Chapter 5

Alternative Estimation Procedures

In this chapter, we present alternative procedures to estimate the unknown parameters of a regression model with an interval-censored covariate. These methods have previously been applied to other models and/or data settings and are now adapted to the particular case of the accelerated failure time model (3.5) of Chapter 3:

$$\ln(Y) = \mu + \beta Z + \sigma W,$$

where the discrete covariate Z is interval-censored in $[Z_l, Z_r]$ and the response variable Y may be either right- and left-censored or doubly censored. The unknown parameter vector is denoted by $\boldsymbol{\theta} = (\mu, \beta, \sigma)'$.

Section 5.1 deals with methods based on imputation techniques. The basic idea of these methods consists in replacing the observed interval $[Z_l, Z_r]$ by an imputed value and the posterior use of standard regression methods. For the implementation of these estimation procedures, we have used the statistical software S-Plus taking advantage of its implemented functions for the evaluation of censored data. Further procedures which could be adapted to the given estimation problem are sketched in Section 5.2.

5.1 Procedures based on data imputation

Imputation techniques have been widely used as methods to overcome the problem of missing data; for an extensive treatise see, for example, Rubin (1987). Applications of these to interval-censored data in survival analysis are reviewed in the following subsection, after which we propose how these methods could be applied to the given estimation problem.

5.1.1 Review of imputation techniques for interval-censored data

In a recent study, Geskus (2001) compares various techniques for the nonparametric estimation of the distribution function of doubly-censored AIDS incubation periods. In his data setting, equal to the one presented in Section 2.2, the moment of infection with HIV is interval-censored, whereas AIDS onset is partially right-censored. Three of the techniques employed use imputation methods for the interval-censored HIV infection times: midpoint, conditional mean and multiple imputation. Once the HIV infection times are imputed, the Kaplan-Meier estimate for right-censored data is used to estimate the distribution function of the AIDS incubation period. In the case of midpoint imputation, the midpoint of the observed interval of HIV infection is imputed, whereas in case of the conditional mean imputation, the expected date of HIV infection based on the nonparametric maximum likelihood estimator of the HIV infection times and conditionally to the observational interval is used. In case of multiple imputation, HIV infection times are imputed multiply, also based on random draws from the NPMLE of HIV infection times, and the subsequent Kaplan-Meier estimates are averaged out.

Concerning semi-parametric models with an interval-censored response variable, Pan (2000a) proposes a multiple imputation approach. His method comprises the iteration of two steps until convergence of the model parameters' estimates is achieved. The first step consists of multiple imputations of the finite interval-censored times, that is, intervals with a finite right-endpoint, and the second of the application of standard statistical procedures for right-censored data. The average of the obtained estimates for every imputed data set in step 2 furnishes the final parameter estimate and is used, together with the Breslow estimate of the baseline survival function, for the data imputation in the step 1.

A similar setting is considered by Pan (2000b) and Goggins, Finkelstein, and Zalavsky (1999). They deal with the proportional hazards model for a doubly-censored response variable: the time origin is interval-censored, the endpoint is possibly right-censored. Pan tackles the problem with a noniterative multiple imputation approach imputing the time origin based on random draws from the NPMLE. Thus, this imputation procedure does not take advantage of the knowledge of the survival time's endpoint. For each imputed data set, the model parameters are estimated and the final estimate is obtained by averaging out these values.

For the same setting, Goggins et al. propose the use of a Monte Carlo EM algorithm. The E-step of this algorithm consists of generating data sets of time origins, for which they use multiple imputation based on random draws from the joint likelihood function of time origin and endpoint. The likelihood function makes use of the parameter estimates obtained in the M-step. In this step, the average log likelihood is used.

5.1.2 Imputation methods

The idea behind these methods is to substitute the observed censored values by an imputed value, and to posteriorly use standard regression methods to estimate the model parameters. We distinguish between response variables that are left- and right-censored (case 1) and responses that are doubly censored (case 2). In both cases, the observed interval $[Z_l, Z_r]$ is replaced by an imputed value, Z' say, and the estimation of the parameter vector $\boldsymbol{\theta}$ is carried out applying standard procedures for the log linear model based on the data vectors: either $(U, Z', \delta_1, \delta_2)$ for case 1 or $(U, Y_{0l}, Y_{0r}, Z', \delta_1, \delta_2)$ for case 2, where U, δ_1, and δ_2 are defined as in (3.6) and (3.7), and $[Y_{0l}, Y_{0r}]$ denote the interval of the response's time origin.

We consider the following imputation methods for the intervals $[Z_l, Z_r]$:

Midpoint imputation For each observation, impute the midpoint of the observed interval:

$$Z' = \tfrac{1}{2}(Z_l + Z_r).$$

Conditional mean imputation The imputation is based on the NPMLE of F_Z, obtained by either the Turnbull estimator (see page 2) or, in case of current status data, the greatest convex minorant algorithm (page 6), and conditioned on the observed interval:

$$Z' = \mathrm{E}_{\hat{F}_Z}(Z|Z_l, Z_r).$$

Multiple imputation For $d = 1, \ldots, D$, impute values $Z_1'^{(d)}, \ldots, Z_n'^{(d)}$ of Z, randomly drawn from the NPMLE $\widehat{F}_Z$ and conditionally to $[Z_{l_i}, Z_{r_i}]$, $i = 1, \ldots, n$. That is, in case of a discrete covariate Z with support $S = \{s_1, \ldots, s_m\}$ where $s_1 < \cdots < s_m$, imputation is carried out using the following probabilities:

$$\mathrm{P}(Z_i'^{(d)} = s_j | Z_{l_i}, Z_{r_i}) = \frac{\alpha_{ij}\hat{\omega}_j}{\sum_{l=1}^m \alpha_{il}\hat{\omega}_l},\ i = 1, \ldots, n,\ j = 1, \ldots, m, \tag{5.1}$$

where $\omega_j = \mathrm{P}(Z = s_j)$. The indicator variables α_{ij} are equal to one if s_j is an admissible value for Z_i and zero otherwise. The final estimate is the average of the values $\hat{\boldsymbol{\theta}}_d$ obtained for each of the D data sets:

$$\hat{\boldsymbol{\theta}} = \frac{1}{D}\sum_{d=1}^{D} \hat{\boldsymbol{\theta}}_d.$$

In case Y is doubly censored, the estimation procedures has to account for the censoring of response variable's time origin, too. However, so far, accelerated failure time models with a doubly censored response variable have not been addressed in the literature and statistical software does not cover that case. To tackle that estimation problem, the imputation techniques above can also be applied to replace the intervals $[Y_{0_l}, Y_{0_r}]$ by an imputed value Y_0'

Comments

In contrast with the simultaneous maximization procedure of Chapter 3, the previously described procedures can be applied to both discrete and continuous covariates. If Z is assumed discrete and contains any right-censored data, a maximum value s_m has to be determined. In case of a continuous Z, the values s_j in formula (5.1) are the values, on which the NPMLE $\hat{F}_Z$ puts positive mass, and the indicator functions α_{ij} correspond to these values.

As long as the majority of the observed intervals of Z is relatively narrow, midpoint imputation might be the preferred method since it is easily accomplished and all three methods would yield similar results. However, for broader intervals, for example in case of current status data, this method might cause biased estimation results when the distribution of Z is not uniform over these intervals.

Some advantages of multiple imputation over single imputation are (Rubin 1987): increase of efficiency, the additional variability as the unknown interval-censored values is taken into account, and it allows for sensitivity analysis. On the other hand, one important disadvantage is the higher computational cost compared with single imputation. If imputation is applied to both intervals, $[Z_l, Z_r]$ and $[Y_{0_l}, Y_{0_r}]$, both advantages and disadvantages of multiple imputation are reinforced.

The imputation techniques using midpoint and conditional mean imputation have been implemented by the author with the programming code of the statistical software package S-Plus; see Section C.3 on page 139. It is one of advantages of this software that it offers two functions to obtain the parameter estimation of the log linear survival model for completely observed covariates, namely the functions `survReg` and `censorReg`, as well as a big variety of different distributions for the response variable.

5.2 Summary of other possible estimation procedures

The methods sketched in this section could also be applied to the given estimation problem, but, in contrast of the procedures above, have not been implemented by the author. The first of these methods is the adaptation of the procedure proposed by Goggins, Finkelstein, and Zalavsky (1999), the following one an approach if the distribution of Z is assumed known, and the remaining two use profile and local likelihood methods, respectively.

5.2.1 The Monte Carlo EM algorithm of Goggins et al.

Herein, we adapt the method of Goggins et al. (1999) to the log linear model when the response is partially left- and right-censored and the covariate is discrete with possible values $s_1, \ldots, s_m$. This Monte Carlo EM algorithm consists of the following steps:

1. Initialization
   Choose initial values $(\hat{\boldsymbol{\theta}}^{(0)}, \hat{\boldsymbol{\omega}}^{(0)})$. For example, for $(\hat{\mu}^{(0)}, \hat{\beta}^{(0)}, \hat{\sigma}^{(0)})'$ choose the maximum likelihood estimates when adjusting model (3.5) to the data using midpoint imputation for the intervals $[Z_l, Z_r]$. For $\hat{\boldsymbol{\omega}}^{(0)}$, choose $\hat{\omega}_j^{(0)} = \frac{1}{m}$, $j = 1, \ldots, m$.

2. The E-step (lth iteration)
   Generate D data sets, each consisting of the observations $(U_i, \delta_{1_i}, \delta_{2_i})$ and the imputed values $Z_i'^{(d)}, i = 1, \ldots, n, d = 1, \ldots, D$. The variable U, defined in Section 3.2, is the random variable of the observed responses and δ_1 and δ_2 are the indicator variables for exact and right-censored observations, respectively. The imputed data are random draws from the joint likelihood function (3.15), each value drawn according to the probabilities

$$\mathrm{P}(Z_i'^{(d)} = s_j | Z_{l_i}, Z_{r_i}) = \frac{\alpha_{ij} f(u_i|s_j)^{\delta_{1_i}} S(u_i|s_j)^{\delta_{2_i}} \big(1 - S(u_i|s_j)\big)^{(1-\delta_{1_i})(1-\delta_{2_i})} \hat{\omega}_j^{(l-1)}}{\sum_{k=1}^{m} \alpha_{ik} f(u_i|s_k)^{\delta_{1_i}} S(u_i|s_k)^{\delta_{2_i}} \big(1 - S(u_i|s_k)\big)^{(1-\delta_{1_i})(1-\delta_{2_i})} \hat{\omega}_k^{(l-1)}},$$

   for $i = 1, \ldots, n$, $j = 1, \ldots, m$, and α_{ij} as in (3.14). Note that the values of the density and the survival functions, f and S, are calculated using the estimates of the previous iteration: $(\hat{\mu}^{(l-1)}, \hat{\beta}^{(l-1)}, \hat{\sigma}^{(l-1)})'$. Their expression is determined by the choice of the distribution of Y.

3. The M-step (lth iteration)
   For each of the D data sets consisting of $\big(U_i, \delta_{1_i}, \delta_{2_i}, Z_i'^{(d)}\big)$, $i = 1, \ldots, n$, the joint likelihood functions are given by

$$L_d(\boldsymbol{\theta}, \boldsymbol{\omega}) = \prod_{i=1}^{n} f(u_i|z_i'^{(d)})^{\delta_{1_i}} S(u_i|z_i'^{(d)})^{\delta_{2_i}} \big(1 - S(u_i|z_i'^{(d)})\big)^{(1-\delta_{1_i})(1-\delta_{2_i})} \mathrm{P}(Z = z_i'^{(d)}).$$

   These likelihood functions can be separated into two factors depending on either $\boldsymbol{\theta}$ or $\boldsymbol{\omega}$. Hence, the updated estimates, $\hat{\boldsymbol{\theta}}^{(l)}$ and $\hat{\boldsymbol{\omega}}^{(l)}$, can be obtained separately:

   (a) Obtain the new estimate $\hat{\boldsymbol{\omega}}^{(l)}$ from the empirical distribution function $\widehat{F}_Z$ of all imputed values $Z_i'^{(d)}$, $i = 1, \ldots, n$, $d = 1, \ldots, D$.

   (b) Maximizing the average log likelihood function, $\frac{1}{D}\sum_{d=1}^{D} l_d(\boldsymbol{\theta})$, furnishes $\hat{\boldsymbol{\theta}}^{(l)}$. The log likelihood functions $l_d(\boldsymbol{\theta}) = \ln L_d(\boldsymbol{\theta})$, $d = 1, \ldots, D$, have the following expressions:

$$l_d(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \Big(f(u_i|z_i'^{(d)})^{\delta_{1_i}} S(u_i|z_i'^{(d)})^{\delta_{2_i}} \big(1 - S(u_i|z_i'^{(d)})\big)^{(1-\delta_{1_i})(1-\delta_{2_i})} \Big).$$

4. Return to step 2 until convergence is achieved.

Alternatively, in step 3b, instead of maximizing the average log likelihood function, the log linear model could be adjusted for each of the D data sets. In this case, the estimate $\hat{\boldsymbol{\theta}}^{(l)}$ would be calculated as the average of the estimates $\hat{\boldsymbol{\theta}}_d^{(l)}$, $d = 1, \ldots, D$. As Goggins et al. remark, the choice of D has to take into account the grid of values of Z, because mass points may disappear during an iteration. The variance of the parameter estimates can be estimated by means of the expected value of the score statistics, which accounts for the additional variability due to the imputation of interval-censored data.

5.2.2 Parametric choice for the covariate's distribution

An alternative to the nonparametric estimation of F_Z is a parametric choice for the distribution of Z. In this case, the particular expression of the density function, f_Z, can be plugged into the likelihood function. For example, the likelihood function (3.13) for a model with a possibly right- and left-censored response variable is then equal to

$$L(\boldsymbol{\theta}, \boldsymbol{\eta}) = \prod_{i=1}^{n} \int_{z_{l_i}}^{z_{r_i}} f(u_i|z)^{\delta_{1_i}} S(u_i|z)^{\delta_{2_i}} \big(1 - S(u_i|z)\big)^{(1-\delta_{1_i})(1-\delta_{2_i})} f_Z(z)\, dz, \tag{5.2}$$

where $\boldsymbol{\eta}$ denotes the parameters of Z's distribution. Maximizing expression (5.2) yields the estimated parameters of both the model and the distribution of Z.

For the maximization of (5.2), the integrals can be approximated by sums applying the trapezoidal method, which is illustrated in Figure 5.1 for a continuous function $f(x)$. A grid of points has to be chosen, say $x_1, \ldots, x_n$. Then, the area under the function between x_1 and x_n is approximated by the sum over the areas of the trapezoids defined by $x_j, x_{j+1}, f(x_j)$, and $f(x_{j+1})$, $j = 1, \ldots, n-1$:

$$\int_{x_1}^{x_n} f(x)\, dx \approx \sum_{j=1}^{n-1} \tfrac{1}{2}(x_{j+1} - x_j)\big(f(x_j) + f(x_{j+1})\big). \tag{5.3}$$

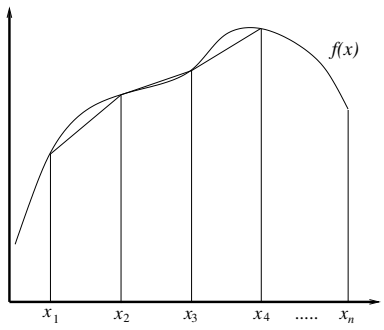

Figure 5.1: Trapezoidal method

The finer the grid of points, the more accurate the approximation, but the higher the computational cost. For the application of the trapezoidal method, to make programming easier, it is recommendable that the interval endpoints are a subset of the chosen grid of points and that these are equidistant. For the implementation of this procedure, AMPL is an appropriate tool and Maple can be used to estimate the variances of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\eta}}$.

As an example, consider the choice of the exponential distribution with the density function $f(z) = \eta \exp(-\eta z)$. Hence, the likelihood function to be maximized has the following expression:

$$L_n(\boldsymbol{\theta}, \eta) = \prod_{i=1}^{n} \int_{z_{l_i}}^{z_{r_i}} f(u_i|z)^{\delta_{1_i}} S(u_i|z)^{\delta_{2_i}} \big(1 - S(u_i|z)\big)^{(1-\delta_{1_i})(1-\delta_{2_i})} \eta \exp(-\eta z)\, dz.$$

Each of the n intergrals could be approximated by formula (5.3), where $f(x_j)$ would be given by

$$f(u_i|x_j)^{\delta_{1_i}} S(u_i|x_j)^{\delta_{2_i}} \big(1 - S(u_i|x_j)\big)^{(1-\delta_{1_i})(1-\delta_{2_i})} \eta \exp(-\eta x_j).$$

5.2.3 Profile likelihood function

To determine the profile likelihood function and maximize it, is an approach to obtain a maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, when the likelihood function also depends on a infinite-dimensional nuisance parameter such as in semi-parametric models (Murphy and van der Vaart 2000). For example, as long as f_Z remains unspecified in likelihood function (5.2), this function depends on such a nuisance parameter.

The idea of this approach is to "profile out" the nuisance parameter, say $\boldsymbol{\eta}$, to obtain the profile likelihood function $L_{prof}(\boldsymbol{\theta})$, which is defined as follows:

$$L_{prof}(\boldsymbol{\theta}) = \sup_{\boldsymbol{\eta}} L(\boldsymbol{\theta}, \boldsymbol{\eta}).$$

The maximization of $L_{prof}(\boldsymbol{\theta})$ furnishes the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, since the profile likelihood function behaves like the ordinary likelihood function in that it has a quadratic expansion (Murphy and van der Vaart).

Staniswalis and Thall (2001) show how the profile likelihood function can be obtained and recommend its use in case the maximization of $\ln L(\boldsymbol{\theta}, \eta)$ is computationally difficult due to the infinite-dimensional parameter $\boldsymbol{\eta}$.

5.2.4 Local likelihood approach

Another approach to tackle the maximization of the above likelihood functions with a continuous covariate is the local likelihood approach as proposed by Betensky, Lindsey, Ryan, and Wand (2002) and Bebchuk and Betensky (2002). They use it for the Cox model with interval-censored survival times and the estimation of the distribution function of a doubly-censored incubation time, respectively.

Applied to model (3.5), this method consists of the use of a local approximation of the covariate's density function f_Z by means of a kernel estimate at a grid of points to be chosen. This estimate is plugged into the (log) likelihood function, which has to be maximized with respect to $\boldsymbol{\theta}$ and the parameters of the kernel function. According to Bebchuk and Betensky, one of the advantages of the local likelihood method is that it avoids the possible biasof a parametric choice of F_Z. However, asymptotic properties are unknown because the local likelihood function it is not derived from the true likelihood.