Analyse de survie : le modèle de Cox

Olivier Bouaziz

Présentation du modèle de Cox

Les observations

- On dispose d'un échantillon $T_1, \ldots, T_n, \Delta_1, \ldots, \Delta_n$,

$$\begin{cases} T_i = \min(\tilde{T}_i, C_i) \\ \Delta_i = 1 \text{ si } \tilde{T}_i \leq C_i, 0 \text{ sinon.} \end{cases}$$

On a maintenant également à disposition des covariables de dimension p : $Z_1, \ldots, Z_n$, où $Z_i = (Z_{i1}, \ldots, Z_{ip})^T$.

- On cherche alors à estimer le risque instantanné conditionnel aux covariables :

$$h(t|Z_{i1}, \ldots, Z_{ip}) = \lim_{dt \to 0} \frac{P[t \leq \tilde{T} < t + dt | \tilde{T} \geq t, Z_{i1}, \ldots, Z_{ip}]}{dt}$$

Soit $S(t|Z_{i1}, \ldots, Z_{ip}) = P[\tilde{T} \geq t|Z_{i1}, \ldots, Z_{ip}]$ et $f(t|Z_{i1}, \ldots, Z_{ip})$ la densité conditionnelle de $\tilde{T}$ sachant $Z_{i1}, \ldots, Z_{ip}$. Alors :

$$h(t|Z_{i1}, \ldots, Z_{ip}) = \frac{f(t|Z_{i1}, \ldots, Z_{ip})}{S(t|Z_{i1}, \ldots, Z_{ip})}$$

Le modèle de Cox est un modèle directement sur **le risque instantané** !

Exemple : les données de mélanome

Les données de mélanome du package **timereg** contiennent les survies relatives à des patients après opération pour un mélanome malin. La base de données contient 205 patients pour 3 covariables

```
library(timereg)
```

```
## Loading required package: survival
```

```
data(melanoma)
head(melanoma)
```

```
##    no status days ulc thick sex
## 1 789      3   10   1   676   1
## 2  13      3   30   0    65   1
## 3  97      2   35   0   134   1
## 4  16      3   99   0   290   0
## 5  21      1  185   1  1208   1
## 6 469      1  204   1   484   1
```

- ulc : l'ulcération, 1 pour présent, 0 pour absent.
- thick : l'épaisseur de la tumeur (1/100 mm)
- sex : 0 pour femme, 1 pour homme.

Exemple : les données PBC

Les données de cirrhose biliaire primitive du package **survival** contiennent les survies relatives de 418 patients atteints de cirrhose biliaire primitive. La base de données contient 17 covariables.

```
library(survival)
head(pbc[,c("time","status","age","edema","bili","protime","albumin")])
```

```
##   time status      age edema bili protime albumin
## 1  400      2 58.76523   1.0 14.5    12.2    2.60
## 2 4500      0 56.44627   0.0  1.1    10.6    4.14
## 3 1012      2 70.07255   0.5  1.4    12.0    3.48
## 4 1925      2 54.74059   0.5  1.8    10.3    2.54
## 5 1504      1 38.10541   0.0  3.4    10.9    3.53
## 6 2503      2 66.25873   0.0  0.8    11.0    3.98
```

- age : l'âge du patient.
- edema : l'oedème, 0 pas d'oedème, 0.5 oedème non traité ou traité avec succès, 1 oedème présent malgré traitement.
- bili : serum bilirunbin (mg/dl)
- protime : temps de coagulation du sang (standardisé)
- albumin : serum albumin (g/dl)
- ...

Un modèle de regression semi-paramétrique

Soit $\theta_0 = (\theta_1, \ldots, \theta_p)$. Le modèle de Cox s'écrit :

$$\begin{aligned} h(t|Z_{i1}, \ldots, Z_{ip}) &= h_0(t) \exp(\theta_1 Z_{i1} + \cdots + \theta_p Z_{ip}) \\ &= h_0(t) \exp(\theta_0 Z_i) \end{aligned}$$

où :

- $h_0(t)$ est une **fonction** inconnue de t. C'est le risque de base, ce risque ne dépend pas de θ_0.
- $\theta_0 = (\theta_1, \ldots, \theta_p) \in R^p$ est un **paramètre** inconnu de dimension p. Ce paramètre ne dépend pas du temps t. Il représente **l'effet des covariables** sur le risque instantané.
- $\exp(\theta_1 Z_{i1} + \cdots + \theta_p Z_{ip})$ est le risque relatif.

Les objectifs :

- Estimation des paramètres $\theta_1, \ldots, \theta_p$.
- Estimation du risque de base $h_0(t)$.

Interprétation du modèle en terme de rapport de risques (pour une variable qualitative)

Si Z_{i1} est une variable **qualitative**, par exemple :

$$Z_{i1} = \begin{cases} 1 \text{ si individu } i \text{ est un homme} \\ 0 \text{ si individu } i \text{ est une femme.} \end{cases}$$

$$\frac{h_0(t)\exp(\theta_1 \times 1 + \cdots + \theta_p Z_{ip})}{h_0(t)\exp(\theta_1 \times 0 + \cdots + \theta_p Z_{ip})} = \exp(\theta_1)$$

$\exp(\theta_1)$ est alors le rapport de risque entre un homme et une femme toutes choses étant égales par ailleurs (c'est à dire après avoir pris en compte la valeur des autres covariables).

- Si $\theta_1 > 0$, $\exp(\theta_1) > 1$ et le risque que l'évènement d'intérêt se produise est plus élevé chez les hommes que chez les femmes.
- Si $\theta_1 < 0$, $\exp(\theta_1) < 1$ et le risque que l'évènement d'intérêt se produise est plus faible chez les hommes que chez les femmes.
- Si $\theta_1 = 0$, $\exp(\theta_1) = 1$ et le risque instantané est le même chez les hommes et chez les femmes.

Interprétation du modèle en terme de rapport de risques (pour une variable continue)

θ_1 est l'effet de Z_{i1} sur le risque instantané après avoir pris en compte la valeur des autres covariables (c'est à dire l'effet de Z_{i1} toutes choses étant égales par ailleurs).

Si Z_{i1} est une variable **continue**, θ_1 peut être interprété en terme de rapport de risque quand la variable Z_{i1} augmente d'une unité :

$$\frac{h_0(t)\exp(\theta_1(Z_{i1}+1)+\cdots+\theta_p Z_{ip})}{h_0(t)\exp(\theta_1 Z_{i1}+\cdots+\theta_p Z_{ip})} = \exp(\theta_1)$$

- Si $\theta_1 > 0$, $\exp(\theta_1) > 1$ et le risque que l'évènement d'intérêt se produise augmente quand Z_{i1} augmente (et diminue quand Z_{i1} diminue).
- Si $\theta_1 < 0$, $\exp(\theta_1) < 1$ et le risque que l'évènement d'intérêt se produise augmente quand Z_{i1} diminue (et diminue quand Z_{i1} augmente).
- Si $\theta_1 = 0$, $\exp(\theta_1) = 1$ et la variable Z_{i1} n'a pas d'impact sur le risque instantané.

Interprétation du risque de base

Le risque de base $h_0(t)$ correspond au risque instantané quand toutes les covariables sont égales à 0 :

$$h_0(t) = h(t|Z_{i1} = 0, \ldots, Z_{ip} = 0)$$

Ce risque dépend du temps t.

En pratique, on cherchera à estimer

$$H_0(t) = \int_0^t h_0(t)dt$$

Attention : le risque de base **ne représente pas** le risque instantané en l'absence de covariables !

Hypothèses du modèle de Cox

Ce modèle fait donc les deux hypothèses suivantes sur les données :

- Le rapport des risques instantanés ("hazard rate" en anglais) de deux patients est **indépendant** du temps. C'est l'hypothèse des **risques proportionnels**.
- $\log(h(t|Z_{i1}, \ldots, Z_{ip})) = \log(h_0(t)) + \theta_0 Z_i$. Le logarithme du risque instantané est une fonction linéaire des Z_{ij}. C'est l'hypothèse de **log-linéarité**.

Exemple : les données mélanome (covariable sexe)

On considère le modèle de Cox pour les données mélanome avec comme variable explicative simplement le sexe :

$$h(t|\text{sexe}) = h_0(t)\exp(\theta_1 \text{sexe})$$

```
fit=coxph(Surv(days,status==1)~ factor(sex),data=melanoma)
fit
```

```
## Call:
## coxph(formula = Surv(days, status == 1) ~ factor(sex), data = melanoma)
##
##                coef exp(coef) se(coef)   z     p
## factor(sex)1 0.662     1.939    0.265 2.5 0.013
##
## Likelihood ratio test=6.15  on 1 df, p=0.0131
## n= 205, number of events= 57
```

Exemple : les données mélanome (covariables le sexe et la tumeur)

On considère le modèle de Cox pour les données mélanome avec comme variables explicatives le sexe et le log de la tumeur :

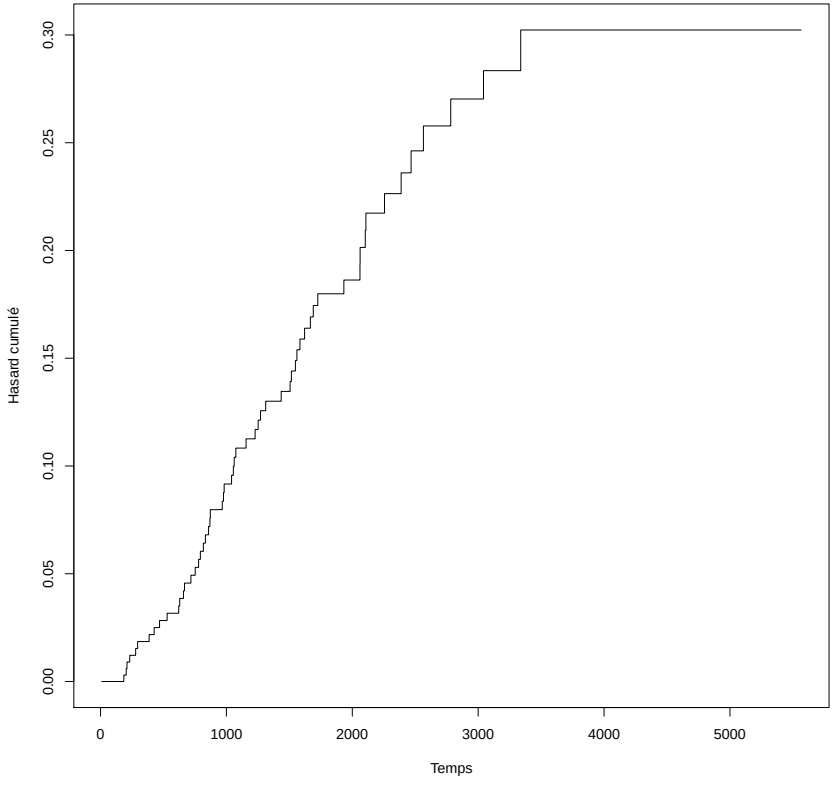
$$h(t|\text{sexe}, \log(\text{tum})) = h_0(t) \exp(\theta_1 \text{sexe} + \theta_2 \log(\text{tum}))$$

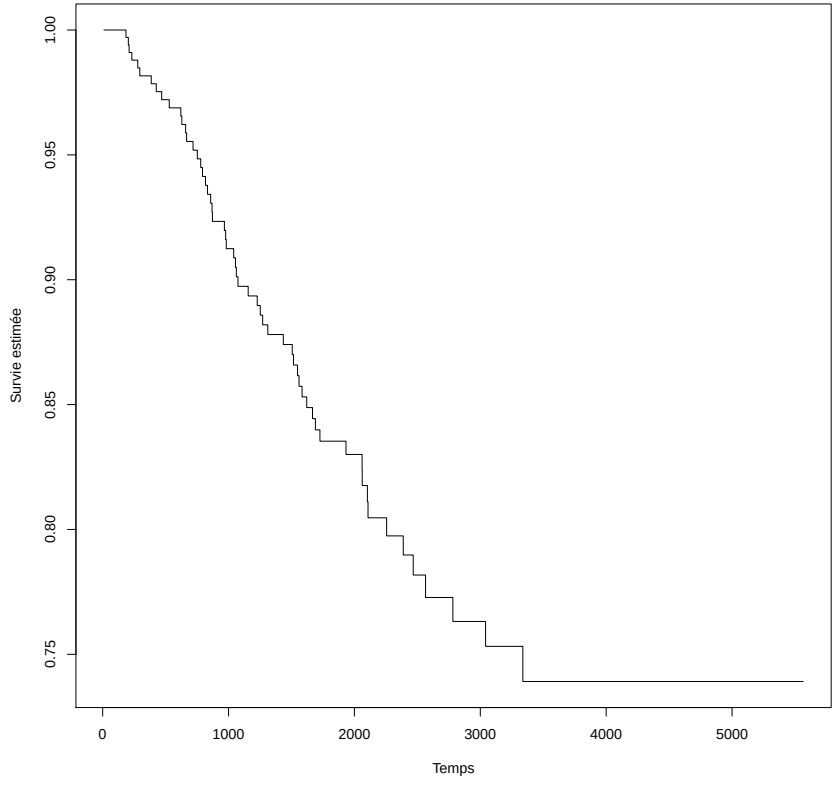
```
melanoma$lthick=log(melanoma$thick)
fit=coxph(Surv(days,status==1)~ factor(sex)+lthick,data=melanoma)
fit
```

```
## Call:
## coxph(formula = Surv(days, status == 1) ~ factor(sex) + lthick,
##     data = melanoma)
##
##                 coef exp(coef) se(coef)    z       p
## factor(sex)1 0.458     1.581    0.269 1.70   0.088
## lthick       0.781     2.183    0.157 4.96 6.9e-07
##
## Likelihood ratio test=33.5  on 2 df, p=5.45e-08
## n= 205, number of events= 57
```

Exemple : les données mélanome (covariables le sexe et la tumeur)

```
fit=coxph(Surv(days,status==1)~ factor(sex)+c(lthick-mean(lthick)),data=melanoma
Hazcum=basehaz(fit,centered=FALSE)
par(mfrow=c(1,2))
plot(Hazcum$time,Hazcum$hazard,type="s",xlab="Temps",ylab="Hasard cumulé")
plot(Hazcum$time,exp(-Hazcum$hazard),type="s",xlab="Temps",ylab="Survie estimée
```





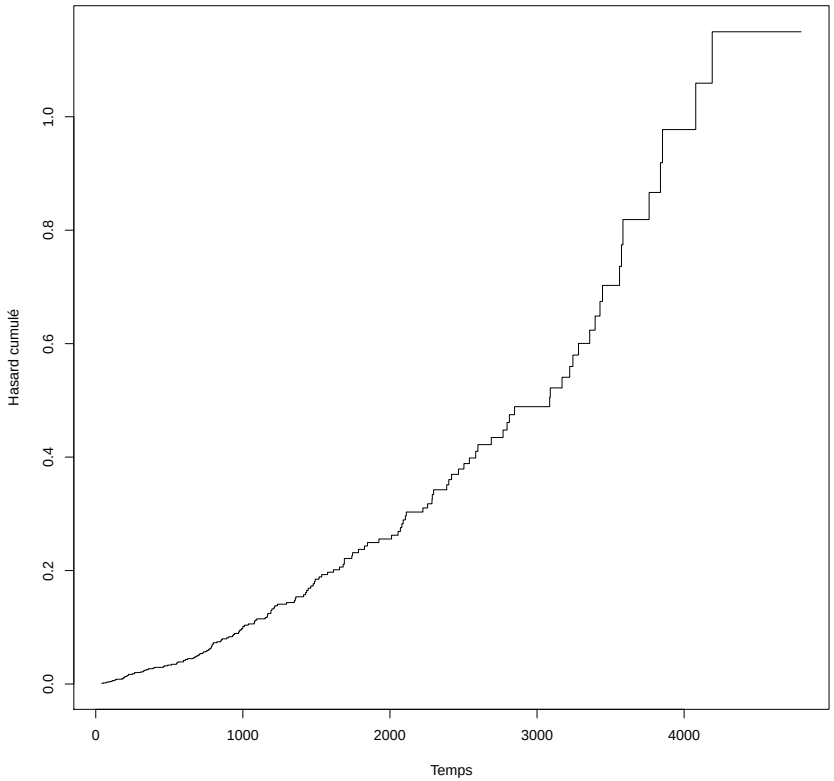
Exemple : les données PBC

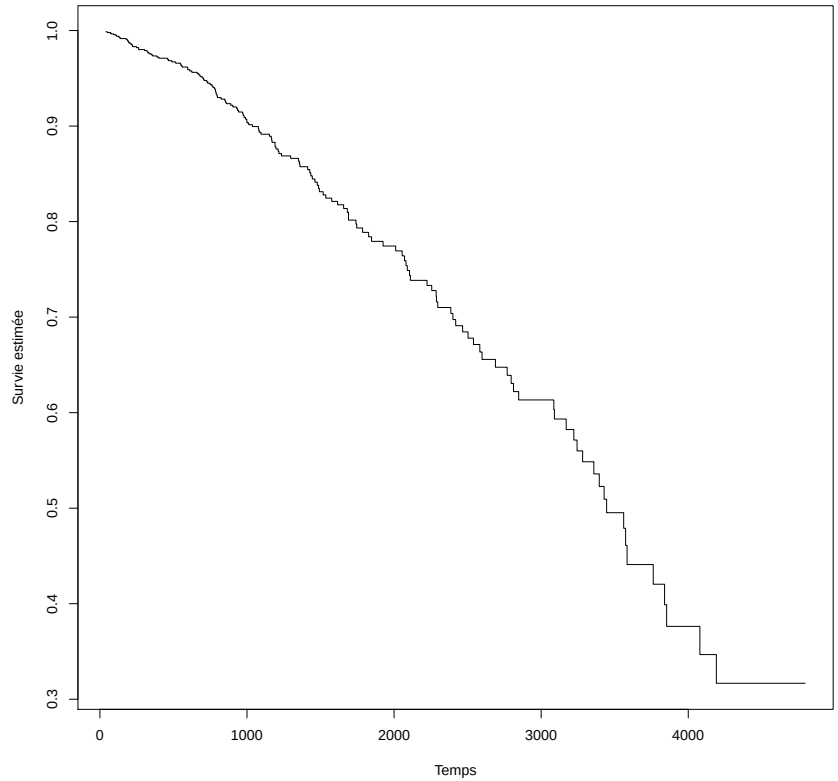
```
library(survival)
fit.pbc<-coxph(Surv(time,status==2) ~ age+factor(edema)+log(bili)
               +log(protime)+log(albumin),data=pbc)
fit.pbc
```

```
## Call:
## coxph(formula = Surv(time, status == 2) ~ age + factor(edema) +
##     log(bili) + log(protime) + log(albumin), data = pbc)
##
##                       coef exp(coef) se(coef)     z       p
## age                0.04045   1.04128  0.00771  5.25 1.6e-07
## factor(edema)0.5   0.28188   1.32563  0.22523  1.25 0.21074
## factor(edema)1     1.01247   2.75240  0.28975  3.49 0.00048
## log(bili)          0.85905   2.36091  0.08324 10.32 < 2e-16
## log(protime)       2.35993  10.59020  0.77314  3.05 0.00227
## log(albumin)      -2.51576   0.08080  0.65262 -3.85 0.00012
##
## Likelihood ratio test=232  on 6 df, p=0
## n= 416, number of events= 160
##    (2 observations deleted due to missingness)
```

Exemple : les données PBC

```
fit.pbc=coxph(Surv(time,status==2)~ c(age-mean(age))+factor(edema)+c(log(bili)-m
Hazcum.pbc=basehaz(fit.pbc,centered=FALSE)
par(mfrow=c(1,2))
plot(Hazcum.pbc$time,Hazcum.pbc$hazard,type="s",xlab="Temps",ylab="Hasard cumul
plot(Hazcum.pbc$time,exp(-Hazcum.pbc$hazard),type="s",xlab="Temps",ylab="Survie
```





Estimation et propriétés des estimateurs dans le modèle de Cox

La vraisemblance du modèle de Cox

La vraisemblance du modèle est égale à :

$$\prod_{i=1}^{n} \left\{ f(T_i|Z_{i1},\ldots,Z_{ip}) S_{C_i}(T_i|Z_{i1},\ldots,Z_{ip}) \right\}^{\Delta_i} \left\{ f_{C_i}(T_i|Z_{i1},\ldots,Z_{ip}) S(T_i|Z_{i1},\ldots,Z_{ip}) \right\}^{1-\Delta_i}$$

On fait les hypothèses :

- Censure **indépendante** de la variable d'intérêt conditionnellement aux covariables : $\tilde{T}$ est indépendant de C conditionnellement à Z.
- Censure **non-informative** : la loi de C ne dépend pas des paramètres du modèle (ni de $\theta_1,\ldots,\theta_p$ ni de h_0).

Alors la vraisemblance du modèle se simplifie par :

$$\prod_{i=1}^{n} \left\{ h(T_i|Z_{i1},\ldots,Z_{ip}) \right\}^{\Delta_i} S(T_i|Z_{i1},\ldots,Z_{ip})$$

$$= \prod_{i=1}^{n} \left\{ h_0(T_i) \exp(\theta_1 Z_{i1} + \cdots + \theta_p Z_{ip}) \right\}^{\Delta_i} \exp\left(- \int_0^{T_i} h_0(t) dt \exp(\theta_1 Z_{i1} + \cdots + \theta_p Z_{ip}) \right)$$

La vraisemblance du modèle de Cox

En se restreignant aux fonctions constantes par morceaux pour le risque de base cumulé $H_0(t)$ et en injectant ce type de fonctions dans l'écriture de la vraisemblance, on peut montrer (hors programme) que la vraisemblance se simplifie par :

$$L_n(\theta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\theta_1 Z_{i1} + \cdots + \theta_p Z_{ip})}{\sum_j I(T_j \geq T_i) \exp(\theta_1 Z_{j1} + \cdots + \theta_p Z_{jp})} \right\}^{\Delta_i}$$

Cette vraisemblance s'appelle la vraisemblance **partielle** de Cox. L'estimateur de θ_0 est alors défini de la façon suivante :

$$\hat{\theta} = \text{argmax}_\theta L_n(\theta).$$

On estime le risque de base cumulé par :

$$\hat{H}_0(t) = \sum_{i=1}^{n} \frac{d_i I(T_i \leq t)}{\sum_j I(T_j \geq T_i) \exp(\hat{\theta} Z_j)}$$

Ecriture de la log-vraisemblance et du vecteur score

La log-vraisemblance pour le paramètre θ est égale à :

$$\mathcal{L}_n(\theta) = \log(L_n(\theta)) = \sum_{i=1}^{n} \Delta_i \left\{ \theta Z_i - \log \left(\sum_{j=1}^{n} I(T_j \geq T_i) \exp(\theta Z_j) \right) \right\}$$

Le vecteur score de la log-vraisemblance vaut 0 en $\theta = \hat{\theta}$ et est égal à :

$$\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} = \sum_{i=1}^{n} \Delta_i \{ Z_i - E_i(\theta) \}$$

où

$$E_i(\theta) = \frac{\sum_{j=1}^{n} I(T_j \geq T_i) Z_j \exp(\theta Z_j)}{\sum_{j=1}^{n} I(T_j \geq T_i) \exp(\theta Z_j)}$$

Ecriture de la matrice hessienne

La matrice hessienne de la log-vraisemblance est égale à :

$$\frac{\partial^2 \mathcal{L}_n(\theta)}{\partial \theta^2} = -\sum_{i=1}^{n} \Delta_i V_i(\theta) = -I(\theta)$$

où

$$V_i(\theta) = \frac{\sum_{j=1}^{n} I(T_j \geq T_i) Z_j Z_j^T \exp(\theta Z_j)}{\sum_{j=1}^{n} I(T_j \geq T_i) \exp(\theta Z_j)} - E_i(\theta) E_i^T(\theta)$$

et $I(\theta)$ représente l'information de Fisher.

Propriétés des estimateurs

Par un développement de Taylor, on a :

$$\frac{\partial \mathcal{L}_n(\hat{\theta})}{\partial \theta} \approx \frac{\partial \mathcal{L}_n(\theta_0)}{\partial \theta} + (\hat{\theta} - \theta_0)\frac{\partial^2 \mathcal{L}_n(\theta_0)}{\partial \theta^2}$$

Or $\partial \mathcal{L}_n(\hat{\theta})/\partial\theta = 0$ et donc,

$$\begin{aligned} 0 &\approx \frac{\partial \mathcal{L}_n(\theta_0)}{\partial \theta} + (\hat{\theta} - \theta_0)\frac{\partial^2 \mathcal{L}_n(\theta_0)}{\partial \theta^2} \\ \sqrt{n}(\hat{\theta} - \theta_0) &\approx \sqrt{n}\left(-\frac{\partial^2 \mathcal{L}_n(\theta_0)}{\partial \theta^2}\right)^{-1}\left(\frac{\partial \mathcal{L}_n(\theta_0)}{\partial \theta}\right) \\ &\approx \left(-\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\theta_0)}{\partial \theta^2}\right)^{-1}\left(\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}_n(\theta_0)}{\partial \theta}\right) \end{aligned}$$

On montre (hors programme) la consistance de $-(1/n)\partial^2 \mathcal{L}_n(\theta_0)/\partial\theta^2$ vers $\Sigma(\theta_0) < \infty$ et la normalité asymptotique du vecteur score : $(1/\sqrt{n})\partial \mathcal{L}_n(\theta_0)/\partial\theta$ converge en loi vers une loi normale centrée et de variance $\Sigma(\theta_0)$.

En conclusion :

$$\sqrt{n}(\hat{\theta} - \theta_0) \longrightarrow \mathcal{N}(0, \Sigma^{-1}(\theta_0))$$

Propriétés des estimateurs

Le processus $\hat{H}_0(t)$ est un estimateur consistant du risque de base cumulé et

$$\sqrt{n}(\hat{H}_0(t) - H_0(t))$$

converge en loi vers un processus centré avec une certaine variance/covariance que l'on peut estimer.

- Cet estimateur est très utile pour estimer la fonction de survie et pour vérifier les hypothèses de validité du modèle.
- La survie conditionnelle est estimée par :

$$\hat{S}(t|Z_{i1} = 0, \ldots, Z_{ip} = 0) = \exp\left(-\hat{H}_0(t)\right)$$

et pour une valeur générale des covariables,

$$\hat{S}(t|Z_{i1}, \ldots, Z_{ip}) = \exp\left(-\hat{H}_0(t)e^{\hat{\theta} Z_i}\right)$$

Comme précédemment, $\hat{S}(t|Z_{i1}, \ldots, Z_{ip})$ est un estimateur consistant de la survie conditionnelle et $\sqrt{n}(\hat{S}(t|Z_{i1}, \ldots, Z_{ip}) - S(t|Z_{i1}, \ldots, Z_{ip}))$ converge en loi vers un processus centré avec une certaine variance/covariance que l'on peut estimer.

Intervalles de confiance et tests pour le paramètre de régression

La variance asymptotique $\Sigma(\theta_0)$ s'estime par $I(\hat{\theta})/n$ et les intervalles de confiance pour θ_0 se déduisent directement de la loi asymptotique de $\hat{\theta}$.

Soit $1 \leq d \leq p$. On veut tester :

(H_0) $\theta_1 = \cdots = \theta_d = 0$ contre (H_1) $\exists j \in \{1, \ldots, d\} : \theta_j \neq 0$.

- Le test du rapport de vraisemblance
- Le test du score
- Le test de Wald

Ces trois tests sont basés sur 3 statistiques de tests différentes qui suivent asymptotiquement, sous (H_0), des lois du χ^2 à d degrés de liberté.

Par exemple, dans le cas $d = 1$, la statistique de test de Wald s'écrit :

$$I(0)\hat{\theta}^2 \longrightarrow \chi^2(1)$$

Retour sur les données PBC

```
fit.pbc<-coxph(Surv(time,status==2) ~ age+factor(edema)+log(bili)
               +log(protime)+log(albumin),data=pbc)
fit.pbc
```

```
## Call:
## coxph(formula = Surv(time, status == 2) ~ age + factor(edema) +
##     log(bili) + log(protime) + log(albumin), data = pbc)
##
##                       coef exp(coef) se(coef)     z       p
## age                0.04045   1.04128  0.00771  5.25 1.6e-07
## factor(edema)0.5   0.28188   1.32563  0.22523  1.25 0.21074
## factor(edema)1     1.01247   2.75240  0.28975  3.49 0.00048
## log(bili)          0.85905   2.36091  0.08324 10.32 < 2e-16
## log(protime)       2.35993  10.59020  0.77314  3.05 0.00227
## log(albumin)      -2.51576   0.08080  0.65262 -3.85 0.00012
##
## Likelihood ratio test=232  on 6 df, p=0
## n= 416, number of events= 160
##    (2 observations deleted due to missingness)
```

Retour sur les données PBC

```
fit.pbc2<-coxph(Surv(time,status==2) ~ age+log(bili)
                +log(protime)+log(albumin),data=pbc)
anova(fit.pbc2,fit.pbc)
```

```
## Analysis of Deviance Table
##  Cox model: response is  Surv(time, status == 2)
##  Model 1: ~ age + log(bili) + log(protime) + log(albumin)
##  Model 2: ~ age + factor(edema) + log(bili) + log(protime) + log(albumin)
##    loglik  Chisq Df P(>|Chi|)
## 1 -756.53
## 2 -751.01 11.031  2  0.004024 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Retour sur les données mélanome

$$h(t|\text{sexe}, \log(\text{tum})) = h_0(t)\exp(\theta_1 \text{sexe} + \theta_2 \log(\text{tum}))$$

```
fit=coxph(Surv(days,status==1)~ factor(sex)+lthick,data=melanoma)
fit
```

```
## Call:
## coxph(formula = Surv(days, status == 1) ~ factor(sex) + lthick,
##     data = melanoma)
##
##                coef exp(coef) se(coef)    z       p
## factor(sex)1 0.458     1.581    0.269 1.70   0.088
## lthick       0.781     2.183    0.157 4.96 6.9e-07
##
## Likelihood ratio test=33.5  on 2 df, p=5.45e-08
## n= 205, number of events= 57
```

Modèle avec interactions

```
fit=coxph(Surv(days,status==1)~ factor(sex)+factor(sex)*lthick,data=melanoma)
```

```
## Call:
## coxph(formula = Surv(days, status == 1) ~ factor(sex) + factor(sex) *
##     lthick, data = melanoma)
##
##                       coef exp(coef) se(coef)     z       p
## factor(sex)1         1.111     3.037    1.817  0.61    0.54
## lthick               0.834     2.302    0.214  3.90 9.8e-05
## factor(sex)1:lthick -0.113     0.893    0.311 -0.36    0.72
##
## Likelihood ratio test=33.6  on 3 df, p=2.43e-07
## n= 205, number of events= 57
```

$$h(t|\text{sex},\text{lt}) = h_0(t)\exp(\theta_1\text{sex} + \theta_2\text{lt} + \theta_3\text{sexlt})$$

$$\theta_1\text{sex} + \theta_2\text{lt} + \theta_3\text{sexlt} = \begin{cases} \theta_1 + (\theta_2+\theta_3)lt, & \text{sex} = 1 \\ \theta_2 lt, & \text{sex} = 0 \end{cases}$$

- Le risque de base est le risque pour qui ?
- Quel est l'effet de l'epaisseur de la tumeur sur les hommes ? Sur les femmes ?
- Y-a-t'il une intéraction entre epaisseur de la tumeur et sexe ?

Validation du modèle

Rappel des hypothèses du modèle

Nous allons essayer de vérifier les hypothèses du modèle de Cox :

- L'hypothèse des risques proportionnels : le rapport de risque entre deux individus n'ayant qu'une covariable qui diffère entre eux est proportionnel au cours du temps. Une façon de vérifier cette hypothèse consiste à vérifier si le paramètre θ_j devrait dépendre du temps.
- L'hypothèse de log-linéarité (difficile à vérifier en pratique).
- La forme des covariables : est-ce que la covariable devrait elle être au carré ? Ou devrait-on prendre son logarithme ? Sa racine carrée ?

L'hypothèse des **risques proportionnels** est l'hypothèse majeure du modèle.

Modèle de Cox stratifié

Si l'hypothèse des risques proportionnels n'est pas vérifiée pour une covariable du modèle on peut, pour s'afranchir de cette hypothèse, **stratifier** le risque de base par rapport à cette covariable. Par exemple, pour les données de mélanome :

$$h(t|sex, ltum, ulc = s) = h_0(t, s)\exp(\theta_1 sex + \theta_2 ltum)$$

autrement dit,

$$h(t|sex, ltum, ulc = 1) = h_0(t, 1)\exp(\theta_1 sex + \theta_2 ltum) \text{ et}$$
$$h(t|sex, ltum, ulc = 0) = h_0(t, 0)\exp(\theta_1 sex + \theta_2 ltum)$$

Le risque de base peut être différent en fonction de l'ulcération mais les paramètres de régression pour les variables sex et tumeur restent les mêmes !

- Ce modèle est utile si il nous semble que l'hypothèse des risques proportionnels n'est pas vérifiée pour la variable ulcération.
- Le problème est que l'on ne peut plus quantifier de manière simple l'effet de l'ulcération sur le risque instantané !

Validation du modèle : l'hypothèse des risques proportionnels

Le modèle de Cox stratifié est utile pour vérifier si l'hypothèse des risques proportionnels est vérifiée pour chaque covariable. On reprend l'exemple des données mélanome.

- Si l'hypothèse des risques proportionnels était vraie, on devrait avoir : $h_0(t,1) = h_0(t,0)\exp(\theta_3)$, avec θ_3 le paramètre correspondant à l'effet de l'ulcération.

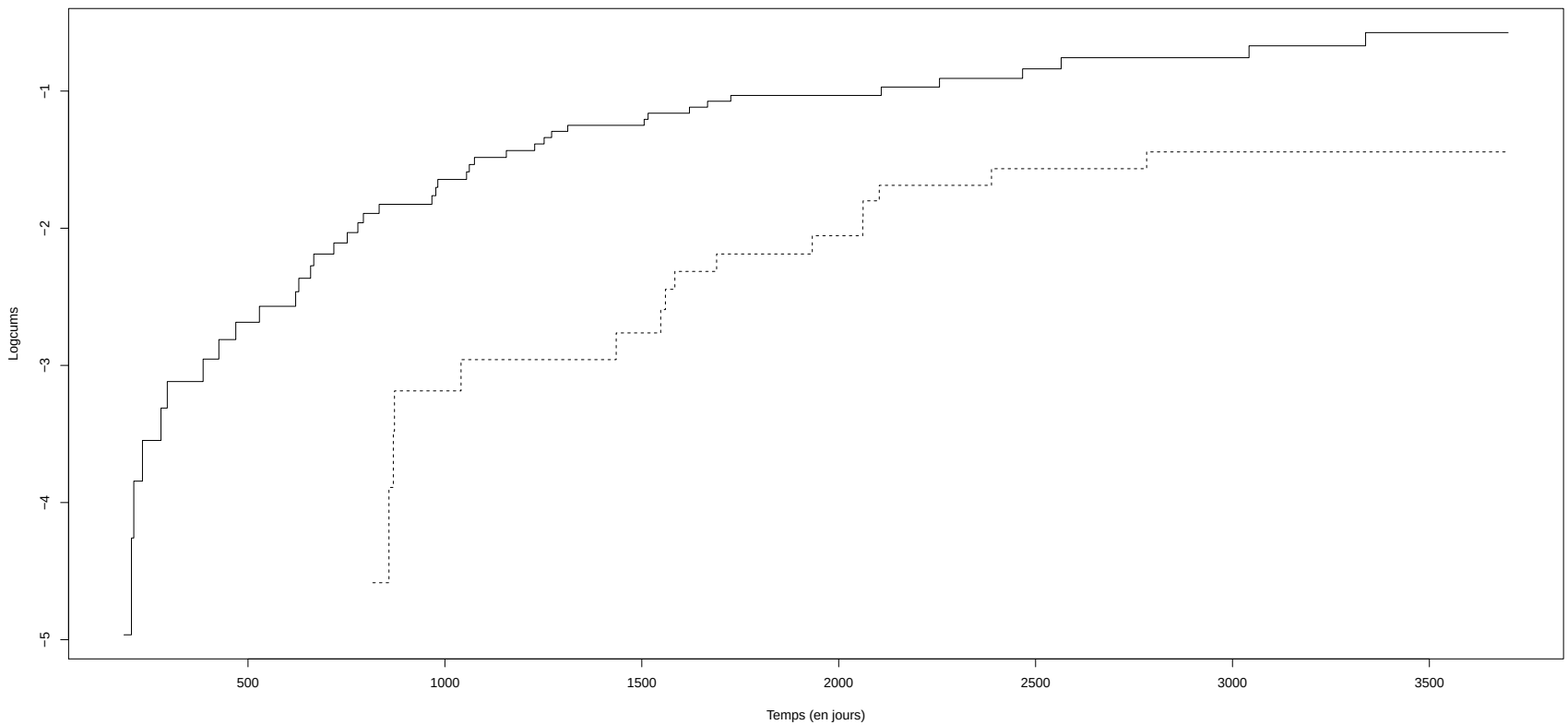
On estime alors dans le modèle stratifié, les risques de bases cumulés $\hat{H}_0(t,1)$ et $\hat{H}_0(t,0)$ et on regarde si ils sont proportionnels ou pas.

- De manière équivalente, on peut estimer les fonctions $\log(-\log(\cdot))$ des survies conditionnelles dans les deux strates et vérifier si ces quantités sont proportionnelles ou pas :

$$\begin{aligned}\log(-\log(\hat{S}(t|sex, ltum, ulc = 1))) &= \log(\hat{H}_0(t,1)) + \hat{\theta}_1 sex + \hat{\theta}_2 ltum \\ &= \log(\hat{H}_0(t,0)) + \hat{\theta}_1 sex + \hat{\theta}_2 ltum + \hat{\theta}_3 \\ &= \log(-\log(\hat{S}(t|sex, ltum, ulc = 0))) + \hat{\theta}_3\end{aligned}$$

Log du hazard cumulé

```
fit=coxph(Surv(days,status==1)~ factor(sex)+lthick+strata(ulc),data=melanoma)
fit.detail=coxph.detail(fit)
logcum1=log(cumsum(fit.detail$hazard[c(17:57)]))
logcum0=log(cumsum(fit.detail$hazard[1:16]))
par(mfrow=c(1,1))
plot(c(fit.detail$time[17:57],3700),c(logcum1,logcum1[41]),type='s',xlab="Temps
lines(c(fit.detail$time[1:16],3700),c(logcum0,logcum0[16]),type='s',lty=2)
```



Tester l'hypothèse des risques proportionnels

- Si l'effet d'une covariable varie au cours du temps, on peut essayer de modéliser les données avec un effet dépendant du temps pour cette variable.

Par exemple, si on veut vérifier l'hypothèse des risques proportionnels sur la variable ulc, on regardera un modèle du type :

$$h(t|sex, ltum, ulc) = h_0(t, s) \exp(\theta_1 sex + \theta_2 ltum + (\theta_{3,1} + \theta_{3,2} g(t)) ulc)$$

- Pour un choix de fonction g on pourra alors tester si $\theta_{3,2} = 0$. Les choix classiques pour la fonction sont l'identité ou la fonction log.
- Le test de $\theta_{3,2} = 0$ s'effectue sous R avec la fonction **cox.zph** pour une fonction donnée de g.

Tester l'hypothèse des risques proportionels

```
fit=coxph(Surv(days/365,status==1)~ factor(sex)+ulc+lthick,data=melanoma)
time.test=cox.zph(fit,transform="log")
time.test
```

```
##                    rho chisq      p
## factor(sex)1 -0.0627 0.230 0.6312
## ulc          -0.1335 0.956 0.3283
## lthick       -0.2942 4.022 0.0449
## GLOBAL            NA 8.773 0.0325
```

Le modèle de Cox est rejeté !

Tester l'hypothèse des risques proportionnels

De manière similaire, il est possible d'implémenter des modèles de Cox où l'effet d'une covariable varie sur des intervalles de temps.

- Par exemple, pour les données mélanome, on peut essayer de voir si l'effet de l'ulcération change après 1400 jours dans l'étude.

$$h(t|sex, ltum, ulc) = h_0(t, s) \exp(\theta_1 sex + \theta_2 ltum + (\theta_{3,1} + \theta_{3,2} I(t > 1400) ulc)$$

- On remarque que

$$(\theta_{3,1} + \theta_{3,2} I(t > 1400)) ulc = \theta_{3,1} ulc + \theta_{3,2} \tilde{Z}(t)$$

avec $\tilde{Z}(t) = I(t > 1400) ulc$.

- Il s'agit donc d'implémenter un modèle de Cox avec une covariable $\tilde{Z}(t)$ qui dépend du temps.
- Ce modèle s'implémente sous R en utilisant la fonction **survSplit**.
- On pourra ensuite tester si $\theta_{3,2} = 0$.

Modèle de Cox avec effet dépendant du temps pour ulc

```
melanoma1=survSplit(melanoma,cut=c(1400),end="days",start="start",event="status
```

```
## Warning in Surv(days, status): Invalid status value, converted to NA
```

```
melanoma1$ulcnew=melanoma1$ulc*as.numeric(melanoma1$days>1400)
fit1=coxph(Surv(start,days,status==1)~ factor(sex)+lthick+factor(ulc)
           + factor(ulcnew), data=melanoma1)
summary(fit1)
```

```
## Call:
## coxph(formula = Surv(start, days, status == 1) ~ factor(sex) +
##     lthick + factor(ulc) + factor(ulcnew), data = melanoma1)
##
##   n= 219, number of events= 57
##    (148 observations deleted due to missingness)
##
##                     coef exp(coef) se(coef)      z Pr(>|z|)
## factor(sex)1      0.2927    1.3401   0.2864  1.022 0.306815
## lthick            0.7806    2.1829   0.2226  3.507 0.000453 ***
## factor(ulc)1      1.5854    4.8813   0.5057  3.135 0.001717 **
## factor(ulcnew)1 -2.4829    0.0835   0.6967 -3.564 0.000366 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Validation du modèle : résumé

- Ces procédures sont les procédures standards de validation du modèle (en anglais “goodness-of-fit”).
- On peut faire des tests contre des déviations spécifiques de l'hypothèse de risques proportionnels en regardant certaines fonctionnelles du temps.
- Ces tests restent difficiles à appliquer en pratique car il faut choisir une fonctionnelle du temps en particulier ou bien choisir un découpage du temps de façon arbitraire. **Différentes fonctionnelles du temps peuvent aboutir à des conclusions contradictoires** !
- Il y a également des méthodes graphiques qui sont utiles en pratique. Cependant, il peut être difficile de réussir à voir si deux courbes sont parallèles ou pas. De plus, ces méthodes ne peuvent pas s'appliquer pour une variable **continue**.
- Un des gros inconvénients de toutes ces méthodes est que quand on teste l'hypothèse de risques proportionnels pour une variable, on doit **supposer que l'hypothèse des risques proportionnels est vérifiée pour toutes les autres covariables** ! Cela peut amener à des résultats surprenants en pratique. . .
- Il existe encore beaucoup d'autres méthodes que nous ne verrons pas ici et qui supposent également que l'hypothèse des risques proportionnels est vérifiée pour toutes les autres covariables du modèle.

Les résidus martingales cumulés

Il existe des méthodes plus récentes de validation du modèle dans le package **timereg**.

- Les résidus martingales cumulés représentent l'erreur entre le modèle et les données cumulée au cours du temps. Ils sont l'équivalent des résidus au modèle de régression linéaire par exemple.
- On peut estimer ces résidus et simuler un grand nombre de trajectoire sous **l'hypothèse que le modèle de Cox est bien vérifié**. On compare alors au résidu observé sur nos données.
- Un test est proposé pour voir si le résidu observé sur nos données est cohérent avec ce que l'on devrait observer si le modèle de Cox est bien vérifié.
- L'avantage de ce test est qu'il n'y a pas besoin de spécifier de **fonctionnelle du temps associé à la covariable**.
- Par contre, ce test a toujours l'inconvénient que quand on teste l'hypothèse des risques proportionnels pour une variable, on doit **supposer que l'hypothèse des risques proportionnels est vérifiée pour toutes les autres covariables** !

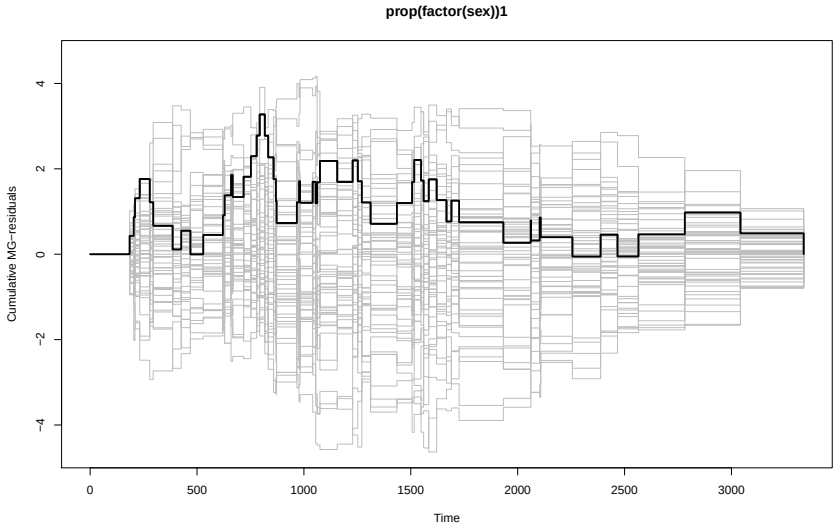
Résidus martingales cumulés (données mélanome)

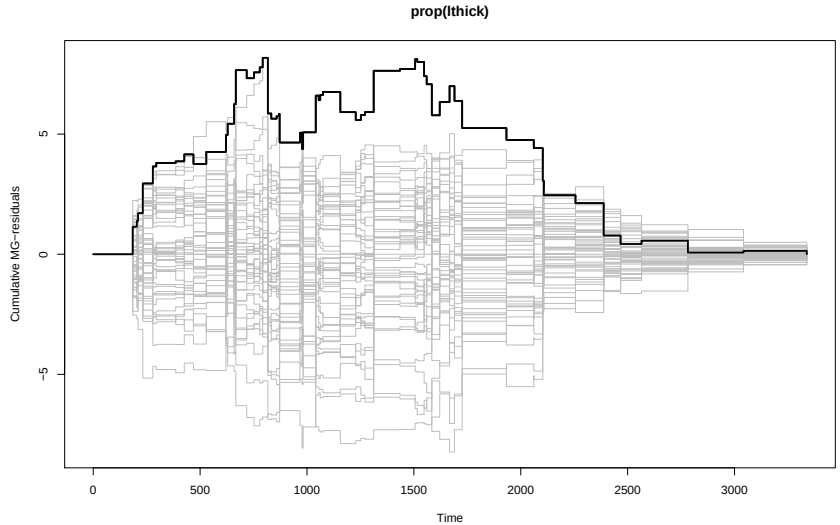
```
library(timereg)
fit=cox.aalen(Surv(days,status==1)~ prop(factor(sex))+prop(lthick)
              +prop(ulc), data=melanoma)
summary(fit)
```

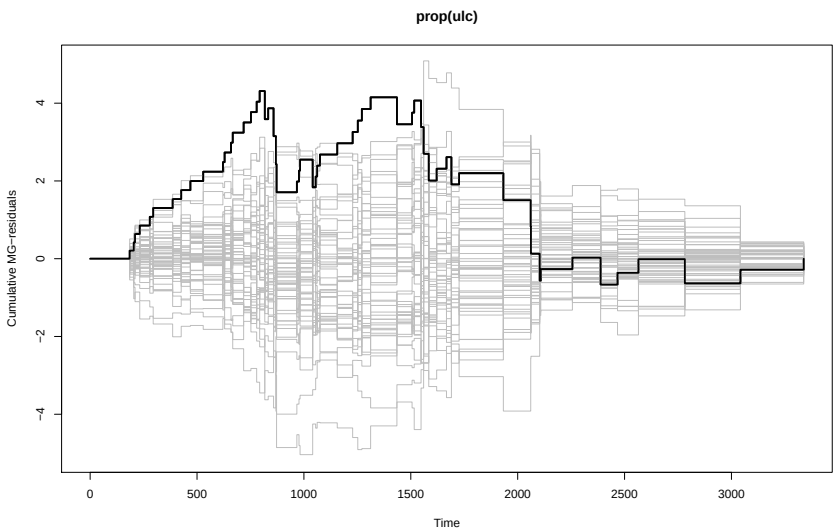
```
## Cox-Aalen Model
##
## Test for Aalen terms
## Test not computed, sim=0
##
## Proportional Cox terms :
##                     Coef.    SE Robust SE D2log(L)^-1    z P-val lower2.5%
## prop(factor(sex))1 0.381 0.274     0.281       0.271 1.36 0.174    -0.156
## prop(lthick)       0.576 0.162     0.172       0.179 3.34 0.001     0.258
## prop(ulc)          0.939 0.315     0.307       0.324 3.06 0.002     0.322
##                     upper97.5%
## prop(factor(sex))1       0.918
## prop(lthick)             0.894
## prop(ulc)                1.560
## Test of Proportionality
##                     sup|  hat U(t) | p-value H_0
## prop(factor(sex))1              3.27        0.312
## prop(lthick)                    8.17        0.028
## prop(ulc)                       4.31        0.038
```

Résidus martingales cumulés (données mélanome)

```
par(mfrow=c(2,2))
plot(fit,score=TRUE)
```







Les résidus martingales cumulés avec la covariable

- Ces résidus sont quasi-identiques aux précédents sauf qu'ils incorporent également la valeur de la covariable. Un test par simulation de trajectoire peut également être implémenté.
- Ils doivent être tracés en fonction de la valeur de la covariable en abscisse.
- Ces résidus ne marchent que pour des variables **continues**.
- Ils permettent de tester la forme de la covariable qui nous intéresse. En utilisant différentes fonctionnelles de la covariable (au carré, racine carré, le logarithme...) on peut apprécier quel est la fonctionnelle la plus vraisemblable dans notre modèle.
- Par contre, ce test a toujours l'inconvénient que quand on teste l'hypothèse des risques proportionnels pour une variable, on doit **supposer que l'hypothèse des risques proportionnels est vérifiée pour toutes les autres covariables** !

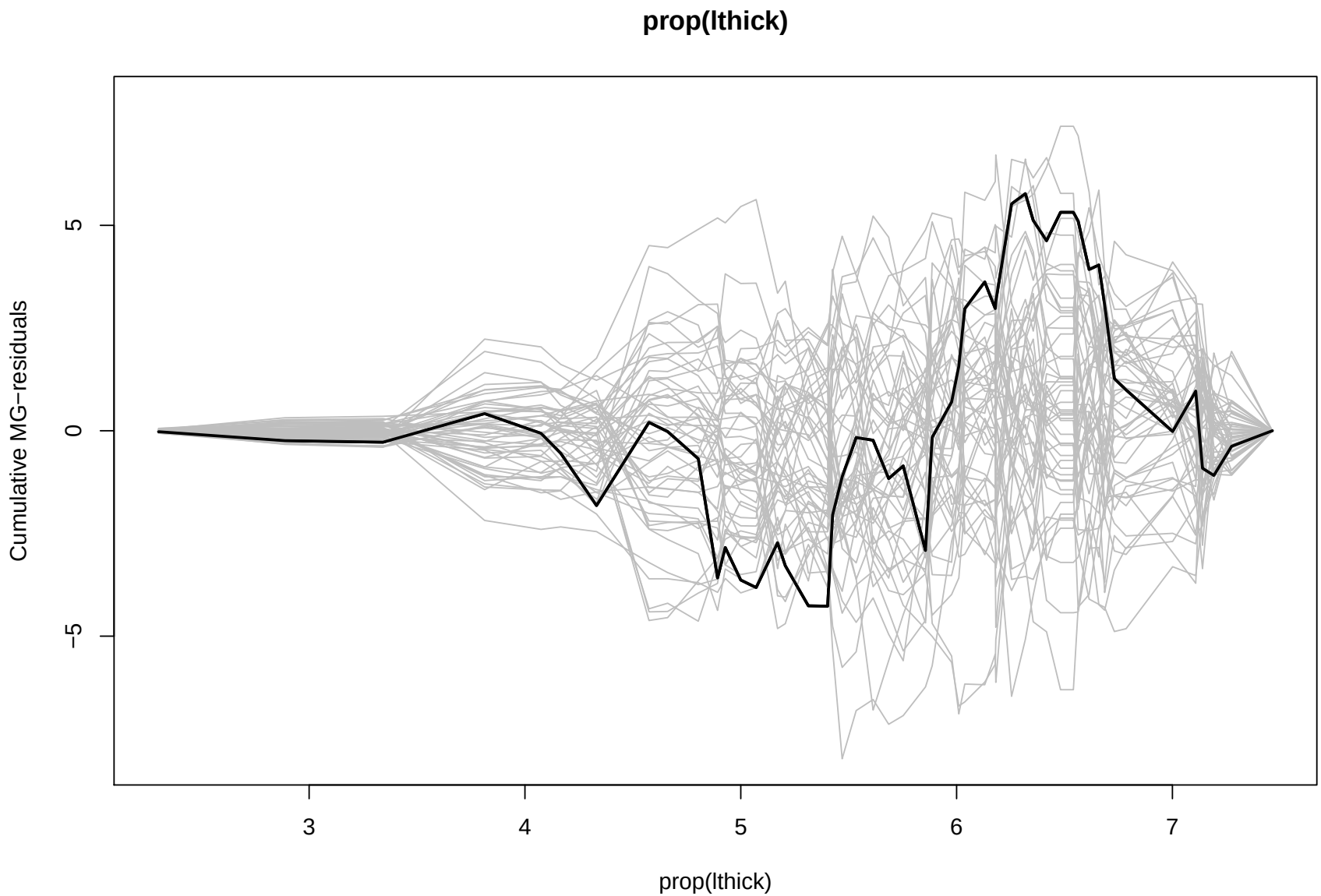
Résidus martingales cumulés avec covariable (mélanome)

```
fit.gof=cox.aalen(Surv(days,status==1)~  prop(lthick),melanoma,residuals=1)
resids=cum.residuals(fit.gof,melanoma,cum.resid=1)
summary(resids)
```

```
## Test for cumulative MG-residuals
##
## Grouped cumulative residuals not computed, you must provide
## modelmatrix to get these (see help)
##
## Residual versus covariates consistent with model
##
##                sup|  hat B(t) | p-value H_0: B(t)=0
## prop(lthick)             5.772                0.176
```

Résidus martingales cumulés avec covariable (mélanome)

```
plot(resids,score=2)
```



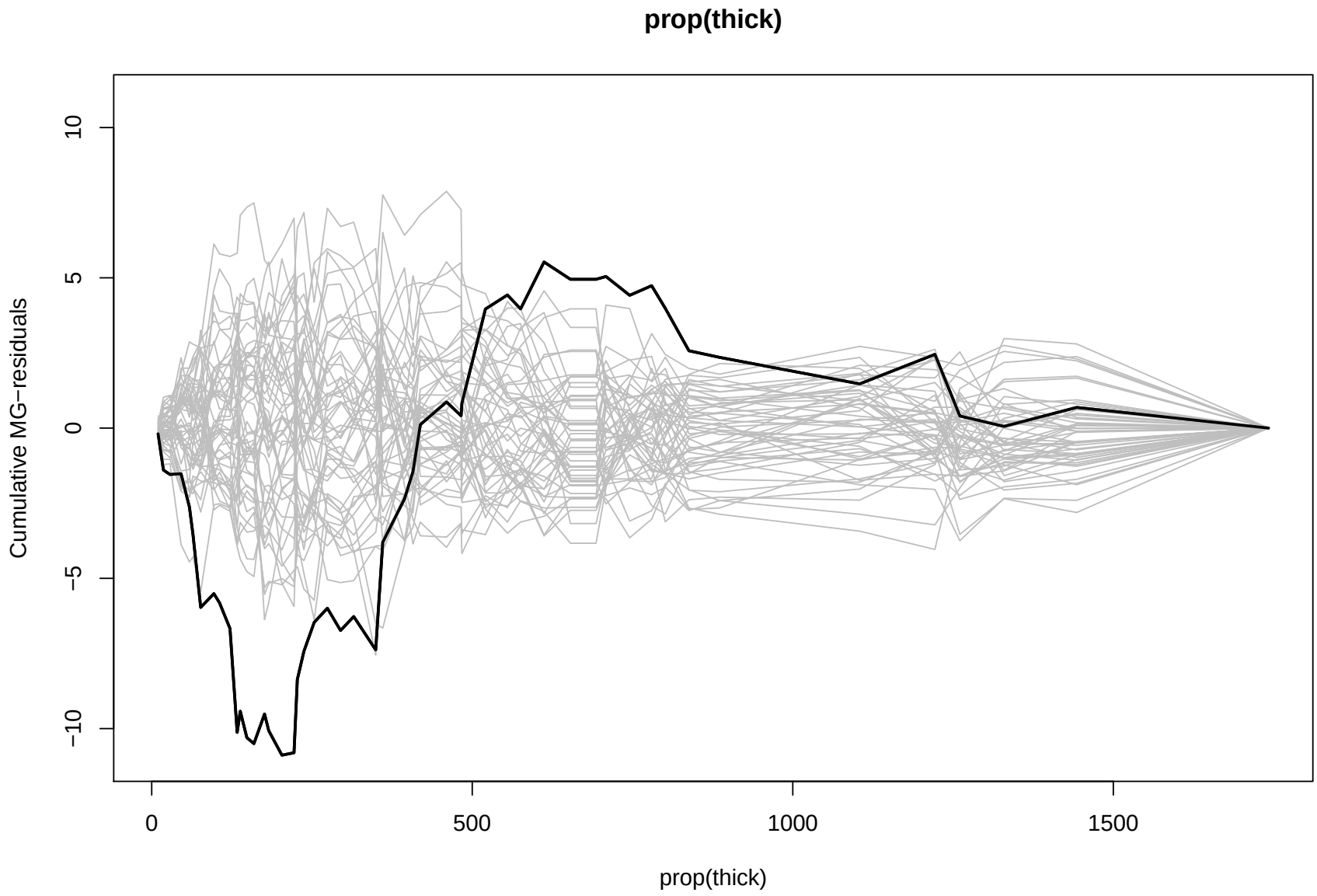
Résidus martingales cumulés avec covariable (mélanome)

```
fit.gof=cox.aalen(Surv(days,status==1)~  prop(thick),melanoma,residuals=1)
resids=cum.residuals(fit.gof,melanoma,cum.resid=1)
summary(resids)
```

```
## Test for cumulative MG-residuals
##
## Grouped cumulative residuals not computed, you must provide
## modelmatrix to get these (see help)
##
## Residual versus covariates consistent with model
##
##              sup|  hat B(t) | p-value H_0: B(t)=0
## prop(thick)           10.885                    0
```

Résidus martingales cumulés avec covariable (mélanome)

```
plot(resids,score=2)
```



Modèle de Cox avec effets dépendant du temps

Le modèle de Cox avec effet des covariables dépendant du temps a été implémenté dans le package **timereg**. Le modèle est le suivant (pour les données mélanome par exemple)

$$h(t|sex, ltum, ulc) = h_0(t) \exp(\theta_1(t)sex + \theta_2(t)ltum + \theta_3(t)ulc)$$

où $\theta_1(t), \theta_2(t)$ et $\theta_3(t)$ dépendent du temps ! On peut estimer les effets cumulés :

$$B_j(t) = \int_0^t \theta_j(s)ds$$

- les tracer, cela permet d'observer comment l'effet de la covariable évolue au cours du temps.
- faire des tests du type

$$(H_0)\ B_j(t) = \text{cste} \times t \text{ contre } (H_0)\ B_j(t) \neq \text{cste} \times t$$

L'avantage de ce modèle est que l'on peut tester si un effet dépend du temps, tout en supposant que les autres effets dépendent également du temps. Ce modèle n'impose donc pas de **supposer que l'hypothèse des risques proportionnels est vérifiée pour toutes les autres covariables du modèle** !

On pourra également regarder des modèles où l'on spécifie certaines variables constantes au cours et d'autres autorisées à varier au cours du temps.

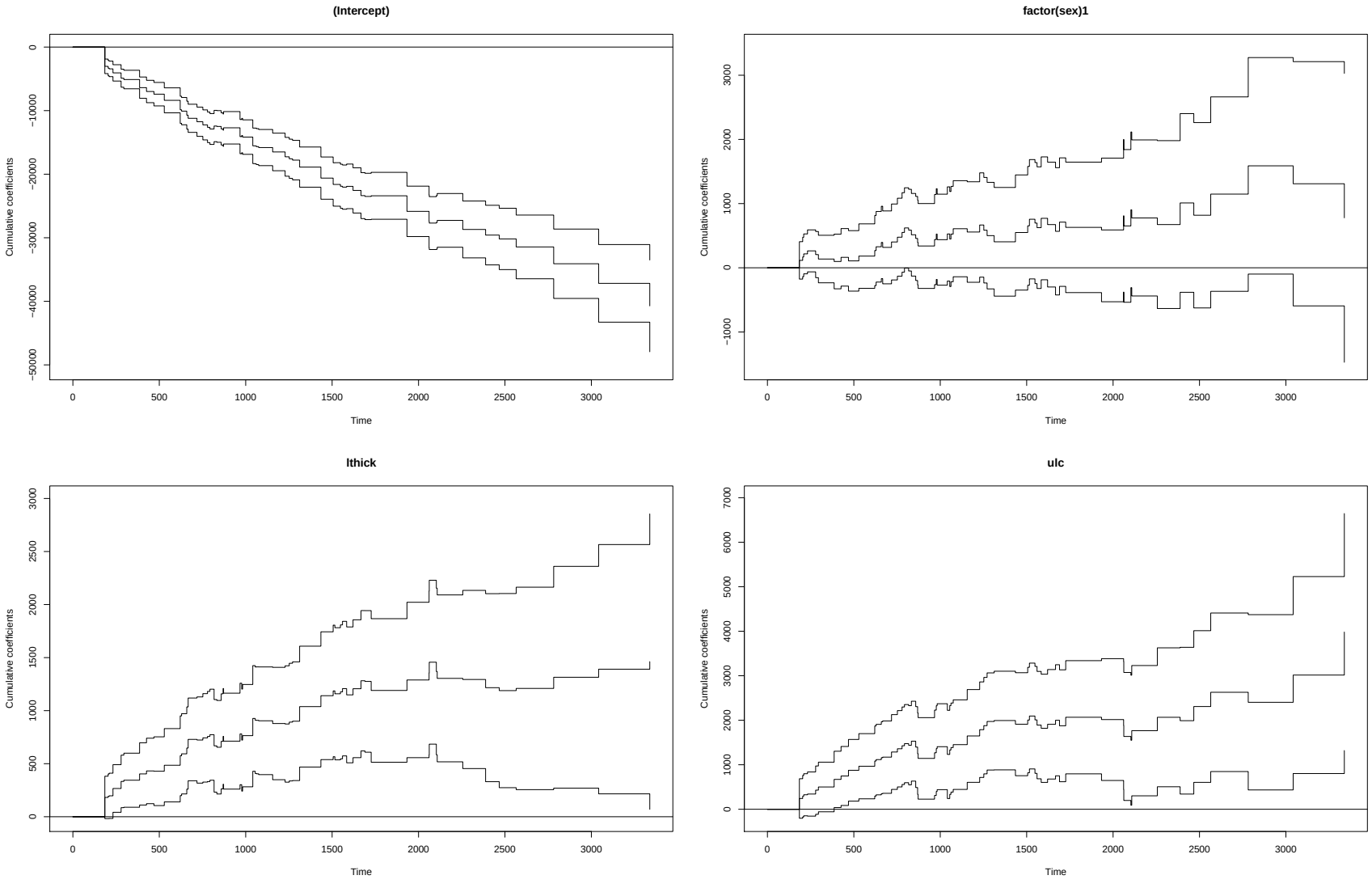
Modèle de Cox avec effet dépendant du temps

```
fittime<-timecox(Surv(days,status==1)~  factor(sex)+lthick+ulc,melanoma)
summary(fittime)
```

```
## Multiplicative Hazard Model
##
## Test for nonparametric terms
##
## Test for non-significant effects
##              Supremum-test of significance p-value H_0: B(t)=0
## (Intercept)                          17.50                0.000
## factor(sex)1                          3.10                0.032
## lthick                                5.95                0.000
## ulc                                   7.59                0.000
##
## Test for time invariant effects
##                   Kolmogorov-Smirnov test p-value H_0:constant effect
## (Intercept)                         3240                        0.008
## factor(sex)1                         940                        0.376
## lthick                               553                        0.023
## ulc                                  958                        0.290
##                     Cramer von Mises test p-value H_0:constant effect
## (Intercept)                     1.20e+10                        0.002
## factor(sex)1                    4.67e+08                        0.495
## lthick                          3.35e+08                        0.006
## ulc                             7.04e+08                        0.261
```

Modèle de Cox avec effet dépendant du temps

```
par(mfrow=c(2,2))
plot(fittime)
```



Modèle de Cox avec effet dépendant du temps

```
fittime<-timecox(Surv(days,status==1)~  const(factor(sex))+lthick+ulc,melanoma)
```

```
## Multiplicative Hazard Model
##
## Test for nonparametric terms
##
## Test for non-significant effects
##             Supremum-test of significance p-value H_0: B(t)=0
## (Intercept)                         16.20                   0
## lthick                               5.76                   0
## ulc                                  8.50                   0
##
## Test for time invariant effects
##                   Kolmogorov-Smirnov test p-value H_0:constant effect
## (Intercept)                         4880                         0.067
## lthick                               717                         0.088
## ulc                                 1240                         0.390
##                     Cramer von Mises test p-value H_0:constant effect
## (Intercept)                     2.22e+10                         0.039
## lthick                          5.21e+08                         0.048
## ulc                             1.23e+09                         0.276
##
## Parametric terms :
##                      Coef.    SE Robust SE   z P-val lower2.5% upper97.5%
## const(factor(sex))1 0.417 0.266     0.297 1.4  0.16    -0.104      0.938
```

Conclusion

- Le modèle de Cox est un modèle facile à interpréter et qui permet de communiquer facilement les résultats.
- Il fait de grosses hypothèses mais reste robuste en pratique.
- Si l'hypothèse des risques proportionnels n'est pas valide en pratique, il se peut que l'effet d'une covariable ne soit pas visible dans le modèle de Cox ou soit atténué.
- Il existe de nombreuses méthodes de validation du modèle. La plupart sont difficiles à utiliser en pratique car elles supposent que le modèle est vrai pour toutes les variables sauf celle que l'on évalue. On recommande d'utiliser les techniques récentes du package **timereg** qui n'ont pas cet inconvénient.
- Il existe d'autres extensions du modèle de Cox : le modèle s'applique bien quand on a des covariables qui dépendent du temps, quand on des données tronquées à gauche, quand on étudie des évènements récurrents. Il s'utilise également dans les modèles à risques compétitifs/modèles multi-états, dans les modèles à fragilité où l'on veut prendre en compte une certaine hétérogénéité au sein des individus...
- Il existe d'autres modèles qui font d'autres hypothèses sur les données : le modèle AFT("Accelerated Failure Time model"), le modèle de Aalen...
- Le modèle de Cox reste le modèle le plus utilisé en biostatistique et dans le domaine médical.