Survival data simulation

Faith Lee

December 11, 2017

1 Objective

- Simulate n Exponentially-distributed survival times with time-invariant covariates

2 Introduction

Let's denote T as a random variable (corresponding to time-to-event). Then, the cumulative distribution function (CDF) will follow a standard uniform distribution. In equation terms,

$$F(T) \sim U(0,1). \tag{1}$$

It is not hard to see that the survival function, which is taken as $S(T) = 1 - F(T)$ follows a uniform distribution as well.
Recall that the probability density function (pdf) of the exponential distribution is as follows:

$$f(T|\lambda) = \lambda e^{-\lambda t} \tag{2}$$

for $t \geq 0$. Here λ is simply the "rate parameter". We simply say that T follows an Exp(λ) distribution. The hazard function, denoted $h(T)$ is simply λ.
Now, in the presence of covariates $\mathbf{X}$, our hazard function, now denoted as $h(T|\mathbf{x})$ becomes

$$h(T|x_1, x_2, ..., x_p) = e^{\beta_0 + \beta_1 x_1 + ... + \beta_p x_p} \tag{3}$$

given $T \geq 0$. Then the survival function of the T given vector of covariates $\mathbf{X}$ can be written as:

$$S(T|\mathbf{X}) = exp(-Te^{\beta_0 + \beta_1 x_1 + ... + \beta_p x_p}) \tag{4}$$

given $T \geq 0$. This is the background information for the simulation that will be outlined below.

3 Simulation steps

For the simplicity of illustrating this simulation, I will just have two covariates, x_1 and x_2 that is generated from the following distribution: $x_1 \sim Bin(n, 0.25)$ and $x_2 \sim Bin(n, 0.70)$. This can be understood as only 25% and 70% of our individuals (or population) has this characteristic, respectively. I will also consider the case of censoring here, specifically, right-censored survival data. **Let's aim for a censoring rate somewhere between 20% and 40%**. We define n = 1000 simulated individuals.

Next, I will outline the steps for the simulation. In the R code, lines corresponding to the steps here have been commented to make it clearer.

1. Generate the covariates x_1 and x_2 as above. For the βs, set it to $\beta_0, \beta_1, \beta_2 = $ (0.5, -1, 1). I'd like to highlight here that based on the choice of coefficients, having x_1 is associated with a lower hazard rate of having the event and having x_2 is associated with an increased hazard rate.

2. Generate $U_i \sim Unif(0, 1)$ for i = 1, ..., n. Subsequently, equate
$$S(T_i|x_{i1}, x_{i2}) = exp(-t_i e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}}) = U_i. \tag{5}$$
By solving for the inverse, we are able to derive the true survival times. Specifically,
$$T_i = \frac{-log(U_i)}{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}}} \tag{6}$$

3. Now, we have the true survival times for each individual. Next we generate censoring times, given by $C_i \sim Unif(0, b)$. Here, b is determined by trial and error to get the desired censoring rate which I will explain in the next step.

4. Define a binary indicator, δ_i. Then $\delta_i = 1$ if $T_i \leq C_i$ and zero otherwise. This represents the survival status. If an individual has $\delta_i = 1$, he/she is said to have the event of interest. Else, he/she has been censored.

5. Compute the number of $\delta_i = 0$ as a proportion of n. This will give us the censoring rate. If it is not between desired censoring rate%, we go back and redefine b in Step 3.

6. If $\delta_i = 1$, then the survival time, $t_i = T_i$. Else, the survival time, $t_i = C_i$.

7. Finally, we then plot the KM curve to look at the non-parametric estimation of survival probability and also to look at the effect of each of the covariate on time to survival.

Codes are in the R Code Section. It is really straightforward.