Published in final edited form as:
Statistics in Medicine (2013), **32**(6): 1004–1015
DOI: 10.1002/sim.5581

Regression with incomplete covariates and left-truncated time-to-event data

HUA SHEN

Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada

RICHARD J. COOK

Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada
E-mail: rjcook@uwaterloo.ca

Summary

Studies of chronic diseases routinely sample individuals subject to conditions on an event time of interest. In epidemiology, for example, prevalent cohort studies aiming to evaluate risk factors for survival following onset of dementia require subjects to have survived to the point of screening. In clinical trials designed to assess the effect of experimental cancer treatments on survival, patients are required to survive from the time of cancer diagnosis to recruitment. Such conditions yield samples featuring left-truncated event time distributions. Incomplete covariate data often arise in such settings, but standard methods do not deal with the fact that individuals' covariate distributions are also affected by left truncation. We describe an expectation-maximization algorithm for dealing with incomplete covariate data in such settings, which uses the covariate distribution conditional on the selection criterion. We describe an extension to deal with subgroup analyses in clinical trials for the case in which the stratification variable is incompletely observed.

Keywords: Incomplete covariates, left-truncation, subgroup analysis, survival analysis

1 INTRODUCTION

Studies of chronic diseases routinely sample individuals subject to specified conditions on an event time of interest. In epidemiology, for example, prevalent cohort studies may aim to evaluate risk factors for death following onset of dementia. Such designs require subjects to have survived from the date of disease onset to the date of the screening assessment [1]. In clinical research, randomized trials are often designed to assess the effect of experimental cancer treatments on survival, and patients

must survive from the time of cancer diagnosis to contact to be recruited; there may be additional conditions imposed on the times of nonfatal events related to the disease process [2]. When the date of disease onset is to be used as the time origin for survival analyses, samples chosen this way feature left truncation, and standard methods of survival analysis can be readily adapted to deal with this feature [3–7].

Incomplete covariate data often arise in studies with time-to-event outcomes [8]. This may be a consequence of the study protocol if resources are limited and a particular subset of individuals are identified for detailed examination of biomarkers, for example. In other cases, it may be due to chance (e.g., noncompliance of study investigators or participants). There is a large literature on the various frameworks and methods for fitting regression models to survival data with incomplete covariate information. Lipsitz and Ibrahim [9], Chen and Little [10] and Herring et al. [11], among others, developed methods based on the expectation-maximization (EM) algorithm. Lipsitz and Ibrahim [12] provided estimating function approaches incorporating inverse probability weights, and Wang and Chen [13] developed augmented estimating equations yielding more efficient estimation. Ibrahim et al. [14] and Bradshaw et al. [15] developed Bayesian approaches for this same problem, and Chen and Little [16] considered an interesting alternative approach for dealing with missing covariates in the context of linear transformation models. These methods do not deal with the setting where individuals are only sampled if they satisfy some response-dependent selection criterion (e.g., truncation). In this setting, the sample covariate distributions are different from the population covariate distribution as a result of selection effects, and in fact, different individuals will have different sample covariate distributions if they have different selection criteria [17–19]. The purpose of this article is to consider this problem and propose a simple strategy for dealing with it.

We describe an EM algorithm [20] for dealing with incomplete discrete covariate data. The algorithm involves the conceptualization of a complete data set, which includes information on both the missing covariates and the number of unsampled individuals in the population who did not satisfy the truncation condition [21]. The maximization step is shown to be easily implemented using standard survival analysis software provided it can accommodate left-censored data. We then develop a generalization of this algorithm for subgroup analyses in clinical trials where information on the stratification variables is missing. We use an application to data from a recently completed trial of patients with metastatic cancer for illustration.

We organize the remainder of this paper as follows. In Section 2, we define notation, give the complete data likelihood, and describe how to carry out the maximization step of the EM algorithm using standard software. We then assess the empirical performance of estimators arising from a complete case analysis, a misspecified likelihood that uses the population rather than the appropriate sample covariate distribution, and the proposed method. We describe extensions to facilitate robust estimation using piecewise-constant baseline hazards in Supplementary Material. We develop the extension dealing with the case of a missing stratification variable to be used in a secondary subgroup analyses in Section 3 and provide an illustrative application in Section 4. We provide concluding remarks and topics for further research in Section 5.

2 NOTATION AND STATEMENT OF THE PROBLEM

2.1 THE OBSERVED DATA LIKELIHOOD

We consider first a cohort study in which a sample of m individuals is obtained by randomly sampling from a population of diseased individuals. Let A denote the calendar time at which subjects are accrued and B denote the calendar time of the end of the study; the duration of the study is then $C = B - A$. Let D_i denote the calendar time of disease onset and E_i denote the calendar time of the event, say death, for individual i; then $T_i = E_i - D_i$ is the corresponding survival time from

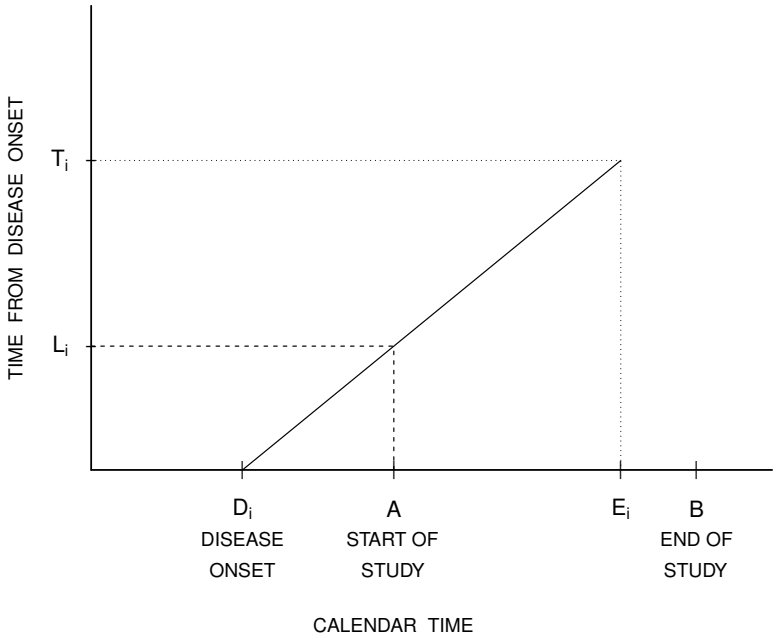

Figure 1: Lexis diagram of calendar event times and left-truncated failure time data.

disease onset. To be included in the study, it is necessary that $T_i > L_i = A - D_i$, and so the survival time of a recruited individual is left-truncated at L_i (Figure 1). If $C_i^\dagger$ ($A < C_i^\dagger < B$) is a random calendar time at which an individual is lost to follow-up, let $C_i = \min(B, C_i^\dagger) - D_i$ denote the censoring time measured from disease onset, $X_i = \min(T_i, C_i)$ denote the observation time, and $\delta_i = I(X_i = T_i)$ indicate whether individual i is observed to die. Consider a proportional hazards model $h(s|Z_i;\theta) = h_0(s;\alpha)\exp(Z_i'\beta)$ specified to assess the effect of a covariate vector Z_i on the survival time, where $h_0(s;\alpha)$ is the baseline hazard function indexed by α, β is a vector of regression coefficients, and $\theta = (\alpha', \beta')'$. Let $H_0(s,t;\alpha) = \int_s^t h_0(u;\alpha)du$, $H(s,t|Z_i;\theta) = \int_s^t h(u|Z_i;\theta)du$, and we denote $H_0(0,t;\alpha)$ and $H(0,t|Z_i;\theta)$ by $H_0(t;\alpha)$ and $H(t|Z_i;\theta)$, respectively. We assume $Z_i \perp D_i$ and $T_i \perp (D_i, C_i^\dagger)|Z_i$, and so the process is stationary and censoring is conditionally independent.

Suppose a sample of m individuals is recruited at the start of the study. For illustration we suppose that the covariate vector is of the form $Z_i = (Z_{i1}, Z_{i2})'$ and contains risk factors at the time of diagnosis, where Z_{i1} is a binary covariate that is not observed for all individuals and Z_{i2} is another binary covariate that is always observed, $i = 1, \ldots, m$; extensions to handle other types of categorical covariates are straightforward. Let $R_i = I(Z_{i1} \text{ is observed})$, $\mathcal{R} = \{i : R_i = 1\}$, and $\bar{\mathcal{R}} = \{i : R_i = 0\}$. The conditional probability mass function for Z_{i1} given Z_{i2} is $P(Z_{i1}|Z_{i2};\eta)$ where $\text{logit}P(Z_{i1} = 1|Z_{i2}) = \eta_0 + \eta_1 Z_{i2}$, $\eta = (\eta_0, \eta_1)'$ and $\psi = (\theta', \eta')'$. We assume that Z_{i1} is missing completely at random according to $P(R_i = 1|D_i, Z_i, T_i, C_i) = P(R_i = 1|Z_{i2})$, where this model does not share any parameters with ψ and hence missingness is non-informative.

In the absence of left truncation (i.e. if $L_i = 0$, $i = 1, \ldots, m$), the observed data likelihood is

$$\begin{aligned} L(\psi) &= \prod_{i\in\mathcal{R}} \left\{ h^{\delta_i}(X_i|Z_i;\theta)\exp\left(-H(X_i|Z_i;\theta)\right) P(Z_{i1}|Z_{i2};\eta) \right\} \\ &\times \prod_{i\in\bar{\mathcal{R}}} \left\{ \sum_{Z_{i1}} h^{\delta_i}(X_i|Z_i;\theta)\exp\left(-H(X_i|Z_i;\theta)\right) P(Z_{i1}|Z_{i2};\eta) \right\} . \end{aligned} \tag{2.1}$$

When a sample features left truncation, the correct probability mass function for the covariate vector

of individual i is $P(Z_i|T_i > L_i;\psi)$, so the likelihood in this setting is

$$\begin{aligned} L(\psi) &= \prod_{i\in\mathcal{R}} \left\{h^{\delta_i}(X_i|Z_i;\theta)\exp\left(-H(L_i,X_i|Z_i;\theta)\right)P(Z_{i1}|Z_{i2},T_i>L_i;\psi)\right\} \\ &\times \prod_{i\in\bar{\mathcal{R}}}\left\{\sum_{Z_{i1}} h^{\delta_i}(X_i|Z_i;\theta)\exp\left(-H(L_i,X_i|Z_i;\theta)\right)P(Z_{i1}|Z_{i2},T_i>L_i;\psi)\right\}, \end{aligned} \tag{2.2}$$

where

$$P(Z_{i1}|Z_{i2},T_i>L_i;\psi) = \frac{P(Z_{i1}|Z_{i2};\eta)\exp\left(-H(L_i|Z_i;\theta)\right)}{\sum_{Z_{i1}} P(Z_{i1}|Z_{i2};\eta)\exp\left(-H(L_i|Z_i;\theta)\right)}. \tag{2.3}$$

The likelihood (2.2) can be maximized directly, but this can be challenging if the dimension of ψ is high. An EM algorithm can alternatively be used with a complete data likelihood analogous to (2.2) where missing covariate values are part of the complete data. The maximization step of such an algorithm, however, would require optimizing a complicated function of ψ because one cannot factor the complete data likelihood to isolate the components θ and η; see (2.3). We propose a computationally more appealing complete data likelihood by incorporating contributions associated with individuals not selected for inclusion in the sample.

2.2 A Turnbull-type Complete Data Likelihood

Corresponding to individual i in the sample with left truncation time L_i, one can conceptualize J_i individuals who are identical in all respects (i.e. with the same covariate vector and disease onset time as individual i), except they did not remain event free (alive) long enough to qualify for inclusion in the sample. Turnbull [21] used the evocative term "ghosts" to refer to such individuals, and we consider a complete data likelihood that includes those individuals. All that is known about these individuals, however, is that their respective survival times are less than L_i, and hence their survival times are left-censored at L_i. The complete data likelihood incorporating these ghosts can be written as follows:

$$L_C(\psi) = L_{C1}(\theta)\cdot L_{C2}(\eta),$$

where

$$\begin{aligned} L_{C1}(\theta) \propto\ & \prod_{i\in\mathcal{R}}\left\{h^{\delta_i}(X_i|Z_i)\,\mathcal{F}(X_i|Z_i)\,\left[F(L_i|Z_i)\right]^{J_i}\right\}\cdot \\ & \prod_{i\in\bar{\mathcal{R}}}\left\{\prod_{z_1=0}^{1}\left\{h^{\delta_i}(X_i|(z_1,Z_{i2}))\,\mathcal{F}(X_i|(z_1,Z_{i2}))\,\left[F(L_i|(z_1,Z_{i2}))\right]^{J_i}\right\}^{I(Z_{i1}=z_1)}\right\}, \end{aligned}$$

and

$$L_{C2}(\eta)\propto\prod_{i\in\mathcal{R}}P(Z_{i1}|Z_{i2})^{J_i+1}\prod_{i\in\bar{\mathcal{R}}}\left\{\prod_{z_1=0}^{1}P(Z_{i1}=z_1|Z_{i2})^{I(Z_{i1}=z_1)}\right\}^{J_i+1},$$

where $\mathcal{F}(t|Z_i)=\exp(-H(t|Z_i))$, $F(t|Z_i)=1-\mathcal{F}(t|Z_i)$, and we suppress the dependence on parameters on the right-hand sides for convenience. The primary appeal of this complete data likelihood is that it does not involve probabilities incorporating truncation, as is the case in (2.3), and as a consequence, one can factor the complete data likelihood and carry out the maximization step much more easily.

Let the observed data for individual i be denoted by $Y_i=\{(Z_i,R_i,L_i,X_i,\delta_i)\}$ if $R_i=1$ or $\{(Z_{i2},R_i,L_i,X_i,\delta_i)\}$ if $R_i=0$, and let $Y=(Y_1',\ldots,Y_m')'$. We let $\ell_C(\psi)=\log L_C(\psi)$ and define $Q(\psi;\psi^r)=E(\ell_C(\psi)|Y;\psi^r)$ as the conditional expectation of the complete data log-likelihood given

the observed data, where the expectation is taken using the estimate ψ^r from the rth iteration of the EM algorithm. We can then write

$$Q(\psi;\psi^r) = Q_1(\theta;\psi^r) + Q_2(\eta;\psi^r) \tag{2.4}$$

with $Q_1(\theta;\psi^r) = E(\ell_{C1}(\theta)|Y;\psi^r)$ given by

$$\begin{aligned} & \sum_{i\in\mathcal{R}} [\delta_i \log h(X_i|Z_i) + \log \mathcal{F}(X_i|Z_i) + \mathcal{J}_i^r \log F(L_i|Z_i)] \\ + & \sum_{i\in\bar{\mathcal{R}}} \zeta_i^r [\delta_i \log h(X_i|(1,Z_{i2})) + \log \mathcal{F}(X_i|(1,Z_{i2})) + \mathcal{J}_i^{1r} \log F(L_i|(1,Z_{i2}))] \\ + & \sum_{i\in\bar{\mathcal{R}}} (1-\zeta_i^r)[\delta_i \log h(X_i|(0,Z_{i2})) + \log \mathcal{F}(X_i|(0,Z_{i2})) + \mathcal{J}_i^{0r} \log F(L_i|(0,Z_{i2}))] \end{aligned} \tag{2.5}$$

with $\mathcal{J}_i^r = E(J_i|Z_i, R_i = 1, T_i > L_i, X_i, \delta_i; \psi^r)$, $\mathcal{J}_i^{zr} = E(J_i|(z,Z_{i2}), R_i = 0, T_i > L_i, X_i, \delta_i; \psi^r)$, and $\zeta_i^r = E(Z_{i1}|Z_{i2}, R_i = 0, T_i > L_i, X_i, \delta_i; \psi^r)$. We provide expressions for these conditional expectations in Appendix A.

Existing software for parametric survival analysis can be used to maximize $Q_1(\theta;\psi^r)$, provided it can handle left-censored observations. This can be achieved by creating pseudo-datasets, in which for each $i \in \mathcal{R}$, two lines are generated. One line corresponds to the observed or right-censored observation depending on whether $\delta_i = 1$ or $\delta_i = 0$, respectively. The second line is introduced to correspond to the left-censored failure time of the "ghosts" and has weight $\mathcal{J}_i^r$. For each $i \in \bar{\mathcal{R}}$, four lines are required. First, a contribution for the observed or right-censored failure time is required with the value $Z_{i1} = 1$ and weight ζ_i^r; a second line corresponding to the left-censored observation time with $Z_{i1} = 1$ will have weight $\zeta_i^r \mathcal{J}_i^{1r}$. A second pair of analogous lines are required to reflect the case in which $Z_{i1} = 0$, where the first will have weight $1-\zeta_i^r$ and correspond to the sampled individual and the second with weight $(1-\zeta_i^r)\mathcal{J}_i^{0r}$ corresponding to the left-censored failure time of the "ghosts". Weibull regression models, for example, can be fitted with right-censored and left-censored data, using standard packages for parametric regression including R (survreg [22]), S-PLUS (survReg or censorReg [23]) and SAS (PROC LIFEREG [24]). Alternatively, a more flexible piecewise constant baseline hazard function can be adopted, in which case the M-step can be carried out using software for fitting generalized linear regression models. We describe the details on how to construct the data frame for this algorithm in the Supplementary Meterial.

The function $Q_2(\eta;\psi^r) = E(\ell_{C2}(\eta)|Y;\psi^r)$ in (2.4) is

$$\sum_{i\in\mathcal{R}} [(\mathcal{J}_i^r + 1)\log P(Z_{i1}|Z_{i2})] + \sum_{i\in\bar{\mathcal{R}}} \sum_{z_1=0}^{1} [\zeta_i^r]^{z_1} [1-\zeta_i^r]^{1-z_1} (\mathcal{J}_i^{z_1 r} + 1)\log P(Z_{i1} = z_1|Z_{i2}) \tag{2.6}$$

and can also be maximized using software for logistic regression by creating a pseudo-dataset with one line for each individual $i \in \mathcal{R}$ with weight $\mathcal{J}_i^r + 1$ and observed value of Z_{i1}. For each $i \in \bar{\mathcal{R}}$ two lines are required: one with weight $\zeta_i^r(\mathcal{J}_i^{1r} + 1)$ and $Z_{i1} = 1$, and one with weight $(1-\zeta_i^r)(\mathcal{J}_i^{0r} + 1)$ and $Z_{i1} = 0$. Specification of a quasi-likelihood model with a logit link function and variance function $V(\mu) = \mu(1-\mu)$ will yield the updated estimate η^{r+1}.

2.3 Empirical Performance of the Proposed Method

Here we evaluate the frequency properties of estimators obtained by the proposed algorithm, and we begin by a description of the method of data generation. We let $P(Z_{ik} = 1) = 0.5$, $k = 1, 2$ and the odds ratio for the association between Z_{i1} and Z_{i2} be 2, so $\eta_0 = -0.347$ and $\eta_1 = \log 2$. Suppose the survival time is Weibull distributed with hazard $h(s|Z_i;\theta) = h_0(s;\alpha)\exp(Z_i'\beta)$, where

Table 1: Empirical biases ($\times 10^2$) and standard errors of estimators from the analysis in the absence of missing data, the complete case analysis, a misspecified model, and the correct missing data model fitted via an EM algorithm of Section 2.2; $\alpha_1 = \log \rho = 0$, $\alpha_2 = \log \kappa = 0.405$, $\beta_1 = 0.693$, $\beta_2 = 0.405$, $\eta_0 = -0.347$, $\eta_1 = 0.693$, $\gamma_1 = 1.386$, 25% net censoring, $m = 500$, $nsim = 500$.

			α_1		α_2		β_1		β_2		η_0		η_1	
T%	M%	METHOD†	BIAS	ESE	BIAS	ESE	BIAS	ESE	BIAS	ESE	BIAS	ESE	BIAS	ESE
50	50	NO MISS	-0.13	0.067	0.40	0.066	-0.19	0.106	0.78	0.106	—	—	—	—
		CC	0.31	0.107	0.70	0.096	0.35	0.149	0.46	0.155	—	—	—	—
		MISSPEC	1.81	0.076	-2.27	0.070	-1.41	0.144	0.16	0.110	-54.65	0.221	-10.99	0.283
		EM	-0.26	0.079	0.700	0.072	0.46	0.150	0.76	0.109	-1.86	0.313	0.54	0.394
50	25	NO MISS	-0.13	0.067	0.40	0.066	-0.19	0.106	0.78	0.106	—	—	—	—
		CC	-0.24	0.078	0.74	0.077	-0.51	0.123	1.17	0.121	—	—	—	—
		MISSPEC	0.84	0.070	-0.97	0.068	-1.23	0.119	0.67	0.109	-53.60	0.164	-11.66	0.217
		EM	-0.07	0.07	0.37	0.068	-0.56	0.121	0.97	0.108	-1.28	0.256	-0.92	0.303
25	50	NO MISS	-0.02	0.059	0.54	0.064	-0.73	0.100	0.97	0.103	—	—	—	—
		CC	0.40	0.098	0.96	0.092	0.12	0.146	0.31	0.156	—	—	—	—
		MISSPEC	0.14	0.068	0.17	0.068	-0.16	0.144	0.64	0.107	-18.56	0.216	-5.56	0.269
		EM	-0.07	0.068	0.89	0.068	0.25	0.145	0.70	0.107	-0.95	0.219	1.50	0.271
25	25	NO MISS	-0.02	0.059	0.54	0.064	-0.73	0.100	0.97	0.103	—	—	—	—
		CC	-0.01	0.071	0.84	0.074	-1.00	0.117	1.34	0.119	—	—	—	—
		MISSPEC	0.07	0.062	0.22	0.065	-1.05	0.115	1.03	0.108	-17.11	0.162	-6.81	0.220
		EM	-0.02	0.062	0.53	0.065	-0.92	0.116	1.07	0.108	0.31	0.168	0.04	0.224

† NO MISS is analysis in the absence of missing data, CC is complete case analysis, MISSPEC is based on a misspecified covariate model ignoring truncation, and EM is the correct algorithm described in Section 2.2.

$h_0(s;\alpha) = \rho\kappa(\rho s)^{\kappa-1}$, $\alpha_1 = \log\rho$, $\alpha_2 = \log\kappa$ and $\alpha = (\alpha_1, \alpha_2)'$; we set $\rho = 1$ and $\kappa = 1.5$. We consider a calendar time origin of zero and suppose disease onset happens according to a stationary process in the population giving $D_i \sim \text{Unif}(0, A)$ where $D_i \perp Z_i$. The desired degree of left truncation is obtained by choosing A to satisfy

$$T\% = 100 \cdot (1 - P(E_i > A)) = 100 \cdot (1 - E_{Z_i}\left[E_{D_i|Z_i} P(T_i > A - D_i | D_i, Z_i)\right])$$

where $T\%$ is the truncation percentage; we consider $T\%$=25 and 50.

To generate covariate data compatible with the sampling requirement, given D_i, we generate Z_i according to $P(Z_i|T_i > L_i)$. We then generate $U_i \sim \text{Unif}(0, 1)$, and solve for the failure time T_i in $U_i = \exp(-H(L_i, T_i|Z_i))$. The probability that an individual included in the study is administratively censored given the disease onset time D_i and covariates Z_i is $P(E_i > B|E_i > A, D_i, Z_i) = P(T_i > B - D_i|D_i, Z_i)/P(T_i > L_i|D_i, Z_i)$. We obtain the administrative censoring rate given Z_i by

$$P(E_i > B|E_i > A, Z_i) = E_{D_i|E_i>A,Z_i}\left[P(E_i > B|E_i > A, D_i, Z_i)\right] ,$$

and solve for B in

$$100 \cdot P(E_i > B|E_i > A) = 100 \cdot E_{Z|E_i>A} P(E_i > B|E_i > A, Z) ,$$

to obtain the desired rate, where $P(Z_i|E_i > A) = P(E_i > A|Z_i)\, P(Z_i)/\sum_{Z_i} P(E_i > A|Z_i)P(Z_i)$. Additional random censoring is incorporated by generating an exponential withdrawal time to give a net censoring rate of 25%.

To simulate incomplete data for Z_1, we assume a missing at random mechanism with $P(R_i = 1|Z_i, D_i, E_i > A, X_i, \delta_i) = P(R_i = 1|Z_{i2})$ and let $\text{logit}P(R_i = 1|Z_{i2}) = \gamma_0 + \gamma_1 z_{i2}$. The net frequency of complete data in the sample is then $P(R_i = 1) = E_{Z_{i2}|E_i>A}(P(R_i = 1|Z_{i2}))$. If we fix $\gamma_1 = \log 4$ and the percentage of missing covariate values at M%, one can solve for γ_0 correspondingly; we set M% $= 25, 50$ (i.e., $P(R_i = 1) = 0.75, 0.50$). We simulated 500 datasets ($\text{nsim} = 500$) of $m = 500$ individuals.

For each simulated dataset, we conducted four analyses: (i) an analysis based on the sample including all values of the covariates (NO MISS), possible because this is a simulation study, (ii) a complete case (CC) analysis, which restricts attention to individuals in $\mathcal{R}$, (iii) an analysis based on a misspecified likelihood (MISSPEC) with the form of (2.2) but with $P(Z_{i1}|Z_{i2}; \eta)$ in place of $P(Z_{i1}|Z_{i2}, T_i > L_i; \psi)$, and (iv) the proposed EM algorithm (EM). The analysis in (ii) is based on a correctly specified model and yields consistent estimates of θ under this missing data mechanism, but it is inefficient because it disregards data from individuals in $\bar{\mathcal{R}}$. The analysis in (iii) is based on the correct model for the survival time given the covariates but an incorrect model for the covariates because the population covariate distribution is used; the estimator for ψ is therefore inconsistent. For this analysis, the asymptotic theory on the behavior of maximum likelihood estimators under misspecified models could be exploited [25–27], but we elect to study this through simulation. The analysis based on (iv) is correct, so a consistent estimator of ψ is obtained, which should be more efficient than the estimator from the complete case analysis. The simulation study sheds light on the bias and efficiency trade-offs for these various approaches. Across all parameter configurations considered here, the proposed EM algorithm converged in between 30 and 60 s on a desktop computer with an Intel Core 2 Duo E7400 processor by Intel operating at 2.80 GHZ, with longer computing times occurring under higher rates of missing data and left truncation.

Table 1 displays the empirical biases and empirical standard errors of the estimators from all four approaches; we do not report performance of estimators of η in the first two rows of each configuration (NO MISS and CC) because the covariate distribution would not typically be modeled in these settings. The analysis based on subjects with complete data yielded estimates that had negligible empirical bias for the parameters of interest, as expected. The complete case analysis leads to estimates

with negligible empirical bias but lower efficiency reflected by the greater empirical standard errors. Under the misspecified model, there were small empirical biases of estimators for θ (most appreciable for the α components) and much larger empirical biases of estimators for η, reflecting misspecification of the covariate model. As expected, the estimates from the proposed EM had negligible empirical biases for the components of θ and η, and empirical standard errors that were smaller than those from the complete case analysis. Note that the efficiency gains from the correct analysis were appreciable for all elements of θ except for β_1, the regression coefficient of the partially observed covariate. Broadly similar conclusions were seen in the case $\eta_1 = 0$ (i.e., when covariates are independent) with slightly lower improvement in efficiency with the proposed EM algorithm (results not reported).

3 SUBGROUP ANALYSIS IN CLINICAL TRIALS

When assessing a treatment effect on a time-to-event response in randomized trials, it is customary to define the time origin as the date of randomization. When this time origin is adopted, one is implicitly making treatment comparisons after marginalizing over the left-truncation times as well as any covariates. The time of randomization is the time at which evidence of a treatment effect could emerge, so from this standpoint, it has face validity. Often however, protocols dictate that analyses be stratified according to risk factors whose effects are manifest at the time of disease onset and hence can influence whether individuals will satisfy the entry criteria for the clinical trial. In cancer trials, for example, it may be appropriate to stratify on tumour type or a tumor marker such as human epidermal growth factor receptor 2 (HER-2) status. Important secondary analyses may in fact be directed at assessing treatment effects by HER-2 status and investigating whether there is evidence of differences in treatment effect between strata defined by HER-2 status. The most sensible time origin for these types of analyses is the time of disease onset, and in fact, this is essential to adopt to ensure valid covariate models when such data are incomplete.

We consider here the problem of conducting prespecified subgroup analyses in which the subgroups are defined by patient characteristics and have biological rationale [28]. We presume that the other criteria for valid subgroup analyses are satisfied, and thus the trial is compliant with the CONSORT statement [29]. Consider the setting of Section 2, with D_i, $(Z_{i1}, Z_{i2})'$ and R_i defined as in Section 2.1, but now suppose that at the time of accrual, individuals are randomized to one of two treatment arms. To accommodate the fact that treatment does not begin until recruitment, we define a time-dependent variable $Z_{i3}(s)$ such that $Z_{i3}(s) = 0$ for $0 < s < L_i$ and for $L_i \leq s$, $Z_{i3}(s) = 1$ if individual i is randomized to receive an experimental treatment, and $Z_{i3}(s) = 0$ otherwise. We then let $Z_i(s) = (Z_{i1}, Z_{i2}, Z_{i3}(s))'$ denote the full covariate vector and $Z_i^*(s) = (Z_{i2}, Z_{i3}(s))'$ denote a subvector containing covariates that are always observed. Next let $\bar{Z}_i(s) = \{Z_i(u), 0 \leq u \leq s\}$ and $\bar{Z}_i^*(s) = \{Z_i^*(u), 0 \leq u \leq s\}$ denote the corresponding histories at s, and $\bar{Z}_i = \bar{Z}_i(\infty)$ and $\bar{Z}_i^* = \bar{Z}_i^*(\infty)$ denote the full paths of the respective covariates.

If interest lies in estimating the effect of treatment according to subgroup defined by Z_{i1}, then a natural model is

$$h(s|Z_i(s);\theta) = h_0(s;\alpha)\exp(Z_{i1}\beta_1 + Z_{i2}\beta_2 + Z_{i3}(s)\beta_3 + Z_{i1}Z_{i3}(s)\beta_4)\ . \tag{3.1}$$

If we let $H_i(t;\theta) = H(t|\bar{Z}_i(t);\theta) = \int_0^t h(s|Z_i(s))ds$, then the complete data likelihood is

$$\begin{aligned} L_C(\psi) \ \propto\ & \prod_{i\in\mathcal{R}} \left\{ h^{\delta_i}(X_i|Z_i(X_i))\ \exp(-H_i(X_i))\ [1-\exp(-H_i(L_i))]^{J_i}\ P(Z_{i1}|\bar{Z}_i^*\})^{J_i+1} \right\} \cdot \\ & \prod_{i\in\bar{\mathcal{R}}} \left\{ h^{\delta_i}(X_i|Z_i(X_i))\ \exp(-H_i(X_i))\ [1-\exp(-H_i(L_i))]^{J_i}\ P(Z_{i1}|\bar{Z}_i^*)^{J_i+1} \right\}^{Z_{i1}} \cdot \\ & \prod_{i\in\bar{\mathcal{R}}} \left\{ h^{\delta_i}(X_i|Z_i(X_i))\ \exp(-H_i(X_i))\ [1-\exp(-H_i(L_i))]^{J_i}\ P(Z_{i1}|\bar{Z}_i^*)^{J_i+1} \right\}^{(1-Z_{i1})} \ . \end{aligned}$$

Table 2: Empirical biases ($\times 10^2$) and standard errors of estimators from the analysis in the absence of missing data, the complete case analysis, and the correct missing data model fitted via the EM algorithm of Section 3.1; $\alpha_1 = \log \rho = 0$, $\alpha_2 = \log \kappa = 0.405$, $\beta_1 = 0.693$, $\beta_2 = 0.405$, $\beta_3 = 0$, $\beta_4 = -0.693$, $\eta_0 = -0.347$, $\eta_1 = 0.693$, $\gamma_1 = 1.386$, 25% net censoring, $m = 500$, $nsim = 500$.

			α_1		α_2		β_1		β_2		β_3		$\beta_3 + \beta_4$	
T%	M%	METHOD†	BIAS	ESE	BIAS	ESE	BIAS	ESE	BIAS	ESE	BIAS	ESE	BIAS	ESE
50	50	NO MISS	-0.35	0.079	1.16	0.066	0.32	0.141	0.97	0.101	-1.23	0.123	-0.92	0.171
		CC	-1.05	0.120	2.21	0.090	1.55	0.201	1.67	0.150	-0.49	0.187	-1.60	0.227
		EM	-0.50	0.089	1.42	0.070	0.48	0.187	0.94	0.105	-1.05	0.141	-0.66	0.202
50	25	NO MISS	-0.35	0.079	1.16	0.066	0.32	0.141	0.97	0.101	-1.23	0.123	-0.92	0.171
		CC	0.26	0.095	1.21	0.076	-0.28	0.169	0.99	0.116	-2.05	0.142	-1.32	0.188
		EM	-0.26	0.083	1.18	0.068	-0.14	0.162	1.06	0.101	-1.29	0.126	-0.70	0.179
25	50	NO MISS	0.15	0.076	1.31	0.065	-0.48	0.142	0.95	0.103	-1.46	0.142	-1.03	0.152
		CC	-0.28	0.118	2.30	0.089	0.74	0.202	1.43	0.156	-0.63	0.204	-2.10	0.206
		EM	0.09	0.084	1.62	0.067	0.10	0.188	1.04	0.108	-1.06	0.166	-1.84	0.188
25	25	NO MISS	0.15	0.076	1.31	0.065	-0.48	0.142	0.95	0.103	-1.46	0.142	-1.03	0.152
		CC	0.57	0.094	1.36	0.075	-0.32	0.172	0.81	0.123	-2.24	0.164	-1.57	0.174
		EM	0.15	0.079	1.43	0.066	-0.27	0.166	0.93	0.105	-1.58	0.148	-1.06	0.165

†NO MISS is analysis in the absence of missing data, CC is complete case analysis, and EM is the correct algorithm described in Section 3.1.

Note that $E(J_i|\bar{Z}_i, T_i > L_i; \psi^r)$ and $E(J_i|Z_{i1} = z, \bar{Z}_i^*, T_i > L_i; \psi^r)$ are given by (A.1) and (A.2) respectively, because the treatment variable is defined to be zero prior to the left truncation time. Here, however, $\zeta_i^r = E(Z_{i1}|\bar{Z}_i^*, R_i = 0, T_i > L_i, X_i, \delta_i; \psi^r)$ is

$$\frac{h^{\delta_i}(X_i|(1, Z_i^*(X_i)); \theta^r) \exp(-H(X_i|(1, \bar{Z}_i^*(X_i)); \theta^r)) P(Z_{i1} = 1|Z_{i2}; \eta^r)}{\sum_{z=0}^{1} h^{\delta_i}(X_i|(z, Z_i^*(X_i)); \theta^r) \exp(-H(X_i|(z, \bar{Z}_i^*(X_i)); \theta^r)) P(Z_{i1} = z|Z_{i2}; \eta^r)}.$$

Calculations such as those of Section 2.3 can be carried out to satisfy the 25% censoring rate and particular truncation and marginal missing data rates.

We carry out analyses based on the full sample with no missing covariates (NO MISS), a complete case analysis (CC), and the proposed EM algorithm. In Table 2, we report the empirical biases and standard errors for truncation and missing data rates of 25% and 50%, respectively, for 500 simulated datasets of $m = 500$ individuals. The estimators of β_3 and $\beta_3 + \beta_4$, the two estimates of treatment effect for individuals with $Z_1 = 0$ and $Z_1 = 1$, respectively, are of greatest interest. As was the case in Section 2, we see small biases in these three analyses with the proposed algorithm giving improved efficiency over the complete case analysis for all parameters.

4 Application to a trial involving metastatic cancer

Here we consider data from a trial of 285 breast cancer patients with skeletal metastases [2] diagnosed within three years of randomization. The primary purpose of this trial was to examine the effect of an experimental bisphosphonate therapy (n=133) compared to the control (standard care) therapy (n=152) on the reduction in skeletal complications arising because of these bone metastases. Secondary interest lies in the the effect of therapy on the time to death; the survival times of 42 (14.7%) of the patients were censored for death. We consider an analysis in which separate estimates of the treatment effect are desired for patients that are estrogen receptor (ER) positive and those that are ER negative, while controlling for whether the patient was 50 years of age or older at the time of diagnosis; the model in (3.1) is therefore suitable to address this question. The ER status is missing for 14.3% of patients in the experimental arm and 17.1% of patients in the control arm, but age of diagnosis was completely observed. Among the 114 individuals in the experimental arm with ER status available, 94 (82%) were ER positive, and among the 126 individuals in the control arm with available ER status, 97 (77%) were ER positive.

Table 3 gives the results of fitting a model based on (3.1) under the complete case analysis and fitting a model based on the proposed EM algorithm; we obtained standard errors on the basis of 500 bootstrap samples. Note that there is no evidence of a treatment effect for any patients irrespective of ER status. This is not surprising because this was a palliative trial in which the aim was to improve quality of life. Among individuals who are ER positive, the relative risks were close to one for both analyses, but the point estimate for ER negative patients suggests a 19.5% relative risk reduction based on the complete case analysis (p=0.491). The proposed EM algorithm, which exploits the information about the missing ER status from the left truncation time, gives a relative risk reduction estimate of 25.9% (95% CI: 0.415, 1.327; p=0.311).

5 Discussion

We have considered issues in the analysis of incomplete covariate data under a form of response-biased sampling, which is widely encountered in epidemiologic research as well as clinical trials. This response bias arises any time that there are conditions imposed on individuals for inclusion in a study, but in prevalent cohort studies, the condition that individuals be event free (e.g. alive) at the time of diagnosis leads to left-truncated event times. Left-truncation can readily be handled using

Table 3: Relative risk estimates from complete case analysis and the proposed EM algorithm for fitting a Weibull proportional hazards model with ER status as the partially observed covariate (Z_1), age at diagnosis ($Z_2 = I(\text{age} \geq 50)$), treatment, and an ER status by treatment interaction; standard errors based on 500 bootstrap samples.

	ER Negative			ER Positive		
Method	RR	95% CI	p-value	RR	95% CI	p-value
Complete Case	0.805	(0.433, 1.493)	0.491	1.048	(0.792, 1.387)	0.741
Proposed EM	0.741	(0.415, 1.322)	0.311	1.029	(0.775, 1.367)	0.842

standard software when covariates are complete [5]. When covariates are incompletely observed, one strategy is to specify an observed data likelihood based on the joint distribution of the response times and the covariates. This can be challenging because the correct covariate distribution must condition on the selection criterion being satisfied and therefore involves parameters of the survival distribution. To address this, we describe an EM algorithm based on a complete data likelihood including contributions from individuals who did not satisfy the truncation condition. Standard software for parametric survival analysis that handles left censoring can then be used at the maximization step. The proposed algorithm is shown to perform well for both the setting of prevalent cohort studies and clinical trials where subgroup analyses are of interest but covariates are incomplete.

We have focused on the setting with two binary covariates for which specification of the population covariate distribution is easy. More complex settings could involve incomplete categorical or continuous covariates and similarly more complex observed covariates. Specification of a model for the joint distribution of the covariates in these settings would be considerably more challenging, and indeed one may be willing to give up the potential efficiency gains from the proposed method in order to ensure robustness of the findings. We have also focused on the simplest kind of missing data mechanism, where missingness is driven by a covariate that is always observed. More elaborate missing data mechanim may require modeling of the missing data process. Standard software can also be used to obtain point estimates of regression coefficients from Cox regression models with incomplete covariates via inverse probability weighted estimating equations. Several authors [30, 31] have considered this approach, and it is of interest to explore this approach in the context of left-truncated data.

In addition to the two settings described in this paper, truncated data arise naturally in studies of multistate Markov processes. Consider a progressive multistate process composed of three states with transitions possible from state 1 to state 2 and from state 2 to state 3. The transition time from state 2 to state 3 is typically treated as left truncated because of the delayed entry time to state 2. When incomplete covariate data arise from such processes, likelihoods may have a different form from those considered here depending on the selection process. For example, individuals may be observed from the start of the process or may be selected for follow-up based on being in state 2; the latter would be more similar to the problem considered in this paper.

Covariates are often imprecisely observed due to misclassification for discrete covariates or measurement error for continuous covariates, and there is a large literature on methods for fitting regression models with covariate measurement error [32]. When a structural modeling approach is taken, models for the latent covariate are adopted, and such models would again require one to specify these models in such a way that the covariate distribution addressed the selection effects arising due to left truncation; this would be necessary for an analysis based on either the observed data likelihood or an EM algorithm.

APPENDIX A DERIVATION OF THE CONDITIONAL EXPECTATIONS FOR SECTION 2.2

For each $i \in \mathcal{R}$, the only "missing" information is J_i, the number of "ghosts" that did not satisfy the truncation condition of the respective individual. If ψ^r denotes the parameter estimate at the rth iteration of the EM algorithm, to take the relevant expectations in (2.5) and (2.6), we note $E(J_i|Z_i, R_i = 1, T_i > L_i, X_i, \delta_i; \psi^r) = E(J_i|Z_i, T_i > L_i; \psi^r)$ and that

$$\mathcal{J}_i^r = E(J_i|Z_i, T_i > L_i; \psi^r) = \frac{P(T_i < L_i|Z_i; \theta^r)}{P(T_i \geq L_i|Z_i; \theta^r)} = \frac{1 - \exp(-H(L_i|Z_i; \theta^r))}{\exp(-H(L_i|Z_i; \theta^r))} , \quad \text{for } i \in \mathcal{R} . \tag{A.1}$$

For $i \in \bar{\mathcal{R}}$, in addition to the number of "ghosts", the value of Z_{i1} is missing. We note $E(J_i|(z, Z_{i2}), R_i = 0, T_i > L_i, X_i, \delta_i; \psi^r) = E(J_i|(z, Z_{i2}), T_i > L_i; \psi^r)$ and let

$$\begin{aligned} \mathcal{J}_i^{zr} &= E(J_i|(z, Z_{i2}), T_i > L_i; \psi^r) \\ &= \frac{P(T_i < L_i|(z, Z_{i2}); \theta^r)}{P(T_i \geq L_i|(z, Z_{i2}); \theta^r)} = \frac{1 - \exp(-H(L_i|(z, Z_{i2}); \theta^r))}{\exp(-H(L_i|(z, Z_{i2}); \theta^r))} , \quad \text{for } i \in \bar{\mathcal{R}}, \end{aligned} \tag{A.2}$$

denote the expectation conditional on a particular value of $Z_i = (z, Z_{i2})'$, $z = 0, 1$. We then note $\zeta_i^r = E(Z_{i1}|Z_{i2}, R_i = 0, T_i > L_i, X_i, \delta_i; \psi^r) = E(Z_{i1}|Z_{i2}, T_i > L_i; \psi^r)$, for $i \in \bar{\mathcal{R}}$, which we obtain through

$$\zeta_i^r = \frac{h^{\delta_i}(X_i|(1, Z_{i2}); \theta^r)\mathcal{F}(X_i|(1, Z_{i2}); \theta^r)\, P(Z_{i1} = 1|Z_{i2}; \eta^r)}{\sum_{z=0}^{1} h^{\delta_i}(X_i|(z, Z_{i2}); \theta^r)\mathcal{F}(X_i|(z, Z_{i2}); \theta^r)\, P(Z_{i1} = z|Z_{i2}; \eta^r)} . \tag{A.3}$$

Standard errors can be obtained using the nonparametric bootstrap as done in the example, or using the approach of Louis [33] which can be implemented as follows. Let $U(\psi) = (U_1'(\theta), U_2'(\eta))'$ where $U_1(\theta) = \partial \log L_C(\psi)/\partial\theta$ and $U_2(\eta) = \partial \log L_C(\psi)/\partial\eta$, and

$$I(\psi) = -\partial U(\psi)/\partial\psi' = \begin{pmatrix} I_1(\theta) & 0 \\ 0 & I_2(\eta) \end{pmatrix} \tag{A.4}$$

where $I_1(\theta) = -\partial U_1(\theta)/\partial\theta'$ and $I_2(\eta) = -\partial U_2(\eta)/\partial\eta'$. Then if $\mathcal{I}(\psi)$ is the information matrix from the observed data likelihood (2.2),

$$\mathcal{I}(\psi) = E_M\{I(\psi)|Y\} - E_M\{U(\psi)U'(\psi)|Y\} \tag{A.5}$$

where M represents the missing data, which is simply the number of "ghosts" J for individuals in $\mathcal{R}$ and is the number of ghosts and the covariate Z_1 for individuals in $\bar{\mathcal{R}}$. The expectations are carried out by individual given their respective observed data. The first term in (A.5), for example, is simply obtained by extracting the usual observed information matrices from the two analyses estimating θ and η at the final iteration of the EM algorithm, and the second term is given by taking the outer product of the stacked score vectors and averaging using the weights estimated at the final iteration.

ACKNOWLEDGEMENTS

This research was supported by the Natural Sciences and Engineering Research Council of Canada, the Canadian Institutes of Health Research, and the Ontario Institute for Cancer Research through a grant from the Division of High Impact Clinical Trials. The authors thank Novartis Pharmaceuticals for permission to use the data from the trial of patients with skeletal metastases. R.J. Cook is a Canada Research Chair in Statistical Methods for Health Research.

REFERENCES

[1] Wolfson C, Wolfson DB, Asgharian M, M'Lan CE, Østbye T, Rockwood K, Hogan M for the Clinical Progression of Dementia Study Group. A re-evaluation of the duration of survival after the onset of dementia. *The New England Journal of Medicine* 2001; **344**:1111-1116.

[2] Hortobagyi GN, Theriault RL, Porter L, Blayney D, Lipton A, Sinoff C, Wheeler H, Simeone JF, Seaman J, Knight RD, Heffernan M, Reitsma DJ, Kennedy I, Allan SG, and Mellars K for the Protocol 19 Aredia Breast Cancer Study Group. Efficacy of pamidronate in reducing skeletal complications in patients with breast cancer and lytic bone metastases. Protocol 19 Aredia Breast Cancer Study Group. *New England Journal of Medicine* 1996; **335**: 1785–1791.

[3] Cox DR, Oakes D. *Analysis of Survival Data*. Chapman and Hall: New York, 1984.

[4] Andersen PK, Borgan O, Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. Springer-Verlag: New York, 1993.

[5] Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Truncated and Censored Data*. Springer: New York, 1997.

[6] Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. Second Edition. John Wiley and Sons: New York, 2002.

[7] Lawless JF. *Statistical Models and Mehods for Lifetime Data*. Second Edition. Wiley: New York, 2002.

[8] Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Second Edition. Wiley: New York, 2002.

[9] Lipsitz SR, Ibrahim JG. Using the EM-algorithm for survival data with incomplete categorical covariates. *Lifetime Data Analysis* 1996; **2**:5-14.

[10] Chen HY, Little RJA. Proportional hazards regression with missing covariates. *Journal of the American Statistical Association* 1999; **94**:896-908.

[11] Herring AH, Ibrahim JG, Lipsitz SR. Non-ignorable missing covariate data in survival analysis: a case-study of an international breast cancer study group trial. *Journal of the Royal Statistical Society* 2004; **53**:293-310.

[12] Lipsitz SR. Ibrahim JG. Estimating equations with incomplete categorical covariates in the Cox model. *Biometrics* 1998; **54**:1002-1013.

[13] Wang CY, Chen HY. Augmented inverse probability weighted estimator for Cox missing covariate regression. *Biometrics* 2001; **57**: 414-419.

[14] Ibrahim JG, Chen MH, Kim S. Bayesian variable selection for the Cox regression model with missing covariates. *Lifetime Data Analysis* 2008; **14**:496-520.

[15] Bradshaw PT, Ibrahim JG, Gammon MD. A Bayesian proportional hazards regression model with non-ignorably missing time-varying covariates. *Statistics in Medicine* 2010; **29**: 3017-3029.

[16] Chen HY, Little RJ. A profile conditional likelihood approach for the semiparametric transformation regression model with missing covariates. *Lifetime Data Analysis* 2001; **7**:207-224.

[17] Begg CB, Gray RJ. Methodology for case-control studies with prevalent cases. *Biometrika* 1987; **4**:191-195.

[18] Bergeron P-J, Asgharian M, Wolfson DB. Covariate bias induced by length-biased sampling of failure times. *Journal of the American Statistical Association* 2008; **103**:737-742.

[19] Cook RJ, Bergeron P-J. Information in the covariate distribution with prevalent cohort samples. *Statistics in Medicine* 2011; in press.

[20] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 1977; **39**:1-38.

[21] Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society* 1976; **38**:290-295.

[22] R Development Core Team. *R: A Language and Environment for Statistical Computing*. 2010. ISBN 3-900051-07-0, http://www.R-project.org

[23] Insightful Corporation, Seattle, WA. *S-PLUS 8 Guide to Statistics*. Volume 2, 2007.

[24] SAS Institute Inc. *SAS/STAT 9.2 Users Guide*. Cary, NC: SAS Institute Inc. 2008.

[25] Cox DR. Tests of separate family of hypotheses. Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability 1961; 105-123.

[26] White HA. Maximum likelihood estimation of misspecified models. *Econometrica* 1982; **50**:1–25.

[27] Rotnitzky A, Wypij D. A note on the bias of estimators with missing data. *Biometrics* 1994; **50**:1163–1170.

[28] Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *Journal of the American Statistical Association* 1991; **266**:93-98.

[29] Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001; **357**:1191-1194.

[30] Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coecients when some regressors are not always observed. *Journal of the American Statistical Association* 1994; **89**:846-866.

[31] Lipsitz SR, Ibrahim JG, Zhao LP. A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association* 1999; **94**:1147-1160.

[32] Carroll RJ, Ruppert D, Stefanski LA and Crainiceanu CM. *Measurement Error in Nonlinear Models: A Moder Perspective*, Second Edition, Chapman and Hall, New York, 2006.

[33] Louis TA. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* 1982; **44**(2): 226–233.

Published in final edited form as:
Statistics in Medicine (2013), **32**(6): 1004–1015
DOI: 10.1002/sim.5581

Supplementary Material for
Regression with incomplete covariates and left-truncated time-to-event data

HUA SHEN

Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada

RICHARD J. COOK

Department of Statistics and Actuarial Science,
University of Waterloo, Waterloo, ON, N2L 3G1, Canada
E-mail: rjcook@uwaterloo.ca

AN EM ALGORITHM FOR REGRESSION MODELS WITH PIECEWISE CONSTANT BASELINE HAZARDS

Here we consider an extension of the algorithm of Section 2.2 of Shen and Cook (2012) to deal with more flexible weakly parametric proportional hazards models with piecewise constant baseline hazard functions. Let $0 = b_0 < b_1 < \ldots < b_{K-1} < b_K = \infty$ denote pre-specified cut points giving K sub-intervals $\mathcal{B}_k = [b_{k-1}, b_k)$, $k = 1, \ldots, K$. The baseline function has the form $h_0(t) = \alpha_k$ if $t \in B_k$, $k = 1, \ldots, K$. Let $\mathcal{A}_i = [L_i, \infty)$ denote the truncation region for individual i, and $\mathcal{A}_i^c = [0, L_i)$. In the observational setting of Section 2.2, a complete data likelihood is given, but here we replace the term $F(L_i|Z_i)^{J_i}$ with $\prod_{j=1}^{J_i} f(t_{ij}|Z_i)$, where t_{ij} is the failure time of the jth "ghost" for individual i known to fall in $\mathcal{A}_i^c$. The reason for considering a different form is that the maximization step of the complete data likelihood becomes trivial under a piecewise constant model if the failure times are observed; this can be exploited in the algorithm that follows.

Let $I_k(t) = I(t \in \mathcal{B}_k)$ and let $w_k(t) = \int_0^t I_k(u)du$ denote the amount of time that a particular subject is at risk in $\mathcal{B}_k$ over the interval $[0, t)$. We can then write $f(t|Z_i) = h(t|Z_i)\exp(-H(t|Z_i))$ as

$$f(t|Z_i) = \left[\prod_{k=1}^{K}\left[\alpha_k \exp(Z_i'\beta)\right]^{I_k(t)}\right] \exp\Big(-\left[\textstyle\sum_{k=1}^{K} w_k(t)\alpha_k\right]\exp(Z_i'\beta)\Big). \tag{S.1}$$

Let $\delta_{ik} = I_k(X_i)$ indicate whether the observation time $X_i = \min(T_i, C_i)$ is in interval $\mathcal{B}_k$ for individual i, and let $S_{ik} = \int_0^{X_i} I(u \in \mathcal{B}_k)du$ denote the total time individual i was at risk of failure during the interval $\mathcal{B}_k$. By replacing $F(L_i|Z_i)^{J_i}$ with $\prod_{j=1}^{J_i} f(t_{ij}|Z_i)$ in the complete data likelihood

of Section 2.2 and by taking the log, we obtain

$$\begin{aligned}
\ell_C(\psi) &= \sum_{i\in\mathcal{R}}\Big\{\sum_{k=1}^{K}\Big[\delta_i\delta_{ik}\Big(\log\alpha_k + Z_i'\beta\Big) - \alpha_k S_{ik}e^{Z_i'\beta}\Big] \\
&\quad + \sum_{j=1}^{J_i}\sum_{k=1}^{K}\Big[I_k(t_{ij})\Big(\log\alpha_k + Z_i'\beta\Big) - w_k(t_{ij})\alpha_k e^{Z_i'\beta}\Big] + (J_i+1)\log P(Z_{i1}|Z_{i2})\Big\} \\
&+ \sum_{i\in\bar{\mathcal{R}}}\Big[Z_{i1}\Big\{\sum_{k=1}^{K}\Big[\delta_i\delta_{ik}\Big(\log\alpha_k + Z_i'\beta\Big) - \alpha_k S_{ik}e^{Z_i'\beta}\Big] \\
&\quad + \sum_{j=1}^{J_i}\sum_{k=1}^{K}\Big[I_k(t_{ij})\Big(\log\alpha_k + Z_i'\beta\Big) - w_k(t_{ij})\alpha_k e^{Z_i'\beta}\Big] + (J_i+1)\log P(Z_{i1}|Z_{i2})\Big\} \\
&+ (1-Z_{i1})\Big\{\sum_{k=1}^{K}\Big[\delta_i\delta_{ik}\Big(\log\alpha_k + Z_i'\beta\Big) - \alpha_k S_{ik}e^{Z_i'\beta}\Big] \\
&\quad + \sum_{j=1}^{J_i}\sum_{k=1}^{K}\Big[I_k(t_{ij})\Big(\log\alpha_k + Z_i'\beta\Big) - w_k(t_{ij})\alpha_k e^{Z_i'\beta}\Big] + (J_i+1)\log P(Z_{i1}|Z_{i2})\Big\}\Big],
\end{aligned}$$

where the event time for the jth ghost corresponding to individual i, t_{ij}, is only known to be in the interval $\mathcal{A}_i^c = [0, L_i)$. As before we can split this likelihood into two parts $\ell_C(\psi) = \ell_{C1}(\theta) + \ell_{C2}(\eta)$, where $\ell_{C1}(\theta)$ is

$$\begin{aligned}
&\sum_{i\in\mathcal{R}}\sum_{k=1}^{K}\Big\{\Big[\delta_i\delta_{ik}\log(\alpha_k e^{Z_i'\beta}) - \alpha_k S_{ik}e^{Z_i'\beta}\Big] + \sum_{j=1}^{J_i}\Big[I_k(t_{ij})\log(\alpha_k e^{Z_i'\beta}) - w_k(t_{ij})\alpha_k e^{Z_i'\beta}\Big]\Big\} \qquad \text{(S.2)} \\
+ &\sum_{i\in\bar{\mathcal{R}}}\Big\{Z_{i1}\sum_{k=1}^{K}\Big\{\Big[\delta_i\delta_{ik}\log(\alpha_k e^{Z_i'\beta}) - \alpha_k S_{ik}e^{Z_i'\beta}\Big] + \sum_{j=1}^{J_i}\Big[I_k(t_{ij})\log(\alpha_k e^{Z_i'\beta}) - w_k(t_{ij})\alpha_k e^{Z_i'\beta}\Big]\Big\} \\
+ &(1-Z_{i1})\sum_{k=1}^{K}\Big\{\Big[\delta_i\delta_{ik}\log(\alpha_k e^{Z_i'\beta}) - \alpha_k S_{ik}e^{Z_i'\beta}\Big] + \sum_{j=1}^{J_i}\Big[I_k(t_{ij})\log(\alpha_k e^{Z_i'\beta}) - w_k(t_{ij})\alpha_k e^{Z_i'\beta}\Big]\Big\}\Big\},
\end{aligned}$$

and $\ell_{C2}(\eta)$ is given by (2.5). Thus $Q(\psi;\psi^r) = E(\ell_C(\psi)|Y;\psi^r) = Q_1(\theta;\psi^r) + Q_2(\eta;\psi^r)$, where as before $Q_1(\theta;\psi^r) = E(\ell_{C_1}(\theta)|Y;\psi^r)$, and $Q_2(\eta;\psi^r) = E(\ell_{C_2}(\eta)|Y;\psi^r)$. At the rth step of the EM algorithm, we need $\mathcal{J}_i^r$, $\mathcal{J}_i^{1r}$, $\mathcal{J}_i^{0r}$ and ζ_i^r, given by (A.1), (A.2) and (A.3) respectively. The expectations regarding t_{ij} are given as follows. If the complement of the truncation interval does not intersect with $\mathcal{B}_k$ (i.e. $\mathcal{C}_{ijk} = \mathcal{A}_i^c \cap \mathcal{B}_k = \emptyset$ because $b_{k-1} > L_i$), then $E(I_k(T_{ij})|Z_i, T_{ij} < L_i, J_i) = 0$. If $b_{k-1} < L_i$, $\mathcal{C}_{ijk} = \mathcal{A}_i^c \cap \mathcal{B}_k = [L_{ijk}, R_{ijk}) \neq \emptyset$, where $L_{ijk} = \max(b_{k-1}, 0) = b_{k-1}$, and $R_{ijk} = \min(b_k, L_i)$. We then take the expectation of (S.2) at the rth step of the EM algorithm, using

$$\begin{aligned}
\iota_{ik}^r &= E(I_k(t_{ij})|Z_i, R_i = 1, T_{ij} < L_i, J_i; \psi^r) = P(T_{ij} \in B_k|Z_i, T_{ij} < L_i, J_i; \psi^r) \\
&= \frac{\mathcal{F}(L_{ijk}|Z_i;\psi^r) - \mathcal{F}(R_{ijk}|Z_i;\psi^r)}{\mathcal{F}(0|Z_i;\psi^r) - \mathcal{F}(L_i|Z_i;\psi^r)} = \frac{\mathcal{F}(b_{k-1}|Z_i;\psi^r) - \mathcal{F}(\min(b_k, L_i)|Z_i;\psi^r)}{1 - \mathcal{F}(L_i|Z_i;\psi^r)},
\end{aligned}$$

and

$$\begin{aligned}
\iota_{ik}^{zr} &= E(I_k(t_{ij})|(z, z_{i2}), R_i = 0, T_{ij} < L_i, J_i; \psi^r) \\
&= \frac{\mathcal{F}(b_{k-1}|(z, z_{i2});\psi^r) - \mathcal{F}(\min(b_k, L_i)|(z, z_{i2});\psi^r)}{1 - \mathcal{F}(L_i|(z, z_{i2});\psi^r)},
\end{aligned}$$

where $z = 0, 1$.

Regarding the time at risk, $\mathcal{C}_{ijk} = \mathcal{A}_i^c \cap \mathcal{B}_k = \emptyset$, (i.e., $L_i < b_{k-1}$), each ghost j corresponding to individual i, $j = 1, \ldots, J_i$ failed before entering interval $\mathcal{B}_k$, and thus they were never at risk of failure in $\mathcal{B}_k$; in that case, $E(w_k(t_{ij})|Z_i, T_{ij} < L_i, J_i) = 0$. If $\mathcal{C}_{ijk} = \mathcal{A}_i^c \cap \mathcal{B}_k = [L_{ijk}, R_{ijk}) \neq \emptyset$, $b_{k-1} < L_i$, it is possible that they could have failed before entering $\mathcal{B}_k$, in which case there is no period at risk corresponding to the interval $[b_{k-1}, b_k)$. At the rth step of the EM algorithm, we have,

$$\omega_{ik}^r = E(w_k(t_{ij})|Z_i, R_i = 1, T_{ij} < L_i, J_i; \psi^r) = \int_{b_{k-1}}^{\min(b_k, L_i)} \frac{\mathcal{F}(u|Z_i; \psi^r) - \mathcal{F}(L_i|Z_i; \psi^r)}{1 - \mathcal{F}(L_i|Z_i; \psi^r)} du,$$

and

$$\begin{aligned} \omega_{ik}^{zr} &= E(w_k(t_{ij})|(z, z_{i2}), R_i = 0, T_{ij} < L_i, J_i; \psi^r) \\ &= \int_{b_{k-1}}^{\min(b_k, L_i)} \frac{\mathcal{F}(u|(z, z_{i2}); \psi^r) - \mathcal{F}(L_i|(z, z_{i2}); \psi^r)}{1 - \mathcal{F}(L_i|(z, z_{i2}); \psi^r)} du, \quad z = 0, 1. \end{aligned}$$

Let $K_i = \max\{k : b_{k-1} < X_i\}$ be the maximum interval over which individual i is known to have been at risk and $K_{ij} = \max\{k : b_{k-1} < L_i\}$ denote the the maximum interval over which the ghosts for individual i could have been at risk. Furthermore, let

$$Q_{i1k}(\theta; \psi^r) = \delta_i \delta_{ik} \left(\log \alpha_k + Z_i' \beta \right) - \alpha_k \exp(Z_i' \beta + \log S_{ik})$$

be the expectation of this kth element of the first term in the first row of (S.2) and let

$$G_{i1k}(\theta; \psi^r) = \mathcal{J}_i^r \left[\iota_{ik}^r \left(\log \alpha_k + Z_i' \beta \right) - \alpha_k \exp(Z_i' \beta + \log \omega_{ik}^r) \right]$$

denote the expectation of the kth element in the second term in the first row of (S.2). Then if $i \in \mathcal{R}$,

$$Q_{i1}(\theta; \psi^r) = \sum_{k=1}^{K_i} Q_{i1k}(\theta; \psi^r) + \sum_{k=1}^{K_{ij}} G_{i1k}(\theta; \psi^r). \tag{S.3}$$

Similarly, for $i \in \bar{\mathcal{R}}$, let

$$Q_{i1k}^z(\theta; \psi^r) = \delta_i \delta_{ik} \left(\log \alpha_k + (z, z_{i2})' \beta \right) - \alpha_k \exp(z\beta_1 + z_{i2}' \beta_2 + \log S_{ik}) \,,$$

and

$$G_{i1k}^z(\theta; \psi^r) = \mathcal{J}_{iz}^r \left[\iota_{ik}^{zr} \left(\log \alpha_k + (z, z_{i2})' \beta \right) - \alpha_k \exp(z\beta_1 + z_{i2}\beta_2 + \log \omega_{ik}^{zr}) \right] ,$$

and then define

$$\bar{Q}_{i1}(\theta; \psi^r) = \sum_{z=0}^{1} (\zeta_i^r)^z (1 - \zeta_i^r)^{1-z} \left[\sum_{k=1}^{K_i} Q_{i1k}^z(\theta; \psi^r) + \sum_{k=1}^{K_{ij}} G_{i1k}^z(\theta; \psi^r) \right]. \tag{S.4}$$

Combining (S.3) and (S.4) we then obtain

$$Q_1(\theta; \psi^r) = \sum_{i \in \mathcal{R}} Q_{i1}(\theta; \psi^r) + \sum_{i \in \bar{\mathcal{R}}} \bar{Q}_{i1}(\theta; \psi^r). \tag{S.5}$$

The function in (S.5) can be maximized using standard software for fitting Poisson or exponential regression models. A sample section of the data frame at the rth iteration is given in Table 4 and 5 for a subject with $R_i = 1$ or 0 respectively. If one creates a factor variable based on column `K`, we could fit a Poisson model with covariates Z_1, Z_2 and factor(K) with response `int-stat` $\times$ `stat`, offset `log(len)`, and weight $\texttt{weight}_z \times \texttt{weight}_J$. The updated estimate of θ is θ^{r+1} and the parameter estimates for the baseline hazard can be obtained from the coefficients of the factor variable K. The updated estimates of η are obtained as described in Section 2.2.

Table 4: The first part of the pseudo-data frame for maximizing $Q_1(\theta;\psi^r)$ with respect to θ for an arbitrary individual $i \in \mathcal{R}$.

R	K	Z_1	Z_2	len	int-stat	stat	weight$_Z$	weight$_J$
1	1	z_{i1}	z_{i2}	S_{i1}	δ_{i1}	δ_i	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	K_i	z_{i1}	z_{i2}	$S_{i\,K_i}$	$\delta_{i\,K_i}$	δ_i	1	1
1	1	z_{i1}	z_{i2}	ω^r_{i1}	ι^r_{i1}	1	1	$\mathcal{J}^r_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	K_{ij}	z_{i1}	z_{i2}	$\omega^r_{iK_{ij}}$	$\iota^r_{iK_{ij}}$	1	1	$\mathcal{J}^r_i$

Table 5: Second part of the pseudo-data frame for maximizing $Q_1(\theta;\psi^r)$ with respect to θ for an arbitrary individual $i \in \bar{\mathcal{R}}$.

R	K	Z_1	Z_2	len	int-stat	stat	weight$_Z$	weight$_J$
0	1	1	z_{i2}	S_{i1}	δ_{i1}	δ_i	ζ^r_i	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
0	K_i	1	z_{i2}	$S_{i\,K_i}$	$\delta_{i\,K_i}$	δ_i	ζ^r_i	1
0	1	1	z_{i2}	ω^{1r}_{i1}	ι^{1r}_{i1}	1	ζ^r_i	$\mathcal{J}^{1r}_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
0	K_{ij}	1	z_{i2}	$\omega^{1r}_{iK_{ij}}$	$\iota^{1r}_{iK_{ij}}$	1	ζ^r_i	$\mathcal{J}^{1r}_i$
0	1	0	z_{i2}	S_{i1}	δ_{i1}	δ_i	$1-\zeta^r_i$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
0	K_i	0	z_{i2}	$S_{i\,K_i}$	$\delta_{i\,K_i}$	δ_i	$1-\zeta^r_i$	1
0	1	0	z_{i2}	ω^{0r}_{i1}	ι^{0r}_{i1}	1	$1-\zeta^r_i$	$\mathcal{J}^{0r}_i$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
0	K_{ij}	0	z_{i2}	$\omega^{0r}_{iK_{ij}}$	$\iota^{0r}_{iK_{ij}}$	1	$1-\zeta^r_i$	$\mathcal{J}^{0r}_i$