

Extras

SURVIVAL ANALYSIS

GBSG2 Data set

The GBSG2 (German Breast Cancer Study Group) data set:

- horTh: hormonal therapy, a factor at two levels (yes and no).
- age: age of the patients in years.
- menostat: menopausal status, a factor at two levels pre (premenopausal) and post (postmenopausal).
- tsize: tumor size (in mm).
- tgrade: tumor grade, a ordered factor at levels I < II < III.
- pnodes: number of positive nodes.
- progrec: progesterone receptor (in fmol).
- estrec: estrogen receptor (in fmol).
- time: recurrence free survival time (in days).
- cens: censoring indicator (0- censored, 1- event).

LASSO

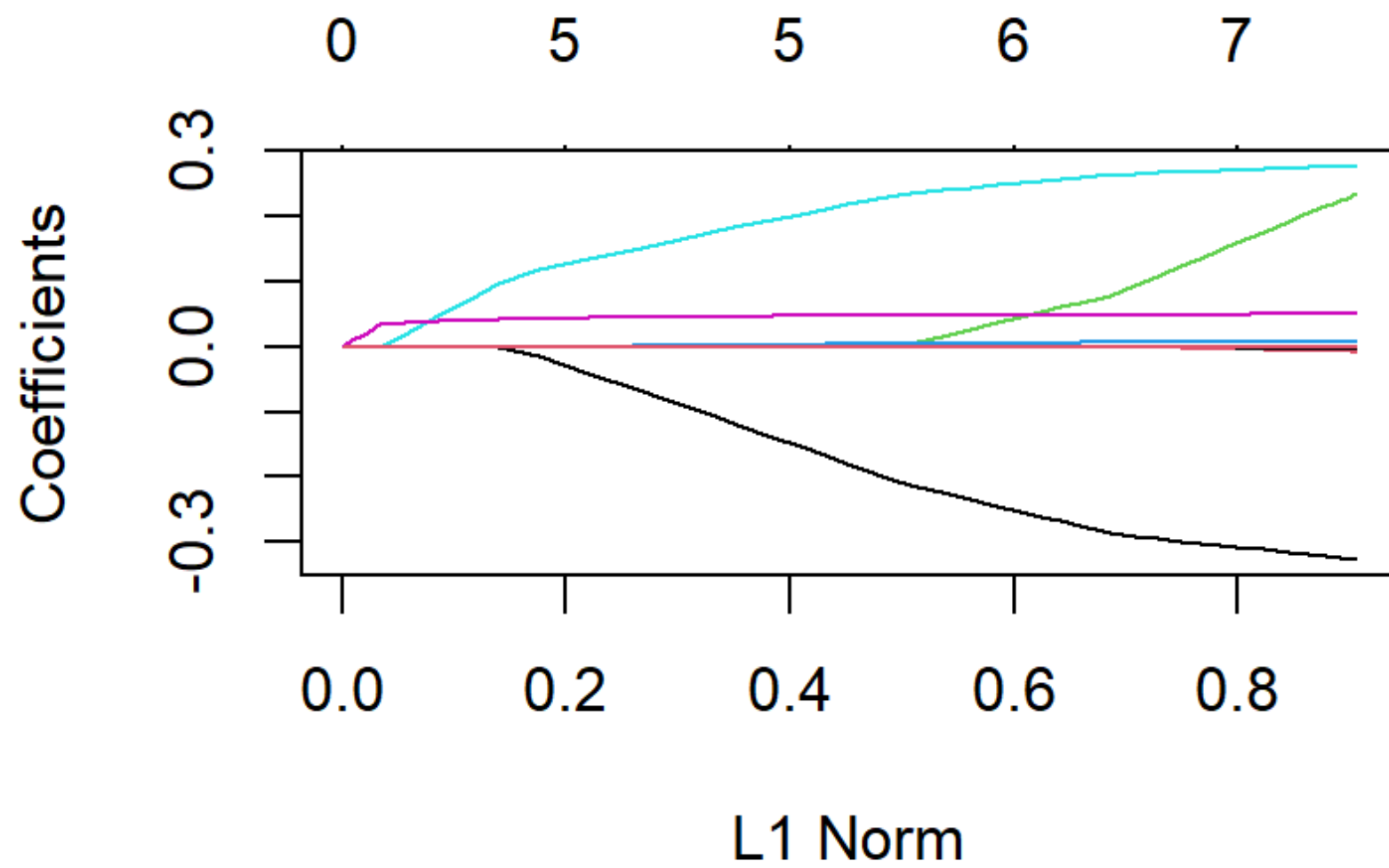
Lasso

The glmnet package in R can be used to perform LASSO (and Ridge) regression for Survival data

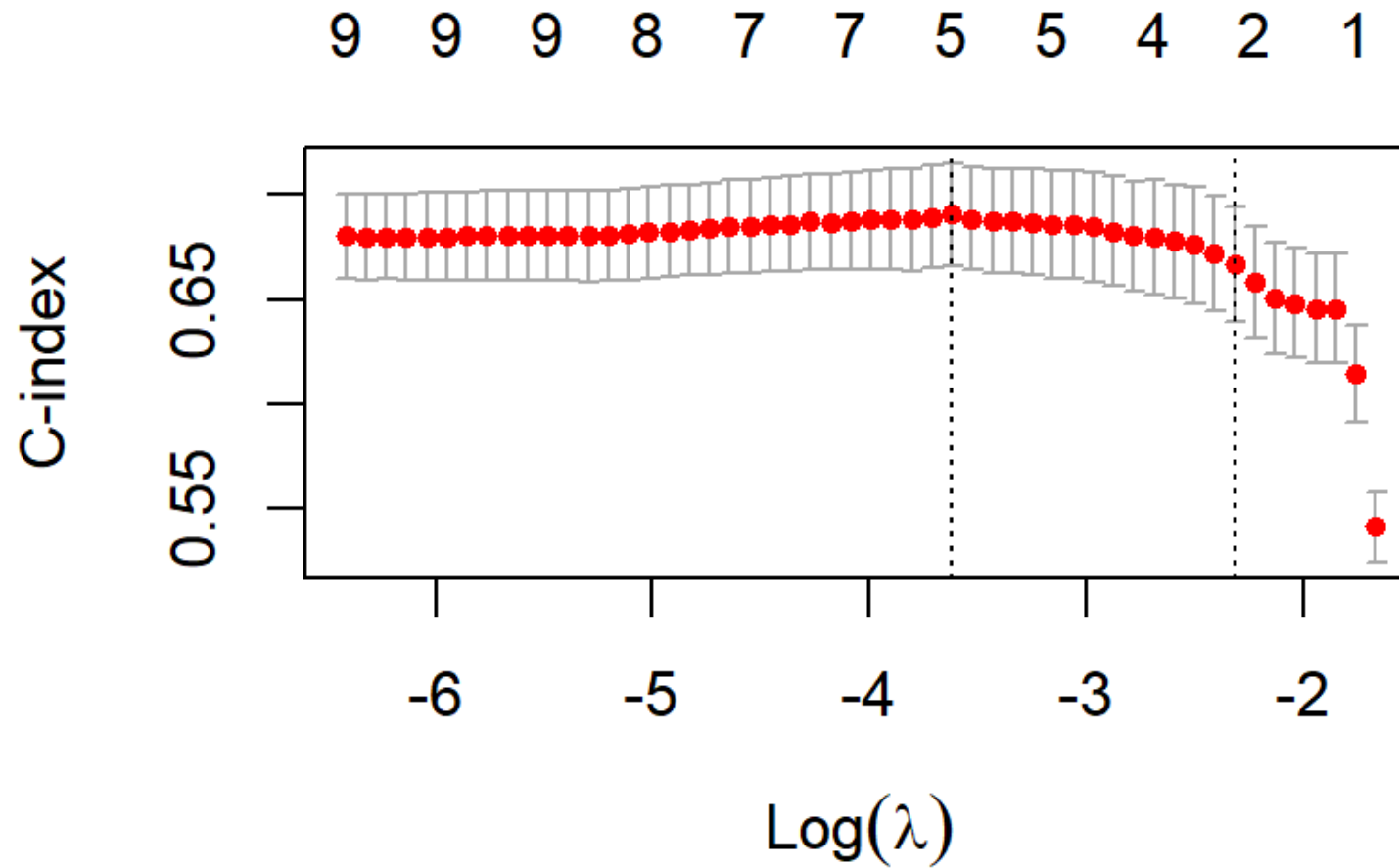
- Underlying model is Cox PH

All the functionalities that you have done previously, you are able to implement here

```
library(glmnet)
x=GBSG2 %>% select(horTh,age,menostat,tsize,
tgrade, pnodes, progrec, estrec)
y=Surv(time=GBSG2$time,event=GBSG2$cens)
model1<-glmnet(x,y,family="cox")
plot(model1)
```



```
x1<-as.matrix(x1) # had to dummy code!!  
set.seed(1287)  
cvfit <- cv.glmnet(x, y, family = "cox", type.measure = "C")  
plot(cvfit)
```



Using Lambda information....

From previous slide, we have that $\log(0.02673542) = -3.621766$

Now we can pull of coef for this lambda value...

```
coef(model1,s=0.03)
```

```
horTh  -0.176785737
age     .
menostat .
tsize   0.004002226
tgrade  0.216889004
pnodes  0.047722334
progrec -0.001325963
estrec  .
```

Can also do repeated events and stratified analysis, see <https://glmnet.stanford.edu/articles/Coxnet.html>.

Decision Trees

Creating the conditional tree

Splitting criteria is based on adjusted p-values from Logrank test (binary splits)

```
tree.surv<- ctree(Surv(time,cens) ~ . ,data=GBSG2)
```

```
tree.surv
```

Model formula:

```
Surv(time, cens) ~ horTh + age + menostat + tsize + tgrade +  
  pnodes + progrec + estrec
```

Fitted party:

[1] root

| [2] pnodes <= 3

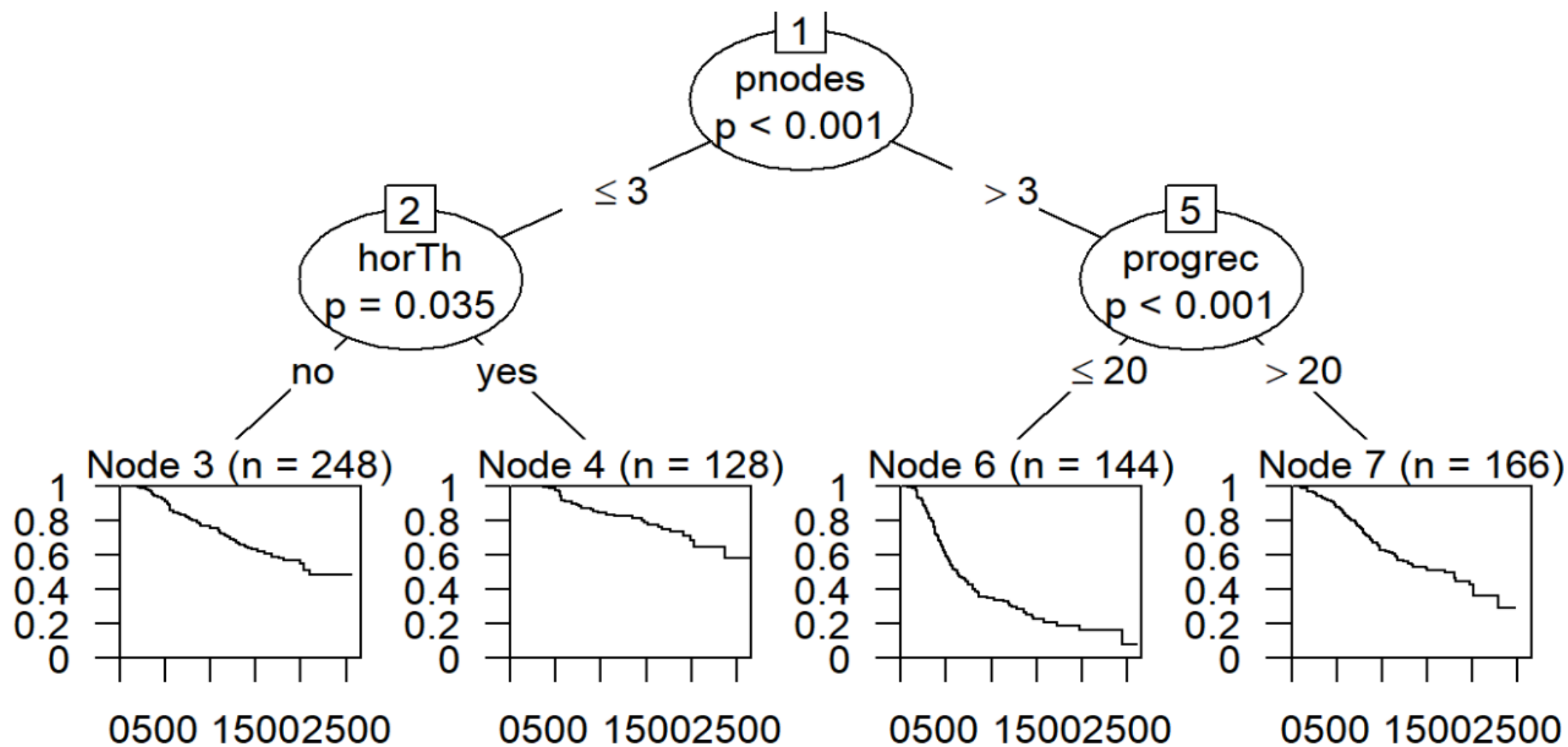
| | [3] horTh in no: 2093.000 (n = 248)

| | [4] horTh in yes: Inf (n = 128)

| [5] pnodes > 3

| | [6] progrec <= 20: 624.000 (n = 144)

| | [7] progrec > 20: 1701.000 (n = 166)



```
predict(tree.surv, newdata = GBSG2[1:2,], type = "node")
```

```
1  2
```

```
3  7
```

```
predict(tree.surv, newdata = GBSG2[1:2,], type = "response")
```

```
1  2
```

```
2093 1701
```

Predicting median survival
time

Nice reference: <https://cran.r-project.org/web/packages/partykit/vignettes/ctree.pdf>

Random Forest

Packages

```
library(randomForestSRC)
```

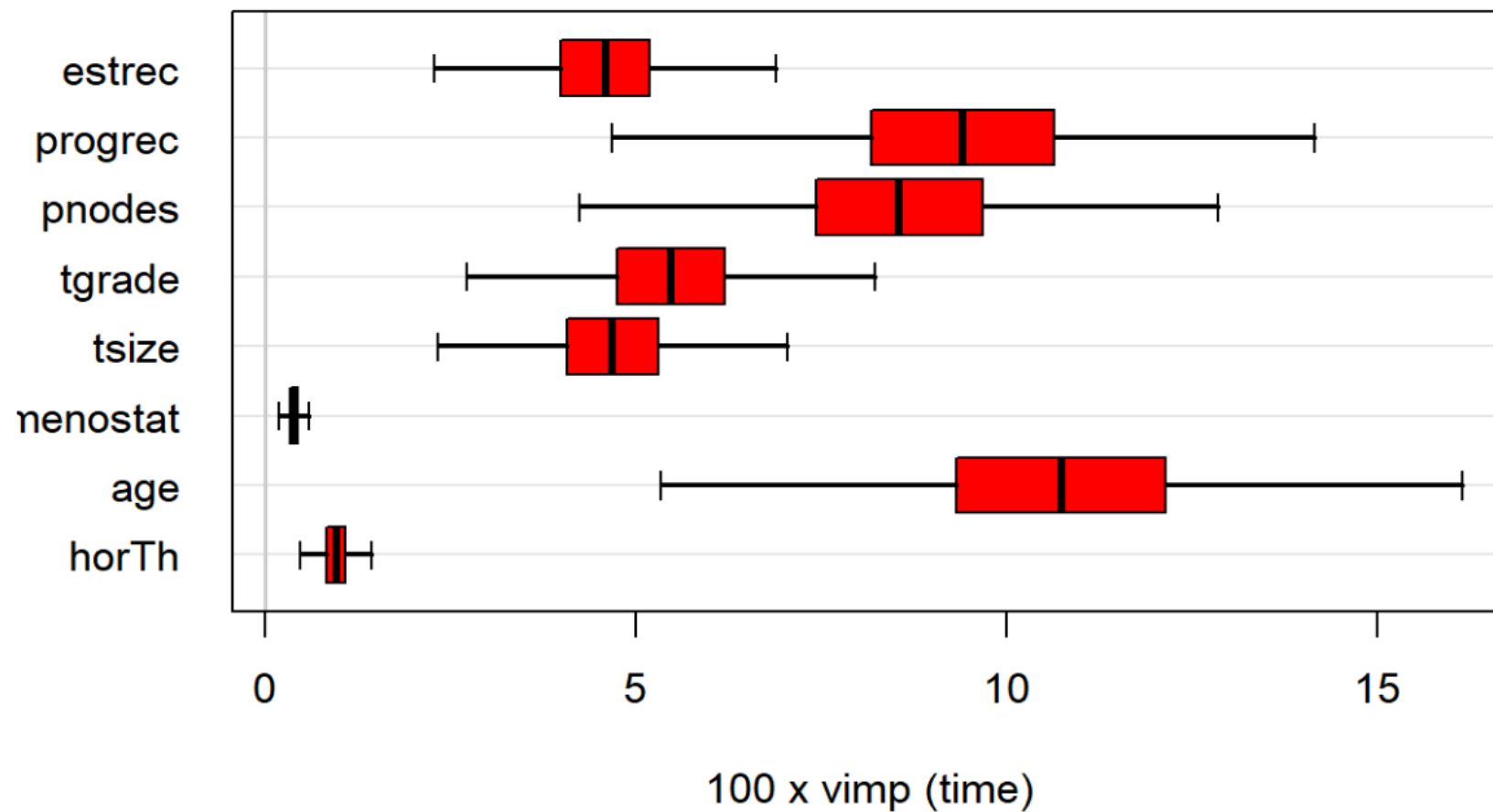
```
library(ggRandomForests)
```

Creating the forest of trees

```
surv.rf <- rfsrc(Surv(time,cens) ~ . ,data=GBSG2,importance = TRUE,splitrule="logrankscore")  
print(surv.rf$importance)
```

| | | |
|-------------|-------------|-------------|
| horTh | age | menostat |
| 0.009596967 | 0.107318858 | 0.003980492 |
| tsize | tgrade | pnodes |
| 0.046818500 | 0.054757864 | 0.085496537 |
| progrec | estrec | |
| 0.094066472 | 0.045879145 | |

Variable importance



Survival curves for first two people in data

Good reference:
<https://www.randomforestsrg.org/articles/survival.html>

