Statistical Methodology 29 (2016) 18–31

Contents lists available at ScienceDirect

Statistical Methodology

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

Regression analysis of competing risks data with general missing pattern in failure types

Anup Dewanji [a,*], P.G. Sankaran [b], Debasis Sengupta [a], Bappa Karmakar [a]

[a] Applied Statistics Unit, Indian Statistical Institute, Kolkata 700108, India

[b] Cochin University of Science and Technology, Cochin 682022, India

ARTICLE INFO

Article history:
Received 19 December 2014
Received in revised form
1 July 2015
Accepted 6 September 2015
Available online 9 October 2015

Keywords:
Competing risks
Cause-specific hazard
Missing failure type
Missing at random
Nelson–Aalen estimator
Semi-parametric model

ABSTRACT

In competing risks data, missing failure types (causes) is a very common phenomenon. In a general missing pattern, if a failure type is not observed, one observes a set of possible types containing the true type along with the failure time. Dewanji and Sengupta (2003) considered nonparametric estimation of the cause-specific hazard rates and suggested a Nelson–Aalen type estimator under such general missing pattern. In this work, we deal with the regression problem, in which the cause-specific hazard rates may depend on some covariates, and consider estimation of the regression coefficients and the cause-specific baseline hazards under the general missing pattern using some semi-parametric models. We consider two different proportional hazards type semi-parametric models for our analysis. Simulation studies from both the models are carried out to investigate the finite sample properties of the estimators. We also consider an example from an animal experiment to illustrate our methodology.

1. Introduction

In survival studies, the failure (or, death) may be attributed to one of several causes or types, known as competing risks. In such situations, for each individual, we observe a random vector (T, J), where T

* Corresponding author.
E-mail address: dewanjia@isical.ac.in (A. Dewanji).

http://dx.doi.org/10.1016/j.stamet.2015.09.002
1572-3127/

is possibly censored survival time and J represents the cause of death (exactly one of m possible causes, say). However, due to inadequacy in the diagnostic mechanism, often there is uncertainty about the true failure type and so the experimentalists are reluctant to report any specific value of J for some individuals. This is usually known as the problem of missing failure type in competing risks and has been addressed by many authors. For example, in carcinogenicity studies, besides deaths (failures) without tumor, there are deaths (with tumor present) due to either the tumor itself or some other cause. Often there is uncertainty in assigning this cause of death even if the presence of tumor can be ascertained [11,19]. In extreme situations, one cannot even ascertain presence or absence of tumor because it is totally cannibalized or autolyzed (see Section 6 for details).

Analysis of competing risks data with missing failure types was first considered by Dinse [12] with the assumption that failure type was either completely available (that is, observed as exactly one of m possible types) or not available at all (that is, unobserved failure type is any one of the m possible types). This problem was subsequently studied by different researchers (see [21,23,24,27]). Goetghebeur and Ryan [16,17] considered the regression problem for two failure types under the assumption that the cause-specific hazards for two failure types are proportional. The method of partial likelihood was employed for estimating the regression parameter. See also Dewanji [9] and Lu and Tsiatis [22] for similar work.

In the above-mentioned works, missingness meant that no information on failure type was available at all. However, in many contexts, one may be able to narrow down to fewer than m causes to be responsible for failure. In the present work, we consider a general missing pattern. Here, for each individual failure, we observe the survival time and a subset $g \subseteq \{1, \ldots, m\}$ of labels of possible failure types, exactly one of which is the true but unobserved cause of failure (see Section 6 for an example). When g is a singleton set, the failure type is exactly observed, and when $g = \{1, \ldots, m\}$, the missingness is total. It is usually said that the true failure type is masked in the set g. Flehinger et al. [13] considered such general pattern of missing failure types for the purpose of estimating survival due to different types, with the strong assumption of proportional cause-specific hazards. They also assumed that, for some of the observations with missing failure type, a second stage diagnosis can be performed to pinpoint the type. Flehinger et al. [14] considered the same problem using a parametric modeling but without assuming proportional hazards. Craiu and Duchesne [7] suggested an estimation procedure using EM algorithm based on piecewise constant cause-specific hazard rates. Under a missing-at-random type assumption and requiring a second stage diagnosis, they developed an EM algorithm to estimate the piecewise constant cause-specific hazard rates and the diagnostic probabilities of the actual cause of failure being j, given the set g of observed possible causes. See also Craiu and Reiser [8]. Dewanji and Sengupta [10], in addition to suggesting a nonparametric estimator using EM algorithm, developed a Nelson–Aalen type estimator of the cumulative cause-specific hazard rates (and also a smooth estimator of the cause-specific hazard rates), when certain information on the diagnostic probabilities are available from the experimentalists, but the missing pattern could be allowed to be non-ignorable and no second stage diagnosis was required.

In this work, we deal with the regression problem, in which the cause-specific hazard rates may depend on some covariates, and consider estimation of the regression coefficients under some proportional hazards type semi-parametric models, when observation on the failure type exhibits the general missing pattern as discussed before. Recently, Chatterjee et al. [6] have considered a similar problem in the context of partially observed disease classification data with possibly large number of types. They have suggested a two-stage modeling in which the first stage involves reducing the number of parameters by imposing a natural structure on the underlying disease types and the second stage involves inference through a general extension of the partial likelihood based estimating equation (see [17]). Apparently, however, they need to make certain assumptions regarding the missing probabilities like most of the work on this issue. Also, Sen et al. [28] have developed a semiparametric Bayesian approach, where the partial information about the cause of death is incorporated by means of latent variables, and proposed a simulation-based method using Markov Chain Monte Carlo (MCMC) techniques to implement the Bayesian methodology.

We also consider estimation of the cumulative baseline hazards in the spirit of Dewanji and Sengupta [10]. In Section 2, we describe the data and two semi-parametric models to study the effect of covariates. In Section 3, we consider estimation of the regression coefficients and the cumulative