Logistic regression with right censored (survival) data: Practicals with R

Paul Blanche (September 2025)

We will practice with `R` and the `rotterdam` data of the `survival` package. These data are observational data coming from the Rotterdam tumour bank. For practicing today, we will pretend that these data have been collected to investigate whether chemotherapy can reduce the 5-year risk of recurrence or death among women treated for breast cancer. Here the time-to-event outcome is recurrence-free survival time, defined as the time from primary surgery to either disease recurrence or death, whichever occurs first. The main analysis will aim to estimate the 5-year "causal" risk difference, that is, the risk difference that we expect if we randomize similar patients to chemotherapy or no chemotherapy.

We will further assume that:

- we have collected enough data about potential confounders to believe that the unmeasured confounding assumption is reasonable.
- the process to collect and register the data makes the independent censoring assumption within each treatment group plausible.
- following **thorough** discussions with oncologists ($\gg$ 10 mins !), we do not believe that interaction terms are needed in the logistic regression model.

Disclaimer: I know very little about these data and I have no idea whether these assumptions make sense, unfortunately.

Before proceeding to the main analysis, we will perform several supplementary/preliminary analyses, for completeness and to practice more with survival data.

Preliminaries

We first load the data and have a look a the first lines.

```
library(survival)
d <- rotterdam # for convenience
head(d)        # print first lines
```

We then create a new `status` variable and change the time scale from days to year (for convenience).

```
d$time <- d$rtime/ 365.25
d$status <- d$recur
```

We get basic descriptive statistics for all variables.

```
summary(d)
```

We then create clinically relevant groups by categorizing some quantitative variables. This will be useful to fit a logistic model that does not rely on strong and questionable linearity assumptions, while keeping the modeling strategy simple. **Remark:** alternative approaches can be more suitable. For instance, using linear splines is often interesting as a good compromise between using a flexible and realistic model and keeping it simple enough for the interpretation and pre-specification (also an interesting bias-variance tradeoff).

```
d$yearcat <- cut(d$year,include.lowest=TRUE,
                 breaks=c(1978,1985,1988,1990,1993))
d$agecat <- cut(d$age,include.lowest=TRUE,
                breaks=c(24,35,40,45,50,55,60,65,90))
d$nodescat <- cut(d$nodes,include.lowest=TRUE,
                  breaks=c(0,1,3,5,10,Inf))
d$pgrcat <- cut(d$pgr,include.lowest=TRUE,
                breaks=c(0,20,40,70,100,150,Inf))
d$ercat <- cut(d$er,include.lowest=TRUE,
               breaks=c(0,7,15,40,60,80,100,140,200,Inf))
```

We print simple descriptive statistics for all the created variables.

```
summary(d[,grep("cat",names(d))],maxsum=9)
```

Transform some variables to factor variables.

```
d$chemo <- factor(d$chemo) # chemotherapy (treatment)
d$grade <- factor(d$grade)
```

Question 1

Produce a Kaplan-Meier plot showing the estimated progression-free survival functions in each treatment group: with and without chemotherapy. We do that with the `prodlim` package (although the `survival` package could have done the job too). We focus on the results at t=5 years.

```
library(prodlim)
fitKM <- prodlim(Hist(time, status) ~ chemo, data = d)
summary(fitKM,time=5)            # print estimated survival at t=5
summary(fitKM,time=5,surv=FALSE) # print estimated risk instead (1-surv)
plot(fitKM,xlim=c(0,6),legend.x="bottomleft")
abline(v=5,lwd=2,col="blue",lty=2)
```

Question 2

Produce a table with descriptive statistic for the following baseline covariates, per treatment group. That is, a usual "Table 1".

- year of inclusion (groups)
- age
- menopausal status
- tumor size
- grade
- number of positive lymph nodes
- progesterone receptors
- estrogen receptors
- hormonal treatment

This can be done using the code below, which computes frequencies and proportions per group for caterogical variables and medians with first and third quartiles for quantitative variables. We will assume that these variables are potential **confounders** that we would like to adjust for. What do you observe? Are the patients similar in the two groups?

Remark about the code: left of symbol `~` is to define the groups (columns in the table), right of symbol `~` to define the rows. Using `Q(age)` instead of simply `age` is to compute median and quartiles instead of mean and sd.

```
library(Publish)
Table1 <- univariateTable(chemo~yearcat + Q(age) + meno +
                              size + factor(grade) + Q(nodes)
                          + Q(pgr) + Q(er) + hormon,
                          data=d,
                          compare.groups = FALSE,
                          show.totals = FALSE)
Table1
```

Question 3

Using the code below, produce a Kaplan-Meier plot showing the estimated censoring cumulative distribution in each treatment group: with and without chemotherapy. We focus on the relevant results within the first 5 years. What do you observe?

Remark about the code: `reverse=TRUE` below tells R that we want to estimate the survival function of the censoring time instead of the time-to-event. Accordingly, it does the same as just changing `status` by `status2` after defining `d$status2=1-d$status`, when

running the code without using the `reverse=TRUE` option (except for the smart handling of ties).

```
fitKMC <- prodlim(Hist(time, status) ~ chemo,
                  data = d,reverse=TRUE)
plot(fitKMC,xlim=c(0,6),
     ylim=c(0,0.25),
     type="cuminc", # to plot 1-surv; ie, cumulative distribution
     ylab="Risk of censoring within s years",
     xlab="time s (years)")
abline(v=5,lwd=2,col="blue",lty=2)
```

Question 4

Just for completeness, look at how many patients are observed:

- with a recurrence or death within 5-years
- lost of follow-up (censored) within 5-years
- alive recurrence free at 5 years

Do you confirm that many patients are lost of follow-up within 5 years?

```
sum(d$time <=5 & d$status==1)
sum(d$time <=5 & d$status==0)
sum(d$time >5)
```

Question 5

Using the code below, fit a logistic regression model for the 5-year risk of recurrence or death, with chemotherapy as covariates as well as all the other variables listed in **Question 2**. Use the categorical version of each quantitative variable, to facilitate the interpretation and, more importantly, to avoid making strong linearity assumptions. To account for right-censoring, we will use the "outcome weighed estimating equations" approach (oipcw) and compute the censoring weights using a Kaplan-Meier estimator stratified on treatment group. What can we conclude about the chemotherapy, from the fitted model?

```
library(mets)
out.oipcw <- binreg(Event(time, status) ~ chemo +
                        yearcat + agecat + meno +
                        size + grade + nodescat +
                        pgrcat + ercat + hormon, # logistic model
                    data=d,
                    time=5,                      # time horizon
                    cens.model=~strata(chemo))   # censoring model
summary(out.oipcw)
```

Additional question (can be skipped)

Fit a logistic model with chemotherapy as the only covariate, using the code below. Are the results consistent with those of **Question 1**? At which slide of the lecture did we mentioned this result? Is it reassuring?

```
out.oipcw.unadj <- binreg(Event(time, status) ~ chemo,
                      data=d,
                      time=5,
                      cens.model=~strata(chemo))
summary(out.oipcw.unadj)      # fitted logistic model
expit <- function(x) exp(x)/(1+exp(x))
expit(coef(out.oipcw.unadj)[1]) # estimated risk for chemo=0
expit(sum(coef(out.oipcw.unadj))) # estimated risk for chemo=1
# reminder of estimated risk in each group (question 1):
summary(fitKM,time=5,surv=FALSE) # print estimated risk instead (1-surv)
```

Question 6

Use the "weighed estimating equations" approach (ipcw-glm) instead as a sensitivity analysis. Is there a substantial difference in the results?

```
out.ipcw.glm <- logitIPCW(Event(time, status) ~ chemo +
                              yearcat + agecat + meno +
                              size + grade + nodescat +
                              pgrcat + ercat + hormon,
                            data=d,
                            time=5,
                            cens.model=~strata(chemo))
# summary(out.ipcw.glm) # summary of model fit
cbind(ipcw.glm=coef(out.ipcw.glm),oipcw=coef(out.oipcw)) # head to head comparison
```

Question 7

Using the code below, use standardization after logistic regression to perform the main analysis and estimate the **marginal** 5-year risk of recurrence or death for a patient randomized to chemotherapy versus that for of a patient randomized to no chemotherapy. We will use the same logistic regression model as above. What is the risk difference? What can we conclude?

```
ateFit <- binregATE(Event(time, status) ~ chemo +
                          yearcat + agecat + meno +
                          size + grade + nodescat +
                          pgrcat + ercat + hormon,
                        data=d,
```

```
                    time=5,
                    treat.model=chemo~1,
                    cens.model=~strata(chemo))
summary(ateFit)
```

Question 8

Just for completeness, we produce the corresponding unadjusted ("crude") results (risk difference with 95-CI and p-value) and we check that they match the plot produced at **Question 1**. We can do that conveniently using the `timeEL` package.

```
library(timeEL)
DiffRisk.unadj <- TwoSampleKaplanMeier(time=d$time,
                                       status=d$status,
                                       group=d$chemo,
                                       t=5,contr = list(method = "Wald"))
print(DiffRisk.unadj, what="Diff")
# reminder of estimated risk in each group:
summary(fitKM,time=5,surv=FALSE) # print estimated risk instead (1-surv)
```

Question 9

To better understand the difference between the results of the adjusted and unadjusted analysis, we look again at the baseline Table (produced at **Question 2**). We see that there is some important imbalance for the number of positive lymph nodes. We therefore do two things. First, we plot the Kaplan-Meier curves for recurrence-free survival per treatment group within the two subgroups of patients: those with ≤ 1 positive lymph node and those with ≥ 2. Second, we compute the proportions of patients with ≤ 1 positive node in the two treatment groups. What do we observe? Can we provide a tentative explanation for the difference between the adjusted and unadjusted results?

```
d$nodes01 <- ifelse(d$nodes<2,"0-1","2+")
tabconf <- round(prop.table(table(nodes=d$nodes01,
                                  chemo=d$chemo),margin=2)*100,1)
fitKMnodes <- prodlim(Hist(time, status) ~ nodes01*chemo, data = d)
par(mfrow=c(1,2))
plot(fitKMnodes,xlim=c(0,7),legend.x="bottomleft")
abline(v=5)
barplot(tabconf[1,],ylab="Pr(n. nodes <=1), in %",xlab="chemo")
```

Appendix: package versions

```
sessionInfo()
```