Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2020:2**
http://www.math.su.se

Survival Analysis and its Application to Childhood Cancer Data

Ida Hed Myrberg*

February 2020

Abstract

Survival analysis denotes a collection of statistical methods, where time to one or several events of interest is considered, for example death, or the onset of a disease. In this thesis, essential concepts and quantities within the field of survival analysis, such as censoring, hazard and survival functions, are introduced. A number of selected methods are presented in detail; the non-parametric Nelson-Aalen and Kaplan-Meier estimators, the semi-parametric Cox proportional hazards model, the fully parametric accelerated failure time, proportional hazards and proportional odds models, the flexible parametric Royston and Parmar proportional hazards and proportional odds models, and the distribution-free quantile regression model. In order to estimate the power of detecting deviations from the proportional hazards assumption, Monte Carlo simulations are used, assuming underlying Weibull distributions. Large deviations are detectable at a high power even for moderate sample sizes, while small deviations are hard to detect even for large sample sizes. The Type I error rates are accurate when the proportional hazards assumption is fulfilled, for all investigated sample sizes and censoring proportions. Furthermore, Cox, Weibull, and Royston and Parmar proportional hazards models are compared, given an underlying Weibull distribution, and given that the proportional hazards assumption is fulfilled, using Monte-Carlo simulations. The three methods show comparable estimates and standard errors. Average coefficient estimates, standard errors, and confidence interval coverage of quantile regression models are evaluated using Monte-Carlo simulations, showing accurate coefficients estimates, but too low confidence interval coverage for small and moderate sample sizes. The proposed methods are used to investigate the association between the so called event-free survival time of children with acute lymphoblastic leukemia (ALL) and a variety of risk factors in a heavily right-censored dataset. Fully parametric distributions do not fit the data well, but coefficient estimates are comparable to semi-parametric and flexible parametric models. Some covariates do not fulfill the proportional hazards assumption, and are better modeled dependent on time in Royston and Parmar models. Quantile regression only works for small probabilities, since the proportion censored observations is high. Even so, this method provides a different perspective that could be useful in a clinical setting.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: ida.hed.myrberg@gmail.com. Supervisor: Mathias Lindholm.