Evaluating the Impact of Baseline Hazard Function Misspecification on Treatment Effect Estimation

Charly Fowler, Hanfei Qi, Robert Tumasian III , and Haoyang Yi

February 24, 2021

Objectives

The goal of this study is to evaluate how misspecifying the baseline hazard function can influence the estimation of treatment effects in survival. This work focuses on conducting simulations to compare the discrete, exponential, Weibull and lognormal proportional hazards models to the semi-parameteric Cox proportional hazards model. We also investigate the impact of utilizing an overly complicated model (e.g., Cox) when a less complex model (e.g., exponential) is sufficient.

Statistical Methods

Survival analysis is used to analyze time-to-event data (e.g., time to symptom onset or time to mortality). Survival functions ($S(t)$) measure the probability of an individual not experiencing an event past a certain time t. Similarly, hazard functions ($h(t)$) measure the instantaneous risk of failure at a certain time t, given that the individual has not experienced an event until that time. The hazard function can be expressed as $\frac{f(t)}{S(t)}$, where $f(t)$ is the distribution of survival times.

The purpose of proportional hazards modeling is to assess the effectiveness of a particular treatment (X) over survival time T, where the hazard ratio for patient i at time t is defined as $h_i(t) = h_0(t)e^{x_i\theta}$. Here, $h_0(t)$ denotes the pre-specified baseline hazard function, x_i indicates treatment allocation (0=control, 1=treatment), and θ represents the log hazard ratio, or the hazard reduction among treated individuals compared to the control group. Thus, the proportional hazard can be expressed as $\frac{h(t|x_0)}{h(t|x_1)} = e^{\beta(x_0-x_1)}$, which is independent of survival time t.

For each of the models we consider implement different baseline hazard functions, except for the Cox model, which estimates θ without this specification. Baseline hazard functions for the exponential, Weibull, and lognormal models are provided in Table 1 below. The baseline hazard function for our discrete model is stepwise, and the Cox model does not utilize a baseline hazard function._____(Are we using lognormal model or we simulate a lognormal situation?)

In order to finding a case that Cox model performs best, we also consider piecewise hazards_____

	Exponential	Weibull	Lognormal
$h_0(t)$	λ	$\lambda\gamma t^{\gamma-1}$	$\dfrac{\frac{1}{t\sigma\sqrt{2\pi}}\exp\left[\frac{-1}{2\sigma^2}\left\{\ln(t)-\mu\right\}^2\right]}{1-\Phi\left\{\frac{\ln(t)-\mu}{\sigma}\right\}}$

Table 1. Baseline hazard functions, $h_0(t)$, for the continuous models considered in this study. μ and σ are mean and standard deviation of normal distribution. The normal CDF is denoted by Φ.

Simulation Design

We simulated survival data

All simulation data was generated using the `simsurv` function in the `simsurv` package. We defined a binomial treatment variable (`trt`) with $p = 0.5$ to ensure equal likelihood of random assignment to the treatment or control group.Parameters are set according to different distributions of hazards.The resulting dataset contains time of event(`eventtime`), status(`status`) and treatment(`trt`).

For the Weibull model, we considered various parameters ($\gamma = 0.5, 1, 2, 5$, where $\gamma = 1$ denotes the exponential model. We used $\beta = -0.5$ as the true effect size. For the lognormal situation, we used $\mu = 0, \sigma = 0.5$ as parameters of normal distribution and $\beta = -0.5$ as the true effect size.

In piecewise simulation, we simulated a set of 20 piecewise hazards specified by seed() function_____

Data Generation

We generated datasets from 6 simulations considering $\gamma = 0.5, 1, 2, 5$, lognormal and piecewise.

We set the sample size as 100 and repeat 1000 times to get the dataframe of results. Each time we fitted exponential, Weibull and cox model to the data and extracted the β provided by three models.

Measuring Performance

To assess model performance, we used mean-squared error (MSE) and _____.

NOTE: shouldn't the tables be part of results? feel free to change it! NOTE2: Values look weird.

The following table presents mean of beta of each simulation model.

```
## # A tibble: 3 x 6
##   Model       Gamma_0.5 Gamma_1 Gamma_2 Gamma_5 Lognormal
##   <chr>           <dbl>   <dbl>   <dbl>   <dbl>     <dbl>
## 1 Exponential     -3.12   -2.31   -1.09  -0.375    -0.484
## 2 Weibull         -4.70   -2.30   -1.15  -0.460    -0.583
## 3 Cox            -0.529  -0.520  -0.501  -0.507    -0.510
```

The following table presents mean squared error of beta of each simulation model.

```
## # A tibble: 3 x 6
##   Model       Gamma_0.5 Gamma_1 Gamma_2 Gamma_5 Lognormal
##   <chr>           <dbl>   <dbl>   <dbl>   <dbl>     <dbl>
## 1 Exponential      7.02    3.32    3.32  0.0167   0.00423
## 2 Weibull          19.0    3.32    3.32 0.00239    0.0121
## 3 Cox             0.309   0.149   0.149  0.0461    0.0423
```

Table #:

Results

Conclusions