Chapter 1

Introduction

The present thesis is motivated by a study on injecting drug users in Badalona, most of whom became infected with the Human Immunodeficiency Virus (HIV) as a consequence of their drug addiction. Our aim has been to examine, whether or not there is an association between the elapsed time from first intravenous drug use until the infection with HIV and the subsequent time until the onset of Acquired Immunodeficiency Syndrome (AIDS), the AIDS incubation period. We have used a survival regression model for this purpose including the latter time as a response variable and the former as a covariate.

Given the fact that the time from first injecting drug use until HIV infection is interval-censored and the AIDS incubation period is doubly-censored, the proposed regression model deals with censoring in both the response variable and one of the regressors. So far, this particular situation has not been addressed in scientific publications and cannot be handled straightforwardly by statistical software packages. For this reason, a major objective of this PhD thesis has been to investigate which methods might be applied to estimate the unknown parameters of such a model.

In the following Section 1.1, we give a survey on parametric and nonparametric methods for the evaluation of interval-censored data, since parts of our methodology take advantage of these methods. Furthermore, we discuss the noninformativity conditions in Section 1.2, on which we base our methodology, and present some aspects from the area of operations research in the subsequent Section 1.3. This is of importance due to the fact that one of our proposed estimation procedures uses optimization tools.

1.1 Interval-censoring: State-of-the-art

Interval censored data both case 1 (current status data) and case 2 have been widely studied in the past. Usually, they arise in longitudinal studies where the time until an event of interest cannot be observed exactly, but is known to lie within an observed time interval.

An important field of medical research, where interval-censored data occur, are studies on HIV/AIDS, for example on the spread of the HIV epidemic or the pathogenesis of AIDS. Whereas the moment of the HIV infection is generally interval-censored, since in the majority of cases it cannot be determined exactly, the AIDS onset is often right-censored, as most studies finish before every individual of the cohort under study has developed AIDS. Hence, studies about the AIDS incubation period deal with doubly-censored survival data.

In this section, we present a general survey on the statistical methodology for interval-censored data starting with the more common case of interval-censored data case 2; current status data are addressed in the subsequent section. We also include methods applied to doubly-censored data which arise when both the origin and the endpoint of the time of interest are censored, may this be left-, right-, or interval-censoring.

In the following, we will emphasize the estimation of the distribution function of an interval-censored random variable, as well as regression models with these kind of data. All the methods we present in this section assume noninformative censoring. This implies that the construction of the corresponding likelihood functions does not need to account for the censoring generation process; see Section 1.2 for further details.

1.1.1 Interval censoring case 2

A random variable T is said to be interval-censored case 2, when only an interval, say $[L, R]$, is observed into which T falls. Case 1 is given if either $L = 0$ or $R = \infty$ and shall be considered in the following section. This definition using closed intervals allows for exact observation, that is when $L = R$, and has been used by Peto (1973) and Turnbull (1974, 1976) for the development of the nonparametric maximum likelihood estimator (NPMLE) of the distribution function of T, $F(t)$. In contrast with that, many authors consider semi-open intervals, $(L, R]$, or open intervals, especially if the probability of observing the event of interest is equal to zero. As an example, consider the moment of infection with a virus which cannot be observed exactly.

One might suppose that the unobserved value of T might be replaced by either the left-endpoint, the right-endpoint, or the midpoint of the observed interval. However, many studies have shown that this causes biased estimation result, especially if big censoring intervals are given. Bias would also be introduced, if the data analysis was based only on the exactly observed values of T. Moreover, the resulting loss in sample size by disregarding the interval-censored observations would decrease the efficiency of estimators and hypothesis testing (Leung, Elashoff, and Afifi 1997). For these reasons, it is obvious that interval-censored data require a particular methodology, part of which is presented in the sequel.

The Turnbull estimator for $F(t)$

As Hougaard (1999) points out, the nonparametric estimate of $F(t)$ can be preferable to a parametric one for various reasons. For example, a wrong parametric choice for the distribution of

T might lead to erroneous conclusions on $F(t)$. Furthermore, it might be difficult to find an appropriate parametric distribution to fit the data. Hougaard gives the example of a population's lifetimes whose hazard function show the so-called "bathtub shape": it decreases the first few years after birth, remains constant throughout many years, and starts increasing with about 20 to 25 years. In this case, the best fit would probably be obtained by a mixture of distributions.

In case of right-censored data, one can use the well known Kaplan-Meier estimator to obtain $\hat{F}(t)$ (Kaplan and Meier 1958). However, with interval-censored data this estimator cannot be applied, and it has been Peto (1973) and Turnbull (1974, 1976) who have developed the NPMLE for these data.

This estimator, the so-called Turnbull estimator, is based on a sample of observational intervals $[L_i, R_i]$, $i = 1, \ldots, n$, which contain the independent random variables $T_1, \ldots, T_n$. As mentioned before, an exact observation T_i is given if the interval's left- and right-endpoint coincide ($L_i = R_i$). Given this sample, the likelihood function to be maximized is the following:

$$L(F) = \prod_{i=1}^{n} \big(F(R_i+) - F(L_i-)\big). \tag{1.1}$$

In order to solve this maximization problem, Peto defines the two sets $\mathcal{L} = \{L_i,\, i = 1, \ldots, n\}$ and $\mathcal{R} = \{R_i,\, i = 1, \ldots, n\}$ containing all left and right interval endpoints, respectively. From these sets, the new intervals $[q_1, p_1], \ldots, [q_m, p_m]$ are derived such that $q_j \in \mathcal{L},\ \ p_j \in \mathcal{R}$, $j = 1, \ldots, m$, and that these intervals contain no other elements of $\mathcal{L}$ and $\mathcal{R}$ other than their left- and right-endpoints.

It can be proved that any function maximizing (1.1) IS constant between the intervals $[q_j, p_j]$ and undefined within them. Note that this implies that $\hat{\mathrm{P}}\big(T \in (p_{j-1}, q_j)\big) = 0$ for any j. As the distribution function is nondecreasing, any function, which was not constant between the intervals, would not maximize $L(F)$. Denoting the increase of F within the intervals $[q_j, p_j]$ by s_j, $j = 1, \ldots, m$, $L(F)$ has to be maximized as a function of $s_1, \ldots, s_m$ subject to $s_j \geq 0$ and $s_m = 1 - \sum_{j=1}^{m-1} s_j$. Peto tackles this maximization problem using the Newton-Raphson algorithm.

In contrast with Peto, Turnbull, in 1976, proposes the use of the self-consistency algorithm for the same maximization problem. A similar procedure he has used before for the estimation of the survival function with doubly-censored data (Turnbull 1974). The idea of self-consistency has first been presented by Efron (1967) and its application to the maximization of the likelihood function in (1.1) is illustrated in the following paragraph.

Let $\alpha_{ij} = \mathbb{1}_{\{[q_j, p_j] \in [L_i, R_i]\}}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$, be the indicator variables of whether or not $[q_j, p_j]$ lies within $[L_i, R_i]$. Then, the probability that T_i lies in the interval $[q_j, p_j]$ given a vector $s = (s_1, \ldots, s_m)'$ is given by

$$\mu_{ij}(s) = \frac{\alpha_{ij} s_j}{\sum_{k=1}^{m} \alpha_{ik} s_k}\,, \tag{1.2}$$

since $\hat{F}$ is constant outside the intervals $[q_j, p_j]$. The proportion of observations in the interval $[q_j, p_j]$ is equal to

$$\pi_j(s) = \frac{1}{n} \sum_{i=1}^{n} \mu_{ij}(s), \tag{1.3}$$

and a vector $s = (s_1, \ldots, s_m)'$ is said to be self-consistent, if

$$s_j = \pi_j(s),\ j = 1, \ldots, m.$$

Following this definition, the self-consistency algorithm of Turnbull for the calculation of the nonparametric maximum likelihood estimator $\hat{F}(t)$ works as follows:

1. Obtain initial estimates for s; for example, $s_j^{(0)} = \frac{1}{m}$, $j = 1, \ldots, m$.
2. For $i = 1, \ldots, n$, $j = 1, \ldots, m$, calculate $\mu_{ij}(s^{(0)})$ according to (1.2), and then $\pi_j(s^{(0)})$ following (1.3).
3. Obtain improved estimates for s by setting $s_j^{(1)} = \pi_j(s^{(0)})$.
4. Return to step 2 replacing $s^{(0)}$ by $s^{(1)}$ and continue until convergence is achieved.

Extensions of the Turnbull estimator

Turnbull shows that the self-consistent estimator is equivalent to the NPMLE under general conditions. Actually, he presents this estimator for the more general case when both interval-censoring and truncated data are given. Later, Frydman (1994) shows that, in general, the Turnbull estimator is not applicable to this case and presents the following correction.

Given the observed intervals $[L_i, R_i]$ as subsets of the truncation intervals $[V_i, U_i]$, $i = 1, \ldots, n$, the corrected likelihood function is the following:

$$L(F) = \prod_{i=1}^{n} \frac{F(R_i+) - F(L_i-)}{F(U_i-) - F(V_i+)}\,.$$

The self-consistency algorithm can be applied as above, but with the slight difference that the endpoints of the intervals $[q_j, p_j]$ are subsets of $\mathcal{L} \cup \{U_i,\ i = 1, \ldots, n\}$ and $\mathcal{R} \cup \{V_i,\ i = 1, \ldots, n\}$, respectively.

Concerning the properties of the NPMLE $\hat{F}$, its consistency is proved by Groeneboom (1991) and further properties, such as the asymptotic normality, are addressed in Yu, Schick, Li, and Wong (1998).

Gentleman and Geyer (1994) show, that it is possible that the NPMLE and the self-consistent estimator do not coincide. In order to assure that the Turnbull estimator furnishes the NPMLE,

it has to be checked that the so-called Kuhn-Tucker conditions are satisfied. In Section 1.3, we shall present these condition.

In a recent work, Ng (2002) addresses the effect on the NPMLE $\hat{F}(t)$ when semi-open intervals $(L_i, R_i]$ are given. This leads to a slight modification of likelihood function (1.1):

$$L(F) = \prod_{i=1}^{n} \big(F(R_i+) - F(L_i+)\big). \tag{1.4}$$

The Turnbull estimator can also be applied to the maximization of (1.4), but results may differ here. As a simple example, consider the two intervals $[a, b]$ and $[b, c]$. In the closed intervals case, we get $\hat{\mathrm{P}}(T = b) = 1$, whereas with semi-open intervals, $\hat{\mathrm{P}}(T \in (a, b]) = \hat{\mathrm{P}}(T \in (b, c]) = 0.5$. Dealing with a continuous random variable T, the latter results are normally more intuitive.

An alternative algorithm, the gradient projection algorithm, is proposed by Pan and Chappell (1998) in order to avoid the possible underestimation of the NPMLE at very early times for left-truncated and interval-censored data. However, unlike with the Turnbull estimator, nondecreasing hazards are required for this monotone NPMLE.

Regression models with an interval-censored response variable

Often one wants to evaluate the possible effect of other variables on the survival time T; for this purpose, mainly parametric and semi-parametric models are used.

Many research articles deal with interval-censored response data within Cox's proportional hazard model (Cox 1972), which models the hazard function of T in dependence of an unspecified underlying baseline hazard function, $\lambda_0(t)$, and a term including the covariate vector $\boldsymbol{Z}$:

$$\lambda(t; \boldsymbol{z}) = \lambda_0(t) \exp(\boldsymbol{\beta' z}). \tag{1.5}$$

To estimate the unknown parameter vector $\boldsymbol{\beta}$, Finkelstein (1986) proposes the maximization of likelihood function (1.1) after substituting $F(t)$ according to (1.5) which implies:

$$F(t; \boldsymbol{z}) = 1 - S(t; \boldsymbol{z}) = 1 - S_0(t)^{\exp(\boldsymbol{\beta' z})},$$

where $S_0(t)$ is the baseline survival function. To achieve maximization with respect to $\boldsymbol{\beta}$, Finkelstein uses the Newton-Raphson algorithm.

Two different approaches are presented by Pan (2000a) and Goetghebeur and Ryan (2000). The former, in a first step, uses multiple imputation to obtain estimated failure times for the (purely) interval-censored observation times. In the second step, he applies standard statistical procedures for right-censored data in order to estimate $\boldsymbol{\beta}$. Goetghebeur and Ryan, on the other hand, propose the use of an approximate likelihood and apply the EM algorithm to estimate the parameters. The M-step consists of fitting model (1.5) to the data to obtain estimates for $\boldsymbol{\beta}$ and

$\lambda_0(t)$, whereas the E-step involves the calculation of individuals at risk and the expected number of events at mass points identified through the use of the Turnbull estimator. Another method to determine $\hat{\boldsymbol{\beta}}$ and $\hat{\lambda}_0$, is the one using local likelihood methodology as proposed by Betensky, Lindsey, Ryan, and Wand (2002). Finally, the extension to the Cox model with multivariate interval-censored response data can be found in a work of Goggins and Finkelstein (2000).

Alternative models for survival data are parametric survival models, such as the accelerated failure time model, which is equivalent to the log linear survival model. These models require the specification of the underlying survival time distribution, but unlike the Cox model, they are not based on the proportional hazards assumption. Lindsey (1998) compares several parametric choices when using an approximation of the exact likelihood function, which is equivalent to the imputation of the interval's midpoint. Denoting the parametric version of the distribution function of T by $F(t;\boldsymbol{\theta})$ and its density by $f(t;\boldsymbol{\theta})$, the likelihood function (1.1) can be written as follows:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \big(F(R_i;\boldsymbol{\theta}) - F(L_i;\boldsymbol{\theta})\big) = \prod_{i=1}^{n} \int_{L_i}^{R_i} f(t;\boldsymbol{\theta})\, dt,$$

an approximation of which is given by

$$L_{appr}(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i;\boldsymbol{\theta})\Delta_i,$$

where t_i is the midpoint of $[L_i, R_i]$ and $\Delta_i = R_i - L_i$. According to Lindsey, who compares nine different parametric choices for $f(t;\boldsymbol{\theta})$, this approximation gives good results in the sense of little bias. As well, the conclusions from the models are remarkably robust with different distributions for T. However, as Lindsey and Ryan (1998) point out, midpoint imputation as well as left- or right-endpoint imputation will tend to underestimate the standard errors of the estimated parameters. This might lead to invalid inference.

1.1.2 Interval censoring case 1: current status data

Interval-censored data case 1 or current status data occur, when a time of interest, T, cannot be observed exactly and only one observation time, U, is available together with the information, whether or not the event of interest has occurred before U. Although similar to left- and right-censored data, there is one substantial difference: in case of left- and right-censored data, the probability for observing the event of interest is positive, whereas for current status data this probability is equal to zero.

We denote the observations by (U_i, δ_i), $i = 1, \ldots, n$, where $\delta_i = \mathbb{1}_{\{T_i \leq U_i\}}$ and obtain the following likelihood function:

$$L(F) = \prod_{i=1}^{n} F(U_i)^{\delta_i} \big(1 - F(U_i)\big)^{1-\delta_i}. \tag{1.6}$$

The Turnbull estimator could be applied to maximize (1.6), however, as Groeneboom (1991) points out, the Turnbull estimator might not coincide with the NPMLE for current status data. He recommends the application of the greatest convex minorant algorithm to determine $\hat{F}$:

1. Order the observation times: $u_{(1)} < u_{(2)} < \cdots < u_{(n)}$, and define $\delta_{(i)} = \delta_j$ if $u_{(i)} = u_j$.
2. Plot $\big(i, \sum_{j=1}^{i} \delta_{(j)}\big)$, $i = 1, \ldots, n$.
3. Form the greatest convex minorant C^* of the points in step 2 on the interval $[0, n]$.
4. $\hat{F}(u_{(i)})$ is the left-derivative of C^* at i, $i = 1, \ldots, n$.

Herein, C^* is defined as follows:

$$\begin{aligned} &C^*\colon [0, n] \longmapsto \mathbb{R}, \\ &C^* = \sup \Big\{ C(t)\colon\ C\colon [0, n] \mapsto \mathbb{R} \text{ convex}, \\ &\qquad\qquad C(i) \leq \sum_{j=1}^{i} \delta_{(j)},\ i = 1, \ldots, n \Big\}. \end{aligned}$$

Groeneboom shows that the left-derivative of the greatest convex minorant can be obtained applying the max-min formula:

$$\hat{F}(u_{(i)}) = \max_{j \leq i} \min_{k \geq i} \frac{\sum_{l=j}^{k} \delta_{(l)}}{k - j + 1}\,.$$

The NPMLE for F from current status data is asymptotically normal at nonstandard $n^{1/3}$-rate and consistent under general conditions (Huang and Wellner 1995; van der Laan 1998). However, according to Pan and Chappell (1999), $\hat{F}$ is inconsistent when both left-truncated and case 1 interval-censored data are present. For this case, they provide a conditional NPMLE given left-truncation and show that it is consistent under reasonable conditions. The corresponding likelihood function is the following, where V_i, $i = 1, \ldots, n$, are the times of left-truncation:

$$L(F) = \prod_{i=1}^{n} \frac{\big(F(U_i) - F(V_i)\big)^{\delta_i} \big(1 - F(U_i)\big)^{1-\delta_i}}{1 - F(V_i)}.$$

Regarding regression models, Huang and Wellner (1998) discuss the properties of the maximum likelihood estimator for the parameters of the Cox model when the response variable is a current

status variable. They show the consistency and asymptotic normality at $n^{1/3}$-rate under general conditions. Other authors deal with the linear regression model for current status data (Murphy, van der Vaart, and Wellner 1999), generalized additive models (Shiboski 1998), and, as a particular case of the latter, with an additive hazards regression model (Lin, Oakes, and Ying 1998).

1.1.3 Doubly-censored data

In the two previous sections, only the case of an interval-censored endpoint is considered. However, it is possible that also the origin of the time of interest is not observed exactly, either due to left- or interval-censoring. An example treated often in scientific publications and already mentioned above, is the AIDS incubation or latency period. Whereas the moment of HIV infection is interval-censored in the majority of cases, the moment of AIDS onset is possibly right-censored in many studies, namely in cases when a study ends and individuals remain AIDS-free.

Given this situation, a possible approach would be to transform the data to simply interval-censored data. Let T denote the AIDS incubation period, $[L, R]$ be the interval which encloses the moment of HIV infection and Z the observed AIDS onset. Then, if the AIDS onset is exactly observed, we have $T \in [Z - R, Z - L]$, and if Z is right-censored, $T > Z - R$. This approach seems to be intuitive, however, as several authors point out, it is not valid, since the independence between T and $[L, R]$ would be lost by this transformation (De Gruttola and Lagakos 1989; Lam 1997; Geskus 2001).

Instead, De Gruttola and Lagakos propose a nonparametric method to estimate the distribution function $F(t)$. It is based on a generalization of Turnbull's self-consistent algorithm and estimates simultaneously both the distribution function of the chronological time of infection with HIV and the subsequent AIDS latency period. The authors mention that the problem with possible multiple maxima of the resulting likelihood function tends to be less common as the sample size increases. The same data set is used by Joly and Commenges (1999) in the framework of a multi-state model with censored and truncated data.

A situation similar to the one of De Gruttola and Lagakos is addressed by Gómez and Lagakos (1994) who also apply the Turnbull estimator. The difference to the former is that they proceed in two steps: first, the infection times' distribution is estimated, secondly, $F(t)$. With this approach they overcome some difficulties of the former estimation procedure such as time-consuming calculations and nonidentifiability problems in case of smaller data sets. A generalization of this method to the case of continuous distributions of HIV infection and the AIDS incubation period is presented by Gómez and Calle (1999).

In another instance, Geskus (2001) compares several approaches to estimate the AIDS incubation period's distribution from doubly-censored data. These include several imputation methods for the interval-censored HIV infection. Once a value for HIV infection is imputed, standard procedures for right-censored data are applied to estimate $F(t)$. In particular, conditional mean

imputation leads to satisfactory results in terms of little bias as shown by simulations. Denoting by X the moment of HIV infection, the conditional mean is defined by $\mathrm{E}_{\hat{F}_X}(X|L,R)$, where the distribution function of X, F_X, is estimated by means of the Turnbull estimator.

The proportional hazards model for doubly-censored data is addressed by Sun, Liao, and Pagano (1999), who use a simple estimation equation approach to estimate the unknown parameters. Finally, the particular case of doubly-censored current status data is dealt with in van der Laan, Bickel, and Jewell (1997) and van der Laan, Jewell, and Peterson (1997).

1.1.4 Interval-censored covariates in regression models

In contrast with regression models for interval-censored response data, only a few studies on the use of interval-censored covariates in regression models have been published. One example is the work of Goggins, Finkelstein, and Zalavsky (1999). It deals with the Cox model with a binary time-varying covariate, where the change point from one state to the other is only known to lie within a given interval.

In a different instance, Gómez, Espinal, and Lagakos (2003) present a simple linear model with a completely observed response variable and a discrete interval-censored covariate:

$$Y = \alpha + \beta Z + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0,\sigma^2)$ and independent of Z. The support of Z is given by $S = \{s_1,\ldots,s_m\}$ with corresponding probabilities $\mathrm{P}(Z = s_j) = \omega_j$, $j = 1,\ldots,m$. The objective consists in estimating the unknown parameter vector $\boldsymbol{\theta} = (\alpha,\beta,\sigma)'$ and $\boldsymbol{\omega} = (\omega_1,\ldots,\omega_m)'$ given the independent observations (y_i, z_{l_i}, z_{r_i}), $i = 1,\ldots,n$, where $\mathrm{P}(z_i \in [z_{l_i}, z_{r_i}]) = 1$.

Defining the indicator variables $\gamma_{ij} = \mathbb{1}_{\{s_j \in [z_{l_i}, z_{r_i}]\}}$, the likelihood function is proportional to:

$$\begin{aligned} L(\boldsymbol{\theta},\boldsymbol{\omega}) &= \prod_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij} f(y_i|s_j;\boldsymbol{\theta})\omega_j \\ &= \prod_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}\frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{(y_i-\alpha-\beta s_j)^2}{2\sigma^2}\Big)\omega_j. \end{aligned} \tag{1.7}$$

Gómez et al. tackle the maximization of likelihood function (1.7) by means of an iterative estimation procedure. In the first step, they use the self-consistent equations of Section 1.1.1 to maximize $L(\boldsymbol{\theta},\boldsymbol{\omega})$ with respect to $\boldsymbol{\omega}$ holding $\boldsymbol{\theta}$ fixed. Here, expression (1.2) has the following form:

$$\mu_{ij}(\boldsymbol{\theta},\boldsymbol{\omega}) = \mathrm{P}(Z_i = s_j|Z_{l_i}, Z_{r_i}) = \frac{\alpha_{ij} f(y_i|s_j;\boldsymbol{\theta})\omega_j}{\sum_{k=1}^{m}\alpha_{ik} f(y_i|s_k;\boldsymbol{\theta})\omega_k},$$

and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_m)'$ is self-consistent if

$$\omega_j = \frac{1}{n}\sum_{i=1}^{n} \mu_{ij}(\boldsymbol{\theta}, \boldsymbol{\omega}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\alpha_{ij} f(y_i|s_j;\boldsymbol{\theta})\omega_j}{\sum_{k=1}^{m} \alpha_{ik} f(y_i|s_k;\boldsymbol{\theta})\omega_k},\ j = 1, \ldots, m.$$

In the second step, $L(\boldsymbol{\theta}, \boldsymbol{\omega})$ is maximized with respect to $\boldsymbol{\theta}$ given $\hat{\boldsymbol{\omega}}$ from the previous step in order to determine $\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\omega}}$. In the resulting expressions for $\hat{\alpha}, \hat{\beta}$, and $\hat{\sigma}$, the unobserved value of Z is replaced by its expected mean given $[Z_l, Z_r]$ and based on $\hat{\boldsymbol{\omega}}$: $\mathrm{E}_{\hat{\boldsymbol{\omega}}}(Z|Z_l, Z_r)$. Both steps are alternated until simultaneous convergence is achieved.

1.2 Noninformativity conditions

The vast majority of methods for the evaluation of interval-censored data are based on the assumption that the censoring mechanism is noninformative. This means that the observed censoring interval $[L, R]$ carries no further information on the survival time T; rather, this time lies in the interval $[L, R]$. This assumption permits evaluating censored data without modeling the censoring generation process.

Turnbull uses the likelihood function (1.1), to which each individual contributes the probability $\mathrm{P}(T \in [L, R]) = F(R+) - F(L)$. In case censoring could not be assumed noninformative, the exact contribution to the likelihood function of each individual would be equal to the following expression (Oller, Gómez, and Calle 2004):

$$\int_L^R f_{T,L,R}(t, L, R)\, dt = \mathrm{P}(T \in [L, R], L \in dL, R \in dR), \tag{1.8}$$

where $f_{T,L,R}$ denotes the joint density of T, L, R, and dL and dR infinitesimal intervals around L and R, respectively. Often, the censoring mechanism is unknown. Hence, it is necessary to make assumptions on the censoring generation process to justify the use of $\mathrm{P}(T \in [L, R])$ instead of expression (1.8) which allows the use of standard statistical methods for interval-censored data. Otherwise, the distribution function of T would be nonidentifiable, that is more than one distribution function of T would be compatible with the data (Leung et al. 1997).

Following the notation in Oller et al., the noninformativity condition is defined by any of the following three conditions. The authors prove their equivalence.

$$f_{T|L,R}(t|l, r) = \frac{f_T(t)}{\mathrm{P}(T \in [l, r])} \mathbb{1}_{\{t:t\in[l,r]\}}(t), \tag{1.9a}$$

$$f_{L,R|T}(l, r|t) = \frac{f_{L,R}(l, r)}{\mathrm{P}(T \in [l, r])} \mathbb{1}_{\{(l,r):t\in[l,r]\}}(l, r), \tag{1.9b}$$

$$f_{L,R|T}(l, r|t) = f_{L,R|T}(l, r|t') \quad \text{for any } t, t' \in [L, R]. \tag{1.9c}$$

Expression (1.9a) implies that the censoring in $[L, R]$ provides the same information as T lying in any interval $[L, R]$. According to definition (1.9b), the observed interval $[L, R]$ is not influenced by the specific unobservable value of T. Finally, expression (1.9c) means that two specific values of T consistent with interval $[L, R]$ provide the same information.

According to Heitjan (1993) and given the example of clinical studies, one noninformative censoring mechanism for interval-censored data are previously fixed monitoring times. The same is true if the examination times are independent of the patient's disease history or if a doctor decides a patient's next visit on the basis of his current observed status.

The exact mathematical expressions of the noninformativity and identifiability conditions for current status data can be found in Betensky (2000).

1.3 Optimization problems in statistics

Statistics and operations research are often seen as two scientific areas with little in common. Statisticians deal with probabilities, estimation, or hypothesis testing, whereas operations research with optimization problems and their exact solutions. However, as Athanari and Dodge (1980) state:

> "... more often than not one finds that the statistician is trying to solve an optimization problem and its related questions, such as existence or uniqueness of the solution."

For example, as mentioned above, Gentleman and Geyer (1994) show that the Turnbull estimator not necessarily furnishes a unique solution to the maximization of the likelihood function (1.1). In order to assure this, the Kuhn-Tucker conditions have to be verified (see below). Barnett (1966) describes another problem when maximizing the likelihood function: it might have several roots, that is local maxima, even though regularity conditions might hold and a unique consistent root is proved to exist.

Given the importance of optimization theory in statistics, in the following, we give a short survey on constrained nonlinear optimization problems including some common algorithms and the Kuhn-Tucker conditions. We close the section presenting some tools to solve these problems in practice.

1.3.1 Nonlinear constrained optimization problems

There are plenty of different optimization problems treated in operations research, of which nonlinear constrained optimization problems might be the most frequent ones in statistics. Just think of the maximization of (mostly nonlinear) likelihood functions given some constraints on the parameters to be estimated. These and other optimization problems are presented, for example, in detail in Nocedal and Wright (1999), or in a more concise way in Moré and Wright (1993).

General form of constrained optimization problems

The general constrained optimization problem consists of minimizing a nonlinear function subject to constraints on the variables which can be linear or nonlinear. Denoting by $f(x)$ the objective function in dependence of the real-valued variables x, the constrained minimization problem can be described by the following expression:

$$\min_{x\in\mathbb{R}^n}\{f(x) : c_i(x) \leq 0, i \in \mathcal{I},\ c_i(x) = 0, i \in \mathcal{E}\}, \tag{1.10}$$

where c_i are mappings from $\mathbb{R}^n$ to $\mathbb{R}$, and $\mathcal{I}$ and $\mathcal{E}$ denote the index sets of the inequality and equality constraints, respectively. Note that maximizing $f(x)$ is equal to minimizing $-f(x)$.

Optimization algorithms

Many algorithms have been developed for the minimization problem (1.10). Herein, we shall briefly present the basic idea of three of them: the Newton method, augmented Lagrangian methods, and sequential quadratic programming. They follow the same strategy in the sense that they generate a sequence of values $\{x_k\}$ which converge to the optimal solution x^*, but differ in the way of determining the search direction and the step length to obtain x_{k+1} given x_k. Actually, the Newton method is used, when no constraints on x are given.

Newton method The idea of the Newton method is to determine the value x_{k+1} as the minimizer of a quadratic model of $f(x)$ around x_k:

$$x_{k+1} = x_k + p_k = x_k - \left(\nabla^2 f(x_k)\right)^{-1}\nabla f(x_k),$$

where $\nabla f(x)$ and $\nabla^2 f(x)$ are the Jacobian vector and the Hessian matrix of f, respectively. If the Hessian matrix is substituted by an approximation to it, the method is called the quasi-Newton method.

Augmented Lagrangian methods These algorithms are based on the successive minimization of the so-called augmented Lagrangian function $\mathcal{L}_A$. Following the notation in (1.10), the iteration value x_{k+1} is the solution to the following minimization problem:

$$x_{k+1} = \arg\min_x\{\mathcal{L}_A(x, \lambda_k; \nu_k) : c_i(x) \leq 0, i \in \mathcal{I}\},$$

where

$$\mathcal{L}_A(x, \lambda; \nu) = f(x) + \sum_{i\in\mathcal{E}} \lambda_i c_i(x) + \frac{1}{2}\sum_{i\in\mathcal{E}} \nu_i c_i^2(x).$$

That is, the Lagrangian function $\mathcal{L}(x, \lambda)$ is augmented by the term $\frac{1}{2}\sum_{i\in\mathcal{E}} \nu_i c_i^2(x)$, which penalizes infeasible solutions for x. After each iteration, the Lagrangian multiplier λ and, possibly, ν are updated.

Sequential quadratic programming This algorithm is a generalization of the Newton method given restrictions on the variables. The algorithm replaces the objective function $f(x)$ by the quadratic approximation

$$q_k(d) = \nabla f(x_k)'d + \tfrac{1}{2}\, d'\nabla^2_{xx}\mathcal{L}(x_k, \lambda_k)d, \tag{1.11}$$

and the constraints by linear approximations. Thus, the iteration value x_{k+1} is obtained by the following formula:

$$\begin{aligned} x_{k+1} &= x_k + d_k \\ &= x_k + \arg\min_d \big\{ q_k(d) : c_i(x_k) + \nabla c_i(x_k)'d \le 0, i \in \mathcal{I}, \\ &\qquad\qquad c_i(x_k) + \nabla c_i(x_k)'d = 0, i \in \mathcal{E} \big\}. \end{aligned}$$

An advantage of this method is that expression (1.11) can be minimized easily with specialized quadratic programming techniques.

Kuhn-Tucker conditions

The Kuhn-Tucker conditions are the usual stopping criterium of the algorithms for constrained optimization problems since they are necessary conditions for x^* to be a minimizer of (1.10). They assure that the derivatives of all possible differentiable curves in x^* are nonnegative.

Technically, if x^* is a minimizer of (1.10), then, there must exist Lagrange multipliers λ and μ, such that the following conditions hold:

$$\begin{aligned} &i) \quad \nabla f(x^*) + \sum_{i\in\mathcal{E}} \lambda_i c_i(x) + \sum_{i\in\mathcal{I}} \mu_i c_i(x) = 0, \\ &ii) \qquad\qquad \sum_{i\in\mathcal{I}} \mu_i c_i(x) = 0, \\ &iii) \qquad\qquad \mu \ge 0. \end{aligned}$$

1.3.2 Optimization tools

Herein, we briefly present some optimization tools which have been of importance for the elaboration of the present PhD thesis: the programming language AMPL, the NEOS server, and SNOPT, a solver available under NEOS.

AMPL: A mathematical programming language

AMPL is a modeling language for mathematical programming which has been designed to solve different kinds of optimization problems; see Fourer, Gay, and Kernighan (2003) for its use and detailed examples. One of AMPL's important features is that optimization problems written in the AMPL code can be sent to the web site 'NEOS: Server for optimization' in order to let them be solved there. This is a big advantage especially if the available solvers of AMPL's local version do not achieve optimization of a given problem or would need a long computing time (Ferris, Mesnier, and Moré 2000).

NEOS: Server for optimization

The NEOS server has been created as a help to optimization problems which one might not be able to solve on his or her own computer. It can be accessed and used free of charge in different manners (Dolan, Fourer, Moré, and Munson 2002). For example, one can directly access its web page[1], choose an appropriate solver and enter the corresponding files written in AMPL code or other programming languages. Several optimization problems can be sent to NEOS and their solution can be received by email.

Another way to use the NEOS solvers is the indirect use by means of the Kestrel interface. This interface, available for download on the web pages of AMPL[2], enables the remote solution of optimization problems within the AMPL modeling language (Dolan and Munson 2001). Its main advantage is that one does not need to leave AMPL's local version since Kestrel invokes the specified NEOS solver. Kestrel also permits the simultaneous submission of several optimization problems.

SNOPT: Solver for nonlinear optimization problems

The solver SNOPT, available under the NEOS server, is a suitable solver for large nonlinearly constrained optimization problems with a modest number of degrees of freedom. It is most efficient if only some of the variables enter nonlinearly in the objective function, or if the number of active constraints is nearly as large as the number of variables. The solver uses a sequential quadratic programming method, within which the Hessian of the Lagrangian function is approximated by means of quasi-Newton methods; see Gill, Murray, and Saunders (1999) for details.

According to the NEOS Guide[3], SNOPT generally requires less evaluations of the nonlinear function than other solvers such as MINOS (Murtagh and Saunders 1978; Murtagh and Saunders 1982) and furnishes convergence for a wide class of optimization problems. Therefore, SNOPT is recommended, for example, in circumstances when the nonlinear functions or their gradients are very costly to evaluate.

[1] URL: `http://www-neos.mcs.anl.gov` [last visited: June 2004]
[2] URL: `http://www.ampl.com/DOWNLOADS/index.html` [June 2004]
[3] URL: `http://www-fp.mcs.anl.gov/otc/Guide` [June 2004]

1.4 Outline of the following chapters

In the following chapter, we present the three data sets analyzed for the elaboration of this dissertation. Two correspond to studies on HIV/AIDS: one is on the survival of Tuberculosis patients co-infected with HIV in Barcelona, the other on injecting drug users from Badalona and surroundings, most of whom became infected with HIV as a result of their drug addiction. The complex censoring patterns in the variables of interest of the latter study have motivated the development of estimation procedures for regression models with interval-censored covariates. The third data set comes from a study on the shelf life of yogurt, a completely different area to the two previous epidemiological studies. We present a new approach to estimate the shelf lives of food products taking advantage of the existing methodology for interval-censored data.

Chapter 3 deals with the theoretical background of an accelerated failure time model with an interval-censored covariate, putting emphasize on the development of the likelihood functions under different censoring patterns of the response variable, as well as on the estimation procedure by means of optimization techniques and tools. Their use in statistics can be an attractive alternative to established methods such as the EM algorithm. In Chapter 4 we present further regression models such as linear and logistic regression with the same type of covariate, for the parameter estimation of which the same techniques are applied as in Chapter 3. All programmes written with the AMPL code are attached in Section C.1 in the appendix.

Other possible estimation procedures are described in Chapter 5. These comprise mainly imputation methods, which consist of two steps: first, the observed intervals of the covariate are replaced by an imputed value, for example, the interval midpoint, then, standard procedures are applied to estimate the parameters. Other possible methods are sketched within the chapter, too.

The application of the proposed estimation procedure for the accelerated failure time model with a interval-censored covariate to the data set on injecting drug users in Badalona is addressed in Chapter 6. Different distributions and covariates are considered and the corresponding results are presented and discussed. The chapter closes with a tentative goodness-of-fit plot based on an adaptation of the Cox-Snell residuals.

To compare the estimation procedure with the imputation based methods of Chapter 5, a simulation study is carried out, whose design and results are the contents of Chapter 7. Finally, in the closing Chapter 8, the main results are summarized and several aspects which remain unsolved or might be approximated in another way are addressed.