Computational Statistics and Data Analysis 117 (2018) 194–206

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Regression analysis for the proportional hazards model with parameter constraints under case-cohort design

Lifeng Deng [a], Jieli Ding [b,*], Yanyan Liu [b], Chengdong Wei [c]

[a] School of Mathematics and Statistics, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China
[b] School of Mathematics and Statistics, Wuhan University, Wuhan, Hubei 430072, China
[c] School of Mathematics and Statistics, Guangxi Teachers Education University, Nanning, Guangxi 530023, China

ARTICLE INFO

Article history:
Received 12 February 2017
Received in revised form 30 August 2017
Accepted 30 August 2017
Available online 6 September 2017

Keywords:
Case-cohort design
Proportional hazards model
Constrained estimation
Karush–Kuhn–Tucker conditions
Minorization–maximization algorithm

ABSTRACT

To reduce the cost and improve the efficiency of cohort studies, case-cohort design is a widely used biased-sampling scheme for time-to-event data. In modeling process, case-cohort studies can benefit further from taking parameters' prior information, such as the histological type and disease stage of the cancer in medical, the liquidity and market demand of the enterprise in finance. Regression analysis of the proportional hazards model with parameter constraints under case-cohort design is studied. Asymptotic properties are derived by applying the Lagrangian method based on Karush–Kuhn–Tucker conditions. The consistency and asymptotic normality of the constrained estimator are established. A modified minorization–maximization algorithm is developed for the calculation of the constrained estimator. Simulation studies are conducted to assess the finite-sample performance of the proposed method. An application to a Wilms tumor study demonstrates the utility of the proposed method in practice.

1. Introduction

In many large cohort studies, the measurements of primary exposure variables can be prohibitively expensive. For time-to-event data, the case-cohort design is one of the most widely used cost-effective strategies, especially when the event rate is low. Under a case-cohort design, the complete information of exposure variables is only assembled for a random sample from the entire cohort (subcohort) and subjects outside the subcohort who experience the event (cases).

In the landmark article of Prentice (1986), case-cohort design was first formally proposed, a pseudo-likelihood method was established for estimation of regression parameters. Since the publication of Prentice (1986), there are numerous and extensive studies on case-cohort design and related statistical methodologies, including likelihood-based methods (Self and Prentice, 1988; Chen, 2001; Lu and Shih, 2006; Tsai, 2009) and estimating equation methods (Chen and Lo, 1999; Kulich and Lin, 2000; Qi et al., 2005; Kang and Cai, 2009), among others.

In practice, some prior information on parameters with a certain of constraints may be available in the modeling process. It is reasonable to take these constraints into account, since ignoring such information may cause an underestimate of parameters' uncertainty and a misled conclusion (Tan et al., 2005; Fang et al., 2006). There are many research on statistical inferences for constrained problems. For completely observed data, Wang (1996, 2000) derived asymptotic properties of restricted estimators in nonlinear regressions. Moore and Sadler (2006) and Moore et al. (2008) established asymptotic

* Corresponding author.
E-mail address: jlding.math@whu.edu.cn (J. Ding).

http://dx.doi.org/10.1016/j.csda.2017.08.013
0167-9473/

theory for the constrained maximum likelihood estimator. For data with censoring, Ding et al. (2015) developed inference methods for the proportional hazards model with parameter constraints.

However, these existed studies are for simple random sampling observations. To the best of our knowledge, statistical methodologies for model parameter with constraints have not yet been explored for failure time data from case-cohort design. In this paper, we develop how to fit the proportional hazards model with parameter restrictions to failure time data from case-cohort studies. The theoretical developments are challenging because of the presence of constraints. To overcome the difficulties, we appeal for the Karush–Kuhn–Tucker conditions (Boyd and Vandenberghe, 2004), a well-known approach in optimization with inequality constraints, to establish the asymptotic properties of the constrained estimator.

Another challenge arises from the numerical implementation of the constrained estimator. To this end, we adopt a minorization–maximization (MM) algorithm, the essential idea of which is to create a surrogate function with computational superiority over the objective function in order to achieve optimization transfer (De Pierro, 1995; Becker et al., 1997; Lange et al., 2000; Hunter and Lange, 2004; Lange, 2004, 2010). It is worth noting that the original MM algorithm cannot be applied to the cases that model parameters are restricted. Ding et al. (2015) proposed a new MM algorithm for the computation of the constrained estimator under the proportional hazards model. Taking the spirit, we develop a modified MM algorithm for the constrained estimator under the case-cohort design by replacing the risk sets of the entire cohort involved in the surrogate function proposed by Ding et al. (2015) with their subcohort counterparts.

The remainder of this paper is organized as follows. We fit data from case-cohort design to the proportional hazards model with constraints in Section 2, and derive the asymptotic properties for the constrained estimator in Section 3. In Section 4, we propose a modified MM algorithm for implementation of the constrained estimation, and present a nonparametric bootstrap approach for standard error estimation. In Section 5, we conduct simulation studies to evaluate the finite-sample performance of the proposed method. An application to a data set from a Wilms tumor study is provided in Section 6. Some concluding remarks are stated in Section 7. All proofs are given in Appendix.

2. Design and estimation

2.1. Model and design

Suppose that there exists a study cohort of N independent subjects. Let $\widetilde{T}_i$ denote the failure time and C_i denote the censoring time or follow-up time for subject i ($i = 1, \ldots, N$). The observed time is $T_i = \min(\widetilde{T}_i,\ C_i)$. Let $\Delta_i = I(\widetilde{T}_i \le C_i)$, $Y_i(t) = I(T_i \ge t)$ and $N_i(t) = \Delta_i I(T_i \le t)$ denote the right-censoring indicator, the at-risk process and the counting process, respectively, where $I(\cdot)$ is the indicator function. Z_i denotes a p-dimensional covariate for subject i, and here we focus our attention on time-independent covariate. Let τ denote the end time for the study.

We assume that $\widetilde{T}_i$ arises from the following proportional hazards model (Cox, 1972):

$$\lambda(t|Z_i) = \lambda_0(t)\exp\left\{Z_i'\beta\right\}, \tag{1}$$

where $\lambda_0(t)$ is the unspecified baseline hazard function, and β is a p-dimensional regression parameter of primary interest. In the cohort studies that covariate information can be assembled for each individual, the following partial likelihood function is widely used for the inference of β (Cox, 1972; Andersen and Gill, 1982):

$$L_F(\beta) = \prod_{i=1}^{N}\left[\frac{\exp\left\{Z_i'\beta\right\}}{\sum_{l\in\mathcal{R}(T_i)}\exp\left\{Z_l'\beta\right\}}\right]^{\Delta_i}, \tag{2}$$

where $\mathcal{R}(t) = \{i:\ T_i \ge t,\ i = 1, \ldots, N\}$ is the risk set.

In the case-cohort studies, the covariate is not completely available for each individual. A subcohort is selected from the full cohort by simple random sampling. The subjects from the subcohort and the additional cases outside the subcohort constitute the case-cohort sample, only for which the measurements of covariate are assembled. Let $\widetilde{\mathcal{C}}$ and $\mathcal{C}$ be the index set of the subcohort and the case-cohort sample, respectively. Let $\widetilde{n}$ and n be the sample size of $\widetilde{\mathcal{C}}$ and $\mathcal{C}$, respectively. Therefore, the observed data structure for such a case-cohort design can be summarized as follows: $(T_i,\ \Delta_i,\ Z_i)$ for $i \in \mathcal{C}$, otherwise $(T_i,\ \Delta_i)$.

Since the covariates are observed incompletely under the case-cohort design, the likelihood function in (2) cannot be calculated. Prentice (1986) proposed the following pseudo-likelihood function:

$$L_P(\beta) = \prod_{i\in\mathcal{C}}\left[\frac{\exp\left\{Z_i'\beta\right\}}{\sum_{l\in\widetilde{\mathcal{R}}(T_i)}\exp\left\{Z_l'\beta\right\}}\right]^{\Delta_i},$$

where the risk set $\widetilde{\mathcal{R}}(t) = \left\{i:\ T_i \ge t,\ i \in \widetilde{\mathcal{C}} \cup \mathcal{D}(t)\right\}$ with $\mathcal{D}(t) = \{i : N_i(t) \ne N_i(t-), i = 1, \ldots, N\}$. Notice that $\mathcal{D}(t)$ is empty unless a failure occurs at time t. The corresponding log-likelihood function takes the following form:

$$l_P(\beta) = \sum_{i\in\mathcal{C}}\Delta_i\left[Z_i'\beta - \log\left\{\sum_{l\in\widetilde{\mathcal{R}}(T_i)}\exp(Z_l'\beta)\right\}\right]. \tag{3}$$