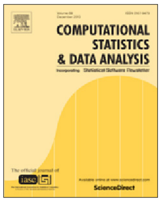
Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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



Analysis of generalized semiparametric regression models for cumulative incidence functions with missing covariates

Unkyung Lee [a], Yanqing Sun [b,*], Thomas H. Scheike [c], Peter B. Gilbert [d,e]

[a] Department of Statistics, Texas A&M University, College Station, TX 77843, USA
[b] Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA
[c] Department of Biostatistics, University of Copenhagen, Øster Farimagsgade 5, DK 1014, Denmark
[d] Department of Biostatistics, University of Washington, Seattle, WA 98195, USA
[e] Vaccine and Infectious Disease and Public Health Sciences Divisions, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

ARTICLE INFO

Article history:
Received 18 April 2017
Received in revised form 10 January 2018
Accepted 12 January 2018
Available online xxxx

Keywords:
Competing risks
Inverse probability weighted complete-case
Semiparametric regression model
RV144 vaccine efficacy trial
Time-varying effects
Two-phase sampling

ABSTRACT

The cumulative incidence function quantifies the probability of failure over time due to a specific cause for competing risks data. The generalized semiparametric regression models for the cumulative incidence functions with missing covariates are investigated. The effects of some covariates are modeled as nonparametric functions of time while others are modeled as parametric functions of time. Different link functions can be selected to add flexibility in modeling the cumulative incidence functions. The estimation procedures based on the direct binomial regression and the inverse probability weighting of complete cases are developed. This approach modifies the full data weighted least squares equations by weighting the contributions of observed members through the inverses of estimated sampling probabilities which depend on the censoring status and the event types among other subject characteristics. The asymptotic properties of the proposed estimators are established. The finite-sample performances of the proposed estimators and their relative efficiencies under different two-phase sampling designs are examined in simulations. The methods are applied to analyze data from the RV144 vaccine efficacy trial to investigate the associations of immune response biomarkers with the cumulative incidence of HIV-1 infection.

1. Introduction

Competing risks data are often encountered in medical researches where study participants are exposed to two or more mutually exclusive causes of failure. Because of the inherited nonidentifiability problem with the competing risks data, the cause-specific hazard function and cumulative incidence function have been used as the primary quantities for analyzing competing risks data (Tsiatis, 1978; Kalbfleisch and Prentice, 1980). The cause-specific hazard function measures the instantaneous failure risk due to a specific cause while the cumulative incidence function describes the probability distribution for a specific cause of failure, both are the measurements of the crude failure risk in the presence of all other risks. The two measures provide different perspectives for cause-specific failure times, cf. Fine and Gray (1999), Katsahian et al. (2006) and Latouche et al. (2007).

The statistical methods developed for failure time hazard regression models such as Cox (1975), Lin and Ying (1994) and Scheike and Zhang (2002) can often be utilized directly for modeling the cause-specific hazard functions. The cumulative

[*] Corresponding author.
E-mail address: yasun@uncc.edu (Y. Sun).

https://doi.org/10.1016/j.csda.2018.01.003
0167-9473/© 2018 Elsevier B.V. All rights reserved.

Please cite this article in press as: Lee U., et al., Analysis of generalized semiparametric regression models for cumulative incidence functions with missing covariates. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.01.003.

incidence function $F_j(t)$ for cause j relates the cause-specific hazard function $\lambda_j(t)$ through the simple formula $F_j(t) = \int_0^t \lambda_j(s)S_T(s)\,ds$, where $S_T(t)$ is the survival function of T of all causes. Cumulative incidence functions can be estimated through modeling the cause-specific hazard functions for all causes (Cheng et al., 1988; Shen and Cheng, 1999). This approach provides a straightforward way to estimate cumulative incidence functions, but the effects of covariates on the cumulative incidence functions are not clear. In addition, indirect modeling $F_j(t)$ through the cause-specific hazard functions requires setting up models for the cause-specific hazard functions of all causes. Alternative approaches that directly model the cumulative incidence functions have been studied by Klein and Andersen (2005) and Scheike et al. (2008). Direct modeling of the cumulative incidence functions allows direct evaluation of covariate effects on the probability of failure over time for a specific cause without the need to model for other causes.

This paper investigates the generalized semiparametric additive model for cumulative incidence functions with missing covariate values. This research is motivated by RV144, a preventive HIV-1 vaccine efficacy trial. RV144 randomized 16,395 HIV-1 negative volunteers in 1:1 allocation to receive vaccine or placebo and followed them for 42 months for occurrence of the primary study endpoint of HIV-1 infection (Rerks-Ngarm et al., 2009), showing partial beneficial efficacy of the vaccine to lower the incidence of HIV-1 infection. An objective of RV144 is to examine the effects of certain biomarkers measuring immune responses to vaccination on the cumulative probability of becoming infected with specific genetic types of HIV-1 (the competing risks of failure). The immune response biomarkers were measured from vaccine recipients at the Week 26 visit, two weeks after the vaccination series. Since the rate of HIV-1 infection was low (fewer than 1% of participants acquired HIV), it would be very expensive and unnecessary to measure the biomarkers from all vaccine recipients. A classical case-cohort design would measure the biomarkers from all participants who experience the event of interest after Week 26 (HIV-1 infection) and from a relatively small random sample from the original cohort of vaccine recipients (Prentice, 1986). However, since the immune response biomarkers were not measured from all failure events after Week 26, we follow the generalized case-cohort design, which is in the form of two-phase sampling data (Breslow et al., 2009a, b; Haneuse et al., 2011), where the phase-one data are variables measured from all participants and the phase-two data are measured in random samples (without replacement or Bernoulli) of participants within each level of a stratification variable defined by the phase-one data. The case-cohort or two-phase sampling of covariate data are common forms of missing covariates. In the application section we describe the specific genetic types of HIV-1 infection that constitute the failure types of interest.

Our estimation procedure for generalized semiparametric additive models is based on the direct binomial regression model of Scheike et al. (2008) using an inverse probability weighting of complete cases (IPW) estimating equation. This approach modified the full data weighted least squares equations by weighting the contributions from participants with fully observed data by the inverses of the estimated sampling probabilities. Under the competing risks situations, the sampling probabilities may be different for different types of events. Previous related methodological research includes the following papers. Kang and Cai (2009) developed methods for fitting failure time data from case-cohort studies with multiple disease outcomes under a marginal proportional hazards model. Kang et al. (2013) developed a weighted estimating equations approach for the marginal additive hazards regression model for case-cohort studies with multiple disease outcomes. Sun et al. (2017) studied a semiparametric additive hazards model for case-cohort and two-phase sampling data. We transport this approach to develop an IPW estimating equations procedure for the generalized semiparametric additive model in Section 2. We develop asymptotic properties of the proposed IPW estimator in Section 3. The finite-sample performances of the IPW estimators and their relative efficiencies are examined in Section 4 for different two-phase sampling designs. The proposed method is applied to analyze data from the RV144 vaccine efficacy trial in Section 5.1 to investigate the association of antibody responses with the vaccine on the cumulative incidence of infection with HIVs of specific genetic types of interest. An additional simulation study based on the RV144 data is conducted in Section 5.2 to examine the performance of the IPW method when the sample size is very large, and both the event rate and sampling percentage of phase-two variables are very small. Proofs of the asymptotic results are given in the Appendix.

2. Estimation of the semiparametric model for the cumulative incidence function with missing covariates

2.1. Generalized semiparametric model and missing data

Let T_i be the failure time and $J_i \in \{1, 2, \ldots, k\}$ denote the k different failure types for the ith subject. Assume that cause $J_i = 1$ is the primary cause of interest and $J_i > 1$ for other competing causes. Let $X_i = (1, X_{i1}, \ldots, X_{ip})^T$ and $Z_i = (Z_{i1}, \ldots, Z_{iq})^T$ be the $(p+1)$- and q-dimensional possibly time-dependent covariate vectors, respectively. Let C_i be the right censoring time. Let $\Delta_i = I(T_i \le C_i)$ be an indicator for uncensored failure time. The observed independent identically distributed (i.i.d.) competing risks data can be represented by $Y_i = (X_i, Z_i, \tilde{T}_i, \tilde{J}_i)$ for $i = 1, 2, \ldots, n$, where $\tilde{T}_i = \min(T_i, C_i)$ and $\tilde{J}_i = J_i\Delta_i$. The value $\tilde{J}_i = J_i$ indicates that the failure time is observed at $\tilde{T}_i$ and the cause of failure is of type J_i. Let $[0, \tau]$ be the follow-up period during which data are collected. We assume that the (T_i, J_i) are independent of the C_i given the covariates (X_i, Z_i). The covariates of subject i are only meaningful in the time interval when the subject is at-risk and still in the study, i.e., $t \le \tilde{T}_i$.

Let $F_1(t; X_i, Z_i) = P(T_i \le t, J_i = 1 | X_i, Z_i)$ be the conditional cumulative incidence function given covariates (X_i, Z_i). We consider the following generalized semiparametric model:

$$F_1(t; X_i, Z_i) = \boldsymbol{h}\{X_i^T\eta(t), g(\gamma, Z_i, t)\} \tag{1}$$

Please cite this article in press as: Lee U., et al., Analysis of generalized semiparametric regression models for cumulative incidence functions with missing covariates. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.01.003.