# Project 1: A simulation study examining three survival models

*Margaret Gacheru, Jin Ge, Sibei Liu, and Adina Zhang*

*2/13/2020*

## Problem Statement

Time-to-event data can be modeled using proportional hazards to investigate the effect of exposure or treatment on time-to-event. A hazard rate describes the likelihood of an event (ie death, disease onset) occuring at time t and is written as the following: $h_i(t) = h_0()exp(xi)$. A proportional hazards model can then used to describe a hazard rate between two treatment or exposure groups: $\frac{h(t|x_1)}{h(t|x_2)} = exp(\beta^T(x_1 - x_2))$. This model is useful as it no longer becomes dependent on time t. However, when fitting a survival model, assumptions must still be made about the baseline hazard function,$h_0(t)$, which is a function that describes the risk of death when all other covariates are equal to zero.

The exponential and Weibull proportional hazards models are parametric functions, whereby the exponential assumes a constant baseline hazard ($h_0(t) = \lambda$) while the Weibull assumes a baseline hazard as a function of time-to-event ($h_0(t) = \lambda\gamma t^{\gamma-1}$). On the other hand, the Cox proportional hazards model is a semi-parametric model, designed to be a more flexible model by not having to specify the baseline hazard function. Depending on the data, choosing the wrong assumptions can lead to inaccurate estimates of the treatment effect . In order to explore the issues related to misspecifying the baseline hazard function, we designed a simulation study to assess three survival models on three different baseline hazard function scenarios. Each scenario generates data based on different baseline hazard functions: exponential, Weibull, and an unspecified Gompertz hazard function. Then an exponential, Weibull, and Cox proportional hazards model are evaluated with each scenario.

## Methods

### Data Generation

Generating survival data involves 5 components: actual event time, censoring time, status indicator, observed time, and covariates, if needed. In order to generate event time, we can utilize the relationship between the survival function, $S(t)$, and hazard function $h(t)$ to obtain a direct relationship between the survival function, $S(t)$, and the baseline hazard function, $h_0(t)$. Survival function is the probability of surviving beyond time $t$

$$S(t) = P(T > t) = 1 - F(t)$$

Hazard function is the instantaneous rate at which the event occurs at time $t$ and be defined as

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} = -\frac{\partial}{\partial y} log[1 - F(t)] = -\frac{\partial}{\partial y} log[S(t)]$$

Isolating $S(t)$, we find

$$S(t) = e^{-H(t)}, where\ H(t) = \int_0^t h(t)dt$$

Additionally, we can find the connection between the cumulative hazard, $H(t)$, and the baseline hazard. Given that $h(t) = h_0(t)e^{x^T\beta}$,

$$H(t) = \int_0^t h_0(t)e^{x^T\beta}dt = e^{x^T\beta}H_0(t), where\ H_0(t) = \int_0^t h_0(t)dt$$

Putting it all together, we obtain

$$S(t) = e^{-H(t)} \qquad = e^{x^T\beta} H_0(t)$$

Finally, utilize the inverse transformation method to obtain $T$, event time

$$T = H_0^{-1}\left(\frac{-log(u)}{e^{x^T\beta}}\right), where\ U \sim U(0,1)$$

Under each of the 3 scenarios, the following steps can be used to generate the dataset:

1. Randomly generate $X_i$, treatment assignment variable, from a bernoulli distribution with $p = 0.5$
2. Generate $T_i$, time to event, using $X_i$ from step 1 and pre-specified $\beta$

$$T = H_0^{-1}\left(\frac{-log(u)}{e^{x^T\beta}}\right)$$

3. Randomly generate $C_i$, censoring time, from an exponential distribution
4. Determine the observe time, $Y_i$ by comparing event and censoring time

$$Yi = min(T_i, C_i)$$

5. Create the status indicator variable, where 1 represents if event is observed and 0 if event is censored

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

**Scenario Simiulation**

We conducted simulation studies to assess the performance of the proposed framework under three scenarios with different covariate forms, censoring time distributions, and baseline hazards. We simulated 1000 data sets with sample size n=400. The parameters in each model were held constant with $\beta = 4$, $\lambda = 0.1$ and $\alpha = 4$. Each time, baseline data was generated to follow the scenario baseline models: exponential, Weibull, and an unspecified Gompertz hazard function. Generated data were fitted to exponential, Weibull, and Cox proportional hazard models. After simulating through each model, a set of $\beta$ were generated and used to calculate the mean and 95% confidence interval to compare with each other. In addition, we simulated different sample sizes from 100 to 500 by 50 as each step. The mean squared errors (MSE) and bias (first-ordered) were calculated to demonstrate the performance of survival models as the sample size increased and show how poorly misspecified models perform. We also change the $\beta$ to show how MSE and bias would change in order to explore how the beta is associated with the model fitting. All the data generation and simulation were performed in R version 3.6.1

## Results

**Association with Beta and MES(or Bias) among three scenarios**

**Association with sample size and MSE(or Bias) among three scenarios**

## Discussion

## Appendix

```r
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(survival)
source("./sim.R")
```