Usage of R package LTRCtrees

Wei Fu, Jeffrey S. Simonoff

2021-01-11

Vignette Info

The `LTRCtrees` package is designed to fit survival trees for left-truncated and right censored (LTRC) data as described in `Fu and Simonoff (2017a)`[1], and interval-censored survival data as described in `Fu and Simonoff (2017b)`[2]. For right censored data, the tree algorithms in `rpart`[3] and `partykit`[4], respectively, can be used to fit a survival tree. The `LTRCtrees` package extends the survival tree algorithms in these two packages to fit trees for LTRC data, as well as interval-censored survival data. Since the package uses the fitting functions in `rpart` and `partykit`, these packages will be installed if they have not already been installed on the user's machine, but they are not loaded into the session when the `LTRCtrees` package is loaded. In addition to fitting trees to LTRC data, a major benefit of these extensions is that they can be used to fit survival trees for data with time-varying covariates.

This document gives several examples of how to use the LTRC trees and ICtree (interval-censored tree) implemented in the `LTRCtrees` package.

Example of fitting survival trees for LTRC data

The assay of serum free light chain data example

The assay of serum free light chain data for 7874 subjects in the R package `survival`[5] is used as a data example to show how to fit survival trees for LTRC data. The original data were obtained from the residents of Olmsted County aged 50 or greater, and therefore the data are left-truncated at age 50. The original analysis of this data set was based on the Cox model with age as a covariate and time after enrollment in the study as response. However, in a mortality study such as this one, since greater age is almost always associated with higher risk of death, it is not that meaningful (or surprising) to have age as a (significant) covariate. Also, it is often the case that the real response of interest should be the subject's life length, not the time from enrollment in the study to death/censoring.

In this example, we use age as a left-truncation point and death/censored time as the event time.

```r
## Adjust data & clean data
library(survival)
set.seed(0)
## Since LTRCART uses cross-validation to prune the tree, specifying the seed
## guarantees that the results given here will be duplicated in other analyses
Data <- flchain
Data <- Data[!is.na(Data$creatinine),]
Data$End <- Data$age + Data$futime/365
DATA <- Data[Data$End > Data$age,]
names(DATA)[6] <- "FLC"
```

The key here is to have a `(start, stop, event)` triplet in the data, where `start` is the left-truncation time, `stop` is the right censored/event time and `event` is the event indicator, where 1 means death and 0 means censored. In our case, the triplet is `(age, End, death)`.

```
## Setup training set and test set
Train = DATA[1:500,]
Test = DATA[1000:1020,]
```

Here we set up a training set and a test set to illustrate predictions using the fitted trees.

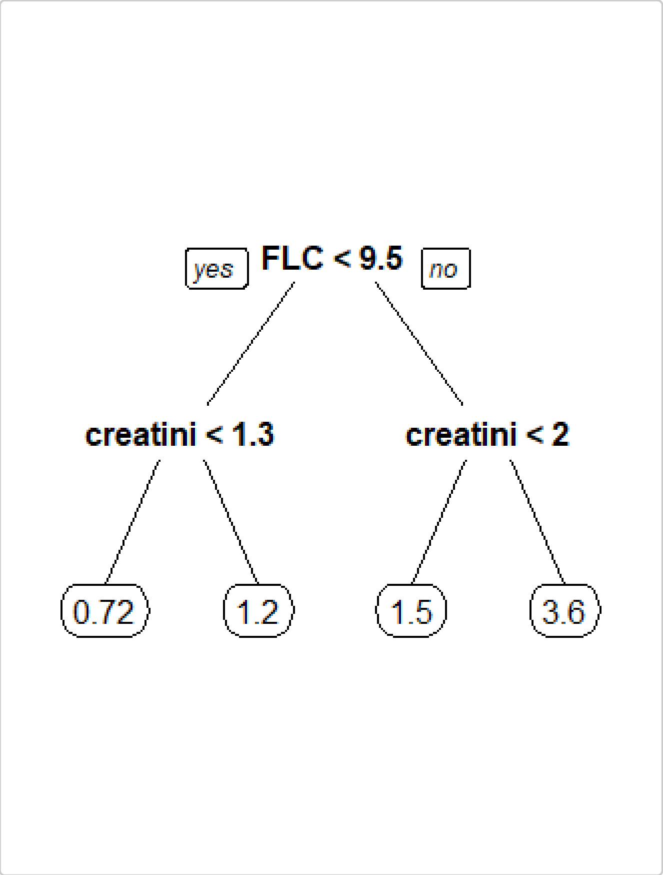
```
## Fit LTRCART and LTRCIT survival tree
library(LTRCtrees)
LTRCART.obj <- LTRCART(Surv(age, End, death) ~ sex + FLC + creatinine, Train)
LTRCIT.obj <- LTRCIT(Surv(age, End, death) ~ sex + FLC + creatinine, Train)
```

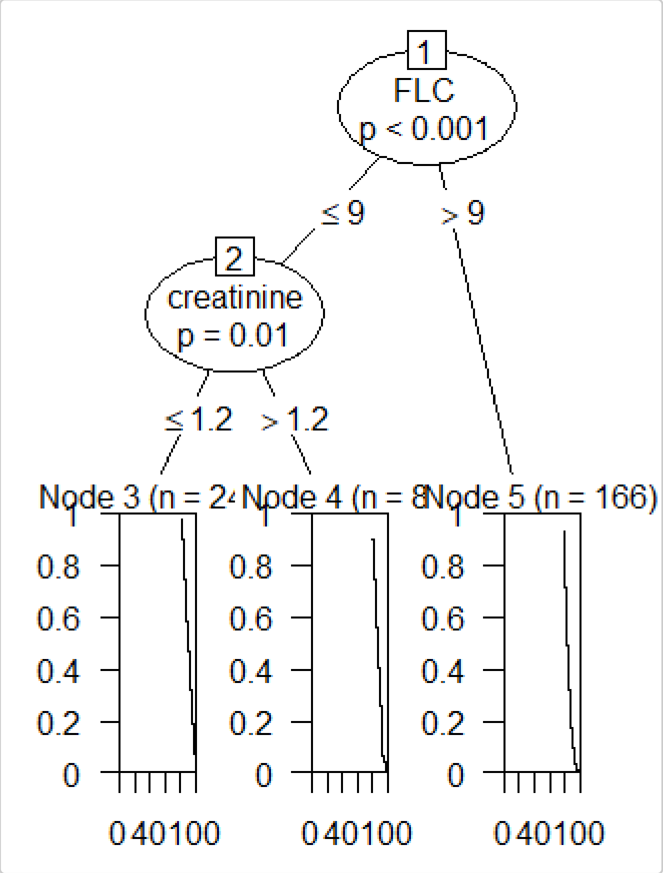
```
## Loading required namespace: inum
```

```
## Putting Surv(End, death) in formula would result an error message
## since both LTRCART and LTRCIT are expecting Surv(time1, time2, event)

## Plot the fitted LTRCART tree using rpart.plot function in rpart.plot[6] package
library(rpart.plot)
```

```
## Loading required package: rpart
```

```
prp(LTRCART.obj, roundint=FALSE)

## Plot the fitted LTRCIT tree
plot(LTRCIT.obj)
```





Note that the terminal node of the LTRCART tree (left panel) displays the relative risk on that node (relative to an estimated baseline hazard), while the terminal node of the LTRCIT tree (right panel) displays the fitted Kaplan-Meier curve for the observations in that node. One can also plot Kaplan-Meier curves on terminal nodes of the LTRCART tree by converting it to `party` object.

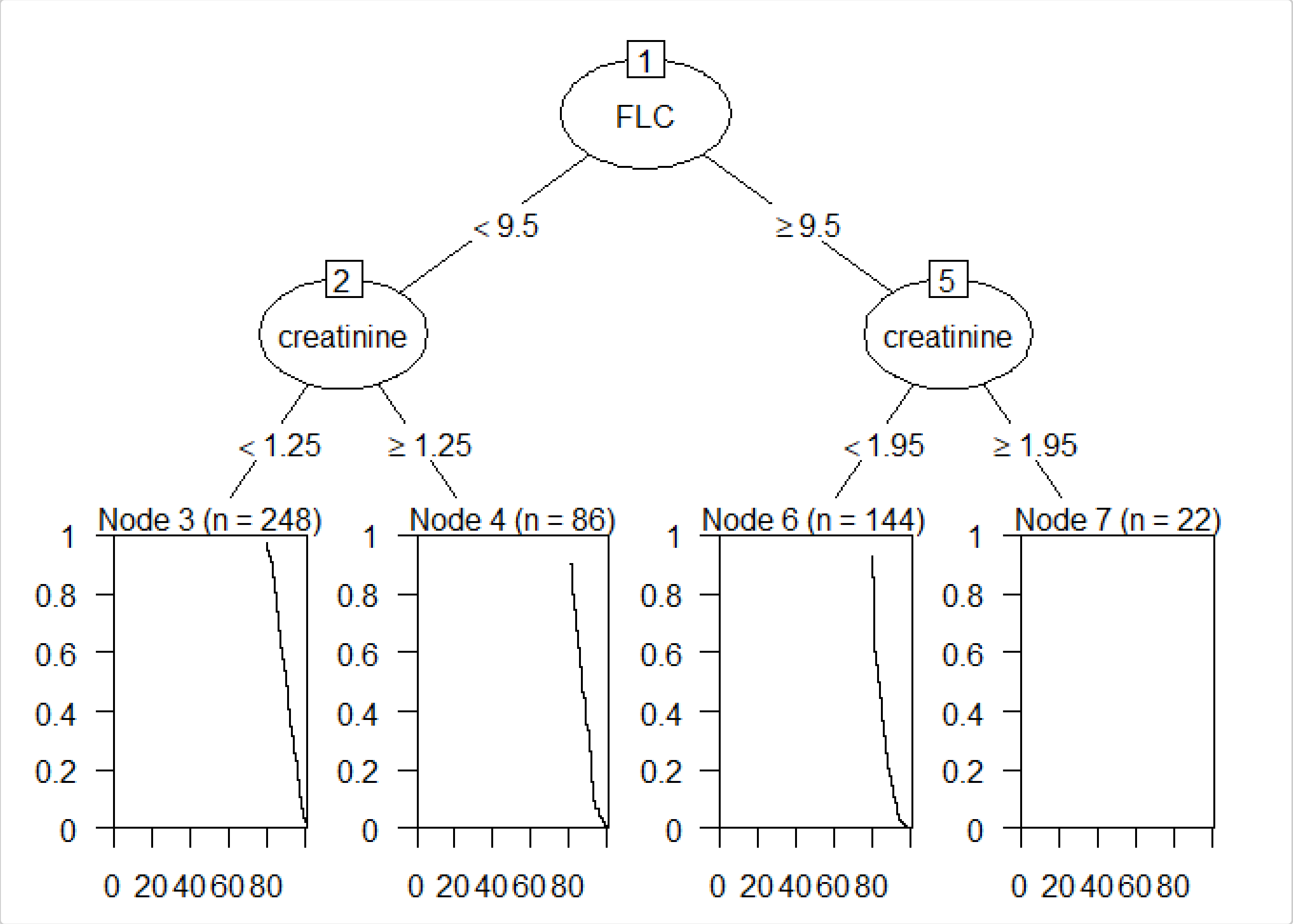
```r
library(partykit)
```

```
## Loading required package: grid
```

```
## Loading required package: libcoin
```

```
## Loading required package: mvtnorm
```

```
LTRCART.obj.party <- as.party(LTRCART.obj)
LTRCART.obj.party$fitted[["(response)"]]<- Surv(Train$age, Train$End, Train$death)
plot(LTRCART.obj.party)
```



Since the class of an LTRCIT object is `party`, the prediction method on an LTRCIT object is the same as those for any trees of class `party` using the `partykit` package.

```
## predict median survival time on test data using fitted LTRCIT tree
LTRCIT.pred <- predict(LTRCIT.obj, newdata=Test, type = "response")
head(LTRCIT.pred)
```

```
##     1057     1058     1059     1060     1061     1062
## 90.47123 90.47123 90.47123 90.47123 90.47123 83.41644
```

```
## predict Kaplan Meier survival curve on test data
## return a list of survfit objects -- the predicted KM curves
LTRCIT.pred <- predict(LTRCIT.obj, newdata=Test, type = "prob")
head(LTRCIT.pred,2)
```

```
## $`1057`
## Call: survfit(formula = y ~ 1, weights = w, subset = w > 0)
## 
## records   n.max n.start  events  median 0.95LCL 0.95UCL 
##   248.0   144.0    36.0   189.0    90.5    88.4    91.4 
## 
## $`1058`
## Call: survfit(formula = y ~ 1, weights = w, subset = w > 0)
## 
## records   n.max n.start  events  median 0.95LCL 0.95UCL 
##   248.0   144.0    36.0   189.0    90.5    88.4    91.4
```

The prediction function on a survival `rpart` object can only predict the relative risk, so we can only predict the relative risk on an LTRCART object.

```
## Predict relative risk on test set
LTRCART.pred <- predict(LTRCART.obj, newdata=Test)
head(LTRCART.pred)
```

```
##      1057      1058      1059      1060      1061      1062 
## 0.7212083 0.7212083 0.7212083 0.7212083 0.7212083 1.5257057
```

In order to predict the median survival time and Kaplan-Meier curves based on an LTRCART object, one can use the `Pred.rpart` function available in the `LTRCtrees` package.

```
## Predict median survival time and Kaplan Meier survival curve
## on test data using Pred.rpart
LTRCART.pred <- Pred.rpart(Surv(age, End, death) ~ sex + FLC + creatinine, Train, Test)
head(LTRCART.pred$KMcurves, 2)  ## list of predicted KM curves
```

```
## [[1]]
## Call: survfit(formula = Formula, data = subset)
## 
## records   n.max n.start  events  median 0.95LCL 0.95UCL 
##   248.0   144.0    36.0   189.0    90.5    88.4    91.4 
## 
## [[2]]
## Call: survfit(formula = Formula, data = subset)
## 
## records   n.max n.start  events  median 0.95LCL 0.95UCL 
##   248.0   144.0    36.0   189.0    90.5    88.4    91.4
```

```
head(LTRCART.pred$Medians)  ## vector of predicted median survival time
```

```
## [1] 90.5 90.5 90.5 90.5 90.5 83.8
```

Note that in order to use the `Pred.rpart` function one needs to pass both the Training and Test sets into the function. If the formula passed to `Pred.rpart` has the form Surv(time1, time2, event), then the function recognizes this as LTRC data and LTRCART is called internally; if the formula has the form Surv(time, event), then the function recognizes this as ordinary right censored data, and the `rpart` function in the `rpart` package is called internally.

Examples of fitting survival trees with time-varying covariates

To fit time-varying covariates survival trees using `LTRCART` or `LTRCIT`, data need to be prepared/transformed into the so-called Andersen-Gill style. This is exactly the same data format needed to fit the Cox model with time-varying covariates.

Patient.ID	Sex	Blood.pressure	Start	End	Death
1	F	100	0	10	0
1	F	89	10	20	0
1	F	120	20	27	1
2	M	110	0	10	0
2	M	105	10	19	0

This table gives an example of the form of Andersen-Gill style data. Information for two patients is presented in this table, with Sex being a time-independent covariate and Blood pressure being a time-varying covariate. For the first patient (patient.ID=1), the event-death is observed at time 27, while the 3 measurements of Sex and blood pressure are recorded at the beginning and time 10 and 20 respectively. Information for the second patient has a similar interpretation except he is censored at time 19.

Note that each row represents a time interval, where within each interval all covariates are assumed to have constant values. The first row means the first patient is a female with blood pressure equal to 100 during the interval (0,10]. Of course, blood pressure is unlikely to keep at a fixed level during this time interval (since it is time-varying), but in practice a patient's blood pressure is measured only at the time he/she visits the clinic, so it is common practice to treat it as constant between the two visits.

Note that as a result each row has the form of LTRC data – left-truncated at Start and right-censored or experience event at End, with Death being the event indicator.

The Mayo Clinic Primary Biliary Cirrhosis Data example

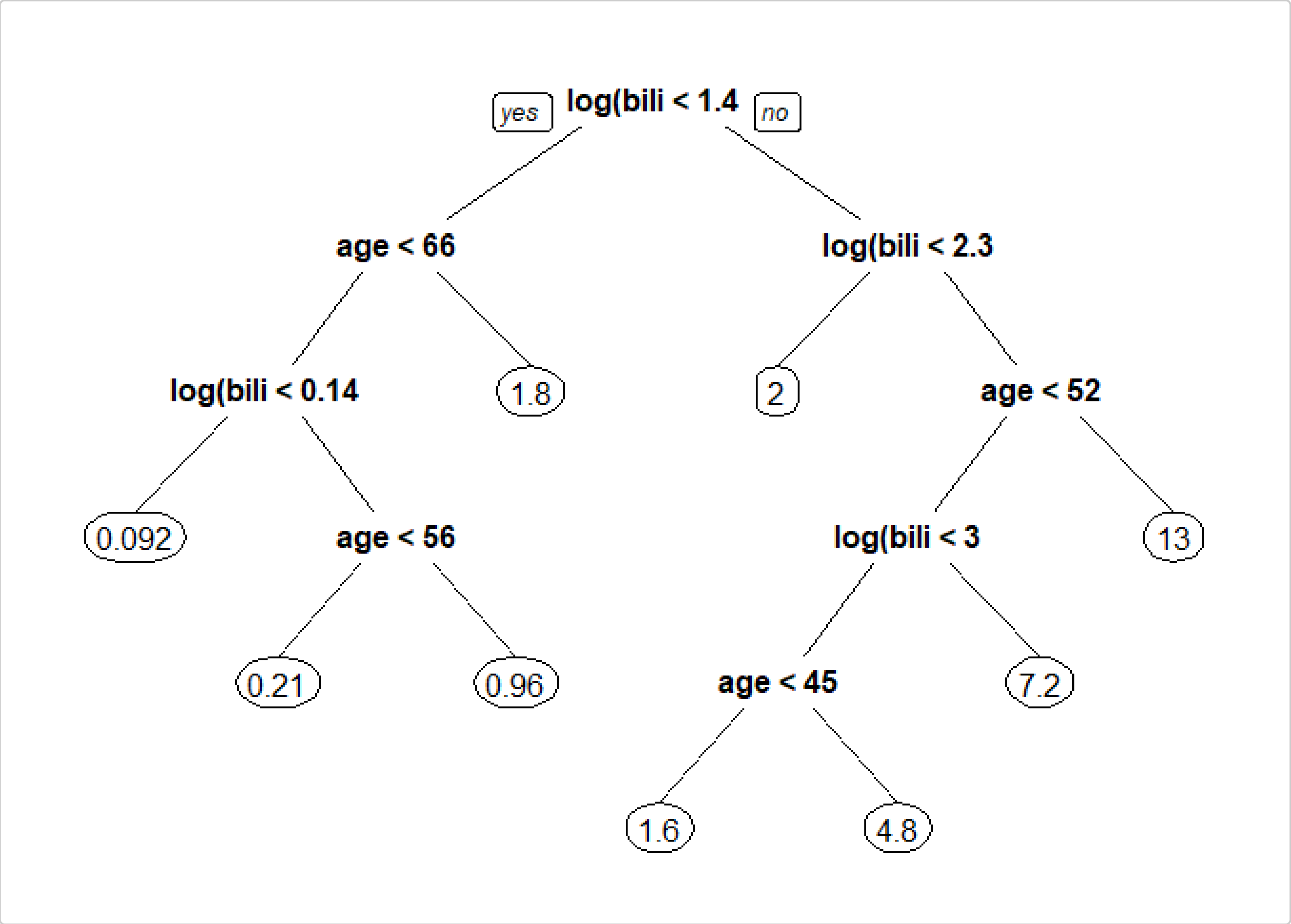
The `pbcseq` dataset in the `survival` package is used as an example to illustrate fitting survival trees with time-varying covariates. These data were obtained from 312 patients with primary biliary cirrhosis (PBC) enrolled in a double-blind, placebo-controlled, randomized trial conducted between January, 1974 and May, 1984 at the Mayo Clinic to evaluate the use of D-penicillamine for treating PBC. A comprehensive clinical and laboratory database was established on each patient. Follow-up was extended to April, 1988, by which time 140 of the patients had died and 29 had undergone orthotopic liver transplantation. These patients generated 1,945 patient visits that enable us to study the change in the prognostic variables of PBC.

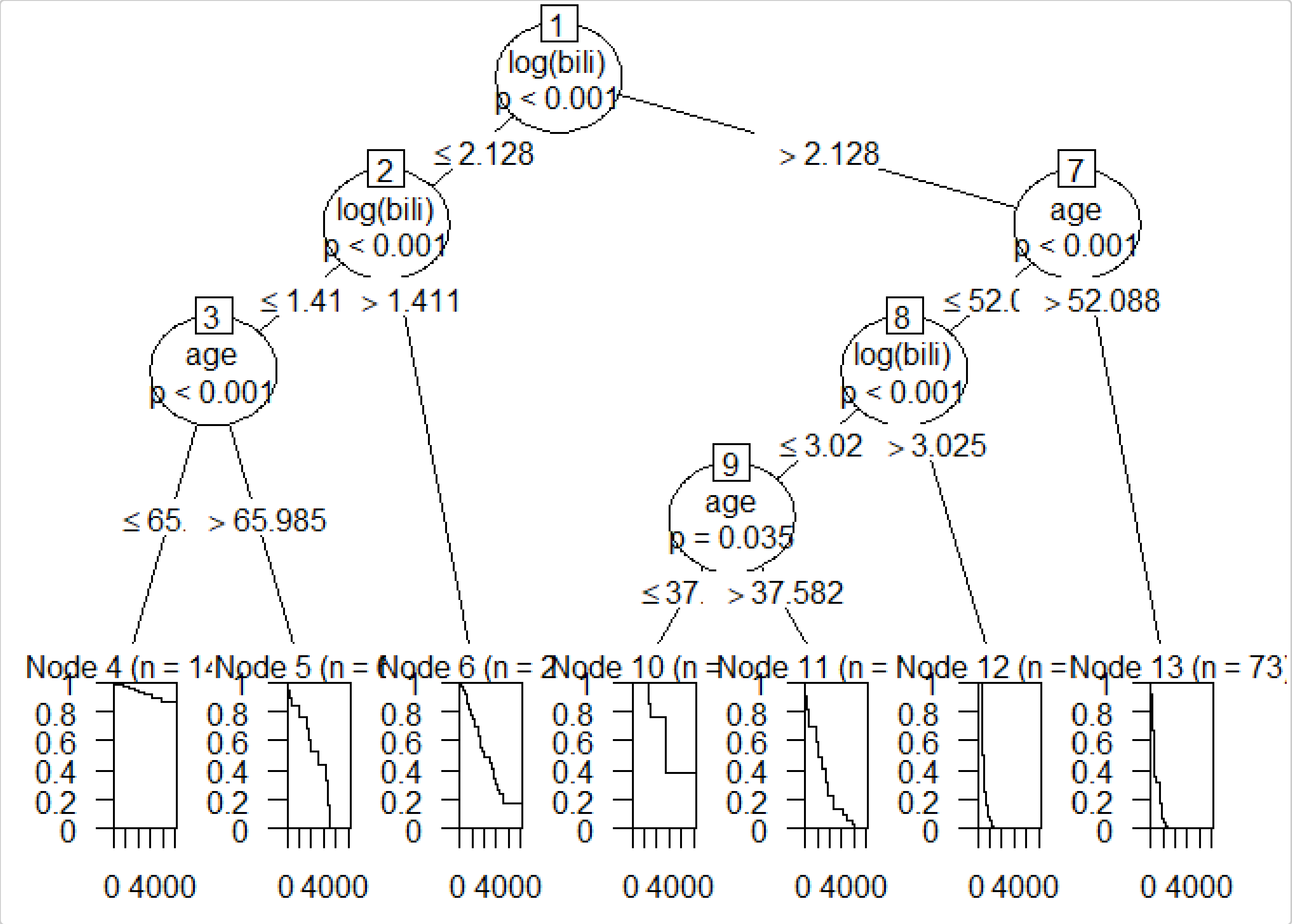
```r
set.seed(0)
library(survival)
## Create the start-stop-event triplet needed for coxph and LTRC trees
first <- with(pbcseq, c(TRUE, diff(id) !=0)) #first id for each subject
last <- c(first[-1], TRUE) #last id
time1 <- with(pbcseq, ifelse(first, 0, day))
time2 <- with(pbcseq, ifelse(last, futime, c(day[-1], 0)))
event <- with(pbcseq, ifelse(last, status, 0))
event <- 1*(event==2)
```

```
pbcseq$time1 <- time1
pbcseq$time2 <- time2
pbcseq$event <-  event

## Fit the Cox model and LTRC trees with time-varying covariates
fit.cox <- coxph(Surv(time1, time2, event) ~ age + sex + log(bili), pbcseq)
LTRCIT.fit <- LTRCIT(Surv(time1, time2, event) ~ age + sex + log(bili), pbcseq)
LTRCART.fit <- LTRCART(Surv(time1, time2, event) ~ age + sex + log(bili), pbcseq)

## Result of the Cox model with time-varying covariates
fit.cox
```

```
## Call:
## coxph(formula = Surv(time1, time2, event) ~ age + sex + log(bili),
##     data = pbcseq)
##
##               coef exp(coef) se(coef)      z        p
## age       0.068285  1.070670 0.008624  7.918 2.42e-15
## sexf      0.182856  1.200641 0.234190  0.781    0.435
## log(bili) 1.465486  4.329646 0.094157 15.564  < 2e-16
##
## Likelihood ratio test=352.8  on 3 df, p=< 2.2e-16
## n= 1945, number of events= 140
```

```
## plots of fitted survival trees with time-varying covariates
prp(LTRCART.fit,type=0, roundint=FALSE)
plot(LTRCIT.fit)
```





Note that both survival trees fitted using LTRCART and LTRCIT, respectively, identify `age` and `log(bili)` as risk factors, which is consistent with the Cox model result.

The Stanford Heart Transplant data example

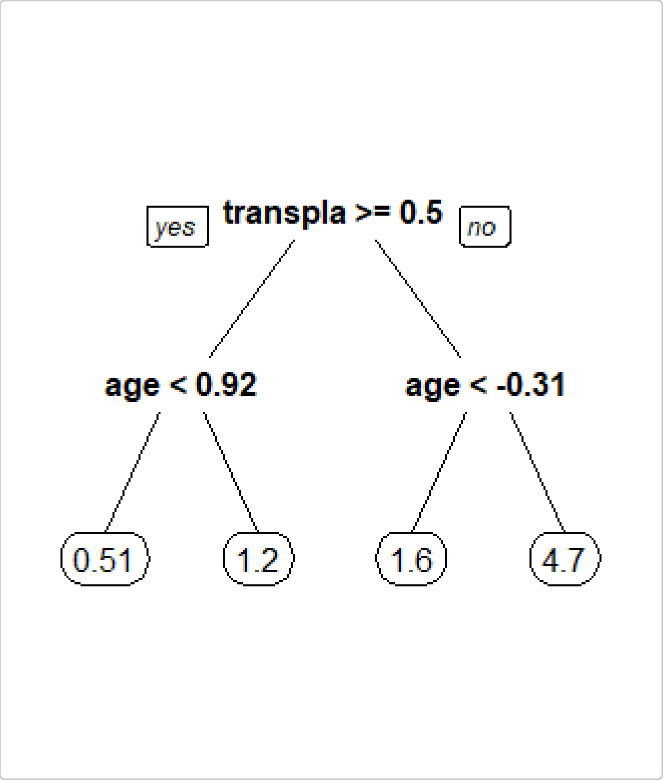
In the previous example, the `pbcseq` data are in the long format – each row represents a start-stop-event triplet, which contains the information in the time interval (start, stop]. Note that all covariates have constant values inside each such interval, while each subject could have multiple rows (multiple intervals) to represent the time-varying covariates effect, i.e. different rows (intervals) of the same subject contain different covariates' value. The long format data can be directly used to fit survival trees with time-varying covariates using either LTRCART or LTRCIT.

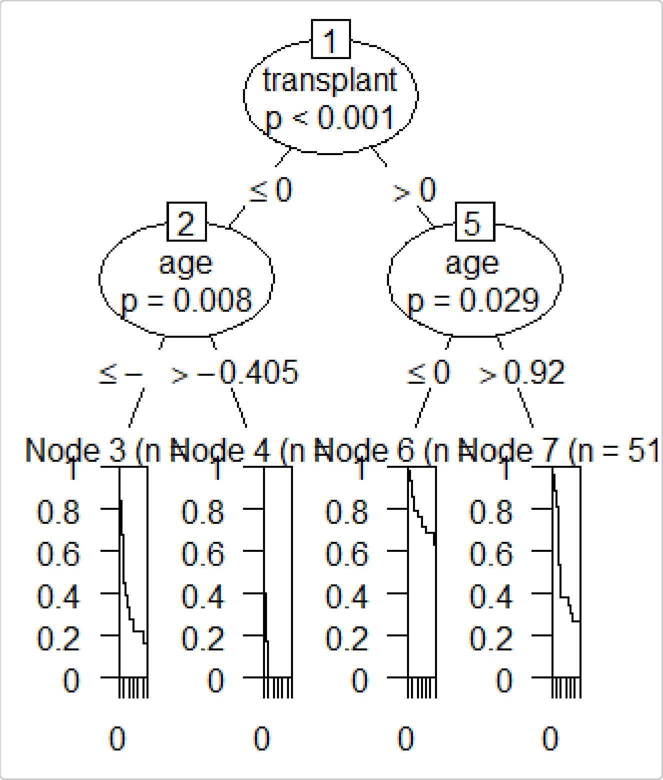
It is common to have time-varying covariates survival data that is in wide format – each row contains all of the information of one subject. To use the tree algorithms in `LTRCtrees`, we first need to transform the wide format into the long format. The `tmerge` function in the `survival` package can be used to serve this purpose.

```r
library(survival)
### transform the wide format data into the long format data using tmerge function
### from survival package on Stanford Heart Transplant data
jasa$subject <- 1:nrow(jasa)

tdata <- with(jasa, data.frame(subject = subject,
    futime= pmax(.5, fu.date - accept.dt),
    txtime= ifelse(tx.date== fu.date,
    (tx.date -accept.dt) -.5,
    (tx.date - accept.dt)),
    fustat = fustat))

sdata <- tmerge(jasa, tdata, id=subject,death = event(futime, fustat),trt = tdc(txtime), options=
  list(idname="subject"))

sdata$age <- sdata$age - 48

sdata$year <- as.numeric(sdata$accept.dt - as.Date("1967-10-01"))/365.25

Cox.fit <- coxph(Surv(tstart, tstop, death) ~ age+ surgery, data= sdata)
LTRCART.fit <- LTRCART(Surv(tstart, tstop, death) ~ age + transplant, data = sdata)
LTRCIT.fit <- LTRCIT(Surv(tstart, tstop, death) ~ age + transplant, data = sdata)

## results
Cox.fit
```

```
## Call:
## coxph(formula = Surv(tstart, tstop, death) ~ age + surgery, data = sdata)
##
##             coef exp(coef) se(coef)     z      p
## age      0.03068   1.03115  0.01364  2.25 0.0245
## surgery -0.77284   0.46170  0.35953 -2.15 0.0316
##
## Likelihood ratio test=10.72  on 2 df, p=0.00471
## n= 170, number of events= 75
```

```r
prp(LTRCART.fit, roundint=FALSE)
plot(LTRCIT.fit)
```





Example of fitting survival trees for interval-censored data

The bcos data example

```
library(interval)
```

```
## Loading required package: perm
## Loading required package: Icens
```

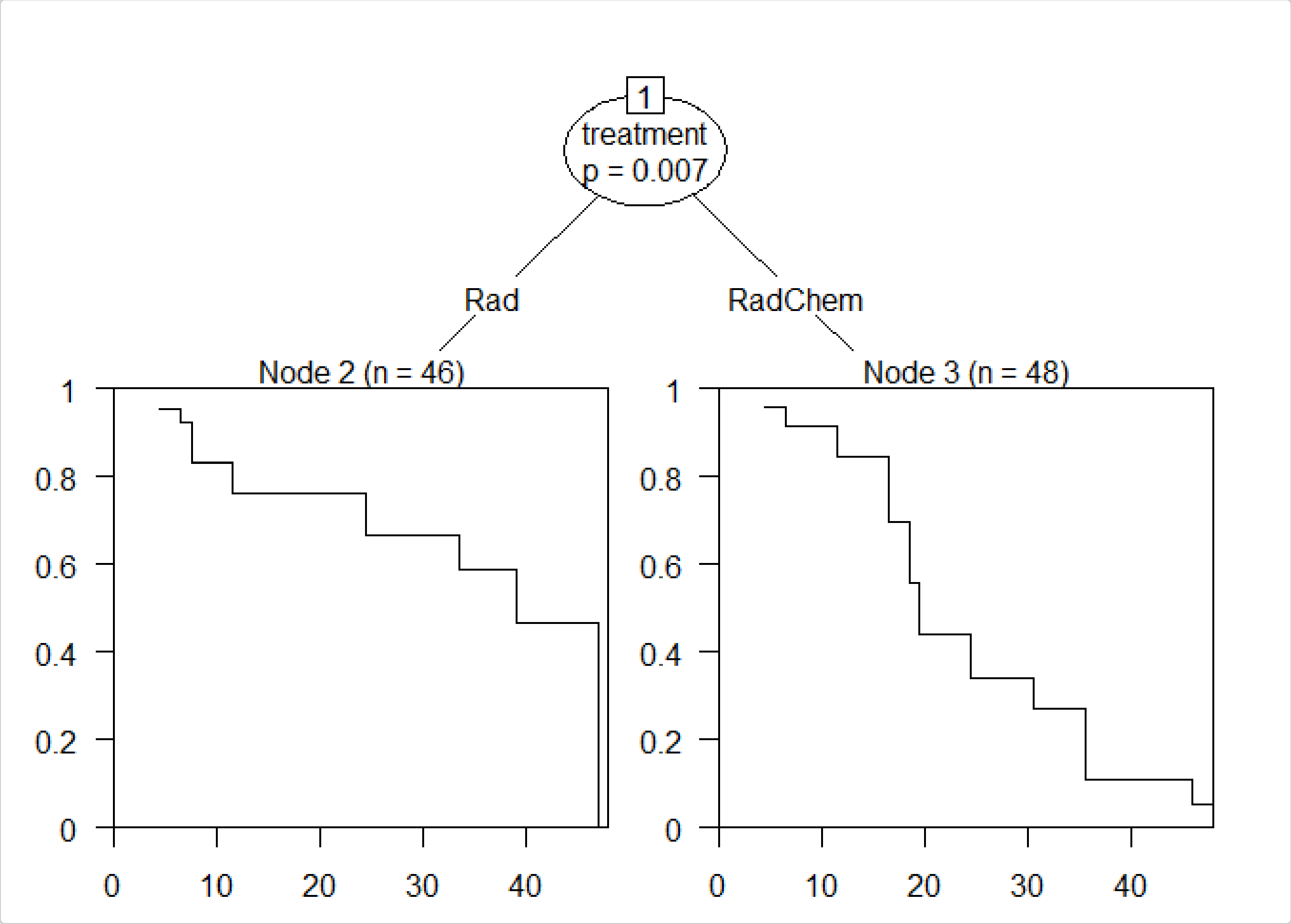
```
## Loading required package: MLEcens
```

```
## Depends on Icens package available on bioconductor.
## To install use for example:
## install.packages('BiocManager')
## BiocManager::install('Icens')
```

```
data(bcos)

## Fit ICtree survival tree
Ctree <- ICtree(Surv(left,right,type="interval2")~treatment, bcos)
```

```
## ans < 0 || ans > 1. t_diff = 12.000000, pStep = 0.055206, intLen = 12.000000, ind = 19, k = 20
```

```
## Plot the fitted tree
plot(Ctree)
```



References

[1] Fu, W. and Simonoff, J.S. (2017a). Survival trees for left-truncated and right-censored data, with application to time-varying covariate data. Biostatistics 18 (2), 352-369

[2] Fu, W. and Simonoff, J.S. (2017b). Survival trees for Interval Censored Survival data. Statistics in medicine 36 (30), 4831-4842

[3] Therneau, T., Atkinson, B., Ripley, B. and Ripley, M.B., (2015). rpart: Recursive Partitioning and Regression Trees.

[4] Hothorn, T. and Zeileis, A., (2015). partykit: A Modular Toolkit for Recursive Partytioning in R.

[5] Therneau, T., (2015). survival: A Package for Survival Analysis in S. version 2.38.

[6] Milborrow, S., (2011). rpart.plot: Plot ‘rpart’ models: An Enhanced Version of ‘plot.rpart’.