Introduction à l'analyse des durées de survie

Philippe SAINT PIERRE [1]

Avril 2021

1. Institut de Mathématiques de Toulouse, Université Paul Sabatier - Toulouse III

Table des matières

Chapitre I

Introduction

1 Introduction

Le terme de durée de survie désigne le temps écoulé jusqu'à la survenue d'un événement précis. L'événement étudié (communément appelé "décès") est le passage irréversible entre deux états (communément nommé "vivant" et "décès"). L'événement terminal n'est pas forcément la mort : il peut s'agir de l'apparition d'une maladie (par exemple, le temps avant une rechute ou un rejet de greffe), d'une guérison (temps entre le diagnostic et la guérison), la panne d'une machine (durée de fonctionnement d'une machine, en fiabilité) ou la survenue d'un sinistre (temps entre deux sinistres, en actuariat).

L'analyse des données (durées) de survie est l'étude du délai de la survenue de cet événement. Dans le domaine biomédical, on étudie ces durées dans le contexte des études longitudinales comme les enquêtes de cohorte (suivi de patients dans le temps) ou les essais thérapeutiques (tester l'efficacité d'un médicament). On cherche alors à estimer la distribution des temps de survie (fonction de survie), à comparer les fonctions de survie de plusieurs groupes ou à analyser la manière dont des variables explicatives modifient les fonctions de survie.

2 Définitions

Quelques définitions sont couramment utilisées dans les études de survie.

— Date d'origine : elle correspond à l'origine de la durée étudiée. Elle peut être la date de naissance, le début d'une exposition à un facteur de risque, la date d'une opération chirurgicale, la date de début d'une maladie ou la date d'entrée dans l'étude. Chaque individu peut donc avoir une date d'origine différente (pas important car c'est la durée qui nous intéresse).

— Date de point : c'est la date au-delà de laquelle on arrêtera l'étude et on ne tiendra plus compte des informations sur les sujets.

— Date des dernières nouvelles : c'est la date la plus récente où des informations sur un sujet ont été recueillies.

3 Distributions de la durée de survie

Supposons que la durée de survie X soit une variable positive ou nulle, et absolument continue, alors sa loi de probabilité peut être définie par l'une des cinq fonctions équivalentes suivantes (chacune des fonctions ci-dessous peut être obtenue à partir de l'une des autres fonctions) :

3.1 Fonction de survie S

La fonction de survie est, pour t fixé, la probabilité de survivre jusqu'à l'instant t, c'est-à-dire

$$S(t) = \mathbb{P}(X > t), \qquad t \geqslant 0.$$

3.2 Fonction de répartition F

La fonction de répartition (ou c.d.f. pour "cumulative distribution function") représente, pour t fixé, la probabilité de mourir avant l'instant t, c'est-à-dire

$$F(t) = \mathbb{P}(X \leqslant t) = 1 - S(t).$$

Remarque 1 *Il est arbitraire de décider que $S(t) = \mathbb{P}(X \geqslant t)$ ou $S(t) = \mathbb{P}(X > t)$. Cela n'a aucune importance quand la loi de X est continue car $\mathbb{P}(X > t) = \mathbb{P}(X \geqslant t)$. Dans les cas où F a des sauts (quand le temps est discret, par exemple, compté en mois ou semaine), on utilise les notations suivantes :*

$$F^-(t) = P(X < t) \quad et \quad F^+(t) = P(X \leqslant t)$$

où F^- est la limite à gauche et F^+ la limite à droite de F (définitions et notations sont identiques pour la fonction S). Remarquons que $F^- \leqslant F^+$ et $S^- \geqslant S^+$.

3.3 Densité de probabilité f

C'est la fonction $f(t) \geqslant 0$ telle que pour tout $t \geqslant 0$

$$F(t) = \int_0^t f(u)du.$$

Si la fonction de répartition F admet une dérivée au point t alors

$$f(t) = \lim_{h \to 0} \frac{\mathbb{P}(t \leqslant X < t + h)}{h} = F'(t) = -S'(t).$$

Pour t fixé, la densité de probabilité représente la probabilité de mourir dans un petit intervalle de temps après l'instant t.

3.4 Risque instantané λ (ou taux de hasard)

Le risque instantané (ou taux d'incidence), pour t fixé caractérise la probabilité de mourir dans un petit intervalle de temps après t, conditionnellement au fait d'avoir survécu jusqu'au temps t (c'est-à-dire le risque de mort instantané pour ceux qui ont survécu) :

$$\lambda(t) = \lim_{h \to 0} \frac{\mathbb{P}(t \leqslant X < t + h \mid X \geqslant t)}{h} = \frac{f(t)}{S(t)} = -\ln(S(t))'.$$

3.5 Taux de hasard cumulé Λ

Le taux de hasard cumulé est l'intégrale du risque instantané λ :

$$\Lambda(t) = \int_0^t \lambda(u)du = -\ln(S(t)).$$

On peut déduire de cette équation une expression de la fonction de survie en fonction du taux de hasard cumulé (ou du risque instantané) :

$$S(t) = \exp(-\Lambda(t)) = \exp\left(-\int_0^t \lambda(u)du\right).$$

On en déduit que

$$f(t) = \lambda(t) \exp\left(-\int_0^t \lambda(u)du\right).$$

3.6 Quantités associées à la distribution de survie

3.6.1 Moyenne et variance de la durée de survie

Le temps moyen de survie $\mathbb{E}(X)$ et la variance de la durée de survie $\mathbb{V}(X)$ sont définis par les quantités suivantes (en utilisant des IPP) :

$$\mathbb{E}(X) = \int_0^\infty S(t)dt,$$
$$\mathbb{V}(X) = 2\int_0^\infty tS(t)dt - (\mathbb{E}(X))^2.$$

Ainsi on peut déduire l'espérance et la variance à partir de n'importe laquelle des fonctions F, S, f, λ, Λ (mais pas l'inverse).

3.6.2 Quantiles de la durée de survie

— La médiane de la durée de survie est le temps t pour lequel la probabilité de survie $S(t)$ est égale à 0.5, c'est-à-dire, la valeur t_m qui satisfait $S(t_m) = 0.5$.
Dans le cas où l'estimateur est une fonction en escalier (ex : Kaplan-Meier), il se peut qu'il y ait un intervalle de temps vérifiant $S(t_m) = 0.5$. Il faut alors être prudent dans l'interprétation, notamment si les deux événements encadrant le temps médian sont éloignés.
Il est possible d'obtenir un intervalle de confiance du temps médian. Soit $[B_i, B_s]$ un intervalle de confiance de niveau α de $S(t_m)$, alors un intervalle de confiance de niveau α du temps médian t_m est

$$[S^{-1}(B_s), S^{-1}(B_i)].$$

— La fonction quantile de la durée de survie est définie par

$$\begin{aligned} q(p) &= \inf(t : F(t) \geqslant p), \qquad 0 < p < 1, \\ &= \inf(t : S(t) \leqslant 1 - p). \end{aligned}$$

Lorsque la fonction de répartition F est strictement croissante et continue alors

$$\begin{aligned} q(p) &= F^{-1}(p), \qquad 0 < p < 1, \\ &= S^{-1}(1 - p). \end{aligned}$$

Le quantile $q(p)$ est le temps où une proportion p de la population a disparu.

4 Censure et troncature

Une des caractéristiques des données de survie est l'existence d'observations incomplètes. En effet, les données sont souvent recueillies partiellement, notamment, à cause des processus de censure et de troncature. Les données censurées ou tronquées proviennent du fait qu'on n'a pas accès à toute l'information : au lieu d'observer des réalisations indépendantes et identiquement distribuées (i.i.d.) de durées X, on observe la réalisation de la variable X soumise à diverses perturbations, indépendantes ou non du phénomène étudié.

4.1 Censure

La censure est le phénomène le plus couramment rencontré lors du recueil de données de survie.

Pour l'individu i, considérons
— son temps de survie X_i,
— son temps de censure C_i,
— la durée réellement observée T_i.

4.1.1 Censure à droite

La durée de vie est dite censurée à droite si l'individu n'a pas subi l'événement à sa dernière observation. En présence de censure à droite, les durées de vie ne sont pas toutes observées ; pour certaines d'entre elles, on sait seulement qu'elles sont supérieures à une certaine valeur connue.

1. La censure de type I

Soit C une valeur fixée, au lieu d'observer les variables $X_1, ..., X_n$ qui nous intéressent, on n'observe X_i uniquement lorsque $X_i \leqslant C$, sinon on sait uniquement que $X_i > C$. On utilise la notation suivante : $T_i = X_i \wedge C = \min(X_i, C)$.

Ce mécanisme de censure est fréquemment rencontré dans les applications industrielles. Par exemple, on peut tester la durée de vie de n objet identiques (ampoules) sur un intervalle d'observation fixé $[0, u]$. En biologie, on peut tester l'efficacité d'une molécule sur un lot de souris (les souris vivantes au bout d'un temps u sont sacrifiées).

2. La censure de type II

Elle est présente quand on décide d'observer les durées de survie des n patients jusqu'à ce que k d'entre eux soient décédés et d'arrêter l'étude à ce moment là. Soient $X_{(i)}$ et $T_{(i)}$ les statistiques d'ordre des variables X_i et T_i. La date de censure est donc $X_{(k)}$ et on observe les variables suivantes

$$\begin{aligned} T_{(1)} &= X_{(1)} \\ &\vdots \\ T_{(k)} &= X_{(k)} \\ T_{(k+1)} &= X_{(k)} \\ &\vdots \\ T_{(n)} &= X_{(k)} \end{aligned}$$

3. La censure de type III (ou censure aléatoire de type I)

Soient $C_1, ..., C_n$ des variables aléatoires i.i.d. On observe les variables

$$T_i = X_i \wedge C_i.$$

L'information disponible peut être résumée par :

— la durée réellement observée T_i,

— un indicateur $\delta_i = \mathbb{1}_{\{X_i \leqslant C_i\}}$

  — $\delta_i = 1$ si l'événement est observé (d'où $T_i = X_i$). On observe les "vraies" durées ou les durées complètes.

  — $\delta_i = 0$ si l'individu est censuré (d'où $T_i = C_i$). On observe des durées incomplètes (censurées).

La censure aléatoire est la plus courante. Par exemple, lors d'un essai thérapeutique, elle peut être engendrée par

(a) la perte de vue : le patient quitte l'étude en cours et on ne le revoit plus (à cause d'un déménagement, le patient décide de se faire soigner ailleurs). Ce sont des patients "perdus de vue".

(b) l'arrêt ou le changement du traitement : les effets secondaires ou l'inefficacité du traitement peuvent entraîner un changement ou un arrêt du traitement. Ces patients sont exclus de l'étude.

(c) la fin de l'étude : l'étude se termine alors que certains patients sont toujours vivants (ils n'ont pas subi l'événement). Ce sont des patients "exclus-vivants". Les "perdus de vue" (et les exclusions) et les "exclus-vivants" correspondent à des observations censurées mais les deux mécanismes sont de nature différente (la censure peut être informative chez les "perdus de vue").

Remarque 2 *Dans la suite de ce cours, nous ferons l'hypothèse que la censure est indépendante de l'événement, c'est-à-dire, que X_i est indépendant de C_i. Cette hypothèse est très utile d'un point de vue mathématique et indispensable aux modèles classiques d'analyse de survie. Il est donc important de voir si elle se justifie. Par exemple, quand la censure est due à un arrêt du traitement ou quand les patients les plus malades ne sont plus suivis, l'hypothèse d'indépendance n'est pas vérifiée car la censure apporte une information sur l'état de santé du patient. Si cette hypothèse n'est pas vérifiée cela entraîne un biais dans l'estimation (par exemple, quand on enlève de l'étude les patients les plus à risque, on surestime la survie). Dans le cas d'une censure causée par un déménagement ou par la fin de l'étude, cette hypothèse est naturelle.*

4.1.2 Censure à gauche

La censure à gauche correspond au cas où l'individu a déjà subi l'événement avant que l'individu soit observé. On sait uniquement que la date de l'événement est inférieure à une certaine date connue. Pour chaque individu, on peut associer un couple de variables aléatoires (T, δ) :

$$T = X \vee C = \max(X, C),$$
$$\delta = \mathbb{1}_{\{X \geqslant C\}}.$$

Comme pour la censure à droite, on suppose que la censure C est indépendante X. Un des premiers exemples de censure à gauche rencontré dans la littérature considère le cas d'observateurs qui s'intéressent à l'horaire où les babouins descendent de leurs arbres pour aller manger (les babouins passent la nuit dans les arbres). Le temps d'événement (descente de l'arbre) est observé si le babouin descend de l'arbre après l'arrivée des observateurs. Par contre, la donnée est censurée si le babouin est descendu avant l'arrivée des observateurs : dans ce cas on sait uniquement que l'horaire de descente est inférieur à l'heure d'arrivée des observateurs. On observe donc le maximum entre l'heure de descente des babouins et l'heure d'arrivée des observateurs (l'heure correspond à une durée).

Remarque 3 *Les modèles présentés dans ce cours traitent le cas de la censure à droite. Très peu de travaux s'intéressent à la censure à gauche car beaucoup moins fréquente. Certains auteurs ont proposé de "renverser" l'échelle de temps, c'est-à-dire de considérer la variable $\tau - T = \tau - (X \vee C) = (\tau - X) \wedge (\tau - C)$ au lieu de la variable T, où τ est une réel positif choisi de sorte que les observations $\tau - T$ restent dans $\mathbb{R}_+$. Cette approche n'est pas entièrement satisfaisante puisqu'elle exclut le cas où les variables X (et T) sont à support*

sur $\mathbb{R}_+$ et le modèle obtenu n'est plus à intensité multiplicative au sens de Aalen (ce qui pose des problèmes d'un point de vue théorique). De plus, cette approche n'est plus possible dans le cas où les données ne sont pas seuleument censurées à gauche.

4.1.3 Censure par intervalle

Une date est censurée par intervalle si au lieu d'observer avec certitude le temps de l'événement, la seule information disponible est qu'il a eu lieu entre deux dates connues. Par exemple, dans le cas d'un suivi de cohorte, les personnes sont souvent suivies par intermittence (pas en continu), on sait alors uniquement que l'événement s'est produit entre ces deux temps d'observations. On peut noter que pour simplifier l'analyse, on fait souvent l'hypothèse que le temps d'événement correspond au temps de la visite pour se ramener à de la censure à droite.

4.2 Troncature

Les troncatures diffèrent des censures au sens où elles concernent l'échantillonnage lui-même. Ainsi, une variable X est tronquée par un sous ensemble éventuellement aléatoire A de $\mathbb{R}_+$ si au lieu de X, on observe X uniquement si $X \in A$. Les points de l'échantillon "tronqué" appartiennent tous à A, et suivent donc la loi de T conditionnée par l'appartenance à A. Il ne faut pas confondre censure et troncature. S'il y a troncature, une partie des individus (donc des X_i) ne sont pas observables et on n'étudie qu'un sous-échantillon (problème d'échantillonage). Le biais de sélection est un cas particulier de troncature.

1. La troncature à gauche

   Soit Z une variable aléatoire indépendante de X, on dit qu'il y a troncature à gauche lorsque X n'est observable que si $X > Z$. On observe le couple (X, Z), avec $X > Z$. Par exemple, si la durée de vie d'une population est étudiée à partir d'une cohorte tirée au sort dans cette population, seule la survie des sujets vivants à l'inclusion pourra être étudiée (il y a troncature à gauche car seuls les sujets ayant survécu jusqu'à la date d'inclusion dans la cohorte sont observables).

2. La troncature à droite

   De même, il y a troncature à droite lorsque X n'est observable que si $X < Z$.

3. La troncature par intervalle

   Quand une durée est tronquée à droite et à gauche, on dit qu'elle est tronquée par intervalle. Par exemple, on rencontre ce type de troncature lors de l'étude des patients d'un registre : les patients diagnostiqués avant la mise en place du registre ou répertoriés après la consultation du registre ne seront pas inclus dans l'étude.

5 Fonction de vraisemblance

Considérons le cas d'une censure aléatoire droite C indépendante de la durée d'intérêt X. Supposons que les variables X et C ont pour densités respectives f et g et pour survies S et G. La distribution de X est définie par un paramètre de dimension finie. Toute l'information est contenue dans le couple (T_i, δ_i), où $T_i = \min(X_i, C_i)$ est la durée observée, et l'indicateur de censure $\delta_i = \mathbb{1}_{\{X_i \leqslant C_i\}}$. Ainsi, la contribution à la vraisemblance pour l'individu i est

$$\begin{aligned} L_i &= P(T_i \in [t_i, t_i + dt], \delta_i = 1 \mid \theta)^{\delta_i} \times P(T_i \in [t_i, t_i + dt], \delta_i = 0 \mid \theta)^{1-\delta_i} \\ &= P(X_i \in [t_i, t_i + dt], C_i \geqslant X_i \mid \theta)^{\delta_i} \times P(C_i \in [t_i, t_i + dt], C_i < X_i \mid \theta)^{1-\delta_i} \\ &= [f(t_i \mid \theta)G(t_i^-)]^{\delta_i} \times [g(t_i)S(t_i \mid \theta)]^{1-\delta_i}. \end{aligned}$$

Par l'hypothèse (de censure non informative), le paramètre d'intérêt θ n'apparaît pas dans la loi de la censure (Il existe des mécanismes indépendants et informatifs). La partie utile de la vraisemblance se réduit alors à

$$L = \prod_{i=1}^{n} f(t_i \mid \theta)^{\delta_i} S(t_i \mid \theta)^{1-\delta_i}.$$

Remarque 4 *Notons que la présence de données censurées doit être prise en compte dans l'écriture de la vraisemblance. En effet, en raisonnant sur le sous échantillon des données non censurées, la vraisemblance est*

$$\bar{L} = \prod_{i=1}^{n} f(t_i \mid \theta)^{\delta_i}.$$

L'estimateur obtenu en maximisant $\bar{L}$ est asymptotiquement biaisé.

Remarque 5 *Si les observations ne sont pas identiquement distribuées (ex : incorporation de covariables), on peut généraliser l'expression précédente en introduisant un indice pour f et S.*

Remarque 6 *Dans le cas d'une censure non aléatoire, on obtient également la vraisemblance L.*

Remarque 7 *Dans le cas d'une censure à droite de type II, la vraisemblance est la suivante :*

$$\tilde{L} = \frac{n!}{k!(n-k)!} \prod_{i=1}^{k} f(t_i \mid \theta) \times S(t_k \mid \theta)^{n-k}.$$

Remarque 8 *Considérons le cas de données tronquées à gauche de manière aléatoire. Les variables X_i sont soumises à troncature par les variables aléatoires Z_i supposées indépendantes des X_i. On dispose d'un échantillon $(X_i, Z_i)_{i=1,\ldots,N}$ où N est aléatoire puisqu'on sélectionne (on observe) les individus pour lesquels $X_i \geqslant Z_i$ dans une population de taille inconnue n. La vraisemblance conditionnelle par rapport à N et aux valeurs de la variable de troncature observées est*

$$L_T = \prod_{i=1}^{N} P(X_i \in [x_i, x_i + dx] \mid X \geqslant z_i) = \prod_{i=1}^{N} \frac{P(X_i \in [x_i, x_i + dx], X \geqslant z_i)}{P(X \geqslant z_i)} = \prod_{i=1}^{N} \frac{f(x_i)}{S(z_i)}.$$

Par construction, on a $x_i \geqslant z_i$ pour tout i. On peut également écrire la vraisemblance conditionnellement à N,

$$\begin{aligned} L_{T_{bis}} &= \prod_{i=1}^{N} \mathcal{L}(X_i, Z_i \mid X \geqslant Z) = \frac{1}{P(X \geqslant Z)^N} \prod_{i=1}^{N} \mathcal{L}(X_i, Z_i) \\ &= \left(\int \mathbb{1}_{\{u \geqslant z\}} \xi(z) f(u) dz du \right)^{-N} \prod_{i=1}^{N} f(x_i) \xi(z_i) \\ &= \left(\int \Phi(u) f(u) du \right)^{-N} \prod_{i=1}^{N} f(x_i) \xi(z_i) \end{aligned}$$

où Φ et ξ sont les fonctions de répartition et de densité de Z. Cette vraisemblance fait intervenir la loi de la troncature (elle utilise l'information complète). Elle ne peut pas être utilisée si on n'a pas d'a priori sur la loi de la troncature. Les deux vraisemblances donnent des estimateurs convergents.

Chapitre II

Estimation non paramétrique

Dans ce chapitre nous nous placerons dans le cadre le plus fréquent d'une censure à droite aléatoire de type I. Si aucun modèle n'est supposé, les principaux estimateurs sont :
— l'estimateur de Kaplan-Meier de la fonction de survie,
— l'estimateur de Nelson-Aalen du risque cumulé.

1 Estimateur de Kaplan-Meier de la survie

1.1 Estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier découle de l'idée suivante : survivre après un temps t c'est être en vie juste avant t et ne pas mourir au temps t, c'est-à-dire, si $t'' < t' < t$

$$\begin{aligned} P(X > t) &= P(X > t', X > t) \\ &= P(X > t \mid X > t') \times P(X > t') \\ &= P(X > t \mid X > t') \times P(X > t' \mid X > t'') \times P(X > t'') \end{aligned}$$

En considérant les temps d'événements (décès et censure) distincts $T_{(i)}$ $(i = 1, ..., n)$ rangés par ordre croissant, on obtient

$$P(X > T_{(j)}) = \prod_{k=1}^{j} P(X > T_{(k)} \mid X > T_{(k-1)}),$$

avec $T_{(0)} = 0$. Considérons les notations suivantes :
— Y_i le nombre d'individus à risque de subir l'événement juste avant le temps $T_{(i)}$,
— d_i le nombre de décès en $T_{(i)}$.
Alors la probabilité p_i de mourir dans l'intervalle $]T_{(i-1)}, T_{(i)}]$ sachant que l'on était vivant en $T_{(i-1)}$, *i.e.* $p_i = P(X \leqslant T_{(i)} \mid X > T_{(i-1)})$, peut être estimée par

$$\hat{p}_i = \frac{d_i}{Y_i}.$$

Comme les temps d'événements sont supposés distincts, on a

$$d_i = 0 \text{ en cas de censure en } T_{(i)}, \textit{ i.e. } \text{quand } \delta_i = 0,$$
$$d_i = 1 \text{ en cas de décès en } T_{(i)}, \textit{ i.e. } \text{quand } \delta_i = 1.$$

On obtient alors l'estimateur de Kaplan-Meier :

$$\hat{S}(t) = \prod_{\substack{i=1,\dots,n \\ T_{(i)} \leqslant t}} \left(1 - \frac{\delta_i}{Y_i}\right) = \prod_{i:T_{(i)} \leqslant t} \left(1 - \frac{\delta_i}{n-(i-1)}\right) = \prod_{i:T_{(i)} \leqslant t} \left(\frac{n-i}{n-i+1}\right)^{\delta_i}.$$

L'estimateur $\hat{S}(t)$ est également appelé Produit Limite car il s'obtient comme la limite d'un produit. On montre que l'estimateur de Kaplan-Meier est un estimateur du maximum de vraisemblance. $\hat{S}(t)$ est une fonction en escalier décroissante, continue à droite. On peut également obtenir un estimateur de Kaplan-Meier dans le cas de données tronquées mais pas dans le cas de données censurées par intervalles (car les temps de décès ne sont pas connus).

Remarque 9 *Dans le cas où il y a des ex-aequo :*

- *si ce sont des événements de nature différente, on considère que les observations non censurées ont lieu avant les censurées,*
- *si il y a plusieurs décès au même temps $T_{(i)}$, alors $d_i > 1$ et on a*

$$\hat{S}(t) = \prod_{\substack{i=1,\dots,n \\ T_{(i)} \leqslant t}} \left(1 - \frac{d_i}{Y_i}\right).$$

Remarque 10 *Estimation empirique :*

Pour un échantillon i.i.d. de durées non censurées $(X_i)_{i=1,\dots,n}$, un estimateur "naturel" de la survie de la variable X est la survie empirique

$$S_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i > x\}}.$$

Cet estimateur a de bonnes propriétés en terme de convergence : convergence p.s (Glivenko-cantelli), convergence en loi du processus empirique associé vers un pont brownien.

Néanmoins, dans le cas des données censurées, la variable d'intérêt n'est plus la variable observée. Ainsi estimer la survie S par la survie empirique des données observées $(T_i)_{i=1,\dots,n}$ ($S_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{T_i > x\}}$) fournit une estimation biaisée de S (les censures (qui ne sont pas des décès) sont considérées comme des décès : il y a une sous estimation de la survie). Il en est de même si on estime la fonction de survie par la survie empirique des données observées non censurées (échantillon tronqué). Notons que quand il n'y a pas de censure, l'estimateur de Kaplan-Meier se réduit à la fonction de survie empirique.

1.2 Estimation de la variance de $\hat{S}(t)$

L'estimateur de Greenwood de la variance de l'estimateur de Kaplan-Meier est

$$\widehat{Var}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{i:T_{(i)}\leqslant t} \frac{d_i}{Y_i(Y_i - d_i)}.$$

Il est obtenu en utilisant l'approximation suivante,

$$\widehat{Var}\left(\log\left(\hat{S}(t)\right)\right) \approx \sum_{i:T_{(i)}\leqslant t} \frac{d_i}{Y_i(Y_i - d_i)},$$

et en appliquant la delta-méthode ($Var(f(Z)) \approx [f'(E(Z))]^2\, Var(Z)$) pour montrer que

$$\widehat{Var}\left(\log\left(\hat{S}(t)\right)\right) \approx \frac{1}{\hat{S}(t)^2} Var\left(\hat{S}(t)\right)).$$

Remarque 11 *Ce résultat s'obtient, de manière théorique, de la propriété de normalité asymptotique de l'estimateur de Kaplan-Meier.*

Théorème 1 En tout point de continuité de S, $t_0 \in [0, \tau]$ et $S(\tau^-) > 0$,

$$\sqrt{n}(\hat{S}(t_0) - S(t_0)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, V^2(t_0)),$$

avec

$$V^2(t_0) = -S^2(t_0) \int_0^{t_0} \frac{S(du)}{S^2(u)G(u)},$$

où $G(t)$ la fonction de survie de la variable C.

Considérons les quantités $H(t) = P(T > t)$ et $H_1(t) = P(T > t, \delta = 1)$. D'après l'hypothèse d'indépendance, on obtient les égalités suivantes

$$\begin{aligned} H(t) &= P(T > t) = P(X > t, C > t) = S(t)G(t) \\ H_1(t) &= P(T > t, \delta = 1) = P(X > t, C \geqslant X) = E\left(\mathbb{1}_{\{X>t\}} G(X^-)\right) \\ &= \int_t^\infty G(u^-) f(u) du = -\int_t^\infty G(u^-) S(du). \end{aligned}$$

Par conséquent, $H_1(dt) = G(t^-)S(dt)$ et on peut ainsi écrire

$$V^2(t_0) = -S^2(t_0) \int_0^{t_0} \frac{H_1(du)}{S(u)H(u)G(u^-)}.$$

En remplaçant les fonctions H et H_1 par leurs équivalents empiriques (calculables car les variables T et δ sont observées),

$$\hat{H}(u) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i>u\}} \text{ et } \hat{H}_1(u) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i>u, \delta_i=1\}}$$

et S par $\hat{S}$, on obtient l'estimateur suivant

$$\hat{V}^2(t) = -\hat{S}^2(t) \int_0^t \frac{\hat{H}_1(du)}{\hat{H}(u)\hat{H}(u^-)}.$$

Un estimateur de la variance de l'estimateur de Kaplan-Meier (qui converge presque sûrement vers la variance asymptotique de $\hat{S}$) est

$$\widehat{Var}(\hat{S}(t)) = \frac{1}{n}\hat{V}^2(t).$$

Avec les notations, Y_i le nombre d'individus à risque de subir l'événement juste avant le temps $T_{(i)}$ et d_i le nombre de décès en $T_{(i)}$, on remarque que

$$\hat{H}(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{T_{(i)}>u\}} = \frac{Y_i - d_i}{n},$$

$$\hat{H}(u^-) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{T_{(i)}\geqslant u\}} = \frac{Y_i}{n},$$

$$\hat{H}_1(du) = -\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{T_{(i)}\in[u;u+du],\delta_i=1\}} = -\frac{d_i}{n}.$$

Ainsi, on obtient

$$\widehat{Var}(\hat{S}(t)) = \hat{S}^2(t) \sum_{i:T_{(i)}\leqslant t}^{n} \frac{d_i}{(Y_i - d_i)\,Y_i}.$$

Dans chaque intervalle de temps, l'estimation de la survie est une proportion. On peut donc, sous certaines conditions, faire une approximation par la loi normale et ainsi obtenir un intervalle de confiance,

$$IC(\alpha) = \left[\hat{S}(t) \pm z_{\frac{\alpha}{2}} \sqrt{\widehat{Var}\left(\hat{S}(t)\right)}\right].$$

Il ne faut pas utiliser cet intervalle quand $\hat{S}(t)$ est proche de 0 ou de 1. En effet, l'intervalle étant symétrique autour de $\hat{S}(t)$, les bornes peuvent dépasser les valeurs 0 ou 1. On préfère utiliser l'intervalle de confiance de Rothman qui contournent cette difficulté :

$$IC(\alpha) = \frac{K}{K + \left(z_{\frac{\alpha}{2}}\right)^2}\left[\hat{S}(t) + \frac{\left(z_{\frac{\alpha}{2}}\right)^2}{2K} \pm z_{\frac{\alpha}{2}}\sqrt{\widehat{Var}\left(\hat{S}(t)\right) + \frac{\left(z_{\frac{\alpha}{2}}\right)^2}{4K^2}}\right],$$

avec $K = \frac{\hat{S}(t)(1-\hat{S}(t))}{\widehat{Var}\left(\hat{S}(t)\right)}$.

Remarque 12 *On montre que, sous certaines conditions, l'estimateur de Kaplan-Meier est uniformément consistant, asymptotiquement normal et presque sans biais quand le nombre d'individu à risque est grand.*

2 Estimateur de Nelson-Aalen du risque cumulé

2.1 Estimateur de Nelson-Aalen

Si la variable X admet une densité, on a par définition du risque cumulé, du risque instantané et de la densité

$$\Lambda(t) = \int_o^t \lambda(u)du = \int_o^t \frac{f(u)}{S(u)}du.$$

Dans le cas où X n'admet pas de dérivée en tout point de $\mathbb{R}^+$, on peut toujours définir le risque cumulé en utilisant la définition de la densité de X,

$$\Lambda(t) = -\int_o^t \frac{S(du)}{S(u-)}.$$

Considérons les quantités $H(t) = P(T > t)$ et $H_1(t) = P(T > t, \delta = 1)$ et introduisons $G(t)$ la fonction de survie de la variable C. D'après l'hypothèse d'indépendance, on obtient les égalités suivantes

$$\begin{aligned} H(t) &= P(T > t) = P(X > t, C > t) = S(t)G(t) \\ H_1(t) &= P(T > t, \delta = 1) = P(X > t, C \geqslant X) = E\left(\mathbb{1}_{\{X>t\}}G(X^-)\right) \\ &= \int_t^\infty G(u^-)f(u)du = -\int_t^\infty G(u^-)S(du). \end{aligned}$$

Par conséquent, $H_1(dt) = G(t^-)S(dt)$ et on obtient l'expression suivante pour le risque cumulé :

$$\Lambda(t) = -\int_o^t \frac{H_1(du)}{H(u^-)}.$$

Un estimateur "naturel" s'obtient en remplaçant les fonctions H et H_1 par leurs équivalents empiriques (calculables car les variables T et δ sont observées). Soient

$$\hat{H}(u) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i>u\}} \text{ et } \hat{H}_1(u) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{T_i>u,\delta_i=1\}}$$

l'estimateur de Nelson-Aalen est donné par les expressions suivantes

$$\hat{\Lambda}(t) = -\int_o^t \frac{\hat{H}_1(du)}{\hat{H}(u^-)} = \sum_{i:T_i\leqslant t} \frac{\sum_{j=1}^n \mathbb{1}_{\{T_j=T_i,\delta_j=1\}}}{\sum_{j=1}^n \mathbb{1}_{\{T_j\geqslant T_i\}}} = \sum_{i:T_i\leqslant t} \frac{d_i}{Y_i},$$

où Y_i représente le nombre d'individus à risque juste avant T_i et d_i représente le nombre de décès en T_i. L'estimateur de Nelson-Aalen est une fonction en escalier qui a un saut de taille d_i/Y_i à chaque instant de décés.

2.2 Estimation de la variance de $\hat{\Lambda}(t)$

En utilisant la théorie des processus de comptage et en faisant une approximation par une loi de Poisson, on montre que la variance de l'estimateur de Nelson-Aalen est,

$$\widehat{Var}(\hat{\Lambda}(t)) = \sum_{i:T_i \leqslant t}^{n} \frac{d_i}{Y_i^2},$$

où d_i et Y_i sont le nombre de décès et d'individus à risque en T_i.

Remarque 13 *On montre que, sous certaines conditions, l'estimateur de Nelson-Aalen est uniformément consistant, asymptotiquement normal et asymptotiquement sans biais (quand* $\mathbb{P}(Y_i = 0) \to 0$*).*

3 Autres estimateurs

3.1 Estimateur de Breslow du risque cumulé

Un estimateur du risque cumulé peut également être obtenu à partir de l'estimateur de Kaplan-Meier en utilisant la relation $\Lambda(t) = -\log(S(t))$:

$$\begin{aligned}\hat{\Lambda}_2(t) &= -\log\left(\hat{S}(t)\right) \\ &= -\sum_{i:T_i \leqslant t} \log\left(1 - \frac{d_i}{Y_i}\right).\end{aligned}$$

La variance de cet estimateur est donnée par

$$\widehat{Var}(\hat{\Lambda}_2(t)) = \sum_{i:T_i \leqslant t} \frac{d_i}{Y_i(Y_i - d_i)}.$$

3.2 Estimateur de Harrington et Fleming de la survie

A partir de la relation $S(t) = \exp(-\Lambda(t))$ et de l'estimateur de Nelson-Aalen, on peut en déduire un autre estimateur de la fonction de survie :

$$\begin{aligned}\hat{S}_2(t) &= \exp(-\hat{\Lambda}(t)) \\ &= \prod_{i:T_i \leqslant t}^{n} e^{-\frac{d_i}{Y_i}} \\ &\approx \prod_{i:T_i \leqslant t}^{n} (1 - \frac{d_i}{Y_i}), \qquad \text{si } \frac{d_i}{Y_i} \to 0,\end{aligned}$$

où d_i et Y_i sont le nombre de décès et d'individus à risque en T_i. En appliquant un développement limité, on retrouve l'estimateur de Kaplan-Meier. En utilisant la delta-méthode ($Var(f(Z)) \approx [f'(E(Z))]^2 Var(Z)$), on peut obtenir un estimateur de la variance de cet estimateur,

$$\begin{aligned}\widehat{Var}(\hat{S}_2(t)) &= \left(\hat{S}_2(t)\right)^2 \widehat{Var}(\hat{\Lambda}(t)) \\ &= \exp\left(-2\sum_{i:T_i\leqslant t}^{n}\frac{d_i}{Y_i}\right) \times \left(\sum_{i:T_i\leqslant t}^{n}\frac{d_i}{Y_i^2}\right).\end{aligned}$$

3.3 Estimation de la survie par la méthode actuarielle

La méthode actuarielle repose sur le même principe de construction que l'estimateur de Kaplan-Meier. La différence est que les probabilités conditionnelles sont estimées sur des intervalles fixés par l'utilisateur et non déterminés par les temps d'événements. Ces intervalles sont généralement de longueur égale, par exemple, un mois, un trimestre, une année.

Considérons, k intervalles de temps $[0, t_1[, [t_1, t_2[, ..., [t_{k-1}, \infty[$, fixés a priori. Définissons,

— d_i le nombre de décés dans le $i^{ème}$ intervalle $[t_{i-1}, t_i[$ (avec $t_0 = 0$ et $t_k = \infty$),
— n_{i-1} le nombre de sujets vivants au temps t_{i-1},
— c_i le nombre de sujets censurés dans l'intervalle $[t_{i-1}, t_i[$,
— r_i le nombre de sujets à risque dans l'intervalle $[t_{i-1}, t_i[$.

Afin de simplifier les calculs, on suppose généralement que les censures sont réparties uniformément dans l'intervalle, c'est-à-dire, que les sujets censurés sont exposés en moyenne un demi-intervalle. Dans le calcul des individus à risque, leur contribution pour l'intervalle $[t_{i-1}, t_i[$ est donc $c_i/2$. Le nombre d'invidus à risque pour l'intervalle $[t_{i-1}, t_i[$ est donc

$$r_i = n_{i-1} - \frac{c_i}{2}.$$

Alors la probabilité $p_i = P(X \leqslant t_i \mid X > t_{i-1})$ de mourir dans l'intervalle $[t_{i-1}, t_i[$ sachant que l'on était vivant en t_{i-1} est estimée par

$$\hat{p}_i = \frac{d_i}{r_i}.$$

L'estimateur de la fonction de survie est donc,

$$\hat{S}_3(t) = \prod_{i:t_i\leqslant t}\left(1 - \frac{d_i}{r_i}\right).$$

La formule de Greenwood permet d'obtenir une estimation de la variance,

$$\widehat{Var}\left(\hat{S}_3(t)\right) = \hat{S}_3(t)^2 \sum_{i:t_i\leqslant t}\frac{d_i}{r_i(r_i - d_i)}.$$

Les intervalles de confiance s'obtiennent de la même manière que pour Kaplan-Meier.

Chapitre III

Modèles semi-paramétriques

Le modèle de Cox est largement utilisé en analyse de données de survie. Il est employé lorsque l'objectif est d'évaluer l'effet de covariables sur la durée de vie.

1 Les modèles à hasards proportionnels

Ces modèles expriment un effet multiplicatif des diverses covariables sur la fonction de hasard (modèle à structure multiplicative). On introduit une fonction de hasard de base qui donne la forme générale du hasard et qui est commune à tous les individus. Les modèles à hasards proportionnels se caractérisent par la relation suivante, pour tout $t > 0$,

$$\lambda(t \mid Z) = \lambda_0(t)h(\beta, Z),$$

où Z est un vecteur de covariables, β le paramètre d'intérêt et h une fonction positive. La fonction de hasard est le produit d'une fonction qui ne dépend que du temps et d'une fonction qui n'en dépend pas. En général, on suppose que l'effet des covariables se résume à une quantité réelle $\beta' Z$, c'est-à-dire $\lambda(t \mid Z) = \lambda_0(t)h(\beta' Z)$.

Ce modèle est dit à risques proportionnels car, quels que soient deux individus i et j qui ont pour covariables Z_i et Z_j, le rapport des fonctions de hasard ne varie pas au cours du temps,

$$\frac{\lambda(t \mid Z_i)}{\lambda(t \mid Z_j)} = \frac{h(\beta' Z_i)}{h(\beta' Z_j)}.$$

Les fonctions de hasard sont donc proportionnelles. C'est une conséquence du modèle mais c'est aussi une hypothèse qu'il faudra vérifier. Le rapport des fonctions de hasard est par définition un risque relatif à l'instant t des sujets de caractéristiques Z_i par rapport aux sujets de caractéristiques Z_j.

Un cas particulier très important est le modèle de Cox, qui suppose que la fonction h est la fonction exponentielle, c'est-à-dire,

$$\lambda(t \mid Z) = \lambda_0(t) \exp(\beta' Z).$$

D'autres choix de fonctions h sont possibles, néanmoins la fonction exponentielle est très souvent utilisée dans la littérature car ses valeurs sont toujours positives et $\exp(0) = 1$.

Remarque 14 *Si λ_0 et/ou h ont une forme inconnue, le modèle est dit semi-paramétrique.*

2 Modèle de Cox

On se place dans le cadre du modèle de Cox,

$$\lambda(t \mid Z) = \lambda_0(t) \exp(\beta' Z),$$

où Z est un vecteur de covariables de dimension $p \times 1$ et β un vecteur $(p \times 1)$ de coefficient de régression.

Considérons,
— D le nombre de décès observés parmi les n sujets à l'étude,
— $T_1 < T_2 < \ldots < T_D$, les temps d'événements (décès) distincts,
— $(1), (2), \ldots, (D)$, les indices des individus décédés respectivement en $T_1, T_2, \ldots, T_D$,
— Z_i la valeur des covariables de l'individu i
— $R(T_i)$ l'ensemble des individus encore à risque à T_i^- (juste avant T_i),

2.1 Vraisemblance partielle de Cox

Le principe de la méthode est d'estimer uniquement le coefficient de régression β en considérant la fonction λ_0 comme un paramètre de nuisance. Par conséquent, on ne cherche pas à estimer λ_0. L'idée de Cox est qu'aucune information ne peut être donnée sur β par les intervalles pendant lesquels aucun événement n'a eu lieu, car on peut concevoir que λ_0 soit nulle dans ces intervalles (On suppose que les moments où se produisent les censures n'apportent peu ou pas d'information sur β). On travaille alors conditionnellement à l'ensemble des instants où un décès a lieu.

Supposons, dans un premier temps, qu'il n'y a qu'un seul décès à chaque temps d'événement (car le raisonnement provient du cas continu). La probabilité qu'il y ait un événement (décès) en T_i (dans l'intervalle $[T_i, T_i + \Delta t]$) est :

$$\sum_{j \in R(T_i)} \lambda_0(T_i) \exp(\beta' Z_j).$$

La probabilité que l'individu i subisse l'événement en T_i sachant qu'un événement a eu lieu en T_i vaut

$$\frac{\lambda_0(T_i) \exp(\beta' Z_{(i)})}{\sum_{j \in R(T_i)} \lambda_0(T_i) \exp(\beta' Z_j)} = \frac{\exp(\beta' Z_{(i)})}{\sum_{j \in R(T_i)} \exp(\beta' Z_j)}.$$

Le point important est que cette probabilité dépend uniquement du paramètre β.

Comme il y a des contributions à la vraisemblance à chaque temps de décès, la vraisemblance partielle de Cox est définie comme le produit sur les temps de décès. La vraisemblance (partielle) totale est donc

$$L_{Cox}(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta' Z_{(i)})}{\sum_{j \in R(T_i)} \exp(\beta' Z_j)}.$$

La vraisemblance partielle ne dépend pas de la fonction de hasard de base $\lambda_0(t)$. On peut donc estimer β, sans connaître la fonction de hasard de base, par maximisation de la vraisemblance partielle de Cox. La vraisemblance partielle n'est pas une vraisemblance dans le sens statistique du terme, mais elle se comporte comme telle. Ainsi, on peut développer une théorie asymptotique similaire et l'utiliser pour estimer et tester les coefficients de régression β.

2.2 Evénements simultanés

Le raisonnement précédent suppose des temps d'événements distincts. Dans le cas des données réelles, cette hypothèse n'est pas toujours vérifiée (ex : mesures tous les mois ou trimestres).

La probabilité que l'individu j décède en T_i est

$$p_j = \frac{\exp(\beta' Z_j)}{\sum_{k \in R(T_i)} \exp(\beta' Z_k)}.$$

En présence de plusieurs événements, la méthode "exacte" consiste à admettre que les événements se produisent les uns à la suite des autres. Cependant, on ne connaît pas l'ordre des événements, il faut donc considérer toutes les possibilités. Dans le cas de deux sujets s_1 et s_2 de caractéristiques Z_1 et Z_2 qui décèdent en T_i, la contribution exacte à la vraisemblance est

$$\frac{\exp(\beta' Z_1)\exp(\beta' Z_2)}{\sum_{j \in R(T_i)} \exp(\beta' Z_j) \times \sum_{j \in R(T_i)\setminus s_1} \exp(\beta' Z_j)} + \frac{\exp(\beta' Z_1)\exp(\beta' Z_2)}{\sum_{j \in R(T_i)} \exp(\beta' Z_j) \times \sum_{j \in R(T_i)\setminus s_2} \exp(\beta' Z_j)}.$$

Le problème de cette méthode est que le temps de calcul devient très long quand il y a beaucoup d'événements simultanés. Ainsi, on utilise le plus souvent l'approximation de Breslow qui consiste à supposer que la contribution des d_i événements en T_i est le produit des probabilités p_j pour les unités décédées en T_i $\left(i.e. \sum_{j \in R(T_i)} \exp(\beta' Z_j) \approx \sum_{j \in R(T_i)\setminus k} \exp(\beta' Z_j) \right)$,

$$L_B(T_i) = \prod_{\substack{j:\ unités \\ décédées\ en\ T_i}} p_j = \frac{\exp\left(\beta' \left(\sum_{\substack{j:\ unités \\ décédées\ en\ T_i}} Z_j\right)\right)}{\left(\sum_{k \in R(T_i)} \exp(\beta' Z_k)\right)^{d_i}}.$$

L'approximation de Breslow de la vraisemblance totale est

$$\prod_{i=1}^{D} L_B(T_i),$$

où D est le nombre de décès observés. La maximisation de cette vraisemblance est rapide. De plus, si le nombre d'événements simultanés n'est pas trop grand alors la méthode est assez précise.

2.3 Estimation

2.3.1 Estimation des coefficients de régression β

A partir de la vraisemblance partielle, on peut obtenir une estimation du vecteur de paramètre β de dimension $p \times 1$. Notons

$$\mathcal{L}(\beta) = \log(L_{Cox}(\beta)) = \sum_{i=1}^{D} \left[\beta' Z_{(i)} - \log \left(\sum_{j \in R(T_i)} \exp(\beta' Z_j) \right) \right],$$

et $U(\beta)$ la fonction score, c'est-à-dire le vecteur $p \times 1$ des dérivées premières de $\mathcal{L}(\beta)$,

$$\begin{aligned} U(\beta) &= \frac{\partial \mathcal{L}(\beta)}{\partial \beta} = \left(\frac{\partial \mathcal{L}(\beta)}{\partial \beta_1}, \ldots, \frac{\partial \mathcal{L}(\beta)}{\partial \beta_p}\right) \\ &= \sum_{i=1}^{D} \left[Z_{(i)} - \frac{\sum_{j \in R(T_i)} Z_j \exp(\beta' Z_j)}{\sum_{j \in R(T_i)} \exp(\beta' Z_j)} \right] \\ &= \left(\sum_{i=1}^{D} \left[Z_{(i),1} - \frac{\sum_{j \in R(T_i)} Z_{j,1} \exp(\beta' Z_j)}{\sum_{j \in R(T_i)} \exp(\beta' Z_j)} \right], \ldots, \sum_{i=1}^{D} \left[Z_{(i),p} - \frac{\sum_{j \in R(T_i)} Z_{j,p} \exp(\beta' Z_j)}{\sum_{j \in R(T_i)} \exp(\beta' Z_j)} \right] \right). \end{aligned}$$

L'estimateur de Cox $\hat{\beta}$ du coefficient de régression est solution de l'équation

$$U(\beta) = 0.$$

Il n'y a pas de solution exacte à ce problème. L'algorithme de Newton-Raphson est souvent utilisé par les logiciels pour obtenir une solution.

Un estimateur consistant de la matrice de variance-covariance de β peut se calculer à partir de l'inverse de la matrice d'information de Fisher,

$$\widehat{Var}(\hat{\beta}) = \left\{ I\left(\hat{\beta}\right) \right\}^{-1}$$

où le terme (i, j) de la matrice $I(\beta)$ est

$$[I(\beta)]_{i,j} = -\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta_i \partial \beta_j}.$$

2.3.2 Estimation du risque cumulé de base Λ_0

Après avoir estimé les coefficients de regression, on peut estimer le risque cumulé de base par l'estimateur de Breslow qui est une extension de l'estimateur de Nelson-Aalen,

$$\hat{\Lambda}_0(t) = \sum_{i:T_i \leqslant t} \frac{d_i}{\sum_{j \in R(T_i)} \exp(\hat{\beta}' Z_j)}$$

où d_i est le nombre de décès en T_i. Si $\hat{\beta} = 0$, on retrouve l'estimateur de Nelson-Aalen. On peut ensuite déduire un estimateur de la fonction de survie pour un vecteur de covariable Z,

$$S(t \mid Z) = \exp\left(-\int_0^t \lambda(u \mid Z)du\right)$$
$$\Longrightarrow \hat{S}(t \mid Z) = \exp\left(-\hat{\Lambda}_0(t)\exp(\hat{\beta}'Z)\right).$$

2.4 Tests

On souhaite souvent vérifier certaines hypothèses sur les coefficients de régression. Les tests, déduits des propriétés asymptotiques de $\hat{\beta}$, portent souvent sur l'hypothèse

$$H_0 : \beta = \beta_0.$$

Les trois statistiques suivantes sont utilisées.

— Statistique du rapport de vraisemblance (mesure la distance entre $\log L_{Cox}(\hat{\beta})$ et $\log L_{Cox}(\beta_0)$) :

$$\chi^2_{LRT} = 2\left[\log L_{Cox}(\hat{\beta}) - \log L_{Cox}(\beta_0)\right] \overset{H_0}{\rightsquigarrow} \chi^2(p)$$

— Statistique de Wald (mesure l'écart entre $\hat{\beta}$ et β_0) :

$$\chi^2_W = (\hat{\beta} - \beta_0)' I(\hat{\beta})(\hat{\beta} - \beta_0) \overset{H_0}{\rightsquigarrow} \chi^2(p)$$

— Statistique du Score (mesure la pente de la tangente en β_0) :

$$\chi^2_S = (U(\beta_0))' \left(I\left(\beta_0\right)\right)^{-1} U(\beta_0) \overset{H_0}{\rightsquigarrow} \chi^2(p)$$

Ces trois statistiques suivent, sous H_0, une loi de χ^2 à p degrés de liberté (où β est un vecteur de dimension p). La statistique du rapport de vraisemblance ne nécessite pas de calculer les dérivées secondes de la log-vraisemblance. La statistique du score ne nécessite pas l'estimation de $\hat{\beta}$.

On peut déduire de ces statistiques des tests partiels permettant de tester des hypothèses concernant certaines coordonnées de β. En particulier, supposons que l'on souhaite tester l'addition d'une nouvelle variable Z_p dans un modèle avec $(p-1)$ variables. On veut savoir si le modèle contenant la variable Z_p apporte plus d'information sur la distribution des durées de vie que le modèle sans cette variable. On teste l'hypothèse

$$H_0 : \beta = \beta_0$$

où $\beta = (\beta_1, ..., \beta_p)$ et $\beta_0 = (\beta_1, ..., \beta_{p-1}, 0)$, c'est-à-dire, on teste

$$H_0 : \beta_p = 0.$$

Dans ce cas,

— la statistique du rapport de vraisemblance :

$$\chi^2_{LRT} = 2\left[\log(L_{Cox}(\hat{\beta})) - \log(L_{Cox}(\hat{\beta}_0))\right] \overset{H_0}{\rightsquigarrow} \chi^2(1)$$

— la statistique de Wald :

$$\chi^2_W = (\hat{\beta} - \hat{\beta}_0)' I\left(\hat{\beta}\right)(\hat{\beta} - \hat{\beta}_0) \overset{H_0}{\rightsquigarrow} \chi^2(1)$$

— la statistique du Score :

$$\chi^2_S = \left(U(\hat{\beta}_0)\right)' \left(I\left(\hat{\beta}_0\right)\right)^{-1} U(\hat{\beta}_0) \overset{H_0}{\rightsquigarrow} \chi^2(1)$$

où $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)$ et $\hat{\beta}_0 = (\hat{\beta}_1, ..., \hat{\beta}_{p-1}, 0)$.

2.5 Interprétation des coefficients de régression

Par définition le risque relatif à l'instant t, pour deux vecteurs de covariables Z_i et Z_j, est égal à :

$$RR(t) = \frac{\lambda(t \mid Z_i)}{\lambda(t \mid Z_j)}.$$

Dans le modèle de Cox, le risque relatif est constant au cours du temps :

$$RR(t) = RR = \exp(\beta'(Z_i - Z_j)).$$

Ainsi, dans un modèle de Cox avec une seule covariable Z, le risque relatif est

— pour une variable binaire codée 0 et 1 : $RR = \exp(\beta)$,

— pour une variable binaire codée a et b : $RR = \exp(\beta(b-a))$,

— pour une variable continue, $\exp(\beta)$ correspond au risque relatif pour une augmentation d'une unité de la variable. Le risque relatif est constant pour une augmentation d'une unité de la variable quelle que soit la valeur de la covariable : c'est une hypothèse de log-linéarité.

Il y a donc deux hypothèses importantes à vérifier dans l'utilisation du modèle de Cox : l'hypothèse de risques proportionnels (risque relatif constant au cours du temps) et l'hypothèse de log-linéarité.

2.6 Adéquation du modèle

Le modèle de Cox suppose que les risques sont proportionnels. Plusieurs méthodes ont été proposées.

— Les méthodes de validation graphique sont peu puissantes mais constituent une première approche intéressante. Par exemple, pour une covariable à deux modalités, on peut estimer la fonction de survie dans les deux groupes et tracer les courbes

$$\ln\left[-\ln(\hat{S}(t))\right] = \ln\left[\int_0^t \hat{\lambda}(u \mid Z)du\right] = \ln\left[\hat{\Lambda}_0(t)\exp(\hat{\beta}Z)\right] = \ln\left(\hat{\Lambda}_0(t)\right) + \hat{\beta}Z.$$

On vérifie ensuite que les courbes obtenues, pour chaque modalité, présentent un écart constant au cours du temps (hypothèse du modèle de Cox). On peut également examiner les résidus de Schoenfeld.

— Une autre approche consiste à considérer une covariable dépendante du temps. Pour vérifier si une covariable Z vérifie l'hypothèse de risques proportionnels, on introduit un terme d'interaction entre le temps et la covariable ; on considère, par exemple, la variable $\beta Z + \gamma Z \log(t)$. Ainsi, la fonction de risque est

$$\lambda(t \mid Z) = \lambda_0(t) e^{\beta Z + \gamma Z \log(t)} = \lambda_0(t) \left[e^{\beta Z} t^{\gamma Z} \right]$$

et le rapport des risques pour deux covariables Z_i et Z_j est

$$\frac{\lambda(t \mid Z_i)}{\lambda(t \mid Z_j)} = e^{\beta(Z_i - Z_j)} t^{\gamma(Z_i - Z_j)}.$$

Pour une covariable binaire, on obtient

$$\frac{\lambda(t \mid Z = 1)}{\lambda(t \mid Z = 0)} = e^{\beta} t^{\gamma}.$$

On teste ensuite si γ est statistiquement différent de 0. Si le coefficient est statistiquement différent de 0, l'hypothèse de proportionnalité n'est pas respectée (le rapport dépend du temps). Quand une covariable ne vérifie pas l'hypothèse de proportionnalité des risques, on peut stratifier suivant les modalités de la covariable (mais perte de puissance) ou utiliser des coefficients de régression dépendants du temps (mais le modèle n'est plus à risques proportionnels).

— Les résidus de Cox-Snell permettent de tester la validité globale du modèle. En effet, on sait que la fonction de survie de la variable X vérifie

$$S(x) = \mathbb{P}(X > x) = \exp(-\Lambda(x \mid Z)) \qquad \text{où } \Lambda(x \mid Z) = \Lambda_0(x) \exp(\beta' Z).$$

Donc la variable la fonction de survie de la variable $V = \Lambda(X \mid Z)$ vérifie

$$\mathbb{P}(V > y) = \mathbb{P}(X > \Lambda^{-1}(y \mid Z)) = \exp(-y),$$

c'est-à-dire que la variable $V = \Lambda(X \mid Z)$ suit une loi exponentielle $\mathcal{E}(1)$. Il y a donc adéquation du modèle si le risque cumulé de la variable V est proche de la droite $y = x$ (qui correspond au risque cumulé d'une loi $\mathcal{E}(1)$).
On procède de la façon suivante :

1. On estime $\Lambda(. \mid Z)$ par l'estimateur semi-paramétrique $\hat{\Lambda}(. \mid Z) = \hat{\Lambda}_0(.) \exp(\hat{\beta}' Z)$.
2. Pour chaque temps observé T_i, $i = 1, ..., n$, on associe la variable r_i (les résidus de Cox-snell)
$$r_i = \hat{\Lambda}(T_i \mid Z) = \hat{\Lambda}_0(T_i) \exp(\hat{\beta}' Z).$$
3. On estime le risque cumulé des variables r_i par l'estimateur de Nelson-Aalen noté $\hat{\Lambda}_r$.
4. On trace donc les fonctions $y = \hat{\Lambda}_r(x)$ et $y = x$ sur un même graphique. En effet, si le modèle est correct, le risque cumulé $\hat{\Lambda}_r$ doit être approximativement égal à la droite $y = x$.

Le modèle fait également une hypothèse de log-linéarité, c'est-à-dire que le logarithme du risque est une fonction linéaire des Z,

$$\log \lambda(t \mid Z) - \log \lambda_0(t) = \beta' Z.$$

Cette hypothèse implique que le risque relatif ($RR = \exp(\beta_k)$) est constant pour une augmentation d'une unité quelle que soit la valeur de la covariable. Cela peut être contraignant dans le cas d'une variable continue. Par exemple, si l'âge est une variable explicative continue et que l'on étudie une maladie qui touche essentiellement les personnes âgées, le modèle supposera que le risque relatif est le même pour une augmentation de 1 an, que ce soit pour un âge de 30 ans ou pour un âge de 70 ans. Dans le cas de variables qualitatives, on peut considérer un codage dichotomique (par exemple, une variable codée 0,1,2 peut être recodée en utilisant deux variables binaires : (0,0), (0,1) et (1,1)).

2.7 Quelques extensions

2.7.1 Covariables dépendantes du temps

Le modèle de Cox permet de prendre en compte des covariables dépendantes du temps (traitement, marqueur biologique,...). Il faut néanmoins que $Z(t)$ soit prédictible, c'est-à-dire connue au temps t. Le traitement statistique est identique néanmoins, on peut faire les remarques suivantes.

- Il est nécessaire de connaître la valeur des covariables pour chaque temps d'événement. En effet, on a $L_{Cox}(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta' Z_{(i)}(T_i))}{\sum_{j \in R(T_i)} \exp(\beta' Z_j(T_i))}$. Ceci peut poser quelques problèmes dans le cas de marqueurs biologiques.
- L'interprétation devient difficile car le risque est spécifique à chaque histoire des covariables.
- L'hypothèse de hasard proportionnel est conservée. En effet, les fonctions de risque pour les différentes modalités d'une covariable restent proportionnelles et leurs rapports sont indépendants du temps. L'effet de la covariable ne varie pas au cours du temps, c'est la variable qui varie.
- L'utilisation de certaines covariables dépendantes du temps permet de tester l'hypothèse de risques proportionnels.

2.7.2 Modèle de Cox stratifié

Dans le cas où une variable qualitative ne vérifie pas l'hypothèse de hasards proportionnels, on peut considérer un modèle de Cox stratifié. Prenons l'exemple, d'une variable binaire Y codée 0 et 1, par exemple, le sexe (0 pour les hommes et 1 pour les femmes). Dans ce modèle, le risque de base est différent dans les deux strates mais les covariables Z agissent de la même manière sur les deux fonctions de hasard, c'est-à-dire,

$$\lambda(t \mid Z, Y = 0) = \lambda_0(t) \exp(\beta' Z),$$
$$\lambda(t \mid Z, Y = 1) = \lambda_1(t) \exp(\beta' Z).$$

L'effet des covariables est le même dans chaque strate. Les estimations obtenues par la méthode de la vraisemblance partielle sont applicables pour obtenir les paramètres (λ_0, λ_1

et β) du modèle. La vraisemblance partielle est calculée dans chacune des strates ; la vraisemblance totale est le produit des vraisemblances de chaque strate.

Le modèle de Cox stratifié fait l'hypothèse que les covariables Z agissent de la même manière dans chaque strate. Cette hypothèse peut être testée en utilisant le test du rapport de vraisemblance :

$$\chi^2_{LRT} = 2\left[\sum_{j=1}^{s} \log\left(L_{Cox}(\hat{\beta}_j)\right) - \log(L_{Cox}(\hat{\beta}))\right] \overset{H_0}{\rightsquigarrow} \chi^2(sp-p)$$

où s est le nombre de strates, p le nombre de covariables (dimension de $\hat{\beta}$), $\sum_{j=1}^{s} \log\left(L_{Cox}(\hat{\beta}_j)\right)$ est la log-vraisemblance en considérant un β différent dans chaque strate et $\log(L_{Cox}(\hat{\beta}))$ est la vraisemblance en considérant le même β dans chacune des strates.

2.7.3 Modèles de fragilité (frailty)

Le modèle de Cox (et les méthodes traditionnelles) suppose que la population est homogène (malgré la prise en compte de covariables). Néanmoins, cette hypothèse n'est pas toujours réaliste, notamment quand des covariables importantes ne sont pas observables ou inconnues. Par exemple, cela peut être des facteurs environnementaux ou génétiques. Les modèles fragilité permettent de prendre en compte l'hétérogénéité des observations.

Considérons une nouvelle covariable non observée Z_0. On suppose, comme dans le modèle de Cox, que l'effet des covariables se résume à une quantité réelle $\exp(\beta_0 Z_0)$, alors la fonction de risque est

$$\lambda(t \mid Z, Z_0) = \lambda_0(t) e^{\beta_0 Z_0} e^{\beta' Z}.$$

En notant $\omega = e^{\beta_0 Z_0}$, la variable aléatoire réelle positive (appelée "fragilité"), la fonction de risque devient

$$\lambda(t \mid Z, \omega) = \lambda_0(t) \omega e^{\beta' Z},$$

et la fonction de survie conditionnelle est

$$S(t \mid Z, \omega) = \exp\left(-\int_0^t \lambda_0(s) \omega e^{\beta' Z} ds\right) = \exp\left(-\omega e^{\beta' Z} \Lambda_0(t)\right).$$

Comme ω est une variable aléatoire, on s'intéresse à la fonction de survie moyennée sur ω. Cette quantité correspond à la fonction de survie marginale pour un individu quelconque

$$S(t \mid Z) = \int_0^\infty \exp\left(-v e^{\beta' Z} \Lambda_0(t)\right) f_\omega(v) dv = \mathcal{L}_\omega\left(e^{\beta' Z} \Lambda_0(t)\right),$$

où f_ω représente la densité de la v.a. ω et $\mathcal{L}_\omega(s) = \mathbb{E}(e^{-\omega s})$ est la transformée de Laplace de la distribution de la fragilité.

Le plus souvent, les modèles de fragilité sont utilisés pour prendre en compte une dépendance entre les temps d'événements de certains individus. En effet, les individus d'un même sous-groupe d'une population peuvent être liés si tous les individus de ce groupe ont des caractéristiques communes non observées. Par exemple, des individus d'une même famille, d'une même région ou d'un même hôpital. Le terme de fragilité est alors commun à chaque individu du groupe (permet de créer la dépendance) mais différent d'un groupe

à l'autre (hétérogénéité entre groupe). On parle de modèle à fragilités partagées (shared frailty model).

Considérons $T_{ij} = \min(X_{ij}, C_{ij})$ où j représente l'indice du $j^{ème}$ individu du groupe i $(i = 1, ..., G)$. Le risque pour l'individu j du groupe i est

$$\lambda_{ij}(t \mid Z_{ij}, \omega_i) = \lambda_0(t)\omega_i e^{\beta' Z_{ij}}, \tag{III.1}$$

où ω_i est la fragilité du groupe i. Les ω_i sont *i.i.d.* et en général, on suppose que $\mathbb{E}(\omega_i) = 1$ et $\mathbb{V}(\omega_i) = \theta$ pour des questions d'identifiabilité du modèle (dans ce cas, si $\omega_i > 1$, le risque du groupe i sera supérieur en moyenne au risque de base et inversement si $\omega_i < 1$). Le paramètre θ permet alors de mesurer l'hétérogénéité entre les groupes (une variance importante entraîne une grande variabilité entre groupes).

- Dans ce modèle, les observations sont indépendantes conditionnellement aux ω_i.
- La loi Gamma est souvent utilisée comme loi des effets aléatoires car elle a de bonnes propriétés mathématiques. Elle fournit de bons résultats en pratique et les $r^{èmes}$ dérivées de la transformée de Laplace ont une écriture simple.
- D'autres distributions sont possibles : loi inverse gaussienne, loi positive stable, ...
- On utilise souvent l'algorithme EM pour estimer les paramètres du modèle. On peut également passer par la transformée de Laplace dans le cas de la loi Gamma.
- Le modèle (III.1) implique que les risques sont proportionnels conditionnellement aux valeurs de la fragilité. Par conséquent, l'interprétation de β est conditionnelle à la fragilité. Par exemple, si $Z_{ij} = 0$ ou 1, cela signifie que e^{β} représente le risque entre un sujet codé 1 et un sujet codé 0 au sein d'un même groupe.
- Plusieurs généralisations sont possibles :
  - modèle de fragilité stratifiée : des risques de base différents dans chaque strate, par exemple, pour différencier les hommes et les femmes au sein d'une même famille ;
  - modèle avec une fragilité qui suit une loi multivariée (par exemple, pour prendre en compte le fait que dans une famille, les frères et les soeurs sont plus proches entre eux que les cousins) ;
  - modèle avec deux fragilités pour prendre en compte une dépendance causée par deux sortes de raisons (facteurs génétiques et facteurs environnementaux) ;
  - modèle avec une fragilité dépendante du temps.

Remarque 15 *Les modèles de fragilité peuvent également être utilisés pour prendre en compte l'hétérogénéité entre les individus d'une population. Dans ce cas, il y a une valeur ω_i par individu $(i = 1, ..., n)$,*

$$\lambda_i(t \mid Z_i, \omega_i) = \lambda_0(t)\omega_i e^{\beta' Z_i}.$$

Le paramètre $\mathbb{V}(\omega_i) = \theta$ mesure l'hétérogénéité entre les individus.

Chapitre IV

Modèles paramétriques

On suppose que la distribution des durées de survie appartient à une famille de loi paramétrique donnée. Ainsi, le modèle paramétrique peut être formulé en précisant la forme de l'une ou l'autre des cinq fonctions équivalentes qui définissent la loi de la durée : λ, Λ, f, S ou F. Néanmoins, on spécifie souvent la forme du risque instantané λ : constant, monotone croissant ou décroissant et en forme de $\cap$ ou de $\cup$.

Les estimateurs des paramètres du modèle sont ensuite obtenus en maximisant la vraisemblance des observations (par l'intermédiaire de méthodes itératives, par exemple l'algorithme de Newton-Raphson).

1 Risque instantané constant (loi exponentielle)

La loi exponentielle $\mathcal{E}(\theta)$, qui ne dépend que d'un paramètre θ, est la seule qui admet un risque instantané constant. Cette loi est aussi dite "sans mémoire" car la probabilité de décès pour un individu dans un certain laps de temps est la même quelle que soit sa durée de vie (*i.e.* $\mathbb{P}(X > s + t \mid X > t) = \mathbb{P}(X > s)$). Les quantités associées à cette loi sont :

$$\begin{aligned} f(t \mid \theta) &= \theta e^{-\theta t}, \qquad t \geqslant 0 \text{ et } \theta > 0, \\ \lambda(t \mid \theta) &= \theta, \\ S(t \mid \theta) &= e^{-\theta t}. \end{aligned}$$

Dans certaines applications, on peut découper le temps en plusieurs intervalles et considérer un θ_i différent pour chacun des intervalles (risque est constant sur chaque période mais varie d'une période à une autre).

2 Risque instantané monotone

2.1 Loi de Weibull

Considérons une loi de Weibull $W(\theta, \nu)$, alors

$$f(t \mid \theta, \nu) = \nu \left(\frac{1}{\theta}\right)^{\nu} t^{\nu-1} \exp\left(-\left(\frac{t}{\theta}\right)^{\nu}\right), \qquad t \geqslant 0 \text{ et } \theta, \nu > 0,$$
$$\lambda(t \mid \theta, \nu) = \nu \left(\frac{1}{\theta}\right)^{\nu} t^{\nu-1},$$
$$S(t \mid \theta, \nu) = \exp\left(-\left(\frac{t}{\theta}\right)^{\nu}\right).$$

Pour $\nu = 1$, on retrouve la loi exponentielle $\mathcal{E}(\frac{1}{\theta})$. Si $0 < \nu < 1$, le risque instantané est monotone décroissant ; si $\nu > 1$ le risque est monotone croissant.

2.2 Loi Gamma

Considérons une loi Gamma $G(\nu, \theta)$, alors

$$f(t \mid \nu, \theta) = \frac{\theta^{\nu}}{\Gamma(\nu)} t^{\nu-1} e^{-\theta t}, \qquad t \geqslant 0 \text{ et } \theta, \nu > 0,$$
$$F(t \mid \nu, \theta) = \frac{1}{\Gamma(\nu)} \int_0^{\theta t} u^{\nu-1} e^{-u} du,$$
$$\lambda(t \mid \nu, \theta) = \frac{f(t \mid \theta, \nu)}{1 - F(t \mid \theta, \nu)},$$

avec $\Gamma(\nu) = \int_0^{\infty} u^{\nu-1} e^{-u} du$. Pour $\nu = 1$, on retrouve la loi exponentielle $\mathcal{E}(\theta)$. Si $0 < \nu < 1$, le risque instantané est décroissant ; si $\nu > 1$ le risque est croissant.

2.3 Autres lois

Il existe de nombreuses lois avec des risques monotones, citons notamment les lois de Gompertz-Makeham, les mélanges de deux distributions exponentielles, les lois de Weibull exponentiées.

3 Risque instantané en $\cap$ et $\cup$

3.1 Lois de Weibull généralisée

La loi de Weibull est intéressante pour modéliser des risques monotones. Cependant elle devient mal adaptée quand les risques sont en forme de cloche. Une alternative est l'utilisation de la loi de Weibull généralisée $GW(\theta, \nu, \gamma)$:

$$\lambda(t \mid \theta, \nu, \gamma) = \left(1 + \left(\frac{t}{\theta}\right)^{\nu}\right)^{\frac{1}{\gamma}-1} \frac{\nu}{\gamma \theta^{\nu}} t^{\nu-1}, \qquad t \geqslant 0 \text{ et } \theta, \nu, \gamma > 0,$$
$$S(t \mid \theta, \nu, \gamma) = \exp\left[1 - \left(1 + \left(\frac{t}{\theta}\right)^{\nu}\right)^{\frac{1}{\gamma}}\right].$$

Pour $\gamma = 1$, on retrouve la loi de Weibull $W(\theta, \nu)$; pour $\gamma = 1$ et $\nu = 1$, on retrouve la loi exponentielle $\mathcal{E}(\frac{1}{\theta})$. En faisant varier les paramètres on peut obtenir des risques constants, monotones croissants et décroissants, avec des formes en $\cap$ (par exemple, $\theta = 10$, $\nu = 0.5$, $\gamma = 4$) et des formes en $\cup$ (par exemple, $\theta = 130$, $\nu = 5$, $\gamma = 0.1$).

Les risques en forme de cloche sont souvent présents dans le domaine du vivant. Par exemple, les risques instantanés en forme de $\cup$ comportent 3 phases : la période de mortalité infantile, la période risque faible et la période de vieillissement durant laquelle le risque augmente.

3.2 Autres lois

Les lois Log-logistiques, Log-normales et gaussienne inverse permettent de considérer des risques instantanés en forme de $\cap$.

4 Introduction de covariables

Dans l'approche paramétrique, les fonctions d'intérêts peuvent dépendre de covariables explicatives succeptibles d'influencer la survie. En plus d'ajuster les fonctions de survie à différents facteurs, ceci permettra de comparer les durées de survie (l'hypothèse nulle sera l'égalité des distributions de survie).

Considérons Z un vecteur de covariables. Notons que ces covariables peuvent dépendre du temps, cependant il est nécessaire de supposer que la valeur des covariables ne change pas entre deux mesures. Afin de simplifier les écritures on supposera dans ce qui suit que les covariables sont fixées au cours du temps. On suppose que les covariables vont modifier les fonctions de risque en suivant un modèle à risques proportionnels "de Cox" (d'autres modèles à risques proportionnels sont possibles), c'est-à-dire

$$\lambda(t \mid Z) = \lambda_0(t) \exp(\beta' Z)$$

où β est le vecteur des coefficients de régression. Les fonctions de survie et de densité correspondant à ces fonctions de risque sont données par

$$S(t \mid Z) = \exp\left(-\int_0^t \lambda(u \mid Z) du\right) = \exp\left(-\int_0^t \lambda_0(u) \exp(\beta' Z) du\right) = S_0(t)^{\exp(\beta' Z)}$$

$$f(t \mid Z) = -S'(t \mid Z) = \lambda(t \mid Z) \exp\left(-\int_0^t \lambda(u \mid Z) du\right) = \lambda_0(u) \exp(\beta' Z) \times S_0(t)^{\exp(\beta' Z)},$$

avec $S_0(t) = \exp\left(-\int_0^t \lambda_0(u) du\right)$.

Les paramètres du modèle s'obtiennent simplement par la méthode du maximum de vraisemblance.

4.1 Comparaison de deux groupes

Considérons la situation où l'on souhaite comparer les durées de survie de deux groupes A et B. On introduit la covariable suivante,

$Z = 0$ si l'individu appartient au groupe $A \Longrightarrow \lambda_A(t) = \lambda_0(t)$

$Z = 1$ si l'individu appartient au groupe $B \Longrightarrow \lambda_B(t) = \lambda_0(t) \exp(\beta)$.

Pour comparer les deux groupes, on estime le coefficient de régression β et on teste l'hypothèse nulle $H_0 : \beta = 0$ c'est-à-dire $H_0 : \lambda_A = \lambda_B$. On peut, à cet effet, utiliser les tests du rapport de vraisemblance, de Wald ou du score qui suivent asymptotiquement une loi de $\chi^2(1)$, sous H_0.

4.2 Exemple

Considérons un risque de base suivant une loi de Weibull $W(\theta, \nu)$, alors

$$\begin{aligned}
\lambda_0(t) &= \nu \left(\frac{1}{\theta}\right)^{\nu} t^{\nu-1}, \qquad t \geqslant 0 \text{ et } \theta, \nu > 0, \\
S_0(t) &= \exp\left(-\left(\frac{t}{\theta}\right)^{\nu}\right), \\
f_0(t) &= \nu \left(\frac{1}{\theta}\right)^{\nu} t^{\nu-1} \exp\left(-\left(\frac{t}{\theta}\right)^{\nu}\right).
\end{aligned}$$

D'après les résultats du début de la section, les fonctions de risque, de survie et de densité dans le cas où il y a des covariables sont

$$\begin{aligned}
\lambda(t \mid Z) &= \nu \left(\frac{1}{\theta}\right)^{\nu} t^{\nu-1} \times \exp(\beta' Z), \qquad t \geqslant 0 \text{ et } \theta, \nu > 0, \\
S(t \mid Z) &= \exp\left(-\left(\frac{t}{\theta}\right)^{\nu}\right)^{\exp(\beta' Z)}, \\
f(t \mid Z) &= \nu \left(\frac{1}{\theta}\right)^{\nu} t^{\nu-1} \times \exp(\beta' Z) \times \exp\left(-\left(\frac{t}{\theta}\right)^{\nu}\right)^{\exp(\beta' Z)}.
\end{aligned}$$

Pour $\nu = 1$, on retrouve la loi exponentielle $\mathcal{E}(\frac{1}{\theta})$. Ainsi, dans le cas d'un risque suivant une loi exponentielle avec des covariables, on obtient

$$\begin{aligned}
\lambda(t \mid Z) &= \frac{1}{\theta} \times \exp(\beta' Z), \qquad \theta > 0, \\
S(t \mid Z) &= \exp\left(-\frac{t}{\theta}\right)^{\exp(\beta' Z)}, \\
f(t \mid Z) &= \frac{1}{\theta} \exp(\beta' Z) \times \exp\left(-\frac{t}{\theta}\right)^{\exp(\beta' Z)}.
\end{aligned}$$

4.3 Modèles de vie accélerée (Accelerated Failure Time model)

Parmi les modèles de régression, les modèles de vie accélerée sont souvent considérés notamment en fiabilité. Ces modèles peuvent être définis de deux manières. La première représentation des modèles de vie accélérée est donnée par la fonction de survie accélérée :

$$S(t \mid Z) = S_0(te^{\beta' Z}),$$

où Z est un vecteur de covariable, β le vecteur des coefficients de régression. Le terme $e^{\beta' Z}$ est un facteur d'accélération car un changement dans les covariables change l'échelle de temps. On peut obtenir une expression de la fonction de risque,

$$\lambda(t \mid Z) = [-\ln(S(t \mid Z))]' = -\frac{[S(t \mid Z)]'}{S(t \mid Z)} = -\frac{-e^{\beta' Z} \times \lambda_0(te^{\beta' Z}) \times S_0(te^{\beta' Z})}{S_0(te^{\beta' Z})} = e^{\beta' Z}\lambda_0(te^{\beta' Z}).$$

En effet, on a les égalités suivantes,

$$S(t \mid Z) = S_0(te^{\beta' Z}) = \exp(-\Lambda_0(te^{\beta' Z})) = \exp\left[-\int_0^t \lambda_0(ue^{\beta' Z})du\right].$$

Si on suppose que $S_0(t)$ est la fonction de survie de la variable $\exp(\mu + \epsilon)$, alors $S_0(t) = P(e^{\mu+\epsilon} > t)$. Ainsi, on obtient que

$$S(t \mid Z) = S_0(te^{\beta' Z}) = P(e^{\mu+\epsilon} > te^{\beta' Z}) = P(e^{\mu-\beta' Z+\epsilon} > t) = P(X > t),$$

est la fonction de survie de la variable X où $\log(X) = \mu - \beta' Z + \epsilon$. En considérant le changement de variable $\alpha = -\beta$, on obtient la deuxième représentation par un modèle de régression log-linéaire pour la durée de survie

$$\log(X) = \mu + \alpha' Z + \epsilon,$$

où X est la durée de survie (pas toujours observée car $T = \min(X, C)$) et ϵ est une variable aléatoire (dans le cas de plusieurs observations, les ϵ_i sont *i.i.d.*).

Plusieurs lois sont possibles pour les variables ϵ, par exemple,
— $\epsilon \sim$ loi aux valeurs extrêmes ($f_\epsilon(y) = \exp(y - e^y)$)
— $\epsilon \sim$ log-logistic
— $\epsilon \sim$ log-normal
— $\epsilon \sim$ generalized gamma
On peut déduire la loi de X et les estimations des paramètres sont obtenues par maximisation de la vraisemblance.

Remarque 16 *On peut remarquer que dans le cas des modèles de vie accélerée, pour une covariable $Z > 0$, un coefficient de régression α négatif entraîne un temps de survie plus petit est donc une survie plus faible. Alors que dans le modèle semi-paramétrique de Cox un coefficient de régression α négatif entraîne un risque d'événement plus faible et donc une survie plus grande.*

Chapitre V

Comparaison de deux ou plusieurs fonctions de survie

1 Comparaison de deux groupes

Dans un premier temps, l'objectif est de comparer les durées de survie de deux groupes notés "A" et "B".

Il est possible de comparer les survies de deux groupes à un instant t donné. En effet, la survie au temps t est une proportion, ainsi, en utilisant l'approximation de la loi binomiale par la loi normale on montre que la statistique suivante

$$\frac{\hat{S}_A(t) - \hat{S}_B(t)}{\sqrt{\widehat{Var}(\hat{S}_A(t)) + \widehat{Var}(\hat{S}_B(t))}}$$

suit asymptotiquement une loi $\mathcal{N}(0,1)$ sous l'hypothèse $H_0 : S_A(t) = S_B(t)$. Néanmoins,

cela ne permet pas de tester (globalement) l'égalité des distributions de survie ce qui limite l'intérêt de la méthode.

1.1 Notations

On s'intéresse dans ce chapitre à une approche non-paramétrique. Le principe des tests consiste à comparer le nombre de décès observés dans chaque groupe au nombre de décès attendus (calculés sous l'hypothèse d'égalité des distributions de survie).

Considérons les notations suivantes,
— $T_1 < ... < T_N$ les temps de décès ordonnés des deux échantillons réunis,
— d_{Ai} et d_{Bi} le nombre de décès observés au temps T_i dans chacun des groupes A et B,
— $d_i = d_{Ai} + d_{Bi}$, le nombre total de décès observés en T_i,
— Y_{Ai} et Y_{Bi} le nombre de sujets à risques en T_i dans les groupes A et B,
— $Y_i = Y_{Ai} + Y_{Bi}$, le nombre total de sujets à risques en T_i.

Pour chaque temps d'événement T_i, l'information peut être résumée sous forme de tableau :

	Décès en T_i	Vivant après T_i	
Groupe A	d_{Ai}	$Y_{Ai} - d_{Ai}$	Y_{Ai}
Groupe B	d_{Bi}	$Y_{Bi} - d_{Bi}$	Y_{Bi}
	d_i	$Y_i - d_i$	Y_i

1.2 Statistiques de test

On cherche à tester l'hypothèse $H_0 : S_A(t) = S_B(t)$ qui est l'égalité des fonctions de survie dans les deux groupes. Ainsi, sous l'hypothèse H_0, la proportion attendue de décès (parmi les sujets à risque) est identique dans les deux groupes pour tous les temps de décès T_i. Pour chaque temps T_i, on peut comparer les pourcentages de décès parmi les sujets à risque dans chacun des groupes en utilisant le test du Chi-2.

Soit D_{Ai} (D_{Bi} et D_i) la variable dont la valeur est d_{Ai} (d_{Bi} et d_i), on peut montrer que D_{Ai} suit une loi hypergéométrique (*cf.* test du Chi-2) d'espérance :

$$\mathbb{E}(D_{Ai}) = \frac{Y_{Ai} \times d_i}{Y_i},$$

et de variance

$$\mathbb{V}(D_{Ai}) = \frac{(Y_i - d_i)}{(Y_i - 1)} \times \frac{d_i Y_{Ai} Y_{Bi}}{Y_i^2},$$

où $\mathbb{E}(D_{Ai})$ correspond au nombre de décès attendus dans le groupe A. Sous H_0, on montre que les variables $D_{Ai} - \mathbb{E}(D_{Ai})$ suivent asymptotiquement des lois $\mathcal{N}(0, \mathbb{V}(D_{Ai}))$ ($\frac{[D_{Ai} - \mathbb{E}(D_{Ai})]^2}{\mathbb{V}(D_{Ai})}$ suivent asymptotiquement des loi de χ_1^2).

Considérons des pondérations w_i, $i = 1, ..., N$, alors par indépendance (*cf.* Remarque) entre les variables D_{Ai} et D_{Aj} (associées aux T_i et T_j), les variables

$$\sum_{i=1}^{N} w_i \left(D_{Ai} - \mathbb{E}(D_{Ai})\right) = \sum_{i=1}^{N} w_i \left(D_{Ai} - \frac{Y_{Ai} \times d_i}{Y_i}\right)$$

suivent asymptotiquement des lois normales de moyennes nulles et de variances $\sum_{i=1}^{N} w_i^2 \mathbb{V}(D_{Ai})$. Par conséquent, sous H_0,les statistiques suivantes

$$\chi_0^2 = \frac{\left[\sum_{i=1}^{N} w_i \left(D_{Ai} - \frac{Y_{Ai} \times d_i}{Y_i}\right)\right]^2}{\sum_{i=1}^{N} w_i^2 \frac{(Y_i - d_i)}{(Y_i - 1)} \frac{d_i Y_{Ai} Y_{Bi}}{Y_i^2}}$$

suivent asymptotiquement des lois de χ^2 à 1 degré de liberté.

Remarque 17 *Le raisonnement précédent est valable pour toutes les "cases" du tableau 2×2, ce qui permet d'obtenir d'autres statistiques de test équivalentes.*

Remarque 18 *Les tests sont établis conditionnellement aux marges des tableaux et en supposant que les temps d'événements sont fixés. Les tableaux peuvent alors être traités comme des tableaux indépendants.*

Plusieurs statistiques de test ont été proposées

— **Test du logrank :** $w_i = 1$,
Cette pondération attribue à chaque décès le même poids quel que soit l'instant où il survient. Le test compare le nombres de décès observés au nombre de décès attendus.

— **Test de Gehan :** $w_i = Y_i$,
La pondération en T_i est égale au nombre d'individus à risque en T_i, donc les poids sont plus élévés pour les décès précoces que tardifs.

— **Test de Peto et Prentice :** $w_i = \prod_{k=1}^{i} \frac{Y_k}{Y_k+d_k}$,
Ces pondérations sont proches de l'estimateur de Kaplan-Meier de la survie. Elles attribuent des poids plus élévés aux décès précoces.

Remarque 19 *Il est important de noter que, par construction, ces tests sont valides uniquement si les fonctions de survie dans les deux groupes ne se croisent pas durant toute la période étudiée. Si les courbes se croisent, on observe une perte de puissance.*

Remarque 20 *Le test du Logrank est le test le plus souvent utilisé.*

Remarque 21 *Il existe un test du Log-rank approché plus facile à calculer à la main.*

2 Comparaison de plusieurs groupes

Les tests de la section précédente se généralisent au cas de la comparaison des fonctions de survie de plusieurs échantillons. Dans ce cours, seule l'extension du test du logrank sera envisagée (car c'est le test le plus utilisé).

Considérons le cas de trois groupes A, B et C; le tableau suivant résume les notations,

	Décès en T_i	Vivant après T_i	
Groupe A	d_{Ai}	$Y_{Ai} - d_{Ai}$	Y_{Ai}
Groupe B	d_{Bi}	$Y_{Bi} - d_{Bi}$	Y_{Bi}
Groupe C	d_{Ci}	$Y_{Ci} - d_{Ci}$	Y_{Ci}
	d_i	$Y_i - d_i$	Y_i

En suivant la même démarche que dans le cas de deux échantillons, on montre que le vecteur suivant

$$V = \begin{pmatrix} \sum_{i=1}^{N} (D_{Ai} - \mathbb{E}(D_{Ai})) \\ \sum_{i=1}^{N} (D_{Bi} - \mathbb{E}(D_{Bi})) \end{pmatrix}$$

avec

$$\mathbb{E}(D_{Ai}) = \frac{Y_{Ai} \times d_i}{Y_i}$$
$$\mathbb{E}(D_{Bi}) = \frac{Y_{Bi} \times d_i}{Y_i}$$

suit asymptotiquement une loi normale dans $\mathbb{R}^2$ d'espérance nulle. On en déduit ensuite que la statistique suivante

$$\chi_0^2 = V' \begin{pmatrix} \sum_{i=1}^{N} \mathbb{V}(D_{Ai}) & \sum_{i=1}^{N} Cov(D_{Ai}, D_{Bi}) \\ \sum_{i=1}^{N} Cov(D_{Ai}, D_{Bi}) & \sum_{i=1}^{N} \mathbb{V}(D_{Bi}) \end{pmatrix}^{-1} V$$

avec

$$\mathbb{V}(D_{Ai}) = \frac{(Y_i - d_i)}{(Y_i - 1)} \times \frac{d_i Y_{Ai}(Y_i - Y_{Ai})}{Y_i^2},$$
$$\mathbb{V}(D_{Bi}) = \frac{(Y_i - d_i)}{(Y_i - 1)} \times \frac{d_i Y_{Bi}(Y_i - Y_{Bi})}{Y_i^2},$$
$$Cov(D_{Ai}, D_{Bi}) = -\frac{(Y_i - d_i)}{(Y_i - 1)} \times \frac{d_i Y_{Ai} Y_{Bi}}{Y_i^2}$$

suit asymptotiquement une loi de de χ^2 à 2 degrés de liberté. Le raisonnement ci-dessus, avec le couple (A, B), est également possible avec les couples (A, C) ou (B, C) ce qui permet d'obtenir d'autres statistiques de test équivalentes.

Dans le cadre général de la comparaison de k groupes, la statistique de test s'obtient de la même façon et suit asymptotiquement une loi de χ^2 à $k - 1$ degrés de liberté.

Remarque 22 *Dans le cas où les différents groupes correspondent à une classification ordonnée, il paraît naturel de regarder l'existence d'une tendance reliant les fonctions de survie. Il est possible de construire une statistique de test (qui suit asymptotiquement une loi de $\chi^2(1)$) permettant de tester une telle hypothèse. Ce test repose sur la statistique permettant de tester la tendance entre plusieurs pourcentages.*

Remarque 23 *Dans la situation où on souhaiterait prendre en compte un facteur de confusion, on peut comparer les fonctions de survie en ajustant sur ce facteur (consiste à comparer les fonctions de survie pour une valeur donnée du facteur de confusion). A cet effet, il existe un test global dit test du logrank "stratifié" (les autres tests peuvent également être adaptés à cette situation), résumant les différences existant entre les deux groupes à l'intérieur de chacune des strates.*

Chaque strate constitue un sous-échantillon à partir duquel il est possible de calculer une différence entre le nombre de décès observés et attendus. On peut ensuite faire la somme sur chacune des strates (v.a. indépendantes car sujets différents).

Chapitre VI

Sujets non (ou partiellement) traités dans ce cours

— L'approche par les processus de comptage qui permet notamment, par l'intermédiaire de la théorie des martingales, de démontrer les propriétés des estimateurs et des tests.

— Le calcul du nombre de sujets nécessaires :

Il est notamment possible de calculer le nombre de sujets nécessaires pour comparer deux distributions de survie,
- — quand on suppose que la distribution de survie dans les deux groupes est exponentielle,
- — et dans un contexte non-paramétrique, quand on utilise le test du logrank.

— L'adéquation des modèles (hypothèse de risques proportionnels, hypothèse de log-linéarité, fonction exponentielle pour l'introduction des covariables, codage des covariables, valeurs manquantes, ...) n'est que partiellement abordée. En particulier, l'examen des résidus n'est pas exposé dans ce cours.

— De même, il existe d'autres tests pour comparer des fonctions de survie.

— Les modèles de régression et les modèles de vie accélérée pour lesquels

$$S(t \mid z) = S_0(te^{\beta' Z}).$$

— Les modèles de fragilité qui permettent de prendre en compte des corrélations entre individus. La fonction de risque est donnée par

$$\lambda(t \mid Z) = \lambda_0(t)\omega e^{\beta' Z},$$

où ω est une variable aléatoire positive appelée fragilité.

— Les modèles additifs :
Plutôt que de faire le produit entre une fonction de risque et une fonction des covariables (comme dans le modèle à risques proportionnels) on peut faire la somme. Dans les modèles additifs, la fonction de risque s'écrit,

$$\lambda(t \mid Z) = \lambda_0(t) + g(\beta' Z).$$

— Les modèles pour événements récurrents (l'événement étudié se répète pour un même individu).

— L'analyse de données de survie multivariée (par exemple, plusieurs événements pour un même individu, temps de survie de jumeaux).

⋮

De nombreux résultats nouveaux sont publiés régulièrement dans les revues scientifiques.

Chapitre VII

Implémentation en SAS et S-Plus

1 Modèles paramétriques

— Avec SAS
LIFEREG

— Avec S-Plus ou R
survreg ou censorreg

— Ces procédures permettent notamment de prendre en compte des durées censurées à droite, à gauche et par intervalles. Attention, les modèles de régression pour les durées de survie proposés par ces procédures n'ont été que partiellement traités dans ce cours (cf. Modèles de vie accélérée).

2 Estimation non-paramétrique et comparaison des courbes de survie

— Avec SAS
LIFETEST pour estimer les fonctions risque et de survie (Kaplan-Meier) et de comparer plusieurs fonctions de survie.

— Avec S-Plus ou R
survfit : pour estimer des fonctions de survie (Kaplan-Meier ou Fleming-Harrington).
survdiff : pour comparer des courbes de survie dans plusieurs groupes.

3 Modèles semi-paramétriques

— Avec SAS
PHREG pour considérer un modèle de Cox à risques proportionnels. Il est notamment possible de choisir la méthode pour les événements simultanés, de considérer un modèle de Cox stratifié (option strata), de prendre en compte des covariables

dépendantes du temps, d'étudier la validité des hypothèses du modèle, construire un modèle Cox à risques non proportionnels,...

— Avec S-Plus ou R

coxph : pour considérer un modèle de Cox à risques proportionnels. Il est notamment possible de choisir la méthode pour les événements simultanés, de considérer un modèle de Cox stratifié (option strata), de prendre en compte des covariables dépendantes du temps, ...

Chapitre VIII

Bibliographie

Quelques liens, notes de cours et livres concernant l'implémentation :

— SAS/STAT User's Guide :
http : *//www.math.wpi.edu/saspdf/stat/pdfidx.htm*

— Analyse des durées de survie avec R :
http : *//www.stat.nus.edu.sg/˜stachenz/Rsurv.pdf*

— Catherine Huber, *Cours de Modélisation Biostatistique en Splus* :
http : *//www.biomedicale.univ−paris*5.*fr/survie/enseign/cours_stat_avec_splus.pdf*

— Allison, Paul D. Survival Analysis Using the SAS System : A Practical Guide, Cary, NC : SAS Institute Inc, 1995.

— Terry M. Therneau, Patricia M. Grambsch. Modeling Survival Data : Extending the Cox Model. Statistics for Biology and Health. Springer, 2000.

Notes de cours sur internet

— Catherine Huber, *Analyse des durées de survie* :
http : *//www.biomedicale.univ* − *paris*5.*fr/survie/enseign/survie_sansi.pdf*

— Jean-David Fermanian, *Modèles de durées*, téléchargeable sur la page :
http : *//www.crest.fr/ses.php?user* = 2975

Livres

— Hill, C. ; Com-Nougue, C. ; Kramar, A. ; Moreau, T. ; O'Quigley, J. ; Senoussi, R. ; Chastang, C. . *Analyse statistique des données de survie.* Collection : Statistique en biologie et en médecine. Flammarion Sciences 1996 ; 3ème édition 2000.

— P. Klein, Melvin L. Moeschberge. *Survival Analysis.* Statistics for Biology and Health. Springer, 2005.

— Hougaard, Philip. *Analysis of multivariate survival data.* Statistics for Biology and Health. Springer, 2000.

⋮