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

Paul Blanche (June 2024)

We will practice with `R` and the `rotterdam` data of the `survival` package. These data are observational data 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 the earlier of disease recurrence or death. 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 on potential confounders to believe that the un-measured 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, 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.

```r
library(survival)
d <- rotterdam # for convenience
head(d)
```

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 summary 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.

```
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.

```
d$chemo <- factor(d$chemo)
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)
plot(fitKM,xlim=c(0,7),legend.x="bottomleft")
abline(v=5,lwd=2,col="blue")
```

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 via the following code, 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?

```r
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

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?

```r
fitKMC <- prodlim(Hist(time, status) ~ chemo, data = d,reverse=TRUE)
plot(fitKMC,xlim=c(0,7),
     ylim=c(0,0.3),
     type="cuminc",
     ylab="Risk of censoring within s years",
     xlab="time s (years)")
abline(v=5,lwd=2,col="blue")
```

Question 4

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

- with a recurrence within 5-years
- lost of follow-up (censored) within 5-years
- 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

Fit a logistic regression model for the 5-year risk of recurrence, with chemotherapy as covariates as well as all the other variables listed in the previous **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,
                    data=d,
                    time=5,
                    cens.model=~strata(chemo))
summary(out.oipcw)
```

Question 6

Use the "weighed estimating equations" approach (ipcw-glm) instead as 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)
```

Question 7

Use standardization after logistic regression to perform the main analysis and estimate the marginal 5-year risk of recurrence 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, produce the corresponding unadjusted ("crude") results (risk difference with 95-CI and p-value) and check that they match the plot produced at **Question 1**.

```
# First we extract the estimates & SEs
fitKM.res <- summary(fitKM,time=5)
fitKM0 <- as.matrix(fitKM.res[fitKM.res$chemo==0,c("surv","se.surv")])
fitKM1 <- as.matrix(fitKM.res[fitKM.res$chemo==1,c("surv","se.surv")])
# Second, we compute the risk difference
diffRisk <- (1-fitKM1[1,"surv"]) - (1-fitKM0[1,"surv"])
# Third, we compute the SE of the difference
seDiffRisk <- sqrt(fitKM1[1,"se.surv"]^2 + fitKM0[1,"se.surv"]^2)
# Then we compute the 95% CI
lowerDiffRisk <- diffRisk - qnorm(1-0.05/2)*seDiffRisk
upperDiffRisk <- diffRisk + qnorm(1-0.05/2)*seDiffRisk
# And the p-value (two-sided test)
pvalDiffRisk <- 2*(1-pnorm(abs(diffRisk/seDiffRisk)))
# Put all the results together and print
ResDiffRisk <- c(Diff=unname(diffRisk),
                 lower=unname(lowerDiffRisk),
                 upper=unname(upperDiffRisk),
                      p=unname(pvalDiffRisk))
ResDiffRisk
```

Question 9

To better understand the difference between the results of the adjusted and unadjusted analysis, we look again at the baseline Table. 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 <2 positive lymph nodes and those with ≥ 2. Second, we compute the proportions of patients with < 2 positive nodes 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")
```