P8160 Project 1 Final Report

A Simulation Study to Compare Three Survival Models

Jingchen Chai, Yi Huang, Ruihan Zhang

Contents

```
library(ggplot2)
```

Abstract

Survival analysis is a statistical method used to analyze time-to-event data, such as the time until a patient's death. It is commonly used in medical research to estimate the probability of an event occurring over time and to identify factors that may affect the risk of the event. The analysis also considers time-dependent covariates that vary over time and can be used to compare the survival times between different groups of patients. The objective of the simulation study is to evaluate the robustness of three survival models against the misspecified baseline hazard functions. To achieve this goal, we assess the accuracy and efficiency of the estimated treatment effects (β) for each model under various baseline hazard functions. This project will perform simulations to compare the parametric regression model (Exponential and Weibull) to the semi-parametric regression model (Cox) for survival data. We apply the inverse transformation method to generate survival data with censored observations from Exponential, Weibull, Gompertz, and Gamma distributions. Despite the different models, we found

Introduction

Survival analysis is a statistical technique used to examine data that measures the time it takes for an event to occur, such as a patient's death. It is commonly used in medical research to determine the likelihood of an event happening over a specific time period and to identify factors that may impact the risk of the event. This method takes into account censoring, which happens when participants do not experience the event before the end of the study or the event occurs after the study period. For instance, a study may track breast cancer patients from diagnosis to death or the end of a five-year period. In survival analysis, uncensored data is recorded when the event occurs at the exact observed event time, and it is coded as 1 in the status indicator variable. Conversely, censored data occurs when patients are lost to follow-up or the event happens after the study period, and it is coded as 0 in the status indicator variable. Time-dependent covariates are also considered in the analysis, which can vary over time and can be used to compare the survival times of different participant groups, such as a treatment group and a control group.

Objective

The objective is to design a simulation study to compare and contrast the efficiency and accuracy of the estimated treatment effects under different baseline hazard functions and assess their robustness against misspecified baseline hazard functions. A practical and effective recommendation will be provided to general users to select a suitable model on the basis of the numerical investigations.

Background

Survival function, $S(t)$ is the probability of observing individual survival time T beyond a certain time t. To analyze survival data, we define a **survival function** $S(t)$ as

$$S(t) = \Pr(T > t) = \int_t^\infty f(s)ds$$

where $T > 0$
The Cumulative distribution function of the random variable survival time T, F(t) is the probability of observing individual survival time less than a certain time t, we define **CDF of survival time** $F(t)$ as

$$F(t) = Pr(T \leq t) = 1 - S(t) = 1 - \int_t^\infty f(s)ds$$

The Hazard function, $h(t)$ measures the instantaneous risk of failure at time t giving that a patient has survived until time t, defined as the ratio of the probability density function of time variable and Survival function.

$$h(t) = \lim_{\Delta_t \to 0} \frac{Pr((T \in (t, t + \Delta_t)|T > t)}{\Delta_t} = \frac{f(t)}{S(t)}$$

Proportional hazard model is the primary regression model to investigate the effectiveness of treatment X over survival time T, where the i-th patient at a time t
is

$$h_i(t) = h_0(t)e^{x_i \beta}$$

where

- $h_0(t)$ is the baseline hazard function
- x_i is is the treatment indicator variable (control = 0, treatment = 1)
- β is the parameter of interest, which is the log hazard ratio for the treatment effect. β measures the relative hazard reduction due to treatment in comparison to the control.

The **proportional hazard** can be expressed as ratio of two hazard functions at time t given individuals in different treatment groups, and does not depend on t.

$$\frac{h(t|x_0)}{h(t|x_1)} = e^{\beta(x_0 - x_1)}$$

There are different ways to formulate the baseline hazard function $h_0(t)$, which lead to different models and estimations.

An exponential proportional-hazards model assumes the baseline hazard function is a constant

$$h_0(t) = \lambda$$

A Weibull proportional-hazards model assumes the baseline hazard function follows Weibull distribution, where

$$h_0(t) = \lambda \gamma t^{\gamma - 1}$$

for $\gamma > 0$

A Cox proportional-hazards model leaves $h_0(t)$ unspecified.

Note that exponential distribution is a special case of Weibull distribution where $\lambda = 1$. Hence, among the three models, the exponential proportional-hazards model is the most restrictive model, while the Cox model is the most general one.

Methodology

Design Simluation

Section 1

Section 2

Plot