Frailty Models: Theory & Practice
Computer Practical

Theodor Balan and Hein Putter

11/3/2017

Introduction

The goal of this practical is to introduce you to fitting frailty models in R. We will try out functions from the `survival` and `frailtyEM` packages. We will not focus on covariate selection or cover the medical implications of the data analysis. Make sure that you have an up-to-date R installation.

The data that we will use comes from a placebo controlled trial of gamma interferon in chronic granulotomous disease (CGD). The main study based on these data was published as Gallin, J. I., et al. *A controlled trial of interferon gamma to prevent infection in chronic granulomatous disease*, in The New England Journal of Medicine 324.8 (1991): 509-516. The data set is available in numerous `R` packages, including the `survival` package.

Variables

The structure of the data is as follows:

- `id`: subject id
- `center`: the centers where the trial took place
- `random`: randomization date
- `treat`: treatment arm: `rIFN-g` or `placebo`
- `sex`: `female` or `male`
- `age`: age in years at time of study entry
- `height`: height in cm at time of study entry
- `weight`: weight in kg at time of study entry
- `inherit`: pattern of inheritance: `autosomal` (autosomal recessive) or `X-linked`
- `propylac`: use of prophylactic antibiotics at time of study entry (`0` for no, and `1` for yes)
- `hos.cat`: institution category: `US:other`, `US:NIH`, `Europe: Amsterdam` and `Europe:other`
- `tstart`, `tstop`, `status`: event times from randomization (time 0) to infections
- `enum` the number of the event

Descriptive analysis

First we load the data:

```
library(survival)
data(cgd)
```

This loads two data sets: `cgd` and `cgd0`. We will use `cgd`, which is already in the Andersen-Gill format that we discussed in the course. You can find a brief description of this data set in the `survival` package:

```
?cgd
head(cgd)
```

(Q1) Which event is the outcome of interest here? How many individuals are there in the data set? How many individuals are within each center?

```
unique(cgd$id)
tapply(cgd[cgd$enum == 1,"id"], cgd[cgd$enum == 1, "center"], length)
```

(Q2) How many events are there in the whole data set?

```
cgd$status
```

(Q3) How many events are there for each individual? Make a histogram of this. What percentage of individuals experience no serious infections? What are the maximum number of observed events / individual?

```
tapply(cgd$status, cgd$id, sum)
```

With recurrent events, often the cumulative intensity is more interesting than the survival. From counting process theory, a counting process $N(t)$ with intensity $\lambda(t)$ may be decomposed into:

$$N(t) = \Lambda(t) + M(t).$$

This classical result is known as the Doobs-Meyer decomposition and it is fundamental in applying martingale theory to survival analysis. The interpretation here is that $N(t)$ is the number of events of an individual up to time t, $M(t)$ is the martingale residual at time t and $\Lambda(t)$ is the cumulative intensity (hazard), which for a proportional intensity (hazard) would be

$$\Lambda(t) = \int_0^t \lambda_i(s)ds = \int_0^t \exp(\beta^\top x_i)\lambda_0(s)ds.$$

We will not dwelve on this here, but $M(t)$ has expectation 0, which implies that $\Lambda(t)$ is equal to the expected number of events of an individual up to time t. When the outcome is a single event (such as death), then $N(t) \leq 1$ and the intensity $\lambda(t)$ is the hazard function.

The survival, in the recurrent event case, would be

$$S(t) = \exp(-\Lambda(t))$$

which is not a quantity that can be easily interpreted.

Time to first event for each individual

First, we will look at the time to the first event for each individual.

```
cgd1 <- cgd[cgd$enum==1,]
plot(survfit(Surv(tstop, status) ~ 1, cgd1))
```

(Q4) What is the interpretation of this curve?

(Q5) Which variables influence the time to the first event, and how? Tip: you can call `summary()` on an object for an easier to read output.

```
coxph(Surv(tstop, status) ~ frailty(id), cgd1)
```

```
coxph(Surv(tstop, status) ~ sex + treat + age, cgd1)
coxph(Surv(tstop, status) ~ sex + treat + age + frailty(id), cgd1)
```

```
mod_univ <- coxph(Surv(tstop, status) ~ sex + treat + age + inherit + steroids +
                    propylac + frailty(id), cgd1)
```

(Q6) Do you think that the `frailty()` statement here is useful? Try to fit the same models without the `frailty()` statement. Does anything change?

(Q7) From the output of `mod_univ`, what do you think that the empirical Bayes estimates of the fraily will look like? Try to find the estimates. You can see the fields corresponding to an object in `R` by calling `str()` on that object. On which scale do you think that the frailty estimates are on?

```
str(mod_univ)
```

(Q8) Do you think that non-proportional covariate effect might be a big problem for this model? Call `cox.zph()` on it and check the results.

Modeling all the observations

(Q8) Now back to analyzing the whole data set. In what time scale are the event times measured in this data set? Do you think that the calendar time or the gap time are more relevant in this case?

We will start with the marginal models with working independence. Below is the fit using the calendar time scale.

(Q9) What is the left hand side in the argument of `coxph`? Why was it all right to use only `(tstop, status)` before?

```
mod_cal_wi <- coxph(Surv(tstart, tstop, status) ~ sex + treat + age + inherit +
                      steroids + propylac + cluster(id),
                    ties = "breslow", cgd)
```

(Q10) We used `+ cluster(id)` to let `coxph` know which observations to take together. What happens if we would not add that statement? Fit the same model without that statment. Do you notice any differences?

(Q11) Suppose now that we want to go in the other direction - and try a fixed effects model. We can do that by adding instead of `+ cluster(id)`, `+ as.factor(id)`. Fit this model and examine the output. Do you see anything problematic? How could you explain that?

```
mod_cal_fe <- coxph(Surv(tstart, tstop, status) ~ sex + treat + age + inherit +
                      steroids + propylac + as.factor(id),
                    ties = "breslow", cgd)
```

(Q12) We can also try to stratify, so that we let each individual have their own baseline hazard. In this case, we would use `+ strata(id)` instead of `+ cluster(id)`. What happens? Any idea why?

```
try(mod_cal_strat <- coxph(Surv(tstart, tstop, status) ~ sex + treat + age + inherit +
                      steroids + propylac + strata(id),
                    ties = "breslow", cgd))
```

(Q13) What do you think is the key for the fixed effects and the stratified models to work? In which case could you use those models?

Shared frailty models

(Q14) We now include the `+ frailty(id)` statement in `coxph`. What kind of frailty model does this fit? Check `?frailty`.

```
mod_cal_fr <- coxph(Surv(tstart, tstop, status) ~ sex + treat + age + inherit +
                      steroids + propylac + frailty(id),
                    ties = "breslow", cgd)
summary(mod_cal_fr)
```

(Q15) How do the estimates of the covariate effects here compare to the ones from the marginal model with `+ cluster(id)`? What is the estimated frailty variance? Do you think it is significant?

(Q16) The output also shows a likelihood ratio test. Which models does this test compare? What type of hazard ratios are shown in the `exp(coef)` column? Are they relevant for a particular individual or for the whole population?

(Q17) We can also fit a model with log-normal frailty. Do you see a big difference in terms of estimates between this distribution and the gamma frailty?

```
mod_cal_fr_lognorm <- coxph(Surv(tstart, tstop, status) ~ sex + treat + age + inherit +
                   steroids + propylac + frailty(id, distribution = "gaussian"),
                 ties = "breslow", cgd)
summary(mod_cal_fr_lognorm)
```

frailtyEM

Although the `coxph` and `frailty` functions are sometimes enough to get an idea of the model fit, we can get some more results with the `frailtyEM` package. If you don't have it already installed, you can do that with `install.packages("frailtyEM")`. The syntax is similar but a bit different from `coxph`.

(Q18) How do you identify the clusters with the `emfrail()` function? What is the default frailty distribution? What frailty distributions can be fitted with this function?

```
library(frailtyEM)
?emfrail
?emfrail_dist
```

(Q19) A gamma frailty fit is very similar to what we have seen with `coxph`. Do you notice any differences in estimates with the `coxph` fit?

```
mod_gam_1 <- emfrail(Surv(tstart, tstop, status) ~  sex + treat + age + propylac + inherit +
                       steroids + cluster(id), cgd)
summary(mod_gam_1)
```

(Q20) You noticed probably that here is more output here. The field `LRT` also shows a likelihood ratio test. From the output shown here, what models does this test contrast? Do you see any other evidence for the presence of frailty from the output? What is θ in this case?

(Q21) Look into the contents of the `emfrail` fit with `str()`. Can you find the empirical Bayes frailty estimates? How are they different from the ones from the `coxph()` fit?

```
mod_cal_fr$frail
mod_gam_1$frail
```

(Q22) We can also calculate the empirical Bayes estimates by hand (only for the gamma frailty). The formula is:

$$\widehat{z}_i = \frac{\theta + N_i}{\theta + \Lambda_i}$$

where N_i is the number of events observed on individual i, Λ_i is the summed up cumulative intensity from cluster i. What do you think θ is? How is it related to the estimated frailty variance? Take a look at the summary of `mod_gam_1`. Here are the ingredients that we need to calculate that:

```
exp(mod_gam_1$logtheta)
mod_gam_1$nevents_id
mod_gam_1$residuals$group
```

(Q23) Calculate the $\widehat{z}_i$ from this. Does it agree with the empirical Bayes estimates from `coxph` or from `frailtyEM`?

(Q24) Calculate the *sample* mean and variance of the estimated $\widehat{z}_i$. Does the sample variance agree with the estimated frailty variance? Should they?

We can make several types of plots from an `emfrail` fit. I will use those based on the `ggplot2` package here. More info on those can be found in `?autoplot.emfrail`. For conventional plots, `?plot.emfrail` describes

what can be obtained. The advantage of the `ggplot2` plots is that they are objects in R and can be easily edited after they are produced.

Start with a histogram of the frailty estimates:

```
autoplot(mod_gam_1, type = "hist")
```

From the histogram it's quite difficult to tell which individual has which value of the frailty. We can make a so-called *catterpillar plot*:

```
autoplot(mod_gam_1, type = "frail")
```

However, it is still difficult to tell to which individual which frailty value belongs to, since there are so many of them. We can use an interactive visualization for this: install the `plotly` package with `install.packages("plotly")` and then try this:

```
library(plotly)
ggplotly(autoplot(mod_gam_1, type = "frail"))
```

(Q25) Which are the individuals with the highest and lowest frailty estimates?

(Q26) From the output of `mod_gam_1` we can already read a number of estimates that correspond to log-hazard ratios. How would a marginal hazard ratio look like? Let's see this, between a treated and an untreated female, with `age = 24`, with baseline values for the other covariates. What do you expect that will happen with the marginal hazard ratio? How drastic do you think that the effect of the frailty will be, given the strength of the evidence for the frailty?

```
sex <- rep("female", 2)
treat <- c("placebo", "rIFN-g")
age <- c(24, 24)
propylac = c(0, 0)
inherit <- rep("X-linked", 2)
steroids <- rep(0, 2)
newdata <- data.frame(sex, treat, age, propylac, inherit, steroids)

autoplot(mod_gam_1, type = "hr", newdata = newdata)
```

(Q27) Now suppose that we want to predict the cumulative intensity (hazard) for a certain individual. There comes a question: which one is more relevant, the conditional one or the marginal one? Which one would we observe at the level of the population? Which one would be more relevant for a patient?

(Q28) Let's say we take the first woman, on placebo, from the previous question. What does this return? Can you tell what are the fields in this result? Hint: check with `?predict.emfrail`

```
predict(mod_gam_1, newdata = newdata[1,])
```

(Q29) We can look at the plot for the cumulative intensity of our patient of interest. How do you expect that the marginal cumulative intensity will compare to the conditional cumulative intensity?

```
autoplot.emfrail(mod_gam_1, type = "pred", newdata = newdata[1,])
```

The dotted lines represent a 95% confidence band.

Other distributions

(Q30) Let's play a bit with the `distribution` argument in `emfrail`. I'll fit 3 more distributions here. Which one is the inverse Gaussian? (hint: `?emfrail_dist()`)

```
mod_pvf1 <- emfrail(formula = Surv(tstart, tstop, status) ~ sex + treat +
    age + propylac + inherit + steroids + cluster(id),
    distribution = emfrail_dist(dist = "pvf"),
    data = cgd)

mod_pvf2 <- emfrail(formula = Surv(tstart, tstop, status) ~ sex + treat +
    age + propylac + inherit + steroids + cluster(id),
    distribution = emfrail_dist(dist = "pvf", pvfm = 0.5),
    data = cgd)

mod_stab <- emfrail(formula = Surv(tstart, tstop, status) ~ sex + treat +
    age + propylac + inherit + steroids + cluster(id),
    distribution = emfrail_dist(dist = "stable"),
    data = cgd)
```

(Q31) Look at the output of the positive stable frailty model. Is the frailty significant here? Compare the estimates with those from the gamma frailty model and the marginal model with `+ cluster(id)`. To which ones is it closer? Is there anything extra in the output? What do you think it means?

```
summary(mod_stab)
```

(Q32) Optional: how do you think that the marginal and conditional hazard ratios would look like, if we did the same thing with `mod_stab`? Try it!

(Q33) How about model `mod_pvf2`? Notice anything in its summary that wasn't there before?

```
summary(mod_pvf2)
```

(Q34) Compare the likelihoods of all the frailty models we fitted here. Which one is the highest? Was this to be expected? Think of one of the first questions, about how many individuals have how many events.

```
lapply(list(mod_gam_1, mod_stab, mod_pvf1, mod_pvf2), logLik)
```

Lastly, we will note that the estimated *mass at 0* depends heavily of the `pvfm` parameter, that defines the large classes of distribution from within the PVF family. This is not estimated by `frailtyEM`, but rather set by the user. We can however try to estimate it, just by trying out a number of values.

(Q35) I collect here the log-likelihood values for every distribution (for different values of `pvfm`). What is the value of `m` for which the highest log-likelihood of the model is obtained?

```
mvals <- seq(from = 0.1, to = 2, by = 0.2)
models <- lapply(mvals, function(x) emfrail(formula = Surv(tstart, tstop, status) ~ sex +
    treat + age + propylac + inherit + steroids + cluster(id), data = cgd,
    distribution = emfrail_dist(dist = "pvf", pvfm = x)) )

likelihoods <- sapply(models, function(x) x$loglik[2])

plot(mvals, likelihoods)
```

Gap time

Another option with recurrent events data is to analyze it in gap-time. For this we will have to add another column in our data:

```
cgd$gap <- cgd$tstop - cgd$tstart
```

The right hand side of the formulas we used before would stay the same, but the left hand side should now be:

```
mod_gap_fr <- coxph(Surv(gap, status) ~ sex + treat + age + inherit +
                    steroids + propylac + frailty(id),
                  ties = "breslow", cgd)
summary(mod_gap_fr)
```

(Q36) Why can't we use `Surv(tstop, status)`? What would that correspond to? What is the time scale now?

Look at the first individual:

```
cgd[cgd$id == 1, c("treat", "sex", "age", "inherit", "gap", "status")]
```

In `mod_gap_fr`, R doesn't know that we are talking about recurrent events, since the syntax is exactly the same as for clustered data. (Q37) Do you think that matters here?

(Q38) Below is a Kaplan-Meier curve. What do you think that this means here? Do you think that, if we analyze the models in gap time, the survival is more meaningful?

```
plot(survfit(coxph(Surv(gap, status) ~ 1, cgd)))
```

(Q39) Compare the estimated conditional hazard ratios (coefficients) between `mod_cal_fr` and `mod_gap_fr`. Are there similarities? Are the differences notable from a clinical point of view, do you think? Would you expect to have large differences between the two models? Do you think anything is lost by looking at the gap times instead of the calendar time?

(Q40) Furthermore, compare the frailty estimates from the gap time model and the calendar time model. Do you expect them to be similar? What is different between them? Take a look again at the formula for the gamma empirical Bayes estimates that we looked at before.

```
exp(mod_gap_fr$frail)
exp(mod_cal_fr$frail)
```