Analyse de survie, Master Bio-info, TP2

Genia Babykina

Analyse de données *lung*

Lectuer de données

1) Lire les données, effectuer les statistiques descriptives, définir les variables qualitatives.

```
data(lung)
```

```
## Warning in data(lung): data set 'lung' not found
```

```
?lung
str(lung)
lung$sex=factor(lung$sex,  labels=c("m", "f"))
```

Estimations

2) Estimer un model de Cox semi-paramétrique pour expliquer le temps jusqu'au décès, en incluant toutes les variables explicatives disponibles. Interpréter les coefficients.

```
model1=coxph(Surv(time, status)~age+sex+ph.ecog+ph.karno+meal.cal+wt.loss, data=lung)
summary(model1)
```

3) Effectuer la sélection automatique des variables en utilisant le modèle complet de la question précédente. Interpréter les résultats.

```
lung1=lung[complete.cases(lung),]
model2=step(coxph(Surv(time, status)~age+sex+ph.ecog+ph.karno+meal.cal+wt.loss, data=lung1))
summary(model2)
```

4) Représenter les courbes de survie $\widehat{S(t)}$, estimées par le modèle pour les hommes et les femmes.

```
model2.sex = survfit(Surv(time, status)~sex, data=lung1)
ggsurvplot(model2.sex,data = lung1)
plot(model2.sex)
```

Analyse des résidus

5) Vérifier l'hypothèse des risques proportionnels : **résidus de Schoenfeld**.

Remarque : résidus de Schoenfeld pour chaque individus ayant eu un événement et pour chaque covariable est calculé comme la différence entre la valeur de sa covariable au temps d'événement et la valeur de cette covariable "prédite" par le modèle.

```
res.schoenf=residuals(model2, type="schoenfeld")
test.prop.hasard = cox.zph(model2)
model2.sex = survfit(Surv(time, status)~sex, data=lung1)
```

6) Vérifier l'hypothèse de proportionnalité des risques graphiquement.

$$h_1(t) = kh_2(t) \Leftrightarrow H_1(t) = kH_2(t) \Leftrightarrow \log(S_1(t)) = k\log(S_2(t))$$

(en effet, $k \times (-H_2(t)) = k\log(S_2(t))$).

Ainsi, $\log(-\log(S_1(t))) = \log(k) + \log(-\log(S_2(t))) \Leftrightarrow$ sur le graphique $\log(t)$ vs $\log(-\log(S_j(t)))$

```
S_f=survfit(Surv(time, status)~1, data=lung1[lung1$sex=="f",])
S_m=survfit(Surv(time, status)~1, data=lung1[lung1$sex=="m",])
plot(log(S_f$time), log(-log(S_f$surv)), type="l", col="red")
lines(log(S_m$time), log(-log(S_m$surv)), type="l", col="blue")
plot(survfit(Surv(time, status)~sex, data=lung1), fun="cloglog")
```

7) Identifier les observations influentes (à l'aide de *dfbetas*) et les observations mal prédites par le modèle (à l'aide des **résidus de deviance**).

Remarque : les résidus de deviance sont symétriques autour de zéro. Valeur positive ⇒ "individu décède trop tôt par rapport à la prédiction", valeur négative ⇒ "individu vit trop longtemps par rapport à la prédiction".

```
res.deviance=residuals(model2, type="deviance")
plot(res.deviance)
res.dfbetas=residuals(model2, type="dfbetas")
plot(res.dfbetas[,1])
text(res.dfbetas[,1],rownames(res.dfbetas[,1]))
ggcoxdiagnostics(model2, type="deviance")
```

8) Vérifier l'hypothèse du lien log-linéaire entre le temps d'événement et les variables quantitatives : **résidus martingales**.

Remarque : si l'hypothèse de log-linéarité est vérifiée, il n'y a pas de lien entre les résidus martingales et les variables quantitatives.

```
res.marting = residuals(model2, type="martingale")
par(mfrow=c(2,2))
plot(lung1$age, res.marting)
lines(lowess(lung1$age, res.marting), col="red")
plot(lung1$ph.karno, res.marting)
lines(lowess(lung1$ph.karno, res.marting), col="red")
plot(lung1$ph.ecog, res.marting)
lines(lowess(lung1$ph.ecog, res.marting), col="red")
plot(lung1$wt.loss, res.marting)
lines(lowess(lung1$wt.loss, res.marting), col="red")

# Choix de la forme de lien possible :
ggcoxfunctional(Surv(time, status)~age+log(age)+age^2, data=lung1)
ggcoxfunctional(Surv(time, status)~ph.karno, data=lung1)
ggcoxfunctional(Surv(time, status)~ph.ecog, data=lung1)
```

Analyser les données myélome

Une étude porte sur l'analyse de survie de 65 patients atteints de myélome. Les donées sont disponibles dans le fichier *myel.csv*. Les donnÈes contiennent les informations suivantes :

Table 1: Variables disponibles

Variables	Description
NOBS	identifiant du sujet
T	temps jusqu'au décès, en mois
DECES	1 si décès, 0 si vivant à la fin d'étude ou sorti d'étude
LOG UREE	concentration d'urée dans le sang (en log)
HB	concentration d'hémoglobine dans le sang
PQ	nombre de plaquettes (0 : faible, 1 : normal)
INGJ0	infection en début d'étude (1 : présence, 0: absence)
AGE	âge du sujet
SEXE	sexe du sujet (1: hommes, 0: femme)
LOG GB	nombre de globules blancs (en log)
FRACTURE	présence (1) ou absence (0) de fracture
LOG GBM	concentration des globules blancs (autre mesure, en log)
P LYMP	variable non utilisée
P MYEL	variable non utilisée
PROT U	Protéinurie dans l'urine
BENCE J	protéine de Bence-Jones dans les urines (1 : présence, 0 : absence)
PROT S	Protéinurie dans le sang
GLOBULINE	concentration de globuline dans le sang
CALCIUM	concentration de calcium dans le sang
TEMP	$T - 1$

9) Analyser ces données. La durée de survie (variable T), est-elle différente que l'on ait, ou non, une protéinurie de type Bence-Jone (variable *BENCE J*).

10) Utilisez le modèle de Cox et d'autres variables disponibles dans ces données afin d'explorer les facteur de risque de décès.