Journal of Multivariate Analysis 149 (2016) 160–176

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

# Composite partial likelihood estimation for length-biased and right-censored data with competing risks

Feipeng Zhang [a], Heng Peng [b,*], Yong Zhou [c,d]

[a] School of Finance and Statistics, Hunan University, Changsha 410082, China
[b] Department of Mathematics, Hong Kong Baptist University, Hong Kong
[c] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China
[d] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

*Article history:*
Received 23 February 2015
Available online 26 April 2016

*AMS subject classifications:*
62N01
62N02
62G09

*Keywords:*
Competing risks
Cumulative incidence function
Composite partial likelihood
Length-biased and right censored data
Resampling

ABSTRACT

This paper considers a competing risks model for survival data from length-biased sampling, where the survival times are left truncated by uniformly distributed random truncation times. We propose a composite partial likelihood estimating procedure for cause-specific failure probabilities using competing risks data. We establish the asymptotic properties of the proposed estimators, and present predictions of the cumulative incidence functions. Furthermore, we show how to construct simultaneous confidence bands for the cause-specific cumulative incidence functions for subjects with given risk factors. A simulation study demonstrates that the proposed estimators have good finite-sample performance. A real data example illustrates the method and the theory.



## 1. Introduction

Prevalent cohort designs are widely used to study time-to-event outcomes, due to their simplicity and cost-effectiveness. In addition to the right censoring caused by the loss of follow-up data, lifetime data sampled in a cross-sectional fashion are subject to left truncation, as those who fail before the recruitment time are unobservable, and hence the lifetime of a selected individual may be longer than the left truncation time. In prevalent cohort studies, the truncation distribution explains patterns of disease incidence and selection bias (see, e.g., [5,21,35]). Hence, failure to account for left truncation will result in a substantial bias in the estimation of the lifetime distribution (see [15,26]). If the incidence of disease onset follows a stationary Poisson process, so that the incidence rate remains constant over time (see [35]), the truncation time will follow a uniform distribution. As a result, the probability of observing a lifetime value is proportional to the value itself. In such a situation, the selection bias due to left truncation is called length bias. We use "length-biased sampling" to refer to prevalent sampling under the assumption of stationarity of disease incidence.

In time-to-event datasets arising from prevalent cohorts, it is often observed that the subjects under study are at risk of more than one exclusive event, such as failure from different causes. Such data are known as competing risks data. For example, in cancer studies, a patient may fail the therapy through a recurrence of the disease (relapse) or by death due to the toxicity of the therapy itself. In this case, the competing causes of treatment failure are cancer relapse and treatment-related

* Corresponding author.
*E-mail address:* hpeng@math.hkbu.edu.hk (H. Peng).

http://dx.doi.org/10.1016/j.jmva.2016.04.002
0047-259X/

mortality (TRM). In the competing risks setting, it is of great interest to estimate the cumulative incidence function, i.e., the cumulative probability of occurrence by time $t$ for a particular type of event in the presence of other risks. In classic survival estimation methods, a natural and simple concept is the marginal hazard of the event of interest, which is the hazard in the situation that the other events do not occur. Then only one event of interest is chosen for analysis and the competing causes of failure are ignored and treated as right-censored observations. This can only be estimated nonparametrically, under the strictly hypothetical assumption that the distribution of the time to the event is independent of the distribution of the time to the other event (see [27]). However, the violation of the assumption of independence, which is nonidentifiable (see [33]), is an important issue in competing risks models. An alternative and commonly used method is to use the Cox model to estimate the covariate effects on the cause-specific hazard function for the event time of interest (see [9]). Under the assumption that there are no constraints on the regression coefficients corresponding to different causes of failure and there are no ties between observations (see [2]), the cause-specific hazard can measure the instantaneous failure rate due to one risk at a time, and it does not need the assumption of independence.

Although for a classic survival analysis with a single type of event, many authors have proposed semiparametric methods for applying the Cox model under length-biased sampling, (see, [12,17,29,30,32,36,37], and others), the goal of this paper is to develop an estimate procedure for competing risks data under length-biased sampling. There are two challenges to estimating the covariate effects on the cause-specific hazard rates under length-biased sampling with competing risks. One challenge is that the length-biased sampling scheme could induce informative censoring and change the model structure. Due to the length-biased sampling, the potential dependence between the failure time and the right-censoring time could be very strong. Moreover, the model structure assumed for a target population is often different from that of observed length-biased data with competing risks. A second challenge is modeling the improper cumulative incidence function for one event of interest with dependent censoring from competing events. It is worth pointing out that there is no longer a one-to-one correspondence between the effect of a covariate on the cause-specific hazard and its effect on the cumulative scale. This is contrary to the classic survival setting with only one type of event. The cumulative incidence function depends on the rate of occurrence of all of the risks, see Gray [13], Putter et al. [28]. The classic regression analysis of competing risks models the cause-specific hazard function based on the proportional hazards model proposed by Cox [9]. There is a rich literature on the estimation of the cause-specific cumulative incidence function under the proportional hazards model with competing risks that are only subject to right censoring (see [6,8,20,27], and their references). Readers are referred to Haller et al. [14] for a comprehensive review of different approaches to the analysis of time-to-event data in the presence of competing risks.

To predict the cumulative incidence function for a subject with certain covariates, the commonly used approach is to combine estimates of the cause-specific hazard functions from the partial likelihood approach with an adjusted risk set generalized to the left-truncated version (see [2]). However, this approach ignores the information of the length-biased data structure, and is expected to be inefficient. This paper introduces a simple and efficient inference procedure for the proportional hazards model with competing risks data from length-biased-sampling. Our approach is motivated by the recent work of Huang and Qin [16]. In that paper, the authors explored the application of the composite likelihood method (see [4,25]) for estimating the Cox model with classic survival data collected under length-biased sampling, but they did not consider the effect of competing risks in the model.

The rest of this paper is structured as follows. In Section 2, we use the cumulative incidence function to extend the approach proposed by Huang and Qin [16] to competing risks data. We propose the partial likelihood method and the composite partial likelihood method for comparison. The large sample properties of the proposed estimators are also derived in this section. In Section 3, we derive formula for predicting a cumulative incidence function. We also evaluate the performance of the proposed estimator through simulation studies and a real data analysis in Section 4. Technical proofs are given in the Appendix.

## 2. Estimation procedures

### 2.1. Data and notation

For the prevalent population with competing risks, let $W^0$ denote the calendar time of the disease incidence, $T^0$ denote the time from the disease incidence to the failure event, and $\mathbf{Z}^0$ denote a $p \times 1$ vector of covariates. Let $\epsilon^0 \in \{1, \ldots, K\}$ be the cause of the failure. We impose the following two assumptions on the incident population throughout the paper.

**Assumption 1.** The variables $(T^0, \epsilon^0, \mathbf{Z}^0)$ are independent of $W^0$, and $\epsilon^0$ is independent of $(T^0, W^0)$ given $\mathbf{Z}^0$.

**Assumption 2.** Disease incidence occurs over calendar time at a constant rate, that is, $W^0$ has a constant density function.

We assume that the sampling time $\xi$ is independent of $(W^0, T^0, \epsilon^0, \mathbf{Z}^0)$. As a prevalent population, those individuals with the disease who have not experienced the failure event at the sampling time are observed. That is, an individual is sampled at time $\xi$ only if $T^0 \geq A^0 > 0$, where $A^0 = \xi - W^0$. Let $(W, T, \epsilon, \mathbf{Z})$ denote the random variables from the prevalent population. Hereafter, the superscript $^0$ is dropped to emphasize that the failure time $T$ in the prevalent population must exceed $A = \xi - W$, which is thus a left-truncated random variable. The probability distribution of $(W, T, \epsilon, \mathbf{Z})$ is the same as the probability distribution of $(W^0, T^0, \epsilon^0, \mathbf{Z}^0)$, conditional on $T^0 \geq A^0 > 0$.