P8160 Project 1 Final Report

A Simulation Study to Compare Three Survival Models

Jingchen Chai, Yi Huang, Ruihan Zhang

March 3, 2023

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, CDF of survival time, hazard function

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.

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

Intro to Proportional hazard model

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.

Three Proportional-hazards model

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

Data Generation: derive survival time T

Given:

$$H(t) = -\log(S(t))$$

$$S(t) = e^{-H(t)} = e^{\int_t^{\infty} h(s)ds}$$

Apply Inverse Transformation Method

$$T = F^{-1}(U) = H_0^{-1}(\frac{-\log{(U)}}{e^{X\beta}})$$

where $U \sim U(0,1)$$

Thus, we can derive the following table

Table 1: Characteristic of Exponential, Weibull and Gompertz distributions

	Distribution		
	Exponential	Weibull	Gompertz
Scale Parameter	$\lambda > 0$	$\lambda > 0$	$\lambda > 0$
Shape Parameter		$\gamma > 0$	$\alpha \in (-\infty, \infty)$
Baseline Hazard function	$h_0(t) = \lambda$	$h_0(t) = \lambda\gamma t^{\gamma-1}$	$h_0(t) = \lambda \exp(\alpha t)$
Cumulative Baseline Hazard Function	$H_0(t) = \lambda t$	$H_0(t) = \lambda t^{\gamma}$	$H_0(t) = \frac{\lambda}{\alpha}(e^{\alpha t} - 1)$
Inverse Cumulative Hazard Function	$H^{-1}(t) = \lambda^{-1}t$	$H^{-1}(t) = (\lambda^{-1}t)^{\frac{1}{\gamma}}$	$H^{-1}(T) = \frac{1}{\alpha}\log{(1 + \frac{\alpha}{\lambda}t)}$
Cumulative Distribution Function	$F(t) = 1 - e^{t\lambda e^{X\beta}}$	$F(t) = 1 - e^{-\lambda t^{\gamma} e^{X\beta}}$	$F(t) = 1 - e^{-\frac{\lambda}{\alpha}(e^{\alpha}-1)e^{X\beta}}$
Survival Time T	$T = -\frac{log(U)}{\lambda e^{X\beta}}$	$T = (-\frac{log(U)}{\lambda e^{X\beta}})^{\frac{1}{\gamma}}$	$T = \frac{1}{\alpha}log[1 - \frac{\alpha\log{(U)}}{\lambda e^{X\beta}}]$

Simulation Design

The goal of this simulation study is to evaluate how misspecifying the baseline hazard function influences the estimated treatment effect from 3 models. We generate 4 types of data, Exponential, Weibull, Gompertz, and Mixture distribution to evaluated the model robustness. All simulation data was generated using inverse transformation method and use `simsurv` function in the `simsurv` package to double check the result. The resulting dataset contains time of event (`eventtime`), status indicator (`status`), and treatment group (`trt`) with censored observations.

- Define Random Variable X, treatment, from a binomial distribution with $p = 0.5$
- Generate survival time T, time to event, using X and β
- Randomly generate censoring time C, from an exponential distribution
- We observe either the survival time T or else the censoring time C. Specifically, we observe the random variable Y takes the minimum value between survival time and censoring time

$$Y = min(T, C)$$

Create a status indicator variable (1 = event, 0 = censored), if the patient dies before censoring time, then status equals one, vice versa.

$$Status = \begin{cases} 1, & T_i \leq C_i \\ 0, & T_i > C_i \end{cases}$$

Define true treatment effect and parameter

Before running the simulation, also need to define the true treatment effect and parameters

1. True treatment effect $\beta = 2$
2. 7 different sample size N ranging from 100 to 400 increasing by 50
3. Simulate exponential distribution with $\lambda = 0.5$
4. Simulate weibull distribution with $\lambda = 0.5$ and $\gamma = 2$
5. Simulate gompertz distribution with $\lambda = 0.5$ and $\alpha = 2$

Simulation Times

A large number of simulations can improve the accuracy and efficiency of estimated treatment effect, thus we simulate 1000 times for N sample. Each datasets will be used to fit in three proportional hazard models to obtain the estimated treatment effects

Performance Measures

To compare accuracy and efficiency of estimated treatment effects, we use performance measures such as Bias, Variance, Mean square error, and confidence interval to compare 3 models.

k: number of independent datasets

Estimation

$$Bias = \frac{1}{k}\sum_{i=1}^{k}(\beta^{\hat{(k)}} - \beta) = \frac{1}{k}\sum_{i=1}^{k}(\beta^{\hat{(k)}} - 2)$$

$$Variance = \frac{1}{k-1}\sum_{i=1}^{k}(\beta^{\hat{(k)}} - \beta)^2 = \frac{1}{k-1}\sum_{i=1}^{k}(\beta^{\hat{(k)}} - 2)^2$$

$$MSE = \frac{1}{k}\sum_{i=1}^{k}(\beta^{\hat{(k)}} - \beta)^2 = \frac{1}{k}\sum_{i=1}^{k}(\beta^{\hat{(k)}} - 2)^2$$

Confidence Interval

$$Coverage : \frac{1}{k}\sum_{i=1}^{k} I\{\beta \in CI^{(k)}\}$$

Plot