Biostatistics 140.623
Third Term, 2017-2018
Problem Set 2 (with R)

Survival in Primary Biliary Cirrhosis

Learning Objectives:

Students who successfully complete this section will be able to:

- To evaluate whether the drug DPCA prolongs life in patients.
- To identify baseline characteristics of patients which predict longer survival.
- Analyze the survival time data (without grouping) by the Kaplan-Meier estimate of the survival function, the log- rank statistic, and Cox proportional hazards model.
- Check the estimated model for its consistency with the observed data; in particular, check the proportional hazards assumption using the complementary log-log plot of the estimated survival function.
- Summarize the findings for public health readers and document and archive the steps of the statistical analysis by creating a script file in R.

Data Set:

Between January 1974 and May 1984, a double-blinded randomized trial on patients with primary biliary cirrhosis (PBC) of the liver was conducted at the Mayo clinic. A total of 312 patients were randomized to either receive the drug D-penicillin (DPCA) or a placebo. Patients were followed until they died from PBC or until censoring, either because of administrative censoring (withdrawn alive at end of study), death not attributable to PBC, liver transplantation, or loss to follow-up. At baseline, a large number of clinical, biochemical, serological and histologic measurements were recorded on each patient. This data set is a subset of the original data, and includes information on each patient's time to death or censoring, treatment, age, gender, serum bilirubin, and histologic disease stage (1-4).

The variables included in this dataset include:
case: unique patient ID number
sex: 0 = male, 1 = female (coded as "Female" and "Male" in the csv file rather than 0/1)
drug: 0 = placebo, 1 = DPCA
bil : serum bilirubin in mg/dl
survyr: time (in years) to death or censoring
death: indicator = 1 if patient died, 0 if censored
ageyr: age in years [continuous variable]
histo: histologic disease stage (1 – 4) [categorical variable]
agecat: age categories, coded as "< 45 yrs", "45 – 55 yrs", and ">= 55 yrs"

Also included in the data set for your possible use are the following indicator (dummy) variables:

Age Indicators (indicator versions of **agecat**):

- **agegr_2**: 1 if patient is 45-55 years old, 0 otherwise
- **agegr_3**: 1 if patient is >= 55 years old, 0 otherwise

Histologic Stage Indicators:

- **hstage2**: 1 if patient is in Stage 2, 0 otherwise
- **hstage3**: 1 if patient is in Stage 3, 0 otherwise
- **hstage4**: 1 if patient is in Stage 4, 0 otherwise

The data are stored in the `csv` data set *`pbctrial.csv`*, which may be downloaded from the course website.

Methods:
Use the data set described above and the appropriate statistical analyses to address the specific learning objectives listed on the first page.

Hints: The hints shown below are based on a dataset with the name `pbcData`, read in with the following code. In the following list of commands, if you want to look at differences by other variables than `drug`, you should change the variable name! Create a new .R file to type/run your commands so that you will have a record of your analysis.

```
library(tidyverse)
pbcData = read_csv("pbctrial.csv")
```

a. Explore the data using descriptive statistics:

```
table()
prop.table()
summary()
etc
```

b. Define a survival object, defining the time variable (`survyr`) and the event (`death == 1`). To do this, you must first install and load the “survival” package:

```
install.packages("survival")  ## only run this the first time
library(survival)

pbcData$SurvObj = with(pbcData, Surv(survyr, death == 1))
```

c. Explore differences in time to death by different baseline variables using graphs and complementary log-log plots.

```
# estimate survival curves for entire sample
km.overall = survfit(SurvObj ~ 1, data = pbcData,
                     type="kaplan-meier", conf.type="log-log")
km.overall
summary(km.overall)

# estimate survival curves for drug group
km.drug = survfit(SurvObj ~ drug, data = pbcData,
                     type="kaplan-meier", conf.type="log-log")
km.drug
summary(km.drug)
```

```
# plot km curves
plot(km.overall)
plot(km.drug)

# log rank test for equality of survivor functions
survdiff(SurvObj ~ drug, data=pbcData)

# complimentary log-log plot
plot(km.drug, fun="cloglog", ylab="log(-log(Survival Probability)",
                        xlab="Analysis time (shown on log scale)")
```

d. Fit several Cox proportional hazards regression models to the ungrouped survival data:

```
model1 = coxph(SurvObj ~ drug, data =  pbcData)
summary(model1)

model2 = coxph(SurvObj ~ sex + bil + as.factor(histo), data = pbcData)
summary(model2)
```

e. Save your R script file that documents and archives the steps of your statistical analysis. This file will make your analysis “reproducible.”

f. **Summarize your findings in a brief report** (less than two pages with at most one table and one figure) as if for a biomedical/public health journal.

A **suggested format** is:

- Introduction – a few sentences about the research question(s)
- Data description – simple tabulations describing patient characteristics
- Results from multiple models that address question(s) (e.g., bivariate and multivariable)
- Graphical display that presents evidence in the data relevant to your scientific question.