Epidemiologic data analysis using R

Practicals 7

Faculty of Social Sciences, University of Tampere

–

Janne Pitkaniemi / Karri Seppä

9.3.2020

# Topics of practical 7

Learning objectives of this practical

* Estimating Kaplan-Meier (survfit) and lifetable (survtab)
* Estimating age-standardised survival (survtab)
* Use proportional hazards model (coxph and glm)
* Test for the assumption of proportional hazards

# Survival analysis: Oral cancer patients

## Description of the data

File oralca2.txt 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:

|  |  |
| --- | --- |
| variable | description |
| sex | a factor with categories: 1 = “Female”, 2 = “Male”, |
| age | in years at the date of diagnosing the cancer, |
| stage | TNM stage of the tumour (factor): 1 = “I”, …, 4 = “IV”, 5 = “unkn”, |
| time | follow-up time (in years) since diagnosis until death or censoring, |
| event | status at the end of follow-up (numeric): 0 = censoring alive, 1 = death from oral cancer, 2 = death from other causes. |

## Loading the packages and the data

Load the R packages Epi, mstate, and survival needed in this exercise.

library(Epi)  
library(mstate)

Loading required package: survival

library(survival)

Read the datafile oralca2.txt into an R data frame named orca.

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

orca <- read.csv("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 *survival object* orca$suob from the event variable and the follow-up time using function Surv(). Look at the structure and summary of orca$suob.

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

'Surv' num [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

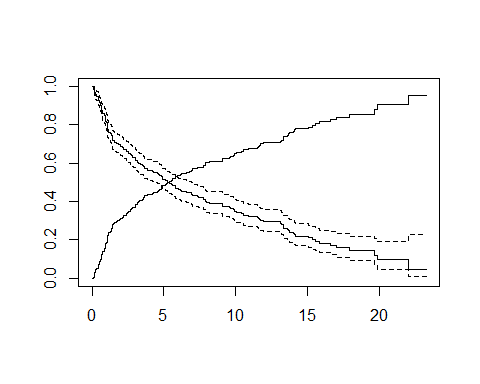
str(s.all)

List of 16  
 $ 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 ...  
 $ std.err : num [1:251] 0.0042 0.00595 0.00847 0.0095 0.00998 ...  
 $ cumhaz : num [1:251] 0.00592 0.01187 0.02385 0.02991 0.03295 ...  
 $ std.chaz : num [1:251] 0.00418 0.00593 0.00843 0.00946 0.00994 ...  
 $ type : chr "right"  
 $ logse : logi TRUE  
 $ conf.int : num 0.95  
 $ conf.type: chr "log"  
 $ lower : num [1:251] 0.986 0.977 0.96 0.953 0.949 ...  
 $ upper : num [1:251] 1 1 0.993 0.989 0.987 ...  
 $ call : language survfit(formula = suob ~ 1, data = orca)  
 - attr(\*, "class")= chr "survfit"

The summary method for a survfit object would return a lengthy life table. However, the plot 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.

Alternatively, use lifetable (actuarial) method in survtab function of popEpi package

library(popEpi)

Attaching package: 'popEpi'

The following object is masked from 'package:survival':  
  
 Surv

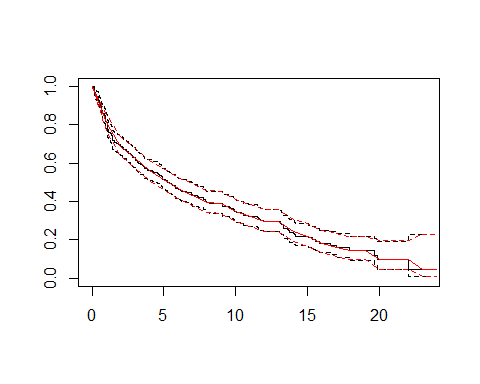
library(Epi)  
  
lex <- Lexis(exit = list(FUT = time),  
 exit.status = (event>0)\*1,  
 data = orca[,-6], merge= T)

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

# lifetable estimated in annual intervals from 0 to 25 years  
st <- survtab(Surv(time = FUT, event = lex.Xst) ~ 1, data = lex,   
 surv.method = "lifetable", surv.type = "surv.obs",   
 conf.type = "log", breaks = list(FUT = seq(0, 25, 1)))

Compare Kaplan-Meier and lifetable estimates

plot(s.all)  
lines(st, col="red")



Lifetable for the first 5 annual intervals:

summary(st, t=1:5)

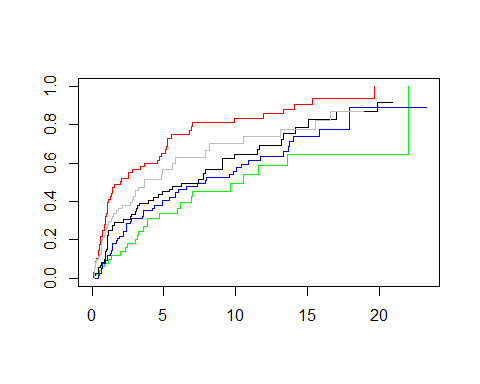
surv.int Tstart Tstop delta d n n.cens surv.obs.lo surv.obs surv.obs.hi  
1: 1 0 1 1 64 338 0 0.7699412 0.8106509 0.8535131  
2: 2 1 2 1 41 274 4 0.6407316 0.6884572 0.7397377  
3: 3 2 3 1 21 229 9 0.5742209 0.6240581 0.6782207  
4: 4 3 4 1 20 199 12 0.5081525 0.5593888 0.6157913  
5: 5 4 5 1 13 167 9 0.4627737 0.5146377 0.5723143  
 SE.surv.obs  
1: 0.02131033  
2: 0.02523540  
3: 0.02650044  
4: 0.02741719  
5: 0.02789207

## 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 )

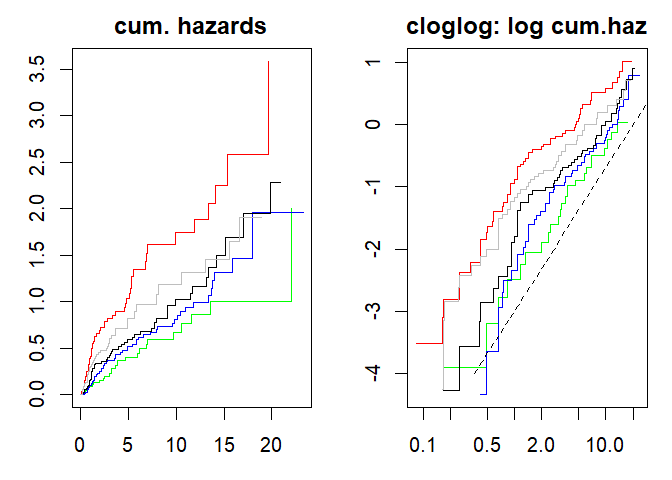


s.stg

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),mar=c(3,3,2,1))  
plot(s.stg, col= col5, fun="cumhaz", main="cum. hazards" )  
plot(s.stg, col= col5, fun="cloglog", main = "cloglog: log cum.haz" )  
#a straight line with a slope of 1  
lines(exp(seq(-4,1,0.1)+3),seq(-4,1,0.1),lty=2)



If the survival times follow the exponential distribution in a given (sub)population, the corresponding cloglog-curve (plotted against time on the logarithmic scale) should be approximately a straight line with a slope of 1 (a straight line with any slope, if survival times follow the Weibull distribution). Could this be the case here in the different stages?

Also, if the survival distributions of the different subpopulations would obey the *proportional hazards* 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?

## Does survival differ between males and females?

coxph(Surv(lex.dur, lex.Xst) ~ sex, data=lex)

Call:  
coxph(formula = Surv(lex.dur, lex.Xst) ~ sex, data = lex)  
  
 coef exp(coef) se(coef) z p  
sexMale 0.1258 1.1340 0.1338 0.94 0.347  
  
Likelihood ratio test=0.89 on 1 df, p=0.3461  
n= 338, number of events= 229

No clear effect in survival between males and females was observed based on this simple analysis.

Next, we shall examine the confounding by age. 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), right=F)  
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.

There is evidence on the effect of sex after adjusting for age at diagnosis.

lex$agegr <- cut(lex$age, br=c(0,55,75,95), right=F)  
m.adj <- coxph(Surv(lex.dur, lex.Xst) ~ sex+agegr, data=lex)  
m.adj

Call:  
coxph(formula = Surv(lex.dur, lex.Xst) ~ sex + agegr, data = lex)  
  
 coef exp(coef) se(coef) z p  
sexMale 0.2994 1.3491 0.1363 2.198 0.027984  
agegr[55,75) 0.6239 1.8661 0.1767 3.531 0.000414  
agegr[75,95) 1.3051 3.6879 0.1940 6.726 1.74e-11  
  
Likelihood ratio test=46.94 on 3 df, p=3.586e-10  
n= 338, number of events= 229

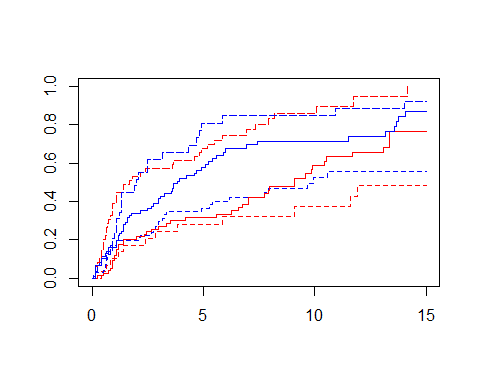
Is proportionality of hazards a reasonable assumption?

cox.zph(m.adj)

chisq df p  
sex 0.106 1 0.744  
agegr 4.767 2 0.092  
GLOBAL 4.776 3 0.189

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.

Now, estimate the age-standardised survival using the distribution of age at diagnosis as the standard:

lex <- Lexis(exit = list(FUT = time),  
 exit.status = (event>0)\*1,  
 data = orca[,-6], merge= T)

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

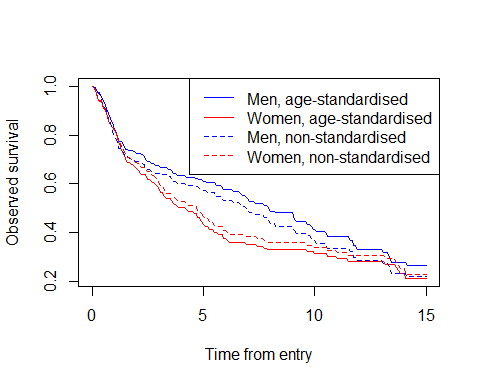
w <- table(orca$agegr)  
  
st <- survtab(Surv(time = FUT, event = lex.Xst) ~ sex + adjust(agegr),   
 data = lex, weights = c(w),   
 surv.method = "lifetable", surv.type = "surv.obs",   
 conf.type = "log", breaks = list(FUT = seq(0, 15, 1/12)))

Some cumulative surv.obs were zero or NA:

sex agegr first.bad.surv.int last.bad.surv.int surv.obs  
1: Female [75,95) 170 170 0

NOTE: Zero surv.obs leads to zero relative survivals as well. Adjusting with weights WILL use the zero surv.obs / relative survival values.

st.non.stand <- survtab(Surv(time = FUT, event = lex.Xst) ~ sex,   
 data = lex,   
 surv.method = "lifetable", surv.type = "surv.obs",   
 conf.type = "log", breaks = list(FUT = seq(0, 15, 1/12)))  
  
plot(st, col=c("blue","red"),conf.int = F)  
lines(st.non.stand, col=c("blue","red"),lty=2,conf.int=F)  
legend("topright",c("Men, age-standardised","Women, age-standardised",  
 "Men, non-standardised","Women, non-standardised"),  
 col=rep(c("blue","red"),2),lty=c(1,1,2,2))



Use piecewise constant hazards model instead of Cox model. Split the follow-up time into annual intervals from 0 to 5 years, and thereafter, into 5-year intervals.

Could hazard be a constant from 0 to 15 years (exponentially distributed survival times)?

lexF <- splitLexis(lex, c(seq(0, 5, 1),seq(10, 15, 5)), time.scale = "FUT")  
#exclude follow-up>15 years  
lexF <- subset(lexF, FUT<15)  
lexF$FUT <- cut(lexF$FUT, c(seq(0, 5, 1),seq(10, 15, 5)),right=F)  
  
# piecewise constant hazards  
m.pc <- glm(lex.Xst ~ FUT + sex + agegr, offset = log(lex.dur),  
 data = lexF, family = "poisson")  
  
# constant hazard model  
m.c <- glm(lex.Xst ~ sex + agegr, offset = log(lex.dur),  
 data = lexF, family = "poisson")  
  
anova(m.c,m.pc,test="LRT")

Analysis of Deviance Table  
  
Model 1: lex.Xst ~ sex + agegr  
Model 2: lex.Xst ~ FUT + sex + agegr  
 Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
1 1416 1156.2   
2 1410 1132.6 6 23.597 0.0006194 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Can we assume proportional hazards between age groups, i.e. no interaction between follow-up time and age at entry?

# main effects only  
m.pc.age <- glm(lex.Xst ~ FUT + sex + agegr, offset = log(lex.dur),  
 data = lexF, family = "poisson")  
  
#interaction included  
m.pc.age.fu <- glm(lex.Xst ~ FUT + sex + agegr + FUT:agegr, offset = log(lex.dur),  
 data = lexF, family = "poisson")  
  
anova(m.pc.age,m.pc.age.fu,test="LRT")

Analysis of Deviance Table  
  
Model 1: lex.Xst ~ FUT + sex + agegr  
Model 2: lex.Xst ~ FUT + sex + agegr + FUT:agegr  
 Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
1 1410 1132.6   
2 1398 1114.4 12 18.154 0.1111

The effect of sex adjusted for age based on the piecewise constant hazards model

summary(m.pc.age)

Call:  
glm(formula = lex.Xst ~ FUT + sex + agegr, family = "poisson",   
 data = lexF, offset = log(lex.dur))  
  
Deviance Residuals:   
 Min 1Q Median 3Q Max   
-1.5067 -0.6083 -0.4510 -0.3503 4.1404   
  
Coefficients:  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) -2.4020 0.2161 -11.113 < 2e-16 \*\*\*  
FUT[1,2) -0.1675 0.2002 -0.837 0.402836   
FUT[2,3) -0.6729 0.2518 -2.672 0.007546 \*\*   
FUT[3,4) -0.5452 0.2565 -2.125 0.033569 \*   
FUT[4,5) -0.8214 0.3046 -2.696 0.007013 \*\*   
FUT[5,10) -0.8049 0.2049 -3.928 8.58e-05 \*\*\*  
FUT[10,15) -0.6531 0.2530 -2.581 0.009841 \*\*   
sexMale 0.2837 0.1389 2.043 0.041101 \*   
agegr[55,75) 0.5986 0.1807 3.314 0.000921 \*\*\*  
agegr[75,95) 1.2888 0.1970 6.543 6.03e-11 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for poisson family taken to be 1)  
  
 Null deviance: 1210.1 on 1419 degrees of freedom  
Residual deviance: 1132.6 on 1410 degrees of freedom  
AIC: 1590.6  
  
Number of Fisher Scoring iterations: 7