Precision of estimators in interval censored parametric survival models

Defen Peng[1] and Gilbert MacKenzie [2]

[1] Centre of Biostatistics, University of Limerick, Ireland & Department of Statistics, Zhongnan University of Economics and Law, China
[2] ENSAI, France & Centre of Biostatistics, University of Limerick, Ireland

Abstract: Recently, several advances have been made in the analysis of interval censored (IC) data mainly in relation to semi-parametric proportional hazard (PH) models (Gómez et al., 2009, Lesaffre et al., 2005). It is arguable, however, that the parametric case has been somewhat neglected, overall, and that more can be learned, especially in relation to non-PH models. Accordingly, we focus on simple parametric models for interval censored survival data arising in longitudinal RCTs. For the exponential regression model we compare the performance of a general likelihood with commonly used proxy likelihoods, which ignore the interval censoring by treating the interval censored times to events as if they were exact. We show *analytically* that use of proxy likelihoods leads to estimators which are artificially precise and we quantify the extent of the resulting biases in a simulation study and by analyzing real data. We also compare the likelihoods using non-PH models and obtain different findings.

Keywords: Artificial precision, Interval Censoring, Longitudinal RCTs; PH & non-PH Survival Models, Proxy likelihoods.

1 Introduction

In longitudinal settings where the response variable, $Y(t)$, is binary typically we observe the ith patient at baseline in a healthy state, i.e., $Y_i(t_0) = 0$. As the process evolves an adverse event may occur, i.e., $Y_i(t_s) = 1$ where $t_s > t_0$. Finkelstien (1986) and Collett (1994) elected to adopt a "time to event" analysis in order to recover information on the treatment effect in the LDA-RCT setting. Moreover, clinicians (Bergink *et al.*, 1998) have adopted a similar approach in which interval censored follow-up times, to the loss of 3 lines of visual acuity (Bailey-Lovie, 1976), were treated as if they were exact times to events. Intuitively, this simple expedient seems sub-optimal and this note investigates the extent of any penalty incurred by comparing a proxy likelihood with the IC likelihood which arises in longitudinal data (MacKenzie, 1999). Here we focus on the use of proxy times (beginning, midpoint and endpoint of intervals) to construct the likelihood rather than treating the lack of exact times as missing data to be imputed. We also focus on simple parametric survival models.

2 Likelihood Construction

Suppose there are $m+1$ *fixed*, scheduled, inspection times, t_0^*, t_1^*,..., t_m^* at which continuous or ordinal responses Y_0, Y_1,..., Y_m, are measured. This arrangement implies m+1 time intervals: $I_1 = (t_0, t_1^*], I_2 = (t_1^*, t_2^*], ..., I_k = (t_{k-1}^*, t_k^*], ..., I_m = (t_{m-1}^*, t_m^*]$ and $I_{m+1} = (t_m^*, \infty]$. Typically, $t_0 = 0$, especially in RCTs where, $t_0 = 0$ represents time of randomization. Hence, let T be a non-negative random variable denoting the time to some outcome of interest defined on the Ys. Let $S(t;\theta)$ and $\lambda(t;\theta)$ be the corresponding survival and hazard functions, respectively, depending on the unknown possibly vector-valued parameter $\theta \in \Theta$. Then, for a sample of n independent subjects subject to non-informative censoring the usual likelihood for the unknown parameters is

$$L_2(\theta) = \Pi_{i=1}^n [\lambda(t_i;\theta)S(t_i;\theta)]^{\delta_i}[S(t_{ic};\theta)]^{1-\delta_i}, \tag{1}$$

where $\lambda(t_i;\theta)S(t_i;\theta) = f(t_i;\theta)$, δ_i is the censoring indicator ($\delta_i = 1$ for an event and 0 otherwise) and t_{ic} is a right censored survival time. Substituting, one of: (a) the beginning point of the interval, t_{ib}, or (b) the interval mid-point, t_{im} or, (c) the interval end-point, t_{ie}, $\forall\, i$, as if it were the exact time at which failure occurred in $L_2(\theta)$ yields the proxy likelihood.
Typically each individual ($i = 1, \ldots, n$) defines their own trajectory over the course of the longitudinal study, thereby generating a person-specific set of intervals. Accordingly, we obtain the following interval censored likelihood

$$L_1(\theta) = \Pi_{i=1}^n \{S(t_{i,k-1};\theta)[1 - S(t_{i,k-1}, t_{ik};\theta)]\}^{\delta_i}[S(t_{ic};\theta)]^{1-\delta_i}. \tag{2}$$

Now, $L_1(\theta)$ and $L_2(\theta)$ may be used for comparative inference. Other authors have reached similar conclusions about the structure of the likelihood in the so-called Case II censoring situation; see Yu et al. (2000) and Schick and Yu (2000), for further details of likelihood construction in related contexts. Note, however, it is unusual to have any exact times to events in a longitudinal study.

3 The Exponential Regression Model

MacKenzie (1999) showed analytically that estimators obtained from the proxy likelihood were artificially precise in the simple Exponential case based on the first order approximation. Here we extend the results to the Exponential Regression case.

3.1 Likelihoods

Armed with some general formulae (not given here) we investigate the Exponential Regression model. Let T follow the exponential regression model

defined by

$$\lambda_{i2} = \lambda(t_i; \alpha_2, \beta_2) = \exp(\alpha_2 + x_i'\beta_2),$$

where $S(t_i; \alpha_2, \beta_2) = \exp[-\lambda_{i2} t_i]$ and α_2 is an unconstrained parameter, β_2 is $p \times 1$ vector of regression coefficients and x_i is a $p \times 1$ vector of fixed covariates. The corresponding proxy likelihood is

$$L_2(\alpha_2, \beta_2) \quad = \quad \prod_{i=1}^{n} \left\{ \lambda_{i2} \mathrm{e}^{-\lambda_{i2} t_i} \right\}^{\delta_i} \left\{ \mathrm{e}^{-\lambda_{i2} t_{ic}} \right\}^{1-\delta_i}, \tag{3}$$

For the IC likelihood we have

$$\lambda_{i1} = \lambda(t_i; \alpha_1, \beta_1) = \exp(\alpha_1 + x_i'\beta_1),$$

where $S(t_{i,k-1}, t_{ik}; \alpha_1, \beta_1) = \exp[-\lambda_{i1} d_i(t_k)]$, and $d_i(t_k) = t_{ik} - t_{i,k-1}$ is the width of the kth interval. Then,

$$L_1(\alpha_1, \beta_1) = \prod_{i=1}^{n} \left\{ \mathrm{e}^{-\lambda_{i1} t_{i,k-1}} \left[1 - \mathrm{e}^{-\lambda_{i1} d_i(t_k)} \right] \right\}^{\delta_i} \left\{ \mathrm{e}^{-\lambda_{i1} t_{ic}} \right\}^{1-\delta_i}, \tag{4}$$

3.2 Comparison of IC and Proxy Approaches

Comparing the Proxy and IC approaches we find that approximate IC mles (i.e., the first order approximation) are identical to those estimated at $t_{ie} = t_{ik}$, the end points of the interval using the proxy likelihood (i.e., $\hat{\alpha}_1 = \hat{\alpha}_2$ and $\hat{\beta}_{1r} = \hat{\beta}_{2r}$) with proxy t_{ie}.

We compared the relative efficiency of the two estimators by examining $V_2(\hat{\alpha}_2)/V_1(\hat{\alpha}_1)$ and $V_2(\hat{\beta}_{2r})/V_1(\hat{\beta}_{1r})$, $r = 1, 2, \cdots, p$. The details are too lengthy to reproduce here. Analytical results are available only for categorical covariates. We have proved the following result for a categorical covariate with $p+1$ categories, modelled by p binary dummy variables, i.e.

$$\begin{aligned} V_2(\hat{\alpha}_{2e})/V_1(\hat{\alpha}_1) \quad &< \quad 1 \\ V_2(\hat{\beta}_{2er})/V_1(\hat{\beta}_{1r}) \quad &< \quad 1 \end{aligned} \tag{5}$$

so that the conjecture that the proxy mles are artificially precise holds, under the first order conditions invoked above, for a single categorical covariate.

We have also proved a similar result for two correlated binary covariates. For higher numbers of correlated binary covariates and for continuous covariates the matrix algebra rapidly becomes intractable. We conjecture that the results hold for two or more categorical variables, but must resort to simulation.

We note in passing that any continuous covariate may be represented in $p \leq n$ distinct categories and hence for such a representation of a continuous covariate the above conjecture holds.

3.3 Information Matrices

The Fisher information matrix based on the IC likelihood with the first order approximation for the Exponential regression model is

$$\mathcal{I}(\alpha,\beta) = \begin{bmatrix} \sum_{i=1}^{n} \mathrm{e}^{\alpha+x_i^T\beta} E(t_i') & \sum_{i=1}^{n} x_i^T \mathrm{e}^{\alpha+x_i^T\beta} E(t_i') \\ \sum_{i=1}^{n} x_i \mathrm{e}^{\alpha+x_i^T\beta} E(t_i') & \sum_{i=1}^{n} x_i x_i^T \mathrm{e}^{\alpha+x_i^T\beta} E(t_i') \end{bmatrix} \tag{6}$$

where $E(t_i') = E\big[\delta_i t_{i,k-1} + (1-\delta_i) t_{ci}\big]$. In general, we have

$$\mathcal{I}(\alpha,\beta) = \mathcal{I}_b(\alpha,\beta) + \mathcal{I}_c(\alpha,\beta)$$

where the subscripts represent the beginning of the interval ($t_{i,k-1}$) and right censored (t_{ci}) components respectively. Fisher Information involves taking the expectation of the negative of the hessian matrix with respect to the random variable T. In this sense it is an averaging or centering operation. Accordingly, in this spirit we may define "general" Fisher information for the IC case by replacing $t_{i,k-1}$ with t_k^* and replacing t_{ci} with its future expectation, as in Buckley & James (1979) yielding

$$\mathcal{I}_{\text{gen}}(\alpha,\beta) = \mathcal{I}_{t_k^*}(\alpha,\beta) + \mathcal{I}_c(\alpha,\beta).$$

Looking at the structure of (6) it is tempting to simplify further by choosing $E(t_i') = E(T_i) = \lambda(t_i)^{-1} = \mathrm{e}^{-(\alpha+x_i^T\beta)}$, whence

$$\mathcal{I}_{\text{ideal}}(\alpha,\beta) = \begin{bmatrix} n & \sum_{i=1}^{n} x_i^T \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i x_i^T \end{bmatrix},$$

an "idealized" form, which is identical to the uncensored solution.
In simulation studies we conduct the exact survival times are known and in these circumstances it is possible to compute an information matrix of the form

$$\mathcal{I}_{\text{rc}}(\alpha,\beta) = \mathcal{I}_u(\alpha,\beta) + \mathcal{I}_c(\alpha,\beta)$$

which we refer to as the "right-censored" version.
In the simulation section we evaluate the performance of all of the above and compare it with the observed information from the IC likelihood which, broadly, we consider should be regarded as the "truth". In the simulation study we found the following relationship between the generalized variances:

$$\det[I_o^{-1}(\hat\alpha,\hat\beta)] > \det[\mathcal{I}_{\text{gen}}^{-1}(\hat\alpha,\hat\beta)] > \det[\mathcal{I}_{\text{ideal}}^{-1}(\alpha,\beta)] > \det[\mathcal{I}_{\text{rc}}^{-1}(\hat\alpha,\hat\beta)].$$

where we have assumed throughout that the δ_i are known.

4 Simulation Study

We conducted a data-directed simulation study mimicking the conduct of a RCT with two arms and a follow-up period of 2 years (Hart et al., 2002). We generated failure times from the Exponential regression model with two covariates: x_1, a binary covariate mimicking the treatment effect (1 = New(50%) and 0 = Old(50%) and x_2 a continuous baseline covariate distributed, $N(0, \sigma^2_{x_2})$, where $\sigma_{x_2} \leq 1$ ($\sigma_{x_2} = 0.5$ in our simulation study). The trajectories for each individual in the study were constructed according to two schedules: an irregular schedule (0.25, 0.5, 1 and 2 years) and a regular schedule (0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75 and 2 years), respectively. Censoring rates of 20% (normal) and 50% (heavy) were considered. The method of creating intervals is non-informative about the survival distribution.

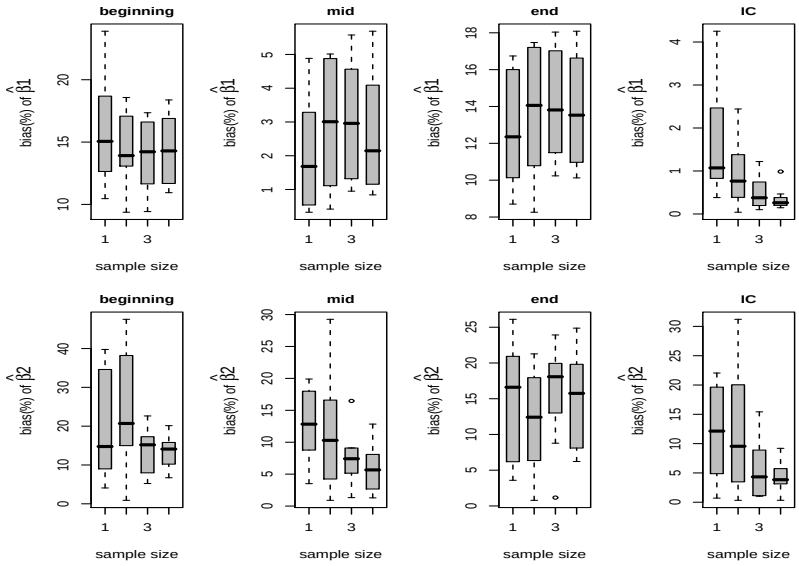

Fig. 1. Percentage bias for 32 scenarios by sample size for β_1 and β_2. The x-axis labels 1-4 represent sample sizes n=100, 200, 500 and 1000 respectively. Boxplot titles: estimates obtained at the beginning, mid, and end points (identical to the first order approximation) by proxy likelihoods and by IC likelihood (NR).

Figure 1 shows the average percentage bias in an exponential regression model by sample size (n=100, 200, 500 and 1000), likelihood method and for two covariates (β_1, β_2) using the irregular schedule. These results show that the estimators from the IC likelihood by using the Newton-Raphson algorithm (NR) have minimum bias and that the bias is asymptotically consistent. However, for the proxy likelihoods this is not the case. Only the mid-point estimator has acceptable levels of bias, but the box-plots (which depict the variation over scenarios) suggest a lack of consistency for β_1.

The findings are similar for the regular schedule. We also considered the Weibull PH model and two non-PH models - the log-logistic the canonical time dependent logistic. For PH models the results showed that, among the IC and proxy likelihoods considered, the estimators in the IC likelihood had the largest variances. This was re-assuring, as á priori, one might reasonably expect the IC likelihood to represent the most uncertainty. Accordingly, this demonstrates that the estimators in all of the proxy likelihoods are artificially precise. However, surprisingly, for non-PH models this finding did not hold. We were able to find immediate contradictions in the non-PH models. The results will be described in detail at the Workshop together with the analysis of two published data sets.

5 Discussion

The analysis of IC data has been reviewed recently by Gómez, et al. (2009). Here, we tried to develop an analytical approach to the analysis of precision of the regression estimator. This was successful in the Exponential Regression model for simple cases. However, for more complicated cases, the algebra rapidly becomes intractable and one must resort to simulation. Our findings support the conjecture that the estimators based on the proxy likelihoods are artificially precise in the PH models studied. Hence proxy approaches should be avoided, especially in RCTs, when the data obey the PH assumption. However, this is apparently not true of non-PH models, a finding which warrants further investigation.

Acknowledgments: The work was supported by the Science Foundation Ireland (SFI, www.sfi.ie) Mathematics Initiative, II, via the BIO-SI (www.ul.ie/bio-si) research programme in the Centre of Biostatistics, University of Limerick, Ireland: grant number 07/MI/012. In addition, Professor Peng is also supported by SFI via a Research Frontiers Programme award, grant number 05/RF/MAT 026.

Key References

Gómez et al. (2009). Tutorial on methods for interval censored data *Statistical Modelling* 9, 4, 259-298.

Lawless J and Babineau (2006). Models for interval censoring and simulated-based inference for lifetime distributions. *Biometrika*, **93**, 671-686.

MacKenzie G (1999). Survival analysis for longitudinal data. Proceedings of the 14th International Workshop on Statistical Modelling, July, Graz, Austria. July 1999, pages 259-264.

Peng D (2009). *Inferences in the Interval Censored Exponential Regression Model.* Masters Thesis MacMaster University, Ontario, Canada.