STAT 215B Assignment 2

Xiaowei Zeng

March 12, 2024

Abstract

The main job in this assignment is to develop deep insights of survival analysis. The first section discusses the parametric survival models with exponential distribution and Weibull distribution. The second section conducts several simulations under a special context of clinical trial with different patterns of censoring.

1 Models for time to failure

Define the hazard function for any disease as

$$H(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

where $F(t)$ and $f(t)$ are the distribution function and density, respectively, for the time to first onset. Thus

$$H(t) = \lim_{\epsilon \to 0} P\left(\text{Get disease during time } t \text{ to } t + \epsilon \mid \text{Never had disease up to time } t\right)$$

1.1 Exponential Setting

If the time to first onset is modeled with the exponential distribution, **the hazard function is constant over time.**

$$h(t) = \lambda$$

The probability density functino (PDF) of exponential distribution is

$$f(t) = \lambda e^{-\lambda t},$$

thus

$$F(t) = 1 - e^{-\lambda t},$$
$$S(t) = 1 - F(t) = e^{-\lambda t}.$$

The hazard function $h(t)$ is given by

$$h(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda.$$

The hazard function for the exponential distribution is constant and equal to the parameter λ over time. Therefore, this model assumes that the hazard of the event is constant over time, indicating that **the conditional probability of the event is the same, no matter when the observation is observed.** This can be attributed to the Memoryless Property of exponential distribution.

Modelling with exponential distribution can be reasonable under certain assumptions and contexts, but it may not always be appropriate. According to the memorylessness of exponential distribution, the probability of an event occurring (hazard) in the future is **independent of how much time has already passed.**

Therefore, for certain scenarios, the model is

- Reasonable when **Events occur randomly and independently over time.**
- Unreasonable when **The occurrence of events depends on the passing time**, e.g., the common knowledge that the probability of death increases as people grow older.

To check whether the exponential distribution fits a dataset with survival function estimate $\hat{S}(t)$, we can plot $-\log[\hat{S}(t)]$ and see whether a straight line through the origin with slope λ fits the shape well, because $\log[\hat{S}(t)]$ is linear in t.

$$\log[S(t)] = -\lambda t$$

1.2 Weibull Setting

The cumulative distribution function (CDF) of Weibull function is

$$G_{\alpha,\beta}(t) := F(t) = 1 - e^{-(t/\alpha)^\beta},$$

where α is the scale parameter, and β is the shape parameter, $t, \alpha, \beta > 0$.

The PDF of Weibull function is

$$f(t) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} e^{-(t/\alpha)^\beta},$$

thus the hazard function $h(t)$ is given by

$$h(t) = \frac{f(t)}{S(t)} = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1}.$$

- If $\beta = 1$, then it is equivalent to the exponential model with **constant hazard** $1/\alpha$.
- If $\beta > 1$, then the hazard function is **monotone increasing**.
- If $\beta < 1$, then the hazard function is **monotone decreasing**.

The exponential setting in Section 1.1 is **a special case** of Weibull setting. Besides, Weibull setting is able to **model the increasing or decreasing trend for hazards**, and thus more flexible and suitable for more contexts then exponential setting.

2 Survival curves

Consider the following stylized model of a study to test the effectiveness of a new surgical procedure. One thousand individuals are enrolled into the study, half receiving the surgery and half serving as the control group – no surgery. The time to death by any cause is measured in years, up to a maximum of five years, at which time the study ends.

The ith participant in the study carries two random values: time to death if assigned to receive the surgery, X_i, and time to death if assigned to the control group, Y_i. Assume that $\{(X_i, Y_i) : i = 1, 2, ..., N\}$ are drawn independently with X_i and Y_i having distributions $G_{3,2}$ and $G_{2,2}$, respectively, as defined in part one.

Two survival functions can be defined as

$$S_X(t) = P(\text{Individual } i \text{ lives past time } t \mid \text{Individual } i \text{ receives surgery}) = P(X_i > t)$$

and

$$S_Y(t) = P(\text{Individual } i \text{ lives past time } t \mid \text{Individual } i \text{ is in control group}) = P(Y_i > t)$$

2.1 $G(\alpha, \beta)$ Simulation

To write an `R` function to simulate draws of a $G_{\alpha,\beta}$ random variable using only `runif`, we need to employ the inverse transform sampling method.

$$\Pr(G^{-1}_{\alpha,\beta}(U) \leq t) = \Pr(U \leq (G^{-1}_{\alpha,\beta})^{-1}(t)) = G_{\alpha,\beta}(t)$$

For $i = 1, ..., n$, the inverse transform sampling method works as follows:

1. Generate a random number u_i from $U \sim \text{Unif}(0, 1)$.
2. Compute $t_i = G^{-1}_{\alpha,\beta}(u_i)$, and thus t_i is the ith sample from $G_{\alpha,\beta}$.

$$\begin{aligned} G_{\alpha,\beta}(G^{-1}_{\alpha,\beta}(u)) &= u \\ 1 - \exp\{-(G^{-1}_{\alpha,\beta}(u)/\alpha)^{\beta}\} &= u \\ (G^{-1}_{\alpha,\beta}(u)/\alpha)^{\beta} &= -\log(1-u) \\ G^{-1}_{\alpha,\beta}(u) &= \alpha[-\log(1-u)]^{1/\beta} \end{aligned}$$

Define the simulation function in `R`. The function takes as input a sample size n as well as α and β and returns a length-n vector of independent realizations of $G_{\alpha,\beta}$.

```r
simulate_G <- function(n, alpha, beta) {
  # Sample U ~ N(0, 1)
  U <- runif(n)
  # Compute Weibull r.v. using inversed CDF
  X <- alpha * (- log(1 - U)) ^ (1 / beta)
  # return a length-n vector
  return(X)
}
```

2.2 Kaplan-Meier Survival Function

Define the `R` function that creates a Kaplan-Meier (KM) survival function estimate. The input is two length-n vectors. The first contains event times. The second vector, parallel to the first, contains logical values: `TRUE` for observed deaths, `FALSE` for censoring events. The return value should be the estimated survival function: an `R` function which, when evaluated at t, returns the Kaplan-Meier estimate of surviving past t.

```
KM_fit <- function(time, delta){
  # Obtain the unique values of event times
  unique_t <- sort(unique(time))
  m <- length(unique_t)
  # Kaplan-Meier Estimates
  KM <- c(1, unique_t)
  for (i in 1:m){
    KM[i + 1] <- KM[i] * (1 - sum(time == unique_t[i] & delta) /
      sum(time >= unique_t[i]))
  }
  # Return the survival function (with precision correction)
  return(function(time){KM[sum(unique_t <= time * (1 - 1e-16)) + 1]})
}
```

Notice that here the KM estimator is calculated as

$$\hat{S}(t) = \prod_{t_i<t} \left(1 - \frac{\#\text{ observed failures at } t_i}{\#\text{ on test during } t_i}\right),$$

where # on test during t_i is the number of observations whose time (event or censored) is $\geq t_i$.

2.3 Clinical Trial Simulation 1

Simulate the performance of the clinical trial by drawing a sample of 500 failure times from $G_{3,2}$ and 500 from $G_{2,2}$. Estimate S_X and S_Y using Kaplan-Meier. Graphically compare these estimates to the true curves. Note: do not be surprised by low survival rates at the five-year horizon. The people in this study are very sick.

The true survival function for $G_{\alpha,\beta}$ is

$$S(t) = e^{-(t/\alpha)^\beta}.$$

Recall that the study has a length of 5 years. However, there are great possibilities that the simulated event times exceed 5 years. Therefore, **I regard the event occurring after 5 years as censored to make sure that the study ends in the end of the fifth year.** Notice that there is no censoring in the middle of the study and it is not a case of rolling entry. Only administrative censoring occurs in this study and thus the Kaplan-Meier estimator should be the same as the traditional empirical estimator.

```
library(ggplot2)
# Define the sample size
n <- 500
```

```r
# Set the random seed for replication
set.seed(0)
# Sample X and Y
time_X <- simulate_G(n, 3, 2)
time_Y <- simulate_G(n, 2, 2)
# Administrative censoring
time_X <- ifelse(time_X < 5, time_X, 5)
time_Y <- ifelse(time_Y < 5, time_Y, 5)
# Compute KM estimates for S_X and S_Y
S_X <- KM_fit(time_X, ifelse(time_X < 5, 1, 0))
S_Y <- KM_fit(time_Y, ifelse(time_Y < 5, 1, 0))
# Define the function for true S_X and S_Y
S_true <- function(t, alpha, beta){exp(- (t / alpha) ^ beta)}
# Compare the KM estimates and the true values graphically
res_X <- data.frame("time" = rep(time_X[time_X <= 5], 2),
                    "S_X" = c(sapply(time_X[time_X <= 5], S_X),
                              sapply(time_X[time_X <= 5],
                                     function(t){S_true(t, 3, 2)})),
                    "Sx_t" = rep(c("KM", "True"),
                                 each = sum(time_X <= 5)))
ggplot(data = res_X) + xlim(c(0, 5)) +
  geom_step(mapping = aes(x = time, y = S_X, color = Sx_t)) +
  theme_bw() + labs(title = "Survival Curves for X (KM & True)")
res_Y <- data.frame("time" = rep(time_Y[time_Y <= 5], 2),
                    "S_Y" = c(sapply(time_Y[time_Y <= 5], S_Y),
                              sapply(time_Y[time_Y <= 5],
                                     function(t){S_true(t, 2, 2)})),
                    "Sy_t" = rep(c("KM", "True"),
                                 each = sum(time_Y <= 5)))
ggplot(data = res_Y) + xlim(c(0, 5)) +
  geom_step(mapping = aes(x = time, y = S_Y, color = Sy_t)) +
  theme_bw() + labs(title = "Survival Curves for Y (KM & True)")
```

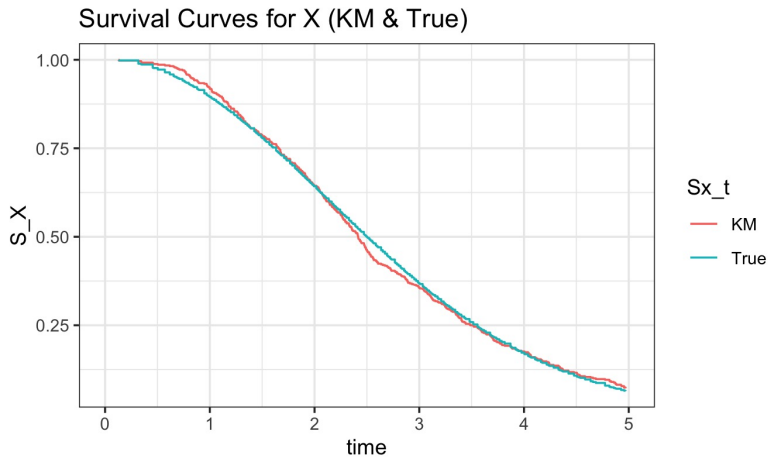

Figure 1: Survival Curves for X

As shown in Figure 1 and 2, the KM estimates for S_X and S_Y are really close to the

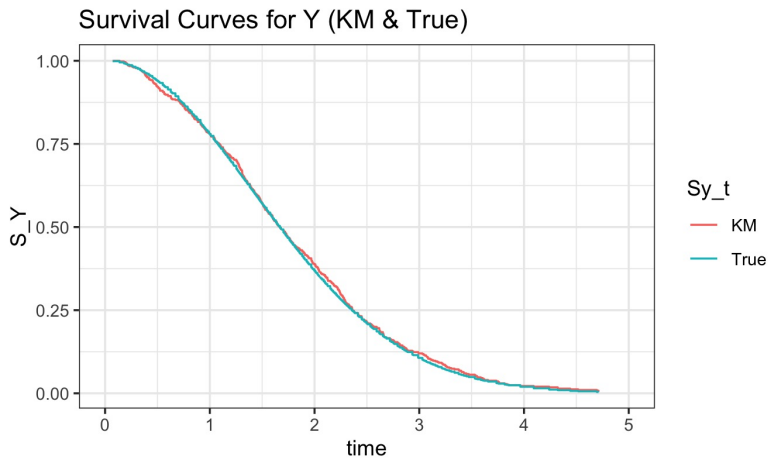

Figure 2: Survival Curves for Y

true survival functions. Notice that the time span of the two survival curves is not exactly $[0, 5]$, because here I only plot the curves starting from the first non-censored event time to the last non-censored event time.

2.4 Clinical Trial Simulation 2

Though the instruction document of this assignment says, "the simulation in Section 2.3 does not include the possibility of censoring", actually I have already taken administrative censoring into consideration. Therefore, in this section, we will assume **a new pattern of censoring**: for individual i there is another random variable, denoted Z_i, that gives the time at which i will be censored – if he lives that long. This involves a combination of administrative censoring and random censoring. Consider the case when the Z_i are i.i.d. exponential random variables with mean 10, chosen independently of X_i and Y_i. Simulate censoring times under this scenario and create new Kaplan-Meier survival curves.

Note that the event times are regarded as censored when $X_i \geq \min\{Z_{Xi}, 5\}$ and $Y_i \geq \min\{Z_{Yi}, 5\}$. We replace line 12 to line 14 of the codes in Section 2.3 with the following codes:

```
# Generate Z_X and Z_Y
Z_X <- rexp(n, 1 / 10)
Z_Y <- rexp(n, 1 / 10)
# Compute KM estimates for S_X and S_Y
S_X <- KM_fit(time_X, ifelse(time_X - Z_X < 0 & time_X < 5, 1, 0))
S_Y <- KM_fit(time_Y, ifelse(time_Y - Z_Y < 0 & time_Y < 5, 1, 0))
```

As shown in Figure 3 and 4, the KM estimates for S_X and S_Y are **slightly larger** than the true survival functions. It makes sense because the times supposed to encounter failure may become censored (Z_i is independent on the event time X_i or Y_i), and then **the survival rate may be overestimated**.

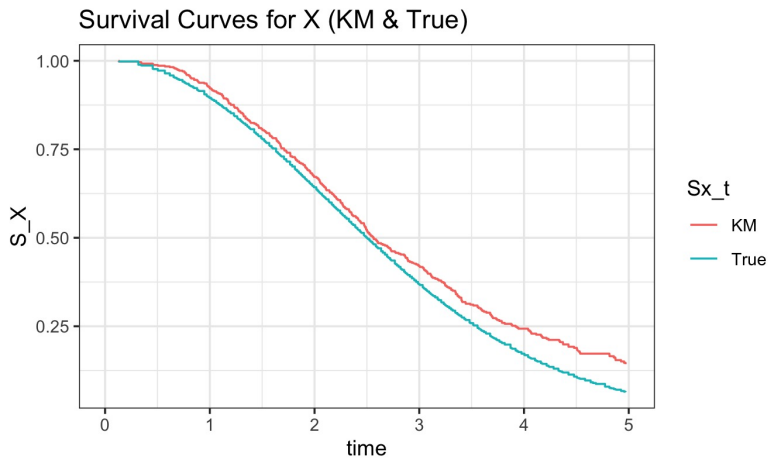

Figure 3: Survival Curves for X

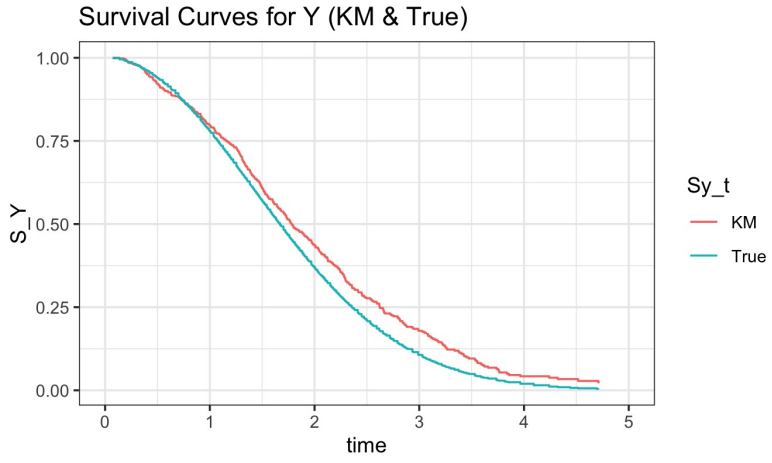

Figure 4: Survival Curves for Y

2.5 Clinical Trial Simulation 3

Now suppose the Z_i are independent exponential random variables **whose mean depends on the individual's time of death**. If the time of death is less than two years, the mean of the distribution of Z_i is 10; otherwise, the mean is 5. This could arise in a study where the sicker patients are more likely to remain under the care of their doctors.

Note that the event times are regarded as censored when $X_i \geq \min\{Z_{Xi}, 5\}$ and $Y_i \geq \min\{Z_{Yi}, 5\}$ because Z_i depends on the time of death and for patient i there are two possible death times (treatment group and control group). We replace line 12 to line 14 of the codes in Section 2.3 with the following codes:

```
# Generate Z_X and Z_Y
Z_X <- sapply(time_X, function(t){rexp(1, ifelse(t < 2, 0.1, 0.2))})
Z_Y <- sapply(time_Y, function(t){rexp(1, ifelse(t < 2, 0.1, 0.2))})
# Compute KM estimates for S_X and S_Y
S_X <- KM_fit(time_X, ifelse(time_X - Z_X < 0 & time_X < 5, 1, 0))
S_Y <- KM_fit(time_Y, ifelse(time_Y - Z_Y < 0 & time_Y < 5, 1, 0))
```

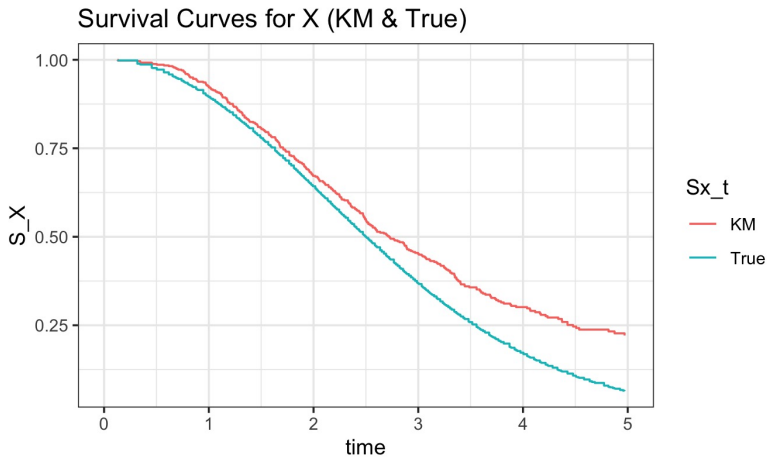

Figure 5: Survival Curves for X

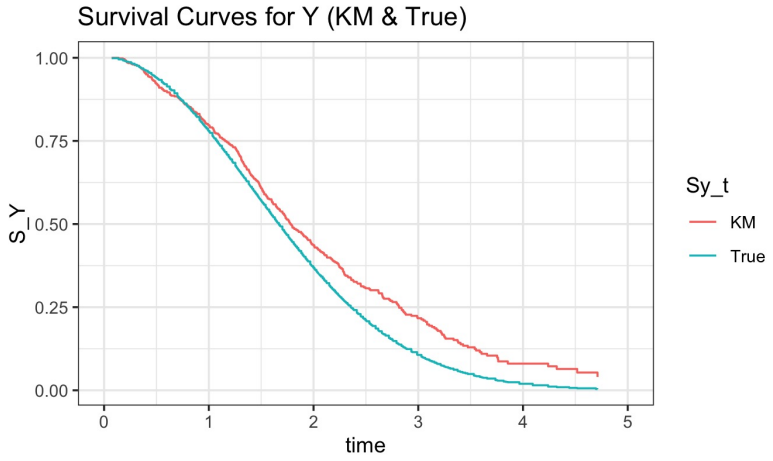

Figure 6: Survival Curves for Y

As shown in Figure 5 and 6, **the gap** between the KM estimates for S_X and S_Y and the

true survival functions **becomes larger as the time increases**. It makes sense because sicker patients, who are more likely to die sooner, have a longer expected censoring time; while less sick patients, who are more likely to survive longer, have a shorter expected censoring time. **This leads to an increased number of censoring data over time, and thus the estimated survival rates are overestimated more and more.**

The key difference between this censoring scenario and the previous one is that this one violates the **Independent Censoring Assumption or Independence of Competing Risk Assumption, which leads to bias in estimation.**