Homework Problems on Survival Analysis

Soc. 761

Fall 2014

1. The file `Canada-mortality.txt` (on the course web site) contains age-specific mortality rates q_x for Canadian females and males for 2009-2011. Using the function `lifeTable` in the file `survival-analysis.R` (also on the web site), compute life tables for females and males.

   (a) Compare the expectation of life at birth for females and males.

   (b) Graph and compare the age-specific mortality rates q_x for females and males.

   (c) Graph and compare the numbers of survivors by age l_x from the life tables for females and males.

   Suggestions: Use log vertical axes for these graphs; note that you can add a line to an existing graph with the `lines` function in R (see the R script from the lecture for examples).

2. The file `Henning.txt` (on the course web site) contains data from yet another study of criminal recidivism, this one by Henning and Frueh (1996), who followed 194 inmates released from a medium-security prison to a maximum of 3 years from the day of their release; during the period of the study, 106 of the released prisoners were rearrested (depressing, eh?). The data set, which is discussed by Judith Singer and John Willet in *Applied Longitudinal Data Analysis* (Oxford, 2003), contains the following variables (using the names employed by Singer and Willet):

   - `months`: The time of re-arrest in months (but measured to the nearest day).
   - `censor`: A dummy variable coded `1` for censored observations and `0` for uncensored observations. Note that this is the opposite of our usual convention, so in R survival time and censoring should be specified as `Surv(months, censor == 0)`; be careful to use the double-equals sign for testing equality!
   - `personal`: A dummy variable coded `1` for prisoners with a record of crime against persons and `0` otherwise.
   - `property`: A dummy variable coded `1` for prisoners with a record of crime against property and `0` otherwise.
   - `cage`: "Centered" age in years at time of release — that is, age − average age.

   (a) Compute and graph the Kaplan-Meier estimate of the survival function for all of the data.

   (b) Compute and graph separate survival curves for those with and without a record of crime against persons; test for differences between the two survival functions.

   (c) Compute and graph separate survival curves for those with and without a record of crime against property; test for differences between the two survival functions.

3. Continuing with the Henning and Frueh data, fit a Cox regression of time to re-arrest on the covariates `personal`, `property`, and `cage`.

   (a) Determine by a Wald test whether each estimated coefficient is statistically significant.

   (b) Interpret each of the estimated Cox-regression coefficients.

4. Now consider the possibility that `personal` and `property` interact in determining time of re-arrest, by specifying the model formula `Surv(months, censor == 0) ~ personal*property + cage`. Test the statistical significiance of the interaction in two ways: (1) by a Wald test of the coefficient for the interaction regressor `personal:property`; and (2) by a partial-likelihood-ratio test.

   You can use the generic `anova` function to perform the likelihood-ratio test: Assuming that the two fitted models are named `mod.1` and `mod.2`, the partial-likelihood-ratio test is obtained by `anova(mod.1, mod.2, test="Chisq")`.

   Are the results of the Wald and likelihood-ratio tests similar? What do you conclude about the interaction between `personal` and `property`?

5. (optional) Return to the Cox regression model that you fit to Henning and Freuh's recidivism data in question 3 (i.e., the additive model with the variables `personal`, `property`, and `cage` as covariates). Examine diagnostics for this model, in particular checking for non-proportional hazards, influential observations, and nonlinearity. If you discover problems, try to fix them.

   Notes: Because the measurement of event time in this data set is to the nearest day, it requires some programing in R to construct a subject $\times$ time-period data set from which you can form time-varying covariates giving interactions with time. You can proceed as follows (the code is in the file `Henning.R` on the course web site):

```
attach(Henning)

unique.times <- sort(unique(months))
times <- c(0, unique.times)

Henning.mat <- as.matrix(Henning)
Henning.long <- matrix(0, length(unique.times)*nrow(Henning.mat), 10)
vnames <- names(Henning)
colnames(Henning.long) <- c(vnames[1], "subrecord", vnames[2:6], "start", "stop", "arrest")

nrec <- 0
for (i in 1:nrow(Henning.mat)) {
    months.i <- Henning.mat[i,"months"]
    record.i <- Henning.mat[i,]
    subrecord <- 0
    for (j in 2:length(times)){
        if (months.i < times[j]) break
        arrest <- 0
        if (months.i == times[j])
            if (Henning.mat[i,"censor"] == 1) break
                else arrest <- 1
        nrec <- nrec + 1
        subrecord <- subrecord + 1
        Henning.long[nrec,] <- c(record.i[1], subrecord, record.i[2:6],
            times[j-1], times[j], arrest)
        }
    }

Henning.long <- as.data.frame(Henning.long[1:nrec,])

remove(arrest, Henning.mat, i, j, months.i, nrec, record.i, subrecord, times,
    unique.times, vnames)  # clean up
detach(Henning)
```

The resulting subject × time-period data set in the data frame `Henning.long` includes the variables `start`, `stop`, and the event indicator `arrest` required to model interactions with time. To check that you've built the new data set properly, it is wise to verify that you get the same results for the original model fit (in question 3) to the time, censoring-indicator data set `Henning` and to the start, stop, event-indicator data set `Henning.long` just constructed.

You can, alternatively (or additionally), attempt to deal with nonproportional hazards by stratification. For example, to divide centered-age in the Henning data set into four approximately equal-size groups (i.e., at the quartiles):

```
Henning$age.cat <- with(Henning, cut(cage,
    breaks=c(-Inf, quantile(cage, c(.25, .5, .75)), Inf)))
table(Henning$age.cat)  # check the count in each age category
```