

Analyses de Survie

Plan du cours

Analyses de survie

1. Données de survie
2. Fonctions de densité, de survie, et de risque
3. Estimateur de Kaplan-Meier de la fonction de survie
4. Test d'une différence de survie entre plusieurs échantillons
5. Modèle de Cox

Plan du cours

Analyses de survie

1. Données de survie
2. Fonctions de densité, de survie, et de risque
3. Estimateur de Kaplan-Meier de la fonction de survie
4. Test d'une différence de survie entre plusieurs échantillons
5. Modèle de Cox

Qu'est-ce qu'on analyse ?

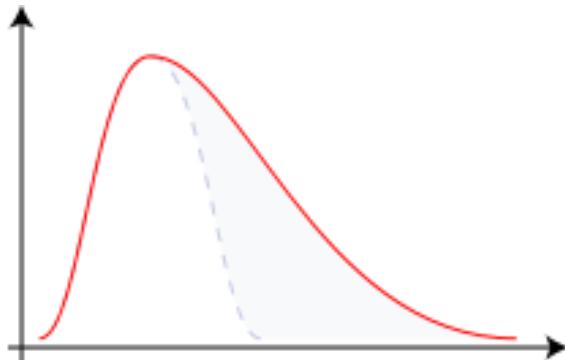
La particularité de cette branche de la statistique est que la variable Y à analyser (variable réponse ou à expliquer) correspond à la durée d'un processus ou à une durée avant la survenue d'un événement :

- Durée de survie de patients ayant eu un infractus du myocarde
- Le temps mis par des rats pour sortir d'un labyrinthe
- L'âge d'entrée dans une schizophrénie
- Durée avant un échec de fonctionnement de moteurs de voiture
- Durée avant une récidive d'anciens détenu de prison
- Durée de mariage
- L'âge d'entrée dans la vie active
- Etc.

Données de survie

Si de telles durées sont d'authentiques variables aléatoires continues, elles présentent néanmoins deux caractéristiques particulières :

- Valeurs uniquement positives et par conséquent la variable Y présente généralement une forte assymétrie positive de sa distribution qui s'écarte fortement de la loi Normale

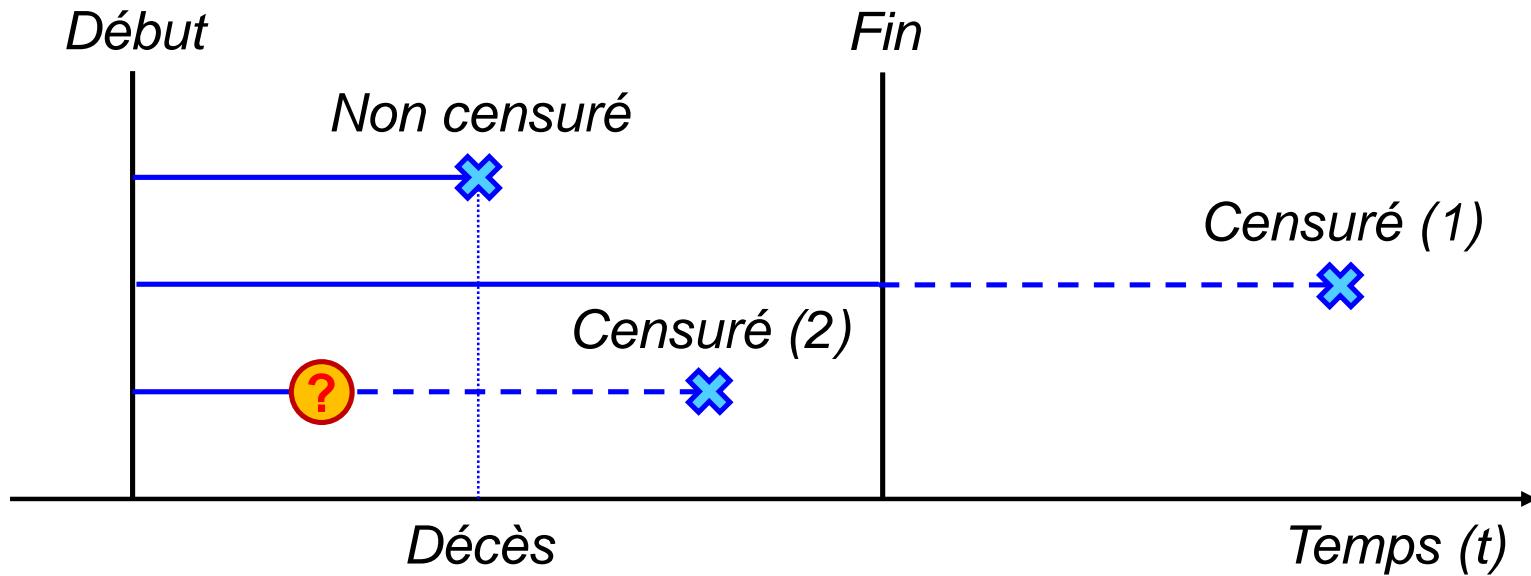


- La variable Y n'est pas forcément observée ou mesurée pendant le laps de temps de l'étude, on dit alors que la variable Y est "censurée"

Exemples de censures

Prenons le cas de la durée de vie de patients atteints de cancers :

- À la fin du laps de temps prévu par l'étude, l'événement étudié qui est le décès n'intervient pas obligatoirement chez tous les patients suivis et pour de tels patients, la durée de vie n'est pas exactement connue (1)
- Au cours du laps de temps prévu par l'étude, le suivi de certains patients peut-être interrompu pour plusieurs raisons indépendantes à l'étude (2)



L'objet de survie

Deux paramètres importants constituent la variable Y :

- Le temps (seconde, jour, semaine, mois, année, etc.) : variable continue
- La survenue ou non de l'évenement étudié : variable discrète codée 1 pour la survenue de l'évenement et 0 quand la donnée est censurée

Sous R ces deux paramètres constituent l'objet de survie et sont gérés par la fonction "Surv()" du package "survival" :

```
> library()
> chooseCRANmirror()
> install.packages("survival")
> library(survival)
> help(package="survival")
> timeOfEvent <- c(3, 6, 6, 8, 9, 10, 14, 16, 17, 18)
> event <- c(1, 1, 1, 0, 1, 1, 0, 1, 1, 1)
> Surv(timeOfEvent, event)
[1] 3   6   6   8+  9   10  14+ 16   17   18
```

Plan du cours

Analyses de survie

1. Données de survie
2. Fonctions de densité, de survie, et de risque
3. Estimateur de Kaplan-Meier de la fonction de survie
4. Test d'une différence de survie entre plusieurs échantillons
5. Modèle de Cox

La fonction f de densité

Soit $t (t > 0)$ le temps et $\mu (\mu > 0)$ la durée moyenne de survie d'un individu, alors le modèle paramétrique de survie le plus simple correspond à un modèle exponentiel dont la fonction de densité est :

$$f(t) = \frac{1}{\mu} e^{-\frac{1}{\mu}t}$$

$$NB : f(t) = \lim_{\Delta t \rightarrow 0^+} \frac{P(t < T \leq t + \Delta t)}{\Delta t} = \frac{dF(t)}{dt}$$

Avec T , la variable aléatoire continue et positive qui représente le temps de survie d'un individus par exemple

Du point de vue du biologiste ou de l'écogiste, la fonction f de densité correspond à la proportion de décès entre t et $t + \Delta t$ rapporté au nombre total d'individus à l'instant initial t_0

La fonction F de répartition

Soit F la fonction de répartition associée à cette fonction de densité :

$$F(t) = P(T \leq t)$$

$$F(t) = \int_{-\infty}^t f(x)dx = \int_{-\infty}^0 f(x)dx + \int_0^t f(x)dx = \int_0^t f(x)dx$$

Car T est positive

$$F(t) = \int_0^t \frac{1}{\mu} e^{-\frac{1}{\mu}x} dx$$

$$F(t) = \left[-e^{-\frac{t}{\mu}} - \left(-e^{-\frac{0}{\mu}} \right) \right]$$

$$F(t) = 1 - e^{-\frac{1}{\mu}t}$$

Rappel : Pour une variable aléatoire continue positive comme le temps ou la durée de survie (T), la fonction F de répartition correspond à l'aire sous la courbe de la fonction f de densité

La fonction S de survie

Soit S la fonction de survie :

$$S(t) = P(T > t) = 1 - P(T \leq t)$$

$$S(t) = 1 - F(t) = 1 - \left(1 - e^{-\frac{1}{\mu}t}\right)$$

$$S(t) = e^{-\frac{1}{\mu}t}$$

$$NB : S(t) = \int_t^{+\infty} f(x)dx$$

$$NB : S(0) = 1$$

$$NB : S(+\infty) = 0$$

La fonction S de survie tout comme la fonction F de répartition correspond à une probabilité, mais cette fois il s'agit de la probabilité qu'à un individu de survivre jusqu'à l'instant t

La fonction h de risque

Soit h la fonction de risque :

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{1}{\mu} e^{-\frac{1}{\mu}t}}{e^{-\frac{1}{\mu}t}}$$

$$h(t) = \frac{1}{\mu}$$

$$NB : h(t) = \lim_{\Delta t \rightarrow 0^+} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t}$$

Du point de vue du biologiste ou de l'écogiste, la fonction h de risque correspond au taux de mortalité instantané entre t et $t + \Delta t$ sachant que le temps de survie T est supérieur à t

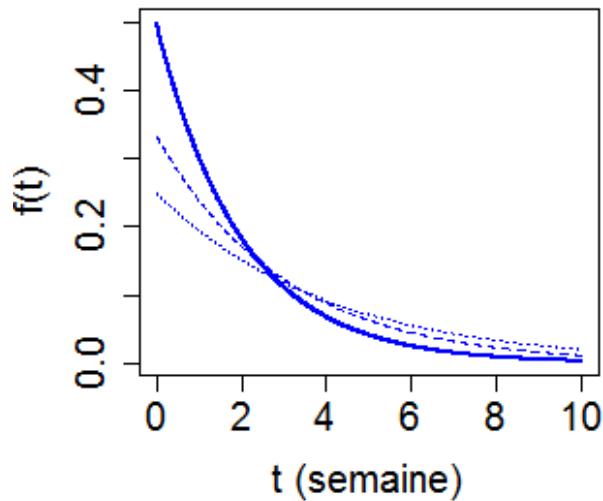
Exemple

A partir de ce modèle exponentiel simple dont la fonction h de risque est constante au cours du temps t , tracez les courbes des fonctions de densité, de répartition, de survie, et de risque pour un organisme dont la durée moyenne de survie est égale à 2 semaines ($\mu = 2$) :

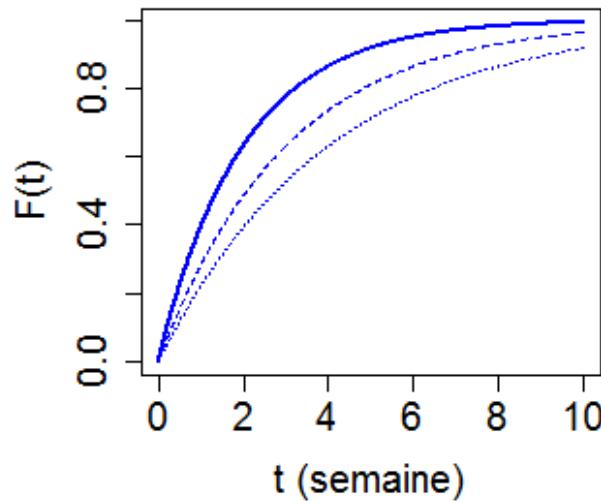
```
> t <- c(seq(0, 10, 0.01))
> f <- (1/2) * (exp(-(1/2)*t))
> F <- 1-exp(-(1/2)*t)
> S <- exp(-(1/2)*t)
> h <- f/S
> par(mfrow=c(2, 2))
> plot(t, f, type="l", col="blue", lwd=2, ylab=c("f(t)"),
main="Fonction de densité")
> plot(t, F, type="l", col="blue", lwd=2, ylab=c("F(t)"),
main="Fonction de répartition")
> plot(t, S, type="l", col="blue", lwd=2, ylab=c("S(t)"),
main="Fonction de survie")
> plot(t, h, type="l", col="blue", lwd=2, ylab=c("h(t)"),
main="Fonction de risque")
```

Représentation graphique

Fonction de densité

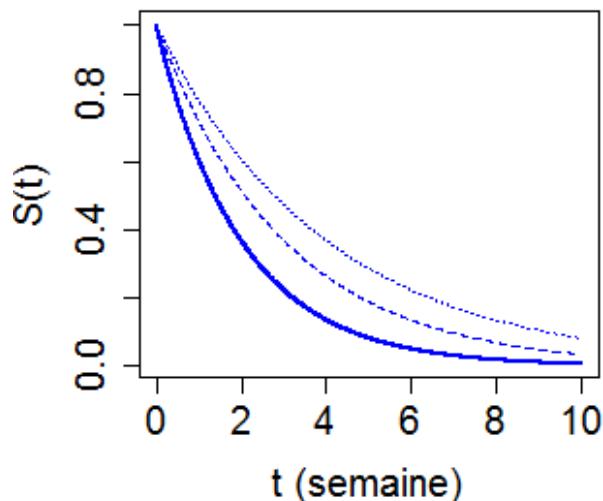


Fonction de répartition

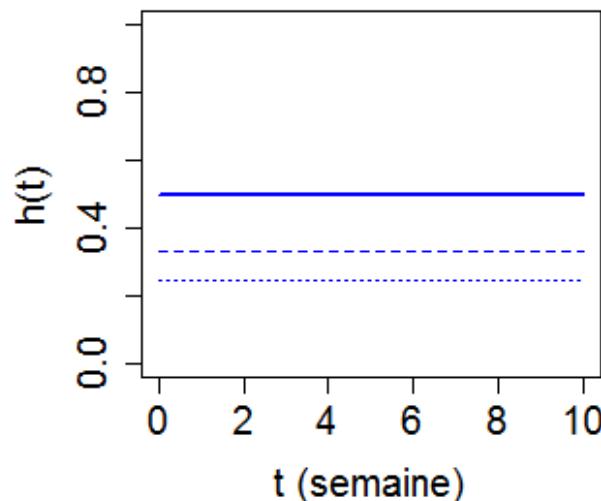


— $\mu = 2$ semaines
- - - $\mu = 3$ semaines
- · - $\mu = 4$ semaines

Fonction de survie



Fonction de risque



Réalisme du modèle exponentiel simple

Attention, le modèle exponentiel présenté précédemment, dont la fonction h de risque est constante au cours du temps est peu réaliste dans le cadre du suivi de la survie des organismes biologiques :

- Pour l'être humain, le risque de décès (h) est relativement faible pendant la période infantile et croît de manière exponentielle avec l'âge pour atteindre un risque maximum chez les personnes âgées
- Chez les salmonidés par exemple, le risque de décès (h) est au contraire à son maximum au début de leurs cycles de vie et tend à décroître avec l'âge

Les modèles de Weibull, de Gompertz, ou de Makeham sont en général utilisés avec des paramètres soit positifs (e.g., être humain) soit négatifs (e.g., salmonidés) pour refléter ces tendances

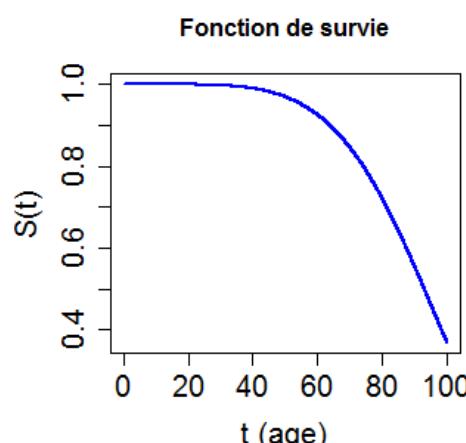
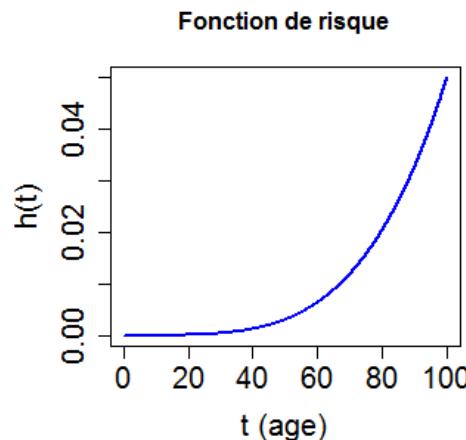
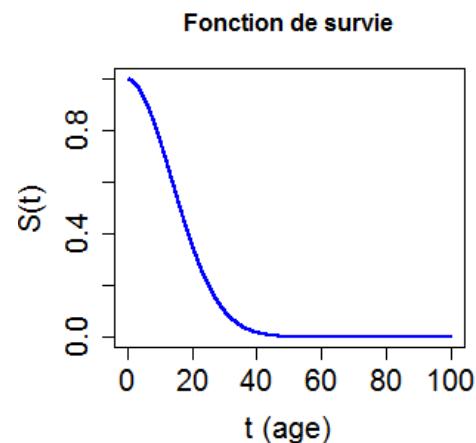
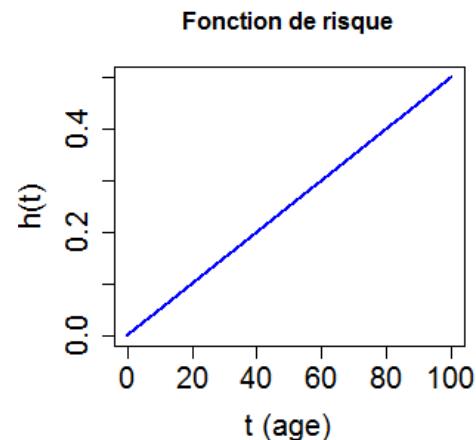
Chez l'homme, le modèle de Makeham est le meilleur car il permet, après la naissance, de prédire un risque constant (cf. accidents et suicides) qui croît exponentiellement ensuite

Exemples de modèles : Rayleigh et Weibull

Selon les modèles dont la fonction h de risque n'est pas constante au cours du temps, l'impact sur la fonction S de survie sera différent :

$$(1) h(t) = 0.001 + 0.005 * t$$

$$(2) h(t) = 5 * 0.01 * (0.01 * t)^{5-1}$$



(1) Modèle de Rayleigh

$$h(t) = a + bt$$

(2) Modèle de Weibull

$$h(t) = \alpha \lambda (\lambda t)^{\alpha-1}$$

NB : La fonction H de risque cumulé

On sait que la fonction h de risque admet comme égalité :

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{dF(t)}{dt}}{\frac{S(t)}{dt}} = \frac{\frac{d(1 - S(t))}{dt}}{S(t)} = \frac{-\frac{dS(t)}{dt}}{S(t)} = -\frac{dS(t)}{S(t)dt}$$

$$h(t)dt = -\frac{dS(t)}{S(t)}$$

$$\int_0^t h(x)dx = - \int_0^t \frac{dS(x)}{S(x)} dx$$

$$\int_0^t h(x)dx = -[\log(S(t)) - \log(S(0))] = -[\log(S(t)) - \log(1)] = -\log(S(t))$$

$$H(t) = -\log(S(t))$$

$$NB : S(t) = e^{-H(t)} = e^{-\int_0^t h(x)dx}$$

Exercice

Un volontaire au tableau pour calculer la formule de la fonction S de survie si la fonction h de risque suit un modèle de Weibull ?

$$h(t) = \alpha\lambda(\lambda t)^{\alpha-1}$$

$$S(t) = e^{-\int_0^t h(x)dx}$$

$$\int_0^t h(x)dx = \int_0^t \alpha\lambda(\lambda x)^{\alpha-1}dx = \int_0^t \alpha\lambda^\alpha x^{\alpha-1}dx$$

$$\int_0^t h(x)dx = \left[\frac{\alpha\lambda^\alpha t^\alpha}{\alpha} - \frac{\alpha\lambda^\alpha 0^\alpha}{\alpha} \right] = (\lambda t)^\alpha$$

$$S(t) = e^{-(\lambda t)^\alpha}$$

Plan du cours

Analyses de survie

1. Données de survie
2. Fonctions de densité, de survie, et de risque
3. Estimateur de Kaplan-Meier de la fonction de survie
4. Test d'une différence de survie entre plusieurs échantillons
5. Modèle de Cox

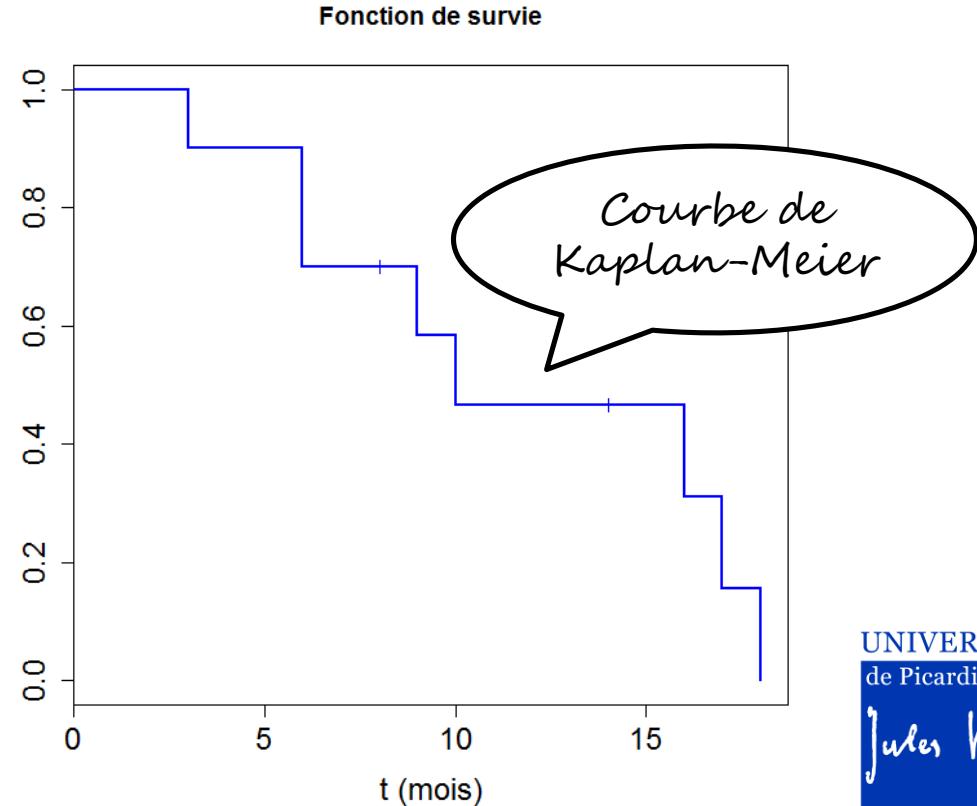
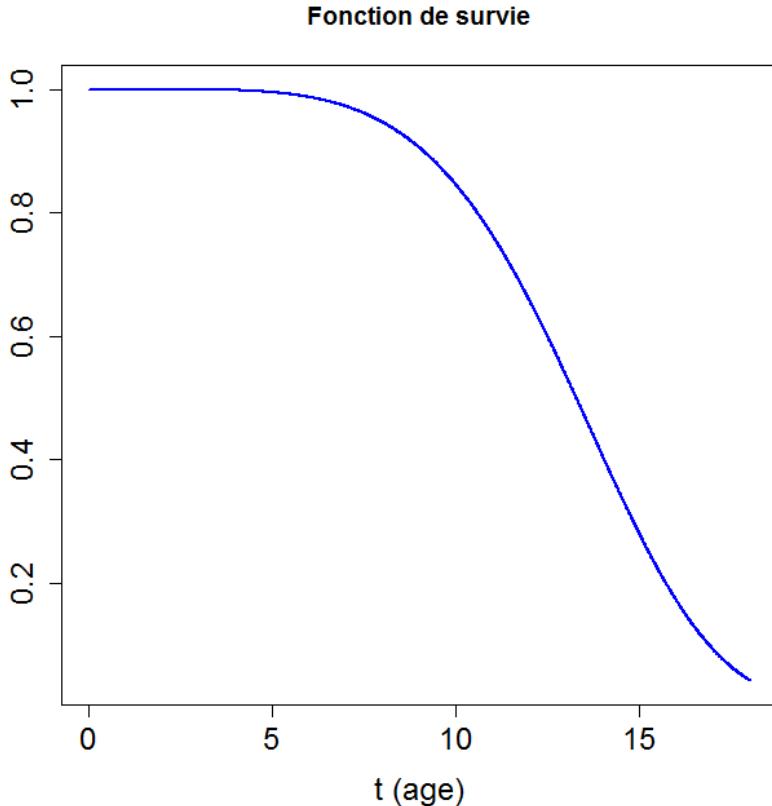
Estimation de la fonction S de survie

En pratique, on estime la fonction S de survie à partir des données :

$$S(t) = P(T > t)$$

Théorie : $S(t)$

Pratique : $\hat{S}(t)$



Estimateur de Kaplan-Meier

Formule de Kaplan-Meier pour le calcul de l'estimation de la fonction S de survie :

$$\hat{S}(t_j) = \prod_{i=1}^j P(T > t_j | T \geq t_j) = \prod_{i=1}^j \frac{n_i - d_i}{n_i}$$

n_i : nb obs restantes non censurées (e.g., nb survivants) juste avant t_i

d_i : nb événements (e.g., nb décès) observés à l'instant t_i

$$\hat{S}(t_j) = \frac{n_j - d_j}{n_j} * \prod_{i=1}^{j-1} \frac{n_i - d_i}{n_i}$$

$$\hat{S}(t_j) = \frac{n_j - d_j}{n_j} * \hat{S}(t_{j-1})$$

Proportion de survivants à t_j

Exemple à partir de données fictives

Soit le jeu de données suivant correspondant au temps (semaine) avant la survenue d'un événement (e.g., mort des individus) :

```
> Surv(timeOfEvent, event)
[1] 3   6   6   8+  9   10  14+ 16  17  18
```

t_j	n_j	d_j	q_j
0	10	0	0
3	10	1	0
6	9	2	1
9	6	1	0
10	5	1	1
16	3	1	0
17	2	1	0
18	1	1	0

$\hat{S}(t_j)$

$$10/10 = 1$$

$$1 * (10 - 1)/10 = 0.9$$

$$0.9 * (9 - 2)/9 = 0.7$$

$$0.7 * (6 - 1)/6 = 0.58$$

$$0.58 * (5 - 1)/5 = 0.47$$

$$0.47 * (3 - 1)/3 = 0.31$$

$$0.31 * (2 - 1)/2 = 0.16$$

$$0.16 * (1 - 1)/1 = 0$$

Proportion de survivants à t_j

q_j : nombre d'observations censurées entre t_j et t_{j+1}

Estimation de la variance par Greenwood

Formule de Greenwood pour estimer la variance de la fonction S de survie :

$$\hat{\sigma}^2(\hat{S}(t_j)) = \hat{S}(t_j)^2 \sum_{i=1}^j \frac{d_i}{n_i(n_i - d_i)}$$

n_i : nb obs restantes non censurées (e.g., nb survivants) juste avant t_i

d_i : nb événements (e.g., nb décès) observés à l'instant t_i

Formule de calcul de l'intervalle de confiance à 95% :

$$IC_{95\%}[S(t_j)] = \hat{S}(t_j) \pm 1.96 * \hat{\sigma}(\hat{S}(t_j))$$

Retour sur l'exemple

A partir de l'estimation de la fonction S de survie et des données, on calcul l'estimation de la variance et donc l'écart type :

t_j	n_j	d_j	q_j	$S(t_j)$
0	10	0	0	1
3	10	1	0	0.9
6	9	2	1	0.7
9	6	1	0	0.58
10	5	1	1	0.47
16	3	1	0	0.31
17	2	1	0	0.16
18	1	1	0	0

$$\widehat{\sigma}(\widehat{S}(t_j))$$

$$0$$

$$\sqrt{0.9^2 * \frac{1}{10 * (10 - 1)}}$$

$$\sqrt{0.7^2 * \left(\frac{1}{10 * (10 - 1)} + \frac{2}{9 * (9 - 2)} \right)}$$

etc.

A partir de ces estimations il est possible de calculer l'intervalle de confiance à 95% autour de la fonction S de survie

R le fait pour vous

Utilisez les fonctions "survfit()" et "summary()" dans R pour obtenir l'estimation de la fonction S de survie par la formule de Kaplan-Meier :

```
> Y <- Surv(timeOfEvent, event)
> summary(survfit(Y~1, conf.type="plain"))
Call: survfit(formula = Y ~ 1, conf.type = "plain")

time n.risk n.event survival std.err lower 95% CI upper 95% CI
  3      10      1     0.900  0.0949      0.714    1.000
  6       9      2     0.700  0.1449      0.416    0.984
  9       6      1     0.583  0.1610      0.268    0.899
 10      5      1     0.467  0.1658      0.142    0.792
 16      3      1     0.311  0.1684      0.000    0.641
 17      2      1     0.156  0.1385      0.000    0.427
 18      1      1     0.000        NaN      NaN    NaN
```

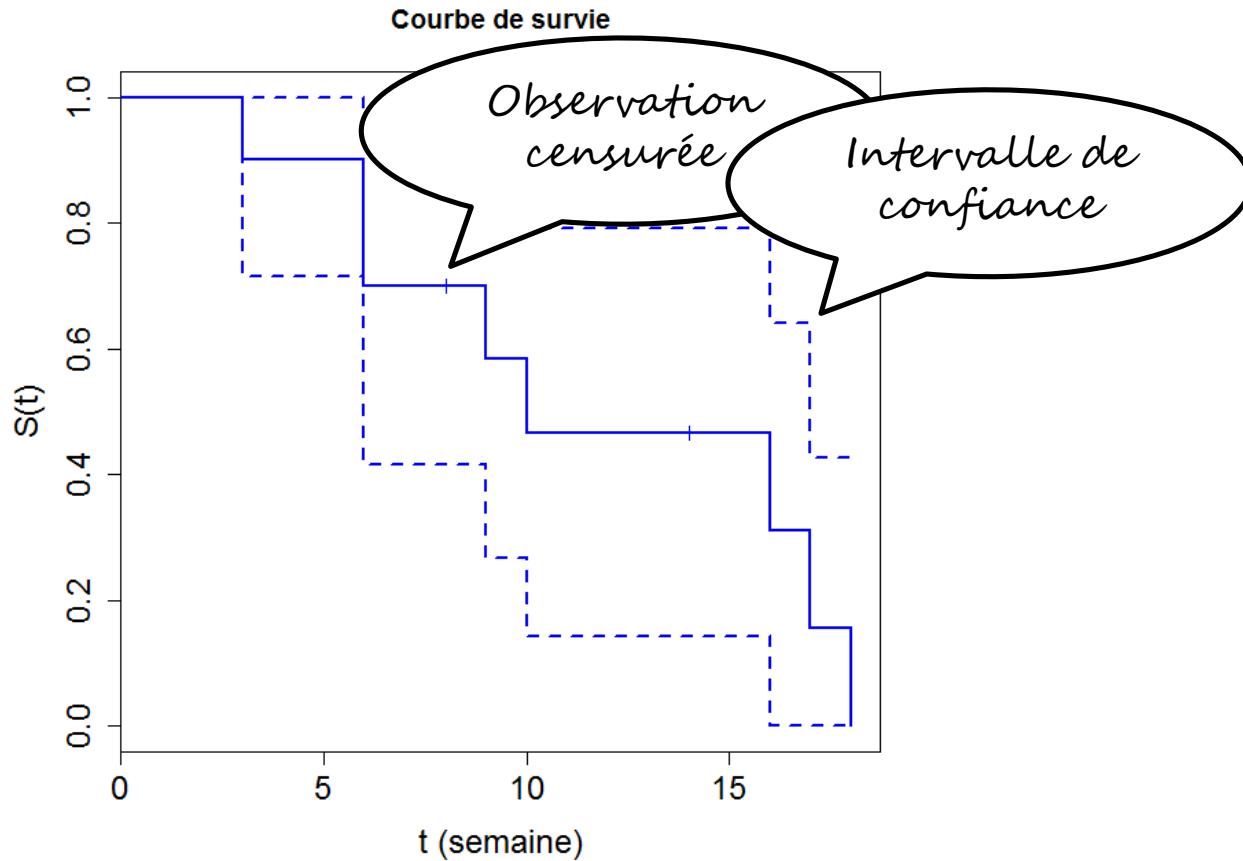
$$(1) IC_{95\%}[S(t_j)] = \hat{S}(t_j) \pm 1.96 * \hat{\sigma}(\hat{S}(t_j))$$

Intervalle de
confiance calculé
à partir de la
formule (1)

Tracé de la courbe de Kaplan-Meier

Utilisez la fonction "plot()" dans R pour afficher la courbe de survie de Kaplan-Meier :

```
> plot(survfit(Y~1, conf.type="plain"), col="blue", lwd=2)
```

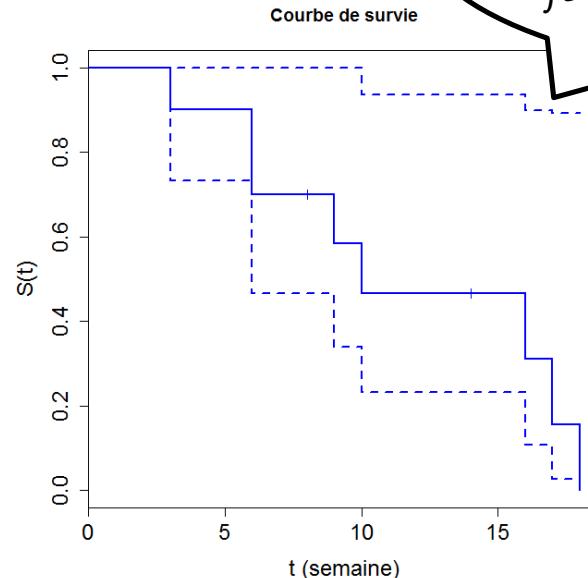
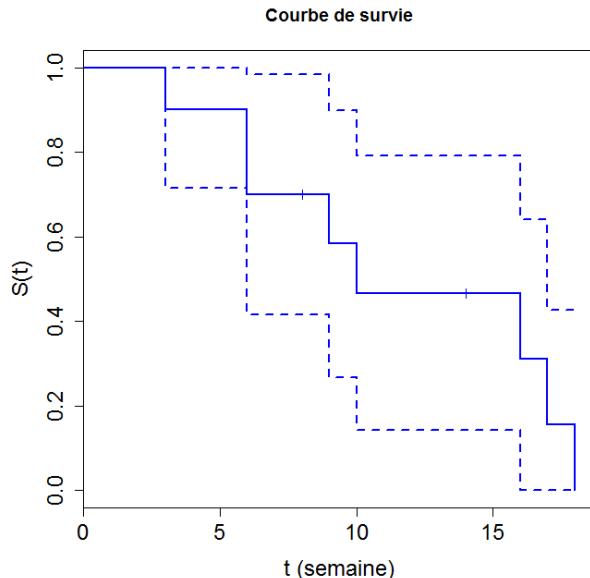


Calcul de l'intervalle de confiance

NB : Par défaut, R calcule un intervalle de confiance à 95% de type "log" calculé sur le *log* de la fonction *S* de survie et qui donne une meilleure estimation de l'intervalle de confiance de la fonction *S* de survie :

$$IC_{95\%}[S(t_j)] = e^{\log(\hat{S}(t_j)) \pm 1.96 * \hat{\sigma}(\log(\hat{S}(t_j)))}$$

$$\hat{\sigma}^2(\log(\hat{S}(t_j))) = \sum_{i=1}^j \frac{d_i}{n_i(n_i - d_i)}$$



Intervalle de confiance calculé sur le *log* de la fonction *S* de survie

Exercice

Soit le temps T (semaine) de rémission chez deux groupes de patients atteints de leucémie :

- Le premier groupe est soumis à un traitement

```
> temps1 <- c(6, 6, 6, 7, 10, 13, 16, 22, 23, 6, 9, 10, 11, 17,  
19, 20, 25, 32, 32, 34, 35)  
> status1 <- c(1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0,  
0, 0, 0, 0)
```

- Le second groupe est soumis à un placebo

```
> temps2 <- c(1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11,  
12, 12, 15, 17, 22, 23)  
> status2 <- c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,  
1, 1, 1, 1, 1)
```

Dans un tableur Excel, créez un tableau à trois colonnes : "temps", "status", et "groupe" afin de pouvoir y rentrer les données ci-dessus

Exercice

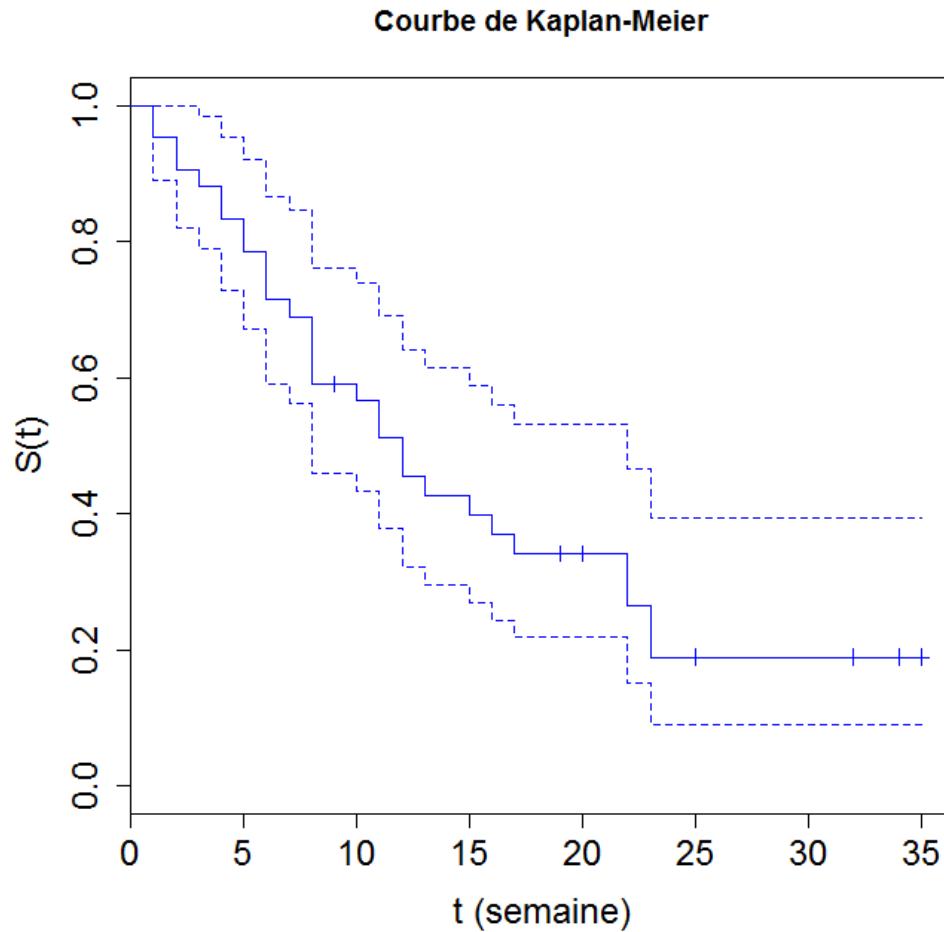
Enregistrez votre fichier au format texte ".txt" avec des séparateurs de type tabulation et utilisez la fonction "read.table()" pour importer ce fichier de données dans R :

```
> data.leu <- read.table("Chap5_Analyses_Survie.txt",
header=TRUE, sep="\t")
> str(data.leu)
'data.frame': 42 obs. of 3 variables:
 $ temps : int 6 6 6 7 10 13 16 22 23 6 ...
 $ statut: int 1 1 1 1 1 1 1 1 1 0 ...
 $ groupe: int 1 1 1 1 1 1 1 1 1 1 ...
> data.leu$temps <- as.numeric(data.leu$temps)
> data.leu$statut <- as.numeric(data.leu$statut)
> data.leu$groupe <- as.factor(data.leu$groupe)
> str(data.leu)
'data.frame': 42 obs. of 3 variables:
 $ temps : num 6 6 6 7 10 13 16 22 23 6 ...
 $ statut: num 1 1 1 1 1 1 1 1 1 0 ...
 $ groupe: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1
```

Exercice

Tracez la courbe de Kaplan-Meier sur l'ensemble des données :

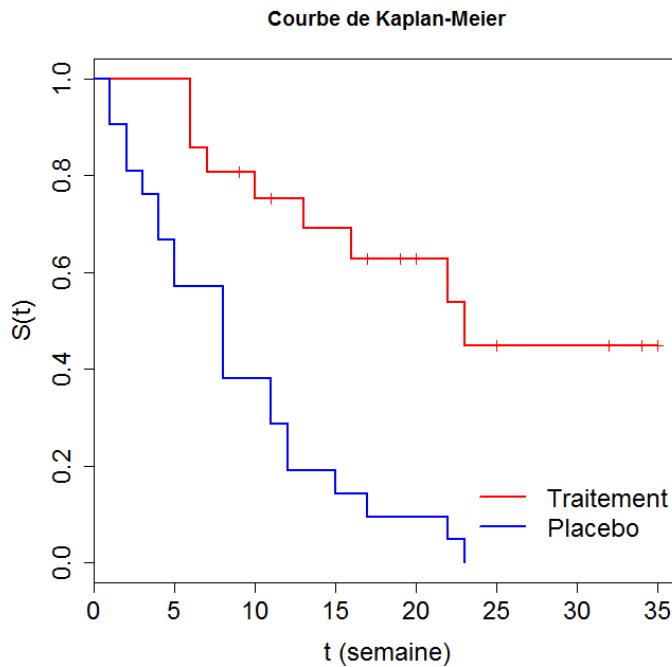
```
> plot(survfit(Surv(data.leu$temps, data.leu$statut)~1))
```



Exercice

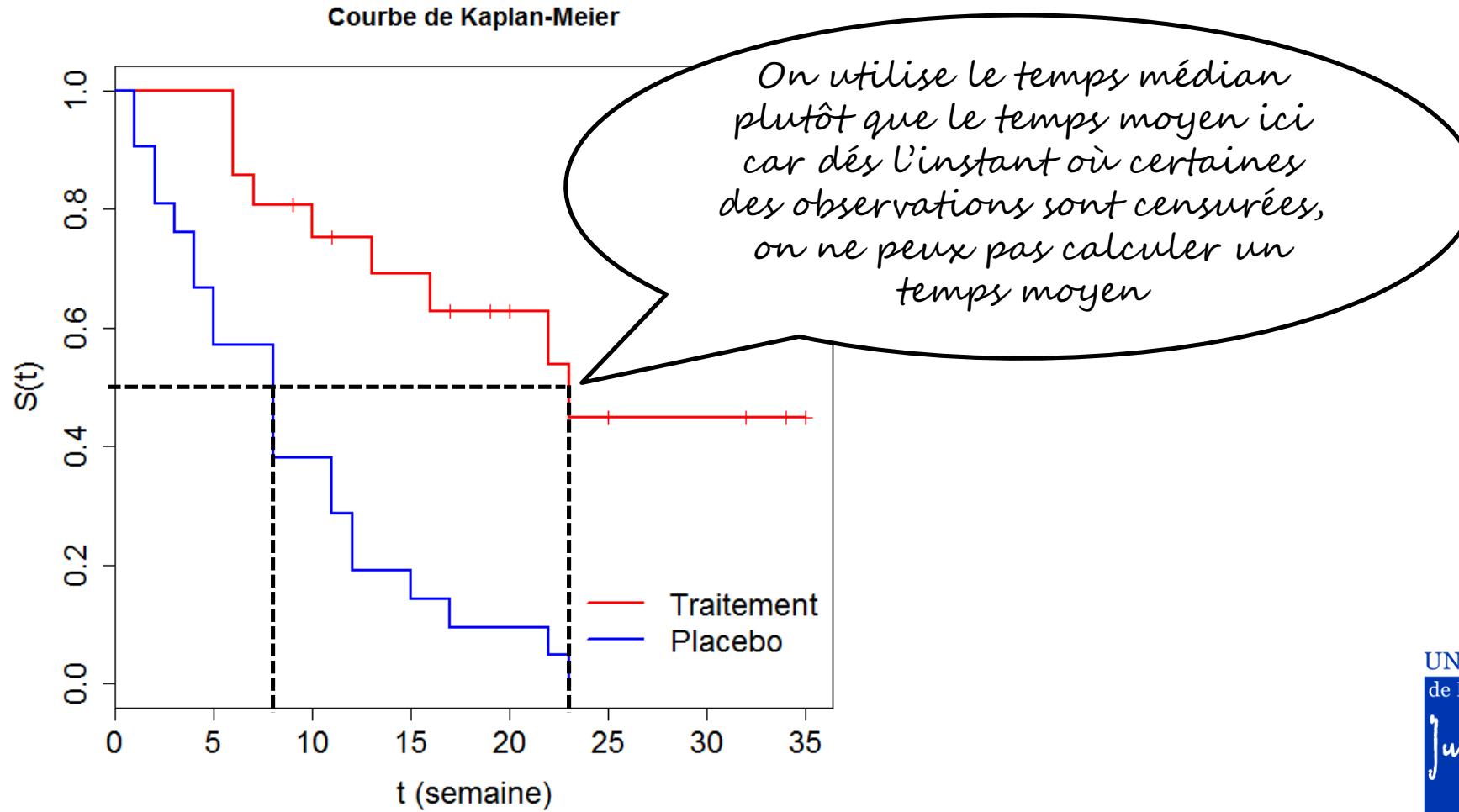
Tracez les courbes de Kaplan-Meier pour chacun des deux groupes :

```
> t1 <- data.leu[which(data.leu$groupe==1), "temps"]
> s1 <- data.leu[which(data.leu$groupe==1), "statut"]
> t2 <- data.leu[which(data.leu$groupe==2), "temps"]
> s2 <- data.leu[which(data.leu$groupe==2), "statut"]
> plot(survfit(Surv(t1, s1)~1), col="red", conf.int=FALSE)
> lines(survfit(Surv(t2, s2)~1), col="blue", conf.int=FALSE)
```



Exercice

Par lecture graphique, donnez le temps médian (médiane) de rémission des patients atteints de leucémie dans chacun des deux groupes :



Exercice

Utilisez la fonction "survfit()" pour obtenir le temps médian :

```
> survfit(Surv(t1, s1)~1)
Call: survfit(formula = Surv(t1, s1) ~ 1)

records    n.max n.start   events   median 0.95LCL 0.95UCL
      21        21       21        9        23        16        NA
```

```
> survfit(Surv(t2, s2)~1)
Call: survfit(formula = Surv(t2, s2) ~ 1)
```

```
records    n.max n.start   events   median 0.95LCL 0.95UCL
      21        21       21       21        8         4        12
```

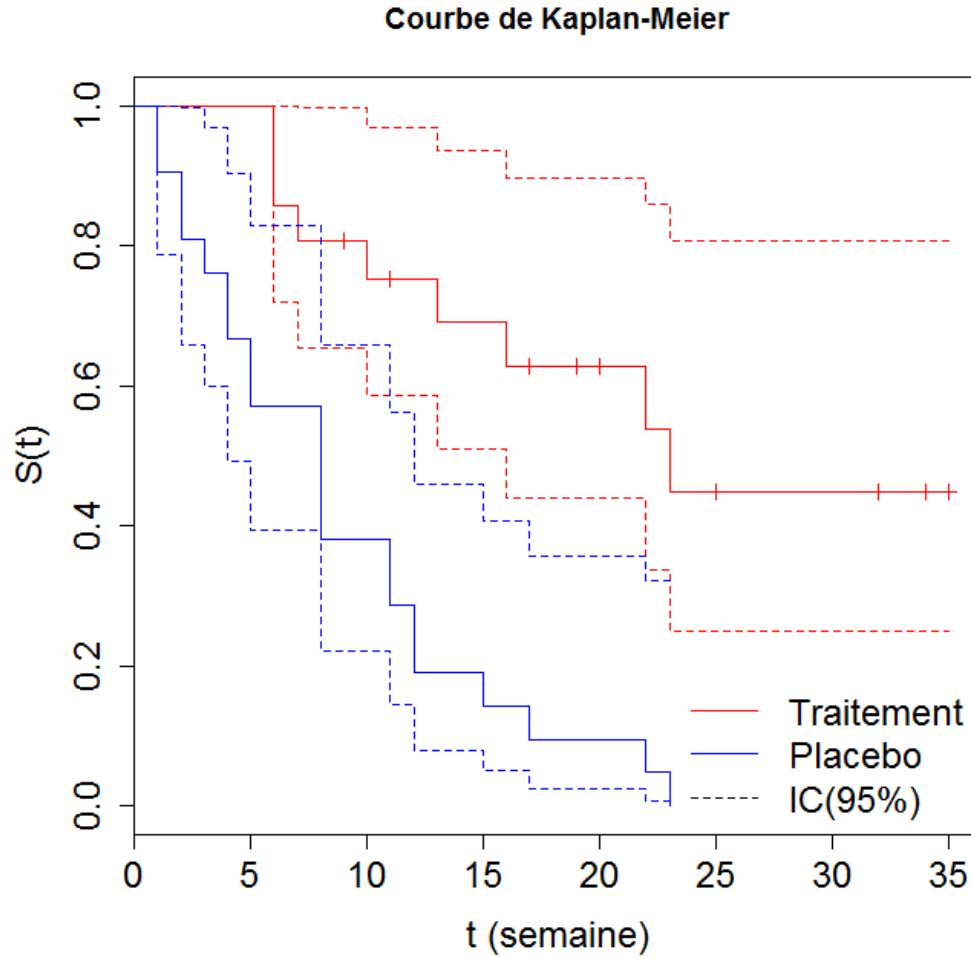
```
> attach(data.leu)
> survfit(Surv(temp, statut)~groupe)
Call: survfit(formula = Surv(temp, statut,
```

Est-ce que cette différence est significative ?

	records	n.max	n.start	events	median	0.95LCL	0.95UCL
groupe=1	21	21	21	9	23	16	
groupe=2	21	21	21	21	8	4	

Exercice

Pourriez-vous conclure sur la significativité de cette différence entre les deux groupes par une simple lecture graphique des IC ?



Plan du cours

Analyses de survie

1. Données de survie
2. Fonctions de densité, de survie, et de risque
3. Estimateur de Kaplan-Meier de la fonction de survie
4. Test d'une différence de survie entre plusieurs échantillons
5. Modèle de Cox

Comparer 2 fonctions S de survie

Il existe plusieurs tests pour comparer les fonctions S de survies de deux échantillons (e.g., 2 groupes de patients), dont deux principaux :

- Le test de **Mantel-Haenszel** encore appelé test du **log-rank** qui est le plus utilisé, le plus simple et le plus performant lorsque les deux courbes de survie ne se croisent pas
- Le test de **Wilcoxon**

Quelques soit le test utilisé, les hypothèses restent les mêmes :

- H₀ : pas de différence de survie entre les deux groupes étudiés
- H₁ : différence de survie entre les deux groupes étudiés

Comparer 2 fonctions S de survie

On réalise ces deux tests à partir des j tables de contingence chacune détaillant pour chaque groupe k le nombre d'événements observés à l'instant t_j (d_{jk}) parmi le nombre d'observations restantes et non censurées juste avant t_j (n_{jk}) :

Évenements à t_j	Groupe 1	Groupe 2	Total
Nb réalisés à t_j	d_{j1}	d_{j2}	d_j
Nb non réalisés à t_j	$n_{j1}-d_{j1}$	$n_{j2}-d_{j2}$	n_j-d_j
Total juste avant t_j	n_{j1}	n_{j2}	n_j

Exemple :

n_j : nb total de survivants juste avant t_j

d_j : nb total de décès observés à l'instant t_j

n_{jk} : nb survivants juste avant t_j dans le groupe k

d_{jk} : nb décès observés à l'instant t_j pour le groupe k

On répète
l'opération
pour chaque
instant t_j

Nombres d'observations attendues ?

À partir de chaque table de contingence obtenue pour l'instant t_j , on calcul le nombre d'observations attendus sous l'hypothèse nulle H0 d'égalité des fonctions S de survie entre les deux groupes étudiés :

Évenements à t_j	Groupe 1	Groupe 2	Total
Nb réalisés à t_j	e_{j1}	e_{j2}	d_j
Nb non réalisés à t_j	$n_{j1} - e_{j1}$	$n_{j2} - e_{j2}$	$n_j - d_j$
Total à t_j	n_{j1}	n_{j2}	n_j

e_{jk} : nb décès attendus sous H0 à l'instant t_j pour le groupe k

$$e_{jk} = \frac{d_j n_{jk}}{n_j} \quad e_{j1} = \frac{d_j n_{j1}}{n_j} \quad e_{j2} = \frac{d_j n_{j2}}{n_j}$$

Sous H0, l'estimation de la variance des e_{jk} s'obtient par la formule :

$$\hat{\sigma}^2(e_{jk}) = \hat{\sigma}^2(e_{j1}) = \hat{\sigma}^2(e_{j2}) = \frac{n_{j1} n_{j2} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$$

Critère utilisé pour le test

À partir des m instants t_j ($1 \leq j \leq m$), on calcul le critère Q_{obs} fonction des observations qui sera comparé à sa valeur critique Q_{crit} :

$$Q_1 = \frac{(\sum_{j=1}^m d_{j1} - \sum_{j=1}^m e_{j1})^2}{\sum_{j=1}^m \hat{\sigma}^2(e_{j1})} \quad Q_2 = \frac{(\sum_{j=1}^m d_{j2} - \sum_{j=1}^m e_{j2})^2}{\sum_{j=1}^m \hat{\sigma}^2(e_{j2})}$$

$$Q_{obs} = Q_1 = Q_2$$

$$Q_{crit} = Chi^2_{1,1-\frac{\alpha}{2}}$$

Se référer à
la table des
quantiles du
 Chi^2

On rejette H_0 si :

$$Q_{obs} > Q_{crit}$$

Retour sur les patients atteints de leucémie

j	t_j	d_{j1}	d_{j2}	n_{j1}	n_{j2}	e_{j1}	e_{j2}
1	1	0	2	21	21	1	1
2	2	0	2	21	19	1.05	0.95
3	3	0	1	21	17	0.55	0.45
4	4	0	2	21	16	1.14	0.86
5	5	0	2	21	14	1.2	0.8
6	6	3	0	21	12	1.91	1.01
7	7	1	0	17	12	0.59	0.41
8	8	0	4	16	12	2.29	1.71
9	10	1	0	15	8	0.65	0.35
10	11	0	2	13	8	1.24	0.76
11	12	0	2	12	6	1.33	0.67
12	13	1	0	12	4	0.75	0.25
13	15	0	1	11	4	0.73	0.27
14	16	1	0	11	3	0.79	0.21
15	17	0	1	10	3	0.77	0.23
16	22	1	1	7	2	1.56	0.44
17	23	1	1	6	1	1.71	0.29

t_1	G1	G2	Total
Nb rémission	2*21/42	2*21/42	2
Nb non rémission	40*21/42	40*21/42	40
Total	21	21	42

t_2	G1	G2	Total
Nb rémission	2*21/40	2*19/40	2
Nb non rémission	38*21/40	38*19/40	38
Total	21	19	40

t_3	G1	G2	Total
Nb rémission	1*21/38	1*17/38	1
Nb non rémission	37*21/38	37*17/38	37
Total	21	17	38

$$e_{jk} = \frac{d_j n_{jk}}{n_j}$$

Retour sur les patients atteints de leucémie

j	t_j	d_{j1}	d_{j2}	n_{j1}	n_{j2}	e_{j1}	e_{j2}	σ^2
1	1	0	2	21	21	1	1	0.49
2	2	0	2	21	19	1.05	0.95	0.49
3	3	0	1	21	17	0.55	0.45	0.25
4	4	0	2	21	16	1.14	0.86	0.48
5	5	0	2	21	14	1.2	0.8	0.47
6	6	3	0	21	12	1.91	1.01	0.65
7	7	1	0	17	12	0.59	0.41	0.24
8	8	0	4	16	12	2.29	1.71	0.87
9	10	1	0	15	8	0.65	0.35	0.23
10	11	0	2	13	8	1.24	0.76	0.45
11	12	0	2	12	6	1.33	0.67	0.42
12	13	1	0	12	4	0.75	0.25	0.19
13	15	0	1	11	4	0.73	0.27	0.2
14	16	1	0	11	3	0.79	0.21	0.17
15	17	0	1	10	3	0.77	0.23	0.18
16	22	1	1	7	2	1.56	0.44	0.3
17	23	1	1	6	1	1.71	0.29	0.2

$$\hat{\sigma}^2(e_{1k}) = \frac{21 * 21 * 2 * (42 - 2)}{42^2 * (42 - 1)}$$

$$\hat{\sigma}^2(e_{2k}) = \frac{21 * 19 * 2 * (40 - 2)}{40^2 * (40 - 1)}$$

$$\hat{\sigma}^2(e_{3k}) = \frac{21 * 17 * 1 * (38 - 1)}{38^2 * (38 - 1)}$$

$$\hat{\sigma}^2(e_{jk}) = \frac{n_{j1}n_{j2}d_j(n_j - d_j)}{n_j^2(n_j - 1)}$$

Retour sur les patients atteints de leucémie

j	t_j	d_{j1}	d_{j2}	n_{j1}	n_{j2}	e_{j1}	e_{j2}	σ^2
1	1	0	2	21	21	1	1	0.49
2	2	0	2	21	19	1.05	0.95	0.49
3	3	0	1	21	17	0.55	0.45	0.25
...
15	17	0	1	10	3	0.77	0.23	0.18
16	22	1	1	7	2	1.56	0.44	0.3
17	23	1	1	6	1	1.71	0.29	0.2
Sommes :		9	21			19.25	10.75	6.26

$$Q_1 = \frac{(\sum_{j=1}^m d_{j1} - \sum_{j=1}^m e_{j1})^2}{\sum_{j=1}^m \hat{\sigma}^2(e_{j1})}$$

$$Q_1 = \frac{(9 - 19.25)^2}{6.26}$$

$$Q_1 = 16.79$$

$$Q_{crit} = ???$$

$$Q_2 = \frac{(\sum_{j=1}^m d_{j2} - \sum_{j=1}^m e_{j2})^2}{\sum_{j=1}^m \hat{\sigma}^2(e_{j2})}$$

$$Q_2 = \frac{(21 - 10.75)^2}{6.26}$$

$$Q_2 = 16.79$$

Une ligne de code dans R !

Utilisez la fonction "survdiff()" dans R qui par défaut réalise un test du log-rank avec H1 bilatéral :

```
> survdiff(Surv(temps, statut)~groupe)
```

Call:

```
survdiff(formula = Surv(temps, statut) ~ groupe)
```

	N	Observed	Expected	(O-E) ^2/E	(O-E) ^2/V
groupe=1	21	9	19.3	5.46	16.8
groupe=2	21	21	10.7	9.77	16.8

Chisq = 16.8 on 1 degrees of freedom, p = 4.17e-05

Remarques sur le test du log-rank

Le test du log-rank peut-être utiliser pour comparer les courbes de survie de k groupes avec $k \geq 2$:

- Le critère statistique Q suit alors une loi du Chi^2 à $k-1$ degrés de liberté

On peut startifier le test à l'aide de la fonction "strata()" dans R :

```
> data.leu[, "sexe"] <- as.factor(c(rep("m", 2), rep("f", 19),
rep("m", 15), rep("f", 6)))
> detach(data.leu)
> attach(data.leu)
> survdiff(Surv(temp, statut) ~ groupe + strata(sexe))
Call:
survdiff(formula = Surv(temp, statut) ~ groupe + strata(sexe))
```

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
groupe=1	21	9	11.7	0.641	2.07
groupe=2	21	21	18.3	0.412	2.07

Chisq = 2.1 on 1 degrees of freedom, p = 0.15

Notes sur le test du log-rank stratifié

Intérêt de stratifier le test du log-rank :

- Permet de tester l'effet de la variable d'intérêt (cf. effet traitement) tout en contrôlant les effets "néfastes" liées à une autre variable dont on sait qu'elle pourrait masquer ou fausser l'effet de la variable d'intérêt

Principe de la stratification :

- Découpage du jeu de données en plusieurs sous-jeux de données ou strates (e.g., strate sexe masculin M vs. strate sexe féminin F)
- Calcul du critère Q_{obs} au sein de chaque strate séparément (e.g., $Q_{obs}[M]$ et $Q_{obs}[F]$)
- Somme des critères Q_{obs} sur les strates (e.g., $Q_{obs}[tot] = Q_{obs}[M] + Q_{obs}[F]$)
- Comparaison des valeurs de $Q_{obs}[tot]$ et Q_{crit}
- Verdict du test

Plan du cours

Analyses de survie

1. Données de survie
2. Fonctions de densité, de survie, et de risque
3. Estimateur de Kaplan-Meier de la fonction de survie
4. Test d'une différence de survie entre plusieurs échantillons
5. Modèle de Cox

Modèle de Cox

Formule générale du modèle de Cox :

$$h(t, X_i) = h_0(t) e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}$$

$$\eta = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

h : risque

h₀ : risque de base

t : temps

X_i : variables explicatives ou prédictives

β_i : coefficients de la régression

Notez que
l'ordonnée à
l'origine (β_0)
est incorporée
dans le risque
de base h_0

Modèle de Cox

En analyse de survie, on modélise donc la fonction h de risque qui dans le cas du modèle de Cox correspond au produit de deux quantités :

$$h(t, X_i) = h_0(t)e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k} = h_0(t)e^{\eta(X_i)}$$

- $h_0 \geq 0$: le risque de base qui est fonction uniquement du temps mais indépendant des variables explicatives
- $e^\eta \geq 0$: l'exponentiel du terme η à modéliser qui est fonction des variables explicatives mais totalement indépendant du temps

Attention, chaque variable explicative dans le modèle de Cox doit être indépendante du temps (sinon, passer au modèle de Cox étendu) et quand toutes les variables explicatives prennent des valeurs nulles alors le risque est égal au risque de base car $e^0 = 1$

Avec le modèle de Cox dit "semi-paramétrique", il n'est pas nécessaire de spécifier le risque h_0 de base d'où sa grande popularité vis-à-vis des modèles purement paramétriques

Modèle de Cox

Le modèle de Cox repose sur la notion de risques proportionnels :

- Si le test du log-rank permet de tester une différence significative de survie entre deux groupes par exemple, il n'est pas possible d'estimer l'étendu de l'impact de cette différence entre ces deux groupes
- Afin de quantifier cet impact, on fait appel au risque instantané de décès de chacun des deux groupes et nous recherchons une fonction simple les reliant
- Pour y arriver, nous nous basons sur une hypothèse essentielle : nous supposons que la proportion des risques instantanés de décès est constante pendant toute la durée d'observation, d'où l'expression de risques proportionnels

Modèle de Cox

Risque instantané h ou hazard rate :

- Le risque instantané décrit la probabilité selon laquelle un événement (décès, diagnostic, maladie, amélioration des symptômes) précis se produira à un instant fixé ou date de point

Risque proportionnel HR ou hazard ratio :

- Le risque proportionnel (HR) pour une variable X_i est le rapport de deux risques instantanés pour un changement d'une unité (ou catégorie) de X_i tout en maintenant les autres variables $X_{j \neq i}$ constantes

$$h(t, X_1 = 1, X_{j \neq 1} = cst) = h_0(t) \times e^{\beta_1 \times 1} \times e^{\beta_2 X_2} \times \cdots \times e^{\beta_k X_k}$$

$$h(t, X_1 = 2, X_{j \neq 1} = cst) = h_0(t) \times e^{\beta_1 \times 2} \times e^{\beta_2 X_2} \times \cdots \times e^{\beta_k X_k}$$

$$HR = \frac{h_0(t) \times e^{\beta_1 \times 2} \times e^{\beta_2 X_2} \times \cdots \times e^{\beta_k X_k}}{h_0(t) \times e^{\beta_1 \times 1} \times e^{\beta_2 X_2} \times \cdots \times e^{\beta_k X_k}} = e^{\beta_1 \times (2-1)} = e^{\beta_1}$$

Modèle de Cox

Interprétations de HR pour une variable X_i :

$$HR = e^{\beta_i} \geq 0$$

- $\beta_i = 0 \Rightarrow HR = 1$: pas d'effet de la variable X_i sur le risque global h
- $\beta_i > 0 \Rightarrow HR > 1$: augmentation du risque global h lié à la variable X_i
- $\beta_i < 0 \Rightarrow HR < 1$: diminution du risque global h lié à la variable X_i

Corresponds à l'impact (en proportion) de l'augmentation d'une unité de X_i , ou à l'impact d'un groupe de traitement proportionnellement au groupe de référence sur le risque globale h

Exemple

Utilisons le jeu de données intitulé "ovarian" dans R :

```
> library(survival)
> data(ovarian)
> str(ovarian)
'data.frame': 26 obs. of 6 variables:
 $ futime   : num  59 115 156 421 431 ...
 $ fustat   : num  1 1 1 0 1 0 1 1 0 1 ...
 $ age       : num  72.3 74.5 66.5 53.4 50.3 ...
 $ resid.ds: num  2 2 2 2 2 1 2 2 2 1 ...
 $ rx        : num  1 1 1 2 1 1 2 2 1 2 ...
 $ ecog.ps  : num  1 1 2 1 1 2 2 2 1 2 ...
> ?ovarian
```

Exemple

Données de survie chez 26 patientes atteintes d'un cancer de l'ovaire :

- *futime* : temps de survenu de l'événement décès en semaine
- *fustat* : statut (0 = survie / 1 = décès)
- *age* : âge des patientes en années
- *resid.ds* : présence résiduel de maladies (1 = non / 2 = oui)
- *rx* : groupe de traitement (1 = placebo / 2 = traitement)
- *ecog.ps* : échelle de performance du patient (1 = moyenne / 2 = faible)

Notre objectif sera de tester l'effet du traitement tout en tenant compte de l'effet âge des patientes sur leurs durées de survie :

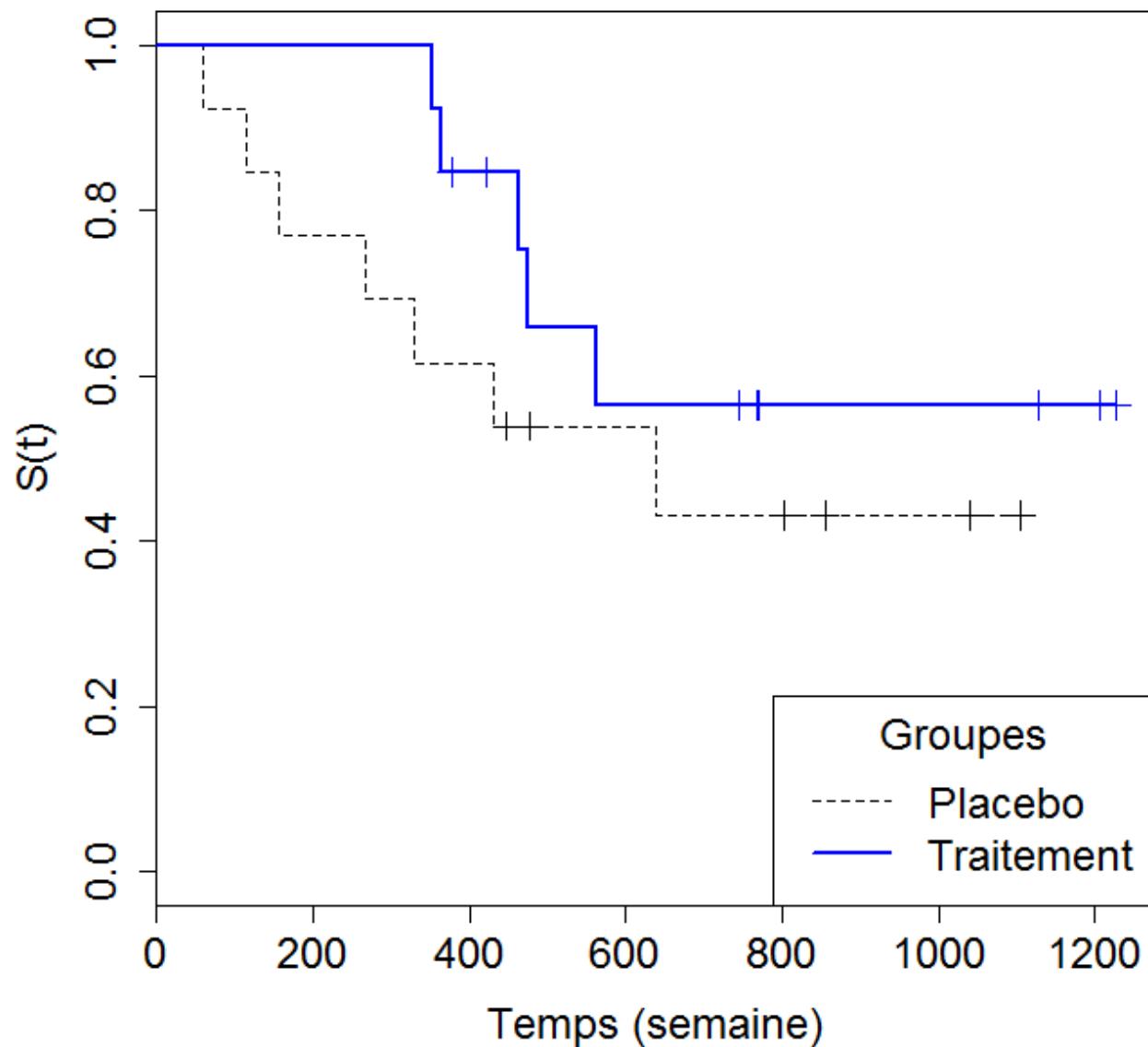
- M1 : *rx*
- M2 : *rx + age*
- M3 : *rx + age + rx:age*

Exemple

Préparons et visualisons les données avant de modéliser :

```
> ovarian$rx <- as.factor(ovarian$rx)
> ovarian$resid.ds <- as.factor(ovarian$resid.ds)
> ovarian$ecog.ps <- as.factor(ovarian$ecog.ps)
> str(ovarian)
'data.frame': 26 obs. of 6 variables:
 $ futime   : num 59 115 156 421 431 ...
 $ fustat   : num 1 1 1 0 1 0 1 1 0 1 ...
 $ age       : num 72.3 74.5 66.5 53.4 50.3 ...
 $ resid.ds: Factor w/ 2 levels "1","2": 2 2 2 2 2 1 2 2 ...
 $ rx        : Factor w/ 2 levels "1","2": 1 1 1 2 1 1 2 2 ...
 $ ecog.ps  : Factor w/ 2 levels "1","2": 1 1 2 1 1 2 2 2 ...
> plot(survfit(Surv(futime, fustat)~1, data=ovarian),
+       xlab=c("Temps (semaine)"), ylab=c("S(t)"))
> plot(survfit(Surv(futime, fustat)~rx, data=ovarian), lty=c(2,
+       1), lwd=c(1, 2), col=c("black", "blue"), conf.int=FALSE,
+       xlab=c("Temps (semaine)"), ylab=c("S(t)"))
> legend("bottomright", c("Placebo", "Traitement"), lty=c(
+       1), lwd=c(1, 2), col=c("black", "blue"), title="Groupes")
```

Exemple



Exemple

Utilisation de la fonction "coxph" :

```
> args(coxph)
function(formula, data, weights, subset, na.action, init,
control, method=c("efron", "breslow", "exact"),
singular.ok=TRUE, robust=FALSE, model=FALSE, x=FALSE, y=TRUE,
...)
NULL
```

Exemple

Cas du modèle nul M1 :

```
> M1 <- coxph(Surv(futime, fustat)~rx, data=ovarian,  
method="breslow")  
> summary(M1)  
n = 26, number of events = 12
```

β_i	coef	exp(coef)	se(coef)	z	Pr(> z)
rx2	-0.5964	0.5508	0.5870	-1.016	0.31

	exp(coef)	exp(-coef)	lower .95	upper .95
rx2	0.5508	1.816	0.1743	1.74

HR	R^2	(se = 0.078)	$e^{\beta_i + 1.96 * se(\beta_i)}$
Concordance = 0.5508	(max possible = 0.932)	(se = 0.078)	$e^{\beta_i - 1.96 * se(\beta_i)}$

Bien que non significatif ($P > 0.05$), l'effet du traitement ($rx = 2$) diminue ($\beta_1 < 0$) le risque h de décès d'un facteur $e^{-0.6} = 0.55$, soit un risque de décès dans le groupe placebo qui est $1/0.55 = 1.81$ fois supérieur à celui du groupe traitement

Exemple

Cas du modèle nul M2 :

```
> M2 <- coxph(Surv(futime, fustat) ~ rx+age, data=ovarian  
method="breslow")
```

```
> summary(M2)
```

n = 26, number of events = 12

	coef	exp(coef)	se(coef)	z	Pr(> z)
rx2	-0.80397	0.44755	0.63205	-1.272	0.2033
age	0.14733	1.15873	0.04615	3.193	0.0014

	exp(coef)	exp(-coef)	lower .95	upper .95
rx2	0.4475	2.234	0.1297	1.545
age	1.1587	0.863	1.0585	1.268

Concordance = 0.798 (se = 0.091)

Rsquare = **0.457** (max possible = 0.932)

Notez également que l'IC à 95% autour de HR dans le groupe traitement est plus ramassé après avoir inclus l'effet de l'âge dans le modèle : gain de précision

Effet très significatif de l'âge sur le risque h de décès qui augmente d'un facteur 1.16 par année tout autre chose étant égales

Exemple

Cas du modèle nul M3 :

```
> M3 <- coxph(Surv(futime, fustat) ~ rx+age+rx:age, data=ovarian,  
method="breslow")  
> summary(M3)  
n = 26, number of events = 12
```

	coef	exp(coef)	se(coef)	z	Pr(> z)
rx2	-9.504e+00	7.453e-05	9.008e+00	-1.055	0.29139
age	1.342e-01	1.144e+00	4.578e-02	2.932	0.00336 **
rx2:age	1.464e-01	1.158e+00	1.498e-01	0.977	0.32833

	exp(coef)	exp(-coef)	lower .95	upper .95
rx2	7.453e-05	1.342e+04	1.602e-12	3468.597
age	1.144e+00	8.744e-01	1.046e+00	1.251
rx2:age	1.158e+00	8.638e-01	8.631e-01	1.553

Concordance = 0.817 (se = 0.091)

Rsquare = **0.479** (max possible = 0.932)

Exemple

Comparaisons de modèles :

```
> anova(M1, M2, test="Chisq")
Analysis of Deviance Table
Cox model: response is Surv(futime, fustat)
Model 1: ~ rx
Model 2: ~ rx + age
  loglik  Chisq Df P(>|Chi|)
1 -34.459
2 -27.042 14.835  1 0.0001174 ***
> anova(M2, M3, test="Chisq")
Analysis of Deviance Table
Cox model: response is Surv(futime, fustat)
Model 1: ~ rx + age
Model 2: ~ rx + age + rx:age
  loglik  Chisq Df P(>|Chi|)
1 -27.042
2 -26.509 1.0648  1      0.3021
```

On conserve donc
le modèle M2

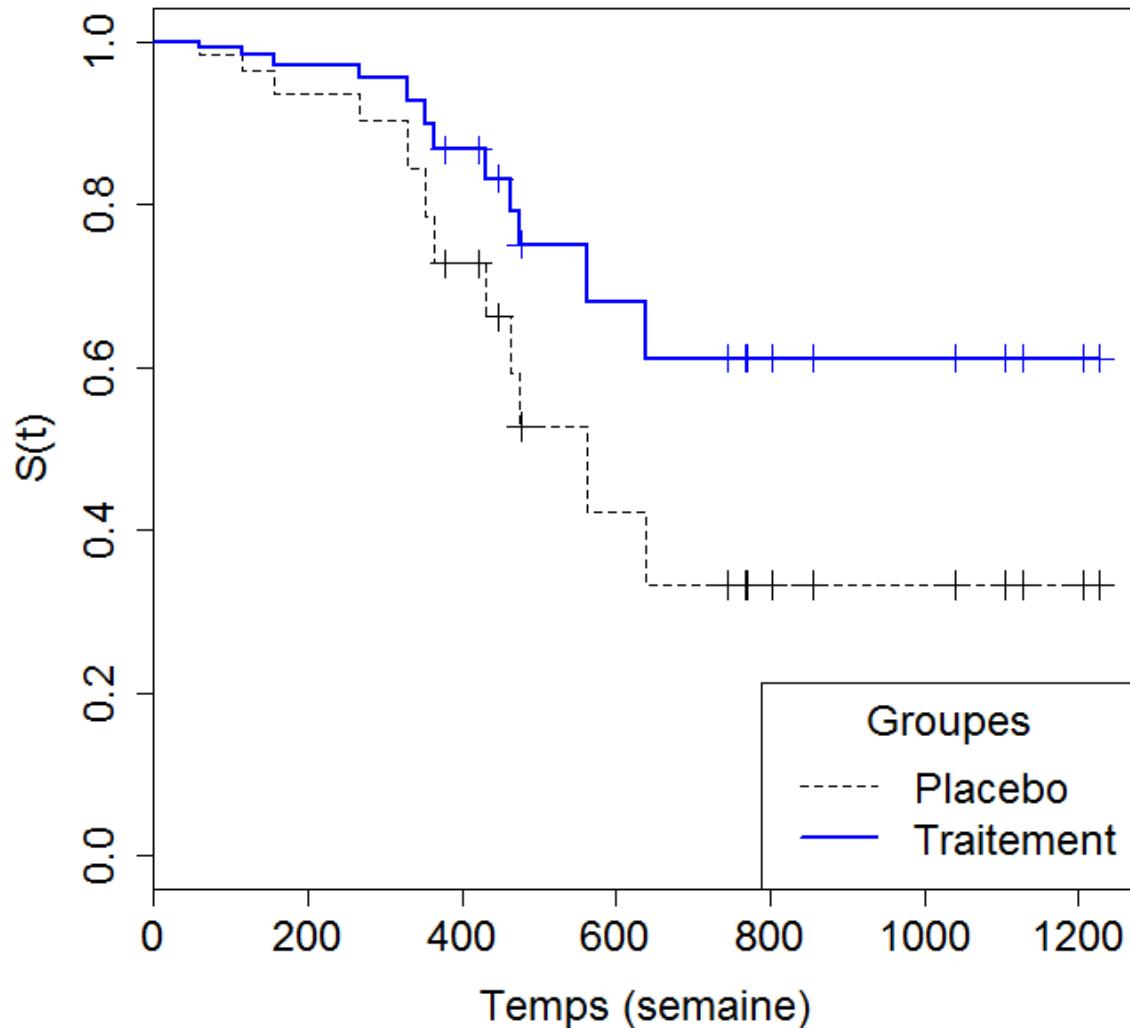
Exemple

Tracer les courbes S de survie à partir du modèle M2 :

```
> ovarian.M2 <- with(ovarian, data.frame(rx=as.factor(c(1, 2)),  
age=rep(mean(age), 2)))  
> plot(survfit(M2, newdata=ovarian.M2), lty=c(2, 1), lwd=c(1,  
2), col=c("black", "blue"), xlab="Temps (semaine)",  
ylab="S(t)", main="Courbes de survie")  
> legend("bottomright", c("Placebo", "Traitement"), lty=c(2,  
1), lwd=c(1, 2), col=c("black", "blue"), title="Groupes")
```

Exemple

Courbes de survie



Exemple

Vérification de la condition de risques proportionnels avec le temps :

```
> cox.zph(M2)
      rho  chisq     p
rx2    0.2072  0.518  0.472
age   -0.0918  0.113  0.736
GLOBAL       NA  0.729  0.695
```

Condition
d'indépendance
vérifiée pour les
variables rx et age du
modèle M2

- La fonction *cox.zph* permet de tester si la condition d'indépendance des variables explicatives X_i en fonction du temps (cf. risque proportionnel) est respecté pour chaque variable X_i en se basant sur l'étude des résidus de Schoenfield
- Toute variable X_i du modèle dont la probabilité du test est inférieur à 0.05 présente alors une variation significative au cours du temps entraînant la violation d'une des conditions du modèle de Cox

Dans ce cas il faudra opter pour un
modèle de Cox étendue afin de tenir
compte de cette dépendance

Exemple

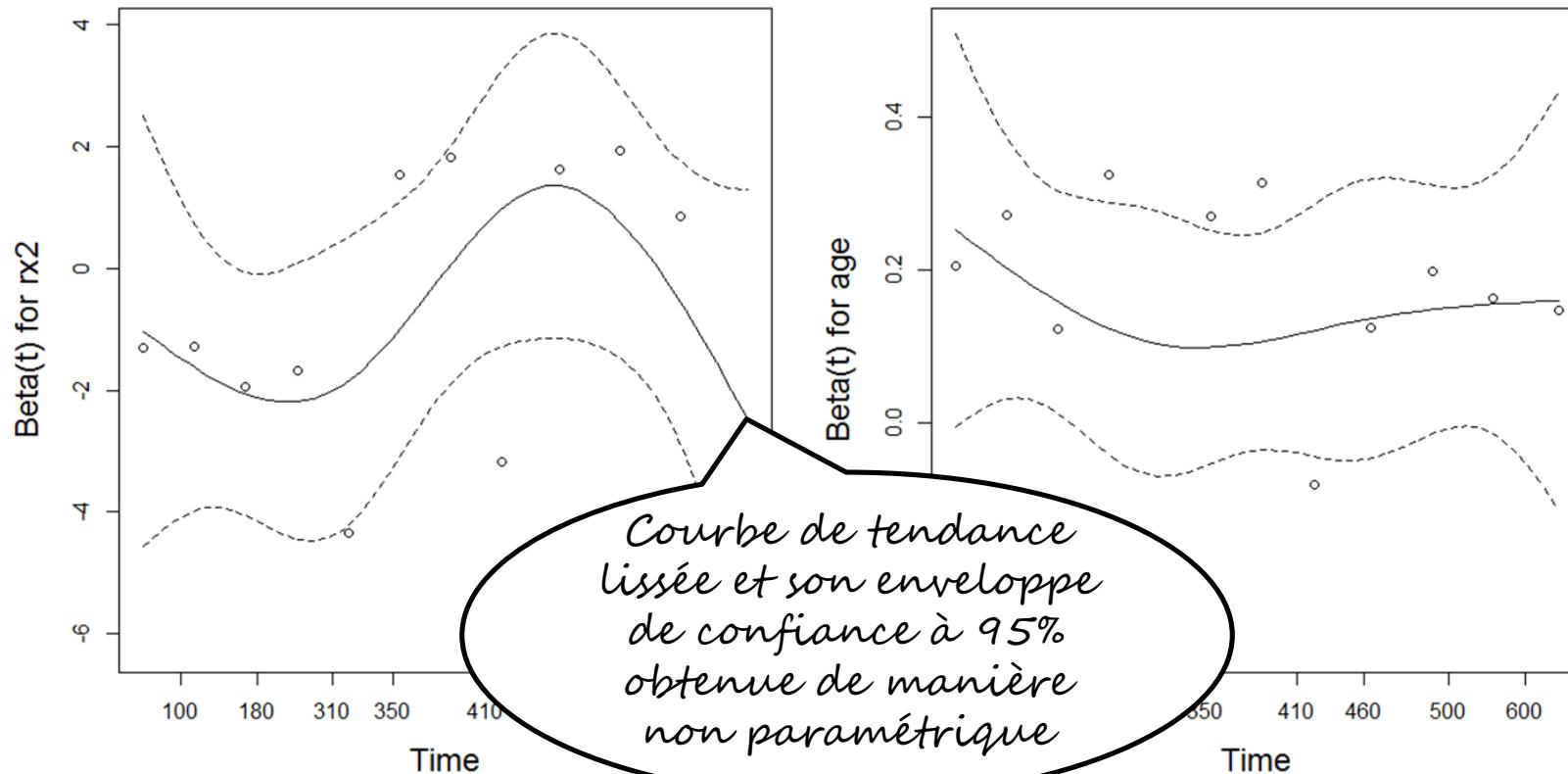
Vérification de la condition de risques proportionnels avec le temps :

- Il est également possible d'afficher le graphique du diagnostic des résidus de Schoenfield en fonction du temps afin de visualiser pour chaque variable X_i la tendance éventuellement détectée par le test de la fonction `cox.zph`

```
> par(mfrow=c(1, 2))  
> plot(cox.zph(M2))
```

Exemple

Vérification de la condition de risques proportionnels avec le temps :



Exemple

Diagnostic des patientes ayant une influence forte sur les coefficients β_i estimés du modèle M2 :

```
> dfbeta.M2 <- residuals(M2, type="dfbeta")
> dfbeta.M2
[,1]      [,2]
1 -0.033744665  0.0039102434
2 -0.020184355  0.0049085637
3 -0.070196131 -0.0005663619
4 -0.017409937  0.0014704996
5 -0.159594456 -0.0146557656
[...]
22  0.009843176 -0.0049572634
23 -0.289645553 -0.0355943808
24  0.123022007  0.0092577874
25  0.106029534  0.0100715614
26 -0.040610413  0.0009906980
```

É

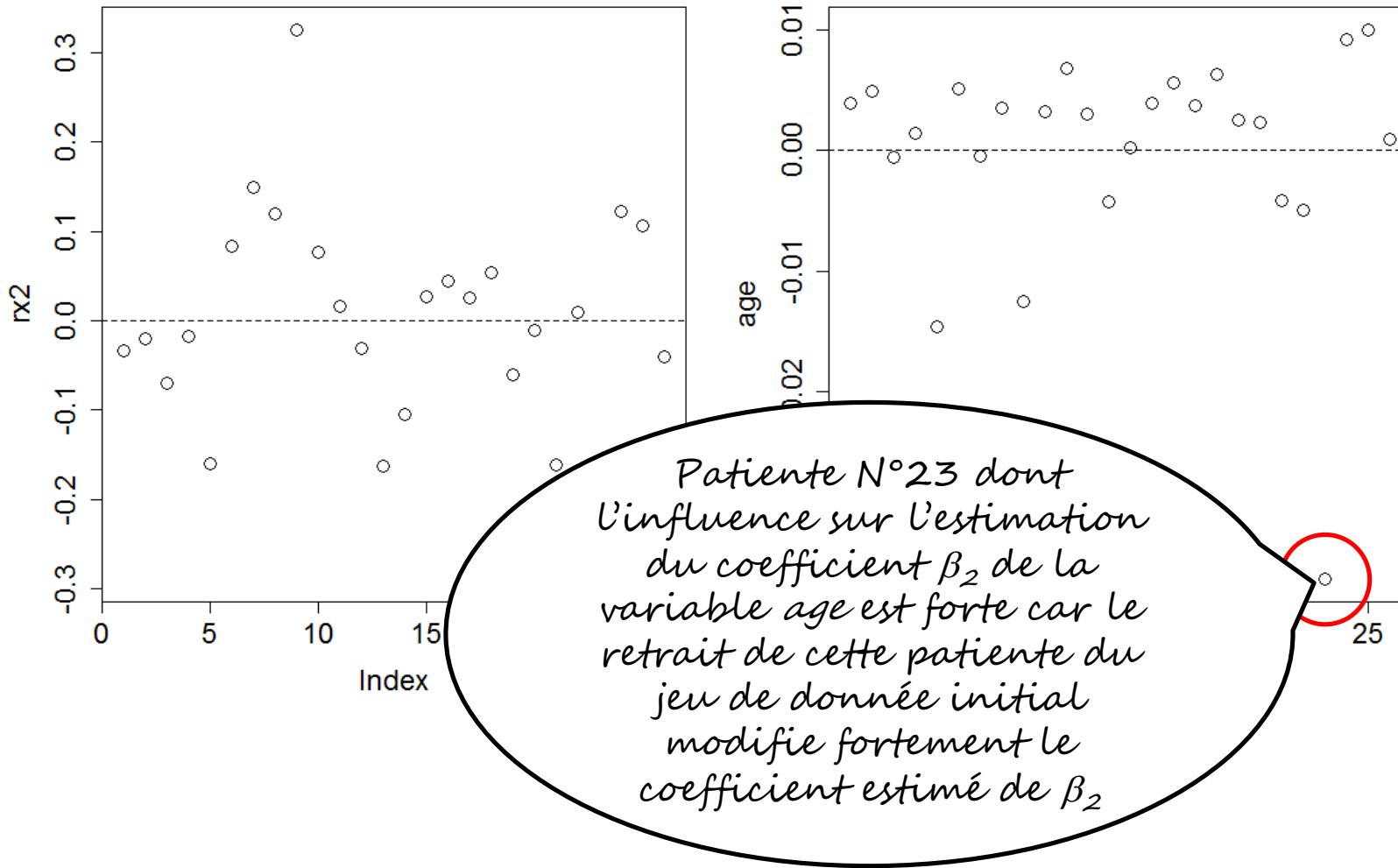
Exemple

Diagnostic des patientes ayant une influence forte sur les coefficients β_i estimés du modèle M2 :

```
> par(mfrow=c(1, 2))
> for (j in 1:2)
{
  plot(dfbeta.M2[, j], ylab=names(coef(M2)) [j])
  abline(h=0, lty=2)
}
```

Exemple

Diagnostic des patientes ayant une influence forte sur les coefficients β_i estimés du modèle M2 :



Exercice

Importez dans R le fichier nommé "Chap5_Rossi.txt" qui contient les données d'une étude sur le récidivisme de 432 prisonniers :

- *week* : temps de survenu de l'événement récidive en semaine
- *arrest* : statut (0 = non arrêté / 1 = arrêté pour récidive)
- *fin* : aide financière aléatoirement attribuée à certains détenus (no / yes)
- *age* : âge du détenu à sa sortie de prison
- *race* : couleur de peau du détenu (black / other)
- *wexp* : activité salarié du détenu avant incarcération (no / yes)
- *mar* : statut marital du détenu à sa sortie de prison (married / not married)
- *paro* : détenu en liberté conditionnelle (no / yes)
- *prio* : nombre de condamnations antérieures
- *educ* : niveau d'éducation sur une échelle ordinaire de 2 à 6

Exercice

A partir des 8 variables explicatives disponibles, modélez le risque h de récidive de ces 432 anciens détenus et tracez les courbes S de « survie » ou « proportions d'anciens détenus toujours en liberté » de votre modèle finale pour les deux groupes de détenus suivant :

- *fin.yes* : ceux ayant reçu une aide financière
- *fin.no* : ceux n'ayant pas reçu d'aide financière