Vietnam Institute for Advanced Study in Mathematics

SURVIVAL ANALYSIS

Practical work 5: Validation of Cox proportional hazards regression model

(Lecturers: Agnès LAGNOUX & Jean-François DUPUY)

We consider the PBC dataset (see [1] and [2] for a full description of the variables). This practical session introduces various tools available in `R` for validating Cox proportional hazards model: goodness-of-fit statistics, residuals, influence, graphical tools (a detailed account on the topic can be found in [3]).

1. Submit to `R` the following lines:

```
library(survival)
data(pbc)
pbc=na.omit(pbc)
pbc=subset(pbc,status!=1)
pbc$death=(pbc$status==2)*1
attach(pbc)
```

2. Martingale residuals are useful for exploring the correct functional form of the covariates. The following code computes martingale residuals from the null model, that is, the model without any covariate:

```
fit.null=coxph(Surv(time,death)~1,data=pbc)
mr=resid(fit.null)
```

The code below plots martingale residuals against `age` and superimpose a scatterplot smooth. What is your conclusion?

```
plot(age,mr,xlab='Age',ylab='Martingale residuals')
lines(lowess(age,mr,iter=0),lwd=2,col="red")
```

Adapt this code to `bilirubin`. What kind of shape would you suggest for this variable? Make the appropriate transformation and plot martingale residuals against the transformed `bilirubin`. Comment the result.

3. We consider scaled Schoenfeld residuals for exploring nonproportionality. The `R` function `cox.zph` can be used to obtain these residuals, along with a test of the proportional hazards assumption based on writing each time-dependent regression coefficient as: $\beta_j(t) = \beta + \theta_j g_j(t)$, where g_j is some known time transformation.

   (a) Submit and comment the following code (in particular, what is the hypothesis tested in the column `chisq`? on the line `GLOBAL`? using simple calculations, find some of the values in the column `p`):

```
cox=coxph(Surv(time/365,death)~age+factor(edema)+logBilirubin+albumin
+protime,data=pbc)
```

```
summary(cox)
time.test=cox.zph(cox,transform="log")
time.test
```

(b) For each covariate, a plot of scaled Schoenfeld residuals against time is obtained by:

```
for (i in 1:6){
plot(time.test[i])
abline(h=coef(cox)[i],col="red",lwd=2)
par(ask=T)
}
```

Each plot includes an horizontal line at the corresponding regression coefficient estimate. Interpret the graphs. Are they coherent with results of 3(a)?

4. Based on AIC stepwise selection of practical work 1, we consider the following variables: `age`, `edema`, `bili`, `albumin` and `protime`. We investigate various graphical tools and goodness-of-fit statistics based on score processes and their supremum over time. These tools are based on the `cox.aalen` function of the package `timereg`.

(a) Submit the code below and comment its output:

```
fit.cox=cox.aalen(Surv(time/365,death)~prop(age)+prop(logBilirubin)
+prop(albumin)+prop(protime),weighted.test=0,pbc)
summary(fit.cox)
```

To plot the score processes, use: `plot(fit.cox,score=T,xlab="Time (years)")`

(b) A weighted version of the supremum test statistics taking the variance of score processes into account can be obtained by slightly modifying the code above:

```
fit.cox.w=cox.aalen(Surv(time/365,death) prop(age)+prop(logBilirubin)
+prop(albumin)+prop(protime),weighted.test=1,pbc)
summary(fit.cox.w)
```

5. Dfbeta residuals allow to assess influence (that is, the impact of each observation on the fit of a model). The code below computes and displays dfbeta residuals for `age`:

```
fit.pbc=coxph(Surv(time,death)~age+factor(edema)+logBilirubin+albumin
+protime,data=pbc)
rdfb=resid(fit.pbc,type='dfbeta')
plot(age,rdfb[,1],xlab='Age',ylab='Influence for Age',xlim=c(25,85))
```

Use `identify(age,rdfb[,1])` to identify the influent observation. Adapt this code to display dfbeta residuals for the other variables in the model and to identify influent individuals[1].

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.
[3] Therneau T. M., Grambsch P. M., 2001. *Modeling Survival Data: Extending the Cox Model*. Springer, New York.

[1] Note that case 210 was discovered to have a true age of 54.4 years rather than 78.4 while case 107 has protime of 10.7 rather than 17.1 seconds. In this dataset, influence residuals have pointed out some problems in the dataset.