Computational Statistics and Data Analysis 74 (2014) 52–63

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Analysis of multivariate survival data with Clayton regression models under conditional and marginal formulations

W. He *

Department of Statistical & Actuarial Sciences, University of Western Ontario, 1151 Richmond Street North, London, Ontario, Canada N6A 5B7

ARTICLE INFO

Article history:
Received 2 October 2012
Received in revised form 9 November 2013
Accepted 1 January 2014
Available online 9 January 2014

Keywords:
Clayton models
Conditional models
Covariate coefficient
Marginal models
Multivariate survival data

ABSTRACT

The Clayton models, also called gamma frailty models, have been widely used for multivariate survival analysis. These models typically appear in either conditional or marginal formulations where covariates are incorporated through regression models. The two formulations provide us the flexibility to delineate various types of dependence of survival times on covariates, along with the availability of directly applying the likelihood method for inferences if the baseline hazard functions are parametrically or weakly parametrically specified. There are, however, fundamental issues pertaining to these models. It is not clear how the covariate effects in the two formulations are related to each other. What is the impact if misusing the conditional formulation when the true form should be marginal, or vice versa? These problems are investigated, and the relationship of the covariate coefficients between conditional and marginal regression models is established. Furthermore, empirical studies are carried out to assess how censoring proportion may affect estimation of covariate coefficients. A real example from the Busselton Health Study is analyzed for illustration.

1. Introduction

Multivariate survival data arise commonly in biomedical research, clinical trials and epidemiological studies. Different from univariate survival analysis, multivariate survival analysis typically deals with various association structures among survival times within same subjects or clusters. A common strategy is to use latent variables to delineate association among multivariate survival times. Conditional on the latent variables, or called frailties, the survival times are assumed independent. These models are flexible and effective in characterizing different types of association structures among multivariate survival times. For instance, Hougaard (1986) investigates a class of frailty models by assuming different distributions for the frailty. Recently, frailty models are extended to incorporate settings such as transformation models (Zeng et al., 2009) and additive hazards models (Martinussen et al., 2011).

Among various frailty models, one of the most popularly used frailty models is the gamma frailty model, or the so-called Clayton model, for which the frailty assumes a gamma distribution (Clayton, 1978; Oakes, 1982, 1986, 1989; Hougaard, 2000). To be specific, let $(T_1, \ldots, T_m)$ be multivariate survival times, and α be a latent variable (or frailty) that features dependence among the $T_j, j = 1, \ldots, m$. That is, conditional on α, the survival times $(T_1, \ldots, T_m)$ are assumed independent. Often the frailty α is assumed to follow a gamma distribution, $\alpha \sim Gamma(\phi, \phi)$, with mean 1 and variance ϕ^{-1} without loss of generality (Lawless, 2003).

Given a frailty α, let the conditional survivor function $S_j(t_j|\alpha)$ for T_j be written as

$$S_j(t_j|\alpha) = P(T_j > t_j|\alpha) = \exp\{-\alpha \cdot \Lambda_j^*(t_j)\},$$

* Tel.: +1 519 661 2111x86982; fax: +1 519 661 3813.
E-mail address: whe@stats.uwo.ca.

0167-9473/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2014.01.001

where $\Lambda_j^*(t_j)$ is referred to as a conditional basic cumulative hazard function. The joint survivor function for the Clayton model is then derived as

$$\begin{aligned} S(t_1,\ldots,t_m) &= \Pr(T_1 > t_1,\ldots,T_m > t_m) \\ &= \int_0^\infty \{S_1(t_1|\alpha)\cdots S_m(t_m|\alpha)\}dG(\alpha;\phi) \\ &= \left[1+\frac{1}{\phi}\left\{\Lambda_1^*(t_1)+\cdots+\Lambda_m^*(t_m)\right\}\right]^{-\phi}, \end{aligned} \tag{1}$$

where $G(\alpha;\phi)$ represents the cumulative distribution function of the gamma distribution for α.

Note that the marginal survivor function for T_j can be obtained by $S_j(t_j) = \int_0^\infty S_j(t_j|\alpha)dG(\alpha;\phi)$, yielding

$$S_j(t_j) = \left\{1+\frac{1}{\phi}\Lambda_j^*(t_j)\right\}^{-\phi}, \quad j = 1,\ldots,m. \tag{2}$$

We can then alternatively write the Clayton model as, using the marginal survivor functions

$$S(t_1,\ldots,t_m) = \{S_1(t_1)^{-\frac{1}{\phi}}+\cdots+S_m(t_m)^{-\frac{1}{\phi}}-(m-1)\}^{-\phi}. \tag{3}$$

Model (1) gives us a conditional representation of the joint survivor function through $\Lambda_j^*(t_j)$ while model (3) provides a marginal expression of the joint survivor function via $S_j(t_j)$. Although models (1) and (3) involve the dependence parameter ϕ differently, they both offer us a way to delineate the dependence of survival times on covariates. Let $\boldsymbol{X}_j$ be the covariates associated with survival times $T_j, j = 1,\ldots,m$. We now employ regression models to feature $\Lambda_j^*(t_j)$ or $S_j(t_j)$ in order to spell out covariate effects on survival processes. In particular, we consider multiplicative regression models

$$\Lambda_j^*(t_j|\boldsymbol{X}_j) = \Lambda_{0j}^*(t_j)\exp(\boldsymbol{X}_j'\beta^*), \tag{4}$$

or

$$\Lambda_j(t_j|\boldsymbol{X}_j) = \Lambda_{0j}(t_j)\exp(\boldsymbol{X}_j'\beta), \tag{5}$$

where $\Lambda_{0j}^*(t_j)$ is the basic baseline cumulative hazard function, $\Lambda_{0j}(t_j)$ is the marginal baseline cumulative hazard function with $S_j(t_j|\boldsymbol{X}_j) = \exp\{-\Lambda_j(t_j|\boldsymbol{X}_j)\}$, and β^* and β are the respective regression parameters.

Conditional representation (1) with (4) and marginal expression (3) with (5) enable us to conveniently modulate the dependence of multivariate survival times on relevant covariates. Based on these model formulations, inference on covariate coefficients becomes straightforward as one can directly invoke the maximum likelihood method when the cumulative hazard functions Λ_{0j}^* or Λ_{0j} are modeled. Due to this reason, the Clayton models with formulations (1) or (3) have been commonly used in practice to analyze multivariate survival data (e.g, Clayton and Cuzick, 1985).

A couple of important issues pertaining to the Clayton models, however, remain unclear. For instance, which representation is preferred, conditional form (1) or marginal form (3)? If (1) with (4) is the correct model, but we misspecify our working model as (3) with (5), or vice versa, what would be the impact? How different are the covariate coefficients β^* and β in terms of their interpretation? Furthermore, how would the association strength and censoring of survival times affect the estimation of covariate effects? In this paper we address these concerns and provide useful insights into the two formulations (1) and (3) for the Clayton models. For ease of exposition, we focus discussion on the bivariate Clayton models. The results can be extended to the general multivariate Clayton models in a straightforward manner.

The remainder of the manuscript is organized as follows. In Section 2, we derive the relationship between the marginal and conditional covariate coefficients for the bivariate Clayton models. In Section 3, we discuss the likelihood method that facilitates the effects of association strength and censoring on estimation of covariate coefficients. In Section 4, we report simulation studies and analysis results for data from the Busselton Health Study. Concluding remarks are included in the last section.

2. Clayton models and covariate coefficients

2.1. Clayton models

For bivariate survival times T_1 and T_2, both conditional model (1) and marginal model (3) can be employed to formulate the joint survivor function $S(t_1, t_2)$. The covariates $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ can be incorporated in $S(t_1, t_2)$ through conditional regression model (4) and marginal regression model (5). Although both (4) and (5) feature covariate effects on survival times, their interpretation is quite different. The covariate coefficients β^* in conditional model (4) emphasize the effects of covariates on individuals, while the β in marginal model (5) give the average covariate effects at the population level.

Employing the conditional proportional hazards (PH) specification (4) to (1) gives a bivariate survivor function:

$$S_c(t_1,t_2|\boldsymbol{X}) = \left\{1+\frac{1}{\phi}\Lambda_{01}^*(t_1)\exp(\boldsymbol{X}_1'\beta_1^*)+\frac{1}{\phi}\Lambda_{02}^*(t_2)\exp(\boldsymbol{X}_2'\beta_2^*)\right\}^{-\phi}, \tag{6}$$