Introduction à l'analyse des durées

Support de formation 2023

Marc Thévenin

2023-08-04

Table des matières

Liste des Figures

Liste des Tables

Le support

MISE A JOUR DU SUPPORT EN COURS

- L'ensemble des sections devraient être disponible la semaine du 07 Aout 2022
- Le support est maintenant intégralement disponible en format pdf en cliquant sur l'icône au dessus de la barre de recherche.
- Quelques mises à jour de contenu seront réalisées d'ici la fin septembre:

Bibliographie vf

Les éléments bibliographiques qui figurent ci-dessous proviennent du champ des sciences sociales. Elle est courte, mais efficace. Quelle que soit la langue, le nombre de cours ou support sont très nombreux en médecine, qui est ici l'espace privilégié de l'ingénierie méthodologique. On trouve également de (trop) nombreux tutoriels à dominante *mise en pratique avec R*.

Accès en ligne

- **Cours Gilbert Colletaz** (Université d'Orléans - master d'économétrie).
  - Le cours est mis à jour tous les ans, applications uniquement avec Sas.
  - Dernière version 2020: lien
- **Document de travail de Simon Quantin** (Insee).
  - Couvre l'ensemble des techniques de base d'analyse des durées en durée dite continue. Il propose surement la meilleure introduction en langue française à la problématique de la *fragilité*.
  - Application en R seulement (attention au passage de la v3 du package `survival`)
  - 2019 - pas de mise à jour: lien

Ouvrage de référence en démographie:
L'analyse démographique des biographies de *Daniel Courgeau* et *Eva Lelièvre* (Edition de l'Ined - 1989).

Outils

- Support réalisé sous Rstudio avec le système de publication Quarto
- Langages utilisés pour la partie programmation:
  - R
  - Stata v18
  - Sas
  - Python

partie I

Introduction

1 L'analyse des durées

1.1 Questions

On dispose de données dites "longitudinales", et on cherche à appréhender l'occurence d'un évènement au sein d'une population. 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*)

1.2 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	

1.3 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…

1.4 Elements nécessaire à l'analyse

1. **Un processus temporel**

   - Une échelle de mesure ou métrique temprelle: 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 avec la fin de la période d'exposition au risque.

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: *Time varying covariates*): Mesurées à tout moment entre le début et la sortie de l'observation: statut matrimonial, taille du ménage, statut d'activité…

partie II

Données et théorie

2 Les Données

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

2.1 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).

2.2 Grille AGEVEN

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…

i Références

- 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<

2.3 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.

2.3.1 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.

2.3.2 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

2.3.3 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.

2.4 Exemples de mise à disposition

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

2.4.1 Enquête biographie et entourage (Ined)

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

Figure 2.1: Biographie et entourage: base caractéristiques individuelle

VIEWTABLE: TMP1.tego

	Identifiant questionniare	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	EMILOWNA	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		309	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

2.4.2 Enquête MAFE (Ined)

Figure 2.2: Biographie et entourage: base biographique logements

	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 2.3: MAFE: base caractéristiques individuelles

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: MAFE: base biographique logement

ident	num_log	q301d	q301f	q302	q303	age_survey	q1a
E1	1	1972	1975	SENEGAL	Namanieque	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

3 La 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**.

3.1 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)**

3.2 Le Risk Set

1. Il s'agit de la population "soumise" ou "exposée" au risque lorsque $T = t_i$.
2. 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.

3.3 La Censure

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

3.3.1 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.

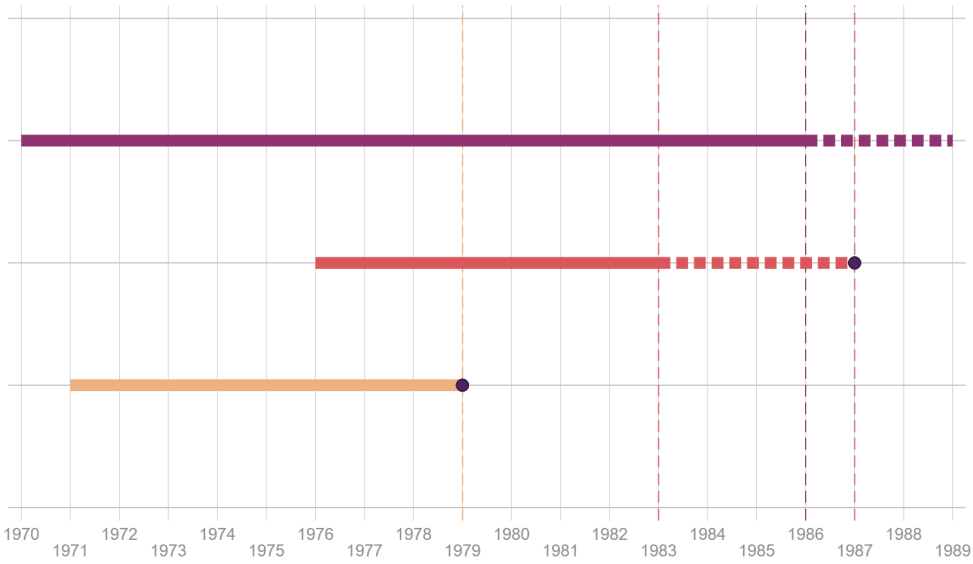
Hypothèse fondamentale
Les observations censurées ont vis à vis du phénomène observé le même comportement que les observations non censurées. On dit que la **censure est non informative**. Elle ne dépend pas de l'évènement analysé. Normalement le problème ne se pose pas dans les recueil retrospectif.

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 (G.Bellis) donne une autre 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, on note qu'ils/elles 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 (voir section dédiée).

Les graphiques suivant représentent, en temps calendaire et après sa transformation en durée, la logique des censures à droite. Le recueil des informations est ici de nature prospectives.

- Trait plein : durée observée
- Pointillés : durée censurée
- Bulle : moment de l'évènement

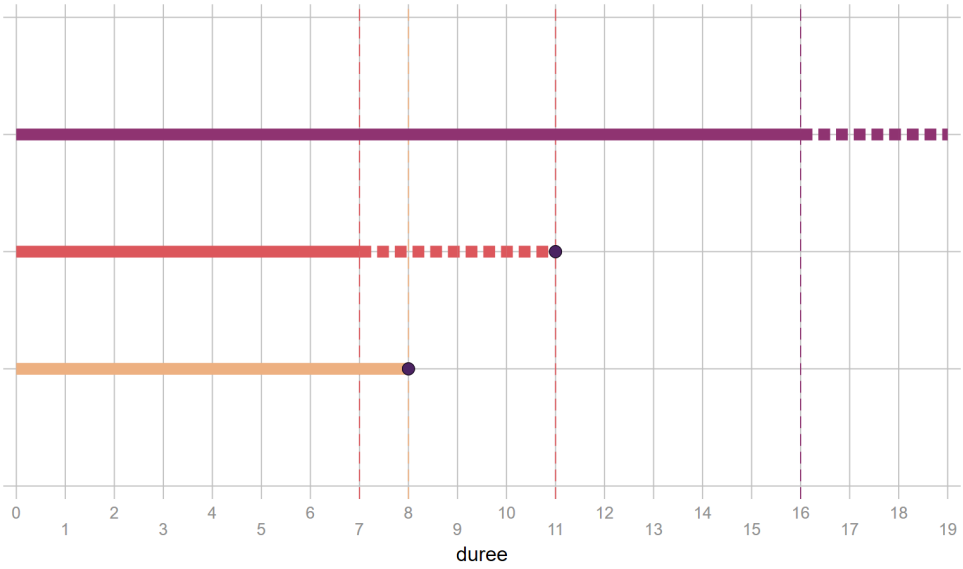
Figure 3.1: Schéma évènement/censure en temps calendaire



3.3.2 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

Figure 3.2: Schéma évènement/censure sous forme de durée



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.

3.4 Les grandeurs

3.4.1 Les grandeurs utilisées

- La fonction de survie: $S(t)$

- La fonction de répartition: $F(t)$
- La fonction de densité: $f(t)$
- Le risque *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 pas des estimateurs, mais des grandeurs dont on précisera les propriétés. Les techniques d'estimations devront respecter ces propriétés .

3.4.2 La fonction de Survie $S(t)$

Dans ce type d'analyse, il est courant d'analyser la courbe dite de survie. Hors contexte de mortalité on peut préférer la notion de **courbe de de séjour** (Courgeau, Lelièvre).

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 donc "survécu".

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

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

Propriétés:

- $S(0) = 1$
- $\lim_{t \to \infty} S(t) = 0$

La fonction de survie est donc strictement non croissante.

3.4.3 La fonction de répartition $F(t)$

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

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

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

La fonction de survie et la fonction de répartition sont donc deux grandeurs strictement complémentaires et décrivent la même information.

Propriétés:

- $F(0) = 0$
- $\lim_{t \to \infty} F(t) = 1$

3.4.4 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:

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

3.5 Le risque instantané $h(t)$

Concept fondamental de l'analyse des durées:

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

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

- En divisant par dt, La quantité obtenue donne alors un nombre moyen d'évènements que connaîtrait un individu durant un très court intervalle de temps.
- 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 remarque que la fonction de risque n'est pas une probabilité car $\frac{f(t)}{S(t)}$ ne peut pas contraindre la valeur obtenue à ne pas être supérieure à 1.

3.5.1 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 réécrire toutes les autres quantités à partir de celle-ci:

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

Exemple avec la loi exponentielle (risque constant)

Si on pose que le risque est strictement constant au cours du temps: $h(t) = a$. Cette forme du risque suit une **loi exponentielle**. Cette situation est, par exemple, typique des processus dits sans mémoire comme la durée de vie des ampoules:

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

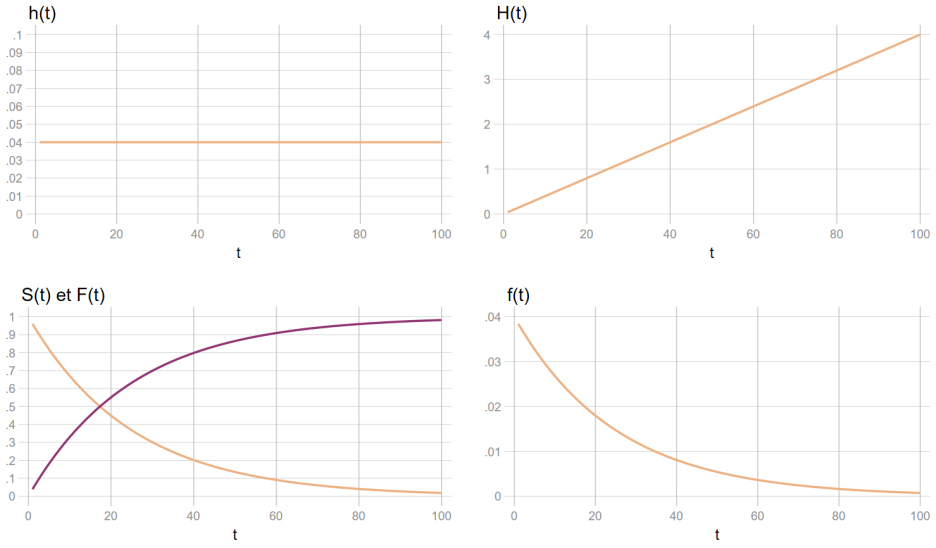
Application: risque et échelles temporelles:

Attention on sort ici très clairement de la durée continue, il s'agit seulement de manipuler les concepts et de voir la dépendance de la mesure du risque à l'échelle temporelle choisie ou 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 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 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'il est 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.

Figure 3.3: Grandeurs de la loi exponentielle avec h(t)=0.04



Le risque étant constant, on peut prendre son inverse 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, ce risque indépendamment de la durée étant par définition égal à 1, mais jusqu'à quand on peut espérer survivre.

- Pour le rhume, la durée moyenne est de $(1.44^{-1} = 0.69$ du trimestre hivernal, soit approximativement le début du mois de mars.
- Pour l'année sabbatique, la durée moyenne de survie est de $3.96^{-1} = 0.25$ 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

Durée	Nombre de décès
9	2
10	1

N=100

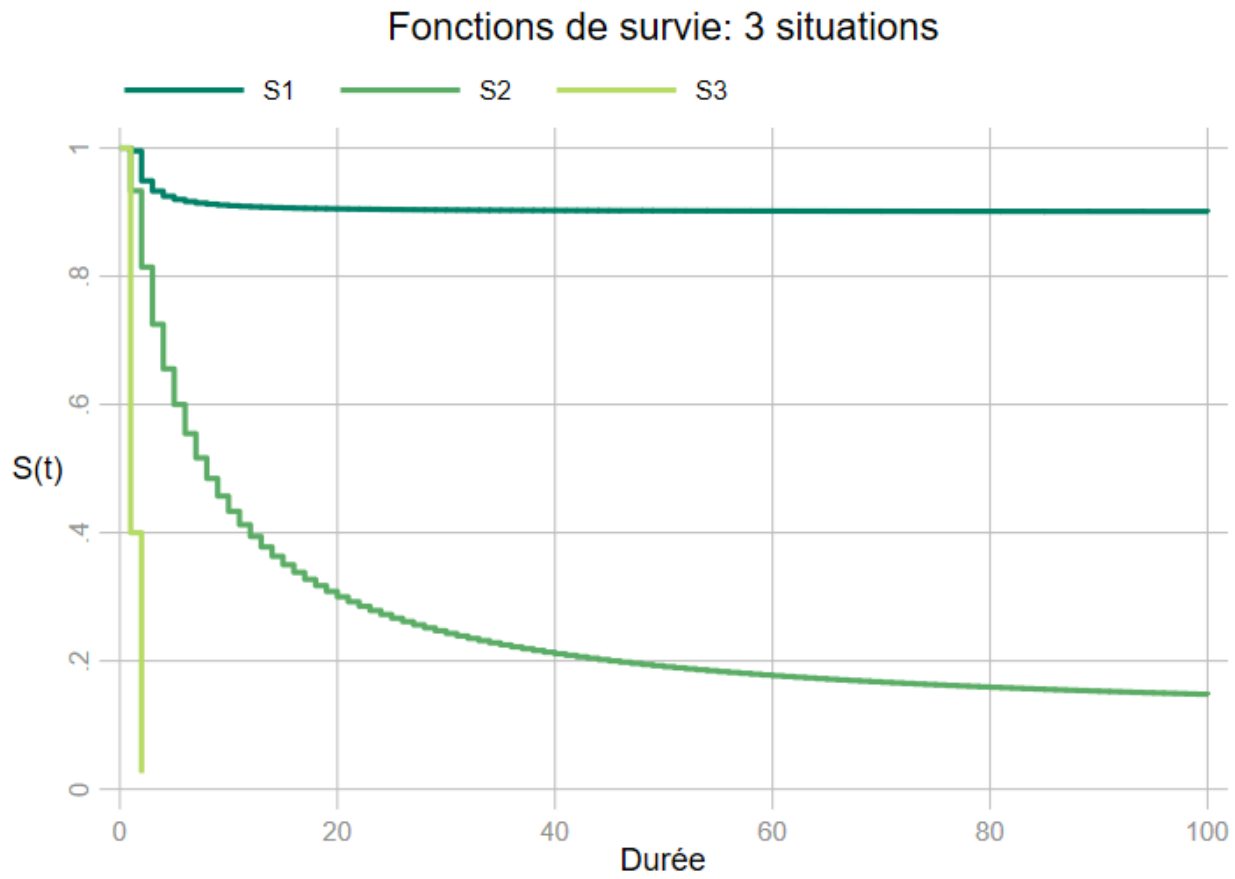
A quel âge le risque de mourir des cochons d'Inde est-il le plus élevé? Quelle est la valeur de ce risque?

3.6 Remarques complémentaires

3.6.1 Formes typiques de la fonction de survie

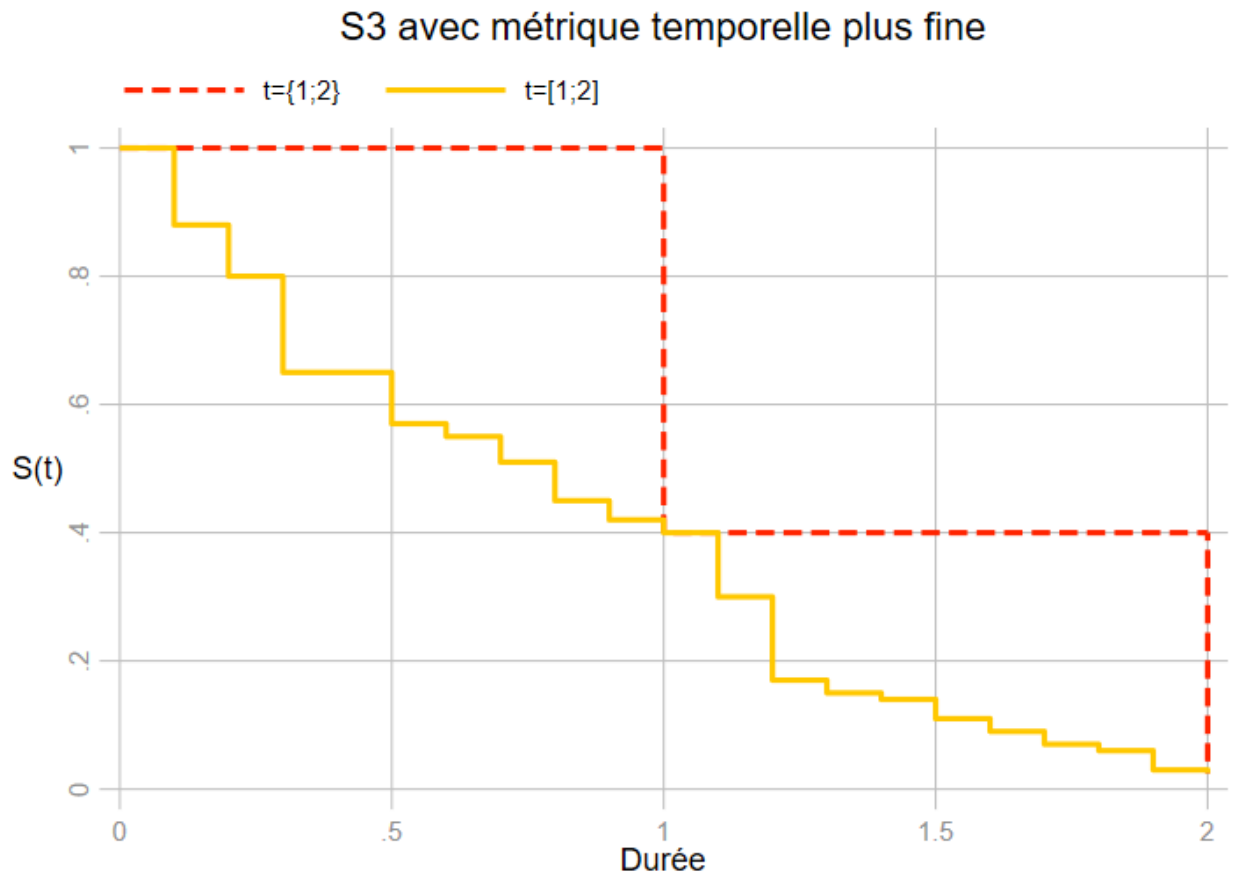
Une des propriétés de la fonction de survie ou de séjour est qu'elles tendent vers 0. A la lecture du graphique suivant, cela peut correspondre à la forme de la courbe S2, bien que le % de survivant tend à baisser de moins en moins à mesure que la durée augmente. Deux cas limites doivent être considéré.

Figure 3.4: Fonction de survie: 3 situtation typiques



- **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 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 3.5: Fonction de survie et modification de la métrique temporelle



3.6.2 Absence de censures à droites

Les méthodes qui vont être présentées plus tard ***gèrent*** la présence de censures à droite. S'il n'y en a pas, 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.

3.6.3 Utilisation des pondérations dans un schema retrospectif avec des biographies longues

Une question assez récurrente concerne l'utilisation des poids de sondage dans les analyses de durées avec longueurs biographiques souvent assez longues. 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 régulièrement 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??????

partie III

Méthodes non paramétrique

4 Estimations des fonctions de survie

Les méthodes non paramétriques portent généralement sur l'analyse des fonctions de survie ($S(t)$) ou sur celles 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 discrète/groupée 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 utilisée, en partie en raison des tests de comparaison, plus ou moins pertinents, qu'elle permet de réaliser.

! Important

- J'insiste sur la nécessité de passer par cette étape avant de se lancer *corps perdu* dans des modèles, comme ceux à durée discrète.
  - Les applications ont des gardes fous permettant d'alerter sur des durées d'exposition incorrecte, en particulier lorsqu'on travaille sur des données prospectives.
  - Egalement très utile, la comparaison graphique de courbes de séjour permet de repérer rapidement des violations fortes de l'hypothèse de proportionalité des risques, ou des situations de quasi *immunité*.
- Concernant les tests non paramétriques, ceux utilsant la technique du *logrank*, présente à mon sens tellement de défauts qu'ils devraient être abandonné. Malheureusement encore très peu diffusée dans les sciences sociales, la comparaison des RMST (*Restricted Mean of Survival Time*) me semble une solution largement supérieure, tant au niveau statistique qu'au niveau interprétatif.

4.1 Les fonctions de survie/séjour

4.1.1 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 $t = 0$.

4.1.2 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 donc avoir survécu en t_{i-1}, t_{i-2}, ..., t_1.
La fonction de survie renvoie donc des probabilités conditionnelles: on survit 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: $S(t_i) = (1 - \frac{d_i}{r_i}) \times (1 - \frac{d_{i-1}}{r_{i-1}}) \times S(t_{i-2})$.

Au final, 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 support

- 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*). Les durées de la variable *stime* ont été regroupée par période de 30 jours pour réaliser des analyses à durée discrete. Cette nouvelle variable de durée a été appelé *mois*.
- L'introduction d'une dimension dynamique, la greffe, est donnée par les informations contenues dans les variables *transplant* et *wait*.
- La variable *compet* est une information simulée pour réaliser des analyses en risques concurrents.
- Les bases en format .csv, .sas7bdat et .dta sont disponibles dans ce dépôt [lien]

Extrait de la base:

id	year	age	died	stime	surgery	transplant	wait	mois	compet
15	68	53	1	1	0	0	0	1	1
43	70	43	1	2	0	0	0	1	1

61	71	52	1	2	0	0	0	1	1
75	72	52	1	2	0	0	0	1	1
102	74	40	0	11	0	0	0	1	0
74	72	29	1	17	0	1	5	1	2

4.2 La méthode actuarielle

- Estimation sur des intervalles définies par l'utilisateur.
- Méthode dite «continue», estimation en milieu d'intevalle.
- Méthode apropriée lorsque la durée est mesurée de manière discrète/groupée.
- Méthode, hélas, quasiment abandonnée dans les sciences sociales où les durées sont plus rarement mesurées de manière exacte. L'absence de test de comparaison des fonctions de survie n'y est pas étranger, tout comme le lien de la méthode suivante (Kaplan-Meier) avec le modèle de Cox.
- Contrairement à la méthode de Cox, la méthode actuarielle permet de calculer les quantiles de la durée.

4.2.1 Estimation

Echelle temporelle

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 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 conseillé de calculer des durées moyennes. On préfère utiliser la médiane ou tout autre quantile lorsqu'ils sont calculables.

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 $S(t_i) > 0.5$.

R-Stata-Sas-Python

4.2.2 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 (programmation pénible), mais les quantiles de la durée ne sont toujours pas estimables, ce qui est bien dommage.

4.2.3 Stata

Commande `ltable`, avec en option la paramétrisation des intervalles de durées. Voir la commande externe `qlt` (MT) qui calcule les durées médianes (+ autres quartiles) et qui recalcule la fonction de séjour avec une définition des intervalles de durées identique à celle de SAS.

4.2.4 Sas

Sous une `proc lifetest` avec en option `method=lifetable`. On peut paramétrer les intervalles d'estimation avec l'option `width`.

4.2.5 Python

A l'heure actuelle, aucune fonction à ma connaissance

4.2.6 Application

Les résultats qui suivent ont été estimés avec Stata en retenant la définition des bornes de Sas, plus pertinente à mon sens, avec des intervalles fixes de 30 jours.

```
     +-------------------------------------------------+
     |   t0     t1   survival CI 95% low  CI 95% up |
     |-------------------------------------------------|
  1. |    0     30          1          .          . |
  2. |   30     60   .7853659   .6925991   .8530615 |
  3. |   60     90   .6461871   .5449008   .7304808 |
  4. |   90    120    .525027   .4232338   .6170507 |
  5. |  120    150   .4740535   .3737563   .5677139 |
     |-------------------------------------------------|
  6. |  150    180   .4636348   .3637283   .5575485 |
```

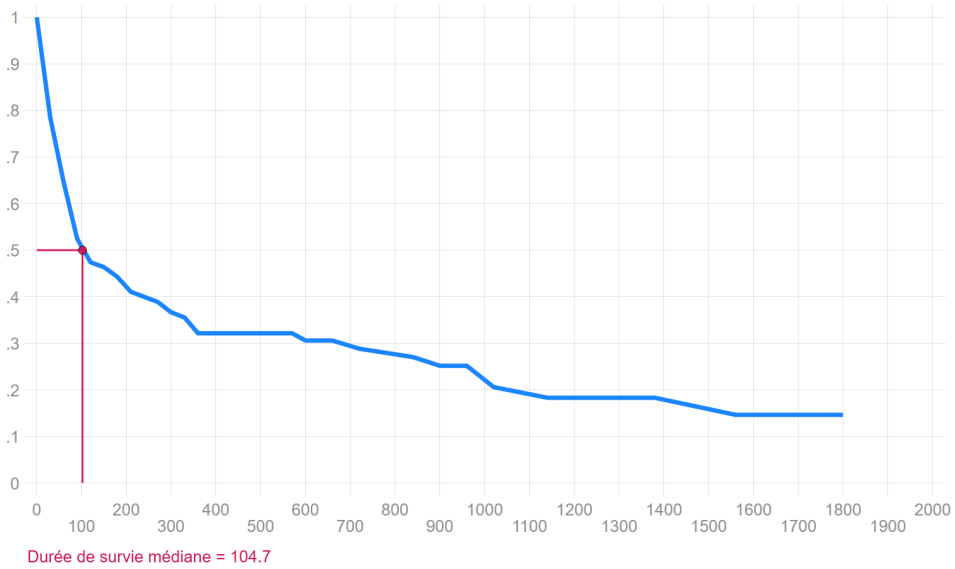
```
 7. |  180    210   .4425605   .3435417   .5368989 |
 8. |  210    240   .4105681   .3132064   .5052779 |
 9. |  240    270   .3997637   .3030412   .4945301 |
10. |  270    300   .3888113   .2927645   .4836136 |
    |-----------------------------------------------|
11. |  300    330   .3665935   .2720434   .4613676 |
12. |  330    360   .3554846   .2617823   .4501585 |
13. |  360    390   .3216289   .2308275   .4157428 |
14. |  390    420   .3216289   .2308275   .4157428 |
15. |  420    450   .3216289   .2308275   .4157428 |
    |-----------------------------------------------|
16. |  480    510   .3216289   .2308275   .4157428 |
17. |  510    540   .3216289   .2308275   .4157428 |
18. |  540    570   .3216289   .2308275   .4157428 |
19. |  570    600   .3216289   .2308275   .4157428 |
20. |  600    630   .3059397   .2154747   .4009653 |
    |-----------------------------------------------|
21. |  660    690   .3059397   .2154747   .4009653 |
22. |  720    750   .2884574   .1981834   .3848506 |
23. |  840    870   .2704288   .1806664   .3680736 |
24. |  900    930   .2517786   .1628919   .3505543 |
25. |  930    960   .2517786   .1628919   .3505543 |
    |-----------------------------------------------|
26. |  960    990   .2517786   .1628919   .3505543 |
27. |  990   1020   .2288896   .1404089   .3303913 |
28. | 1020   1050   .2060007   .1191749   .3093143 |
29. | 1140   1170   .1831117   .0991601   .2873401 |
30. | 1320   1350   .1831117   .0991601   .2873401 |
    |-----------------------------------------------|
31. | 1380   1410   .1831117   .0991601   .2873401 |
32. | 1560   1590   .1464894   .0645215   .2602391 |
33. | 1770   1800   .1464894   .0645215   .2602391 |
34. | 1800      .   .1464894   .0645215   .2602391 |
    +-----------------------------------------------+
```

Table 4.2: Quantiles de la fonction de séjour type actuarielle - Bornes Sas

S(t)	**t**
0.90	13.977
0.75	37.623
0.50	104.729
0.25	906.993
0.10	.

Lecture des résultats: 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% sont décédés.

Figure 4.1: Courbe de survie: estimation méthode actuarielle



4.3 La méthode de 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 estimation 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.

4.3.1 Estimation

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 de la décrire dans ce support.

Récupération de la médiane
Il n'y a pas de méthode pour calculer directement la durée médiane (ou tout autre quantile) contrairement à l'approche actuarielle.
La définition retenue est conventionnelle. 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 $S(t) \leq 0.5$. Attention, il n'est pas impossible que le % de survivant.e.s soit bien en deçà de 50% pour l'obtention cette durée médiane.

R-Stata-Sas-Python

4.3.2 R

Les estimateurs sont obtenus avec fonction **`survfit`** du package `survival`. On peut obtenir des rendus graphiques de meilleures qualité avec le package `survminer` (fonction `ggsurvplot`)

4.3.3 Stata

Après avoir appelé les variables de durée et de censure en mode **survival** avec **`stset`**), le tableau des estimateurs est obtenu avec la commande **`sts list`** et le graphique avec **`sts graph`**.

4.3.4 SAS

L'estimation de Kaplan-Meier est affichée par défaut par la `proc lifetest`. ***Warning*** : le tableau affiché par SAS est particulièrement pénible à lire voire illisible, en particulier lorsque le nombre de censures est élevé, une ligne étant ajoutée pour chaque observation censurée. Je conseille de ne pas afficher cette partie de l'output (se reporter à la section SAS du chapitre programmation). On récupère pour le reste de l'output les valeurs de la durée pour S(t) =(.75,.5,.25) ainsi que le graphique, ce qui est suffisant.

4.3.5 Python

Les resultats sont donnés dans la librairie `lifeline` par des fonctions au nom interminable. Je conseille plutôt l'utilisation de la librairie **`statmodels`** (se reporter à la section dédiée à Python).

4.3.6 Application

On reprend l'exemple précédent.

Time	Total	Fail	Lost	Function	Error	[95% Conf.	Int.]
1	103	1	0	0.9903	0.0097	0.9331	0.9986
2	102	3	0	0.9612	0.0190	0.8998	0.9852
3	99	3	0	0.9320	0.0248	0.8627	0.9670
5	96	2	0	0.9126	0.0278	0.8388	0.9535
6	94	2	0	0.8932	0.0304	0.8155	0.9394
8	92	1	0	0.8835	0.0316	0.8040	0.9321
9	91	1	0	0.8738	0.0327	0.7926	0.9247
11	90	0	1	0.8738	0.0327	0.7926	0.9247
12	89	1	0	0.8640	0.0338	0.7811	0.9171
16	88	3	0	0.8345	0.0367	0.7474	0.8937
17	85	1	0	0.8247	0.0375	0.7363	0.8857
18	84	1	0	0.8149	0.0383	0.7253	0.8777
21	83	2	0	0.7952	0.0399	0.7034	0.8614
28	81	1	0	0.7854	0.0406	0.6926	0.8531
30	80	1	0	0.7756	0.0412	0.6819	0.8448
31	79	0	1	0.7756	0.0412	0.6819	0.8448
32	78	1	0	0.7657	0.0419	0.6710	0.8363
35	77	1	0	0.7557	0.0425	0.6603	0.8278
36	76	1	0	0.7458	0.0431	0.6495	0.8192
37	75	1	0	0.7358	0.0436	0.6388	0.8106
39	74	1	1	0.7259	0.0442	0.6282	0.8019
40	72	2	0	0.7057	0.0452	0.6068	0.7842
43	70	1	0	0.6956	0.0457	0.5961	0.7752
45	69	1	0	0.6856	0.0461	0.5855	0.7662
50	68	1	0	0.6755	0.0465	0.5750	0.7572
51	67	1	0	0.6654	0.0469	0.5645	0.7481
53	66	1	0	0.6553	0.0472	0.5541	0.7390
58	65	1	0	0.6452	0.0476	0.5437	0.7298
61	64	1	0	0.6352	0.0479	0.5333	0.7206
66	63	1	0	0.6251	0.0482	0.5230	0.7113
68	62	2	0	0.6049	0.0487	0.5026	0.6926
69	60	1	0	0.5948	0.0489	0.4924	0.6832
72	59	2	0	0.5747	0.0493	0.4722	0.6643
77	57	1	0	0.5646	0.0494	0.4621	0.6548
78	56	1	0	0.5545	0.0496	0.4521	0.6453
80	55	1	0	0.5444	0.0497	0.4422	0.6357
81	54	1	0	0.5343	0.0498	0.4323	0.6261
85	53	1	0	0.5243	0.0499	0.4224	0.6164
90	52	1	0	0.5142	0.0499	0.4125	0.6067
96	51	1	0	0.5041	0.0499	0.4027	0.5969
100	50	1	0	0.4940	0.0499	0.3930	0.5872

```
   102       49       1       0              0.4839    0.0499     0.3833    0.5773
   109       48       0       1              0.4839    0.0499     0.3833    0.5773
   110       47       1       0              0.4736    0.0499     0.3733    0.5673
   131       46       0       1              0.4736    0.0499     0.3733    0.5673
   149       45       1       0              0.4631    0.0499     0.3632    0.5571
   153       44       1       0              0.4526    0.0499     0.3531    0.5468
   165       43       1       0              0.4421    0.0498     0.3430    0.5364
   180       42       0       1              0.4421    0.0498     0.3430    0.5364
   186       41       1       0              0.4313    0.0497     0.3327    0.5258
   188       40       1       0              0.4205    0.0497     0.3225    0.5152
   207       39       1       0              0.4097    0.0495     0.3123    0.5045
   219       38       1       0              0.3989    0.0494     0.3022    0.4938
   263       37       1       0              0.3881    0.0492     0.2921    0.4830
   265       36       0       1              0.3881    0.0492     0.2921    0.4830
   285       35       2       0              0.3660    0.0488     0.2714    0.4608
   308       33       1       0              0.3549    0.0486     0.2612    0.4496
   334       32       1       0              0.3438    0.0483     0.2510    0.4383
   340       31       1       1              0.3327    0.0480     0.2409    0.4270
   342       29       1       0              0.3212    0.0477     0.2305    0.4153
   370       28       0       1              0.3212    0.0477     0.2305    0.4153
   397       27       0       1              0.3212    0.0477     0.2305    0.4153
   427       26       0       1              0.3212    0.0477     0.2305    0.4153
   445       25       0       1              0.3212    0.0477     0.2305    0.4153
   482       24       0       1              0.3212    0.0477     0.2305    0.4153
   515       23       0       1              0.3212    0.0477     0.2305    0.4153
   545       22       0       1              0.3212    0.0477     0.2305    0.4153
   583       21       1       0              0.3059    0.0478     0.2156    0.4008
   596       20       0       1              0.3059    0.0478     0.2156    0.4008
   620       19       0       1              0.3059    0.0478     0.2156    0.4008
   670       18       0       1              0.3059    0.0478     0.2156    0.4008
   675       17       1       0              0.2879    0.0483     0.1976    0.3844
   733       16       1       0              0.2699    0.0485     0.1802    0.3676
   841       15       0       1              0.2699    0.0485     0.1802    0.3676
   852       14       1       0              0.2507    0.0487     0.1616    0.3497
   915       13       0       1              0.2507    0.0487     0.1616    0.3497
   941       12       0       1              0.2507    0.0487     0.1616    0.3497
   979       11       1       0              0.2279    0.0493     0.1394    0.3295
   995       10       1       0              0.2051    0.0494     0.1183    0.3085

[Résultats non reportés à partir de t=1000 ]
--------------------------------------------------------------------------------
```

La durée durée médiane de survie est $t = 100$. Elle correspond à $S(t) = 0.4940$.

Table 4.3: Quantiles de la fonction de séjour type Kaplan-Meier

S(t)	t
0.90	6
0.75	36
0.50	100
0.25	979
0.1	.

Figure 4.2: Courbe de survie: estimation méthode actuarielle

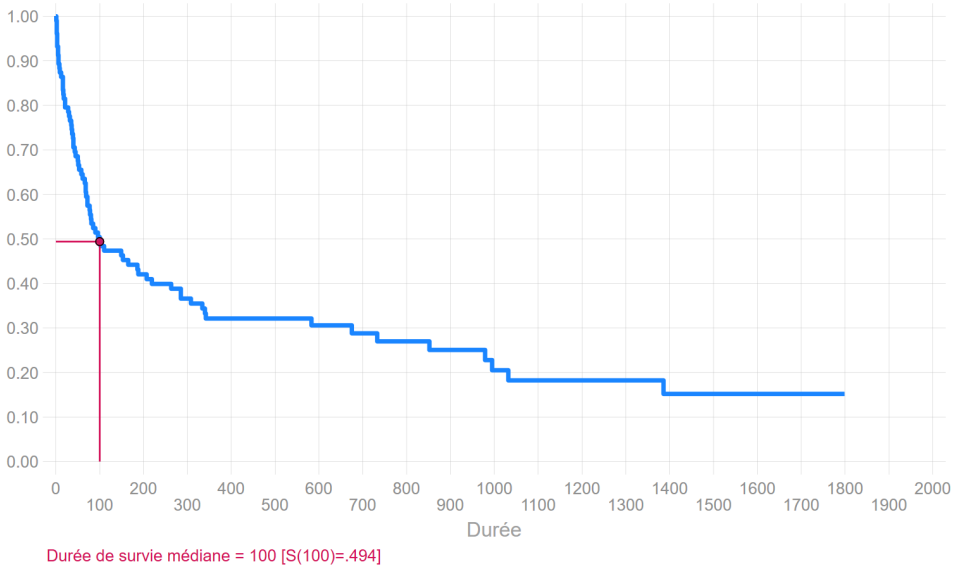
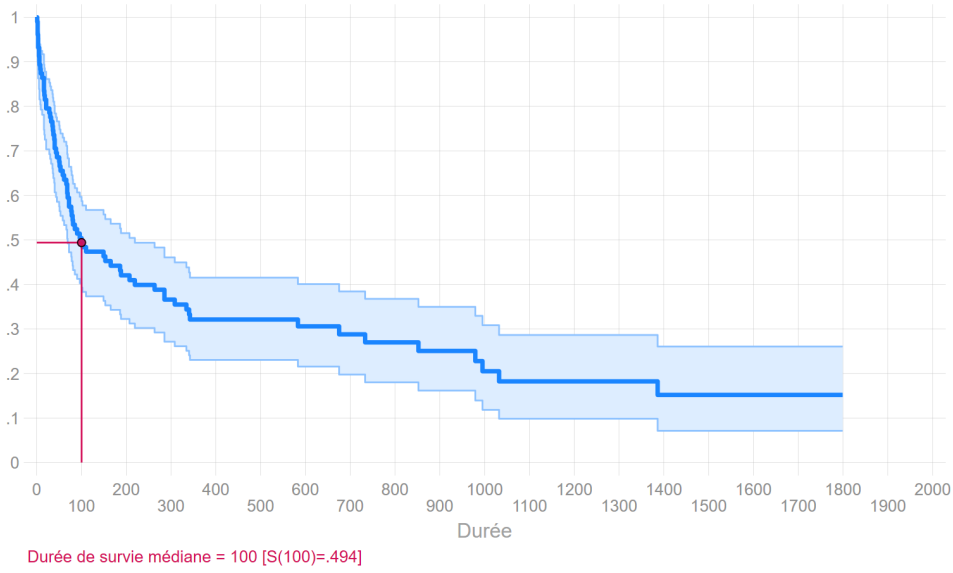


Figure 4.3: Courbe de survie: estimation méthode actuarielle + CI



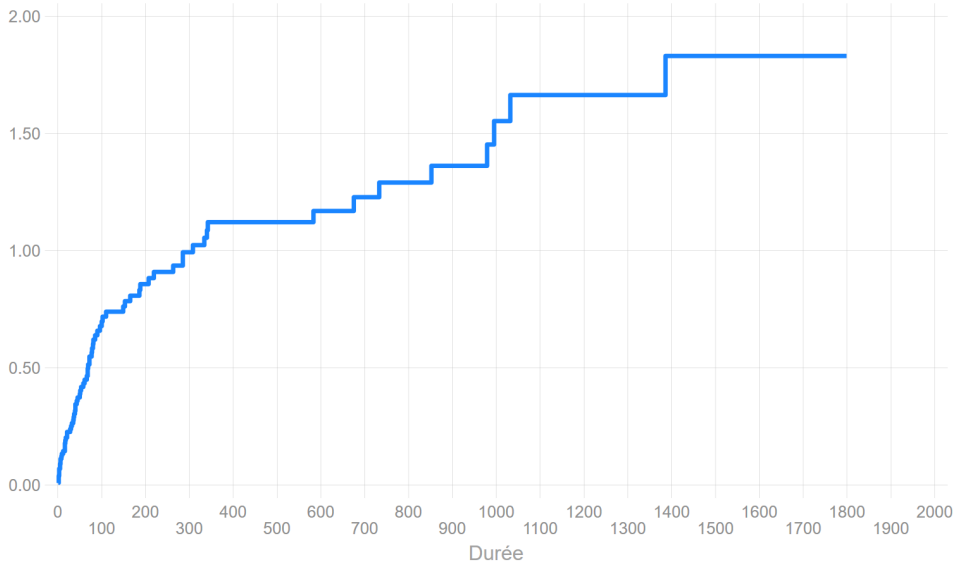
4.3.7 Quantités associées à l'estimateur Kaplan-Meier..

Le risque cumulé: estimateur de Nelson AAlen

Il est simplément égal à:

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

Figure 4.4: Risque cumulé: estimateur Nelson-Aalen

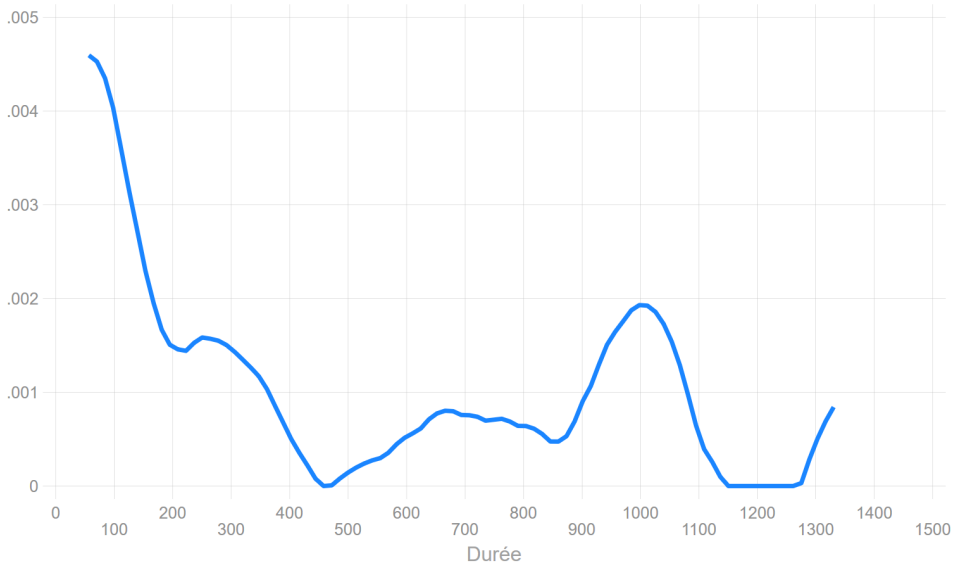


Le risque instantané

Nécessite l'estimateur de risque cumulé de Nelson-Aalen. Le risque est obtenu en lissant les différences - toujours positive - entre $H(t)$ par la méthode dite du **kernel** (cf estimation de la densité des distributions). 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.

> **Note**
>
> Il n'est pas inutile de noter qu'il n'y a pas de *formule* toute faite pour obtenir des valeurs du risque instantané. Ce type de méthode par lissage est pleinement paramétrable, par exemple sa fenêtre, ce qui implique que son profil varie d'un paramétrage à l'autre. Le graphique précédent a été fait avec Stata, si on utilisait le package `muhaz` les différences de paramétrage par défaut font que les courbes ne se confondent pas.

Figure 4.5: Risque instantané: estimateur du Kernel



5 Tests de comparaison

- 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. Il faut donc prévoir des regroupements pour les transformer en variable ordinale.

Deux méthodes sont utilisées:

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

5.1 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 et à la non proportionalité des risques.

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 moment d'évènement. La principale différence réside dans le calcul de la variance de la statistique du test qui, ici, suit assez logiquement une loi hypergéométrique [proche loi binomiale mais avec tirage avec remise].

5.1.1 Principe de calcul de la statistique de test

- **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 $\chi^2(1)$. Si on teste simultanément la différence de g fonctions de survie, ce qui n'est pas une bonne idée en passant, la statistique de test suis un $\chi^2(g-1)$.

5.1.2 Les principaux tests 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 de base non pondéré.

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, il est fortement déconseillé de les utiliser. Cela ne signifie pas que si les courbes ne se croisent pas, l'hypothèse de proportionalité des risques est respectée : des rapports de risque peuvent au cours du temps s'intensifier, se réduire ou, le cas échant s'inverser, ce qui est 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 ou disponibles si on utilise R.

R-Stata-Sas-Python

5.1.3 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`** du package **`survminer`**. Cette fonction permet également d'obtenir des tests 2 à 2.

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`**.

5.1.4 Stata

On utilise la commande **`sts test`** avec le nom de la version du test: `peto`, `wilcoxon` . Sans préciser le nom de la variante, le test non pondéré est exécuté.

5.1.5 Sas

Le test non pondéré et la version Wilcoxon sont données avec l'option **`strata`** de la `proc lifetest`. Attention : ne jamais utiliser la version *LR Test* qui est biaisée. Pour obtenir d'autres versions du test du log-rank, on ajoute **`/test=all`** à l'option `strata`.

5.1.6 Python

Avec la librairie `lifelines`, on utilise la fonction **`logrank_test`**. Quatre variantes sont disponibles (Wilcoxon, Tarone-Ware, Peto-Peto et Fleming-Harrigton). On peut également utiliser la fonction **`duration.survdiff`** de `statmodels` (non pondéré, Wilcoxon - appelé ici Breslow- et Tarone-Ware).

5.1.7 Application

On compare ici l'effet du pontage coronarien sur le risque de décéder depuis l'inscription dans le registre de greffe.

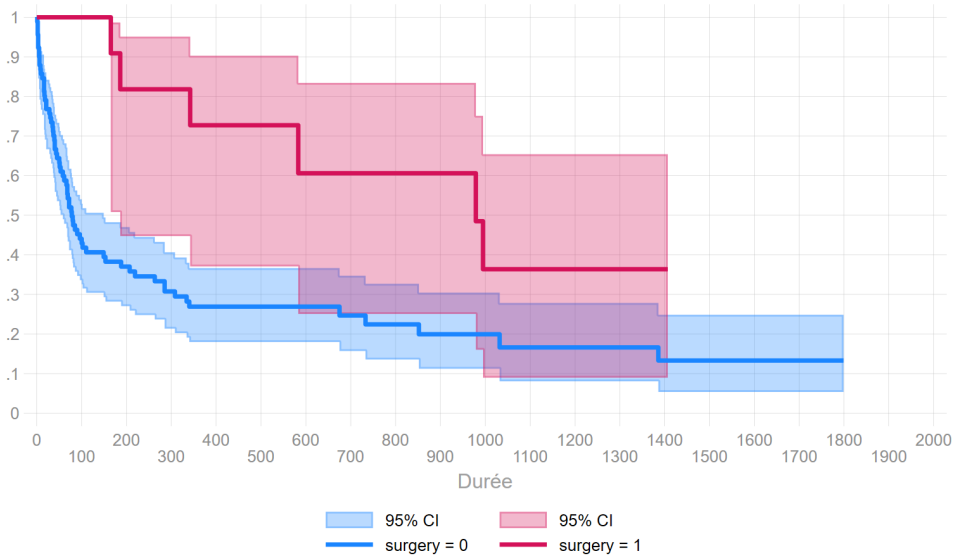


Table 5.1: Résultats des tests du logrank

Test	df	Chi2	P>Chi2
Non pondéré	1	6.59	0.0103
Wilcoxon (Breslow)	1	8.99	0.0027
Tarone-Ware	1	8.46	0.0036
Peto-Peto		8.66	0.0033

Les résultats font apparaître que l'opération permet d'augmenter la durée de survie des personnes. Il apparait que la p-value est plus élevée pour test non pondérée. Cela peut-il s'expliquer en regardant les deux courbes de séjours? Qu'en est-il de la proportionalité des risques ???? …. Réponse pendant la formation.

5.2 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.

Principe

- 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-Stata-Sas-Python

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.

5.2.1 R

Librairie `SurvRm2`. Programmée par les mêmes personnes que la commande Stata, la fonction proposée n'est pas très souple.

5.2.2 Stata

Commande externe `strmst2`. La plus ancienne fonction proposée par les logiciels. Au final plus limitée que la solution Sas. J'ai programmé une commande, `diffrmst`, qui représente graphiquement les estimations des Rmst pour chaque temps d'évènement, leurs différences et les p-value issues des comparaisons.

5.2.3 SAS

Disponible depuis la version 15.1 de SAS/Stat (fin 2018). Les estimations et le résultat du test de comparaison sont récupérables très simplement dans une `proc lifetest`, avec en option **`plots=(rmst)`** . Bien que sortie tardivement par rapport Stata et R, les résultats sont particulièrement complets.

5.2.4 Python

Estimation un peu pénible. A partir de l'estimateur KM obtenu avec la fonction `KaplanMeierFitter` de `lifelines`, on peut obtenir les RMST avec la fonction `restricted_mean_survival_time`. On peut tracer les fonctions, en revanche le test de comparaison n'est pas implémenté.

Application

Avec $tmax = 1407$:

Table 5.2: Estimation des Rmst pour la variable surgery

Groupes	RMST	Std. Err	95% CI
$surgery = 1$	884.576	187.263	517.546 - 1251.605
$surgery = 0$	379.148	61.667	258.282 - 500.014

Table 5.3: Différences entre Rmst pour la variable surgery

Types de contraste	Ecarts RMST	P>|z|	95% CI
$Rmst(surgery1 - surgery0)$	505.428	0.010	517.546 - 1251.605
$Rmst\left(\frac{surgery1}{surgery0}\right)$	2.333	0.002	1.383 - 3.937

Ici t^* est égal à 1407 jours, soit la durée qui correspond au dernier décès observé.
Sur un horizon de 1407 jours, ces individus opérés d'un pontage peuvent espérer vivre 884 jours en moyenne, contre 379 jours pour les autres. La durée moyenne de survie est donc 2.3 fois plus importante pour les personnes opérées (rapport des Rmst = 2.3), ce qui correspond à une différence de 379 jours.

Le tableau et le graphique suivant donnent les valeurs des Rmst et les écarts de la variable *surgery* en faisant varier *tmax* sur chaque jour où au moins un décès a été observé. Il a été réalisé avec Stata, la durée maximale utilisée a été paramétrée à 1407 jours (idem R, Sas).

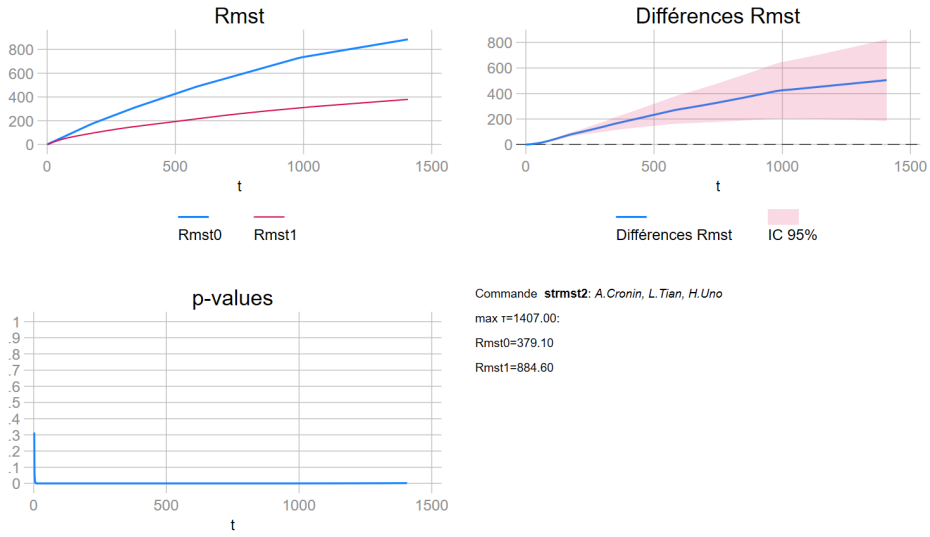
Comme le premier décès observé pour les personnes opéré se situe le 165eme jours, il est tout à fait normal que pour ce groupe de personnes la valeur de la Rmst soit identique au jour de décès des individus non opérés.

> **Note**
>
> Pour la version pdf, seulement une dizaine de points a été sélectionné en raison de la longueur du tableau

_time	_rmst1	_rmst0	_diff	95%CI lower	95%CI upper	pvalue
1	1	1	0	0	0	.
2	2	1.989011	.010989	-.0104304	.0324084	.3146368
3	3	2.945055	.0549451	-.0009099	.1108	.0538507
5	5	4.791209	.2087912	.0549289	.3626535	.0078217
6	6	5.692307	.3076923	.0995576	.5158269	.0037617
8	8	7.45055	.5494505	.2224352	.8764658	.0009908
9	9	8.318682	.6813186	.2913915	1.071246	.0006156
50	50	38.90242	11.09758	7.539261	14.6559	9.80e-10
515	437.5454	197.5971	239.9483	150.1031	329.7935	1.65e-07
995	734.7576	310.1678	424.5898	204.0643	645.1152	.0001609
1032	748.2121	317.5443	430.6678	202.7468	658.5889	.0002127

1141	787.8485	335.6531	452.1953	200.7097	703.681	.0004248
1321	853.303	365.5577	487.7454	191.5434	783.9473	.0012492
1386	876.9394	376.3565	500.5829	186.9499	814.2158	.0017585
1400	882.0303	378.2173	503.813	186.4392	821.1869	.0018625
1407	884.5757	379.1476	505.4281	186.1745	824.6817	.0019162

Figure 5.1: Comparaison des Rmst à chaque jour où au moins un décès est observé



partie IV

Modèles à risques proportionnels

6 Introduction

7 Proprortionalité des risques

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 (ou taux de risque).
- $h_0(t)$ est une fonction qui dépend de la durée mais pas des caractéristiques individuelles. Il définiera le risque de base, et jouera donc le rôle de la constante dans un modèle classique.
- $e^{X'b}$ est une fonction qui ne dépend pas de la durée, mais des caractéristiques individuelles $X'b = \sum_{k=1}^{p} b_k X_k$. La forme exponentielle assurera sa positivité [1].

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 la durée du *rapport des risques* (**hazard ratio**).

Exemple:
Avec une seule covariable X introduite au modèle, et 2 observations disons A et B:

- $h_A(t) = h_0(t)e^{bX_A}$
- $h_B(t) = h_0(t)e^{bX_B}$.

Le rapport des risques entre A et B est simplement égal à:

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

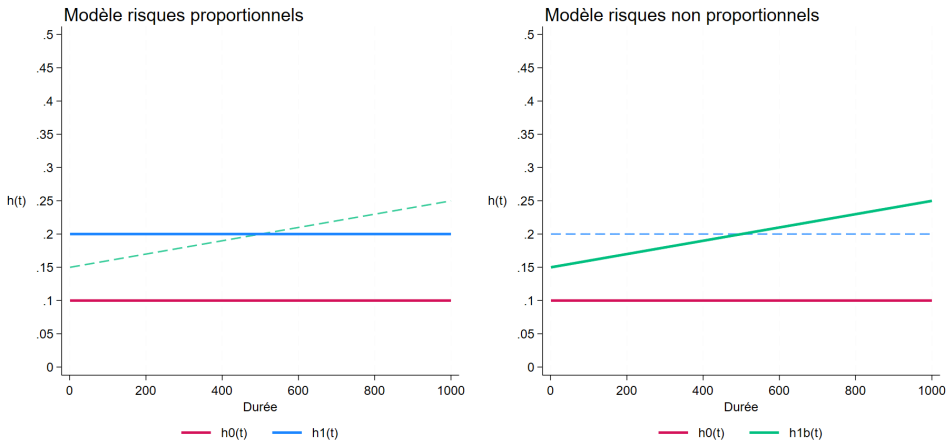
Autrement dit, cette proportionnalité des risques est la traduction d'une absence d'interaction entre les rapports de risques estimés par un modèle à risque proportionnel et la durée (ou une fonction de celle-ci).

Si on part d'un modèle tel que $h_0(t) = 0.1$ quelque soit t (baseline à risque constant).

Si $h_1(t)$ est lui même constant, le rapport entre $h_1(t)$ et $h_0(t)$ sera lui même constant dans la durée. On dit que les risques sont proportionnels. Ici, $h_1(t) = 0.2$ quel que soit t, le rapport des risques est toujours égal à $\frac{0.2}{0.1} = 2 = e^b$. Le paramètre estimé par un modèle à risque proportionnel sera égal à $log(2) = 0.69$.

[1] On rappelera qu'en durée continue, seule positivité du risque doit être assurée, d'où l'expression *hazard rate*

Figure 7.1: L'hypothèse de proportionalite des risques



Pour $h_{1b}(t)$, le risque augmente de manière à un rythme constant (linéaire): $h_{1b}(1) = 0.15$ et $h_{1b}(1000) = 0.25$. Comme $h_0(t)*$ est constant, le rapport des risques s'accroît également. On dit que les risques ne sont pas proportionnels.

Si on est dans le deuxième cas de figure, un modèle à risque proportionnel estimera un rapport toujours égal à 2. Il estimera un *rapport moyen* sur la période d'observation.

8 Les modèles

- **Modèle semi-paramétrique de Cox (1972)**
  Le modèle estime directement les b indépendamment de $h_0(t)$. C'est pour cela qu'il est appelé modèle ***semi-paramétrique de Cox***. Les rapports de risque (e^b) seront utilisés dans un deuxième temps 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 proportionnalité est donc importante et doit donc être analysée.

- ***Modèle à durée discrète*** Sa spécification diffère quelque peu 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 taux de risque est estimée simultanément aux autres paramètres. Il est comme son nom l'indique, particulièrement adapté au durées discrètes ou groupées. Avec une spécification logistique, les Odds vont sous certaines conditions (souvent respectée), se confondre avec des probabilités/risques. Lorsque le nombre de points d'observations (t) n'est pas trop faible, les résultats obtenus sont très proches de ceux issus directement d'un modèle de Cox. On peut souligné que ce modèle a été à l'origine proposé par Cox lui même à la fin des années 60.

- **Les modèles paramétriques standards**
  Les modèles dits de Weibull, exponentiel, Gompertz ont une spécification sous hypothèse de risque proportionnel. Ils seront traités brièvement dans les compléments. Historiquement, le modèle de Cox est une réponse à une possible difficulté dans l'ajustement du risque par une loi de distribution du risque a priori.

- **Modèle paramétrique de Parmar-Royston** (non traité)
  $h_0(t)$, via le risque cumulé $H(t)$, est estimé simultanément avec les rapports de risques en utilisant la méthode des *splines cubiques*. Il est maintenant implémenté dans les logiciels standards (R, Stata, Sas). Les rapports de risque obtenus sont très proches de ceux estimés par le modèle classique de Cox. Il offre donc une alternative surement intéressante au Cox standard, et il s'est maintenant largement diffusé dans l'analyse des effets cliniques.

- **Modèle à non proportionnalité**: on a bien évidemment les modèles paramétriques de type AFT (Accelerated Failure Time), le peut-être très prometteur modèle à *pseudo observations* d'Andersen (j'espère en faire une courte présentation rapidement après avoir évaluer son passage en durée discrète). Dans le champ du machine learning, il y a depuis son origine une version modèle de survie dans les *forêts aléatoires*.

9 Le modèle de Cox

On peut ignorer la partie sur l'estimation du modèle. On retiendra tout de même qu'il est déconseillé d'utiliser 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, très fréquentes dans les sciences sociales.

9.1 Le modèle semi-paramétrique de Cox

9.1.1 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 pour la censure).

On peut représenter le processus aléatoire d'une analyse de survie en présence de censure à droite, avec l'équation de vraisemblance suivante:

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

- $f(t_i)$ 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: $L_i = S(t_i)$

Vraisemblance partielle de Cox

Comme $f(t_i) = h(t_i) \times S(t_i)$ [1], 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 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 observations 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 seulement le premier terme de la vraisemblance, soit:

[1] Se reporter à la définition des grandeurs dans la section *Théorie*

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

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 à droite 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 donc bien une probabilité, et la vraisemblance partielle est donc bien un produit de probabilités. Pour un individu ayant connu l'évènement, la contribution à la vraisemblance partielle est **la probabilité que l'individu observe l'évènement en t_i sachant qu'un évènement (et un seul) s'est produit**.

- Si $\delta_i = 0$: $PL_i = 1$
- Si $\delta_i = 1$: $PL_i = \frac{e^{X_i' b}}{\sum_{j \in R_i} e^{X_j' b}}$

Condition nécessaire: pas d'évènement simultané: en présence d'évènements mesurés simultanément, 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 doit pas y avoir d'évènement simultané, on va introduire à la vraisemblance partielle toutes les permutations possibles des évènements observés au même moment. Bien qu'en t_i on observe au même moment l'évènement pour 2 observations (A,B) une métrique temporelle plus précise permettrait de savoir si A s'est produit avant B ou B s'est produit avant A (2 permutations). Comme le nombre de permutations est calculé à l'aide d'une factorielle [2], avec 3 évènements mesurés simultanément, on obtient 6 permutations ($3 \times 2 \times 1$). 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.*

[2] $n! = (n) \times (n-1) \times (n-2) \times .... \times 3 \times 2 \times 1$

- 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.

9.1.2 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 loin).

- 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 (sans évènement simultané), la vraisemblance partielle s'ecrit:

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

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

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

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

Comme $\frac{e^{X_j b}}{\sum_{j \in R} e^{X_j b}}$ est une probabilité, et $\sum_{j \in R} X_{ik} \times p_i$ 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 va permettre d'analyser le respect ou non de l'hypothèse de risques proportionnels via les *résidus de Schoenfeld*.

9.1.3 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 (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, le risque de connaitre l'évènement est e^b fois inférieur/supérieur à celui d'une personne qui a un an de moins.

Exemple pour les insuffisances cardiaques

- Correction de la vraisemblance: méthode d'Efron
- Nombre d'observations: 103
- Nombre de décès: 75
- Log-Vraisemblance: -289.30639

Table 9.1: Cox: log Hazard Ratio (Risks Ratio)

Variables	logRR	Std.Err	z	$P > \|z\|$	95% IC
year	-0.119	0.0673	-1.78	0.076	-0.2516;+0.0124
age	+0.0296	0.0135	2.19	0.029	+0.0031;+0.0561
surgery	-0.9873	0.4363	-2.26	0.024	-1.8424;-0.1323

Table 9.2: Cox: Hazard Ratio (Risks Ratio)

Variables	RR	Std.Err	z	$P > \|z\|$	95%CI
year	0.8872	0.0597	-1.78	0.076	0.7775; 1.0124
age	1.0300	0.0139	2.19	0.029	1.0031; 1.0577
surgery	0.3726	0.1625	-2.26	0.024	0.1584; 0.8761

On retrouve les 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, renvoie à une réduction plutôt modérée du risque journalier de décès durant l'attente d'une greffe.

R-Stata-Sas-Python

9.1.4 R

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

9.1.5 Stata

Le modèle est estimé avec la commande **`stcox`**.

9.1.6 SAS

Le modèle est estimé avec la **`proc phreg`**.

9.1.7 Python

Avec la librairie `lifelines`, le modèle est estimé avec la fonction **`CoxPHFitter`**. Avec la librairie `statmodels`, il est estimé avec la fonction `smf.phreg`.

9.2 Analyse de la constance des rapports de risque

- Les rapports de risque (RR) estimés par le modèle sont contraints à être constant sur toute la période d'observation. C'est une hypothèse forte.
- Le respect de cette hypothèse doit être analysé, en particulier pour le modèle de Cox où la baseline du risque est habituellement estimée à l'aide de ces rapports (par exemple la méthode dite de Breslow, non traitée). En post-estimation, les valeurs estimées du risque pourront présenter des valeurs aberrantes si on dévie trop de constance, en particulier en obtenant des négatives des taux de risque.
- Analyser cette hypothèse revient à introduire une interaction entre les rapports et la durée ou plutôt précisément une fonction de la durée).
- Plusieurs méthodes disponibles, on traitera celles basées sur les **résidus de Schoenfeld**, et l'introduction directe d'une intéraction entre une fonction la durée et les covariables du modèle. Cette dernière fait également office de méthode de correction lorsque la violation de l'hypothèse est jugée trop importante ou problématique du point de vue des résultats obtenus.
- Si on regarde les courbes de Kaplan-Meier, leurs croisement non tardif impliquera nécessairement un problème sur cette hypothèse.

9.2.1 Test de Grambsch-Therneau sur les résidus de Schoenfeld

Ce test a été proposé par P.Grambsch et T.Therneau [3] dans un cadre à durée strictement continue. Il repose originellement sur une régression linéaire estimé avec les moindres carrés généralisés (GLS) correction de l'autocorrélation des erreurs avec des sér) . Dans un premier temps pour des raisons plutôt pratiques (informatique), le test a une version moindres carrés ordinaires (OLS). Jusqu'en 2020, tous les logiciels ne proposaient que le test OLS. T.Therneau avec la V3 de package `survival` a substitué - assez brutalement - le test GLS au test OLS. Si les résultats sont proches dans le cadre d'une durée continue et que le test GLS peut être considéré comme un test *exact*, cela devient problématique dans une situation de durée discrète/groupée [4]. Le test OLS reste, à mon sens, la méthode à privilégier dans le cas discret.

Il est également important de souligner que pour P.Grambsch et T.Therneau [5] n'est qu'un moyen parmi d'autres d'analyser une violation de l'hypothèse de proportionnalité. Ce n'est pas *the solution* (comme tout autre test au passage). Le croisement des courbes de séjours peut-être suffisant pour alerter sur cette violation.

Principe du test: consiste à regarder la corrélation entre les **résidus de Schoenfeld** obtenus directement avec la fonction de score de la vraisemblanc partielle de Cox et une fonction de la durée.

Principe de calcul des résidus

- Les résidus *bruts* sont directement calculés à partir des équations de scores [voir section estimation].
- Ils ne sont calculés que pour les observations qui ont connues l'évènement, au moment où un évènement s'est produit.
- La somme des résidus pour chaque covariable est égale à 0. Il s'agit de la propriété de l'équation de score à l'équilibre.
- On utilise généralement les *résidus standardisés* (*remis à l'échelle* / scaled) - par leur variance -. C'est la mesure de cette variance qui distingue le test OLS du test GLS.

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:

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

- Ce résidu est formellement la contribution d'une observation ou d'un moment d'évènement au score. Il se lit comme la différence entre la valeur observée d'une covariable et sa valeur espérée au moment où l'évènement s'est produit.
- Si la constance des rapports de risque varie peu les résidus ne doivent pas suivre une tendance précise localement ou globalement, à la hausse ou à la baisse.

[3] Il s'agit bien de la personne qui maintient le package **`survival`** dans R
[4] Pour les personnes utilisant R, je donne un moyen pour récupérer et exécuter le test OLS sous R
[5] Se reporter à leur ouvrage *Modeling Survival Data: Extending the Cox Model* (2001)

Pourquoi?

Par l'exemple, sans censure à droite et en ne considérant que les résidus bruts: Avec un rapport de risque strictement égal à 1 en début d'exposition, une population soumise au risque $R_i = 100$ 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 en toute logique très proche de 0.

De manière encore plus simple, cette proportionnalité avec un risque ratio égal à 1 suggère qu'au cours de la durée d'observation, on observe une succession d'un même nombre d'hommes et de femmes qui connaissent l'évènement. Si tous les hommes ou presque avaient observés l'évènement plutôt en début d'éxposition et si toutes les femmes ou presque avaient observé l'évènement plutôt en fin d'exposition, l'hypothèse de proportionnalité pourraient fortement remise en cause.

On trouvera des éléments de calcul du test OLS ici

Avertissement

- **Test omnibus**: Ne pas l'utiliser bien qu'il figure généralement en bas des output. Il n'a pas d'interprétation directe, et les p-value peuvent présenter des valeurs très faibles alors que ce n'est pas le cas pour les covariables prises une à une. Rester comme c'est souvent le cas à un test à un degré de liberté.
- **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 $g(t) = 1 - S(t)$ où $S(t)$ 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 : $g(t)=t rend le test relativement sensible aux évènements tardifs lorsque la population restant soumise est peu nombreuse (outliers).
- Par défaut Stata, Sas, Python: $g(t) = t$
- Par défaut R: $g(t) = 1 - S(t)$

Pour des raisons de reproductibilité dans l'espace des logiciels et dans le temps pour les différentes versions du package `survival` de R, on ne présente ici que la version OLS.

Test OLS avec $g(t) = t$

Table 9.3: Test OLS Grambsch-Therneau avec $g(t) = t$

Variables	chi2	df	P>Chi2
year	0.80	1	0.3720
age	1.61	1	0.2043
surgery	5.54	1	0.0186

Ici l'hypothèse de proportionnalité des risques est questionnable pour la variable *surgery*. Le risque ratio pourrait ne pas constant dans le temps. Ce n'est pas du tout étonnant, le premier décès pour les personnes opérées d'un pontage n'est observé qu'au bout de 165 jours. Au final, un test était-il bien nécessaire pour arriver à ce constat ???????

Test OLS avec $g(t) = 1 - S(t)$

Table 9.4: Test Grambsch-Therneau avec $g(t) = 1 - S(t)$

Variables	chi2	df	P>Chi2
year	1.96	1	0.162
age	1.15	1	0.284
surgery	3.96	1	0.046

R-Stata-Sas-Python

9.2.2 R

Attention seulement version GLS du test depuis le V3 de survival.

- Après avoir créer un objet à l'estimation du modèle de Cox, on utilise la fonction **`cox.zph`**. Cette fonction utilise par défaut $g(t) = 1-S(t)$ où $S(t)$ sont les estimateurs de la courbe de Kaplan-Meier. On peut modifier cette fonction. Il est préférable de conserver cette fonction par défaut.
- Test OLS: j'ai récupéré le programme du test antérieur, renommé **`cox.zphold`**. On peut le charger simplement, et il est facilement exécutable. Pour le charger: `source("https://raw.githubusercontent.c`

9.2.3 Stata

Le test (OLS) est obtenu avec la commande **`estat phtest, d`**. Par défaut Stata utilise $g(t) = t$. On peut modifier cette fonction.

9.2.4 SAS

Le test (OLS) est disponible depuis quelques années avec l'argument **`zph`** sur la ligne `proc lifetest`. Par défaut SAS utilise $g(t) = t$. On peut modifier cette fonction.

9.2.5 Python

Le test (OLS) est donné avec la fonction **`proportional_hazard_test`** de la librairie `lifelines`. La fonction utilise par défaut $g(t) = t$, mais on peut afficher les résultats pour toutes les transformations de t disponibles avec l'option `time_transform='all'`.

9.2.6 Intéraction avec la durée

Petit retour sur l'estimation du modèle

Pour estimer le modèle de Cox, les données sont dans un premier temps splitées aux moment où au moins un évènement a été observé.

Sur l'application, avec 2 individus avec la covariable *age* (rappel: il s'agit de l'âge en t_0:

Table 9.5: Base spittées sur les intervals d'évènement

id	age	died	t_0	t
2	51	0	0	1
2	51	0	1	2
2	51	0	2	3
2	51	0	3	5
2	51	1	5	6
3	54	0	0	1
3	54	0	1	2
3	54	0	2	3
3	54	0	3	5
3	54	0	5	6
3	54	0	6	8
3	54	0	8	9
3	54	0	9	12
3	54	1	12	16

Les bornes des intervalles $[t_0; t]$ présentent 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 $t = 4$ pour $id = (2, 3)$et $t = 7, 10, 11, 13, 14, 15$ pour $id = 3$.
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.

Sur cette base *splitée* aux moments d'évènement (n=3573), on pourra vérifier facilement que les résultats obtenus par le modèle de Cox sont identiques à ceux obtenus précédemment.

Introduction d'une intéraction avec une fonction de la durée

On a une variable de durée (on prendra $g(t) = t$) qui sera croisée avec la variable *surgery*.

Le modèle s'écrit:

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

Le modèle avec cette intéraction donne les résultats suivants:

Table 9.6: Modèle de Cox avec une intéraction entre une fonction de la durée et la variable *surgery

Variable	e^b	Std.err	z	P>|z|	95% IC
year	0.884	0.059	-1.84	0.066	0.776 ; 1.008
age	1.029	0.014	+2.15	0.032	1.003 ; 1.057
$surgery(t_{0+})$	0.173	0.117	-2.60	0.009	0.046 ; 0.649
$surgery \times t$	1.002	0.001	+2.02	0.043	1.000 ; 1.004

On retrouve donc un résultat proche de celui obtenu à partir du test OLS sur les résidus de Schoenfeld pour la variable *surgery*. Et c'est normal. Avec $g(t) = t$, il a le mérite de pouvoir être interprété directement. Ce qui ne veut pas dire qu'il s'agit de la meilleure solution.

Donc, malgré une hypothèse plutôt forte sur la forme fonctionnelle de l'intéraction, et dans les faits surement pas pertinente, on peut dire que chaque jour le rapport des risques entre personnes opérées et personnes non opérées augmente de +0.2%. Pour plus précis, étant à l'origine <1, l'écart se modère. L'effet de l'opération sur la survie des individus s'estompe donc avec le temps.

A noter

- 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.

Si $surgery = 0$

id	surgery	died	t_0	t	surgery*t
2	0	0	0	1	0
2	0	0	1	2	0
2	0	0	2	3	0
2	0	0	3	5	0
2	0	1	5	6	0

Si $surgery = 1$ (jusqu'à $t = 6$ car aucun décès précoce pour ce groupe)

id	surgery	died	t_0	t	surgery*t
40	1	0	0	1	1
40	1	0	1	2	2
40	1	0	2	3	3
40	1	0	3	5	5
40	1	1	5	6	6

Exemple pour une variable quantitative (*age*)

id	age	died	t_0	t	age*t
2	51	0	0	1	51
2	51	0	1	2	102
2	51	0	2	3	153
2	51	0	3	5	255
2	51	1	5	6	306

- L'altération des rapports de risque dépend de la forme fonctionnelle de l'intéraction choisie. Ici la variation dans la durée 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, qui devient lui même strictement constant.

9.2.7 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 respectent pas l'hypothèse, l'estimation de la baseline du risque pourrait être sensiblement affectée si l'analyse a des visée prédictives. Il convient tout de même lors de l'interprétation, de préciser les variables qui seront analysées sous cette forme très « 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 écarts de risque, s'accentuent à la hausse comme à la baisse, on peut conserver cet estimateur moyen. Mais si cette non proportionnalité conduit à un changement du sens des rapport de risque je suis moins convaincu de la pertinence de cette stratégie. Encore une fois, et il faut le rappeler, l'estimation des courbes de survie doit permette d'anticiper ce dernier cas de figure.

Il faut également 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 [**important**]… 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 observable. Des estimations, plus complexes, sont possibles dans ce cas, et sont en mesure malgré leur interprétation plutôt difficile de régler le problème. 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 prédites ajustées, et si une seule covariable (binaire) présente un problème. Les HR ne seront pas estimés pour la variable qui ne respecte pas l'hypothèse.

Intéraction

Introduire une interaction avec la durée. Cela peut permettre en plus d'enrichir le modèle au niveau de l'interprétation. Valable si peu de covariables présentent des problèmes de stabilité des rapports de risque, 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…. on introduit de nouveau une contrainte de proportionnalité dans le modèle.

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 ou un modèle à temps discret. Pour la dernière solution, on peut également corriger la non proportionnalité avec l'introduction d'une intéraction. Si on ne le fait pas, les risques prédits, par définition des 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 (non traité). Mais ces modèles, dont les résultats sont présentés par des graphiques, se commentent assez difficilement.

Forêt aléatoire
Autre méthode : les forêts aléatoires. 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.

10 Modèle à durée discrète

On va principalement traiter du ***modèle logistique à durée discrète***.

- Par définition ce n'est pas un modèle à risques proportionnels, mais à **Odds proportionnels**. Toutefois en situation de rareté (p<10%), l'Odds converge vers une probabilité, qui est une mesure du risque.
- Le modèle à durée discrète est de type pleinement paramétrique, il est moins contraignant que le modèle de Cox si l'hypothèse de proportionnalité n'est pas respectée, car le modèle est ajusté par une fonction de la durée.
- Pour être estimé, la base de données doit être transformée en format long: aux durées d'observation ou sur des intervalles de durée choisis. C'est une des principales différences avec le modèle de Cox qui est une estimation aux moments d'évènement. Néanmoins avec une bonne forme fonctionnelle de la durée traitée comme une variable quantitative, et si le nombre de points d'observation est suffisamment grand, les deux modèles aboutissent à des résultats quasiment identiques.
- Ce modèle permet d'introduire de manière plutôt souple un ensemble de covariables no fixes.

Avec un lien logistique, le modèle à durée discrète, avec seulement des covariables fixes, peut s'écrire:

$$log\left[\frac{P(Y=1 \mid t_p, X_k)}{1-P(Y=1 \mid t_p, X_k)}\right] = a_0 + \sum_p a_p f(t_p) + \sum_k b_k X_k$$

10.1 Organisation des données

Format long

Les données doivent être en format long: pour chaque individu on a une ligne par durée observée ou par intevalle de durées jusqu'à l'évènement ou la censure. On retrouve le *split* des données du modèle de Cox, mais généralisé à des intervalles où aucun évènement n'est observé. Avec des données de type discrètes ou groupées, phénomène classique en sciences sociales, il y a souvent peu de différence entre un allongement aux temps d'évènement et aux temps d'observation.

Durée

La durée est dans un premier temps construite sous forme d'un simple compteur, par exemple $t = 1, 2, 3, 4, 5...$ (des valeurs non entières sont possibles). Le choix de la forme fonctionnelle de la durée sera présentée plus tard.

Variable évènement/censure

Si l'individu a connu l'évènement, elle prend la valeur 0 avant celui-ci. Au moment de l'évènement sa valeur est égale à 1. Pour les observations censurées, la variable prend toujours la valeur 0.

Application
On reprend les données de la base *transplantation*, mais les durées ont été regroupées par période de 30 jours. Il n'y a pas de durée mesurée comme nulle, on a considéré que les 30 premiers jours représentaient, le premier mois d'exposition. Cette variable de durée se nomme *mois*.

Format d'origine

Table 10.1: Durée discrète: données en format d'origine

id	year	age	surgery	mois	died
1	67	30	0	2	1

La personne décède lors du deuxième intervalle de 30 jours

Format long et variables pour l'analyse

Table 10.2: Durée discrète: données en format long

id	year	age	surgery	mois	died	t
1	67	30	0	2	0	1
1	67	30	0	2	1	2

10.2 Ajustement de la durée

Un des principaux enjeux réside dans la paramétrisation de la durée:

- Elle peut-être modélisée sous forme de fonction d'une variable de type quantitative/continue.
- Elle peut-être modélisée comme variable discrète, de type indicatrice $0; 1$, sur tous les points d'observation, ou sous forme de regroupements. Il doit y avoir au moins un évènement observé dans chaque intervalle.

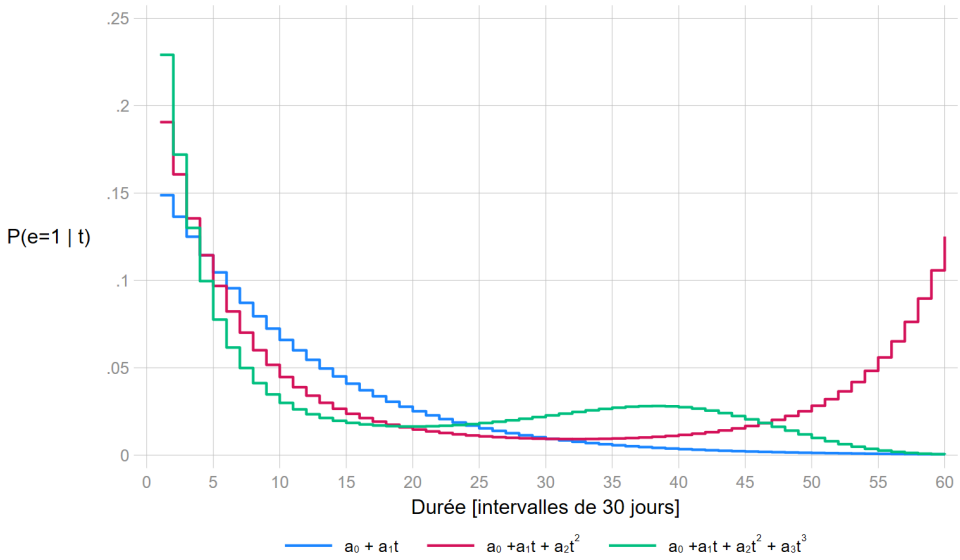
10.2.1 Ajustement avec une durée en continu

Le modèle étant paramétrique, on doit trouver une fonction qui ajuste le mieux les données. Toutes transformations de la variable est possible: $f(t) = a \times t$, $f(t) = a \times ln(t)$……formes quadratiques. Les ajustements sous forme de **splines** (cubiques) tendent à se développer ces dernières années.

Pour sélectionner cette fonction, on peut tester différents modèles sans covariable additionnelle, et sélectionner la forme dont le critère d'information de type **AIC** (vraisemblance pénalisée) est le plus faible.

Exemple:
On va tester les paramétrisations suivante:s une forme linéraire stricte $f(t) = a \times t$ et des effets quadratiques d'ordres 2 et 3: $f(t) = a_1 \times t + a_2 \times t^2$ et $f(t) = a_1 \times t + a_2 \times t^2 + a_3 \times t^3$.

Figure 10.1: Probabilité de décéder avec 3 ajustements de la durée



Critères AIC

$f(t)$	AIC
$a \times t$	504
$a_1 \times t + a_2 \times t^2$	492
$a_1 \times t + a_2 \times t^2 + a_3 \times t^3$	486

On peut utiliser la troisième forme à savoir $a_1 \times t + a_2 \times t^2 + a_3 \times t^3$.

Estimation du modèle avec toutes les covariables

Table 10.4: Modèle logistique à durée discrète ($f(t)$ continue)

Variables	OR - RR	Std. err	z	P>|z|	95% IC
t	0.678	0.057	-4.52	0.000	0.587 ; 0.810
t^2	1.014	0.005	+2.83	0.005	1.004 ; 1.024
t^3	1.000	0.000	-2.11	0.035	1.000 ; 1.000
year	0.876	0.015	-1.80	0.072	0.758 ; 1.012
age	1.034	0.163	+2.27	0.023	1.005 ; 1.064
surgery	0.364	0.110	-2.25	0.024	0.151 ; 0.877
Constante	*0.440*	*0.110*	*-3.29*	*0.001*	*0.270 ; 0.718*

Remarque: les variables *year* et *age* ont été centrée sur leur moyenne pour rendre la constante interprétable. La constante reporte donc l'Odds de décéder lors des 30 premiers jours d'une personne dont

l'âge et l'année à l'entrée dans le registre est égal à l'âge et à l'année moyenne et qui n'a pas été opéré préalablement.

Si maintenant on estime un modèle de Cox sur ces données journalières groupées, on remarque que les résultats obtenus sont très proches

Table 10.5: Modèle de Cox

Variables	OR - RR	Std. err	z	P>|z|	95% IC
year	0.878	0.059	-1.93	0.053	0.769 ; 1.002
age	1.029	0.014	+2.13	0.033	1.002 ; 1.057
surgery	0.379	0.165	-2.22	0.026	0.111 ; 0.892

10.2.2 Ajustement discret

- Il s'agit d'introduire la variable de durée dans le modèle comme une variable catégorielle (indicatrices).
- Démarche pas conseillé si on a beaucoup de points d'observation, ce qui est le cas ici.
- A l'inverse, si peu de points d'observation la paramétrisation avec une durée continue n'est pas conseillé.
- La correction de la non proportionnalité peut être plus compliquée à mettre en oeuvre.

On va supposer que l'on ne dispose que de 4 intervalles d'observation. Pour l'exemple, on va créer ces points à partir des quartiles de la durée, et conserver pour chaque personne une seule observation par intervalle.

- $t = 1$: Entre le début de l'exposition et 4 mois.
- $t = 2$: Entre 5 mois et 11 mois .
- $t = 3$: Entre 12 mois et 23 mois.
- $t = 4$: 24 mois et plus.

On va estimer le risque globalement sur l'intervalle. La base sera plus courte que la précédente (197 observations pour 103 individus). Il ne sera plus possible ici d'interpréter les résultats en termes de rapport de probabilité, l'évènement devenant trop fréquent à l'intérieur de chaque intervalle.

Table 10.6: Modèle logistique à durée discrète ($f(t)$ indicatrices)

Variables	OR - RR	Std. err	z	P>|z|	95% IC
$0-4mois$	2.811	1.177	+2.47	0.014	1.237 ; 6.387
$5-11mois$	ref	-	-	-	-
$12-23mois$	0.559	0.346	-0.94	0.347	0.166 ; 1.881
$24-46mois$	1.741	1.159	+0.83	0.405	0.472 ; 6.417
year	0.816	0.076	-2.18	0.029	0.680 ; 0.980
age	1.048	0.019	+2.53	0.011	1.011 ; 1.087
surgery	0.330	0.166	-2.21	0.027	0.123 ; 0.882

Variables	OR - RR	Std. err	z	P>|z|	95% IC
Constante	*0.407*	*0.151*	*2.43*	*0.015*	*0.198 ; 0.840*

On trouve des résultats proches de ceux éstimés avec un ajustement continu de la durée. C'est normal, la dirée fait office de variable d'ajustement peu ou pas corrélée avec les autres variables introduites.

Variables	Ajustement discret	Ajustement continu
year	0.816	0.876
age	1.048	1.034
surgery	0.330	0.364

10.3 Proportionnalité des risques

- Formellement un modèle logistique à temps discret repose sur une hypothèse d'Odds proportionnel [Odds ratios constants pendant la durée d'observation]. Contrairement au modèle de Cox, l'estimation des probabilités (risque) n'est pas biaisée si l'hypothèse PH n'est pas respectée, les paramètres estimés sont considérés au pire comme des approximation.
- Comme pour le modèle de Cox, la correction de la non proportionnalité peut se faire en intégrant une interaction avec la durée dans le modèle.

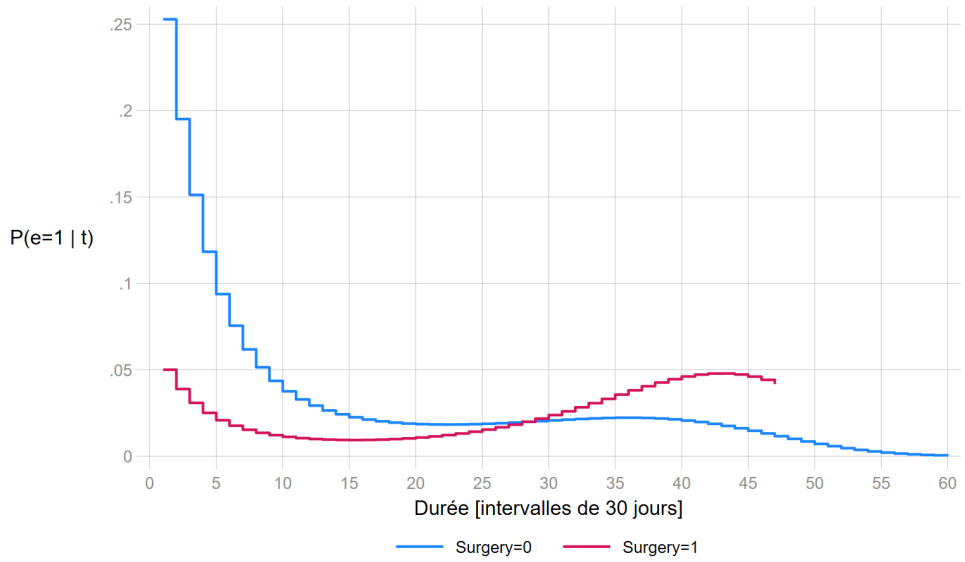
Avec un ajustement continue, on remarque de nouveau que le résultat du modèle est de nouveau très proche de celui estimé avec un modèle de Cox.

Table 10.8: Modèle logistique à durée discrète avec correction de la non proportionnalité

Variables	OR - RR	Std. err	z	P>|z|	95% CI
t	0.702	0.059	-4.2	0.000	0.595 ; 0.828
$surgery(t=0)$	0.155	0.108	-2.67	0.008	0.039 ; 0.609
$surgery \times t$	1.072	0.036	2.08	0.037	1.004 ; 1.145
t^2	1.013	0.005	2.37	0.018	1.002 ; 1.023
t^3	1.00	0.000	-1.71	0.086	1.000 ; 1.000
year	0.872	0.064	-1.86	0.062	0.755 ; 1.007
age	1.033	0.015	2.23	0.026	1.004 ; 1.063
constante	*0.445*	*0.112*	*-3.22*	*0.001*	*0.272 ; 0.728*

Si on avait omis les variables *year* et *age* du modèle:

Figure 10.2: Probabilité de décéder après correction de la non proportionnalité pour la variable surgery



11 Variables dynamiques

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 ce type de variable.
- 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 en 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.

11.1 Facteur dynamique traitée de manière fixe

On reprend l'exemple sur malformation cardiaque, en ajoutant la variable relative à la greffe. La question est donc de savoir si une transplantation du coeur réduitle 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 avec la variable fixe *transplant*.

Table 11.1: Modèle de cox avec une variable dynamique (binaire) traitée de manière fixe (estimation biaisée

Variables	HR	Std. err	z	P>|z|	95% CI
year	0.910	0.060	-1.42	0.155	0.799 ; 1.036
age	1.054	0.015	3.71	0.000	1.025 ; 1.084
surgery	0.541	0.243	-1.37	0.171	0.224 ; 1.304
transplant	0.278	0.088	-4.06	0.000	0.150 ; 0.515
wait	0.992	0.005	-1.50	0.134	0.982 ; 1.002

Interprétation: traitée de manière fixe, la greffe réduit donc sensiblement le risque journalier de décéder (RR=0.278). De même on peut admettre une certaine cohérence pour la durée jusqu'à la transplantation: plus elle est précoce et plus les personnes survivent (HR=0.992).

Sauf que…..

Au niveau des données le modèle à été estimé, pour une personne greffée (ici id=70), à partir de ce mapping:

Table 11.2: Mapping de la base avec une variable dynamique binaire traitée de manière fixe

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

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).

11.2 Estimation avec une variable dynamique

Il convient donc de modifier l'information avec le délai d'attente jusqu'à la greffe. Le principe de construction de la variable dynamique, quelle que soit le logiciel utilisé, doit suivre la logique suivante:

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

11.2.1 Modèle de Cox

Table 11.3: Mapping correct de la base avec une variable dynamique binaire

id	year	age	surgery	transplant	wait	died	t_0	t	**TVC**
70	72	52	0	1	5	0	0	1	**0**
70	72	52	0	1	5	0	1	2	**0**
70	72	52	0	1	5	0	2	3	**0**
70	72	52	0	1	5	0	3	5	**0**
70	72	52	0	1	5	0	5	6	**1**

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

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

Table 11.4: Modèle de Cox avec une variable dynamique binaire

Variables	HR	Std. err	z	P>|z|	95% CI
year	0.887	0.060	-1.79	0.074	0.777 ; 1.012
age	1.031	0.014	2.19	0.029	1.003 ; 1.059
surgery	0.374	0.163	-2.25	0.024	0.159 ; 0.880
TVCtransplantation	0.921	0.281	-0.27	0.787	0.507 ; 1.674

L'impact de la greffe apparaît maintenant bien plus modéré sur la survie des individus. 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. Il faut également garder en tête que l'état de santé des personnes est particulièrement dégradé, cette opération étant celle de la *dernière chance*.

R - Stata - Sas - Python

11.2.2 Sas

La base n'est pas modifiée et la création de la TVC est faite *en aveugle* dans la procédure `phreg`, après l'instruction `model`. Ce n'est franchement pas super.

11.2.3 R - Stata, Python

La base doit être transformée en format long aux temps d'évènement (`survsplit` avec R, `stsplit` avec Stata) avant la création de la variable dynamique.

11.2.4 Modèle à temps discret

Même principe pour la construction de la variable dynamique. Pour rappel l'échelle temporelle est le mois, on a créé en amont une variable qui regroupe les valeurs de la variable *wait* en périodes de 30 jours.

Table 11.5: Modèle logistique à durée discrète avec variable dynamique binaire

Variables	OR - RR	Std. err.	z	P>|z|	95% IC
t	0.686	0.070	-3.71	0.000	0.562 ; 0.837
t^2	1.015	0.006	2.53	0.011	1.003 ; 1.026
t^3	1.000	0.000	-1.97	0.049	1.000 ; 1.000
$year$	0.876	0.065	-1.79	0.073	0.758 ; 1.012
age	1.034	0.015	2.22	0.027	1.004 ; 1.064
$surgery$	0.363	0.163	-2.25	0.024	0.151 ; 0.876
$TVC\ greffe$	1.029	0.355	0.08	0.934	0.524 ; 2.022
Constante	*0.440*	*0.110*	*-3.29*	*0.001*	*0.270 ; 0.718*

11.3 Précautions

- Rappel: la cause doit précèder l'effet.
- Lorsque l'évènement étudié n'est pas intrinsèquement de type absorbant comme le décès, la *cause* peut se manifester ou plutôt être observée après la survenue de l'évènement étudié. Les modèles de durée standards ne peuvent pas gérer ces situations car l'observation sort du risque après la survenue de l'évènement. Il y a d'autres techniques, par exemple de type économétrique, qui sont plus à même de traiter ce genre de situations.
- Même si la cause est bien mesurée avant l'évènement d'intérêt, un *choc* n'est peut-être qu'un point final d'un processus causal antérieur: une séparation est rarement un évènement ponctuel, une phase plus ou moins longue de mésentente dans le couple lui a vraisemblablement préexister. La datation du début d'un processus causal n'est donc pas toujours facile à mesurer.
  - **Logique d'adaptation**: la *cause* identifiée est mesurée avant l'évènement étudié.
  - **Logique d'anticipation**: la *cause* identifiée est mesurée après l'occurrence de l'évènement étudié. L'origine causale est bien antérieure à l'évènement, mais elle n'est pas directement observable.
- Lorsque les variables dynamiques sont de type quantitatives/continues, le problème on doit aussi considérer avec des phénomènes d'anticipation sur les valeurs attendues de ces variables, observées postérieurement à l'évènement étudié. On peut introduire des « lags » dans le modèle pour saisir ce phénomène : par exemple $x_t = x_{t+1}$. Ce décalage des durées d'occurrence peut être aussi introduite pour les variables discrètes (naissance d'un enfant par exemple).

partie V

Compléments

partie VI

Annexe