finite range of possible bandwidths has to be selected. It is possible that no optimal bandwidth can be found (neither global nor local) in this selection. This occurs for example when the estimated MSE is a monotone function of

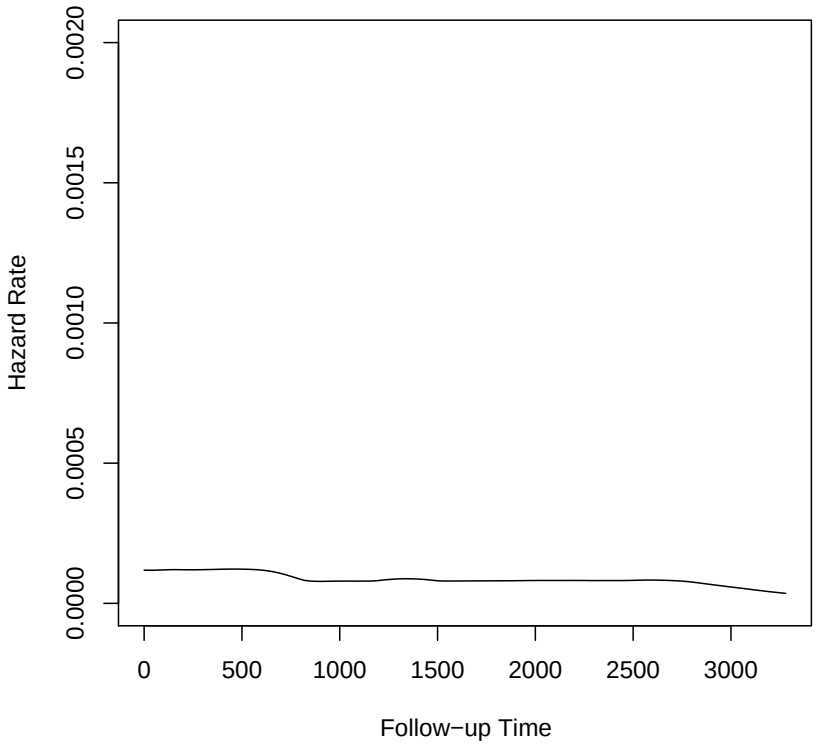

Figure 5: Using optimal local bandwidth

the bandwidth. In such cases the largest value of the preselected bandwidth range is chosen (Hess et al. [1999]).

5.4 Boundary Effects

Choosing an optimal bandwidth (local or global) may not provide satisfactory estimates near the endpoints of the observation period. Such complications near the endpoints are called boundary effects. As mentioned, boundary effects occur when the support of the kernel exceeds the available range of data. Then kernel estimates usually provide biased results. Mathematically the order of the bias is $O(b)$ instead of $O(b^2)$ at boundary points (Karunamuni and Alberts [2004]). These effects are even worse when the hazard rate changes rapidly near the endpoints. Following here the notation of Mueller and Wang [1994] the left and the right boundary regions are defined as $B_l = \{t : 0 \leq t < b(t)\}$ and $B_r = \{t : R-b(t) < t \leq R\}$. The interior is defined as $I = \{t : b(t) \leq t \leq R - b(t)\}$. In the interior region no boundary effects should occur. Remember, R is the maximal observation time.

There are several possibilities to handle such boundary effects. One is just not to estimate the hazard on B_r and B_l, i.e. to restrict hazard estimation on the interior. The problem of this possibility is that the boundary regions can be quite large and may cover important information of the hazard function. Looking at Figure 4 and 5 we see that if our medical data would be cutted near the left endpoint mostly all of the information about the hazard on early time points would be lost. Many different methods to remove boundary effects exist. One is the boundary kernel method (Gasser and Müller, 1979; Gasser, Müller and Mammitzsch, 1985; Jones, 1993; Müller, 1991; Zhang and Karunamuni, 2000). This method uses different polynomial kernels within the boundary regions, so called boundary kernels. It is the one that we will use in the following explanations. Short information and papers about other methods can be found in Karunamuni and Alberts [2004].

A general boundary kernel hazard estimate looks as follows:

$$\widehat{\lambda}(t) = \frac{1}{b(t)} \sum_{i=1}^{n} K_t(\frac{t - t_{(i)}}{b(t)}) \frac{\delta_{(i)}}{r_i} \tag{39}$$

Hence, K_t depends on the time point t where the estimate is to be computed. Boundary kernels have, compared to the interior kernels, asymmetric supports, to avoid the above mentioned problems. Using boundary kernels an increase of the variance is accepted in order to reduce the bias. In this case the application of locally optimal bandwidths can be used to diminish this effect and therefore keep the variance in acceptable limits.

Mueller and Wang [1994] presented a new class of polynomial boundary kernels with advantages over other boundary kernels. They lead, for example, to smaller asymptotic variance. With this approach the inner and outer kernels are defined as:

$$K_t(u) = K_+(1, u) \text{ for } t \in I \tag{40}$$

$$K_t(u) = K_+(\frac{t}{b(t)}, u) \text{ for } t \in B_l \tag{41}$$

$$K_t(u) = K_-(\frac{R - t}{b(t)}, u) \text{ for } t \in B_r \tag{42}$$

where $K_+, K_- : [0, 1] \times [-1, 1]$ are boundary kernel functions. The support of $K_+(q, .)$ is $[-1, q]$ for $0 \leq q \leq 1$ and $K_-(q, t) = K_+(q, -t)$. Both boundary kernel functions have to satisfy $(22) - (25)$ (page 17). Furthermore $K_+(q, .)$ has to be continuously differentiable on its support $(-1, q)$ and $K_+(q, -1)$ has to be zero (see Mueller and Wang [1994]). For $0 \leq q < 1$

(see (41) and (42)) kernels $K_+(q,.)$ are called boundary kernels. For $q = 1$, that is in the inner I (see (40)), the boundary kernel function is equal to the according symmetric kernel function. Note, that due to the construction of these boundary kernels negative values are possible, which could lead to negative hazard rate estimates near the endpoints. It may be reasonable that such estimates are then set to zero, so we consider $\widehat{\lambda}(t) = max(\widehat{\lambda}(t), 0)$ as in Hess et al. [1999] and Mueller and Wang [1994].

Besides the construction of this new class of kernels, Mueller and Wang [1994] also presented explicit formulas for the left boundary kernels for all of the above mentioned kernels (see (26)-(29) on page 17). The left Epanechnikov boundary kernel has, as the general Epanechnikov kernel, optimal mathematical properties (see Wand and Chris [1995]) and is given as:

$$K_+(q,t) = \frac{12(t+1)}{(1+q)^4}\left[(1-2q)t + \frac{3q^2-2q+1}{2}\right] \text{ with support on } [-1,q], q \leq 1 \qquad (43)$$

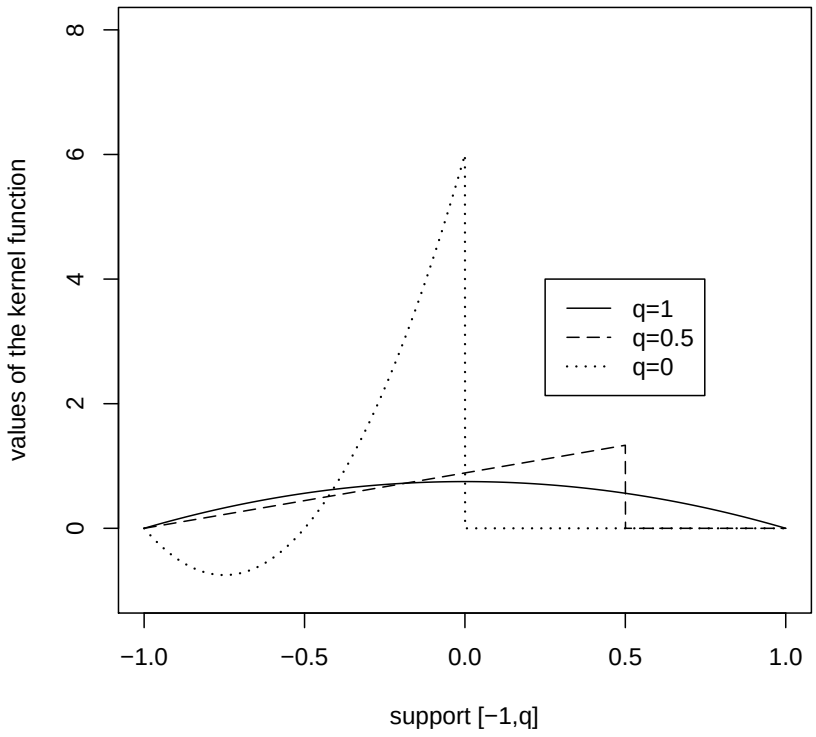

Figure 6: Left Epanechnikov boundary kernel for different q

Figure 6 shows the different left boundary Epanechnikov kernels for different q. For $q = 0$, which only occurs for left boundary kernels if $t = 0$ as $q = \frac{t}{b(t)}$, the kernel exists only

at the negative part of the support. There the kernel takes negative values on the interval $[-1, -0.5]$ of the support. For q between 0 and 1 the kernel changes until for $q = 1$ the general Epanechnikov kernel (compare (27) and Figure 3) is achieved. The kernel is positive on the whole support in the latter case.

As the decreasing number of patients near the right endpoint generally leads to an increase in both variance and bias, it occurs that correcting on the right side leads to a small improvement. The decrease in bias can even be accompanied by an excessive increase in variance. Therefore only left boundary correction is assumed in this work. Nevertheless, to not disregard the problems near the right endpoint remember that the maximal observed time R is chosen to be the maximal time at which at least ten patients are at risk Hess et al. [1999] .

Using a left boundary kernel the Epanechnikov-kernel-smoothed Nelson-Aalen estimate for the hazard rate in the left boundary region (for $t \in B_l$) with optimal local bandwidths looks as follows:

$$\widehat{\lambda}(t, b(t)) = \frac{1}{b(t)} \sum_{i=1}^{n} \frac{12}{(1+q)^4} \left(\frac{t - t_{(i)}}{b(t)} + 1 \right) \left[\frac{t - t_{(i)}}{b(t)} (1 - 2q) + \frac{3q^2 - 2q + 1}{2} \right] I_{\{|\frac{t-t_{(i)}}{b(t)}| \leq 1\}} \frac{\delta_{(i)}}{r_i} \tag{44}$$

For all other time points ($\{t : b(t) \leq t \leq R\}$) the normal Epanechnikov-kernel is used.

In our example the hazard rate was of interest on the time-interval $[0, 3286]$, where 3286 is the longest follow-up period in days. To consider possible right boundary effects the follow-up time is cut when only ten patients are at risk, which occurs on day 3280. Therefore the new time-interval is $[0, 3280]$. Left boundary effects may therefore occur in particular outside the interior $I = \{t : b(t) \leq t \leq 3280\}$. Additionally in this problematic left boundary region $B_l = \{t : 0 \leq t < b(t)\}$ the hazard rate may change quickly from one time point to another, as the number of events decreases fast in the starting period of the data. Therefore, when estimating the Epanechnikov-kernel-smoothed Nelson-Aalen estimate for the hazard rate, left boundary correction was done in the left boundary region using the left Epanechnikov boundary kernel. See Figure 7. Again an initial bandwidth of 101.3020 was chosen. The estimated optimal local bandwidths once more take values between 20 and 2026. While corrected bandwidths near the left endpoint are smaller than the bandwidths calculated without boundary correction, bandwidths farther from the left endpoint are in a large part identical to those calculated without boundary correction. Optimal bandwidths are then smoothed at each point of the estimation grid.

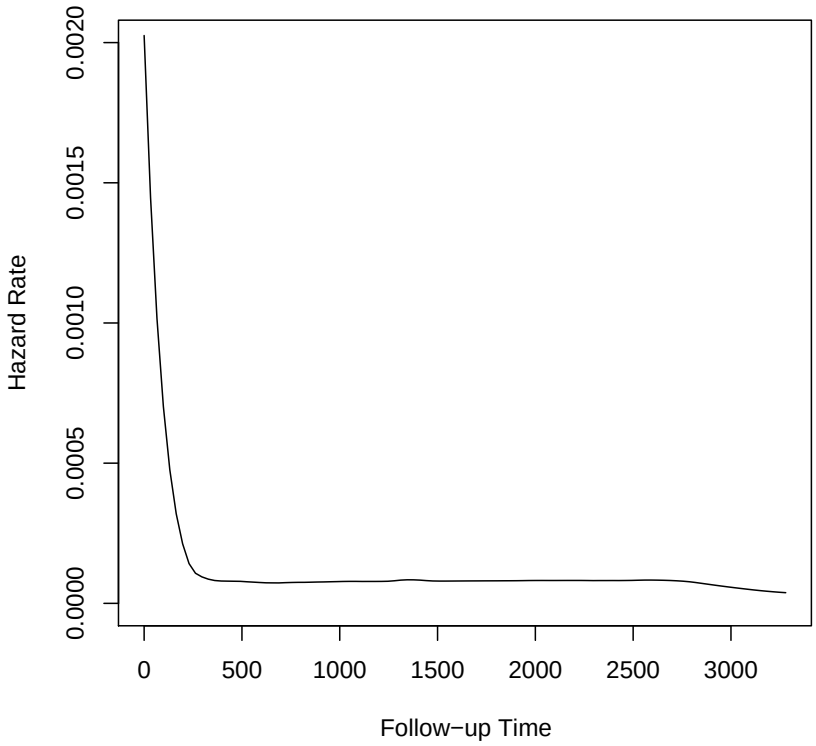

Figure 7: Using optimal local bandwidth and left boundary correction

Figure 8 shows the comparison of all three mentioned kernel-smoothing methods (global bandwidth, local bandwidth and local bandwidth with left boundary correction) for an Epanechnikov-kernel-smoothed Nelson-Aalen estimate. The estimate using an optimal global bandwidth (solid line) shows a bathtub-shape at the left boundary region. This effect is unplausible from a medical point of view and is most probable only a calculational effect. The local optimal bandwidths estimate (dashed line) smooths the global bandwidth estimate. By this, the effect of the high hazard rate at the beginning disappears almost completely. Looking at the estimate using local optimal bandwidths and additionally left boundary correction (dotted line) the strong smoothing disappears, but the bathtub-shape is eliminated. Away from the left boundary region the latter is almost the same as the estimate using only local optimal bandwidths. In published simulation studies the median relative improvement in mean square error for locally optimal bandwidths with left boundary correction over fixed-bandwidth estimates is about 66% (see Hess et al. [1999]). The hazard estimate with local bandwidth and left boundary kernels reflects the medical assumptions for our example data. It has been assumed that patients have a high risk to die immediately after the surgery but if they survive the first time post-surgery the risk decreases rapidly. Therefore the estimate using locally optimal bandwidths and left boundary correction seems to be superior over the

two other possibilities for our data.

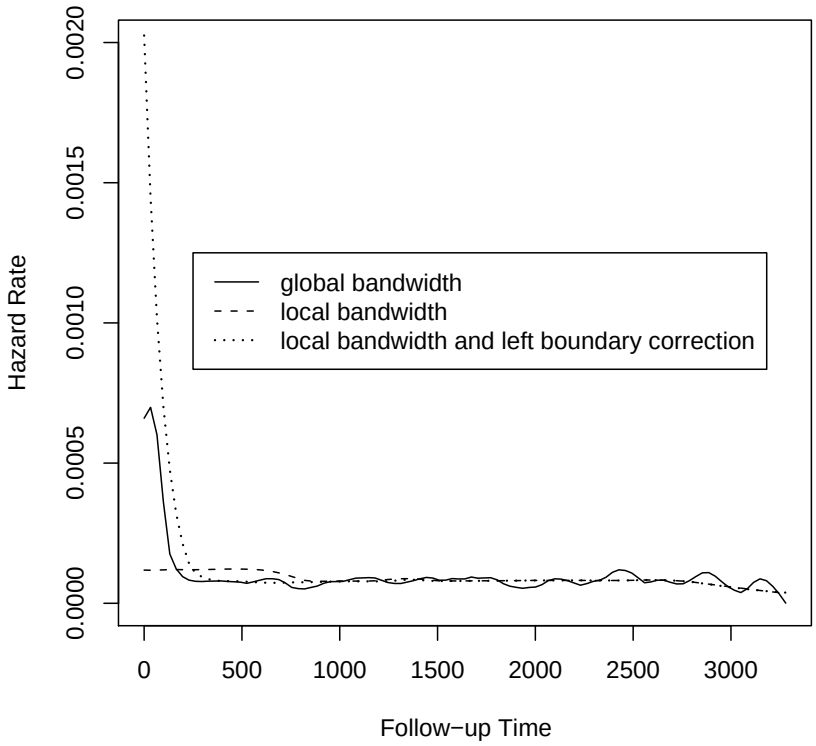

Figure 8: Comparison of the three different hazard estimates

6 Relative Survival Models

There exist many different extensions of the usual survival analysis. One example is robust Cox-Regression, where a weight function is included to the partial likelihood to handle outliers. Another example is the introduction of time-dependent covariables in a survival model, where the survival time is split into time intervals on which all covariables are constant. The extension we are interested in is relative (or excess) survival analysis, where the information of the background mortality is accounted and excess over this background mortality is measured.

The motivation of relative survival analysis is based on the problem that in many studies the cause of death is not definitely known. For example for data including old patients it is possible that the true cause of death is simply high age or some other disease than the one of interest. Due to missing information on the cause of death it is not possible to estimate the relation of death and disease directly. Relative survival analysis tries to handle this problem by including information on the mortality of an adequate background population, gained from population mortality tables of the appropriate region. So it examines whether mortality in the observed data differs from the standard population mortality.

This chapter follows articles like Pohar and Stare [2006], Pohar and Stare [2007] and Andersen et al. [1985].

Usually a population mortality table includes information about age, sex, year and the according mortality rate. As an example a short abstract of the population mortality table of Austria of the year 2005 for people from 60 to 70 years from Statistik Austria is given in Table 2. Mortality is given separately for males and females for each age.

6.1 Important functions and methods

Relative survival models are based on the so called cumulative relative survival function $S_R(t)$. The idea of this relative function is to compare the survival function of the observed patients $S_O(t)$ with the population survival function $S_P(t)$ by building the fraction:

$$S_R(t) = \frac{S_O(t)}{S_P(t)} \tag{45}$$

Here the population survival function $S_P(t)$ is, as mentioned before, estimated on the basis of the population mortality table of the appropriate region. It is clear that, as $S_O(t)$ and $S_P(t)$ both take values between 0 and 1 at each time point, $S_R(t)$ could reach any non-negative

```
Rate table with dimension(s): age sex
            male       female
60 3.433774e-05 1.548616e-05
61 3.110002e-05 1.531903e-05
62 3.450132e-05 1.618591e-05
63 3.734504e-05 1.764572e-05
64 3.858765e-05 1.708584e-05
65 4.296931e-05 2.097186e-05
66 4.762572e-05 2.176463e-05
67 5.103703e-05 2.387501e-05
68 5.433921e-05 2.674088e-05
69 6.227641e-05 2.994550e-05
70 6.764458e-05 3.662960e-05
```

Table 2: Mortality Table

value at each time point t, so $S_R(t) \geq 0$. As the population survival function is often higher than the observed survival function based on patients with a disease, the relative survival function is often less than 1.

Figure 9 shows the connection between the three survival functions (observed, population and relative) in our example. The relative survival function based on censored data is estimated using the Kaplan-Meier method for the observed and the Hakulinen method for the expected (population) survival (see Hakulinen and Tenkanen [1987]). While the population survival function has an almost constant slope, the observed survival function decreases rapidly at the beginning. The difference of these effects can be seen in the estimated relative survival function. While in the first time period the observed survival is worse than the expected survival and the relative survival function decreases, in later time periods the observed survival becomes closer to the population survival and the relative survival function is constant or even increases slightly. The effect of cardiac surgery patients having better relative survival at later time points may occur because the "bad" patients already died in the first time period and only superior patients stay in the study.

The hazard rate is another important tool to observe relative survival models. In contrast to usual survival models there is not a single definition of the relative hazard, but several ones that are based on different assumptions. The three most common definitions are:

- Additive: It is assumed that the observed hazard rate $\lambda_O(t)$ can be split into a population hazard rate $\lambda_P(t)$ and a non-negative relative (or excess) hazard rate $\lambda_R(t)$:

$$\lambda_O(t) = \lambda_P(t) + \lambda_R(t) \tag{46}$$

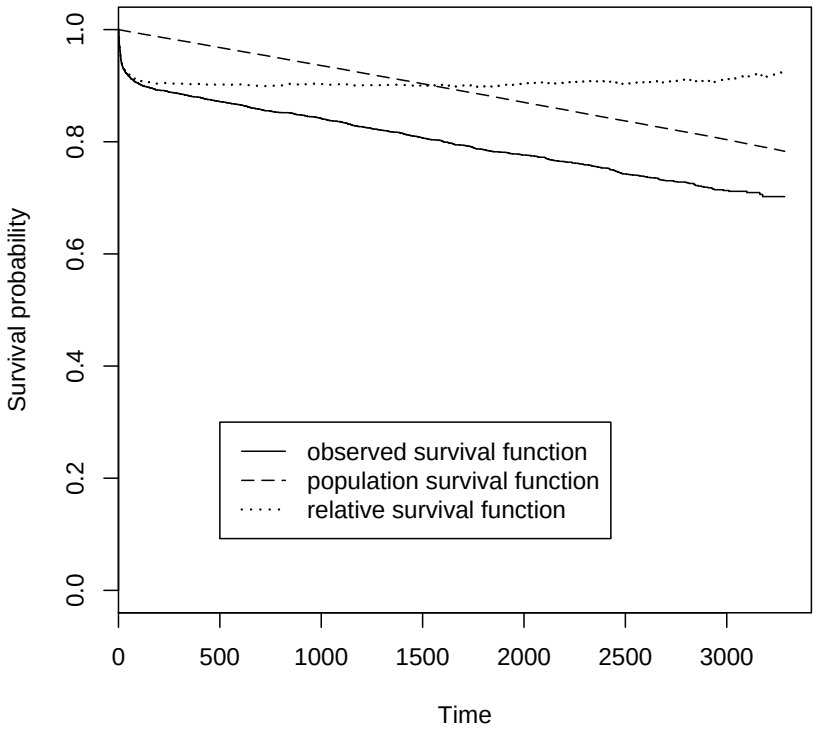

Figure 9: Observed, expected and relative survival function

The relative (or excess) hazard is specific for the disease in question. It is the difference between population and observed hazard rate. This approach assumes that the observed hazard rate is greater than the population hazard rate.

- Multiplicative: The assumption in this definition is that the explained hazard coefficients, population and relative hazard rate, are associated by a multiplicative equation:

$$\lambda_O(t) = \lambda_P(t) * \lambda_R(t) \tag{47}$$

Here the unit-free relative hazard can be seen as relative mortality. Hence, models of this type are sometimes called relative mortality models. Even if the interpretation of this multiplicative assumption is less obvious than of the additive assumption, the main advantage is that the connection between observed and population hazard rate is not predetermined by the equation.

- Transformation: No assumption about the relation between the three hazard rates is made. The individual survival times are transformed by taking the general population mortality into account so that they can be analyzed by any of the ordinary survival

models.

In the past, the additive and multiplicative definitions have been used mostly, while the transformation approach is new (Stare et al. [2005]).

The relative cumulative hazard rate then is the integrated relative hazard independent of the choice of the definition:

$$\Lambda_R(t) = \int_0^t \lambda_R(s)\, ds \tag{48}$$

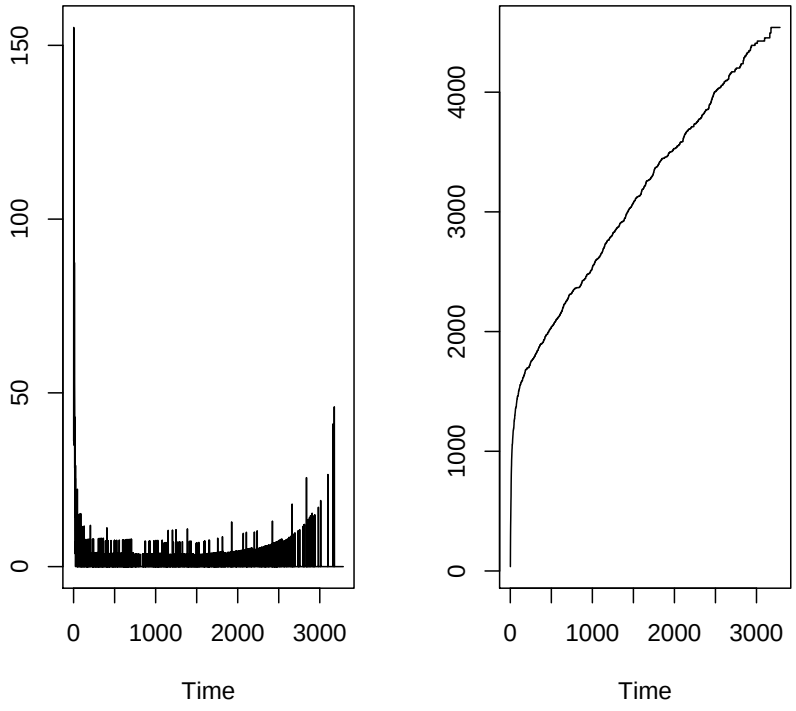

Figure 10: Relative hazard and cumulative relative hazard rate

Figure 10 shows the estimates of the relative hazard and the cumulative relative hazard rate according to the multiplicative definition in our example. Again the hazard is high at the beginning and has an approximately uniform slope afterwards. The values are much larger than 1 in the boundary regions because the observed mortality in these time periods is higher than the mortality of the matched population.

6.2 Andersen multiplicative model

As for usual survival analysis an important tool to estimate the influence of different risk factors on relative survival is regression analysis. Regression models for relative survival are based on the above explained definitions, like Cox-Regression models for usual survival are based on the proportional-hazards assumption (see page 14). As in our example the assumption for the additive model ($\lambda_P < \lambda_O$) seems not to be fulfilled (see Figure 9) this paper concentrates on the so called Andersen multiplicative model. Like the usual hazard function in the Cox proportional-hazards model, the relative hazard function $\lambda_R(t)$ for the relative multiplicative model consists of an unknown relative baseline hazard $\lambda_{R0}(t)$ and an exponential part based on the unknown regression coefficients β and the risk factors x. Thus the observed hazard rate λ_O looks as follows:

$$\lambda_O(t) = \lambda_P(t)\lambda_R(t) = \lambda_P(t)\lambda_{R0}(t)\exp^{\beta' x} = \lambda_{R0}(t)\exp^{\beta' x + 1\ln(\lambda_P(t))} \tag{49}$$

The population hazard rate $\lambda_P(t)$ is assumed to be known and should be calculated using the population mortality table. The model supposes that the mortality of an observed patient is equal to the product of mortality of a person of the same age and sex from the underlying population and a factor explained by the risk indicators. The unspecified relative baseline hazard $\lambda_0(t)$ explains the unknown time-dependent relative mortality. Looking at the transformations the right side is a usual Cox-model including an additional time-dependent risk factor with a known associated regression coefficient of 1. Therefore the unknown regression coefficients β can again, as for the usual Cox model (compare page 14), be calculated by a partial likelihood procedure. The logarithmized partial likelihood function for the relative model looks as follows:

$$PL_R(\beta) = \sum_{i=1}^{n} \delta_{(i)} \left[\beta' x - \log(\sum_{j \in R_i} \lambda_{Pj} \exp^{\beta' x}) \right] \tag{50}$$

Here R_i is the set of patients who are at risk just before $t_{(i)}$ and λ_{Pj} is the population hazard for patient j in R_i.

To fit the model using the Cox-model procedure survival data has to be adapted. The observation period of each patient has to be split into time intervals on which the matched population hazard is constant. Since in the mortality table the hazard changes at the end of every calendar year the follow-up time has to be split to annual intervals. Additionally, since the values in the mortality table also depend an age, further splits have to be done at every anniversary of a patient. It is assumed that a patient gets one year older at every anniversary of his study inclusion day. Mathematically, the time interval $I_i = (0, T_i)$ of patient i is split

into smaller time intervals $I_i = \bigcup_{l=1}^{m_i} I_{il}$. Here I_{il} are the time intervals on which the matched population hazard for patient i is constant and m_i is the number of time intervals of patient i.

The first twenty lines of the adapted data from our medical study are shown in Table 3. The excerpt shows start and stop times of the splitted intervals, the status of the patient and the matched population hazard, different for each time interval. The first patient has 13 time intervals from 0 to 2194 days and is then censored. The second patient has only 5 time intervals and died at 786 days after surgery. For each of these time intervals a different population hazard is calculated out of the according mortality table matched to the appropriate year and to current age and sex of the patient.

	PatId	Start	Stop	Status	Population Hazard
1	1	0.00	2.00	0	8.141881e-05
2	1	2.00	365.24	0	8.035651e-05
3	1	365.24	368.00	0	9.805078e-05
4	1	368.00	730.48	0	9.136157e-05
5	1	730.48	733.00	0	1.027212e-04
6	1	733.00	1095.72	0	9.963158e-05
7	1	1095.72	1098.00	0	1.218403e-04
8	1	1098.00	1460.96	0	1.099600e-04
9	1	1460.96	1463.00	0	1.366344e-04
10	1	1463.00	1826.20	0	1.305344e-04
11	1	1826.20	1829.00	0	1.426533e-04
12	1	1829.00	2191.44	0	1.475200e-04
13	1	2191.44	2194.00	0	1.678518e-04
14	2	0.00	95.00	0	5.288480e-06
15	2	95.00	365.24	0	5.330725e-06
16	2	365.24	461.00	0	6.294603e-06
17	2	461.00	730.48	0	5.495600e-06
18	2	730.48	786.00	1	5.602322e-06
19	3	0.00	66.00	0	1.150060e-04
20	3	66.00	365.24	0	1.043206e-04

Table 3: Relative survival data split into time intervals

Note, that similar time splits are done to handle time dependent covariables in a usual Cox-regression analysis. Hence, in a relative Cox-regression analysis one can easily account for time dependent covariables by adding further splits where the time dependent covariables change.

The Andersen multiplicative model in our example, including all 27 risk factors like in the usual Cox-Regression (see page 6), shows similar results as the usual Cox-Regression

model. Again congestive heart failure, diabetes mellitus, COPD, extracardiac arteriopathy, LVEF<50%, atrial fibrillation and chronic renal failure increase mortality significantly. Also some surgery-related and therapy variables show statistically significant influence once again. What is interesting is that the influence of sex and age change. While higher age and being a male patient increased the risk in the usual Cox-Regression, lower age and being a female patient increases the relative hazard in the relative model. This effect occurs because the population mortality is matched by age and sex. Hence, the higher mortality of old people is fully covered by the population mortality and the reversed effect in the relative survival model indicates that the effect of age is lower in the study population than it is in the general Austrian population. The same idea is imperative for sex. Furthermore weight is not statistically significant any longer.

7 Smoothing Relative Survival Models

In this section we show how the general ideas of smoothing hazard rates (see page 16) can be applied to relative survival models. To this end we need to extend the existing methods on the one hand for left censored data and on the other hand for relative hazard rates. We apply the same kernel functions as for usual survival models, namely those of Mueller and Wang [1994]. We thereby adapt the estimate of the cumulative hazard rate for relative survival by using a Nelson-Aalen type estimate for the cumulative relative (or excess) hazard rate (see (48)):

$$\widehat{\Lambda_R}(t) = \sum_{i=1, t_{(i)} \leq t}^{n} \frac{\delta_{(i)}}{\sum_{j=1}^{n} \sum_{l=1}^{m_j} I_{\{t_{(i)} \in I_{jl}\}} \lambda_{Pjl}(t_{(i)})} \tag{51}$$

Since the time intervals I_{jl} of a patient are disjoint (see page 33) there is at most one time interval from each patient in $\sum_{j=1}^{n} \sum_{l=1}^{m_j} I_{\{t_{(i)} \in I_{jl}\}} \lambda_{Pjl}(t_{(i)})$. If the patient died or was censored before $t_{(i)}$ (e.g. if he is not at risk) there is of course no such time interval. Accordingly, the λ_{Pjl} are the matched population hazards for the l-th time interval of patient j.

A relative kernel hazard estimate is a convolution of a kernel function $K(u)$ and the Nelson-Aalen type estimate $\widehat{\Lambda_R}(t)$:

$$\widehat{\lambda_R}(t) = \frac{1}{b} \int K(\frac{t-s}{b})\, d\widehat{\Lambda_R}(s) = \frac{1}{b} \sum_{i=1}^{n} K(\frac{t - t_{(i)}}{b}) \frac{\delta_{(i)}}{\sum_{j=1}^{n} \sum_{l=1}^{m_j} I_{\{t_{(i)} \in I_{jl}\}} \lambda_{Pjl}(t_{(i)})} \tag{52}$$

Again b means a fix positive bandwidth of the kernel estimate. It is clear that also for relative survival the trade-off between variance and bias is influenced by the choice of the bandwidth. Like for usual survival models a global or a local bandwidth can be chosen. Besides the choice of the bandwidth we also consider boundary correction to advance the estimation. As we have seen before (compare page 28) boundary correction is particularly important for our data example.

For estimating optimal bandwidths and boundary corrected estimates for relative survival models formal analogons as for smoothing usual survival models are used. The asymptotic properties of these relative estimates are not yet investigated. This is a topic for further research.

7.1 Bandwidth Choice

It is already known that the bandwidth is responsible for the degree of smoothing as it impacts the trade-off between variance and bias. Therefore a good choice of the bandwidth is important to fit the data in an appropriate way. An optimal global bandwidth can be achieved by minimizing an estimate of the integrated mean square error, while an optimal local bandwidth can be achieved by minimizing an estimate of the mean square error. To estimate both functions, IMSE and MSE, estimates of the relative variance and the relative bias are needed:

$$\widehat{var}_R(t,b) = \frac{1}{b}\int K^2(u)\frac{\widehat{\lambda_R}(t-bu)}{\sum_{i=1}^{n}\sum_{l=1}^{m_i} I_{\{t\in I_{il}\}}\lambda_{Pi}}\,du \tag{53}$$

$$\widehat{bias}_R(t,b) = \int \widehat{\lambda_R}(t-bu)K(u)\,du - \widehat{\lambda_R}(t) \tag{54}$$

Different to the usual survival case where only uncensored observations are used for the estimation of the empirical survival function $L_n(t)$ (compare Mueller and Wang [1994]) we use all observations here as in Mueller and Wang [1990]. This produces a larger variance and is therefore the more conservative approach. The relative variance estimate is generated by observing that $\bar{L}_n(t)$ in the variance estimate from the usual case (see (33) on page 20) can be approximated in the following way:

$$n\bar{L}_n(t) = n\left[1 - \frac{1}{n+1}\sum_{i=1}^{n} I_{\{T_i\le t\}}\right] \approx \sum_{i=1}^{n} I_{\{T_i>t\}} = \sum_{i=1}^{n} I_{\{t\in I_i\}} = \sum_{i=1}^{n}\sum_{l=1}^{m_i} I_{\{t\in I_{il}\}} \tag{55}$$

For the relative estimate we then change the number of patients at risk $\sum_{i=1}^{n}\sum_{l=1}^{m_i} I_{\{t\in I_{il}\}}$ to $\sum_{i=1}^{n}\sum_{l=1}^{m_i} I_{\{t\in I_{il}\}}\lambda_{Pi}$ as in the Nelson-Aalen type estimate for the relative cumulative hazard rate in (51).

Using above results, optimal global and optimal local bandwidths for relative hazard rates can be estimated. If no optimal bandwidth can be found out of the finite range of preselected bandwidths the largest value of these possible bandwidths is chosen. It is already known (see page 21) that the disadvantage of global bandwidths is that they tend to oversmooth in regions with many observations and undersmooth in regions with few observations. Local bandwidths adjust the data at each time point. Nevertheless, relative hazard rate estimates, as usual hazard rate estimates, are given for both possibilities.

The global-bandwidth Epanechnikov-kernel-smoothed Nelson-Aalen-type estimate for the

relative hazard rate is defined as:

$$\widehat{\lambda_R}(t) = \frac{1}{b}\sum_{i=1}^{n}\frac{3}{4}\left[1-(\frac{t-t_{(i)}}{b})^2\right] I_{\{|\frac{t-t_{(i)}}{b}|\leq 1\}}\frac{\delta_{(i)}}{\sum_{j=1}^{n}\sum_{l=1}^{m_j} I_{\{t_{(i)}\in I_{jl}\}}\lambda_{Pjl}(t_{(i)})} \tag{56}$$

where b is chosen to minimize an estimation of the IMSE (see (32) on page 19) based on relative variance (see (53)) and relative bias (see (54)).

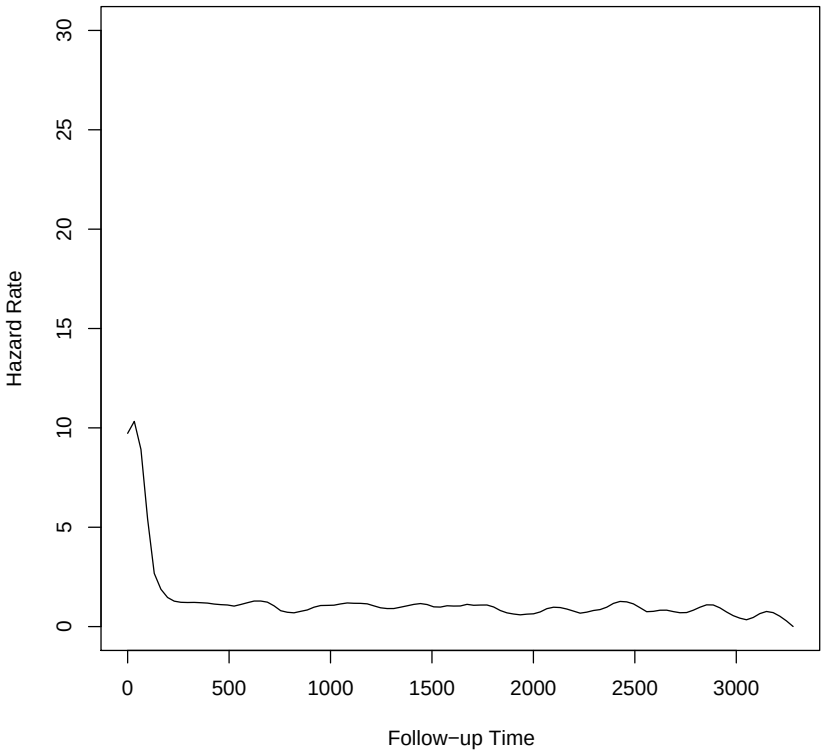

Figure 11: Using optimal global bandwidth (relative)

Figure 11 shows the relative hazard rate estimate for our medical data example using an Epanechnikov kernel, the Nelson-Aalen-type estimate for the relative cumulative hazard rate and a global bandwidth of 103.83. The optimal global bandwidth from the usual hazard rate estimation is used as numerical problems did not allow to estimate optimal bandwidths for relative models in time. Like for usual survival models the maximal observation time is the one at which at least ten patients are at risk.

The Epanechnikov-kernel-smoothed Nelson-Aalen-type estimate for the relative hazard

rate with locally adapted bandwidths looks as follows:

$$\widehat{\lambda_R}(t, b(t)) = \frac{1}{b(t)} \sum_{i=1}^{n} \frac{3}{4} \left[1 - (\frac{t - t_{(i)}}{b(t)})^2 \right] I_{\{|\frac{t-t_{(i)}}{b(t)}| \leq 1\}} \frac{\delta_{(i)}}{\sum_{j=1}^{n} \sum_{l=1}^{m_j} I_{\{t_{(i)} \in I_{jl}\}} \lambda_{Pjl}(t_{(i)})} \tag{57}$$

Here the bandwidth $b(t)$ belongs on the actual time point t and minimizes an estimation of the MSE (see (37) on page 21) based on (53) and (54) (page 37).

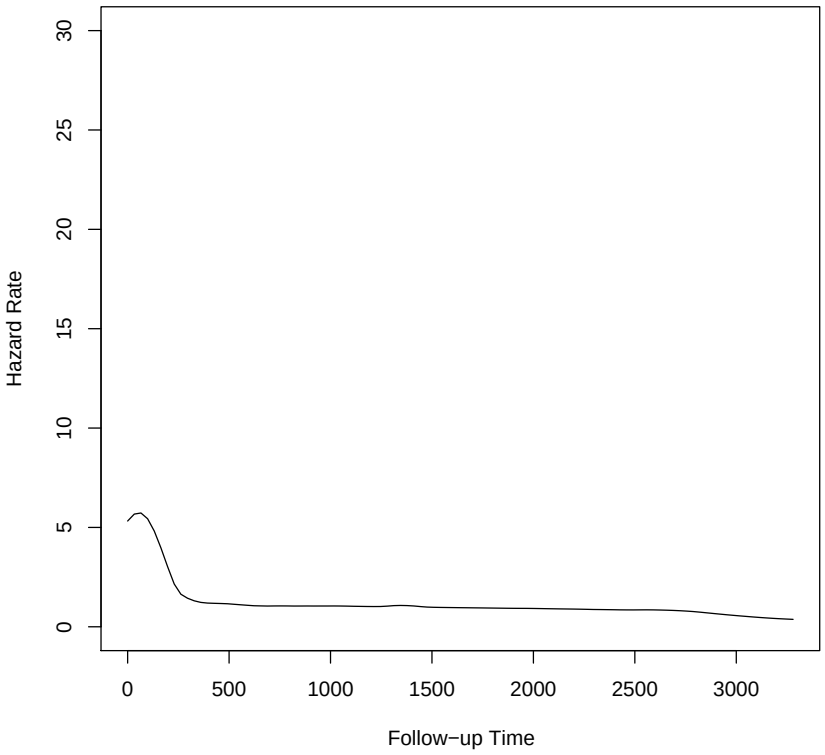

Figure 12: Using optimal local bandwidth (relative)

The Nelson-Aalen-type estimate of the relative hazard rate with Epanechnikov kernel and local bandwidth choice based on our medical data example is given in Figure 12. Here we used the local optimal bandwidths of the usual hazard rate estimate. Hence, the local bandwidths change from 20 to 2026, where lower bandwidths occur near the left time endpoint and higher bandwidths occur over the remaining time interval. The time period is cut when at least ten patients are at risk.

7.2 Boundary Effects

As for usual hazard rate estimation boundary effects may also occur for relative hazard rate estimations when only the bandwidth changes but the kernel itself remains constant over time. Boundary effects occur when the kernel exceeds the available range of data. To handle this problem we apply boundary kernels as introduced by Mueller and Wang [1994] (see page 24 ff.).

Accordingly, we use Epanechnikov kernels for the interior $I = \{t : b(t) \leq t \leq R - b(t)\}$ where R is the maximal observation time, while we favor left Epanechnikov boundary kernels (see (42) on page 25) for the left boundary region $B_l = \{t : 0 \leq t < b(t)\}$. Boundary kernels for the right boundary regions often lead to only a little better results. Hence, we do not use boundary kernels at the right time endpoint but cut the observed time when at least ten patients are at risk.

For $t \in B_l$ a left Epanechnikov-kernel-smoothed boundary estimate for the relative hazard rate with local optimal bandwidths is given as follows:

$$\widehat{\lambda}(t, b(t)) = \frac{1}{b(t)} \sum_{i=1}^{n} \frac{12}{(1+q)^4} \left(\frac{t - t_{(i)}}{b(t)} + 1\right) \left[\frac{t - t_{(i)}}{b(t)}(1-2q) + \frac{3q^2 - 2q + 1}{2}\right] \cdot \tag{58}$$

$$\cdot\, I_{\{|\frac{t-t_{(i)}}{b(t)}| \leq 1\}} \frac{\delta_{(i)}}{\sum_{j=1}^{n} \sum_{l=1}^{m_j} I_{\{t_{(i)} \in I_{jl}\}} \lambda_{Pjl}(t_{(i)})}$$

For interior points the normal locally adapted Epanechnikov-kernel (see (57)) is used.

Based on our data an Epanechnikov-kernel-smoothed Nelson-Aalen-type estimate of the hazard rate with locally adapted bandwidths and left boundary correction can be seen in Figure 13. While right boundary correction is considered by cutting the time interval when at least ten patients are at risk (day 3280 in our medical data example), left boundary correction is considered using left Epanechnikov boundary kernels. The local optimal bandwidths from the usual hazard rate estimation with left boundary correction are used. Once more they take values between 20 and 2026, where bandwidths near the left endpoint are smaller than the bandwidths calculated without boundary correction.

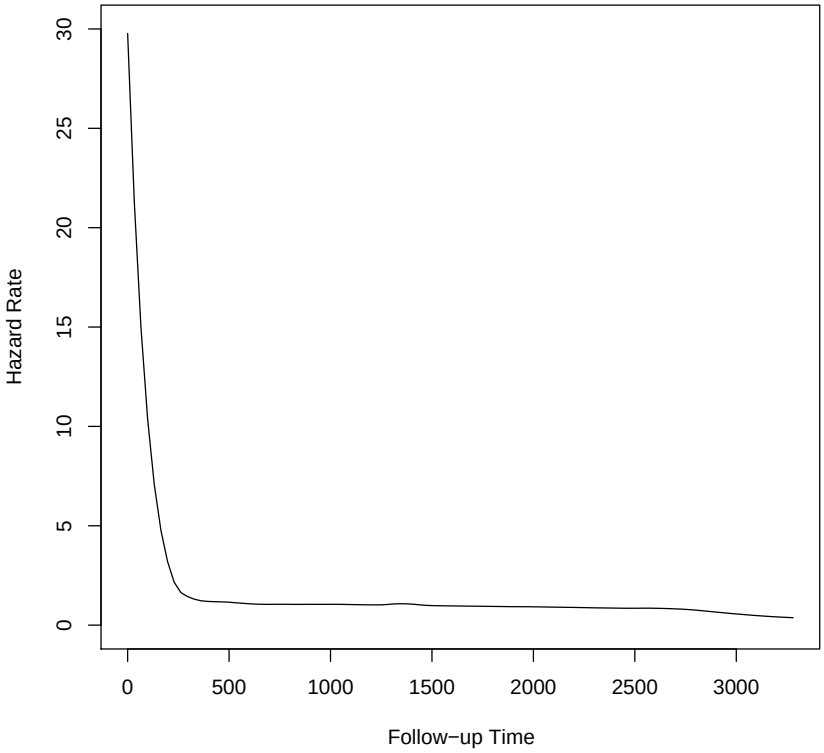

Figure 13: Using optimal local bandwidth and left boundary correction (relative)

Figure 14 shows the comparison of the three relative hazard rate estimates. The estimates using global bandwidth (solid line) and locally adapted bandwidths (dashed line) are very similar. Both show increased bathtub-shapes at the beginning and decrease after about one year to a level about 1. While the global bandwidth curve is higher at the beginning and varies more during the remaining time period, the locally adapted curve is lower at the beginning and almost horizontal afterwards. The relative hazard rate estimate using locally adapted bandwidths and left boundary correction (dotted line) is very high at the beginning and removes the probably calculational effect of the bathtub-shape. It decreases quickly within the first year and is then almost equal to the estimate using only locally adapted bandwidths. As the population hazard is assumed to be 1 (line alternating dots and dashes) in relative hazard rate estimations this would mean that for patients who survive the first time post-surgery the risk to die does not only decrease rapidly, but it even decreases until it is almost equal to the mortality in the normal population after one to two years.

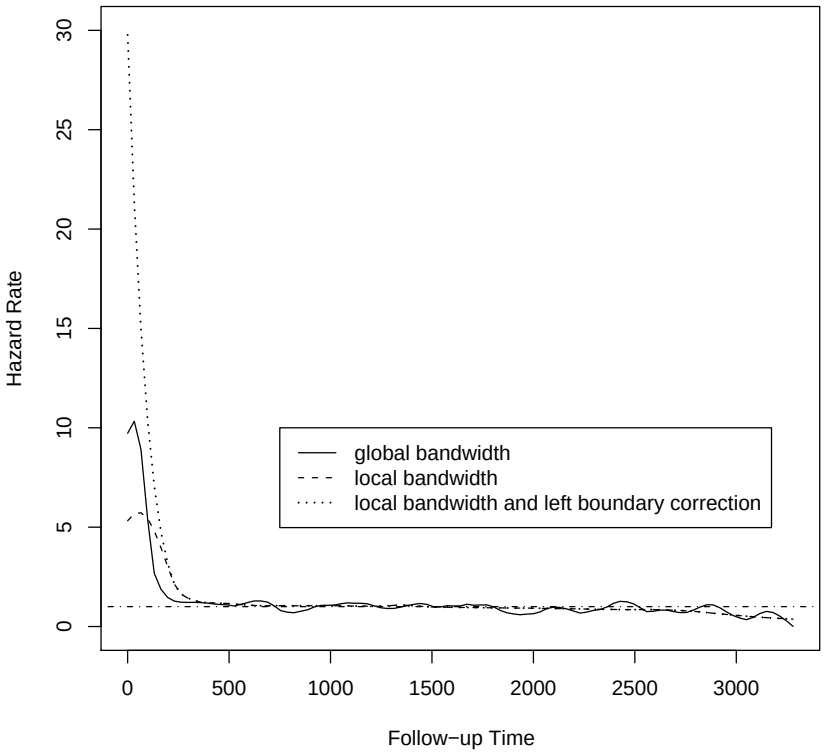

Figure 14: Comparison of the three different relative hazard estimates

8 Discussion

The results of Hess et al. [1999] confirmed the advantages of boundary correction and optimal bandwidths for smoothing usual hazard rates. The median relative improvement over fixed-bandwidth estimates without boundary correction was highest for locally optimal bandwidths with left boundary correction. Similar results can be seen when smoothing relative hazard rates for our data example using adapted bandwidths and left boundary correction. Also in our example the best result can be achieved using locally adapted bandwidths and left boundary correction. Hence, while Hess et al. [1999] shows that locally optimal bandwidths plays the decisive role in smoothing usual hazard rates, the effect of locally adapted bandwidths seems not to be that important in our case as it only smooths the fixed-bandwidth estimate a little more. Instead left boundary correction shows the most important change in our estimates as it can handle the unplausible bathtub-shape effect near the left endpoint. Although left boundary correction also further improved the results in Hess et al. [1999], it only plays a minor role in their simulations. As Hess et al. [1999] demonstrates that estimates using left and right boundary correction sometimes upgrade and sometimes degrade the results, the problem of right boundary correction in our data example is considered by limiting the estimate to time points when at least ten patients are at risk.

There are some points that should be done furthermore to approve our results and to upgrade relative survival theory. The most important point is of course to implement the missing parts of the estimation of global and local optimal bandwidths which was not possible here because of numerical problems. Further asymptotic results should be investigated to see whether our assumption, that smoothing hazard rate estimation formulas can easily be adapted for relative survival by using the relative hazard rate and the relative empirical survival function instead of the usual functions, can be confirmed or not. It would be interesting to do simulations, like Hess et al. [1999] for usual hazard function estimations, to compare the influence of adapted bandwidths and boundary correction on the MSE (respectively the IMSE), hence the relative improvement of the different estimations.

References

P. Andersen, K. Borch-Johnson, T. Deckert, A. Green, P. Hougaard, N. Keiding, and S. Kreiner. A cox regression model for the relative mortality and its application to diabetes mellitus survival data. *Biometrics*, 41(4):921–932, 1985.

P. Andersen, O. Borgan, R. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. Springer, 1993. ISBN 0-387-97872-0.

R. Cao, P. Janssen, and N. Veraverbeke. Relative hazard rate estimation for right censored and left truncated data. *Sociedad de Estadistica e Investigacion Operativa*, 14:257–280, 2005.

D. Collett. *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC, 2 edition, 2005. ISBN 1-58488-325-1.

J. Fox. Appendix to an r and s-plus companion to applied regression. `http://cran.r-project.org/doc/contrib/Fox-Companion/appendix.html`, 2002.

W. Haerdle, Y. Mori, and P. Vieu. *Statistical Methods for Biostatistics and Related Fields*. Springer, 2007. ISBN 978-3-540-32690-8.

T. Hakulinen and L. Tenkanen. Regression analysis of relative survival rates. *Journal of the Royal Statistical Society, Series C*, 36:309–317, 1987.

K. Hess, D. Serachitopol, and B. Brown. Hazard function estimators: A simulation study. *Statistics in Medicine*, 18:3075–3088, 1999.

J. Kalbfleisch and R. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley Interscience, 2 edition, 2002. ISBN 0-471-36357-X.

R. Karunamuni and T. Alberts. On boundary correction in kernel density estimation. Technical report, Fifth Biennial IISA International Conference on Statistics, Probability and Related Areas, University of Georgia, Athens, Georgia, 2004.

A. Lassnigg, M. Hiesmayr, S. Frantal, W. Brannath, M. Mouhieddine, C. Isetta, and D. Schmidlin. Cardiac surgical mortality, a population based study. Unpublished paper, 2010.

H. Mueller and J. Wang. Locally adaptive hazard smoothing. *Probability Theory and Related Fields*, 85:523–538, 1990.

H. Mueller and J. Wang. Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics*, 50:61–76, 1994.

M. Pohar and J. Stare. Relative survival analysis in r. *Elsevier Science, Irlande*, 81(3): 272–278, 2006.

M. Pohar and J. Stare. Making relative survival analysis relatively easy. *Elsevier Science, Irlande*, 37:1741–1749, 2007.

J. Stare, R. Henderson, and M. Pohar. An individual measure of relative survival. *Journal of the Royal Statistical Society, Series C*, 54:115–126, 2005.

E. Vittinghoff, D. Glidden, S. Shiboski, and C. McCulloch. *Regression methods in Biostatistics: Linear, Logistic, Suvival and Repeated Measures Models*. Springer, Germany, 2005. ISBN 0-387-20275-7.

M. Wand and J. Chris. *Kernel smoothing*. Chapman and Hall, 1995. ISBN 0-412-55270-1.

J. Wang. *Encycolpedia of Biostatistics*, chapter Smoothing Hazard Rates. Wiley, 2005.

M. Wells. Nonparametric kernel estimation in counting processes with explanatory variables. *Biometrika*, 81(4):795–801, 1994.

List of Tables

List of Figures

R-Code

All analyses are calculated using R 2.8.1. Besides the general R-features there exist different R-packages which can be installed and loaded directly from R. There exist useful packages for many mathematical or statistical problems, also libraries for survival- and relative survival analysis. Principally used in this work is the package survival for survival analysis and the package *relsurv* for relative survival analysis. For different smoothing options and bandwidth selection for usual survival analysis the package *muhaz* is available. Look at the R online help for more information: file:///C:/Program%20Files/R/R-2.8.1/doc/html/index.html.

```
library(survival)
library(relsurv)
library(muhaz)
```

For the Kaplan-Meier estimation the estimated survival function can be given by using the $Surv()$ and $survfit()$ functions. $Surv()$ creates a survival object, usually used as a response variable in a model formula. $Surv()$ objects are implemented as a matrix of 2 or 3 columns, giving the survival times and the censoring variable. The $survfit()$ function computes an estimate of a survival curve for censored data using either the Kaplan-Meier or the Fleming-Harrington method. The Kaplan-Meier estimate can be used overall or for different strata, in this example male and female.

```
attach(data)
KM<-survfit(Surv(Death_SurvivalDays,Death),data=data)
KM2<-survfit(Surv(Death_SurvivalDays,Death)~SexCode,data=data)
```

To plot such Kaplan-Meier curves the simple $plot()$ function can be used. It is a generic function for plotting R-objects where many different graphical parameters can be chosen. Some interesting ones that are used for the following plots will be explained. For further information see "par" in the R-help. Compare figure 1 on page 11.

```
par(pin=c(2,4),mfrow=c(1,2))
plot(KM,conf.int=FALSE,xlab="Time",ylab="Survival Probability")
plot(KM2, lty=c(1,3),xlab="Time",ylab="Survival Probability")
legend(500,0.2,legend=c("Male","Female"),lty=c(1,3))
par(mfrow=c(1,1))
```

To get the hazard rate the connection between the cumulative hazard rate and the survival function is used. First the cumulative hazard is calculated using a Nelson-Aalen estimate

by using information from the above estimated Kaplan-Meier plot and typical R-functions. The number of events and the number at risk at each time point is needed. Then numeric derivation is used to get the hazard from the cumulative hazard rate.

```
num.ab <- function(time,y){
  x <- integer(length(y)-1)
  x[[1]] <- y[[1]]
  for(i in 2:length(y)){
    x[[i-1]] <- (y[[i]]-y[[i-1]])/(time[[i]]-time[[i-1]])}
  time <- time[1:(length(y)-1)]
  list(time=time,x=x)}
time1<-KM$time
surv2<-KM$surv
cumhaz1<--log(surv2)
hazmatrix1<-num.ab(time1,cumhaz1)
time1a<-KM$time[KM$n.event>0]
NA1<-unique(cumsum(KM$n.event/KM$n.risk))
```

The hazard rate is plotted based on the calculation with the numeric derivation. Same *plot*() function as above. Compare figure 2 on page 12.

```
par(pin=c(2,4),mfrow=c(1,2))
plot(hazmatrix1[[1]],hazmatrix1[[2]],type="s",xlab="Time",sub="Hazard")
plot(time1a,NA1,type="l",xlab="Time",sub="cumulative hazard")
par(mfrow=c(1,1))
```

To test the difference between males and females, as seen in the Kaplan-Meier-plot, using the Log-rank test the R-function $survdiff()$ can be used. The function tests if there is a difference between the two survival curves.

```
LogRank<-survdiff(Surv(Death_SurvivalDays,Death)~SexCode,data=data)
SumLogRank<-summary(LogRank)
```

The R-output of the Log Rank test first gives the underlying formula. The small table presents the number of subjects in each group ”n” and the number of observed and expected events per group. The ”chisq” shows the test statistic for a test of equality with the according degrees of freedom and with the depending p-value.

Using Cox-Regression the influence of all explained, interesting risk factors can be calculated. The used function is $coxph()$, again based on the $Surv()$ function. A Cox proportional-hazards regression model is considered which is usually expressed in terms of a single survival time value for each person, with possible censoring.

```
CoxReg<-coxph(Surv(Death_SurvivalDays,Death)~Age+Weight+SexCode+OPCABGplus+
  OPCABGOff+OPValv+OPTAA+Revision48+Revisionspaeter+CardialeDekomp+Asthma+
  COPD+Diabetes+ANiereninsufchron+MACE+MDiuretika+UrgentOP+HLMIABP+Ery+TK+
  HF+EF1neu+EF2neu+EKG_VHFL+Heartpump+Infarct+CaroPAVK,data=data)
SumCoxReg<-summary(CoxReg)
```

Again the output shows the underlying formula and the number of observed patients. "Coef" shows the regression coefficients β. "exp(coef)" is the hazard ratio, so the ratio of hazard rates of different groups (e.g.: male or female patients). It is essential that an effect bigger than 1 means higher risk and accordingly an effect smaller than 1 means lower risk. Note: R always tests for the highest value of a risk factor. For metric variables it tests for "the higher the variable" and for nominal variables it tests for the highest value e.g. for a dichotomous variable with values 0 and 1 it tests for 1. "Lower" and "upper" are the confidence bounds for the regression function. "se(coef)" displays the standard error. "z" shows the underlying test statistic and "p" the associated p-value. Normally the α-value is set to 0.05 and all p-values smaller than this value are considered as statistically significant. The last section shows the R-squared which indicates the illustrated variance of the model. It reaches values between 0 and 1, the higher, the better. The three tests are so called overall tests as they test the whole model with the null hypothesis $H_0 : \beta = (\beta_1, ..., \beta_F) = 0$. If they show no significant result and the tests on the individual variables do, then the model is not good and it does not fit the appropriate reality in an acceptable way.

To draw the distribution functions of the different introduced kernels only general R-functions are used. Also the distribution function of the left Epanechnikov boundary kernel depending on the choice of q only uses simple R-functions. Each formula is applied to each grid point of a grid between -1 and 1.

```
x<-seq(-1,1,0.0005)
RK<-c(0,rep(1/2,length(x)-2),0)
EK<-3/4*(1-x^2)
BK<-15/16*(1-x^2)^2
TK<-35/32*(1-x^2)^3
bk<-function(q){
```

```
  x<-seq(-1,q,0.0005)
  K<-12/(1+q)^4*(x+1)*(x*(1-2*q)+(3*q^2-2*q+1)/2)}
```

Again the normal *plot*() function is used to get both figures. The first figure compares the four general kernel functions (rectangle kernel, Epanechnikov kernel, biquadratic kernel and triquadratic kernel), while the second figure shows the left Epanechnikov boundary kernel. To plot all four general kernel distributions in one window the function *lines*() can be used. It is a generic function taking coordinates given in various ways and joining the corresponding points with line segments to a plot. Compare figure 3 on page 18 and figure 6 on page 25.

```
plot(x,TK,type="l",lty=3)
lines(x,EK,lty=5)
lines(x,BK,lty=2)
lines(x,RK,lty=1)
legend(-0.5,0.3,lty=c(1,5,2,3),legend=c("Rectangle kernel","Epanechnikov
  kernel","Biquadratic kernel","Triquadratic kernel"))

plot(seq(-1,1,0.0005),bk(1),type="l",ylim=c(-1,8),ylab="values of the kernel
  function",xlab="support [-1,q]")
lines(seq(-1,1,0.0005),c(bk(0.5),rep(0,length(seq(-1,1,0.0005))-
  length(bk(0.5)))), lty=5)
lines(seq(-1,1,0.0005),c(bk(0),rep(0,length(seq(-1,1,0.0005))-length(bk(0)))),
  lty=3)
legend(0.25,4,legend=c("q=1","q=0.5","q=0"),lty=c(1,5,3))
```

The different smooth hazard rate estimates using Epanechnikov kernels and alternatively global or local bandwidths and none or left boundary correction are calculated using the function *muhaz*(). It estimates the hazard function from right-censored data using kernel-based methods based on the survival time and the censoring status. Options include three types of bandwidth functions, three types of boundary correction, and four shapes for the kernel function. The statistical properties of many of these estimates are reported and compared in Hess et al. [1999]. The algorithm of *muhaz*() is explained in Mueller and Wang [1994].

```
test1<-muhaz(Death_SurvivalDays,Death,bw.method="global",b.cor="none",
  kern="epanechnikov")
test2<-muhaz(Death_SurvivalDays,Death,bw.method="local",b.cor="none",
  kern="epanechnikov")
```

```
test3<-muhaz(Death_SurvivalDays,Death,bw.method="local",b.cor="left",
  kern="epanechnikov")
```

To plot the three estimated hazard functions and the comparison of all three the function *plot.muhaz*() is used. It plots functions from an object of class muhaz. Default time limits are those provided to muhaz, which default to zero and the time corresponding to when ten patients remain at risk. Default y-axis limits are 0 and the maximum estimated hazard rate. Additional lines can be added to the same set of axes using the general *lines*() function. Compare figure 4 on page 21, figure 5 on page 23, figure 7 on page 27 and figure 8 on page 28.

```
plot.muhaz(test1,ylim=c(0,0.002))
plot.muhaz(test2,ylim=c(0,0.002))
plot.muhaz(test3,ylim=c(0,0.002))

plot.muhaz(test1,ylim=c(0,0.002),lty=1)
lines(test2,lty=2)
lines(test3,lty=3)
legend(750,0.00125,legend=c("global bandwidth","local bandwidth","local
  bandwidth and left boundary correction"),lty=c(1,2,3))
```

To compare the observed and the expected survival function and to look at the relative survival function the three curves have to be first calculated again. For the observed survival the normal Kaplan-Meier-estimate is used. To get an overall survival curve for the expected survival of a cohort of subjects the *survexp*() function is chosen. In the used Hakulinen method the cohort is recommended to be censored at the anticipated censoring time of each patient.

Including no risk factor, but the ratetable object, gives just the overall information. The ratetable object, given by the ratetable() function, matches the names of the data with those used in the mortality table. Therefore beside the information of the used data set "data=", also the used mortality table "ratetable=" and the used follow-up times "times=" has to be given. The mortality table in the form that is needed by R is built up using the transrate() function, which assists in reorganizing certain types of data into a ratetable object. Here two matrices containing the yearly (conditional) probabilities of one year survival (!!!) for male and female and the interesting time period "yearlim=" has to be given. So actually the mortality tables from Statistik Austria have to be transformed into survival probabilities (1-mortality) to be used in R ("male","female2"). Note, that mortalities of 0 are not possible and have to be replaced by sufficiently small mortalities. All time values have to be given in days.

Also the relative survival is calculated. the function *rs.surv*() computes an estimate of the relative survival curve using the Kaplan-Meier method for the observed and the Hakulinen method for the expected survival. Again the ratetable object is used.

```
mortality<-transrate(as.matrix(male),as.matrix(female2),yearlim=c(1997,2005))
expsurv<-survexp(Surv(Death_SurvivalDays,Death)~ratetable(age=Age*365.24,
  sex=SexCode,year=OPDate),data=data,ratetable=mortality,times=KM$time)
relsurv<-rs.surv(Surv(Death_SurvivalDays,Death)~ratetable(age=Age*365.24,
  sex=SexCode,year=OPDate),data=data,ratetable=mortality)
```

For visual comparison between the subjects and the population at large, both, the observed and the expected survival function, are added to the plot of the relative survival function. Once again the normal *plot*() function with additional *lines*() functions and a legend is used. Compare figure 9 on page 31.

```
plot(relsurv,xlab="Time",ylab="Survival probability",conf.int=FALSE)
lines(KM,lty=5)
lines(expsurv,lty=3)
legend(500,0.3,legend=c("relative survival function", "observed survival
  function","expected survival function"),lty=c(1,5,3))
```

To get the relative hazard rate the same idea as later for relative regression and smoothing relative hazard rates is consulted (see page 33). Only the simple hazard estimation is applied, no kernel smoothing and no optimal bandwidth and boundary correction are used. The relative cumulative hazard rate is then simply calculated as the cumulative sum of the relative hazard rate.

```
core<-"Surv(Death_SurvivalDays,Death)~"
core1<-"+ratetable(age=Age*365.24,sex=SexCode,year=OPDate)"
vari1<-"1"
form1<-as.formula(paste(core,vari1,core1))
model1<-rsmul(form1,ratetable=mortality,data=data,method="mul1")
a<-data.frame(start=model$y[,1],stop=model$y[,2],cens=model$y[,3],lambdap=
  model$lambda)
b<-a[order(a$stop),]
t<-unique(b$stop)
d<-integer(length(t))
s<-integer(length(t))
```

```
q<-integer(length(t))
for (j in 1:length(t)){
  d[[j]]=length(b[(b$stop==t[[j]])&(b$cens==1),"stop"])
  s[[j]]=sum(exp(b[((b$stop>=t[[j]])&(b$start<t[[j]])),"lambdap"]))
  q[[j]]=d[[j]]/s[[j]]}
ch<-cumsum(q)
```

Similar to the usual survival case relative hazard and relative cumulative hazard rate can easily be plotted using the known *plot*() function. See figure 10 on page 32.

```
par(pin=c(2,4),mfrow=c(1,2))
plot(t,q,type="l",xlab="Time",ylab="Relative hazard")
plot(t,ch,type="l",xlab="Time",ylab="cumulative relative hazard")
par(mfrow=c(1,1))
```

To see the influence of the 27 interesting risk factors on relative survival, so on the relative hazard rate, the Andersen multiplicative regression model is used. It is implemented in the *rsmul*() function which is just an extension of the *coxph*() function using relative survival. Again the method uses the *ratetable*() object. The used method "mul1" is just more accurate than the default method "mul", although it can be more computationally expensive. "Mul1" uses the ratetable to determine the time points when hazard changes, whereas "mul" assumes the hazard to be constant on yearly intervals.

```
vari<-"Age+Weight+SexCode+OPCABGplus+OPCABGOff+OPValv+OPTAA+Revision48+
  Revisionspaeter+CardialeDekomp+Asthma+COPD+Diabetes+ANiereninsufchron+MACE+
  MDiuretika+UrgentOP+HLMIABP+Ery+TK+HF+EF1neu+EF2neu+EKG_VHFL+Heartpump+
  Infarct+CaroPAVK"
form <-as.formula(paste(core,vari,core1))
model<-rsmul(form,ratetable=mortality,data=data,method="mul1")
Summodel<-summary(model)
```

The output is, looking at the present arguments, very similar to those from the usual Cox-regression. Again after the model formula the small table presents the regression coefficients, the hazard ratio, the standard errors, the test statistics and the p-values. Also one overall test, here the Likelihood ratio test, with degrees of freedom, number of observations and the according overall p-value is given. The important difference is the interpretation, as for the relative survival analysis the values explain the influence on the relative hazard and not the observed hazard rate.

For smooth estimates of the relative hazard rate the function *muhaz* (for usual survival models) is adapted to *my.muhaz*(). Hazard rate, cumulative hazard rate and survival function are replaced by relative analogons.

The needed information for relative hazard rate estimation can be obtained out of a relative regression model only with a time-dependent offset variable, hence the *ratetable*() object, as a risk factor using the function *rsmul*. The output gives the survival time in the appropriate way to use it for the relative hazard rate estimation. Hence, it builds time intervals for each patient with cut points at the end of every year and on every anniversary of the patients starting day of the study as it is assumed that the patient gets one year older on this day. Those intervals are given by a *start* and a *stop* vector. The censoring indicator is adjusted on this new time format. Additionally, in the output a factor *lambda* is dumped which contains the logarithmized population hazard for each time interval of a matched patient. Hence, the exponential of *lambda* gives the wanted individual population hazard. This also explains why time intervals are needed. As the matched population hazard changes when the used population mortality table information changes, hence every new year and every anniversary of the starting day of the study of a patient as it is assumed that the patient gets one year older on this day, time has to be split up to include all available information. With the information of *start*, *stop*, the censoring indicator and *lambda* the relative hazard rate can be estimated using kernel-based methods.

```
corerel<-"Surv(Death_SurvivalDays,Death)~"
core1rel<-"+ratetable(age=Age*365.24,sex=SexCode,year=OPDate)"
varirel<-"1"
formrel<-as.formula(paste(corerel,varirel,core1rel))
modelrel<-rsmul(formrel,ratetable=mortality,data=data,method="mul1")
arel<-data.frame(start=modelrel$y[,1],stop=modelrel$y[,2],cens=modelrel$y[,3],
  lambdap=exp(modelrel$lambda))
attach(arel)
```

As a numerical problem did not allow to adapt the optimal bandwidth part of the code in time, optimal bandwidths from the usual hazard rate estimations are used. The R-Code accesses a Fortran-Code using the function *.Fortran*() which contains most of the estimation procedure. As this Fortran-Code is very long and complex and therefore not easy to understand only the most important part is shown here. If an R-package is constructed later the full code will be presented there. Further only simple R-functions are used in *my.muhaz*().

```
my.muhaz <- function(Start, Stop, cens, lambdap, subset, max.time,
```

```
n.est.grid=101, bw.method="local", b.cor="left",kern="epanechnikov"){
method <- pmatch(bw.method, c("global", "local"))
if (is.na(method))
  stop("bw.method MUST be one of: 'global' or 'local'")
b.cor <- pmatch(b.cor, c("none", "left", "both"))
if (is.na(b.cor))
  stop("b.cor MUST be one of: 'none', 'left', or 'both'")
b.cor <- b.cor - 1
kern <- pmatch(kern, c("rectangle", "epanechnikov", "biquadratic",
  "triquadratic"))
if (is.na(kern))
  stop("kern MUST be one of: 'rectangle', 'epanechnikov', 'biquadratic',
    or 'triquadratic'")
kern <- kern - 1
if (missing(Start))
  stop("Parameter Start is missing")
if (missing(Stop))
  stop("Parameter Stop is missing")
nobsrel <- length(Start)
if (missing(cens))
  stop("Parameter cens is missing")
if (missing(lambdap))
  stop("Parameter lambdap is missing")
if ((length(Start)!=length(Stop))|(length(Stop)!=length(cens))|
  (length(cens)!=length(lambdap)))
  stop("Start, Stop, cens and lambdap MUST have the same length")
if (missing(subset)){
  subset <- rep(TRUE, nobsrel)}
else{
  if (is.logical(subset)){
    if (length(subset)!=nobsrel){
      stop("Start, Stop and subset MUST have the same length")}}
  else{
    stop("subset MUST contain ONLY logical values (T or F)")}}
Start <- Start[subset]
Stop <- Stop[subset]
cens <- cens[subset]
lambdap <- lambdap[subset]
```

```
ix <- order(Stop)
Stop <- Stop[ix]
Start <- Start[ix]
cens <- cens[ix]
lambdap <- lambdap[ix]
nobsrel <- length(Start)
min.time<-0
startz<-0
if (missing(max.time)){
  sfit <- survfit( Surv(Stop,cens) ~ 1 )
  endz <- approx(sfit$n.risk,sfit$time,xout=10)$y}
else{
  if (max.time>sort(Stop)[nobsrel]){
    warning("maximum time > maximum Survival Time")
    endz <- sort(Stop)[nobsrel]}
  else{
    endz <- max.time}}
if (startz>endz)
  stop("min.time MUST be < max.time")
zz <- seq(startz, endz, len=n.est.grid)
endl <- startz
endr <- endz
pin.commonrel <- list(Start=Start, Stop=Stop, cens=cens, lambdap=lambdap,
  nobsrel=nobsrel, min.time=startz, max.time=endz, n.est.grid=n.est.grid,
  method=method, b.cor=b.cor, kernel.type=kern)
globlb <-  103.8345
if (b.cor=="none"){
  bopt1 <-  c(921.3571, 946.5878, 961.3875, 970.72, 975.1696, 987.1706,
    1006.7249, 1021.3064, 1031.2227, 1035.5131, 1033.9420, 1028.5556,
    1018.7354, 1008.3878, 996.9470, 986.8583, 970.9826, 951.2432, 926.1377,
    901.3377, 880.4040, 858.6295, 841.9516, 827.6614, 821.1640, 820.9857,
    827.5678, 841.0215, 865.1290, 901.2831, 940.3310, 978.7671, 1016.0658,
    1053.2695, 1091.0683, 1130.1363, 1181.0030, 1241.5018, 1302.9253,
    1364.1990, 1421.5096, 1465.8620, 1494.0606, 1508.3388, 1508.1993,
    1496.7892, 1473.7016, 1452.2885, 1433.2107, 1410.3563, 1389.2931,
    1371.1061, 1352.3895, 1337.5558, 1325.3366, 1310.7854, 1293.4467,
    1269.0846, 1238.4245, 1201.4892, 1161.7462, 1120.3384, 1074.7389,
    1031.7316, 993.9291, 958.2457, 927.5505, 902.2724, 884.1975, 872.8971,
```

```
    869.7208, 869.5342, 861.6094, 847.1068, 823.2409, 793.2076, 755.4633,
    720.4299, 701.8562, 691.4267, 684.5246, 674.6764, 663.1611, 650.4745,
    642.4029, 649.2992, 669.0931, 691.7647, 712.5609, 732.0729, 748.0092,
    760.9437, 769.5876, 774.2550, 781.9211, 792.7502, 804.0325, 816.0951,
    830.7423, 848.6320, 871.8349)}
else{
  bopt1 <- c(222.4234, 214.7177, 210.0819, 207.8286, 208.7858, 213.1771,
    221.1854, 232.1833, 245.6193, 259.0146, 271.1009, 281.4459, 289.6169,
    297.9875, 306.9334, 318.0347, 329.7942, 343.2480, 363.3282, 389.5641,
    418.3821, 449.7215, 484.2686, 519.4815, 555.2097, 590.8753, 626.7471,
    663.7304, 702.2620, 742.9479, 786.3804, 831.0511, 876.8096, 925.7164,
    979.0034, 1035.2577, 1095.1273, 1156.4936, 1218.5144, 1281.5553,
    1342.0808, 1390.2977, 1423.2035, 1442.8491, 1448.9072, 1444.3693,
    1428.9677, 1415.9336, 1406.0283, 1393.0615, 1377.2882, 1360.3870,
    1343.1057, 1329.8319, 1319.3182, 1306.6013, 1291.2380, 1268.9847,
    1240.5697, 1206.0182, 1166.4040, 1123.1611, 1075.5870, 1030.4700,
    992.2124, 957.5151, 927.8747, 903.7201, 885.6420, 873.3554, 869.7208,
    869.5342, 861.6094, 847.1068, 823.2409, 793.2076, 755.4633, 720.4299,
    701.8562, 691.4267, 684.5246, 674.6764, 663.1611, 650.4745, 642.4029,
    649.5860, 672.0325, 699.7382, 727.5975, 755.9076, 782.6921, 809.0579,
    835.2656, 863.7771, 898.7201, 939.2661, 984.8250, 1035.8915, 1091.2982,
    1148.2846, 1200.4018)}
if (method<3){
  m1 <- method - 1
  dyn.load("C:/Users/Meins/Diplomarbeit/muhazrelative/Versuch12OhneOptBW/
    muhazrelative12.dll")
  ans <- .Fortran("newhad",as.integer(nobsrel),as.double(Start),as.double(Stop),
    as.integer(cens),as.double(lambdap),as.integer(kern),as.integer(m1),
    as.double(zz),as.integer(n.est.grid),as.double(globlb),as.double(bopt1),
    as.double(endl),as.double(endr),as.integer(b.cor),fzz = double(n.est.grid))
  if (method==1){
    ansrel <- list(pinrel=pin.commonrel,bw=globlb,est.grid=zz,haz.est=ans$fzz)}
  else{
    ansrel <- list(pinrel=pin.commonrel,bw=bopt1,est.grid=zz,haz.est=ans$fzz)}}
else{
  stop("method MUST be 1 or 2")}
class(ansrel) <- "muhazrelative"
ansrel}
```

The most important change from *muhaz*() is to adapt the hazard rate estimation itself. The relative hazard is achieved by implicating the population hazard calculated from the relative regression analysis for each time interval.

```
lambdap1 = ZERO
DO 200, j=1, n
  IF ((start(j).LE.stopp(i)).AND.(stopp(i).LT.stopp(j))) THEN
    lambdap1 = lambdap1 + lambdap(j)
  END IF
CONTINUE
```

As for usual hazard rates three different estimations are carried out. Estimations using global bandwidth and no boundary correction, locally adapted bandwidths and no boundary correction and locally adapted bandwidths and left boundary correction are implemented in *my.muhaz*().

```
testR1<-my.muhaz(start, stop, cens, lambdap, bw.method="global", b.cor="none",
  kern="epanechnikov")
testR2<-my.muhaz(start, stop, cens, lambdap, bw.method="local", b.cor="none",
  kern="epanechnikov")
testR3<-my.muhaz(start, stop, cens, lambdap, bw.method="local", b.cor="left",
  kern="epanechnikov")
```

The *plot.muhaz*() function can also be used here to demonstrate the relative results. For the comparison of the three relative hazard rate estimates again the functions *lines*() and *legend*() are used. Additionally *abline*() is applied to draw a horizontal line at 1 to make it easier to compare the estimates with the population hazard. Compare figure 11 on page 38, figure 12 on page 39, figure 13 on page 41 and the comparison in figure 14 on page 42.

```
plot.muhaz(testR1,ylim=c(0,30))
plot.muhaz(testR2,ylim=c(0,30))
plot.muhaz(testR3,ylim=c(0,30))

plot.muhaz(testR1,ylim=c(0,30),lty=1)
lines(testR2$est.grid,testR2$haz.est,lty=2)
lines(testR3$est.grid,testR3$haz.est,lty=3)
abline(h=1,lty=4)
legend(750,10,legend=c("global bandwidth","local bandwidth","local bandwidth
  and left boundary correction"),lty=c(1,2,3))
```