**Intro to Survival analysis, practical session with R**

# Introduction to survival analysis workshop

# practical session with R

# CBB, 2024/01/19

# ***Output with notes***

*# Integrated and adapted by Davide Valentini (CBB @KI /CAST @KS)*

*# from STHDA educational guides (*[*http://www.sthda.com/*](http://www.sthda.com/)*)*

# BEFORE THE START: install packages survival and surminer

# install.packages(c("survival", "survminer"))

install.packages("survival")

install.packages("survminer")

library("survival")

library("survminer")

# IMPORT OR LOAD DATA

# One can use the drop-down menu or the R command

# the same dataset can be loaded from a folder, local or from the web. A csv file

# can be found at

# lung <- read.csv("https://raw.githubusercontent.com/davval-ki/IntroSurv/main/lung\_cancer.csv")

# We’ll use the lung cancer data available in the survival package.

data("cancer")

# how is the data structured? Variables, variable types, observations

? help

head(lung)

inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss

1 3 306 2 74 1 1 90 100 1175 NA

2 3 455 2 68 1 0 90 90 1225 15

3 3 1010 1 56 1 0 90 90 NA 15

4 5 210 2 57 1 1 90 60 1150 11

5 1 883 2 60 1 0 100 90 NA 0

6 12 1022 1 74 1 1 50 80 513 0

* inst: Institution code
* time: Survival time in days
* status: censoring status 1=censored, 2=dead
* age: Age in years
* sex: Male=1 Female=2
* ph.ecog: ECOG performance score (0=good 5=dead)
* ph.karno: Karnofsky performance score (bad=0-good=100) rated by physician
* pat.karno: Karnofsky performance score as rated by patient
* meal.cal: Calories consumed at meals
* wt.loss: Weight loss in last six months

1. **Survival function and KM curves**

#First, it is necessary to create a "Survival" object

?Surv

# It can do much more than this (i.e. time dependent variables, etc)

Surv(lung$time, lung$status)

# the output contains a surv object, with times and censoring status. Takes 0/1, T/F or 1/2

# one can also specify the event of interest... Here in this dataset, censored data are == 1

Surv(lung$time, lung$status==1)

# We want to compute the survival curves: survfit()

? survfit

fit1 <- survfit(Surv(time, status), data = lung)

print(fit1)

Call: survfit(formula = Surv(time, status) ~ 1, data = lung)

n events median 0.95LCL 0.95UCL

[1,] 228 165 310 285 363

* n: total number of subjects in each curve.
* time: the time points on the curve.
* n.risk: the number of subjects at risk at time t
* n.event: the number of events that occurred at time t.
* n.censor: the number of censored subjects, who exit the risk set, without an event, at time t.
* lower,upper: lower and upper confidence limits for the curve, respectively.
* strata: indicates stratification of curve estimation. If strata is not NULL, there are multiple curves in the result. The levels of strata (a factor) are the labels for the curves.

For example:

fit1$time

fit1$n.risk

#what we can simply do now is plotting the KM curve

plot(fit1)

A graph of a graph

Description automatically generated with medium confidence

#But let's try a nicer system, using survminer

ggsurvplot(fit1)

A graph showing a curve

Description automatically generated

# now let's compute the survival separately for two groups, for example by sex at birth and and

# plot the corresponding curves,

fit <- survfit(Surv(time, status) ~ sex, data = lung)

print(fit)

summary(fit)

summary(fit)$table

# and draw a even more informative plot

ggsurvplot(fit,

pval = TRUE, conf.int = TRUE,

risk.table = TRUE, # Add risk table

risk.table.col = "strata", # Change risk table color by groups

linetype = "strata", # Change line type by groups

surv.median.line = "hv", # Specify median survival

ggtheme = theme\_bw(), # Change ggplot2 theme

palette = c("#E7B800", "#2E9FDF"))

A graph with numbers and a number at risk

Description automatically generated

# many possibilities, among the different parameters:

# 1) xlim, to shorten your plot

ggsurvplot(fit,

conf.int = TRUE,

risk.table.col = "strata", # Change risk table color by groups

ggtheme = theme\_bw(), # Change ggplot2 theme

palette = c("#E7B800", "#2E9FDF"),

xlim = c(0, 600))

A graph of a graph

Description automatically generated with medium confidence

# 2) fun="event" , to plot the event function (relapses, infections, etc)

ggsurvplot(fit,

conf.int = TRUE,

risk.table.col = "strata", # Change risk table color by groups

ggtheme = theme\_bw(), # Change ggplot2 theme

palette = c("#E7B800", "#2E9FDF"),

fun = "event")

A graph of a graph

Description automatically generated

# Log-Rank test comparing survival curves: survdiff()

surv\_diff <- survdiff(Surv(time, status) ~ sex, data = lung)

surv\_diff

Call:

survdiff(formula = Surv(time, status) ~ sex, data = lung)

N Observed Expected (O-E)^2/E (O-E)^2/V

sex=1 138 112 91.6 4.55 10.3

sex=2 90 53 73.4 5.68 10.3

Chisq= 10.3 on 1 degrees of freedom, p= 0.001

# What about age? we can compare survival by different age groups..

lung$agegrp <- NA

lung$agegrp[lung$age < 55] <- "less than 55"

lung$agegrp[lung$age >= 55 & lung$age < 70] <- "55-69"

lung$agegrp[lung$age >= 70] <- "70 and above"

fitAge <- survfit(Surv(time, status) ~ agegrp, data = lung)

ggsurvplot(fitAge,

conf.int = F,

risk.table.col = "strata",

ggtheme = theme\_bw())

A graph with different colored lines

Description automatically generated

As exercise, turn on the confidence intervals and see what happens...

1. **Cox proportional hazards model**

##### R function to compute the Cox model: coxph()

# 1) Compute the Cox model

# We’ll fit the Cox regression using the following covariates:

# age, sex, ph.ecog and wt.loss.

# We start by computing univariate Cox analyses for all these variables;

# then we’ll fit multivarable cox analyses using two variables

# to describe how the factors jointly impact on survival.

# Univariate Cox regression

res.cox <- coxph(Surv(time, status) ~ sex, data = lung)

res.cox

Call:

coxph(formula = Surv(time, status) ~ sex, data = lung)

coef exp(coef) se(coef) z p

sex -0.5310 0.5880 0.1672 -3.176 0.00149

Likelihood ratio test=10.63 on 1 df, p=0.001111

n= 228, number of events= 165

#again for a more complete report

summary(res.cox)

Call:

coxph(formula = Surv(time, status) ~ sex, data = lung)

n= 228, number of events= 165

coef exp(coef) se(coef) z Pr(>|z|)

sex -0.5310 0.5880 0.1672 -3.176 0.00149 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

exp(coef) exp(-coef) lower .95 upper .95

sex 0.588 1.701 0.4237 0.816

Concordance= 0.579 (se = 0.021 )

Likelihood ratio test= 10.63 on 1 df, p=0.001

Wald test = 10.09 on 1 df, p=0.001

Score (logrank) test = 10.33 on 1 df, p=0.001

The Cox regression results can be interpreted as follow:

1. Statistical significance. The column marked “z” gives the Wald statistic value. It corresponds to the ratio of each regression coefficient to its standard error (z = coef/se(coef)). The wald statistic evaluates, whether the beta coefficient of a given variable is statistically significantly different from 0. From the output above, we can conclude that the variable sex have highly statistically significant coefficients.
2. The regression coefficients. The second feature to note in the Cox model results is the the sign of the regression coefficients (coef). A positive sign means that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable. The variable sex is encoded as a numeric vector. 1: male, 2: female. The R summary for the Cox model gives the hazard ratio (HR) for the second group relative to the first group, that is, female versus male. The beta coefficient for sex = -0.53 indicates that females have lower risk of death (lower survival rates) than males, in these data.
3. Hazard ratios. The exponentiated coefficients (exp(coef) = exp(-0.53) = 0.59), also known as hazard ratios, give the effect size of covariates. For example, being female (sex=2) reduces the hazard by a factor of 0.59, or 41%. Being female is associated with good prognostic.
4. Confidence intervals of the hazard ratios. The summary output also gives upper and lower 95% confidence intervals for the hazard ratio (exp(coef)), lower 95% bound = 0.4237, upper 95% bound = 0.816.
5. Global statistical significance of the model. Finally, the output gives p-values for three alternative tests for overall significance of the model: The likelihood-ratio test, Wald test, and score logrank statistics. These three methods are asymptotically equivalent. For large enough N, they will give similar results. For small N, they may differ somewhat. The Likelihood ratio test has better behavior for small sample sizes, so it is generally preferred.

Doing the same for the variables age, ph.ecog, wt.loss, we’ll have

age 0.019 1 (1-1) 4.1 0.042

sex -0.53 0.59 (0.42-0.82) 10 0.0015

ph.ecog 0.48 1.6 (1.3-2) 18 2.7e-05

wt.loss 0.0013 1 (0.99-1) 0.05 0.83

res.cox <- coxph(Surv(time, status) ~ age + sex + ph.ecog, data = lung) summary(res.cox)

Call:

coxph(formula = Surv(time, status) ~ age + sex + ph.ecog, data = lung)

n= 227, number of events= 164

(1 observation deleted due to missingness)

coef exp(coef) se(coef) z Pr(>|z|)

age 0.011067 1.011128 0.009267 1.194 0.232416

sex -0.552612 0.575445 0.167739 -3.294 0.000986 \*\*\*

ph.ecog 0.463728 1.589991 0.113577 4.083 4.45e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

exp(coef) exp(-coef) lower .95 upper .95

age 1.0111 0.9890 0.9929 1.0297

sex 0.5754 1.7378 0.4142 0.7994

ph.ecog 1.5900 0.6289 1.2727 1.9864

Concordance= 0.637 (se = 0.025 )

Likelihood ratio test= 30.5 on 3 df, p=1e-06

Wald test = 29.93 on 3 df, p=1e-06

Score (logrank) test = 30.5 on 3 df, p=1e-06

The p-value for all three overall tests (likelihood, Wald, and score) are significant, indicating that the model is significant. These tests evaluate the omnibus null hypothesis that all of the betas are 0. In the above example, the test statistics are in close agreement, and the omnibus null hypothesis is soundly rejected.

In the multivariate Cox analysis, the covariates sex and ph.ecog remain significant (p < 0.05). However, the covariate age fails to be significant (p = 0.23).

The p-value for sex is 0.000986, with a hazard ratio HR = exp(coef) = 0.58, indicating a strong relationship between the patients’ sex and decreased risk of death. The hazard ratios of covariates are interpretable as multiplicative effects on the hazard. For example, holding the other covariates constant, being female (sex=2) reduces the hazard by a factor of 0.58, or 42%. We conclude that, being female is associated with good prognostic.

Similarly, the p-value for ph.ecog is 4.45e-05, with a hazard ratio HR = 1.59, indicating a strong relationship between the ph.ecog value and increased risk of death. Holding the other covariates constant, a higher value of ph.ecog is associated with a poor survival.

By contrast, the p-value for age is now p=0.23. The hazard ratio HR = exp(coef) = 1.01, with a 95% confidence interval of 0.99 to 1.03. Because the confidence interval for HR includes 1, these results indicate that age makes a smaller contribution to the difference in the HR after adjusting for the ph.ecog values and patient’s sex, and only trend toward significance. For example, holding the other covariates constant, an additional year of age induce daily hazard of death by a factor of exp(beta) = 1.01, or 1%, which is not a significant contribution.

A graph of a graph

Description automatically generated

# Plot the hazard ratios with c.i.

ggforest(res.cox, data=lung)

A graph of a hazard ratio

Description automatically generated

**c)** **Testing Cox assumptions**

#Testing the proportional hazards assumption.

test.ph <- cox.zph(res.cox)

test.ph

chisq df p

age 0.188 1 0.66

sex 2.305 1 0.13

ph.ecog 2.054 1 0.15

GLOBAL 4.464 3 0.22

# Plot the scaled Schoenfeld residuals against the transformed time

plot(test.ph[1], lwd=1.5, col="red")

A graph of a line graph

Description automatically generated with medium confidence

plot(test.ph[2], lwd=1.5, col="red")

A graph of numbers and lines

Description automatically generated

plot(test.ph[3], lwd=1.5, col="red")

A graph of numbers and points

Description automatically generated with medium confidence

# Checking influential observations

ggcoxdiagnostics(res.cox, type = "dfbeta",

linear.predictions = FALSE, ggtheme = theme\_bw())

A screenshot of a graph

Description automatically generated