Turnbull's Nonparametric Estimator for Interval-Censored Data

Suely Ruiz Giolo

Department of Statistics, Federal University of Paraná

81531-990 - Curitiba, Paraná, Brazil

e-mail: suely@est.ufpr.br

Technical Report, August, 2004

Summary

In most applications, the data may be interval-censored. By interval-censored data, we mean that a random variable of interest is known only to lie in an interval, instead of being observed exactly. In such cases, the only information we have for each individual is that their event time falls in an interval, but the exact time is unknown. A nonparametric estimate of the survival function can also be found in such interval-censored situations. The survival function is perhaps the most important function in medical and health studies. In this work we describe and illustrate the iterative procedure proposed by Turnbull (1976) to estimate such function. This procedure has been implemented in the software *R* and the code used is provided in this work.

Key-Words: nonparametric estimator, interval censoring, survival

1 Introduction

Situations where the observed response for each individual under study is either an exact survival time or a censoring time are common in practice. Other situations, however, can occur, and amongst them we find the longitudinal studies, where the individuals are followed for a pre-fixed time period or visited periodically for a fixed number of times. In this context, the time T_i $(i = 1, \cdots, n)$ until the occurrence of the event of interest for each individual is only known (whenever it occurs) to be within the interval between visits, ie, between the visit in time L_i and the visit in time U_i. Note that in such studies, the survival times T_i are no longer known exactly. It is only known that the event of interest has occurred within the interval $(L_i, U_i]$ with

$L_i < T_i \leq U_i$. Furthermore, note that if the event occurs exactly at the moment of a visit, which is very little probable but can happen, then we have an exact survival time. In this case it is assumed that $T_i = L_i = U_i$.

On the other hand, it is known for the individuals that are right censored that the event of interest did not occur until the last visit but it can happen at any time from that moment on. We therefore assumed in this case that T_i can occur within the interval (L_i, ∞) with L_i being equal to the period of time from the beginning of the study until the last visit and $U_i = \infty$.

Similarly, it is known for the individuals that are left censored, that the event of interest have occurred before the first visit and, hence, we assume that T_i has falls in the interval $(0, U_i]$ with $L_i = 0$ representing the beginning of the study and U_i is the period of time from the beginning of the study until the first visit.

Note from what we have presented so far that exact survival times as well as right and left censored data, are all special cases of interval survival data with $L_i = U_i$ for exact times, $U_i = \infty$ for right censoring and $L_i = 0$ for left censoring. We can therefore state that interval survival data generalize any situation with combinations of survival times (exact or interval) and right and left censoring that can occur in survival studies.

As usual in the analysis of non-interval survival data, here it is also of interest to estimate the survival function $S(t)$ and to assess the importance of potencial prognostic factors.

Few statistical softwares allow for such data, and for this reason a common practice amongst data analysts is to assume that the event occurring within the interval $(L_i, U_i]$, has occurred either at the upper/lower limit of the interval or, else at the middle point of each interval. Some authors, amongst them Rücker and Messerer (1988), Odell et al.(1992) and Dorey et al.(1993), state that assuming interval survival times as exact times can lead to biased estimates as well as results and conclusions that are not fully reliable.

In this work we describe a nonparametric procedure for estimation of the survival function for interval survival data. This procedure has been implemented in R[1] and it is available in the appendix.

2 Nonparametric interval-censored survival estimation

As the main objective is to estimate the survival function and investigate the importance of potential prognostic factors upon interval survival times, the number of factors under study should depend on the purpose of the study. The nonparametric procedure described here should be seen simply as an initial investigation tool and

[1]The R project is a free and open source software for statistical data analysis and can be dowloaded from http://www.r-project.org

descriptive of the survival times. It is therefore indicated for situations where one or two prognostic factors are of interest or for investigation of several factors one by one as an attempt to select those of greater importance. Such factors should preferably be qualitative and with few levels.

This does not imply that quantitative variables cannot be used, since these are categorised. Ages, for example, can be classified into three or four categories such as 0 to 5 years, 5 to 10 years and so on.

Since the event of interest is not observed for all individuals, an indicator variable for censoring should be defined. Some lines of a dataset are presented bellow to illustrate how a dataset should be organized for an analyses in R. In this example, it is assumed that for each individual it is known the upper and lower limits of the intervals within which the event of interest has occurred as well as the therapy (1 or 0) assigned to each individual. The censoring indicator variable is also assumed to be known.

Table 1: Lines of a dataset.

left	right	therapy	cens
8	12	1	1
24	31	1	1
17	27	1	1
17	NA	1	0
13	NA	0	0
11	15	0	1
..	..	..	..

Note that "NA" means that $U_i = \infty$ and that the ith observation is a right censoring.

In this section we shall present an analog Product-Limit estimator of the survival function for interval-censored data. This estimator, which has no closed form, is based on an iterative procedure and has been suggested by Turnbull (1976).

To construct the estimator, let $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_m$ be a grid of time which includes all the points L_i and U_i for $i = 1, \ldots, n$. For the ith observation, define a weight α_{ij} to be 1 if the interval (τ_{j-1}, τ_j) is contained in the interval $(L_i, U_i]$ and 0, otherwise. The weight α_{ij} indicates whether the event which occurs in the interval $(L_i, U_i]$ could have occurred at τ_j. An initial guess at $S(\tau_j)$ is made and the Turnbull´s algorithm is as follows:

- **Step 1:** Compute the probability of an event occurring at time τ_j by

$$p_j = S(\tau_{j-1}) - S(\tau_j) \qquad j = 1, \ldots, m;$$

- **Step 2:** Estimate the number of events which occurred at τ_j by

$$d_j = \sum_{i=1}^{n} \frac{\alpha_{ij} p_j}{\sum_{k=1}^{m} \alpha_{ij} p_k} \qquad j = 1, \ldots, m;$$

- **Step 3:** Compute the estimated number at risk at time τ_j by $Y_j = \sum_{k=j}^{m} d_k$;

- **Step 4:** Compute the updated Product-Limit estimator using the pseudo data found in Steps 2 and 3. If the updated estimate of S is close to the old version of S for all τ_j´s, stop the iterative process, otherwise repeat Steps 1-3, using the updated estimate of S.

As no standard statistical package produces the survival curve estimate based on Turnbull´s algorithm we have implemented the algorithm in R which is available in the appendix. The code uses as an initial guess of $S(\tau_j)$ the estimates obtained from the Kaplan-Meier estimator.

2.1 Example

The study we use for illustrating the method and the usage of the R code is a retrospective study presented by Klein and Moeschberger (1997) which was carried out to compare the cosmetic effects of radiotherapy alone versus radiotherapy and adjuvant chemotherapy on women with early breast cancer.

To compare the two treatments, a retrospective study of 46 radiation only and 48 radiation plus chemotherapy patients was conducted. Patients was observed initially every 4-6 months, but, as their recovery progressed, the interval between visits lengthened. The event of interest was the time to first appearance of moderate or severe breast retraction. As the patients were observed only at some random times, the exact time, T_i, of breast retraction is known only to fall within the interval between visits.

Patients with no moderate or severe breast retraction until the last visit were classified as right-censored and then the end point of their intervals were assumed to be $U_i = \infty$ as well as L_i were assumed as the time from the begining to the last visit. The data are shown in Table 2.

Using the Turnbull's algorithm we obtained the estimated survival functions for radiotherapy only and radiation plus chemotherapy groups showed in Figure 1. Note that the estimated survival curves do not show striking differences from 0 to 18 months. From 18 onwards, however, a fast decay of the curve is seen for patients given radiotherapy plus chemotherapy whilst this does not happen for those given only radiotherapy. Note, for instance, that only 11.06% of the patients in the radiotherapy plus chemotherapy group is estimated to be free of any evidence of breast retraction at time $t = 40$ months against 47.37% in the radiotherapy group at this very same point

Table 2: Time to cosmetic deterioration in breast cancer patients with two treatments.

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

in time. A longer time to cosmetic deterioration for patients given only radiotherapy is, therefore, indicated from the estimated survival curves presented in Figure 1.

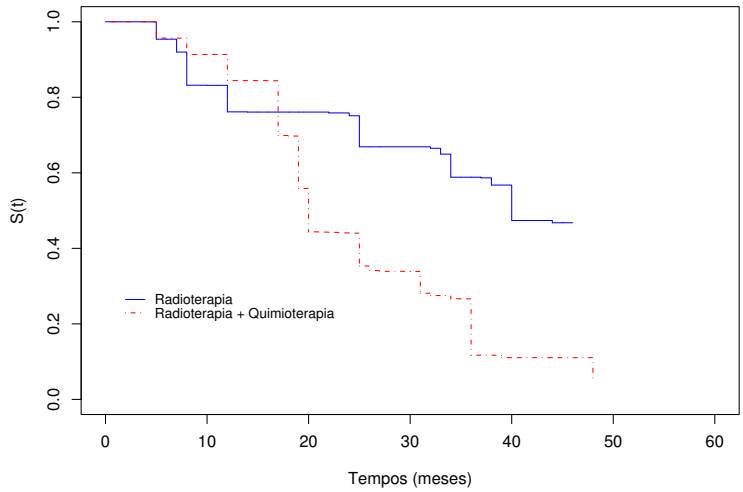

Figure 1: Estimated survival based on interval-censored data.

Using the midpoint of each interval, which is a common practice amongst analysts due to the lack of well-known statistical methodology and available software, and then applying the Kaplan-Meier method we obtained the estimated survival curves presented in Figure 2. The curves estimated previously are also shown in the figure. Comparing the curves we can see that the estimates obtained using both, the midpoints and the intervals, are very similar from each other at several times but they trend to be under or overestimed at others. Although not shown here, under or overestimation become more evident if it is assumed that the event occurred at the end or at the beginning of each interval instead of at the midpoint. The range of each interval also contributes for the magnitude of these differences. They are more

accentuated as the range of each interval increases.

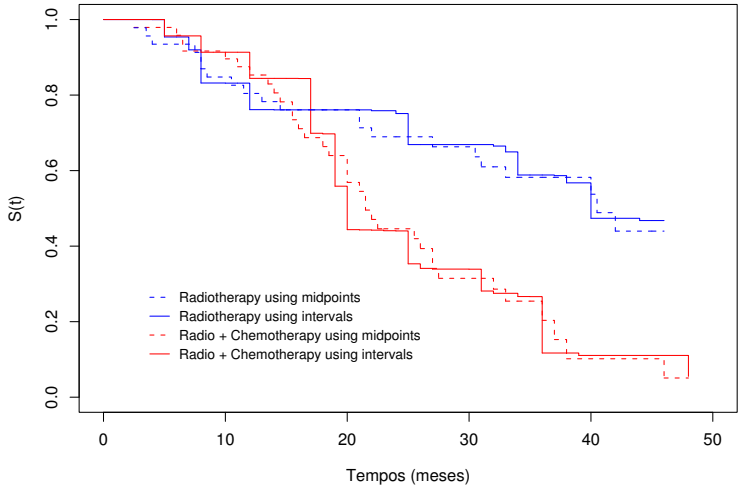

Figure 2: Estimated survival functions using midpoints and intervals.

From these results and according to some authors such as Lindsey et al. (1998), the analysis of interval-censored data assuming that the event occurred at the end (or beginning or midpoint) of each interval, and then applying methods for standard time-to-event data can lead to invalid inferences. Thus, analysts that have been using methods for standard time-to-event data for analyzing interval-censored data should be not too confident at their conclusions.

Acknowledgement

I would like to thank Silvia E. Shimakura and Elias T. Krainski for their helpful suggestions in some details of the Turnbull.R function.

References

DOREY, F.J., LITTLE, R.J., SCHENKER, N. Multiple imputation for threshold-crossing data with interval censoring. *Statistics in Medicine*, 12, 1589-1603, 1993.

KLEIN, J. P, MOESCHBERGER, M. *Survival Analysis*. New York: Springer Verlag, 1997.

LINDSEY, J.C., RYAN, L.M. Tutorial in Biostatistics: methods for interval-censored data. *Statistics in Medicine*, 17, 219-238, 1998.

ODELL P.M., ANDERSON, K.M., D'AGOSTINHO, R.B. Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics*, 48, 951-959, 1992.

RÜCKER G., MESSERER D. Remission duration: an example of interval-censored observation. *Statistics in Medicine*, 7, 1139-1145, 1988.

APPENDIX A - R commands for obtaining the survival curves

```
================================================
 Survival curves using intervals - Figure 1
================================================

require(survival)
source("Turnbull.R")                        # Turnbull.R available in the Appendix B

dat <- read.table('breast.txt',header=T)    # breast.txt available in the Appendix C
dat1 <- dat[dat$ther==1,]
dat1$right[is.na(dat1$right)] <- Inf
tau <- cria.tau(dat1)
p <- S.ini(tau=tau)
A <- cria.A(data=dat1,tau=tau)
tb1 <- Turnbull(p,A,dat1)
tb1

dat1 <- dat[dat$ther==0,]
dat1$right[is.na(dat1$right)] <- Inf
tau <- cria.tau(dat1)
p <- S.ini(tau=tau)
A <- cria.A(data=dat1,tau=tau)
tb2 <- Turnbull(p,A,dat1)
tb2

plot(tb1$time,tb1$surv,lty=1, col = 4,type="s",ylim=c(0,1),xlim=range(c(0,60)),
     xlab="Tempos (meses)",ylab="S(t)")
lines(tb2$time,tb2$surv,lty=4,col=2,type="s")
legend(1,0.3,lty=c(1,4),col=c(4,2),c("Radioterapia","Radioterapia + Quimioterapia"), bty="n",cex=0.8)

================================================
 Survival curves using midpoints - Figure 2
================================================

p <-dat$left+((dat$right-dat$left)/2)
pm <-ifelse(is.finite(p),p,dat$left)
cens <- ifelse(is.finite(p),1,0)
ekm<-survfit(Surv(pm,cens)~ther,type=c("kaplan-meier"),data=dat)

plot(tb1$time,tb1$surv,lty=1,type="s",col=4,ylim=c(0,1),xlim=c(0,50),xlab="Tempos (meses)",ylab="S(t)")
lines(tb2$time,tb2$surv,lty=1,col=2,type="s")
lines(ekm[1]$time,ekm[1]$surv,type="s",col=2,lty=2)
lines(ekm[2]$time,ekm[2]$surv,type="s",col=4,lty=2)
legend(3,0.30,lty=2,col=4, "Radiotherapy using midpoints", bty="n",cex=0.8)
legend(3,0.25,lty=1,col=4, "Radiotherapy using intervals", bty="n",cex=0.8)
legend(3,0.2,lty=2,col=2,"Radio + Chemotherapy using midpoints", bty="n",cex=0.8)
legend(3,0.15,lty=1,col=2,"Radio + Chemotherapy using intervals", bty="n",cex=0.8)
```

APPENDIX B - Turnbull.R function

```
cria.tau <- function(data){
  l <- data$left
  r <- data$right
  tau <- sort(unique(c(l,r[is.finite(r)])))
  return(tau)
}

S.ini <- function(tau){
  m<-length(tau)
  ekm<-survfit(Surv(tau[1:m-1],rep(1,m-1)))
  So<-c(1,ekm$surv)
  p <- -diff(So)
  return(p)
}

cria.A <- function(data,tau){
  tau12 <- cbind(tau[-length(tau)],tau[-1])
  interv <- function(x,inf,sup) ifelse(x[1]>=inf & x[2]<=sup,1,0)
  A <- apply(tau12,1,interv,inf=data$left,sup=data$right)
  id.lin.zero <- which(apply(A==0, 1, all))
  if(length(id.lin.zero)>0) A <- A[-id.lin.zero, ]
  return(A)
}

Turnbull <- function(p, A, data, eps=1e-3, iter.max=200, verbose=FALSE){
  n<-nrow(A)
  m<-ncol(A)
  Q<-matrix(1,m)
  iter <- 0
  repeat {
    iter <- iter + 1
    diff<- (Q-p)
    maxdiff<-max(abs(as.vector(diff)))
    if (verbose)
      print(maxdiff)
    if (maxdiff<eps | iter>=iter.max)
      break
    Q<-p
    C<-A%*%p
    p<-p*((t(A)%*%(1/C))/n)
  }
    cat("Iterations = ", iter,"\n")
    cat("Max difference = ", maxdiff,"\n")
    cat("Convergence criteria: Max difference < 1e-3","\n")
  dimnames(p)<-list(NULL,c("P Estimate"))
  surv<-round(c(1,1-cumsum(p)),digits=5)
  right <- data$right
   if(any(!(is.finite(right)))){
    t <- max(right[is.finite(right)])
    return(list(time=tau[tau<t],surv=surv[tau<t]))
  }
  else
    return(list(time=tau,surv=surv))
}
```

APPENDIX C - breast.txt

left	right	ther	cens
0	7	1	1
0	8	1	1
0	5	1	1
4	11	1	1
5	12	1	1
5	11	1	1
6	10	1	1
7	16	1	1
7	14	1	1
11	15	1	1
11	18	1	1
15	NA	1	0
17	NA	1	0
17	25	1	1
17	25	1	1
18	NA	1	0
19	35	1	1
18	26	1	1
22	NA	1	0
24	NA	1	0
24	NA	1	0
25	37	1	1
26	40	1	1
27	34	1	1
32	NA	1	0
33	NA	1	0
34	NA	1	0
36	44	1	1
36	48	1	1
36	NA	1	0
36	NA	1	0
37	44	1	1
37	NA	1	0
37	NA	1	0
37	NA	1	0
38	NA	1	0
40	NA	1	0
45	NA	1	0
46	NA	1	0
46	NA	1	0
46	NA	1	0
46	NA	1	0
46	NA	1	0
46	NA	1	0
46	NA	1	0
0	22	0	1
0	5	0	1
4	9	0	1
4	8	0	1
5	8	0	1
8	12	0	1
8	21	0	1
10	35	0	1
10	17	0	1
11	13	0	1
11	NA	0	0
11	17	0	1
11	NA	0	0
11	20	0	1
12	20	0	1
13	NA	0	0
13	39	0	1
13	NA	0	0
13	NA	0	0
14	17	0	1
14	19	0	1
15	22	0	1
16	24	0	1
16	20	0	1
16	24	0	1
16	60	0	1
17	27	0	1
17	23	0	1
17	26	0	1
18	25	0	1
18	24	0	1
19	32	0	1
21	NA	0	0

```
22      32      0       1
23      NA      0       0
24      31      0       1
24      30      0       1
30      34      0       1
30      36      0       1
31      NA      0       0
32      NA      0       0
32      40      0       1
34      NA      0       0
34      NA      0       0
35      NA      0       0
35      39      0       1
44      48      0       1
48      NA      0       0
```