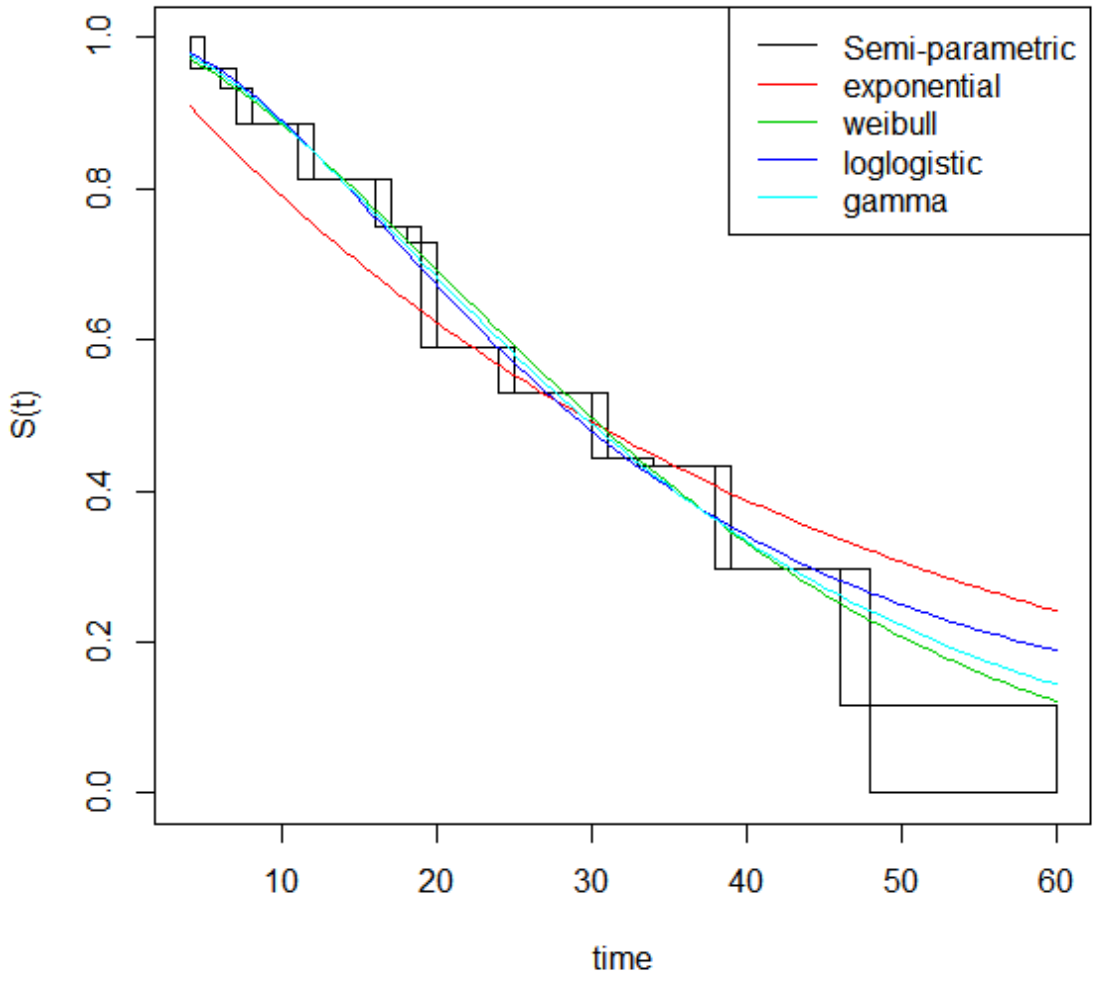
Survival Analysis for Interval Censored Data

Anubhab Biswas

Presidency University, Kolkata

June 2021



Instructor:

Dr. Biswabrata Pradhan

ISI, Kolkata

Abstract

In this project we focus on the survival analysis of interval censored data. We have described the methodology to fit a curve for interval censored data both in a parametric, non-parametric and semi-parametric way. We took the famous "chemotherapy vs radiotherapy" data and applied those methodologies there to heuristically infer about the effectiveness of the two treatments. It was comparative study to comment about which one of the treatments is more effective.

1. Introduction

The very primary question that one might ask before doing any analysis is the basic fundamental question - why do we need to study it? Similarly in survival analysis one might be asking the same question in the beginning - why at all we need to study survival analysis separately? How is it different from any kind of statistical analysis?

As the saying goes - "Everything in this earth is mortal", it is obvious that death (in case of a living thing) and failure to work/function (in case of a non-living thing) is the ultimate fate of an object in this earth as far as its utility to the human civilization is concerned. Ofcourse, this death or failure is a random phenomenon since no one can know or even predict when the event of interest that is death/failure is going to happen. This leads to the fact that survival time or the time until failure/death is also random. We statisticians deal with random phenomena. Hence it is not a surprise that survival analysis or lifetime analysis or the analysis of time until this random event of interest i.e. failure has been a topic of interest of the statisticians from a long time.

This survival time or lifetime analysis plays undoubtedly an important role in various branches of study including biomedical, clinical, engineering and even social sciences. For example, in a biomedical experiment, one may be interested in studying or rather analyzing how a drug or medicine influences the chances of survival of an individual or a health camp influences the chances of survival of a community on an average depending on its demography. In both of these examples the common point is survival and survival methodology will be inevitable in both the cases.

2. Some basic definitions and terminologies

- **Survival distribution** : Let T denote the survival time or time to failure of an item/individual since the occurrence of a particular event (like birth, accident etc). The probability distribution of T is known as **survival distribution** or **life-time distribution**. Some of the standard life-time distributions are :
  - **Exponential distribution** : $f_T(t) = \frac{1}{\theta} e^{-\frac{t}{\theta}} \ t > 0, \ \theta > 0$
  - **Weibull distribution** : $f_T(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}} \ t > 0, \ \alpha, \beta > 0$
  - **Gamma distribution** : $f_T(t) = \frac{\alpha^{\beta}}{\Gamma(\beta)} e^{-\alpha t} t^{\beta-1} \ t > 0, \ \alpha, \beta > 0$
- **Survival function** : Survival function at any time point t gives the probability that the item will survive for at-least t units of time. It is usually denoted by $S(t)$. Thus

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F_T(t)$$

- **Hazard rate** or **Failure rate** : Let $h(t)$ denote the hazard rate of an item at the time point t. It is the instantaneous probability rate at the time point t of the item for its failure or experience of the event of interest(death) given that the item has already survived for at-least t units of time.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f_T(t)}{1 - F_T(t)} \qquad t > 0$$

Note: The exponential distribution has a constant hazard rate.

3. Censoring

In the introduction we had some preliminary idea about survival analysis. But one question - the basic question is still not resolved. Why do we need to study survival analysis in detail separately? Why don't we just treat is as a simple problem of statistical analysis and proceed accordingly making survival the response variable and the other factors as factors or covariates and perform the problem of inference of the general regression analysis?

Well, the second case is not so simple. The reason behind this is Censoring. So let us see what do we mean by censoring and how does it make the survival analysis a different problem than the general statistical methodologies. We shall also see after that, how to tackle censoring and develop a methodology which deals with censoring.

But before that, in this context, it should be noted that the standard likelihood methodology applies to the models where the parameter is finite dimensional vector, i.e., parametric models. Also, semi-parametric and non-parametric models can also be developed by the likelihood methodology. In survival analysis too these are applicable when we want to find the the probability distribution of the random variable denoting the lifetime of an individual.

Censoring : Censoring is a key issue in survival analysis which distinguishes it from regular statistical problems. Censoring is when an observation is incomplete due to some random cause. The cause of censoring is usually assumed to be independent of the event of interest for standard survival analysis methodology.
This incompleteness in the observation may come in various ways. Depending on the direction through which this incompleteness in the observation comes, censoring can be broadly classified into 3 categories -

1. Right censoring : It is the most common form of censoring. Here the lifetime of an item is followed until some time at which the event of interest is yet to occur; but the individual takes no further part in the study after that time.

2. Left censoring : This occurs when the event of interest has already taken place at the time of the observation; but the exact time of occurrence of the event is unknown.

3. Interval censoring : Here the exact time of the event of interest is not known precisely, but an interval bounding this time is known.

In this project, we shall work with interval censored data.

4. The given data

The data for this project is given as :

"Finkelstein (1986) and Undsey and Ryan (1998) discussed interval-censored data from a study of patients with breast cancer. The response variable of interest was the time, T. to cosmetic deterioration of the breast, and whether there was a difference in the distribution of T for women who received radiation therapy alone versus a combination of radiation and chemotherapy. The data are shown in Table 3.10 for the two groups."

Table 3.10. Interval Censored Times to Cosmetic Deterioration

Radiotherapy			Radiotherapy and Chemotherapy		
(45, ∞]	(25, 37]	(37, ∞]	(8,12]	(0, 5]	(30, 34]
(6, 10]	(46, ∞]	(0, 5]	(0, 22]	(5, 8]	(13, ∞]
(0, 7]	(26, 40]	(18, ∞]	(24, 31]	(12, 20]	(10, 17]
(46, ∞]	(46, ∞]	(24, ∞]	(17, 27]	(11, ∞]	(8, 21]
(46, ∞]	(27, 34]	(36, ∞]	(17, 23]	(33, 40]	(4, 9]
(7, 16]	(36, 44]	(5, 11]	(24, 30]	(31, ∞]	(11, ∞]
(17, ∞]	(46, ∞]	(19, 35]	(16, 24]	(13, 39]	(14, 19]
(7, 14]	(36, 48]	(17, 25]	(13, ∞]	(19, 32]	(4, 8]
(37, 44]	(37, ∞]	(24, ∞]	(11, 13]	(34, ∞]	(34, ∞]
(0,8]	(40, ∞]	(32, ∞]	(16, 20]	(13, ∞]	(30, 36]
(4,11]	(17, 25]	(33, ∞]	(18, 25]	(16, 24]	(18, 24]
(15, ∞]	(46, ∞]	(19, 26]	(17, 26]	(35, ∞]	(16,60]
(11, 15]	(11, 18]	(37, ∞]	(32, ∞]	(15, 22]	(35, 39]
(22, ∞]	(38, ∞]	(34, ∞]	(23, ∞]	(11, 17]	(21, ∞]
(46, ∞]	(5, 12]	(36, ∞]	(44, 48]	(22, 32]	(11, 20]
(46, ∞]		(14, 17]	(10, 35]	(48, ∞]	

5. The framework of interval censoring

Lifetime data occur over chronological time and a variety of schemes are used to obtain data according to prevailing time and resource constraints. This can produce a common occurance where individuals in a study are observed intermittently at discrete time points.

Consider a framework where each individual $i = 1(1)n$ is observed at a pre-specified set of times $0 = a_{i0} < a_{i1} < ... < a_{in_i} < \infty$.

If an individual has not failed by time $a_{i,j-1}(j = 1(1)m_i)$, they are observed next at a_{ij}, and it is determined whether or not failure occurred in the interval $(a_{i,j-1}, a_{ij}]$.

Hence, the observed data consists of an interval $(U_i, V_i]$ for each individual with the information that $U_i < T_i \leq V_i$, and the lifetime is said to be interval censored; where T_i is the survival time of the i-th individual.

If failure has not occurred by time a_{im_i}, then $V_i = \infty$ and $U_i = a_{im_i}$ is a right censoring time for T_i.

The observed likelihood function from a sample of N independent individuals under this observation scheme is -

$$L = \prod_{i=1}^{n} [F_i(V_i) - F_i(U_i)],$$

where $F_i(t)$ is cdf of T_i and $F_i(0) = 0$.....(I)

Here we might consider $T_i \sim < Multinomial(1; p_{i1}, p_{i2}, ..., p_{im_i})$ where $p_{ij} = F_i(a_{ij}) - F_i(a_{i,j-1})$.

6. Inference

6.1 Non-parametric estimation of survival function

Here we use a version of E-M algorithm to find the estimate of the survival function as proposed by Turnbull (1976). In case of interval-censored data, the generalized maximum likelihood estimator can be shown to take a form similar to the closed form product limit (Kaplan-Meier) estimator; but it has jumps on a discrete set of equivalent classes defined through intervals. Turnbull derived this estimator or rather an approximation to this estimator, using a iterative self-consistency algorithm described below.

The search of the NPMLE of the survival function under interval censoring requires the definition of a set of intervals, called Turnbull intervals, denoted by $I = \{(q_1, p_1], (q_2, p_2], ..., (q_m, p_m]\}$ .These intervals are obtained from the set of all left and right interval endpoints in such a way that q_j is a left endpoint, p_j is a right endpoint and there is no other left or right endpoint between q_j and p_j . Turnbull proved that a maximum likelihood estimator of the survival function under interval censoring concentrates its mass on this set of intervals. Specifically, he stated that the search of the non-parametric MLE of S should be performed within the class of survival curves which are constant outside the set of Turnbull intervals and that the estimated survival curve is unspecified within each $(q_j, p_j]$.

Now let $w_j = P(q_j < T \leq p_j) = S(q_j) - S(p_j)$ be the weight of the $j-th$ Turnbull interval, $j = 1(1)m$. In order to get the NPMLE of survival function we hence need to maximize :

$$L_T(w_1, w_2, ..., w_m) = \prod_{i=1}^{n} (\sum_{j=1}^{m} \alpha_{ij} w_j)$$

where, $\alpha_{ij} = I\{(q_j, p_j] \subseteq (U_i, V_i]\}$ indicates whether or not the interval $(q_j, p_j]$ is contained in $(U_i, V_i]$ and the parameters are subject to the constraints $w_j \geq 0 \ \forall j = 1(1)m$ and $\sum_{j=1}^{m} w_j = 1$.

We will be using the R-programming (interval and Icens package) for the iteration purpose. So let us see the algorithm.

- Let us see the notations first :
  - T_i : event time for $i-th$ subject.
  - $(U_i, V_i]$: observed interval for $i-th$ subject.
  - $s_1, s_2, .., s_m$: set of possible observation times where NPMLE may change.
  - Let $s_0 = 0$ and $s_{m+1} = \infty$.
  - $p(s_j) = P[s_{j-1} < X \leq s_j]$
  - $\underset{\sim}{p} = [p(s_1), p(s_2), ..., p(s_{m+1})]$
  - $p_i(s_j) = P[s_{j-1} < T_i \leq t_j | T_i \epsilon (U_i, V_i]]$
- start with initial estimate of $\underset{\sim}{\hat{p}}$.
  - Each of m+1 elements should be positive, and all should sum to 1.
  - Example : $\underset{\sim}{\hat{p}} = [\frac{1}{m+1}, \frac{1}{m+1}, ..., \frac{1}{m+1}]$
- E-step: for each i

- $\hat{p_i}(s_j) = \frac{\hat{p}(s_j) I\{s_j \epsilon (U_i, V_i]\}}{\sum\limits_{s_k \epsilon (U_i, V_i]} \hat{p}(s_k)}$
- Example: $U_i = s_1$, $V_i = s_3$, $\hat{p_i} = [0, \frac{\hat{p_i}(s_2)}{\hat{p_i}(s_2)+\hat{p_i}(s_3)}, \frac{\hat{p_i}(s_3)}{\hat{p_i}(s_2)+\hat{p_i}(s_3)}, 0, ..., 0]$

- M-step: update $\underset{\sim}{\hat{p}}$For each j,
  - $\hat{p}(s_j) = \frac{1}{n} \sum\limits_{i=1}^{n} \hat{p_i}(s_j)$
- Iterate until convergence.

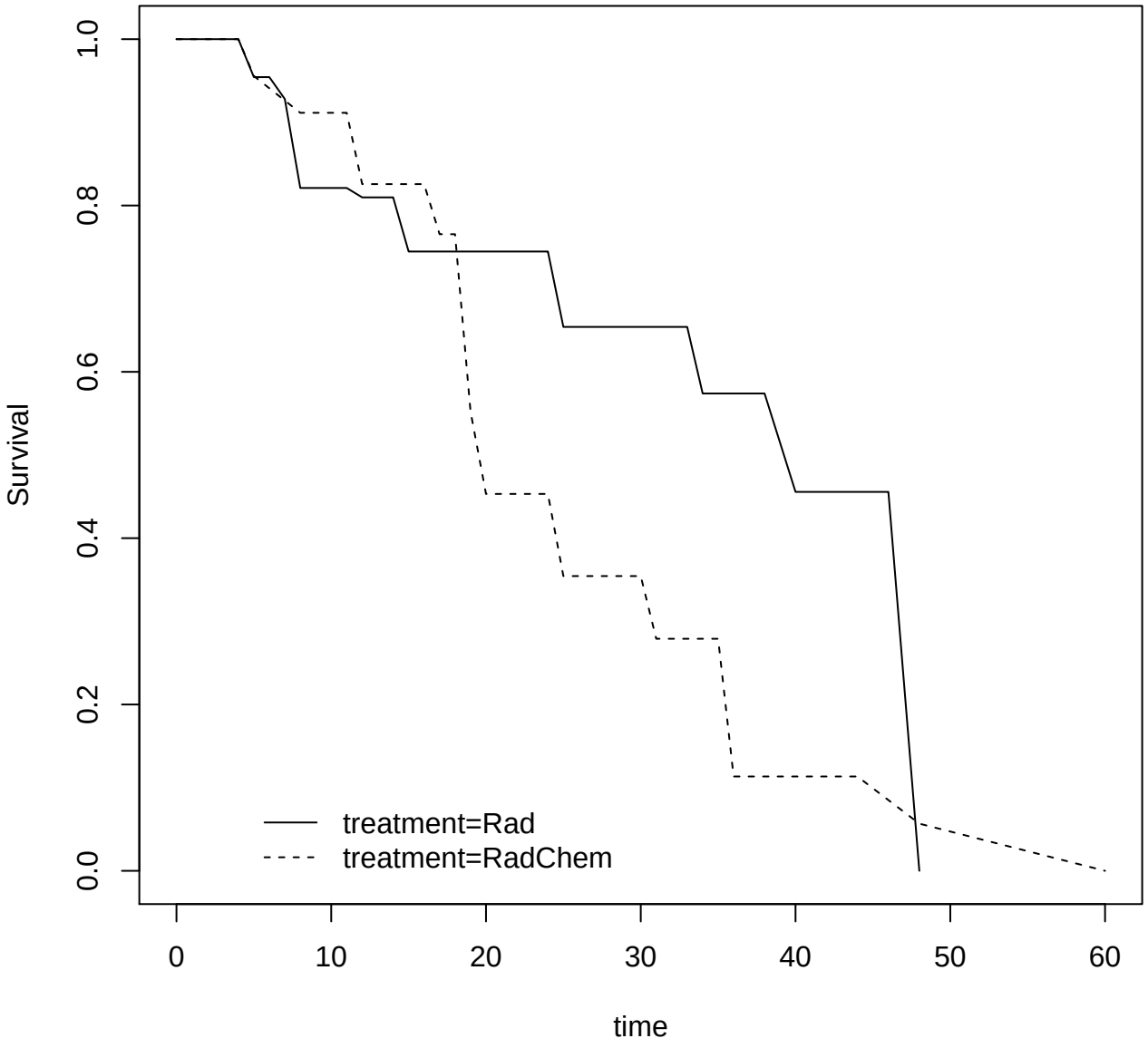
This procedure is very time consuming as we might need huge number of iterations. Here the time consumption was drastically reduced by the introduction of Turnbull intervals.

Now let us find our NPMLE by this method.

```
Loading required package:  survival
Loading required package:  perm
Loading required package:  Icens
Loading required package:  MLEcens
Depends on Icens package available on bioconductor.
To install use for example:
install.packages('BiocManager')
BiocManager::install('Icens')

treatment=Rad:
  Interval Probability
1    (4,5]      0.0456
2    (6,7]      0.0263
3    (7,8]      0.1070
4  (11,12]      0.0113
5  (14,15]      0.0651
6  (24,25]      0.0907
7  (33,34]      0.0800
8  (38,40]      0.1183
9  (46,48]      0.4557
treatment=RadChem:
   Interval Probability
1     (4,5]      0.0443
2     (5,8]      0.0443
3   (11,12]      0.0857
4   (16,17]      0.0602
5   (18,19]      0.2131
6   (19,20]      0.0992
7   (24,25]      0.0988
8   (30,31]      0.0754
9   (35,36]      0.1656
10  (44,48]      0.0567
11  (48,60]      0.0567
```

We can also graphically observe how the estimates of survival time differ in case of different treatments :



6.2 Fitting known probability distributions to the survival time

Interval censored data, generate likelihood functions of the form :

$$L(\underset{\sim}{\theta}) = \prod_{i=1}^{n} [F(V_i; \underset{\sim}{\theta}) - F(U_i; \underset{\sim}{\theta})]$$

where $F(t; \underset{\sim}{\theta})$ is the c.d.f, for lifetime and the $i-th$ lifetime has been observed to lie in the interval $(U_i, V_i]$.
The parametric estimation of the survival function can be done by the direct maximization of the log-likelihood of this function involving the parameter $\underset{\sim}{\theta}$.

Now these computations are really cumbersome and hence we have to take help of different software for fitting a specific family of distribution. In our case we take the help of the package "icenReg" in R-program for this computation purpose.

We shall be using a particular class of models of the survival function (or failure time) known as Accelerated Failure Time (AFT) models for a variety of distributions (as described in section 2 of this project) to interval censored data. The AFT model is defined by the transformation : $T_z = T_0 e^{-\beta z}$, where T_z is the failure time random variable for an individual with co-variate z and T_0 is the failure time the individual would have if they had co-variate value 0. The effect of changing co-variates is to shrink or stretch the time to event. If β is negative, then the co-variate has the effect of "speeding up time" so that individuals with larger values of z have higher failure rates and hence shorter survival times.

The survival function can be written as :

$$S(t; z) = P(T_z \geq t|z) = P(T_0 \geq te^{\beta z}) = S_0(te^{\beta z})$$

where $S_0(t)$ is the survival function for an individual with co-variate value 0. Taking natural logarithms, the AFT model can be expressed as

$$logT = logT_0 - \beta z$$

If we assume that $logT_0$ can be expressed as $\mu + \sigma W$ where W is a random variable, then the model can be written in a linear model-like form :

$$logT = \mu - \beta z + \sigma W$$

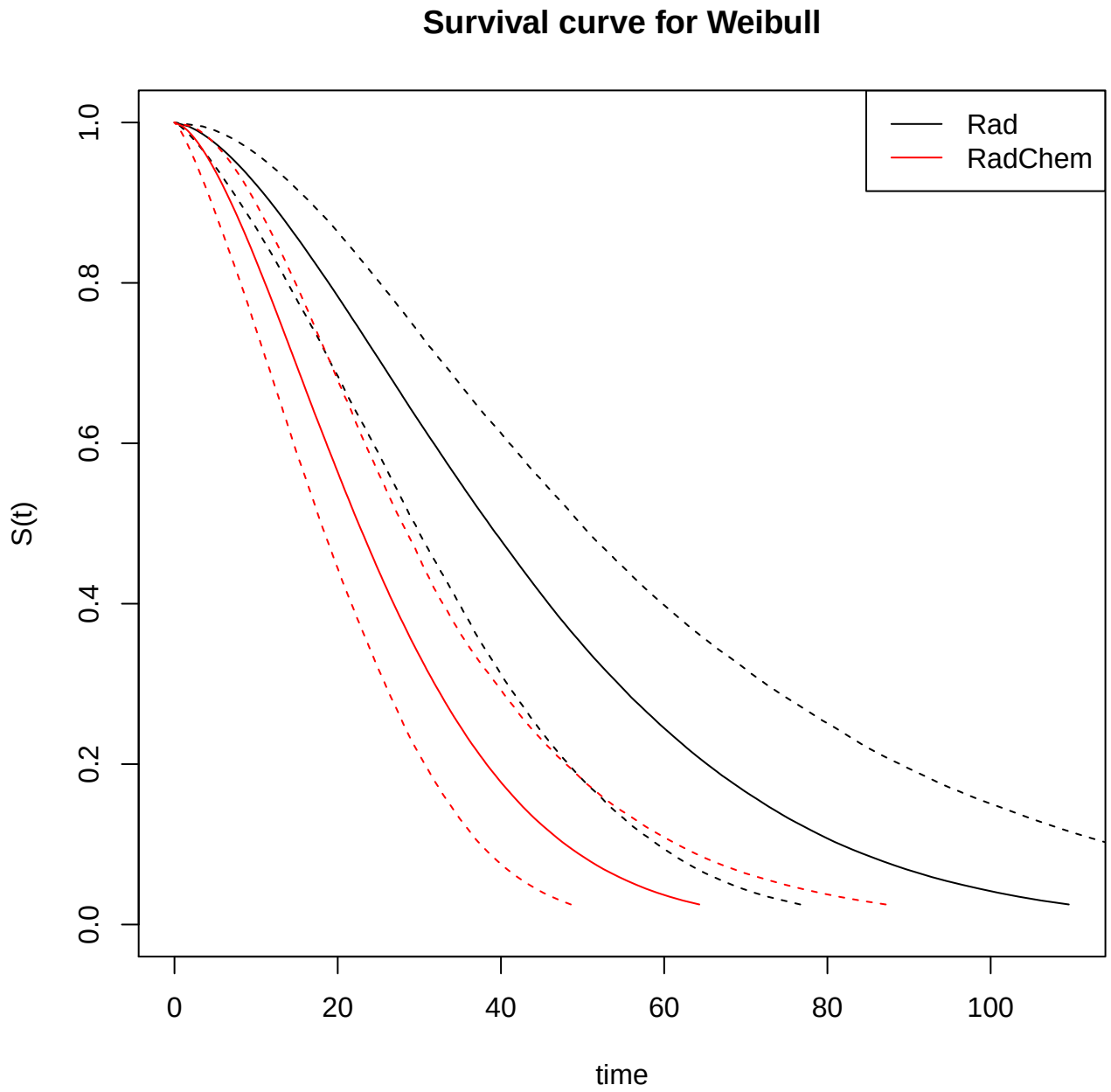
What we are really going to do here is allowing a variety of distributions to be placed on the error term W, including the log of the exponential, log-normal and log-gamma distributions. The intercept parameter μ and the scale parameter σ are usually not of direct interest, although for some distributions, there is a relationship between the AFT model and a proportional hazards model through the scale parameter. For example, if W is an extreme value distribution (log of a unit exponential), then T has a Weibull distribution.

Now let us fit various distributions as described in section 2.

(i) **Weibull distribution** :

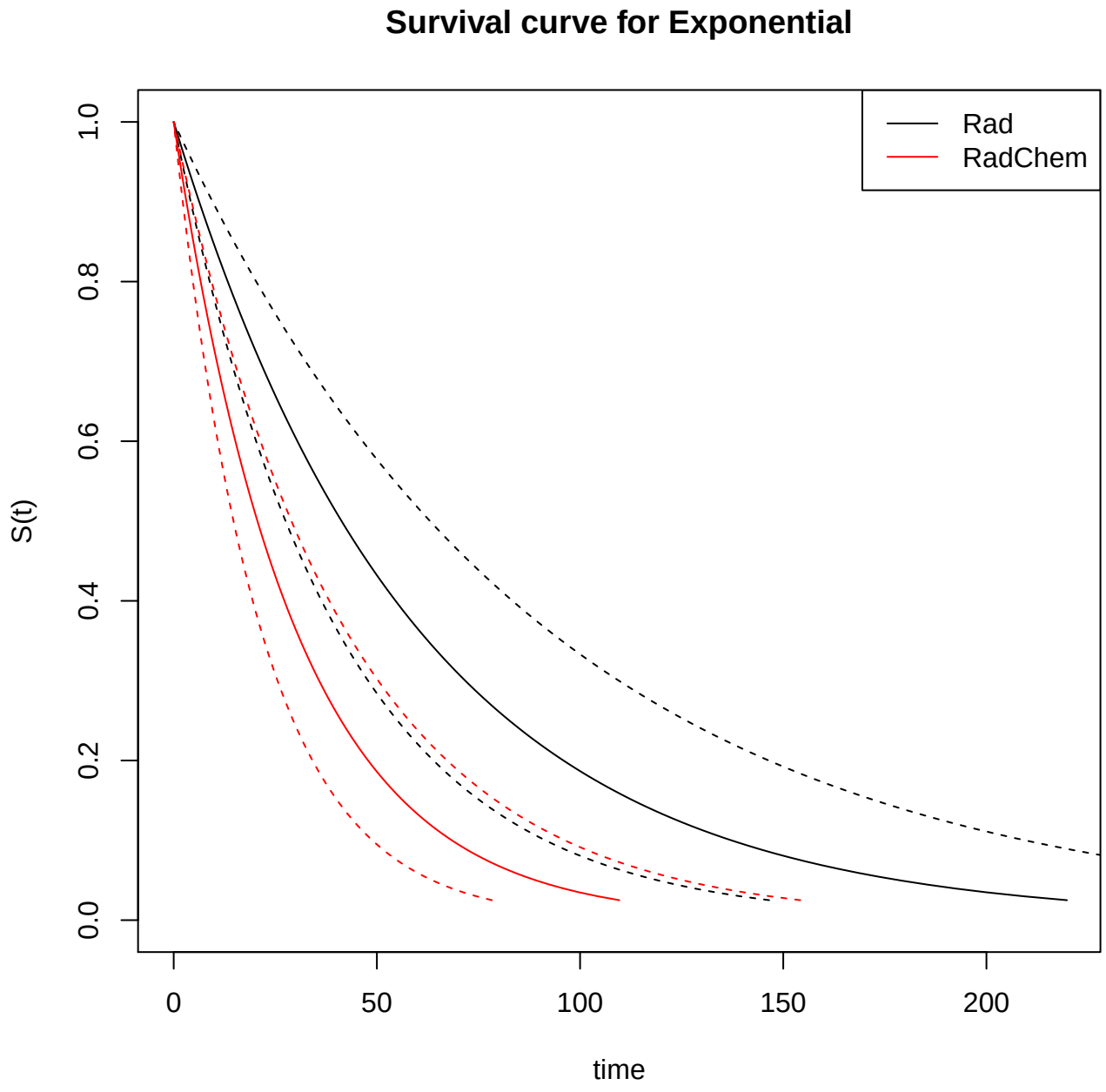
```
Loading required package:  Rcpp
Loading required package:  coda

[1] "Iterations = 6"
```



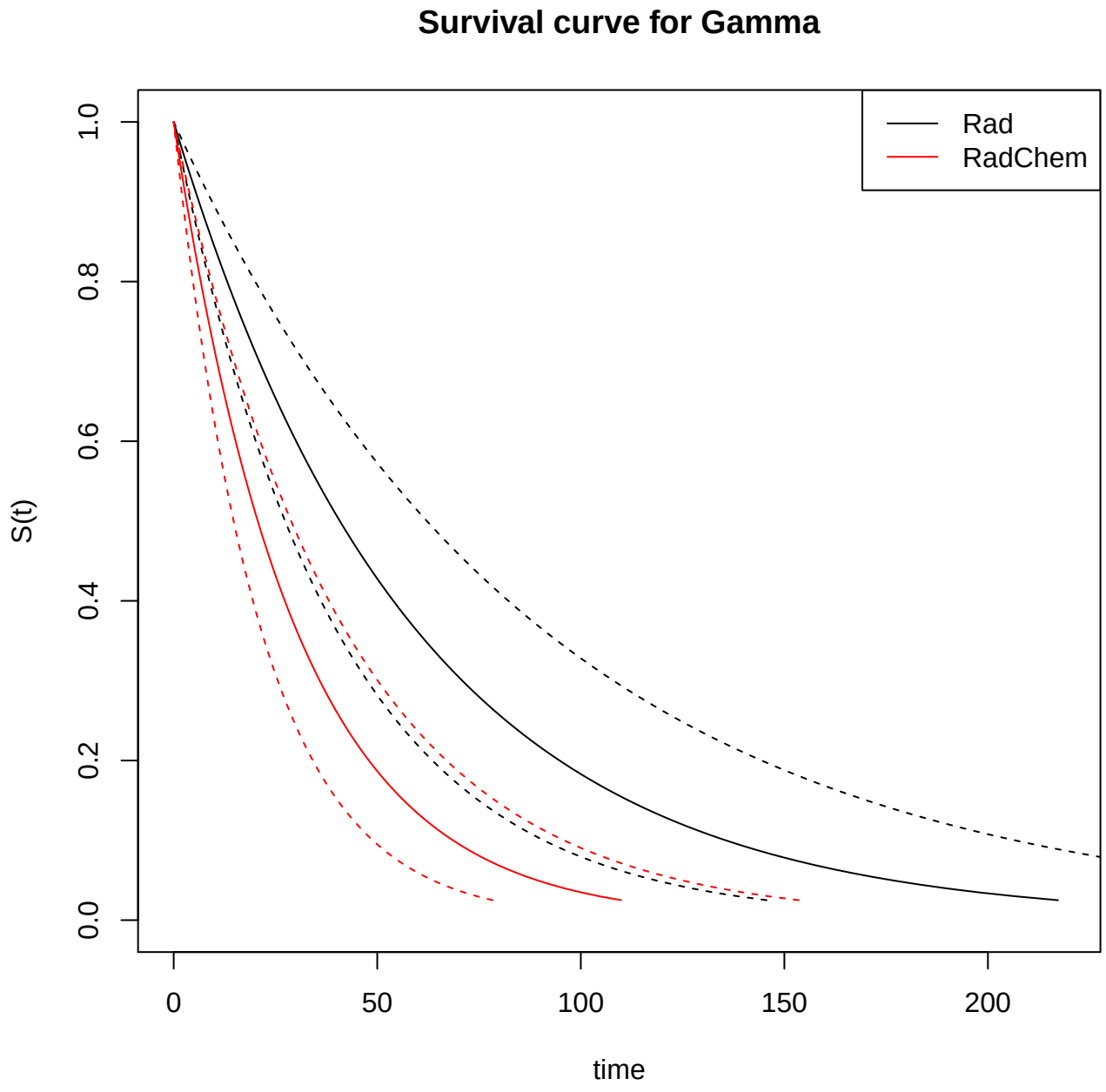
(ii) **Exponential distribution** :

```
[1] "Iterations = 5"
```

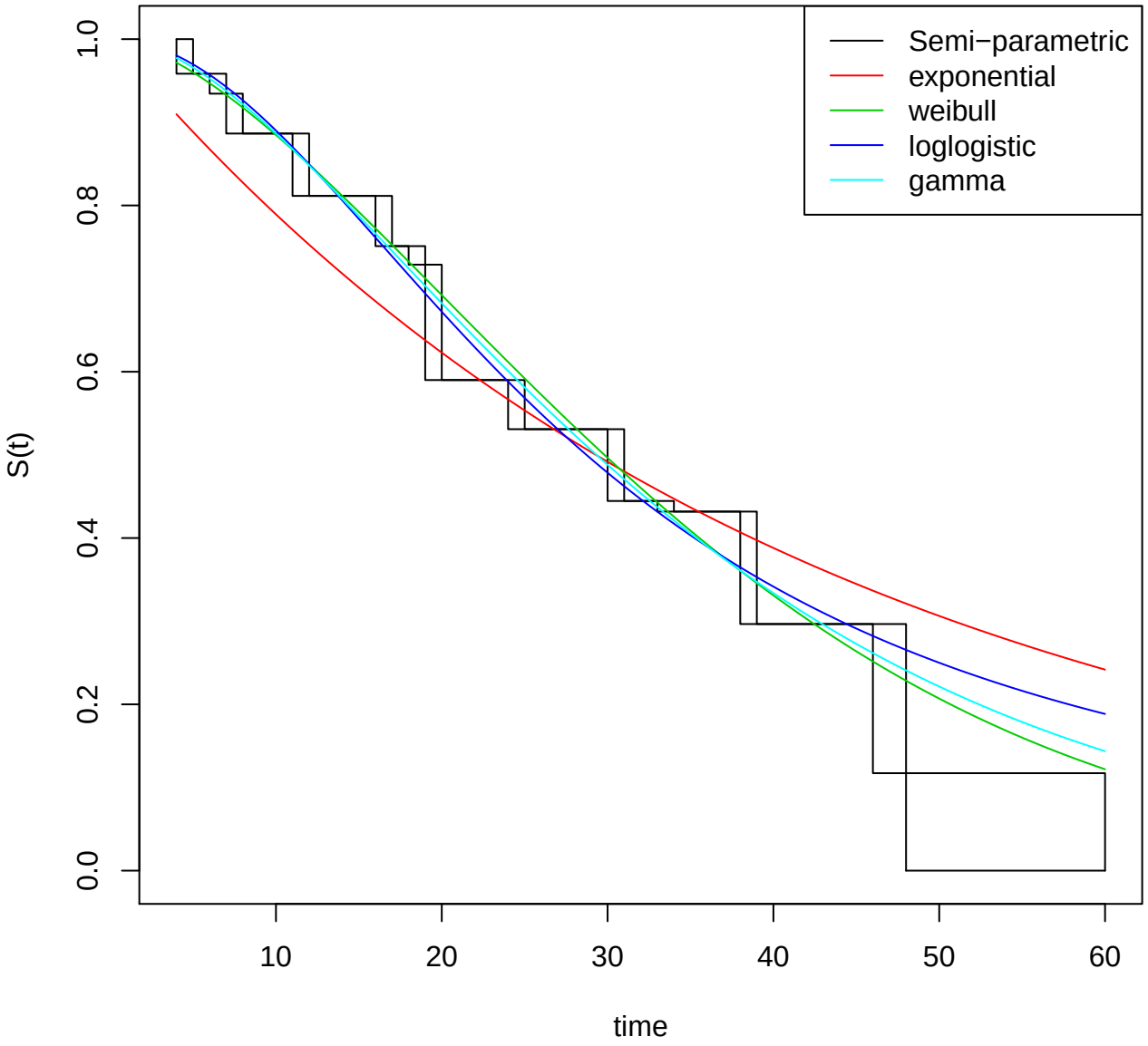


(iii) **Gamma distribution** :

```
[1] "Iterations = 5"
```



Comparing all (proportional odds model since aft model computation cannot be done by the package for semi-parametric case):



Hence we can see that the closest survival function estimated by parametric method to the one estimated by a semi-parametric method is the Weibull distribution. So in intuitionally we can say the survival function has a Weibull distribution among the ones which can be fitted by this package. Also the exponential once gave the worst fit among all since it is rigid in the sense that the hazard rate has to be constant which might not be the case for a real-world scenario.

7.Conclusion

Based on the given data we can conclude that -

- In the non-parametric setup from the estimated survival curves, we can say that, the chance of surviving more is better for the ones who received only radiotherapy than the women than radiotherapy and chemotherapy together.
- In the parametric setup, for each and every fitted family of distributions, we can say that, radiotherapy gives a better chance of survival to the women than radiotherapy and chemotherapy together.
- Note that we do not do the goodness of fit test since this is done in a heuristic setup and the mathematics behind that is cumbersome for interval censored data.

Acknowledgment

I would like to thank Dr. Biswabrata Pradhan from the SQC-OR unit in the Indian Statistical Institute (ISI), Kolkata for giving me the opportunity with this wonderful project. The project has helped me to have a more detailed knowledge over the survival analysis which had been taught in our regular course. I also want to extend my acknowledgment to the professors of the Department of Statistics, Presidency University, Kolkata who have inspired me in this topic in the first place.

References

1. Statistical Models and Methods for Lifetime Data - Jerald F. Lawless
2. TUTORIAL IN BIOSTATISTICS METHODS FOR INTERVAL-CENSORED DATA - JANE C. LINDSEY AND LOUISE M. RYAN
3. Tutorial on methods for interval-censored data and their implementation in R - Guadalupe Gomez, M Luz Calle, Ramon Oller and Klaus Langohr