Cours de Master 2 - Analyse des Durées de Vie

Elodie Brunel

Université de Montpellier

2024-2025

STATISTIQUE
SCIENCE DES DONNÉES
UNIVERSITÉ DE MONTPELLIER

Modalités du Contrôle des Connaissances (MCC)

L'UE est évaluée en 100% Contrôle Continu : pas de Contrôle Terminal, pas de Session 2.

- 30 % TP1 : travail individuel
- 30 % TP2 : travail individuel
- 40 % TP3 : travail en binôme.

Voir les dates de remise des travaux sur Moodle.

Programme du Cours

Partie 1 (Séances 1 et 2) : Différents types de censure, estimation de la fonction de survie, tests de rangs.
TP1 à rendre.

Partie 2 (Séances 3 et 4) : Modèles paramétriques pour l'ADV et régression log-linéaires en présence de covariables (Accelerated Failure Time models)
TP2 à rendre.

Partie 3 (Séances 5 et 6) : Modèle de Cox.
TP3 à rendre.

Partie 1

1. Introduction et Vocabulaire
2. Définitions des différents types de censure
3. Estimation non paramétrique de la fonction de survie
4. Intervalles de confiance / Bandes de confiance
5. Tests de rangs

1. Introduction et Vocabulaire

En analyse de survie, la variable d'intérêt est la durée qui s'écoule à partir d'une date d'origine et jusqu'à la survenue d'un événement d'intérêt :

dans le domaine biomédical

- durée de survie d'un patient ayant eu un infarctus,
- durée de rémission d'un patient (rémission = disparition momentanée des symptômes d'une maladie),
- durée de séropositivité sans symptôme d'un patient infecté par le VIH,

autres domaines d'applications

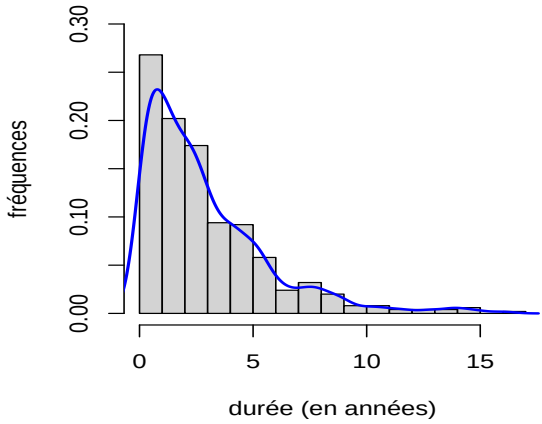
- durée de fonctionnement d'un matériel avant la panne
- durée de chômage jusqu'au réemploi
- durée d'assurance avant sinistre ...

...

1. Introduction et Vocabulaire

On dit que l'individu survit au temps t, si au bout de la durée t, l'événement d'intérêt n'a pas encore eu lieu. Ainsi, si l'on reprend les exemples précédents, le patient survit au temps t lorsque :

- le patient est toujours en vie après une durée t depuis la date de l'infarctus (événement d'intérêt=décès)
- le patient est toujours en rémission après une durée t (événement d'intérêt=rechute),
- le patient séropositif est toujours asymptomatique depuis une durée t (événement d'intérêt=apparition de symptômes marqueur du SIDA)
- le matériel est toujours en état de fonctionnement depuis sa mise en service après une durée t (événement d'intérêt=panne)
- l'individu est toujours au chomâge après une durée t (événement d'intérêt=réemploi)



Les durées sont des variables aléatoires à valeurs positives et à distribution en général très asymétrique (distribution gaussienne inadéquate).

Principaux objectifs d'une analyse de survie :

- quantifier l'évolution du risque de survenue de l'événement d'intérêt au cours du temps.
- comparer des groupes de patients lors d'un essai thérapeutique : durées de survie des patients qui reçoivent un nouveau traitement à ceux recevant le traitement usuel ou un placebo.
- identifier (lors d'une étude épidémiologique ou prospective) les facteurs de risque de la survenue de l'événement et quantifier leur effet.

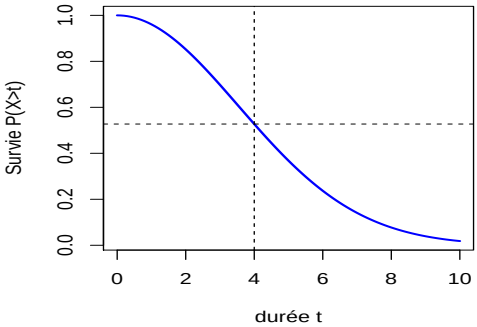
Fonctions d'intérêt en analyse de survie : Si X est la variable aléatoire qui représente durée d'intérêt :

- la fonction de survie $S(t) = P(X > t)$ représente la probabilité que l'événement d'intérêt se produise au-delà d'une durée t.
- la fonction de risque instantané notée $\lambda(t)$
- la fonction de risque cumulé notée $\Lambda(t)$.

Si X désigne une variable de durée continue (admettant une loi de probabilité à densité [1]). La fonction de survie $S(t)$ est la probabilité de survivre au moins une durée t :

$$S(t) = \mathbb{P}(X > t)$$

Elle est représentée graphiquement par la courbe de survie.
La courbe est décroissante et continue sur $\mathbb{R}^+$.
La probabilité de survivre au delà de $t = 4$ est d'environ 53%.



[1] Il existe des modèles discrets mais nous ne les abordons pas dans ce cours

Autres fonctions d'intérêt :

- La fonction de risque instantané (hazard rate) $\lambda(t)$

$$\lambda(t) = \lim_{\delta t \to 0^+} \frac{1}{\delta t} \mathbb{P}\left(t < X \leq t + \delta t | X > t\right) = \lim_{\delta t \to 0^+} \frac{1}{\delta t} \frac{\mathbb{P}\left(t < X \leq t + \delta t\right)}{\mathbb{P}(X > t)}$$

  Elle s'interprète comme le "risque" de survenue de l'événement d'intérêt (ou de décès) en t sachant qu'il ne s'est pas produit avant t (ou que l'individu a survécu jusqu'à t).

- La fonction de risque cumulé $\Lambda(t)$:

$$\Lambda(t) = \int_0^t \lambda(u) du$$

Autre spécificité des variables de durée : elles sont souvent incomplètement observées : problème de la censure.

$\hookrightarrow$ Il est nécessaire de développer des traitements statistiques spécifiques : analyse des durées de vie.

Nous allons aborder ces différents types de censures.

2. Définitions des différents types de censure

La durée de vie sera complètement observée si l'on connaît précisément la date du début et la date de survenue de l'événement d'intérêt :
Si l'une des deux est manquante alors nous n'avons qu'une information partielle sur la durée de vie et la durée de vie est dite censurée. Il peut s'agir de :

Censure à droite

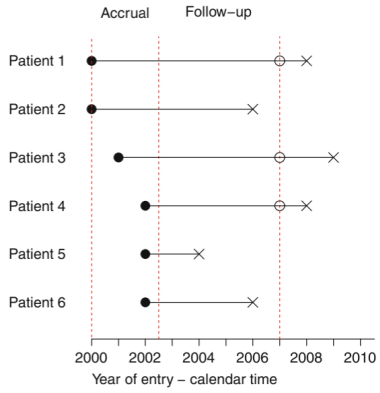
Censure à gauche

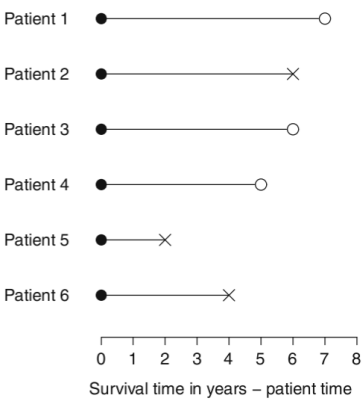
Censure par intervalle

d'autres mécanismes comme la troncature peuvent aussi être à l'orgine d'une observation partielle des durées d'intérêt.

2. Définitions des différents types de censure

Censure droite fixe : à la fin de l'étude, l'événement n'a pas encore été observé pour les patients 1, 3 et 4. Figures d'après Moore [2016]

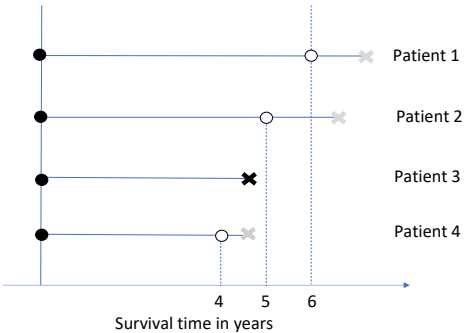




2. Définitions des différents types de censure

Censure droite aléatoire : Le patient est perdu de vue au cours du suivi avant la fin de l'étude : on ne peut pas observer l'événement d'intérêt.

Les patients 1 et 2 sont censurés par la fin de l'étude (censure fixe à droite : durée de l'étude = 6 ans).
La durée de vie du patient 3 est complètement observée.
Le patient 4 est perdu de vue au bout de 4 ans (censure aléatoire à droite).



2. Définitions des différents types de censure

Censure droite aléatoire : Une durée de vie aléatoire X est dite censurée à droite par une durée aléatoire de censure C si on observe parfois C au lieu de X. L'information donnée par la durée C sur X est : $X > C$.

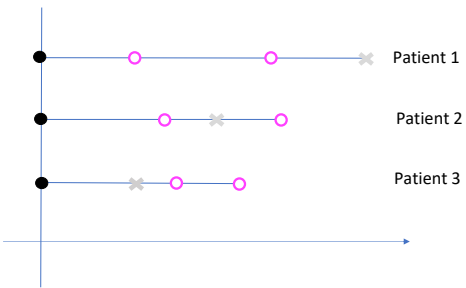
Censure gauche aléatoire : De même, une durée de vie aléatoire X est dite censurée à gauche par une durée aléatoire de censure C si l'information donnée par la durée C sur X est : $X < C$.

Exemple : On étudie la durée nécessaire à l'acquisition d'une tâche (comme la lecture=événement d'intérêt) chez de jeunes enfants. Certains savent déjà lire au début de l'étude. La durée X est inconnue mais elle est inférieure à l'âge C des enfants. Pour plus de détails sur cet exemple, cf. Huber-Carol [1994] [2].

[2] Huber-Carol C., (1994), Durées de survie tronquées et censurées, Journal de la société statistique de Paris, tome 135, 4 (1994), p. 3-23

Censure par intervalle : Si, au lieu de la durée de vie d'intérêt X, on observe $C_1 < C_2$ qui définissent l'intervalle dans lequel a lieu l'événement d'intérêt (X est non observée), il y a censure par intervalle.

pour le patient 1, $X > C_2$.
pour le patient 2 : $C_1 < X < C_2$.
pour le patient 3 : $X < C_1$.
On sait simplement si l'événement d'intérêt a lieu avant C_1, entre C_1 et C_2, ou après C_2.
Si $C_1 = C_2$, on parle de censure par intervalle de type I ou "current status data".



Troncature : Il ne faut pas confondre les mécanismes de censure et de troncature. Il y a troncature si l'observation de la variable d'intérêt X n'a lieu que conditionnellement à un événement B.

- **Troncature à gauche :** On dit qu'il y a troncature gauche lorsque la variable d'intérêt X n'est observable que si elle est supérieure à T. X n'est observée que si l'événement $X > T$ est réalisé.

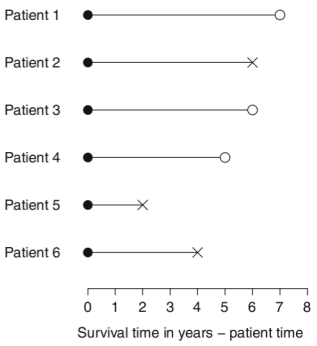
  Exemple : Durée de vie après la retraite : les sujets entrent dans l'enquête à la suite d'un tirage au sort dans une caisse de retraite. La durée X d'un sujet n'est donc observée que si cette durée de vie après la retraite excède le délai T entre sa prise de retraite et l'instant de l'enquête. La durée de vie après la retraite X est donc tronquée à gauche par ce délai T : $X > T$.
  On ne peut estimer, avec ce type de données, que la loi de la durée X conditionnellement à l'événement $(X > T)$.

3. Estimation non paramétrique de la fonction de survie en présence de

Nous nous consacrons désormais à l'analyse de durées de vie X censurée à droite qui représente le phénomène de censure le plus usuellement rencontré.

Modélisation pour des données censurées à droite : on introduit une variable indicatrice $D_i = 1_{(X_i \leq C_i)}$ de l'événement d'intérêt (décès)

patient i	durée observée $T_i = \min(X_i, C_i)$	statut $D_i = 1_{(X_i \leq C_i)}$
1	7	0 (censuré)
2	6	1 (décès)
3	6	0 (censuré)
4	5	0 (censuré)
5	2	1 (décès)
6	4	1 (décès)



On observe pour chaque individu i :

$$T_i = \min(X_i, C_i) = \begin{cases} X_i & \text{si le "décès" est observé} \quad D_i = 1 \\ C_i & \text{si l'individu est censuré} \quad D_i = 0 \\ & \text{avant qu'on ait pu observer le "décès"} \end{cases}$$

Hypothèse fondamentale sur le mécanisme de la censure à doite :
Pour pouvoir faire de l'inférence (écriture de la vraisemblance, propriétés de convergence des estimateurs, etc) On doit supposer que pour tout i :

- **La durée jusqu'au décès X_i et la durée de censure C_i sont indépendantes**.
- En présence de covariables explicatives, on peut supposer que :
  X_i et C_i sont **indépendantes conditionnellement aux covariables Z_i**.

On note $t_1 < t_2 < \cdots < t_k$ les durées distinctes et ordonnées correspondant à des événements d'intérêt (décès). On a $1 \leq k \leq n$.
L'estimateur de Kaplan-Meier (1958) de la fonction de survie $S(t)$ est défini par :

$$\hat{S}_{KM}(t) = \begin{cases} \prod_{i:t \geq t_i} \left(1 - \frac{m_i}{n_i}\right) & \text{si } t \geq t_1 \\ 1 & \text{si } t < t_1 \end{cases}$$

avec :
m_i le nombre d'événements observés en t_i
n_i le nombre d'individus à risque en t_i (ni décédés, ni censurés)

Construction heuristique de l'estimateur de Kaplan-Meier

La probabilité de survivre au-delà d'une durée $t > s$ peut s'écrire :

$$P(X > t) = P(X > t|X > s)P(X > s)$$

si on choisit comme durées de conditionnement les durées $t_1 < t_2 < \cdots < t_k$ où se sont produit un ou des événements, on a alors des probabilités de la forme:

$$\pi_i = P(X > t_i|X > t_{i-1}) \quad \text{probabilité de survivre entre } t_{i-1} \text{ et } t_i \text{ sachant qu'on était vivant en } t_{i-1}$$

Et ainsi, de proche en proche :

$$P(X > t_i) = \pi_i \times \pi_{i-1} \times \cdots \times \pi_1 \times \pi_0 \quad \text{avec } \pi_0 = P(X > 0) = 1$$

Les π_i sont estimées par $\hat{\pi}_i = 1 - \frac{m_i}{n_i}$, $i \geq 1$ et on obtient :

$$\hat{S}_{KM}(t) = \prod_{i:t \geq t_i} \hat{\pi}_i$$

Exemple historique : les données de Freireich

La variable de durée est ici une durée de rémission (en semaines) obtenue par des stéroïdes chez des patients atteints de leucémie aiguë, traités soit par placebo soit par 6-mercaptopurine (6-MP).

6-MP	6	6	6	6^+	7	9^+	10	10^+	11^+	13	
	16	17^+	19^+	20^+	22	23	25^+	32^+			
	32^+	34^+	35^+								
Placebo	1	1	2	2	3	4	4	5	5	8	8	8
	8	11	11	12	12	15	17	22	23			

Le signe $+$ correspond à des patients qui ont quitté l'étude à la date considérée. Ils sont donc censurés. Par exemple le 4ème patient est perdu de vue au bout de 6 semaines de traitement avec le 6-MP : il a donc une durée de rémission supérieure à 6 semaines. Par convention, on suppose que les censures ont lieu après les événements d'intérêts.

Exemple historique : les données de Freireich

Dans le groupe placebo, il n'y a aucune censure, la fonction de survie sans rémission se calcule donc de la façon suivante :

$$\hat{S}_{KM}(t) = \hat{S}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(T_i > t)$$

semaine i	nombre de rémissions à la semaine i	proportion de rémissions à la semaine i
1	19	19/ 21= 0,90
2	17	17/ 21= 0,81
3	16	0,76
4	14	0,67
5	12	0,57
8	8	0,38
11	6	0,29
12	4	0,19
15	3	0,14
17	2	0,10
22	1	0,05
23	0	0,00

Exemple historique : les données de Freireich

Dans le groupe 6-MP :

durée de rémission observées t_i	sujets en rémission n_i au début de la semaine i	rechutes observées m_i à la semaine i	prob. $\hat{\pi}_i$ de ne pas rechuter à la semaine i sachant qu'on est en rémission à la semaine $(i-1)$	prob. d'être en rémission à la semaine i
6	21	3		
7	17	1		
10	15	1		
13	12	1		
16	11	1		
22	7	1		
23	6	1		

Exemple historique : les données de Freireich

durée de rémission observées t_i	sujets en rémission n_i au début de la semaine i	rechutes observées m_i à la semaine i	prob. $\hat{\pi}_i$ de ne pas rechuter à la semaine i sachant qu'on est en rémission à la semaine $(i-1)$	prob. d'être en rémission à la semaine i
6	21	3	$18/21 = 0{,}857$	1* 18 /21= 0,857
7	17	1	$16/17 = 0{,}941$	0,857 * 16/17 =0,807
10	15	1	$14/15 = 0{,}933$	0,807 * 14/15=0,753
13	12	1	$11/12 = 0{,}917$	0,753* 11/12= 0,690
16	11	1	$10/11 = 0{,}909$	0,690*10/11=0,627
22	7	1	$6/7 = 0{,}857$	0,627*6/7=0,538
23	6	1	$5/6 = 0{,}833$	0,538*5/6=0,448

L'estimateur de Kaplan-Meier est asymptotiquement gaussien et sa variance asymptotique peut être estimée par la formule de Greenwood :

$$\widehat{V(\hat{S}_{KM}(t))} = \hat{S}_{KM}(t)^2 \left(\sum_{t \geq t_i} \frac{m_i}{n_i(n_i - m_i)} \right)$$

On peut construire alors des intervalles de confiance asymptotiques ponctuels (à t fixé) dont les bornes sont, pour un niveau de confiance $1 - \alpha$:

$$\hat{S}_{KM}(t_i) \left(1 \pm q_{1-\alpha/2} \sqrt{\sum_{t \geq t_i} \frac{m_i}{n_i(n_i - m_i)}} \right) \quad (1)$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la gaussienne standardisée.

Idée de preuve de la formule de Greenwood :

Etape 1 : Ecrivons l'estimateur de Kaplan-Meier :

$$\hat{S}(t) = \prod_{j=1}^{i} \hat{\pi}_j \quad \text{où } \hat{\pi}_j = \frac{n_j - m_j}{n_j} \quad \text{et } t_i \leq t < t_{i+1}$$

$\hat{\pi}_j$ représente la proportion d'individus qui survivent entre t_j et t_{j+1}, sachant qu'ils ont survécu jusqu'en t_j.

Etape 2 : On cherche à déterminer $\mathbb{V}(\hat{\pi}_j)$
Conditionnellement au passé de t_j, le nombre d'individus $n_j - m_j$ qui survivent sur l'intervalle $[t_j, t_{j+1}[$ suit une loi binomiale $\mathcal{B}(n_j, \pi_j)$ où π_j est la probabilité (inconnue) de survivre sur l'intervalle $[t_j, t_{j+1}[$ sachant qu'on a survécu jusqu'en t_j donc

$$n_j - m_j \sim \mathcal{B}(n_j, \pi_j) \Longrightarrow \mathbb{V}(n_j - m_j) = n_j \pi_j (1 - \pi_j)$$

Enfin, $\mathbb{V}(\hat{\pi}_j) = \mathbb{V}\left(\frac{n_j - m_j}{n_j}\right) = \frac{\mathbb{V}(n_j - m_j)}{n_j^2} = \frac{\pi_j(1-\pi_j)}{n_j}$ que l'on peut estimer par :

$$\widehat{\mathbb{V}(\hat{\pi}_j)} = \frac{\hat{\pi}_j(1-\hat{\pi}_j)}{n_j}$$

Etape 3 : On prend le logarithme de $\hat{S}(t)$, on a pour $t_i \leq t < t_{i+1}$

$$\mathbb{V}(\log \hat{S}(t)) = \mathbb{V}\left(\sum_{j=1}^{i} \log \hat{\pi}_j\right) = \sum_{j=1}^{i} \mathbb{V}\left(\log \hat{\pi}_j\right)$$

(ici on suppose que les $\hat{\pi}_j$ sont indépendantes, ce qui peut être discutable).
On peut l'estimer par :

$$\mathbb{V}(\log \hat{S}(t)) = \sum_{j=1}^{i} \frac{1}{\pi_j^2} \mathbb{V}(\hat{\pi}_j),$$

en appliquant la δ-méthode, $\mathbb{V}(g(X_n)) \approx (g'(\mu))^2 \mathbb{V}(X_n)$ avec $g = \log$ et $X_n = \hat{\pi}_j$ et $\mu = \pi_j$.

Puis, en appliquant encore une fois la δ-méthode, avec $g = \log$ et $X_n = \hat{S}(t)$, on a :

$$\mathbb{V}(\log \hat{S}(t)) \approx \frac{1}{S(t)^2} \mathbb{V}(\hat{S}(t))$$

d'où l'on déduit :

$$\mathbb{V}(\hat{S}(t)) \approx S(t)^2 \mathbb{V}(\log \hat{S}(t))$$

Finalement, comme $\hat{\pi}_j = (n_j - m_j)/n_j$ et en remplaçant $S(t)$ par son estimation $\hat{S}(t)$,

$$\widehat{\mathbb{V}(\hat{S}(t))} = \hat{S}(t)^2 \widehat{\mathbb{V}(log\hat{S}(t))} = \hat{S}(t)^2 \sum_{j=1}^{i} \frac{m_j}{n_j(n_j - m_j)}$$

Application à la construction d'intervalles de confiance :

L'estimateur de Kaplan-Meier $\hat{S}(t)$ est asymptotiquement normal de moyenne $S(t)$ et sa variance estimée est donnée par la formule de Greenwood.
On en déduit un intervalle de confiance de niveau $1-\alpha$ pour la survie $S(t)$ de la forme :

$$[\hat{S}(t) - \hat{\sigma}(\hat{S}(t))q_{1-\alpha/2}; \hat{S}(t) + \hat{\sigma}(\hat{S}(t))q_{1-\alpha/2}]$$

où $q_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$.

Attention: Comme on a à un intervalle "symétrique" dont les bornes peuvent être négative ou supérieure à 1, on prendra en pratique :

$$[max(\hat{S}(t) - \hat{\sigma}(\hat{S}(t))q_{1-\alpha/2}; 0); min(\hat{S}(t) + \hat{\sigma}(\hat{S}(t))q_{1-\alpha/2}; 1)]$$

Exemple : pharmacoSmoking

- Medical therapies to help smokers Randomized trial of triple therapy vs. patch for smoking cessation.
- Data frame with 125 observations and 14 variables [3]:
  - id: patient ID number
  - ttr: Time in days until relapse
  - relapse: Indicator of relapse (return to smoking)
  - grp: Randomly assigned treatment group with levels combination or patchOnly
  - etc

```
## id ttr relapse grp
## 1 21 182 0 patchOnly
## 2 113 14 1 patchOnly
## 3 39 5 1 combination
## 4 80 16 1 combination
```

[3] Steinberg et al. (2009), Triple-combination pharmacotherapy for medically ill smokers: A randomized trial. Annals of Internal Medicine 150, 447-454.

Exemple : pharmacoSmoking

Mise en oeuvre de l'estimateur de Kaplan-Meir avec bornes de l'Intervalle de Confiance de niveau 95%

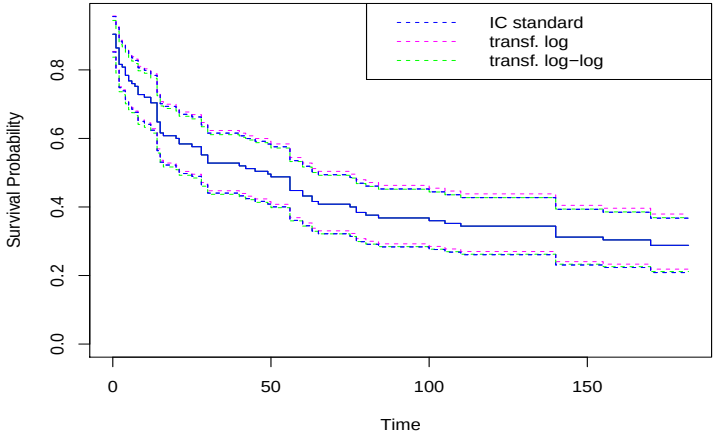
```
fit_KM<-survfit(Surv(ttr, relapse) ~ 1, data = pharmacoSmoking)
summary(fit_KM)

## time n.risk n.event survival std.err lower 95%CI upper 95%CI
##    0    125      12    0.904  0.0263      0.854        0.957
##    1    113       5    0.864  0.0307      0.806        0.926
##    2    108       6    0.816  0.0347      0.751        0.887
##    3    102       1    0.808  0.0352      0.742        0.880
##    4    101       3    0.784  0.0368      0.715        0.860
##    5     98       2    0.768  0.0378      0.697        0.846
↪      t_i     n_i       m_i   Ŝ_KM(t_i)   σ̂(Ŝ_KM(t_i))
```

Les bornes de l'intervalle de confiance asymptotique "standard" peuvent être inférieure à 0 ou supérieure à 1. Diverses améliorations ont été proposées via des transformations de $\hat{S}_{KM}(t)$ (en utilisant une δ-méthode) cf. Section 4.3 dans Klein & Moeshberger (2005) [4] :

- transformation "log"(bornes $\in \mathbb{R}^+$).
- transformation "log-log" (bornes $\in [0, 1]$)
- transformation "arcsin" (bornes $\in [0, 1]$)

[4] John P Klein and Melvin L Moeschberger. Survival analysis: techniques for censored and truncated data. Springer Science & Business Media, 2005.

Intervalles de confiance ponctuels pour différentes transformations



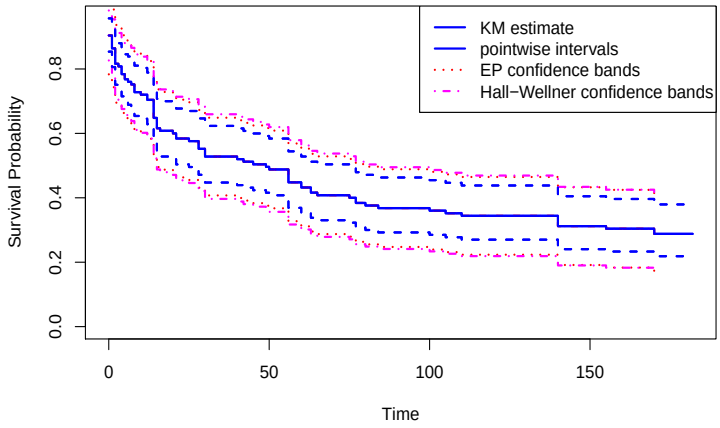
4. Intervalles de confiance *vs* Bandes de confiance

Pour obtenir une bande de confiance de niveau $1-\alpha$ qui garantit que la courbe de survie se trouve dans la bande de confiance pour tout t dans un intervalle donnée $[t_L, t_U]$, c'est-à-dire :

$$P(L(t) \leq S(t) \leq U(t), \forall t_L \leq t \leq t_U) \approx 1-\alpha$$

• une première approche est proposée par Nair (1984) : Equal Probabililty (EP) confidence bands. Les bornes obtenues par cette méthode sont proportionnelles aux bornes de l'IC ponctuel.
• une autre approche est proposée par Hall-Wellner (1980).
Les deux méthodes fournissent des bandes de confiance qui sont plus larges que les intervalles de confiance ponctuels : cf. Section 4.4 dans Klein & Moeshberger (2005) pour plus de détails et pour les formules explicites.

Bandes de confiance



Autres Estimateurs non paramétriques

- Estimateur de **Nelson-Aalen** de la fonction de risque cumulé $\Lambda(t)$:

$$\hat{\Lambda}_{NA}(t) = \sum_{i:t \geq t_i} \frac{m_i}{n_i}$$

- Avec la relation, $S(t) = \exp(-\Lambda(t))$, on construit l'estimateur de **Harrington-Fleming** de la fonction de survie :

$$\hat{S}_{HF}(t) = \exp(-\hat{\Lambda}_{NA}(t)) = \exp(-\sum_{i:t \geq t_i} \frac{m_i}{n_i}) = \prod_{i:t \geq t_i} \exp(-\frac{m_i}{n_i})$$

$$\text{et } \hat{S}_{KM}(t) \leq \hat{S}_{HF}(t), \forall t \geq 0$$

car $\ln[\hat{S}_{KM}(t)] - \ln[\hat{S}_{HF}(t)] = \sum_{i:t \geq t_i} \left[\ln(1 - \frac{m_i}{n_i}) + \frac{m_i}{n_i}\right]$ et $\ln(1-x) + x \leq 0$, donc $\ln[\hat{S}_{KM}(t)] \leq \ln[\hat{S}_{HF}(t)]$.

On peut aussi montrer que l'estimateur de Harrington-Fleming a un biais de surestimation.

5. Tests de rangs

La comparaison de deux échantillons ou plus est un objectif majeur dans les études cliniques.
Dans le cas de la censure à droite, la majorité des tests non-paramétriques sont basés sur des statistiques de rang étendues au cas censuré; parmi lesquels :

- le test du log-rank et ses variantes sont les plus utilisés : MANTEL [1966] étend le test de Mantel-Haenszel pour données non-censurées, cf. aussi PETO et PETO [1972].
- le test de Wilcoxon

5 6

[5] MANTEL, N. 1966, ≪Evaluation of survival data and two new rank order statistics arising in its consideration ≫, Cancer Chemotherapy Reports. Part 1, vol. 50, no 3, p. 163–170, ISSN 0069-0112. 24, 26

[6] PETO, R. et J. PETO. 1972, ≪Asymptotically Efficient Rank Invariant Test Procedures≫, Journal of the Royal Statistical Society. Series A (General), vol. 135, no 2, doi :10.2307/2344317, p. 185, ISSN 00359238.

5. Tests de rangs : comparaison de deux groupes

L'hypothèse nulle est $\mathcal{H}_0 : S_1 = S_2$ Egalité des survies dans les deux groupes.
On note $t_1 < \cdots < t_k$, les décès distincts et ordonnés dans les deux groupes.
À chaque t_i, on peut résumer les observations par le tableau suivant :

en t_i	décédés	vivants après t_i	total
groupe 1	m_{i1}	$n_{i1} - m_{i1}$	n_{i1}
groupe 2	m_{i2}	$n_{i2} - m_{i2}$	n_{i2}
total	m_i	$n_i - m_i$	n_i

A chaque temps t_i, on compare $m_{i\ell}$, le nombre de décès observé dans le groupe ℓ, à $e_{i,\ell}$, le nombre de décès que l'on aurait dû observer sous l'hypothèse $\mathcal{H}_0$ d'égalité des survie dans les deux groupes, $e_{i\ell} = n_{i\ell} \times \frac{m_i}{n_i}$.

$$U = \sum_{i=1}^{k} w_i (m_{i\ell} - e_{i\ell}) = \sum_{i=1}^{k} w_i \left(m_{i\ell} - m_i \times \frac{n_{i\ell}}{n_i} \right), \ \ell = 1 \text{ ou } 2.$$

où w_i est une pondération à choisir.

Le nombre de décès $m_{i\ell}$ en t_i est une v.a. qui suit une loi hypergéométrique (tirages sans remise des décès qui ont lieu en t_i) sous H_0 d'espérance

$$e_{i\ell} = \frac{m_i n_{i\ell}}{n_i} \quad \text{et de variance } v_i = m_i \times \frac{n_i - m_i}{n_i - 1} \times \frac{n_{i1} n_{i2}}{n_i^2}$$

.

Ainsi, sous $\mathcal{H}_0$,
$\mathbb{E}(U) = \sum_{i=1}^k w_i \mathbb{E}(m_{i\ell} - e_{i\ell}) = 0$ et $V(U) = \sum_{i=1}^k w_i^2 v_i$

$$\frac{U - E(U)}{\sqrt{V(U)}} \sim \mathcal{N}(0,1) \text{ (asymptotiquement)}$$

Et donc, $U^2/V(U) \sim \chi^2(1)$ (asymptotiquement).

En pratique, les choix les plus usuels des poids w_i:

- le test du log-rank proposé par Peto (1972): $w_i = 1, \forall i = 1, \cdots, k$. C'est la pondération la plus simple qui attribue le même poids à chaque décès.
- le test de Wilcoxon généralisé ou de Gehan (1965) $w_i = n_i, \forall i = 1, \cdots, k$ où n_i est le nombre d'individus à risque de décès à la date t_i. Les décès précoces ont un poids plus important que les décès tardifs dans la comparaison des 2 groupes. Ce test est donc utile pour montrer une différence entre les courbes de survie portant sur les décès les plus précoces.

Le test de Wilcoxon généralisé fait partie d'une famille de tests proposés par Fleming & Harrington [7] où le poids $w_i = \hat{S}(\tau)^\rho$. Ces tests appelés $G - rho$ tests sont implémentés dans la fonction `survdiff`.

[7] FLEMING, T. R. et D. P. HARRINGTON. 1991, Counting processes and survival analysis, 2e éd.,Wiley series in probability and mathematical statistics,Wiley, New York.

Rapport des risques instantanés

Pour quantifier la différence de risque de décès dans les 2 groupes, on peut compléter avec la quantité :

$$RR = \frac{O_1/E_1}{O_2/E_2}$$

qui est une estimation du rapport des risques instantanés de décès dans les deux groupes, avec

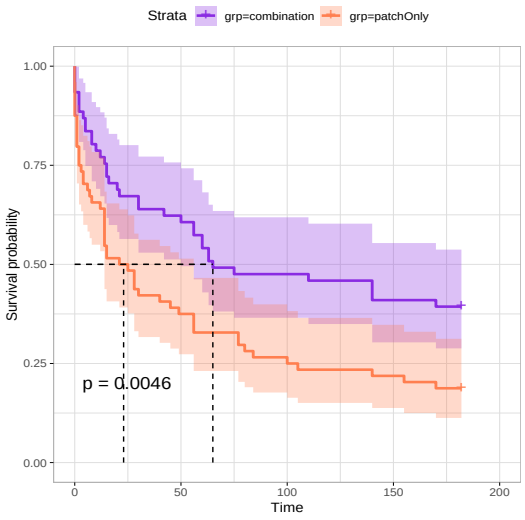
$$O_1 = \sum_{i=1}^{k} w_i m_{i1} \quad \text{et } E_1 = \sum_{i=1}^{k} w_i e_{i1}$$

et des expressions analogues pour O_2 et E_2.

Ainsi, si l'on a pu mettre en évidence une différence significative de survie entre les deux groupes, on peut interpréter ce rapport qui donne le risque de survenue de l'événement d'un groupe par rapport à l'autre.

```
survdiff(Surv(ttr,relapse)~grp,data=pharmacoSmoking)

                 N  Observed Expected (O-E)^2/E  (O-E)^2/V
grp=combination 61     23.1     32.1      2.53       8.01
grp=patchOnly   64     35.8     26.8      3.04       8.01

 Chisq= 8  on 1 degrees of freedom, p= 0.005
```

5. Tests de rangs : comparaison de deux groupes




```
                    N  Observed Expected   (O-E)^2/V
grp=combination    61  23.1     32.1         8.01
grp=patchOnly      64  35.8     26.8         8.01

 Chisq= 8  on 1 degrees of freedom, p= 0.005
```

La statistique du test du log-tank est égale à 8.01. On rejette l'égalité des survies dans les deux groupes de traitement combination *vs* PatchOnly avec une p-value de 0,0046.

Comparaison de L groupes avec $L > 2$: on souhaite tester
$\mathcal{H}_0 : S_1 = S_2 = \cdots = S_L$. (par ex. si le nombre de traitements est supérieur à 2)
Pour chaque groupe $\ell \in \{1, L\}$, on a :

$$e_{i\ell} = m_i \times \frac{n_{i\ell}}{n_i}, \quad i = 1, \cdots, k$$

et $\hat{\Sigma}$ = matrice de variance-covariance estimée de $\begin{pmatrix} \sum_{i=1}^k (m_{i1} - e_{i1}) \\ \cdots \\ \sum_{i=1}^k (m_{iL} - e_{iL}) \end{pmatrix}$

Sous $\mathcal{H}_0 : S_1 = S_2 = \cdots = S_L$, la statistique du log-rank pour L échantillons (avec $n_1, \cdots, n_L \to +\infty$:

$$\begin{pmatrix} \sum_{i=1}^k (m_{i1} - e_{i1}) \\ \cdots \\ \sum_{i=1}^k (m_{iL} - e_{iL}) \end{pmatrix}^T \hat{\Sigma}^{-1} \begin{pmatrix} \sum_{i=1}^k (m_{i1} - e_{i1}) \\ \cdots \\ \sum_{i=1}^k (m_{iL} - e_{iL}) \end{pmatrix} \overset{\mathscr{L}}{\longrightarrow} \chi^2(L-1)$$

Extension des tests de rangs pour des données censurées par intervalles :
Trois packages proposent des tests adaptés à la censure par intervalle :

- `interval` par Fay et Shaw [2010],
- `glrt` par Zhao et Sun [2015],
- `FHtest` par Oller et Langhor [2015].

Les différentes généralisations de tests de rangs (et leur implémentation sous R) pour la comparaison de groupes sont très bien documentées dans la thèse de Sarah Flora Jonas [2018][8].

[8] Jonas, S. (2018) Méthodes de comparaisons de deux ou plusieurs groupes de données censurées par intervalle avec application en immunologie clinique. Thèse de doctorat de l'Université Paris Saclay.

Références

- Hall, W. J. and Wellner,(1980), J. A. Confidence Bands for a Survival Curve from Censored Data. Biometrika 67 : 133–143.
- Huber-Carol C., (1994), Durées de survie tronquées et censurées, Journal de la société statistique de Paris, tome 135, 4 (1994), p. 3-23
- Jonas, S. (2018) Méthodes de comparaisons de deux ou plusieurs groupes de données censurées par intervalle avec application en immunologie clinique. Thèse de doctorat de l'Université Paris Saclay.
- J.P. Klein and M. L. Moeschberger., (2005) Survival analysis: techniques for censored and truncated data. Springer Science & Business Media.
- Moore, Dirk, F., (2016) Applied Survival Analysis Using R, Springer.
- Nair, V. N., (1984) Confidence Bands for Survival Functions with Censored Data: A Comparative Study. Technometrics 14 : 265–275.
- Steinberg et al. (2009), Triple-combination pharmacotherapy for medically ill smokers: A randomized trial. Annals of Internal Medicine 150, 447-454.

Programme du Cours

Partie 1 (Séances 1 et 2) :Différents types de censure, estimation de la fonction de survie, tests de rangs.
TP1 à rendre.

Partie 2 (Séances 3 et 4) : Modèles paramétriques pour l'ADV et régression log-linéaires en présence de covariables (Accelerated Failure Time models)
TP2 à rendre.

Partie 3 (Séances 5 et 6) : Modèle de Cox.
TP3 à rendre.

Partie 2

1. Modèles paramétriques pour une variable de durée

2. Modèles de régression ou de survie accélérée

Rappel : La fonction de survie

La fonction de survie est la probabilité que l'événement d'intérêt pour un individu, se produise au-delà d'une durée x :

$$S(x) = P(X > x)$$

La fonction de survie est donc aussi $S(x) = 1 - F(x)$ où $F(x)$ désigne la fonction de répartition de la v.a. X et lorsque X est une v.a. continue, la fonction de survie s'exprime aussi à l'aide de la fonction de densité $f(x)$ de la façon suivante :

Exemple de fonctions de survie

La fonction de survie d'une loi de Weibull est $S(x) = \exp(-\lambda x^{\alpha})$.

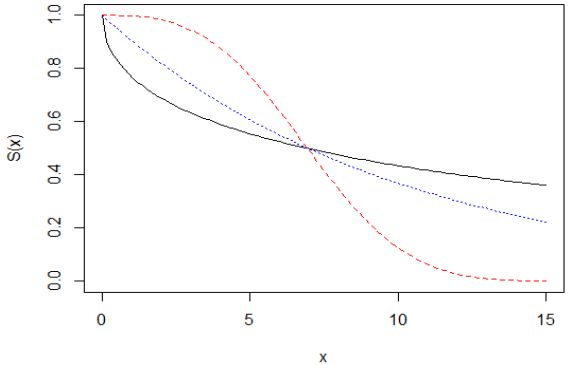


Figure 1: $\alpha = 0.5$, $\lambda = 0.26$ (noir), $\alpha = 1$, $\lambda = 0.1$ (pontillés bleus), $\alpha = 3$, $\lambda = 0.002$ (tirets rouges).

Propriétés de la fonction de survie

La fonction de survie $S(x)$ est :

- monotone décroissante
- $S(0) = 1$ et $\lim_{x \to +\infty} S(x) = 0$.

La vitesse de décroissance vers 0 dépend du risque de l'apparition de l'événement d'intérêt au point x. Il est difficile en pratique de l'évaluer par un examen graphique de la courbe de survie.

Rappel : La fonction de risque instantané

Une fonction fondamentale en analyse de durées de vie est la fonction de risque instantané. Elle est désignée aussi sous le terme taux de défaillance (en fiabilité), taux de mortalité (si l'événement d'intérêt est le décès) ou taux de hasard (anglicisme pour "hazard rate"). Elle est définie par :

$$\lambda(x) = \lim_{\Delta x \to 0} \frac{P\left(x \leq X < x + \Delta x | X \geq x\right)}{\Delta x}$$

$\lambda(x)\Delta x \approx P\left(x \leq X < x + \Delta x | X \geq x\right)$ s'interprète comme la probabilité que l'événement d'intérêt se produise juste après x sachant qu'il ne s'est pas produit auparavant.

Exemples de fonctions de risque

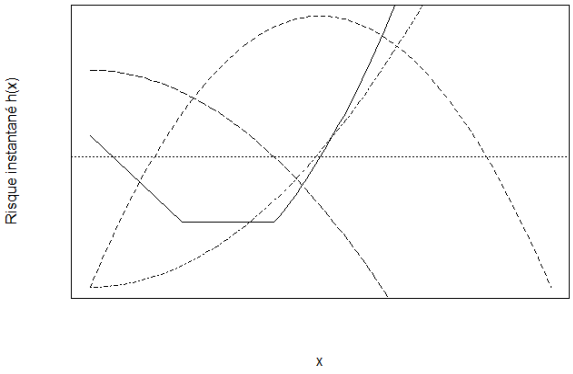


Figure 2: Formes pour différentes fonctions de risque instantané.

La fonction de risque instantané donne une meilleure information qualitative sur le mécanisme qui génère l'apparition de l'événement d'intérêt que la fonction de survie. Pour cette raison, estimer la fonction de risque intantané $\lambda(x)$ est une question centrale en analyse des durées de vie.

Soit X une variable aléatoire de durée de densité f, de fonction de survie S et de fonction de risque instantané λ.

Modèle exponentiel $\gamma(1, \lambda)$, $\lambda > 0$

- fonction de risque constant $\lambda(t) = \lambda$
- densité $f(t) = \lambda \exp(-\lambda t)\mathbb{I}_{(t \geq 0)}$
- fonction de survie $S(t) = \exp(-\lambda t)$
- $\mathbb{E}(T) = 1/\lambda$ $\mathbb{V}(T) = 1/\lambda^2$

Paramétrage dans R : $f(t) = \lambda \exp(-\lambda t)$, pour $t \geq 0$

Modèle Gamma $\gamma(\beta, \lambda)$, $\beta, \lambda > 0$

Elle généralise la loi exponentielle avec $\beta = 1$.

β : paramètre de forme, λ : paramètre d'échelle

- densité $f(t) = \dfrac{\lambda^\beta}{\Gamma(\beta)} t^{\beta-1} e^{-\lambda t}, \quad t > 0$

  avec $\Gamma(\beta) = \int_0^{+\infty} x^{\beta-1} e^{-x} dx$
- fonction de risque $\lambda(t) = \dfrac{t^{\beta-1} e^{-\lambda t}}{\int_0^{+\infty} x^{\beta-1} e^{-\lambda x} dx}, \quad t > 0$
- fonction de survie $S(t) = \int_t^{+\infty} \frac{\lambda^\beta}{\Gamma(\beta)} x^{\beta-1} e^{-\lambda x}$
- $\mathbb{E}(T) = \beta/\lambda$ $\mathbb{V}(T) = \beta/\lambda^2$

Paramétrage dans R : $f(t) = \frac{1}{\sigma^\alpha \Gamma(\alpha)} t^{\alpha-1} e^{-t/\sigma}$, pour $t \geq 0$, $\alpha > 0$ (shape) et $\sigma > 0$ (scale). $(\beta, \lambda) \mapsto (\alpha, \frac{1}{\sigma})$

Pour le modèle Gamma:

- Si $\beta > 1$, $\lambda(t)$ est monotone croissante
- Si $\beta < 1$, $\lambda(t)$ est monotone décroissante
- Si $\beta = 1$, on retrouve le modèle exponentiel de paramètre $\lambda > 0$: $\lim_{t \to +\infty} \lambda(t) = \lambda$: La droite $y = \lambda$ est asymptote de la courbe de risque.

Modèle Weibull $\mathcal{W}(\alpha, \lambda)$, $\alpha, \lambda > 0$

La v.a. de durée X suit une loi de Weibull $\mathcal{W}(\alpha, \lambda)$ si X^α suit une loi exponentielle $\gamma(1, \lambda)$.

- densité $f(t) = \alpha\lambda t^{\alpha-1} e^{-\lambda t^\alpha}, \quad t > 0$
- fonction de survie $S(t) = e^{-\lambda t^\alpha}$
- fonction de risque $\lambda(t) = \alpha\lambda t^{\alpha-1}, \quad t > 0$

Paramétrage dans R : $f(t) = \frac{a}{b}(t/b)^{a-1} e^{-(t/b)^a}$, pour $t \geq 0$, $a > 0$ (shape) et $b > 0$ (scale). $(a, b) \mapsto (a, \frac{1}{b^a}) = (\alpha, \lambda)$

Pour le modèle Weibull:

- Si $\alpha > 1$, $\lambda(t)$ est monotone croissante:
- Si $\alpha < 1$, $\lambda(t)$ est monotone décroissante
- Si $\alpha = 1$, on retrouve le modèle exponentiel.
- pas d'asymptote.

Modèle Log-normal $\mathcal{LN}(m, \sigma^2)$

La v.a. de durée X suit une loi Log-normale $\mathcal{LN}(m, \sigma^2)$ si $\log X$ suit une loi normale $\mathcal{N}(m, \sigma^2)$ avec m: paramètre de position et σ: paramètre d'échelle.

Si φ et Φ désignent respectivement la densité et la f.d.r. d'une v.a. $\mathcal{N}(0, 1)$ alors on a

- densité $f(t) = \dfrac{1}{t\sigma}\varphi\left(\dfrac{\log t - m}{\sigma}\right), \quad t > 0$
- fonction de survie $S(t) = 1 - \Phi\left(\frac{\log t - m}{\sigma}\right)$
- fonction de risque $\lambda(t) = \dfrac{1}{t\sigma}\varphi\left(\dfrac{\log t - m}{\sigma}\right) / \left[1 - \Phi\left(\dfrac{\log t - m}{\sigma}\right)\right], \quad t > 0$

Modèle Log-logistique

La v.a. de durée X suit une loi Log-logistique de paramètres $\alpha, \lambda > 0$ si sa densité est donnée par :

- $f(t) = \dfrac{\alpha\lambda t^{\alpha-1}}{(1+\lambda t^{\alpha})^2}, \quad t > 0$
- fonction de survie $S(t) = \dfrac{1}{1+\lambda t^{\alpha}}$
- fonction de risque $\lambda(t) = \dfrac{\alpha t^{\alpha-1}}{1+\lambda t^{\alpha}}, \quad t > 0$

1.2 Construction de la vraisemblance avec censure à droite

Observations : $(T_1, \delta_1) \cdots, (T_n, \delta_n)$ i.i.d avec $T_i = \min(X_i, C_i)$ et $\delta_i = 1(X_i \leq C_i)$.

Sous l'hypothèse que X et C sont indépendantes (censure non informative) et si l'on note g la densité de C et G la survie de C, alors la vraisemblance s'exprime :

$$\begin{aligned} L(\theta) &= \left\{ \prod_{i=1}^{n} G(T_i)^{\delta_i} g(T_i)^{1-\delta_i} \right\} \left\{ \prod_{i=1}^{n} f_\theta(T_i)^{\delta_i} S_\theta(T_i)^{1-\delta_i} \right\} \\ &\propto \prod_{i=1}^{n} f_\theta(T_i)^{\delta_i} S_\theta(T_i)^{1-\delta_i} \end{aligned}$$

preuve : voir chap. 3 p. 76 dans Klein & Moeschberger (2005).

cas des données complètes :

On dispose de l'échantillon complètement observé des durées de vie $X_1, \cdots, X_n$. Le paramètre à estimer est $\theta = \lambda > 0$:

$$\log L(\lambda) = n \log(\lambda) - \lambda \sum_{i=1}^{n} X_i$$

L'E.M.V. (estimateur du maximum de vraisemblance) est :

$$\hat{\lambda}_n = \frac{n}{\sum_{i=1}^{n} X_i}$$

cas des données complètes :
D'après le T.C.L, on a :

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{\lambda}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \frac{1}{\lambda^2})$$

D'où l'on déduit, grâce à la δ-méthode que :

$$\sqrt{n}\left(\hat{\lambda}_n - \lambda\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \lambda^2)$$

Rappel : delta méthode

La δ-méthode permet de déduire un T.C.L pour une suite de v.a. $(\ell(X_n))$ à partir d'un T.C.L sur la suite (X_n) à condition d'avoir des hyp. de régularité sur ℓ :
Soit (X_n) une suite de v.a. définies sur le même espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $\mathbb{R}^d$. Soient $x_0 \in \mathbb{R}^d$, Σ matrice $d \times d$ et (a_n) une suite réelle qui tend vers $+\infty$ tels que :

$$a_n(X_n - x_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma)$$

Alors pour toute fonction ℓ définie de $\mathbb{R}^d$ dans $\mathbb{R}^p$ de classe $\mathcal{C}^1$ au voisinage de x_0, avec $\ell = (\ell_1, \cdots, \ell_p)$ on a :

$$a_n\left(\ell(X_n) - \ell(x_0)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, D_\ell(x_0)\Sigma D_\ell(x_0)^T)$$

avec $D_\ell(x)$ la matrice $p \times d$ définie par $[D_\ell(x)]_{i,j} = \frac{\partial \ell_i}{\partial x_j}(x)$

cas des données complètes :

Application de la $\delta -$ *méthode* :

$$g(X_n) = \hat{\lambda}_n = n/\sum_{i=1}^{n} X_i$$

$$g(x) = \frac{1}{x}$$
$$x_0 = 1/\lambda$$

Et on montre la normalité asymptotique de $\hat{\lambda}_n$ à partir de celle de $\bar{X}_n$.

à faire en exercice.

cas des données censurées :

On dispose de l'échantillon censuré à droite des durées de vie : les observations sont $(T_1, \delta_1) \cdots, (T_n, \delta_n)$ i.i.d avec $T_i = \min(X_i, C_i)$ et $\delta_i = 1(X_i \leq C_i)$. Le paramètre à estimer est $\theta = \lambda > 0$:

$$\log L(\lambda) = \left(\sum_{i=1}^{n} \delta_i \right) \log(\lambda) - \lambda \sum_{i=1}^{n} T_i + cste$$

L'E.M.V. (estimateur du maximum de vraisemblance) est :

$$\hat{\lambda}_n = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} T_i}$$

cas des données censurées :

On a la convergence de l'estimateur $\hat{\lambda}_n$ vers λ p.s.

à faire en exercice.

cas des données censurées :

On a le résultat suivant (T.C.L) multidimensionnel de v.a. i.i.d :
Soient

$$U_n = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} T_i \\ \frac{1}{n}\sum_{i=1}^{n} \delta_i \end{pmatrix} \quad \text{et} \quad U = \begin{pmatrix} \mathbb{E}[T_1] \\ \mathbb{E}[\delta_1] \end{pmatrix}$$

Alors

$$\sqrt{n}(U_n - U) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Gamma)$$

où

$$\Gamma = \begin{pmatrix} V(T_1) & \mathrm{cov}(T_1, \delta_1) \\ \mathrm{cov}(T_1, \delta_1) & V(\delta_1) \end{pmatrix}$$

cas des données censurées :

On déduit du T.C.L multidimensionnel et de la δ-méthode :

$$\sqrt{n}\left(\hat{\lambda}_n - \lambda\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\lambda^2}{\mathbb{P}(\delta_1 = 1)}\right)$$

On remarque que si $\mathbb{P}(\delta_1 = 1) = 1$, on retrouve la variance asymptotique du cas non censuré.

à faire en exercice.

Considérons une durée de vie $X > 0$ et un vecteur $\mathbf{Z}^t = (Z_1, \cdots, Z_p)$ de variables explicatives qui lui sont associées. Le vecteur $\mathbf{Z}^t$ peut inclure :

- des variables quantitatives (pression artérielle, âge, poids)
- et/ou qualitatives (genre, traitement, stade de la maladie)
- et/ou des variables dépendant du temps x (série de mesures au cours du traitement).

Deux approches sont devenues très populaires pour modéliser l'effet des covariables sur la durée de vie : les **modèles de survie accélérée** et les **modèles à risques proportionnels ou de Cox**.

On généralise le modèle linéaire classique en posant $Y = \log X$. C'est une transformation naturelle pour étalonner des observations positives sur la droite réelle toute entière. Puis, on pose :

$$Y = \mu + \gamma^t \mathrm{Z} + \sigma W$$

où $\mu \in \mathbb{R}$, $\sigma > 0$ et $\gamma^t = (\gamma_1, \cdots, \gamma_p)$ est un vecteur de coefficients réels et W est l'erreur aléatoire. Des choix possibles pour la loi de l'erreur W sont une loi gaussienne, une loi de Weibull , une loi logistique, etc.

Notons $S_0(x)$ la fonction de survie de la v.a. $\exp(\mu + \sigma W)$.
On peut écrire la survie de X étant fixé un vecteur de covariables Z :

$$\begin{aligned} P(X > x|\mathrm{Z}) &= P(Y > \log x|\mathrm{Z}) \\ &= P\left(\mu + \sigma W > \log x - \gamma^t \mathrm{Z}|\mathrm{Z}\right) \\ &= P\left(e^{\mu+\sigma W} > x\exp(-\gamma^t \mathrm{Z})|\mathrm{Z}\right) \\ &= S_0(x\exp(-\gamma^t \mathrm{Z})) \end{aligned}$$

L'effet des covariables Z est de multiplier la durée x par un facteur $\exp(-\gamma^t \mathrm{Z})$. Ainsi, en fonction du signe de $\gamma^t \mathrm{Z}$, la durée x est "accélérée" ou "ralentie" par un facteur constant. D'où le terme de **Accelerated failure time models (AFT)**.

Exemple : écrire les fonctions de survie dans chaque groupe si $\mathrm{Z} = 0$ (groupe 1) et $\mathrm{Z} = 1$ (groupe 2).

Soit X la durée de vie d'intérêt et Z un vecteur de covariables fixes au cours du temps.

Autrement dit :

$$S(x|\mathrm{Z}) = S_0\left(\exp(-\gamma^t \mathrm{Z})x\right), \quad \text{pour tout } x.$$

La survie $S(x|\mathrm{Z})$ d'un individu avec une covariable Z au temps x est la même que celle d'un individu avec une survie S_0 au temps $x\exp(-\gamma^t \mathrm{Z})$ (propriété de survie accélérée). La fonction de risque instantané s'écrit alors :

$$\lambda(x|\mathrm{Z}) = \exp(-\gamma^t \mathrm{Z})\lambda_0[\exp(-\gamma^t \mathrm{Z})x], \quad \text{pour tout } x.$$

2.1 Modèle Weibull

Une grande variété de modèles paramétriques peut être utilisée pour modéliser la loi de X. La loi de Weibull est un modèle flexible qui permet d'avoir une fonction de risque monotone croissante, décroissante ou constante. On pose :

$$\ln(X) = \mu + \gamma^t \mathrm{Z} + \sigma W \quad \text{où } W \sim f_W(w) = \exp(w - e^w)$$
$$\text{et } S_W(w) = \exp(-e^w)$$
$$\text{(loi des valeurs extrêmes).}$$

Alors, la fonction de survie $S(x|\mathrm{Z})$ de X est la survie d'une loi de Weibull de paramètres λ et α :

$$S(x|\mathrm{Z}) = \exp(-\lambda x^\alpha)$$

avec $\lambda = \exp[(-\mu - \gamma^t \mathrm{Z})/\sigma]$ et $\alpha = 1/\sigma$.

ou bien $S(x|\mathrm{Z}) = \exp(-(\frac{x}{b})^a)$ avec $a = \frac{1}{\sigma}$ et $b = \exp(\mu)$.

à faire en exercice.

2.1 Modèle Weibull

La loi d'une durée de vie X, étant donné un vecteur de covariables Z satisfait l'hypothèse des risques proportionnels si pour deux individus i_1 et i_2 différents pour une seule des covariables Z_k, les autres étant égales par ailleurs, alors pour tout x

$$\frac{\lambda(x|z_{k,i_1})}{\lambda(x|z_{k,i_2})} = cste \text{ qui ne dépend pas de } x.$$

Propriété

Le modèle log-linéaire $Y = \ln(X) = \mu + \gamma^t Z + \sigma W$ où W suit la loi des valeurs extrêmes satisfait l'hypothèse des risques proportionnels pour X et sa fonction de risque est donnée par :

$$\lambda(x|Z) = \frac{1}{\sigma} \exp\left(\frac{-\mu - \gamma^t Z}{\sigma}\right) x^{1/\sigma - 1} = (\alpha \lambda x^{\alpha-1}) \exp(-\gamma^t Z/\sigma)$$

où $\alpha = 1/\sigma$, $\lambda = \exp(-\mu/\sigma)$.

La vraisemblance du modèle pour des observations censurées à droite s'écrit :

$$\begin{aligned} L &= \prod_{i=1}^{n} [f_Y(y_i)]^{\delta_i} [S_Y(y_i)]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[\frac{1}{\sigma} f_W \left(\frac{y_i - \mu - e^{\gamma^t \mathbf{Z}}}{\sigma} \right) \right]^{\delta_i} \left[S_W \left(\frac{y_i - \mu - e^{\gamma^t \mathbf{Z}}}{\sigma} \right) \right]^{1-\delta_i} \end{aligned}$$

Les estimateurs du maximum de vraisemblance de μ et σ s'obtiennent par résolution numérique du problème d'optimisation.

Les fonctions `survreg` et `flexsurvreg` de R permettent d'obtenir les estimations des paramètres par maximum de vraisemblance. en l'absence de covariables, elles peuvent être utilisées pour estimer les paramètres de divers modèles paramétriques classiques en survie (cf. TP2).

2.2 Modèle Log-logistique

La fonction de survie et la fonction de risque cumulé d'une loi Log-logistique sont définies par :

$$S_X(x) = \frac{1}{1+\lambda x^\alpha} \quad \text{et} \quad \Lambda_X(x) = \ln(1+\lambda x^\alpha).$$

On pose

$$\ln(X) = \mu + \gamma^t Z + \sigma W \qquad \text{où } W \sim f_W(w) = e^w/(1+e^w)^2$$
$$\text{et } S_W(w) = 1/(1+e^w) \text{ (loi logistique).}$$

Alors la fonction de survie de X est log-logistique et s'écrit, en posant à nouveau $\alpha = 1/\sigma$ et $\lambda = \exp[(-\mu - \gamma^t Z)/\sigma]$:

$$S(x|Z) = \frac{1}{1+\lambda x^\alpha}$$

Pour passer de (μ, σ) à (α, λ), on effectue le même changement de variables que pour la loi de Weibull.

En présence de covariable Z, on peut définir dans les modèles AFT la notion de résidus en mimant la notion de résidus en régression linéaire classique de la façon suivante :

$$R_j = \hat{\Lambda}_{(\hat{\mu},\hat{\sigma},\hat{\gamma})}(T_j|\mathrm{Z}_j) \qquad \text{où } \hat{\Lambda}_{(\hat{\mu},\hat{\sigma},\hat{\gamma})} \text{ est l'estimateur paramétrique de } \Lambda.$$

Heuristique : Si Λ est la fonction de risque cumulé des X_j alors les $R_j = \Lambda(T_j|\mathrm{Z}_j)$ forment un échantillon censuré de loi exponentielle de paramètre 1.
Les résidus R_j sont appelés résidus de Cox-Snell.

On construit alors l'estimateur de Nelson-Aalen de la fonction de risque cumulé Λ des R_j et on le compare graphiquement à $\Lambda(x) = x$. (cf. TP2)

Références

– J.P. Klein and M. L. Moeschberger., (2005) Survival analysis: techniques for censored and truncated data. Springer Science & Business Media.

Programme du Cours

Partie 1 (Séances 1 et 2) : Différents types de censure, estimation de la fonction de survie, tests de rangs.
TP1 à rendre.

Partie 2 (Séances 3 et 4) : Modèles paramétriques pour l'ADV et régression log-linéaires en présence de covariables (Accelerated Failure Time models)
TP2 à rendre.

Partie 3 (Séances 5 et 6) : Modèle de Cox.

Partie 3

1. Définition du Modèle de Cox
2. Vraisemblance partielle
3. Distribution asymptotique
4. Tests sur les paramètres
5. Modèle de Cox : vérification des hypothèses

Notons $h(t|\boldsymbol{Z})$ la fonction de risque instantané de décès au temps t pour un individu dont les covariables sont $\boldsymbol{Z}^t = (Z_1, \cdots, Z_p)$ (vecteur de taille p). Dans le modèle à risques proportionnels, la fonction de risque instantané (hazard rate) s'écrit :

$$h(t|\boldsymbol{Z}) = h_0(t)\exp(\beta^t \boldsymbol{Z}) = h_0(t)\exp\Big(\sum_{j=1}^{p}\beta_j Z_j\Big)$$

où

– h_0 est une fonction de risque instantané (inconnue) appelée risque de base (baseline hazard rate).

– $\beta = (\beta_1, \cdots, \beta_p)^t$ est un vecteur de paramètres inconnu de $\mathbb{R}^p$.

Propriété fondamentale du modèle de Cox

Notons i_1 et i_2 deux individus possédant les mêmes covariables sauf la k-ème alors :

$$\frac{h(t|\boldsymbol{Z}_{i_1})}{h(t|\boldsymbol{Z}_{i_2})} = \frac{h_0(t)\exp(\beta^t \boldsymbol{Z}_{i_1})}{h_0(t)\exp(\beta^t \boldsymbol{Z}_{i_2})} = \exp\Big((Z_{k,i_1} - Z_{k,i_2})\beta_k\Big)$$

où $\boldsymbol{Z}_i^t = (Z_{1,i}, \cdots, Z_{p,i})$ est le vecteur des covariables de l'individu i.

Le modèle de Cox vérifie l'hypothèse des risques proportionnels.

Si les individus i_1 et i_2 ont une seule covariable $Z_{k,i_1} \neq Z_{k,i_2}$ qui diffère, toutes les autres étant égales par ailleurs ($Z_{j,i_1} = Z_{j,i_2}$ pour $j \neq k$) alors,

$$\frac{h(t|\boldsymbol{Z}_{i_1})}{h(t|\boldsymbol{Z}_{i_2})} = \exp\Big((Z_{k,i_1} - Z_{k,i_2})\beta_k\Big)$$

En particulier, pour une covariable Z_k binaire qui code le traitement : par exemple, l'individu i_1 reçoit le placebo $Z_{k,i_1} = 1$ et l'individu i_2 reçoit le traitement $Z_{k,i_2} = 0$:

$$\frac{h(t|\boldsymbol{Z}_{i_1})}{h(t|\boldsymbol{Z}_{i_2})} = \exp(\beta_k)$$

C'est le risque de survenue de l'événement du groupe placebo par rapport au groupe traité.

Cox Model pour les données PharmacoSmoking : traitement

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

#Call:
#coxph(formula = Surv(ttr, relapse) ~ grp, data = pharmacoSmoking)

# n= 125, number of events= 89

#               coef exp(coef) se(coef)    z     Pr(>|z|)
#grppatchOnly 0.6050    1.8313   0.2161   2.8   0.00511 **
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#             exp(coef) exp(-coef) lower .95 upper .95
#grppatchOnly     1.831     0.5461     1.199     2.797
```

La valeur $\exp(\beta_j)$ associée à la covariable k représente le rapport de risque :

- lorsque $Z_{k,i_1} = Z_{k,i_2} + 1$
- et toutes les autres covariables étant égales par ailleurs : $Z_{j,i_1} = Z_{j,i_2}$ pour $j \neq k$

le coefficient $\exp(\beta_j)$ s'interprète comme le rapport de risque associé à une augmentation de la $k$ème covariable de 1 unité.

Cox Model pour les données PharmacoSmoking : traitement et âge

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

#Call:
#coxph(formula = Surv(ttr, relapse) ~ grp + age, data = pharmacoSmokin

# n= 125, number of events= 89

#                  coef exp(coef)  se(coef)      z    Pr(>|z|)
#grppatchOnly  0.558663  1.748334  0.216674  2.578    0.00993 **
#age          -0.023018  0.977245  0.009605 -2.397    0.01655 *
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#             exp(coef) exp(-coef) lower .95   upper .95
#grppatchOnly    1.7483      0.572     1.143    2.6734
#age             0.9772      1.023     0.959    0.9958
```

En l'absence d'ex-æquo :
la vraisemblance partielle de Cox s'écrit :

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp\left(\beta^t \boldsymbol{Z}_{(i)}\right)}{\sum_{j \in \mathcal{R}_i} \exp\left(\beta^t \boldsymbol{Z}_j\right)}$$

- $t_1 < t_2 < \cdots < t_k$ sont les k durées distinctes où se produisent les D décès. (comme il n'y a pas d'ex-æquo, ici $D = k$).
- $\boldsymbol{Z}_{(i)}$ est le vecteur des covariables du patient qui décède en t_i. $\boldsymbol{Z}_j$ est le vecteur des covariables du patient j pour $j \in \mathcal{R}_i$.
- $\mathcal{R}_i$ l'ensemble des individus exposés au risque de décès en t_i.

Remarque : Le numérateur dépend seulement de l'individu pour lesquel le décès est observé en t_i, alors que le dénominateur utilise toute l'information des individus qui ne sont ni décédés, ni censurés en t_i.

2. Vraisemblance partielle

On parle de vraisemblance partielle, car seuls les sujets subissant l'événement étudiés entrent dans le calcul, les sujets censurés n'étant considérés qu'indirectement, voir ([9])

La vraisemblance partielle considère la probabilité qu'en t_i, un individu donné subisse l'événement plutôt qu'un autre individu exposé au risque au même instant. Elle est obtenue en prenant le produit de toutes ces probabilités sur tous les décès observés :

$$L(\beta) = \prod_{i=1}^{D} P(\text{un individu donné décède en } t_i | \text{un décès se produit en } t_i)$$

[9] Klein & Moeshberger [2005], cf. paragraphe 8.3, p. 257 : "Theoretical Notes".

En l'absence d'ex-æquo :

Soient les événements $\mathcal{A}_j$: "l'individu j décède en t_i" $(j \in \mathcal{R}_i)$
$\mathcal{S}_j$: "l'individu j a survécu jusqu'en t_i et n'est pas censuré" donc $\mathcal{S}_j = (T_j > t_i)$,
Notons $i_0 \in \mathcal{R}_i$, l'indice de l'individu qui décède en t_i alors on peut écrire :

$$
\begin{aligned}
& P(\text{l'individu } i_0 \text{ décède en } t_i | \text{un décès se produit en } t_i) \\
= \quad & P(\mathcal{A}_{i_0} | \bigcup_{j \in \mathcal{R}_i} \mathcal{A}_j) = \frac{P(\mathcal{A}_{i_0} \cap \bigcup_{j \in \mathcal{R}_i} \mathcal{A}_j)}{P(\bigcup_{j \in \mathcal{R}_i} \mathcal{A}_j)} = \frac{P(\mathcal{A}_{i_0})}{P(\bigcup_{j \in \mathcal{R}_i} \mathcal{A}_j)} \quad \text{car } \mathcal{A}_{i_0} \subset \bigcup_{j \in \mathcal{R}_i} \mathcal{A}_j
\end{aligned}
$$

En l'absence d'ex-æquo

$$
\begin{aligned}
& P(\text{l'individu } i_0 \text{ décède en } t_i | \text{un décès se produit en } t_i) \\
= \; & = \frac{P(\mathcal{A}_{i_0})}{\sum_{j \in \mathcal{R}_i} P(\mathcal{A}_j)} \text{ car les } \mathcal{A}_j \text{ disjoints} \\
= \; & \frac{P(\mathcal{A}_{i_0} \cap \mathcal{S}_{i_0})}{\sum_{j \in \mathcal{R}_i} P(\mathcal{A}_j \cap \mathcal{S}_j)} \text{ car } \mathcal{A}_j \subset \mathcal{S}_j \, \forall j \\
= \; & \frac{P(\mathcal{A}_{i_0} | \mathcal{S}_{i_0})}{\sum_{j \in \mathcal{R}_i} P(\mathcal{A}_j | \mathcal{S}_j)} \text{ car } P(\mathcal{S}_j) = P(T \geq t_i) \, \forall j \\
\approx \; & \frac{h(t_i | \boldsymbol{Z}_{i_0})}{\sum_{j \in \mathcal{R}(t_i)} h(t_i | \boldsymbol{Z}_j)} = \frac{h_0(t_i) \exp(\beta^t \boldsymbol{Z}_{i_0})}{\sum_{j \in \mathcal{R}(t_i)} h_0(t_i) \exp(\beta^t \boldsymbol{Z}_j)} \text{ car } P(\mathcal{A}_j | \mathcal{S}_j), \approx h(t_i | \boldsymbol{Z}_j) \delta t_i \, \forall j \\
= \; & \frac{\exp(\beta^t \boldsymbol{Z}_{(i)})}{\sum_{j \in \mathcal{R}(t_i)} \exp(\beta^t \boldsymbol{Z}_j)} \text{ car } \boldsymbol{Z}_{i_0} = \boldsymbol{Z}_{(i)}
\end{aligned}
$$

En l'absence d'ex-æquo

La vraisemblance partielle est obtenue en prenant le produit de toutes ces probabilités conditionnelles sur tous les décès observés :

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{D} P(\text{un individu décède en } t_i | \text{un décès se produit en } t_i) \\ &= \prod_{i=1}^{D} \frac{\exp(\beta^t \boldsymbol{Z}_{(i)})}{\sum_{j \in \mathcal{R}(t_i)} \exp(\beta^t \boldsymbol{Z}_{(i)})} \end{aligned}$$

Généralisation en présence d'ex-æquo [10] :

- Breslow [1974] :

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp\left(\sum_{j\in\mathcal{D}_i} \beta^t \boldsymbol{Z}_{(j)}\right)}{\sum_{j\in\mathcal{R}_i} \exp\left(\beta^t \boldsymbol{Z}_j\right)^{m_i}}$$

où $\mathcal{D}_i$ est l'ensemble des indices des individus qui décèdent en t_i et $m_i \geq 1$ le nombre de décès observés en t_i, et $D = \sum_{i=1}^{k} m_i$.

- Efron [1977] :

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp(\sum_{j\in\mathcal{D}_i} \beta^t \boldsymbol{Z}_{(j)})}{\prod_{\ell=1}^{m_i} \left(\sum_{j\in\mathcal{R}_i} \exp(\beta^t \boldsymbol{Z}_j) - \frac{\ell-1}{m_i} \sum_{j\in\mathcal{D}_i} \exp(\beta^t \boldsymbol{Z}_{(j)})\right)}$$

Remarque : Lorsque le nombre d'ex-æquo est faible, les deux expressions sont très proches.

[10] Klein & Moeshberger [2005], cf. paragraphe 8.4, p. 259 : "Partial Likelihood when ties are present".

L'estimateur $\hat{\beta}$ du maximum de vraisemblance (partielle) du paramètre β est défini par:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\max}\, L(\beta)$$

En dérivant la log-vraisemblance partielle par rapport à β, on peut exprimer le vecteur de score :

$$U_p(\beta) = \sum_{i=1}^{D} \left[\boldsymbol{Z}_{(i)} - \frac{\sum_{j \in \mathcal{R}_i} \boldsymbol{Z}_j \exp\left(\beta^t \boldsymbol{Z}_j\right)}{\sum_{j \in \mathcal{R}_i} \exp\left(\beta^t \boldsymbol{Z}_j\right)} \right]$$

Pour maximiser la vraisemblance partielle (soit annuler le vecteur score partiel), et ainsi déterminer les paramètres $\hat{\beta}$, l'algorithme de Newton-Raphson est très souvent utilisé (fonction `coxph` de R) et on peut aussi calculer numériquement une approximation $\mathcal{I}_n(\hat{\beta})$ de la matrice d'information de Fisher :

$$\mathcal{I}(\beta) = -\mathbb{E}\left[\frac{\partial^2 \log(L(\beta))}{\partial \beta_j \partial \beta_k} \right]_{1 \leq j,k \leq p}$$

3. Distribution asymptotique

On a les résultats asymptotiques suivants qui permettent de construire des tests :

- $$\frac{1}{\sqrt{n}} U_p(\beta) \underset{n \to \infty}{\overset{\mathscr{L}}{\longrightarrow}} \mathcal{N}(\vec{0}; \mathcal{I}(\beta)) \text{ (test du score)}$$
- $$\sqrt{n}(\hat{\beta} - \beta) \underset{n \to \infty}{\overset{\mathscr{L}}{\longrightarrow}} \mathcal{N}(\vec{0}; \mathcal{I}(\beta)^{-1}) \text{ (test de Wald)}$$
- $$-2\left(\log L(\hat{\beta}) - \log L(\beta)\right) \underset{n \to \infty}{\overset{\mathscr{L}}{\longrightarrow}} \chi^2_p \text{ (test du rapport de vraisemblance)}$$

Soit $\hat{\sigma}_j^2$ le j-ème élément diagonal de la matrice $\mathcal{I}_n(\hat{\beta})$.

Le test de l'hypothèse nulle $\mathcal{H}_0 : \beta_j = 0$ au risque α s'appuie sur la statistique de Wald:

$$W = \frac{(\hat{\beta}_j - \beta_j)^2}{\hat{\sigma}_j^2}$$

qui suit une loi de χ_1^2 sous $\mathcal{H}_0$.

On rejette $\mathcal{H}_0$ si la valeur de la statistique de test dépasse le quantile d'ordre $1 - \alpha$ de la loi χ_1^2.

Interprétation du test

	$\hat{\beta}_j$		$\hat{\sigma}_j$	$\frac{\hat{\beta}_j}{\hat{\sigma}_j}$	$\Pr(W > z^2)$
	coef	exp(coef)	se(coef)	z	$\Pr(>\|z\|)$
grppatchOnly	0.558663	1.748334	0.216674	2.578	0.00993
age	-0.023018	0.977245	0.009605	-2.397	0.01655

Dans ce modèle :

• Recevoir le traitement par Patch versus le traitement combiné augmente le risque instantané de reprise du tabac de 74,8 % .

• Une année d'âge supplémentaire diminue le risque instantané de reprise du tabac de 2,3% .

4. Tests sur les paramètres –Tests globaux

Test de nullité simultanée des coefficients : $\mathcal{H}_0 : \beta_1 = \cdots = \beta_p = 0$ versus $\mathcal{H}_1 : \beta_{j_0} \neq 0$ pour au moins un j_0.

- Rapport de vraisemblance ;
- Test de Wald ;
- Score (ou logrank) ;

Loi asymptotique des trois tests : χ^2_p, avec p le nombre de paramètres du modèle.

```
# Likelihood ratio test= 13.82  on 2 df,   p=0.001
# Wald test            = 13.48  on 2 df,   p=0.001
# Score (logrank) test = 13.74  on 2 df,   p=0.001
```

Remarque : variables explicatives binaires

Les variables explicatives ou covariables ou facteurs de risque peuvent être quantitatives ou qualitatives. Une attention particulière doit être portée aux variables qualitatives.
Soit une variable binaire comme le sexe. Le codage choisi est arbitraire mais l'interprétation en dépend : par exemple, si

$$Z_1 = \begin{cases} 1 & \text{si l'individu est un homme} \\ 0 & \text{si l'individu est une femme} \end{cases}$$

alors pour un homme $h(t|Z_1 = 1) = h_0(t)\exp(\beta)$ et pour une femme $h(t|Z_1 = 0) = h_0(t)$. Ici, le risque relatif des hommes par rapport aux femmes est égal à e^{β}. Mais si l'on choisit de coder par $Z_1 = 1$ pour une femme et $Z_1 = 0$ pour un homme, alors c'est le risque relatif des femmes par rapport aux hommes qui vaut e^{β}.

Remarque : variables explicatives binaires

Considérons une variable catégorielle possédant 3 modalités . On utilise des variables indicatrices comme dans le cas d'une variable binaire :

$$\begin{aligned} Z_1 = 1 \quad & \text{si l'individu appartient à la catégorie } 1, 0 \text{ sinon} \\ Z_2 = 1 \quad & \text{si l'individu appartient à la catégorie } 2, 0 \text{ sinon} \end{aligned}$$

On ne rajoute pas de variable Z_3 qui coderait la catégorie 3 car les 3 variables Z_1, Z_2, Z_3 seraient alors dépendantes linéairement puisque $Z_1 + Z_2 + Z_3 = 1$. Ainsi, on a :

$$\begin{aligned} h(t|Z_1 = 1, Z_2 = 0) &= h_0(t)\exp(\beta_1) \quad \text{pour un individu de la cat. 1} \\ h(t|Z_1 = 0, Z_2 = 1) &= h_0(t)\exp(\beta_2) \quad \text{pour un individu de la cat. 2} \\ h(t|Z_1 = 0, Z_2 = 0) &= h_0(t) \quad \text{pour un individu de la cat. 3} \end{aligned}$$

Le risque relatif $RR(1/3)$ d'un individu dans la cat. 1 vs cat. 3 est e^{β_1}. $RR(2/3) = e^{\beta_2}$ et $RR(1/2) = e^{\beta_1 - \beta_2}$. Avec ce codage, la cat. 3 est la "catégorie de référence" .

Remarque : variables explicatives binaires

Il serait inadéquat de coder une variable catégorielle à 3 modalités par une seule variable avec $Z = i$, pour $i = 1, 2, 3$. En effet, on aurait alors le risque relatif d'un individu dans la cat. i qui serait égal :

$$h(t|Z = i) = h_0(t) \exp(\beta \times i)$$

et donc

$$RR(2/1) = RR(3/2) = e^{\beta} \quad \text{et} \quad RR(3/1) = e^{2\beta}$$

Ces relations n'ont aucune raison d'être vraies.

4. Tests sur les paramètres – une seule variable explicative binaire

On considère une seule covariable Z qui est l'appartenance à un groupe :

$$\begin{cases} Z = 0 \text{ pour les patients du groupe A} \\ Z = 1 \text{ pour les patients du groupe B} \end{cases}$$

Le modèle de Cox devient $h(t) = h_0(t)\exp(\beta Z)$, $\beta \in \mathbb{R}$ et la log-vraisemblance partielle s'écrit :

$$\log L(\beta) = \beta m_B - \sum_{i=1}^{k} m_i \ln[n_{A,i} + n_{B,i}\ \exp(\beta)]$$

où m_B est le nombre total de décès observés dans le groupe B, m_i le nombre de décès observés en t_i, $n_{A,i}$ et $n_{B,i}$ sont les nombres de patients exposés au risque dans les groupes A et B en t_i.

à faire en exercice dans le cas où il n'y a pas d'ex æquo.

Pour tester $H_0 : \beta = 0$, on évalue la fonction de score (à partir de la vraisemblance de Breslow (sans ex æquo)

$$U(\beta) = \partial \log L(\beta)/\partial\beta = \sum_{i=1}^{k} m_{B,i} - \sum_{i=1}^{k} m_i \frac{n_{B,i} e^{\beta}}{n_{A,i} + n_{B,i} e^{\beta}}$$

et on calcule l'information de Fisher $\mathcal{I}(\beta) = -\mathbb{E}(\partial^2 \log L(\hat{\beta})/\partial^2\beta)$:

$$\partial^2 \log L(\beta)/\partial^2\beta = -\sum_{i=1}^{k} m_i \frac{n_{B,i} e^{\beta}(n_{A,i} + n_{B,i} \exp(\beta)) - (n_{B,i} e^{\beta})^2}{(n_{A,i} + n_{B,i} \exp(\beta))^2}$$

On peut alors déterminer la statistique du test du score Sous $H_0 : \beta = 0$, la statistique du test du score

$$\frac{U(0)^2}{\hat{\mathcal{I}}(0)} = \frac{\left(\sum_{i=1}^{k}\left(m_{B,i} - n_{B,i}\frac{m_i}{n_i}\right)\right)^2}{\sum_{i=1}^{k}\frac{m_i n_{A,i} n_{B,i}}{n_i^2}} \sim \chi^2(1)$$

pour un nombre d'individus assez grand. On remarque qu'elle coïncide avec la statistique de test du log-rank vue dans la Partie I.

Hypothèse des risques proportionnels :comparaison graphique

Nous considérons la transformation suivante des courbes de survie : $\ln(-\ln(S(t|\boldsymbol{Z}))$
Cette transformation a la propriété suivante : Si l'hypothèse de risques proportionnels est valide, alors :

$$S(t|\boldsymbol{Z}) = S_0(t)\exp(\beta^t \boldsymbol{Z}) \Longleftrightarrow ln(-\ln S(t|\boldsymbol{Z})) = \ln(-\ln S_0(t)) + \beta^t \boldsymbol{Z}$$

Ainsi pour deux individus de caractéristiques différentes $\boldsymbol{Z}_1$ et $\boldsymbol{Z}_2$ la différence entre les courbes "LML" $(\ln(-\ln))$ vaut

$$\beta^t(\boldsymbol{Z}_2 - \boldsymbol{Z}_1)$$

Cette quantité est indépendante du temps. Les courbes de survie après transformation "LML" sont donc parallèles pour différentes valeurs de t.

Il suffit alors de tracer les courbes "LML" correspondant aux différents niveaux d'une covariable, les autres covariables restant telles quelles, et de les comparer. S'il est possible de superposer les différentes courbes par simple translation, alors l'hypothèse de proportionnalité est vérifiée.

Hypothèse des risques proportionnels : les résidus de Schoënfeld [11]

Les résidus R_{ki} représentent la différence entre la valeur observée $Z_{k(i)}$ de la k-ième covariable de l'individu qui décède en t_i et la valeur attendue de cette covariable pour le patient décédé sous le modèle avec hypothèse des risques proportionnels.

Cette valeur attendue est une moyenne pondérée de la covariable Z_k par le risque de décès des patients à risque en t_i. Si l'hypothèse des risques proportionnels est satisfaite, les résidus ne doivent pas dépendre du temps.

On cherche alors à tester la non-corrélation des résidus de Schoënfeld avec le temps pour une covariable Z_k donnée.

[11] Schoenfeld D. (1982). Partial residuals for the proportional hazards regression model. Biometrika.Schoenfeld D. Partial residuals for the proportional hazards regression model. Biometrika. 69 (1): 239-241.

5. Modèle de Cox : vérification de l'hypothèse des RP

Plus formellement, pour chaque t_i, on calcule la différence entre les caractéristiques de l'individu décédé (en cas d'ex-aequo, on calcule un résidu pour chaque individu et on somme les résidus) et une moyenne pondérée des caractéristiques des individus à risque de décès au temps t_i ous l'hypothèse des risques proportionnels :

$$R_{ki} = Z_{k(i)} - \bar{Z}_k(t_i)$$

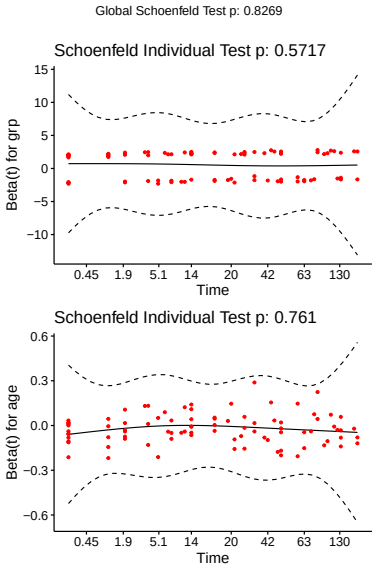
Avec :
R_{ki} : résidu au temps t_i
$Z_{k(i)}$: valeur de la covariable k pour l'individu (i) décédé au temps t_i.
$\bar{Z}_k(t_i)$: moyenne pondérée de la covariable k pour tous les individus à risque au temps t_i
soit $\bar{Z}_k(t_i) = \sum_{j \in \mathcal{R}(t_i)} Z_{k,j} \frac{e^{\beta^t \mathbf{z}_j}}{\sum_{j \in \mathcal{R}(t_i)} e^{\beta^t \mathbf{z}_j}}$.

On utilise les résidus standardisés c'est-à-dire les résidus divisés par leur variance.

5. Modèle de Cox : vérification de l'hypothèse des RP

- Les résidus de Schoenfeld peuvent être analysés au moyen de graphiques afin de repérer d'éventuelles violations de l'hypothèse de proportionnalité.
- L'approche consiste à représenter graphiquement les résidus en fonction du temps (ou d'une transformation du temps.)
- Il est également possible d'ajouter sur le même graphique une courbe de régression illustrant l'évolution moyenne des résidus en fonction du temps. Toute déviation par rapport à une ligne horizontale sur ce graphique signale une divergence par rapport à l'hypothèse de proportionnalité

5. Modèle de Cox : vérification de l'hypothèse des RP



Examen graphique des résidus de Schoënfeld :

On représente les résidus de Schoënfeld en fonction du logarithme du temps ici, pour les deux covariables `grp` et `age` et on ajoute une courbe de leur tendance (cf. TP).

L'hypothèse des risque proportionnels ne semble pas être remise en cause

5. Modèle de Cox : vérification de l'hypothèse des RP

Une méthode pour tester l'hypothèse des risques proportionnels consiste à calculer la corrélation entre les résidus et les durées t_i où l'on observe un événement.

- Si l'hypothèse nulle d'absence de corrélation n'est pas rejetée, alors l'hypothèse de proportionnalité n'est pas remise en cause. Sinon, elle est rejetée. (cf. `cox.zph`) La statistique de test est un peu complexe à décrire, elle est construite à partir des résidus de Schoënfeld.
- Il est possible aussi de construire un test "à la main" : à partir d'une régression linéaire expliquant les résidus par le temps ou une transformation (logarithme) du temps, on teste simplement si la pente de la droite de régression est bien nulle.

Hypothèse des risques proportionnels (coefficients invariants au temps)

Pour chaque variable `grp` et `age` on teste si les résidus de Schoënfeld sont corrélés au temps.

```
> cox.zph(coxph(Surv(ttr,relapse)~grp+age,data=pharmacoSmoking))
        chisq df    p
grp    0.3198  1 0.57
age    0.0925  1 0.76
GLOBAL 0.3803  2 0.83
```

Ici pour les deux covariables testées, il n'y a pas lieu de rejeter l'hypothèse des risque proportionnels.

5. Modèle de Cox : vérification de l'hypothèse des RP

Dans le cas où la proportionnalité des risques doit manifestement être rejetée par rapport à une ou plusieurs covariables, deux options sont possibles :

- Des effets d'interaction entre ces covariables et le temps peuvent être introduits explicitement dans le modèle de Cox. Par exemple : `age* log(t)`
- On utilise le modèle de Cox stratifié :
  La stratification consiste à calculer un modèle de Cox en attribuant une valeur différente du risque de base $h_0(t)$ à chaque catégorie de la variable de stratification. En revanche, l'influence des variables explicatives, et donc les valeurs estimées des paramètres β, est commune à toutes les catégories.

Transformation d'une variable continue en variable catégorielle et modèle stratifié selon ses modalités

Si une covariable continue semble ne pas vérifier l''' hypothèse des risques porportionnels, on peut tenter de transformer la variable continue en une variable catégorielle :

Soit une variable de stratification avec catégories indicées $s = 1, 2, ...$
Exemple : variable `age` : $[21 - 34]$ $[35 - 49]$ $[50 - 64]$ $[\geq 65 -]$

- Le modèle de Cox est estimé avec un risque de base $h_{s,0}(t)$ différent pour chaque strate s.
  Les risques sont toujours supposés proportionnels pour individus d'une même strate, mais pas entre les strates.
- Les effets des covariables sont identiques dans toutes les strates : il n'y a pas d'effet d'interaction entre la variable de stratification et les variables explicatives du modèle.

La vraisemblance partielle où l'on a stratifié une covariable en $s = 1, 2, \cdots$ catégories s'écrit :

$$L_{strat}(\beta) = \prod_{s} \prod_{i \in I_s} \frac{\exp\left(\beta^t \boldsymbol{Z}_{(i)}\right)}{\sum_{j \in \mathcal{R}_{i,s}} \exp\left(\beta^t \boldsymbol{Z}_j\right)}$$

avec I_s : les indices des individus de la strate s non censurés et $\mathcal{R}_{i,s}$: les indices des individus de la strate s à risque en t_i.

Les estimations sont différentes de celles obtenues sans stratification.

Après avoir estimé le vecteur des paramètres β par $\hat{\beta}$, on peut calculer l'estimateur de Breslow de la fonction de risque cumulé pour une nouvelle valeur $\boldsymbol{Z}^\star$ du vecteur des covariables :

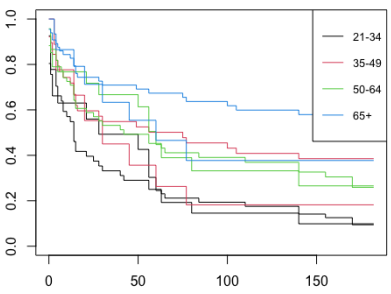
$$\hat{H}(t|Z^\star) = \hat{H}_0(t)\exp(\hat{\beta}^t \boldsymbol{Z}^\star),\ t \geq 0$$

avec $\hat{H}_0(t) = \sum_{i:t\geq t_i} \frac{1}{\sum_{j\in\mathcal{R}_i}\exp(\beta^t \boldsymbol{Z}_j)}$ (estimateur de Breslow de la fonction de risque cumulé de base dans le cas sans ex æquo, cf. [12]) et on peut en déduire :

$$\hat{S}(t|\boldsymbol{Z}^\star) = \exp\left(-\hat{H}(t|\boldsymbol{Z}^\star)\right) = \hat{S}_0(t)^{\exp(\beta^t \boldsymbol{Z}^\star)}\ t \geq 0$$

[12] Klein & Moeshberger [2005], cf. paragraphe 8.3, p. 257 : "Theoretical Notes".

5. Modèle de Cox stratifié



Fonctions de survie estimées pour chaque groupe de traitement et dans chaque strate.

Même β estimé pour chaque strate de l'âge, mais risques $h_{s,0}(t)$ différents d'où fonctions de survie estimées différentes.

Conclusion

- Le modèle de Cox est un modèle facile à interpréter.
- plusieurs modélisations sont possibles selon la question adressée : pas de méthode complètement "automatique" (bien connaître les données).
- Si l'hypothèse des risques proportionnels n'est pas vérifiée en pratique, il se peut que l'effet d'une covariable passe inaperçu ou soit très atténué.
- package timereg pour tests récents de la dépendance des covariables au temps.
- autres extensions du modèle de Cox :
  - pour des données tronquées à gauche,
  - pour des évènements récurrents.
  - modèles à risques compétitifs/modèles multi-états,
  - dans les modèles à fragilité (modéliser l'hétérogénéité des individus)
- Il existe d'autres modèles qui font d'autres hypothèses sur les données : le modèle AFT ("Accelerated Failure Time model") par exemple.

Références

- Cox, David R. "Regression models and life tables (with discussion)". In: Journal of the Royal Statistical Society 34 (1972), pp. 187–220.
- Hall, W. J. and Wellner,(1980), J. A. Confidence Bands for a Survival Curve from Censored Data. Biometrika 67 : 133–143.
- Huber-Carol C., (1994), Durées de survie tronquées et censurées, Journal de la société statistique de Paris, tome 135, 4 (1994), p. 3-23
- Jonas, S. (2018) Méthodes de comparaisons de deux ou plusieurs groupes de données censurées par intervalle avec application en immunologie clinique. Thèse de doctorat de l'Université Paris Saclay.
- J.P. Klein and M. L. Moeschberger., (2005) Survival analysis: techniques for censored and truncated data. Springer Science & Business Media.
- Lin, D. Y., Wei, L. J., Ying, Z. (1993), Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals, Biometrika, Vol. 80, No. 3, pp. 557-572.
- Moore, Dirk, F., (2016) Applied Survival Analysis Using R, Springer.
- Nair, V. N., (1984) Confidence Bands for Survival Functions with Censored Data: A Comparative Study. Technometrics 14 : 265–275.
- Steinberg et al. (2009), Triple-combination pharmacotherapy for medically ill smokers: A randomized trial. Annals of Internal Medicine 150, 447-454.

Références

- Therneau, T. M., Grambsch, P. M., Fleming T. R., (1990), Martingale-based residuals for survival models, Biometrika, Volume 77, Issue 1, Pages 147–160.
- Therneau,T., Crowson, C., and Atkinson, E. "Using time dependent covariates and time dependent coefficients in the cox model". In: Survival Vignettes (2017).