A Logistic PH Regression Model for Interval Censored Survival Data

Yasin Al-Tawarah[1] & Gilbert MacKenzie[1]

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

Abstract

The survival function of the logistic proportional hazards (PH) regression model, MacKenzie (1996), is obtained. A general likelihood for interval censored survival data which depends only on various closed forms of the survival function of a continuous underlying failure time distribution is introduced. The use of the logistic PH survival distribution is proposed for the analysis of interval censored data following a PH distribution. The corresponding interval censored likelihood is developed and compared with a mis-specified likelihood which treats the interval censored data as if they were exact. A simulation study is used to compare the two likelihoods.

Keywords: Interval censoring, Logistic Survival, Proportional Hazards, Mis-specified Likelihood

1 Introduction

In classical survival analysis, the exact time to event is usually known. However, in longitudinal clinical trials where outcome is a continuous or ordinal variable measured repeatedly at scheduled follow-up times, the exact time-to-event may be unknown. Such situations arise when the outcome is classified according to threshold of clinical interest. Then scientific interest is focused on the time at which the threshold is crossed. In these studies recruitment is staggered in time and, increasingly, survival-type methods (Kaplan Meier, 1958; Peto & Peto, 1972 and 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. Application of conventional methods to interval 'end' or 'mid' points can lead to bias (Lindsey and Ryan, 1998) and optimistic precision (MacKenzie, 1999). Finkelstein

(1986) introduced a method for fitting a PH model to interval censored data and developed the corresponding score test for $\beta = 0$. By comparison, we develop the parametric PH time dependent logistic (PHTDL) model (MacKenzie, 1996) in which the baseline hazard follows the time-dependent logistic (TDL) survival model. We derive the corresponding likelihoods and compare inference from the correct model with that from the mis-specified model which treats the interval censored data as if they were exact.

2 Model Formulation

2.1 The Logistic PH Survival Model

The hazard function for the logistic proportional hazards (LPH) model (MacKenzie, 1996) is:

$$\lambda(t;x) = \lambda_0(t)\exp(x'\beta) \quad (1)$$

$$\lambda_0(t) = \frac{\exp(t\alpha+\gamma)}{1+\exp(t\alpha+\gamma)} \quad (2)$$

where t is a nonnegative random variable, $x'=(x_1,...,x_p)$ is a row vector of covariates and $\beta'=(\beta_1,...,\beta_p)$ is a row vector unknown regression parameters to be estimated. The survival function for time dependent logistic can be obtained directly by integrating the hazard function in equation (3), viz:

$$S(t;x) = S_0(t)^{exp(x'\beta)} \quad (3)$$

$$S_0(t) = \exp\left[-\int_0^t \lambda_0(u)du\right] \quad (4)$$

$$= \left[\frac{1+\exp(t\alpha+\gamma)}{1+\exp(\gamma)}\right]^{-1/\alpha} \quad (5)$$

the latter survival function is that of the TDL survival model. Another quantity, the conditional event survival function:

$$S(t_{(k-1)},t_k;x) = \exp\left[-\int_{t_{k-1}}^{t_k} \lambda_0(u)du.\exp(x'\beta)\right]$$

$$= [S_0(t_k,t_{k-1})]^{\exp(x'\beta)}$$

$$= [S_0(t_k)/S_0(t_{(k-1)})]^{\exp(x'\beta)} \quad (6)$$

representing the conditional probability of an event in $[t_{(k-1)},t_k)$, is required later.

When the censoring mechanism is non-informative the likelihood for the logistic PH survival model may be written as usual:

$$\Pi_{i=1}^n\{\lambda(t;x)\}^{\delta_i}S(t;x) \quad (7)$$

2.2 A General Interval-Censored Likelihood

Suppose there are $m+1$ scheduled inspection times, $t_o^+, t_1^+, ..., t_m^+$ at which continuous or ordinal responses $Y_0, Y_1, ..., Y_m$ are measured. Let T be a non negative 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 hazards functions, respectively, depending on the unknown vector parameter $\theta' = (\alpha, \gamma, \beta)'$. Then for a sample of n independent subjects it may be shown that the true censored likelihood for the unknown parameters is:

$$L_1(\theta) = \prod_{i=1}^{n} \left\{ S(t_{i_{(k-1)}};\theta) \left[1 - S(t_{i_{k-1}}, t_{i_k};\theta)\right] \right\}^{\delta_i} [S(t_i^*;\theta)]^{1-\delta_i} \tag{8}$$

where typically n_k patients fail between scheduled examination times $t^+_{(k-1)}$ and t_k^+ for $k = 1, ..., m$ and n_c patients censored or withdrawn at specific times such that $n_c + \sum_{k=1}^{m} n_k = n$. Here $\delta_i = 1$ denotes an event and $\delta_i = 0$ denotes a censored observation. Notice that by denoting $t_{i_{k-1}}$ and t_{i_k} with nested subscripts we mean that the intervals are individual-specific and are *not* fixed at the scheduled inspection times $t_o^+, t_1^+, ..., t_m^+$, in keeping with practical observation in longitudinal trials. Moreover, we write t_i^* to denote a right censoring time, recognizing that it may not correspond to any scheduled inspection time - for example, an early withdrawal from the trial.

3 Comparative Inference

The mis-specified censored likelihood resulting from treating the end-points of the observed intervals as if they were exact failure times is:

$$L_2(\theta) = \prod_{i=1}^{n} [\lambda(t_{i_k};\theta) S(t_{i_k};\theta)]^{\delta_i} [S(t_{i_k};\theta)]^{1-\delta_i} \tag{9}$$

Equations (8) and (9) therefore enable us to investigate the effect of mis-specification for any model which has a closed form survival function.

We regard (8) as the natural vehicle for inference and of key interest is the null hypothesis, $H_o : \beta = 0$. A variety of tests may be constructed based on (8) and, for example, when this null hypothesis obtains, (8) reduces to a function of $S_o(t)$, depending on α & γ, viz:

$$L_o(\alpha, \gamma) = \prod_{i=1}^{n} \left\{ S_0(t_{i_{(k-1)}}) \left[1 - [\frac{S_0(t_{i_k})}{S_0(t_{i_{k-1}})}]^{exp(x'\beta)}\right] \right\}^{\delta_i} [S_0(t_i^*)]^{1-\delta_i} \tag{10}$$

The logistic PH model is particularly convenient for analyzing PH data as it avoids non-parametric estimation (Turnbull, 1976) of the so-called "baseline" survival function, equation (4), inherent in Cox's PH model in which

$\lambda_0(t)$ is unknown. Moreover, the properties of (5) have been described in detail elsewhere, MacKenzie(1996).

By routine calculation of the first and second derivatives of (8) we may obtain $U(\theta)$ and $I(\theta)$ and then compute $U(\hat{\theta}_0)'I(\hat{\theta}_0)U(\hat{\theta}_0) \sim asym\ \ \chi^2_p$, a modified score test for $H_o : \beta = 0$, where $\hat{\theta}_0'$=$(\hat{\alpha}, \hat{\gamma}, \beta = 0)'$. This statistic may be compared with the corresponding score test based on (9) which treats the times as if they were exact.

4 Results

4.1 Logistic PH

In order to demonstrate the utility of the logistic PH model, which has never previously been fitted, we analyze time-to-event data from a population-based prospective study of incident cases of lung cancer diagnosed in Northern Ireland in one year. Time from diagnosis to death or censoring is analyzed in relation to clinical and demographic factors measured on each patient.

Table 1: Comparison of Models: Age and Sex
Maximum Likelihood Estimates & SEs

		LPH		Cox	
Factors	Parameters	Estimate	(se)	Estimate	(se)
	$\hat{\gamma}$	-2.980	(0.301)		
Age	$\hat{\beta}1$	+0.017	(0.003)	+0.017	(0.004)
	$\hat{\alpha}$	-0.057	(0.011)		
	$\hat{\gamma}$	-1.651	(0.103)		
Sex	$\hat{\beta}1$	+0.029	(0.096)	+0.031	(0.081)
	$\hat{\alpha}$	-0.064	(0.011)		
	$\hat{\gamma}$	-2.988	(0.339)		
Age &	$\hat{\beta}1$	+0.017	(0.004)	+0.017	(0.004)
Sex	$\hat{\beta}2$	+0.014	(0.087)	+0.017	(0.082)
	$\hat{\alpha}$	-0.057	(0.010)		

We illustrate the model in relation to Age and Sex - factors which are usually considered be determinants of lung cancer *incidence* rather than *survival*. Table 1 shows the Maximum likelihood estimates and their standard errors for Age, Sex, and Age & Sex for the LPH and Cox Model. Of the factors studied only Age reached statistical significance. Figure 1 shows that the conclusion in relation to Sex is corroborated by the data. In later multi-factor analyzes (not shown) the effect of Age on survival was abolished by the influence of other factors related to the clinical stage of the disease, while that of Sex did not emerge. The similarity of the LPH results with Cox model should be noted.

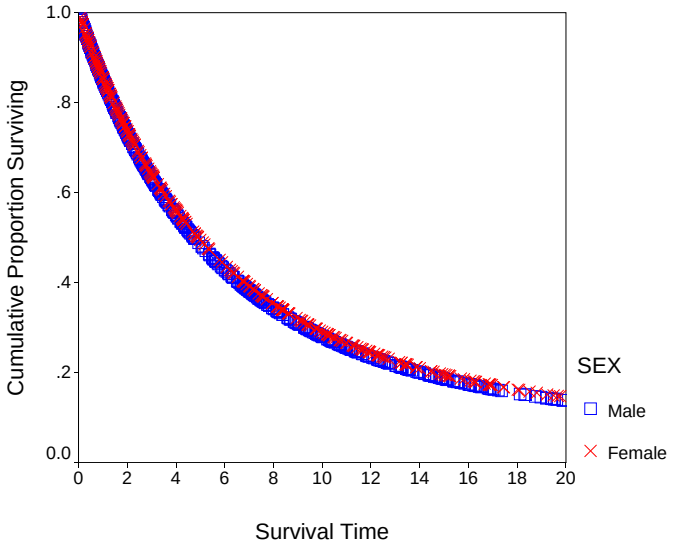

FIGURE 1. Observed and Fitted (LPH) Survivor Functions by Sex.

4.2 Simulation

Method In addition, we conducted a simulation study in order to quantify the degree to which inference about the parameters in the logistic PH model, especially β, is compromised by the use of the mis-specified likelihood (9), when (8) obtains. For simplicity we investigated the 2-sample case, mimicking a RCT in which scientific interest is focused on estimating the treatment effect and its associated standard error. Times to events were generated according to the LPH model and we set the maximum time to event in the reference group (x=0) to 24 months in order mimic a trial with a two year follow-up period. Accordingly, negative values of β correspond to longer times to events in the intervention group (x=1), ie, to benefit.

The parameters studied in the simulation included: sample size (100, 200, 500); percentage censored within the trial (0, 5, 10, 30). Pattern of scheduled follow-up visits: irregular and regular intervals, for example, (3,6,12,24) and (3,6,9,12,15,18,21,24) respectively. A range of parameter values for (α, γ, β). In the mis-specified likelihood we substituted the mid-points and end-points (not shown) as if they were exact. The latter procedure is the one which is usually adopted in many types of clinical trials. The analysis was based on 1000 simulations for each model and the results were analyzed using conventional statistical methods.

Table 2: Comparison of Mis-specified and True Models Estimators
follow up = $(3, 6, 12, 24)$ Mid-point

		Mis-specified			True		
	n	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\beta}$
		$\alpha = 0.03$, $\gamma = -1.3$, $\beta = 0$, % within censoring=0					
Mean	100	0.040	−1.456	0.000	0.049	−1.351	0.000
(se.)		(0.032)	(0.225)	(0.230)	(0.047)	(0.272)	(0.238)
Mean	200	0.032	−1.430	0.000	0.037	−1.317	0.000
(se.)		(0.019)	(0.140)	(0.143)	(0.029)	(0.170)	(0.150)
Mean	500	0.029	−1.423	0.000	0.033	−1.307	0.000
(se.)		(0.012)	(0.093)	(0.093)	(0.019)	(0.114)	(0.097)
		$\alpha = 0.03$, $\gamma = -1.3$, $\beta = -0.1$, % within censoring=0					
Mean	100	0.039	−1.441	−0.100	0.048	−1.334	−0.100
(se.)		(0.033)	(0.215)	(0.216)	(0.046)	(0.259)	(0.223)
Mean	200	0.030	−1.423	−0.100	0.037	−1.310	−0.100
(se.)		(0.020)	(0.149)	(0.140)	(0.031)	(0.182)	(0.146)
Mean	500	0.029	−1.426	−0.090	0.033	−1.310	−0.090
(se.)		(0.011)	(0.089)	(0.091)	(0.019)	(0.108)	(0.096)
		$\alpha = 0.03$, $\gamma = -1.3$, $\beta = -0.25$, % within censoring=0					
Mean	100	0.038	−1.451	−0.240	0.046	−1.342	−0.250
(se.)		(0.030)	(0.218)	(0.212)	(0.044)	(0.255)	(0.220)
Mean	200	0.033	−1.431	−0.240	0.038	−1.315	−0.250
(se.)		(0.018)	(0.149)	(0.147)	(0.028)	(0.179)	(0.154)
Mean	500	0.030	−1.429	−0.241	0.033	−1.308	−0.250
(se.)		(0.011)	(0.091)	(0.091)	(0.017)	(0.108)	(0.095)
		$\alpha = 0.03$, $\gamma = -1.3$, $\beta = -0.50$, % within censoring=0					
Mean	100	0.038	−1.452	−0.482	0.041	−1.323	−0.501
(se.)		(0.026)	(0.209)	(0.213)	(0.039)	(0.247)	(0.220)
Mean	200	0.035	−1.442	−0.481	0.037	−1.316	−0.500
(se.)		(0.018)	(0.148)	(0.154)	(0.026)	(0.173)	(0.160)
Mean	500	0.032	−1.442	−0.479	0.032	−1.306	−0.500
(se.)		(0.011)	(0.089)	(0.094)	(0.015)	(0.102)	(0.099)

Results We report only a subset of the complete simulation using mid-points in the mis-specified likelihood. Table 2 shows the MLE's for the three parameters using 4 follow-up visits scheduled at at (3,6,12,24) months. The true likelihood provides consistently better estimates, although the difference using mid-points is small. It is also clear that the mis-specified likelihood is consistently over-optimistic in terms of precision. The standard

error is underestimated on average by 4.3% an amount which is unlikely to be important except when the effect of intervention is small - the frequency of which scenario is increasing as large medical trials investigate smaller and smaller benefits.

The previous analysis did not allow for drop-outs during the trial period. In Table 3, we permit within term censoring to be 10% and present the result for $\beta = 0$ and $\beta = -0.5$. The results are similar to those shown above although estimates obtained from true likelihood are less biased and have more accurate standard errors - the mis-specified likelihood again producing standard errors which are artificially precise.

Table 3: Comparison of Mis-specified and True Models Estimators
follow up $= (3, 6, 12, 24)$ Mid-point

		Mis-specified			True		
	n	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\gamma}$	$\hat{\beta}$
		$\alpha = 0.03$, $\gamma = -1.3$, $\beta = 0$, % within censoring =10					
Mean	100	0.036	−1.564	0.000	0.035	−1.456	0.000
(se.)		(0.032)	(0.220)	(0.226)	(0.043)	(0.251)	(0.230)
Mean	200	0.032	−1.563	0.000	0.030	−1.452	0.000
(se.)		(0.021)	(0.155)	(0.159)	(0.030)	(0.181)	(0.164)
Mean	500	0.028	−1.551	0.000	0.025	−1.432	0.000
(se.)		(0.012)	(0.094)	(0.093)	(0.017)	(0.109)	(0.097)
		$\alpha = 0.03$, $\gamma = -1.3$, $\beta = -0.5$, % within censoring=10					
Mean	100	0.037	−1.574	−0.490	0.035	−1.464	−0.511
(se.)		(0.028)	(0.207)	(0.220)	(0.037)	(0.234)	(0.228)
Mean	200	0.033	−1.582	−0.478	0.029	−1.462	−0.491
(se.)		(0.019)	(0.147)	(0.159)	(0.024)	(0.164)	(0.165)
Mean	500	0.032	−1.571	−0.475	0.027	−1.456	−0.490
(se.)		(0.011)	(0.091)	(0.099)	(0.014)	(0.101)	(0.102)

In further analyzes to be presented at the Workshop we show that the use of end-points in the mis-specified likelihood leads to significantly worse results in terms of bias and precision for some parameters.

5 Final Remarks

In this short paper we have developed the Logistic PH model introduced by MacKenzie (1996) showing that the model has currency in the analysis of medical survival data. The advantages of the model stem from the closed form survivor function and the fact that when $\beta = 0$ the underlying

survival function has a testable parametric form - unlike the PH model of Cox (1972). We have also demonstrated, by means of a simulation, its use in the analysis of interval censored survival data arising in longitudinal randomized controlled trials.

References

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

Finkelstein D (1986). A proportional hazards model for interval-censored failure time data. *Biometrics.* 42,845-854.

Kaplan Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association.* 53,457-481.

Lindsey & Ryan (1998). Tutorial in biostatistics methods for interval-censored data. *Statistics in Medicine.* Vol. 17,219-238.

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

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

Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *J.R. Statist. Soc.* A,135,185-206.

Turnbull BW (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *JRSS B.* 38, 290-295.