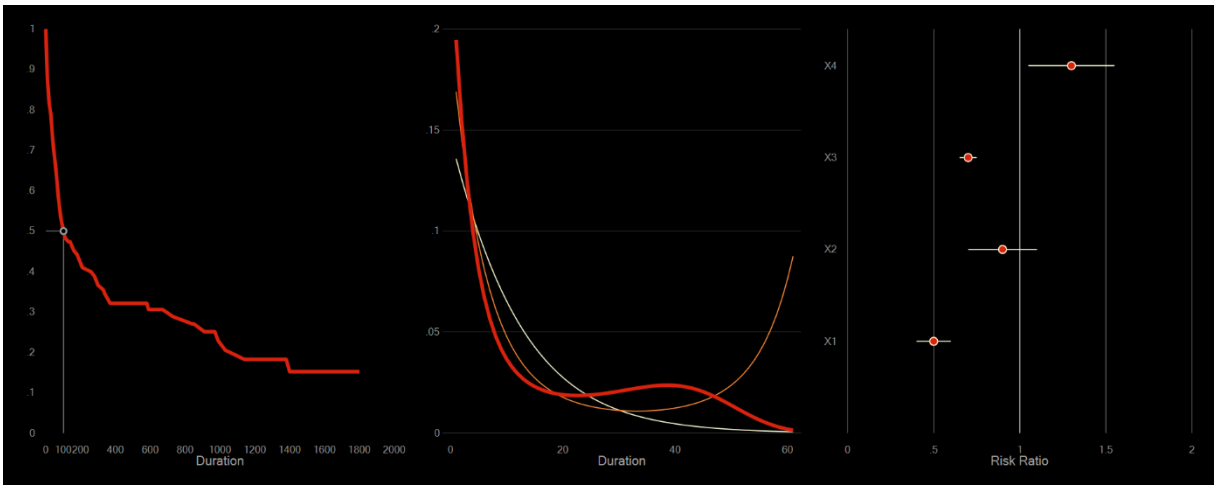
FORMATION ANALYSE DES DUREES 2021



Marc Thévenin

Version du 11 mars 2021 [données et théorie – analyse non paramétrique – modèles à risque proportionnels]

Table des matières

Introduction

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 et changer au cours de la période d'observation.

Terminologies

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

Exemples d'analyse

Nuptialité, Mise en couple
cohabiter, décohabiter, se marier, rompre une union …

Logement
Changement de statut (locataire/propriétaire), mobilité résidentielle, migration …

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…

Eléments nécessaires à l'analyse

Un processus temporel :

- Une échelle de mesure (minutes, heures, jours, mois, années...)
- Une origine commune définissant un évènement de départ (âge au mariage, âge aux premières règle...
- Une définition précise de l'évènement d'étude.
- Une durée entre le début et la fin de la période d'observation.

Une population soumise au risque de connaître l'évènement. Elle est appelé **Risk Set**.

Des variables « explicatives » ou **covariables** :

- Fixes: genre, génération, niveau de diplôme, CSP…
- Dynamiques: Mesurées à tout moment entre le début et la sortie de l'observation: statut matrimonial, taille du ménage, niveau de revenu, statut d'activité… Elles sont souvent dénommées **Time Varying Covariates**.

Plan de la formation

Objectifs du document

[A faire après reprise du doc]

Bibliographie

[A faire après le reprise du doc]

Données et théorie

Les données biographiques

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

Données prospectives et rétrospectives

Les données prospectives

- Individus suivis à des dates successives.
- Instrument de mesure identique à chaque vague (si possible).
- **Avantages**: qualité des données.
- **Inconvénients**: coût important, délais pour les exploiter dans une analyse, mêmes hypothèses entre deux passages pas forcément respectées, problèmes d'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 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 liés, par exemple, à l'inférence.

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, faible coût.
- **Inconvénients** Questionnaire long, informations datées qui font appel à la mémoire de l'enquêté.e. A de rares exceptions (enfant.s, mariage.s), il est difficile d'aller chercher des datations trop fines.

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

Grille AGEVEN

Pour recueillir des informations biographiques retrospectives, on utilise généralement la méthode des grilles AGEVEN

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

Références:

- 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

Enregistrement des données

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

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

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

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

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

Semi-long [format individu-évènements]

C'est le format le plus courant de mise à disposition des enquêtes biographiques. Si l'évènement est de type continu, par exemple on a toujours un lieu de résidence, 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 en cours au moment de l'enquête la date de fin est souvent une valeur manquante. Pour les trajectoire discontinues dont une date de début ou de fin d'épisode serait oubliée on peut, si le logiciel le permet (Sas, Stata, Spss), utiliser une valeur manquante informative.

Exemple précédent (trajectoires discontinues):

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

Long [format individu-périodes]

Typique des recueils prospectifs. Ils engendrent des lignes sans informations supplémentaires.

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 (niveau "statut"). 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.

Exemples de mise à disposition

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

Enquête biographie et entourage (Ined)

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

Base sur les caractéristiques individuelles

VIEWTABLE: TMP1.tego

	Identifiant questionniaire	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

Module biographique sur le logement et les lieux de résidence

	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

Enquête MAFE (Ined)

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

Base sur les 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

Module biographique sur les lieux de résidence

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

La théorie de l'analyse des durées

L'analyse des durées peut être vue comme l'étude d'une variable aléatoire T qui décrit le temps 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, c'est-à-dire dire lorsque la durée d'observation est inférieure à la durée d'exposition au risque.

Temps et durée

Le temps est une dimension (la quatrième), la durée est sa mesure. La durée est tout simplement calculer 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 durée : la ***durée continue*** et la ***durée 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 strictement continu car deux évènements ne peuvent pas avoir 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 de la théorie de la relativité, une minute (voire une seconde) est une mesure très discrète pour ne pas dire grossière du temps, pour un géologue c'est une mesure continue. Pour ces deux disciplines, cette échelle de mesure n'est pas adaptée à leur activité. Le choix de l'échelle temporelle doit être pertinent par rapport aux objectifs de l'analyse.
Il existe néanmoins des cas où les durées sont par nature discrète, lorsqu'un évènement ne peut avoir lieu qu'à un moment précis.

A retenir

Durée continue : absence (ou très peu) d'évènements simultanés. Durée discrète/groupée : présence d'évènements simultanés (en grand nombre).

Le Risk Set

Il s'agit de la population « soumise » ou « exposée » au risque lorsque $T = t_i$.
Cette population varie dans le temps car:

- Certaines personnes ont connu l'évènement, donc peuvent ne plus être soumises au risque (exemple: 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, perdus de vue, fin de l'observation à une durée peu avancée dans un recueil rétrospectif.

Exemples:
Les individus célibataires sont soumis au risque ….…[remplir]

Les individus mariés sont soumis au risque ……[remplir]

Les individus au chômage sont soumis au risque ……..[remplir]

Les individus qui travaillent sont soumis au risque ….[remplir]

Les individus vivants sont soumis au risque …….…..[remplir]

La Censure

Définition de la censure

Une observation est dite censurée lorsque la durée d'observation est inférieure à la durée d'exposition au risque.

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 un 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 sans connaître 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 et on souhaite mesurer l'âge de la naissance de leur premier enfant. Au moment de l'enquête, une femme est âgée de 29 ans et 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.

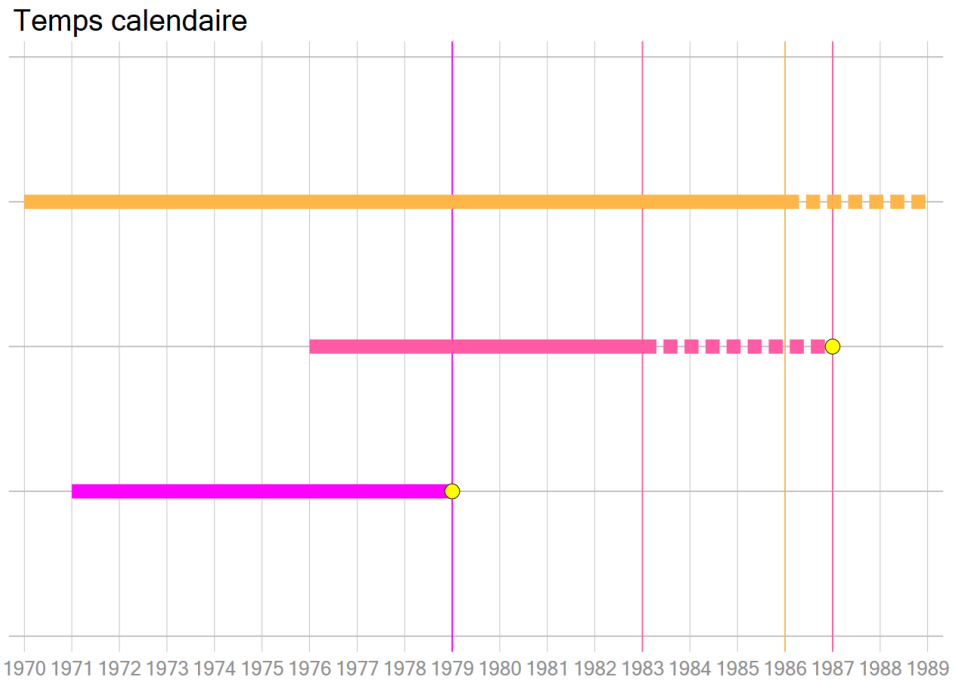
Problème posé par la censure informative

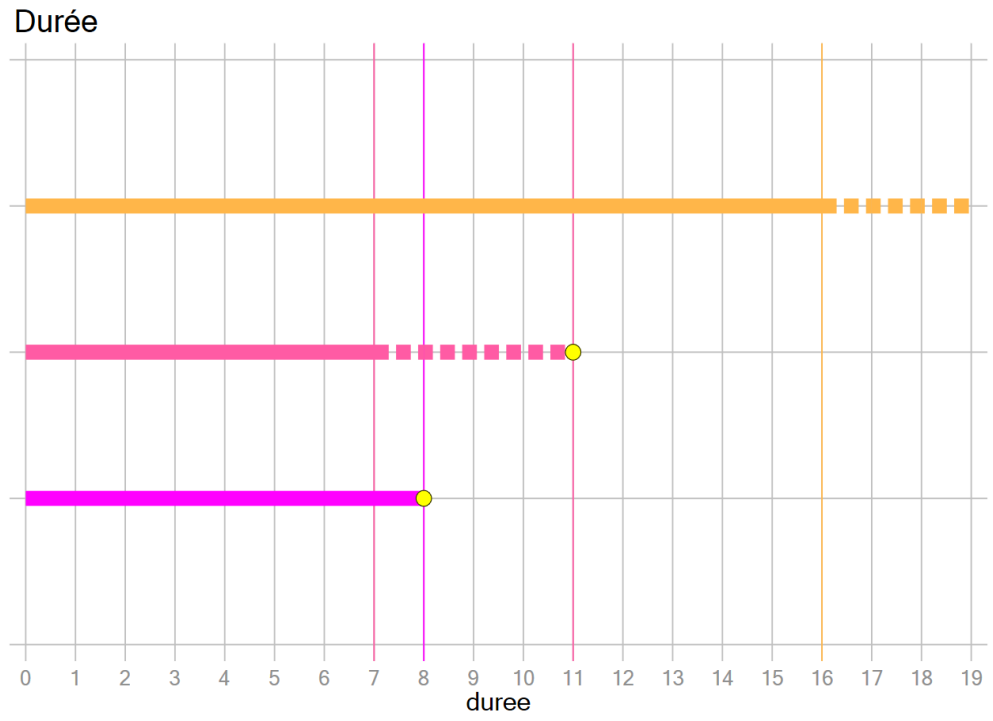
Par exemple en analysant des décès avec un recueil prospectif, si un individu est perdu de vue en raison d'une dégradation de son état de santé, l'indépendance entre la cause de la censure et le décès ne peut plus être assurée.

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 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 informative, 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 censurés
- Bulle : moment de l'évènement





Censure à gauche, troncature et censure par intervalle

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

Censure par intervalle
Un évènement peut avoir lieux entre 2 temps d'observations sans qu'on puisse les observer (ex: en criminologie récidive d'un délit entre deux arrestations).

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

Troncature à gauche (late-entry)
Par l'exemple, on analyse la survie d'une population. Seule la survie des individus vivants à l'inclusion peut être analysée. On peut également trouver un phénomène de troncature lorsqu'on mesure la durée à partir d'un certain seuil niveau temporel (ce qui autorise aussi des phénomènes de troncature à droite).

Ce type de phénomène peut affecter directement le phénomène étudié. Si on souhaite analyser la question de ou des tentatives de suicide de la personne interrogée via une enquête, par définition ont pourra seulement récolter l'information des survivant.e.s. Cela rend difficile une analyse sur le suicide en général.

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.

Les grandeurs

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 techniques de calcul, mais des grandeurs dont on précisera seulement les propriétés et qui feront l'objet de techniques d'estimation.

La fonction de Survie $S(t)$

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

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

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

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

Propriétés: $S(0) = 1$ et $\lim_{t \to \infty} S(t) = 0$

La fonction de survie strictement non croissante.

La fonction de répartition $F(t)$

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

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

Comme $t \leq 0$, $F(t) = 1 - S(t)$. Fonction de survie et fonction de répartition sont donc deux grandeurs strictement complémentaires.

Propriétés: $F(0) = 0$ et $\lim_{t \to \infty} F(t) = 1$

La fonction de répartition strictement non décroissante.

La fonction de densité $f(t)$

Pour une valeur de t donnée, la fonction de densité de l'évènement donne la probabilité de connaitre l'évènement dans un petit intervalle de temps après t. Si dt est proche de 0 alors cette probabilité tend également vers 0. On norme donc cette probabilité par dt.
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:

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

Dans les analyses on utilise pas ou peu les densités, qui sont plutôt mobilisées lors des calculs des estimateurs.

Le risque instantané $h(t)$

Concept fondamental de l'analyse des durées:

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

- $P(t \leq T < t + dt | T \geq t)$ donne la probabilité de survenue de l'évènement sur l'intervalle $[t, t + dt[$ ***conditionnellement à la survie au temps t***.
- La quantité obtenue donne un nombre moyen d'évènements que connaîtrait un individu durant une unité de temps choisie.
- A priori *cette quantité n'est pas une probabilité*. C'est la nature de l'évènement, en particulier sa non récurrence, et la métrique temporelle choisie ou disponible qui peut la rendre assimilable à une probabilité.

On peut écrire également:

$$h(t) = \frac{f(t)}{S(t)} = \frac{F\prime(t)}{S(t)} = -\frac{S\prime(t)}{S(t)}$$

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

Le risque cumulé $H(t)$

Le risque cumulé est égal à :

$$H(t) = \int_0^t h\,(u)du = -log(S(t))$$

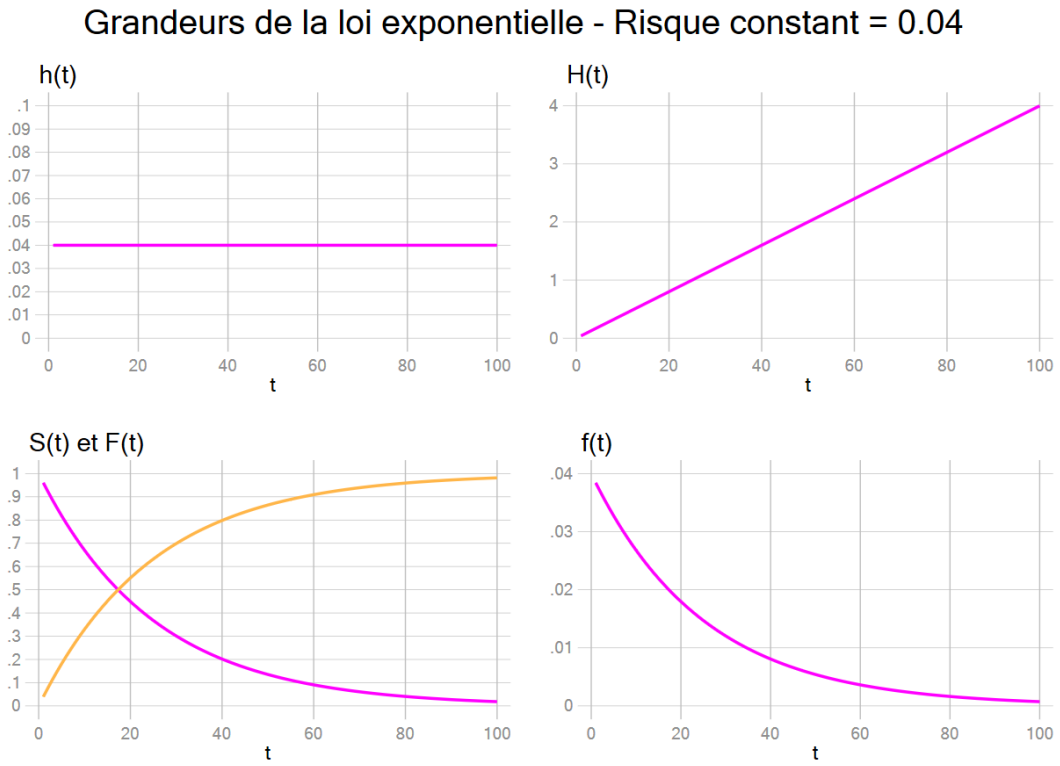
On peut alors le réécrire toutes les autres quantités:

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

Exemple

Si on pose que le risque instantané est strictement constant au cours du temps: $h(t) = a$ (loi dite exponentielle, typique des processus sans mémoire):

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



Application: risque et échelles temporelles

Fortement inspiré, pour ne pas dire copié, de l'*excellent* cours de **Gilbert Colletaz**: https://www.univ-orleans.fr/deg/masters/ESA/GC/sources/Econometrie%20des%20Donnees%20de%20Survie.pdf

Attention on sort ici très clairement du temps continu, il s'agit seulement de manipuler les concepts, et de voir la dépendance de la mesure du risque à l'échelle temporelle. Par ailleurs on inverse plutôt la logique de « l'instantanéité » en augmentant les intervalles de durée (du mois au trimestre ou à l'année).

- Durant les mois d'hiver, disons 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 à attraper 1.44 rhume durant la période d'hiver.

- 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

  $\frac{0.33}{1/12} = 3.96$. On peut donc s'attendre à mourir près de 4 fois durant les 12 mois.

Le risque peut donc être supérieur à 1. En soit cela ne pose pas de problème comme il s'agit d'un nombre moyen d'évènements espérés durant une unité de temps, mais pour des évènements qui ne peuvent pas se répéter, évènements dits « *absorbants* », l'interprétation n'est pas très intuitive.

On peut donc prendre l'inverse du risque qui mesure la durée moyenne (espérée) jusqu'à l'occurence de l'évènement.

On retrouve ici 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, mais jusqu'à quand on peut espérer survivre.

- Pour le rhume, la durée espérée d'attraper le premier rhume est de $1/1.44 = 0.69$ du trimestre hivernal, soit approximativement le début du mois de mars.
- Pour l'année sabbatique, la durée moyenne de survie est de $1/3.96 = 0.25$ d'une année, soit 3 mois après l'arrivée dans cette région visiblement peu accueillante.

Exercice

- On a une population de 100 cochons d'Inde.
- On analyse leur mortalité (naturelle).
- Ici l'analyse est en durée discrète/groupée.
- La durée représente le nombre d'années de vie.
- Il n'y a pas de censure à droite.

Durée	Nombre de décès
1	1
2	1
3	3
4	9
5	30
6	40
7	10
8	3
9	2
10	1
N=100	

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

Méthodes non paramétriques

Les méthodes non paramétriques portent généralement sur l'analyse des fonctions de survie (séjour) ou sur celle des fonctions de répartitions, 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 méthodes sont adaptées à des mesures différentes de la durée : plutôt discrète pour la technique actuarielle et plutôt continue pour Kaplan-Meier (KM). Cela induit un traitement différent de la censure dans l'estimation. La seconde est de très très loin la plus diffusée, surement en raison des tests de comparaison qu'elle est en mesure de fournir.

Introduction

Les variables d'analyse

On a un échantillon aléatoire de n individus avec:

- Des indicateurs de fin d'épisode $e_1, e_2, \ldots, e_k$ 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 $t_1, t_2, \ldots, t_k$ jusqu'à l'évènement ou la censure.
- En théorie, il ne peut pas y avoir d'évènement en $t = 0$.

Exemple (temps continu, pas d'évènement simultané):

t_i	d_i	Commentaires
1	0	Censure
2	1	Evènement
5	1	Evènement
10	0	Censure
11	1	Evènement

Calculer le Risk set pour $t = 2$, $t = 5$ et $t = 11$. En $t = 0$, $R = 5$.

Calcul de la fonction de survie (séjour)

Rappel: La fonction de survie donne la probabilité que l'évènement survienne après t_i, soit $S(t_i) = P(T > t_i)$.

Pour survivre en t_i , il faut avoir survécu en t_{i-1} , t_{i-2} , …., t_1 . La fonction de survie rapporte donc des probabilités conditionnelles: survivre en t_i conditionnellement au fait d'y avoir survécu avant. Il s'agit donc d'un produit de probabilités.

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

$S(t_i) = (1 - \frac{d_i}{r_i}) \times S(t_{i-1}) = S(t_i) = (1 - q(t_i)) \times S(t_{i-1})$.

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

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

Application pour la suite de la formation

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*)

L'âge à l'entrée dans le registre (*age*)

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 est le décès (*died*).

La méthode actuarielle

- Estimation sur des intervalles définis par l'utilisatrice/utilisateur.
- Au niveau technique, la méthode est dite «continue», avec une estimation en milieu d'intervalle.
- Méthode adaptée lorsque la durée est mesurée de manière discrète.

Echelle temporelle

La durée est divisée en J intervalles, en choisissant J points: $t_0 < t_1 < \ldots < t_J$ 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

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

Calcul de la durée médiane (ou autre quantiles)

Rappel

Compte tenu des censures à droite, le dernier intervalle étant ouvert, il est déconseillé voire proscris de calculer des durées moyennes. On utilise la médiane ou tout autre quantile lorsqu'ils sont estimables.

Définition:

Il s'agit de la valeur 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 $S(t_i) > 0.5$.

Logiciels

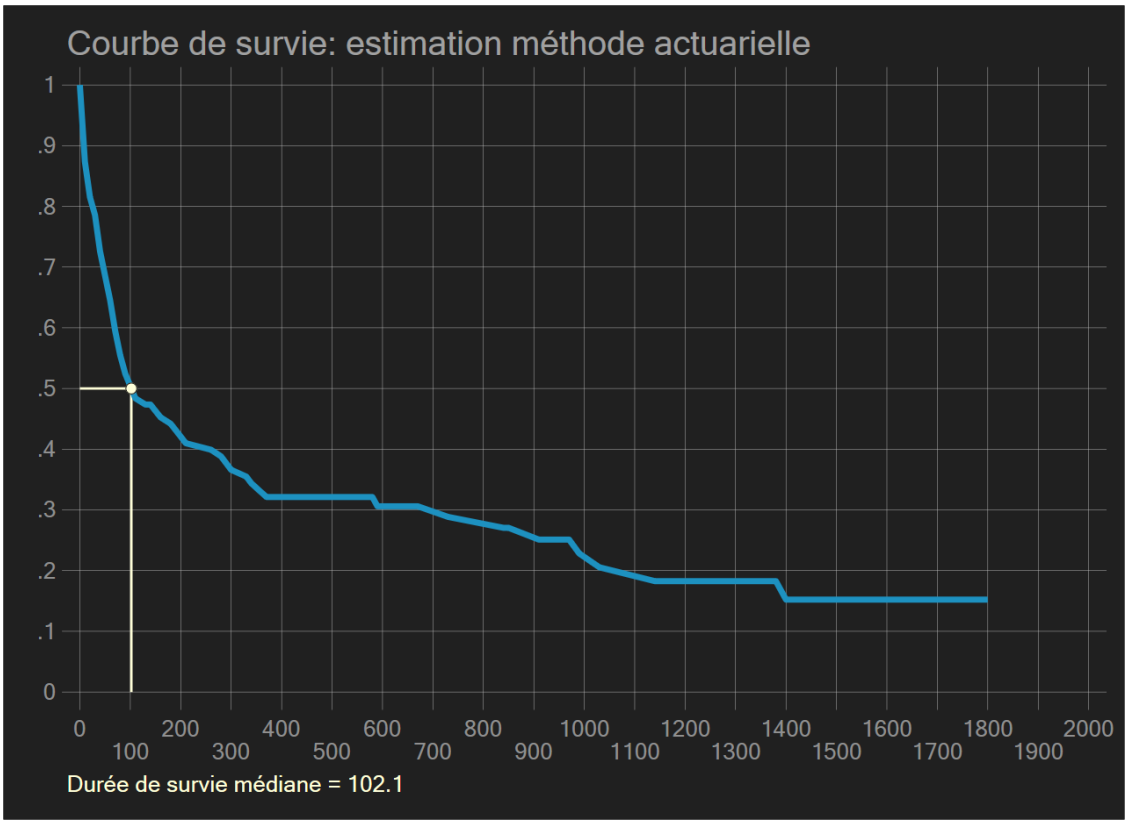
SAS: incluse dans ***proc lifetest***.

Stata: commande ***ltable***. Voir la commande externe ***qlt*** (MT) qui calcule les durées médianes (+ quartiles) et qui cale la définition des intervalles avec celle de Sas.

R: une fonction programmée par un utilisateur (package ***discSurv*** => fonction ***lifeTable***), mais pas convaincante car pas d'estimation sur les quantiles, et estimation avec des intervalles toujours fixés à $dt = 1$. D'un intérêt très limité.

Python: à l'heure actuelle, aucune fonction à ma connaissance.

Exemple



Interval		Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
0	10	103	13	0	0.8738	0.0327	0.7926	0.9247
10	20	90	6	1	0.8152	0.0383	0.7257	0.8779
20	30	83	3	0	0.7857	0.0405	0.6931	0.8533
30	40	80	6	2	0.7261	0.0441	0.6284	0.8020
40	50	72	4	0	0.6857	0.0461	0.5857	0.7664
50	60	68	4	0	0.6454	0.0476	0.5439	0.7299
60	70	64	5	0	0.5950	0.0489	0.4926	0.6834
70	80	59	4	0	0.5546	0.0496	0.4523	0.6454
80	90	55	3	0	0.5244	0.0499	0.4225	0.6165
90	100	52	2	0	0.5042	0.0499	0.4029	0.5971
100	110	50	2	1	0.4838	0.0500	0.3831	0.5773
110	120	47	1	0	0.4735	0.0499	0.3732	0.5673
130	140	46	0	1	0.4735	0.0499	0.3732	0.5673
140	150	45	1	0	0.4630	0.0499	0.3631	0.5570
150	160	44	1	0	0.4525	0.0499	0.3530	0.5467
160	170	43	1	0	0.4420	0.0498	0.3429	0.5364
180	190	42	2	1	0.4207	0.0496	0.3227	0.5154
200	210	39	1	0	0.4099	0.0495	0.3125	0.5047
210	220	38	1	0	0.3991	0.0494	0.3024	0.4939
260	270	37	1	1	0.3882	0.0492	0.2921	0.4830
280	290	35	2	0	0.3660	0.0489	0.2714	0.4608
300	310	33	1	0	0.3549	0.0486	0.2612	0.4496
330	340	32	1	0	0.3438	0.0483	0.2510	0.4383
340	350	31	2	1	0.3213	0.0477	0.2305	0.4153
370	380	28	0	1	0.3213	0.0477	0.2305	0.4153
390	400	27	0	1	0.3213	0.0477	0.2305	0.4153
420	430	26	0	1	0.3213	0.0477	0.2305	0.4153
440	450	25	0	1	0.3213	0.0477	0.2305	0.4153
480	490	24	0	1	0.3213	0.0477	0.2305	0.4153
510	520	23	0	1	0.3213	0.0477	0.2305	0.4153
540	550	22	0	1	0.3213	0.0477	0.2305	0.4153
580	590	21	1	0	0.3060	0.0478	0.2156	0.4008

```
  590   600        20        0      1     0.3060    0.0478     0.2156    0.4008
  620   630        19        0      1     0.3060    0.0478     0.2156    0.4008
  670   680        18        1      1     0.2885    0.0482     0.1983    0.3847
  730   740        16        1      0     0.2705    0.0484     0.1808    0.3680
  840   850        15        0      1     0.2705    0.0484     0.1808    0.3680
  850   860        14        1      0     0.2511    0.0487     0.1622    0.3501
  910   920        13        0      1     0.2511    0.0487     0.1622    0.3501
  940   950        12        0      1     0.2511    0.0487     0.1622    0.3501
  970   980        11        1      0     0.2283    0.0493     0.1398    0.3299
  990  1000        10        1      0     0.2055    0.0494     0.1187    0.3088
 1030  1040         9        1      0     0.1826    0.0489     0.0988    0.2869
 1140  1150         8        0      1     0.1826    0.0489     0.0988    0.2869
 1320  1330         7        0      1     0.1826    0.0489     0.0988    0.2869
 1380  1390         6        1      0     0.1522    0.0493     0.0715    0.2609
 1400  1410         5        0      2     0.1522    0.0493     0.0715    0.2609
 1570  1580         3        0      1     0.1522    0.0493     0.0715    0.2609
 1580  1590         2        0      1     0.1522    0.0493     0.0715    0.2609
 1790  1800         1        0      1     0.1522    0.0493     0.0715    0.2609
-------------------------------------------------------------------------------

(Heart transplant data)

Durée pour differents quantiles de la fonction de survie
Définition des bornes Sas-lifetest
S(t)=0.90: t=    7.923
S(t)=0.75: t=   35.989
S(t)=0.50: t=  102.068
S(t)=0.25: t=  913.968
S(t)=0.10: t=        .
```

102 jours après leur inscription dans le registre d'attente pour une greffe, 50% des malades sont toujours en vie. Au bout de 914 jours, 75% des personnes sont décédées.

La méthode de Kaplan-Meier

- La méthode e qui exploite toute l'information disponible est celle dite de **Kaplan-Meier** (*KM*).
- Il y a autant d'intervalles que de moment 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 donc une fonction en escalier (stairstep), d'où une méthode dite « discrète ».
- Pour chaque intervalle, on compte le nombre d'évènements et le nombre de censures.
- Méthode adaptée pour une mesure de la durée de type continue.

Définition du Risk Set (r_i)

S'il y a à au même moment des évènements et des censures, 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, nommée appelée également l'estimateur « product-limit »:

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

Calcul de $S(t_i)$

On applique la méthode de la section précédente avec :

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

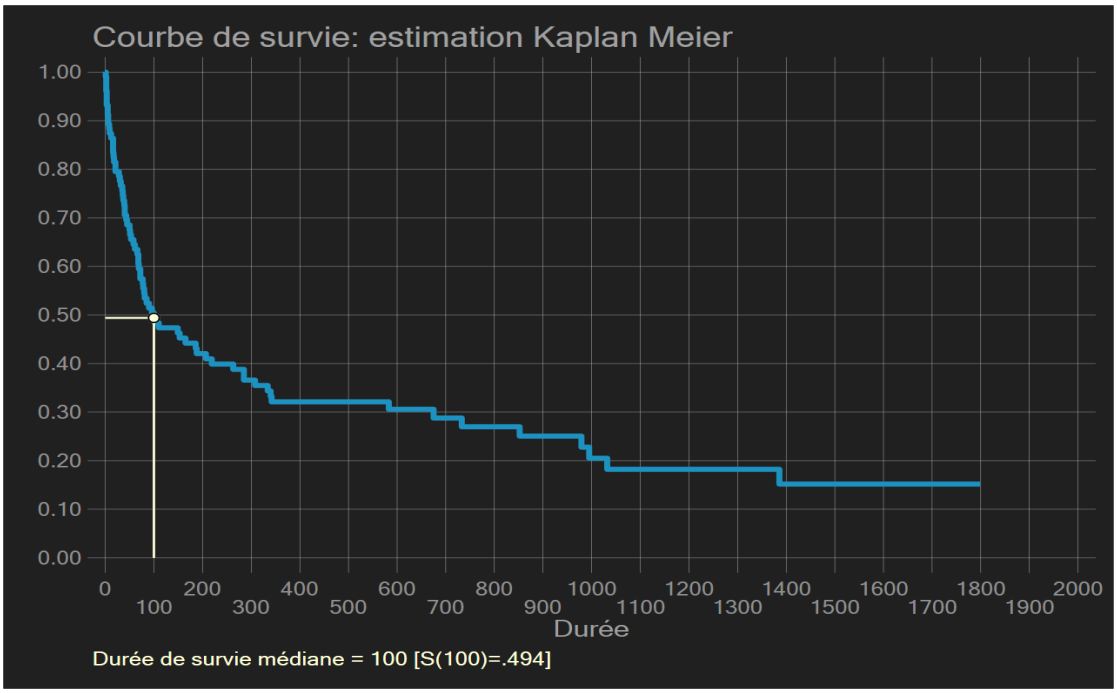
Récupération de la médiane

Il n'y a pas de méthode pour calculer directement la durée médiane (ou tout autre quantile de la durée).

On va seulement lire la valeur de la durée qui se situe juste « en dessous » de 50% de survivant.e.s. Elle est donc définie tel que $S(t) \leq 0.5$. Pas de formule savant, c'est une convention. Attention, il n'est pas impossible que le % de survivant.e.s soit bien en deçà de 50% pour définir cette durée.

Exemple

Durée médiane: $t = 100$ et correspond à $S(t) = 0.4940$.



Beg.	Net	Survivor	Std.

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
1032	9	1	0	0.1823	0.0489	0.0985	0.2865
1141	8	0	1	0.1823	0.0489	0.0985	0.2865
1321	7	0	1	0.1823	0.0489	0.0985	0.2865
1386	6	1	0	0.1519	0.0493	0.0713	0.2606
1400	5	0	1	0.1519	0.0493	0.0713	0.2606
1407	4	0	1	0.1519	0.0493	0.0713	0.2606
1571	3	0	1	0.1519	0.0493	0.0713	0.2606
1586	2	0	1	0.1519	0.0493	0.0713	0.2606
1799	1	0	1	0.1519	0.0493	0.0713	0.2606

Logiciels

SAS : l'estimation de Kaplan-Meier est affichée par défaut par la **`proc lifetest`**. Attention : le tableau proposé par SAS est particulièrement pénible à lire voire illisible, en particulier lorsque le nombre de censure est élevé, une ligne est ajoutée pour chaque observation censurée. Je conseille de ne pas afficher cette partie de l'output (voir chapitre SAS). 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.

Stata : en mode survie (`stset`), le tableau des estimateurs est donnée par la commande **`sts list`** et le graphique par **`sts graph`**.

R : les estimateur sont donnés par la fonction survfit de la librairie survival.

Python: les resultats sont donnés dans la librairie lifeline par des fonctions dont le nom est interminable. Je conseille plutôt l'utilisation de la librairie statmodels (se reporter à la session dédiée à Python.

Exercice

Calculer la fonction de survie S avec un tableur.

t	d	c	r	q	S
0	0	0			
6	1	0			
19	1	0			
32	1	0			

```
       42 |          2          0
       43 |          0          1
       94 |          1          0
      126 |          0          2
      207 |          1          0
      227 |          0          2
      253 |          1          0
      255 |          0          1
-------------------------------------------------------------------------
```

d= nombre d'évènements en t

c = nombre de censure en t

r = risk set en t

q= quotient mesurant l'intensité de l'évènement en t

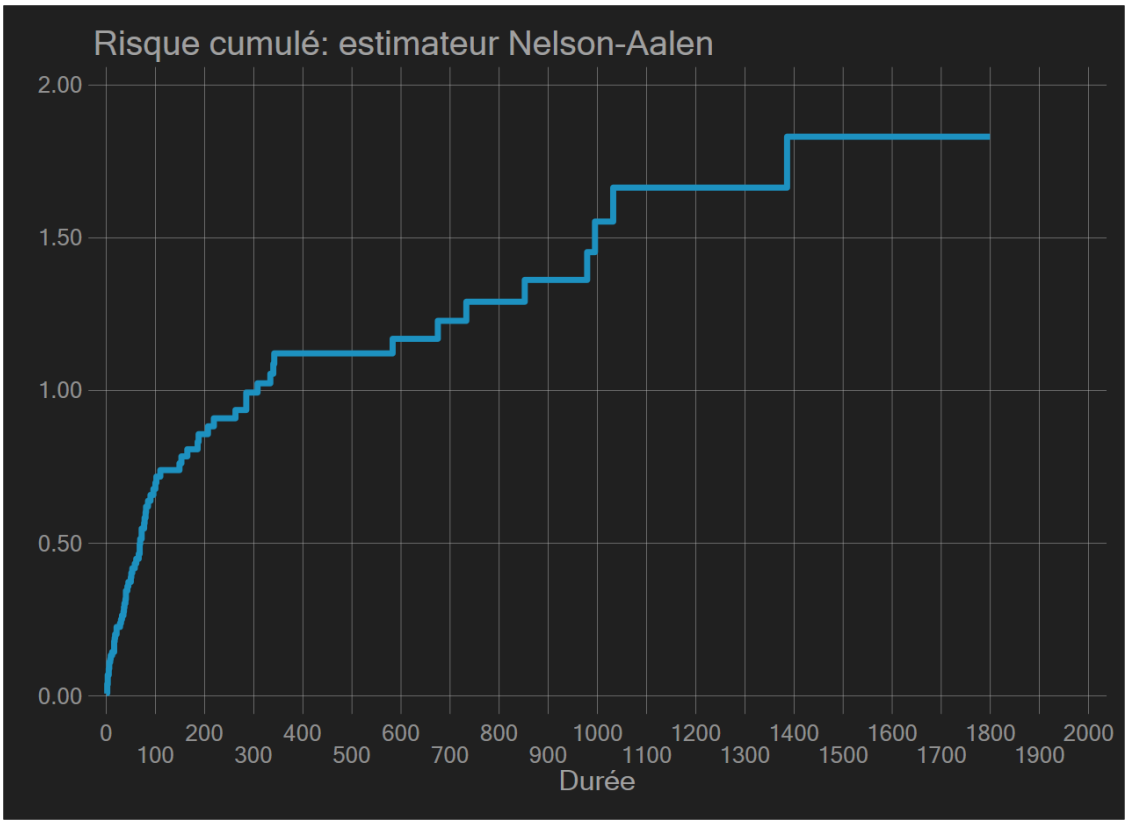
S= valeur de la survie en t

Quantités associées

Le risque cumulé

On utilise habituellement l'estimateur de Nelson-Aalen. Il est simplement égal à:

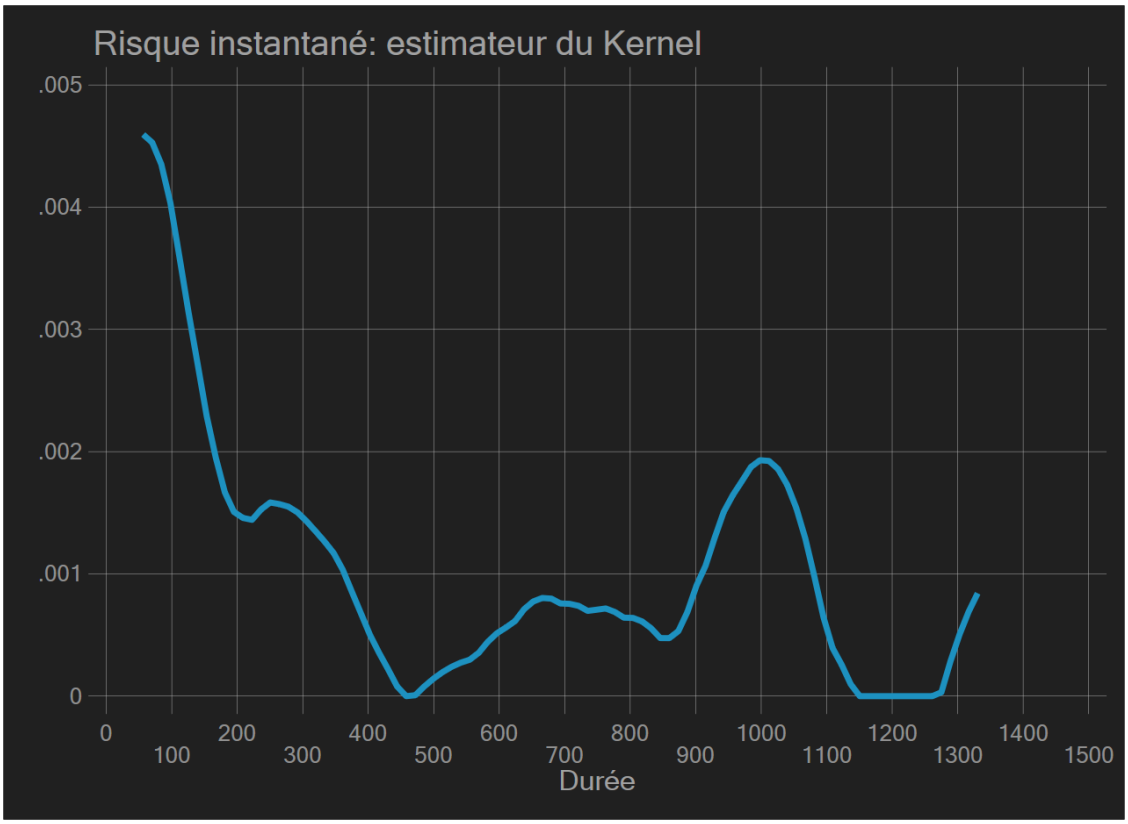
$$H(t) = \sum_{t_i \leq k} q\,(t_i)$$



On peut également utiliser l'estimateur de Breslow qui tire directement les Valeurs de $H(t)$ de la relation $S(t) = e^{-H(t)}$. Donc $H(t) = -\log S(t)$.

Le risque instantané

Nécessite l'estimateur de l'incidence cumulée ($H(t)$). L'estimation du risque instantané est obtenu en lissant les différences – par définition positive - entre $H(t)$ par la méthode du noyau (**kernel**). Elle permet d'obtenir une fonction continue avec la durée (paramétrables sur les largeurs des fenêtres de lissage). D'autres méthodes de lissage sont maintenant possibles, et de plus en plus utilisées, en particulier celles utilisant des splines restreintes (paramétrable sur un nombre de degré de liberté – nœuds-).



Tester l'égalité des courbes de survie (méthode KM)

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 d'un test correspond à la nécessité de déterminer si une même distribution gouverne les évènements observés dans les différentes strates ou les différents échantillons.

Attention: pas de test possibles sur des variables continues. Il faudra donc prévoir des regroupements pour les transformer en variable ordinale.

Deux méthodes sont utilisées:

- La plus ancienne et la plus diffusée: tests sur les rangs ou tests tests du **log-rank**.
- Plus récente et moins difusée: comparaison des **RMST** (*Restricted Mean of Survival Time*).

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. Ils entrent dans le cadre des tests d'ajustement dits du Chi2, même si formellement ils relèvent des techniques dites de rang, d'où leur nom.

Il s'agit donc de comparer des effectifs observés à des effectifs espérés à chaque temps d'évènement. La différence réside dans le calcul de la variance de la statistique du test qui, ici, suit une loi hypergéométrique (proche de la loi binomiale).

Principe de calcul

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 espérés (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 , soit $e_{i1} = R_{i1} \times \frac{d_i}{R_i}$ et $e_{i2} = R_{i2} \times \frac{d_2}{R_2}$. Leur somme pour tous les temps d'évènement est égale à E_1 et E_2.

Le principe de calcul des effectifs espérés repose donc sur l'hypothèse d'un rapport des risques toujours égal à 1 au cours de la période d'observation (hypothèse fondamentale de risque proportionnel).

Les écarts entre effectifs observés et espérés doivent également respecter cette hypothèse. Si les courbes de séjour ne sont pas homogènes, alors cette non homogénéité doit reposer sur des écarts constants au cours du temps. La validité de ce test repose aussi sur cette hypothèse.

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

Les principaux tests de type log-rank

Le principe de construction des effectifs observés et espérés reste le même dans chaque test, les différences résident dans les pondérations (w_i) qui prennent en compte, de manière différente, la taille de la population soumise au risque à chaque durée où au moins un évènement est observé. Outre le problème de proportionnalité, la validité du test du log-rank repose également sur une distribution des censures homogènes entre les différents groupes qui sont comparés, car la présence de censures affecte la valeur du risk set au cours du temps.

- **Test du log-rank (standard)**: $w_i = 1$

Il accorde le même poids à toutes les durées d'évènement. C'est le teststandard. Il est très sensible aux distribution des censures.

- **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 des observations censurées dans chaque groupe.

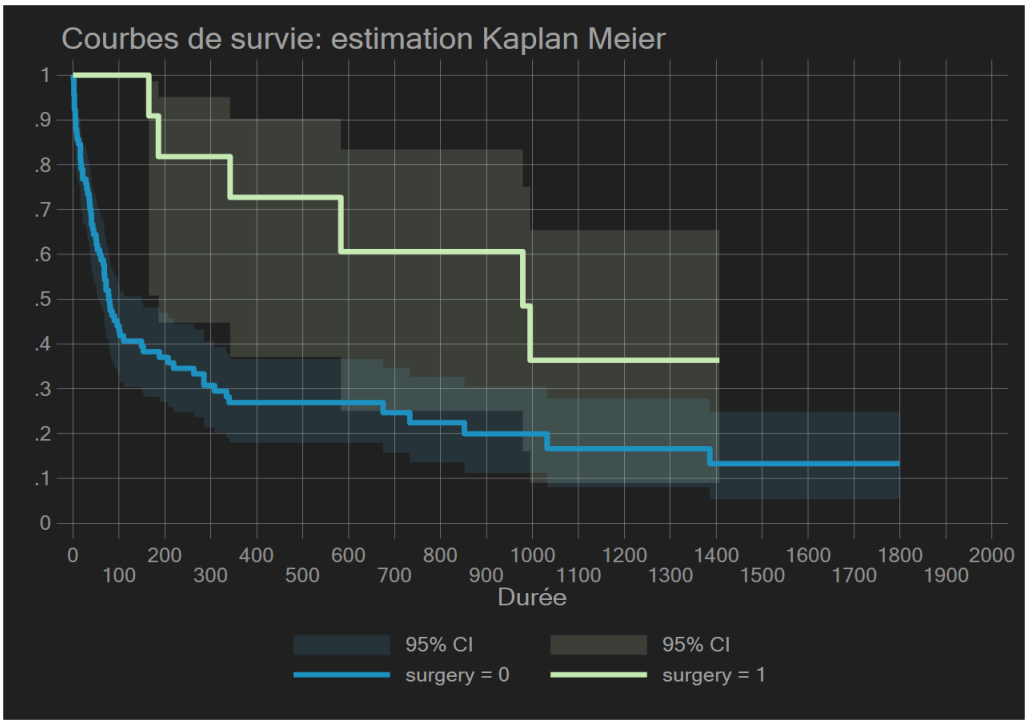
- **Test de Tarone-Ware**: $w_i = \sqrt{R_i}$
  Variante du test précédent, il atténue le poids accordé aux évènements au début de la période d'observation. Il est par ailleurs moins sensible au problème de la distribution des censures entre les groupes. A utiliser si le nombre de censures est faivle et qu'on ne privilégie pas trop le début de la durée d'observation.

- **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. Il accorde un poids important au début de la période d'observation. A utiliser lorsque le nombre de censures est élevé.

- **Test de Feming-Harington**: $w_i = (S_i)^p \times (1 - S_i)^q$ avec $0 \leq p \leq 1$ Il permet de paramétrer le poids accordé au début où à la fin de temps d'observation. Si $p = q = 0$ on retrouve le test du log-rank.

Exemple

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



Log-rank test for equality of survivor functions

surgery	Events observed	Events expected
0	69	60.34
1	6	14.66
Total	75	75.00

chi2(1) = 6.59
Pr>chi2 = 0.0103

Wilcoxon (Breslow) test for equality of survivor functions

surgery	Events observed	Events expected	Sum of ranks
0	69	60.34	623
1	6	14.66	-623
Total	75	75.00	0

chi2(1) = 8.99
Pr>chi2 = 0.0027

Tarone-Ware test for equality of survivor functions

surgery	Events observed	Events expected	Sum of ranks
0	69	60.34	73.111827
1	6	14.66	-73.111827
Total	75	75.00	0

```
              chi2(1) =       8.46
              Pr>chi2 =     0.0036


Peto-Peto test for equality of survivor functions

        |   Events         Events        Sum of
surgery |  observed       expected        ranks
--------+-------------------------------------
0       |        69          60.34    6.0529913
1       |         6          14.66   -6.0529913
--------+-------------------------------------
Total   |        75          75.00            0

              chi2(1) =       8.66
              Pr>chi2 =     0.0033
```

Les résultats font apparaître, que l'opération permet d'allonger la durée de survie des personnes.

Logiciels

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`.

Stata: on utilise la commande **`sts test`** avec le nom de la version du test si on ne souhaite pas récupère toutes les variantes.

R: on utilise la fonction **`survdiff`**. Le résultat du test Peto-Peto est affiché par défaut (`rho=1`). Si on souhaite utiliser le test non pondéré, on ajoute l'option `rho=0`. En particulier pour des tests multiples, on peut utiliser la fonction **`pairwise_survdiff`** de la librairie `survminer`.

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).

En pratique/remarques:

- Les tests du log-rank sont sensibles à l'hypothèse de risque proportionnel (voir modèle **Semi-paramétrique de Cox**). En pratique si des courbes de séjours se croisent, il est déconseillé de les utiliser. Cela ne signifie pas que si les courbes ne se croisent pas, l'hypothèse de proportionnalité des risques est respectée : des rapports de risque peuvent au cours du temps s'accentuer, se réduire ou le cas échant s'inverser (typique d'un croisement).

- Effectuer un test global (multiple/omnibus) sur un nombre important de groupes (ou >2) peut rendre le test très facilement significatif. Il peut être intéressant de tester des courbes deux à deux (idem qu'une régression avec

covariable discrète), en conservant un seul degré de liberté. Des méthodes de correction du test multiple sont possibles.

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

Logiciels

SAS : 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`**. Bien que sortie tardivement par rapport aux autres application standard, les résultats sont particulièrement complets.

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.

R : librairie **`SurvRm2`**. Programmée par les mêmes personnes que la commande Stata, la fonction est peu souple.

Python : estimation avec une fonction de la librairie **`lifelines`**. Pas de test de comparaison.

```
Restricted Mean Survival Time (RMST) by arm
-----------------------------------------------------------
   Group |  Estimate    Std. Err.      [95% Conf. Interval]
---------+-------------------------------------------------
   arm 1 |   734.758     133.478      473.145      996.370
   arm 0 |   310.169      43.158      225.581      394.757
-----------------------------------------------------------

Between-group contrast (arm 1 versus arm 0)
------------------------------------------------------------------------
           Contrast  |  Estimate       [95% Conf. Interval]     P>|z|
---------------------+--------------------------------------------------
RMST (arm 1 - arm 0) |   424.589      149.641      699.537      0.002
RMST (arm 1 / arm 0) |     2.369        1.513        3.710      0.000
------------------------------------------------------------------------
```

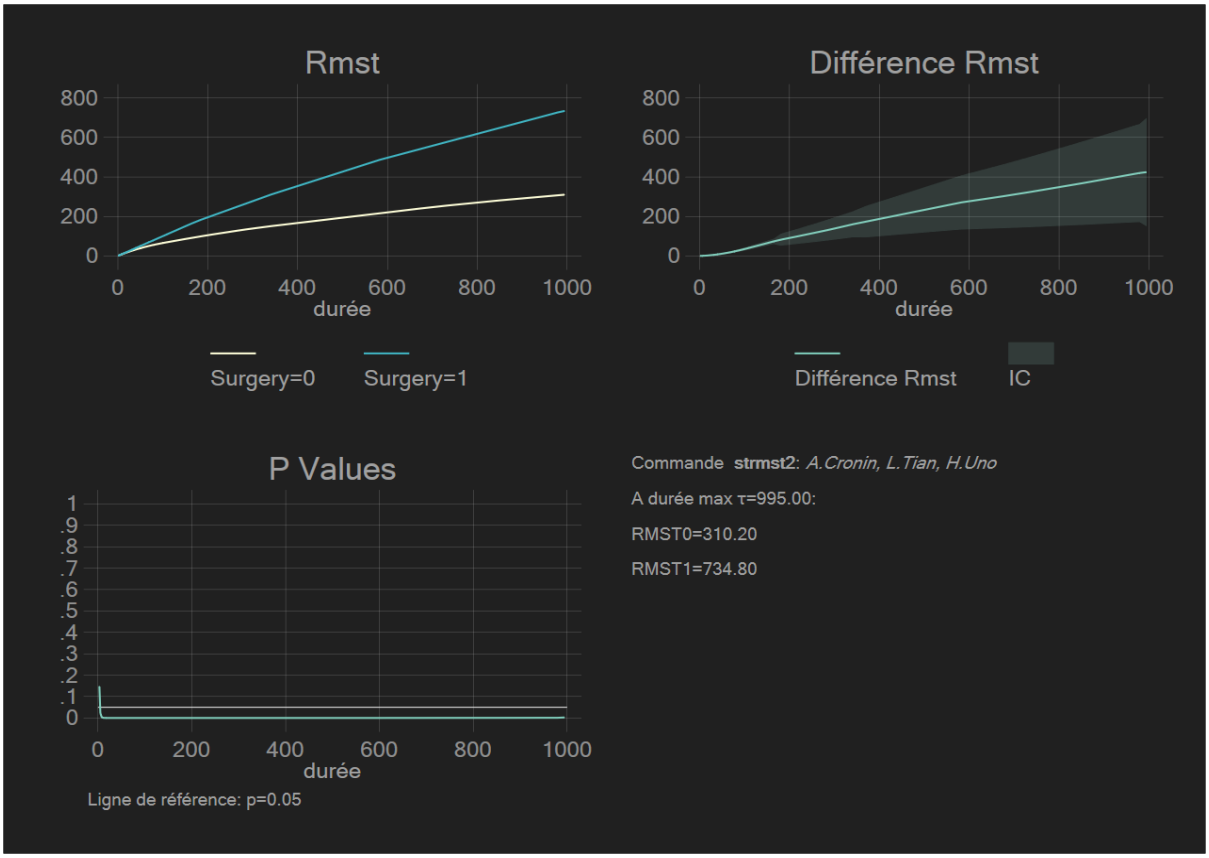
Ici t^* est égal à 995 jours, soit la durée qui correspond au dernier décès observé lorsqu'une personne a été opérée pour un pontage (surgery=1). Sur cet horizon, les personnes qui ont été opérées peuvent espérer vivre 735 jours en moyenne, contre 310 jours pour les autres. La durée moyenne de survie est donc deux fois plus importante (rapport des rmst = 2.3), soit une différence de 424 jours.

Rmst et différences de Rmst à tous les points d'évènement jusqu'à $tmax$

_time	_rmst1	_rmst0	_diff	_l	_u	_p
1	1	1	0	0	0	.
2	2	1.989011	.010989	.010989	.010989	.
3	3	2.945055	.0549451	-.0196757	.1295658	.1489731
5	5	4.791209	.2087912	.0256584	.3919241	.0254456
5.1	5.1	4.882418	.2175824	.0240373	.4111275	.0275679
6	6	5.693407	.3065934	.0643487	.5488381	.0131162
8	8	7.451648	.5483516	.1860869	.9106163	.0030096
9	9	8.31978	.6802198	.2523926	1.108047	.0018318
11	11	10.03407	.965934	.4072903	1.524578	.0007017
12	12	10.89121	1.108791	.4836155	1.733967	.0005087
16	16	14.27525	1.724747	.8259398	2.623554	.0001692
17	17	15.08787	1.912131	.9277063	2.896555	.0001407
18	18	15.88935	2.110646	1.05301	3.168283	.0000918
21	21	18.26041	2.739589	1.458002	4.021176	.0000279
28	28	23.63703	4.362966	2.526501	6.199431	3.22e-06
30	30	25.15095	4.849051	2.842812	6.85529	2.17e-06
31	31	25.89677	5.103226	3.014868	7.191583	1.67e-06
32	32	26.6426	5.3574	3.186886	7.527915	1.31e-06
35	35	28.84618	6.153824	3.736433	8.571216	6.06e-07
36	36	29.5694	6.4306	3.929635	8.931564	4.67e-07
37	37	30.28132	6.718675	4.135427	9.301923	3.44e-07
39	39	31.68257	7.317427	4.569757	10.0651	1.79e-07
40	40	32.37189	7.628103	4.797789	10.45842	1.28e-07
43	43	34.37207	8.627934	5.552385	11.70349	3.83e-08

45	45	35.68291	9.31709	6.07508	12.5591	1.77e-08
50	50	38.90352	11.09648	7.431942	14.76102	2.94e-09
51	51	39.53634	11.46366	7.711818	15.2155	2.12e-09
53	53	40.77938	12.22061	8.298259	16.14297	1.02e-09
58	58	43.83049	14.16951	9.81571	18.52331	1.79e-10
61	61	45.62725	15.37275	10.75502	19.99047	6.81e-11
66	66	48.56535	17.43465	12.37503	22.49426	1.44e-11
68	68	49.71799	18.28201	13.04371	23.5203	7.90e-12
69	69	50.27171	18.72829	13.40325	24.05333	5.45e-12
72	72	51.89897	20.10103	14.51838	25.68369	1.70e-12
77	77	54.49805	22.50194	16.49285	28.51104	2.14e-13
78	78	55.00657	22.99343	16.89797	29.08889	1.43e-13
80	80	56.00101	23.99899	17.73478	30.26321	5.97e-14
81	81	56.48692	24.51307	18.16526	30.86089	3.77e-14
85	85	58.38539	26.61461	19.93458	33.29464	5.77e-15
90	90	60.70197	29.29803	22.1984	36.39766	6.66e-16
96	96	63.41406	32.58594	24.97681	40.19506	0
100	100	65.17693	34.82308	26.87198	42.77418	0
102	102	66.03575	35.96425	27.84368	44.08482	0
109	109	68.96255	40.03745	31.32724	48.74766	0
110	110	69.38067	40.61933	31.82339	49.41528	0
131	131	77.91717	53.08283	42.46146	63.7042	0
149	149	85.23417	63.76583	51.49939	76.03227	0
153	153	86.81235	66.18765	53.54893	78.82638	0
165	165	91.4034	73.5966	59.87845	87.31474	0
180	178.6364	97.14223	81.49413	51.34782	111.6404	1.17e-07
186	184.0909	99.43776	84.65315	53.34505	115.9613	1.16e-07
188	185.7273	100.2029	85.52434	53.56977	117.4789	1.56e-07
207	201.2727	107.2376	94.0351	58.18815	129.8821	2.73e-07
219	211.0909	111.5325	99.55843	61.16676	137.9501	3.72e-07
263	247.0909	126.7373	120.3536	72.25138	168.4559	9.40e-07
265	248.7273	127.4037	121.3235	72.75741	169.8897	9.77e-07
285	265.0909	134.0682	131.0227	77.89536	184.1501	1.34e-06
308	283.9091	141.1427	142.7664	84.36629	201.1664	1.66e-06
334	305.1818	148.8068	156.375	91.96277	220.7872	1.95e-06
340	310.0909	150.4986	159.5923	93.78695	225.3977	2.00e-06
342	311.7273	151.0369	160.6904	94.42397	226.9568	2.01e-06
370	**332.0909**	**158.5728**	**173.5181**	**93.67896**	**253.3572**	**.0000205**
397	351.7273	165.8396	185.8876	98.91358	272.8617	.000028
427	373.5454	173.9138	199.6316	104.6545	294.6087	.0000379
445	386.6364	178.7584	207.878	108.0686	307.6874	.0000446
482	413.5454	188.7166	224.8289	115.0297	334.6281	.0000599
515	437.5454	197.5982	239.9472	121.1866	358.7078	.000075
545	459.3636	205.6725	253.6912	126.7507	380.6316	.0000897
583	487	215.8998	271.1002	133.7623	408.438	.0001093
596	494.8788	219.3987	275.4801	134.5264	416.4339	.0001279
620	509.4243	225.858	283.5662	136.4692	430.6632	.0001579
670	539.7273	239.3151	300.4122	140.0713	460.7531	.0002405
675	542.7576	240.6608	302.0968	140.4026	463.7909	.0002504
733	577.9091	254.9701	322.939	145.3689	500.509	.0003645
841	643.3636	279.1928	364.1708	155.9437	572.3979	.0006085
852	650.0303	281.6599	368.3704	156.9483	579.7925	.000638
915	688.2121	294.2198	393.9923	164.2457	623.7389	.0007762
941	703.9697	299.4033	404.5664	167.1596	641.9732	.0008378
979	727	306.9791	420.0209	171.3309	668.7109	.0009321

```
|   995    734.7576    310.1689    424.5887    149.6407    699.5366    .0024726 |
+-------------------------------------------------------------------------------+
```



Si on se limite a un horizon d'un an après l'inscription dans le registre (ici 370 jours), les personnes qui ont bénéficiées d'un pontage peuvent espérer survivre en moyenne 173 jours de plus que les personnes qui n'ont pas été opérées.

Remarques:

- Comme l'estimateur de la Rmst est calculé comme une aire sous la fonction de séjour KM, on peut calculer l'aire au dessus. L'estimateur obtenu est appelé **Restricted Mean of Time Loss** (Rmtl). Les logiciels proposent également cette estimation.

- Un modèle basé sur les Rmst a été proposé. Après quelques tests il me semble mal supporter la complexité souvent inhérente aux modèles dans les sciences sociales. Il supporte visiblement un nombre très limité de covariables (en médecine une variable d'intérêt de type traitement, et un nombre limité de contrôle type âge sexe).

Les modèles à risques proportionnels

Introduction aux modèles à risques proportionnels

La spécification usuelle est:

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

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

Le risque de base

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

Risques proportionnels

Cette hypothèse stipule l'invariance dans le temps des « rapports de risque" (**Hazard Ratios**).

Exemple:

Une seule covariable X est introduite, et soit 2 individus A et B: $h_A(t) = h_0(t)e^{bX_A}$ et $h_B(t) = h_0(t)e^{bX_B}$.

Les rapport des risques entre A et B est égal à:

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

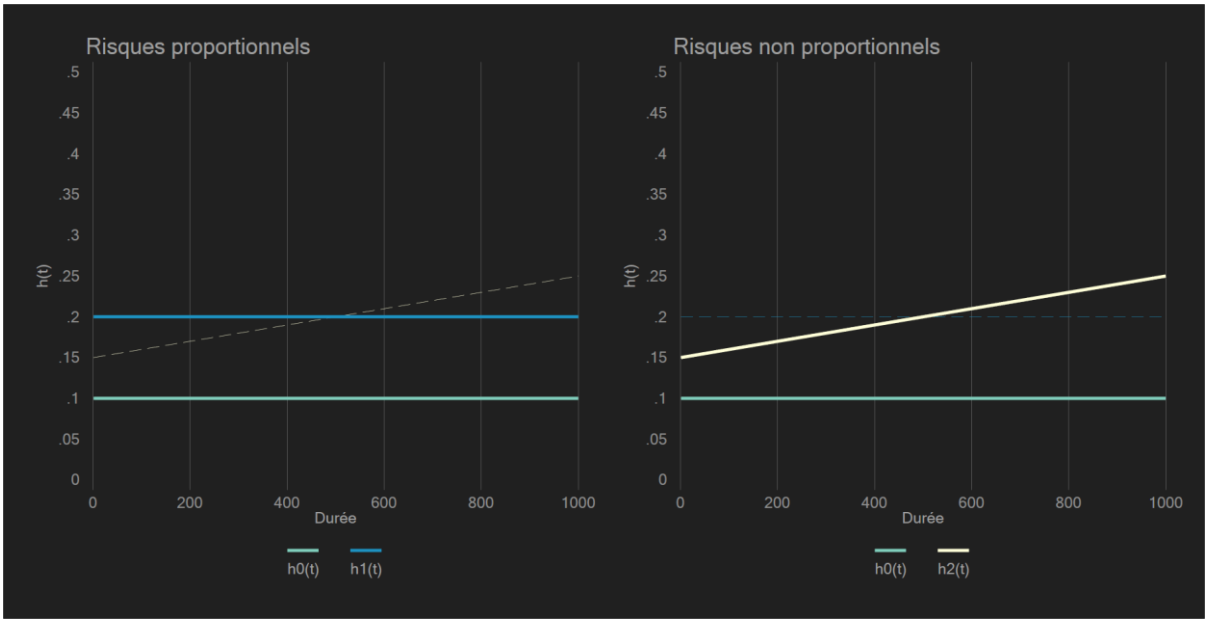
Pour une caractéristique binaire: $X_A = 1$ et $X_B = 0$:

$$\frac{h_A(t)}{h_B(t)} = e^{b}$$

Autrement dit, la proportionnalité des risques traduit l'absence d'une interaction significative entres les rapports de risque estimés et la durée.

Mais également, la proportionnalité des risques implique que toutes les observations partagent le même profil de risque durant la période d'observation.

Illustration graphique



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

Comme $h_1(t) = 0.2$ quel que soit t, le rapport de risque est toujours égal à $\frac{0.2}{0.1} = 2 = e^b$. Le coefficient estimé sera égal à $log(2) = 0.69$.

Pour $h_{1b}(t)$, le rapport de risque augmente avec le temps: $t = 1$, $h_{1b}(1) = 0.15$ et $h_{1b}(1000) = 0.25$ l'hypothèse de proportionnalité n'est donc pas respectée.

Néanmoins, estimé par un modèle à risque proportionnel comme celui de Cox, l'estimateur sera égal à .69 (rapport de risque =2).

Les modèles

Le modèle semi-paramétrique de Cox

Le modèle estime directement les b indépendement de $h_0(t)$, c'est pour cela qu'il est semi-paramétrique. Les rapports de risque (e^b) sont utilisés pour estimer la baseline $h_0(t)$ nécessaire si on souhaite calculer des fonctions de survie ajustée. Le respect de l'hypothèse de proportionalité va alors s'avérer importante et donc être testée.

Les modèles à temps discret

De type paramétrique. Peut être estimé à l'aide d'un modèle logistique, probit ou complémentaire log-log. La première est la plus courante, la dernière a l'avantage d'être directement relié au modèle de Cox (modèle de Cox à temps discret).

Cas particulier car sa forme diffère de la présentation usuelle d'un modèle à risque proportionnel. Toutefois, il suit également une hypothèse de proportionnalité. Le non respect de l'hypothèse est moins critique car la baseline du « risque » est estimée simultanément. Il est comme son nom l'indique, particulièrement adapté au durées discrètes ou groupées.

Avec une spécification logistique, la plus courante, les Odds vont sous certaines conditions, se confondre avec des probabilités/risques.

Les modèles paramétriques standard

Les modèles dits de **Weibull**, **exponentiel** ou **Gompertz** ont une spécification sous hypothèse de risque proportionnel. Ils seront traités brièvement dans les compléments.

Modèle paramétrique de Parmar-Royston (non traité)

$h_0(t)$, via le risque cumulé $H(t)$, est estimé simultanément avec les risques ratios en utilisant la très populaire méthode des splines cubiques. Il est implémenté dans les logiciels standards (R, Stata, Sas). Les rapports de risque sont très proches de ceux estimés par le modèle classique de Cox.

Il offre donc une alternative particulièrement intéressante à celui-ci, et il est maintenant largement diffusé dans l'analyse des effets cliniques.

Le modèle semi-paramétrique de Cox

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 les censures).

Pour une observation quelconque en t_i, la vraisemblance s'écrit:

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

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

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

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

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

La vraisemblance partielle retient le premier terme de la vraisemblance, qui intervient comme un paramètre de nuisance, soit:

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

Une fois remplacée la valeur de $h(t_i)$ par son expression en tant que modèle à risque proportionnel, la vraisemblance partielle ne dépendra plus de la durée. Mais elle va dépendre de l'ordre d'arrivée des évènements.

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

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

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

Condition nécessaire: pas d'évènement simultané

Ici le temps est strictement continu, il ne doit pas y avoir d'évènement simultané. L'estimation de la vraisemblance doit alors être corrigée.

Correction de la vraisemblance avec des évènements simultanés

- La méthode dite **exacte**: Comme en réalité il n'y a pas d'évènement simultané, on va intégrer à la vraisemblance toutes les permutations possibles des évènements observés simultanément: si en t_i on observe au « même moment » l'évènement pour A et B, une échelle temporelle plus précise nous permettrait de savoir si A a eu lieu avant B ou B avant A.
  Le nombre de permutations étant calculé avec une factorielle, si 3 évènements sont mesurés simultanément, il y a 6 permutations possibles : $3 \times 2 \times 1$
  Problème: le nombre de permutations pour chaque t_i peut devenir très vite particulièrement élevé. Par exemple pour 10 évènements simultanés, le nombre de permutations est égal à 3.628.800 ($!10 = 10 \times 9 \times 8 \times 7 \times ... \times 2 \times 1$). Le temps de calcul devient particulièrement long, et ce type de correction totalement inopérant.

- La méthode dite de **Breslow**: il s'agit d'une approximation de la méthode exacte permettant de ne pas avoir à intégrer chaque permutation. Cette approximation est utilisée par défaut par les logiciels Sas et Stata.

- La méthode dite d'**Efron**: elle corrige l'approximation de Breslow, et est jugée plus proche de la méthode exacte. Le temps de calcul avec les ordinateurs actuels est quasiment identique à celle de Breslow. C'est la méthode utilisée par défaut avec le logiciel R. Elle est disponible dans les autres applications.

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 validiter de l'hypothèse de constance des rapports de risque (hazard ratio) pour calculer les **résidus dits de Schoenfeld** (voir test de l'hypothèse de risque proportionnel).

- Conditions de second ordre: calcul des dérivées secondes qui permettent d'obtenir la matrice d'information de Fisher et la matrice des variances-covariances des paramètres.
- Comme il n'y a pas de solution numérique directe, on utilise un algorithme d'optimisation (ex: Newton-Raphson) à partir des équations de score et de la matrice d'information de Fisher.

Eléments de calcul

En logarithme, la vraisemblance partielle s'écrit:

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

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

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

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

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

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

Comme $\frac{e^{X_{ik}b}}{\sum_{j\in R} e^{X_{jk}b}}$ est une probabilité $\sum_{j\in R_i} X_{ik} \frac{e^{X_{ik}'b_k}}{\sum_{j\in R_i} e^{X_{jk}'b_k}}$ est l'espérance (la moyenne) $\bar{X_k}$ d'avoir la caractéristique X_k lorsqu'un évènement a été observé. Finalement:

$$\frac{\partial lp(b)}{\partial b_k} = \sum_{i=1}^{n} \delta_i \, (X_{ik} - \bar{X_k})$$

Cette expression permet de tester le respect ou non de l'hypothèse de risque proportionnel.

Lecture des résultats

Comme il s'agit d'un modèle à risques proportionnels, **leurs rapports sont constants pendant toute la période d'observation**.

Covariable binaire

$X = (0,1)$ et $\frac{h(t|X=1)}{h(t|X=0)} = e^b$.
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 continue (mais fixe dans le temps)

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).

$\frac{h(t|X=a+c)}{h(t|X=a)} = e^{c\times b}$.

Si $c = 1$ (résultat de l'estimation): A un âge donnée en début d'exposition, le risque de connaitre l'évènement est e^b fois inférieur/supérieur à celui d'une personne qui a un an de moins .

Si on regarde une différence de 5 ans en âge ($c = 5$), le risque est $e^{5\times b}$ inférieur/supérieur à celui d'une personne qui a 5 ans de moins.

On peut facilement exprimer les coefficients multiplicateurs (rapports de risque) en %.

Exemple pour les insuffisances cardiaques

Estimateurs: b

```
Cox regression -- Efron method for ties

No. of subjects =          103                 Number of obs    =         103
No. of failures =           75
Time at risk    =        31938
                                               LR chi2(3)       =       17.63
Log likelihood  =   -289.30639                 Prob > chi2      =      0.0005

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |    -0.1196     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
------------------------------------------------------------------------------
```

Rapports de risque: e^b

```
Cox regression -- Efron method for ties

No. of subjects =          103                 Number of obs    =         103
No. of failures =           75
Time at risk    =        31938
                                               LR chi2(3)       =       17.63
Log likelihood  =   -289.30639                 Prob > chi2      =      0.0005

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        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 résultats des tests non paramétriques, à savoir que le pontage réduit les risques journalier de décès pendant la période d'observation (augmente la durée de vie) de -62% ($((0.37 - 1) \times 100)$.

De la même manière, plus on entre à un âge élevé plus le risque journalier de survie baisse (le rapport de risque de décédé = 1.3 => +3% pour une différence d'âge d'un an). La variable year, qui traduit les améliorations en médecines n'est quant à elle que très faiblement significative.

En pratique / à savoir

Avec les modèles de durée, les valeurs estimées peuvent, en lecture, faire apparaître de forts gradients en terme de risque. Il faut garder en mémoire que ces rapports sont lus par rapport à une baseline, conditionnelle à l'échelle temporelle, dont les valeurs sont par nature souvent très faibles, en particulier avec une métrique plutôt fine. Avec un risque de base de 0.1%, le multiplier par 3 (rapport des risques) aboutit à une valeur de 0.3%. C'est une lecture des survies ou de l'incidence cumulée ajustée qui permet d'appréhender les résultats sur une périodicité plus longue.

Logiciels

SAS : le modèle est estimé avec la **`proc phreg`** .

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

R : le modèle est estimé avec la fonction **`coxph`** de la librairie `survival`.

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`.

Qualité de l'ajustement

On peut évaluer la qualité prédictive du modèle (goodness of fit) avec **l'indice de concordance d'Harell** ou des **courbes de ROC** évaluées à chaque temps d'évènement.

Un indice au moins égal à .8 peut laisser à penser que le modèle possède de bonnes qualités au niveau de la prédiction.

Cela n'est pas le cas dans notre exemple :

```
          Harrell's C = (E + T/2) / P =   0.6581
                              Somers' D =   0.3162
```

Logiciels

SAS: on utilise l'option concordance sur la ligne proc phreg. Pour afficher des courbes de ROC à certains points de la durée d'observation ainsi qu'un résumé à tous les temps d'évènement, on utilise l'option plots=roc toujours sur la même ligne.

Stata: on utilise la commande estat concordance après avoir exécuté le modèle

R,Python: l'indice de concordance est directement donné dans l'output du modèle (seulement la librairie lifelines pour Python).

L'hypothèse de constance des rapports de risque

- Les rapports de risque (RR) estimés par le modèle sont contraints à être constant pendant toute la période d'observation. C'est une hypothèse forte.
- Le respect de cette hypothèse doit être testé, en particulier pour un modèle de Cox où la baseline du risque est habituellement estimée à l'aide de ces rapports (méthode dite de Breslow, non traitée). En post estimations, les valeurs estimées du risque pourront présenter des valeurs aberrantes, en particulier négatives.
- Tester cette hypothèse revient à tester une interaction entre les rapports et la durée (ou plutôt une fonction de la durée).
- Plusieurs méthodes disponibles, celle sur les résidus de martingales, réservée aux covariables continues, et le « test » graphique ne seront pas traités.

Tests sur les résidus de Schoenfeld

- Les résidus « bruts » sont directement calculés à partir des équations de scores (voir section estimation).
- Ce résidu n'est calculé que pour les observations qui ont observées l'évènement.
- Il est calculé au moment où l'évènement s'est produit.
- La somme des résidus pour chaque covariable est égale à 0 (propriété de l'équation de score à l'équilibre).

- On utilise généralement les résidus de Schoenfeld « standardisés » - par leur variance - pour tenir compte du fait que le risk set diminue au cours du temps.
- Pour une observation dont l'évènement s'est produit en t_i, le résidu brut de Schoenfeld pour la covariable X_k, après estimation du modèle, est égal à:

$$rs_{ik} = X_{ik} - \sum_{j \in R_i} X_{ik} \frac{e^{X'_{ik}b}}{\sum_{j \in R_i} e^{X'_{jk}b}} = X_{ik} - \bar{X_k}$$

- Ce résidu est formellement la contribution d'un individu au score. Il se lit comme la différence entre la valeur observée d'une covariable et sa valeur espérée au moment où un évènement se produit.

- Si l'hypothèse de constance des risques ratio est respectée, les résidus ne doivent pas suivre une tendance précise.

- Intuitivement sans censure à droite et en ne considérant que les résidus bruts: on a un RR strictement égal à 1, en début d'exposition $R_i = 100$ avec 50 hommes ($X_k = 0$) et 50 femmes ($X_k = 1$). Si l'hypothèse PH est 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 X_k est toujours égal à 0.5 et les résidus brut prendront toujours la valeur -.5 si $X = 0$ et .5 si $X = 1$. En faisant une simple régression linéaire entre les résidus qui alternent les deux valeurs comme le RR=1 et t , le coefficient estimé sera non significativement différent de 0.

- On peut donc tester l'hypothèse sur les résidus par une régression entre ces résidus pour chaque covariable et la durée (ou une fonction dérivée de la durée). La solution la plus utilisée est le test dit de **Grambsch-Therneau** implémenté dans tous les logiciels. On peut montrer que le test de Grambsch-Therneau consiste à introduire une interaction entre les covariables et une fonction de la durée dans le modèle.

```
Test of proportional-hazards assumption

Time:  Time
------------------------------------------------------------------
             |       rho           chi2       df       Prob>chi2
-------------+----------------------------------------------------
year         |      0.10162         0.80        1          0.3720
age          |      0.12937         1.61        1          0.2043
surgery      |      0.29664         5.54        1          0.0186
-------------+----------------------------------------------------
global test  |                      8.76        3          0.0327
------------------------------------------------------------------
```

Ici l'hypothèse de proportionnalité des risques peut être rejetée au seuil de 5% pour la variable *surgery*. Le risque ratio ne serait donc pas constant dans le temps.

Remarques / à savoir

- Test multiple: de nouveau, il convient de se méfier du résultat du test multiple lorsque le nombre de degrés de liberté est élevé. Le risque de premier espèce peut être inférieur à 5% alors que les tests pour chaque covariables présentent des p-value élevé. Il es considéré par certain.e.s comme un indicateur de l'ampleur du biais qui affecte la baseline du risque.
- Les transformations de la durée : n'importe quelle fonction de la durée peutêtre utilisée pour effectuer le test. On retient généralement les fonctions suivantes: $f(t) = t$ (« identity »), $f(t) = \log(t)$, $f(t) = \mathrm{KM}(t)$ ou $f(t) = 1 - \mathrm{KM}(t)$ où $KM(t)$ est l'estimateur de Kaplan-Meieir, enfin une transformation appelée par les logiciels « rank » utile seulement pour les durées strictement continue avec des valeurs décimales, par exemple $t = (.1, .5, 1, 2.3)$ donne une transformation $t = (1,2,3,4)$. A savoir : la fonction « identity » rend le test sensible sensibles aux évènements très tardifs et rares (outliers).

Logiciels

SAS: le test est disponible depuis quelques années avec l'argument **zph** sur la ligne `proc lifetest`. Par défaut SAS utilise $f(t) = t$.

Stata: le test est donné par la commande **estat phtest, d**. Par défaut SAS utilise $f(t) = t$

R : après avoir créer l'objet lié à l'estimation du modèle de cox, on utilise la fonction `cox.zph`. La fonction utilise par défaut $f(t) = 1 - KM(t)$ où $KM(t)$ sont les estimateurs de la courbe de Kaplan-Meier.

Python : après avoir créer l'objet lié à l'estimation du modèle de Cox, on utilise la fonction `proportional_hazard_test`. La fonction utilise par défaut $f(t) = t$, mais on peut afficher les résultats pour toutes les transformations de t disponibles avec l'option `time_transform='all'`.

Test avec introduction d'une interaction avec la durée

Petit *retour sur l'estimation du modèle*

Pour estimer le modèle de Cox, les données sont dans un premier temps splitées au temps d'évènement. A l'exception de Sas, les fonctions des logiciels prennent en charge cette opération.

```
     +-----------------------------+
     | id   surgery   _d   _t   _t0 |
     |-----------------------------|
24.  |  2         0    0    1     0 |
25.  |  2         0    0    2     1 |
26.  |  2         0    0    3     2 |
27.  |  2         0    0    5     3 |
28.  |  2         0    1    6     5 |
     |-----------------------------|
29.  |  3         0    0    1     0 |
30.  |  3         0    0    2     1 |
```

```
 31. |  3         0    0    3     2 |
 32. |  3         0    0    5     3 |
 33. |  3         0    0    6     5 |
     |------------------------------|
 34. |  3         0    0    8     6 |
 35. |  3         0    0    9     8 |
 36. |  3         0    0   12     9 |
 37. |  3         0    1   16    12 |
     +------------------------------+
```

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

On vérifie que les paramètres estimés sont identiques

```
Cox regression -- Breslow method for ties

No. of subjects =          103                 Number of obs    =       3,573
No. of failures =           75
Time at risk    =        31938
                                                LR chi2(3)       =       17.56
Log likelihood  =   -289.54474                  Prob > chi2      =      0.0005

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -.1195075   .0673691    -1.77   0.076    -.2515486    .0125336
         age |   .0295539   .0135341     2.18   0.029     .0030275    .0560803
   1.surgery |   -.984869   .4362881    -2.26   0.024    -1.839978   -.1297601
------------------------------------------------------------------------------
```

Introduction de l'interaction avec une fonction de la durée

On a une variable de durée (on prendra t avec $f(t) = t$) qui sera croisée avec la variable surgery. Le modèle à risque proportionnel va s'écrire:

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

Remarque : il est plutôt d'usage d'utiliser $f(t) = \log(t)$, qui permet d'homogénéiser l'échelle entre la baseline et le terme d'interaction.

$log(h(t|X,t)) = logh_0(t) + b_4(surgery \times log(t)) + \ b_1 age + b_2 year + b_3 surgery$

En revanche, l'avantage de $f(t) = t$ réside dans la lecture du résultat où le terme d'interaction est directement interprétable comme un rapport de rapports de risques, qui exprime la variation constante du rapport au cours du temps.

Estimation du modèle

On présentera le modèle avec le log des paramètres estimées (le terme d'interaction n'étant pas un rapport de risque mais un rapport de rapport de risque).

Important: le modèle estimé n'est plus un modèle à risques proportionnels.

```
Cox regression -- Breslow method for ties

No. of subjects =          103                 Number of obs    =         103
No. of failures =           75
Time at risk    =        31938
                                               LR chi2(4)       =       21.50
Log likelihood  =   -287.57352                 Prob > chi2      =      0.0003

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
main         |
        year |  -.1229512   .0668619    -1.84   0.066    -.2539981    .0080958
         age |   .0288597   .0134588     2.14   0.032      .002481    .0552384
   1.surgery |  -1.751567   .6744632    -2.60   0.009    -3.073491   -.4296435
-------------+----------------------------------------------------------------
tvc [interaction]          |
     surgery |   .0022277   .0011025     2.02   0.043     .0000669    .0043886
------------------------------------------------------------------------------
Note: Variables in tvc equation interacted with _t.
```

L'interaction $surgery \times t$ est ici significative ($p < 0.05$). On retrouve le résultat du test sur les résidus de Schoenfeld.

Interprétation:

- Le paramètre (logRR) pour la variable *surgery* donne le risque ratio au début de l'exposition au risque ($t = 0 + \epsilon$): le risque de décéder en début d'observation est $(e^{-1.27} - 1) \times 100 = -82\%$ plus faible pour les personnes qui ont eu un pontage avant leur inscription dans le registre.
- Le terme d'interaction étant positif, le gain en survie pour les personnes qui ont eu un pontage va diminuer avec le temps. Le RR augmente donc avec le temps, ici de +.2% par jour.

t	Calcul	Risk ratio
$0 + \epsilon$	$e^{-1.27+0.002\times0}$	0.28
1	$e^{-1.27+0.002\times1}$	0.281
2	$e^{-1.27+0.002\times2}$	0.282
.	.	
.	.	
10	$e^{-1.27+0.002\times10}$	0.286
100	$e^{-1.27+0.002\times100}$	0.34
365	$e^{-1.27+0.002\times365}$	0.58
730	$e^{-1.27+0.002\times365}$	1.20

Que faire si l'hypothèse n'est pas respectée?

Ne rien faire

On interprète le risque ratio comme un ratio moyen pendant la durée d'observation (P.Allison). Difficilement soutenable pour l'analyse des effets cliniques, elle peut être envisagée dans d'autres domaines. Attention au nombre de variables qui ne respecte pas l'hypothèse, l'estimation de la baseline du risque pourrait être sensiblement affectée. Il convient tout de même lors de l'interprétation, de préciser les variables qui seront analysées sous cette forme « moyenne » sur la période d'observation.

On peut également adapter cette stratégie du « ne rien faire » selon sens de l'altération des rapports de risque. Si aux cours du temps les écarts déjà significatifs en début d'observation s'accentuent, à la hausse comme à la baisse, on peut conserver cet estimateur moyen. Mais si l'effet est modéré : RR>1 qui baisse ou RR<1 qui augmente au cours du temps, je suis bien moins convaincu de la pertinence de ce rien faire.

Egalement il faut tenir compte de l'intérêt portée par les variables qui présentent un problème par rapport à l'hypothèse. Il n'est peut-être pas nécessaire de complexifier le modèle pour des variables introduites comme comme contrôle.

Mais plus problématique… On sait qu'une des causes du non respect de l'hypothèse peut provenir d'effets de sélection liées à des variables non observables omises. En analyse de durée ce problème prend le nom de **frailty** (fragilité). Des estimations, plutôt complexes, sont possibles dans ce cas, et sont en mesure malgré leur interprétation plutôt difficile de régler le problème. Si l'hypothèse est sensible aux problèmes d'omission, il convient donc de bien spécifier le modèle au niveau des variables de contrôle.

Cox stratifié

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

Interaction

Introduire une interaction avec la durée comme ce qui a été fait plus haut. Cela peut permettre d'enrichir le modèle au niveau de l'interprétation. Valable si peu de covariables présentent des problèmes de stabilité des RR. Attention tout de même à la forme de la fonction, dans l'exemple on a contraint l'effet d'interaction à être linéaire (strictement proportionnelle), ce qui est une hypothèse plutôt forte.

Modèles alternatifs

Utiliser un modèle alternatif: modèles paramétriques de type risque proportionnel si la distribution du risque s'ajuste bien, le modèle paramétrique « flexible » de **Parmar-Royston** (non traité) ou un **modèle à temps discret**.

Utiliser un modèle non paramétrique additif dit d'*Aalen* ou une de ses variantes (non traité). Mais ces modèles, dont les résultats sont des visuels graphiques, se commentent difficilement.

Remarque finale sur l'estimation du modèle de Cox

Le modèle a été estimé par la méthode de la vraisemblance partielle. On peut montrer que le modèle de Cox est estimable à partir d'un modèle de Poisson. Cette estimation est appelée « Constant Piecewise Exponential PH model ».

```
Poisson regression                              Number of obs     =      3,573
                                                LR chi2(90)       =     122.42
                                                Prob > chi2       =     0.0131
Log likelihood = -344.95318                     Pseudo R2         =     0.1507

------------------------------------------------------------------------------
          _d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -.1195204    .067374    -1.77   0.076     -.251571    .0125302
         age |   .0295531    .013535     2.18   0.029      .003025    .0560811
     surgery |    -.98486   .4363162    -2.26   0.024    -1.840024   -.1296961
             |
       stime |
          2  |   .4223592   1.154708     0.37   0.715    -1.840826    2.685544
          3  |   .0495983   1.154704     0.04   0.966     -2.21358    2.312776
          5  |   -.828855   1.224781    -0.68   0.499    -3.229383    1.571673
          6  |  -.9922643   1.224779    -0.81   0.418    -3.392786    1.408258
          8  |  -1.940546   1.414215    -1.37   0.170    -4.712356    .8312648
          9  |   -2.03868   1.414233    -1.44   0.149    -4.810526    .7331649
```

11	-12.59255	2179.957	-0.01	0.995	-4285.23	4260.045
12	-2.30438	1.414226	-1.63	0.103	-5.076213	.4674526
16	-1.481512	1.154713	-1.28	0.199	-3.744708	.7816839
17	-2.591968	1.41426	-1.83	0.067	-5.363868	.1799312
18	-2.642535	1.414254	-1.87	0.062	-5.414421	.1293516
21	-2.095489	1.224819	-1.71	0.087	-4.496091	.3051133
28	-3.055745	1.414246	-2.16	0.031	-5.827616	-.2838729
30	-3.103495	1.41426	-2.19	0.028	-5.875395	-.3315961
31	-13.89462	2179.957	-0.01	0.995	-4286.532	4258.743
32	-3.144133	1.414253	-2.22	0.026	-5.916018	-.3722474
35	-3.219157	1.414258	-2.28	0.023	-5.991053	-.447262
36	-3.227893	1.414267	-2.28	0.022	-5.999806	-.4559796
37	-3.246139	1.414263	-2.30	0.022	-6.018042	-.4742351
39	-3.26923	1.414303	-2.31	0.021	-6.041213	-.4972472
40	-2.581465	1.224839	-2.11	0.035	-4.982106	-.1808239
43	-3.311454	1.414341	-2.34	0.019	-6.083512	-.5393966
45	-3.327473	1.414407	-2.35	0.019	-6.099661	-.5552857
50	-3.421206	1.414399	-2.42	0.016	-6.193377	-.6490357
51	-3.425081	1.414444	-2.42	0.015	-6.197341	-.6528208
53	-3.44536	1.414475	-2.44	0.015	-6.21768	-.673039
58	-3.514664	1.414489	-2.48	0.013	-6.287011	-.7423175
61	-3.543748	1.414557	-2.51	0.012	-6.316229	-.7712681
66	-3.602062	1.414546	-2.55	0.011	-6.374522	-.8296028
68	-2.90779	1.225106	-2.37	0.018	-5.308954	-.5066257
69	-3.580111	1.414549	-2.53	0.011	-6.352577	-.8076461
72	-2.914206	1.22505	-2.38	0.017	-5.31526	-.5131524
77	-3.624716	1.414601	-2.56	0.010	-6.397284	-.8521489
78	-3.593983	1.414779	-2.54	0.011	-6.366898	-.8210686
80	-3.597845	1.414765	-2.54	0.011	-6.370734	-.8249558
81	-3.589639	1.414745	-2.54	0.011	-6.362489	-.8167895
85	-3.598044	1.414948	-2.54	0.011	-6.371291	-.824798
90	-3.62163	1.415099	-2.56	0.010	-6.395173	-.8480873
96	-3.658159	1.415134	-2.59	0.010	-6.43177	-.884548
100	-3.672641	1.415162	-2.60	0.009	-6.446308	-.8989739
102	-3.662767	1.415232	-2.59	0.010	-6.436571	-.8889631
109	-14.65089	2179.957	-0.01	0.995	-4287.288	4257.986
110	-3.704316	1.415125	-2.62	0.009	-6.47791	-.9307209
131	-14.68697	2179.957	-0.01	0.995	-4287.324	4257.95
149	-3.976368	1.415012	-2.81	0.005	-6.749742	-1.202995
153	-3.97938	1.414995	-2.81	0.005	-6.752718	-1.206041
165	-4.013449	1.415156	-2.84	0.005	-6.787104	-1.239793
180	-15.09339	2179.957	-0.01	0.994	-4287.731	4257.544
186	-4.112806	1.414994	-2.91	0.004	-6.886144	-1.339468
188	-4.11316	1.414915	-2.91	0.004	-6.886342	-1.339978
207	-4.178467	1.414929	-2.95	0.003	-6.951678	-1.405256
219	-4.206547	1.41493	-2.97	0.003	-6.97976	-1.433334
263	-4.342286	1.415099	-3.07	0.002	-7.11583	-1.568742
265	-16.10078	2179.957	-0.01	0.994	-4288.738	4256.536
285	-3.688074	1.225534	-3.01	0.003	-6.090076	-1.286072
308	-4.409256	1.414925	-3.12	0.002	-7.182459	-1.636054
334	-4.432237	1.415143	-3.13	0.002	-7.205866	-1.658607
340	-4.422918	1.415095	-3.13	0.002	-7.196453	-1.649383
342	-4.365805	1.41511	-3.09	0.002	-7.13937	-1.592239
370	-16.64142	2179.957	-0.01	0.994	-4289.279	4255.996
397	-16.53454	2179.957	-0.01	0.994	-4289.172	4256.103
427	-16.28492	2179.957	-0.01	0.994	-4288.922	4256.352
445	-16.76689	2179.957	-0.01	0.994	-4289.404	4255.87
482	-15.8041	2179.957	-0.01	0.994	-4288.441	4256.833
515	-16.91429	2179.957	-0.01	0.994	-4289.552	4255.723
545	-16.10426	2179.957	-0.01	0.994	-4288.742	4256.533
583	-4.641434	1.415524	-3.28	0.001	-7.41581	-1.867057
596	-16.41019	2179.957	-0.01	0.994	-4289.047	4256.227
620	-17.07029	2179.957	-0.01	0.994	-4289.708	4255.567
670	-17.14785	2179.957	-0.01	0.994	-4289.785	4255.489
675	-4.631958	1.416377	-3.27	0.001	-7.408006	-1.855911
733	-4.60698	1.416405	-3.25	0.001	-7.383082	-1.830878
841	-17.05138	2179.957	-0.01	0.994	-4289.689	4255.586
852	-4.566243	1.417712	-3.22	0.001	-7.344907	-1.787579
915	-17.40169	2179.957	-0.01	0.994	-4290.039	4255.236
941	-17.34105	2179.957	-0.01	0.994	-4289.978	4255.296
979	-4.426862	1.419414	-3.12	0.002	-7.208862	-1.644862
995	-4.401387	1.418922	-3.10	0.002	-7.182422	-1.620352
1032	-4.390393	1.418666	-3.09	0.002	-7.170928	-1.609859
1141	-16.4898	2179.957	-0.01	0.994	-4289.127	4256.148

```
       1321  |  -17.02179   2179.957    -0.01   0.994    -4289.659    4255.616
       1386  |  -4.427898    1.41857    -3.12   0.002    -7.208243   -1.647553
       1400  |  -17.74095   2179.957    -0.01   0.994    -4290.378    4254.896
       1407  |  -17.17352   2179.957    -0.01   0.994    -4289.811    4255.464
       1571  |  -18.15173   2179.957    -0.01   0.993    -4290.789    4254.486
       1586  |  -18.39765   2179.957    -0.01   0.993    -4291.035     4254.24
       1799  |  -18.08037   2179.957    -0.01   0.993    -4290.718    4254.557
             |
       _cons |   2.482922   4.946271     0.50   0.616    -7.211591    12.17744
  ln(stime) |          1  (exposure)
------------------------------------------------------------------------------
```

Modèles à temps discret

On va principalement traiter le modèle **logistique à temps discret**.

- Par définition ce n'est pas un modèle à risque proportionnel, mais à Odds proportionnels. Toutefois en situation de rareté (p<10%), l'Odds converge vers une probabilité, qui est une mesure du risque (ici une probabilité conditionnelle).
- Le modèle à temps discret est de type 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.
- Formellement, le modèle est estimable avec des évènements mesurés à une durée nulle (même si cela n'a pas grand sens).
- La base de données doit être transformée en format long: aux temps d'observation ou sur des intervalles de temps. C'est une des principales différences avec le modèle de Cox qui est une estimation aux temps d'évènement.
- Permet d'introduire de manière plutôt souple un ensemble de covariables dynamiques.

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

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

Organisation des données

Format long

Les données doivent être en format long: pour chaque individu on a une ligne par durée observée ou par intervalle de durées jusqu'à l'évènement ou la censure. On retrouve le système de *splittage* des données du modèle de Cox. Avec des données de type discrète, qui se traduisent par des nombres élevés d'évènement simultanés, classique en science sociale, il y a souvent peu de différence entre un allongement aux temps d'évènement et aux temps d'observation.

A savoir: Si on construit des intervalles, on doit s'assurer qu'au moins un évènement s'est produit dans chaque intervalle, sinon on rencontrera un problème de séparabilité parfaite.

Durée

La durée est dans un premier temps construite sous forme d'un simple compteur. La paramétrisation de la durée dans le modèle sera présentée par la suite.

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.

Exemple avec les malformations cardiaques

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

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

Format long et variables pour l'analyse

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

Estimation et ajustement de la durée

L'enjeu principal réside dans la paramétrisation de la durée:

- Elle peut être modélisée sous forme de fonction d'une variable de type continue.
- Elle peut être modélisée comme variable discrète, de type indicatrice (0,1), sur tous les points d'observation, ou sous forme de regroupements (rappel: au moins un évènement observé dans chaque intervalle).

Ajustement avec une durée en continu

Le modèle étant paramétrique, on doit trouver une fonction qui ajuste le mieux les données. Toute transformation de la variable de durée est possible: $f(t) = t$, $f(t) = log(t)$……formes quadratiques. Les ajustements sous forme de **splines** tendent à se développer ces dernières années.
Pour sélectionner cette fonction, on peut tester différents modèles sans covariable

additionnelle, et sélectionner la forme qui minimise un critère d'information de type **AIC** ou **BIC** (vraisemblance pénalisée).

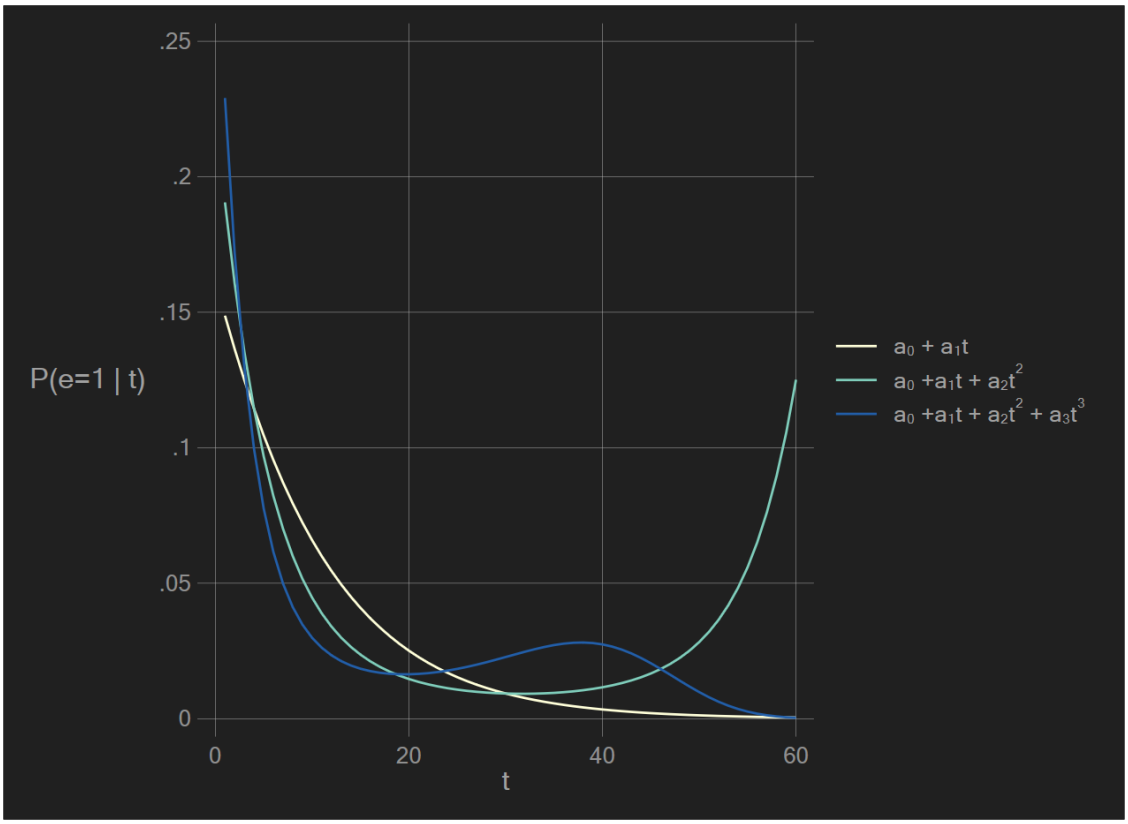
Exemple avec les malformations cardiaques

On va tester les paramétrisations suivantes :

- Une forme linéraire stricte $f(t) = a \times t$
- Des effets quadratiques d'ordres 2 et 3: $f(t) = a_1 \times t + a_2 \times t^2$ et $f(t) = a_1 \times t + a_2 \times t^2 + a_3 \times t^3$.

Remarques / à savoir:

- Les effets quadratiques consiste à introduires des intéractions entre une variable continue et elle même. Cela permet d'estimer des effets non linéaires.
- Attention aux effets quadratiques d'ordre supérieur à 2, ils sont très sensible aux *outliers*. En analyse des durées en fin de période d'observation si peu de personnes restent soumises au risque, ce dernier peut-être fortement surestimé.



Lecture : t = mois

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 :

```
Logistic regression                                     Number of obs     =      1,127
                                                        LR chi2(6)        =      90.69
                                                        Prob > chi2       =     0.0000
Log likelihood = -230.33671                             Pseudo R2         =     0.1645

------------------------------------------------------------------------------
           e |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |  -.3720566   .0823946    -4.52   0.000    -.5335471   -.2105661
          t2 |   .0142379    .005023     2.83   0.005     .0043929    .0240828
          t3 |  -.0001659   .0000785    -2.11   0.035    -.0003198    -.000012
        year |  -.1326693   .0737755    -1.80   0.072    -.2772666     .011928
         age |   .0333413   .0146876     2.27   0.023     .0045541    .0621285
     surgery |  -1.010918    .448598    -2.25   0.024    -1.890154   -.1316821
       _cons |   7.082657   5.307737     1.33   0.182    -3.320316    17.48563
------------------------------------------------------------------------------
```

Maintenant si on estime le modèle avec la méthode de Cox (avec des durées mesurées sur une échelle de 30 jours) :

```
Cox regression -- Efron method for ties

No. of subjects =          103                  Number of obs    =         103
No. of failures =           75
Time at risk    =         1127
                                                LR chi2(3)       =       17.97
Log likelihood  =   -289.81242                  Prob > chi2      =      0.0004

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -.1304397   .0674344    -1.93   0.053    -.2626087    .0017293
         age |   .0288141   .0134981     2.13   0.033     .0023583    .0552698
     surgery |  -.9695805   .4361069    -2.22   0.026    -1.824334   -.1148266
------------------------------------------------------------------------------
```

On remarque que les coefficients estimés sont particulièrement proche.

Ajustement discret

- Il s'agit d'introduire la variable de durée dans le modèle comme une variable catégorielle (factor).
- Elle n'est pas conseillée si beaucoup de points d'observation, ce qui est le cas dans l'application précédente, et surtout en présence de points d'observation sans évènement.
- A l'inverse, si on ne dispose que peu de points d'observation, la paramétrisation avec une durée continue n'est pas conseillé (idem Cox).

On va supposer qu'on dispose seulement de quatre intervalles d'observations. Pour l'exemple, on va créer ces intervalles 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), on ne conserve qu'une ligne par intervalle d'observation pour chaque individu soumis au risque de décéder.

```
        4 |
 quantiles |           e
     of t |         0          1 |     Total
----------+----------------------+----------
        1 |        50         53 |       103
        2 |        35         11 |        46
        3 |        27          5 |        32
        4 |        10          6 |        16
----------+----------------------+----------
    Total |       122         75 |       197


Logistic regression                             Number of obs     =        197
                                                LR chi2(6)        =      39.30
                                                Prob > chi2       =     0.0000
Log likelihood = -111.23965                     Pseudo R2         =     0.1501

------------------------------------------------------------------------------
           e |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ct4 |
          2  |  -1.033368   .4188719    -2.47   0.014    -1.854342   -.2123944
          3  |  -1.615245    .544858    -2.96   0.003    -2.683147   -.5473433
          4  |  -.4789305   .5992969    -0.80   0.424    -1.653531    .6956698
             |
        year |  -.2032436   .0931956    -2.18   0.029    -.3859036   -.0205835
         age |   .0468518   .0184958     2.53   0.011     .0106006     .083103
     surgery |  -1.110163   .5025594    -2.21   0.027    -2.095161   -.1251644
       _cons |   12.44666   6.653694     1.87   0.061      -.59434    25.48766
------------------------------------------------------------------------------
```

Au niveau de l'interprétation, avec 37% d'évènements sur l'ensemble des observations, il n'est plus possible d'interpréter le modèle en terme de rapport de probabilité. La lecture en termes d'Odds Ratio s'impose ici.

Le tableau suivant présente les probabilités estimées de décéder à partir d'un modèle avec la durée seulement (sous forme discrète).

Durées	p
0 à 4 mois	0.51
4 à 11 mois	0.24
11 à 23 mois	0.16
23 à 61 mois	0.37

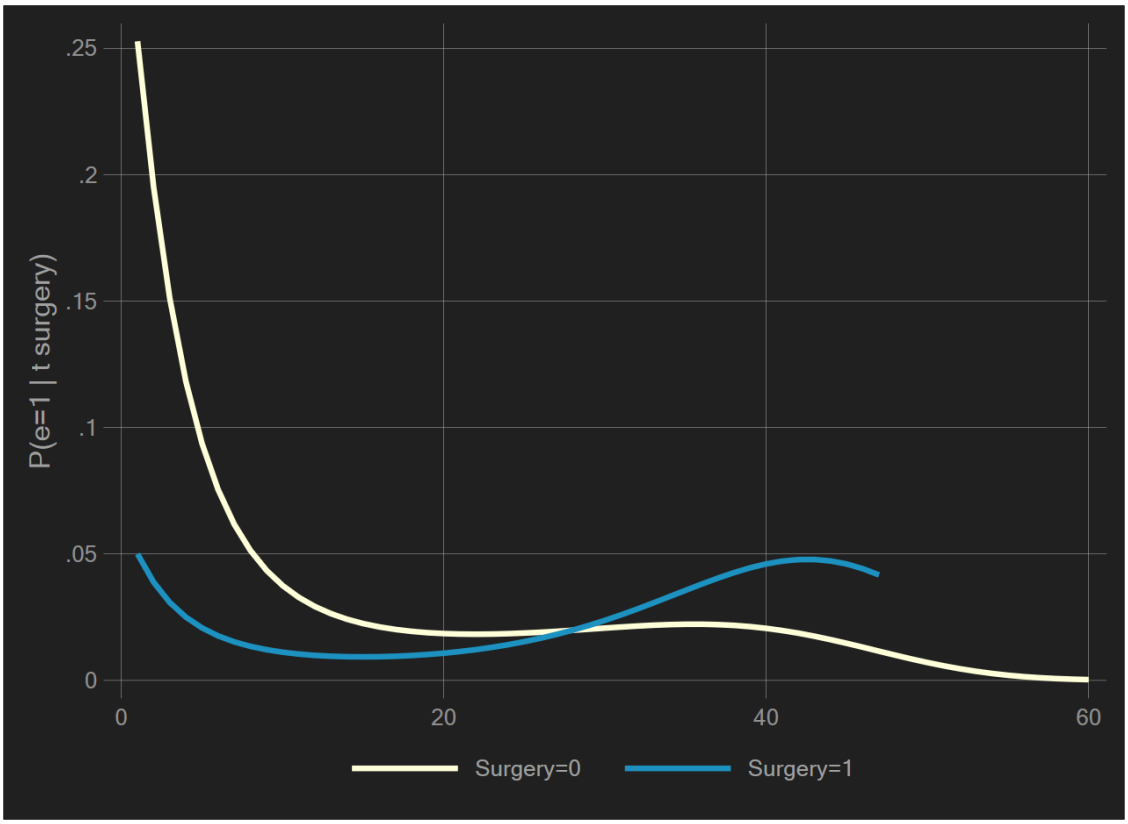
Modèle à temps discret et hypothèse PH

- 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.
- Comme pour le modèle de Cox, la correction de la non proportionalité peut se faire en intégrant une interaction avec la durée dans le modèle.

Les variables *year* et *age* seront omises pour faciliter la représentation graphique.

```
Logistic regression                             Number of obs     =      1,127
                                                LR chi2(5)        =      84.78
                                                Prob > chi2       =     0.0000
Log likelihood = -233.29204                     Pseudo R2         =     0.1538

------------------------------------------------------------------------------
           e |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |   -.373826    .083913    -4.45   0.000    -.5382924   -.2093595
  1.surgery  |  -1.929061   .6920142    -2.79   0.005    -3.285383   -.5727377
             |
 surgery#c.t |
          1  |   .0690069   .0333128     2.07   0.038      .003715    .1342987
             |
          t2 |   .0137676   .0052405     2.63   0.009     .0034964    .0240388
          t3 |  -.0001596   .0000828    -1.93   0.054    -.0003218    2.62e-06
       _cons |   -.723378   .2445386    -2.96   0.003    -1.202665   -.2440911
------------------------------------------------------------------------------
```


.25
.2
.15
.1
.05
0
P(e=1 | t surgery)
0
20
40
60
Surgery=0
Surgery=1

Introduction de variables dynamiques

Cette section sera principalement traitée par l'exemple, et on ne s'intéressera qu'aux variables de type discrete.

- Dans un modèle de durée, une variable dynamique peut-être appréhendée comme une interaction entre la durée et une variable.
- Pour un modèle de Cox, l'hypothèse de risque proportionnel ne peut donc pas être testée.
- Ne pas tenir compte du caractère dynamique d'une dimension peut conduire à des interprétations erronées.
- La façon de modéliser les dimensions dynamiques en analyse des durées peut conduire à des biais de causalité, en particulier dans 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.

Facteur dynamique traitée de manière fixe

On reprend l'exemple sur malformation cardiaque, en ajoutant la variable relative à la greffe : la transplantation réduit-elle le risque (journalier de décéder) / augment la durée de survie.

On a dans la base 2 variables: une variable binaire pour savoir si l'individu à été greffé ou non, **tranplant**, et une variable continue tronquée donnant la durée en jour jusqu'à la greffe greffe (0 si pas de greffe), **wait**.

On va dans un premier temps estimer le modèle (de Cox) avec la variable fixe transplant.

```
Cox regression -- Efron method for ties

No. of subjects =          103                 Number of obs    =         103
No. of failures =           75
Time at risk    =        31938
                                               LR chi2(4)       =       49.81
Log likelihood  =   -273.21499                 Prob > chi2      =      0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |    -0.0909     0.0659    -1.38   0.168      -0.2201      0.0383
         age |     0.0579     0.0147     3.95   0.000       0.0292      0.0866
   1.surgery |    -0.6547     0.4475    -1.46   0.143      -1.5318      0.2224
1.transplant |    -1.6484     0.2792    -5.90   0.000      -2.1957     -1.1011
------------------------------------------------------------------------------
```

Interprétation?

Au niveau des données le modèle à été estimé, pour une personne greffée, à partir de ce mapping:

id	year	age	surgery	transplant	wait	_d	_t	_t0
10	68	42	0	1	12	0	1	0
10	68	42	0	1	12	0	2	1
10	68	42	0	1	12	0	3	2
10	68	42	0	1	12	0	5	3
10	68	42	0	1	12	0	5.0999999	5
10	68	42	0	1	12	0	6	5.0999999
10	68	42	0	1	12	0	8	6
10	68	42	0	1	12	0	9	8
10	**68**	**42**	**0**	**1**	**12**	**0**	**12**	**9**
10	**68**	**42**	**0**	**1**	**12**	**0**	**16**	**12**
10	**68**	**42**	**0**	**1**	**12**	**0**	**17**	**16**
10	**68**	**42**	**0**	**1**	**12**	**0**	**18**	**17**
10	**68**	**42**	**0**	**1**	**12**	**0**	**21**	**18**
10	**68**	**42**	**0**	**1**	**12**	**0**	**28**	**21**
10	**68**	**42**	**0**	**1**	**12**	**0**	**30**	**28**
10	**68**	**42**	**0**	**1**	**12**	**0**	**32**	**30**
10	**68**	**42**	**0**	**1**	**12**	**0**	**35**	**32**
10	**68**	**42**	**0**	**1**	**12**	**0**	**36**	**35**
10	**68**	**42**	**0**	**1**	**12**	**0**	**37**	**36**
10	**68**	**42**	**0**	**1**	**12**	**0**	**39**	**37**
10	**68**	**42**	**0**	**1**	**12**	**0**	**40**	**39**
10	**68**	**42**	**0**	**1**	**12**	**0**	**43**	**40**
10	**68**	**42**	**0**	**1**	**12**	**0**	**45**	**43**
10	**68**	**42**	**0**	**1**	**12**	**0**	**50**	**45**
10	**68**	**42**	**0**	**1**	**12**	**0**	**51**	**50**
10	**68**	**42**	**0**	**1**	**12**	**0**	**53**	**51**
10	**68**	**42**	**0**	**1**	**12**	**1**	**58**	**53**

Problème: la personne est codée transplantée avant le jour de la greffe. L'effet causal est donc mal mesuré si sa dimension temporelle a été ignorée.

Estimation avec une variable dynamique

Modèle de Cox

Il convient donc de modifier l'information avec le délai d'attente jusqu'à la greffe. On doit générer une variable qui prend la valeur 0 avant la greffe et la valeur 1 à partir du jour où la personne a été opérée:

- $tvc = 1$ si $transplant = 1$ et $t \geq wait$
- $tvc = 0$ sinon

C’est le même principe qu’avec la variable évènement/censure. La personne est marquée « death » le jour de sa mort et non avant (le modèle serait pas estimable).

id	year	age	surgery	tvc	wait	_d	_t	_t0
10	68	42	0	0	12	0	1	0
10	68	42	0	0	12	0	2	1
10	68	42	0	0	12	0	3	2
10	68	42	0	0	12	0	5	3
10	68	42	0	0	12	0	5.0999999	5
10	68	42	0	0	12	0	6	5.0999999
10	68	42	0	0	12	0	8	6
10	68	42	0	0	12	0	9	8
10	**68**	**42**	**0**	**1**	**12**	**0**	**12**	**9**
10	**68**	**42**	**0**	**1**	**12**	**0**	**16**	**12**
10	**68**	**42**	**0**	**1**	**12**	**0**	**17**	**16**
10	**68**	**42**	**0**	**1**	**12**	**0**	**18**	**17**
10	**68**	**42**	**0**	**1**	**12**	**0**	**21**	**18**
10	**68**	**42**	**0**	**1**	**12**	**0**	**28**	**21**
10	**68**	**42**	**0**	**1**	**12**	**0**	**30**	**28**
10	**68**	**42**	**0**	**1**	**12**	**0**	**32**	**30**
10	**68**	**42**	**0**	**1**	**12**	**0**	**35**	**32**
10	**68**	**42**	**0**	**1**	**12**	**0**	**36**	**35**
10	**68**	**42**	**0**	**1**	**12**	**0**	**37**	**36**
10	**68**	**42**	**0**	**1**	**12**	**0**	**39**	**37**
10	**68**	**42**	**0**	**1**	**12**	**0**	**40**	**39**
10	**68**	**42**	**0**	**1**	**12**	**0**	**43**	**40**
10	**68**	**42**	**0**	**1**	**12**	**0**	**45**	**43**
10	**68**	**42**	**0**	**1**	**12**	**0**	**50**	**45**
10	**68**	**42**	**0**	**1**	**12**	**0**	**51**	**50**
10	**68**	**42**	**0**	**1**	**12**	**0**	**53**	**51**
10	**68**	**42**	**0**	**1**	**12**	**1**	**58**	**53**

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

```
Cox regression -- Efron method for ties

No. of subjects =          103                 Number of obs    =       3,668
No. of failures =           75
Time at risk    =      31938.1
                                               LR chi2(4)       =       17.70
Log likelihood  =   -289.27058                 Prob > chi2      =      0.0014

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |  -.1202612   .0673414    -1.79   0.074     -.252248    .0117256
         age |   .0304498   .0138998     2.19   0.028     .0032068    .0576929
   1.surgery |  -.9829386   .4365524    -2.25   0.024    -1.838566   -.1273116
      1.tvc |  -.0826682   .3047751    -0.27   0.786    -.6800165      .51468
------------------------------------------------------------------------------
```

Interprétation?

Logiciels

SAS: la base n'est pas modifiée et la création de la TVC est faite "en aveugle" dans la procédure *phreg*.
Stata, **R**, **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.

Modèle à temps discret

Même principe pour la construction de la variable dynamique (rappel : l'échelle temporelle est le mois)

```
  +---------------------------------------------+
  | id   year   age   surgery   tvc   mwait   t |
  |---------------------------------------------|
  | 13     68    54         0     0       2   1 |
  | 13     68    54         0     1       2   2 |
  | 13     68    54         0     1       2   3 |
  +---------------------------------------------+


Logistic regression                             Number of obs     =      1,127
                                                LR chi2(7)        =      90.73
                                                Prob > chi2       =     0.0000
Log likelihood = -230.32152                     Pseudo R2         =     0.1645

------------------------------------------------------------------------------
           e |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           t |   -.365048   .0915105    -3.99   0.000    -.5444052   -.1856907
          t2 |   .0139226   .0053256     2.61   0.009     .0034846    .0243606
          t3 |   -.000162   .0000815    -1.99   0.047    -.0003217   -2.27e-06
        year |  -.1324928   .0737516    -1.80   0.072    -.2770433    .0120577
         age |    .033829   .0149503     2.26   0.024      .004527    .0631311
     surgery |  -1.007795   .4490177    -2.24   0.025    -1.887854   -.1277365
         tvc |  -.0543011   .3114096    -0.17   0.862    -.6646528    .5560505
       _cons |   7.060589   5.305278     1.33   0.183    -3.337566    17.45874
------------------------------------------------------------------------------
```

Quelques remarques sur les problèmes de causalité avec les variables dynamiques

Ce qui suit est important.

- Rappel: la cause précède toujours l'effet (problématique quantique excepté).
- Lorsque l'évènement étudié n'est pas intrasèquement de type absorbant comme le décès, la « cause » peut se manifester ou 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.

- Logique d'*adaptation* **[OK]**: la « cause » identifiée est mesurée avant l'évènement étudié.
- Logique d'*anticipation* **[Problème]**: la « cause » identifiée est mesurée après l'évènement étudié. L'origine causale est bien antérieure à l'évènement, mais elle n'est pas observable.