Vietnam Institute for Advanced Study in Mathematics

SURVIVAL ANALYSIS

Practical work 4: Inference in Cox proportional hazards regression model

(Lecturers: Agnès LAGNOUX & Jean-François DUPUY)

The PBC dataset (see [1] and [2] for a full description of the variables) comes from a clinical trial in the field of primary biliary cirrhosis conducted at the Mayo Clinic between 1974 and 1984. Primary biliary cirrhosis is a fatal chronic liver disease. A total of 418 PBC patients were randomized to either a placebo or a drug called D-penicillamine. Each of them was followed until death or censoring (the duration is measured in days). The status at endpoint is coded as follows: 0/1/2 for censored, transplant and dead respectively. In addition, 17 covariates are recorded for this study. These include a treatment variable, patient age, gender and clinical, biochemical and histologic measurements made at the time of randomization. In this work, we will mainly consider the following variables: age (in years), serum albumin (g/dl), serum bilirubin (mg/dl), edema (0 if no edema, 0.5 if untreated or successfully treated and 1 if edema despite diuretic therapy) and prothrombin time (standardised blood clotting time).

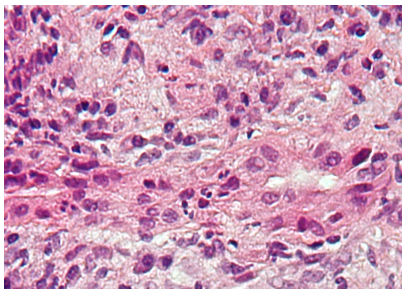

Figure 1: Photograph of primary biliary cirrhosis showing bile duct inflammation

1. Load the `R` package `survival` with: `library(survival)`. Submit `data(pbc)` and `help(pbc)`. Attach the dataset using `attach(pbc)`.

2. Submit `head(pbc,10)` and `pbc[1:10,c("time","status","age","edema","albumin","bili","protime")]`. Comment the outputs.

3. The following code produces a plot of observed survival times:

```
nr=50
plot(c(0,time[1]),c(1,1),type="l",ylim=c(0,nr+1),xlim=c(0,max(time[1:nr])+10),
ylab="nr",xlab="Survival time")
for (i in 2:nr) lines(c(0,time[i]),c(i,i))
for (i in 1:nr){
if (status[i]==0) points(time[i],i,col="red",pch=20) #censored
```

```
if (status[i]==1) points(time[i],i)
}
```

4. Determine the numbers of patients who died, got transplant and were censored. We exclude patients with transplant from further analysis: `pbc=subset(pbc,status!=1)`.
   Finally, the command `pbc=transform(pbc,status=as.logical(status))` changes the status variable to a logical one.

5. We consider first the single factor `edema`. The following code calculates the Kaplan-Meier estimate of the survival function within each level of `edema` and plots the three estimates on a same graph:

   ```
   pbcSurv=survfit(Surv(time,status)~edema,data=pbc)
   plot(pbcSurv,mark.time=FALSE,conf.int=FALSE,col=c("black","grey","red"),
   lty=1:3)
   legend(2500,1,c("no edema","untreated or successfully treated","edema
   despite diuretic therapy"),col=c("black","grey","red"),lty=1:3)
   ```

   Comment this plot and test (using the logrank) the null hypothesis that the survival distribution is the same across the three levels of `edema`. What is the null asymptotic distribution of the test statistic? What is your conclusion?

6. The command `fit=coxph(Surv(time,status)~factor(edema),data=pbc)` fits a Cox proportional hazards model to the data, with `edema` as the single (factor) covariate. Interpret the coefficients in this model. Comment the output of `summary(fit)`.
   In particular, use some simple calculations to find the values in the columns `z` and `Pr(>|z|)`. What is the null hypothesis tested by the likelihood ratio, Wald and score tests at the bottom of the output? What is the asymptotic distribution of these test statistics under the null? What can be said about the value of the score test (compare with the logrank test of question 5)?

7. The command

   ```
   fit.pbc=coxph(Surv(time,status)~age+factor(edema)+log(bili)+log(albumin)+
   log(protime),data=pbc)
   ```

   fits a Cox proportional hazards model to the data, with `age`, `edema` and log transformations of `bilirubin`, `albumin` and `protime` as covariates.

   (a) Write the corresponding hazard function. Interpret the coefficients in this model.

   (b) Submit `fit.pbc` and `summary(fit.pbc)` and comment the results. In particular, find some of the values in the columns `z` and `Pr(>|z|)`. What is the null hypothesis tested by the likelihood ratio (LRT), Wald and score tests at the bottom of the output? What is the asymptotic distribution of these test statistics under the null?

   (c) Submit and comment `fit.pbc$loglik`. Use the results to find the value of the overall LRT.

   (d) Submit and comment `fit.pbc$coeff` and `fit.pbc$var`. Use the results to find the value of the overall Wald test (recall that in `R`, the inverse of a matrix M can be obtained by `solve(M)`).

   (e) Test the null hypothesis that `edema` has no significant effect on survival (use a LRT and calculate the p-value).

   (f) Test the null hypothesis that levels "0.5" and "1" of `edema` have the same effect on survival (this effect being eventually different from 0 and from the effect of level "0") (hint: use a contrast to construct a Wald-type test).

8. (a) What is the estimate of the hazard ratio of two patients having covariates `(age,edema)=(40,0.5)` and `(age,edema)=(30,0)` respectively (all other covariates being equal)? Is this ratio significantly different from 1 (hint: use a Wald test)?

(b) The previous question can easily be answered by using the function `contrast` of the package `rms`[1]. Submit the code below and comment the results:

```
require(rms)
fit.cont=cph(Surv(time,status)~age+as.factor(edema)+log(bili)
+log(albumin)+log(protime),data=pbc)
contrast(fit.cont,list(age=40,edema=.5,bili=3,albumin=3,protime=2),
list(age=30,edema=0,bili=3,albumin=3,protime=2))
```

In particular, explain how the results given as "`Contrast`", "`S.E.`", "`Lower`", "`Upper`", "`Z`" and "`Pr(>|z|)`" can be obtained from the results of question 8(a). Submit and comment:

```
contrast(fit.cont,list(age=56,edema=.5,bili=10,albumin=10,protime=10),
list(age=46,edema=0,bili=10,albumin=10,protime=10))
```

(c) Use the `contrast` function to test the hypothesis of question 7(f).

9. We now perform stepwise model selection by AIC and BIC. Submit and comment the following code:

```
pbc=na.omit(pbc)
pbc$edema=as.factor(pbc$edema)
fit.max=coxph(Surv(time,status)~.-id,data=pbc)

library(MASS)
fit.aic=stepAIC(fit.max,direction="both")
summary(fit.aic)

n=dim(pbc)[1]
fit.bic=stepAIC(fit.max,direction="both",k=log(n))
summary(fit.bic)
```

Using some simple calculations, find the AIC and BIC for both maximal and final models.

10. The following code computes and plots the predicted survival function of a new patient (that is, a patient who does not belong to the original dataset), based on a given estimated Cox model. The new patient is defined by specifying some values for the covariates. Comment and submit this code:

```
fit.pbc=coxph(Surv(time,status)~age+as.factor(edema)+log(bili)+log(albumin)
+log(protime),data=pbc)
newpatient=data.frame(age=40,edema=1,bili=5,albumin=3,protime=10)
plot(survfit(fit.pbc,newdata=newpatient),xscale=365.25,mark.time=FALSE,
xlab="Years",ylab="Survival")
```

Adapt this code to create a patient with: `(age,edema,bili,albumin,protime)=(40,0,5,3,10)` and add his estimated survival function to the previous graph.

References

[1] Fleming T. R., Harrington D. P., 1991. *Counting Processes and Survival Analysis*. Wiley, New York.
[2] Martinussen T., Scheike T. H., 2006. *Dynamic Regression Models for Survivial Data*. Springer, New York.

[1] rms: Regression Modeling Strategies, see http://cran.r-project.org/web/packages/rms/