A Logistic Regression Model for Survival Data

Gilbert MacKenzie[1]

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

Abstract

The near non-identifiability of one of the parameters in the Generalized Time-Dependent Logistic (GTDL) Survival Model (MacKenzie, 1996, 1997) is discussed. A new canonical 3-parameter logistic model survival model, in which all of the parameters are identifiable, is obtained. A direct connection with Fisher's Z distribution is established. The properties of this non-PH model are contrasted briefly with Cox's PH model. The new model is used to investigate survival from lung cancer in a population study.

Keywords: Canonical Model, Fisher' Z, Generalized Time-Dependent Logistic, Non-PH Survival

1 Introduction

The 3-parameter generalized time dependent logistic regression survival model (GTDL), introduced by MacKenzie (1996) generalizes the relative risk in the proportional hazards (PH) model of Cox (1972) to time-dependent form. The model, which may be regarded from some perspectives as a wholly parametric competitor for the PH model, has several interesting properties including a frailty interpretation. In applications, a reduced (2-parameter) form of the GTDL - the time dependent logistic (TDL) - has been used in studies of survival from lung cancer and other medical conditions (MacKenzie 1996, 1997).

However, although the existence of the 3-parameter form, the GTDL, has been demonstrated, parametric estimation using standard Maximum Likelihood methods has proved difficult. The problem arises because in the original formulation, the parameter, λ, is apparently aliased with the intercept, β_0, in the linear predictor of the model. The resulting near non-identifiability of λ was un-anticipated but the effect has been to limit applications to the TDL model which is not as general.

In this paper we develop the model class and relax the non-identifiability constraint to reveal a new survival model - which we refer to as the canonical logistic regression survival model. The structure of the paper is as follows. In section § 2 we define the GTDL, and derive of the model illustrating its connection with Fisher's Z distribution. In section § 3 we relax the non-identifiability and discuss and in § 4 we analyze survival in a population-based study of 900 incident cases of lung cancer.

2 Model Formulation

2.1 *Definition of the GTDL*

Denote by T a non-negative random variable representing failure time and let the instantaneous failure rate, or hazard, be defined as $\lambda(t)$ =lim$\{$pr$(t \leq T < t + \delta t | T \geq t)/\delta t\}$, where the limit is taken as $\delta t \rightarrow 0^+$. Then, the generalized time-dependent logistic regression model is defined by the hazard function:

$$\lambda(t;x) = \lambda_o \theta(t;x) \tag{1}$$

where $\lambda_o > 0$ is a scalar, $\theta(t;x) = \exp(t\alpha + x\beta/)/[1 + \exp(t\alpha + x\beta)]$ is a linear logistic function in time, α is a scalar measuring the effect of time and $\beta' = (\beta_0, ..., \beta_p)$ is a column vector of $p+1$ unknown regression parameters measuring the influence of p+1 covariates $x = (x_o, ..., x_p)$. The covariate vector includes an intercept term ($x_o = 1$). The density of the failure time model is:

$$f(t;x) = \lambda(t;x) S(t,x) \tag{2}$$

and it may be shown that the survivor function, $S(t;x)$, is given by:

$$S(t;x) = \Big\{ \frac{1 + \exp(t\alpha + x\beta)}{(1 + \exp(x\beta)} \Big\}^{-(\lambda_o/\alpha)} \tag{3}$$

Consider a sample of n independent individuals with data (t_i, x_i, δ_i) where $\delta_i = 1$ for a failure and 0 otherwise. When the censoring mechanism is non-informative the likelihood may be written as:

$$\Pi_{i=1}^{n} \{\lambda(t;x)\}^{\delta_i} S(t;x) \tag{4}$$

A reduced model - the Time Dependent Logistic (TDL) - emerges when $\lambda_o = 1$ and the likelihood is then a function of the $p+2$ parameters (α, β').

More extensive details, including the score functions and the observed information matrix will be given in the full paper.

2.2 *A connection with Fisher's Z*

The model may be derived in a variety of ways. A direct route is via the 4-parameter logistic model presented by Ahuja and Nash (1967), viz:

$$f(z \mid \rho, \sigma, \psi, \theta) = \frac{1}{\sigma\beta(\psi, \theta)}(\rho e^{-\frac{z}{\sigma}})^{\psi}(1 + \rho e^{-\frac{z}{\sigma}})^{-(\psi+\theta)} \tag{5}$$

which is a generalization of Fisher's Z distribution, $z = \frac{1}{2}\log_e F$, where F is the variance ratio and $-\infty < z < \infty$. Fisher's Z distribution may be recovered immediately by writing: $\rho = \nu_0/\nu_1, \sigma = 1/2, \psi = \nu_0/2$ and $\theta = \nu_1/2$, where ν_1 and ν_0 are, respectively, the degrees of freedom associated with the numerator and denominator of the sample variances being compared. Now the probability density of the GTDL model, may be written as:

$$f(t) = \lambda_0\psi(\gamma)^{\frac{\lambda_0}{\alpha}}\left\{e^{-(\alpha t+\gamma)}\right\}^{\frac{\lambda_0}{\alpha}}\left\{1 + e^{-(\alpha t+\gamma)}\right\}^{-(1+\frac{\lambda_0}{\alpha})} \tag{6}$$

Apart from a constant, the functional form of (6) is the same as (5), but the range is $0 < t < \infty$. It follows immediately that $t = \mid z \mid$. Thus we reach the important conclusion that the GTDL is the *modulus* of a special case of Ahuja and Nash's 4-parameter logistic, or, equivalently, of a special case of Fisher's Z distribution.

3 Canonical Form

From (1), $\partial\lambda(t;x)/\partial\lambda_0 = \theta(t;x)$ and $\partial\lambda(t;x)/\partial\beta_0 = \lambda_0.x_{oi}[\theta(t;x) - \theta^2(t;x)]$. Since $\theta(t;x) < 1$ and small, $\theta^2(t;x) \approx 0$ and it follows that $\partial\lambda(t;x)/\partial\lambda_0 \approx \lambda_0\partial\lambda(t;x)/\partial\beta_0$ for each i and $\forall \;\; t$, showing that there is no new information contained in β_0 These conditions often occur in practice (Aalen, 1988), and since the roles of the parameters λ and β_0 are then inter-changeable only one is required in the model.

The rationale for the original extension from the TDL to the GTDL model was predicated on the need to remove the inconvenient constraint on the hazard imposed by the TDL model, namely that $0 < \lambda(t;x) \leq 1$, for $\forall \;\; t \geq 0$. This led to the incorporation of $\lambda_0 > 0$ as a multiplier on the hazard function in (1). For this essential generalization to succeed, λ_0 would need to be retained and β_0 dropped. Thus, the canonical form of the regression model has a hazard function given by:

$$\lambda^*(t;x) = \lambda_0^*\theta^*(t;x^*) \tag{7}$$

where $\lambda_0^* > 0$ is a scalar, $\theta^*(t;x^*) = \exp(t\alpha^* + x^*\beta^*/)/[1 + \exp(t\alpha^* + x^*\beta^*)]$ is a linear logistic function in time, α^* is a scalar measuring the effect of time and $\beta^* = (\beta^*{}_1, ..., \beta_p^*)$ is a column vector of p unknown regression

parameters measuring the influence of p covariates $x^* = (x_1^*, ..., x_p^*)$. The time-dependent *relative risk* for two distinct covariate profile vectors x_1^* and x_o^* is:

$$\rho(t; x_1^*, x_0^*) = \exp[(x_1^* - x_0^*)\beta]\nu(t, x_1^*, x_0^*) \tag{8}$$

where $\nu(t; {x_1}^*, x_0^*) = [1 + \exp(t\alpha + x_0^*\beta)]/[1 + \exp(t\alpha + x_1^*\beta)]$. The leading term on the right, Cox's relative risk, is therefore moderated by $\nu(.)$, a function of time and the covariates.

This modification frees the hazard function from the unit interval, renders all parameters estimable and provides the statistical community with a simple parametric non-PH competitor for Cox's PH model.

4 Example

The data analyzed form part of a population-based prospective study of incident cases of lung cancer diagnosed in Northern Ireland in one year. This multi-source study identified 900 incident cases during the year studied. Outcome was missing in 25 cases and another 20 cases were diagnosed at post-mortem. Thus, there were 855 valid cases and we analyzed 'Time from Diagnosis to Death or Censoring' in relation to a range of clinical and demographic factors measured on each patient.

We implemented the new model in SPSS (V11) and in SPLUS (V6) and in the main paper we compare the results and fits obtained using the canonical form of the logistic regression model with those from the PH model, especially in high dimensional covariate space where the lung cancer data are non-PH. Some preliminary findings have been presented but a fuller account will be included in the main paper. We will also include the results of a small simulation study to demonstrate the advantages of (7) over the PH model.

References

Aalen O.O (1988). Heterogeneity in Survival Analysis *Statistics In Medicine* 7, 1121-1137.

Ahuja JC & Nash SW (1967). The generalized Gompertz-Verhulst family of distributions. *Sankhya A* 29, 141-156.

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.