On the Impact of Socioeconomic Status on Mortality in the Presence of Interval Censored Status Change

Göran Broström and Joakim Malmdin

March 4, 2009

Abstract

Survival analysis with interval-censored data has been studied extensively in the past, but almost exclusively with interval censored survival. In this paper, we investigate the consequences of interval censoring of status change of a time-varying dichotomous covariate. First an imputation method based on an assumption of a parametric hazard for the time to status change is proposed and evaluated, both in a simulation study and with real data. Then the problem is attacked with the aid of the EM algorithm, and some comparisons are made.

1 Introduction

In survival analysis, notably Cox regression, a common problem is: An explanatory variable can take only a finite number of values, and it changes value over time. The exact time of status change is not recorded, but only a time interval, containing the status change. This paper is concerned with the special case where the covariate is dichotomous, with one state absorbing. Our application in this paper is from historical demography where the covariate socio-economic status (Ses), in many data sources is of secondary interest, and only registered when some other, primary, vital event is occurring. For the timing of status shifts in Ses, thus only an interval where it happened is known.

This problem has been considered mainly in the context of joint modeling of survival and longitudinal data. Danardono (2005) gives a fairly complete review of this literature, for instance Goggins, Finkelstein & Zaslavsky (1999*b*), Goggins, Finkelstein & Zaslavsky (1999*a*), Wulfsohn &

Tsiatis (1997), Henderson, Diggle & Dobson (2000), Lin, Turnbull, McCulloch & Slate (2002), McCulloch, Lin, Slate & Turnbull (2002), Xu & Zeger (2001*b*), Xu & Zeger (2001*a*), Tsiatis, DeGruttola & Wulfsohn (1995), Tsiatis, Boucher & Kim (1995), Rabinowitz, Tsiatis & Aragon (1995), and Pawitan & Self (1993). Bruijne, Cessie, Kluin-Nelemans & Houwelingen (2001) suggested using *time elapsed since the last measurement* (TEL) as a complement to an imputed value.

The problem may also be formulated as a three-state *illness-death* model, which is a very useful, and often used, model in biostatistics. For reference, see the paper by Andersen (1988) and the monograph by Andersen, Borgan, Gill & Keiding (1993). The three-state formulation to our problem is described in Section 2.

The undesirable properties of the traditional approach are demonstrated in Section 3. It is done by simulating a two-sample case, where the survival in the two samples have the same survival probabilities. Then the data is artificially mangled trough an interval censoring mechanism, and, as expected, the estimate of sample difference becomes more and more biased, as the lengths of the censoring intervals increase.

In Section 4 an imputation method is suggested. It is first described in the constant hazard case (for the transition from the `lower` to the `upper` class), then in the general case, but still within the framework of a parametric family of distributions. As an example, we show the calculations for a Weibull family of distributions. The imputation method gives too optimistic standard errors, and we discuss a method to correct them by simulation.

How to solve the problem with the aid of the EM algorithm is explained in Section 5. Then, in Section 6, in a simulation study, we investigate the possible loss in terms of bias, MSE, and coverage probabilities of confidence intervals that the loss of information due to interval censored status changes may result in. First, we investigate the properties of the standard method, which is to assume that the status change occurs at the end of the interval, by simulating data with known times of status change. This data set is then filtered through an interval censoring mechanism. Then we run a Cox regression on both data sets and compare results. This procedure is repeated many times, and bias and coverage probabilities of 95% confidence intervals can thus be estimated and compared for the two situations. Then the imputation method is evaluated by simulation, and it is shown to compare very favourably. However, the standard errors, given by treating the imputed values as real ones, are too optimistic. We use the method from Section 4 to correct for this.

Finally, we use our method on real data from the Demographic Data

Base, Umeå University, Sweden. The setup is survival in the ages 20 to 50, with a time-varying covariate *socio-economic status*, for males in a nineteenth century parish in northern Sweden. See Section 7 for details. The main part of the paper finishes with conclusions in Section 8.

All the numerical analyses were performed in the statistical environment **R** (R Development Core Team 2008). In addition to using the **R** package `eha` (Broström 2007), we wrote a new package `inD` with some utility functions for writing this paper. It also contains some functions, maybe of general interest, related to the Weibull distribution. They are described in Appendix A.

2 The illness-death model

The problem may be formulated as an *illness-death* model, see Figure 1. Two durations are measured, denoted by t and τ. The duration t is simply

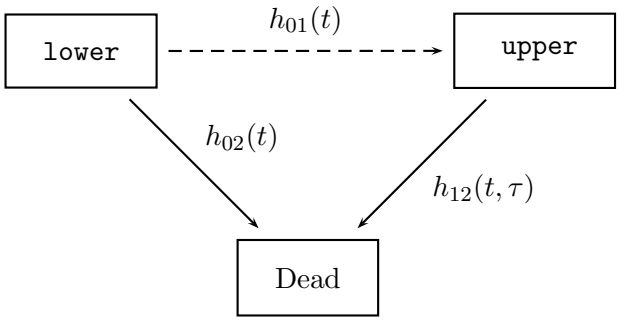

Figure 1: The unidirectional illness-death model applied to social mobility and mortality.

age, while τ is time measured from entering state `upper`. For individuals starting in state `upper`, they are the same, i.e., $t = \tau$.

We are interested in a comparison of h_{02} and h_{12}, i.e., do individuals in the `upper` class have a different mortality compared to those in the `lower` class? We are assuming that h_{12} depends on t only, and not on τ. We are following all individuals from a common start age, until they die (or get censored). Some individuals start in the `lower` class, some in the `upper`. Therefore, the problem cannot be solved directly by methods suggested by Frydman (1995) or Joly, Commenges, Helmer & Letenneur (2002).

3 Properties of the traditional method

The "traditional" method is to use the date when the change first was noticed as the date of change. This will of course overestimate the true date of change, while we in fact have an interval-censoring of the status change.

As an example, consider a two-sample situation where individuals are categorised into one of two possible states, `lower` and `upper`. An individual is allowed to move from `lower` to `upper`, but not the other way around. We are interested in whether mortality in the two states differ. When in reality there is no difference, we may get the results shown in Figure 2 with varying degree of censoring interval length. See also Table 1 for different results of Cox regressions on the same data sets. It is obvious from Table 1 and Figure 2 that too sparse surveillance together with a neglect of the need to view data as interval-censored will ultimately lead to severely biased analyses.

Table 1: Six Cox regressions of a simulated data set with varying degree of interval censoring of status change. True coefficient value is zero.

censoring ivl length	prop. time in `upper`	coef	R.R.	se(coef)	p-value
0	0.628	-0.006	0.994	0.052	0.901
1	0.615	0.046	1.047	0.052	0.379
2	0.603	0.100	1.105	0.052	0.057
3	0.590	0.155	1.167	0.052	0.003
5	0.568	0.261	1.299	0.053	0.000
10	0.514	0.542	1.719	0.053	0.000

The reason for this phenomenon is obvious. What happens when the status change time is censored, in the naive way of treating the problem, is that the average time spent in the `upper` class decreases, while it increases in the `lower class`, and the number of events (deaths) in the two classes remain constant. Therefore, estimated mortality in the `lower` class will successively be lower and lower, as the censoring gets more and more severe, while the opposite will happen in the `upper` class.

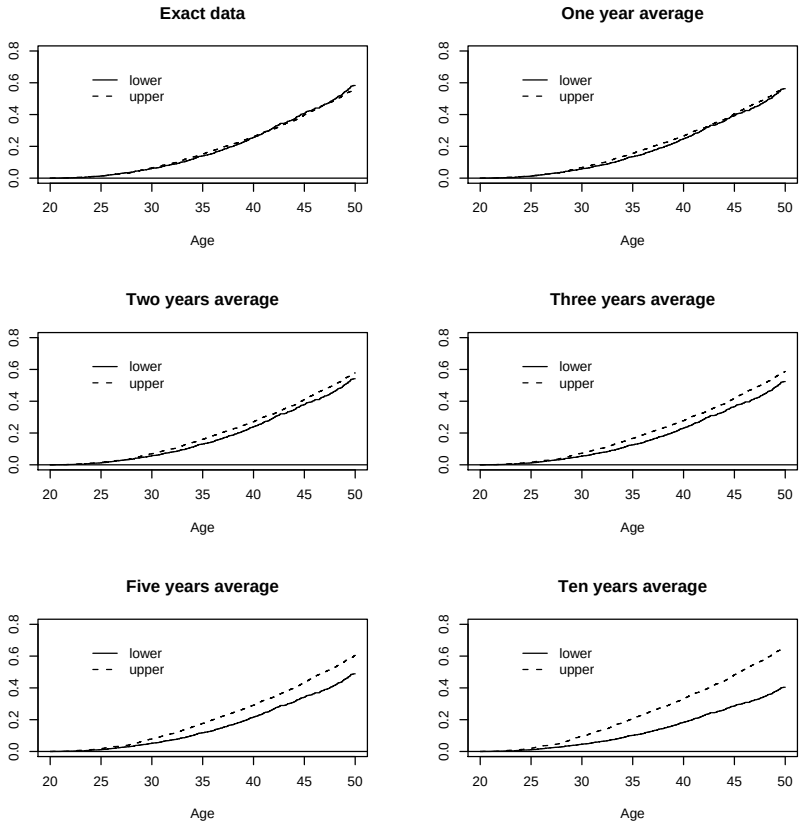

Figure 2: The effect of interval censoring of status change, the two-sample case (simulated data). The effect is increasing with increasing average interval length. Nelson-Aalen plots.

4 Imputation methods

The simplest possible method of imputation given that a status change occurred in a certain time interval is to take the midpoint of the interval. Here, two different approaches are taken. First, we assume that there is a constant intensity r_u of a transition from the lower to the upper class. Second, we assume a general parametric family of distributions. In both cases, parameters are estimated by maximum likelihood from data. Then expected values are imputed.

4.1 Exponential imputation

There are two types of intervals, either there is no transition in the interval, or there is one. In the first case, the contribution to the likelihood function is

$$L_0(r_u;\ell_i) = P(T > \ell_i \mid r_u) = \exp(-r_u\ell_i), \quad i \in S_0,$$

and in the second case

$$L_1(r_u;\ell_j) = P(T \leq \ell_j \mid r_u) = 1 - \exp(-r_u\ell_j), \quad j \in S_1,$$

where ℓ_i is the length of the ith interval, S_0 and S_1 are the sets of intervals with no or one event, respectively. The full likelihood function thus becomes

$$L(r_u;\boldsymbol{\ell}) = \Big\{\prod_{i\in S_0} \exp(-r_u\ell_i)\Big\} \prod_{j\in S_1} \big(1 - \exp(-r_u\ell_j)\big), \tag{1}$$

from which we get, by numerical maximisation, the ML estimate $\hat{r}_u$. Now, the imputed values for the intervals with one event is given by the conditional expectation

$$\begin{aligned} \hat{t}_j &= \frac{\int_0^{\ell_j} x\hat{r}_u \exp(-x\hat{r}_u)dx}{1 - \exp(-\hat{r}_u\ell_j)} \\ &= \frac{1}{\hat{r}_u} - \frac{\ell_j \exp(-\ell_j\hat{r}_u)}{1 - \exp(-\ell_j\hat{r}_u)}, \quad j \in S_1. \end{aligned} \tag{2}$$

4.2 General distribution imputation

We now introduce a general distribution for the time to status change. Since we no longer can utilise the exponential distribution property of lack of memory, we need to introduce the start age t_i of an interval together with its length ℓ_i, $i = 1, \ldots, n$. Thus the contributions to the likelihood become

$$L_0\big(\boldsymbol{\theta};(t_i,\ell_i)\big) = P_{\boldsymbol{\theta}}(T > t_i + \ell_i \mid T_i \geq t_i) = \frac{S(t_i+\ell_i;\boldsymbol{\theta})}{S(t_i;\boldsymbol{\theta})}$$

in the no-transition case, and

$$L_1(\boldsymbol{\theta};(t_i,\ell_i)) = P_{\boldsymbol{\theta}}(T \leq t_i + \ell_i \mid T_i \geq t_i) = 1 - \frac{S(t_i+\ell_i;\boldsymbol{\theta})}{S(t_i;\boldsymbol{\theta})}$$

in the one-transition case. In analogy with (1), we get

$$L\{\boldsymbol{\theta};(\boldsymbol{t},\boldsymbol{\ell})\} = \Big\{\prod_{i\in S_0} \frac{S(t_i+\ell_i;\boldsymbol{\theta})}{S(t_i;\boldsymbol{\theta})}\Big\} \prod_{j\in S_1} \Big\{1 - \frac{S(t_j+\ell_j;\boldsymbol{\theta})}{S(t_j;\boldsymbol{\theta})}\Big\} \tag{3}$$

In the Weibull case, we have $\boldsymbol{\theta} = (p, \lambda)$, and

$$S(t; (p, \lambda)) = \exp\left\{-\left(\frac{t}{\lambda}\right)^p\right\}, \quad t > 0, \tag{4}$$

and an application of the imputing method with the Weibull distribution requires first order partial derivatives of the Weibull survivor function S, corresponding to the hazard function given by (17). It is convenient to reparametrise according to

$$\begin{aligned} \gamma &= \log(p), \\ \alpha &= \log(\lambda), \end{aligned}$$

or

$$\begin{aligned} p &= e^{\gamma}, \\ \lambda &= e^{\alpha}, \end{aligned} \tag{5}$$

which leads to the following expressions for the log survivor function and its first order partial derivatives, for $t > 0$:

$$\begin{aligned} \log S(t; (e^{\gamma}, e^{\alpha})) &= -\left(\frac{t}{\exp(\alpha)}\right)^{\exp(\gamma)}, \\ \frac{\partial}{\partial \gamma} \log S(t; (e^{\gamma}, e^{\alpha})) &= p \log\left(\frac{t}{\lambda}\right) \log S(t; (p, \lambda)), \\ \frac{\partial}{\partial \alpha} \log S(t; (e^{\gamma}, e^{\alpha})) &= -\lambda \log S(t; (p, \lambda)). \end{aligned} \tag{6}$$

Now, the relations

$$\frac{\partial}{\partial \theta} S(t; (p, \lambda)) = S(t; (p, \lambda)) \frac{\partial}{\partial \theta} \log S(t; (p, \lambda)), \quad \theta = \gamma, \alpha$$

together with (6) are all we need in order to estimate the parameters in the Weibull model (3) with a quasi-newton method.

4.3 Variance estimation

The imputation method may give too small variance estimates, because a variance component is removed by imputing an expected value instead of the corresponding random variable. One way to correct for that is to impute a random draw from the estimated conditional distribution of the time to transition, i.e., to draw random numbers from truncated versions of the estimated distribution. If this procedure is repeated n times, the result is n estimates of the regression coefficient, and the sample variance of these is added to the variance given by Cox regression procedure.

5 The EM algorithm

5.1 The likelihood function

Given data $(s_i, T_i, u_i, c_i, d_i, x_i, \mathbf{z}_i),\ i = 1, \ldots, n$, where s_i is a left truncation time point, u_i is a failure or right censoring time point, c_i is a status change indicator, d_i is an event indicator, x_i is the status indicator, T_i is the potential ($c_i = 1$) time point for status change in (s_i, u_i), and $\mathbf{z}_i$ is a vector of covariates.

Suppose that $T_i, i : c_i = 1$ is fully observed. Then the full likelihood function is

$$\begin{aligned} L(\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\phi}) = \prod_{i=1}^{n} \Bigg\{ \big(h_{\boldsymbol{\gamma}}(u_i) e^{x_i\alpha + \mathbf{z}_i\boldsymbol{\beta}}\big)^{d_i} \left(\frac{S_{\boldsymbol{\gamma}}(u_i)}{S_{\boldsymbol{\gamma}}(s_i)}\right)^{e^{x_i\alpha + \mathbf{z}_i\boldsymbol{\beta}}} \\ \times \left(\frac{S_{\boldsymbol{\gamma}}(u_i)}{S_{\boldsymbol{\gamma}}(T_i)}\right)^{c_i e^{\alpha}} R_{\boldsymbol{\phi}}(s_i, u_i, T_i, c_i) \Bigg\} \end{aligned} \tag{7}$$

where $h_{\boldsymbol{\gamma}}$ is the hazard function of the survival distribution of survival and g_ϕ is the density of time to status change. The function $R_{\boldsymbol{\phi}}$ is the contribution to the likelihood function from interval i regarding the process of status change. Note that each individual possibly is represented by several intervals. Either $c_i = 1$ for exactly one of the individual's intervals, in which case we have an interval-censored observation, or $c_i = 0$ for all intervals, in which case the status change either never happens or happens after the time of last seen, which will constitute a right censored observation of status change.

The full log likelihood function is, with $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\phi})$,

$$\begin{aligned} \ell(\boldsymbol{\theta}) = & \sum_{i:d_i=1} \big(x_i\alpha + \mathbf{z}_i\boldsymbol{\beta} + \log h_{\boldsymbol{\gamma}}(u_i)\big) \\ & - \sum_{i=1}^{n} e^{x_i\alpha + \mathbf{z}_i\boldsymbol{\beta}} \big(H_{\boldsymbol{\gamma}}(u_i) - H_{\boldsymbol{\gamma}}(s_i)\big) \\ & - \sum_{i:c_i=1} e^{\alpha} \big(H_{\boldsymbol{\gamma}}(u_i) - H_{\boldsymbol{\gamma}}(T_i)\big) + \sum_{i=1}^{n} \log R_{\boldsymbol{\phi}}(s_i, u_i, T_i, c_i) \end{aligned} \tag{8}$$

Here $H_{\boldsymbol{\phi}}(x) = -\log S_{\boldsymbol{\phi}}(x), x > 0$ is the cumulative hazard function of survival.

5.2 Implementing the EM algorithm

The EM algorithm consists of two steps. In the first, the E step, the conditional expected value of (8) with respect to available information and given parameter vector $\boldsymbol{\theta} = \boldsymbol{\theta}^{(j)}$ is calculated. Then, in the M step, the calculated expectation is maximized with respect to $\boldsymbol{\theta}$. However, from (8) it is obvious that the updating of $\boldsymbol{\phi}$ will run unaffected by the updating of the rest of the parameters. Therefore, it is possible to proceed in two stages. In the first stage, $\boldsymbol{\phi}$ is estimated, and in the second stage, the EM algorithm is run on

$$\begin{aligned}\ell_r(\boldsymbol{\theta}) = &\sum_{i:d_i=1} \big(x_i\alpha + \mathbf{z}_i\boldsymbol{\beta} + \log h_{\boldsymbol{\gamma}}(u_i)\big) \\ &- \sum_{i=1}^{n} e^{x_i\alpha+\mathbf{z}_i\boldsymbol{\beta}}\big(H_{\boldsymbol{\gamma}}(u_i) - H_{\boldsymbol{\gamma}}(s_i)\big) \\ &- \sum_{i:c_i=1} e^{\alpha}\big(H_{\boldsymbol{\gamma}}(u_i) - H_{\boldsymbol{\gamma}}(T_i)\big)\end{aligned} \tag{9}$$

To begin with, we assume that the fist phase is carried out, giving $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}$.

5.2.1 The E step

The last sum in (9) contains the unobservables, $T_i, i : d_i = 1$. For each such record, we have to calculate the expectation

$$\begin{aligned}&E_{\boldsymbol{\theta}^{(j)}}\big\{e^{\alpha}\big(H_{\boldsymbol{\gamma}}(u_i) - H_{\boldsymbol{\gamma}}(T_i)\big) \mid s_i < T_i < u_i\big\} \\ &\qquad = e^{\alpha}\big\{H_{\boldsymbol{\gamma}}(u_i) - E_{\boldsymbol{\theta}^{(j)}}\big(H_{\boldsymbol{\gamma}}(T_i) \mid s_i < T_i < u_i\big)\big\} \\ &\qquad\qquad = e^{\alpha}\big\{H_{\boldsymbol{\gamma}}(u_i) - E_{\hat{\boldsymbol{\phi}}}\big(H_{\boldsymbol{\gamma}}(T_i) \mid s_i < T_i < u_i\big)\big\}\end{aligned} \tag{10}$$

5.2.2 The M step

In iteration $(j + 1)$ the M step consists of maximizing

$$\begin{aligned}E_{\boldsymbol{\theta}^{(j)}}\big(\ell(\boldsymbol{\theta})\big) = &\sum_{i:d_i=1} \big(x_i\alpha + \mathbf{z}_i\boldsymbol{\beta} + \log h_{\boldsymbol{\gamma}}(u_i)\big) \\ &- \sum_{i=1}^{n} e^{x_i\alpha+\mathbf{z}_i\boldsymbol{\beta}}\big(H_{\boldsymbol{\gamma}}(u_i) - H_{\boldsymbol{\gamma}}(s_i)\big) \\ &- \sum_{i:c_i=1} e^{\alpha}\big\{H_{\boldsymbol{\gamma}}(u_i) - E_{\hat{\boldsymbol{\phi}}}\big(H_{\boldsymbol{\gamma}}(T_i) \mid s_i < T_i < u_i\big)\big\}\end{aligned} \tag{11}$$

with respect to $\boldsymbol{\theta}$, which gives $\boldsymbol{\theta}^{(j+1)}$. From this it is obvious that the EM algorithm converges in one step. The remaining problem is to calculate the conditional expectation in the last sum in (11). One option is to do it by numerical integration.

In order to be able to use an efficient Newton procedure we need the score vector of (11).

$$\begin{aligned}\frac{\partial}{\partial\alpha}E_{\boldsymbol{\theta}^{(j)}}\big(\ell(\boldsymbol{\theta})\big) = &\sum_{i:d_i=1} x_i \\ &- \sum_{i=1}^{n} x_i e^{x_i\alpha+\mathbf{z}_i\boldsymbol{\beta}}\big(H_{\boldsymbol{\gamma}}(u_i) - H_{\boldsymbol{\gamma}}(s_i)\big) \\ &- \sum_{i:c_i=1} e^{\alpha}\big\{H_{\boldsymbol{\gamma}}(u_i) - E_{\hat{\boldsymbol{\phi}}}\big(H_{\boldsymbol{\gamma}}(T_i) \mid s_i < T_i < u_i\big)\big\},\end{aligned} \tag{12}$$

$$\begin{aligned}\frac{\partial}{\partial\beta_j}E_{\boldsymbol{\theta}^{(j)}}\big(\ell(\boldsymbol{\theta})\big) = &\sum_{i:d_i=1} z_{ij} \\ &- \sum_{i=1}^{n} z_{ij} e^{x_i\alpha+\mathbf{z}_i\boldsymbol{\beta}}\big(H_{\boldsymbol{\gamma}}(u_i) - H_{\boldsymbol{\gamma}}(s_i)\big), \quad j = 1,\ldots,p,\end{aligned} \tag{13}$$

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\gamma}}E_{\boldsymbol{\theta}^{(j)}}\big(\ell(\boldsymbol{\theta})\big) = &\sum_{i:d_i=1} \frac{\frac{\partial}{\partial\boldsymbol{\gamma}}h_{\boldsymbol{\gamma}}(u_i)}{h_{\boldsymbol{\gamma}}(u_i)} \\ &- \sum_{i=1}^{n} e^{x_i\alpha+\mathbf{z}_i\boldsymbol{\beta}}\frac{\partial}{\partial\boldsymbol{\gamma}}\big(H_{\boldsymbol{\gamma}}(u_i) - H_{\boldsymbol{\gamma}}(s_i)\big) \\ &- \sum_{i:c_i=1} e^{\alpha}\frac{\partial}{\partial\boldsymbol{\gamma}}\big\{H_{\boldsymbol{\gamma}}(u_i) - E_{\hat{\boldsymbol{\phi}}}\big(H_{\boldsymbol{\gamma}}(T_i) \mid s_i < T_i < u_i\big)\big\}\end{aligned} \tag{14}$$

As an example, and the application in this paper, let the survival distribution be Weibull. Then, for $t > 0$,

$$\begin{aligned}H_{\boldsymbol{\gamma}}(t) &= \left(\frac{t}{e^{\gamma_2}}\right)^{e^{\gamma_1}} \\ h_{\boldsymbol{\gamma}}(t) &= e^{\gamma_1-\gamma_2}\left(\frac{t}{e^{\gamma_2}}\right)^{e^{\gamma_1}-1},\end{aligned} \tag{15}$$

and the partial derivatives are, for $t > 0$

$$\begin{aligned}
\frac{\partial}{\partial\gamma_1} H_{\boldsymbol{\gamma}}(t) &= H_{\boldsymbol{\gamma}}(t) \log H_{\boldsymbol{\gamma}}(t) \\
\frac{\partial}{\partial\gamma_2} H_{\boldsymbol{\gamma}}(t) &= -e^{\gamma_1} H_{\boldsymbol{\gamma}}(t) \\
\frac{\partial}{\partial\gamma_1} h_{\boldsymbol{\gamma}}(t) &= h_{\boldsymbol{\gamma}}(t) + \frac{e^{\gamma_1}}{t} H_{\boldsymbol{\gamma}}(t) \log H_{\boldsymbol{\gamma}}(t) \\
\frac{\partial}{\partial\gamma_2} h_{\boldsymbol{\gamma}}(t) &= -e^{\gamma_1} h_{\boldsymbol{\gamma}}(t)
\end{aligned} \tag{16}$$

6 A simulation study

6.1 Layout

In order to quantify the bias, we perform a simulation study in the two-sample situation, where individuals are allowed to move from sample `lower` to sample `upper`, but not the other way around. The layout of the simulation study is as follows.

The initial conditions are:

1. We want to study the effect of Ses a dichotomous covariate, on mortality in a certain age interval $(a, b]$, common to all individuals.
2. We assume that Ses takes the two values `lower` and `upper`, and that
   (a) at most one transition occurs for each individual,
   (b) only *upward* transitions are allowed,
   (c) the intensity of dying changes by a factor $\gamma = \exp(\beta)$ at the age of a transition, and
   (d) With probability p, each individual is born in the `upper` class, independently of all the other individuals.
3. Each individual is "peeked at" at regular time points in calendar time with constant period. It is only at these "peeking" ages that Ses is directly observed. All individual are "peeked at" at death, i.e. the true value of Ses is known at the death age.
4. Birth dates follow a Poisson process with constant intensity between given dates. This assumption is not important, but present to ensure a random distribution of peeking ages over individuals.

There are two important objectives in the simulation layout: The simulation of *exact* data, including ages of change in SES and the creation of "peeked" data, from exact data.

Survival times are drawn from the Weibull proportional hazards model

$$h(t;\lambda,p,\boldsymbol{\beta}) = \frac{p}{\lambda}\left(\frac{t}{\lambda}\right)^{p-1}\exp(\beta x(t)), \quad 20 < t \leq 50, \tag{17}$$

where $x(t)$ is an indicator function:

$$x(t) = \begin{cases} 0 & \text{if in } \texttt{lower} \text{ SES at t} \\ 1 & \text{if in } \texttt{upper} \text{ SES at t} \end{cases}, \quad 20 < t \leq 50.$$

We assume, without loss of generality, that throughout all simulations we have

$$\begin{aligned} p &= 2 \\ \lambda &= 50 \\ p_u &= 0.2 \\ r_u &= 0.02 \end{aligned}$$

where p_u is the probability of starting in `upper` at age 20, and r_u is the intensity of moving from `lower` to `upper`. The regression parameter β is taking the values $0, \pm 0.5, \pm 1$ in the simulations. The periods at which peeking is done is taken as 5, 10, 20. Studied sample sizes are $n = 50, 100$.

6.2 Results

First, we investigate the effect of increasing severity of the interval censoring, i.e., the effect of different lengths of time between consecutive observations.

As can be seen in Figure 3, with sparse surveillance comes bias in parameter estimates and too low coverage probabilities for confidence intervals calculated by standard asymptotic methods. This is of course exactly as expected.

For comparing the imputation method to the exact and naive methods, the simulation is carried out in exactly the same way as in the previous case. The situation with peeking every tenth year is considered, and the three situations with continuous, peeked, and imputed information is compared. As can be seen in Figure 4, the imputation performs very well, almost as well as if full information was available.

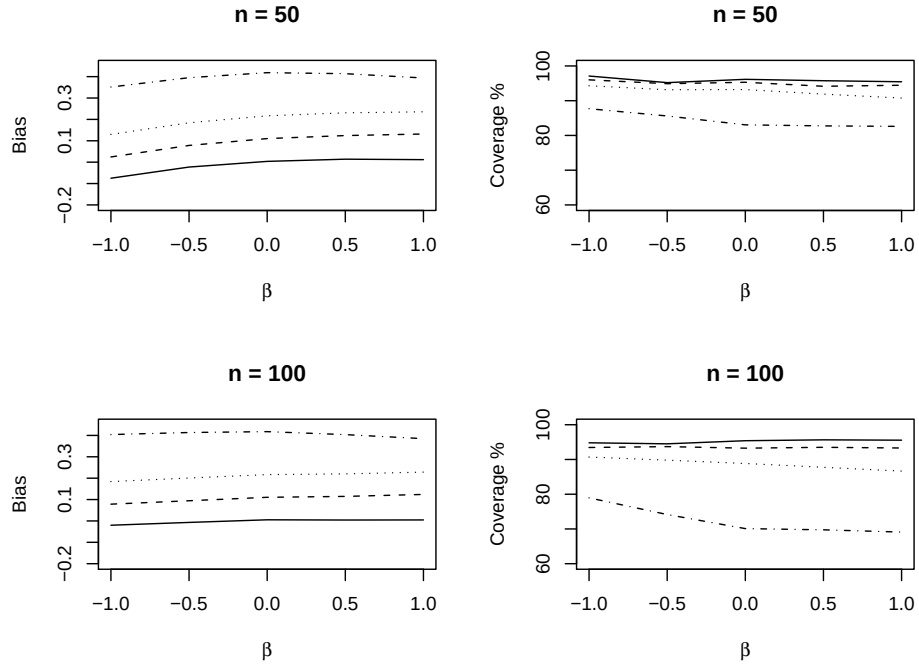

Figure 3: Bias (left panels) and coverage (right panels), nominal confidence is 95%) for sample sizes 50 and 100. Continuous surveillance (solid lines), every fifth year (dashed), every tenth year (dotted), and every twentieth year (dashed-dotted).

7 Real data: Socio-Economic Status

A data set from the Demographic Data Base, Umeå University, contains survival data from the Skellefteå region in northern Sweden for the years 1840–1870. The time-varying dichotomous covariate Ses (Socio-Economic Status) is of special interest as a factor determining mortality in the ages 20–50.

7.1 Weibull regression with peeked data

Results in

```
Covariate          Coef Exp(Coef)  se(Coef)    Wald p
ses, upper       -0.146     0.864     0.145     0.313
```

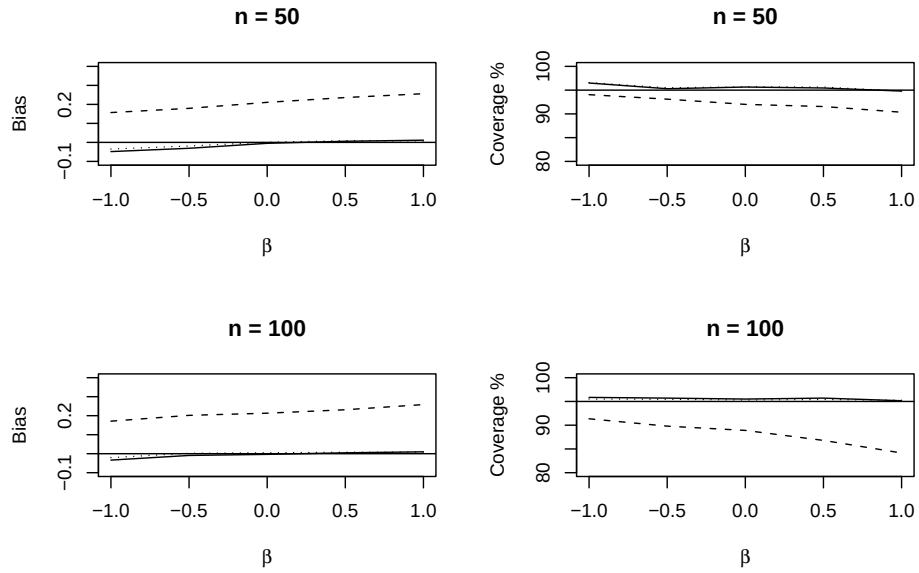

Figure 4: Bias (left panels) and coverage (right panels), nominal confidence is 95%) for sample sizes 50 and 100. Continuous surveillance (solid), every tenth year (dashed), and imputed (dotted).

```
log(scale)          4.561    95.723     0.098     0.000
log(shape)          0.251     1.285     0.063     0.000

Events                    255
Total time at risk      35783
Max. log. likelihood   -1508.1
```

7.2 Weibull regression with imputed data

Imputation was performed by assuming a Weibull distribution for time to promotion, with the result that the parameter ϕ was estimated by $\hat{\phi} = (0.3435, 2.7334)$. The result from the survival analysis:

```
Covariate           Coef Exp(Coef)  se(Coef)    Wald p
ses               -0.788     0.455     0.135     0.000
```

```
log(scale)        4.383    80.118     0.080     0.000
log(shape)        0.334     1.396     0.058     0.000

Events                     255
Total time at risk       35978
Max. log. likelihood    -1493.3
```

7.3 Weibull regression with the EM algorithm

Results:

```
Covariate          Coef Exp(Coef)   se(Coef)    Wald p
ses              -0.402     0.728      0.136    0.0007

log(scale)        4.483    88.500      0.098    0.0000
log(shape)        0.292     1.339      0.061    0.0000

Events                     255
Total time at risk       35978
Max. log. likelihood    -1524.6
```

The analysis was performed in both the traditional way and using two methods of correction, imputation and the EM algorithm. Only upward changes were considered. From Figure 5 it is obvious that the effect of not correcting for interval censoring may be dramatic.

8 Conclusion

The imputation method performs very well in our simulation example. One reason to be careful with conclusions, though, is that the distributional assumption under which the imputations were estimated is exactly the same as was used in the simulation.

From the real data example, it is obvious that different methods, like the EM algorithm and the imputation methods, can give rather different results. More research is needed in order to shed light on this issue.

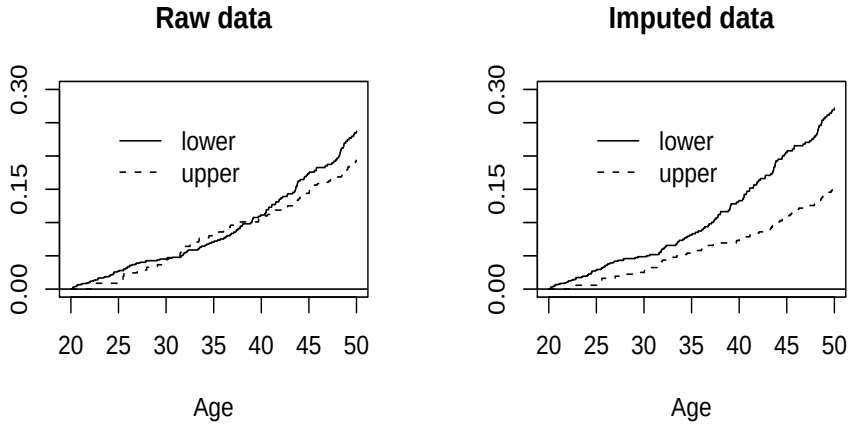

Figure 5: Cumulative hazards functions for raw and imputed data and for lower and upper social class.

References

Andersen, P., Borgan, Ø., Gill, R. & Keiding, N. (1993), *Statistical Models Based on Counting Processes*, Springer, New York.

Andersen, P. K. (1988), 'Multistate models in survival analysis: a study of nephropathy and mortality in diabetes', *Statistics in Medicine* **7**, 661–670.

Broström, G. (2007), *eha: Event History Analysis*. R package version 1.0. *http://www.stat.umu.se/∼goran.brostrom/eha

Bruijne, M. H. J. d., Cessie, S. l., Kluin-Nelemans, H. C. & Houwelingen, H. C. v. (2001), 'On the use of Cox regression in the presence of an irregularly observed time-dependent covariate', *Statistics in Medicine* **20**(24), 3817–3829.

Danardono (2005), Multipe time scales and longitudinal measurements in event history analysis, PhD thesis, Department of Statistics, Umeå University, Umeå, Sweden.

Frydman, H. (1995), 'Nonparametric estimation of a Markov 'illness-death' process from interval-censored observations, with application to diabetes survival data', *Biometrika* **82**, 773–789.

Goggins, W., Finkelstein, D. & Zaslavsky, A. (1999*a*), 'Applying the Cox proportional hazards model for analysis of latency data with interval censoring', *Statistics in Medicine* **18**, 2737–2748.

Goggins, W., Finkelstein, D. & Zaslavsky, A. (1999*b*), 'Applying the Cox proportional hazards model when the change time of a time-varying covariate is interval-censored', *Biometrics* **55**, 445–451.

Henderson, R., Diggle, P. & Dobson, A. (2000), 'Joint modelling of longitudinal measurements and event time data', *Biostatistics* **1**(4), 465–480.

Joly, P., Commenges, D., Helmer, C. & Letenneur, L. (2002), 'A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia', *Biostatistics* **3**, 433–443.

Lin, H., Turnbull, B. W., McCulloch, C. E. & Slate, E. H. (2002), 'Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer', *Journal of the American Statistical Association* **97**(457), 53–65.

McCulloch, C. E., Lin, H., Slate, E. H. & Turnbull, B. W. (2002), 'Discovering subpopulation structure with latent class mixed models', *Statistics in Medicine* **21**(3), 417–429.

Pawitan, Y. & Self, S. (1993), 'Modeling disease marker processes in AIDS', *Journal of the American Statistical Association* **88**, 719–726.

R Development Core Team (2008), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*http://www.R-project.org

Rabinowitz, D., Tsiatis, A. & Aragon, J. (1995), 'Regression with interval-censored data', *Biometrika* **82**, 501–513.

Tsiatis, A. A., Boucher, H. & Kim, K. (1995), 'Sequential methods for parametric survival models', *Biometrika* **82**, 165–173.

Tsiatis, A. A., DeGruttola, V. & Wulfsohn, M. S. (1995), 'Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS', *Journal of the American Statistical Association* **90**, 27–37.

Wulfsohn, M. S. & Tsiatis, A. A. (1997), 'A joint model for survival and longitudinal data measured with error', *Biometrics* **53**, 330–339.

Xu, J. & Zeger, S. L. (2001*a*), 'The evaluation of multiple surrogate endpoints', *Biometrics* **57**(1), 81–87.

Xu, J. & Zeger, S. L. (2001*b*), 'Joint analysis of longitudinal data comprising repeated measures and times to events', *Applied Statistics* **50**(3), 375–387.

A An R package for simulation and analysis of interval censored status change

As a preparation for writing this paper, an R package for simulation and analysis of interval censored status change was written. We call it `inD`, and the current (February 2008) version is 0.8. It contains the following functions.

pcweibull Calculates a conditional Weibull cumulative distribution function, given survival up to a certain age.

pjweibull The cdf for a jump Weibull distribution with constant shape parameter, but with a jump in the scale parameter at a certain age.

rcweibull Generates random numbers from a conditional Weibull distribution, see `pcweibull`.

rjweibull Generates random numbers from a jump Weibull distribution, see `pjweibull`.

simData Simulates survival data with a time-dependent covariate.

peekData From a data frame with exact date for status change, this function creates a data frame with interval censored status change.

impute For a data frame with interval censored status change, it imputes an exact value for the status change and outputs a data frame accordingly.