Introduction à l'analyse biographique des durées

Université de Strasbourg - Master 2 Démographie

Marc Thévenin [Ined-Sms]

01/12/2022

Table of contents

List of Figures

List of Tables

1 Introduction

On dispose de données dites ***longitudinales***, et on cherche à analyser l'occurence d'un évènement au sein d'une population, conditionnellement à la durée.
Les problématiques se basent sur les questions suivantes:

- Observe-t-on la survenue de l'évènement pour l'ensemble des individus?
- Quelle est la durée jusqu'à la survenue de l'évènement?
- Quels sont les facteurs qui favorisent la survenue de cet évènement? Facteurs fixes ou facteurs pouvant apparaitre/changer au cours de la période d'observation: variables dynamiques (**TVC**: *Time Varying Covariate*)

Terminologies

Français	Anglais
Modèles de durée	Duration analysis (Econométrie)
Analyse de survie	Survival analysis (Epidémiologie, médecine, démographie)
Analyse de fiabilité	Failure time data analysis (Statistiques industrielles)
Analyse des transitions	Event-history analysis (Démographie, Sociologie)
Données de séjour	Transition analysis (Sociologie)
Histoires de vie	

Exemples d'analyse

- **Nuptialité, Mise en couple**: cohabiter, décohabiter, se marier, Rompre une union …
- **Logement**: Changement de statut (locataire <=> propriétaire), mobilité résidentielle …
- **Emploi**: Trouver un 1er emploi, changer d'emploi, entrée ou sortie du chômage …
- **Fécondité**: Avoir un premier enfant, avoir un nouvel enfant …
- **Mortalité**: Décéder après diagnostic, survivre après l'administration un traitement…

Elements nécessaire à l'analyse

1. **Un processus temporel**

- Une échelle de mesure (metrique temporelle): minutes, heures, jours, mois, années….
- Une origine définissant un évènement de départ

- Une définition précise de l'évènement d'étude.

- Une durée entre le début et la fin de la période d'observation, si nécessaire la fin de la période d'exposition.

2. **Une population soumise au risque** de connaître l'évènement (**Risk Set**)
3. ***Des variables "explicatives"*** ou ***covariables***

   - Fixes: genre, génération, niveau de diplôme, csp,……
   - Dynamiques (TVC pour *Time varying covariates*): Mesurées à tout moment entre le début et la sortie de l'observation: statut matrimonial, taille du ménage, statut d'activité…

Mini bibliographie / autres éléments de formation

Les éléments bibliographiques qui figurent ci-dessous proviennent du champ des sciences sociales/économie. Quelle que soit la langue, le nombre de cours ou documents sont très nombreux dans le domaine de la médecine.

Cours Gilbert Colletaz (Université d'Orléans). Le cours est mis à jour tous les ans. Il n'est plus possible d'accéder directement à la dernière version du document, mais il est néanmoins possible de télécharger la version 2016 (lien). Applications avec Sas.

Document de travail de Simon Quantin (Insee). Egalement un excellent document, qui couvre l'ensemble des techniques de base d'analyse des durées en temps dit continu Lien. Il propose surement la meilleure introduction en langue française à la problématique de la fragilité. Applications avec R (package `survival` version 2).

2 Données biographiques et éléments théoriques

On distingue deux types de données : les données prospectives et rétrospectives:

2.1 Données

On distingue deux types de données : les données ***prospectives*** et ***rétrospectives***.

2.1.1 Les données prospectives

- Individus suivis à des dates successives.
- Instrument de mesure identique à chaque vague (si possible).
- Avantages: qualité des données (moins de biais de mémoire).

- Inconvénients: délais pour les exploiter dans une analyse, mêmes hypothèses entre deux passages pas forcément respectées, attrition, problèmes liés aux âges d'inclusion.

A noter l'exploitation croissante des données administratives qui peuvent regorger d'informations biographiques. Déjà disponibles, le problème du coût de collecte est contourné. Ce type de données comprend par exemple les informations issues des fichiers des Ressources Humaines des entreprises, qui sont par exemples actuellement exploitées à l'Ined, par exemple dans le cadre du projet « worklife » (https://worklife.site.ined.fr/). Elles engendrent en revanche des questionnements techniques liésà l'inférence ((on ne travaille directement pas sur des échantillons).

2.1.2 Les données rétrospectives

- Individus interrogés une seule fois.
- Recueil de biographies thématiques depuis une origine jusqu'au moment de l'enquête.
- Recueil d'informations complémentaires à la date de l'enquête (âge, sexe, csp au moment de l'enquête et/ou csp représentative).
- Avantages: Information longitudinale immédiatement disponible.
- Inconvénients: questionnaire long, informations datées qui font appel à la mémoire de l'enquêté.e. A de rares exceptions (enfant, mariage), il est difficile d'aller chercher des datations trop fines avec une retrospectivité assez longue.

Les deux types de recueil peuvent être mixés avec des enquêtes à passages répétés (prospectifs) comprenant des informations retrospectives entre 2 vagues (Exemple: la cohorte Elfe de l'Ined-Inserm ou la Millenium-Cohort-Study en Grande Bretagne).

Grille AGEVEN (données rétrospectives)

Pour recueillir des informations biographiques retrospectives, on utilise généralement la méthode des grilles AGEVEN.
Il s'agit d'une grille âge-évènement, de type chronologique, avec des repères temporels en ligne (âge, année). En colonne, sont complétés de manière progressive et relative, les évènements relatifs à des domaines, par exemple la biographie professionnelle, familiale, résidentielle…

Références:

> **i** Note
>
> - Antoine P., X. Bry and P.D. Diouf, 1987 "**La fiche Ageven : un outil pour la collecte des données rétrospectives**", Statistiques Canada 13(2).
> - Vivier G, "**Comment collecter des biographies ? De la fiche Ageven aux grilles biographiques, Principes de collecte et Innovations récentes**", Acte des colloques de l'AIDELF, 2006.
> - GRAB, 1999, "**Biographies d'enquêtes : bilan de 14 collectes biographiques**", Paris, INED.
>
> Exemple grille Ageven page 121: http://retro.erudit.org/livre/aidelf/2006/001404co.pdf

2.1.3 Enregistrement des données

La question du format des fichiers biographiques mis à disposition n'est pas neutre, en particulier au niveau des manipulations pour créer le fichier d'analyse, opération qui pourra s'avérer particulièrement chronophage et complexe si plusieurs modules doivent être appariés. On distingue trois formats d'enregistrement.

Large [format individu]

Une ligne par individu, qui renseigne sur une même ligne tous les évènements liés à un domaine : les datations et les caractéristiques des évènements.

Exemple: domaine : unions - échelle temporelle: année - fin de l'observation en 1986:

id	debut1	fin1	cause1	début2	fin2	cause2
A	1979	1982	décès conjoint	1985	.	.
B	1983	1984	Séparation	.	.	.

Inconvénients: peut générer beaucoup de vecteurs colonnes avec de nombreuses valeurs manquantes. Le nombre de colonnes va dépendre du nombre maximum d'évènements. Si ce nombre concerne un seul individu, on va multiplier le nombre de colonnes pour un niveau d'information très limité. Situation classique, le nombre d'enfants, où les naissances de rang élevé deviennent de plus en plus rares.

Semi-long [format individu-évènements]

C'est le format le plus courant de mise à disposition des enquêtes biographiques. Si les transitions sont de type continu, par exemple le lieu de résidence (on habite toujours quelque part), la date de fin de la séquence correspond à la date de début de la séquence suivante. Les dates de fin ne sont pas forcément renseignées sur une ligne pour des trajectoires continues, l'information peut être donnée sur la ligne suivante avec la date de début.
Pour la séquence qui se déroule au moment de l'enquête, la date de fin est souvent une valeur manquante, une fin de séquence pouvant se produire juste avant l'enquête au cours d'une même année. Il est également possible d'avoir une information qui ne s'est pas encore produite au moment de l'enquête, mais qui aura lieu peu de temps après (personne enceinte, donc une naissance probable la même année).

Exemple précédent (trajectoires discontinues):

id	debut	fin	cause	Numero séquence
A	1979	1982	décès conjoint	1
A	1985	.	.	2
B	1983	1984	Séparation	1

Long [format individu-périodes]

Typique des recueils prospectifs. Ils engendrent des lignes sans information supplémentaire par rapport à la ligne précédente.

Exemple précédent:

id	Année	cause	Numero séquence
A	1979	.	1
A	1980	.	1
A	1981	.	1
A	1982	Décès conjoint	1
A	1985	.	2
A	1986	.	2
B	1983	.	1
B	1984	Séparation	1

Ici les trajectoires ne sont pas continues. Une forme continue présenterait toute la trajectoire, avec l'ajout d'un statut du type être en couple ou non. Pour ID=A, en 1983 et 1984, deux lignes « pas couple » (cohabitant ou non) pourraient être insérées avec au total 3 séquences.

Remarque : pour certaines analyses (par exemple analyse en temps discret), on doit transformer passer d'un format large ou semi-long à un format long, sur les durées observées ou sur des intervalles de durées construits.

2.1.4 Deux Exemples de mise à disposition

Deux enquêtes biographiques de type rétrospectives produite par l'Ined, avec un fichier donnat des informations générales sur les individus (une ligne par individu), et une série de modules biographiques en format individus-évènements (donc de type séquence).

Enquête biographie et entourage (Ined)

https://grab.site.ined.fr/fr/enquetes/france/biographie_entourage/

VIEWTABLE: TMP1.tego

	Identifiant questionnaire	prénom d'ego	sexe d'ego	Date de naissance	Département de naissance	Commune ou pays de naissance	Pays ou DOM-TOM de naissance	Numéro INSEE de la commune de naissance	Nationalité actuelle en clair
1	101	ANDREE	2	06/19/1938	93	LIVRY-GARGAN		46	FRANCAISE
2	102	JEANINE	2	06/11/1934	37	TOURS		261	FRANCAISE
3	103	MANUEL	1	08/20/1942	99	NR	PORTUGAL	99139	PORTUGAISE
4	104	LEON	1	01/13/1933	93	BONDY		10	FRANCAISE
5	105	FRANCOIS	1	12/27/1932	99	ALGER	ALGERIE	99352	FRANCAISE
6	106	EVELYNE	2	11/21/1950	99	NR	ALGERIE	99352	FRANCAISE
7	107	MICHEL	1	05/23/1949	75	PARIS-20E_ARRONDISSEMENT		120	FRANCAISE
8	108	JEANNINE	2	05/21/1948	94	PERREUX-SUR-MARNE		58	FRANCAISE
9	109	BEATRICE	2	06/09/1949	59	LOUVROIL		365	FRANCAISE
10	110	THANH-CUA	1	03/16/1941	99	TRAVINH	VIET NAM	99243	FRANCAISE
11	111	MAXIME	1	07/31/1950	77	LAGNY-SUR-MARNE		243	FRANCAISE
12	112	JACQUELINE	2	09/25/1934	54	SAINT-MAX		482	FRANCAISE
13	113	YVETTE	2	09/09/1937	19	CORNIL		61	FRANCAISE
14	114	ZOFIA	2	06/11/1935	99	EMILIOWNA	POLOGNE	99122	POLONAISE
15	115	ANTONIO	1	09/19/1932	99	SEVILLE	ESPAGNE	99134	ESPAGNOL
16	116	JEAN PIERRE	1	04/18/1930	75	PARIS-12E_ARRONDISSEMENT		112	FRANCAISE
17	117	JOSETTE	2	04/20/1939	75	PARIS-6E_ARRONDISSEMENT		106	FRANCAISE
18	118	RADA	2	12/18/1945	99	ZAGREB	YOUGOSLAVIE	99121	CROATE
19	119	JACQUELINE	2	03/23/1933	92	CLICHY		24	FRANCAISE
20	120	CLAUDE	1	09/11/1942	83	TOULON		137	FRANCAISE
21	121	MARIE-NOELLE	2	07/06/1944	21	SEMUR-EN-AUXOIS		603	FRANCAISE
22	122	ROGER	1	12/03/1935	62	ESQUERDES		359	FRANCAISE
23	123	DANIEL	1	06/12/1948	75	PARIS-14E_ARRONDISSEMENT		114	FRANCAISE
24	124	JEAN-CLAUDE	1	08/31/1936	92	NEUILLY-SUR-SEINE		51	FRANCAISE
25	125	GHISLAINE	2	01/20/1944	60	BRETEUIL		104	FRANCAISE
26	126	JOCELYNE	2	06/28/1949	28	BOULLAY-LES-DEUX-EGLISES		53	FRANCAISE
27	127	MARIE-JOSE	2	10/31/1949	76	MONT-SAINT-AIGNAN		451	FRANCAISE

Figure 2.1: Caractéristiques

	Identifiant questionnaire	Age en début de période	Code des événements familiaux	Etape	Département	Liste de communes ou pays ou DOM-TOM	INSEEL3	Type de logement (appartement, maison, ...)	Nombre de pièces dans le logement	Confort sanitaire	Détenteur du statut
1	101	0		1	93	LIVRY-GARGAN	46	21	3	1	P M
2	101	18	M1	2	93	LIVRY-GARGAN	46	22	3	0	2
3	101	23		2M	93	LIVRY-GARGAN	46	22	3	4	2
4	101	49	DCC1	2M	93	LIVRY-GARGAN	46	22	3	4	1
5	102	0		1	37	TOURS	261	12	99	99	P M
6	102	5		2	37	TOURS	261	22	4	1	P M
7	102	7		3T	.	.	.	.	.	.	
8	102	7		3	37	TOURS	261	12	99	1	P M
9	102	10	NF3	4	75	PARIS-18E_ARRONDISSEMENT	118	41	2	0	P M
10	102	22	M1	5	93	BOBIGNY	8	22	1	1	1 2
11	102	26		6	93	BOBIGNY	8	21	4	4	1 2
12	102	37		7	93	LIVRY-GARGAN	46	21	3	4	1 2
13	103	0		1	99	PORTUGAL	99139	22	2	0	P M
14	103	20		2T	.	.	.	.	.	.	
15	103	20		2	92	NANTERRE	50	43	1	88	1
16	103	22		3	93	DRANCY	29	43	1	88	1
17	103	24	M1	4	93	LIVRY-GARGAN	46	22	2	2	1
18	103	27		5	93	LIVRY-GARGAN	46	21	3	4	1 2

Figure 2.2: Séquences logement

Figure 2.3: Bases Biographie et Entourage

Enquête MAFE (Ined)

https://mafeproject.site.ined.fr/

ident	q1	q1a	statu_mig	year	age_survey
E1	Man	1972	Migrant	2008	37
E10	Man	1966	Migrant	2008	43
E100	Man	1972	Migrant	2008	37
E101	Woman	1977	Migrant	2008	32
E102	Woman	1966	Migrant	2008	43
E103	Woman	1978	Migrant	2008	31
E104	Woman	1958	Migrant	2008	51
E105	Man	1968	Migrant	2008	41
E106	Man	1961	Migrant	2008	48
E107	Woman	1965	Migrant	2008	44
E108	Man	1972	Migrant	2008	37
E109	Woman	1966	Migrant	2008	43
E11	Man	1979	Migrant	2008	30
E110	Man	1966	Migrant	2008	43
E111	Woman	1983	Migrant	2008	26
E112	Man	1972	Migrant	2008	37
E113	Man	1977	Migrant	2008	32
E114	Man	1964	Migrant	2008	45
E115	Woman	1983	Migrant	2008	26
E116	Man	1951	Migrant	2008	58
E117	Man	1963	Migrant	2008	46
E118	Woman	1965	Migrant	2008	44
E119	Woman	1968	Migrant	2008	41
E12	Woman	1977	Migrant	2008	32
E120	Woman	1973	Migrant	2008	36

Figure 2.4: Caractéristiques

ident	num_log	q301d	q301f	q302	q303	age_survey	q1a
E1	1	1972	1975	SENEGAL	Namanleque	37	1972
E1	2	1975	2001	SENEGAL	Madina Aly	37	1972
E1	3	2001	2007	SPAIN	Santa Maria De Palautordera	37	1972
E1	4	2007	.	SPAIN	Santa Maria De Palautordera	37	1972
E10	1	1966	1996	SENEGAL	Anambe	43	1966
E10	2	1996	1997	SPAIN	Pineda De Mar	43	1966
E10	3	1997	1999	SPAIN	Granollers	43	1966
E10	4	1999	2006	SPAIN	Figueres	43	1966
E10	5	2006	.	SPAIN	Figueres	43	1966
E100	1	1972	2004	SENEGAL	Dakar	37	1972
E100	2	2004	2007	SENEGAL	Fass / Colobane / Gueule Tapee	37	1972
E100	3	2007	.	SPAIN	Murcia	37	1972
E101	1	1977	1997	SENEGAL	Mandegane	32	1977
E101	2	1997	2006	SENEGAL	Dakar	32	1977
E101	3	2006	2007	SPAIN	Rubi	32	1977
E101	4	2007	.	SPAIN	Rubi	32	1977
E102	1	1966	2005	SENEGAL	Bignona	43	1966
E102	2	2005	.	SPAIN	Mataro	43	1966
E103	1	1978	1992	SENEGAL	Medina Yero	31	1978
E103	2	1992	1995	SPAIN	Calella	31	1978
E103	3	1995	1997	SENEGAL	Medina Yero	31	1978
E103	4	1997	.	SPAIN	Barcelona	31	1978
E104	1	1958	2004	SENEGAL	Dakar	51	1958
E104	2	2004	2007	SPAIN	Salou	51	1958
E104	3	2007	.	SPAIN	Salou	51	1958

Figure 2.5: Séquences logement

Figure 2.6: Bases MAFE (Sénégal)

2.2 Eléments de théorie

L'analyse des durées peut être vue comme l'étude d'une variable aléatoire T qui décrit la durée d'attente jusqu'à l'occurence d'un évènement.

- La durée $T = 0$ est le début de l'exposition au risque (entrée dans le **Risk set**).
- T est une mesure non négative de la durée.

La principale caractéristique de l'analyse des durées est le traitement des informations dites **censurées**, lorsque la **durée d'observation est plus courte que la durée d'exposition au risque**.

2.2.1 Temps, durée et Risk set

Temps et durée

Le temps est une dimension (la quatrième), la durée est sa mesure. La durée est tout simplement calculée par la différence, pour une échelle temporelle donnée, entre la fin et le début d'une période d'exposition ou d'observation.
On distingue généralement deux types de mesure de la durée : ***continue*** et ***discrète/groupée***. Ces deux notions ne possèdent pas réellement de définition, la différence s'explique plutôt par la présence ou non de simultanéité dans l'occurrence des évènements. Le temps étant intrasèquement continu car deux évènements ne peuvent pas avoir lieu en « même temps ». C'est donc l'échelle temporelle choisie ou imposée par l'analyse et les données qui pourra rendre cette mesure continue ou discrète/groupée.
Pour un physicien travaillant sur la théorie de la relativité avec des horloges atomiques, une minute

(voire une seconde) est une mesure très groupée pour ne pas dire grossière du temps, alors que pour un géologue c'est une mesure continue. Pour ces deux disciplines, cette échelle de mesure n'est pas adaptée à leur domaine. Le choix de l'échelle temporelle doit être pertinent par rapport aux objectifs de l'analyse même si on dispose des informations très fines (dates de naissance exactes). Etudier la fécondité avec une métrique journalière n'aurait pas de sens.

Il existe des situations où les durées sont par nature discrète, lorsqu'un évènement ne peut avoir lieu qu'à un moment précis (date d'anniversaire des contrats pour l'analyse des résiliation). Généralement dans les sciences sociales avec un recueil de données de type rétrospectif, les mesures dites discrètes sont plutôt de nature groupées. Pour une même année, on considèrera indifféremment des évènements qui se produiront un premier janvier et un 31 décembre d'une même année.

> ! Important
>
> - **Durée continue : absence (ou très peu) d'évènements simultanés**
> - **Durée discrète/groupée : présence d'évènements simultanés (en grand nombre)**

Le Risk Set

- Il s'agit de la population "soumise" ou "exposée" au risque lorsque $T = t_i$.
- Cette population varie dans le temps car:
  - Certaines personnes ont connu l'évènement, donc ne peuvent plus être soumises au risque (ex: décès si on analyse la mortalité).
  - Certaines personnes sortent de l'observation sans avoir (encore) observé l'évènement: décès si on analyse un autre type d'évènement, perdu.e.s de vue, fin de l'observation dans un recueil rétrospectif.

2.2.2 La Censure

> ! Important
>
> **Une observation est dite censurée lorsque la durée d'observation est inférieure à la durée d'exposition au risque**

2.2.2.1 Censure à droite

Définition
Certains individus n'auront pas (encore) connu l'évènement à la date de l'enquête après une certaine durée d'exposition. On a donc besoin d'un marqueur permettant de déterminer que les individus n'ont pas observé l'évènement sur la période d'étude.

Pourquoi une information est-elle censurée (à droite)?

- Fin de l'étude, date de l'enquête.
- Perdu de vue, décès si autre évènement étudié.

En pratique (important)

- **Ne pas exclure ces observations**, sinon on surestime la survenue de l'évènement.
- **Ne pas les considérer a-priori comme sorties de l'exposition sans avoir connu l'évènement**. Elles peuvent connaître l'évènement après la date de l'enquête ou en étant perdues de vue. Sinon on sous-estime la durée moyenne de survenue de l'évènement.

Exemple
On effectue une enquête auprès de femmes : On souhaite mesurer l'âge à la première naissance. Au moment de l'enquête, une femme est âgée de 29 ans n'a pas (encore) d'enfant.
Cette information sera dite «censurée».
Elle est clairement encore soumise au risque après la date de l'enquête. Au niveau de l'analyse, elle sera soumise au risque à partir de ses premières règles jusqu'au moment de l'enquête.

Hypothèse fondamentale
Les observations censurées ont vis à vis du phénomène observé le même comportement que les observations non censurées. On dit que la **censure est non informative**. Elle ne dépend pas de l'évènement analysé. Normalement le problème ne se pose pas dans les recueil retrospectif.

Problème posé par la censure informative
Par exemple en analysant des décès avec un recueil prospectif, si un individu est perdu de vue en raison d'une dégradation de son état de santé, l'indépendance entre la cause de la censure et le décès ne peut plus être assurée.
A l'Ined l'exploitation du registre des personnes atteintes de mucoviscidose (*Gil Bellis*) donne une illustration de ce phénomène. Chaque année un nombre significatif de personnes sortent du registre (pas de résultats aux examens annuels). Si certain.e.s perdu.e.s de vue s'expliquent par des déménagements, émigration ou par un simple problème d'enregistrement des informations, ces personnes sont nombreu.se.s à présenter une forme « légère » de la maladie. L'information pouvant être donnée ici par la mutation du gène. Comme il n'est pas recommandé de supprimer ou de traiter ces observations comme des censures à droite non informatives, on peut les appréhender comme un risque concurrent au décès ou à tout autre évènement analysé à partir de ce registre (section formation interne Ined).

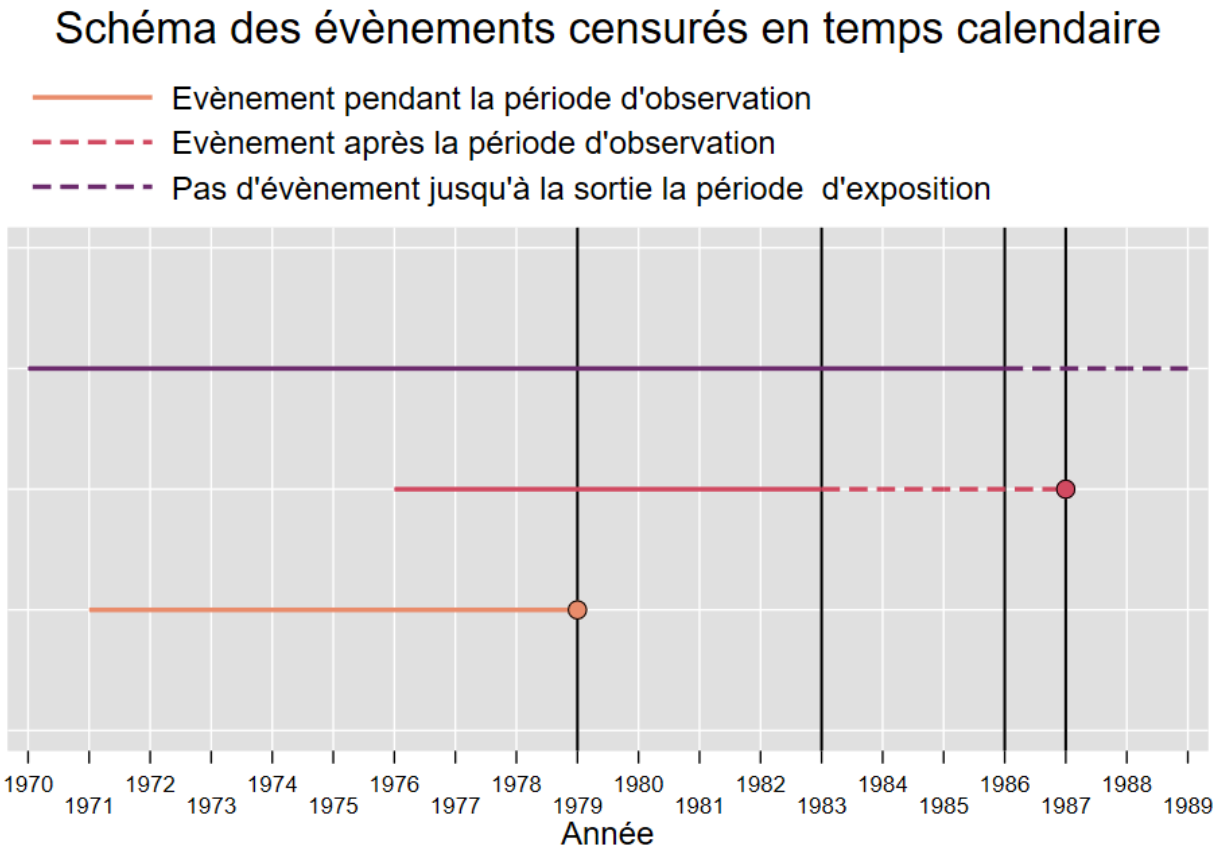

Figure 2.7: Censure - Temps calendaire

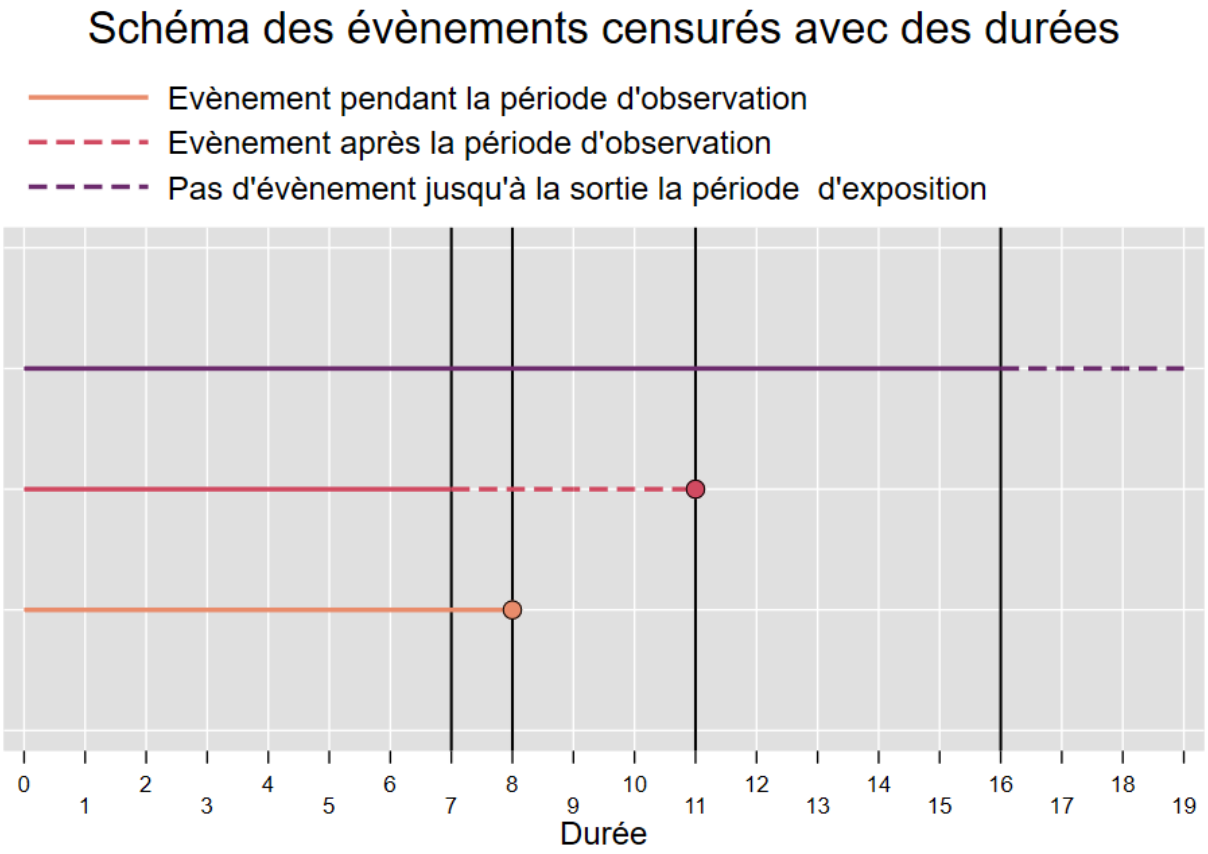

Figure 2.8: Censure - Durée

2.2.2.2 Censure à gauche, troncature et censure par intervalle

Censure à gauche
L'évènement s'est produit avant le début période d'observation. Typique des données prospectives, de type registre, avec des âges d'inclusion différenciés.

Censure par intervalle
Un évènement peut se produire entre 2 temps d'observations sans qu'on puisse les observer (ex: en criminologie récidive d'un delit entre deux arrestations). Un phénomène de censure à droite avec perdu.e de vue peut se transformer en censure par intervalle lorsque la personne réapparait et est de nouveau incluse dans les données.

Troncature
Par l'exemple, on analyse la survie d'une population. Seule la survie des individus vivants à l'inclusion peut être analysée (*troncature à gauche*). On peut également trouver un phénomène de troncature lorsqu'on mesure la durée à partir ou jusqu'à un certain seuil niveau.

Ces situations sont généralement plutôt bien contrôlées dans les recueils rétrospectifs. Elles sont assez courantes lorsque le recueil est de type prospectif.

Durée d'observation supérieure à la durée d'exposition
A l'inverse des individus peuvent sortir de l'exposition avant la fin de la période d'observation, et il convient donc de corriger la durée de cette sortie.
Un exemple simple : si au moment de l'enquête une femme sans enfant a 70 ans, cela n'a pas de sens de continuer de l'exposer au risque au-delà d'un certain âge. Si on ne dispose pas d'information sur l'âge à la ménopause on peut tronquer la durée un peu au-delà de l'âge le plus élevé à la première naissance observée dans les données.

2.2.3 Les grandeurs

- La fonction de survie $S(t)$ et la fonction de répartition $F(t)$
- La fonction de densité $f(t)$
- Le risque ("hazard") "instantané" $h(t)$
- Le risque "instantané" cumulé $H(t)$

Remarques:

- Toutes ces grandeurs sont mathématiquement liées les unes par rapport aux autres. En connaître une permet d'obtenir les autres.

- Au niveau formel on se placera ici du point de vue où la durée mesurée est strictement continue. Cela se traduit, entre autre, par l'absence d'évènements dits "simultanés".

- Les expressions qui vont suivre ne sont ne donnent de techniques de calcul, mais des grandeurs dont on précisera seulement les propriétés. Les estimateurs, en particuliers ceux des fonctions de survie/séjour sont données dans la partie **méthode non paramétrique**.

2.2.3.1 La fonction de Survie $S(t)$

Dans ce type d'analyse, il est courant d'analyser la courbe de survie (ou de séjour).

La fonction de survie donne la proportion de la population qui n'a pas encore connue l'évènement après une certaine durée t. Elle y a "survécu".

Formellement, la fonction de survie est la probabilité de survivre au-delà de t, soit:

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

Propriétés: $S(0) = 1$ et $\lim_{t \to \infty} S(t) = 0$
La fonction de survie est strictement non croissante.

2.2.3.2 La fonction de répartition $F(t)$

C'est la probabilité de connaitre l'évènement jusqu'en t, soit:

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

$F(t) = 1 - S(t)$

La fonction de survie et la fonction de répartition sont deux grandeurs strictement complémentaires.
Propriétés: $F(0) = 0$ et $\lim_{t \to \infty} F(t) = 1$

2.2.3.3 La fonction de densité $f(t)$

- Pour une valeur de t donnée, la fonction de densité de l'évènement donne la distribution des moments où les évènement ont eu lieu. Elle est donnée dans un premier temps par la probabilité de connaitre l'évènement dans un petit intervalle de temps dt. Si dt est proche de 0 (temps continu) alors cette probabilité tend également vers 0. On norme donc cette probabilité par dt. Rappel: on est toujours ici dans la théorie.

- En temps continu, la fonction de densité est donnée par la dérivée de la fonction de répartition: $f(t) = F'(t) = -S'(t)$. Formellement la fonction de densité $f(t)$ s'écrit:

$$f(t) = \frac{P(t \leq T < t + dt)}{dt}$$

2.2.3.4 Le risque instantané $h(t)$

Concept fondamental de l'analyse des durées:

$$h(t) = \frac{P(t \leq T < t + dt | T \geq t)}{dt}$$

- $P(t \leq T < t + dt | T \geq t)$ donne la probabilité de survenue de l'évènement sur l'intervalle $[t, t + dt[$ *conditionnellement à la survie au temps* t.

- La quantité obtenue donne alors un nombre moyen d'évènements que connaîtrait un individu durant une unité de temps choisie.
- A priori cette quantité n'est pas une probabilité. C'est la nature de l'évènement, en particulier sa non récurrence, et la métrique temporelle choisie ou disponible qui peut la rendre assimilable à une probabilité. Tout comme la densité, on est plutôt dans la définition d'un *taux* (d'où l'expression ***hazard rate*** en anglais).

On peut également écrire: $h(t) = \frac{f(t)}{S(t)} = \frac{F'(t)}{S(t)} = -\frac{S'(t)}{S(t)}$

On voit ici clairement que la fonction de risque n'est pas une probabilité : $\frac{f(t)}{S(t)}$ ne peut pas contraindre la valeur à ne pas être supérieure à 1.

> Info
>
> Formellement pour $f(t)$ et $h(t)$, les expressions sont la limite lorsque dt tend vers 0 des deux quantités présentées ci-dessus. Son omission est volontaire de ma part en raison de l'application qui est donnée un peu plus loin qui, justement, regarde l'effet de l'élargissement des intervalles de durée sur la mesure du taux de risque.

2.2.3.5 Le risque cumulé $H(t)$

Le risque cumulé est égal à : $H(t) = \int_0^t h(u)du = -log(S(t))$
On peut alors le réécrire toutes les autres quantités:

- $S(t) = e^{-H(t)}$
- $F(t) = 1 - e^{-H(t)}$
- $f(t) = h(t) \times e^{-H(t)}$

Exemple avec la loi exponentiel (risque constant)
Si on pose que le risque est strictement constant au cours du temps: $h(t) = a$ (on parle de **loi exponentielle** - cf partie sur les modèles AFT - typique des processus sans mémoire comme la durée de vie des ampoules):

- $h(t) = a$
- $H(t) = a \times t$
- $S(t) = e^{-a \times t}$
- $F(t) = 1 - e^{-a \times t}$
- $f(t) = a \times e^{-a \times t}$

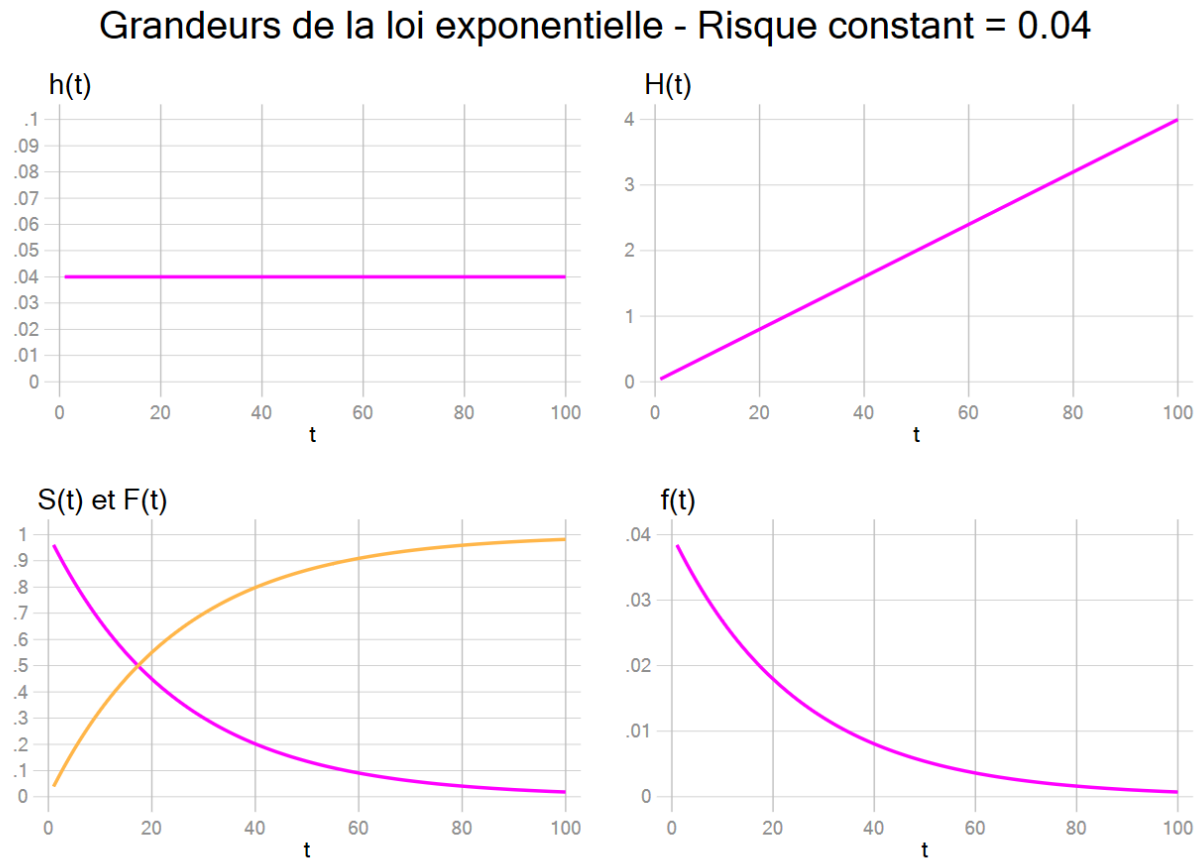

Figure 2.9: Grandeurs de la loi exponentielle

2.2.3.6 Application: risque et échelles temporelles

Fortement inspiré, pour ne pas dire copié, de l'excellent cours de Gilbert Colletaz.
Attention on sort ici très clairement du temps continu, il s'agit seulement de manipuler les concepts, et de voir la dépendance de la mesure du risque à l'échelle temporelle choisie/disponible.

1. Durant les mois d'hiver, entre le 1er janvier et le 1er avril (3 mois), la probabilité d'attraper un rhume chaque mois est de 0.48 (il s'agit bien d'un risque). Quelle est le risque d'attraper le rhume durant la saison froide?
   $\frac{0.48}{1/3} = 1.44$. On peut donc s'attendre a attraper 1.44 rhume durant la période d'hiver.
2. On passe une année en vacances dans une région où la probabilité de décéder chaque mois est évaluée à 0.33. Quelle est le risque de décéder pendant cette année sabbatique? $\frac{0.33}{1/12} = 3.96$

Le risque peut donc être supérieur à 1 (c'est donc plutôt un taux tel qu'on le défini généralement). En soit cela ne pose pas de problème comme il s'agit d'un nombre moyen d'évènements espérés (exemple: "taux" de fécondité), mais pour des évènements qui ne peuvent pas se répéter, évènements dits "absorbants", l'interprétation n'est pas très intuitive.

On peut donc prendre l'inverse du risque qui mesure la durée moyenne (espérée) jusqu'à l'occurence de l'évènement.
On retrouve donc un concept classique en analyse démographique comme l'espérance de vie (survie): la question n'est pas de savoir si "on" va mourir ou non, le risque indépendement du temps étant par définition égal à 1, mais jusqu'à quand on peut espérer survivre.

- Pour le rhume, la durée moyenne est de $1/1.44 = 0.69$ du trimestre hivernal, soit approximativement le début du mois de mars.
- Pour l'année sabbatique, la durée moyenne de survie (l'espérance de vie) est de $1/3.96 = 0.25$ d'une année soit 3 mois après l'arrivée dans la région.

Exercice

- On a une population de 100 cochons d'Inde.
- On analyse leur mortalité (naturelle).

- Ici l'analyse est en temps discret.
- La durée représente le nombre d'année de vie.
- Il n'y a pas de censure à droite.

Durée	Nombre de décès
1	1
2	1
3	3
4	9
5	30
6	40
7	10
8	3
9	2
10	1

N=100

A quel âge le risque de mourir des cochons d'Inde est-il le plus élevé? Quelle est la valeur de ce risque?

2.2.3.7 Compléments

Forme des fonctions de survie

Une des propriétés de la fonction de survie ou de séjour est qu'elles tendent vers 0. A la lecture du graphique suivant, cela peut correspondre à la forme de la courbe S2, bien que le % de survivant tend à baisser de moins en moins à mesure que la durée augmente. Deux cas limites doivent être considéré.

- **S1**: très peu d'évènements et la fonction de séjour suit une asymptote nettement supérieur à 0 ($\lim_{t \to \infty} S(t) = a$ avec $a > 0$). La question est plus délicate car on interroge l'exposition au risque d'une partie de l'échantillon ou, dit autrement on peut penser qu'une fraction est immunisé au risque. Cette problématique est rapidement posée en fin de formation.
- **S2**: la situation attendue
- **S3**: La survie tombe à 0 très/trop rapidement : il n'y a donc pas ou presque pas de durée (par exemple presque tout l'échantillon observe l'évènement la première année de l'exposition). Les méthodes en temps continue ne sont a priori pas adaptées à ce genre de situation. Si on dispose d'une information plus fine pour dater les évènements, la fonction de séjour pourra reprendre une

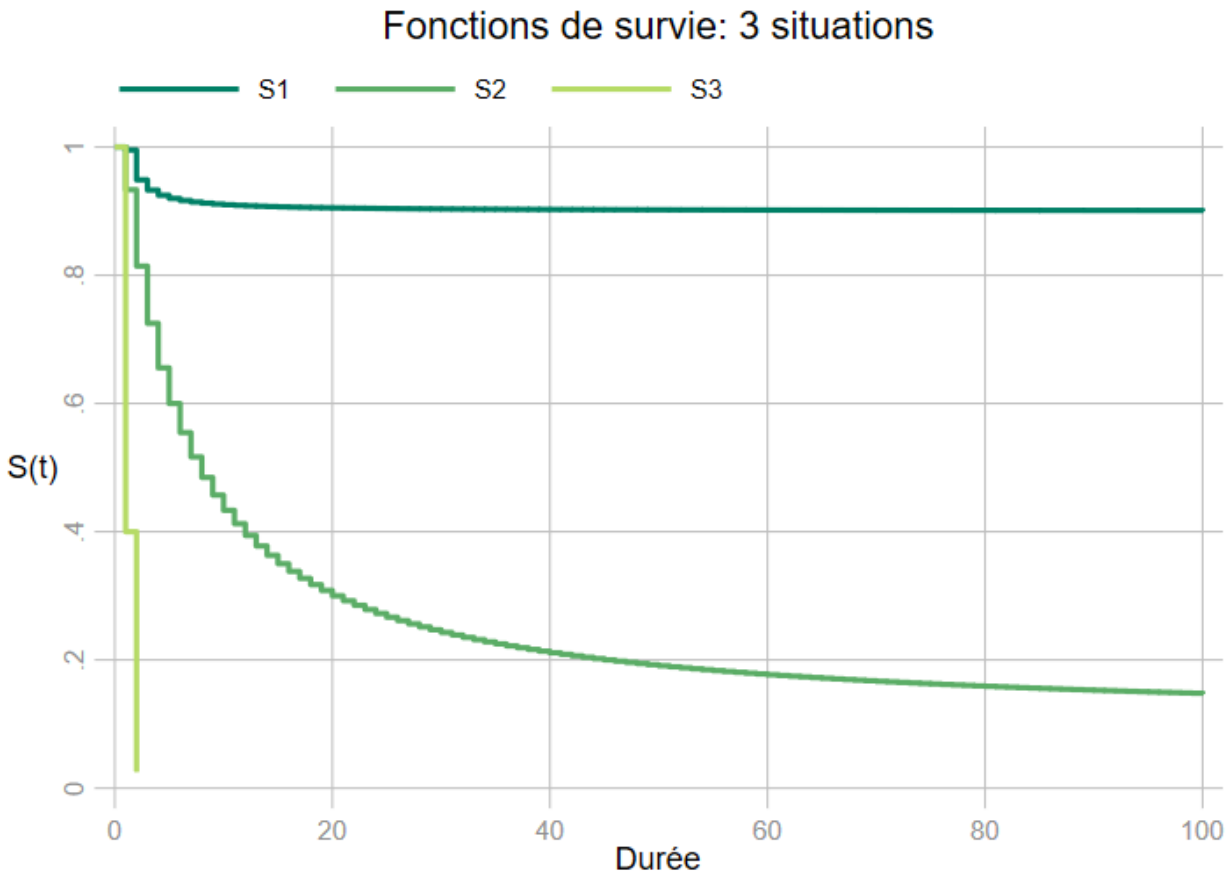

Figure 2.10: $S(t)$: 3 situations types

forme plus “standard”. Dans le graphique, $S(t = 1) = 0.4$, $S(t = 2) = 0.025$, mais si on dispose par exemple de 10 points d’observations supplémentaires dans chaque durée groupée:

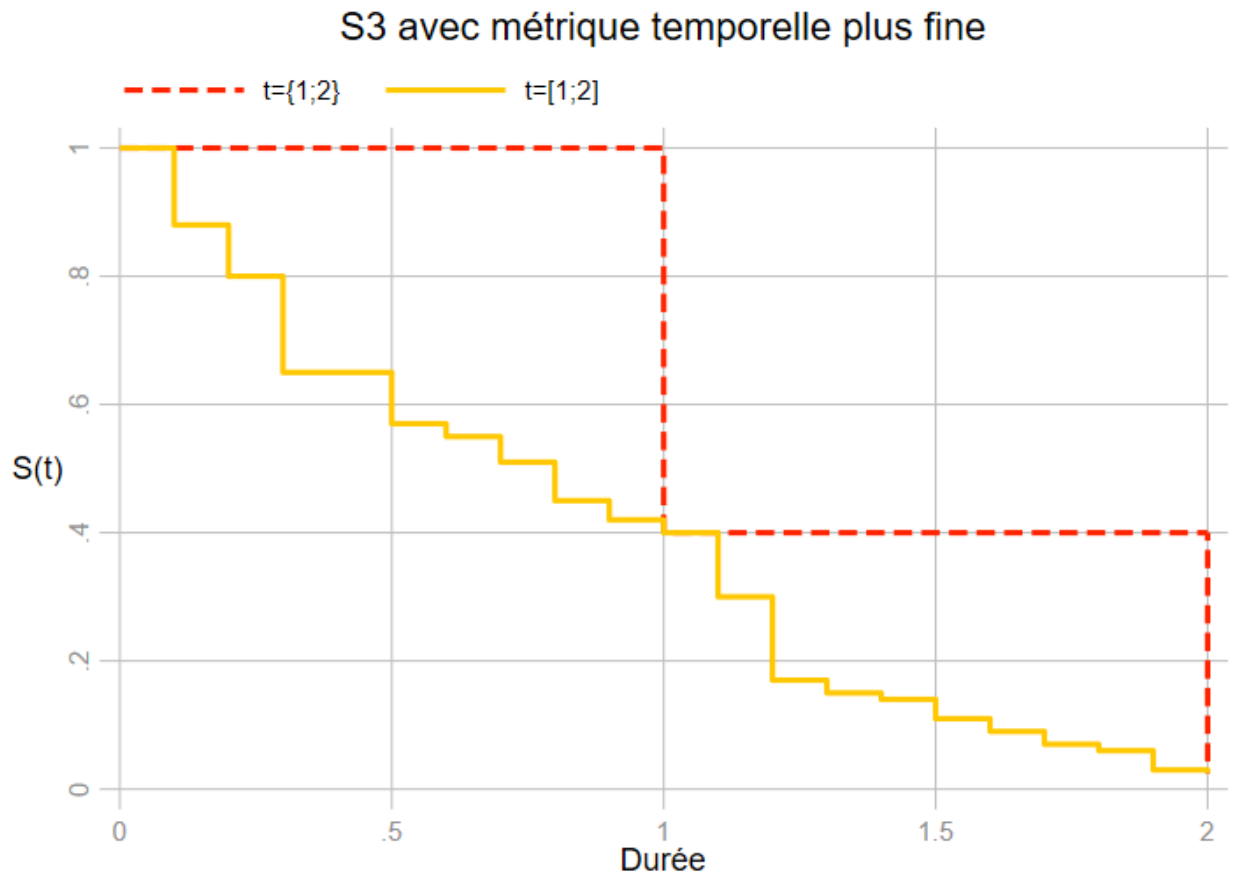

Figure 2.11: $S(t)$: changement de l’échelle temporelle

Absence de censures à droites

Les méthodes qui vont être présentées ***gèrent*** la présence de censures à droite. En leur absence, elle restent néanmoins parfaitement valables. L'absence de censure facilite certaines analyses, par exemple celles des fonctions de séjour où le calcul direct des durées moyennes est rendu possible.

Utilisation des pondérations

Une question assez récurrente concerne l'utilisation des poids de sondage dans les analyses de durées avec longueurs biographiques souvent assez longues. Appartenant à l'*école du bon sens* (Eva Levièvre), leur utilisation ne me semble pas recommandée voire à exclure sauf exceptions. En effet les pondérations sont générées au moment de l'enquête, alors que les évènements étudiés peuvent remonter dans un passé plus ou moins lointain pour une partie de la population analysée. Si on regarde de plus près, la création de poids longitudinaux ne résoudrait pas grand chose ,les pondérations devant être recalculées à chaque moment d'observation ou à chaque moment où des évènements se produisent. Par ailleurs on mélangerait à un instant donné des personnes issues de générations différentes ce qui rend impossible tout calage sur des caractéristiques d'un population. Supposons une personne âgée de 25 ans et un personne âgée de 70 ans au moment de l'enquête en 2022, avec un début d'observation à l'âge de 18 ans . A 20 ans ($t = 2$), pour la première personne les caractéristiques de la population sont celles de 2017, pour celle de 70 ans celles de 1972. On fait comment??????

3 Analyse non paramétrique

Les méthodes non paramétriques portent généralement sur l'analyse des fonctions de survie ($S(t)$) ou sur celle des fonctions de répartitions ($F(t)$), plus rarement sur les mesures d'incidence données par le risque cumulé. Deux méthodes d'estimations sont proposées : la méthode dite actuarielle et la méthode dite de Kaplan & Meier. Ces deux approches sont adaptées à des mesures différentes de la durée : plutôt discrèt/groupées pour la technique actuarielle et plutôt continue pour Kaplan-Meier (KM). Cela induit un traitement différent de la censure dans l'estimation. La seconde est de très très loin la plus diffusée, surement en raison des tests de comparaison qu'elle propose.

3.1 Les fonctions de survie/séjour

3.1.1 Les variables d'analyse

On a un échantillon aléatoire de n individus avec:

- Des indicateurs de fin d'épisode $e_1, e_2, ..., e_k$ avec $e_i = 0$ si censure à droite et $e_i = 1$ si évènement observé pendant la période d'observation.
- Des durées d'exposition au risque $t_1, t_2, ..., t_k$ jusqu'à l'évènement ou la censure.
- En théorie, il ne peut pas y avoir d'évènement en $t = 0$.

3.1.2 Calcul de la fonction de survie

Rappel:
La fonction de survie donne la probabilité que l'évènement survienne après t_i, soit $S(t_i) = P(T > t_i)$. Pour survivre en t_i, il faut avoir survécu en t_{i-1}, t_{i-2}, ..., t_1.
La fonction de survie rapporte donc des probabilités conditionnelles: survivre en t_i conditionnellement au fait d'y avoir survécu avant. Il s'agit donc d'un produit de probabilités:

Soit $d_i = \sum e_i$ le nombre d'évènements observé en t_i et r_i la population encore soumise au risque en i. On peut mesurer l'intensité de l'évènement en t_i en calculant le quotient $q(t_i) = \frac{d_i}{r_i}$. Si le temps est strictement continu on devrait toujours avoir $q(t_i) = \frac{1}{r_i}$.

$S(t_i) = (1 - \frac{d_i}{r_i}) \times S(t_{i-1}) = S(t_i) = (1 - q(t_i)) \times S(t_{i-1})$.
En remplaçant $S(t_{i-1})$ par sa valeur:

$S(t_i) = (1 - \frac{d_i}{r_i}) \times (1 - \frac{d_{i-1}}{r_{i-1}}) \times S(t_{i-2})$.
En remplaçant toutes les expressions de la survie jusqu'en t_0 ($S(0) = 1$):

$$S(t_i) = \prod_{t_i \leq k} (1 - q(t_i))$$

i Application pour la suite du cours

- On va analyser le risque de décéder (la survie) de personnes souffrant d'une insuffisance cardiaque. Le début de l'exposition est leur inscription dans un registre d'attente pour une greffe du coeur.
- Les covariables sont dans un premier temps toutes fixes: l'année (*year*) et l'âge (*age*) à l'entrée dans le registre, et le fait d'avoir été opéré pour un pontage aorto-coronarien avant l'inscription (*surgery*). Le début de l'exposition au risque est l'entrée dans le registre, la durée est mesurée en jour (*stime*). La variable évènement/censure est le décès (*died*).
- L'introduction d'une dimension dynamique, la greffe, est donnée par les informations contenues dans les variables *transplant* et *wait*.

Extrait de la base:

id	year	age	died	stime	surgery	transplant	wait
15	68	53	1	1	0	0	0
43	70	43	1	2	0	0	0
61	71	52	1	2	0	0	0
75	72	52	1	2	0	0	0
102	74	40	0	11	0	0	0
74	72	29	1	17	0	1	5

3.1.3 La méthode actuarielle

- Estimation sur des intervalles définies par l'analyste.
- Méthode dite «continue», estimation en milieu d'intervalle.
- Méthode apropriée lorsque la durée est mesurée de manière discrète/groupée.
- Méthode, hélas, quasiment abandonnée dans les sciences sociales même si les durées sont plus rarement mesurées de manière exacte. L'absence de test de comparaison des fonctions de survie n'y est pas étranger.

La durée est divisée en J intervalles, en choisissant J points: $t_0 < t_1 < ... < t_J$ avec $t_{J+1} = \infty$.

Calcul du Risk set

- A $t_{min} = 0$, $n_0 = n$ individus soumis au risque: $r_0 = n_0$.
- Le nombre d'exposé.e.s au risque sur un intervalle est calculé en soustrayant la moitié des cas censurés sur la longueur de l'intervalle: $r_i = n_i - 0.5 \times c_i$, avec n_i le nombre de personnes soumises au risque au début de l'intervalle et c_i le nombre d'observations censurées sur la longueur de l'intervalle. On suppose donc que les observations censurées c_i sont sorties de l'observation uniformément sur l'intervalle. Les cas censurés le sont en moyenne au millieu de l'intervalle.

Calcul de $S(t_i)$
On applique la méthode générique de la section précédente avec:

$$q(t_i) = \frac{d_i}{n_i - 0.5 \times c_i}$$

Calcul de la durée médiane (ou autre quantiles)

Rappel: en raison de la présence de censures à droite, le dernier intervalle étant ouvert jusqu'à la dernière sortie d'observation, il n'est pas du tout conseillé de calculer des durées moyennes. On préfère utiliser la médiane ou tout autre quantile lorsqu'ils sont calculables, c'est à dire jusqu'à la limite de la dernière valeur estimé de $S(t)$

Définition: il s'agit de la durée telle que $S(t_i) = 0.5$.

Calcul: Comme on applique une méthode continue et monotone à l'intérieur d'intervalles, on ne peut pas calculer directement un point de coupure qui correspond à 50% de survivants. On doit donc trouver ce point par interpolation linéaire dans l'intervalle $[t_i; t_{i+1}[$ avec $S(t_{i+1}) \leq 0.5$ et $S(t_i) > 0.5$.

Avec R

Les fonctions de survie avec la méthode dite actuarielle sont estimables avec le package **`discSurv`**. Avec le temps, il s'est étoffé, on peut maintenant paramatrer des intervalles, mais les quantiles de la durées ne sont toujours pas estimables, ce qui est fort dommage.

Avec des intervalles de 10 jours

Figure 3.1: $S(t)$: méthode actuarielle

Ce graphique a été produit avec R (`discSurv`). En abscisse sont reporté les bornes des intervalles. Comme la longueur des intervalles est égale à 10, il est facile de convertir en jours. Cela ne serait pas le cas avec des intervalles de 7 ou 30 jours par exemple.

Table 3.2: Quantiles de la durée (Stata avec définition des bornes de Sas)

$S(t)$	t
0.90	8
0.75	36
0.50	102
0.25	914
0.10	.

102 jours après leur inscription dans le registre d'attente pour une greffe, 50% des malades sont toujours en vie. Au bout de 914 jours, 75% des personnes sont décédées (en prenant une lecture par la fonction de répartition).

3.1.4 La méthode Kaplan-Meier

- L'approche qui exploite toute l'information disponible est celle dite de **Kaplan-Meier** (KM).
- Il y a autant d'intervalles que de durées où l'on observe au moins un évènement.
- Au lieu d'utiliser des intervalles prédéterminés, l'estimateur KM va définir un intervalle entre chaque évènement enregistré.
- La fonction de survie estimée par la méthode KM est une fonction en escalier (*stairstep*), d'où une méthode dite "discrète".
- Pour chaque intervalle, on compte le nombe d'évènements et le nombre de censures.
- Méthode adaptée pour une mesure de la durée de type continue.

Définition du Risk Set (r_i)

S'il y a à la fois des évènements et des censures à une durée t_i, les observations censurées sont considérées comme exposées au risque à ce moment, comme si elles étaient censurées très rapidement après. C'est la principale caractéristique de cette méthode, appelé également l'estimateur « product-limit »

$$r_i = r_{i-1} - d_{i-1} - c_{i-1}$$

Calcul de q_i
On applique la méthode de la section précédente avec:

$$q_i = \frac{d_i}{r_{i-1} - d_{i-1} - c_{i-1}}$$

Remarque: la variance de l'estimateur est obtenu par la méthode dite de "Greenwood"". Il n'y a pas d'intérêt particulier pour ce cours de la décrire.

"Récupération" de la médiane
Il n'y a pas de méthode pour calculer directement la durée médiane (ou tout autre quantile).
On va prendre la valeur de la durée qui se situe juste "en dessous" de 50% de survivant.e.s. Elle est donc définie tel que $S(t) \leq 0.5$. donc pas de formule savante pour obtenir ce résultat, c'est une convention. Attention, il n'est pas impossible que le % de survivant.e.s soit bien en deçà de 50% pour l'obtention cette durée médiane.

Avec R

Les estimateurs sont obtenus avec fonction **`survfit`** de la librairie `survival`. On peut obtenir des rendus graphiques supérieurs avec la librairie `survminer` (fonction `ggsurvplot`)

Table 3.3: Quantiles de la durée (Kaplan Meier)

$S(t)$	t
0.90	6
0.75	36
0.50	100
0.25	979
0.10	.

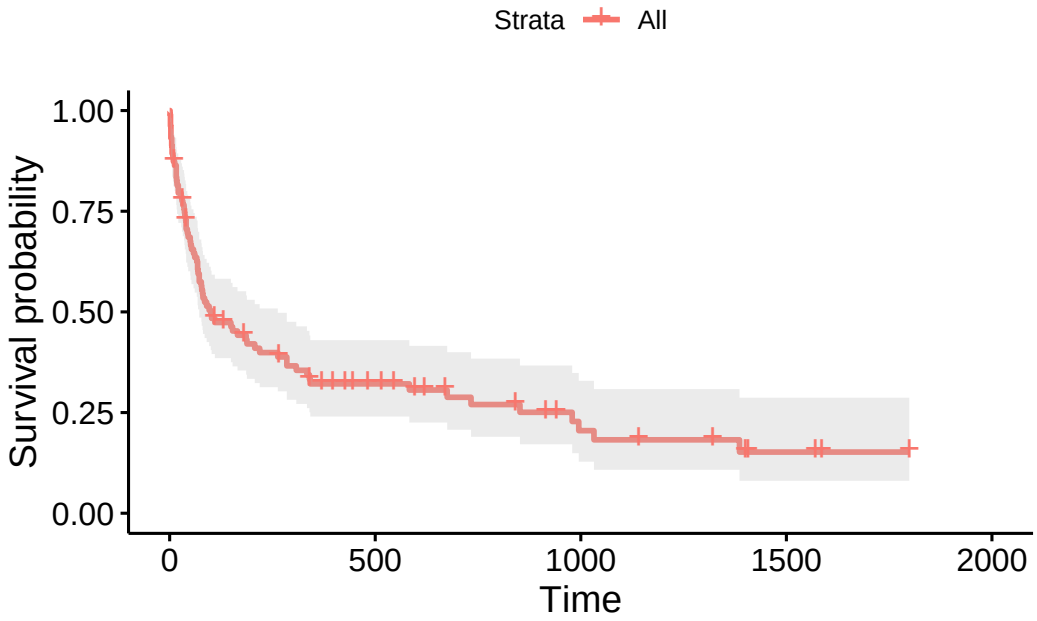

Figure 3.2: $S(t)$: méthode Kaplan Meier

La durée médiane est ici égale à 100 jours et correspond à la valeur de la durée pour $S(t) = 0.494$. Les valeurs des quantiles sont proches de ceux calculé par interpolation linéaire avec la méthode actuarielle sur des intervalles fixes de 10 jours.

Quantités associées

A partir de l'estimateur Kaplan-Meier..

Le risque cumulé: estimateur d'AAlen Il est simplément égal à:

$$H(t) = \sum_{t_i \leq k} q(t_i)$$

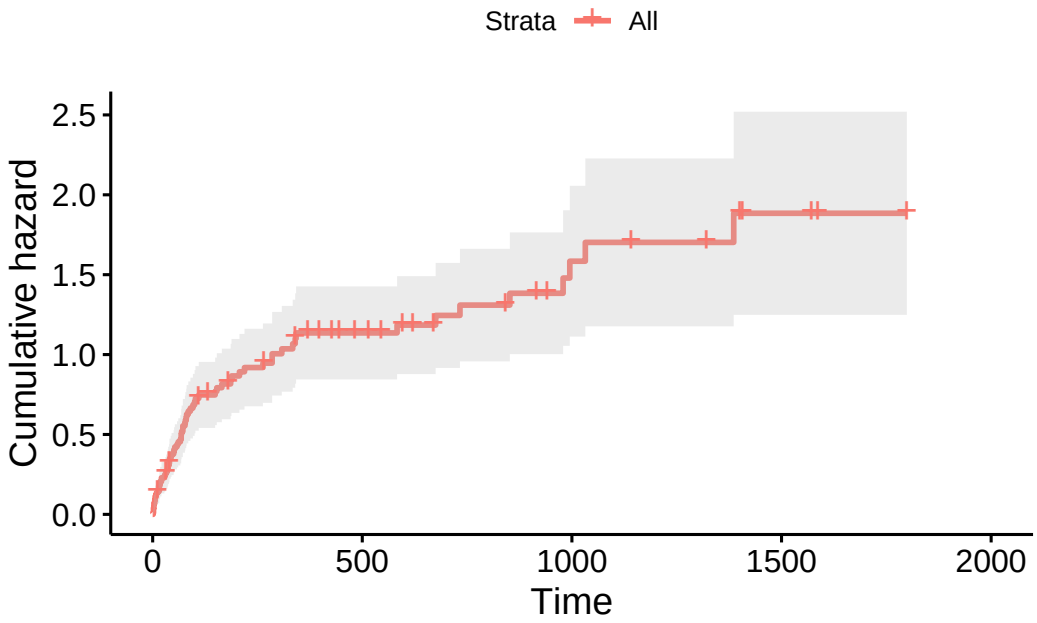

Figure 3.3: $H(t)$: méthode d'AAlen

Le risque instantané
Nécessite l'estimateur de Nelson-Aalen. Le risque est obtenu en lissant les différences - toujours positive - entre $H(t)$ par la méthode dite du **kernel**. Elle permet d'obtenir une fonction continue avec la durée (paramétrables sur les largeurs des fenêtres de lissage). D'autres méthodes de lissage sont maintenant possibles, et de plus en plus utilisées, en particulier celles utilisant des *splines*.

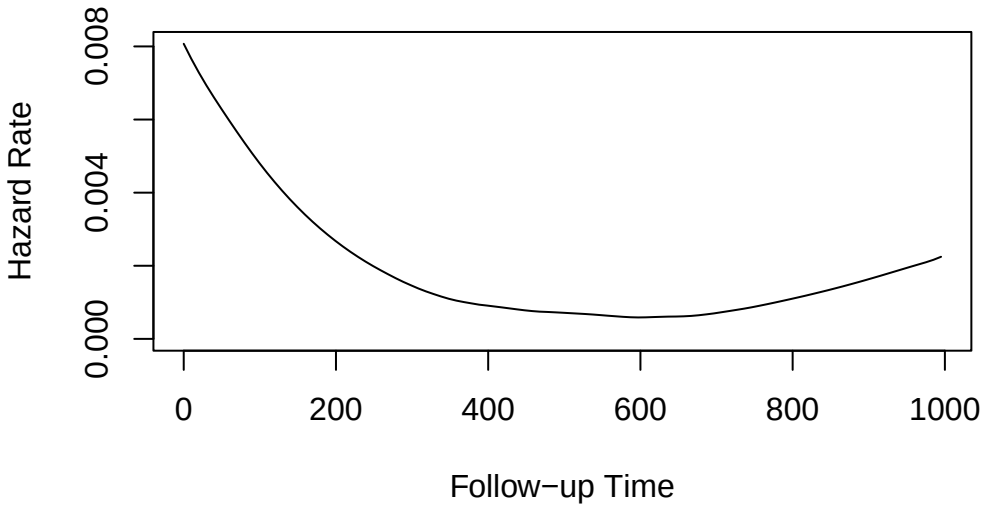

Figure 3.4: $h(t)$: méthode du Kernel

3.2 Comparaison des fonctions de survie/séjour

Les tests d'égalités des fonctions de survie entre différentes valeurs d'une covariable sont calculés à partir de la méthode de Kaplan Meier.
L'utilisation du test correspond à la nécessité de déterminer si une même distribution gouverne les évènements observés dans les différentes strates.
Attention: pas de test possible sur des variables quantitatives/continues. Il faut donc prévoir des regroupements pour les transformer en variables ordinales.

Deux méthodes sont utilisées:

- La plus ancienne et la plus diffusée: test dits du **log-rank**).
- Plus récente et (hélas) moins difusée: comparaison des **RMST** (*Restricted Mean of Survival Time*).

3.2.1 Tests du log-rank

Il s'agit d'une série de tests qui répondent à la même logique, la seule différence réside dans le poids accordé au début ou à la fin de la période d'observation. Par ailleurs ces différents tests sont plus ou moins sensibles à la distribution des censures à droites entre les sous échantillons.

Dans leur logique, ces tests entrent dans le cadre des tests d'indépendance du Khi2, même si formellement ils relèvent des techniques dites de rang.
Il s'agira donc de comparer des effectifs observés à des effectifs espérés à chaque temps d'évènement. La principale différence réside dans le calcul de la variance de la statistique du test qui, ici, suit une loi hypergéométrique (proche loi binomiale mais avec tirage sans remise).

Principe de calcul des effectifs - évènements - observés et espérés pour deux groupes

- **Effectifs observés en** t_i: o_{i1} et o_{i2} sont égaux à d_{i1} et d_{i2}, et leur somme pour tous les temps d'évènement à O_1 et O_2.
- **Effectifs expérés** (hypothèse nulle H_0): comme pour une statistique du χ^2 on se base sur les marges, avec le risque set (R_i) en t_i pour dénombrer les effectifs, soit $e_{i1} = R_{i1} \times \frac{d_i}{R_i}$ et $e_{i2} = R_{i2} \times \frac{d_2}{R_i}$. Leur somme pour tous les temps d'évènement est égale à E_1 et E_2. Le principe de calcul des effectifs observés reposent donc sur l'hypothèse d'un rapport des risques toujours égal à 1 au cours du temps (*hypothèse fondamentale de risques proportionnels*).
- **Statistique du log-rank**: $(O_1 - E_1) = -(O_2 - E_2)$.
- **Statistique de test**: sous H_0, $\frac{(O_1-E_1)^2}{\sum v_i}$, avec v_i la variance de $(o_{i1} - e_{i2})$, suis un $\chi^2(1)$. Ce n'est pas forcément à retenir, la variance suis une loi géométrique. Elle est particulièrement adaptée à ce type d'analyse, comme elle correspond à une variante de la loi binomiale avec des tirages sans remise.

Si on teste la différence de g fonctions de survie, la statistique de test suis un $\chi^2(g-1)$.

Les principaux tests de type log-rank

Le principe de construction des effectifs observés et espérés reste le même dans chaque test, les différences résident dans les pondérations (w_i) qui prennent en compte, de manière différente, la taille de la population soumise au risque à chaque durée où au moins un évènement est observé.

- **Test du log-rank**: $w_i = 1$
  Il accorde le même poids à toutes les durées d'évènement. C'est le test standard, le plus utilisé.
- **Test de Wilconxon-Breslow-Grehan**: $w_i = R_i$
  Les écarts entre effectifs observés et espérés sont pondérés par la population soumise à risque en t_i. Le test accorde plus de poids au début de la période analysée, et il est sensible aux différences de distributions entre les strates des observations censurées.
- **Test de Tarone-Ware**: $w_i = \sqrt{R_i}$
  Variante du test précédent, il atténue le poids accordé aux évènements au début de la période d'observation. Il est par ailleurs moins sensible au problème de la distribution des censures entre les strates.

- **Test de Peto-Peto** : $w_i = S_i$
  La pondération est une variante de la fonction de survie KM (avec $R_i = R_i + 1$). Le test n'est pas sensible au problème de distribution des censures.
- **Test de Fleming-Harington**: $w_i = (S_i)^p \times (1 - S_i)^q$ avec $0 \leq p \leq 1$ Il permet de paramétrer le poids accordé au début où à la fin de temps d'observation. Si $p = q = 0$ on retrouve le test du log-rank.

En pratique/remarques:

- Les tests du log-rank sont sensibles à l'hypothèse de risques proportionnels (voir **modèle semi-paramétrique de Cox**). En pratique si des courbes de séjours se croisent alors que la population soumise au risque reste relativement nombreuse, il est déconseillé de les utiliser. Cela ne signifie pas que si les courbes ne se croisent pas, l'hypothèse de proportionnalité des risques est respectée : des rapports de risque peuvent au cours du temps s'intensifier, se réduire ou le cas échant s'inverser (typique d'un croisement).
- Effectuer un test global (multiple/omnibus) sur un nombre important de groupes (ou >2) peut rendre le test très facilement significatif. Il peut être intéressant de tester des courbes deux à deux (idem qu'une régression avec covariable discrète), en conservant un seul degré de liberté. Des méthodes de correction du test multiple sont possibles.

R

On utilise la fonction **`survdiff`** de la librairie `survival`. Le résultat du test de Peto-Peto est affiché par défaut (`rho=1`). Si on souhaite utiliser le test non pondéré, on ajoute l'option `rho=0`.
Pour obtenir le résultat d'un test multiple corrigé (plus d'un degré de liberté), on peut utiliser la fonction **`pairwise_survdiff`** de la librairie **`survminer`**. Cette fonction permet d'obtenir des tests 2 à 2 si une variable a plus de deux groupes.
Je conseille de rester sur l'option ***Peto-Peto*** et dans le cas d'une variable à plus de deux modalités, d'utiliser la fonction de `survminer` **`pairwise_survdiff`**.

Application

On compare ici l'effet du pontage sur le risque de décéder depuis l'inscription dans le registre de greffe (variable *surgery*).

Table 3.4: Résultats des tests du logrank

Variante du test	$Chi2(1)$	$p > chi2$
Log-Rank	6.59	0.0103
Wilcoxon	8.99	0.0027
Taron-Ware	8.46	0.0036
Peto-Peto	8.66	0.0033

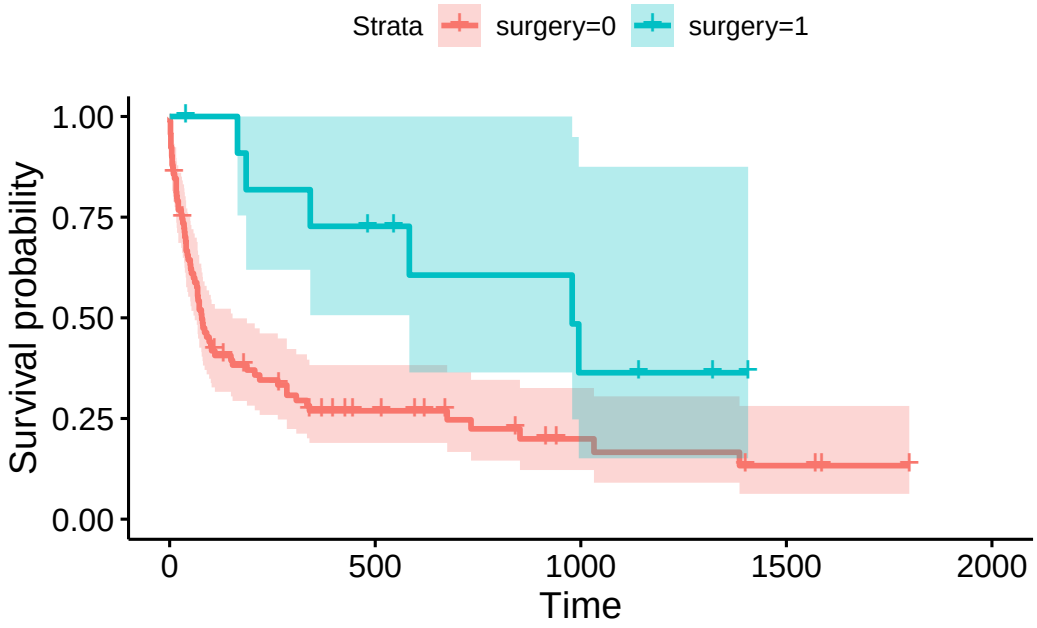

Figure 3.5: $S(t)$: méthode Kaplan Meier

Question: En visualisant la fonction de survie, doutez-vous ou non de que l'hypothèse de proportionalité des risques soit respectée.

Même si les courbes ne se croisent pas, on peut établir un premier diagnositic de proportionalité avec une double transformation *log* de $S(t)$.

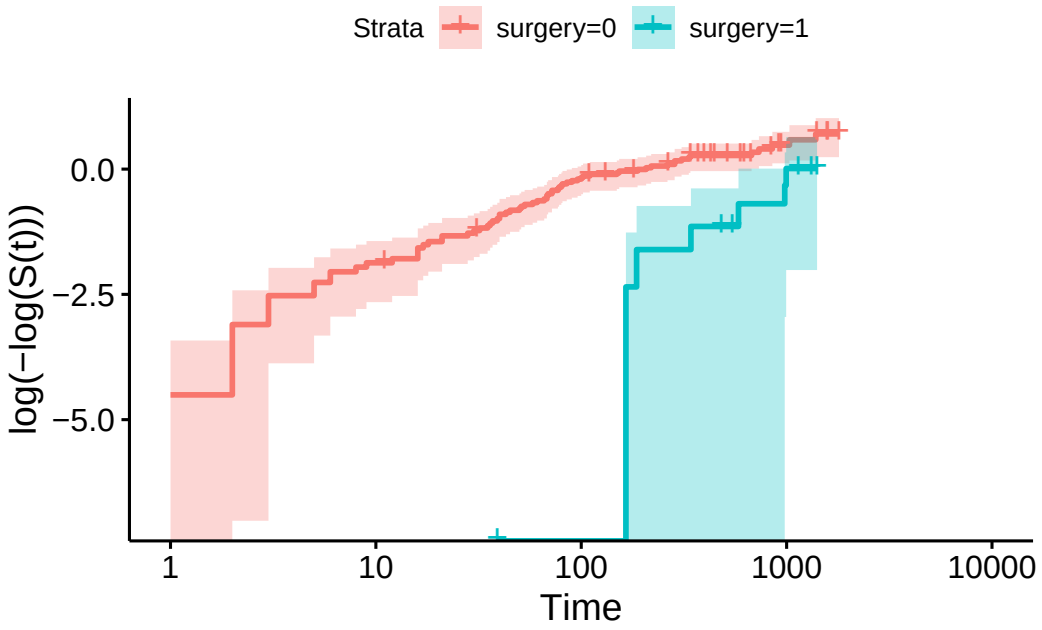

Figure 3.6: $log(log(S(t))$: proportionalité des risques

3.2.2 Comparaison des RMST

RMST: *Restricted Mean of Survival Time*

La comparaison des RMST est une alternative pertinente aux tests du log-rank car elle ne repose pas sur des hypothèses contraignantes (proportionnalité des risques, distribution des censures), et permet une lecture vivante basée sur des espérances de séjour et non sur la lecture d'une simple p-value traduisant l'homogénéité ou non des fonctions de séjour. Par ailleurs les comparaisons sont souples, on peut choisir un ou plusieurs points d'horizon pour alimenter l'analyse.

- L'aire sous la fonction de survie représente la durée moyenne d'attente jusqu'à l'évènement, soit une espérance de survie.
- En présence de censure à droite, il faut borner la durée maximale $t^* < \infty$. L'espérance de survie s'interprète donc sur un horizon fini. On est très proche d'une mesure en analyse démographique type « espérance de vie partielle ».
- $RMST = \int_0^{t^*} S(t)dt$.
- On peut facilement comparer les RMST de deux groupes, en termes de différence ou de ratio.
- Par défaut on définit généralement t^* à partir le temps du dernier évènement observé. Il est néanmoins possible de calculer le RMST sur des intervalles plus court, ce qui lui permet une véritable souplesse au niveau de l'analyse.

R

On utilise la librairie **SurvRm2**.Elle demande à ce que la variable de comparaison soit appelé **arm** et codée sous forme d'indicatrice $(0, 1)$. La fonction, issue d'une commande de Stata, n'est pas très souple.
Attention, selon les logiciels la durée max par défaut n'est pas la même. Pour R et Sas, il s'agit du dernier évènement observé sur l'ensemble de l'échantillon, alors que Stata prend la durée qui correspond au dernier évènement observé le plus court des deux groupes . Cela affectera légèrement la valeur des Rmst estimées par défaut.
Pour l'exemple, la durée maximale utilisée par R est de 1407 jours alors que pour Stata elle est de 995 jours.

Table 3.5: RMST

	RMST	Lower .95	Upper 0.95
arm=0	884.576	586.802	1182.450
arm=1	379.148	264.283	494.012

Table 3.6: Différences des RMST

	Est	p
(arm=1) - (arm=0)	505.428	0.002
(arm=1) /(arm=0)	2.333	0.000

Sur un horizon de 1407 jours, les personnes opérées peuvent espérer vivre 884 jours depuis leur inscription dans le registre de greffe , contre 379 jours pour les autres. Cette durée moyenne de survie est donc deux fois plus élevée pour les personnes opérées d'un pontage (rapport des Rmst = 2.3), soit une différence de 505 jours.

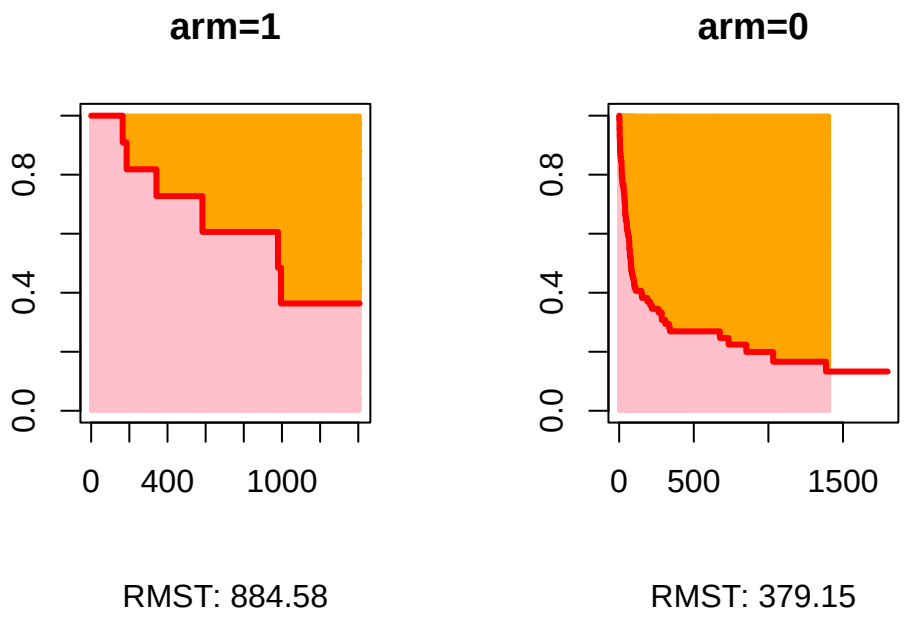

Figure 3.7: Estimation des Rmst

Le graphique suivant, donne les valeurs des Rmst pour la variable *surgery* en faisant varier t_max sur chaque durée où un décès a été observé. Il a été réalisé avec Stata, la durée maximale utilisée est ici de 995 jours.

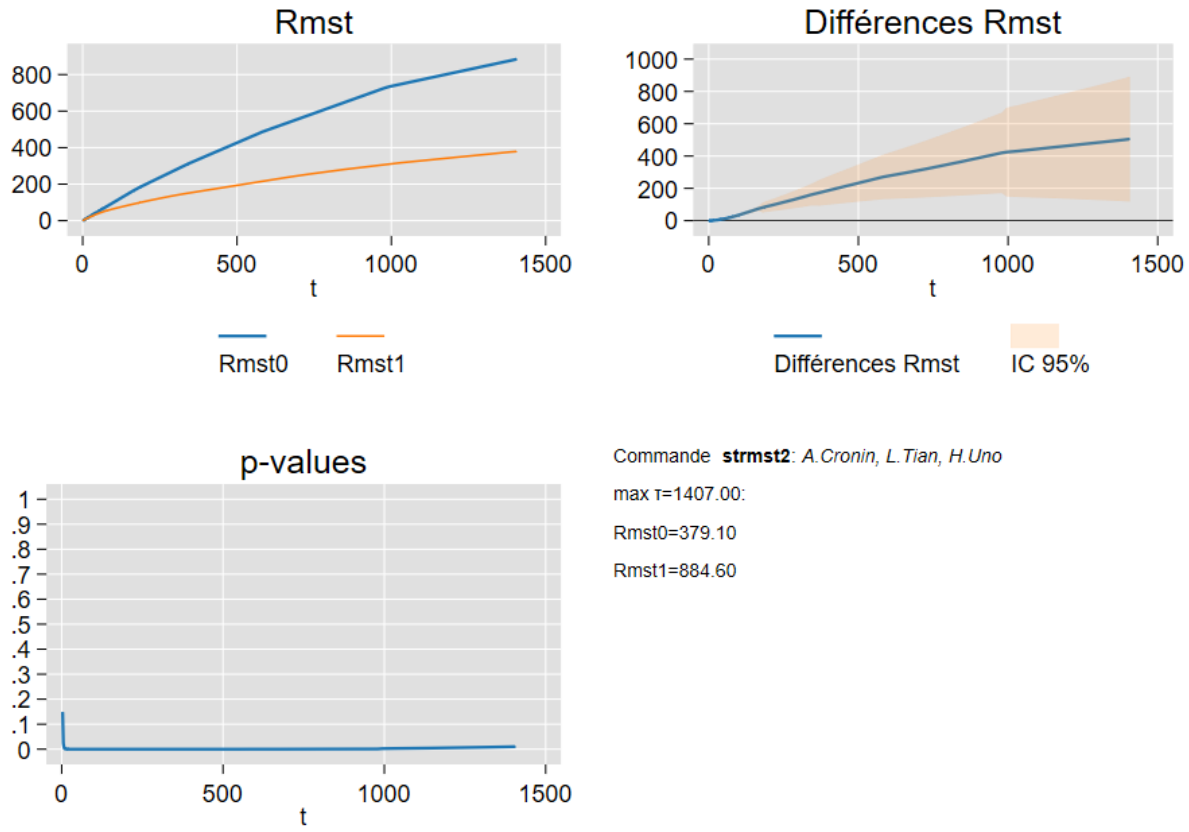

Figure 3.8: Variation des Rmst (Graphique Stata)

4 Cox: un Modèle à risques proportionnels

4.1 Introduction

La spécification usuelle d'un modèle à risque proportionnel est:

$$h(t) = h_0(t) \times e^{X^{'}b}$$

- $h(t)$ est une fonction de risque.
- $h_0(t)$ est une fonction qui dépend du temps mais pas des caractéristiques individuelles. Il définiera le risque de base (baseline).
- $e^{X^{'}b}$ est une fonction qui ne dépend pas du temps, mais des caractéristiques individuelles $X^{'}b = \sum_{k=1}^{p} b_k X_k$. La forme exponentielle assurera la positivité du risque.

Le risque de base

- $h(t) = h_0(t)$ donc $e^{X^{'}b} = 1$
- Observations pour lesquelles $X = 0$

Risques proportionnels
Cette hypothèse stipule l'invariance dans le temps du "rapport des risques" (**hazard ratio**).

Avec une seule covariable X introduite au modèle, et 2 individus A et B: $h_A(t) = h_0(t)e^{bX_A}$ et $h_B(t) = h_0(t)e^{bX_B}$.
Le rapport des risques entre A et B est égal à:

$$\frac{h_A(t)}{h_B(t)} = \frac{e^{bX_A}}{e^{bX_B}} = e^{b(X_A - X_B)}$$

Pour une caractéristique binaire: $X_A = 1$ et $X_B = 0$: $\frac{h_A(t)}{h_B(t)} = e^b$.

Autrement dit, la proportionnalité des risques peut traduire l'absence d'une interaction significative entres les rapports de risques estiméspar un modèle à risque proportionnel et la durée (ou une fonction de celle-ci).

On part d'un modèle à risque constant avec $h_0(t) = 0.1$.

Comme $h_1(t) = 0.2$ quel que soit t, le rapport des risques est toujours égal à $\frac{0.2}{0.1} = 2 = e^b$. Le coefficient estimé est égal à $log(2) = 0.69$.
Pour $h_{1b}(t)$, le rapport de risques augmente avec le temps: $t = 1$, $h_{1b}(1) = 0.15$ et $h_{1b}(1000) = 0.25$ l'hypothèse de proportionalité n'est donc pas respectée.

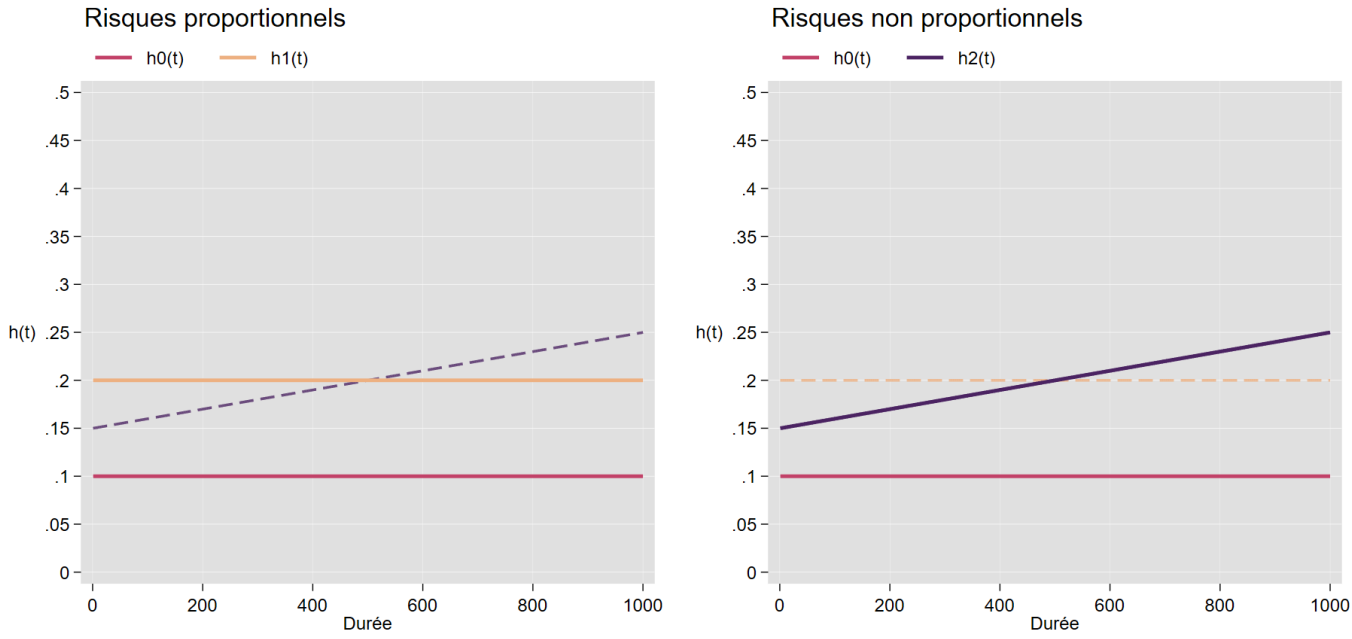

Figure 4.1: Modèle à risque proportionnel

Les modèles

- ***Modèle semi-paramétrique de Cox***: Le modèle estime directement les b indépendement de $h_0(t)$, c'est pour cela qu'il est semi-paramétrique. Les rapports de risque (e^b) seront utilisés pour estimer la baseline $h_0(t)$, qui peut s'avérer nécessaire pour calculer des fonctions de survie ajustées. Le respect de l'hypothèse de proportionalité est donc important et donc être testé. Très, peut-être trop, populaire, c'est ce modèle qui va être étudier pendant le cours.
- ***Modèle à temps discret***: De type paramétrique. Peut être estimé à l'aide d'un modèle logistique, probit ou complémentaire log-log. Le premier est le plus courant, le dernièr a l'avantage d'être directement relié au modèle de Cox (modèle de Cox à temps discret). Sa forme diffère de la présentation usuelle d'un modèle à risque proportionnel. Toutefois, il est régi par une hypothèse de proportionnalité. Le non respect de l'hypothèse est moins critique car la baseline du « risque » est estimée simultanément. Il est comme son nom l'indique, particulièrement adapté au durées discrètes ou groupées. Avec une spécification logistique, les Odds vont sous certaines conditions, se confondre avec des probabilités/risques. Si on a un peu de temps, je le présenterais brièvement Support formation interne Ined.

- ***Les modèles paramétriques standard***:
  - les modèles dits de *Weibull*, *exponentiel* ou *Gompertz* ont une spécification sous hypothèse de risques proportionnels. Ils ne sont pas traité dans ce cours . Historiquement, le modèle de Cox est une réponse à une possible difficulté dans l'ajustement du risque par une loi de distribution a priori. formation interne Ined.
  - Le modèle de *Parmar-Royston*. $h_0(t)$, via le risque cumulé $H(t)$, est estimé simultanément avec les risques ratios en utilisant des *splines cubiques*. Il est implémenté dans les logiciels standards (R, Stata, Sas). Les rapports de risque sont très proches de ceux estimés par le modèle classique de Cox. Il offre donc une alternative particulièrement intéressante à celui-ci, et il s'est maintenant largement diffusé dans l'étude des effets cliniques.

4.2 Le modèle semi-paramétrique de Cox

On peut ignorer la partie sur l'estimation du modèle. On retiendra tout de même qu'il est déconseillé de tester la méthode dite exacte pour la correction de la vraisemblance, qui ne peut matériellement fonctionner qu'avec un nombre très limité d'évènements observés simultanément, ce qui est plutôt rare avec des données à durées discrètes ou groupées, classiques dans les sciences sociales.

4.2.1 La vraisemblance partielle et estimation des paramètres

On se situe dans une situation où la durée est mesurée sur une échelle strictement continue. Il ne peut donc y avoir qu'un seul évènement observé en t_i (idem si censure):

$$L_i = f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

Vraisemblance partielle de Cox

- $f(t_i)$ est la valeur de la fonction de densité en t_i
- $S(t_i)$ est la valeur de la fonction de survie en t_i
- $\delta_i = 1$ si l'évènement est observé: $L_i = f(t_i)$
- $\delta_i = 0$ si l'observation est censurée: $L_i = S(t_i)$

Comme $f(t_i) = h(t_i) \times S(t_i)$, on obtient: $L_i = [h(t_i)S(t_i)]^{\delta_i} S(t_i)^{1-\delta_i} = h(t_i)^{\delta_i} S(t_i)$.
Pour $i = 1, 2, ..., n$, la vraisemblance totale s'ecrit donc: $L_i = \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i)$.

On peut réécrire cette vraisemblance en la multipliant et en la divisant par: $\sum_{j \in R_i} h(t_i)$, où $j \in R_i$ est l'ensemble des observation soumises au risque en t_i.

$$L = \prod_{i=1}^{n} \left[h(t_i) \frac{\sum_{j \in R} h(t_i)}{\sum_{j \in R} h(t_i)} \right]^{\delta_i} S(t_i) = \prod_{i=1}^{n} \left[\frac{h(t_i)}{\sum_{j \in R_i} h(t_i)} \right]^{\delta_i} \sum_{j \in R_i} h(t_i)^{\delta_i} S(t_i)$$

La vraisemblance partielle retient le premier terme de la vraisemblance, soit:

$$PL = \prod_{i=1}^{n} \left[\frac{h(t_i)}{\sum_{j \in R} h(t_i)} \right]^{\delta_i}$$

Une fois remplacée la valeur de $h(t_i)$ par son expression en tant que modèle à risques proportionnels, la vraisemblance partielle ne dépendra plus de la durée. **Mais elle va dépendre de l'ordre d'arrivée des évènements, c'est à dire leur rang**.

Remarque: pour les observations censurées($\delta_i = 0$), $PL = 1$. Toutefois, ces censures entrent dans l'expression $\sum_{j \in R} h(t_i)$ tant qu'elles sont soumises au risque.

En remplaçant donc $h(t_i)$ par l'expression $h_0(t)e^{X_i' b}$:

$$PL = \prod_{i=1}^{n} \left[\frac{h_0(t)e^{X_i' b}}{\sum_{j \in R_i} h_0(t)e^{X_j' b}} \right]^{\delta_i} = \prod_{i=1}^{n} \left[\frac{e^{X_i' b}}{\sum_{j \in R_i} e^{X_j' b}} \right]^{\delta_i}$$

L'expression $\frac{e^{Xb}}{\sum_{j \in R} e^{Xb}}$ est une probabilité, la vraisemblance partielle est donc bien un produit de probabilités. **Il s'agit de la probabilité qu'un individu observe l'évènement en t_i sachant qu'un évènement (et un seul) s'est produit**.

Condition nécessaire: pas d'évènement simultané: On rappelle que la durée est mesuré de manière strictement continue, il ne doit pas y avoir d'évènement simultané. Sinon, l'estimation de la vraisemblance doit faire l'objet d'une correction.

Correction de la vraisemblance avec des évènements simultanés:

- La ***méthode dite exacte***: Comme il ne dois pas y avoir d'évènement simultané, on va intégrer à la vraisemblance toutes les permutations possibles des évènements observés simultanément: si en t_i on observe au « même moment » l'évènement pour A et B, une échelle temporelle plus précise nous permettrait de savoir si A s'est produit avant B ou B s'est produit avant A. Comme le nombre de permutations se calcule par une factorielle, avec 3 évènements mesurés simultanément, il y a 6 permutations possibles ($3 \times 2 \times 1$).
  Problème: le nombre de permutations pour chaque t_i peut devenir très vite particulièrement élevé. Par exemple pour 10 évènements simultanés, le nombre de permutations est égal à 3.628.800. Le temps de calcul devient extrêmement long, et ce type de correction totalement inopérant.
- La ***méthode dite de Breslow***: il s'agit d'une approximation de la méthode exacte permettant de ne pas avoir à intégrer chaque permutation. Cette approximation est utilisée par défaut par les logiciels Sas et Stata.
- La ***méthode dite d'Efron***: elle corrige l'approximation de Breslow, et est jugée plus proche de la méthode exacte. C'est la méthode utilisée par défaut avec le logiciel `R`, et elle est disponible avec les autres applications. On ne touche donc à rien sur `R`.

Estimation des paramètres

On utilise la méthode habituelle, à savoir la maximisation de la log-vraisemblance (ici partielle).

- Conditions de premier ordre: calcul des équations de score à partir des dérivées partielles. Solution: $\frac{\partial log(PL)}{\partial b_k} = 0$. On ne peut pas obtenir de solution numérique directe.
  Remarque: les équations de score sont utilisées pour tester la validité de l'hypothèse de constance des rapports de risque pour calculer les **résidus de Schoenfeld** (voir plus bas).
- Conditions de second ordre: calcul des dérivées secondes qui permettent d'obtenir la matrice d'information de Fisher et la matrice des variances-covariances des paramètres.
- Comme il n'y a pas de solution numérique directe, on utilise un algorithme d'optimisation (ex: Newton-Raphson) à partir des équations de score et de la matrice d'information de Fisher.

Eléments de calcul

En logarithme, la vraisemblance partielle s'ecrit:

$$pl(b) = log(pl(b)) = log\left(\prod_{i=1}^{n}\left[\frac{e^{X_i'b}}{\sum_{j\in R_i} e^{X_j b}}\right]^{\delta_i}\right)$$

$$pl(b) = \sum_{i=1}^{n}\delta_i log\left(\frac{e^{X_i'b}}{\sum_{j\in R_i} e^{X_j' b}}\right)$$

$$pl(b) = \sum_{i=1}^{n}\delta_i\left(log(e^{X_i'b}) - log\sum_{j\in R_i} e^{X_i'b}\right)$$

$$pl(b) = \sum_{i=1}^{n}\delta_i\left(X_i'b - log\sum_{j\in R_i} e^{X_i'b}\right)$$

Calcul de l'équation de score pour une covariable X_k:

$$\frac{\partial pl(b)}{\partial b_k} = \sum_{i=1}^{n}\delta_i\left(X_{ik} - \sum_{j\in R_i} X_{ik}\frac{e^{X_i'b}}{\sum_{j\in R_i} e^{X_j' b}}\right)$$

Comme $\frac{e^{X_{ik}b}}{\sum_{j\in R} e^{X_{jk}b}}$ est une probabilité, $\sum_{j\in R} X_{ik} \times p_i$ est l'espérance (la moyenne) $E(X_k)$ d'avoir la caractéristique X_k lorsqu'un évènement a été observé. Au final:

$$\frac{\partial lp(b)}{\partial b_k} = \sum_{i=1}^{n}\delta_i\left(X_{ik} - E(X_{j\in R_i,k})\right)$$

Cette expression est assez importante, elle va permettre de tester le respect ou non de l'hypothèse de risques proportionnels via les *résidus de Schoenfeld* .

4.2.2 Lecture des résultats

Comme il s'agit d'un modèle à risque proportionnel, **les rapports de risques sont constants pendant toute la période d'observation**. Il s'agit d'une propriété de l'estimation.

Covariable binaire (indicatrice):
$X = (0; 1)$: $RR = \frac{h(t \mid X=1)}{h(t \mid X=0)} = e^b$.
A* **chaque moment de la durée** t, le risque d'observer l'évènement est e^b fois plus important/plus faible pour $X = 1$ que pour $X = 0$.

Covariable quantitative (mais fixe dans le temps):
$RR = \frac{h(t \mid X=a+c)}{h(t \mid X=a)} = e^{c \times b}$. On prendra pour illustrer une variable type âge au début de l'exposition au risque (a) et un delta de comparaison avec un âge inférieur (c).
Si $c = 1$ (résultat de l'estimation): A un âge donnée en début d'exposition, le risque de connaitre l'évènement est e^b fois inférieur/supérieur à celui d'une personne qui a un an de moins au début de l'exposition.

Avec l'application sur les insuffisance cardiaque:

Table 4.1: Modèle de Cox

	Hazard Ratio	**95% CI**	**p-value**
year	0.89	[0.78-1.01]	0.076
age	1.03	[1.00-1.06]	0.029
surgery	0.37	[0.16-0.88]	0.024

On retrouve les résultats des tests non paramétriques pour l'opération, à savoir qu'un pontage réduit les risques journaliers de décès pendant la période d'observation (augmente la durée de survie).
De la même manière, plus on entre à un âge élevé dans la liste d'attente plus le risque de décès augmente.
La variable *year*, qui traduit des progrès en médecine, exprime une réduction relativement modérée du risque journalier de décès durant l'attente de la greffe du coeur.

R

Le modèle est estimé avec la fonction `coxph` de la librairie `survival`. Hors options, la syntaxe est identiques aux fonctions `survfit` et `survdif`.

4.2.3 L'hypothèse de constance des rapports de risque

- Les rapports de risque (RR) estimés par le modèle sont contraints à être constant pendant toute la période d'observation. C'est une hypothèse forte.

- Le respect de cette hypothèse doit être testé, en particulier pour un modèle de Cox où la baseline du risque est habituellement estimée à l'aide de ces rapports (méthode dite de Breslow, non traitée). En post-estimation, les valeurs estimées du risque pourront présenter des valeurs aberrantes, en particulier négatives.
- Tester cette hypothèse revient à tester une interaction entre les rapports et la durée (ou plutôt une fonction de la durée).
- Plusieurs méthodes disponibles, celle sur les résidus de martingales, réservée aux covariables continues, et le « test » graphique sur des variables catégorielles ne seront pas traités. On traitera celle basée sur les **résidus de Schoenfeld** puis l'introduction d'une intéraction avec la durée dans le modèle. Cette dernière peut faire également office de méthode de correction lorsque l'hypothèse n'est pas respectée.
- Si on regarde les courbes de Kaplan-Meier, leurs croisement non tardif impliquera nécessairement un problème sur cette hypothèse.

Tests sur les résidus de Schoenfeld

- Les résidus « bruts » sont directement calculés à partir des équations de scores (voir section estimation).
- Ce résidu n'est calculé que pour les observations qui ont connues l'évènement.
- Il est calculé au moment où l'évènement s'est produit.
- La somme des résidus pour chaque covariable est égale à 0 (propriété de l'équation de score à l'équilibre).
- On utilise généralement les résidus de Schoenfeld « standardisés » ou plutôt "remis à l'échelle" - par leur variance - pour tenir compte du fait que la population soumise au risque diminue au cours du temps.
- Pour une observation dont l'évènement s'est produit en t_i, le *résidu brut de Schoenfeld* pour la covariable X_k, après estimation du modèle, est égal à:

$$rs_{ik} = X_{ik} - \sum_{j \in R_i} X_{ik} \frac{e^{X_i' b}}{\sum_{j \in R_i} e^{X_j' b}} = X_{ik} - E(X_{j \in R_i, k})$$

- Ce résidu est formellement la contribution d'une observation au score. Il se lit comme la différence entre la valeur observée d'une covariable et sa valeur espérée au moment où un évènement se produit.
- Si l'hypothèse de constance des risk ratios est respectée, les résidus ne doivent pas suivre une tendance précise, en particulier à la hausse ou à la baisse.
- Intuitivement sans censure à droite et en ne considérant que les résidus bruts: on a un RR strictement égal à 1 en début d'exposition $R_i = 100$ avec 50 hommes et 50 femmes. Si l'hypothèse PH (strictement) respectée, lorsqu'il reste 90 personnes soumises au risque, on devrait avoir 45 hommes et 45 femmes. Avec $R_i = 50$, 25 hommes et 25 femmes,…….avec $R_i = 10$, 5 hommes et 5 femmes. Au final l'espérance d'avoir la caractéristique X est toujours égal à 0.5 et les résidus bruts prendront toujours la valeur -.5 si $X = 0$ et .5 si $X = 1$. En faisant une simple régression linéaire

entre les résidus, qui alternent ces deux valeurs, et t, le coefficient estimé sera non significativement différent de 0.

- On peut tester l'hypothèse sur les résidus par une régression entre ces résidus pour chaque covariable et la durée (ou fonction dérivée de la durée, par exemple t ou $log(t)$)). Cette solution prend forme avec le test de **Grambsch-Therneau** sous sa forme simplifiée - OLS - (R jusqu'à v3 de `survival`, Sas, Stata, Python) ou exacte - GLS - (R depuis la V3).

Vous trouverez des éléments de calcul du test simplifié ici

! survival v2 versus v3

Depuis la v3 de la librairie `Survival`, il s'agit non pas pas d'une nouvelle version du test qui est implémentée, mais de la version d'origine . Tous les autres logiciels utilisent la version *simplifiée* de ce test, proposée elle même en 2002 par les auteur.e.s. A ce jour, tous les autres applications statistiques utilisent la version simplifiée, qui est donc parfaitement reproductible dans l'espace des logiciels et dans le temps.
Il y a visiblement peu de conséquences en durée strictement continue, situation plutôt rares dans les sciences sociales. Je conseille donc, en attendant une investigation méthodologique plus poussée, d'utiliser le test de la version précédente. Je donne une procédure simple d'exécuter ce test dans le chapitre *programmation avec R*, en récupérant directement la fonction. Seul le nom de la fonction a été modifié (**`cox.zphold`**).

Table 4.2: Test Grambsch-Therneau simplifié (v2 de survival) avec $g(t) = t$

	rho	chi2	df	$p > Chi2$
year	0.102	0.80	1	0.3720
age	0.129	1.61	1	0.2043
surgery	0.297	5.54	1	0.0186
global test	-	8.76	3	0.0327

Ici l'hypothèse de proportionalité des risques est questionnable pour la variable *surgery*. Le risque ratio pourrait ne pas constant dans le temps. Ce n'est pas peut-être pas étonnant, le premier décès pour les personnes opérées d'un pontage n'est observé qu'au bout de 165 jours.

i Remarques / à savoir

- **Test multiple**: de nouveau il convient de se méfier du résultat du test multiple. Le risque de premier espèce peut-être assez faible alors que les tests pour chaque covariables prises une à une présenteraient des valeurs plutôt élevées (>.1 par exemple). Le résultat de ce test multiple est considéré par certain.e.s comme un indicateur de l'ampleur du biais qui affecte la baseline du risque.

- **Transformations de la durée**: n'importe quelle fonction de la durée peut être utilisée

pour réaliser le test. On retient généralement les fonctions suivantes: $g(t) = t$ (« identity »), $g(t) = log(t)$, $g(t) = KM(t)$ ou $g(t) = 1 - KM(t)$ où $KM(t)$ est l'estimateur de Kaplan-Meier. Enfin une transformation appelée « rank », est utilisée seulement pour les durées strictement continue ou suffisamment dispersées. Par exemple $t = (0.1, 0.5, 1, 2.6, 3)$ donne une transformation $t = (1, 2, 3, 4)$. A savoir : la fonction « identity », utilisée ici, rend le test relativement sensible aux évènements tardifs lorsque la population restant soumise est peu nombreuse (outliers).

Intéraction avec la durée

Petit retour sur l'estimation du modèle.

Pour estimer le modèle de Cox, les logiciels *splittent* dans un premier temps les données aux durées d'évènement.

Table 4.3: Format splitté de la base aux temps d'évènement

id	surgery	died	stime	t	t_0
2	0	0	6	1	0
2	0	0	6	2	1
2	0	0	6	3	2
2	0	0	6	5	3
2	0	1	6	6	5
3	0	0	16	1	0
3	0	0	16	2	1
3	0	0	16	3	2
3	0	0	16	5	3
3	0	0	16	6	5
3	0	0	16	8	6
3	0	0	16	9	8
3	0	0	16	12	9
3	0	1	16	16	12

Les bornes des intervalles $[t_0; t]$ ont des valeurs seulement lorsqu'un évènement s'est produit (principe de la vraisemblance partielle). Il n'y a donc pas de valeurs pour t et t_0 en $t = 4$ ($id = 2, 3$), $t = 7, 10, 11, 13, 14, 15$ ($id = 3$).
Les deux individus observent l'évènement en $t = 6$ pour $id = 2$, et en $t = 16$ pour $id = 3$. Avant ce moment la valeur de la variable évènement/censure (ici d) prend toujours la valeur 0, et prend la valeur 1 le jour du décès.

On pourra vérifier facilement que les paramètres estimés sont identiques en estimant le modèle avec cette base splittée.

Introduction d'une intéraction avec une fonction de la durée

On a une variable de durée avec une forme fonctionnelle simple $f(t) = t$. La variablr sera croisée avec la variable *surgery*, et le le modèle va s'écrire:

$$h(t|X,t) = h_0(t)e^{b_1 age + b_2 year + b_3 surgery + b_4 (surgery \times t)}$$

Table 4.4: Modèle de Cox - Intéraction avec la durée

	Hazard Ratio	**95% CI**	**p-value**
year	0.89	[0.78-1.01]	0.076
age	1.03	[1.00-1.06]	0.029
$surgery(t_{0^+})$	0.17	[0.03-0.65]	0.009
$surgery(\frac{(HR)_t}{(HR)_{t-1}})$	1.002	[1.000-1.004]	0.043

On retrouve donc un résultat proche de celui obtenu à partir du test simplifié sur les résidus de Schoenfeld pour la variable *surgery* (et c'est normal). Il a le mérité de pouvoir être interprété directement.
Malgré une hypothèse plutôt forte sur la forme fonctionnelle de l'intéraction, et dans les fait surement pas pertinente, on peut dire que chaque jour le HR entre personnes opérées et personnes non opérées augmente de 0.2%. L'effet de l'opération sur la survie des individus s'estompe donc avec le temps.

! Le terme d'intéraction est un rapport de rapports de risque

Attention, la colonne **HR** est censée présenter des rapports de risque. Le terme d'intéraction n'est pas un rapport de risque, mais un rapport de rapports de risque, plus précisément le rapport entre le rapport de risque en t et le rapport de risque en $t-1$, quel que soit t (car $g(t) = t$).
C'est la même chose dans un modèle logistique , ou le terme d'intéraction sous sa forme exponentielle est un rapport d'Odds Ratios (lui même un rapport). Sur l'échelle d'estimation (logarithmique), le terme d'intéraction est donc une double différence.

Important:

- Le modèle n'est plus un modèle à risque proportionnel. La variable *surgery* n'est plus une variable **fixe** mais une variable tronquée dynamique qui prend la valeur de t pour les personnes qui ont été opérées d'un pontage avant leur entrée dans le registre de greffe.
- L'altération des rapports de risque dépend de la forme fonctionnelle de l'intéraction choisie. Ici la modification dans le temps du rapport des risque est constante, ce qui est une hypothèse assez forte. On a, en quelques sorte, réintroduit une hypothèse de proportionnalité, ici sur le degré d'altération des écarts de risques dans le temps.

Que faire

Ne rien faire

On interprète le risque ratio comme un ratio moyen pendant la durée d'observation (P.Allison). Difficilement soutenable pour l'analyse des effets cliniques, elle peut être envisagée dans d'autres domaines. Attention au nombre de variables qui ne respecte pas l'hypothèse, l'estimation de la baseline du risque pourrait être sensiblement affectée. Il convient tout de même lors de l'interprétation, de préciser les variables qui seront analysées sous cette forme « moyenne » sur la période d'observation.
On peut également adapter cette stratégie du « ne rien faire » selon sens de l'altération des rapports de risque. Si aux cours du temps des différences de risque, déjà assez important en début d'observation s'accentuent, à la hausse comme à la baisse, on peut conserver cet estimateur moyen. Mais si l'effet est modéré : $RR > 1$ qui baisse ou $RR < 1$ qui augmente au cours du temps, je suis moins convaincu de la pertinence de cette option.
Il faut tenir compte de l'intérêt portée par les variables qui présentent un problème par rapport à l'hypothèse. Il n'est peut-être pas nécessaire de complexifier le modèle pour des variables introduites comme simples contrôles.
Mais plus problématique… On sait qu'une des causes du non respect de l'hypothèse peut provenir d'effets de sélection liées à des variables omises ou non observables. En analyse de durée ce problème prend le nom de ***frailty*** (fragilité) lorsque cette non homogénéité n'est pas observables. Des estimations, plus complexes, sont possibles dans ce cas, et sont en mesure malgré leur interprétation plutôt difficile de régler le problème. Si l'hypothèse est sensible aux problèmes d'omission, il convient donc de bien spécifier le modèle au niveau des variables de contrôle observables et disponibles.

Modèle de Cox stratifié

Utiliser la méthode dite de « Cox stratifiée » (non traitée). Utile si l'objectif est de présenter des fonctions de survie ajustées, et si une seule covariable (binaire) présente un problème. Les RR ne seront pas estimés pour la variable qui ne respecte pas l'hypothèse.

Intéraction

Introduire une interaction avec la durée, ce qui a été fait juste haut dessus. Cela peut permettre d'enrichir le modèle au niveau de l'interprétation. Valable si peu de covariables présentent des problèmes de stabilité des RR, dans l'idéal une seule variable. Attention tout de même à la forme de la fonction, dans l'exemple on a contraint l'effet d'interaction à être strictement linéaire, ce qui est une hypothèse plutôt forte…. et on remets en quelque sort une contrainte de proportionalité.

Modèles alternatifs

Utiliser un modèle alternatif: modèles paramétriques à risques proportionnels si la distribution du risque s'ajuste bien, le modèle paramétrique « flexible » de Parmar-Royston (non traité) ou un modèle à temps discret. Dans ce dernier cas, on peut également corriger la non proportionalité avec une intéraction. Si on ne le fait pas, les risques prédits sous forme de probabilités conditionnelles resteront toujours dans les bornes contrairement au modèle de Cox.
Utiliser un modèle non paramétrique additif dit d'Aalen ou une de ses variantes. Mais ces modèles, dont les résultats sont des visuels graphiques, et se commentent difficilement.

Forêt aléatoire
Autre méthode : les forêts aléatoires [à présenter un jour tout de même]. L.Breiman a dès le départ proposé une estimation des modèles de survie par cette méthode. Par définition, pas sensible à l'hypothèse PH. Mais cela reste des méthodes à finalité prédictive, moins riche en interprétation si ce n'est sur les facteurs d'importance.

4.2.4 Introduction d'une variable dynamique

Cette section sera principalement traitée par l'exemple, et on ne s'intéressera qu'aux variables de type discrète, avec un seul changement d'état.

- Dans un modèle de durée, une variable dynamique peut-être appréhendée comme une intéraction entre la durée et une variable quantitative.
- Pour un modèle de Cox, l'hypothèse de risque proportionnel ne peut donc pas être testée sur la variable d'origine.
- Ne pas tenir compte du caractère dynamique d'une dimension peut conduire à des interprétations erronées.
- Warning: La façon de modéliser les dimensions dynamiques en analyse des durées peut conduire à des biais de causalité, en particulier dans les sciences sociales, en omettant les effets d'anticipation. C'est une situation classique avec des covariables dynamiques de type discrètes. Les techniques standards ne peuvent modéliser que des effets d'adaptation : la cause - observée - précède l'effet.

Variable dynamique traitée de manière fixe

On reprend l'exemple sur malformation cardiaque, en ajoutant la variable relative à la greffe. La question est donc: une transplantation du coeur réduit-elle le risque journalier de décéder (ou augmente la durée de survie).
On a dans la base 2 variables: une variable binaire pour savoir si l'individu à été greffé ou non, **transplant**, et la variable *wait* de type continue tronquée donnant la durée en jour jusqu'à l'opération depuis l'inscription dans le registre (0 si $transplant = 0$).
On va dans un premier temps estimer le modèle (de Cox) en introduisant les variables *transplant* et *wait*
.

Table 4.5: Variable dynamique traitée de manière fixe

	Hazard Ratio	**95% CI**	**p-value**
year	0.93	[0.81-1.1]	0.246
age	1.06	[1.03-1.09]	0.000
surgery	0.54	[0.17-1.3]	0.167
transplant	0.25	[0.13-0.47]	0.000
wait	0.99	[0.98-1.00]	0.139

Interprétation: traité de manière fixe, la greffe réduit donc sensiblement le risque journalier de décéder.

Au niveau des données le modèle à été estimé, pour une personne greffée (ici id=70), à partir de ce mapping (base splitté aux durée d'évènement):

id	year	age	surgery	transplant	wait	died	t	t_0
70	72	52	0	1	5	0	1	0
70	72	52	0	1	5	0	2	1
70	72	52	0	1	5	0	3	2
70	72	52	0	1	5	0	5	3
70	72	52	0	1	5	0	6	5
70	72	52	0	1	5	0	8	6
70	72	52	0	1	5	0	9	8
70	72	52	0	1	5	0	12	9
70	72	52	0	1	5	0	16	12
70	72	52	0	1	5	0	17	16
70	72	52	0	1	5	0	18	17
70	72	52	0	1	5	0	21	18
70	72	52	0	1	5	0	28	21
70	72	52	0	1	5	1	30	28

Problème: une personne est codée greffée avant le jour de la transplantation. L'effet " "" *causal* "" " est donc mal mesuré si sa dimension temporelle a été ignorée, ici le jour exact de l'opération. C'est le même principe pour l'évènement, la personne est codée décédée (1) le jour du décès, et vivante avant (0).

Estimation avec une variable dynamique

Il convient donc de modifier l'information avec le délai d'attente jusqu'à la greffe. Quel que soit le logiciel, le principe de construction de la variable suit la logique suivante:

$tvc = transplant$ **, si** $transplant = 1$ **et** $t < wait$ **alors** $tvc = 0$

id	year	age	surgery	transplant	wait	**TVC**	died	t	t_0
70	72	52	0	1	5	**0**	0	1	0
70	72	52	0	1	5	**0**	0	2	1
70	72	52	0	1	5	**0**	0	3	2
70	72	52	0	1	5	**0**	0	5	3
70	72	52	0	1	5	**1**	0	6	5
70	72	52	0	1	5	**1**	0	8	6
70	72	52	0	1	5	**1**	0	9	8
70	72	52	0	1	5	**1**	0	12	9
70	72	52	0	1	5	**1**	0	16	12
70	72	52	0	1	5	**1**	0	17	16
70	72	52	0	1	5	**1**	0	18	17
70	72	52	0	1	5	**1**	0	21	18
70	72	52	0	1	5	**1**	0	28	21
70	72	52	0	1	5	**1**	1	30	28

Maintenant, si on estime le modèle avec cette variable dynamique qui indique clairement le moment de la transition (jour de la greffe):

Table 4.8: Estimation avec une variable dynamique

	Hazard Ratio	**95% CI**	**p-value**
year	0.90	[0.78-1.01]	0.07
age	1.03	[1.00-1.06]	0.028
surgery	0.37	[0.16-0.88]	0.024
Greffe	0.92	[050.-0.67]	0.787

L'impact de la greffe apparaît maintenant bien plus contestable sur la survie des individus, tout du moins elle ne l'augmente pas. Cela ne signifie pas non plus que des personnes ont pu être "sauvée" grâce à cette opération (ou plutôt leur durée de vie augmentée), mais des complications lors de l'opération ou post-opératoire, surtout à une époque où ces techniques étaient à leurs balbutiements, ont pu également accélérer la mortalité de quelques jours ou quelques semaines par rapport à une personne non (encore) greffée.

R

La base doit être transformée en format long aux temps d'évènement (`survsplit`) avant la création de la variable dynamique. Le modèle est estimé de manière classique mais en précisant dans la formule les variables qui définissent le début et la fin de chaque intervalle.

5 Programmation avec R

Programme de cette section: Lien

```
options(scipen=999)  # empêcher le format scientifique
options(show.signif.stars=FALSE)
```

5.1 Packages et fonctions

Analyse	Packages - fonctions
Non paramétrique	• `discsurv` – `lifetable` – `contToDisc` • `survival` – `survfit` – `survdif` • `survRM2` – `rmst2`
Modèles à risques proportionnel	• `survival` – `coxph` – `cox.zph` (v3) `cox.zphold` (récupération v2) – `survsplit` • base et `tydir` – `uncount` – `glm`
Modèles paramétriques (ph ou aft)	• `survival` – `survreg` • `flexsurv` – `survreg`

Analyse	Packages - fonctions
Risques concurents	• `cmprsk` – `cuminc` • `nnet` – `multinom`
Autres (graphiques - mise en forme)	• `survminer` • `jtools` • `stargazer` - `gtsummary`

Installation des packages

```
#install.packages("survival")
#install.packages("survminer")
#install.packages("survRM2")
#install.packages("gtsummary")
#install.packages("muhaz")
#install.packages(discSurv)
library(survival)
library(survminer)
library(survRM2)
library(muhaz)
library(discSurv)
library(gtsummary)
```

! Survival v2 versus v3: test de Grambsch-Therneau

Se reporter à la partie sur le test de Grambsch-Therneau (hypothèse de proportionalité des risques). La solution la plus simple est de récupérer et d'exécuter la fonction du test de la version 2 du package **survival**.
La marche à suivre, qui est plutôt simple, est donnée dans la section dédié à ce test.

Chargement de la base transplantation

```
library(readr)
trans <- read_csv("https://raw.githubusercontent.com/mthevenin/
analyse_duree/master/bases/transplantation.csv")
```

5.2 Analyse Non paramétrique

5.2.1 Méthode actuarielle

La fonction disponible du paquet `discsurv`, *`lifetable`*, a des fonctionalités plutôt limitées. Si on peut depuis une MAJ récente définir des intervalles de durée, il n'y a toujours pas d'estimateurs les différents quantiles de la courbe de survie.

Pire, la programmation est rendue un peu compliquée pour pas grand chose. On doit s'assurer que la base est bien en format `data.frame`.
Je donne les codes pour info, sans plus de commentaires.

```
trans = as.data.frame(trans)
```

Fonction `lifeTable`

Intervalle par defaut $dJ = 1$

```
lt = lifeTable(dataShort=trans, timeColumn="stime", eventColumn = "died")

plot(lt, x = 1:dim(lt$Output)[1], y = lt$Output$S, xlab = "Intervalles t = journalier", ylab=
```

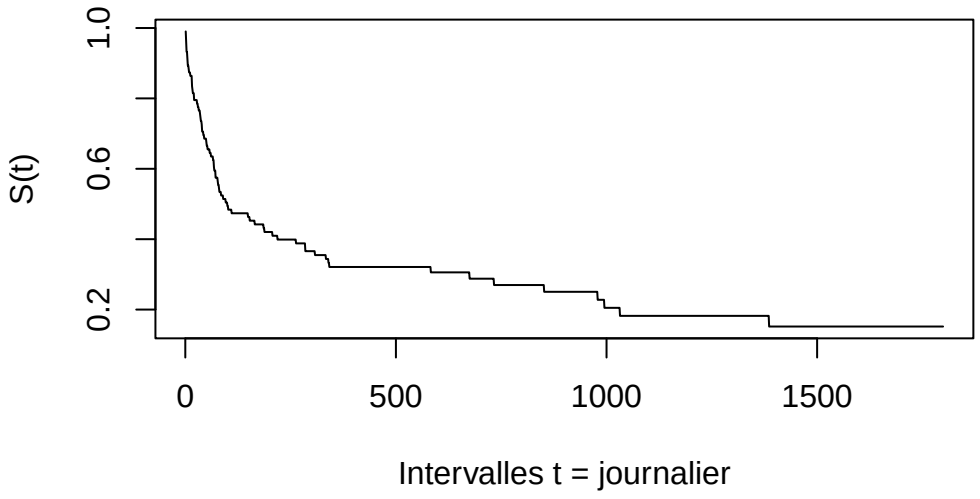

Figure 5.1: S(t) méthode actuarielle avec `discSurv` (1)

Intervalle $dJ = 30$

```
#  Vecteur qui définit la longueur des intervalles (il n'y avait pas plus simple????)
J <- 1:ceiling(max(trans$stime)/30)*30

# Base dis avec une nouvelle variable de durée => timeDisc
dis <- contToDisc(dataShort=trans, timeColumn="stime", intervalLimits = J )

lt <- lifeTable(dataShort=dis, timeColumn="timeDisc", eventColumn = "died")

plot(lt, x = 1:dim(lt$Output)[1], y = lt$Output$S, xlab = "Intervalles t = 30 jours", ylab="S
```

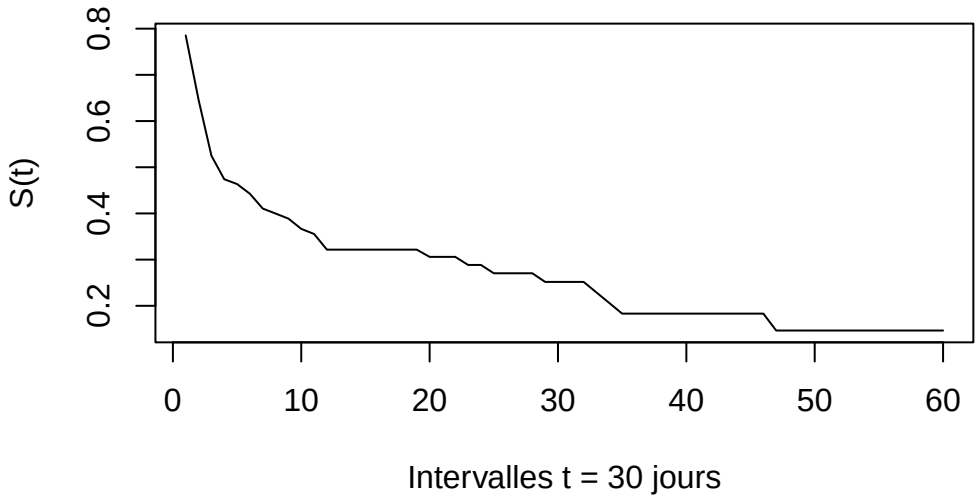

Figure 5.2: S(t) méthode actuarielle avec `discSurv` (2)

Sur les abcisses, ce sont les valeurs des intervalles qui sont reportés: 10=300 jours. Ce n'est vraiment pas terrible. Pour ce type d'estimateurs, il est préférable d'utiliser Sas ou Stata.

5.2.2 Méthode Kaplan-Meier

Le package **`survival`** est le principal outil d'analyse des durée. Le package **`survminer`** permet d'améliorer la présentation des graphiques des fonctions de survie, et il a également quelques ajouts intéressant pour les tests du log-rank. En particulier, il permet d'effectuer des tests deux à deux lorsqu'une variable catégorielle a plus de deux modalités.
survminer1
survminer2

Les RMST sont estimées avec le package **`survRM2`** et si on souhaite visualiser les estimateurs lissés de la fonction de risque, on peut utiliser **`cumhaz`**.

5.2.2.1 Estimation des fonctions de survie

Fonction **`survfit`**

Syntaxe:

```
fit <- survfit(Surv(time, status) ~ x, data = base)
```

On peut renseigner directement les variables permettant de calculer la durée et non la variable de durée elle-même. On en aura besoin pour l'introduction d'une variable dynamique dans un modèle semi-paramétrique (`coxph`).

```
fit <- survfit(Surv(variable_start, variable_end, status) ~ x, data = nom_base)
```

Sans comparaison de groupes:

```
fit <- survfit(Surv(stime, died) ~ 1, data = trans)
fit
```

```
Call: survfit(formula = Surv(stime, died) ~ 1, data = trans)

       n events median 0.95LCL 0.95UCL
[1,] 103     75    100      72     263
```

```
summary(fit)
```

```
Call: survfit(formula = Surv(stime, died) ~ 1, data = trans)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1    103       1    0.990 0.00966        0.9715        1.000
    2    102       3    0.961 0.01904        0.9246        0.999
    3     99       3    0.932 0.02480        0.8847        0.982
    5     96       2    0.913 0.02782        0.8597        0.969
    6     94       2    0.893 0.03043        0.8355        0.955
    8     92       1    0.883 0.03161        0.8237        0.948
    9     91       1    0.874 0.03272        0.8119        0.940
   12     89       1    0.864 0.03379        0.8002        0.933
   16     88       3    0.835 0.03667        0.7656        0.910
   17     85       1    0.825 0.03753        0.7543        0.902
   18     84       1    0.815 0.03835        0.7431        0.894
   21     83       2    0.795 0.03986        0.7208        0.877
   28     81       1    0.785 0.04056        0.7098        0.869
   30     80       1    0.776 0.04122        0.6989        0.861
   32     78       1    0.766 0.04188        0.6878        0.852
   35     77       1    0.756 0.04250        0.6769        0.844
   36     76       1    0.746 0.04308        0.6659        0.835
   37     75       1    0.736 0.04364        0.6551        0.827
```

39	74	1	0.726	0.04417	0.6443	0.818
40	72	2	0.706	0.04519	0.6225	0.800
43	70	1	0.696	0.04565	0.6117	0.791
45	69	1	0.686	0.04609	0.6009	0.782
50	68	1	0.675	0.04650	0.5902	0.773
51	67	1	0.665	0.04689	0.5796	0.764
53	66	1	0.655	0.04725	0.5690	0.755
58	65	1	0.645	0.04759	0.5584	0.746
61	64	1	0.635	0.04790	0.5479	0.736
66	63	1	0.625	0.04819	0.5374	0.727
68	62	2	0.605	0.04870	0.5166	0.708
69	60	1	0.595	0.04892	0.5063	0.699
72	59	2	0.575	0.04929	0.4857	0.680
77	57	1	0.565	0.04945	0.4755	0.670
78	56	1	0.554	0.04958	0.4654	0.661
80	55	1	0.544	0.04970	0.4552	0.651
81	54	1	0.534	0.04979	0.4451	0.641
85	53	1	0.524	0.04986	0.4351	0.632
90	52	1	0.514	0.04991	0.4251	0.622
96	51	1	0.504	0.04994	0.4151	0.612
100	50	1	0.494	0.04995	0.4052	0.602
102	49	1	0.484	0.04993	0.3953	0.592
110	47	1	0.474	0.04992	0.3852	0.582
149	45	1	0.463	0.04991	0.3749	0.572
153	44	1	0.453	0.04987	0.3647	0.562
165	43	1	0.442	0.04981	0.3545	0.551
186	41	1	0.431	0.04975	0.3440	0.541
188	40	1	0.420	0.04966	0.3336	0.530
207	39	1	0.410	0.04954	0.3233	0.519
219	38	1	0.399	0.04940	0.3130	0.509
263	37	1	0.388	0.04923	0.3027	0.498
285	35	2	0.366	0.04885	0.2817	0.475
308	33	1	0.355	0.04861	0.2713	0.464
334	32	1	0.344	0.04834	0.2610	0.453
340	31	1	0.333	0.04804	0.2507	0.442
342	29	1	0.321	0.04773	0.2401	0.430
583	21	1	0.306	0.04785	0.2252	0.416
675	17	1	0.288	0.04830	0.2073	0.400
733	16	1	0.270	0.04852	0.1898	0.384
852	14	1	0.251	0.04873	0.1712	0.367
979	11	1	0.228	0.04934	0.1491	0.348
995	10	1	0.205	0.04939	0.1279	0.329
1032	9	1	0.182	0.04888	0.1078	0.308
1386	6	1	0.152	0.04928	0.0804	0.287

```
plot(fit)
```

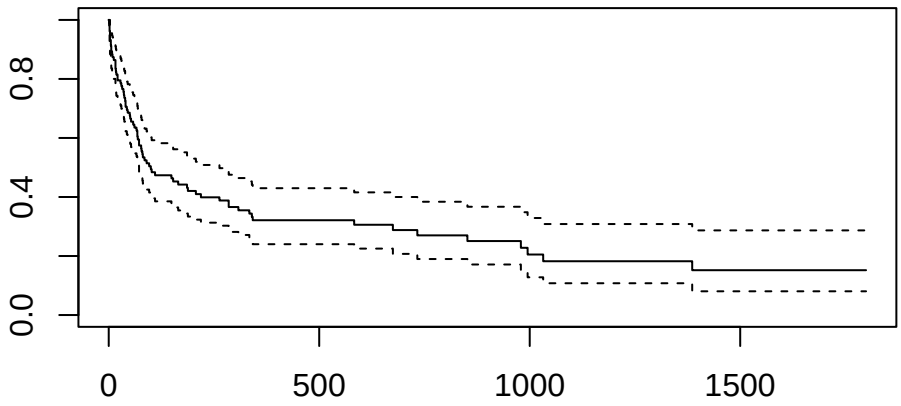

Figure 5.3: S(t) méthode Kaplan-Meier avec `survfit` (1)

Le premier output `fit` permet d'obtenir la durée médiane, ici égale à 100 ($S(100) = 0.494$). Le second avec la fonction **`summary`** permet d'obtenir une table des estimateurs. La fonction de survie peut être tracée avec la fonction **`plot`** (en pointillés les intervalles de confiance).

On peut obtenir des graphes de meilleur qualité avec la librairie **`survminer`**, avec la fonction **`ggsurvplot`**

```
ggsurvplot(fit, conf.int = TRUE)
```

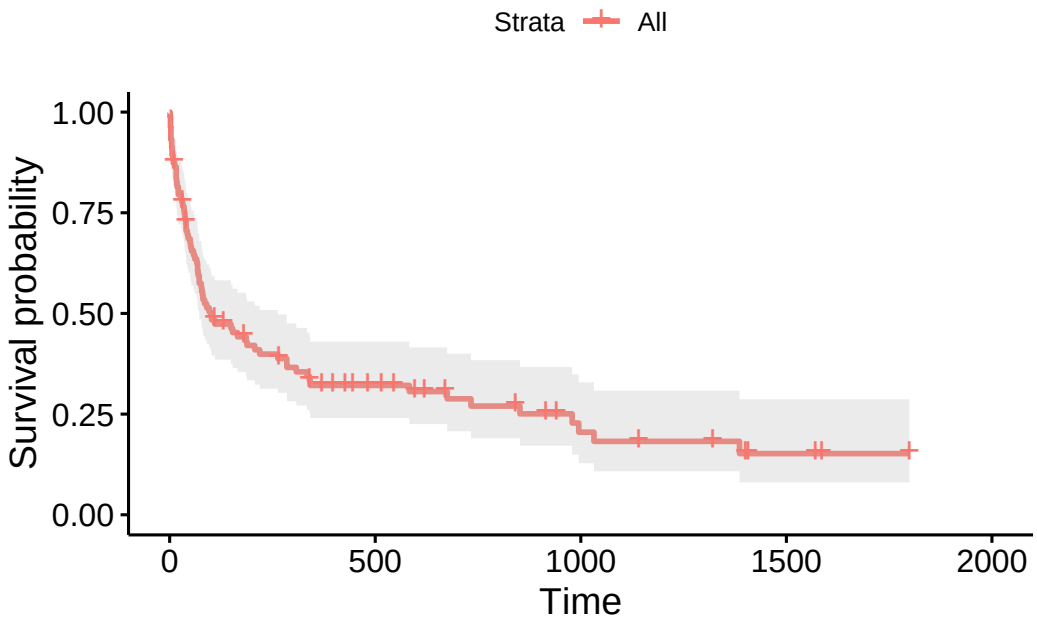

Figure 5.4: S(t) méthode Kaplan-Meier avec `survfit` (2)

Visualisation de la fonction de risque

On utilise la fonction `muhaz` du package du même nom et on trace la courbe avec la fonction `plot`. Les estimateurs étant lissés, on peut paramétrer la méthode et les fenêtre de lissage. Ici tout est estimés avec les options par défaut.

```
library(muhaz)

haz = muhaz(trans$stime,trans$died)
plot(haz)
```

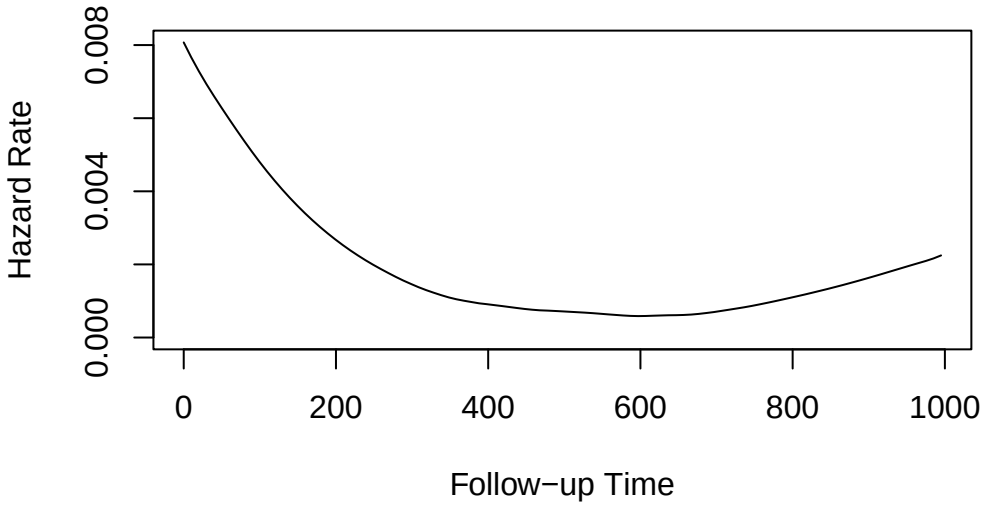

Figure 5.5: h(t) avec `muhaz`

5.2.2.2 Comparaison des fonctions de survie

On va comparer les deux fonctions de survie pour la variable *surgery*, celle pour les personnes non opérées et celle pour les personnes opérées.

```
fit <- survfit(Surv(stime, died) ~ surgery, data = trans)
fit
```

```
Call: survfit(formula = Surv(stime, died) ~ surgery, data = trans)

           n events median 0.95LCL 0.95UCL
surgery=0 91     69     78      61     153
surgery=1 12      6    979     583      NA
```

```
ggsurvplot(fit, conf.int = TRUE, risk.table = TRUE)
```

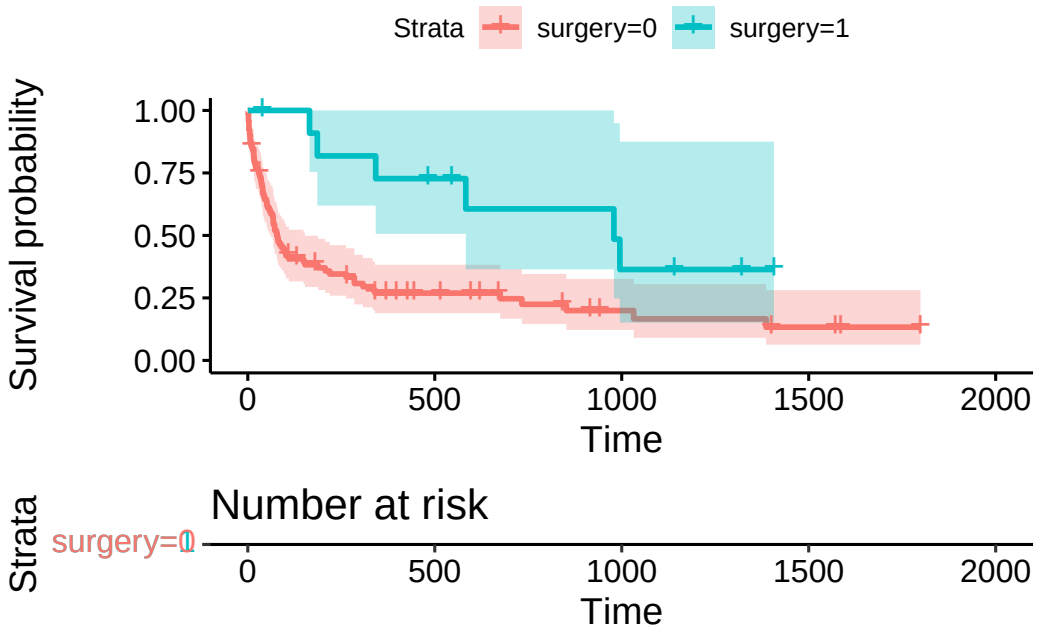

Figure 5.6: Comparaison de S(t) avec `survfit`

Tests du log-rank

On utilise la fonction **`survdiff`**, avec comme variante le test de Peto-Peto (`rho=1`).
La syntaxe est quasiment identique à la fonction `survdiff`.

```
survdiff(Surv(stime, died) ~ surgery, rho=1, data = trans)
```

```
Call:
survdiff(formula = Surv(stime, died) ~ surgery, data = trans,
    rho = 1)

           N Observed Expected (O-E)^2/E (O-E)^2/V
surgery=0 91    45.28    39.12     0.968      8.65
surgery=1 12     2.03     8.18     4.630      8.65

 Chisq= 8.7  on 1 degrees of freedom, p= 0.003
```

Ici la variable est binaire. Si on veux tester deux à deux les niveaux d'une variable catégorielle à plus de deux modalités, on utilise la fonction **`pairwise_survdiff`** de `survminer` (syntaxe identique que `survdiff`). [Voir avec le TP]

Comparaison des RMST

La fonction **rmst2** du package **survRM2** permet de comparer les RMST entre 2 groupes (et pas plus). La strate pour les comparaisons doit être renommée *arm*. La fonction, issue d'une commande de Stata, n'est donc pas très souple.

```
trans$arm=trans$surgery
a=rmst2(trans$stime, trans$died, trans$arm, tau=NULL)
print(a)
```

```
The truncation time, tau, was not specified. Thus, the default tau  1407  is used.

Restricted Mean Survival Time (RMST) by arm
                Est.      se lower .95 upper .95
RMST (arm=1) 884.576 151.979   586.702  1182.450
RMST (arm=0) 379.148  58.606   264.283   494.012


Restricted Mean Time Lost (RMTL) by arm
                 Est.      se lower .95 upper .95
RMTL (arm=1)  522.424 151.979   224.550   820.298
RMTL (arm=0) 1027.852  58.606   912.988  1142.717


Between-group contrast
                        Est. lower .95 upper .95     p
RMST (arm=1)-(arm=0) 505.428   186.175   824.682 0.002
RMST (arm=1)/(arm=0)   2.333     1.483     3.670 0.000
RMTL (arm=1)/(arm=0)   0.508     0.284     0.909 0.022
```

```
plot(a)
```

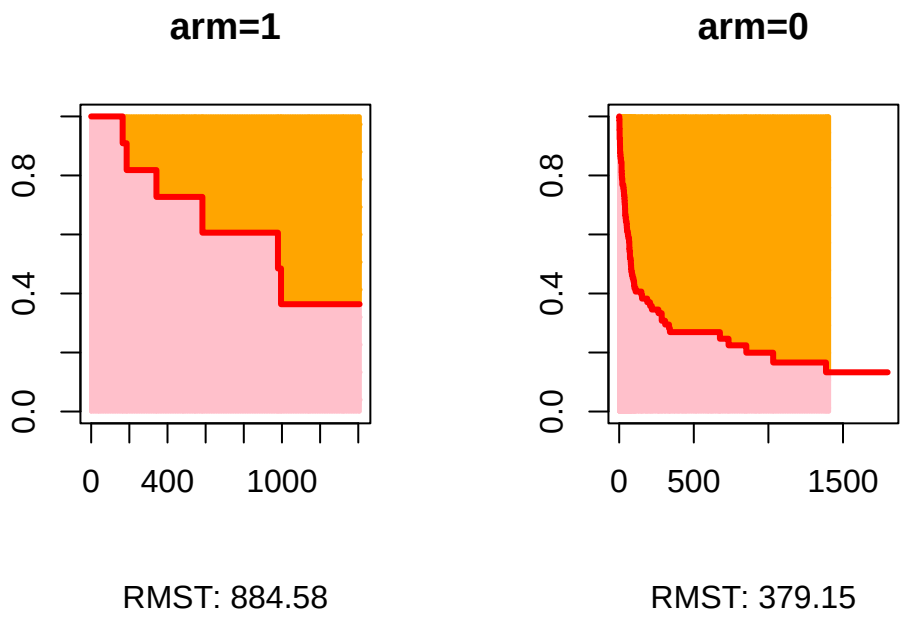

Figure 5.7: Rmst avec `rmst2`

5.3 Modèle de Cox

Ici tout est estimé avec des fonctions du package `survival`:

- Estimation du modèle: `coxph`.
- Test de Grambsch-Therneau: `cox.zph`.
- Introduction d'une variable dynamique: `survsplit`.

5.3.1 Estimation du modèle

Par défaut, R utilise la correction d'Efron pour les évènements simultanés. Il est préférable de ne pas la modifier.

Syntaxe:

```
coxph(Surv(time, status) ~ x1 + x2 + ....., data=base, ties="nom_correction"))
```

```
coxfit = coxph(formula = Surv(stime, died) ~ year + age + surgery, data = trans)
summary(coxfit)
```

```
Call:
coxph(formula = Surv(stime, died) ~ year + age + surgery, data = trans)

  n= 103, number of events= 75

            coef exp(coef) se(coef)      z Pr(>|z|)
year    -0.11963   0.88725  0.06734 -1.776   0.0757
age      0.02958   1.03002  0.01352  2.187   0.0287
surgery -0.98732   0.37257  0.43626 -2.263   0.0236

        exp(coef) exp(-coef) lower .95 upper .95
year       0.8872     1.1271    0.7775    1.0124
age        1.0300     0.9709    1.0031    1.0577
surgery    0.3726     2.6840    0.1584    0.8761

Concordance= 0.653  (se = 0.032 )
Likelihood ratio test= 17.63  on 3 df,   p=0.0005
Wald test            = 15.76  on 3 df,   p=0.001
Score (logrank) test = 16.71  on 3 df,   p=0.0008
```

- Le tableau des résultats reporte le logarithme des Risques Ratios (coef) ainsi que les RR (exp(coef)). Il est intéressant de regarder la valeur de concordance (Harrel's) qui donne des indications sur la qualité de l'ajustement (proche de l'AUC/ROC).
- Le tableau par défaut offre également un basculement des points de référence (exp(-coef)), c'est intéressant lorsqu'on veut éviter de lire des RR<1.
  - pour les variables quantitatives, $dx = 1$ devient $dx = -1$: par exemple au lieu de comparer le risque d'une personne d'un âge donnée à celui d'une personne âgée d'un an de moins, on compare le risque d'une personne d'un âge donné à une personne âgée d'un an de plus.
  - pour les variables discrètes, on permute la modalité de référence pour chaque variable: ici exp(-coef) pour la variable surgery donne le RR comparant les personnes non opérées aux personnes opérées. on peut dire dans ce cas que le risque de décéder chaque jour des personnes non opérées est 2.7 fois plus élevé que celui des personnes opérées.

Fonction ggforest de survminer

```
ggforest(coxfit)
```

```
Warning in .get_data(model, data = data): The `data` argument is not provided.
Data will be extracted from model fit.
```

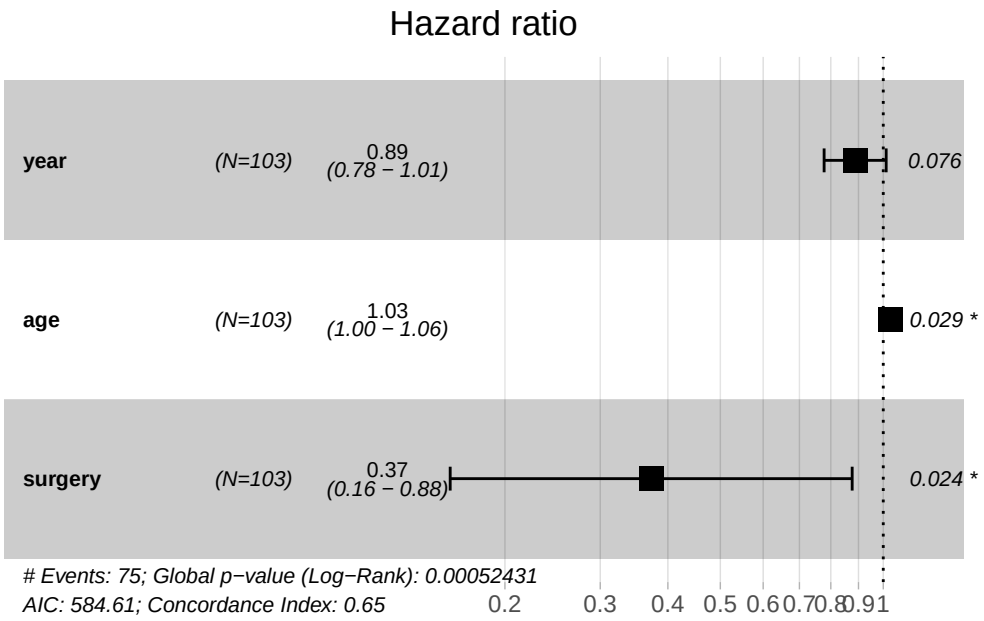

Figure 5.8: Modèle de Cox avec `ggforest`

5.3.2 Hypothèse PH

5.3.2.1 Résidus de Schoenfeld

> **Warning**
>
> Attention aux résultats entre la v2 et la v3 de survival.
> Je conseille vivement d'utiliser le test de la v2 avec des durées discrète, dont j'ai récupéré et renommé la fonction. Ce n'est pas forcément le cas avec l'exemple du cours. Les tests des deux versions sont donc présentées.
> A ce jour, il est important si on présente des résultats reposant sur ce test de préciser s'il s'agit du test exact (v3) ou du test simplifié (v2 et identique ***Sas***, ***Stata***, ***Python***).

Test de la v3: `cox.zph()`

On utilise la fonction **`cox.zph`** pour le test de *Grambsch-Therneau*.
Le test peut utiliser plusieurs fonctions de la durée. Par défaut la fonction utilise $1 - KM$, soit le complémentaire de l'estimateur de Kaplan-Meier (option `transform="km"`).

Avec `transform="km"`

```
cox.zph(coxfit)
```

```
        chisq df     p
year    3.309  1 0.069
age     0.922  1 0.337
surgery 5.494  1 0.019
GLOBAL  8.581  3 0.035
```

Avec `transform="identity"` ($f(t) = t$) [remarque: solution de Stata par défaut].

```
cox.zph(coxfit, transform="identity")
```

```
        chisq df     p
year     4.54  1 0.033
age      1.71  1 0.191
surgery  4.92  1 0.027
GLOBAL   9.47  3 0.024
```

Avec la v3 l'option `terms=FALSE` permet d'avoir un test par modalité si la variable catégorielle en a plus de deux, ce que je conseille (par défaut la nouvelle fonction donne un test multiple sur chaque variable de ce type).

Test de la v2: `cox.zphold()`

- J'ai récupéré la fonction de la version précédente de survival qui a été renommée **`cox.zphold`**. Elle est téléchargeable à cette adresse: **https://github.com/mthevenin/analyse_duree/tree/main/cox.zphold)**.

- Une fois enregistré, charger dans le programme d'analyse la fonction avec la fonction `source(path/cox.zphold.R)`.
- Après avoir estimé le modèle de Cox, exécuter la fonction **`cox.zphold()`** :::

```
source("D:/D/Marc/SMS/FORMATIONS/2022/Durée2/a distribuer/cox.zphold.R")

coxfit = coxph(formula = Surv(stime, died) ~ year + age + surgery,   data = trans)
```

```
cox.zphold(coxfit)
```

```
          rho chisq      p
year    0.159  1.96 0.1620
age     0.109  1.15 0.2845
surgery 0.251  3.96 0.0465
GLOBAL     NA  7.99 0.0462
```

```r
cox.zphold(coxfit, transform="identity")
```

```
          rho chisq      p
year    0.102 0.797 0.3720
age     0.129 1.612 0.2043
surgery 0.297 5.539 0.0186
GLOBAL     NA 8.756 0.0327
```

i Régression linéaire sur les résidus de Schoenfeld

Pour information:

```r
resid= resid(coxfit, type="scaledsch")
varnames <- names(coxfit$coefficients)
coln = c(varnames)
colnames(resid) = c(coln)

times    = as.numeric(dimnames(resid)[[1]])

resid = data.frame(resid)
resid = cbind(resid, t=times)

year    = summary(lm(year~t, data=resid))
age     = summary(lm(age~t, data=resid))
surgery = summary(lm(surgery~t, data=resid))

#####################
# p-values de l'OLS #
#####################

paste("p-value pour year:",    year$coefficients[2,4])
```

```
[1] "p-value pour year: 0.38565336851861"
```

```r
paste("p-value pour age:",     age$coefficients[2,4])
```

```
[1] "p-value pour age: 0.268640363261224"
```

```r
paste("p-value pour surgery:", surgery$coefficients[2,4])
```

```
[1] "p-value pour surgery: 0.00975820981508909"
```

On retrouve bien les résultats de la version *simplifiée* du test

5.3.2.2 Introduction d'une intéraction

Lorsque la covariable n'est pas continue, elle doit être transformée en indicatrice. Vérifier que les résultats du modèle sont bien identiques avec le modèle estimé précédemment (ne pas oublier d'omettre le niveau en référence).
Ici la variable *surgery* est déjà sous forme d'indicatrice (0,1).
La variable d'intéraction est **`tt(nom-variable)`**, la fonction de la durée (ici forme linéaire simple) est indiquée en option de la fonction: **`tt = function(x, t, ...) x*t`**.

```
coxfit2 = coxph(formula = Surv(stime, died) ~ year + age + surgery + tt(surgery),
          data = trans, tt = function(x, t, ...) x*t)
summary(coxfit2)
```

```
Call:
coxph(formula = Surv(stime, died) ~ year + age + surgery + tt(surgery),
    data = trans, tt = function(x, t, ...) x * t)

  n= 103, number of events= 75

                coef exp(coef)  se(coef)      z Pr(>|z|)
year        -0.123074  0.884198  0.066835 -1.841  0.06555
age          0.028888  1.029310  0.013449  2.148  0.03172
surgery     -1.754738  0.172953  0.674391 -2.602  0.00927
tt(surgery)  0.002231  1.002234  0.001102  2.024  0.04299

            exp(coef) exp(-coef) lower .95 upper .95
year           0.8842     1.1310   0.77564    1.0080
age            1.0293     0.9715   1.00253    1.0568
surgery        0.1730     5.7819   0.04612    0.6486
tt(surgery)    1.0022     0.9978   1.00007    1.0044

Concordance= 0.656  (se = 0.032 )
Likelihood ratio test= 21.58  on 4 df,   p=0.0002
Wald test            = 16.99  on 4 df,   p=0.002
Score (logrank) test = 19  on 4 df,   p=0.0008
```

Rappel: le paramètre estimé pour **`tt(surgery)`** ne reporte pas un Risques Ratio, mais un rapport de Risques Ratios.

5.3.3 Variable dynamique (binaire)

La dimension dynamique est le fait d'avoir été opéré pour une greffe du coeur.

- **Etape 1**: créer un vecteur donnant les durées aux temps d'évènement.
- **Etape 2**: appliquer ce vecteurs de points de coupure à la fonction `survsplit`.
- **Etape 3**: modifier la variable transplant (ou créer une nouvelle) à l'aide de la variable **wait** qui prend la valeur 1 à partir du jour de la greffe, 0 avant.

Etape 1 Création de l'objet cut (vecteur)

```
cut= unique(trans$stime[trans$died == 1])
```

Etape 2

```
tvc = survSplit(data = trans, cut = cut, end = "stime", start = "stime0", event = "died")
```

Remarque: pour estimer le modèle de Cox de départ avec cette base longue.

```
coxph(formula = Surv(stime0, stime, died) ~ year + age + surgery, data = tvc)
```

```
Call:
coxph(formula = Surv(stime0, stime, died) ~ year + age + surgery,
    data = tvc)

            coef exp(coef) se(coef)      z      p
year    -0.11963   0.88725  0.06734 -1.776 0.0757
age      0.02958   1.03002  0.01352  2.187 0.0287
surgery -0.98732   0.37257  0.43626 -2.263 0.0236

Likelihood ratio test=17.63  on 3 df, p=0.0005243
n= 3573, number of events= 75
```

Etape 3

```
tvc$tvc=ifelse(tvc$transplant==1 & tvc$wait<=tvc$stime,1,0)
```

Estimation du modèle
En format long, on doit préciser dans la formule l'intervalle de durée avec les variables stime0 (début) et stime(fin)

```
tvcfit = coxph(formula = Surv(stime0, stime, died) ~ year + age + surgery + tvc,
         data = tvc)
summary(tvcfit)
```

```
Call:
coxph(formula = Surv(stime0, stime, died) ~ year + age + surgery +
    tvc, data = tvc)

  n= 3573, number of events= 75

            coef exp(coef) se(coef)      z Pr(>|z|)
year    -0.12032   0.88664  0.06734 -1.787   0.0740
age      0.03044   1.03091  0.01390  2.190   0.0285
surgery -0.98289   0.37423  0.43655 -2.251   0.0244
tvc     -0.08221   0.92108  0.30484 -0.270   0.7874

        exp(coef) exp(-coef) lower .95 upper .95
year       0.8866      1.128    0.7770    1.0117
age        1.0309      0.970    1.0032    1.0594
surgery    0.3742      2.672    0.1591    0.8805
tvc        0.9211      1.086    0.5068    1.6741

Concordance= 0.659  (se = 0.032 )
Likelihood ratio test= 17.7  on 4 df,   p=0.001
Wald test            = 15.79  on 4 df,   p=0.003
Score (logrank) test = 16.74  on 4 df,   p=0.002
```

```
ggforest(tvcfit)
```

```
Warning in .get_data(model, data = data): The `data` argument is not provided.
Data will be extracted from model fit.
```

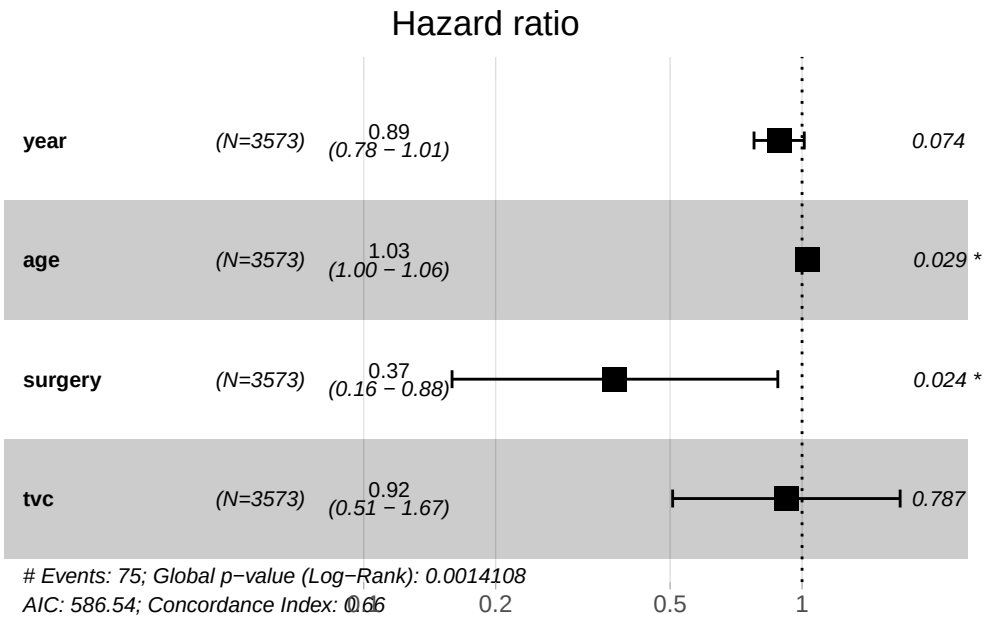

Figure 5.9: Modèle de Cox avec variable dynamique avec `ggforest`