Journal of Statistical Planning and Inference 151-152 (2014) 1–16

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Review

An overview of semiparametric models in survival analysis

Shaojun Guo [a], Donglin Zeng [b,*]

[a] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, People's Republic of China
[b] Department of Biostatistics, University of North Carolina, United States

A R T I C L E I N F O

Article history:
Received 12 April 2013
Received in revised form
15 October 2013
Accepted 17 October 2013
Available online 28 October 2013

Keywords:
Proportional hazards model
Partial likelihood
Proportional odds model
Transformation model
Frailty model
Right censoring
Interval censoring
Nonparametric maximum likelihood
Counting process
Empirical process

A B S T R A C T

We provide an overview of semiparametric models commonly used in survival analysis, including proportional hazards model, proportional odds models and linear transformation models. The applications of these models to different types of censored data, either univariate or multivariate survival analysis, are given. For each case, inference procedures using censored observations are discussed.

Contents

* Corresponding author. Tel.: +1 919 966 7273.
E-mail addresses: guoshaoj@amss.ac.cn (S. Guo), dzeng@email.unc.edu (D. Zeng).

0378-3758/$ - see front matter
http://dx.doi.org/10.1016/j.jspi.2013.10.008

1. Introduction

In survival analysis, one of the most important goals is to estimate the association between risk factors and time-to-event and make future prediction of subject's survival probabilities. One major challenge in achieving this goal is that censored data are usually observed in studying time-to-event, where only partial information instead of precise measurement of time-to-event is available for some subjects. Since the seminal paper of Cox (1972) on proportional hazards model, numerous semiparametric models have been developed to analyze censored data in survival analysis, including proportional odds model and linear transformation models as well as their generalization from univariate survival analysis to multivariate survival analysis. At the same time, inefficient and efficient inference procedures have been proposed to derive the estimation of model parameters which are used to predict future survival probabilities. These procedures have been largely helped by the substantial development of mathematical theories including counting process theory (Fleming and Harrington, 1991) and empirical process theory (van der Vaart and Wellner, 1996).

In this paper, we aim to provide an overview of semiparametric models in survival analysis as well as theoretical development for these models. Since the work in this area is very extensive (maybe the most extensive area in statistics), our review is never meant to be complete and comprehensive. The paper is organized as follows: in Section 2, we describe the challenge of data structure specific to survival analysis. We then start to review semiparametric models for each type of survival data, including univariate survival data with right censored observations (Section 3), multivariate survival data with right censored observations (Section 4), and interval censored data and joint analysis of multiple outcomes (Section 5). Two real examples are analyzed in Section 5 to illustrate the power of semiparametric models. In Section 7, we conclude the paper by further listing some other semiparametric models and work in survival analysis.

2. Challenges in analysis of survival data

One of the major features that distinguishes the statistical analysis of survival data from studying other outcomes such as continuous or categorical outcomes is the presence of censoring. For censored subjects, only partial information instead of accurate measurement of survival time is observed. Usually, the censoring mechanism is classified into either right censoring or interval censoring (Kalbfleisch and Prentice, 2002) as follows: let T be the survival time of interest. For right-censored data, we either observe T or observe T greater than a time C. Mathematically, if we let C be potential censoring time, then the observed data can be expressed as $(Y=\min(T,C),\ \Delta=I(T\le C))$, where $I(\cdot)$ is the indicator function and Δ is called censoring indicator. For interval censored data, one only observes that T belongs to an interval, denoted by $[L,U]$, and the observed data can be summarized as $(L, U, I(T\le L), I(T\le U))$. When $L=U$, i.e., T is observed to be less or larger than L, this special interval censored data are called current status data. One key difference between right-censored data and interval censored data is that for the former, there is still some chance that we can accurately measure the survival time T. Therefore, right-censored data should provide more information regarding T as compared to interval censored data.

In survival analysis, another complication is different nature of outcome of interest. Typical time-to-event is a single outcome, such as patient's cancer survival or time to disease. However, multivariate survival outcomes are often investigated in many studies. They can be the survival events from a clustered group of subjects (clustered survival time), different types of survival outcomes from the same subject (multiple types of survival events), difference cause-specific survival time (competing risks), the repeated occurrence of the same type of event in the same subject (recurrent event), and the presence of multiple outcomes of which survival event is one (joint outcomes). In addition to event outcomes, another level of the complexity in survival analysis involves time-dependent covariates and time-varying effects. Furthermore, in practice, different study designs may be adopted in the study of time-to-event. All these pose very different challenges in survival analysis from analysis of non-censored observations.

Before introducing all kinds of semiparametric models, we list a few real examples below to illustrate different data structure in survival analysis.

Example 1 (*Lung cancer study*). This study was from the Veterans Administration lung cancer study group (Prentice, 1973; Kalbfleisch and Prentice, 2002). In a subgroup of 97 patients without prior therapy, survival times ranges from 1 to 587 days and 6 of them are censored. This is a typical survival analysis with right-censored data.

Example 2 (*rhDNase study*). A randomized clinical trial was conducted to assess the efficacy of rhDNae, a highly purified recombinant enzyme, in reducing exacerbations of respiratory symptoms for patients with cystic fibrosis. See Therneau and Hamilton (1997). In this clinical trial, a total of 321 patients were assigned to rhDNase and 324 were assigned to placebo. By the