TP : Modèle de Cox pour la prise en compte de covariables

Nous étudions les données `pharmacoSmoking` disponibles dans le package `saur` d'un essai clinique décrit par Steinberg [1] et al. (2009). Il s'agit d'étudier l'effet sur la durée de sevrage du tabac (en jours) d'une combinaison de traitements vs le patch nicotinique seul.

1 Analyse descriptive des covariables

Maintenant nous considérons toutes les covariables du jeu de données `pharmacoSmoking`, soit 9 covariables auxquelles sont ajoutées 2 variables issues de la variable continue `age` regroupées en deux classes `ageGroup2` et quatre classes `ageGroup4`.

- **ttr** Time in days until relapse
- **relapse** Indicator of relapse (return to smoking)
- **grp** Randomly assigned treatment group with levels combination or patchOnly
- **age** Age in years at time of randomization
- **gender** Female or Male
- **race** black, hispanic, white, or other
- **employment** ft (full-time), pt (part-time), or other
- **yearsSmoking** Number of years the patient had been a smoker
- **levelSmoking** heavy or light
- **ageGroup2** Age group with levels 21-49 or 50+
- **ageGroup4** Age group with levels 21-34, 35-49, 50-64, or 65+
- **priorAttempts** The number of prior attempts to quit smoking
- **longestNoSmoke** The longest period of time, in days, that the patient has previously gone without smoking

– Il est utile de faire une analyse descriptive rapide des différentes covariables disponibles. Proposez des visualisations ou des résumés numériques adéquats pour chacune d'entre elles.

2 Estimation d'un modèle de Cox

2.1 La fonction 'coxph'

Renseignez-vous sur la fonction `coxph`.

Différentes stratégies pour choisir les variables pertinentes à inclure dans le modèle sont possibles.

- avec la covariable `grp` :

```
cox_treatment<-coxph(Surv(ttr,relapse)~grp,data=pharmacoSmoking)
summary(cox_treatment)
```

- avec les covariables `grp` et `age` :

1. Steinberg, M.B. Greenhaus, S. Schmelzer, A.C. Bover, M.T., Foulds, J., Hoover, D.R., and Carson, J.L. (2009) Triple-combination pharmacotherapy for medically ill smokers : A randomized trial. Annals of Internal Medicine 150, 447-454.

```
cox_age<-coxph(Surv(ttr,relapse)~grp+age,data=pharmacoSmoking)
summary(cox_age)
```

► avec toutes les covariables :

```
cox_complet<-coxph(Surv(ttr,relapse)~grp+age+gender+race+employment+yearsSmoking
+levelSmoking+priorAttempts+longestNoSmoke,data=pharmacoSmoking)
summary(cox_complet)
```

– Observez les tests globaux du rapport de vraisemblance, de Wald et du score. Comparez avec le test de log-rank du TP1. Observez la façon dont les covariables catégorielles sont traitées. Interprétez les coefficients estimés le cas échéant.

Un graphique synthétique qui reprend les informations du `summary` peut être obtenu avec la fonction : `ggforest(cox_complet)` de la librairie `survminer`.

– Accompagnez vos commentaires de ce graphique.

2.2 Sélection de modèles

Pour choisir les covariables pertinentes, on peut s'aider de méthodes de sélection de variables automatiques comme les critères pénalisés AIC et BIC :

$$AIC(\beta) = -2L(\beta_1, \cdots, \beta_p) + \text{pen}(p) \quad \text{avec pen}(p) = 2p$$

Le terme de vraisemblance (partielle de Cox) augmente lorsque p augmente, donc $-2L(\beta_1, \cdots, \beta_p)$ diminue et le terme de pénalité $\text{pen}(p) = 2p$ augmente. avec p. On cherche donc un compromis de la valeur de p qui minimise le critère pénalisé $AIC(\beta)$.

Avec le choix $\text{pen}(p) = 2p \times \ln(n)$ on obtient le critère $BIC(\beta)$ qui pénalise davantage les modèles avec beaucoup de covariables.

```
library(MASS)
stepAIC(cox_complet,k=2) #AIC
stepAIC(cox_complet,k=log(nrow(pharmacoSmoking))) #BIC
```

– Expliquez comment est effectuée la sélection de variables avec `stepAIC` et donnez les modèles sélectionnés avec les critères BIC et AIC.

Si le nombre de covariables p est très grand ($p \geq n$) on peut utiliser des pénalités LASSO. On va alors chercher à maximiser la vraisemblance partielle pénalisée :

$$L(\beta_1, \cdots, \beta_p) - \lambda \sum_{j=1}^{p} |\beta_j|$$

où λ est le paramètre de "régularisation" à calibrer. La fonction `cv.glmnet` de la librairie `glmnet` permet d'implémenter ce type de pénalité (cf. Friedman, Hastie, Tibshirani, 2009 [2]) et d'obtenir un modèle parcimonieux lorsqu'on a un grand nombre de covariables..

– Après avoir implémenté les méthodes classiques AIC/BIC, effectuez la sélection LASSO (même si pour ce jeu de données, on n'est pas vraiment dans la "grande dimension").

2. https ://web.stanford.edu/ hastie/Papers/glmnet.pdf

2.3 Vérification des hypothèses

► **Résidus de Schoënfeld :**Tester l'hypothèse des risque proportionnels. Reprenons le modèle sélectionné par BIC.

```
bic_model<-coxph(Surv(ttr,relapse)~grp+age,data=pharmacoSmoking)
cox.zph(bic_model)

residus<-cox.zph(bic_model)
residus$y

library(survminer)
ggcoxzph(cox.zph(bic_model))
```

La fonction `cox.zph` réalise le test de la non corrélation des résidus avec le temps (ou une transformation du temps) pour chacune des covariables présentes dans le modèle en utilisant une statistique de test basée sur le score. Les résidus sont accessibles dans `residus$y`. Il est possible de tester la nullité de la pente d'un ajustement linéaire des résidus (à faire "à la main"). La fonction `ggcoxzph` de la librairie `survminer` permet de visualiser les résidus.

► Graphique "LML" ("Log-Minus-Log") : Afin de valider graphiquement l'hypothèse des risques proportionnel pour une variable catégorielle, il est possible de représenter les courbes "LML" dans chaque catégorie de cette variable, les autres covariables étant supposées égales par ailleurs.

```
bic_model
indiv_new <- data.frame(grp = c("patchOnly","combination"),age=c(50,50))
survie_pred <- survfit(bic_model, newdata = indiv_new,conf.type="none")
plot(survie_pred,fun="cloglog",col=c(1,2))
```

– Rappelez comment sont construites les courbes "LML3". Expliquez ce que produit le code ci-dessus. Représentez les courbes "LML" associées à la variable `grp` pour différentes valeurs de `age`. Conclure.

3 Extension du modèle de Cox

► **Stratifier une variable continue :** L'avantage de la stratification est que la variable n'a pas besoin de vérifier l'hypothèse des risques proportionnels. On utilise la variable `ageGroup4` qui est déjà présente dans le jeu de données. Cette variable correspond à une stratification de la variable continue `age` en quatre classes : $21-34$; $35-49$; $50-64$ et $65+$.

```
## Variable age continue
cox_ageGroup4=coxph(Surv(ttr,relapse)~age+grp, data=pharmacoSmoking)

## Variable ageGroup4
cox_ageGroup4=coxph(Surv(ttr,relapse)~ageGroup4+grp, data=pharmacoSmoking)

## Variable age stratifiée en 4 classes
cox_ageGroup4_st=coxph(Surv(ttr,relapse)~grp:strata(ageGroup4), data=pharmacoSmoking)
```

– Comparez les trois façons d'introduire l'âge dans la modélisation. Donnez l'expression de la fonction de risque instantané avec chacune des façons d'introduire l'effet de l'âge (variable continue, variable catégorielle, variable de stratification). Expliquez les avantages/inconvénients de chaque approche.

► **Introduire une dépendance au temps :** Si l'on détecte une forme fonctionnelle de dépendance au temps dans les résidus de Schoënfeld, on peut essayer de la modéliser. Par exemple, l'option directement utilisable pour modéliser

une dépendance au temps linéaire peut se faire en ajoutant `tt(age)` dans le modèle. La variable `age` intervient alors dans le modèle sous la forme $\exp(\beta_{age} \times age \times t)$.

```
cox_grp_age_tt=coxph(Surv(ttr,relapse)~age+tt(age)+grp, data=pharmacoSmoking)
summary(cox_grp_age_tt)
```

– Commentez la sortie R de ce modèle.

4 Prédiction de la fonction de risque (pour un individu de caractéristiques données)

Prédire la fonction de survie d'un individu d'âge donné et ayant reçu l'un des deux traitements (combiné ou le patch seul). Par exemple un individu :

– d'âge 25 ans, ayant reçu le traitement combiné,

– d'âge 25 ans ayant reçu le patch seul

– d'âge 55 ans, ayant reçu le traitement combiné,

– d'âge 55 ans ayant reçu le patch seul

Exemple avec le modèle sélectionné par BIC :

```
indiv_new <- data.frame(age = rep(20,times=2),
                        grp = c("patchOnly","combination"))
surv.new <- survfit(bic_model, newdata = indiv_new)
```

– Comparez les durées médianes de rechute pour ces différents profils d'individus prédites avec le modèle où la variable `age` est introduite comme une covariable continue et avec le modèle où elle est utilisée comme variable de stratification.