Epidemiologic data analysis using R

Practicals 7

Faculty of Social Sciences, University of Tampere

–

Janne Pitkaniemi

09.04.2018

# Topics of practical 7

Learning objectives of this practical

* Estimating CIF
* use proportional hazards model when competing risk

# Survival analysis: Oral cancer patients

## Description of the data

File , that you may access from a url address to be given in the practical, contains data from 338 patients having an oral squamous cell carcinoma diagnosed and treated in one tertiary level oncological clinic in Finland since 1985, followed-up for mortality until 31 December 2008. The dataset contains the following variables:

## Loading the packages and the data

Load the R packages , and needed in this exercise.

library(Epi)

Warning: package 'Epi' was built under R version 3.4.4

library(mstate)

Loading required package: survival

library(survival)

Read the datafile {oralca2.txt} from a website, whose precise address will be given in the practical, into an R data frame named .

Look at the head, structure and the summary of the data frame. Using function count the numbers of censorings as well as deaths from oral cancer and other causes, respectively, from the variable.

orca <-   
 read.csv("C:/Users/janne.pitkaniemi/Projects/TRE2018/oralca2.txt",   
 sep="")  
head(orca)

sex age stage time event  
1 Male 65.42274 unkn 5.081 0  
2 Female 83.08783 III 0.419 1  
3 Male 52.59008 II 7.915 2  
4 Male 77.08630 I 2.480 2  
5 Male 80.33622 IV 2.500 1  
6 Female 82.58132 IV 0.167 2

## Total mortality: Kaplan–Meier analyses

We start our analysis of total mortality pooling the two causes of death into a single outcome. First, construct a {} from the event variable and the follow-up time using function . Look at the structure and summary of .

# all deaths  
orca$suob <- Surv(orca$time, 1\*(orca$event > 0) )  
str(orca$suob)

Surv [1:338, 1:2] 5.081+ 0.419 7.915 2.480 2.500 0.167 5.925+ 1.503 13.333 7.666+ ...  
 - attr(\*, "dimnames")=List of 2  
 ..$ : NULL  
 ..$ : chr [1:2] "time" "status"  
 - attr(\*, "type")= chr "right"

summary(orca$suob)

time status   
 Min. : 0.085 Min. :0.0000   
 1st Qu.: 1.333 1st Qu.:0.0000   
 Median : 3.869 Median :1.0000   
 Mean : 5.662 Mean :0.6775   
 3rd Qu.: 8.417 3rd Qu.:1.0000   
 Max. :23.258 Max. :1.0000

Create a {survfit} object {s.all}, which does the default calculations for a Kaplan–Meier analysis of the overall (marginal) survival curve.

s.all <- survfit(suob ~ 1, data=orca)

See the structure of this object and apply {print()} method on it, too. Look at the results; what do you find?

s.all

Call: survfit(formula = suob ~ 1, data = orca)  
  
 n events median 0.95LCL 0.95UCL   
 338.00 229.00 5.42 4.33 6.92

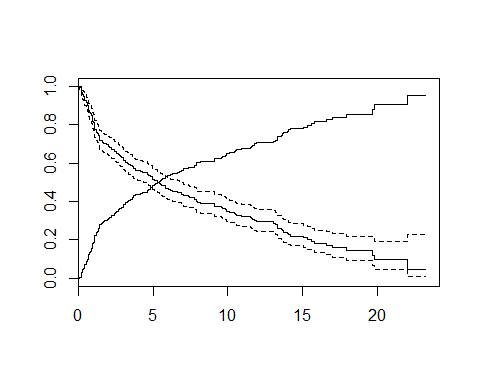
str(s.all)

List of 13  
 $ n : int 338  
 $ time : num [1:251] 0.085 0.162 0.167 0.17 0.246 0.249 0.252 0.329 0.334 0.413 ...  
 $ n.risk : num [1:251] 338 336 334 330 328 327 326 323 322 321 ...  
 $ n.event : num [1:251] 2 2 4 2 1 1 3 1 1 1 ...  
 $ n.censor : num [1:251] 0 0 0 0 0 0 0 0 0 0 ...  
 $ surv : num [1:251] 0.994 0.988 0.976 0.97 0.967 ...  
 $ type : chr "right"  
 $ std.err : num [1:251] 0.0042 0.00595 0.00847 0.0095 0.00998 ...  
 $ upper : num [1:251] 1 1 0.993 0.989 0.987 ...  
 $ lower : num [1:251] 0.986 0.977 0.96 0.953 0.949 ...  
 $ conf.type: chr "log"  
 $ conf.int : num 0.95  
 $ call : language survfit(formula = suob ~ 1, data = orca)  
 - attr(\*, "class")= chr "survfit"

The method for a object would return a lengthy life table. However, the method with default arguments offers the Kaplan–Meier curve for a conventional illustration of the survival experience in the whole patient group.

Alternatively, instead of graphing survival proportions, one can draw a curve describing their complements: the cumulative mortality proportions. This curve is drawn together with the survival curve as the result of the second command line below.

plot(s.all)  
lines(s.all, fun = "event", mark.time=F, conf.int=F)



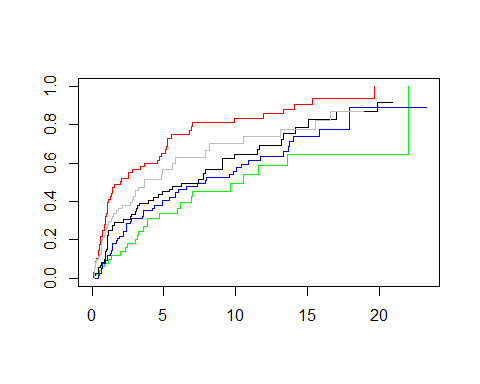
The effect of option {mark.time=F} is to omit marking the times when censorings occurred.

## Total mortality by stage

Tumour stage is an important prognostic factor in cancer survival studies.

Plot separate cumulative mortality curves for the different stage groups marking them with different colours, the order which you may define yourself. Also find the median survival time for each stage.

s.stg <- survfit(suob ~ stage, data= orca)  
col5 <- c("green", "blue", "black", "red", "gray")  
plot(s.stg, col= col5, fun="event", mark.time=F )

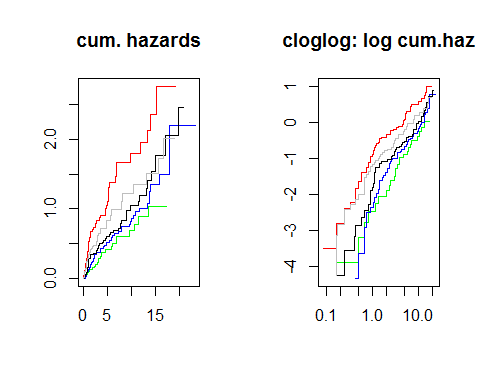


s.stg

Call: survfit(formula = suob ~ stage, data = orca)  
  
 n events median 0.95LCL 0.95UCL  
stage=I 50 25 10.56 6.17 NA  
stage=II 77 51 7.92 4.92 13.34  
stage=III 72 51 7.41 3.92 9.90  
stage=IV 68 57 2.00 1.08 4.82  
stage=unkn 71 45 3.67 2.83 8.17

Create now two parallel plots of which the first one describes the cumulative hazards and the second one graphs the log-cumulative hazards against log-time for the different stages. Compare the two presentations with each other and with the one in the previous item.

par(mfrow=c(1,2))  
plot(s.stg, col= col5, fun="cumhaz", main="cum. hazards" )  
plot(s.stg, col= col5, fun="cloglog", main = "cloglog: log cum.haz" )



If the survival times were {} distributed in a given (sub)population the corresponding cloglog-curve should follow an approximately linear pattern. Could this be the case here in the different stages?

Also, if the survival distributions of the different subpopulations would obey the {} model, the vertical distance between the cloglog-curves should be approximately constant over the time axis. Do these curves indicate serious deviation from the proportional hazards assumption?

it has been observed that the crude contrast between males and females in total mortality appears unclear, but the age-adjustment in the Cox model provided a more expected hazard ratio estimate. We shall examine the confounding by age somewhat closer. First categorize the continuous age variable into, say, three categories by function {cut()} using suitable breakpoints, like 55 and 75 years, and cross-tabulate sex and age group:

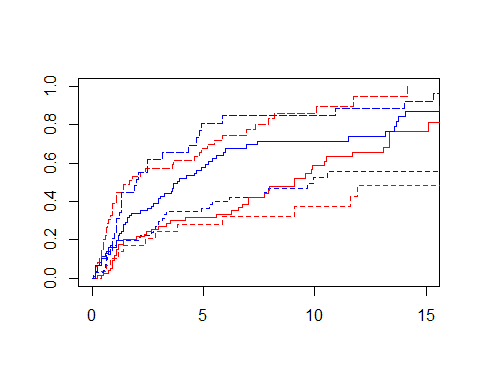
orca$agegr <- cut(orca$age, br=c(0,55,75, 95))  
stat.table( list( sex, agegr), list( count(), percent(agegr) ),margins=T, data = orca )

-----------------------------------------   
 --------------agegr--------------   
 sex (0,55] (55,75] (75,95] Total   
 -----------------------------------------   
 Female 29 74 49 152   
 19.1 48.7 32.2 100.0   
   
 Male 71 86 29 186   
 38.2 46.2 15.6 100.0   
   
   
 Total 100 160 78 338   
 29.6 47.3 23.1 100.0   
 -----------------------------------------

Male patients are clearly younger than females in these data.

Now, plot Kaplan–Meier curves jointly classified by sex and age.

s.agrx <- survfit(suob ~ agegr + sex, data=orca)  
par(mfrow=c(1,1))  
plot(s.agrx, fun="event", mark.time=F, xlim = c(0,15),  
 col=rep(c("red", "blue"),3), lty=c(2,2, 1,1, 5,5))



In each ageband the mortality curve for males is on a higher level than that for females.

## Lexis object with multi-state set-up

Before entering to analyses of cause-specific mortality it might be instructive to apply some Lexis tools to illustrate the competing-risks set-up.

Form a object from the data frame and print a summary of it. We shall name the main (and only) time axis in this object as {stime}.

orca.lex <- Lexis(exit = list(stime = time), exit.status = factor(event,  
 labels = c("Alive", "Oral ca. death", "Other death")),  
 data = orca)

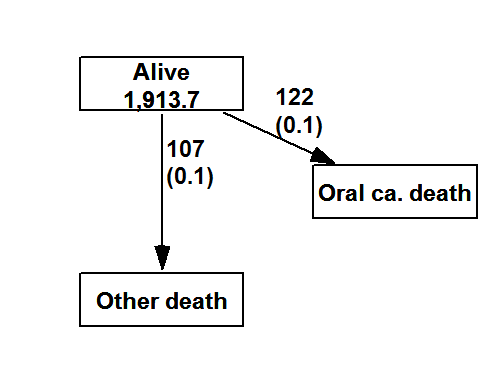
NOTE: entry.status has been set to "Alive" for all.  
NOTE: entry is assumed to be 0 on the stime timescale.

summary(orca.lex)

Transitions:  
 To  
From Alive Oral ca. death Other death Records: Events: Risk time:  
 Alive 109 122 107 338 229 1913.67  
   
Transitions:  
 To  
From Persons:  
 Alive 338

Draw a box diagram of the two-state set-up of competing transitions. Run first the following command line

boxes( orca.lex,boxpos=T )



Now, move the cursor to the point in the graphics window, at which you wish to put the box for Alive'', and click. Next, move the cursor to the point at which you wish to have the box forOral ca. death’‘, and click. Finally, do the same with the box for ``Other death’’. If you are not happy with the outcome, run the command line again and repeat the necessary mouse moves and clicks.

## Event-specific cumulative mortality curves

We move on to analysing cumulative mortalities for the two causes of death separately, first overall and then by prognostic factors.

Use function in package and view the structure of the thus created object.

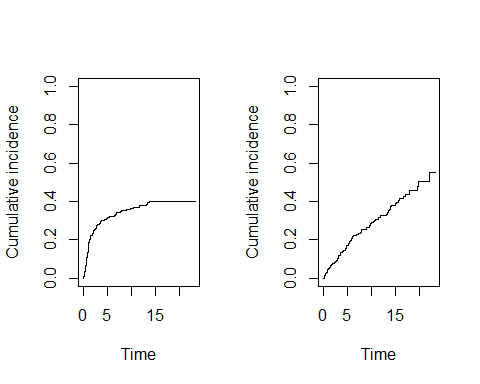
cif1 <- Cuminc( time = "time", status= "event", data = orca)  
str(cif1)

Classes 'Cuminc' and 'data.frame': 160 obs. of 7 variables:  
 $ time : num 0.085 0.162 0.167 0.17 0.246 0.249 0.252 0.329 0.334 0.413 ...  
 $ Surv : num 0.994 0.988 0.976 0.97 0.967 ...  
 $ CI.1 : num 0.00592 0.01183 0.01775 0.02071 0.02367 ...  
 $ CI.2 : num 0 0 0.00592 0.00888 0.00888 ...  
 $ seSurv: num 0.00417 0.00588 0.00827 0.00922 0.00965 ...  
 $ seCI.1: num 0.00417 0.00588 0.00718 0.00775 0.00827 ...  
 $ seCI.2: num 0 0 0.00417 0.0051 0.0051 ...  
 - attr(\*, "survfit")=List of 18  
 ..$ n : int 338  
 ..$ time : num 0.085 0.162 0.167 0.17 0.246 0.249 0.252 0.329 0.334 0.413 ...  
 ..$ n.risk : int [1:251, 1:3] 0 0 0 0 0 0 0 0 0 0 ...  
 ..$ n.event : int [1:251, 1:3] 2 2 2 1 1 0 2 1 1 1 ...  
 ..$ n.censor : int 0 0 0 0 0 0 0 0 0 0 ...  
 ..$ pstate : num [1:251, 1:3] 0.00592 0.01183 0.01775 0.02071 0.02367 ...  
 ..$ p0 : num [1:3(1d)] 0 0 1  
 .. ..- attr(\*, "dimnames")=List of 1  
 .. .. ..$ : chr "1" "2" ""  
 ..$ cumhaz : num [1:3, 1:3, 1:251] 0 0 0.00592 0 0 ...  
 ..$ std.err : num [1:251, 1:3] 0.00417 0.00588 0.00718 0.00775 0.00827 ...  
 ..$ sp0 : num 0 0 0  
 ..$ transitions: 'table' int [1:3, 1:2] 0 0 122 0 0 107  
 .. ..- attr(\*, "dimnames")=List of 2  
 .. .. ..$ from: chr "1" "2" ""  
 .. .. ..$ to : chr "1" "2"  
 ..$ lower : num [1:251, 1:3] 0 0.000238 0.003573 0.00541 0.007327 ...  
 ..$ upper : num [1:251, 1:3] 0.0141 0.0233 0.0317 0.0358 0.0397 ...  
 ..$ conf.type : chr "log"  
 ..$ conf.int : num 0.95  
 ..$ states : chr "1" "2" ""  
 ..$ type : chr "mright"  
 ..$ call : language survfit(formula = Surv(time, statuscr) ~ 1, data = tmp)  
 ..- attr(\*, "class")= chr "survfitms" "survfit"

Function thus creates an ordinary data frame with quite self-explanatory column names. Unfortunately, no handy method is provided in the package, but in Epi package there is funciont plotCIF

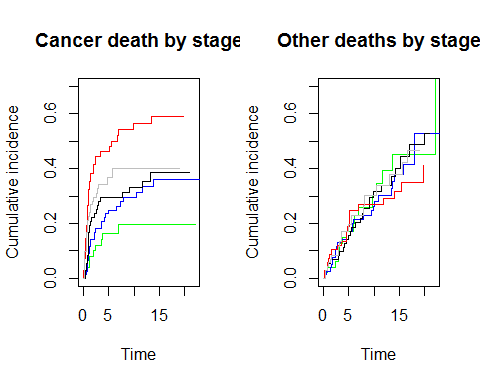
Draw two parallel plots describing the overall cumulative incidence curves for both causes of death

par(mfrow=c(1,2))  
cif1 <- survfit( Surv( time, event, type="mstate") ~ 1,  
 data = orca.lex)  
plotCIF(cif1,event=1)  
plotCIF(cif1,event=2)



Compute the estimated cumulative incidences by stage for both causes of death. Now you have to add argument when calling . See the structure of the resulting object, in which you should observe the first column containing the grouping variable. Plot the pertinent curves in two parallel graphs. Cut the -axis for a more efficient graphical presentation

par(mfrow=c(1,2))  
cif2 <- survfit( Surv( time, event, type="mstate") ~ stage,  
 data = orca.lex)  
 plotCIF(cif2, 1, main = "Cancer death by stage",  
 col=col5, ylim = c(0, 0.7) )  
 plotCIF(cif2, 2, main= "Other deaths by stage",  
 col=col5, ylim = c(0, 0.7) )



Compare the two plots. What would you conclude about the effect of stage on the two causes of death?

par(mfrow=c(1,2))  
cif2 <- survfit( Surv( time, event, type="mstate") ~ stage,  
 data = orca.lex)  
 plotCIF(cif2, 1, main = "Cancer death by stage",  
 col=col5, ylim = c(0, 0.7) )  
 plotCIF(cif2, 2, main= "Other deaths by stage",  
 col=col5, ylim = c(0, 0.7) )

