

Méthodologie statistique

M 2018/02

**Modèles semi-paramétriques de survie
en temps continu sous **

**Simon Quantin
(DMCSI)**

Document de travail



Institut National de la Statistique et des Études Économiques

M 2018/02

Modèles semi-paramétriques de survie en temps continu sous 

Simon Quantin (DMCSI)


Cet article n'aurait pu être mené à son terme sans la motivation et le soutien indéfectible de Pauline Givord tout au long de son écriture, la relecture attentive d'Elise Coudin, Mathilde Poulhes, Patrick Sillard et Stéfan Lollivier et la mise en page précieuse dans des délais très courts de Gaëlle Cordani. Qu'ils et elles en soient ici tous remerciés infiniment.

Modèles semi-paramétriques de survie en temps continu sous



Simon Quantin*

Résumé

Ce document se veut une introduction pratique à la mise en œuvre sous  des modèles de survie en temps continu dans la cadre semi-paramétrique, souvent appelé modèle de Cox. Après avoir explicité la spécificité des modèles de survie, nous présentons comment mettre en œuvre le modèle à hasards proportionnels (et notamment comment questionner sa validité). Enfin, un chapitre est aussi consacré à la prise en compte de l'hétérogénéité individuelle inobservée.

Mots-clés : Modèles de survie, fragilité

Classification JEL : C24, C41

* INSEE. Auteur correspondant

Adresse : 88, avenue Verdier, CS 70058 92541 Montrouge cedex

Téléphone : (+33) 1 87 69 55 26

E-mail : simon.quantin@insee.fr

Table des matières

Introduction

1 Spécificité de l'analyse de la survie

- 1.1 Collecter les données
- 1.2 Données complètes, censurées et tronquées
 - 1.2.1 Censure 5
 - 1.2.2 Troncature 5
- 1.3 Durées de vie continues, discrètes ou groupées
- 1.4 Formalisation de l'analyse

2 Décrire la survie

- 2.1 Fonctions de survie et de hasard
- 2.2 Estimer la fonction de survie
- 2.3 Estimer la fonction de hasard
 - 2.3.1 L'estimateur de Nelson-Aalen de la fonction de risque cumulé 12
 - 2.3.2 Estimation par noyau de convolution de la fonction de hasard 12
- 2.4 Décrire la durée de vie
- 2.5 Mise en œuvre sous R
 - 2.5.1 Estimateur de Kaplan-Meier de la fonction de survie 15
 - 2.5.2 Estimateur de Nelson-Aalen de la fonction de hasard cumulé 16
 - 2.5.3 Estimateur par noyau de convolution de la fonction de hasard 17

3 Modéliser la durée de vie

- 3.1 Le modèle de Cox
 - 3.1.1 Hasards proportionnels et interprétation des paramètres 19
 - 3.1.2 Estimation et prise en compte des temps non distincts 20
- 3.2 Mise en œuvre sous R
 - 3.2.1 Interprétation des paramètres estimés 23
 - 3.2.2 Prédire la fonction de survie individuelle et illustrer les résultats 24
- 3.3 Modèles stratifiés et covariables dépendant du temps
 - 3.3.1 Modèles stratifiés 26
 - 3.3.2 Covariables dépendant du temps 28

4 Choix de la forme fonctionnelle

- 4.1 Résidus de martingale
- 4.2 *Smoothing splines*

5 Validité de l'hypothèse de proportionalité

5.1	Représentations graphiques	
5.2	Résidus de Schoenfeld (standardisés)	
5.3	Tenir compte de la non-proportionnalité	
5.3.1	Stratification	43
5.3.2	Modélisation d'un effet dépendant du temps	43
6	Hétérogénéité individuelle inobservée	
6.1	Le modèle de Cox avec fragilité individuelle	
6.2	Les fonctions de hasard et de survie marginales	
6.3	Distributions du paramètre de fragilité et conséquences	
6.3.1	Distributions gamma et inverse gaussienne de la fragilité	53
6.3.2	Distribution positive stable de la fragilité	54
6.3.3	Distribution mélangée de Poisson (<i>compound Poisson</i>) de la fragilité	54
6.4	Mise en œuvre sous R	

Bibliographie

Introduction

La *durée de vie* désigne le temps écoulé jusqu'à l'apparition d'un évènement précis (communément appelé « décès »), qui correspond à un changement d'état (typiquement passage de l'état « vivant » à « décédé »). Les modèles dits de *survie*, couramment utilisés en biostatistique, permettent ainsi d'analyser *si* un évènement particulier a lieu et *quand* celui-ci survient. En pratique, on s'intéresse à la distribution des durées de vie (*fonctions de survie*) et à la manière dont des variables explicatives les modifient.

La mise en œuvre d'une telle analyse nécessite dans un premier temps de comprendre et questionner la nature du phénomène étudié (par exemple l'évènement peut-il survenir à n'importe quel moment?) et des données collectées (comment sont sélectionnées les unités suivies? avec quelle précision la durée est-elle enregistrée?). Ces spécificités propres à l'analyse de la survie sont présentées dans le premier chapitre, dont l'objectif est d'expliquer au chargé d'étude à quelle problématique peut répondre la mise en œuvre des modèles de survie *en temps continu* présentés dans ce document.

Le deuxième chapitre introduit les notions de fonctions de survie, de hasard et de hasard cumulé, trois outils complémentaires pour décrire la durée de vie en temps continu, en lieu et place des plus classiques densité ou fonction de répartition. Si différents estimateurs non-paramétriques pour ces fonctions sont bien-sûr présentés, ce chapitre permet surtout de comprendre l'interprétation de la fonction de hasard. En effet, c'est à partir de cette fonction qu'est modélisé le lien de la durée de vie avec des covariables dans le modèle semi-paramétrique de Cox détaillé dans le troisième chapitre. Si ce modèle ne fait aucune hypothèse sur la dépendance au temps du risque de connaître l'évènement, il impose une proportionnalité des fonctions de hasard des individus observés qui implique une attention particulière lors de l'interprétation des paramètres du modèle. Dans ce chapitre enfin, deux extensions classiques sont introduites : le modèle stratifié et celui intégrant des covariables dépendant du temps.

Les quatrième et cinquième chapitres questionnent la validité du modèle spécifié en détaillant des méthodes pour aider le chargé d'étude à choisir la forme fonctionnelle du lien entre variable explicative et durée de vie et à vérifier la validité de l'hypothèse de proportionnalité propre au modèle de Cox. Enfin, le dernier chapitre explicite comment tenir compte de l'hétérogénéité individuelle inobservée dans le modèle de Cox par l'introduction d'une variable aléatoire dite de « fragilité » (*univariate frailty model*). Plus précisément, nous détaillons comment s'interprètent les paramètres dans ce type de modèle, mais aussi les conséquences du choix (nécessaire) de telle ou telle loi pour la variable de fragilité.

Si tout formalisme n'est pas exclu, ce document ne prétend pas se substituer à une présentation théorique complète des modèles de survie en temps continu telle que brillamment exposés dans les ouvrages de [Therneau et Grambsch \(2000\)](#) et [Duchateau et Janssen \(2010\)](#) dont ce document s'inspire

grandement. Il se veut plutôt un appui à une mise en œuvre pratique de ces modèles. L'implémentation sous **R** des méthodes est détaillée à chaque chapitre. Pour cela, nous utilisons deux packages; le package **survival** (Therneau et Grambsch, 2000) est à privilégier pour implémenter les modèles de durée, et donc de survie en temps continu et le package **frailtyEM** (Balan et Putter, 2018) nous apparaît à ce jour comme le plus complet pour tester différentes spécifications de modèle de fragilité. Si le chapitre 2 illustre l'estimation non paramétrique des fonctions de survie, de hasard et de hasard cumulé à partir de données simulées, les autres exemples s'appuient tous sur la base **pbc** du package **survival** qui est présentée à la section 3.2 du chapitre 3.

1 Spécificité de l'analyse de la survie

Les modèles de survie sont utilisés pour étudier *si* un évènement particulier a lieu et *quand* celui-ci survient, le cas échéant. Ainsi, par exemple, [Cooney et al. \(1991\)](#) étudient le risque de rechute de personnes alcooliques ayant suivi une cure de désintoxication dans un hôpital, et le nombre de jours qui séparent la sortie de l'établissement de la première consommation d'alcool, le cas échéant. De son côté, [Singer \(1993\)](#) s'interroge sur le nombre d'années qui séparent l'embauche d'un éducateur spécialisé de son départ éventuel de l'établissement. De même, [Bolger et al. \(1989\)](#) ont mené une enquête pour savoir si des étudiants avaient déjà eu des idées suicidaires et si oui, à quel âge.

Dans les trois études évoquées, on s'intéresse tout autant à *l'évènement d'intérêt* (début de la reprise de consommation d'alcool, fin de l'activité d'un enseignant spécialisé, première pensée suicidaire) qu'*au temps écoulé avant l'apparition de l'évènement*. Plus précisément, la durée dite *de survie* qui y est étudiée désigne le temps écoulé entre *deux états*, c'est-à-dire entre un état initial (abstinent à la sortie de l'hôpital, embauché, naissance) et la survenue d'un évènement d'intérêt final (rechute, fin d'activité, première pensée suicidaire). Implicitement, cela suppose donc que :

- L'occurrence d'un évènement est définie précisément, c'est-à-dire que *chaque état est exclusif l'un de l'autre* et fournit une *description complète des états possibles*. Ainsi, un ancien alcoolique est abstinent (état 1) jusqu'à ce qu'il ait recommencé à boire (état 2). De même, un éducateur spécialisé est en poste (état 1) jusqu'à ce qu'il ait quitté l'établissement (état 2).
- *Au début, tous les individus sont dans un seul et même état*. Ainsi, en sortant de l'hôpital, toutes les personnes sont abstinentes. De même, le jour de leur embauche, tous les éducateurs spécialisés enseignent. À la naissance, aucun nourrisson n'a de pensée suicidaire.
- *Une métrique du temps est spécifiée*, qu'il s'agisse du nombre de jours ou d'années, dans les deux premières études, ou de l'âge dans la troisième.

Par ailleurs, à chaque personne ne correspond qu'une seule période dans un état donné, et donc une durée mesurée.

De fait, contrairement aux modèles plus généraux dits de « durée », les modèles de « survie » explicités dans ce document n'étudient pas les durées de transitions entre plusieurs états (au moins trois) ou les différentes durées passées dans un état donné sur différentes périodes (par exemple, la durée passée sur plusieurs postes d'enseignement). Ils visent cependant tout autant à décrire la distribution des temps passés dans un état donné (par exemple, l'abstinence avant la rechute), à les comparer entre plusieurs groupes de personnes (ceux ayant suivi un traitement particulier pendant leur séjour à l'hôpital) ou à analyser la manière dont des variables explicatives la modifient (comme l'âge du patient par exemple). Si les méthodes qu'il faut mettre en œuvre seront décrites ultérieurement, ce chapitre vise à présenter plusieurs éléments caractéristiques à toute analyse des durées de survie. La première partie rappelle les différentes méthodes de collecte d'information qui peuvent être mises en œuvre. Dans une deuxième partie, les notions de censure et de troncature qui peuvent engendrer

des données incomplètes sont explicitées. Puis, nous reviendrons sur les différences entre durée de vie continue et discrète. L'objectif de ces trois parties est de permettre au lecteur de comprendre quel type d'analyse est mené dans ce document. En effet, la dernière partie pose plus formellement le cadre d'analyse, classique et fréquent en pratique, des données de survie continue en présence de censure aléatoire à droite qui fera l'objet des chapitres suivants.

1.1 Collecter les données

Il existe de nombreuses façons de constituer des bases de données de survie. Tout d'abord, les personnes enquêtées peuvent être sélectionnées selon différents processus :

1. *Échantillonnage de stock* : la base de données est constituée à partir (d'un échantillon) des personnes qui se trouvent dans l'état d'intérêt (par exemple, les personnes inscrites à Pôle Emploi) à un instant donné (par exemple, le 1^{er} janvier 2018). En général, la date d'entrée est connue (dans notre exemple, la date d'inscription à Pôle Emploi), et les personnes sont interrogées par la suite sur leur date de sortie.
2. *Échantillonnage de flux entrant* : la base de données est obtenue à partir (d'un échantillon) de toutes les personnes qui entrent dans l'état d'intérêt entre deux instants donnés. Par exemple, les nouveaux inscrits à Pôle Emploi entre le 1^{er} janvier et le 31 décembre 2018. Ces personnes sont alors suivies pendant un laps de temps donné ou jusqu'à leur sortie des listes de Pôle Emploi.
3. *Échantillonnage de flux sortant* : la base de données est constituée à partir des personnes qui quittent l'état d'intérêt entre deux instants donnés (par exemple, les sortants des listes de Pôle Emploi entre le 1^{er} janvier et le 31 décembre 2018).

Les données collectées peuvent aussi provenir d'une combinaison de ces types d'échantillonnage. Par exemple, si l'on retient toutes les périodes d'inscription à Pôle Emploi comprises entre deux dates. Certaines périodes correspondront à des inscriptions à Pôle Emploi commencées avant la date de début d'échantillonnage et toujours en cours (comme dans le cas de l'échantillonnage de stock) d'autres à des inscriptions qui commenceront après (comme dans l'échantillonnage de flux entrant).

Il est préférable cependant de disposer d'un échantillonnage de flux, car, dans le cas d'un échantillonnage de stock, il faut tenir compte de la sur-représentation des durées longues. Nous privilégions donc dans ce document *l'analyse de durée de survie issues d'échantillonnages de flux (entrant)*.

1.2 Données complètes, censurées et tronquées

Expliciter comment constituer les bases de données nécessaires pour étudier la durée qui sépare deux états permet de comprendre la nature spécifique des données qui seront utilisées. En effet, il coexiste, de fait, quatre types de dates engendrées par la collecte. La *date d'origine* marque le début de la période de suivi et donc de l'état initial. Elle peut correspondre à la date d'un évènement particulier (la sortie de l'hôpital ou la date d'embauche) ou la date de naissance de l'individu, ce qui implique que chaque personne peut donc avoir une date d'origine différente¹. La *date d'évènement* correspond à l'instant où l'évènement d'intérêt se produit et où cesse de fait le suivi. Enfin, la *date de fin de suivi* désigne la date à laquelle s'arrête le suivi alors que la *date de dernière nouvelle* désigne celle à compter de laquelle on ne dispose plus d'informations sur les personnes.

1. ce qui est de peu d'importance, puisque c'est la durée qui nous intéresse

Parce que les enquêtes ou remontées d'information sont souvent limitées dans le temps, une date de fin de suivi est souvent fixée pour tous les individus enquêtés. Bien évidemment, rien ne garantit que la date d'évènement ne se produise avant. Par ailleurs, à cause des problèmes de suivi propres aux enquêtes prolongées dans le temps (dû par exemple au déménagement de l'enquêté), il se peut aussi que la date de dernière nouvelle ne coïncide ni avec la date d'évènement, ni avec la date de fin de suivi. L'existence d'observations incomplètes, (évènement de début et/ou de fin inobservé), est de fait une des spécificités des modèles de survie. On parle alors de **données censurées ou tronquées**. Dans ce document, nous présenterons l'estimation de modèles de survie *censurées à droite aléatoirement*, qui est le cas le plus fréquemment rencontré. Néanmoins, il nous semble important à ce stade de préciser les différentes censures et troncatures auxquels peut être confronté le chargé d'études.

1.2.1 Censure

Une durée de vie est dite *censurée* si tout ce que l'on sait est qu'elle commence ou se termine dans un intervalle de temps particulier, et souvent en dehors de la période de suivi. La durée exacte n'est donc pas connue.

On distingue usuellement trois types de censure :

- *Censure à droite* : à la fin de la période de suivi, l'évènement d'intérêt ne s'est pas encore produit. On ne connaît donc pas la durée de vie T , mais seulement que $T > t$.
- *Censure à gauche* : il s'agit du cas où la date d'origine n'est pas observée de telle sorte que la durée de vie n'est, là encore, pas connue, que l'évènement d'intérêt se soit produit ou non.
- *Censure par intervalle* : la date de changement d'état n'est pas renseignée, mais un intervalle de temps est connu. Ce phénomène est caractéristique des enquêtes où le suivi est réalisé avec des rendez-vous réguliers. La seule information disponible sur la durée de survie est caractérisée par les dates des rendez-vous entre lesquelles l'évènement d'intérêt s'est produit.

Les différentes censures peuvent bien-sûr être présentes simultanément.

Au-delà de leur impact sur la date d'origine ou d'évènement, on distingue aussi - surtout - les censures par leur *mécanisme générateur*.

Censure de type I (fixée) : la durée n'est pas observable au-delà d'une durée maximale fixe (ou avant une date fixe, *identique pour tous les individus*). Ce type de censure provient donc de l'arrêt du recueil d'informations à une date fixée *a priori*.

Censure aléatoire (de type III) : il s'agit le plus souvent d'une information incomplète liée à un évènement non fixé par le protocole de suivi. Dans le cas d'une censure aléatoire à droite, il peut s'agir de l'apparition d'un évènement qui entraîne la sortie de l'étude avant la fin de la période de suivi fixée préalablement, comme le déménagement de l'individu, l'arrêt du traitement qui occasionne la sortie de l'étude, etc.

Censure de type II (attente) : le protocole de collecte suppose d'observer les durées de vie de n individus jusqu'à ce que R individus aient vu l'évènement d'intérêt se produire. Ainsi au lieu d'observer T_1, T_2, \dots, T_n , on observe seulement :

$$T_1 \leq T_2 \leq \dots \leq T_R$$

1.2.2 Troncature

Les données tronquées diffèrent complètement des données censurées. Elles correspondent à des durées qui ne sont pas observées en dessous d'un certain seuil (*troncature à gauche*) ou au dessus

d'un certain seuil (*troncature à droite*), lesquels peuvent être aléatoires². Contrairement aux données censurées, on ne dispose donc même pas de l'information sur l'existence d'un minimum (ou d'un maximum) pour la durée de survie.

Un exemple classique de troncature est celui des femmes toximanes enceintes, dont le suivi de la grossesse par un service spécialisé commence au premier rendez-vous pris. On retrouve alors la date de début de grossesse rétrospectivement. Mais les femmes qui ont un avortement avant le premier rendez-vous ne sont, elles, de fait pas suivies (la date du rendez-vous n'étant pas nécessairement fixé par échantillonnage...). Un échantillonnage de stock peut aussi mener à des problèmes de troncature. En effet, étudier la durée de vie d'un groupe de personnes vivants à une date donnée implique nécessairement que seule *leur* durée de vie est étudiée.

1.3 Durées de vie continues, discrètes ou groupées

Jusqu'ici nous avons implicitement considéré que l'évènement d'intérêt pouvait survenir à n'importe quel moment, ce qui signifie que le processus sous-jacent à la durée de vie est *continu*. Formellement, cela se traduit en considérant que la durée de vie T est une variable aléatoire continue, prenant ses valeurs sur \mathbb{R}^+ . Si ce formalisme correspond au continuum du temps, il pose la question de son adéquation au problème considéré et aux données. Deux raisons peuvent en effet nous conduire à envisager une durée de vie discrète.

La première raison de considérer des durées de vie discrètes est lorsque le processus sous-jacent est *intrinsèquement discret*. Par exemple, dans l'article de [Singer \(1993\)](#) sur la durée de l'activité des enseignants spécialisés, la date de fin ne peut avoir lieu qu'à des dates précises correspondant aux derniers jours de chaque année scolaire.

La deuxième raison découle du processus d'enregistrement de l'information. Les durées de vie sont souvent *groupées*³ (et comptées en nombre de mois, d'années) de telle sorte que les valeurs possibles se résument souvent à un ensemble réduit de valeurs discrètes, alors même que le processus sous-jacent est continu. De fait, il n'est pas possible d'être suffisamment précis dans la mesure du temps pour affirmer que les données ne sont jamais regroupées, et l'on envisage la distinction entre durées de vie continues et censurées/regroupées par intervalle par le nombre d'occurrences simultanées enregistrées.

Les méthodologies à mettre en œuvre dans le cas de durées de vie discrètes ne dépendent pas du processus sous-jacent qui les a engendrées (continu dans les données censurées par intervalle ou intrinsèquement discret). De telle sorte que l'on parle souvent de durées de vie discrètes sans plus de précision. Ces données n'en restent pas moins fréquentes et un chapitre particulier leur sera consacré.

1.4 Formalisation de l'analyse

Dans la plupart des cas, on disposera de données combinant des observations pour lesquelles la durée de vie est connue avec des observations censurées à droite de manière déterministe (censure de type I) ou aléatoire. C'est l'analyse de ce type de données qui seront présentées dans ce document. Comme nous l'avons précisé, elles seront obtenues à partir d'un échantillonnage en flux entrant.

2. la troncature par intervalles survient lorsque la durée est tronquée simultanément à gauche et à droite.

3. on parle parfois aussi de données regroupées par intervalles.

Formellement, comme nous l'avons déjà mentionné, la durée de vie est assimilée à la réalisation d'une variable aléatoire T , continue, qui ne prend que des valeurs positives. En présence de censure aléatoire à droite, on considère aussi une variable aléatoire latente, C , dont les réalisations correspondent à la durée écoulée avant la censure de l'information, comme la durée qui s'écoule avant le déménagement de la personne enquêtée qui engendre la fin du suivi.

Dès lors, à l'issue de la collecte des données, nous disposons, pour chaque individu, des données suivantes :


$$T_i^* = \min(T_i, C_i) \text{ et } \delta_i = \mathbb{1}_{T_i \leq C_i}$$

Or l'estimation dans le cas des modèles de survie s'appuie sur la vraisemblance statistique. D'un point de vue pratique, sa maximisation en présence de censure (et/ou de troncature) suppose de considérer que les processus de durée T et de censure (et/ou de troncature) C sont *indépendants*⁴. En présence de variables explicatives, cette hypothèse est remplacée par une indépendance conditionnellement aux covariables introduites. Au-delà de la simplification possible de l'écriture de la vraisemblance, cette hypothèse d'exogénéité rend possible d'un point de vue théorique à partir des observations T^* l'identification à une unique loi pour T .

Cette hypothèse nécessaire n'est cependant pas valide dans toutes les situations et doit être justifiée. Par exemple, dans le cas de l'évaluation d'un médicament, si la censure est due à l'arrêt du traitement, ou si les patients les plus malades ne sont plus suivis. À l'inverse, ce n'est pas le cas, si la censure est liée à la fin de l'étude ou occasionnée par un déménagement sans lien avec l'état de santé de l'enquêté.

4. La censure est parfois dite « non informative ». De fait, dans ce cadre, les « informations » en provenance de la loi de la censure peuvent être considérées comme constantes.

2 Décrire la survie

Décrire la durée de vie consiste à analyser la distribution des temps de survie. En effet, chaque durée de vie (censurée par la date de fin d'étude ou réellement observée) est assimilée à la réalisation t d'une variable aléatoire positive T , que l'on considèrera continue si l'évènement peut survenir à n'importe quel instant. Les valeurs possibles de T présentent une distribution qui peut être caractérisée par sa fonction de densité et sa fonction de répartition. Toutefois, dans le cas des modèles de survie, on privilégie plutôt deux notions statistiques particulières : la *fonction de survie* et la *fonction de hasard* (appelé aussi *taux de hasard instantané*). Dans ce chapitre, nous présentons donc d'abord ces deux notions (section 2.1) et leurs estimateurs respectifs (section 2.2 et 2.3). Puis nous montrons comment l'analyse conjointe de ces deux fonctions permet de décrire la durée de vie en temps continu (section 2.4) avant d'illustrer notre propos sous  (section 2.5).

2.1 Fonctions de survie et de hasard

La durée écoulée t avant la survenue d'un évènement correspond aux réalisations d'une variable aléatoire T continue et positive, de fonction de répartition $F(t)$ et de densité $f(t)$. Par définition, la fonction de répartition correspond à :

$$F(t) = \int_0^t f(u) du = \mathbf{P}(T \leq t)$$

et s'interprète comme la probabilité que l'évènement se réalise avant la date t considérée. La densité $f(t)$ est définie par

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

et correspond à la probabilité dite *instantanée* que l'évènement survienne dans l'intervalle de temps infinitésimal $[t, t+dt]$.

Si ces deux fonctions caractérisent la distribution de T , on s'intéresse cependant plutôt à la fonction de survie $S(t)$ qui représente la probabilité d'avoir survécu au-delà d'un instant t . Elle traduit donc la proportion de personnes encore susceptibles de connaître l'évènement d'intérêt à une date donnée. À la date d'origine ($t = 0$), tout le monde a « survécu » puisque personne n'a connu l'évènement : sa valeur est égale à 1. Au fur et à mesure que le temps passe et que le nombre d'individus qui ont connu l'évènement augmente, sa valeur décroît. La fonction $S(t)$ est donc toujours décroissante. Son expression est reliée à la fonction de répartition $F(t)$ par :

$$S(t) \equiv \mathbf{P}(T > t) = 1 - F(t)$$

Une autre grandeur permet de décrire la distribution de la durée T . En effet, la fonction de hasard évalue le risque que l'évènement survienne à un instant précis, *sachant qu'il n'a jamais été observé auparavant*. Elle permet d'étudier *si et quand* un individu a changé d'état depuis le début de l'étude et

se retrouve donc au cœur de l'analyse des durées de survie. Ses variations permettent d'identifier les instants où le risque est élevé que l'individu change d'état, mais aussi d'étudier comment ce risque évolue au fur et à mesure que le temps s'écoule. Formellement, son expression correspond à la limite d'un ratio et l'on parle parfois de *taux de hasard instantané* :

$$h(t) = \lim_{dt \rightarrow 0^+} \frac{P(t \leq T < t + dt | T \geq t)}{dt} = \frac{f(t)}{S(t)}$$

Cette expression suscite deux remarques importantes. Tout d'abord, en temps continu, on fait l'hypothèse que l'évènement peut survenir à n'importe quel instant. On suppose donc une infinité de dates d'évènements possibles (contrairement au cas dit discret) que seule limite la précision des données (seconde, heure, jour, semaine etc.). Cette hypothèse a deux conséquences :

- la probabilité d'observer un évènement à une date précise est nulle, ou plus intuitivement tend vers zéro au fur et à mesure que l'on affine la métrique du temps. Cela explique pourquoi on définit *par unité de temps* (dt) la fonction de hasard, mais aussi les profils heurtés que l'on obtient souvent dans les estimations.
- la probabilité d'observer des individus ayant *précisément* la même date d'évènement est aussi nulle. Bien évidemment, la présence de « jumeaux » est néanmoins possible dans le cas continu, puisque l'enregistrement des dates d'évènements ne peut jamais être assez précis pour se prémunir d'une telle éventualité. Mais cette hypothèse est un fondement théorique de l'analyse des durées de vie en temps continu et oblige à une prise en compte particulière des « jumeaux ».

Par ailleurs, la fonction de hasard est une probabilité *conditionnelle* (puisque l'évènement ne doit pas avoir eu lieu avant l'instant considéré) *par unité de temps*. En cela, elle dépend donc de la précision avec laquelle les dates d'évènement sont enregistrées (en seconde, en jour etc.). Parce qu'elle se rapporte à une unité de temps, il est possible de « lire » la fonction de hasard comme le nombre attendu d'évènements qui se produiront dans une période donnée. Par exemple, si le hasard *mensuel* de retour à l'emploi des chômeurs est constant et de 0.10, alors on peut s'attendre à ce que 10 % des personnes *encore au chômage* retrouvent *chaque mois* un emploi. De plus, son expression comme ratio autorise ses valeurs à être plus grandes que 1, au contraire d'une probabilité¹.

La fonction de survie et de hasard (comme la fonction de répartition ou la densité) caractérisent chacune entièrement la loi de la variable aléatoire continue T supposée décrire la durée de vie. En pratique, cependant, les deux sont utiles pour décrire efficacement et simplement la durée de vie.

2.2 Estimer la fonction de survie

L'estimateur le plus fréquemment utilisé de la fonction de survie est l'estimateur de Kaplan-Meier. Supposons que l'on dispose d'un échantillon de n durées, qu'elles soient censurées (à droite) ou observées, et $0 = t_{(0)} < t_{(1)} < t_{(2)} < \dots < t_{(j)} < \dots < t_{(m-1)} < t_{(m)} < \infty$, les dates distinctes et ordonnées auxquelles on observe un changement de status (et non une censure). En notant, d_j le nombre d'évènements à la date $t_{(j)}$ et n_j le nombre de personnes encore *à risque juste avant* $t_{(j)}$ (i.e. qui sont toujours suivies, et donc notamment toujours non censurées), l'estimateur de Kaplan-Meier (parfois appelé estimateur produit limite) est :

$$\hat{S}(t) = \prod_{j: t_{(j)} \leq t} \left(1 - \frac{d_j}{n_j} \right) \quad (2.1)$$

1. tout comme, bien évidemment, la densité $f(t)$ d'une variable aléatoire continue, à laquelle elle est liée.

Son principe est le suivant. Pour survivre au temps t , il faut avoir survécu au temps $t_{(1)}$. Il faut aussi survivre entre $t_{(1)}$ et $t_{(2)}$ sachant que l'on a survécu au-delà de $t_{(1)}$ et ainsi de suite. Par ailleurs, la probabilité de connaître l'évènement à l'instant $t_{(j)}$, sachant qu'on ne l'a pas connu avant, peut être estimé par $\frac{d_j}{n_j}$. Dès lors, la probabilité conditionnelle de survivre au-delà de $t_{(j)}$ est $1 - \frac{d_j}{n_j}$ et la probabilité (non conditionnelle) de survivre au-delà de la date t est obtenue par le produit des probabilités conditionnelles de survie sur les intervalles antérieurs².

L'estimateur de Kaplan-Meier est donc une fonction en escalier qui présente des « sauts » aux seules dates $t_{(j)}$ où l'on observe un évènement. De plus, il n'est pas possible sans hypothèse supplémentaire d'obtenir d'estimation de la fonction de survie au-delà de la dernière date d'évènement observé. Enfin, cet estimateur est l'estimateur non-paramétrique du maximum de vraisemblance. Peterson (1977) a montré que cet estimateur est convergent et Breslow et Crowley (1974) que $\sqrt{n}(\hat{S}(t) - S(t))$ converge en loi vers un processus gaussien³.

Un estimateur (convergent) de la variance de l'estimateur de Kaplan-Meier est donné par la formule de Greenwood :

$$\widehat{Var}(\hat{S}(t)) = \hat{S}^2(t) \sum_{j:t_{(j)} \leq t} \frac{d_j}{(n_j - d_j)n_j}$$

Plusieurs méthodes sont utilisées pour déterminer un intervalle de confiance de $\hat{S}(t)$. En pratique, elles visent surtout à s'assurer que les intervalles de confiance sont bien compris entre 0 et 1, notamment lorsque la fonction de survie estimée est proche de 0 ou de 1. Ainsi, un intervalle de confiance peut être calculé en s'appuyant sur la loi normale suivie asymptotiquement par l'estimateur de Kaplan-Meier ou plus fréquemment en considérant que $\log\{-\log\{\hat{S}\}\}$ suit approximativement une loi normale, ce qui conduit à construire l'intervalle de confiance :

$$\hat{S}(t)^{\exp\{\pm z_{1-\alpha/2}\sigma_S(t)/\ln(\hat{S}(t))\}}$$

Le calcul par défaut de l'intervalle de confiance diffère entre les différents logiciels, mais les deux méthodes présentées ici sont toujours implémentées.

2.3 Estimer la fonction de hasard

Après avoir obtenu une estimation de la fonction de survie, il est assez naturel de souhaiter obtenir une estimation de la fonction de hasard. La méthode développée par Kaplan-Meier s'appuie sur une décomposition du temps en intervalles dont les bornes sont les différentes dates d'évènement observées. Chaque intervalle commence donc à la date d'un évènement et s'arrête juste avant la date d'évènement suivante. Par construction, la longueur de l'intervalle dépend du temps qui s'écoule entre

2. En l'absence de données censurées, l'estimateur de Kaplan-Meier coïncide avec l'estimateur de la fonction de survie empirique :

$$\hat{S}_n(t) = \frac{1}{n} \sum_{i=1, \dots, n} 1_{[t_{(i)} > t]}$$

3. Dès lors que les variables aléatoires de durée de vie T et de censure C sont indépendantes - comme supposé ici - et que leurs lois ne présentent aucune discontinuité commune - ce qui est assuré dès lors qu'il s'agit de lois continues. Enfin, il convient de mentionner que de telles analyses de cet estimateur sont désormais supplantées par l'approche par processus de comptage qui ont enrichi considérablement l'analyse des modèles de durée.

deux évènements successifs. Dès lors, estimer directement la fonction de hasard conduit à observer des variations erratiques selon la durée de la période considérée rendant leur interprétation difficile.

La fonction de hasard cumulé (également appelé risque cumulé) permet de contourner cette difficulté. Elle correspond au « total » des risques (hasards) instantanés auquel l'individu a été confronté depuis la date d'origine. Dans le cas continu, il s'agit formellement de :

$$H(t) = \int_0^t h(u) du$$

La fonction de hasard h (qui est donc la dérivée de $H(t)$) correspond ainsi à la « pente » de la fonction de risque cumulé. Dès lors une démarche en deux temps peut être envisagée pour déterminer une estimation de la fonction de hasard :

1. estimer la fonction de hasard cumulé $H(t)$
2. en déduire par différences successives de $H(t_j)$ la « pente » $h(t_j)$ et lisser les estimations obtenues

2.3.1 L'estimateur de Nelson-Aalen de la fonction de risque cumulé

Il existe deux méthodes pour estimer la fonction de hasard cumulé. La première découle de la relation qui lie la fonction de survie et la fonction de risque cumulé :

$$\begin{aligned} S(t) &= \exp[-H(t)] \\ \implies \widehat{H}_{Br}(t) &= -\log \widehat{S}(t). \end{aligned}$$

Cet estimateur est appelé l'estimateur de Breslow.

On lui préfère cependant l'estimateur de Nelson-Aalen :

$$\widehat{H}_{NA}(t) = \sum_{j:t_{(j)} \leq t} \left(\frac{d_j}{r_j} \right) \quad (2.2)$$

où d_j représente le nombre de décès à $t_{(j)}$ et r_j , le nombre de personnes à risque à $t_{(j)}$. Cet estimateur permet ensuite d'obtenir un autre estimateur de la fonction de survie, dit de Fleming-Harrington $\widehat{S}_{FH}(t) = \exp(-\widehat{H}_{NA}(t))$. Les deux estimateurs sont asymptotiquement équivalents.

Pour des échantillons de taille plus réduite, l'estimateur de Kaplan-Meier est un meilleur estimateur de la fonction de survie, alors que l'estimateur de Nelson-Aalen est un meilleur estimateur de la fonction de risque cumulé.

2.3.2 Estimation par noyau de convolution de la fonction de hasard

Puisque la fonction de hasard correspond à la « pente » du risque cumulé, une estimation pourrait se faire à partir des différences successives de $\widehat{H}(t_j)$, obtenues par l'estimateur de Nelson-Aalen ou se déduire de l'estimateur de Kaplan-Meier de la fonction de survie. Toutefois, l'estimation de la fonction de hasard ainsi obtenue (basée sur deux points!) présente encore un profil heurté que l'estimation par noyau de convolution présentée dans cette section permet de lisser.

Schématiquement, il ne s'agit pas d'obtenir une valeur précise du hasard à chaque instant t , mais plutôt une valeur correspondant à la moyenne des hasards au voisinage de t , communément appelée *fenêtre de lissage*. Cette méthode nécessite de choisir la longueur de la fenêtre de lissage. Il n'existe pas de « bonne » longueur qui permettrait d'obtenir une « vraie » estimation de la fonction de hasard. Plus la longueur de la fenêtre de lissage est petite, plus le biais est faible. La variance est cependant plus

grande et l'estimé de la courbe n'est donc pas très lisse. À l'inverse, plus la fenêtre de lissage est grande, plus le biais est élevé, mais la variance est faible et l'estimé de la courbe est très lisse. Enfin, accroître la taille de la fenêtre de lissage restreint aussi la fenêtre temporelle sur laquelle les estimations de la fonction de hasard peuvent être obtenues, à cause d'effets de bord. Pour toutes ces raisons, il est souvent utile de comparer différentes tailles de fenêtre de lissage.

2.4 Décrire la durée de vie

Pour décrire une durée de vie en temps continu, il est nécessaire d'examiner simultanément :

- l'estimateur de Kaplan-Meier de la fonction de survie,
- l'estimateur de Nelson-Aalen de la fonction de risque cumulé,
- l'estimateur par noyau de convolution de la fonction de hasard.

L'analyse de l'estimation de la fonction de hasard est souvent mise en avant. En effet, contrairement aux fonctions de risque cumulé et de survie qui « agrègent » l'information sur les risques passés encourus, seule la fonction de hasard identifie l'ampleur du risque à une date donnée. Son profil permet donc d'identifier les périodes où le risque est élevé de celles où le risque est plus faible. On s'intéresse donc souvent aux « pics » qui apparaissent au fur et à mesure que le temps s'écoule. Au-delà de ces considérations, il est cependant utile de rappeler l'importance de la métrique du temps. Dans le cas continu, comme nous l'avons précisé, la fonction de hasard est une probabilité conditionnelle *par unité de temps*. Les valeurs estimées à chaque date dépendent donc de la métrique retenue. Cela implique d'une part que pour comparer deux estimations de fonction de hasard, il convient de s'assurer au préalable que la métrique du temps est identique dans les deux études. D'autre part, il est parfois utile de ne pas conserver la métrique d'enregistrement des données pour décrire la durée de vie. À partir du suivi journalier disponible dans le fichier historique de Pôle Emploi, il peut par exemple être préférable de commenter un taux de retour à l'emploi de 3% par semaine qu'un taux de retour à l'emploi journalier de 0,43% (= 0,03/7). Rappelons aussi qu'un changement de métrique n'implique pas d'arrondir dans le même temps les durées de vie observées, puisque cela diminuerait la qualité de l'information utilisée, augmentant par ailleurs le nombre d'évènements simultanés.

Par définition, toutes les fonctions de survie présentent un profil monotone et décroissant. À la date d'origine, elles sont toutes égales à 1 et diminuent au fur et à mesure que le temps s'accroît. Parce que dans la période d'étude, tous les individus ne connaîtront pas l'évènement et/ou certains seront censurés, il est fréquent que la fonction de survie n'approche pas 0 à la date de fin de l'étude, voire que la durée ne soit pas observée pour plus de 50 % des individus⁴. C'est donc sur la vitesse de décroissance que porte l'analyse du profil de la fonction de survie. Tout d'abord, celui-ci témoigne du profil de la fonction de hasard. Ainsi :

- lorsque le hasard est élevé, la fonction de survie décroît rapidement ;
- lorsque le hasard est faible, la décroissance est légère ;
- lorsque le hasard est nul, la fonction de survie est stable.

Étudier la fonction de survie permet donc d'illustrer aussi l'enchaînement des périodes de risque élevé et de risque faible, mais avec une grandeur plus intuitive que le hasard. Par ailleurs, elle offre aussi une estimation de la proportion de personnes encore confrontées à un risque à chaque date. Dès lors, si la fonction de survie a une valeur élevée quand le risque est grand, de nombreux individus

4. ce qui excluerait de pouvoir estimer la durée de vie médiane, par exemple.

connaîtront l'évènement. À l'inverse, une valeur faible de la fonction de survie, même en présence d'un risque élevé, relève que peu de personnes sont susceptibles de changer d'état. Il est donc nécessaire de coupler l'analyse du profil de la fonction de survie et de hasard pour offrir une description pertinente de la durée de vie⁵.

L'apport de la fonction de risque cumulé est plus subtil. Tout d'abord, l'estimation de la fonction de hasard par noyau de convolution réduit la fenêtre temporelle à cause des effets de bord inhérents à ce type de méthode. Ainsi, seule l'analyse des inflexions du profil de la fonction de hasard cumulé fournit une information sur l'ampleur du risque encouru au début et à la fin de la période d'étude. D'autre part, il faut rappeler que l'estimation de la fonction de hasard obtenue par lissage ne correspond pas exactement à une estimation de la fonction de hasard de la population.

2.5 Mise en œuvre sous R

Nous explicitons dans cette section comment estimer sous R la fonction de survie, de hasard cumulé et de hasard pour décrire une durée de vie continue. Pour cela, nous allons tout d'abord simuler un jeu de données de durée de vie, censurées à droite, suivant une distribution exponentielle. L'intérêt d'utiliser, dans ce chapitre, des données suivant une distribution connue est de pouvoir illustrer les relations qui lient ces différentes fonctions. En effet, dans le cas d'une durée de vie T qui suit une loi exponentielle de paramètre θ :

- la fonction de hasard est constante : $h(t) = \theta$,
- la fonction de hasard cumulé croissante et linéaire : $H(t) = \theta t$
- et la fonction de survie exponentielle : $S(t) = \exp^{-\theta t}$

La fonction `rexp` permet de simuler des données suivant une loi exponentielle de paramètre (rate) donné. Pour censurer des observations, le principe consiste à déterminer aussi une valeur aléatoire positive correspondant à la réalisation d'une durée avant censure dont la loi est indépendante de celle de la durée. Dans l'exemple, ci-dessous, nous utilisons une loi uniforme sur l'intervalle $[0;5]$ (dont les réalisations sont générées par la fonction `runif`). Les données de durées effectivement observées correspondent alors au minimum des deux valeurs simulées, et une indicatrice `status` est créée pour distinguer les données censurées.

Code R : Simulation d'une distribution exponentielle avec données censurée

```
## Durée de vie suivant une loi exponentielle de taux 0.9
event.times <- rexp(n=1000, rate=0.9)

## Génération de la loi de censure pour les 100 observations
cens.times <- runif(100,0,5)

## Construction du temps observé (minimum entre event.times et cens.times)
obs.times <- pmin(event.times,cens.times)
## Le statut est un vecteur logique qui prend la valeur TRUE
## si l'évènement est observé
status <- as.numeric(event.times<=cens.times)
```

5. On retrouve là l'approche épidémiologique qui questionne autant l'incidence d'une maladie que sa prévalence pour étudier la fréquence et la vitesse d'apparition d'une pathologie.

2.5.1 Estimateur de Kaplan-Meier de la fonction de survie

Sous **R**, l'analyse de modèles de survie peut être réalisée avec le package **survival**. La fonction `survfit` est dédiée à l'estimation non paramétrique de la fonction de survie.

```
## Principales options de la fonction survfit
survfit(formula,
        data, weights, subset,
        type, error, conf.type, ...)
```

Sous **R**, l'argument `formula` désigne un objet de « survie » déclaré avec la fonction `Surv` propre au package **survival**, comme suit `Surv(time, status)` où `time` est la variable de la base de données contenant les durées observées ou censurées et `status`, une variable valant 0 si l'observation est censurée et 1 si l'évènement d'intérêt est effectivement observé⁶. Par défaut, l'estimateur de la variance $var(\hat{S}(t))$ calculé correspond à l'estimateur de Greenwood, mais il est aussi possible d'estimer la variance avec l'estimateur de Tsiatis avec l'argument `error = "tsiatis"`⁷. Enfin, l'estimateur de Fleming-Harrington de la fonction de survie (qui découle de l'estimateur de Nelson-Aalen de la fonction de hasard cumulé) peut être obtenue avec l'argument `type="fleming-harrington"`.

Code **R**: Estimateur de Kaplan-Meier de la survie

```
library(survival)
## le paramètre formula permet de déclarer l'analyse de l'objet de survie
## ici, sans explicative
fit.surv <- survfit(formula = Surv(obs.times, status) ~ 1)
```

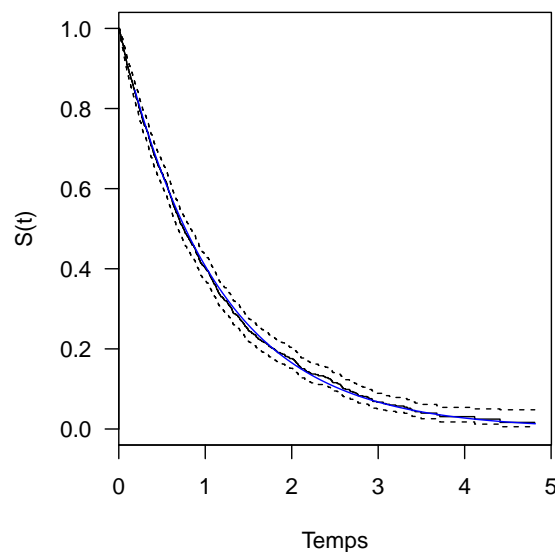
```
summary(fit.surv, time = seq(0, 0.15, 0.01))
```

```
Call: survfit(formula = Surv(obs.times, status) ~ 1)
```

time	n.risk	n.event	survival	std.err	lower	95% CI	upper	95% CI
0.00	1000	0	1.000	0.00000		1.000		1.000
0.01	994	6	0.994	0.00244		0.989		0.999
0.02	985	9	0.985	0.00384		0.977		0.993
0.03	967	8	0.977	0.00474		0.968		0.986
0.04	956	11	0.966	0.00575		0.955		0.977
0.05	939	7	0.959	0.00630		0.947		0.971
0.06	927	12	0.947	0.00715		0.933		0.961
0.07	916	11	0.935	0.00782		0.920		0.951
0.08	887	10	0.925	0.00840		0.909		0.942
0.09	880	7	0.918	0.00878		0.901		0.935
0.10	869	11	0.906	0.00932		0.888		0.925
0.11	864	5	0.901	0.00956		0.882		0.920
0.12	854	10	0.890	0.01000		0.871		0.910

6. Dans le cas simple des modèles de survie avec données censurées à droite étudié dans ce document.

7. On notera dans ce cas que l'estimateur de la fonction de survie reste l'estimateur de Kaplan-Meier, et que seule l'estimateur de l'écart-type diffère



Graphique 2.1 – Estimateur de Kaplan-Meier de la fonction de survie

0.13	849	5	0.885	0.01021	0.865	0.906
0.14	839	10	0.875	0.01061	0.854	0.896
0.15	834	5	0.870	0.01080	0.849	0.891


L'appel de l'objet `fit.surv` créé par la fonction `survfit` avec la fonction `plot` permet de représenter la fonction de survie estimée par Kaplan-Meier. Dans le code ci-dessous, l'argument `mark.time = FALSE` supprime de la représentation graphique la représentation sur les courbes des temps correspondant à des censures. Enfin, nous comparons la fonction de survie estimée à la distribution théorique avec la fonction `curve` (cf. Graphique 2.1).

Code : Représentation graphique de la fonction de survie, réelle et estimée

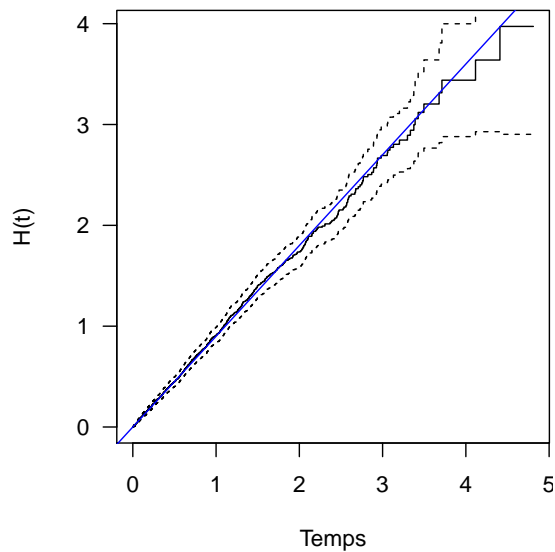
```
plot(fit.surv, xlab = "Temps", ylab = "S(t)", mark.time= FALSE)
curve(exp(-0.9*x), add = TRUE, col="blue")
```

Dans notre exemple, la fonction de survie estimée décroît continûment témoignant comme attendu qu'à chaque instant la fonction de hasard est positive. Mais il apparaît difficile sur ce seul graphique de pouvoir distinguer les dates où le hasard est élevé de celles où le hasard est faible.

2.5.2 Estimateur de Nelson-Aalen de la fonction de hasard cumulé

L'obtention sous  de l'estimateur de Nelson-Aalen de $H(t)$ avec la fonction `survfit` nécessite de calculer tout d'abord l'estimateur de Fleming-Harrington de la fonction de survie, avant d'en prendre l'opposé du logarithme. Pour en obtenir une représentation graphique, il suffit cependant de renseigner simplement l'argument `fun = "cumhaz"` lors de l'exécution de la fonction `plot` sur l'objet créé par la fonction `survfit` comme dans l'exemple ci-dessous.

Code : Estimateur de Nelson-Aalen du hasard cumulé



Graphique 2.2 – Estimateur de Nelson-Aalen du hasard cumulé

```
fit.survFH <- survfit(Surv(obs.times,status) ~ 1, type = "fleming-harrington")
plot(fit.survFH,fun="cumhaz")
abline(0,0.9, col="blue")
```

Comme précédemment, le Graphique 2.2 représente la fonction de hasard cumulé théorique (avec l'instruction `abline`) et l'estimateur de Nelson-Aalen correspondant. Comme attendu, la fonction de hasard cumulé estimée présente un profil linéaire au cours du temps dont la pente est égale au paramètre de la loi exponentielle.

2.5.3 Estimateur par noyau de convolution de la fonction de hasard

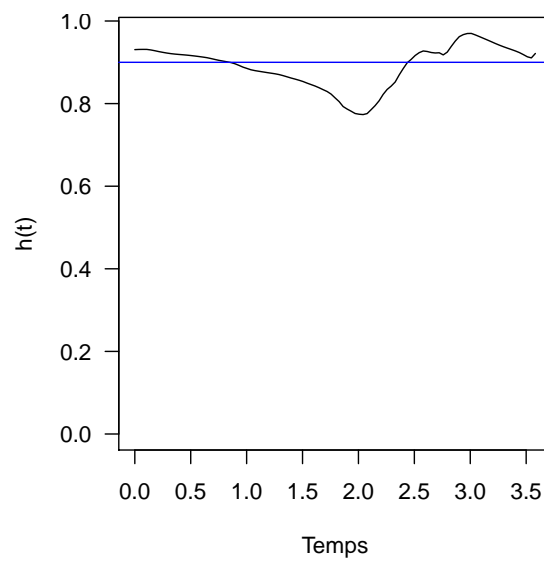
Pour obtenir une estimation de la fonction de hasard par noyau de convolution avec **R**, on peut utiliser le package **muhaz**. Dans ce cas, le calcul de l'estimateur se fait par la fonction `muhaz`.

```
## Syntaxe complète de la fonction muhaz
muhaz(times, delta, subset, min.time, max.time, bw.grid, bw.pilot,
       bw.smooth, bw.method="local", b.cor="both", n.min.grid=51,
       n.est.grid=101, kern="epanechnikov")
```

Comme pour les fonctions précédentes, il faut préciser le vecteur contenant les durées observées (`times`) et l'indicatrice permettant de distinguer les observations non censurées (`status`). Le choix de la fonction de noyau (argument `kern`) est souvent de peu d'incidence en pratique.

Code **R** : Estimateur par noyau de convolution de la fonction de hasard

```
library(muhaz)
fit <- muhaz(obs.times,status)
plot(fit, xlab = "Temps",ylab = "h(t)")
abline(h=0.9, col="blue")
```



Graphique 2.3 – Estimateur par noyau de convolution de la fonction de hasard

La représentation graphique obtenue (cf. Graphique 2.3) révèle que la fonction de hasard estimée est presque constante et proche de 0.9, sauf au niveau des bornes de la période d'étude.

3 Modéliser la durée de vie

Dans la majorité des applications pratiques, on souhaite étudier le lien entre la durée de vie et des variables explicatives. Dans le cas des modèles de survie, cela revient à étudier la distribution des durées de vie conditionnellement aux covariables. Le modèle le plus couramment utilisé est le modèle de Cox (aussi appelé modèle à hasards proportionnels)¹. La première partie présente ce modèle dans le cas de covariables fixes dans le temps. Ce faisant, elle précise notamment comment interpréter les paramètres. La deuxième partie illustre sa mise en œuvre sous R. Enfin, la troisième partie présente les modèles à hasards proportionnels stratifiés mais aussi comment intégrer des covariables dépendant du temps, deux extensions importantes du modèle de Cox classique.

3.1 Le modèle de Cox

Le modèle de Cox est un modèle de régression qui permet de modéliser l'effet de covariables sur la distribution de la durée de vie. En effet, contrairement au modèle de régression linéaire classique, les covariables ne sont pas directement reliées à la durée de vie, mais à sa fonction de hasard. Le modèle de Cox postule que la fonction de hasard pour l'individu i peut s'écrire :

$$h_i(t|x) = h_0(t)e^{x_i'\beta}, \forall t \geq 0 \quad (3.1)$$

où $h_0(t)$ est appelée fonction de risque de base et X un ensemble de covariables supposées, dans cette partie, fixes dans le temps. C'est un modèle semi-paramétrique puisqu'aucune supposition n'est faite sur h_0 . La fonction de risque de base ne dépend pas de l'individu considéré et traduit donc l'hypothèse que la *dépendance au temps* du risque de connaître l'évènement est *identique pour tous les individus*.

3.1.1 Hasards proportionnels et interprétation des paramètres

L'interprétation des résultats d'un tel modèle diffère de celle des modèles linéaires plus classiques. Elle repose sur la notion de *risque relatif*. En effet, si l'on considère deux individus, alors

$$\frac{h_i(t|x_i)}{h_j(t|x_j)} = \frac{h_0(t) \exp(x_i'\beta)}{h_0(t) \exp(x_j'\beta)} = \exp((x_i - x_j)'\beta) \quad (3.2)$$

Ainsi, ce ratio mesure le risque relatif à *tout moment* t de connaître l'évènement pour l'individu i par rapport au même risque pour l'individu j . Ce ratio est constant dans le temps, et c'est pourquoi le modèle de Cox est aussi appelé *modèle à hasards proportionnels*.

1. D'autres modèles existent, tels le modèle à durée de vie accélérée, les modèles paramétriques (pour lesquels on suppose connue la loi suivie par T) ou les modèles à hasards constants par morceaux. Ils ne sont pas présentés dans ce document.

Dans le cas d'une variable dichotomique, par exemple, si $x = 1$ correspond à un individu qui reçoit un traitement et 0 sinon, alors e^β représente le risque relatif constant dans le temps de connaître l'évènement pour un individu recevant le traitement par rapport à un individu qui ne le reçoit pas. Dans le cas d'une variable continue, e^β représente le risque relatif constant dans le temps lorsque la variable augmente d'une unité, *quelle que soit la valeur x de référence*².

Quelle que soit la covariable, il est important de retenir que l'effet mesuré par chaque paramètre est un effet *relatif* sur la fonction de hasard instantané. Il ne dit donc rien sur la valeur absolue du risque de décès $h(t)$, à un instant donné, et ce, quelle que soit l'ampleur du paramètre β . De plus, dans le cas des variables continues, il serait erroné de conclure à l'importance de l'impact sur la seule valeur estimée de β , sans analyser aussi la distribution de la covariable dans l'échantillon analysé. Ce point sera détaillé dans l'exemple analysé dans ce chapitre.

3.1.2 Estimation et prise en compte des temps non distincts

L'estimation du modèle (3.1) repose sur la fonction de *vraisemblance partielle* introduite par Cox (1972, 1975). Son expression s'appuie sur deux hypothèses déjà évoquées au premier chapitre qui permettent de simplifier la vraisemblance initiale du modèle. Tout d'abord, la censure et la durée de vie sont deux variables *indépendantes*³. De plus, la censure est *non informative* : elle ne renseigne en rien sur la loi de T et sa densité ne dépend pas du paramètre β . Formellement, dans le cas de la censure à droite, elle s'exprime en fonction des seules densité, f , et fonction de répartition, F, de la loi de T de paramètre θ :

$$L_n = \prod_{i=1}^n [f_\theta(T_i, x_i)]^{\delta_i} [\bar{F}_\theta(T_i, x_i)]^{1-\delta_i}$$

où δ_i est une indicatrice qui vaut 1 si la durée de vie de l'individu i n'est pas censurée et $\bar{F} = 1 - F$.

Les estimateurs obtenus par maximisation de la vraisemblance partielle⁴ sont *convergen*t, *efficaces* et suivent asymptotiquement une distribution normale. Les tests d'hypothèses sur les paramètres β ainsi estimés sont les tests de Wald, de score et de ratio de vraisemblance. Par défaut, le logiciel R s'appuie sur la z -statistique (test de Wald) pour tester la significativité des coefficients estimés.

De cette approche, il est important de retenir deux éléments d'un point de vue pratique. Tout d'abord, *seul l'ordre, pas la date précise, des évènements importe* pour calculer la vraisemblance partielle et donc permettre l'estimation. Il en résulte qu'une modification des dates d'évènements qui ne perturberait pas leur ordre ne change pas la valeur des estimations.

Cependant, *la présence d'évènements concomitants (i.e. à des dates « enregistrées » identiques) nécessite de modifier l'estimateur de vraisemblance partielle*. En effet, la vraisemblance partielle utilisée pour estimer le modèle de Cox repose sur l'hypothèse de temps distincts, c'est-à-dire que la durée de vie est effectivement continue. En pratique, il est souvent inévitable que l'enregistrement des temps d'évènements ne soit pas assez précis et conduise à observer, pour des individus différents, des dates identiques d'évènements et/ou de censures. Ainsi, lorsqu'une date de censure est identique à une date d'évènement, on fait l'hypothèse que l'évènement *précède* la censure. Lorsque des individus connaissent l'évènement à la même date observée, trois méthodes sont souvent implémentées pour calculer la vraisemblance partielle, qui sont autant de possibilités offertes en pratique par les logiciels

2. Dans les deux cas, ceci s'entend toutes choses égales par ailleurs.

3. ce qui permet d'écrire la densité jointe qui intervient dans l'expression de la vraisemblance comme produit des densités

4. avec l'algorithme de Newton-Raphson sous R.

statistiques. La méthode *exacte* (Peto, 1972; Kalbfleisch et Prentice, 1980) calcule la contribution à la vraisemblance partielle de chaque individu, ayant une date d'évènement commune à un ou plusieurs autres individus, pour tous les rangs possibles (son évènement est le premier à survenir, le second, etc. parmi les évènements survenant à la même date). La méthode est de fait assez chronophage. Pour remédier à ce problème, deux approximations ont été proposées : l'approximation de *Breslow-Peto* et d'*Efron*. En pratique, la méthode exacte, bien que chronophage, est dans la mesure du possible à privilégier. Sinon, plusieurs simulations ont montré qu'il fallait privilégier la méthode d'*Efron* sur la méthode de *Breslow-Peto*. On notera dans tous les cas, que si le nombre d'évènements à dates identiques est supérieur aux nombres d'évènements à dates uniques, il est préférable de considérer les données comme groupées par intervalles.

3.2 Mise en œuvre sous R

Pour illustrer la mise en œuvre du modèle de Cox⁵, nous utiliserons la base de données `pbcc` disponible dans le package `survival`. Il s'agit d'une base constituée au cours d'une expérience aléatoire menée entre 1974 et 1984, pour étudier l'efficacité de la D-pénicillamine sur la cholangite (ou cirrhose) biliaire primitive (*Primary Biliary Cirrhosis*) du foie. Cette maladie dégénérative, probablement auto-immune, engendre une inflammation des voies biliaires pouvant conduire au développement d'une cirrhose du foie, voire à la mort du patient.

Code R : Chargement de la base de données

```
data(pbc, package="survival")
```

La base comporte 418 observations de patients, dont 312 ont participé à une expérience aléatoire afin de tester la validité du médicament (`drug` : 1 = D-pénicillamine ; 2 = placebo). Les 106 autres patients ont accepté, en parallèle de ce protocole, de fournir plusieurs informations⁶ et de se faire suivre régulièrement. Qu'ils participent ou non à l'expérience, le suivi des patients prend fin en juillet 1986. Au cours de la période d'étude, certains patients ne développeront pas de complications (`status=0`), se feront greffer un nouveau foie (`status=1`) ou décéderont (`status=2`). La variable `time` précise ainsi le nombre de *jours*⁷ écoulés entre le début du suivi du patient et juillet 1986, la date de la transplantation ou du décès. Dans ce document, nous nous intéressons à la durée de vie *avant le décès*; les personnes greffées seront considérées comme censurées. Sous R, il faut pour cela déclarer un objet de « survie » avec la fonction `Surv` comme suit `Surv(time, status==2)`⁸.

La maladie, peu fréquente, touche principalement des femmes (elles constituent 89% des patients de l'échantillon) et débute en moyenne vers 50 ans. Elle peut ne pas causer de symptômes dans sa phase initiale de telle sorte que l'âge⁹ est souvent endogène à la durée de la maladie. Néanmoins, plusieurs examens peuvent être pratiqués pour diagnostiquer la maladie, notamment des analyses sanguines. Parmi les anomalies qui peuvent alors être constatées, on note, par exemple, une présence d'autant plus élevée de bilirubine (`bil.i` en mg/dl) dans le sang que la maladie est en phase avancée, et

5. mais aussi pour discuter de sa validité et tenir compte de l'hétérogénéité inobservée dans les chapitres suivants.

6. La base comporte cependant des données manquantes pour ces personnes sur plusieurs variables médicales

7. Ce point sera important dans l'interprétation des résultats

8. Si, au contraire, on s'intéresse à la durée écoulée avant le décès *ou* la greffe du patient, il faut déclarer `Surv(time, status > 0)`

9. enregistré dans cette étude au début du suivi

des altérations du taux d'albumine (albumin en mg/dl) et de taux de prothrombine (prottime). Enfin, la présence d'œdèmes plus ou moins sévères est aussi le signe d'une maladie déjà avancée (edema : 0 = absence, 0.5 = modéré, 1 = présent malgré l'utilisation d'un traitement diurétique). Toutes ces informations sont présentes dans la base de données et enregistrées au début du suivi, de telle sorte que les covariables du modèle sont *constantes dans le temps*.

L'estimation du modèle de Cox peut être effectuée sous **R** par la fonction `coxph` du package **survival**, dont la syntaxe complète est donnée ci-dessous.

```
## Principaux arguments de la fonction coxph
coxph(formula, data=, weights, subset,
      na.action, ties=c("efron", "breslow", "exact"),...)
```

L'argument `formula` permet de préciser les covariables présentes dans la base de données utilisées dans le modèle. Classiquement, l'objet de survie étudié (cf. section 2.5) et les covariables du modèle sont séparés par un `~`. La méthode choisie pour tenir compte des dates d'évènements ou de censure concomitantes (cf. section 3.1.2) se fait avec l'argument `ties` qui privilégie la méthode d'Efron par défaut.

Code **R** : Estimation d'un modèle de Cox avec covariables constantes

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

Call:

```
coxph(formula = Surv(time, status == 2) ~ age + factor(edema) +
      log(bili) + log(prottime) + log(albumin), data = pbc, ties = "efron")
```

```
n= 416, number of events= 160
(2 observations deleted due to missingness)
```

	coef	exp(coef)	se(coef)	z	Pr(> z)	
age	0.040453	1.041283	0.007712	5.246	1.56e-07	***
factor(edema)0.5	0.281884	1.325625	0.225233	1.252	0.210743	
factor(edema)1	1.012472	2.752396	0.289755	3.494	0.000475	***
log(bili)	0.859048	2.360912	0.083244	10.320	< 2e-16	***
log(prottime)	2.359929	10.590195	0.773143	3.052	0.002270	**
log(albumin)	-2.515764	0.080801	0.652616	-3.855	0.000116	***

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

	exp(coef)	exp(-coef)	lower .95	upper .95
age	1.0413	0.96035	1.02566	1.0571
factor(edema)0.5	1.3256	0.75436	0.85252	2.0613
factor(edema)1	2.7524	0.36332	1.55981	4.8568

log(bili)	2.3609	0.42357	2.00550	2.7793
log(prottime)	10.5902	0.09443	2.32704	48.1953
log(albumin)	0.0808	12.37607	0.02249	0.2904

Concordance= 0.835 (se = 0.025)
 Rsquare= 0.427 (max possible= 0.985)
 Likelihood ratio test= 231.9 on 6 df, p=<2e-16
 Wald test = 238.9 on 6 df, p=<2e-16
 Score (logrank) test = 327.5 on 6 df, p=<2e-16

Les coefficients β estimés sont renseignés dans la colonne coef. Leur interprétation est cependant facilitée en lisant leur valeur exponentialisée renseignée dans la colonne exp(coef).

3.2.1 Interprétation des paramètres estimés

Dans notre exemple, le taux de bilirubine a été intégré au modèle sous la forme logarithmique. Chaque point supplémentaire de log(bili) est associé, toutes choses égales par ailleurs, à un risque journalier¹⁰ de décès 2,4 fois ($\exp(\text{coef})=2.37$) supérieur, *quel que soit le nombre de jours passés depuis le diagnostic*. Dans l'échantillon étudié, le taux de bilirubine dans le sang est inférieur à 0,8 mg/dl pour 25% des patients et supérieur à 3,4 mg/dl dans 25% des cas. Dès lors, la probabilité de décès est environ 3,5 fois ($\exp(0.86[\log(3.4) - \log(0.8)])$) supérieure pour les patients du dernier quartile par rapport à celui du premier quartile.

Être âgé au moment du diagnostic est aussi associé à un risque de décès plus important. Ainsi, chaque année supplémentaire est associée à une augmentation de 4% ($\exp(\text{coef})=1.04$) du risque journalier de décès. Dès lors, dix ans d'intervalle entre deux patients accroissent le risque de près de 50% ($\exp(0.0396 * 10) = 1.486$). Comme nous l'avons cependant déjà précisé, l'âge est une variable très endogène au développement de la maladie qui ne peut souvent être détectée qu'à un âge avancé.

La présence d'œdèmes est mesurée par une variable qualitative. Une personne présentant des œdèmes sévères a un risque journalier de décès 2,75 fois plus élevé que celui d'une personne sans œdème (toutes choses égales par ailleurs). Un patient avec des œdèmes moins problématiques présente un risque similaire à celui d'une personne sans œdème. Le ratio de risque associé à un accroissement du taux de prothrombine est le plus élevé; une augmentation d'un point du (log) du taux de prothrombine est associé, toutes choses égales par ailleurs, à un risque de décès 10,6 fois plus élevé. Cependant, le taux de prothrombine varie peu dans la population étudiée, de telle sorte qu'entre les personnes du premier et du dernier quartile, le ratio de risques est de 28%.

Enfin, seul le taux d'albumine est associé à un risque moindre. En effet, l'albumine étant produit en partie par le foie, un taux faible en révèle le dysfonctionnement. Ainsi, chaque point perdu de log(albumin) est associé un risque journalier de décès 12 fois supérieur ($\exp(-\text{coef})=12.27$). Néanmoins, comme pour le taux de prothrombine, la dispersion du taux d'albumine dans l'échantillon étudié est faible et son importance en comparaison des autres covariables est à relativiser.

10. puisque le temps est mesuré en nombre de jours.

3.2.2 Prédire la fonction de survie individuelle et illustrer les résultats

Les effets mesurés par chaque paramètre peuvent aussi être reliés à la fonction de survie. Si β est positif, le risque de connaître l'évènement, par exemple pour les individus présentant des œdèmes sévères, augmente : leur fonction de survie $S(t)$ est donc plus faible. Si β est négatif, la fonction de survie est plus élevée. Dans tous les cas, elle est cependant décroissante (cf. section 2.1). Dès lors, il est souvent utile de prédire les fonctions de survie pour des individus « types » afin d'illustrer l'impact des différentes covariables.

Alors même que la fonction de hasard de base $h_0(t)$ n'est pas spécifiée dans le modèle, il est en effet possible à partir des paramètres estimés de représenter la fonction de survie $S(t|x_i)$ individuelle. En effet, le modèle 3.1 assure que :

$$S(t|x_i) = e^{-\exp(x'_i\beta) \int_0^t h_0(u) du} = S_0(t)^{\exp(x'_i\beta)} \quad (3.3)$$

Il existe deux estimateurs de $S_0(t)$: celui de Breslow et de Kalbfleisch-Prentice. Ils peuvent être utilisés indifféremment car en pratique, ils sont proches. Une fois calculé un des estimateurs, une estimation de la fonction de survie pour un individu i donné peut être obtenue par $\hat{S}(t|x_i) = \hat{S}_0(t)^{\exp(x'_i\hat{\beta})}$.

Sous **R**, l'estimation de la fonction de survie pour des individus aux caractéristiques prédéfinies par le chargé d'études est obtenue avec la fonction `survfit` appliquée à l'objet créé par `coxph`. Par défaut, celle-ci estime la fonction de survie pour un individu moyen¹¹. Pour illustrer son utilisation, nous allons représenter les fonctions de survie de trois personnes d'âge moyen, présentant toutes un taux de bilirubin, d'albumine et un temps de prothrombine correspondant à la moyenne des valeurs observées dans l'échantillon donné. Elles diffèrent uniquement par la présence (plus ou moins sévère) ou non d'œdèmes (`edema` : 0 = absence, 0.5 = modéré, 1 = présent malgré l'utilisation d'un traitement diurétique). Un nouveau jeu de données (`mean.pbc` dans le code ci-dessous) comportant trois observations est donc créée préalablement.

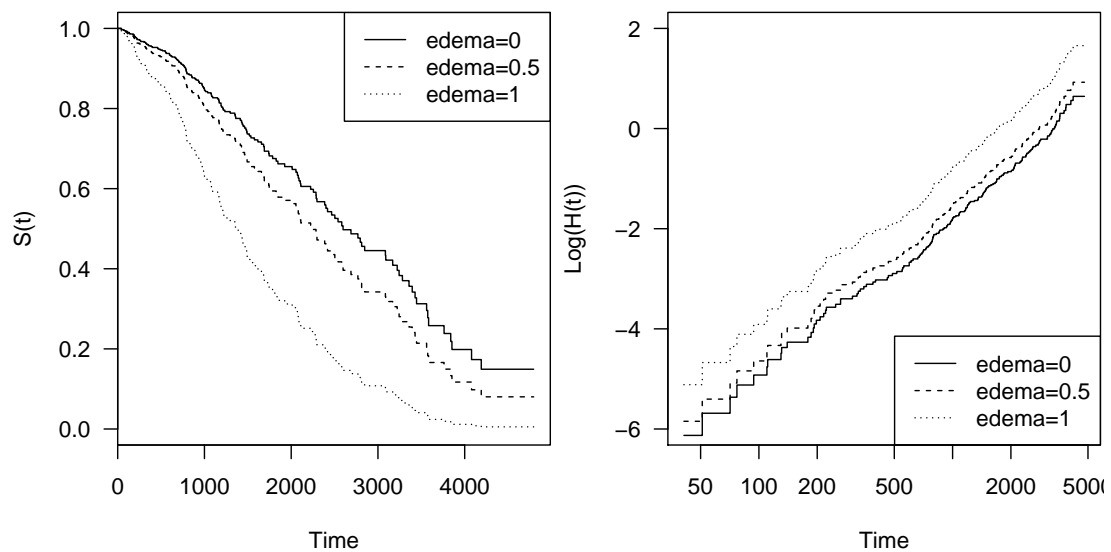
Code **R** : Fonction de survie prédite par le modèle

```
mean.pbc <- data.frame(age = rep(mean(pbc$age), times=3),
                      edema = c(0, 0.5, 1),
                      bili = rep(mean(pbc$bili), times=3),
                      protime = rep(mean(pbc$protime, na.rm=TRUE), times=3),
                      albumin = rep(mean(pbc$albumin), times=3))
surv.pbc <- survfit(fit.pbc, newdata = mean.pbc)
```

Les fonctions de survie pour les 3 individus types sont toutes décroissantes (cf. Graphique 3.1). Cependant, on notera que la présence d'œdèmes malgré l'utilisation d'un traitement réduit sensiblement la probabilité de survie quel que soit le temps écoulé depuis la découverte de la maladie (`edema` : 1). Il est intéressant aussi de mettre en évidence l'hypothèse de proportionnalité qui sous-tend la modélisation retenue. En effet, l'équation 3.3 peut aussi s'écrire en fonction du hasard cumulé de base $H_0(t)$:

$$\log\{-\log[S(t|x_i)]\} = \log(H_0(t)) + x'_i\beta = \log(H(t|x_i))$$

11. La pertinence d'un tel calcul est évidemment sujette à caution en présence de covariables qualitatives. Les auteurs du package **survival** n'en recommandent d'ailleurs pas le calcul.



Graphique 3.1 – Fonctions de survie et de hasard cumulé

Il s'ensuit qu'une transformation « log-log complémentaire » des 3 fonctions de survie assure la représentation de droites parallèles dont l'espacement correspond à l'ampleur du paramètre estimé correspondant (cf. Graphique 3.1).

Code : Représentation graphique des fonctions de survie et de hasard cumulé

```
## Survie prédite
plot(surv.pbc, lty = c(1:3), xlab = "Time", ylab = "S(t)")
legend("topright", legend=c("edema=0", "edema=0.5", "edema=1"), lty=c(1:3))

## (log) Hasard cumulé prédit
plot(surv.pbc, fun="cloglog",
     lty = c(1:3), xlab = "Time", ylab = "Log(H(t))", ylim=c(-6,2))
legend("bottomright", legend=c("edema=0", "edema=0.5", "edema=1"), lty=c(1:3))
```

```
## Survie prédite
plot(surv.pbc, lty = c(1:3), xlab = "Time", ylab = "S(t)")
legend("topright", legend=c("edema=0", "edema=0.5", "edema=1"), lty=c(1:3))

## (log) Hasard cumulé prédit
plot(surv.pbc, fun="cloglog",
     lty = c(1:3), xlab = "Time", ylab = "Log(H(t))", ylim=c(-6,2))
legend("bottomright", legend=c("edema=0", "edema=0.5", "edema=1"), lty=c(1:3))
```

3.3 Modèles stratifiés et covariables dépendant du temps

3.3.1 Modèles stratifiés

Le modèle 3.1 à hasards proportionnels suppose que le risque de décès des patients présentant des œdèmes est proportionnel à celui des patients sans œdèmes. Dans les trois groupes, on fait donc l'hypothèse que les fonctions de hasard présentent toutes la même forme fonctionnelle de dépendance au temps. Il est cependant possible de relâcher une telle hypothèse de sorte que la dépendance au temps de la fonction de hasard de chaque groupe diffère, mais dans ce cas l'effet de la présence d'œdèmes sur la probabilité de décès ne peut plus être estimé. Ce type de modèle est appelé *modèle à hasards proportionnels stratifié*. Formellement, si l'on considère G groupes distincts, alors le modèle stratifié classique correspondant pose que :

$$h_g(t|x) = h_{0,g}(t) \exp(\beta'x), g = 1, \dots, G \quad (3.4)$$

On notera que dans cette modélisation, les variables exogènes X ont le même effet dans tous les groupes.

Sous **R**, l'estimation de ce modèle s'effectue en déclarant simplement `strata(var)` comme une variable exogène dans l'option `formula` ¹².

Code **R** : Estimation d'un modèle de Cox stratifié

```
strata.pbc <- coxph(formula = Surv(time,status==2) ~ age + log(bili) +
  log(prottime) + log(albumin) + strata(edema),
  data = pbc)
```

Call:

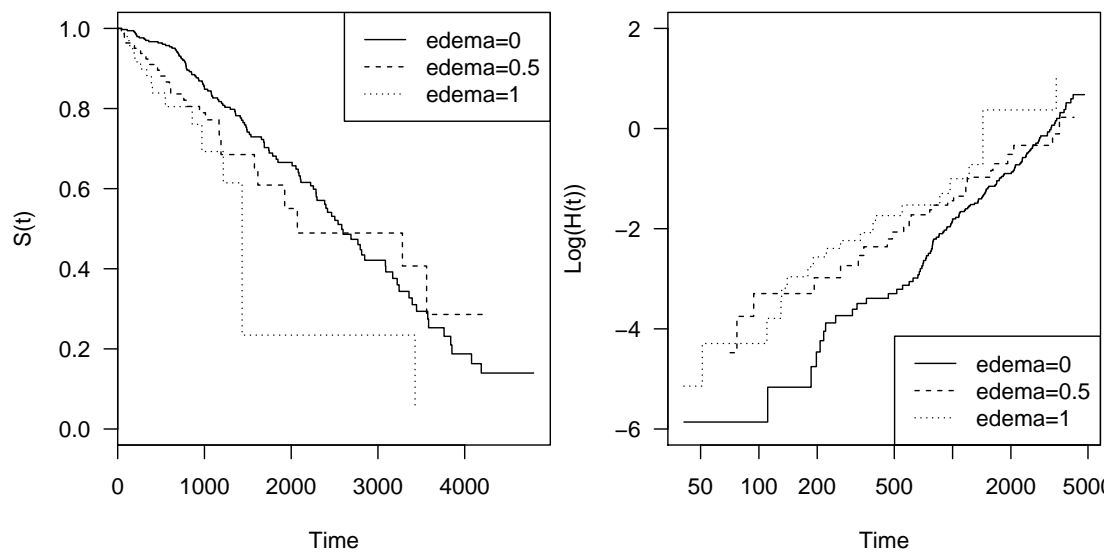
```
coxph(formula = Surv(time, status == 2) ~ age + log(bili) + log(prottime) +
  log(albumin) + strata(edema), data = pbc)
```

	coef	exp(coef)	se(coef)	z	p
age	0.04093	1.04178	0.00787	5.20	2e-07
log(bili)	0.84855	2.33625	0.08358	10.15	< 2e-16
log(prottime)	2.43819	11.45231	0.76674	3.18	0.00147
log(albumin)	-2.45630	0.08575	0.65591	-3.74	0.00018

```
Likelihood ratio test=160.8 on 4 df, p=<2e-16
n= 416, number of events= 160
(2 observations deleted due to missingness)
```

Les fonctions de survie de base et de hasard cumulé correspondant à chaque groupe peuvent être représentées graphiquement, en procédant comme dans la section 3.2.2. Dans notre exemple, les fonctions de survie dans les 3 groupes se croisent (cf. graphique 3.2) et leurs fonctions de hasard cumulé ne sont plus parallèles puisque cette hypothèse a été relâchée, contrairement aux représentations obtenues précédemment avec le modèle non stratifié.

12. Il est possible de stratifier selon plusieurs variables, comme le sexe et la présence d'œdèmes. Le code à renseigner est alors `strata(sex, edema)`. Dans ce cas, les différents groupes sont constitués par toutes les combinaisons possibles des différentes variables.



Graphique 3.2 – Fonctions de survie et de hasard cumulé pour le modèle stratifié

Usuellement, les modèles stratifiés ne distinguent pas l'effet des covariables selon le groupe. Il n'est cependant pas toujours possible de considérer une telle hypothèse comme pertinente. Il est dès lors nécessaire d'intégrer dans le modèle l'interaction d'une covariable avec l'appartenance à chaque groupe. Si toutes les covariables sont interagies avec la variable de stratification, les résultats obtenus correspondent à des estimations séparées sur chaque sous-population.

Sous **R**, il suffit de déclarer `var1*strata(var2)`. Dans l'exemple ci-dessous, l'effet de l'âge est supposé différer selon la présence d'œdèmes plus ou moins sévères.

Code **R : Estimation d'un modèle de Cox stratifié avec effets de l'âge différenciés par strates d'œdèmes**

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

Call:

```
coxph(formula = Surv(time, status == 2) ~ log(bili) + log(protime) +
      log(albumin) + age * strata(edema), data = pbc)
```

	coef	exp(coef)	se(coef)	z	p
log(bili)	0.88850	2.43148	0.08744	10.16	< 2e-16
log(protime)	2.36640	10.65900	0.76329	3.10	0.00193
log(albumin)	-2.38749	0.09186	0.67744	-3.52	0.00042
age	0.03414	1.03473	0.00873	3.91	9.2e-05
age:strata(edema)edema=0.5	0.02752	1.02790	0.02366	1.16	0.24477
age:strata(edema)edema=1	0.05746	1.05914	0.03411	1.68	0.09212

```
Likelihood ratio test=164.7 on 6 df, p=<2e-16
n= 416, number of events= 160
(2 observations deleted due to missingness)
```

Pour les personnes sans œdèmes, chaque année supplémentaire au moment du diagnostic est associée à un risque de décès (journalier) plus élevé de 3,5%. Un patient présentant des œdèmes peu problématiques présente un risque relatif pour une année supplémentaire similaire (le coefficient associé, `age : strata (edema) edema=0.5`, n'est pas significatif). Enfin, en présence d'œdèmes sévères, le risque de décès associé à une année supplémentaire au moment du diagnostic est significativement plus fort que dans les deux autres groupes : il serait de 9,6% ($\exp(0.03414 + 0.05746)$)¹³.

3.3.2 Covariables dépendant du temps

Les 312 patients ayant participé à l'expérience aléatoire ont été suivis régulièrement pendant toute la durée de l'étude. Leurs taux de bilirubine, d'albumine et leur temps de coagulation ont ainsi été enregistrés à différentes dates, pour connaître l'évolution de ces caractéristiques au cours du temps. L'objet de cette partie est de présenter comment intégrer ces covariables, dont les valeurs dépendent du temps, au modèle à hasards proportionnels étudié jusqu'à présent¹⁴.

La démarche consiste simplement à modifier la structure de la base de données. En effet, il s'agit, pour chaque personne, de définir autant d'observations qu'il y a de valeurs différentes au cours du temps de ces variables. La base `pbseq` du package **survival** contient ainsi pour les 312 patients autant d'observations qu'ils ont effectué de visites au cours de l'étude. Pour chaque visite, le nombre de jours écoulés depuis le diagnostic est enregistré (`day`) avec la valeur correspondante observée pour les différentes covariables.

	id	status	day	age	edema	bili	albumin	prottime
3	2	0	0	56.44627	0.0	1.1	4.14	10.6
4	2	0	182	56.44627	0.0	0.8	3.60	11.0
5	2	0	365	56.44627	0.0	1.0	3.55	11.6
6	2	0	768	56.44627	0.0	1.9	3.92	10.6
7	2	0	1790	56.44627	0.5	2.6	3.32	11.3
8	2	0	2151	56.44627	1.0	3.6	2.92	11.5
9	2	0	2515	56.44627	1.0	4.2	2.73	11.5
10	2	0	2882	56.44627	1.0	3.6	2.80	11.5
11	2	0	3226	56.44627	1.0	4.6	2.67	11.5

Ainsi, par exemple, le patient 2 a été suivi pendant 3226 jours au terme desquels il était toujours en vie (`status=0`). Entre le jour du diagnostic et la dernière visite, son taux de bilirubine, son taux d'albumine et son temps de coagulation ont été enregistrés à plusieurs reprises, au 182^e jour, au 365^e, au 768^e, etc. On notera que son âge (noté au début de l'étude) n'a pas été mis à jour. Cette covariable présente donc pour chaque observation des valeurs identiques.

13. Ce résultat peut être obtenu directement en déclarant comme covariable `age : strata (edema)` dans l'option `formula`.

14. Dans cette partie, nous considérons des variables dont la dépendance au temps ne peut être prédite. En effet, dans le cas contraire, comme l'âge, celles-ci peuvent être utilisées pour définir une autre métrique du temps. Les paramètres du modèle de Cox voient alors leur interprétation modifiée.

D'un point de vue pratique, chaque observation doit cependant être identifiée par une date de début (`tstart`) et une date de fin (`tstop`), et le statut de la personne à *la fin de l'intervalle de temps* (`death`). Le code ci-dessous illustre comment mettre à jour la base initiale (`pbc`) ne contenant que des covariables fixes dans le temps¹⁵ à partir des données de la base `pbcseq` qui contient les caractéristiques dont les valeurs sont mises à jour à chaque rendez-vous (à savoir le taux de bilirubine, d'albumine et de prothrombine). Pour cela, nous utilisons la fonction `tmerge` du package **survival**. Celle-ci doit être exécutée une première fois pour répliquer les observations fixes dans le temps, créer les date de début et de fin d'intervalle, ainsi que la variable de statut à la fin de chaque intervalle. La seconde exécution permet d'intégrer à la nouvelle base de données ainsi constituée, les covariables évoluant dans le temps et qui sont déclarées comme telles avec la fonction `tdc`.

Code : Ajout de covariables non fixes dans le temps

```
## Sélection des patients participant à l'essai clinique
temp <- subset(x = pbc, subset = id<=312,select = c(id:sex,edema))

## Création des dates de début (tstart) et de fin (tstop)
## des intervalles de temps
## et le statut du patient à la fin de chaque intervalle (death)
pbc2 <- tmerge(data1 = temp, data2 = temp, id = id,
               death = event(time,status))

## Ajout des variables
pbc2 <- tmerge(data1 = pbc2, data2 = pbcseq, id = id,
               ## déclaration des covariables dépendant du temps
               bili = tdc(day,bili),
               albumin = tdc(day,albumin),
               protime = tdc(day,protime))
```

	id	tstart	tstop	death	bili	albumin	protime
1	1	0	192	0	14.5	2.60	12.2
2	1	192	400	2	21.3	2.94	11.2

Dans la base de données ainsi constituée, le patient 1 était encore vivant à sa première visite le 192^e jour (`death=0`). Mais il est décédé le 400^e jour qui a suivi son diagnostic (`death=2`).

Sans entrer dans les détails, ce changement de structure des données pour estimer le modèle à hasards proportionnels est justifié par l'analogie des modèles de durée avec les processus de comptage qui s'appuie sur la théorie des martingales. Schématiquement, cette approche permet d'explicitier l'état du patient en fonction des valeurs *passées* des covariables, mais exclut de le relier aux valeurs futures. Ainsi, le taux de bilirubin de 14,5 mg/dl enregistré au début de l'étude est considéré comme un prédicteur du risque de décès sur l'intervalle (0,192], contrairement à la valeur du taux de bilirubin enregistrée précisément le 192^e jour. Par ailleurs, il n'est pas possible de construire une observation du taux de bilirubin correspondant au 100^e jour (par exemple par interpolation linéaire), puisque la

15. pour illustrer notre propos nous ne conservons que les données relatives au sexe et à la présence plus ou moins sévère d'œdèmes

valeur du 192^e jour n'est *effectivement* pas connue à cette date.

À l'inverse, cette structure des données permet d'estimer des modèles de survie où les covariables évoluant au cours du temps sont « retardées ». Dans l'exemple de code ci-dessous, nous constituons une nouvelle base de données `pb3`, où les variables évoluant dans le temps sont retardées de 14 jours (`delai = 14`).


Code : Ajout de covariables non fixes dans le temps et retardées

```
pb3 <- tmerge(data1 = temp, data2 = temp, id = id, death = event(time,status))

pb3 <- tmerge(data1 = pb3, data2 = pb3seq, id = id,
              ## déclaration des covariables dépendant du temps
              ascites = tdc(day,ascites),
              bili = tdc(day,bili),
              albumin = tdc(day,albumin),
              protime = tdc(day,protime),
              options = list(delay = 14))
```

	id	time	edema	tstart	tstop	death	ascites	bili	albumin	protime
1	1	400	1	0	206	0	1	14.5	2.60	12.2
2	1	400	1	206	400	2	1	21.3	2.94	11.2

Ainsi, le taux de bilirubine enregistré au début de l'étude pour le patient 1 est supposé un bon prédicteur sur l'intervalle de temps (0,206], soit 14 jours de plus que précédemment¹⁶. Pour le dire autrement, on étudie la corrélation entre l'absence ou la présence d'évènements *au 206^e jour* (`death=0`) et le taux de bilirubine *valable jusqu'au 192^e jour*, soit 14 jours avant. L'intérêt de retarder les covariables est de se prémunir des effets de causalité inverse, notamment lorsque les intervalles de temps sont courts.

Une fois les données structurées, la syntaxe sous  pour estimer le modèle à hasards proportionnels est similaire à celle présentée à la section 3.2.1. La seule différence réside dans la déclaration de l'objet de survie avec la fonction `Surv` où il faut renseigner les variables définissant les bornes de chaque intervalle (`tstart, tstop`], ainsi que la variable qui précise le statut à la fin de chaque intervalle (ici `death`).

Code : Estimation d'un modèle de Cox avec covariables non fixes dans le temps

```
## Estimation du modèle à hasards proportionnels
fit.pb2 <- coxph(Surv(tstart,tstop,death== 2) ~ log(bili) + age +
                 factor(edema) + log(protime) + log(albumin),
                 data = pb2)
```

Le tableau ci-dessous compare les résultats obtenus en considérant respectivement les valeurs des covariables le jour du diagnostic (X)¹⁷, en intégrant leurs évolutions dans le temps ($X(t)$) et leurs évolutions dans le temps retardées de 14 jours ($X(t-14)$).

16. La date de décès, et donc de fin du 2^e intervalle, n'est cependant pas modifiée.

17. et donc fixes dans le temps

	X	X(t)	X(t-14)
age	0.040	1.236	1.143
factor(edema)0.5	0.282	0.044	0.044
factor(edema)1	1.012	0.894	0.903
log(bili)	0.859	1.132	1.238
log(protime)	2.360	2.702	1.971
log(albumin)	-2.516	-4.260	-3.764

4 Choix de la forme fonctionnelle

Dans le modèle de Cox, la fonction de hasard est supposée vérifier :

$$\log(h(t|x)) = \log(h_0(t)) + x'\beta \quad (4.1)$$

Pour une variable continue fixe dans le temps, comme l'âge dans le modèle 3.1 estimé précédemment, cela implique, par exemple, que le risque relatif entre une personne âgée de 45 ans et une autre de 50 ans est identique à celui entre une personne de 80 ans et une autre de 85 ans. Le ratio de risques est ainsi supposé constant *quelque soit l'âge de référence*. Dit autrement, la covariable a un effet linéaire sur le logarithme de la fonction de hasard. Or, il n'est pas exclu que l'effet de l'âge soit, par exemple, nul avant 65 ans, ou croissant jusqu'à un certain âge avant de diminuer ensuite. La forme fonctionnelle pour chaque covariable retenue dans le modèle doit donc faire l'objet d'une attention particulière.

Ce chapitre présente ainsi deux méthodes couramment utilisées pour aider le chargé d'études à déterminer la forme fonctionnelle la plus adéquate. La première méthode repose sur l'analyse des *résidus dits de martingale* (section 4.1). En effet, bien que dans son expression le modèle (3.1) ne dispose d'aucun résidu (à l'instar de ceux que l'on retrouve dans le modèle de régression linéaire), l'approche des modèles de durée par la théorie des processus de dénombrement¹ permet d'utiliser les résultats sur les martingales pour étudier la pertinence de la forme fonctionnelle retenue dans le modèle pour une variable exogène continue. La deuxième méthode s'appuie sur les *fonctions splines* (section 4.2). Ces dernières permettent de considérer des formes fonctionnelles définies par morceaux par des polynômes en la covariable considérée et donc de modéliser une relation plus flexible que la relation linéaire. Pour chaque méthode, sa mise en œuvre pratique sous R est illustrée à partir de l'exemple discuté au chapitre précédent.

4.1 Résidus de martingale

Dans le modèle de régression linéaire classique, les résidus correspondent à l'écart entre les valeurs observées de la variable expliquée et celles prédites par le modèle. Avec les modèles de survie, une telle approche ne peut être mise en œuvre à cause des données censurées (la durée écoulée jusqu'à la survenue de l'évènement n'est pas connue). L'obtention de résidus dans le cadre des modèles de survie se base sur l'analogie avec les processus de comptage (voir Barlow et Prentice, 1988; Therneau *et al.*, 1990). Sans rentrer dans les détails très techniques, il s'agit de comparer le nombre d'évènements connus par un individu i (donc 0 ou 1) à celui prédit par le modèle durant la période de suivi.

Formellement, on s'appuie sur le résidu de martingale défini par :

$$M_i(t) \equiv N_i(t) - \int_0^t Y_i(s) e^{\beta'X_i(s)} h_0(s) ds$$

1. dont la formalisation théorique complexe n'est pas présentée ici.

où $N_i(t)$ est un processus de dénombrement, c'est-à-dire qu'il correspond au nombre d'évènements connus par l'individu i , 1 ou 0, et $Y_i(t)$, une indicatrice d'être encore à risque à cette même date.

On s'intéresse alors à la quantité :

$$\hat{M}_i \equiv \hat{M}_i(\infty) = N_i(\infty) - \int_0^{\infty} Y_i(t) e^{\beta' X_i(t)} h_0(t) dt$$

où $N_i(\infty)$ vaut 1 si l'individu i n'est pas censuré, 0 sinon.

Si le résidu est positif, cela signifie que l'individu a connu l'évènement et qu'il l'a connu *avant* ce qui était prédit. Le résidu est négatif, (a) si l'individu est censuré, ou (b) s'il a connu l'évènement, mais *plus tard* que prévu.

Les résidus de martingale sont d'espérance nulle, non corrélés entre deux individus si les observations sont i.i.d. [Therneau et al. \(1990\)](#) montrent que si le vrai modèle vérifie $h(t|x) = h_0(t) \exp(f(x)\beta)$ alors les résidus de martingale M du modèle sans covariable vérifient $E[M|x] \approx cf(x)$, où c ne dépend pas de x , mais du nombre de personnes censurées. Dès lors, l'utilisation la plus simple pour investiguer la forme fonctionnelle consiste à représenter les résidus de martingale du modèle *sans covariables* en fonction des valeurs de chaque covariable. Pour faciliter l'analyse, on ajoute souvent au graphique une estimation non paramétrique lissée de la relation entre le résidu de martingale et la variable explicative. Ainsi, par exemple, si la forme fonctionnelle adéquate pour la covariable X est linéaire, l'estimation lissée doit être une droite².

Sous **R**, l'estimation des résidus de martingale peut être obtenue par la fonction `resid` du package **survival**, une fois le modèle sans covariable estimé. La fonction `lowess` permet d'obtenir une représentation graphique lissée de $E[M|x]$ par régression locale³.

Code **R** : Calcul des résidus de martingale et représentation graphique

```
## Estimation du modèle sans covariable
fit.pbc0 <- coxph(formula = Surv(time,status==2) ~ 1, data = pbc, ties = "efron")
residus <- resid(fit.pbc0)

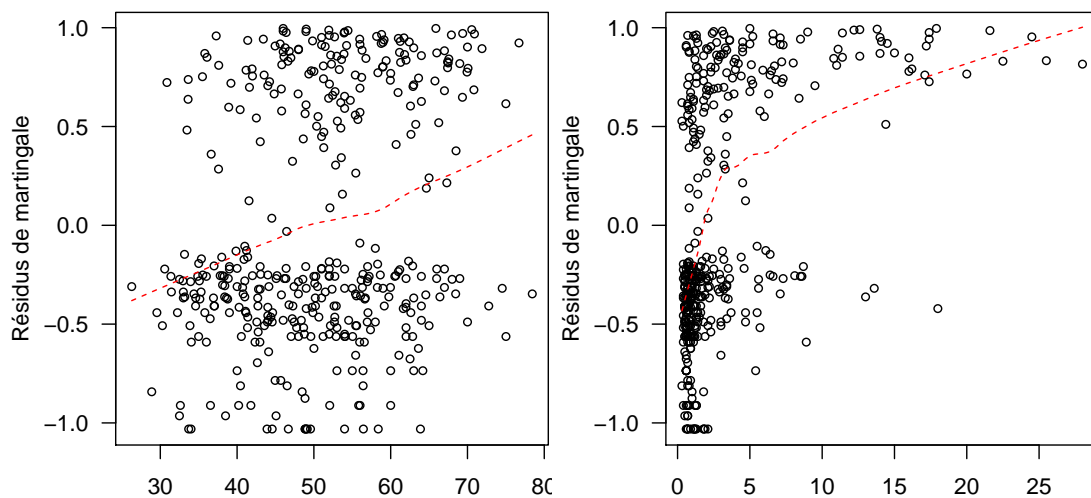
## Représentation graphique pour l'âge
plot(x = pbc$age, y = resid(fit.pbc0), xlab='', ylab="Résidus de martingale", cex=0.75)
lines(lowess(x=pbc$age, y=resid(fit.pbc0), iter=0), lty=2, col="red")
```

Le graphique 4.1 trace les résidus de martingale en fonction de l'âge et de la bilirubin. En ce qui concerne l'âge, la relation est linéaire, alors que pour la bilirubin, une forme fonctionnelle logarithmique apparaît plus adaptée.

Cette approche simple à mettre en œuvre ne tient cependant pas compte des corrélations possibles entre covariables, et peut, dès lors, conduire à retenir des formes fonctionnelles erronées. En pratique,

2. De même, l'absence de tendance de la forme lissée de la représentation graphique des résidus de martingale du modèle *avec covariables*, tracés en fonction des covariables, valide visuellement la forme fonctionnelle retenue.

3. En présence de nombreuses données censurées, la mise en œuvre de la régression `lowess` - qui recherche et rejette toute valeur « aberrante » (*outliers*) - peut conduire à éliminer les données *non censurées*. L'argument `iter=0` de la fonction `lowess` permet d'éviter que les observations non censurées ne soient considérées comme aberrantes lors de la régression lissée.



Graphique 4.1 – Résidus de martingale en fonction de l'âge (gauche) et du taux de bilirubine (droite)

il est alors plus facile de déclarer directement dans le modèle une forme fonctionnelle flexible, par exemple en introduisant un polynôme en la covariable. Une méthode plus couramment utilisée en analyse des survie s'appuie sur les fonctions splines dont la mise en œuvre sous **R** est particulièrement aisée.

4.2 *Smoothing splines*

L'utilisation des fonctions splines permet de modéliser un effet non-linéaire d'une covariable sur le logarithme de la fonction de hasard. Schématiquement, la fonction spline est une fonction définie par morceaux par des polynômes, en ayant au préalable déterminer le nombre de points de jonctions des différents polynômes. Ainsi, une fonction spline de degré 1 correspond à une ligne polygônale constituée de segments reliant les points entre eux. Le cas le plus courant est la fonction spline cubique.

L'utilisation de fonctions splines sous **R** est facilitée par la fonction `pspline` qui détermine à partir des données le nombre de points de jonction (pour plus de détails, voir [Therneau et Grambsch, 2000](#)). Seul, le degré des polynômes doit être renseigné.

Pour illustrer son utilisation, nous allons chercher la forme fonctionnelle la plus adéquate à la variable `age` du modèle de Cox étudié à la section 3.2. L'argument `df` de la fonction `pspline` permet de préciser le degré des polynômes : il correspond au nombre de liberté `df` moins 1.

Code **R** : Estimation du modèle avec fonction spline

```
## Smoothing splines avec un polynôme d'ordre 3 pour l'âge
fit.splines <- coxph(Surv(time,status==2) ~ pspline(age, df=4)
  + log(bili),data = pbc)
```

4. CHOIX DE LA FORME FONCTIONNELLE

```
Call:
coxph(formula = Surv(time, status == 2) ~ pspline(age, df = 4) +
      log(bili), data = pbc)

n= 418, number of events= 161

              coef      se(coef) se2      Chisq DF  p
pspline(age, df = 4), lin 0.04438 0.008072 0.008072  30.23 1.00 3.8e-08
pspline(age, df = 4), non              3.24 3.08 3.7e-01
log(bili)              1.02096 0.077929 0.077762 171.64 1.00 3.2e-39

      exp(coef) exp(-coef) lower .95 upper .95
ps(age)3      1.637    0.61094    0.3284    8.157
ps(age)4      2.673    0.37409    0.1852   38.583
ps(age)5      4.198    0.23819    0.1787   98.619
ps(age)6      5.846    0.17106    0.2359  144.901
ps(age)7      7.744    0.12913    0.3333  179.917
ps(age)8     10.458    0.09562    0.4632  236.115
ps(age)9     12.634    0.07915    0.5560  287.060
ps(age)10    14.681    0.06812    0.6436  334.891
ps(age)11    20.422    0.04897    0.8856  470.915
ps(age)12    21.317    0.04691    0.8929  508.889
ps(age)13    14.696    0.06805    0.5499  392.751
ps(age)14     9.456    0.10575    0.2309  387.268
log(bili)     2.776    0.36025    2.3827    3.234

Iterations: 4 outer, 12 Newton-Raphson
Theta= 0.7931191
Degrees of freedom for terms= 4.1 1.0
Concordance= 0.815 (se = 0.025 )
Likelihood ratio test= 191.4 on 5.08 df, p=<2e-16
```

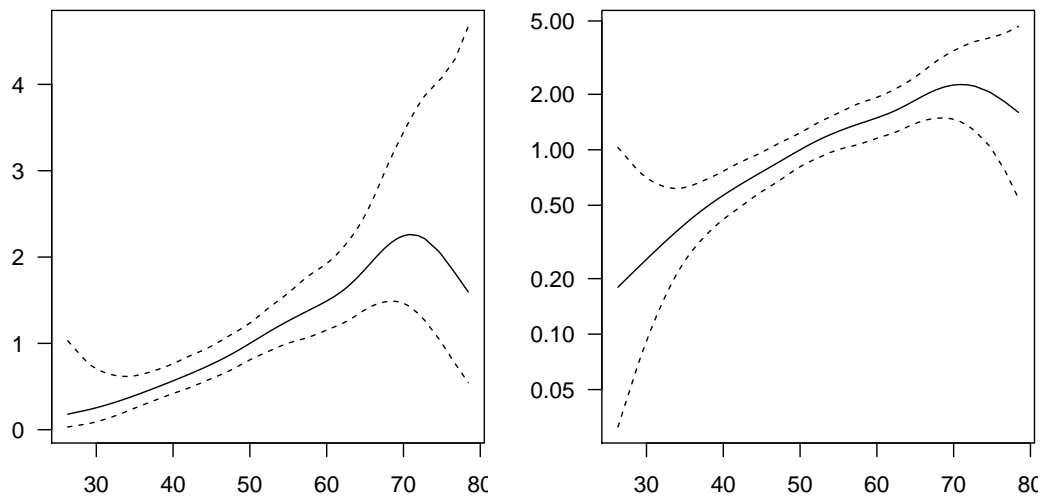
Les résultats de l'estimation comportent deux parties. Dans la première partie, la sortie **R** présente le coefficient relatif à la partie linéaire en l'âge de la forme fonctionnelle retenue (`pspline(age, df = 4)`) qui est ici significativement différent de 0, et la contribution non linéaire d'ensemble des paramètres, qui est aussi ici significativement différente de 0. Dans la seconde partie, les coefficients obtenus sont exponentialisés, mais ceux relatifs à la covariable `age` (`ps(age)3`, `ps(age)4`, etc.) ne sont pas interprétables directement facilement.

Une présentation plus pertinente des résultats consiste à représenter graphiquement l'effet de la covariable sur le risque relatif, en choisissant une valeur de référence. Dans l'exemple détaillé ci-dessous, l'âge de référence est fixé à 50 ans. Sous **R**, la démarche consiste à récupérer les termes prédits avec la fonction `termplot` et à centrer les résultats autour de l'âge de référence retenu ⁴.

Code **R**: Graphique du risque relatif avec l'utilisation de fonctions splines

```
## On récupère les résultats prédits pour l'âge
ptemp <- termplot(fit.splines, se=TRUE, plot=FALSE)
ageterm <- ptemp$age
## Âge de référence
center <- with(ageterm, y[round(x, digits=1)==50])
## Calcul des intervalles de confiance
ytemp <- ageterm$y + outer(ageterm$se, c(0, -1.96, 1.96), '*')
## Représentation graphique
```

4. Dans notre exemple, il n'existe pas dans la base de données d'observation dont l'âge est exactement 50 ans; nous sélectionnons ici la personne présentant l'âge le plus proche.



Graphique 4.2 – Risque relatif lié à l'âge (âge de référence : 50 ans) en log ou non (resp. à gauche et à droite)

```
matplot(agetterm$x, exp(ytemp - center),
        type='l', lty=c(1,2,2), col=1,ann=FALSE)

matplot(agetterm$x, exp(ytemp - center), log='y',
        type='l', lty=c(1,2,2), col=1,ann=FALSE)
```

Le graphique 4.2 de gauche montre que le logarithme du risque relatif est une fonction globalement linéaire de l'âge⁵, conformément à l'équation 4.1. Le graphique 4.2 de droite montre lui que le risque relatif entre une personne de 60 ans et de 50 ans est proche de 1.5. Ce résultat est conforme à l'estimation obtenue à la section 3.2.1; en effet l'effet relatif de l'âge pour 10 années supplémentaires était estimé à $\exp(0.040453 * 10) \approx 1,499$.

5. la partie non-linéaire de la forme fonctionnelle, bien que significative, est peu marquée.

5 Validité de l'hypothèse de proportionalité


L'hypothèse de hasards proportionnels suppose qu'avec des covariables constantes dans le temps, le ratio de risques entre deux personnes est aussi constant dans le temps. Dans le cas multivarié, cette hypothèse doit tenir pour toutes les covariables. Il convient donc d'en vérifier la validité avant toute interprétation des résultats. Dans ce chapitre, nous présentons tout d'abord une approche qui s'appuie sur des représentations graphiques de (transformation de) la fonction de survie qui peuvent être utiles dans le cas de variables explicatives qualitatives (section 5.1). Puis, nous détaillons l'utilisation des résidus de Schœnfeld standardisés couramment utilisés dans le cas des variables continues (section 5.2). Enfin, dans une dernière section, nous suggérons deux méthodes qui peuvent être envisagées si l'hypothèse de proportionnalité du modèle estimé n'est pas vérifiée (section 5.3) : la stratification du modèle et la prise en compte d'effets dont l'ampleur évolue dans le temps.

5.1 Représentations graphiques

Dans le cas de variables qualitatives ou discrètes présentant un nombre restreint de valeurs différentes, un test graphique permet de vérifier simplement la validité de l'hypothèse de hasards proportionnels. En effet, si cette hypothèse est vérifiée, alors :

$$\log\{-\log[S(t|x_i)]\} = \log(H_0(t)) + x_i'\beta$$

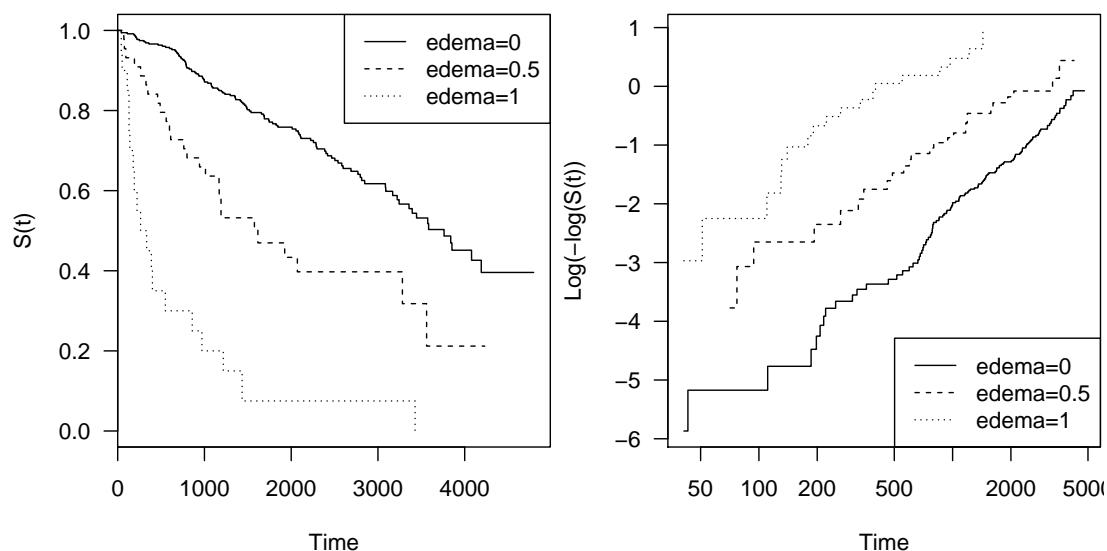
Dès lors, les représentations graphiques des estimations de Kaplan-Meier des fonctions de survie pour chaque valeur x distincte doivent être des courbes approximativement parallèles sur une échelle log-log complémentaire.

L'estimation de chaque fonction de survie peut être obtenue sous  avec la fonction `survfit` en déclarant comme `factor` la caractéristique considérée. Les représentations graphiques sont réalisées avec la fonction `plot` sur l'objet contenant les résultats (`surv.edema` dans l'exemple ci-dessous). L'obtention du graphique sur l'échelle log-log complémentaire est obtenue en renseignant l'option `fun="cloglog"`.

Code : Fonctions de survie par groupes

```
surv.edema <- survfit(formula = Surv(time,status==2) ~ factor(edema), data=pbcc)
```

Dans notre exemple, les 3 fonctions de survie estimées ne se croisent pas (cf. graphique 5.1 de gauche). Elles révèlent par ailleurs que la présence d'œdèmes réduit d'autant plus la probabilité de survie qu'ils ne disparaissent pas malgré l'utilisation d'un traitement (`edema : 1`). Enfin, sur une échelle

Graphique 5.1 – Fonctions de survie $S(t)$ et $\text{cloglog}(S(t))$ selon la présence d'oedèmes

log-log complémentaire, les courbes obtenues sont approximativement parallèles (cf. graphique 5.1 de droite) ce qui conforte la validité de l'hypothèse de proportionnalité du modèle.

Dans le cas continu, l'approche est similaire; elle consiste à partitionner la variable et à la traiter comme une variable qualitative. Dans l'exemple ci-dessous, l'âge est partitionné en 3 classes à l'aide de la fonction `cut`.

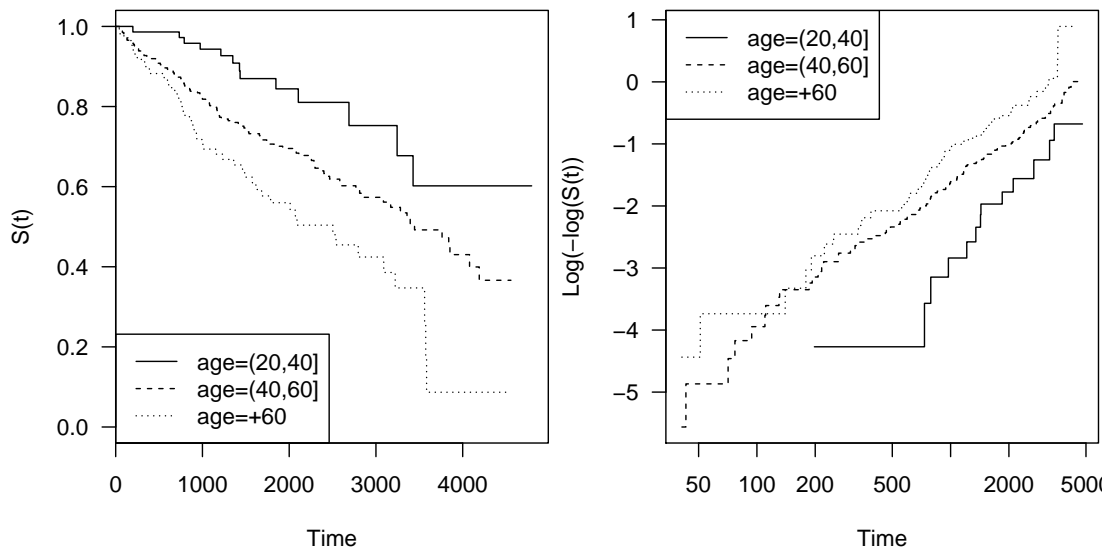
```
strata.age <- cut(pbc$age, breaks=c(20, 40, 60, Inf))
```

Cette fois encore, les courbes des fonctions de survie pour chaque classe ne se coupent pas (cf. graphique 5.2) et leurs représentations sur une échelle log-log complémentaire sont encore approximativement parallèles. Néanmoins, une telle représentation est moins probante de la validité de l'hypothèse de proportionnalité puisqu'elle dépend de la partition réalisée.

Enfin, dans tous les cas, il faut retenir que l'espacement entre les courbes parallèles *ne correspond pas* à l'estimation du risque relatif comme dans la représentation graphique des paramètres estimés présentée à la section 3.2.2. En effet, dans l'estimation des fonctions de survie avec `survfit`, on ne contrôle pas des différences de composition, de chaque classe, en les autres covariables.

5.2 Résidus de Schœnfeld (standardisés)

Dans le cas d'une covariable continue, la validité de l'hypothèse de proportionnalité peut être investiguée à partir des résidus de Schœnfeld standardisés, *dès que l'on se trouve en présence de plusieurs variables explicatives*. Contrairement aux résidus de martingale, ils ne comparent pas la valeur observée à une valeur prédite du nombre d'évènements, mais la valeur observée et prédite de la *variable explicative* lorsque l'évènement a lieu. Plus précisément, pour calculer le résidu de Schœnfeld s_i d'une variable X donnée, on compare sa valeur à t_i pour un individu i à la valeur moyenne parmi les individus encore à risques à cette date (et donc aussi potentiellement censurés plus tard). Les résidus de Schœnfeld ne sont donc calculés que pour les individus non censurés, à chaque date d'évènement,




Graphique 5.2 – Fonctions de survie $S(t)$ et $\text{cloglog}(S(t))$ selon la classe d'âge

et il existe un résidu de Schoenfeld pour chaque variable explicative.

Sous l'hypothèse de proportionnalité, [Therneau et Grambsch \(1994\)](#) montrent que les résidus standardisés de Schoenfeld vérifient :

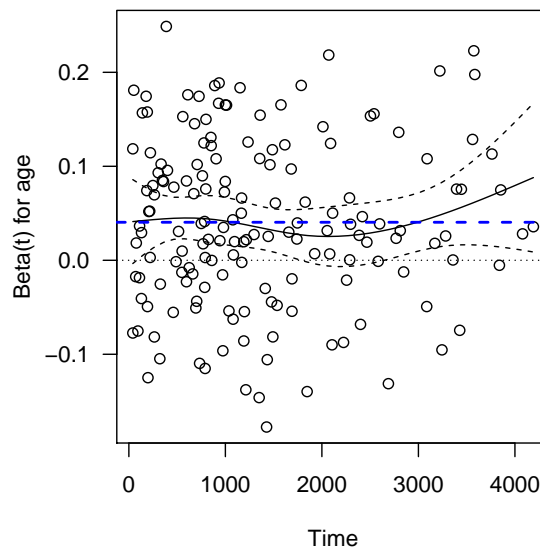
$$E[s_{kj}] + \hat{\beta}_j \approx \beta_j(t_k)$$

où s_{kj} est le résidu de Schoenfeld standardisé à la k^{e} date d'évènement (pour la j^{e} covariable) et $\hat{\beta}_j$ l'estimation obtenue à partir d'un modèle de Cox où le paramètre ne dépend pas du temps. Cela suggère de tester l'existence d'une corrélation (linéaire) entre s_i et le temps t_i , ou une fonction du temps $g(t_i)$. Les résultats du test sont souvent associés à une représentation graphique de $s_{kj} + \hat{\beta}_j$ en fonction du temps¹, avec une fonction lissée et ses intervalles de confiance. Si une ligne horizontale peut être tracée dans l'intervalle de confiance, alors il est raisonnable de penser que le risque est constant dans le temps, c'est-à-dire que l'on ne rejette pas l'hypothèse de proportionnalité. Le cas échéant, une telle représentation graphique renseigne sur la nature et l'ampleur de la non-proportionnalité.

Sous , le test de corrélation peut être effectué avec la fonction `cox.zph` appliqué aux résultats du modèle de Cox estimé (ici `fit.pbc`, voir section 3.2). L'option `transform` permet de préciser la fonction du temps que l'on souhaite appliquer. En pratique, on privilégie souvent $g(t) = t$ (`transform="identity"`), $g(t) = \log(t)$ (`transform="log"`) ou le rang de t_i (`transform="rank"`). Des résultats cohérents entre ces trois transformations permettent notamment de se prémunir de conclure de rejeter l'hypothèse de proportionnalité à cause de quelques valeurs aberrantes². Dans notre exemple, le temps de coagulation (`protime`) a la statistique de test la plus élevée et est la seule covariable pour laquelle l'hypothèse de hasards proportionnels est rejetée au seuil de 5%.

1. ou d'une fonction du temps.

2. Comme classiquement dans le cas de la corrélation linéaire.



Graphique 5.3 – Résidus de Schoenfeld pour le temps de coagulation (en log)

Code R : Test de corrélation des résidus de Schoenfeld avec le temps

```
zph.pbc <- cox.zph(fit.pbc, transform="identity")
```

	rho	chisq	p
age	0.01889	4.75e-02	0.82743
factor(edema)0.5	-0.11824	2.27e+00	0.13217
factor(edema)1	0.00197	6.08e-04	0.98033
log(bili)	0.10271	1.48e+00	0.22439
log(protime)	-0.30395	9.93e+00	0.00162
log(albumin)	0.01982	6.69e-02	0.79595
GLOBAL	NA	1.44e+01	0.02593

La fonction `plot` appliquée à l'objet créé par la fonction `cox.zph` (ici `zph.pbc`) permet d'obtenir la représentation graphique de $s_{kj} + \hat{\beta}_j$ en fonction du temps, pour la variable renseignée dans l'option `var`.

Code R : Représentation graphique des résidus de Schoenfeld

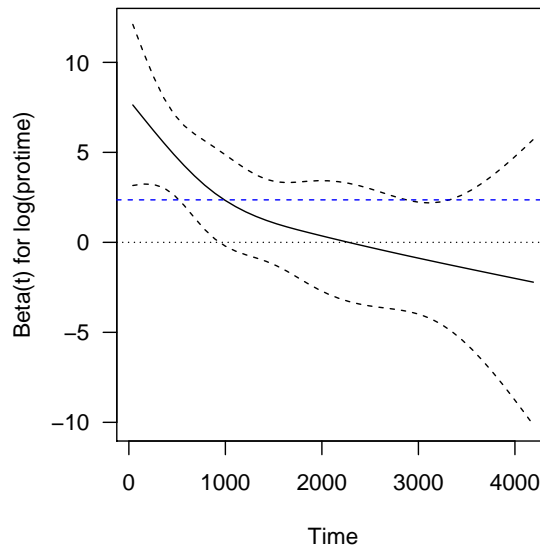
```
plot(zph.pbc, var="age")
abline(h=0, lty=3)

abline(h=fit.pbc$coefficients["age"], lty=2, lw=2, col="blue")
```

L'âge est associé à un risque plus élevé de décès, mais cet effet ne varie pas avec le temps (cf. graphique 5.3). On notera, par ailleurs, que l'effet estimé (ligne bleue) par le modèle de Cox classique se trouve quel que soit l'instant t , dans l'intervalle de confiance de la fonction lissée.

À l'inverse, l'impact du temps de coagulation évolue clairement au cours du temps³ (cf. graphique

3. L'option `resid=FALSE` dans la fonction `plot` permet de ne pas représenter les points correspondant à $s_{kj} + \hat{\beta}_j$.



Graphique 5.4 – Résidus de Schoenfeld pour le temps de coagulation (en log)

5.4). Au début, un temps élevé de coagulation est associé à un risque plus important de décès. Cependant, cet effet diminue avec le temps et n'est plus significatif trois ans (environ 1000 jours) après le diagnostic de la maladie. L'hypothèse de proportionnalité n'est donc pas satisfaite. Lorsqu'on l'impose par la spécification du modèle de Cox, on estime une valeur (représentée par une ligne horizontale bleue sur le graphique) plus faible et qui s'apparente à une valeur moyenne dans le temps.

5.3 Tenir compte de la non-proportionnalité

Plusieurs méthodes peuvent être envisagées pour tenir compte de la non-proportionnalité. Nous en présentons ici deux dont la pertinence dépend bien évidemment du cas traité.

5.3.1 Stratification

Dans le cas des variables qualitatives pour lesquelles l'hypothèse de proportionnalité n'est pas respectée, il peut être envisagé de stratifier le modèle en la variable (cf. section 3.3.1). Comme nous l'avons déjà expliqué, les effets des autres covariables sont alors considérés comme identiques dans chaque strate. Cette méthode peut aussi être mise en œuvre avec des variables quantitatives, mais sa pertinence dépendra alors de la partition préalable retenue.

5.3.2 Modélisation d'un effet dépendant du temps

Une covariable dont l'effet sur le taux de hasard évolue dans le temps induit le rejet de l'hypothèse de proportionnalité.

$$h(t|x) = h_0(t)e^{x'\beta(t)} \Rightarrow \frac{h(t|x_i)}{h(t|x_j)} = \exp((x_i - x_j)'\beta(t))$$

Si la forme fonctionnelle de la dépendance de β au temps est connue, sa prise en compte dans la modélisation s'effectue aisément en s'appuyant sur l'analogie :

$$h(t|x) = h_0(t)e^{x'\beta(t)} = h_0(t)e^{\tilde{x}'(t)\tilde{\beta}}$$

La démarche consiste donc à intégrer au modèle des covariables dépendant du temps comme évoqué dans la section 3.3.2. Dans le cas de covariables initialement fixes dans le temps, la constitution du nouveau jeu de données nécessaires est cependant facilitée par les fonctions disponibles dans le package **survival**. Nous présentons ici leur utilisation dans le cas où la forme fonctionnelle de $\beta(t)$ peut s'expliquer facilement.

Prenons un exemple : l'impact du temps de coagulation (*prottime*) semble dépendre linéairement du temps (cf. graphique 5.4 de droite), de telle sorte que l'on peut supposer que $\beta(t) = \alpha + \alpha_1 t$. La méthode consiste donc à transformer $\beta(t)X$ en $\alpha X + \alpha_1 t \times X$. Sous **R**, la déclaration d'une telle variable $X(t)$ peut être réalisée par l'option *time-transform* *tt* de *coxph*. Sa prise en compte dans le modèle s'effectue lors de sa déclaration avec *formula* en ajoutant *tt*(*X*) comme covariable. Ainsi, dans le code ci-dessous, *tt*(*prottime*) permet de déclarer la covariable supplémentaire $t \times \log(\text{prottime})$ ⁴.

Code **R** : Effet dépendant du temps dans le modèle de Cox

```
pfit.timedep <- coxph(formula = Surv(time,status==2) ~ log(bili) +
  age + log(prottime) + tt(prottime) +
  log(albumin) + factor(edema),
  data = pbc,
  tt = function(x, t, ...) log(x)*t)
```

Call:

```
coxph(formula = Surv(time, status == 2) ~ log(bili) + age + log(prottime) +
  tt(prottime) + log(albumin) + factor(edema), data = pbc, tt = function(x,
  t, ...) log(x) * t)
```

	coef	exp(coef)	se(coef)	z	p
log(bili)	8.35e-01	2.31e+00	8.33e-02	10.03	< 2e-16
age	3.96e-02	1.04e+00	7.79e-03	5.09	3.6e-07
log(prottime)	6.00e+00	4.02e+02	1.35e+00	4.43	9.3e-06
tt(prottime)	-2.42e-03	9.98e-01	8.62e-04	-2.80	0.0050
log(albumin)	-2.66e+00	6.96e-02	6.57e-01	-4.06	5.0e-05
factor(edema)0.5	3.10e-01	1.36e+00	2.26e-01	1.37	0.1712
factor(edema)1	8.77e-01	2.40e+00	2.91e-01	3.01	0.0026

```
Likelihood ratio test=241.6 on 7 df, p=<2e-16
n= 416, number of events= 160
(2 observations deleted due to missingness)
```

Dans notre exemple, le paramètre $\beta(t)$ est estimé par $6 - 0.00242 \times t$. La pertinence de cette estimation peut être mise en évidence sur le graphique réalisé à partir des résidus de Schoenfeld (cf. graphique 5.5 de gauche).

4. On notera que la déclaration sous **R** de la transformation évite de devoir ici constituer une nouvelle base de données incluant une covariable évoluant dans le temps ($\alpha_1 t \times X$) comme nous l'avons vu dans la section 3.3.2

Code R: Représentation graphique des résidus de Schoenfeld et du paramètre estimé

```
plot(zph.pbc, var="log(prottime)", resid=FALSE)
abline(coef(pfit.timedep)[3:4], col="red")
```

Lorsque la forme fonctionnelle de $\beta(t)$ n'est pas si évidente, celle-ci peut être approchée par une fonction en escalier, c'est-à-dire par des coefficients différents par intervalle de temps. Cette fois, il s'agit donc d'intégrer au modèle des interactions entre la variable et une indicatrice propre à chaque intervalle. De telle sorte, que la démarche s'apparente à nouveau à la prise en compte de variables évoluant dans le temps.

Sous R, cette stratégie impose cette fois de créer une nouvelle base de données comme dans la section 3.3.2, en générant pour chaque personne, une observation par intervalle. Mais son implémentation est facilitée par l'utilisation de la fonction `survSplit` qui nécessite toutefois de prédéfinir les intervalles de temps. Dans l'exemple ci-dessous, nous considérons la fonction en escalier définie sur les intervalles (0,1000], (1000,2000] et (2000,inf] (option `cut=`). Dans la nouvelle base, chaque intervalle est identifié par la variable `tgroup` renseignée dans l'option `episode`.

Code R: Partition des observations par intervalles

```
pbcsplit <- survSplit(Surv(time,status==2) ~ ., data = pbc,
                      cut=c(1000,2000),
                      episode = "tgroup")
```

Dans la base initiale (`pbcsplit`), l'individu 4 décède (`status=2`) 1925 jours après s'être vu diagnostiqué la maladie.

```
id time status
4 1925      2
```

Dans la nouvelle base, l'observation correspondante est scindée en deux observations. Sur l'intervalle (0,1000] (défini par les valeurs de `tstart` et `time`), elle ne connaît pas d'évènement (`event = 0`). Dans le second intervalle (`tgroup=2`), la personne décède (`event=1`) le 1925^e jour. Aucune observation sur le 3^e intervalle n'est générée puisque l'individu n'est plus à risque.

```
id tstart time event tgroup
4     0 1000     0     1
4  1000 1925     1     2
```

L'estimation du modèle est ensuite réalisée par la fonction `coxph`, en ajoutant un effet par intervalle `log(prottime):strata(tgroup)` au modèle.

Code R: Effet dépendant du temps par une fonction en escalier

```
split.fit <- coxph(formula = Surv(time,event) ~ log(bili) + age +
                  log(prottime):strata(tgroup) +
                  log(albumin) + factor(edema),
                  data = pbcsplit)
```

5. VALIDITÉ DE L'HYPOTHÈSE DE PROPORTIONALITÉ

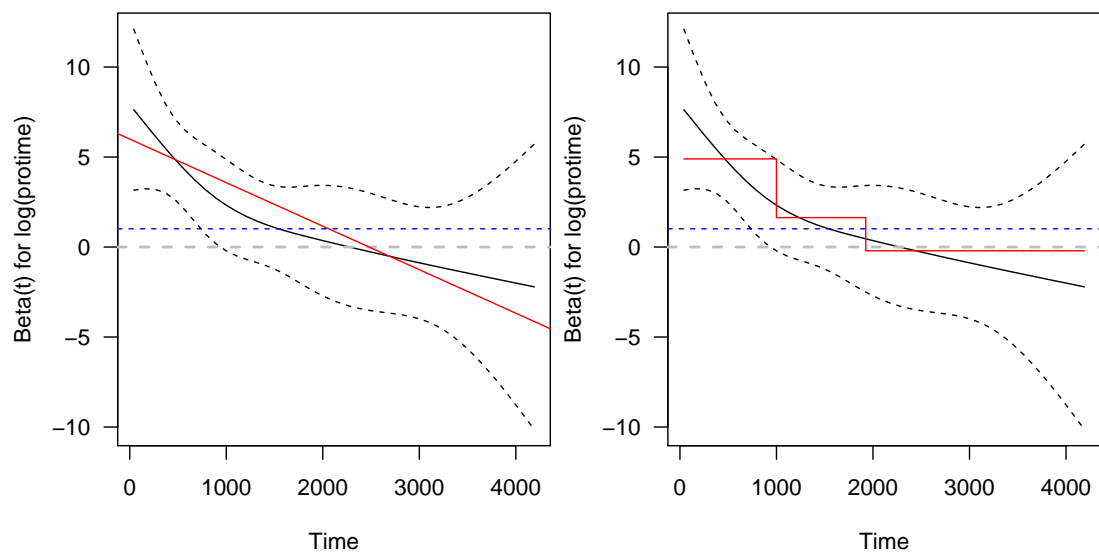
Call:

```
coxph(formula = Surv(time, event) ~ log(bili) + age + log(protime):strata(tgroup) +  
      log(albumin) + factor(edema), data = pbc.split)
```

	coef	exp(coef)	se(coef)	z
log(bili)	0.83933	2.31482	0.08335	10.07
age	0.04017	1.04099	0.00779	5.15
log(albumin)	-2.64656	0.07089	0.65766	-4.02
factor(edema)0.5	0.30023	1.35016	0.22608	1.33
factor(edema)1	0.88262	2.41724	0.29128	3.03
log(protime):strata(tgroup)tgroup=1	4.89543	133.67732	1.12969	4.33
log(protime):strata(tgroup)tgroup=2	1.63350	5.12176	1.66818	0.98
log(protime):strata(tgroup)tgroup=3	-0.20947	0.81101	1.57511	-0.13
	p			
log(bili)	< 2e-16			
age	2.6e-07			
log(albumin)	5.7e-05			
factor(edema)0.5	0.1842			
factor(edema)1	0.0024			
log(protime):strata(tgroup)tgroup=1	1.5e-05			
log(protime):strata(tgroup)tgroup=2	0.3275			
log(protime):strata(tgroup)tgroup=3	0.8942			

```
Likelihood ratio test=240 on 8 df, p=<2e-16  
n= 919, number of events= 160  
(4 observations deleted due to missingness)
```

Dans notre exemple, $\beta(t)$ vaudrait 4,9 sur $(0,1000]$, 1,6 sur $(1000,2000]$ et -0.2 au-delà, comme en témoigne aussi le graphique 5.5 de droite.



Graphique 5.5 – Différentes modélisations d'effets dépendants du temps de la prothrombine (en log) - linéaire et en escaliers

6 Hétérogénéité individuelle inobservée

Ce chapitre est consacré à la prise en compte de l'hétérogénéité *individuelle* dans les modèles à hasards proportionnels. En effet, si certains individus sont plus susceptibles de connaître l'évènement étudié, rien ne garantit que les raisons de cette hétérogénéité soit intégralement saisie par les covariables observées. La présence de facteurs de risques individuels inobservés (ou inobservables), et l'hétérogénéité inobservée qui en découle, est appelée *fragilité (frailty)* dans les modèles de survie ; les personnes les plus « fragiles » présentant un risque plus élevé.

Les conséquences de la présence d'hétérogénéité individuelle ont été depuis longtemps identifiées. Si des facteurs individuels inobservés impactent la probabilité de connaître l'évènement étudié, la *forme* de la fonction de hasard observée au niveau de la population pourra être différente de celle observée au niveau individuel. Par exemple, même si la fonction de hasard de chaque individu est constante au cours du temps, la fonction de hasard de la population pourra être dépendante du temps, généralement décroissante, à cause de *l'effet de sélection* induit par le terme de fragilité individuelle. Schématiquement, les individus les plus à risques « disparaissent » plus rapidement, la population restante étant au fur et à mesure du temps constituée de personnes présentant un risque plus faible. Cette sélection systématique entraîne ainsi une distorsion de la forme des fonctions de hasard (et de survie) observée sur la population considérée. La non prise en compte de l'hétérogénéité individuelle conduit donc à sous-estimer une dépendance positive au temps de la fonction de hasard et à surestimer une dépendance négative au temps. Enfin, les paramètres des covariables sont aussi sous-estimés, même si l'hétérogénéité inobservée n'est pas corrélée aux covariables du modèle¹.

Dans une première partie, nous détaillons comment intégrer la prise en compte de cette hétérogénéité inobservée au modèle de Cox (section 6.1)². La démarche amène à définir les fonctions de hasard et de survie marginales dont l'interprétation diffère des fonctions de hasard et de survie présentées jusqu'ici (section 6.2). Pour pouvoir estimer les paramètres, il est cependant nécessaire de préciser la distribution du paramètre de fragilité intégré au modèle. Les différentes distributions possibles (et actuellement implémentées sous **R**) et leurs conséquences sur la modélisation retenue sont présentées dans la section 6.3. Enfin, la dernière partie est consacrée à la mise en œuvre sous **R** des différentes spécifications discutées.

6.1 Le modèle de Cox avec fragilité individuelle

La prise en compte de l'hétérogénéité individuelle repose sur une extension des modèles à hasards proportionnels auxquels on introduit une variable aléatoire non observée, appelée *fragilité (frailty)*.

1. Comme nous le verrons, les estimations des paramètres de modèles de fragilité ont aussi *une interprétation différente* de celle du modèle à hasards proportionnels sans prise en compte de l'hétérogénéité individuelle

2. [Elbers et Ridder \(1982\)](#) ont montré que le modèle avec fragilité n'était identifiable, dans le cas des modèles de survie étudiés dans ce document, que si d'autres covariables sont présentes dans le modèle.

La relation qui lie la fonction de survie de la population aux fonctions de survie individuelles dépend alors de la distribution du terme de fragilité, et notamment de sa variance qui détermine le degré d'hétérogénéité de la population étudiée.

Formellement, il s'agit d'introduire un effet aléatoire aux modèles à hasards proportionnels ayant un effet *multiplicatif*³ sur la fonction de hasard de base (Vaupel *et al.*, 1979). Cet effet est supposé *ne pas dépendre du temps*⁴ et *ne pas être corrélé aux covariables observées*⁵. Plus précisément, la fonction de hasard d'un individu à l'instant t est donnée par :

$$h(t|u,x) = uh_c(t|x) = uh_0(t) \exp(x^t\beta) \quad (6.1)$$

où u est le terme de fragilité individuelle et $h_c(t|x)$ correspond à une caractérisation de la fonction de hasard individuelle sans prise en compte de la fragilité et dont $h_0(t)$ est la fonction de hasard de base.

La variable aléatoire u doit prendre des valeurs positives (par définition de la fonction de hasard) et, sans restriction, il peut être convenu que $E[u] = 1$ ⁶. Dès lors, le risque individuel se réduit si $u < 1$ ou s'accroît si $u > 1$. La variance σ^2 de u traduit l'hétérogénéité de la population étudiée. Si la variance est élevée, les valeurs du terme de fragilité sont très dispersées et la population étudiée très hétérogène. À l'inverse, si σ^2 est faible (voire nulle comme dans le cas du modèle à hasards proportionnels sans hétérogénéité) les valeurs de u sont concentrées autour de 1 et la population considérée comme homogène. Enfin, la densité f_U de la variable caractérise la distribution du terme de fragilité dans la population à l'instant $t = 0$. En effet, comme nous le verrons par la suite, la distribution de la fragilité sur la population encore à risques à la date t évolue avec le temps. En particulier, la fragilité moyenne de la population encore à risque à la date t diminue, puisque les individus les plus « fragiles » ne sont plus présents.

La fonction de hasard individuelle s'interprète comme une fonction de hasard *conditionnelle* à u , à laquelle correspond une fonction de survie individuelle, elle aussi *conditionnelle* à u :

$$S(t|u,x) = \exp(-uH_c(t|x)) \quad (6.2)$$

où $H_c(t|x) = H_0(t) \exp(x^t\beta)$ correspond à la fonction de hasard cumulé sans prendre en compte l'hétérogénéité individuelle.

La prise en compte de l'hétérogénéité individuelle complexifie cependant l'interprétation des coefficients estimés. Dans le cas d'une covariable X dont l'effet est mesuré par le paramètre β , $\exp(\beta)$ correspond à un risque relatif (quel que soit l'instant t) entre deux individus qui diffèrent seulement sur cette covariable (voir section 3.1.1). Ainsi, dans le cas des modèles de fragilité, cette comparaison s'effectue notamment entre deux personnes *présentant la même valeur du paramètre de fragilité*, ce qui réduit l'intérêt de l'analyse des coefficients estimés. Par exemple, si en absence d'hétérogénéité individuelle, il est possible de quantifier l'impact d'être mariée pour deux femmes ayant un niveau de formation identique sur la probabilité d'avoir un enfant, une interprétation aussi parlante n'est plus

3. Cet effet multiplicatif est une hypothèse faite communément. Il existe cependant des modèles de survie accélérée avec fragilité, ou des modèles de fragilité additifs. Ils ne sont pas étudiés dans ce document.

4. Des modèles avec fragilité non constante dans le temps ont été proposés par de nombreux auteurs en présence de données groupées (*clustered*) ou récurrentes (on parle alors de *shared frailty* ou *multivariate frailty*), mais ne font pas l'objet d'une présentation dans ce document (voir par exemple Manda *et Meyer*, 2005; Harkanen *et al.*, 2003).

5. Une hypothèse qui peut être relâchée si l'on dispose de données groupées (*clustered*).

6. Le facteur d'échelle étant, le cas échéant, intégré au hasard de base.

possible dans le cas des modèles de fragilité puisqu'ils supposent de comparer deux femmes ayant aussi le même facteur de risque, qui est inobservable et de fait non défini explicitement.

En cela, dans les modèles de fragilité, le paramètre β est souvent assimilé à un effet individuel de X ou plus communément *conditionnel à la fragilité*. Il est donc utile pour l'analyse et la présentation des résultats de comprendre comment sont reliées les fonctions de survie et de hasard conditionnelles des modèles (6.1) et (6.2) aux fonctions de hasard et de survie dites *marginales* (ou *non conditionnelles*), c'est-à-dire moyenne sur la sous-population définie seulement par des caractéristiques observables communes.

6.2 Les fonctions de hasard et de survie marginales

La fonction de survie *marginale* est obtenue en intégrant les fonctions de survie individuelles selon l'effet aléatoire inobservé. Si l'on note f_U la densité du terme de fragilité, son expression correspond donc à :

$$S(t|x) = \int_0^\infty S(t|u,x) f_U(u) du$$

Ainsi, la fonction de survie marginale correspond à la moyenne des fonctions de survie individuelles sur *l'ensemble de la population*.

$$S(t|x) = \int_0^\infty \exp(-uH_c(t|x)) f_U(u) du = \mathcal{L}(H_c(t|x)) \quad (6.3)$$

Il est intéressant de noter qu'elle peut s'exprimer en fonction de la transformée de Laplace, \mathcal{L} , associée à f_U ⁷.

De manière analogue, la fonction de hasard marginale est obtenue en intégrant les fonctions de hasard individuelles, mais *sur la seule sous-population des individus encore à risque à la date t* .

$$\begin{aligned} h(t|x) &= \int_0^\infty h(t|x,u) f_U(u|T > t,x) du \\ &= h_c(t|x) E[u|T \geq t,x] \end{aligned}$$

La fonction de hasard marginale est le *produit* de la fonction de hasard du modèle sans fragilité par le terme de fragilité *moyen* de la (sous) population encore à risque à la date t . Cette expression quelque peu complexe vise surtout à mettre en évidence, tout d'abord, que *la fonction de hasard de la population dépend éventuellement du temps* (même en présence d'une fonction de hasard hors hétérogénéité inobservée $h_c(t|x)$ constante). Intuitivement, cette fragilité moyenne diminue avec le temps puisque les plus « fragiles » tendent à connaître l'évènement assez vite. Ainsi, la fonction de hasard de la (sous)population décroît plus vite (ou augmente moins vite) que la fonction de hasard conditionnelle; la forme de la fonction de hasard marginale peut donc être complètement différente de celle des fonctions de hasard conditionnelles. Par ailleurs, on notera aussi que cette dépendance temporelle *dépend de la densité du terme de fragilité*.

7. L'avantage de cette écriture, mise en évidence par Hougaard (1984), est que la transformée de Laplace pour de nombreuses distributions est souvent connue et d'expression simple. Comme nous le verrons dans la partie 6.3, ces considérations pratiques ont guidé de nombreux praticiens dans le choix de la distribution supposée de u .

Comme pour la fonction de survie, la fonction de hasard marginale peut aussi s'exprimer en fonction de la transformée de Laplace \mathcal{L} ⁸ :

$$h(t|x) = -\frac{\mathcal{L}^{(1)}(H_c(t|x))}{\mathcal{L}(H_c(t|x))} h_c(t|x) \quad (6.4)$$

où $\mathcal{L}^{(1)}$ correspond à la dérivée première de \mathcal{L} , par rapport à t , et $h_c(t|x) = h_0(t) \exp(x^t \beta)$ désigne, comme dans l'équation (6.1), la fonction de hasard sans prise en compte de l'hétérogénéité.

Comme nous l'avons souligné, l'interprétation des paramètres dans les modèles de fragilité est plus complexe que dans le cas usuel vu jusqu'à présent. Si, *conditionnellement au terme de fragilité*, le ratio des fonctions de hasard permet d'exprimer le risque relatif entre deux individus aux caractéristiques observables semblables, il est préférable d'analyser le ratio des fonctions de hasard marginales, pour caractériser l'effet moyen sur la sous population (hétérogène) des individus présentant les mêmes caractéristiques observables.

Pour cela, il faut intégrer le ratio des fonctions de hasard individuelles sur la distribution des fragilités à la date t . Ainsi, l'effet d'une covariable sur le ratio des fonctions de hasards marginales dépend de la distribution du paramètre de fragilité *au cours du temps*, et donc en pratique de l'hypothèse faite sur sa distribution. Les conséquences de ce choix sont détaillées pour différentes distributions dans la section suivante.

6.3 Distributions du paramètre de fragilité et conséquences

Pour pouvoir estimer les paramètres dans un modèle de fragilité individuelle, il est nécessaire de préciser la distribution du terme de fragilité⁹ afin de pouvoir expliciter la vraisemblance partielle. Les distributions (présentées) les plus souvent utilisées sont la distribution gamma et inverse gaussienne (section 6.3.1), positive stable (section 6.3.2) et composée de Poisson (section 6.3.3)¹⁰. Il n'existe pas de justification théorique permettant de privilégier telle ou telle distribution pour l'ensemble des cas, leur utilisation étant essentiellement liée à la disponibilité de leur implémentation avec les logiciels statistiques standards. En pratique, il est ainsi courant de comparer les résultats obtenus avec différentes distributions. Au-delà, il nous semble surtout important de comprendre les conséquences sur les résultats de la densité choisie. Pour chaque distribution, nous précisons ainsi la fonction de hasard et de survie marginales, mais aussi le ratio des fonctions de hasard marginales dont l'importance dans l'interprétation des résultats a été soulignée précédemment et qui présentent des propriétés

8. à partir de la relation $h(t|x) = -\frac{d \log S(t|x)}{dt}$

9. Plus précisément, les modèles de survie introduisant un terme de fragilité peuvent être estimés si la fonction de hasard de base suit une distribution connue (cas paramétrique) ou si la densité du terme de fragilité est donnée (cas du modèle à hasards proportionnels). Il n'est pas possible de relâcher simultanément ces deux hypothèses.

10. Heckman et Singer (1982), dans un papier très influent, ont cherché à mettre en évidence la forte dépendance des paramètres estimés à la distribution retenue pour la fragilité. Si leurs résultats ont été contestés (voir par exemple Klein et al., 1992), Heckman et Singer (1984) ont aussi proposé comme solution d'estimer non paramétriquement la distribution de la fragilité par maximum de vraisemblance non paramétrique (*non-parametric maximum likelihood estimator NPML*). Plus précisément, leur estimateur conduit à une distribution de la fragilité qui prend des valeurs u_1, \dots, u_k avec des probabilités π_1, \dots, π_k , k restant à déterminer de manière itérative (au contraire des modèles à fragilité discrète où le nombre de groupes est fixé *a priori*). Il faut cependant préciser que cette méthode suppose connue *la forme paramétrique de la fonction de hasard de base*. Par ailleurs, Trussel et Richards (1985) ont souligné que l'estimation des paramètres obtenus par cette méthode restait *sensible au choix de la distribution de la fonction de hasard de base*.

différentes selon de la densité retenue ¹¹.

6.3.1 Distributions gamma et inverse gaussienne de la fragilité

La distribution gamma est la plus souvent usitée dans la littérature car sa transformée de Laplace est facile à calculer et permet ainsi de déterminer facilement la forme des fonctions de survie et de hasard marginales. Néanmoins, il faut rappeler qu'il n'existe pas de justification théorique pour privilégier ce choix dans toutes les situations concrètes, même si [Abbring et Van den Berg \(2007\)](#) montrent que la distribution des termes de fragilité parmi les personnes encore à risques converge dans de nombreux cas vers une distribution gamma sous certaines hypothèses de régularité.

Dans le cas d'une distribution gamma, la fonction de hasard et de survie marginale sont :

$$h(t|x) = \frac{h_c(t|x)}{1 + \sigma^2 H_c(t|x)} \text{ et } S(t|x) = (1 + \sigma^2 H_c(t|x))^{-1/\sigma^2}$$

où σ^2 désigne la variance de u .

Il en découle que le ratio des fonctions de hasard marginales dépend du temps. Par exemple, dans le cas d'une covariable x binaire :

$$\frac{h(t|x=1)}{h(t|x=0)} = \left(\frac{1 + \sigma^2 H_0(t)}{1 + \sigma^2 H_0(t) e^\beta} \right) e^\beta \quad (6.5)$$

Cette expression est telle que le ratio de leurs fonctions de hasard marginales est égal au ratio des fonctions de hasard conditionnelles au début de la période d'étude, soit e^β à $t = 0$, puis diminue avec le temps. Cet effet d'atténuation dépend de l'ampleur de l'hétérogénéité des individus (σ^2) et du paramètre β , mais aussi de la fonction de hasard cumulé de base, c'est-à-dire de la forme de la dépendance temporelle. On notera cependant que :

$$\lim_{t \rightarrow \infty} \frac{h(t|x=1)}{h(t|x=0)} = 1 \quad (6.6)$$

Cette propriété est appelée *effet d'atténuation de la fragilité*.

La distribution inverse gaussienne est une alternative souvent utilisée à la distribution gamma. Les fonctions de hasard et de survie marginales correspondantes sont données par les expressions suivantes :

$$h(t|x) = \frac{h_c(t|x)}{\sqrt{1 + 2\sigma^2 H_c(t|x)}}$$

$$S(t|x) = \exp\left(\frac{1}{\sigma^2} (1 - [1 + 2\sigma^2 H_c(t|x)]^{1/2})\right)$$

Le ratio des fonctions de hasard marginales décroît toujours avec le temps et à $t = 0$ il est aussi égal au coefficient de proportionnalité e^β , des fonctions de hasard conditionnelles. Cependant,

$$\lim_{t \rightarrow \infty} \frac{h(t|x=1)}{h(t|x=0)} = \sqrt{e^\beta} \quad (6.7)$$

de telle sorte que l'effet de la fragilité ne disparaît pas complètement avec le temps.

11. Le lecteur intéressé pourra compléter cette présentation sommaire de ces distributions avec l'ouvrage "The Frailty Model" de [Duchateau et Janssen \(2010\)](#).

6.3.2 Distribution positive stable de la fragilité

Avec les distributions gamma et gaussienne inverse, l'hypothèse de proportionnalité des fonctions de hasard n'est plus respectée au niveau marginal, comme nous l'avons vu, puisque le ratio des fonctions de hasard dépend du temps. *Les distributions positives stables permettent, elles, de « conserver » au niveau marginal cette hypothèse de proportionnalité.* Pour attractive que soit cette propriété, il faut néanmoins garder à l'esprit qu'il s'agit d'une hypothèse qui doit être validée par les données.

Une distribution positive stable de paramètre γ , compris entre $]0; 1[$ ¹² a pour densité¹³ :

$$f_U(u) = -\frac{1}{\pi u} \sum_{k=1}^{\infty} \frac{\Gamma(k\gamma + 1)}{k!} (-u^{-\gamma})^k \sin(\gamma k\pi)$$

Cette fonction a une *espérance et une variance infinies*. Cette propriété, qui pourrait rendre cette distribution difficile à exploiter, assure cependant de l'indépendance du terme de fragilité avec les covariables (Hougaard, 1986).

Les fonctions de hasard et de survie marginales qui en découlent sont données par :

$$h(t|x) = \gamma H_c(t|x)^{\gamma-1} h_c(t|x) \text{ et } S(t|x) = \exp(-H_c(t|x))^{\gamma}$$

Dès lors, le ratio des fonctions de hasard marginales est :

$$\frac{h(t|x=1)}{h(t|x=0)} = \exp(\gamma\beta) \quad (6.8)$$

Il ne dépend pas du temps et l'hypothèse de proportionnalité est ainsi respectée par les fonctions de hasard marginales, mais avec un coefficient différent. De plus, les paramètres restent affectés par l'effet d'atténuation.

6.3.3 Distribution mélangée de Poisson (*compound Poisson*) de la fragilité

Cette distribution a été introduite dans le cadre des modèles de fragilité par Aalen (1988, 1992). L'intérêt de cette distribution est qu'elle *modélise l'existence d'un sous-groupe de personnes qui ne connaîtront jamais (ou ne peuvent connaître) l'évènement considéré, c'est-à-dire pour lesquelles le terme de fragilité est nulle*¹⁴. Par exemple, Aalen (1992) illustre l'utilisation de cette distribution pour étudier la fertilité des femmes en Norvège. Le sous-groupe correspond alors aux femmes ne pouvant concevoir d'enfants.

Formellement, la distribution mélangée de Poisson est définie par une densité continue pour $u > 0$ et un point de masse en 0.

$$P(u=0) = \exp\left(\frac{-(\nu+1)}{\sigma^2(\nu)}\right)$$

où ν est un paramètre de la loi qui est positif¹⁵.

12. pour que les valeurs de la variable soient positives

13. Plus généralement, une distribution positive stable a pour propriété que la *distribution de la somme* de n variables X_1, \dots, X_n aléatoires et identiquement distribuées a la même distribution que $n^{1/\gamma} X_1$. La distribution normale est, par exemple, une distribution positive stable de paramètre 2.

14. Les modèles de fragilité avec une distribution mélangée de Poisson font partie des modèles dits de guérison (*cure models*).

15. Cette écriture s'appuie ici sur l'hypothèse que $E[U] = 1$ pour réduire le nombre de paramètre renseigné.

6.4 Mise en œuvre sous R

Il existe plusieurs méthodes pour estimer les modèles de fragilité : l'approche par l'algorithme d'Espérance-Maximisation (EM), l'approche par vraisemblance partielle pénalisée et l'approche bayésienne. Le package **frailtyEM** (Balan et Putter, 2018) que nous présentons dans ce document s'appuie sur l'algorithme EM. Il est actuellement le seul à pouvoir traiter l'ensemble des distributions paramétriques présentées précédemment ¹⁶.

```
library(frailtyEM)
```

L'estimation d'un modèle avec hétérogénéité individuelle est réalisée avec la fonction **emfrail** dont la syntaxe complète est renseignée ci-dessous :

```
emfrail(formula, data, distribution = emfrail_dist(),
control = emfrail_control(), model = FALSE, model.matrix = FALSE, ...)
```

La loi du paramètre de fragilité est indiquée par l'argument `distribution = emfrail_dist(...)` où `emfrail_dist` précise la distribution choisie ¹⁷. Dans l'exemple ci-dessous, le paramètre de fragilité est supposé suivre une loi gamma. La déclaration de l'intégration au modèle d'un terme de fragilité est réalisée lors de sa déclaration dans l'argument `formula` en ajoutant `cluster(id)` où `id` désigne un identifiant présent dans la base de données.

Code : Modèle à hétérogénéité individuelle - Loi Gamma


```
gamma.pbc <- emfrail(formula = Surv(time,status==2) ~ age +
                    factor(edema) + log(bili) + log(protime) +
                    log(albumin) + cluster(id),
                    data = pbc,
                    distribution = emfrail_dist(dist = "gamma"))
```

Call:

```
emfrail(formula = Surv(time, status == 2) ~ age + factor(edema) +
        log(bili) + log(protime) + log(albumin) + cluster(id), data = pbc,
        distribution = emfrail_dist(dist = "gamma"))
```

Regression coefficients:

	coef	exp(coef)	se(coef)	adj. se	z	p
age	0.0514	1.0527	0.0104	0.0111	4.6114	0.00
factor(edema)0.5	0.4405	1.5535	0.3053	0.3341	1.3184	0.19
factor(edema)1	1.2861	3.6185	0.4171	0.4350	2.9563	0.00
log(bili)	1.0461	2.8466	0.1115	0.1425	7.3415	0.00

16. D'autres packages  peuvent aussi être utilisés, comme le package **survival** utilisé dans les chapitres précédents. Celui-ci s'appuie sur la méthode de vraisemblance pénalisée (Therneau *et al.*, 2003) - ce qui implique que les résultats obtenus avec le package présenté ici puissent différer - et peut implémenter les distributions gamma et log-normal pour la fragilité. Cet algorithme plus rapide ne peut cependant pas être appliqué aux autres distributions (et aux données tronquées à gauche).

17. et si les données sont tronquées à gauche.


```
log(protime)      3.3312  27.9724  1.1218  1.2367  2.6937  0.01
log(albumin)     -2.7518   0.0638  0.8175  0.8328 -3.3043  0.00
Estimated distribution: gamma / left truncation: FALSE
```

Fit summary:

```
Commenges-Andersen test for heterogeneity: p-val  0.0385
no-frailty Log-likelihood: -751.165
Log-likelihood: -748.844
LRT: 1/2 * pchisq(4.64), p-val 0.0156
```

Frailty summary:

	estimate	lower 95%	upper 95%
Var[Z]	0.569	0.042	1.380
Kendall's tau	0.221	0.021	0.408
Median concordance	0.218	0.020	0.413
E[logZ]	-0.311	-0.830	-0.021
Var[logZ]	0.759	0.043	2.685
theta	1.758	0.724	23.634

Confidence intervals based on the likelihood function

Outre la présentation classique des estimations des paramètres, la fonction `emfrail` teste la présence d'hétérogénéité inobservée quelle que soit la distribution retenue (Fit summary). Plus précisément, le test réalisé est :

$$H_0 : \sigma^2 = 0 \text{ versus } H_A : \sigma^2 > 0,$$

où σ^2 correspond à la variance du terme de fragilité. En pratique, le package **frailtyEM** réalise deux tests de présence d'hétérogénéité inobservée (quelle que soit la distribution retenue) : le test de Commenges-Andersen¹⁸ (Commenges et Andersen, 1995) et le test du ratio de vraisemblance. Dans notre exemple, les deux tests conduisent à ne pas rejeter la présence d'hétérogénéité inobservée au seuil de 5%, mais à la rejeter à 1%.

Enfin, plusieurs mesures concernant le paramètre de fragilité sont présentées (Frailty summary) dépendant en partie de la distribution retenue. On trouvera ainsi une estimation des paramètres de la loi. Dans le cas de la distribution gamma, par exemple, sont estimés la variance (Var [Z]) et le paramètre $\theta = \frac{1}{\sigma^2}$ de la loi. Plus précisément, la loi gamma retenue pour le terme de fragilité, $\Gamma(\theta, \theta)$, est définie par un seul paramètre (*one parameter gamma distribution*), un choix classique dans les modèles de durée¹⁹.

Le tableau ci-dessous compare les (exponentielles) des estimations des paramètres avec ou sans prise en compte de l'hétérogénéité individuelle, en distinguant différentes lois pour le paramètre de fragilité.

18. qui ne dépend pas de la distribution retenue et est donc identique quelle que soit celle renseignée.

19. Cet exemple doit donc inciter le lecteur à se renseigner sur la saisie (éventuelle) des arguments nécessaires à l'implémentation par `emfrail` de la distribution souhaitée, mais aussi sur leurs liens avec les paramètres classiques des lois (voir la vignette correspondante du package, Balan et Putter, 2018). Le code comparant plusieurs modèles de fragilité dans ce document illustre ainsi les différents paramètres de `emfrail_dist` à saisir pour étudier le cas de la distribution mélangée de Poisson.

Code R: Estimation de plusieurs modèles de fragilité

```
list.frailty <- list("gamma","pvf","stable")
results <- sapply(X = list.frailty,
  FUN = function(frailty.type)
    exp(coef(emfrail(formula = Surv(time,status==2) ~ age +
      factor(edema) + log(bili) + log(protime) +
      log(albumin) + cluster(id),
      data = pbc,
      distribution = emfrail_dist(dist = frailty.type))))
)

## Distribution mélangée de Poisson
## Cette distribution nécessite la saisie de plusieurs paramètres :
## - dist="pvf"
## - une valeur initiale pour pvf (>0)
## - une valeur initiale pour theta (>0)

compound.poisson.pbc <- emfrail(Surv(time, status==2) ~ age +
  factor(edema) + log(bili) + log(protime) +
  log(albumin) + cluster(id),
  data = pbc,
  distribution = emfrail_dist(dist="pvf",theta=1.5,pvfm=0.3))
compound.poisson <- exp(coef(compound.poisson.pbc))

sansfrailty <- exp(coef(fit.pbc))
results <- cbind(sansfrailty,results,compound.poisson)

colnames(results) <- c("Sans frailty","Gamma",
  "Inv. gaussian","Stable",
  "Compound Poisson")
```

	Sans frailty	Gamma	Inv. gaussian	Stable
age	1.0413	1.0527	1.0512	1.0413
factor(edema)0.5	1.3256	1.5535	1.5581	1.3247
factor(edema)1	2.7524	3.6185	3.5376	2.7480
log(bili)	2.3609	2.8466	2.8378	2.3597
log(protime)	10.5902	27.9724	24.2489	10.5772
log(albumin)	0.0808	0.0638	0.0591	0.0816
	Compound Poisson			
age	1.0525			
factor(edema)0.5	1.5027			
factor(edema)1	3.5469			
log(bili)	2.8009			
log(protime)	26.3315			
log(albumin)	0.0666			

Lorsque la dépendance au temps est positive ($\exp(\hat{\beta}) > 1$), l'effet de la covariable considérée, si l'on ne tient pas compte de l'hétérogénéité inobservée (Sans frailty) est sous-estimé. Quelle que soit la distribution retenue pour le paramètre de fragilité, l'effet estimé attendu devrait donc être supérieur. Cela est vérifié, dans notre exemple, pour les distributions gamma, inverse gaussienne et mélangée de Poisson. Les résultats avec la distribution positive stable sont proches de ceux sans hétérogénéité,

ce qui n'est pas surprenant car elle fait l'hypothèse de hasards proportionnels au niveau marginal, comme le modèle sans hétérogénéité (cf. section 6.3.2). Dans le cas où la dépendance au temps est négative ($\exp(\hat{\beta}) < 1$), la non prise en compte de l'hétérogénéité individuelle inobservée surestime l'effet. En présence d'hétérogénéité inobservée, le coefficient obtenu est donc plus faible. Enfin, dans notre exemple, les estimations des paramètres dépendent peu de la distribution retenue.

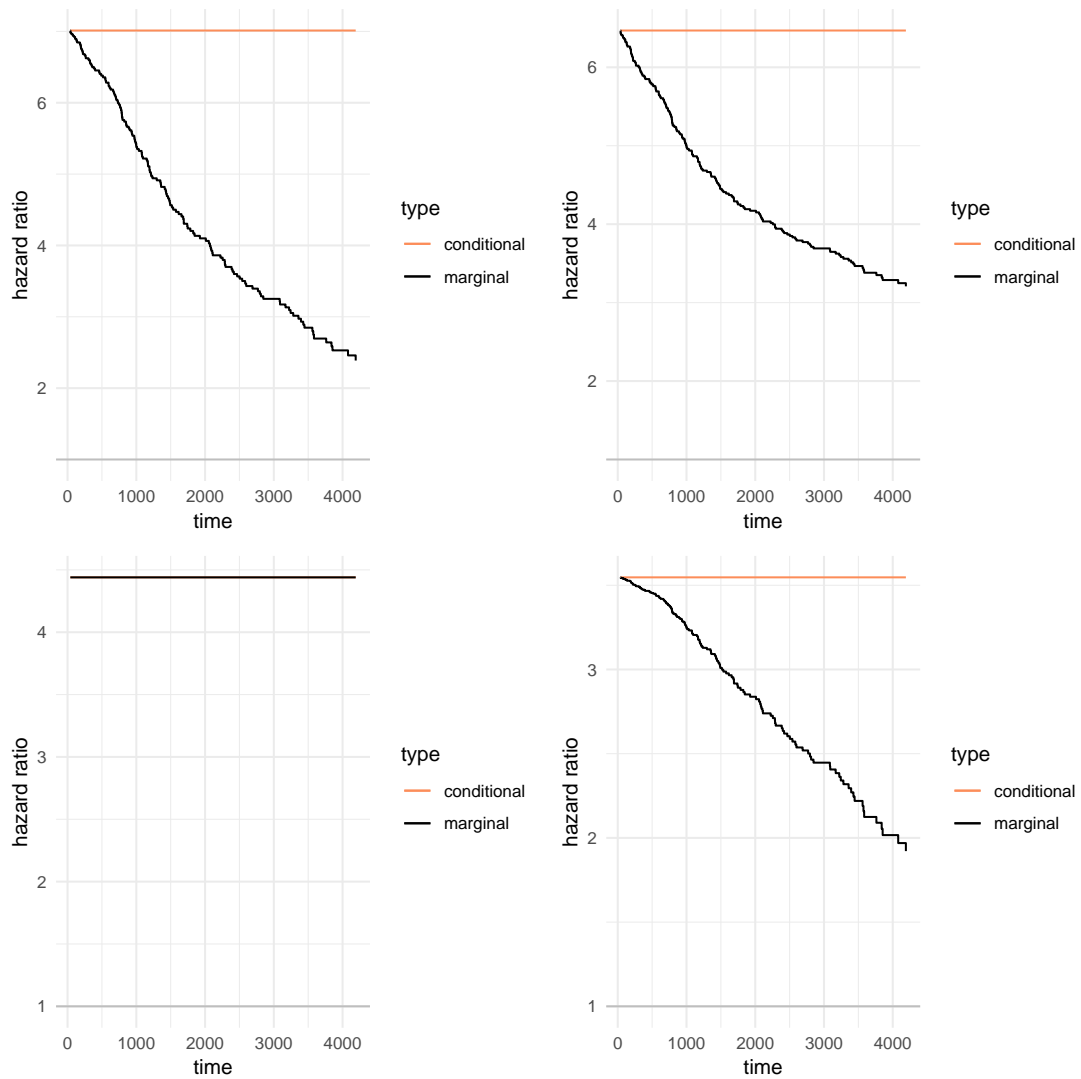
Il est souvent utile d'illustrer l'effet de la prise en compte de l'hétérogénéité en comparant les ratios de hasard marginaux et conditionnels entre deux individus types. Dans l'exemple ci-dessous, on s'intéresse au ratio de hasard entre deux personnes qui ne diffèrent que par la présence ou non d'œdèmes (edema : 1 ou 0). Un nouveau jeu de données (newdata) comportant deux observations est créé préalablement. La représentation graphique du ratio de hasard marginal et conditionnel (type="hr") est ensuite obtenue avec la fonction autoplot.

Code R: Représentation graphique des ratios de hasards marginaux et conditionnels

```
newdata <- data.frame(age = rep(mean(pbc$age), 2),
                      edema = c(1, 0),
                      bili = rep(mean(pbc$bili), 2),
                      protime = rep(mean(pbc$protime, 2)),
                      albumin = rep(mean(pbc$albumin, 2))
                      )

autoplot(gamma.pbc, type="hr", newdata=newdata)
```

Les graphiques 6.1 représentent les ratios de hasards marginaux (en noir) et conditionnels (en rouge). Par définition, le ratio de hasard conditionnel correspond au coefficient (exponentialisé) estimé par le modèle (6.1). Il est indépendant du temps quelle que soit la distribution de la fragilité retenue. Les ratios de hasards marginaux dépendent du temps dans le cas d'une distribution gamma, inverse gaussienne et mélange de Poisson. Ils sont décroissants dans nos trois graphiques. À l'inverse, la distribution positive stable « conserve » au niveau marginal l'hypothèse de proportionnalité, et le ratio de hasard correspondant est donc constant au cours du temps (cf. Graphique 6.1 correspondant).



Graphique 6.1 – Ratios de hasards marginaux et conditionnels selon la distribution de la fragilité retenue (Gamma, inverse gaussienne, stable, mélangée de Poisson)

Bibliographie

- Odd O. AALEN : Heterogeneity in survival analysis. *Statistics in Medicine*, 7(11):1121–1137, 1988.
- Odd O. AALEN : Modelling heterogeneity in survival analysis by the compound poisson distribution. *The Annals of Applied Probability*, 2(4):951–972, 1992.
- Jaap H. ABBRING et Gerald J. Van den BERG : The unobserved heterogeneity distribution in duration analysis. *Biometrika*, 94(1):87–99, 2007.
- Theodor Adrian BALAN et Hein PUTTER : *frailtyEM : Fitting Frailty Models with the EM Algorithm*, 2018. URL <https://CRAN.R-project.org/package=frailtyEM>. R package version 0.8.8.
- W.E. BARLOW et R.L. PRENTICE : Residuals for relative risk regression. *Biometrika*, 75:65–74, 1988.
- N. BOLGER, G. DOWNEY, E. WALKER et P. STEININIGER : the onset of suicide ideation in childhood and adolescence. *Journal of Youth and Adolescence*, 18:175–189, 1989.
- N. BRESLOW et J. CROWLEY : A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship. *The Annals of Statistics*, 2(3):437–453, 1974.
- D. COMMENGES et P.K. ANDERSEN : Score test of homogeneity for survival data. *Lifetime Data Analysis*, 1(2):145–160, 1995.
- N. L. COONEY, R. M. KADDEN, M. D. LITT et H. GETTER : Matching alcoholics to coping skills or interactional therapies : Two-year follow-up results. *Journal of Consulting and Clinical Psychology*, 59:598–601, 1991.
- D.R. COX : Regression Models and Life Tables. *Journal of the Royal Statistical Society*, 34:187–220, 1972.
- D.R. COX : Partial Likelihood. *Biometrika*, 62:269–276, 1975.
- Luc DUCHATEAU et Paul JANSSEN : *The Frailty Model*. Springer Publishing Company, Incorporated, 1st édition, 2010. ISBN 144192499X, 9781441924995.
- C. ELBERS et G. RIDDER : True and spurious dependence : the identifiability of the proportional hazard model. *Review of Economic Studies*, pages 403–409, 1982.
- T. HARKANEN, H. HAUSEN, J.I. VIRTANEN et E. ARJAS : A non-parametric frailty model for temporally clustered multivariate failure times. *Scand. J. Stat.*, 30:523–533, 2003.
- J. HECKMAN et B SINGER : *Multidimension Mathematical Demography*, chapitre Population heterogeneity in demographic models, pages 567–599. Academic Press, 1982.
- J. HECKMAN et B SINGER : A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. *Econometrica*, 52(2):271–320, march 1984.

- P. HOUGAARD : Life table methods for heterogeneous populations : Distributions describing the heterogeneity. *Biometrika*, 71:75–83, 1984.
- P. HOUGAARD : Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73:387–396, 1986.
- J.D. KALBFLEISCH et R.L. PRENTICE : *The Statistical Analysis of Failure Time Data*. New York :Wiley, 1980.
- J.P. KLEIN, M. MOESCHBERGER, Y.H. LI et S.T. WANG : *Survival Analysis : State of the Art*, chapitre Estimating random effects in the Framingham Heart Study. Springer Netherlands, 1 édition, 1992.
- S.O.M. MANDA et R. MEYER : Bayesian inference for recurrent events data using time-dependent frailty. *Stat. Med.*, 24:1263–1274, 2005.
- Arthur V. PETERSON : Expressing the Kaplan-Meier Estimator as a Function of Empirical Subsurvival Functions. *Journal of the American Statistical Association*, 72(360):854–858, 1977.
- R. PETO : Discussion of Professor Cox's paper. *Journal of the Royal Statistical Society - Series B*, 34:205–207, 1972.
- J.D. SINGER : Are special educators' careers special? *Exceptional Children*, 59:262–279, 1993.
- Terry M. THERNEAU et Patricia M. GRAMBSCH : Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81:515–526, 1994.
- Terry M. THERNEAU et Patricia M. GRAMBSCH : *Modeling Survival Data : Extending the Cox Model*. Springer-Verlag New York, 2000. ISBN 978-0-387-98784-2.
- T.M. THERNEAU, P.M. GRAMBSCH et T.R. FLEMING : Martingale based residuals for survival models. *Biometrika*, 77:147–160, 1990.
- T.M. THERNEAU, P.M. GRAMBSCH et Pankratz V.S. : Penalized Survival Models and Frailty. *Journal of Computational and Graphical Statistics*, 12(1):156–175, 2003.
- J. TRUSSEL et T. RICHARDS : Correcting for unmeasured heterogeneity in hazard models using the Heckman-Singer procedure. *Sociological Methodology*, 15:242–276, 1985.
- James W. VAUPEL, Kenneth G. MANTON et Eric STALLARD : The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3), August 1979.

Série des Documents de Travail « Méthodologie Statistique »

- 9601** : Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population.
G. DECAUDIN, J.-C. LABAT
- 9602** : Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises.
N. CARON, P. RAVALET, O. SAUTORY
- 9603** : La procédure **FREQ** de **SAS** - Tests d'indépendance et mesures d'association dans un tableau de contingence.
J. CONFAIS, Y. GRELET, M. LE GUEN
- 9604** : Les principales techniques de correction de la non-réponse et les modèles associés.
N. CARON
- 9605** : L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration.
P. RAVALET
- 9606** : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (**LOGIT**, **PROBIT**).
S. LOLLIVIER, M. MARPSAT, D. VERGER
- 9607** : Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes.
N. CARON, D. LE BLANC
- 9701** : Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?
J.-C. DEVILLE
- 9702** : Modèles univariés et modèles de durée sur données individuelles.
S. LOLLIVIER
- 9703** : Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises.
N. CARON, J.-C. DEVILLE
- 9704** : La faisabilité d'une enquête auprès des ménages.
1. au mois d'août.
2. à un rythme hebdomadaire
C. LAGARENNE, C. THIESSET
- 9705** : Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine.
P. GIRARD.
- 9801** : Les logiciels de désaisonnalisation **TRAMO & SEATS** : philosophie, principes et mise en œuvre sous **SAS**.
K. ATTAL-TOUBERT, D. LADIRAY
- 9802** : Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation.
J.-C. DEVILLE
- 9803** : Pour essayer d'en finir avec l'individu Kish.
J.-C. DEVILLE
- 9804** : Une nouvelle (encore une !) méthode de tirage à probabilités inégales.
J.-C. DEVILLE
- 9805** : Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish.
J.-C. DEVILLE
- 9806** : Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel **POULPE**.
N. CARON, J.-C. DEVILLE, O. SAUTORY
- 9807** : Estimation de données régionales à l'aide de techniques d'analyse multidimensionnelle.
K. ATTAL-TOUBERT, O. SAUTORY
- 9808** : Matrices de mobilité et calcul de la précision associée.
N. CARON, C. CHAMBAZ
- 9809** : Échantillonnage et stratification : une étude empirique des gains de précision.
J. LE GUENNEC
- 9810** : Le Kish : les problèmes de réalisation du tirage et de son extrapolation.
C. BERTHIER, N. CARON, B. NEROS
- 9901** : Perte de précision liée au tirage d'un ou plusieurs individus Kish.
N. CARON
- 9902** : Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen.
N. CARON
- 0001** : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (**LOGIT**, **PROBIT**) (version actualisée).
S. LOLLIVIER, M. MARPSAT, D. VERGER
- 0002** : Modèles structurels et variables explicatives endogènes.
J.-M. ROBIN
- 0003** : L'enquête 1997-1998 sur le devenir des personnes sorties du RMI - Une présentation de son déroulement.
D. ENEAU, D. GUILLEMOT
- 0004** : Plus d'amis, plus proches ? Essai de comparaison de deux enquêtes peu comparables.
O. GODECHOT
- 0005** : Estimation dans les enquêtes répétées : application à l'Enquête Emploi en Continu.
N. CARON, P. RAVALET
- 0006** : Non-parametric approach to the cost-of-living index.
F. MAGNIEN, J. POUGNARD
- 0101** : Diverses macros **SAS** : Analyse exploratoire des données, Analyse des séries temporelles.
D. LADIRAY
- 0102** : Économétrie linéaire des panels : une introduction.
T. MAGNAC
- 0201** : Application des méthodes de calages à l'enquête EAE-Commerce.
N. CARON
- C 0201** : Comportement face au risque et à l'avenir et accumulation patrimoniale - Bilan d'une expérimentation.
L. ARRONDEL, A. MASSON, D. VERGER
- C 0202** : Enquête Méthodologique Information et Vie Quotidienne - Tome 1 : bilan du test 1, novembre 2002.
J.-A. VALLET, G. BONNET, J.-C. EMIN, J. LEVASSEUR, T. ROCHER, P. VRIGNAUD, X. D'HAULTFOEUILLE, F. MURAT, D. VERGER, P. ZAMORA
- 0203** : General principles for data editing in business surveys and how to optimise it.
P. RIVIERE
- 0301** : Les modèles logit polytomiques non ordonnés : théories et applications.
C. AFSA ESSAFI
- 0401** : Enquête sur le patrimoine des ménages - Synthèse des entretiens monographiques.
V. COHEN, C. DEMMER
- 0402** : La macro **SAS CUBE** d'échantillonnage équilibré
S. ROUSSEAU, F. TARDIEU
- 0501** : Correction de la non-réponse et calage de l'enquêtes Santé 2002
N. CARON, S. ROUSSEAU

0502 : Correction de la non-réponse par ré pondération et par imputation
N. CARON

0503 : Introduction à la pratique des indices statistiques - notes de cours
J-P BERTHIER

0601 : La difficile mesure des pratiques dans le domaine du sport et de la culture - bilan d'une opération méthodologique
C. LANDRE, D. VERGER

0801 : Rapport du groupe de réflexion sur la qualité des enquêtes auprès des ménages
D. VERGER

M2013/01 : La régression quantile en pratique
P. GIVORD, X. D'HAULTFOEUILLE

M2014/01 : La microsimulation dynamique : principes

généraux et exemples en langage R
D. BLANCHET

M2015/01 : la collecte multimode et le paradigme de l'erreur d'enquête totale
T. RAZAFINDROVONA

M2015/02 : Les méthodes de Pseudo-Panel
M. GUILLERM

M2015/03 : Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse
E. GROS, K. MOUSSALAM

M2016/01 : Le modèle Logit Théorie et application.
C. AFSA

M2016/02 : Les méthodes d'estimation de la précision de l'Enquête Emploi en Continu
E. GROS

K.MOUSSALAM

M2016/03 : Exploitation de l'enquête expérimentale Vols, violence et sécurité.
T. RAZAFINDROVONA

M2016/04 : Savoir compter, savoir coder. Bonnes pratiques du statisticien en programmation.
E. L'HOURL, R. LE SAOUT, B. ROUPPERT


M2016/05 : Les modèles multiniveaux
P. GIVORD, M. GUILLERM


M2016/06 : Econométrie spatiale : une introduction pratique
P. GIVORD, R. LE SAOUT

M2016/07 : La gestion de la confidentialité pour les données individuelles
M. BERGEAT

M2016/08 : Exploitation de l'enquête expérimentale Logement internet-papier
T. RAZAFINDROVONA

M2017/01 : Exploitation de l'enquête expérimentale Qualité de vie au travail
T. RAZAFINDROVONA

M2018/01 : Estimation avec le score de propension sous 
S. QUANTIN

M2018/02 : Modèles semi-paramétriques de survie en temps continu sous 
S. QUANTIN