Chapter 8

Discussion and Future Research

8.1 Conclusions

8.1.1 Methodology

The regression models with an interval-censored covariate presented in this PhD thesis are of importance in any scientific area, in which interval-censored data arise. An example are epidemiological or clinical studies on HIV/AIDS, where patients are often under treatment and observation for a long time. Events of interest, such as the time of HIV infection or the moment a medical marker of the HIV infection passes a certain threshold, cannot be determined exactly, but are known to lie within observed intervals. Nonetheless, these models have received little attention, so far.

We have developed an estimation procedure for such regression models. Its novel approach lies in the use of optimization techniques and tools for statistical procedures. The combined use of the mathematical programming language AMPL, the NEOS solvers, and the Kestrel interface as described in Section 3.5 is an attractive alternative to the use of the EM algorithm developed by Ayer, Brunk, Ewing, Reid, and Silverman (1954). It allows the maximization of very cumbersome (log) likelihood functions with a large number of variables and restrictions. Its disadvantage, that only the maximum likelihood estimators are obtained, but not their variances, is shared by other algorithms which are not covered by standard statistical software. The need of further software for the computation of the parameters' confidence intervals is compensated by the rapid maximization of the objective function. With the data set from the hospital Can Ruti, the maximum likelihood estimates for more than 200 variables, have been obtained in less than 25 seconds on a Pentium III computer with 871 MHz. Besides, AMPL's programming code is intuitive and, hence, easy to learn.

Another important advantage of the proposed methodology is the simultaneous estimation of the covariate's distribution function, which makes its estimation apart unnecessary. In the simulation study, its implementation in the respective AMPL programmes has shown a somewhat

smaller bias than the Turnbull estimate, which has been obtained by means of the S-Plus function `qkaplanMeier`.

Apart from these models, in Section 2.3, we have described the use of statistical methods for interval-censored data for the analysis of the shelf life of food products. This approach offers new and flexible tools to examine shelf lives and their predictors.

8.1.2 Further aspects of interest

In the sequel, we briefly address several aspects related to the applied methodology which have not been mentioned before.

The response variable Y of the accelerated failure time model in Chapter 3 is either right- and left-censored or doubly censored, whereas the covariate Z is interval-censored including left- and right-censored data as particular cases. The inclusion of interval-censored observations of Y, on one hand, and exact observations of Z, on the other, could easily be accomplished. In the former case, the likelihood contribution of an individual with observed interval $[Y_l, Y_r]$ would be equal to $S(Y_l|Z) - S(Y_r|Z)$ accounting for all possible values of $Z \in [Z_l, Z_r]$. Concerning Z, if an exact observation of Z was given, the integrals over the corresponding conditional density or survival functions given Z would reduce to either $f(Y|Z), S(Y|Z), 1 - S(Y|Z)$, or $S(Y_l|Z) - S(Y_r|Z)$ depending on censoring in Y. However, data are often either completely interval-censored or partly right- and left-censored. Generally, interval-censoring occurs when an event of interest cannot be observed exactly, whereas the left- and right-censoring arise, when a study finishes before the event of interest occurs.

For the data analysis in Chapter 6, we have used the solver SNOPT, one of the available solvers under the NEOS server. As described briefly in Section 1.3.2, this solver is recommended, for example, in circumstances when the nonlinear functions or their gradients are very costly to evaluate. This has been the case with the log likelihood functions (6.3) and (6.5), each with more than 200 parameters, which could be maximized in less than 30 seconds. Given the practical experience of the simulation study, we believe that SNOPT is also an appropriate solver for the maximization of the likelihoods corresponding to the models in Chapter 4.

We have applied other solvers for nonlinearly constraint optimization problems to the data set on injecting drug users, which are available at the NEOS server. None of them has been as efficient as SNOPT: the solvers LOQO (Vanderbrei 2000) and LANCELOT (Conn, Gould, and Toint 1992) have obtained the same results as SNOPT, but the computing time has been longer, seven minutes and nearly four hours, respectively. The solvers MINOS (Murtagh and Saunders 1978) and PENNON (Kočvara and Stingl 2003) have not been able to solve the maximization problem.

In every analysis, we have assumed noninformative censoring following the definition in Oller, Gómez, and Calle (2004). This avoids modeling the censoring generation process, but might not always be justified. Concerning the evaluated data set from Badalona, informative censoring

cannot not be ruled out, but should be less present with the data after 1985. Since then, HIV tests have been available for all injecting drug users.

Finally, in order to avoid even more than the 200 parameters, we have not taken into account the calendar time of HIV infection. Different studies show that the expansion of the HIV epidemic has varied along the years (Gómez and Lagakos 1994; Joly and Commenges 1999; Geskus 2001) as it is reported also by the Joint United Nations Programme on HIV/AIDS (UNAIDS 2002). This fact implies different distribution functions of HIV infection in dependence of calendar time. We could account for that by stratifying Z according to the year of first intra-venous drug use. Denoting by a the index of calendar year, we would have covariates Z_a with discrete support $S_a = \{s_{a_1}, \ldots, s_{a_m}\}$ and corresponding probabilities summarized by $\boldsymbol{\omega}_a = (\omega_{a_1}, \ldots, \omega_{a_m})'$. For the accelerated failure time model, we might consider two possibilities: the parameter vector $\boldsymbol{\theta}$ is same for any covariate Z_a or varies with calendar time. This might be an interesting aspect of future research

Other possible methods to evaluate these data comprise the joint study of longitudinal data, such as CD4 count and viral load, or a multi-state models approach considering the infection with HIV, the presence of AIDS, and death as three different states. For a concise summary on multi-state models, see, for example, Commenges (1999).

8.1.3 Epidemiological results

The main epidemiological result, according to the model adjustments in Chapter 6, is the positive association between time from first intra-venous drug use until HIV infection and the subsequent AIDS incubation period. That implies that the longer an injecting drug user remains seronegative, the longer s/he remains AIDS-free once s/he is infected with HIV. One could think of several possible explanations for this finding: for example, a genetically based strong immune system resisting both HIV infection and AIDS onset, or the hygienic precautions taken by the individuals.

Nonetheless, apart from the statistical considerations above, there are other epidemiological factors which have to be taken into account in necessary further studies before final conclusions can be drawn. For example, other covariates such as type of drugs or treatment medication should be reported and considered in an analysis, as well as the possible exposure to other risk factors for HIV infection. For example, female injecting drug users have partly been working as prostitutes.

Longer AIDS incubation times of female injecting drug users, as indicated by model (6.6) on page 90, have also been reported by Pérez-Hoyos, del Amo, Muga, del Romero, García de Olalla, Guerrero, Hernàndez-Aguado, and GEMES (2003) in the context of a multicentric study on the effectiveness of highly active anti-retroviral treatment (HAART) in Spain. The same study shows the positive effects of HAART on the survival of HIV infected persons. These treatments are available since 1997, before, the therapies administered to HIV-infected injecting drug users were a monotherapy with AZT (until 1992) and a dual combination therapy, respectively. It

is, precisely, these new therapies prolonging the AIDS incubation periods, which nowadays most probably reduce the effect of the duration of time from first intra-venous drug use on the AIDS incubation period.

8.2 Future research

A very important aspect of future research is the study of adequate residuals to check the goodness-of-fit for the accelerated failure time model with an interval-censored covariate. This aspect has not been resolved satisfactorily by now. Our adaptation of the Cox-Snell residuals in Section 6.5 is a first tentative approach, but lacks theoretical rigorousness concerning the distribution of the adapted residuals. Following the work of Topp and Gómez (2004) in the framework of a linear regression model with an interval-censored covariate, we could determine the lower and upper bound of the Cox-Snell residuals, substituting Z by $[Z_l, Z_r]$ in expression (6.7). The resulting residual would be equal to the estimated mean given the residual's lower and the upper bound and based on the model error distribution.

Another point of interest is the possible improvement of simultaneous maximization. Although, in general, the simulation study in Chapter 7 has shown superiority of this method over the imputation techniques employed, midpoint estimation has partly shown less biased results. Our conjecture is, that his has to do with the large number of parameters in case of simultaneous maximization. Midpoint imputation, on the other hand, deals only with the model parameters. It might be interesting to see, whether a reduction of the number of parameters would improve the estimation of the model parameters with simultaneous maximization; this can be achieved by grouping the covariate's observations into categories, even though the estimation of the covariate's distribution function would become more imprecise.

An alternative to the discrete assumption, is the parametric approach as sketched in Section 5.2. Programming might become more cumbersome since integrals must be approximated by a sum over rectangles, but the number of parameters is much lower than with a discrete covariate. This could improve their estimates' preciseness, whenever the parametric assumption is justified.

Like other solvers, SNOPT cannot completely guarantee the localization of the global maximum of the objective function. With the evaluated data set, we believe that the obtained estimates are, indeed, the wanted maximum likelihood estimates, since different starting values have all resulted in the same maximum of the log likelihood function. This observation might confirm the thesis of Pewsey (2000), who states that the problem of encountering local rather than the global maximum is more frequent with small sample sizes of less than 50 observations. His study is on the parameter estimation of the skew-normal distribution.

Gan and Jiang (1999) present an approach to this problem in the context of maximum likelihood estimation. They provide the necessary and sufficient conditions for consistency and

asymptotic optimality of a likelihood's maximum, and supply a test for global maximization. However, we have to have in mind that[1]:

> Global optimization algorithms try to find an x^* that minimizes f over all possible vectors x. This is a much harder problem to solve. We do not discuss it here because, at present, no efficient algorithm is known for performing this task. For many applications, local minima are good enough, particularly when the user can draw on his/her own experience and provide a good starting point for the algorithm.

Although the detection of the global maximum is an important topic for statistics, too, it is rather the objective of experts in optimization theory to find a solution to this problem in the future.

[1] `http://www-fp.mcs.anl.gov/otc/Guide/OptWeb/continuous/unconstrained/` [June 2004]