Survival Analysis for Longitudinal Data

Gilbert MacKenzie[1]

[1] Centre for Medical Statistics, Keele University, Keele, Staffordshire ST5 5BG, UK. E-mail: g.mackenzie@keele.ac.uk

Abstract

In longitudinal studies with a set of continuous or ordinal repeated response variables it may be convenient to summarise the outcome as a threshold event. Then, the time to this event becomes of interest. This is particularly true of recent Ophthalmological trials evaluating the effect of treatment on the loss of visual acuity over time. However, the practice of employing conventional survival analysis methods for testing the null hypothesis of no treatment effect in these types of studies is intrinsically flawed as the exact time to the threshold event is not measured. In this paper we obtain a general Likelihood for the unknown parameters when the underlying survival model is parametric. We also recover the actual information available in repeated measures data for a variety of models and compare the results with those obtained using a mis-specified model, which assumes the time to the event is one of the possibly irregularly spaced inspection times.

Keywords Longitudinal Data, Survival Analysis, Model Mis-specification, Grouped Likelihood.

1 Introduction

In longitudinal studies in which the response is continuous or ordinal, clinicians often find it convenient to categorise the outcome. If the response is change from baseline it may be natural to think in term of a threshold effect. Opthalmological studies usually investigate visual loss over time in terms of distance visual acuity (DVA) measured on an ordinal scale due to Bailey- Lovie (B-L) and analyses are frequently performed in terms of numbers of lines of visual acuity lost (MPSG, 1994). For example, Bergink et al (1998) in their RCT of teletherapy in ARMD analysed the outcome <3 or $3+$ lines of visual acuity lost. In these studies recruitment is staggered over time and increasingly survival-type methods, such as Kaplan Meier curves, the Log-Rank test, and the Proportional Hazards model of Cox (1972) are being pressed into service.

These methods are appropriate for right censored 'time to event data' when the exact time of occurrence is known, but strictly inappropriate when the 'time to event' is known only to lie in an interval. Thus, for example, the loss of 3+ BL lines observed at a scheduled inspection time could have occurred at any time since the previous examination. Thus, the use of inspection times to rank the data is merely approximate. It is obvious that this procedure must lead to many ties in the so-called times to the events. Such ties are often ignored or inappropriately broken by the variation in examination times, rather than by the actual times when the threshold was crossed.

In this paper we (a) identify the correct Likelihood for pseudo 'time to event data' arising in repeated measures studies (b) obtain the maximum Likelihood estimating equations for some parametric models and (c) compare inference under the correct model with the mis-specified model, which utilises the inspection times as if they were exact.

2 Model Formulation

Suppose there are $m+1$ scheduled inspection times, t_o^*, t_1^*,..., t_m^* at which continuous or ordinal responses Y_o, Y_1,..., Y_m, are measured. 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 vector parameter $\theta \epsilon \Theta$. Then for a sample of n independent subjects it may be shown that the true censored Likelihood for the unknown parameters is:

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

where typically n_k patients fail between scheduled examination times $t^*_{(k-1)}$ and t_k^* for $k = 1,..,m$ and n_c patients are censored or withdrawn at specific times such that $n_c + \sum_{k=1}^{m} n_k = n$. Here $\delta_i = 1$ denotes an event. We may compare (1) directly with the mis-specified censored likelihood resulting from treating the inspection times as if they were exact:

$$L_2(\theta) = \Pi_{i=1}^{n}[\lambda(t_{ik};\theta)S(t_{ik};\theta)]^{\delta_i}[S(t_{ik};\theta)]^{1-\delta_i} \tag{2}$$

Equations (1) and (2) enable us to investigate the effect of mis-specification for any survival model where the survivor function takes a closed form. A popular choice is the PH model. With a binary covariate, possibly a treatment indicator, the test of the null hypothesis ($\beta = 0$) is equivalent to the use of the log-rank test. However, in general we note that:

$$S(t_{ik};\theta) = exp\big[-\int_0^{t_{ik}} \lambda_o(u)exp(x'\beta)du\big] \tag{3}$$

and the survivor function cannot be obtained in closed form without further assumption, as the baseline hazard function $\lambda_o(t)$ is unknown. Accordingly, below, we prefer to utilise the wholly parametric comptetitor for the PH model given by MacKenzie (1996) which is known to track the Cox model closely when the data obey the PH assumption. The survivor function for the time-dependent logistic (TDL) is:

$$S(t;\theta) = \Big[\frac{1+exp(t\alpha + x'\beta)}{1+exp(x'\beta)}\Big]^{-1/\alpha} \tag{4}$$

where α is a scalar parameter and β is a vector of covariates. In the TDL model, with a binary treatment indicator, the test of the null hypothesis is similar, but the additional parameter α must be estimated simultaneously.

In the next section we illustrate some of the underlying ideas using the Exponential and TDL survival distributions, but first we re-write (1) in the more convenient form:

$$L_1(\theta) = \Pi_{i=1}^{n}\{S(t_{i(k-1)};\theta)[1 - S(t_{i(k-1)}, t_{ik};\theta)]\}^{\delta_i}[S(t_{ik};\theta)]^{1-\delta_i} \tag{5}$$

where:

$$S(t_{i(k-1)}, t_{ik};\theta) = exp\Big[-\int_{t_{i(k-1)}}^{t_{ik}} \lambda(u;\theta)du\Big] \tag{6}$$

is the conditional event-specific survival function, whence the second member of (5) is the conditional probability of failure in $(t_{i(k-1)}, t_{ik}]$ given survival to time $t_{i(k-1)}$.

3 Two Parametric Examples

Our focus is to compare the MLEs and standard errors obtained from the grouped Likelihood (5) with those from equation (2) when the underlying Exponential and TDL models hold.

(a) Exponential

If T follows the Exponential distribution with parameter ϕ, then $\lambda(t;\theta)=\phi$, $S(t;\theta)=exp(-\phi.t)$ and $S(t_{i(k-1)}, t_{ik};\theta)= exp[-\phi.(t_{ik} - t_{i(k-1)})]$. From (2), using the inexact $t_{ik}s$ and writing $l_2(\phi)$ for $log_e L_2(\phi)$, the first derivative is given by:

$$U_{2,\phi} = \frac{\partial l_2(\phi)}{\partial \phi} = \sum_{i=1}^{n}[\delta_i.\phi^{-1} - \delta_i.t_{ik} - (1-\delta_i).t_i^*] \tag{7}$$

and solving $U_{2,\phi} = 0$ yields the closed form MLE:

$$\hat{\phi} = n_u/(T_u + T_c) \tag{8}$$

where T_u and T_c are the sums of the uncensored t_{ik} and censored t_i^*, respectively, and $n_u = \sum_{i=1}^{n} \delta_i$ is the total number of uncensored events. Differentiating again we find:

$$I_{2,\phi} = -\frac{\partial^2 l_2(\phi)}{\partial \phi^2} = \sum_{i=1}^{n} \delta_i . \phi^{-2} \tag{9}$$

Thus, the variance of $\hat{\phi}$ is given by $V_{2,\phi} = \phi^2/n_u$ which may be consistently estimated by subsituting $\hat{\phi}/(T_u + T_c)$. The corresponding equations for (5) are:

$$U_{5,\phi} = \sum_{i=1}^{n} [\delta_i . d_i(t) . \psi(t_{ik}, t_{i(k-1)}) - \delta_i . t_{i(k-1)} - (1 - \delta_i) . t_i^*] \tag{10}$$

where $d_i(t) = (t_{ik} - t_{i(k-1)})$ and $\psi(t_{ik}, t_{i(k-1)}) = S(t_{i(k-1)}, t_{ik}; \phi)/[1 - S(t_{i(k-1)}, t_{ik}; \phi)]$ the conditional odds on survival in the interval $(t_{i(k-1)}, t_{ik}]$. Since $\psi(.)$ is non-linear in ϕ, the ML estimating equation, $U_{5,\phi} = 0$, must be solved iteratively for $\hat{\phi}$. However, it may be shown that the MLE given by (8) is the approximate (i.e., first order) solution of $U_{5,\phi} = 0$. The observed information is given by:

$$I_{5,\phi} = \sum_{i=1}^{n} \delta_i .\ d_i^2(t) \psi(t_{ik}, t_{i(k-1)}) [1 + \psi(t_{ik}, t_{i(k-1)})] \tag{11}$$

and the asymptotic variance of $\hat{\phi}$ is given by $V_{5,\phi} = \phi^2/(n_u - \phi . d)$ where $d = \sum_{i=1}^{n} \delta_i . d_i(t)$. We may compare the relative efficiency of the two estimators by examining $V_{2,\phi}/V_{5,\phi} = 1 - (\phi . d/n_u) < 1$, which shows that the estimator based on (2) under-estimates the true variance $V_{5,\phi}$ when the observed inspection times are analysed as if they were exact. We may gain further insight by substituting $\hat{\phi}$ from (8) to show that the relative efficiency is approximately $(T_u + T_c - d)/(T_u + T_c)$, a factor which artificially increases the precision of the estimator based on (2), as the time intervals between visits coarsen.

(b) Time Dependent Logistic

When T follows the Time Dependent Logistic regression model the misspecified censored Likelihood (MacKenzie, 1997) is:

$$L_2(\alpha, \beta) = \prod_{i=1}^{n} \Big[\frac{exp(t_{ik}\alpha + x_i^{'}\beta)}{1 + exp(t_{ik}\alpha + x_i^{'}\beta)} \Big]^{\delta_i} . \Big[\frac{1 + exp(t_{ik}^{*}\alpha + x_i^{'}\beta)}{1 + exp(x_i^{'}\beta)} \Big]^{-(1/\alpha)} \tag{12}$$

where the first member of (12) is the hazard function $\lambda(t; \alpha, \beta)$, the second member is the survival function given at (4), the t_{ik} are inexact and $\delta_i = 1$ for an event and zero otherwise.

The only additional quantity required is the conditional event specific survival function:

$$S(t_{i,(k-1)}, t_{ik}) = \Big[\frac{1 + exp(t_{ik}\alpha + x_i^{'}\beta)}{1 + exp(t_{i(k-1)}\alpha + x_i^{'}\beta)}\Big]^{-1/\alpha} \tag{13}$$

whence the grouped likelihood is given by:

$$\begin{aligned} L_5(\alpha, \beta) = \prod_{i=1}^{n} \Big\{ \Big[\frac{1 + exp(t_{i(k-1)}\alpha + x_i^{'}\beta)}{1 + exp(x_i^{'}\beta)}\Big]^{-1/\alpha} \times \\ \Big[1 - \Big[\frac{1 + exp(t_{ik}\alpha + x_i^{'}\beta)}{1 + exp(t_{i(k-1)}\alpha + x_i^{'}\beta)}\Big]^{-1/\alpha}\Big]\Big\}^{\delta_i} \times \\ \Big\{\Big[\frac{1 + exp(t_{ik}^{*}\alpha + x_i^{'}\beta)}{1 + exp(x_i^{'}\beta)}\Big]^{-(1/\alpha)}\Big\}^{(1-\delta_i)} \end{aligned} \tag{14}$$

An advantage of the Time Dependent Logistic family of models is that when α is small the behaviour of the PH model is reproduced closely and when $\alpha = 0$ the Exponential distribution emerges. Accordingly, the TDL model provides a flexible basis for comparisons. However, formal comparison of $L_2(\alpha, \beta)$ and the obviously more complicated $L_5(\alpha, \beta)$ is less straightforward than for the Exponential distribution. Details of the derivation $U_{2,\alpha,\beta}$ and $V_{2,\alpha,\beta}$ may be found in MacKenzie(1996) and we omit details of $U_{5,\alpha,\beta}$ and $V_{5,\alpha,\beta}$ in the interests of space. In summary, numerical comparisons show findings similar to those described in the previous section.

4 Discussion

In this short paper we have highlighted some of the difficulties which can arise from the application of survival analysis methods to the inexact 'time to event' data recorded in longitudinal studies. Perhaps the key result is the identification of the grouped Likelihood (5) as the preferred vehicle for recovering the actual time to event information available in repeated measures data. Although, the investigation is at a preliminary stage, the analytical findings in relation to the Exponential distribution show that the precision of the estimator is artificially increased by analysing the observed inspection times as if they were exact. Moreover it is clear that this apparent gain in precision increases with coarsening intervals. Similar results have been obtained for the Time Dependent Logistic distribution in numerical studies. We conjecture that our findings are more general and that over-optimistic results may well be obtained in repeated measures situations where conventional survival methods are (mis-)applied. An extensive simulation study, designed, *inter alia,* to evaluate the performance of the Log-rank test is now well under way and the results will be described at

the forthcoming workshop. We are currently applying our new methodology to analyse data from the MRC's current trial of Teletherapy in ARMD in order to compare the results with those obtained using the specially extended Laird-Ware model of Reeves and MacKenzie (1998), which allows for fixed and random effects in the presence of serially correlated and irregularly spaced inspection times. Whilst the reasons for wishing to adopt survival-type methods are clear, especially when recruitment is staggered in time, the wisdom of using conventional approaches is in question.

References

Bergink, et al (1998). A Randomised Controlled Clinical Trial on the efficacy of radiation therapy in the control of subfoveal choroidal neovascularisation in age-related macular degeneration: radiation versus observation. *Graefe's Arch. Clin. Exp.Ophthalmology.* 236, No 752, 1-5.

Macular Photocoagulation Study Group. (1994). Visual outcome after laser photocoagulation for sub-foveal choroidal neovascular secondary to age-related macular degeneration. *Arch. Ophthalmol.* 112, 480-488.

Cox DR (1972). Regression models and life tables (with Discussion) *JRSS B.* 34, 187-220.

MacKenzie G (1996). Regression models for survival data. *JRSS D.* 45, 1, 21-34.

MacKenzie G (1997). On a non-proportional hazards regression model for repeated medical random counts. *Statistics in Medicine.* 16, 1831-1843, 21-34.

Reeves J and MacKenzie G (1998). A bivariate regression model with serial correlation. *JRSS D.* 47, 4, 607-615.