Exercises day 9

Basic Statistics for health researchers 2025

March 24, 2025

Warming up

Before starting the exercise below, learn from the R-demo of Lecture 9 (available from the course webpage):

1. Read and run the code.
2. Check that the output matches the results presented on the slides.
3. Do not hesitate to add your own comments into the script.

Notes: the last part of the R-demo is not relevant for the main exercise of today (Exercise A). This last part exemplifies R codes for plotting estimated survival curves from a multiple Cox model and for the analysis of competing risks data. These two topics are not included in Exercise A.

Exercise A

For this exercise we will work with the "colon cancer" data, available from the course webpage. The data come from a three arms randomized clinical trial, which aimed to evaluate adjuvant chemotherapy for colon cancer. For this exercise, we will focus on two of the three treatment groups and we will proceed as if the third group did not exist (to make it simpler). We will assume that the main endpoint was all-cause death and that we have prespecified a multiple Cox regression analysis adjusted on a few baseline covariates (listed at question 11). This is a typical situation with randomized clinical trials, because i) adjusting on a few prognostic variables is expected to increase the power, ii) adjusting is necessary to obtain correct inference when using some popular types of randomizations (e.g. stratified randomization) and iii) randomization provides some reassuring protection against "imperfect" modeling. Before proceeding to the main analysis, we will perform several supplementary/preliminary analyses, which will give us the opportunity to practice with several statistical methods.

Question 1

Read the description of the data and load the data into R. Visualize the first lines of the data and a summary of the data. **Hint:** to make the coding simpler, you can assign the data to a simpler name using the code `d <- colon2`.

Question 2

In this exercise, we will compare the survival chances in the two groups of patients who have:

- received "nothing" (coded "`Obs`")
- received levamisole plus fluorouracil (coded "`Lev(amisole)+5-FU`")

Keep only the observations corresponding to patients of these two groups, i.e., delete the others. **Hint:** you can use this code

```
d <- d[which(d$rx %in% c("Obs","Lev+5FU")),]
d$rx <- droplevels(d$rx)
```

How many observations do we have in each treatment group?

Question 3

It is common to present a "Table 1", to summarize the baseline variable distributions in each treatment group. This helps to better understand the population from which the sample should be representative. It also helps to see the similarities and differences between the two groups of patients that we aim to compare.

1. Before creating the table, first create factor variables for all the categorical variables, using appropriate labels. **Hint:** you can use this code:

```
d$sex <- factor(d$sex,levels=c(1,0),labels=c("male","female"))
d$obstruct <- factor(d$obstruct,levels=c(0,1),labels=c("no","yes"))
d$perfor <- factor(d$perfor,levels=c(0,1),labels=c("no","yes"))
d$adhere <- factor(d$adhere,levels=c(0,1),labels=c("no","yes"))
d$differ <- factor(d$differ,levels=c(1,2,3),labels=c("well","moderate","poor"))
d$extent <- factor(d$extent,levels=c(1,2,3,4),labels=c("submucosa",
                                                       "muscle",
                                                       "serosa",
                                                       "contiguous structures"))
d$surg <- factor(d$surg,levels=c(0,1),labels=c("short","long"))
d$node4 <- factor(d$node4,levels=c(0,1),labels=c("<=4",">4"))
```

2. Create an appropriate "Table 1". **Hint:** you can use the `univariateTable()` function from the `Publish` package and this code:

```
library(Publish)
tab1 <- univariateTable(rx~Q(age) + Q(nodes) + sex +
                            obstruct + perfor + adhere +
                            differ + extent + surg + node4,
                        data=d,
                        compare.groups = FALSE,
                        show.totals = FALSE)
tab1
```

3. Are the variable distributions similar in the two treatment groups? Was it expected?

Question 4

Use the Kaplan-Meier method to estimate the survival curves in each treatment group and plot the curves. Does the treatment seem to work? **Hint:** use the appropriate "time" and "status" variables for the code, here and in what follows. Their names in this dataset are `timeD` and `statusD`. Do not use `time` and `status` as in the R-demo !

Question 5

Compute the median survival times (with 95%-CI) for each treatment group. Could you directly read the results from the plot (approximately)?

Question 6

1. What are the estimated survival probability at $t = 7$ years (and 95%-CI) in each group? Check that the results match the plot produced at question 4.

2. Compute the estimated survival difference between the two groups $t = 7$ years, a 95%-CI and a p-value to conclude whether the difference is statistically different from zero. **Hint:** you can copy-paste the relevant code example from the R-demo and:

   - replace `KM2` by the appropriate object name (you have probability defined this name in the code for question 4).
   - replace `trt` by `rx`, i.e., use the correct name of the variable that describes the treatment groups.
   - replace `time=2*365` by what is appropriate, i.e., `time=7*365`
   - repace `` `trt=0` `` by by `` `rx=Obs` `` and `` `trt=1` `` by `` `rx=Lev+5FU` ``, to have the correct variable name and labels for each group.

Question 7

Before we move on to the main analysis (at question 11), let's first compare the survival curves of the two treatment groups via a simple log-rank test (just to practice and explore the data further).

1. Compute the p-value of the test.

2. Restate the null hypothesis of the log-rank test, interpret the result and conclude.

3. Were the results surprising, given that you have previously seen the two Kaplan-Meier curves at question 4?

Question 8

To accompany the p-value of the log-rank test, it is considered good practice to report an estimated "effect size" and 95%-CI. That is, the hazard ratio obtained from a univariate Cox model and its 95%-CI.

1. Compute those and interpret the results.

2. Moertel et al. (1995), who analyzed an almost identical version of these data, reported that *"patients receiving fluorouracil plus levamisole were found to have a significant survival advantage when compared with patients assigned to observation only; they had a 33% reduction in mortality rate (95%-CI, 16% to 47%)"*. Do your results approximately match their findings?

Question 9

Have a look at Figure 1 on page 7 below. The R code to produce the plot will be provided in the solution. This plot is a simple plot to visually compare the survival curves for each treatment group, when estimated via Kaplan-Meier or via a univariate Cox model (as in the lecture). Does the proportional hazard assumption of the univariate Cox model seem fine? Conclude whether the hazard ratio that we have previously estimated has a relevant interpretation.

Note: if (and only if) you finish both exercises early, try to write the code that produces this plot after you have finished the exercise. A template for writing the code is provided from the R-demo.

Question 10

Before we move on to the main analysis (at question 11), let's first compare the restricted mean survival times (RMST) at $t = 7$ years for the two treatment groups (just to practice and explore the data further).

1. What are the estimated RMST in each group and 95%-CIs? Interpret the results and write clear conclusion sentences. **Hint:** you can use the following code. Note that we first need to create a 0/1 binary variable for the treatment group to use the `rmst2` function of the package survRM2.

```
d$trt <- as.numeric(d$rx=="Lev+5FU")
library(survRM2)
ResRMST <- rmst2(time=d$timeD/365, # trick to have a time unit in years
```

```
                    status=d$statusD,
                    arm=d$trt,
                    tau=7)            # Beware of the time unit!
ResRMST
```

2. What are estimated difference in RMST between the two groups and the corresponding 95%-CI and p-value? Interpret the result and write a clear conclusion sentence.

Question 11

We now proceed to the main analysis. We fit a multiple Cox regression model to compare the survival chances of a patient who received "levamisole plus fluorouracil" to that of a patient who did not, when both patients are similar with regards to:

- `obstruct`: obstruction of the colon by the tumor,
- node4: number of positive lymph nodes (more or less than 4)
- surg: time from surgery to inclusion into the trial (short vs long)
- age
- sex

We assume that the following was prespecified. We do not model any interaction and age is used in the model as a continuous variable with a (usual/simple) log-linear effect on the hazard rate. Estimate the Cox model, interpret the results and conclude. **Hint:** you can use this code:

```
cox2 <- coxph(Surv(timeD,statusD)~rx+obstruct+node4+surg+age+sex,data=d)
summary(cox2)
```

Question 12 ("For those who need more")

One could wonder whether modeling a log-linear effect of age on the hazard rate was a good idea. It is, after all and to some extent, an arbitrary choice. It can be interesting to compare the previous conclusions to those obtained when modeling the effect of age via age groups, which does not require to assume a log-linear effect. Use four age groups (18-50, 50-60, 70-85), refit the Cox model and conclude. **Hint:** you can use this code

```
d$agec <- cut(d$age,breaks=c(18,50,60,70,85),include.lowest=TRUE)
table(d$agec,useNA="always")
cox2c <- coxph(Surv(timeD,statusD)~rx+obstruct+node4+surg+agec+sex,data=d)
summary(cox2c)
```

Exercise B: "For those who need more"

(Only the code will be provided in the solutions)

Question 1

Plot and visually compare the cumulative incidence function of cancer recurrence between the two treatment groups (accounting for the competing risk of death).

Question 2

Same as questions 4 and 11 of Exercise A, but instead of all-cause death, consider the composite endpoint "cancer recurrence or death". That is, compare the "Progression-Free" Survival (PFS) between the two groups.

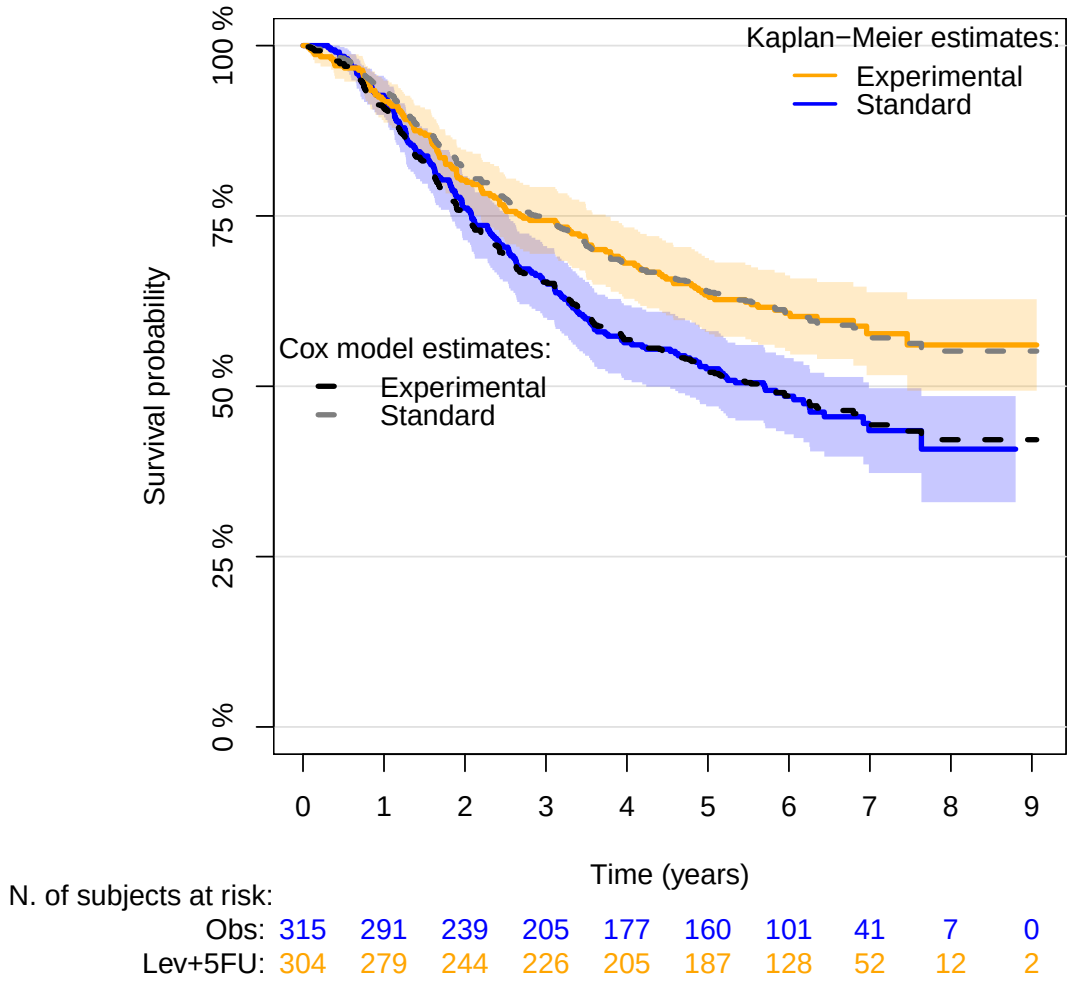

Figure 1: Plot produced following the intructions of item 1 at Question 9, Exercise A.