Chapter 7

Simulation Study

Simultaneous maximization with AMPL has been shown to be a very valuable method for the parameter estimation of a regression model, whenever standard statistical software is not available to carry out the estimation due to particular data and/or censoring patterns. However, a few disadvantages remain. One of these is the need of a second software for the computation of the parameters' confidence intervals. In contrast with that, the imputation based procedures presented in Chapter 5 can be implemented in statistical software such as S-Plus or R, which accomplish both parameter estimation and computation of the confidence intervals. This has motivated a simulation study, in order to compare imputation techniques with simultaneous maximization with respect to precision of the estimation results.

The model we consider is the accelerated failure time model (6.1), which relates the logarithm of the response variable with the one of the discrete interval-censored covariate:

$$\ln(Y) = \mu + \beta \ln(Z) + \sigma W. \tag{7.1}$$

As before, the support of Z is denoted by $S = \{s_1, \ldots, s_m\}$ and the corresponding probabilities by $\omega_j = \mathrm{P}(Z = s_j),\ j = 1, \ldots, m$, summarized in the vector $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_m)'$. Instead of Z, the intervals $[Z_l, Z_r]$ are given, where $\mathrm{P}(Z \in [Z_l, Z_r]) = 1$. The response variable Y is a positive real-valued random variable partly exactly observed, partly censored. We mainly consider right-censoring and, less frequently, left-censored data. All censored values are generated following distributions independently of the true values Y and Z.

Within this chapter, we first present the three estimation procedures, simultaneous maximization, on one hand, and midpoint and conditional mean imputation, on the other. The different simulation settings —a total of 48— and the evaluation criteria used for the comparison of the methods are addressed in Sections 7.2 and 7.3, respectively. Finally, the simulation results are summarized in Section 7.4; the corresponding tables with the results for each simulation setting are attached in Chapter D in the appendix.

7.1 Estimation procedures

Three estimation procedures are compared within the framework of the present simulation study: midpoint and conditional mean imputation, both implemented with S-Plus, and simultaneous maximization accomplished with AMPL.

7.1.1 Imputation based methods

We consider two of the three imputation methods described in Section 5.1.2: midpoint and conditional mean imputation. The third, multiple imputation, could also be implemented with S-Plus, however, the cost of implementation is higher and makes it, hence, less attractive in practise. In contrast with multiple imputation, midpoint and conditional mean imputation require less programming, especially the former, which replaces the intervals $[Z_l, Z_r]$ by their midpoint: $Z' = \frac{1}{2}(Z_l + Z_r)$. The latter needs two steps: application of the S-Plus function `kaplanMeier` to determine the Turnbull estimate $\hat{F}_Z$ and the posterior computation of the conditional mean:

$$Z_i' = \mathrm{E}_{\hat{F}_Z}(Z_i | Z_{l_i}, Z_{r_i}) = \sum_{j=1}^{m} \alpha_{ij} s_j \omega_j \bigg/ \sum_{l=1}^{m} \alpha_{il} \omega_l, \; i = 1, \ldots, n,$$

where $\alpha_{ij} = \mathbb{1}_{\{s_j \in [z_{l_i}, z_{r_i}]\}}$.

After the imputation step, in both cases, model (7.1) is adjusted using the imputed variable Z':

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

For this purpose, the S-Plus function `censorReg` is applied, which allows for left-, right-, and interval-censored data in the response variable.

The S-Plus programme, which carries out both midpoint and conditional mean imputation for the simulation, study is shown in Section C.3.2 in the appendix.

7.1.2 Simultaneous maximization

With AMPL it is possible to maximize the log likelihood function of model (7.1) simultaneously with respect to $\boldsymbol{\theta} = (\mu, \beta, \sigma)'$ and $\boldsymbol{\omega}$. This is carried out for each of the settings with the solver SNOPT. The log likelihood function is similar to the one in Chapter 3 on page 46 for a left- and right-censored response:

$$l(\boldsymbol{\theta}, \boldsymbol{\omega}) = \sum_{i=1}^{n} \ln\Big(\sum_{j=1}^{m} \alpha_{ij} f_W\big(\ln(u_i) | \ln(s_j)\big)^{\delta_{1_i}} S_W\big(\ln(u_i) | \ln(s_j)\big)^{\delta_{2_i}} \big(1 - S_W\big(\ln(u_i) | \ln(s_j)\big)\big)^{(1-\delta_{1_i})(1-\delta_{2_i})} \omega_j \Big), \tag{7.2}$$

where f_W and S_W are the density and survival function of the error term distribution W, and U is defined as in (3.6) on page 42. When Y follows a Weibull distribution, W is the Gumbel distribution and (7.2) is equal to the log likelihood function (3.25); in case of the log logistic model, it is equal to (3.30). The maximization of the log likelihood is subject to the inequality constraints $\sigma > 0$ and $\omega_j \geq 0$, $j = 1, \dots, m$, as well as to the equality constraint $\sum_{j=1}^{m} \omega_j = 1$.

The AMPL programmes, consisting of a model, a data, and a programme file, are given in Section C.1.3. As commented in detail in Section 7.3 below, the simulation study does not require the use of Maple for the computation of the variances. They are estimated based on the parameters' estimates for each the data sets of a given simulation setting.

7.2 Simulation settings and data generation

7.2.1 Parameters defining simulation settings

The settings are defined by the following parameters chosen according to the ones of the data set from the hospital Can Ruti in the previous chapter. A total of $2 \cdot 2 \cdot 2 \cdot 2 \cdot 3 = 48$ combinations are considered and for each of these settings, $D = 500$ data sets are generated.

- **Distribution of the response variable**
  Weibull and log logistic.

- **Distribution of the covariate**
  Normal with mean $\mu = 50$ and variance $\sigma^2 = 100$, and Weibull with parameters $\lambda = 0.0004$ and $\alpha = 2$.

- **Parameter values**
  $(\mu, \beta, \sigma) \in \{(3, 0.45, 0.65), (4, 0.25, 0.45)\}$.

- **Average length of intervals** $[Z_l, Z_r]$
  On one hand, on average, intervals of length equal to 16 units, on the other, current status data.

- **Number of observations**
  $n \in \{50, 150, 300\}$.

With the two combinations of the parameter values, the relative risk/odds ratio amounts to 0.5 and 0.574, respectively. The choice of the Weibull distribution parameters implies a right-skewed distribution with mean and standard deviation approximately equal to 44.3 and 23.2, respectively. Both density functions are shown in Figure 7.1. On average, with the Weibull choice, the generated intervals $[Z_l, Z_r]$ cover a broader range of values on the support of Z than with the Normal distribution.

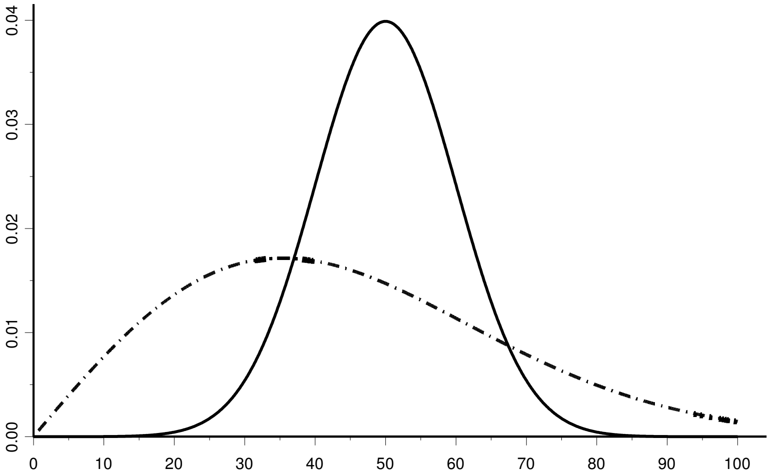

Figure 7.1: Normal and Weibull distribution densities of the covariate

7.2.2 Generation of data sets

For each of the D data sets of a particular setting, n data vectors of the following form are generated: $(U, Z_l, Z_r, \delta_1, \delta_2)$. This is carried out accomplishing the instructions in the following steps:

1. Generation of Z

   According to the setting, either $Z \sim \mathcal{N}(50, 10)$ or $Z \sim \mathcal{W}(0.0004, 2)$. Round Z to the nearest integer and force it to be greater or equal to one. To avoid very large values of Z with the Weibull choice, we set an upper limit of $\max(Z) = 100$.

2. Generation of Y

   Compute $Y = \exp\big(\mu + \beta \ln(Z) + \sigma W\big)$, where W is a random number of either the standard Gumbel or the standard logistic distribution.

3. Generation of $I_Z = [Z_l, Z_r]$

   (a) Average width of I_Z equal to 16

   Generate random numbers C_l and C_r, both uniformly distributed over $[1, 15]$ and compute $Z_l = \max(1, Z - C_l)$ and $Z_r = Z + C_r$. Round both Z_l and Z_r to the nearest integer.

(b) Current status data

Generate two random numbers: C_Z, uniformly distributed over $[1, 15]$, and P_Z, Bernoulli distributed with $p = 0.65$. Then, determine I_Z by:

$$Z_l = \begin{cases} 1 & \text{if } P_Z = 1, \\ \max(1, Z - C_Z) & \text{if } P_Z = 0, \end{cases} \qquad Z_r = \begin{cases} Z + C_Z & \text{if } P_Z = 1, \\ s_m & \text{if } P_Z = 0, \end{cases}$$

rounding $Z - C_Z$ and $Z + C_Z$ to the nearest integers. In each of the D data sets, set $s_m = \max\big\{\{Z + C_Z | P_Z = 1\} \cup \{Z - C_Z | P_Z = 0\}\big\}$, that is equal to the maximum value of the "observed" current status data.

4. Generation of U, δ_1, and δ_2

Generate two random numbers: C_U uniformly distributed over $[1, 15]$, and P_U uniformly distributed over $[0, 1]$. Then, calculate U, δ_1, and δ_2:

$$U = \begin{cases} Y & \text{if } P_U \leq 0.45, \\ \max(1, Y - C_U) & \text{if } 0.45 < P_U \leq 0.9, \\ Y + C_U & \text{otherwise.} \end{cases}$$

$$\delta_1 = \begin{cases} 1 & \text{if } P_U \leq 0.45, \\ 0 & \text{otherwise.} \end{cases}$$

$$\delta_2 = \begin{cases} 1 & \text{if } 0.45 < P_U \leq 0.9, \\ 0 & \text{otherwise.} \end{cases}$$

Note that in step 3(b), right- and left-censored values of Z are generated in proportion 65:35. In the last step concerning U, on average, 45% of exact observations, 45% of right-censored values, and 10% of left-censored values are generated. Since Y is a continuous variable, it is not necessary to round U to an integer, however, it must be a positive value, since the logarithm of Y is modeled. Different to the data from Can Ruti, we do not consider the case of missing values of the response variable.

The generation of the data sets is accomplished by the function `simgendat` programmed with the S-Plus code. Its code is shown on page 139.

7.3 Evaluation criteria

7.3.1 Estimation of parameters, relative risk, and conditional median

Given any of the simulation settings, $D\,(= 500)$ data sets are generated as described before, and for each, the parameter vector is estimated. Thus, we obtain $\hat{\boldsymbol{\theta}}_d$, $d = 1, \ldots, D$. Based on these estimations and given the true parameter vector $\boldsymbol{\theta}_0$, we calculate the mean, variance, bias, and mean square error (MSE) of the estimates. These measures are used to judge the precision of each of the three estimators:

$$\bar{\hat{\boldsymbol{\theta}}} = \frac{1}{D} \sum_{d=1}^{D} \hat{\boldsymbol{\theta}}_d, \tag{7.3a}$$

$$\widehat{\text{Var}}(\hat{\boldsymbol{\theta}}) = \frac{1}{D-1} \sum_{d=1}^{D} \big(\hat{\boldsymbol{\theta}}_d - \bar{\hat{\boldsymbol{\theta}}}\big)^2, \tag{7.3b}$$

$$\widehat{\text{Bias}}(\hat{\boldsymbol{\theta}}) = \bar{\hat{\boldsymbol{\theta}}} - \boldsymbol{\theta}_0, \tag{7.3c}$$

$$\widehat{\text{MSE}}(\hat{\boldsymbol{\theta}}) = \widehat{\text{Var}}(\hat{\boldsymbol{\theta}}) + \widehat{\text{Bias}}(\hat{\boldsymbol{\theta}})^2. \tag{7.3d}$$

Contrary to the evaluation of a single data set with AMPL, the estimation of the estimator's variance by means of formula (7.3b) does not require the use of Maple to compute $\widehat{\text{Var}}(\hat{\boldsymbol{\theta}}_d)$, $d = 1, \ldots, D$. The whole simulation process can be carried out by AMPL once the data are generated with S-Plus.

Concerning the imputation methods with S-Plus, $\text{Var}(\hat{\boldsymbol{\theta}}_d)$ could easily be estimated within each of the D runs and their mean could serve also as an estimator for $\text{Var}(\hat{\boldsymbol{\theta}})$. However, formula (7.3b) is more appropriate, since $\widehat{\text{Var}}(\hat{\boldsymbol{\theta}}_d)$ usually underestimates the true variance $\text{Var}(\hat{\boldsymbol{\theta}})$, because both estimation procedures based on imputation do not account for the additional variance due to the imputation of Z.

Besides comparing the estimation of the single parameters, we are also interested in specific terms including at least two of them, since the precision of the estimators might vary from parameter to parameter. In particular, our measures of interest are the relative risk/odds ratio and the conditional median of Y given Z as inference from the accelerated failure time model is often based on these values. They are estimated as follows:

$$\widehat{\text{RR}} = \widehat{\text{OR}} = \exp\big(-\hat{\beta}/\hat{\sigma}\big), \tag{7.4}$$

$$\widehat{\text{Med}}(Y|Z) = \begin{cases} \exp\big(\hat{\mu} + \hat{\beta}\ln(Z)\big)\ln(2)^{\hat{\sigma}} & \text{Weibull regression model,} \\ \exp\big(\hat{\mu} + \hat{\beta}\ln(Z)\big) & \text{Log logistic regression model.} \end{cases} \tag{7.5}$$

Estimation of the relative risk and the conditional median is carried out for each of the D data sets of a given setting, and as with the single parameters, the mean, variance, bias, and MSE are calculated as in (7.3a) to (7.3d). Regarding the conditional median, we consider the values $Z = 20$ and $Z = 50$.

7.3.2 Estimation of the covariate's distribution function

Another point of interest refers to the estimation of F_Z, the distribution function of Z. Concerning the imputation based methods, midpoint imputation does not require any knowledge on F_Z, whereas with conditional mean imputation, the imputation step is based on the Turnbull estimator of F_Z. On the other hand, simultaneous maximization furnishes an estimation of Z's distribution function characterized by $\hat{\boldsymbol{\omega}}$. It is, hence, interesting to see, whether that estimate differs from the Turnbull estimate.

For the purpose of comparison, we estimate five quantiles for each data set and calculate their mean, variance, bias, and the MSE. The chosen quantiles are the 10%-, 25%-, 50%-, 75%-, and the 90%-quantile.

The software package S-Plus offers the command `qkaplanMeier` for the calculation of quantiles based on the Turnbull estimates. This function uses linear interpolation to compute the quantiles. Regarding AMPL, the estimation of the quantiles z_p, $p = 0.1, 0.25, 0.5, 0.75, 0.9$, is accomplished by including the following estimation in the programme file:

$$\hat{z}_p = \min\big\{s_j \big| \sum_{l=1}^{j} \hat{\omega}_l \geq p\big\}. \tag{7.6}$$

7.4 Simulation results

In the sequel, we first summarize the main findings concerning the estimation of the parameters μ, β, and σ. The estimation of the relative risk and the conditional median as well as the distribution function of Z is addressed in Sections 7.4.2 and 7.4.3, respectively.

7.4.1 Single parameter estimation

The estimation results for all 48 settings, including the values of $\bar{\hat{\boldsymbol{\theta}}}$, $\widehat{\mathrm{Bias}}(\hat{\boldsymbol{\theta}})$, and $\widehat{\mathrm{MSE}}(\hat{\boldsymbol{\theta}})$ calculated according to formulas (7.3a)–(7.3d), are shown in the tables of Section D.1 on pages 150 through 157.

Herein, for the purpose of illustration, we mainly compare the three estimation procedures with respect to the number of parameter estimates with least bias. The results are summarized in Tables 7.1 and 7.2. For example, in Table 7.1 we see, that simultaneous maximization (SIMAX) estimates the parameter μ with less bias than the imputation methods in 29 of the 48 settings. In total, this procedures yields parameter estimates with least bias in the majority of all cases: 84 out of 144 cases (58.3%), compared to 43 cases (29.9%) of midpoint imputation (MID), and 17 cases (11.8%) of conditional mean imputation (COND). However, we can observe some differences with respect to each of the three parameters: whereas midpoint imputation estimates the parameter β best in most cases, SIMAX yields least bias for the parameters μ and, especially, σ. This latter parameter is estimated nearly identically by both imputation methods and with larger bias than by SIMAX. Concerning the parameter μ, COND furnishes somewhat better results than MID.

Table 7.1: Number of least biased parameter estimations

Parameter	**MID**[a]	**COND**[b]	**SIMAX**[c]
$\boldsymbol{\mu}$	9	10	29
$\boldsymbol{\beta}$	30	7	11
$\boldsymbol{\sigma}$	4	0	44
Total	43	17	84

[a] Midpoint imputation

[b] Conditional mean imputation

[c] Simultaneous maximization

Table 7.1 does not contain any information on the mean square error of the estimators, neither quantifies it the differences of the bias between the three methods. All this information can be found in the tables in Section D.1. An illustration for two of these tables is shown in Figures 7.2 and 7.3 below, where the relative bias of the estimation procedures is compared for μ, β, and σ. The relative bias is defined as $\widehat{\text{Bias}}(\hat{\boldsymbol{\theta}})/\boldsymbol{\theta}_0$ and allows to compare the bias of different parameters on the same scale. In both figures, the values from one through six on the abscises correspond to: 1: $n = 50$, $[Z_l, Z_r]$ narrow; 2: $n = 50$, $[Z_l, Z_r]$ current status data; 3: $n = 150$, $[Z_l, Z_r]$ narrow; 4: $n = 150$, $[Z_l, Z_r]$ CSD; 5: $n = 300$, $[Z_l, Z_r]$ narrow; 6: $n = 300$, $[Z_l, Z_r]$ CSD.

For example, we can see in the first of both figures, that some of the values of midpoint imputation for $\hat{\mu}$ and $\hat{\beta}$ are far from the true values, precisely when the covariate's observations are current status data. On the other hand, the estimates of conditional mean imputation and simultaneous maximization, generally, do not differ that much from MID, when this procedure yields least biased estimates.

Concerning mean square error, simultaneous maximization yields least MSE for the estimator $\hat{\sigma}$, midpoint imputation mostly for $\hat{\mu}$ and $\hat{\beta}$. This is also true for cases, where its bias is larger than the one of COND and SIMAX. We conjecture that this is due to the fact that each imputed value does only depend on the corresponding interval $[Z_l, Z_r]$, but not on the estimation of F_Z, which is based on all intervals. In contrast with that, conditional mean imputation and simultaneous maximization estimate F_Z, which introduces more variability into the estimation procedure. Also, SIMAX tackles the maximization of the corresponding likelihoods with respect to up to 100 unknown parameters. Naturally, a small mean square error is desirable, whenever the bias is low; however, a small variance in combination with a large bias implies, that the estimator is generally far from the true value.

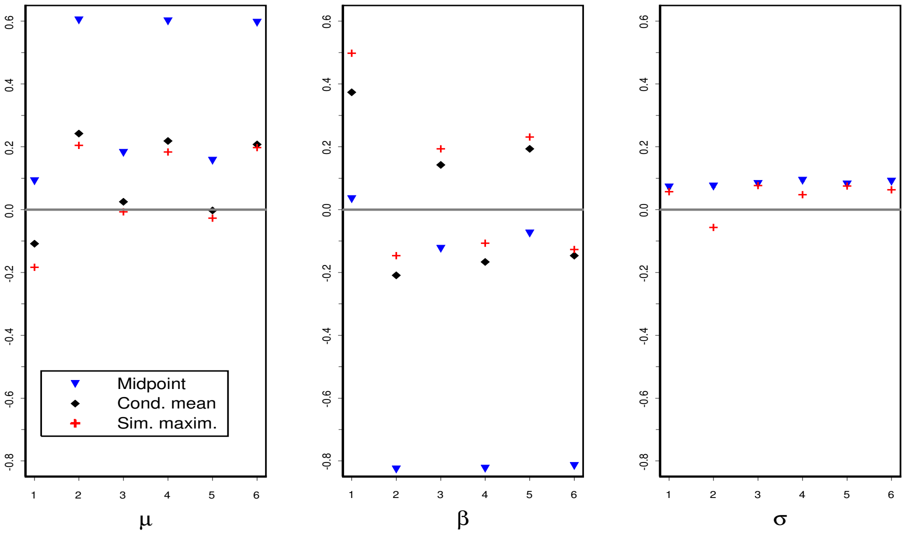

Figure 7.2: Comparison of relative bias of estimation procedures for the Weibull regression model with a normally distributed covariate and parameters equal to $\mu = 3, \beta = 0.45, \sigma = 0.65$

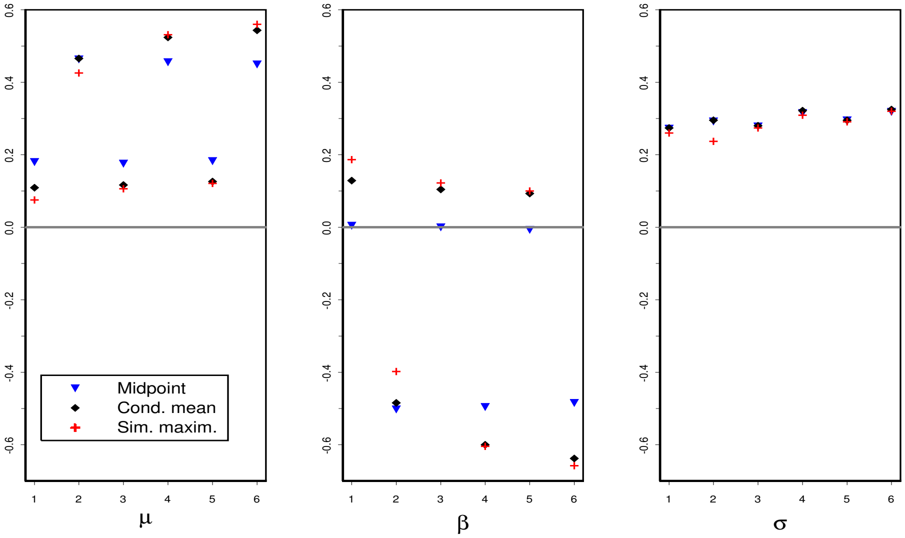

Figure 7.3: Comparison of relative bias of estimation procedures for the log logistic regression model with a Weibull-distributed covariate and parameters equal to $\mu = 3, \beta = 0.45, \sigma = 0.65$

Table 7.2 does also show the number of least biased estimates by each of the estimation procedures, but now subdivided according to the response' and the covariate's distribution as well as to the average width of the intervals $[Z_l, Z_r]$. Neither sample size nor parameter values are considered, since the comparison of MID, COND, and SIMAX with respect to these factors has not shown substantial differences. In particular, given any of the settings, increasing the sample size has hardly changed the differences between the three estimation procedures.

We can observe similar results for the Weibull and the log logistic regression model, but differences regarding the covariate's distribution and the width of the intervals. For example, conditional mean imputation shows better results with the normally distributed covariate than with a Weibull-distributed one. With midpoint imputation it is the opposite. With any distribution of the response and the covariate, simultaneous maximization estimates most parameters with least bias. However, an interesting difference is observed. With the normally distributed covariate, SIMAX and COND furnish better results, if the intervals $[Z_l, Z_r]$ correspond to current status data, with the Weibull-distributed covariate it is just the other way round.

Naturally, these differences are not due to the parametric choice of the covariate, but are due to the distribution of the intervals over the support of Z. According to the covariate's density functions, illustrated in Figure 7.1, on average, the current status data lie closer around the mean with the normal distribution. In contrast with that, there is much more dispersion with the chosen Weibull distribution. The former case causes a larger bias for midpoint imputation, whereas the latter case introduces more variability into the estimation of F_Z. This is confirmed by the estimation of the corresponding quantiles as described in Section 7.4.3 and shown in Tables D.7a through D.10b in Section D.3. With current status data, the quantiles are estimated with far larger bias in case of the Weibull distribution. This is true for both the Turnbull estimator, which is used with conditional mean imputation, and simultaneous maximization.

Table 7.2: Number of least biased parameter estimations according to simulation settings

		Distribution of response variable					
		Weibull			**Log logistic**		
Covariate	$[\boldsymbol{Z_l, Z_r}]$	**MID**[a]	**COND**[b]	**SIMAX**[c]	**MID**	**COND**	**SIMAX**
Normal	Narrow	6	4	8	7	2	9
	CSD[d]	0	3	15	0	7	11
Weibull	Narrow	5	1	12	6	0	12
	CSD	10	0	8	9	0	9
Total		21	8	43	22	9	41

[a] Midpoint imputation

[b] Conditional mean imputation

[c] Simultaneous maximization

[d] Current status data

Summarizing these findings, simultaneous maximization yields least biased estimates in the majority of cases. Even when the parameters are estimated with bias larger than the one of midpoint imputation, the differences are not that big. An exception are the settings with a Weibull distributed covariate and current status data for Z. For these settings, midpoint imputation yields as many least biased estimates as simultaneous maximization. Here, MID might be a valuable alternative to simultaneous maximization given the fact that it requires less programming. However, both imputation based procedures have the disadvantage, that they underestimate the variance of $\hat{\boldsymbol{\theta}}$ since the variability caused by the imputation step is not taken into account by the subsequent model adjustment. This underestimation has been confirmed by the simulation results (without tables) and would have to be overcome, for example, by the use of bootstrapping methods.

General findings

The model constant μ is mostly overestimated by all methods. This is especially true in case of current status data of the covariate, where a clearly larger bias is observed than with narrow intervals $[Z_l, Z_r]$. In these cases, the values of $\hat{\beta}$ are smaller, clearly underestimating β, as illustrated in the two figures above for abscises values 2, 4, and 6. This fact indicates that the less precise the information on the covariate, the less exact its effect on the response can be estimated.

With a normally distributed covariate, conditional mean imputation and simultaneous maximization can yield estimates far from the true values, whenever the sample size is small ($n = 50$). This is reflected by the high mean square error for $\hat{\mu}$, for example, in Tables D.1a or D.3a, even though the bias yielded by SIMAX is still smaller than the others. These extreme estimates, which affect especially the transformations of the parameters (see below) do not occur with the estimation of the scale parameter σ and disappear with bigger sample sizes.

7.4.2 Relative risk, odds ratio, and conditional median

For each of the $D = 500$ generated data sets of a setting, the relative risk, the odds ratio as well as the conditional median for $Z = 20$ and $Z = 50$ have been estimated according to equations (7.4) and (7.5), respectively. The respective results are summarized in the tables of Section D.2 on pages 159 through 164.

As mentioned before, in case of the normally distributed covariate and with $n = 50$, conditional mean imputation and simultaneous maximization have yielded some parameter estimates far from the true values. In case of the relative risk/odds ratio and the conditional median, these have caused extreme estimates in about 1%–5% of a setting's data sets. With the relative risk/odds ratio, this has occurred for $\hat{\beta} < 0$ and $\hat{\sigma}$ very close to zero. For this reason, values of $\widehat{\mathrm{RR}}$ ($\widehat{\mathrm{OR}}$) bigger than 75 as well as estimated conditional median values bigger than 2000 have been disregarded for the evaluation. This has to be taken into account, when looking at the tables for estimates with $n = 50$ observations. Such extreme values have not appeared with $n = 150$ and $n = 300$.

Relative risk/Odds ratio

Given the, generally, least biased estimation result of $\hat{\sigma}$ by simultaneous maximization and the slightly better results for the estimation of the parameter β of the midpoint imputation method, generally, one of both methods furnishes the least biased estimate of the relative risk or the odds ratio, respectively. The latter more often in log logistic regression models, simultaneous maximization more often with the Weibull choice.

Conditional median times

Two values for Z are chosen: 20 and 50. Given the least biased estimates for μ for simultaneous maximization, the estimated median times are generally best estimated by this method for both the Weibull and the log logistic models. However, there are settings, for which simultaneous maximization yields the least biased estimation of each of the parameters, but not for the conditional mean for both values of Z. This has to do, on one hand, with the direction of the bias as β, contrary to μ, is often underestimated, and, on the other, on the exponential transformation of the parameters.

Table 7.3 on the following page contains the estimated median times of one of the Weibull regression models with a normally distributed covariate. For sample sizes above 50, the median given $Z = 20$ is estimated with least bias by simultaneous maximization, whereas MID shows better results for $Z = 50$ when current status data are present.

In general, all median times are clearly overestimated due to the fact that the parameter μ is always overestimated, even more with currents status data in the covariate.

7.4.3 Distribution function of the covariate

The comparison of the estimation of the quantiles of $\hat{F}_Z$ by either the Turnbull estimator or simultaneous maximization is summarized in the tables of Section D.3 on pages 166 through 173.

Whereas the quantile estimation according to formula (7.6) has always furnished an estimation result, the S-Plus function `qkaplanMeier` applying the Turnbull estimator has not; especially for the 75%- and the 90%-quantile with current status data. To our knowledge, this has to do with the internals of the function's algorithm, which does not calculate a quantile z_p, whenever this lies beyond the largest value z_r^*, for which $\hat{F}_Z(z_r^*) < 1$. For this reason, the calculation of the mean of the respective quantiles is based on less than $D = 500$ values.

For nearly all settings and independently of the chosen distribution of Z, the implemented formula (7.6) has furnished more precise results with less mean square error than the Turnbull estimator. However, differences are small, generally around 0.5 units, which might be due to the applied interpolation of the S-Plus function `qkaplanMeier`. The 50%-, 75%- and 90%-quantiles are, generally, underestimated, the remaining two are slightly overestimated.

Table 7.3: Conditional median estimation in the Weibull regression model with a normally distributed covariate and parameters equal to $\mu = 3, \beta = 0.45, \sigma = 0.65$

	Midpoint Imputation		**Cond. Mean Imputation**		**Simultaneous Maximization**	
	Mean	**Bias**	**Mean**	**Bias**	**Mean**	**Bias**
			$\mathbf{Med}(Y\|Z = 20) = 60.94$			
$\boldsymbol{n = 50}$[a]	98.84	37.9	105.09	44.15	106.33	45.4
	124.66	63.72	145.48	84.54	154.83	93.9
$\boldsymbol{n = 150}$	91.92	30.99	85.68	24.75	84.42	23.48
	119.75	58.82	103.94	43.0	102.41	41.48
$\boldsymbol{n = 300}$	88.38	27.44	80.55	19.61	79.22	18.28
	119.1	58.16	96.12	35.19	96.4	35.46
			$\mathbf{Med}(Y\|Z = 50) = 92.04$			
$\boldsymbol{n = 50}$	127.37	35.33	127.4	35.37	127.09	35.05
	130.97	38.93	132.25	40.21	136.68	44.64
$\boldsymbol{n = 150}$	125.49	33.46	125.32	33.28	124.95	32.91
	127.9	35.87	129.84	37.81	132.01	39.97
$\boldsymbol{n = 300}$	126.32	34.28	126.19	34.16	125.67	33.64
	128.14	36.1	130.21	38.18	131.65	39.61

[a] The first line refers always to narrow intervals $[Z_l, Z_r]$, the second to current status data

As mentioned before, with current status data, the bias is far bigger when Z follows a Weibull distribution, whereas little difference is observed with narrow intervals $[Z_l, Z_r]$.

In case of the simultaneous maximization procedure, no substantial differences are observed between the different scenarios with respect to the values of $\boldsymbol{\theta}$ the distribution of the response variable. It seems as if the distribution of Y and/or the values of the model parameters have only little influence on $\hat{F}_Z$. Naturally, this is a desirable property.

7.5 Conclusions

Summarizing the observed simulation results, we can conclude that the proposed simultaneous maximization method is an adequate and valuable tool for the parameter estimation in an accelerated failure time model with an interval-censored covariate. On average, it has furnished least biased estimates for both the model parameters and the covariate's distribution function, F_Z.

However, it has been somewhat surprising to observe the relatively good results of midpoint estimation for the parameter β and its smallest mean square error for estimators $\hat{\mu}$ and $\hat{\beta}$ in most of the settings. The latter observation is probably due to the fact that midpoint estimation does not include the estimation of F_Z. In contrast with that, its estimation, in case of conditional mean imputation and simultaneous maximization, introduces more variability. This can be observed especially for the small sample sizes with only 50 observations. In this case, due to the discretization of Z, simultaneous maximization has dealt with more unknown parameters than observations.

Similar observations have been reported by Geskus (2001) and Robins and Ritov (1997). The former, as mentioned in Section 5.1.1, compares several procedures for the nonparametric estimation of doubly-censored AIDS incubation periods including imputation methods and a full likelihood approach. This method provides also the nonparametric estimation of the distribution function of the HIV infection times. Contrary to the expectation of the author, this approach has not been superior to the imputation techniques. Robins and Ritov, on the other hand, address the inference in designed studies with randomly missing data when the missingness depends on a high-dimensional vector of variables. To overcome this problem, they propose what they call a curse of dimensionality appropriate (CODA) asymptotic theory.

In our case, the dimension of the nuisance parameter could be reduced by choosing a broader grid for Z. This might reduce variance of the parameter estimates, however, this loss of information would affect the estimation of F_Z, which has shown less bias than the Turnbull estimate by means of the S-Plus function `qkaplanMeier`.

More simulations could be carried out. On one hand, the results of simultaneous maximization with a broader grid for the support of Z might be of interest. On the other, further parameter settings could be chosen. What if the model constant μ was smaller and the effect of the covariate larger? And what if each of the three employed methods are applied under a wrong parametric assumption regarding the distribution of the response? These aspect have been beyond the scope of the present simulation studies and might be of interest in the future.