

# **FORMATION**

# **ANALYSE DES DUREES**

# **2022**

**Marc Thévenin**

## Table des matières

Introduction .....	4
<b>Questions</b> .....	4
<b>Terminologies</b> .....	4
<b>Exemples d'analyse</b> .....	4
<b>Eléments nécessaires à l'analyse</b> .....	5
<b>Bibliographie</b> .....	5
Données et théorie .....	6
<b>Les données biographiques</b> .....	6
<b>Données prospectives et rétrospectives</b> .....	6
<b>Enregistrement des données</b> .....	7
<b>Exemples de mise à disposition</b> .....	9
<b>La théorie de l'analyse des durées</b> .....	11
<b>Temps et durée</b> .....	11
<b>Le Risk Set</b> .....	12
<b>La Censure</b> .....	12
<b>Les grandeurs</b> .....	16
<b>Compléments</b> .....	21
<i>Absence de censures à droites</i> .....	22
<b>Méthodes non paramétriques</b> .....	24
<b>Introduction</b> .....	24
<b>Les variables d'analyse</b> .....	24
<b>Calcul de la fonction de survie (séjour)</b> .....	24
<b>La méthode actuarielle</b> .....	25
<b>La méthode de Kaplan-Meier</b> .....	28
<b>Tester l'égalité des courbes de survie (méthode KM)</b> .....	34
<b>Tests du log-rank</b> .....	34
<b>Comparaison des RMST</b> .....	39
<b>Les modèles à risques proportionnels</b> .....	43
<b>Introduction aux modèles à risques proportionnels</b> .....	43
<b>Le modèle semi-paramétrique de Cox</b> .....	46

<b>Vraisemblance partielle et estimation des paramètres.....</b>	46
<b>Estimation des paramètres .....</b>	48
<b>Lecture des résultats .....</b>	49
<b>L'hypothèse de constance des rapports de risque .....</b>	51
<b>Modèles à temps discret.....</b>	59
<b>Organisation des données .....</b>	59
<b>Estimation et ajustement de la durée .....</b>	60
<b>Modèle à temps discret et hypothèse PH.....</b>	64
<b>Introduction de variables dynamiques.....</b>	66
<b>Facteur dynamique traitée de manière fixe .....</b>	66
<b>Estimation avec une variable dynamique .....</b>	68
<b>Compléments.....</b>	71
<b>Modèles paramétriques.....</b>	71
<b>Les modèles paramétriques usuels.....</b>	71
<b>Hypothèse AFT: Accelerated Failure Time .....</b>	71
<b>Principe de construction des modèles AFT .....</b>	72
<b>Quelques lois de distribution .....</b>	72
<b>Risques concurrents .....</b>	76
<b>Risques « cause-specific » et biais sur les estimateurs KM .....</b>	76
<b>Estimations en présence de risques concurrents.....</b>	78
<b>Estimation non paramétrique .....</b>	78
<b>Modèles semi paramétrique et à temps discret .....</b>	83
<b>Fragilité et immunité .....</b>	86
<b>Applications logiciels .....</b>	87
<b>SAS .....</b>	87
<b>R .....</b>	115
<b>Stata .....</b>	139
<b>Python.....</b>	162

# Introduction

## Questions

On dispose de données dites « longitudinales », et on cherche à appréhender l'occurrence 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 apparaître ou dont la valeur peut changer durant la période d'observation.

## Terminologies

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

## Exemples d'analyse

### Nuptialité, Mise en couple

cohabiter, décohabiter, se marier, rompre une union ...

### Logement/mobilité

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 un diagnostic, survivre après l'administration d'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 règles...)
- Une définition précise de l'évènement d'étude.
- Une durée entre le début et la fin de la période d'observation, si nécessaire la fin de la période d'exposition au risque.

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

Des variables « *explicatives* » ou covariables :

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

## Bibliographie

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

Cours Gilbert Colletaz (Université d'Orléans).

Le cours est mis à jour tous les ans. Il n'est plus possible d'accéder directement à la dernière version du document, mais il est néanmoins possible de télécharger la version 2016 (<https://docplayer.fr/69359088-Modeles-de-survie-notes-de-cours-master-2-esa-voies-professionnelle-et-recherche-gilbert-colletaz.html>). Applications avec Sas.

Document de travail de Simon Quantin (Insee).

Egalement un excellent document, qui couvre l'ensemble des techniques de base d'analyse des durées en temps dit continu (<https://www.insee.fr/fr/statistiques/3695681>). Il propose sûrement la meilleure introduction en langue française à la problématique de la *fragilité*. Applications avec R. On peut juste regretter que les risques concurrents ne soient pas abordés.

# 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).
- **Avantage:** qualité des données (moins de biais de mémoire).
- **Inconvénients:** coût important, délais pour mettre en oeuvre 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. Ces informations sont par exemple actuellement exploitées à l'Ined dans le cadre du projet « **worklife** » : <https://worklife.site.ined.fr/> ). Elles engendrent en revanche des interrogations techniques liés à l'inférence (on ne travaille directement pas sur des échantillons).

#### 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, mariage), il est difficile d'aller chercher des datations trop fines avec une retrospectivité assez longue.

Les deux types de recueil peuvent être mixés avec des enquêtes à passages répétés comprenant des informations rétrospectives 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-événement, de type chronologique, avec des repères temporels en ligne (âge, année). En colonne, sont complétés de manière progressive et relative, les événements 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 pourra s'avérer particulièrement chronophage et complexe si plusieurs modules doivent être appariés. On distingue trois formats d'enregistrement.

### **Large [format individu]**

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

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-événements]

C'est le format le plus courant de mise à disposition des enquêtes biographiques. Si l'évènement est de type continu, par exemple le 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, une fin de séquence peut se produire juste avant l'enquête la même année. Il est également possible d'avoir une information qui ne s'est pas encore produite au moment de l'enquête, mais qui aura lieu peu de temps après (personne enceinte, donc une naissance probable la même année).

Exemple précédent (trajectoires discontinues):

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

### Long [format individu-périodes]

Typique des recueils prospectifs. Ils engendrent des lignes sans information supplémentaire.

Exemple précédent:

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

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

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

## Exemples de mise à disposition

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

### Enquête biographie et entourage (Ined)

[https://grab.site.ined.fr/fr/enquetes/france/biographie\\_entourage/](https://grab.site.ined.fr/fr/enquetes/france/biographie_entourage/)

#### Base sur les caractéristiques individuelles

VIEWTABLE: TMP1.tego								
	Identifiant questionnaire	prénom d'ego	sexe d'ego	Date de naissance	Département de naissance	Commune ou pays de naissance	Pays ou DOM-TOM de naissance	Numéro INSEE de la commune de naissance
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			99139 PORTUGAISE
4	104 LEON		1	01/13/1933	93 BONDY			10 FRANCAISE
5	105 FRANCOIS		1	12/27/1932	99 ALGER			99352 FRANCAISE
6	106 EVELYNNE		2	11/21/1950	99 NR			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	06/08/1949	59 LOUVRIOIL			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	YUGOSLAVIE		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 ESQUERDSES			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 PM	
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 PM	
6	102	5		2	37	TOURS	261	22	4	1 PM	
7	102	7		3T							
8	102	7		3	37	TOURS	261	12	99	1 PM	
9	102	10 NF3		4	75	PARIS-18E_ARRONDISSEMENT	118	41	2	0 PM	
10	102	22 M1		5	93	BOBIGNY	8	22	1	1 12	
11	102	26		6	93	BOBIGNY	8	21	4	4 12	
12	102	37		7	93	LIVRY-GARGAN	46	21	3	4 12	
13	103	0		1	99	PORTUGAL	99139	22	2	0 PM	
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 12	

## Enquête MAFE (Ined)

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

### *Base sur les caractéristiques individuelles*

ident	q1	q1a	statut_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 la durée d'attente jusqu'à l'occurrence d'un événement.

- 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 durée est sa mesure. La durée est tout simplement calculer par la différence, à 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 strictement continu car deux évènements ne peuvent pas avoir lieu en « même temps » ; c'est donc l'échelle temporelle choisie ou imposée par l'analyse et les données qui pourra rendre cette mesure continue ou discrète/groupée.

Pour un physicien travaillant sur la théorie de la relativité avec des horloges atomiques, une minute (voire une seconde) est une mesure très discrète pour ne pas dire grossière du temps, alors que pour un géologue c'est une mesure continue. Pour ces deux disciplines, cette échelle de mesure n'est pas adaptée à leur recherche. 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. Généralement dans les sciences sociales avec un recueil de données de type rétrospectif, les mesures discrètes sont plutôt de nature groupées. Pour une même année, on considérera indifféremment des évènements qui se produiront un premier janvier et un 31 décembre d'une même année.

#### A retenir

Durée continue : absence (ou très peu) d'évènements simultanés.

Durée discrète/groupée : présence d'évènements simultanés (en nombre élevé).

## 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 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 de l'exposition sans avoir connu l'évènement. Elles peuvent connaître l'évènement après la date de l'enquête ou en étant perdues de vue. Sinon on sous-estime la durée moyenne de survenue de l'évènement.

### *Exemple*

On effectue une enquête auprès de femmes 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 recueils rétrospectifs.

## *Problème posé par la censure informative*

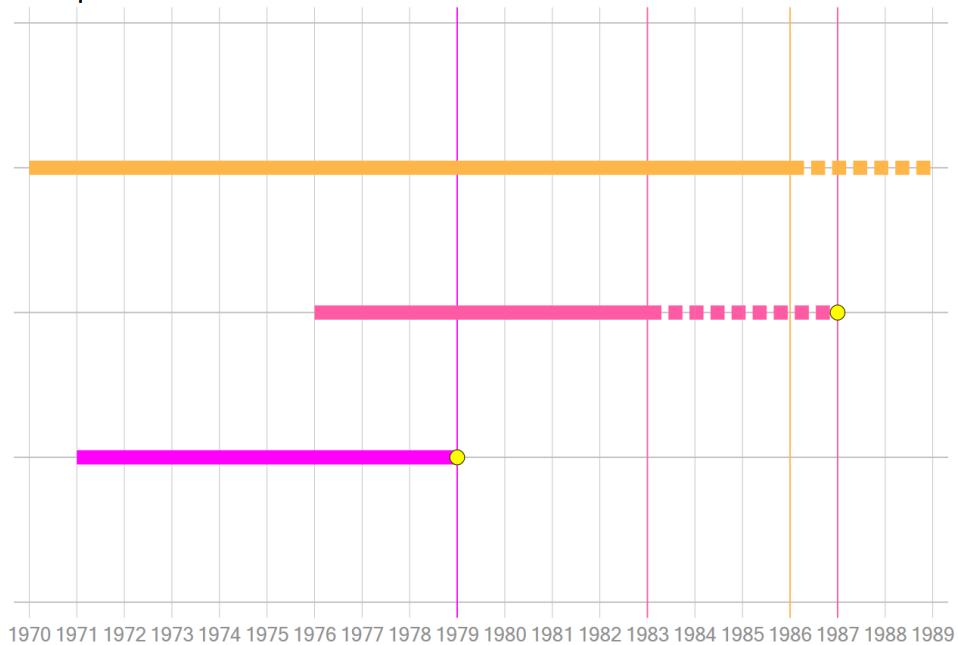
Par exemple en analysant des décès avec un recueil prospectif, si un individu est perdu de vue en raison d'une dégradation de son état de santé, l'indépendance entre la cause de la censure et le décès ne peut plus être assurée.

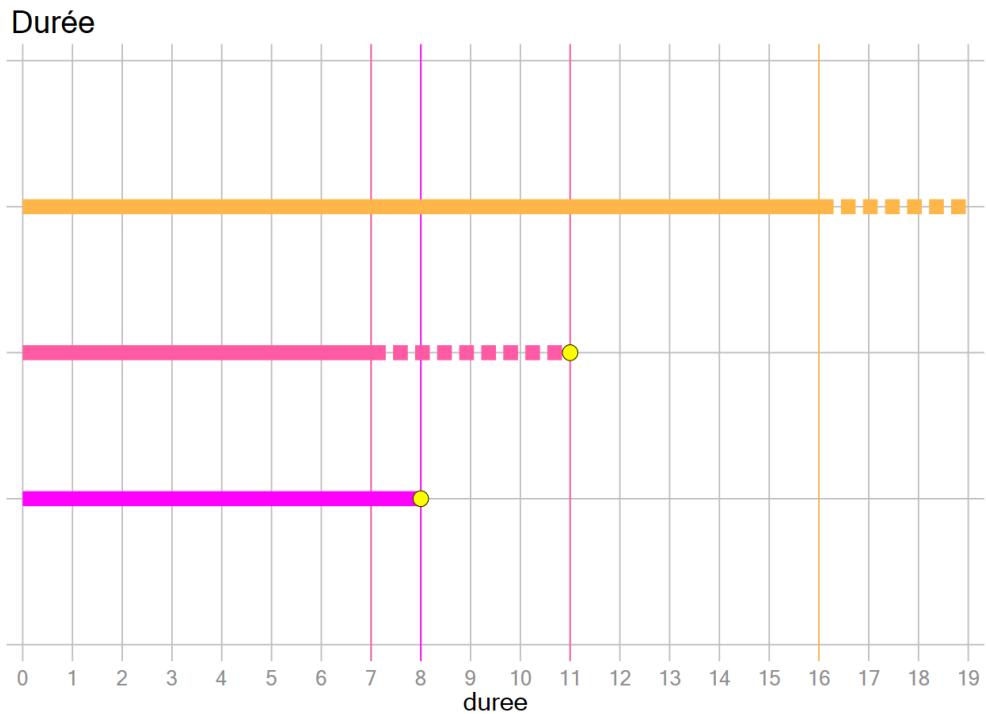
A l'Ined l'exploitation du registre des personnes atteintes de mucoviscidose (G.Bellis) donne une autre illustration de ce phénomène. Chaque année un nombre significatif personnes sortent du registre (pas de résultats aux examens annuels). Si certain.e.s perdu.e.s de vue s'expliquent par des déménagements, 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

Temps calendaire





## Censure à gauche, troncature et censure par intervalle

### *Censure à gauche*

L'événement 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 (late-entry)*

Par l'exemple, on analyse la survie d'une population. Seule la survie des individus vivants à l'inclusion peut être analysée. Des phénomènes de sélection peuvent être rencontré. 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).

La troncature peut affecter directement le phénomène étudié. Si on souhaite analyser le problème des tentatives de suicide de personnes interrogées via une enquête, par définition ont pourra seulement récolter l'information des survivant.e.s. Cela rendra 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.

### 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  $\lim_{t \rightarrow \infty} S(t) = 0$

La fonction de survie strictement non croissante.

## La fonction de répartition $F(t)$

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

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

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

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

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 distribution des moments où les évènements ont lieu. Elle est donnée dans un premier temps par la probabilité de connaître l'évènement dans un petit intervalle de temps  $dt$ . 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:

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

## 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 connaît 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 également écrire :

$$h(t) = \frac{f(t)}{S(t)} = \frac{F'(t)}{S(t)} = -\frac{S'(t)}{S(t)}$$

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

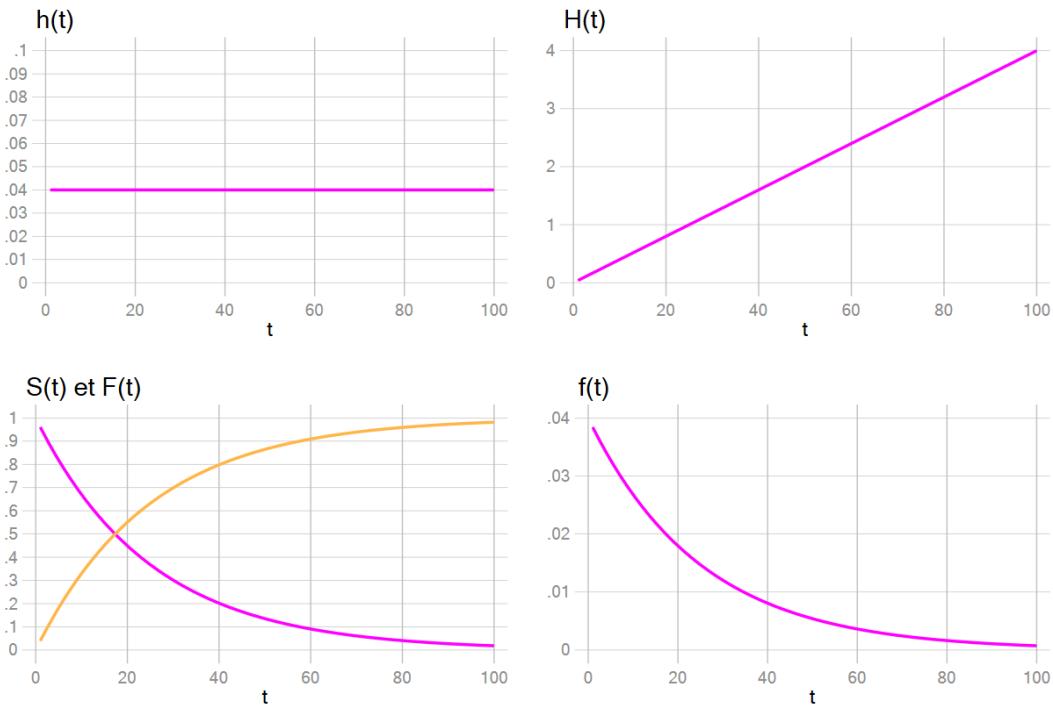
- $S(t) = e^{-H(t)}$
- $F(t) = 1 - e^{-H(t)}$
- $f(t) = h(t) \times e^{-H(t)}$

### *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 comme la durée de vie des ampoules):

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

## Grandeurs de la loi exponentielle - Risque constant = 0.04



### Application: risque et échelles temporelles

Fortement inspiré, pour ne pas dire copié, de l'excellent cours de **Gilbert Colletaz**:

<https://www.univ-orleans.fr/dcg/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 (c'est donc plutôt un taux tel qu'on le définit généralement). En soit cela ne pose pas de problème comme il s'agit d'un nombre moyen d'événements espérés durant une unité de temps, mais pour des événements qui ne peuvent pas se répéter, événements 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'occurrence de l'événement.

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.

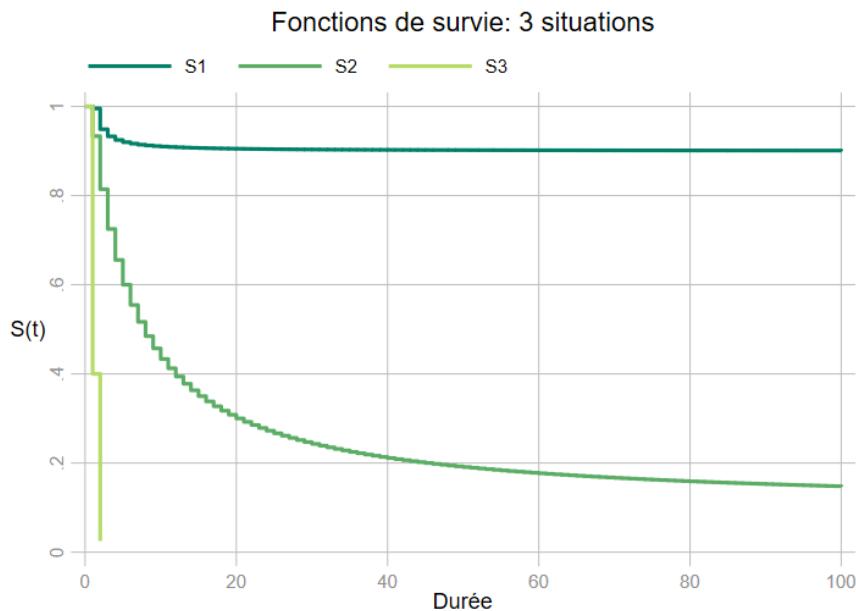
Durée	Nombre de décès
1	1
2	1
3	3
4	9
5	30
6	40
7	10
8	3
9	2
10	1
N=100	

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

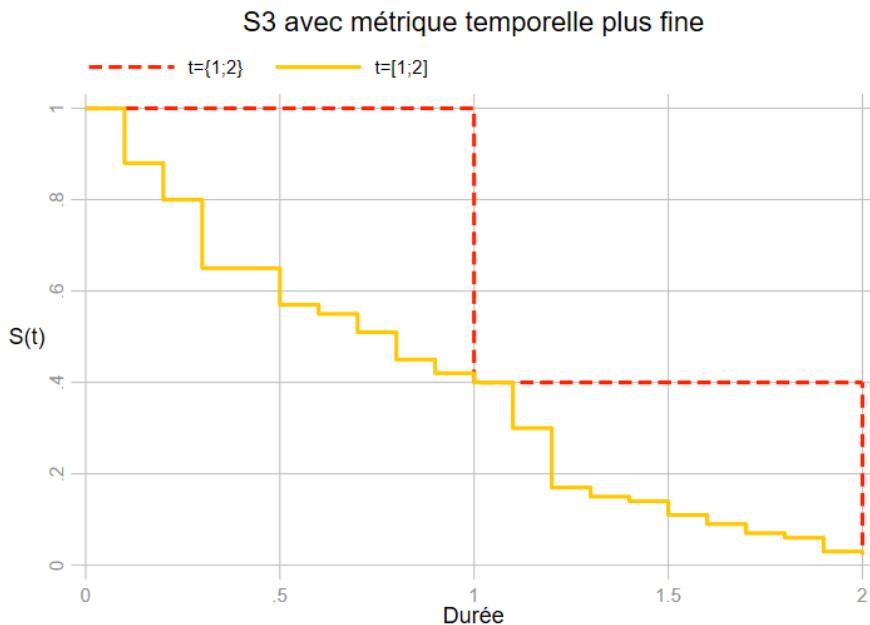
## Compléments

### Forme des fonctions de survie

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



- La survie tombe à 0 très/trop rapidement (courbe S3) : il n'y a donc pas ou presque pas de durée (par exemple presque tout l'échantillon observe l'évènement la première année de l'exposition). Les méthodes en temps continu ne sont a priori pas adaptées à ce genre de situation. Si on dispose d'une information plus fine pour dater les évènements, la fonction de séjour pourra reprendre une forme plus « standard ». Dans le graphique,  $S(t = 1) = 0.4$  et  $S(t = 2) = 0.025$ , mais si on dispose par exemple de 10 points d'observations supplémentaires dans chaque durée groupée:



- Très peu d'événements et la fonction de séjour suit une asymptote nettement supérieur à 0 ( $\lim_{t \rightarrow \infty} S(t) = a$  avec  $a > 0$ ). La question est plus délicate car on interroge l'exposition au risque d'une partie de l'échantillon ou, dit autrement on peut penser qu'une fraction est immunisé au risque. Cette problématique est rapidement posée en fin de formation.

### Absence de censures à droites

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

### Utilisation des pondérations

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

caractéristiques de la population sont celles de 2017, pour celle de 70 ans celles de 1972. On fait comment ??????

# 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 loin la plus diffusée en raison, par exemple, 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  $e_1, e_2, \dots, e_k$  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  $t_1, t_2, \dots, t_k$  jusqu'à l'évènement ou la censure.
1. En théorie, il ne peut pas y avoir d'évènement en  $t = 0$ .

### 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  $q(t_i) = \frac{1}{r_i}$ .

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

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  $t_0$  ( $S(0) = 1$ ):

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

### 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:  $t_0 < t_1 < \dots < t_J$  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 milieu de l'intervalle.

### Calcul de $S(t_i)$

On applique la méthode de la section précédente avec

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

### Calcul de la durée médiane (ou autre quantiles)

#### Rappel

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

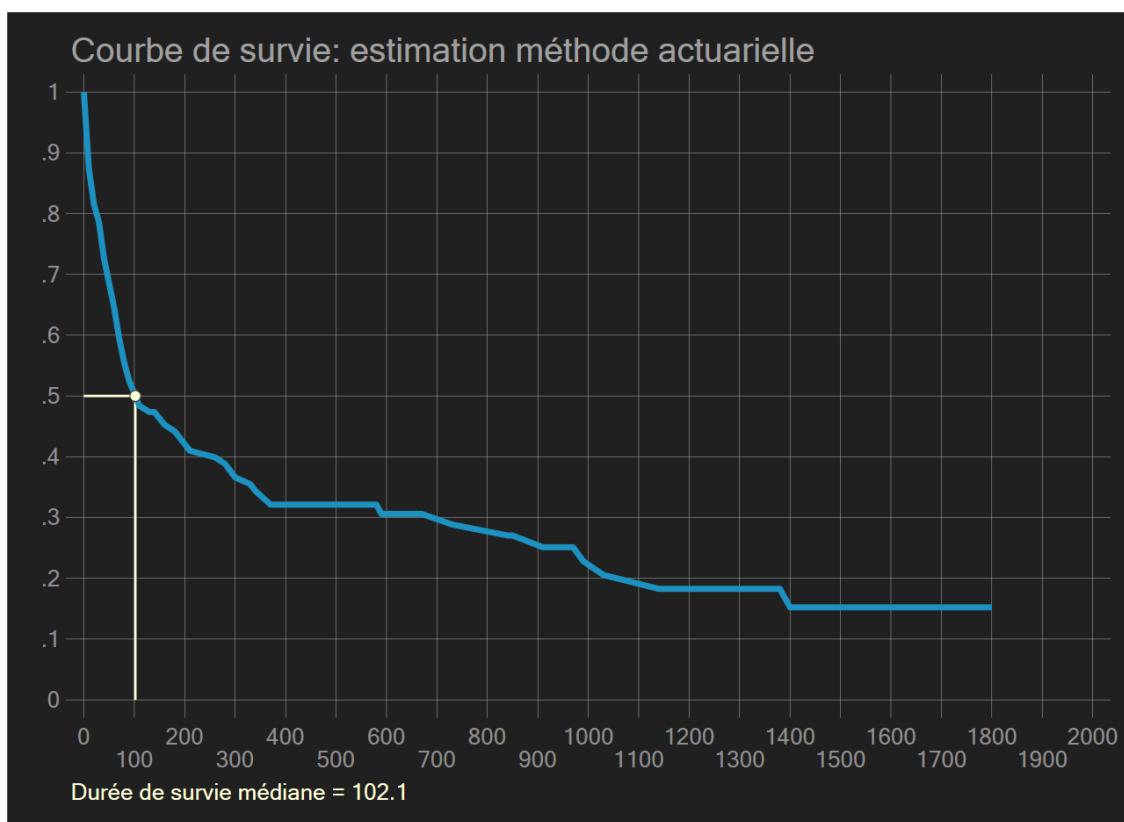
**SAS:** incluse dans *proc lifetest* avec l'option `method=act`.

**Stata:** commande *Itable*. 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
<b>90 100</b>	<b>52</b>	<b>2</b>	<b>0</b>	<b>0.5042</b>	<b>0.0499</b>	<b>0.4029 0.5971</b>
<b>100 110</b>	<b>50</b>	<b>2</b>	<b>1</b>	<b>0.4838</b>	<b>0.0500</b>	<b>0.3831 0.5773</b>
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 différents 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 qui exploite toute l'information disponible est celle dite de **Kaplan-Meier (KM)**.
- Il y a autant d'intervalle 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, 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 :

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

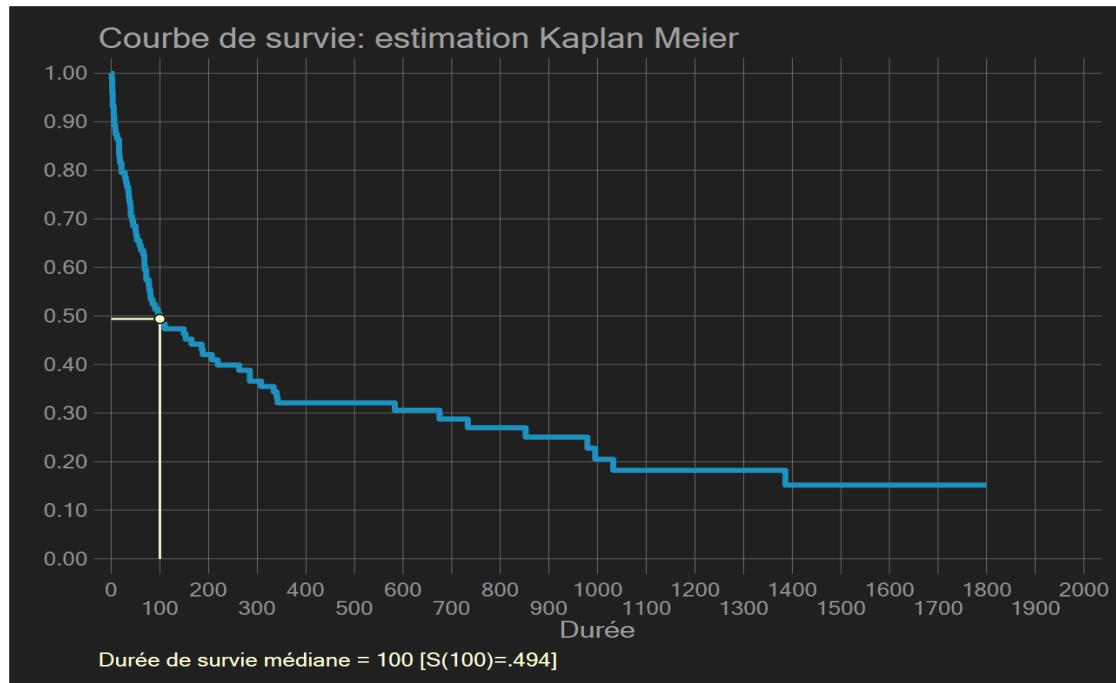
### Récupération de la médiane

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 savante pour obtenir ce résultat, c'est une convention. Attention, il n'est pas impossible que le % de survivant.e.s soit bien en deçà de 50% pour obtenir cette durée médiane.

### Exemple

La durée médiane est égale à 100 jours (102 pour la méthode actuarielle) et correspond à  $S(t) = 0.4940$ .



Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. 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 donné 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 estimateurs sont donnés par la fonction `survfit` de la librairie `survival`.

**Python**: les résultats sont donnés dans la librairie `lifeline` par des fonctions dont le nom est tout bonnement 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

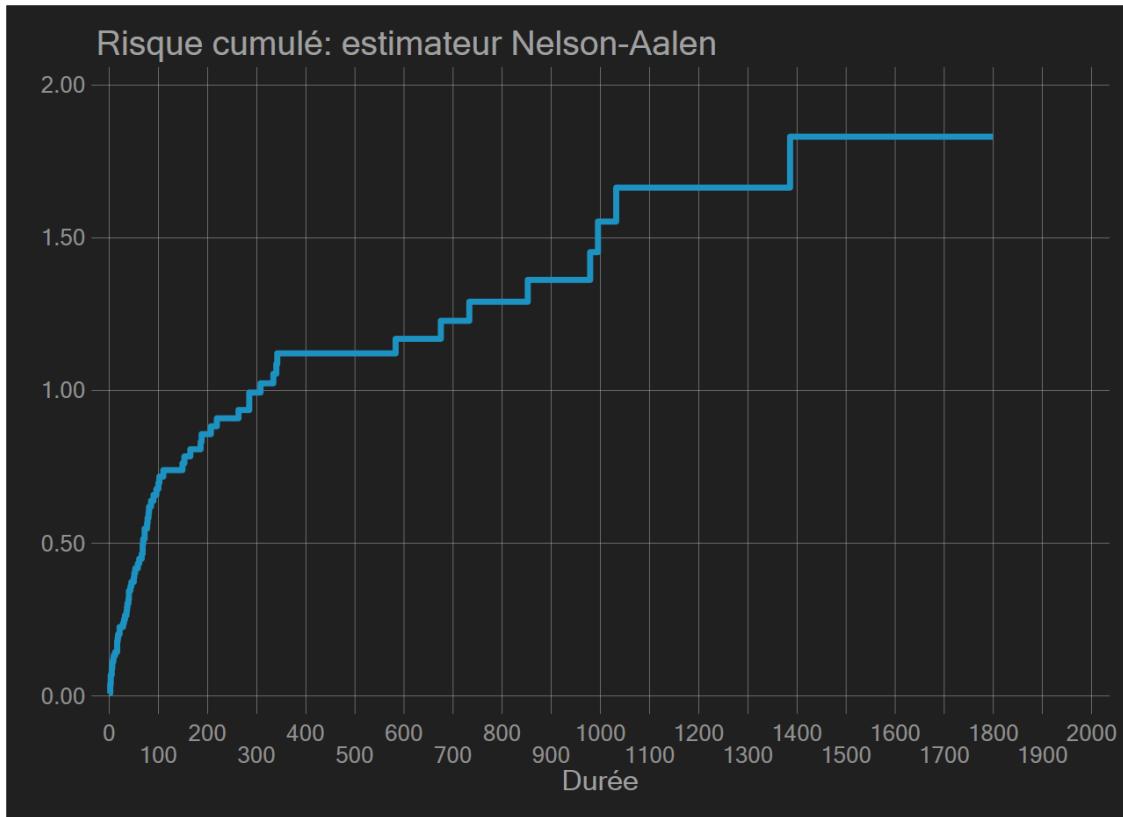
S= valeur de la survie en t

### Quantités associées

#### Le risque cumulé

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

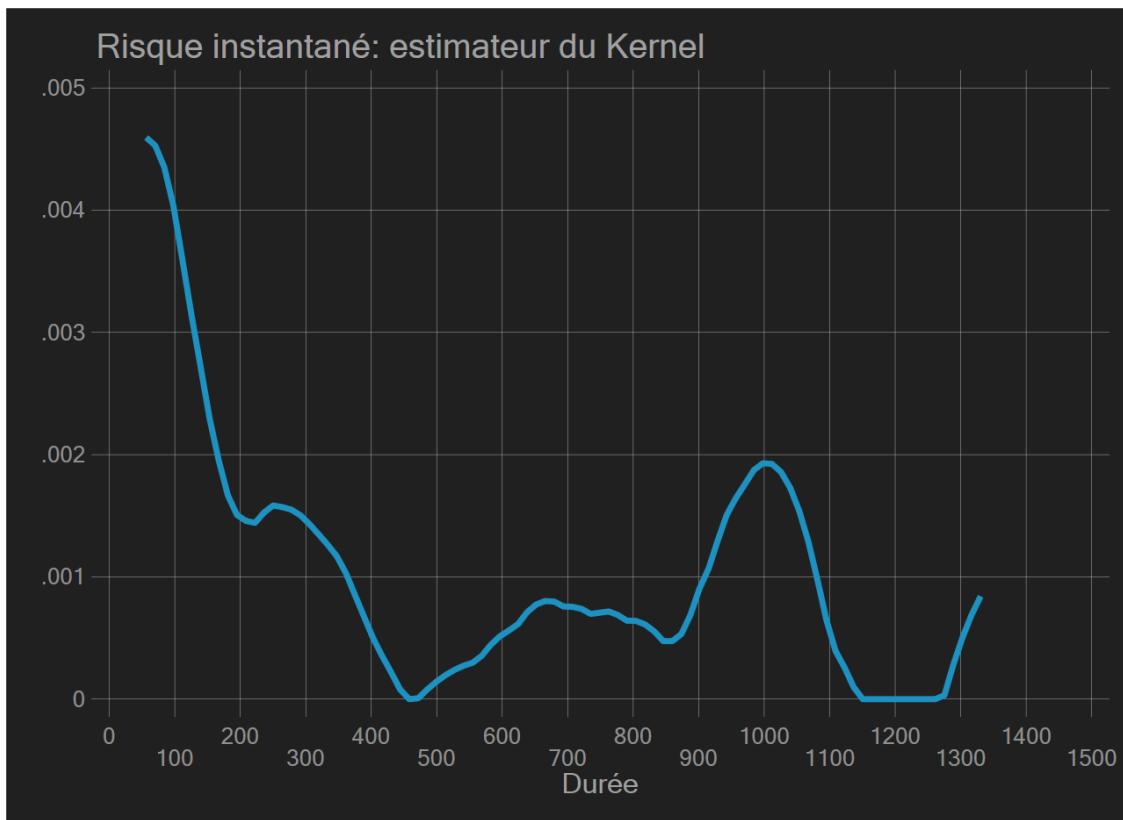
$$H(t) = \sum_{t_i \leq k} q(t_i)$$



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 - noeuds-).



## 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 événements observés dans les différentes strates ou les différents échantillons.

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

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 dits du Khi2, 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 principale 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 la statistique du test standard du  $\chi^2$  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  $e_{i2} = R_{i2} \times \frac{d_i}{R_i}$ . Leur somme pour tous les temps d'évènement est égal 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  $\chi^2(g - 1)$ .

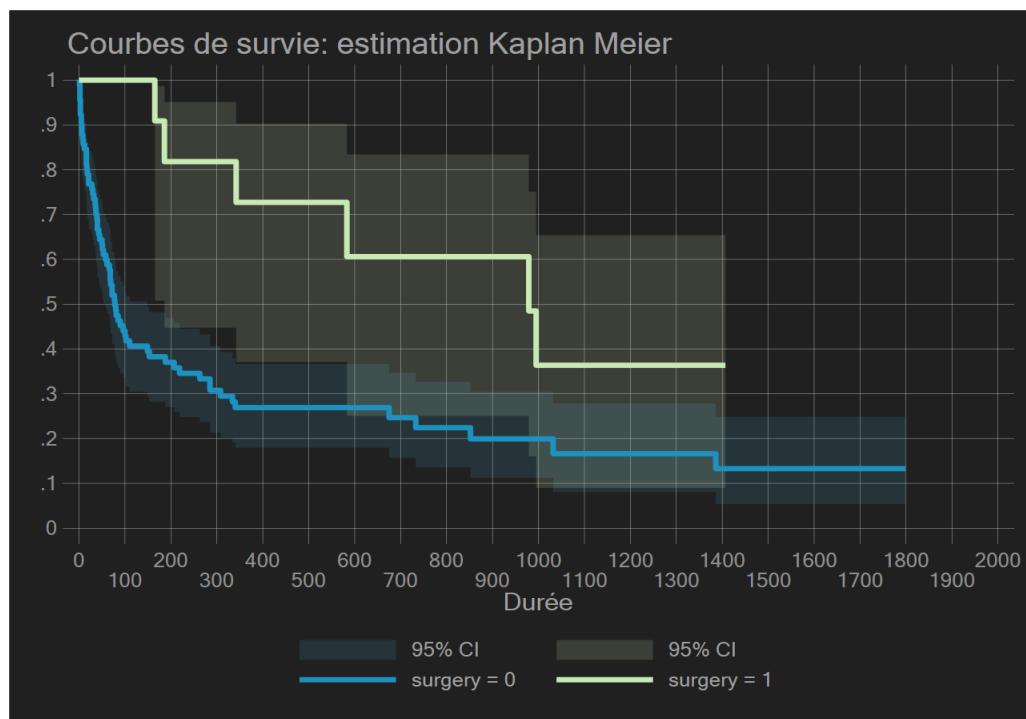
### **Les principaux tests de type log-rank**

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 test standard. Il est très sensible aux distribution des censures.
- **Test de Wilcoxon-Breslow-Gehan:**  $w_i = R_i$   
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 faible 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, mais moins qu'avec les deux versions précédentes. A utiliser lorsque le nombre de censures est élevé.
- **Test de Fleming-Harington:**  $w_i = (S_i)^p \times (1 - S_i)^q$  avec  $0 \leq p \leq 1$  Il permet de paramétriser le poids accordé au début ou à la fin de temps d'observation. Si  $p = q = 0$  on retrouve le test du log-rank.

### Exemple

On compare ici « l'effet » d'un 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
$\text{chi2}(1) = 6.59$ $\text{Pr}>\text{chi2} = 0.0103$		

#### 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
$\text{chi2}(1) = 8.99$ $\text{Pr}>\text{chi2} = 0.0027$			

#### 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



| surgery | Events observed | Events expected | Sum of 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 récupérer toutes les variantes.

**R:** on utilise la fonction `survdiff`. Le résultat du test de Peto-Peto est affiché par défaut (`rho=1`). Si on souhaite utiliser le test non pondéré, on ajoute l'option `rho=0`. En particulier pour les tests multiples (plus d'un degré de liberté), 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-Harrington). 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 échéant 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 (proportionnalité des risques, distribution des censures), et permet une lecture vivante basée sur des espérances de séjour et non sur la lecture d'une simple *p-value* traduisant l'homogénéité ou non des fonctions de séjour. Par ailleurs les comparaisons sont souples, on peut choisir un ou plusieurs points d'horizon pour alimenter l'analyse.

#### Principe

- L'aire sous la fonction de survie représente la durée moyenne d'attente 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, en termes de différence ou de ratio.
- Par défaut on définit généralement  $t^*$  à partir le temps du dernier évènement observé. Il est néanmoins possible de calculer 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)

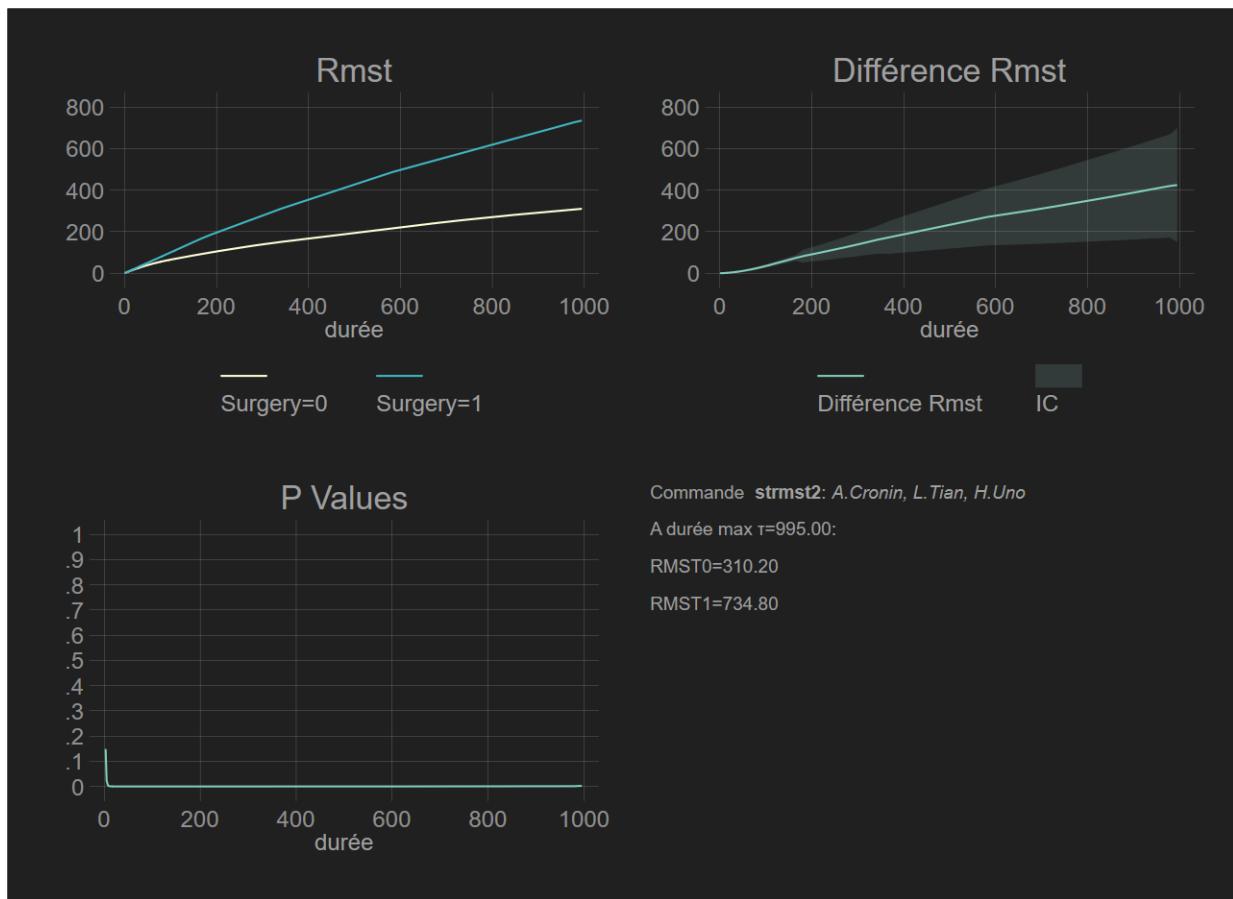
Contrast	Estimate	[95% Conf. Interval]	P> z
RMST (arm 1 - arm 0)	424.589	149.641 699.537	0.002
RMST (arm 1 / arm 0)	2.369	1.513 3.710	0.000

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	
	<b>370</b>	<b>332.0909</b>	<b>158.5728</b>	<b>173.5181</b>	<b>93.67896</b>	<b>253.3572</b>	<b>.0000205</b>	
	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	
	<b>995</b>	<b>734.7576</b>	<b>310.1689</b>	<b>424.5887</b>	<b>149.6407</b>	<b>699.5366</b>	<b>.0024726</b>	



Si on se limite à 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:

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

- $h(t)$  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 assure la positivité du risque.

### Le risque de base

- $h(t) = h_0(t)$  donc  $e^{X'b} = 1$
- Observations pour lesquelles  $X = 0$
- Il joue donc le rôle de constante

### 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:

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

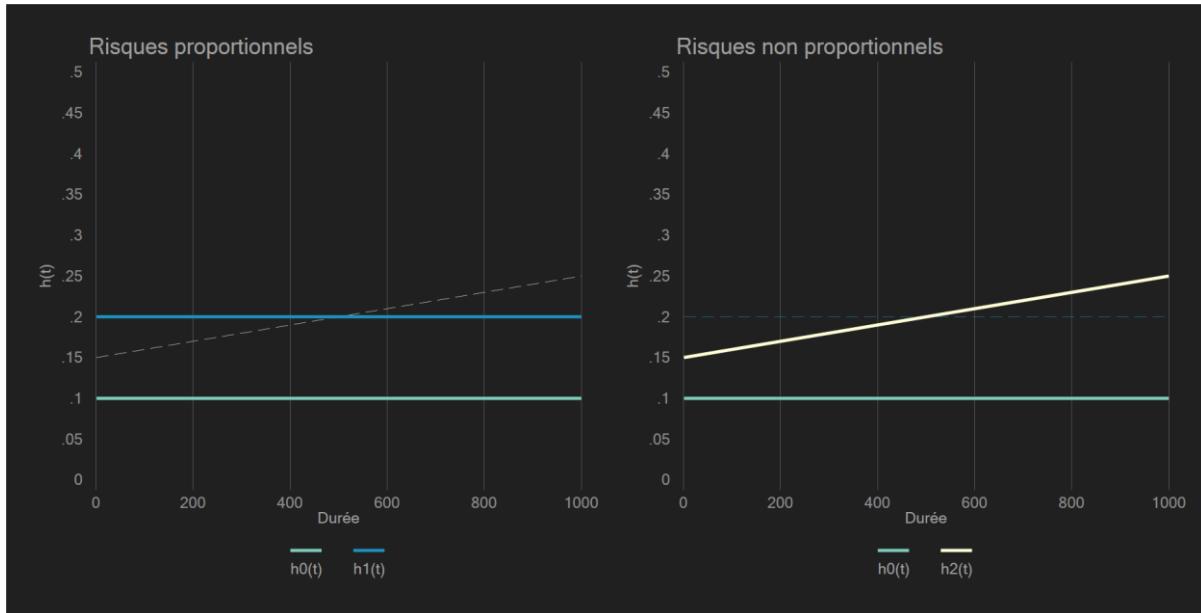
Pour une caractéristique binaire:  $X_A = 1$  et  $X_B = 0$ :

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

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

proportionnalité des risques implique que toutes les observations partagent un profil homogène 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 des risques 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 des risques 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 risques =2).

### Les modèles

#### *Le modèle semi-paramétrique de Cox*

Le modèle estime directement les  $b$  indépendamment de  $h_0(t)$ , c'est pour cela qu'il est semi-paramétrique. Les rapports des risques ( $e^b$ ) 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 proportionnalité 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 est régi par une hypothèse de proportionnalité. Le non respect de l'hypothèse est moins critique car la baseline du « risque » est estimée simultanément. Il est comme son nom l'indique, particulièrement adapté au durées discrètes. Le modèle de Cox est une réponse à une possible difficulté dans l'ajustement du risque par une loi a priori (modèles paramétriques standards).

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 risk ratios (RR) en utilisant la populaire méthode des splines cubiques. Il est implémenté dans les logiciels standards (R, Stata, Sas). Les RR 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

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

## 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 événement observé en  $t_i$  (idem pour les censures).

Pour une observation quelconque en  $t_i$  qui, pour uniquement un seul individu correspond à un événement ou une censure,, la vraisemblance peut s'écrire:

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

$f(t_i)$  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'événement est observé:  $L_i = f(t_i)$ .

$\delta_i = 0$  si l'observation est censurée:  $L_i = S(t_i)$ .

Comme  $f(t_i) = h(t_i) \times S(t_i)$ , on obtient:

$$L_i = [h(t_i)S(t_i)]^{\delta_i} S(t_i)^{1-\delta_i} = h(t_i)^{\delta_i} S(t_i).$$

Pour  $i = 1, 2, \dots, n$ , la vraisemblance totale s'écrit donc:  $L_i = \prod_{i=1}^n h(t_i)^{\delta_i} S(t_i)$ .

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

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

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

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

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, c'est-à-dire leur rang.

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

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

$$PL = \prod_{i=1}^n \left[ \frac{h_0(t_i)e^{X_i'b}}{\sum_{j \in R_i} h_0(t_j)e^{X_j'b}} \right]^{\delta_i} = \prod_{i=1}^n \left[ \frac{e^{X_i'b}}{\sum_{j \in R_i} e^{X_j'b}} \right]^{\delta_i}$$

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

### Condition nécessaire : pas d'évènement simultané

On rappelle que le temps est strictement continu, il ne doit pas y avoir d'évènement simultané.

Sinon, l'estimation de la vraisemblance doit 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 s'est produit avant B ou si B s'est produit avant A. Le nombre de permutations étant calculé par 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 seulement 10 évènements mesurés simultanément, le nombre de permutations est égal à 3.628.800 ( $10! = 10 \times 9 \times 8 \times 7 \times \dots \times 2 \times 1$ ). 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. 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 validité 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 seconde 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:

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

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

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

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

$$\frac{\partial \ln p(b)}{\partial b_k} = \sum_{i=1}^n \delta_i (X_{ik} - E(X_k))$$

Cette expression va permettre 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, **les rapports des risques sont constants pendant toute la période d'observation**. Il s'agit d'une propriété de l'estimation.

### Covariable binaire (indicatrice)

$$X = (0,1) \text{ et } RR = \frac{h(t|X=1)}{h(t|X=0)} = e^b$$

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).

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

Si  $c = 1$  (résultat de l'estimation): A un âge donnée en début d'exposition, le risque de connaître 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.

\* Par exemple si on s'intéresse à la durée d'une migration, il s'agira de l'âge à la migration.

## 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	<b>-0.1196</b>	0.0673	-1.78	0.076	-0.2516	0.0124
age	<b>0.0296</b>	0.0135	2.19	0.029	0.0031	0.0561
surgery	<b>-0.9873</b>	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 qu'un pontage réduit les risques journaliers de décès pendant la période d'observation (augmente la durée de survie).

De la même manière, plus on entre à un âge élevé dans la liste d'attente plus le risque de décès augmente. La variable year, qui traduit des progrès en médecine, implique une réduction relativement modérée du risque de décès durant l'attente de la greffe.

## Logiciels

SAS : le modèle est estimé avec la proc phreg .

Stata : Le modèle est estimé avec la commande stcox .

R : le modèle peut être estimé avec la fonction coxph de la librairie survival (a utilisé de préférence).

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 .

## L'hypothèse de constance des rapports de risque

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

## Tests sur les résidus de Schoenfeld

- Les résidus « bruts » sont directement calculés à partir des équations de scores (voir section estimation).
- Ce résidu n'est calculé que pour les observations qui ont observées l'événement.
- Il est calculé au moment où l'événement 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'événement s'est produit en  $t_i$ , le résidu brut de Schoenfeld pour la covariable  $X_k$ , après estimation du modèle, est égal à:

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

- 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 événement 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 bruts prendront toujours la valeur -.5 si  $X = 0$  et .5 si  $X = 1$ . En faisant une simple régression linéaire entre les résidus, qui alternent ces deux valeurs , et  $t$  , le coefficient estimé sera non significativement différent de 0.
- On peut 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			3	0.0327

Ici l'hypothèse de proportionnalité des risques est questionable pour la variable *surgery*. Le risque ratio n'est vraisemblablement 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 (ou tout simplement supérieur à 1). Le risque de premier espèce peut-être assez faible alors que les tests pour chaque covariables prises une à une présentent des risques élevées (>.1 par exemple). Dans l'exemple sur les trois variables, seulement une est questionable et le test sur un seuil courant de 5% pourrait être jugé significatif.

- Le test est 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 peut être utilisée pour effectuer le test. On retient généralement les fonctions suivantes:  $f(t) = t$  (« identity »),  $f(t) = \log(t)$ ,  $f(t) = KM(t)$  ou  $f(t) = 1 - KM(t)$  où  $KM(t)$  est l'estimateur de Kaplan-Meier. Enfin une transformation appelée « rank » utilisée seulement pour les durées strictement continue ou suffisamment dispersées. Par exemple  $t = (0.1, 0.5, 1, 2, 6, 3)$  donne une transformation  $t = (1, 2, 3, 4)$ . A savoir : la fonction « identity » rend le test relativement sensible aux événements 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, pour les autres logiciels des fonctions 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	
<hr/>						
_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 va alors 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 souvent d'usage d'utiliser  $f(t) = \log(t)$ , qui permet d'homogénéiser l'échelle entre la baseline et le terme d'interaction :

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

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 exprimera la variation constante de ce double 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	
<hr/>						
	_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
<hr/>						
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
<hr/>						
tvc [interaction]						
surgery	.0022277	.0011025	2.02	0.043	.0000669	.0043886
<hr/>						
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.

<i>t</i>	Calcul	Risk ratio
$0 + \epsilon$	$e^{-1.27+0.002 \times 0}$	0.28
1	$e^{-1.27+0.002 \times 1}$	0.281
2	$e^{-1.27+0.002 \times 2}$	0.282
.	.	
.	.	

10	$e^{-1.27+0.002 \times 10}$	0.286
100	$e^{-1.27+0.002 \times 100}$	0.34
365	$e^{-1.27+0.002 \times 365}$	0.58

### Important :

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

### Que faire 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 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 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 omises ou non observables. En analyse de durée ce problème prend le nom de **frailty** (fragilité) lorsque une non homogénéité n'est pas observables. 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 observables et disponibles.

#### *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, 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, dans l'idéal corriger une seule variable. Attention tout de même à la forme de la fonction, dans l'exemple on a contraint l'effet d'interaction à être linéaire, 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 à ce jour) 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.

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

## 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	
<hr/>							
_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. Il est donc pleinement paramétrique.
- Formellement, le modèle est estimable avec des événements 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. Néanmoins avec une bonne forme fonctionnelle de la durée traitée de manière continue, les deux modèles aboutissent à des résultats quasiment identiques.
- 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 ou *individus-périodes*: 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 *split* des données du modèle de Cox. Avec des données de type discrète, qui se traduisent par des nombres élevés d'évènement simultanés, classique en sciences sociales, il y a souvent peu de différence entre un allongement aux temps d'évènement et aux temps d'observation.

## Durée

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

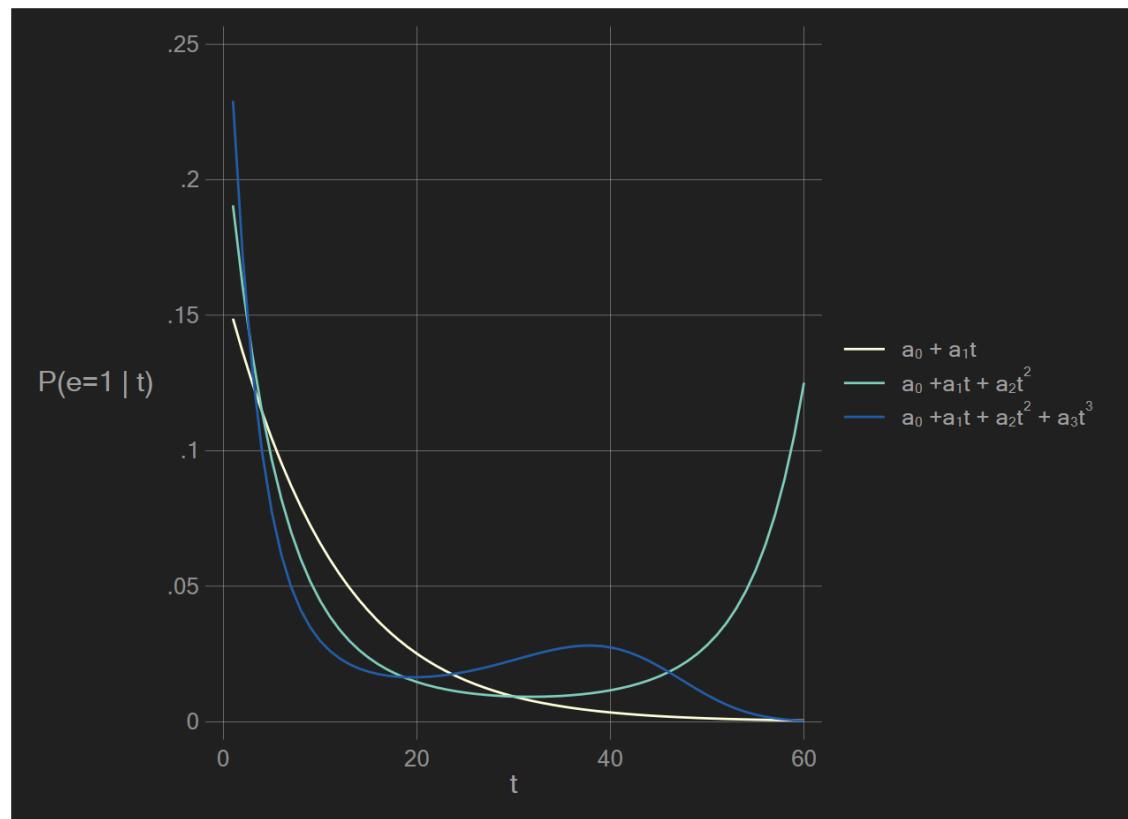
## Ajustement avec une durée en continu

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** (cubique) tendent à se développer ces dernières années, et présentent beaucoup d'avantages par rapport aux formes quadratiques. 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 formes suivantes :

- Une forme linéaire 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  $f(t) = a_1 \times t + a_2 \times t^2 + a_3 \times t^3$ .



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	

Remarque : La constante n'est pas reportée, les valeurs de la référence n'ayant pas grand sens (année et âge à 0). On peut centrer les deux variables par rapport à leur moyenne si on souhaite reporter une constante plus « parlante »

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 proches.

Remarque: il est tout à fait possible d'estimer ce modèle avec la durée mesurée en jour.

## Ajustement discret

- Il s'agit d'introduire la variable de durée dans le modèle comme une variable catégorielle.
- La méthode n'est pas conseillée avec beaucoup de points d'observation, ce qui est le cas dans l'application, avec en plus un nombre très (trop élevé) élevé de degrés de liberté.
- A l'inverse, si on ne dispose que peu de points d'observation, la paramétrisation avec une durée continue n'est pas un choix pertinent (et ne pas faire un modèle de 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.

quantiles of t	e		Total
	0	1	
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

*Remarque : La constante n'est pas reportée, les valeurs de La référence n'ayant pas grand sens (année et âge à 0). Préférer par exemple un centrage des variables quantitatives par rapport à Leur moyenne si on veut afficher La constante.*

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 probabilités. La lecture en termes d'Odds Ratio s'impose ici. Mais on peut utiliser les effets marginaux pour présenter les écarts en termes de points de %.

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

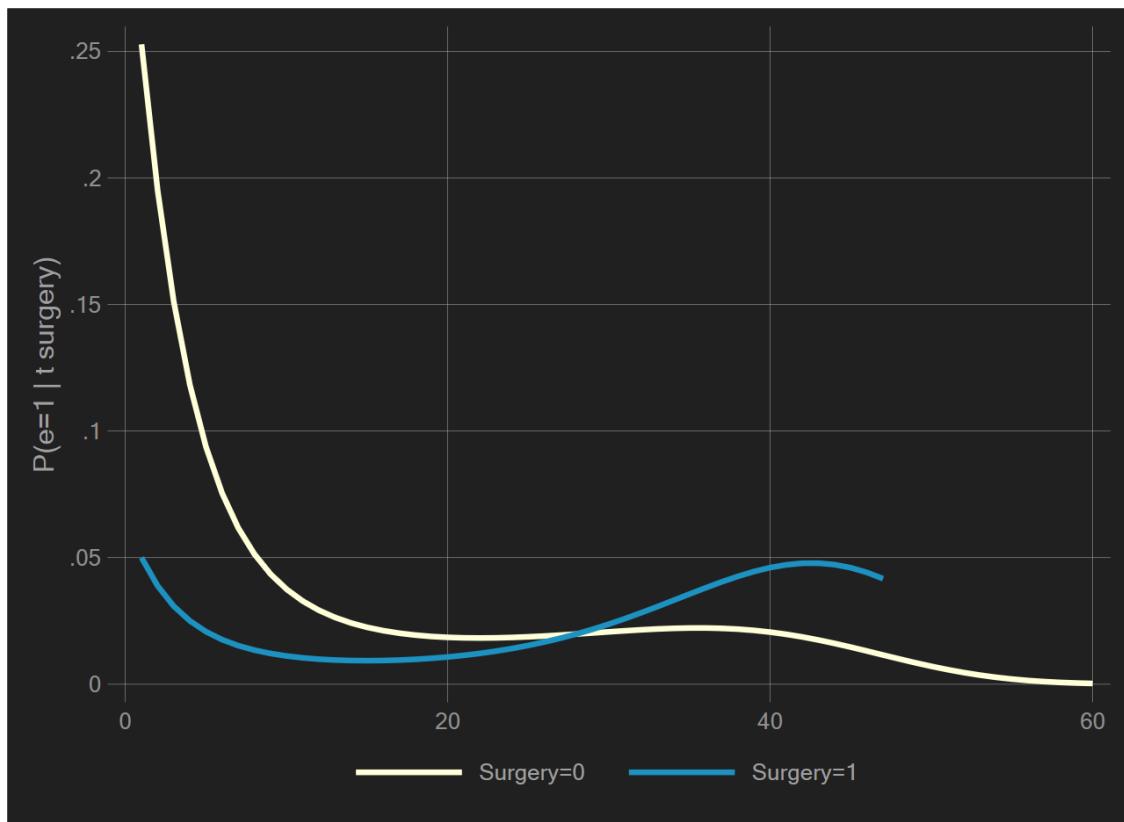
### Modèle à temps discret et hypothèse PH

- 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 proportionnalité 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				
<hr/>						
e	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
t	-.373826	.083913	-4.45	0.000	-.5382924	-.2093595
1.surgery	<b>-1.929061</b>	<b>.6920142</b>	<b>-2.79</b>	<b>0.005</b>	<b>-3.285383</b>	<b>-.5727377</b>
surgery#c.t						
1	<b>.0690069</b>	<b>.0333128</b>	<b>2.07</b>	<b>0.038</b>	<b>.003715</b>	<b>.1342987</b>
t2	.0137676	.0052405	2.63	0.009	.0034964	.0240388
t3	-.0001596	.0000828	-1.93	0.054	-.0003218	2.62e-06

*Remarque : La constante n'est pas reportée, les valeurs de la référence n'ayant pas grande sens (année et âge à 0)*



# Introduction de variables dynamiques

Cette section sera principalement traitée par l'exemple, et on ne s'intéressera qu'aux variables de type discrètes. Pour une variable dynamique quantitative on peut penser sur une durée annuelle à une variable de type revenu qui est susceptible de varier d'une année sur l'autre.

- 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 les sciences sociales, en omettant les effets d'anticipation. C'est une situation classique avec des covariables dynamiques de type discrètes. Les techniques standards ne peuvent modéliser que des effets d'adaptation : la cause - observée - précède l'effet.

## 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 (ou augmente la durée de survie).

On a dans la base 2 variables : une variable binaire pour savoir si l'individu a été greffé ou non, **transplant**, et une variable continue tronquée donnant la durée en jour jusqu'à la 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 : traité de manière fixe, la greffe réduit le risque journalier de décéder de près de  $-81\%: (1 - e^{1.65}) \times 100$

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, ici le jour exact de la greffe. C'est le même principe pour l'évènement, la personne est codée décédée (1) le jour du décès, et vivante avant (0).

## Estimation avec une variable dynamique

### 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. Quel que soit le logiciel, le principe de construction de la variable suit la logique suivante:

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

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

Maintenant 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	
<hr/>						
_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	<b>-0.27</b>	<b>0.786</b>	<b>-.6800165</b>	<b>.51468</b>

La greffe n'a plus d'impact significatif sur la survie des individus, tout du moins elle ne l'augmente pas. Cela ne signifie pas non plus que des personnes ont pu être « sauvée » grâce à cette opération (ou plutôt la durée de vie augmenté), mais des complications lors de l'opération ou post-opératoires, des rejets du greffon, à une époque où ces techniques étaient encore à leurs balbutiements, ont provoqué une mortalité peut-être un peu plus précoce (ici mesuré en jour il faut le rappelé).

## 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 (*survspli* avec R, *stspli* avec Stata) avant la création de la variable dynamique.

## Modèle à temps discret

Même principe pour la construction de la variable dynamique. Pour rappel l'échelle temporelle est le mois, et on a créé en amont une variable qui transforme les jours d'attente en mois (*mwait*)

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
						Pseudo R2
						= 0.1645
Log likelihood = -230.32152						

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	<b>-.0543011</b>	<b>.3114096</b>	<b>-0.17</b>	<b>0.862</b>	<b>-.6646528 .5560505</b>

Remarque : la constante n'est pas reportée, les valeurs de la référence n'ayant pas grand sens (année et âge à 0)

## Remarques (important)

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

# Compléments

## Modèles paramétriques

**Objectifs:** présenter (très/trop rapidement) la logique des modèles de type AFT, pour Accelerated Failure Time, principalement le modèle de Weibull (et exponentiel). Je n'ai pas forcément de pratique sur les modèles paramétriques, et à terme plutôt intéressé pour explorer de manière approfondie et présenter le modèle de Parmar-Royston.

### Les modèles paramétriques usuels

- Dans les modèles paramétriques, la durée de survie est distribuée selon une loi dont la densité  $f(t)$  s'exprime en fonction de paramètres (de la loi).
- Pour utiliser l'approche paramétrique, il faut avoir de bonnes raisons de penser que les temps de survie sont approximativement distribués selon une certaine loi connue plutôt qu'une autre.
- La majorité des distributions reposent sur une hypothèse dite **AFT (Accelerated Failure Time)**. Une seule repose seulement sur l'hypothèse PH (Gompertz), certaines peuvent selon la paramétrisation reposer sur les deux (exponentiel et Weibull).

### Hypothèse AFT: Accelerated Failure Time

L'hypothèse AFT signifie que l'effet des covariables est multiplicatif par rapport à la durée de survie. Par opposition, les modèles PH décrivent un effet multiplicatif par rapport au risque. Selon les caractéristiques des individus, *le temps ne s'écoulent pas à la même vitesse*, ils ne partagent plus la même métrique temporelle. Remarque: on a souvent des explications de type \*dilation/contraction\* du temps, par analogie à la théorie de la relativité.

Exemple populaire: la durée de vie d'un être humain et d'un chien. On dit qu'une année de vie d'un être humain équivaut à 7 années de vie d'un chien. C'est typiquement une hypothèse d'AFT.

$S_h(t) = S_c(7 \times t)$ . C'est ce facteur multiplicatif qu'estime un modèle paramétrique de type AFT.

$$S(t_i|X_1) = S(\phi t_i|X_0)$$

Remarque: si un modèle s'estime également sous hypothèse PH (ex Weibull):  
 $h(t_i|X_1) = -\rho \phi h(t_i|X_0)$

- Avantage: l'interprétation des modèles est directement liée aux fonctions de survie. Pratique après une analyse non paramétrique.
- Inconvénient: ne permet pas l'introduction de variables dynamiques.

Un être humain versus un chien: la probabilité qu'un être humain survive 80 ans est égale à la probabilité qu'un chien survive 11 ans (80/7). Le temps s'écoule donc plus vite pour le chien que pour l'être humain. Ce raisonnement peut s'appliquer aux quantiles du temps de survie: le temps de survie médian d'un être humain est 7 fois plus élevé que celui d'un chien. En terme d'interprétation des paramètres estimés, si le temps de survie est plus court le risque est plus élevé.

## Principe de construction des modèles AFT

Le raisonnement mathématique est ici bien plus complexe. On donnera juste quelques pistes en début de raisonnement.

On part d'une expression proche du modèle linéaire (à une transformation logarithmique près de la variable dépendante). En imposant la contrainte  $t_i > 0$ , en ne posant qu'une seule covariable  $X$  de type binaire, et en se situant de nouveau dans une logique de durée strictement continue (pas d'évènement simultané):

$$\log(t_i) = \alpha_0 + \alpha_1 X_i + b \times u_i$$

$b$  est un paramètre d'échelle identique pour toutes les observations et  $u_i$  un terme de terme d'erreur qui suit une loi de distribution de densité  $f(u)$ . C'est la forme de  $f(u)$  qui définit le type de modèle paramétrique.

Par simple réécriture:  $f(u_i) = f\left(\frac{\log(t_i) - \alpha_0 - \alpha_1 X_i}{b}\right)$

## Quelques lois de distribution

### Modèles de Weibull et modèle exponentiel

#### Weibull

- Peut estimer un modèle PH ou AFT, d'où sa popularité.
- Distribution monotone des temps d'évènement (toujours croissante/décroissante).
- $f(t) = \lambda \alpha t^{\alpha-1} e^{-\lambda t^\alpha}$  et  $h(t) = \lambda \alpha (\lambda t)^{\alpha-1}$ ,  $\alpha > 0$  et  $\lambda > 0$ . Le risque est croissant si  $\lambda > 1$ , décroissant si  $\lambda < 1$ , et est égal à la loi exponentielle si  $\lambda = 1$ .

#### Exponentiel

- Processus sans mémoire, utilisé pour étudier par exemple la durée de vie de composants électriques ou électroniques.

- La fonction de risque est une constante.
- Cas limite de la loi de Weibull. Un modèle de type exponentiel peut-être de type AFT ou PH.
- Pour contourner la stricte constance du risque dans le temps, on peut estimer un modèle en scindant la durée en plusieurs intervalles. Le risque sera constant à l'intérieur de ces intervalles, il s'agit d'un modèle "exponential piecewise" (exponentiel par morceau). Ce type de modèle est assez populaire, en particulier en démographie, il est estimable à partir d'un modèle de Poisson, et quasiment assimilable à un modèle de Cox.

### **Log-logistique**

- Estime un modèle de type AFT seulement. Proche du modèle log-normal (plus difficile à estimer).
- Permet une interprétation en terme d'Odds de survie.
- La fonction du risque peut être "U-shaped" (unimodale croissante puis décroissante).

**Autres lois:** Gompertz (risques proportionnel seulement), Gamma et Gamma généralisé..... Le modèle de Gompertz a été beaucoup utilisé en démographie, en particulier dans l'analyse de la mortalité.

**Sélection de la loi** On peut sélectionner la loi en comparant les AIC où les BIC des modèles. Pour le modèle de Weibull, on peut voir s'il peut bien ajuster les données en regardant la linéarité de la transformation  $\log(-\log(S(t_i)))$  par rapport à  $\log(t_i)$ .

### **Exemple avec les malformations cardiaques**

Estimation sans covariables

#### **Comparaison des AIC**

Weibull: 400.1

Exponentiel: 461.0

Gompertz: 409.6

*Log-logistique:* 391.8

### **Exemple avec le modèle de Weibull**

Ce n'est pas le meilleur modèle (cf AIC du modèle log-logistique), mais il est assez standard, et peut être paramétré sous risques proportionnels, approche décrites précédemment.

## AFT

### Weibull AFT regression

No. of subjects =	103	Number of obs	=	103
No. of failures =	75			
Time at risk =	31938			
		LR chi2(3)	=	18.87
Log likelihood =	-188.6278	Prob > chi2	=	0.0003
<hr/>				
_t	Coef.	Std. Err.	z	P> z  [95% Conf. Interval]
year	0.1620	0.1218	1.33	0.184 -0.0768 0.4008
age	-0.0615	0.0247	-2.49	0.013 -0.1100 -0.0130
surgery	1.9703	0.7794	2.53	0.011 0.4427 3.4980
_cons	-3.0220	8.7284	-0.35	0.729 -20.1294 14.0854
/ln_p	-0.5868	0.0927	-6.33	0.000 -0.7685 -0.4051
p	0.5561	0.0516		0.4637 0.6669
1/p	1.7983	0.1667		1.4995 2.1566

Une journée de survie d'une personne qui n'a pas été opérée d'un pontage correspond à 7 jours (soit une semaine) de survie d'une personne opérée ( $e^{1.9703}$ ). Cette remise à l'échelle de la métrique temporelle entre les deux groupes exprime bien le gain en durée de survie médiane pour les personnes opérées, soit des risques journaliers de décès plus faibles (et plus faibles à valeurs constantes, proportionnalité oblige).

## PH

### Weibull PH regression

No. of subjects =	103	Number of obs	=	103
No. of failures =	75			
Time at risk =	31938			
		LR chi2(3)	=	18.87
Log likelihood =	-188.6278	Prob > chi2	=	0.0003
<hr/>				
_t	Coef.	Std. Err.	z	P> z  [95% Conf. Interval]
year	-.0900736	.0663972	-1.36	0.175 -.2202097 .0400626
age	.034214	.0138509	2.47	0.014 .0070667 .0613613
surgery	-1.095685	.4341312	-2.52	0.012 -1.946566 -.2448033
_cons	1.680511	4.823645	0.35	0.728 -7.77366 11.13468
/ln_p	-.5868247	.0927049	-6.33	0.000 -.768523 -.4051264
p	.5560902	.0515523		.4636974 .6668925
1/p	1.798269	.1667084		1.499492 2.156579

Remarque:  $b_{ph} = -\rho \times b_{aft}$ . Ici  $-0.556 \times (1.97) = -1.096$

### Modèle de Cox précédemment estimé

Cox regression -- Breslow method for ties						
No. of subjects =	103			Number of obs	= 103	
No. of failures =	75					
Time at risk =	31938			LR chi2(3) =	17.56	
Log likelihood =	-289.54474			Prob > chi2 =	0.0005	
<hr/>						
_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
surgery	-.984869	.4362881	-2.26	0.024	-1.839978	-.1297601
<hr/>						

**Attention:** on ne peut pas comparer la qualité d'un modèle paramétrique à celle d'un modèle de Cox par des critères type AIC ou BIC. Les deux méthodes d'estimation sont totalement différentes.

## Risques concurrents

Le problème des événements multiples dans les analyses de survie a été posée dans les années 1970 avec la notion de « **risques concurrents** » (*competing risks*) : il s'agit d'événements survenant pendant la période d'observation et qui « empêchent » l'occurrence de l'évènement d'intérêt.

- On étudie un processus dont l'occurrence peut être scindée en plusieurs modalités, « causes » ou « types »: la mortalité par cause de décès, les types de sortie du chômage (formation, emploi, radiation), les types de sortie de l'emploi (chômage, longue maladie, sortie du marché du travail hors retraite), les lieux de migration ou les espaces de mobilité résidentielle, type de rupture d'union (séparation-divorce, veuvage). Déjà abordé dans la partie théorie, avec un recueil de données de type prospectif des « perdu.e.s de vue » peuvent difficilement être assimilés à des sorties d'observation non informatives (censures).
- L'analyse des risques concurrents est un cas particulier des modèles « multi-états » avec différents risques considérés comme absorbants.
- En présence de risques concurrents, l'estimation de Kaplan-Meier ne peut se faire que sous **l'hypothèse d'indépendance entre chacun des risques**. Sinon l'estimateur n'est plus une probabilité. Il n'est pas possible de mesurer cette hypothèse.
- Une estimation de type KM d'un évènement en concurrence avec d'autres impose que ces derniers soient traités comme des censures à droites non informatives.
- En terme de lecture des résultats, on utilise rarement les fonctions de survie. On survit ou non à un évènement, mais il est en revanche peu intuitif de dire qu'on survit à une cause.

## Risques « cause-specific » et biais sur les estimateurs KM

- Si les risques ne sont pas indépendants les uns par rapport aux autres, la somme des estimateurs de  $(1-KM)$  pour chaque risque n'est pas égale - elle est **supérieure** - à l'estimateur de  $(1-KM)$  où les risques concurrents sont regroupés en un évènement unique (par exemple le décès).

- Le risque calculé en considérant les risques concurrents comme des censures à droite est appelé « **cause-specific risk** ».

Pour le risque de type  $k$ , le risque *cause-spécifique* (traduction?) en  $t_i$  est égal à:

$$h_k(t_i) = \frac{d_{i,k}}{R_i}$$

Où  $d_{i,k}$  est le nombre d'évènement de type  $k$  survenu en  $t_i$  et  $R_i$  la population soumise en  $t_i$ .

Conséquence: si les risques ne sont pas indépendants, la fonction de survie estimée avec la méthode Kaplan Meier n'exprime plus une probabilité.

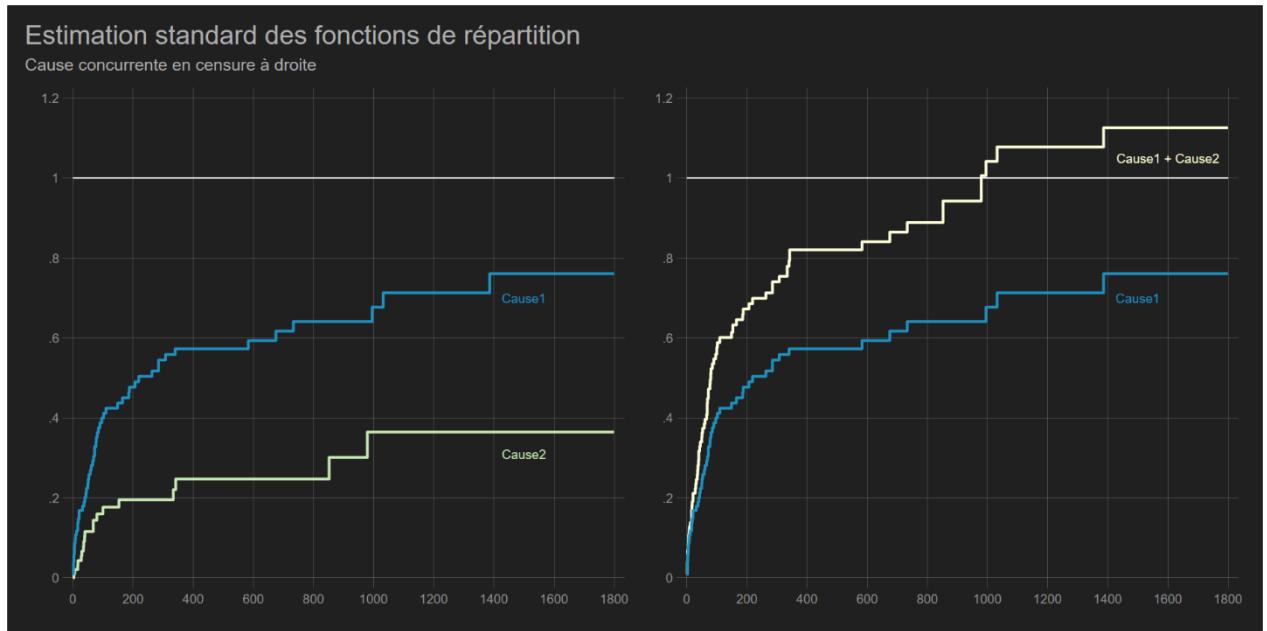
#### **Exemple sur les décès causés par une malformation cardiaque aigüe**

Dans la base d'origine, il n'y a pas directement cette dimension de risque concurrent, même si on trouve dans la littérature médicale des études prenant le décès rapide post greffe comme un risque de ce type. Les données étant assez anciennes, avec beaucoup de décès post-opératoire, je ne me suis pas « risquer » à générer directement un risque concurrent. Une sortie concurrente a donc été simulée sans plus de précision (cause2), que l'on considérera non strictement indépendante à la cause d'intérêt. Ce risque entre en concurrence avec la cause du décès directement liée à la malformation cardiaque, que la personne ait été transplanté ou non.

		Survival Status (1=dead)		Total
compet		0	1	
0		28	0	28
1		0	56	56
2		0	19	19
Total		28	75	103

Lorsqu'on a analysé le décès par la méthode KM, la proportion de survivant.e.s était de 15%.

Si on applique la méthode de Kaplan Meier à la cause 1 en traitant la cause 2 comme une censure à droite, et à la cause 2 en traitant la cause 1 comme une censure à droite, puis en sommant les deux estimateurs, la fonction de répartition excède 100% au bout de 1000 jours environ. La proportion de survivant.e.s est donc négative.



## Estimations en présence de risques concurrents

### Estimation non paramétrique

- Utiliser l'estimateur de Nelson Aalen: il s'agit du risque instantané cumulé. Comme il ne s'agit pas d'une probabilité, il a été longtemps utilisé comme mesure de l'incidence en présence d'au moins un risque concurrent dans une logique dite « cause spécifique »:

$$H_k(t_i) = \sum_{t_i \leq t} \left( \frac{e_{i,k}}{n_i} \right)$$

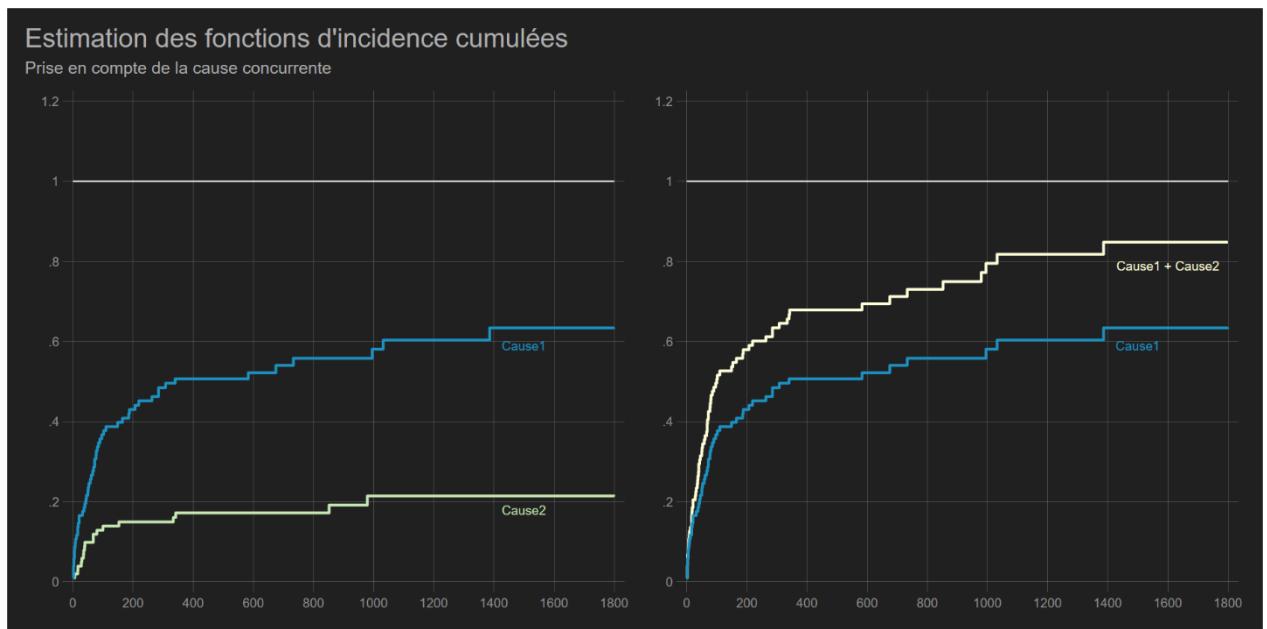
- De nos jours, l'estimateur le plus utilisé est la fonction dite **d'incidence cumulée - CIF** - (Kalbfleisch-Prentice, Marubini-Valscchi):
  - Il repose sur une probabilité tout en supportant la non indépendance des risques.
  - Son interprétation est identique à la fonction de répartition  $F(t) = 1 - S(t)$ . Cette fonction est donc croissante.
  - Il est possible de tester les différences entre CIF: test de Gray ou test de Pepe-Mori.

### La fonction d'incidence cumulée (CIF)

- Si  $h_k(t_i)$  est le risque « cause-spécifique » en  $t_i$  et  $S(t_i - 1)$  l'estimateur de Kaplan-Meier en  $t_i - 1$  lorsque tous les risques sont regroupés en un événement unique, la fonction d'incidence cumulée pour le risque  $k$  en  $t_i$  est égale à:

$$IC_k(t_i) = \sum_{t_i \leq t} S(t_i - 1) h_k(t_i)$$

- Les valeurs prises par cette fonction en  $t_i$  ne dépendent donc pas seulement des individus ayant observé l'évènement à partir de cette seule cause, mais aussi du nombre de personnes qui n'ont pas encore observés l'évènement à partir des autres causes identifiées. Cette dernière information est donnée par  $S(t_i - 1)$ . Cela permet à la grandeur de conserver une propriété de probabilité.
- L'incidence cumulée peut ainsi s'interpréter simplement comme la proportion d'individus qui sont sortis du risque jusqu'en  $t_i$  en raison de la cause  $k$ .



```

failure:  competit == 1
competing failures:  competit == 2

      Time      CIF        SE    [95% Conf. Int.]
-----
```

Time	CIF	SE	[95% Conf. Int.]
1	0.0097	0.0097	0.0009 0.0477
2	0.0194	0.0136	0.0038 0.0619
3	0.0485	0.0212	0.0181 0.1022
5	0.0680	0.0248	0.0300 0.1273
6	0.0874	0.0278	0.0429 0.1515
8	0.0971	0.0292	0.0497 0.1634
9	0.1068	0.0304	0.0566 0.1751
12	0.1166	0.0316	0.0638 0.1868
16	0.1264	0.0328	0.0711 0.1984
18	0.1362	0.0338	0.0785 0.2099
21	0.1559	0.0358	0.0937 0.2325
30	0.1657	0.0367	0.1014 0.2437
32	0.1756	0.0376	0.1093 0.2550
35	0.1856	0.0384	0.1173 0.2662
37	0.1955	0.0392	0.1253 0.2773

39	0.2055	0.0400	0.1335	0.2884
40	0.2156	0.0407	0.1418	0.2996
45	0.2256	0.0414	0.1502	0.3107
50	0.2357	0.0421	0.1586	0.3217
53	0.2458	0.0427	0.1671	0.3327
58	0.2559	0.0433	0.1757	0.3436
61	0.2660	0.0439	0.1843	0.3544
66	0.2761	0.0445	0.1930	0.3652
68	0.2861	0.0450	0.2018	0.3759
69	0.2962	0.0454	0.2106	0.3866
72	0.3063	0.0459	0.2195	0.3973
77	0.3164	0.0463	0.2284	0.4079
78	0.3265	0.0467	0.2374	0.4184
81	0.3365	0.0471	0.2464	0.4289
85	0.3466	0.0474	0.2554	0.4393
90	0.3567	0.0478	0.2645	0.4497
96	0.3668	0.0481	0.2737	0.4601
100	0.3769	0.0484	0.2829	0.4704
102	0.3870	0.0486	0.2921	0.4807
110	0.3972	0.0489	0.3016	0.4911
149	0.4078	0.0491	0.3112	0.5019
153	0.4183	0.0494	0.3209	0.5125
186	0.4291	0.0496	0.3309	0.5235
188	0.4399	0.0498	0.3409	0.5343
207	0.4506	0.0500	0.3509	0.5451
263	0.4614	0.0502	0.3610	0.5559
285	0.4836	0.0505	0.3818	0.5780
308	<b>0.4947</b>	<b>0.0506</b>	<b>0.3923</b>	<b>0.5890</b>
340	0.5058	0.0507	0.4028	0.5999
583	0.5211	0.0513	0.4162	0.6158
733	0.5391	0.0524	0.4313	0.6351
852	0.5584	0.0535	0.4475	0.6555
995	0.5811	0.0550	0.4657	0.6801
1032	0.6039	0.0561	0.4850	0.7036
1386	0.6343	0.0584	0.5084	0.7362

```

failure:  competit == 2
competing failures:  competit == 1

```

Time	CIF	SE	[95% Conf. Int.]	
2	0.0194	0.0136	0.0038	0.0619
16	0.0391	0.0191	0.0128	0.0897
17	0.0489	0.0213	0.0182	0.1029
28	0.0587	0.0232	0.0240	0.1157
36	0.0686	0.0250	0.0302	0.1286
40	0.0787	0.0267	0.0368	0.1413
43	0.0888	0.0283	0.0436	0.1539
51	0.0989	0.0297	0.0506	0.1663
68	0.1090	0.0310	0.0578	0.1785
72	0.1190	0.0323	0.0651	0.1905
80	0.1291	0.0334	0.0726	0.2024
165	0.1396	0.0346	0.0804	0.2149
219	0.1504	0.0358	0.0886	0.2276
334	0.1615	0.0370	0.0970	0.2406
342	0.1730	0.0383	0.1058	0.2540
675	0.1910	0.0414	0.1177	0.2777
979	0.2138	0.0457	0.1321	0.3086

En présence du risque concurrent, et traité comme tel, 50% des personnes sont décédées directement de la malformation cardiaque au bout de 308 jours (200 jours avec une estimation de type « cause specific »).

On peut vérifier que la somme des estimateurs permet d'obtenir la survie toutes causes confondues. Il n'y a pas de surprise à cela, dans l'estimateur Marubini-Valscchi la survie d'ensemble intervient comme un facteur de pondération de la mesure d'intensité dite « cause-specific ».

## Logiciels

L'estimation avec des risques de type « cause-specific » demande juste de recoder la variable évènement/censure, en glissant les risques concurrents en censure à droite.

**Sas** : maintenant directement estimable avec **proc lifetest**. Il suffit d'indiquer le ou les risques d'intérêt dans l'instruction précisant la variable de durée et de censure avec l'option **failcode=valeur**.

**Stata** : Estimation avec la commande externe **stcompet**. La commande génère des variables qui demande des manipulations supplémentaires pour afficher les résultats sous forme de tableau par exemple. On peut utiliser la commande externe **stcomlist** pour afficher directement les résultats.

**R** : la librairie **cmprsk** permet d'estimer simplement les incidences cumulées avec la fonction **cuminc**.

**Python** : le wrapper de R (cmprsk) ne fonctionne plus à ce jour à défaut de mise à jour.

## Comparaison des courbes

- Test d'homogénéité de **Gray**: est basé sur une autre mesure du risque en évènement concurrent. Il s'agit du « subdistribution risks » (« risque de sous-répartition », A.Latouche). Son interprétation n'est pas aisée car les personnes ayant observé un risque concurrent sont remises dans le Risk Set. Mais il est directement lié à l'estimation des CIF. Disponible avec SAS, R et Python. Il est également sensible l'hypothèse de proportionnalité et à la distribution des censures à droites entre les groupes comparés. A ma connaissance il n'y a pas de variantes pondérées.
- Test de **Pepe & Mori**: teste directement deux courbes d'incidences et seulement 2. Il y a une version du test qui repose sur une autre fonction directement tirée des incidences, la CPF (Conditional Propability Function), que je ne traite pas afin de ne pas rajouter un concept de supplémentaire.

- On peut toujours tester les différences entre les risques « causes spécifiques » avec un test du log-rank. Mais on ne teste pas des différences entre CIF. Déconseillé.

### Test de Gray pour la variable surgery

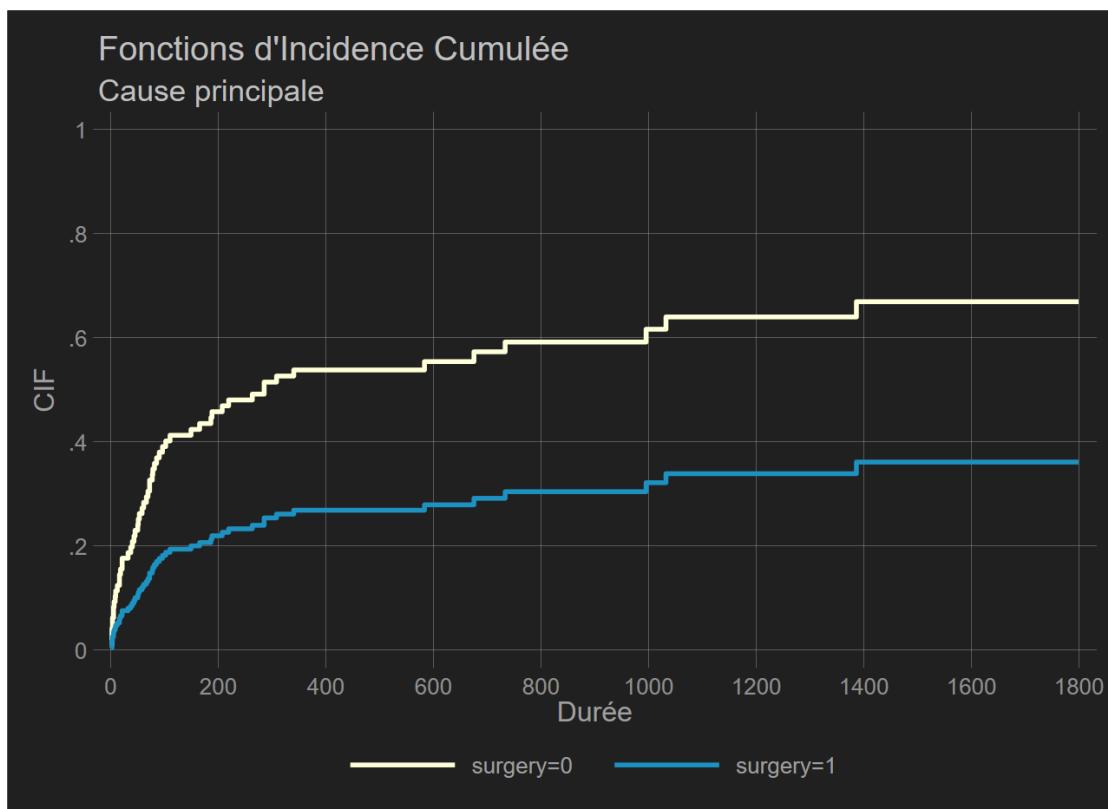
Tests:

	stat	pv	df
cause1	5.7834605	0.0161	1
cause2	0.1293076	0.7191	1

### Test de Pepe-Mori pour la variable surgery

```
Main event failure:  compet == 1
Chi2(1) = 6.2028 - p = 0.01275
```

```
Competing event failure:  compet == 2
Chi2(1) = 1.8796 - p = 0.17038
```



## Logiciels

**Sas :** le test de Gray est estimé si on ajoute l'option `strata=nom_variable` à la proc `lifetest` sous risque concurrent (voir encadré précédent). Le test de Pepe-Mori est disponible via une macro externe (%`compcif`) : <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470870709.app2>

**Stata:** Le test de Gray n'est pas disponible, il faut passer par une exécution de la fonction `cuminc` de la librairie R `cmprsk` directement dans stata (voir la commande `rsource`). Pour faire plus simple, on peut estimer le modèle de Fine-Gray avec une seule variable (discrète). Le résultat est comparable à celui du test (voir plus bas). Le test de Pepe-Mori est disponible via la commande externe `stpepmori`.

**R :** On ajoute une variable à la fonction `cuminc` de la librairie `cmprsk`. Pas de test de Pepe-Mori sur les fonctions d'incidence à ma connaissance.

**Python :** Ne pas essayer d'utiliser la librairie `cmprsk` qui n'est pas mis à jour et ne fonctionne plus.

## Modèles semi paramétrique et à temps discret

Cette présentation sera plutôt brève. Dans le domaine des sciences sociales, je préconise plutôt l'utilisation d'un modèle multinomial à temps discret de type logistique. Le modèle de Cox en présence de risques concurrent n'est valable que dans une logique de risques « cause-specific », le modèle de Fine et Gray bien que directement relié à l'estimation des incidences cumulées, repose sur une définition du risque (de sous répartition) dont l'interprétation n'est pas naturelle. Il est également soumis à l'hypothèse de proportionnalité des risques. Il est globalement très critiqué (cf blog d'Allison).

### Modélisation des risques « cause-specific » : Cox

Modèle de Cox «standard» pour chaque événement, les évènements concurrents sont traités comme des censures à droite. Aucune interprétation sur les fonctions d'incidence ne peut-être faite.

### Modèle de Fine-Gray: subdistribution hazard regression

Modèle de type semi-paramétrique avec une redéfinition du risque lié à l'estimation des fonctions d'incidence (voir test de Gray). La différence avec le Cox classique réside dans le calcul du risk-set : les évènements concurrents ne sont pas considérés comme des censures, on laisse les individus leur « survivre » jusqu'à la durée maximale observée dans l'échantillon. L'interprétation n'est donc pas très intuitive (Fine et Gray le soulignent). Ce modèle est relativement controversé.

Pour les questions liées à l'interprétation de ces deux types de modèles, se reporter à: <https://onlinelibrary.wiley.com/doi/epdf/10.1002/sim.7501>

## Logiciels (modèle de Fine & Gray)

**Sas:** même principe que pour l'estimation non paramétrique, on ajoute l'option `eventcode=valeur` à l'instruction `model` de la proc `phreg`.

**Stata:** on utilise la commande interne `stcrreg`.

**R:** on utilise la fonction `crr` du package `cmprsk`.

**Python :** ne pas essayer d'utiliser la librairie `cmprsk` qui n'est pas mis à jour et ne fonctionne plus.

## Modèle à temps discret

- Il s'agit d'une extension du modèle à temps discret à évènement unique (toutes causes regroupées) avec ici le **modèle logistique multinomial**.
- S'il ne permet pas une interprétation sur les fonctions d'incidences, les risques concurrents ne sont pas traitées comme des censures à droite et sont estimés simultanément à la cause d'intérêt.
- Le modèle multinomial repose sur une hypothèse dite « d'indépendance » des alternatives non pertinentes (IIA). Cela peut donc paraître contradictoire d'utiliser ce modèle pour des évènements qui sont supposés non indépendants. Néanmoins la dépendance entre risques concurrents n'est pas non plus stricte. L'hypothèse d'IIA est souvent illustrée par l'exemple des couleurs des bus dans le choix du mode de transport (G. Debreu). On est loin ici d'un tel niveau de dépendance.
- En terme de lecture, le modèle logistique multinomial les estimateurs peuvent directement s'interpréter comme des rapports de risque (ou relative risk ratio). Ceci est l'objet d'une petite controverse liée à la lecture en Odds Ratio dans un modèle multinomial, qui devrait être strictement réservé pour certain.e.s à un signal binaire.
- En sciences sociales, il me semble que ce type de modèle soit à privilégier.
- On peut également envisager un modèle de type probit multinomial, mais on peut rencontrer des problèmes d'estimations (repose sur la loi normale multivariée). Prévoir un regroupement des causes concurrentes, et dans tous les cas de figure ne pas dépasser trois causes. Niveau lecture, il conviendra d'utiliser une méthode de standardisation, de type « effets marginaux ».

Pour l'exemple, j'ai utilisé comme plus haut pour les modèles à temps discret à évènement unique, le mois comme métrique temporelle.

Multinomial logistic regression	Number of obs	=	1,127
	LR chi2(10)	=	86.25
	Prob > chi2	=	0.0000
Log likelihood = -275.00542	Pseudo R2	=	0.1356

e	RRR	Std. Err.	z	P> z	[95% Conf. Interval]
<hr/>					
0   (base outcome)					
<hr/>					
1					
t   0.8159 0.0338 -4.91 0.000 0.7522 0.8850					
t2   1.0032 0.0009 3.53 0.000 1.0014 1.0049					
age   1.0449 0.0183 2.51 0.012 1.0097 1.0813					
year   0.8795 0.0718 -1.57 0.116 0.7494 1.0321					
surgery   0.3175 0.1711 -2.13 0.033 0.1104 0.9129					
<hr/>					
2					
t   0.8168 0.0565 -2.93 0.003 0.7134 0.9353					
t2   1.0030 0.0015 1.94 0.052 1.0000 1.0060					
age   1.0111 0.0248 0.45 0.654 0.9635 1.0610					
year   0.8158 0.1127 -1.47 0.141 0.6223 1.0695					
surgery   0.5412 0.4221 -0.79 0.431 0.1173 2.4959					

Remarque : les constantes ne sont pas reportées, les valeurs de la référence n'ayant pas grand sens (année et âge à 0)

## Fragilité et immunité

A faire un jour.... mais le plus rapidement sera le mieux.

Quelques remarques tout de même.

Pour la fragilité (« frailty »), lire la dernière section du document de travail de Simon Quantin, je n'ai pas vu de meilleure présentation du problème que la sienne. Très important, car une des sources de la non proportionnalité des risques se trouvent dans l'omission de variables. Ici on va être confronté une omission sur des traits non observables. Certaines de ces caractéristiques vont « accélérer » dès le début de la période d'exposition la survenue de l'évènement. L'introduction d'un facteur de fragilité se fait par l'introduction d'un effet aléatoire dans le modèle, de nature plus complexe, reposant sur des hypothèses fortes et rendant l'interprétation des modèles plus compliquée (on doit passer par des graphiques, le tableau de régression ne pouvant présenter des résultats en début d'exposition).

Pour l'immunité, qui est un cas particulier du précédent, on va interroger l'exposition au risque d'une partie observation. Visuellement on peut se poser cette question lorsque la courbe de séjour ne tend pas vers 0 mais présente une longue asymptote sur une valeur nettement supérieure à 0 :  $\lim_{t \rightarrow \infty} S(t) = a$ . On trouve déjà des applications en démographie en analyse de la fécondité (analyse des naissances de rang supérieur à 1). Cette problématique affecte les modèles de durées avec des évènements récurrents (non présentés). Pour traiter cette question, les modèles peuvent être de type mixte (probabilité d'être immunisée associée à un modèle de durée) ou non mixte de type bayésien. Il n'y a pas de méthode unifiée à ce jour, elles reste très dépendante du champ d'analyse d'origine.

# Applications logiciels

## SAS

### *Remarque: Sélection des outputs*

Selon le type d'analyse la totalité des outputs ne seront pas reproduits (*ods include* ou *ods exclude* pour la sélection). Un problème spécifique s'observe pour le tableau des estimateurs de Kaplan-Meier qui est particulièrement illisible en présence d'un nombre important d'observations censurées.

Exemple pour *proc lifetest*: noms des outputs récupérés dans la log

Output Added:

```
-----  
Name:      ProductLimitEstimates  
Label:     Product-Limit Estimates  
Template:  Stat.Lifetest.ProductLimitEstimates  
Path:      Lifetest.Stratum1.ProductLimitEstimates  
-----
```

Output Added:

```
-----  
Name:      Quartiles  
Label:    Quartiles of the Survival Distribution  
Template:  Stat.Lifetest.Quartiles  
Path:      Lifetest.Stratum1.TimeSummary.Quartiles  
-----
```

Output Added:

```
-----  
Name:      Means  
Label:    Mean  
Template:  Stat.Lifetest.Means  
Path:      Lifetest.Stratum1.TimeSummary.Means  
-----
```

Output Added:

```
-----  
Name:      SurvivalPlot  
Label:    Survival Curve  
Template:  Stat.Lifetest.Graphics.ProductLimitSurvival  
Path:      Lifetest.Stratum1.SurvivalPlot  
-----
```

Output Added:

```
-----  
Name:      CensoredSummary
```

Label:	Censored Summary
Template:	Stat.Lifetest.CensoredSummary
Path:	Lifetest.CensoredSummary

Utiliser de préférence le nom figurant dans la ligne **path**: (si comparaison de deux strates, le nom figurant dans la ligne **name** est identique).

### Récupération de la base

On peut télécharger la base d'exemple à cette adresse (onglet download ou cliquer sur view raw) :

[https://github.com/mthevenin/analyse\\_duree/blob/main/bases/trans.sas7bdat](https://github.com/mthevenin/analyse_duree/blob/main/bases/trans.sas7bdat)

## Analyse non paramétrique

### *Méthode actuarielle*

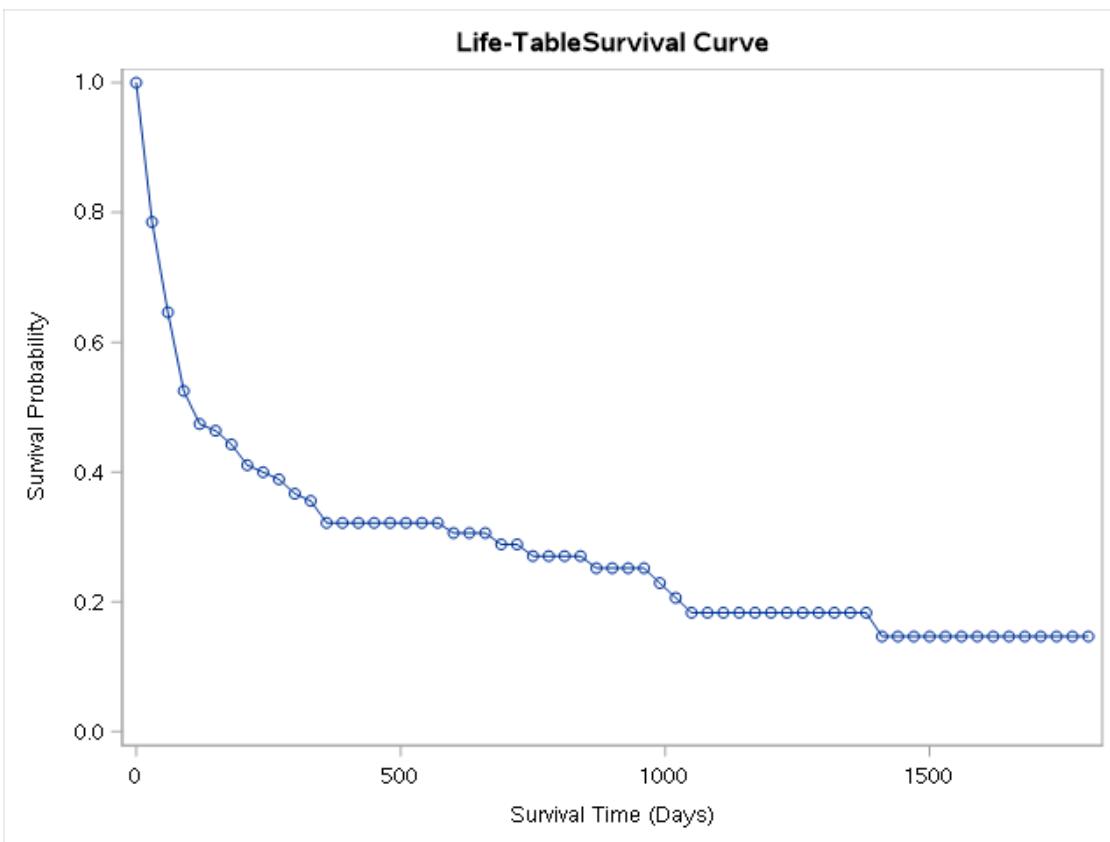
Avec une longueur d'intervalle fixe égale à 30 jours.

La durée médiane est donnée par la colonne **résidual median time**. Sur la première ligne, il s'agit de la durée médiane sur toutes les personnes exposées au risque. Dans les lignes suivantes, cette durée médiane est recalculée pour les personnes restant exposées au risque dans chaque intervalle.

```
proc lifetest data=trans method=lifetable width=30;
time stime*died(0);run;
```

Life Table Survival Estimates												
Interval		Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival	Failure	Survival Standard Error	Median Residual Lifetime	Median Standard Error	
[Lower	Upper)											
0	30	22	1	102.5	0.2146	0.0406	1.0000	0	0	104.7	29.0660	
30	60	14	2	79.0	0.1772	0.0430	0.7854	0.2146	0.0406	229.4	121.0	
60	90	12	0	64.0	0.1875	0.0488	0.6462	0.3538	0.0475	298.7	35.7872	
90	120	5	1	51.5	0.0971	0.0413	0.5250	0.4750	0.0498	762.7	58.8416	
120	150	1	1	45.5	0.0220	0.0217	0.4741	0.5259	0.0499	859.3	46.0561	
150	180	2	0	44.0	0.0455	0.0314	0.4636	0.5364	0.0499	836.2	45.8052	
180	210	3	1	41.5	0.0723	0.0402	0.4426	0.5574	0.0498	820.0	45.0209	
210	240	1	0	38.0	0.0263	0.0260	0.4106	0.5894	0.0495	810.9	43.6474	
240	270	1	1	36.5	0.0274	0.0270	0.3998	0.6002	0.0494	788.0	43.3633	
270	300	2	0	35.0	0.0571	0.0392	0.3888	0.6112	0.0492	765.2	43.0695	
300	330	1	0	33.0	0.0303	0.0298	0.3666	0.6334	0.0488	749.8	41.8209	
330	360	3	1	31.5	0.0952	0.0523	0.3555	0.6445	0.0486	1054.4	25.9424	
360	390	0	1	27.5	0	0	0.3216	0.6784	0.0478	1038.3	25.1208	
390	420	0	1	26.5	0	0	0.3216	0.6784	0.0478	1008.3	25.5904	
420	450	0	2	25.0	0	0	0.3216	0.6784	0.0478	978.3	26.3469	
450	480	0	0	24.0	0	0	0.3216	0.6784	0.0478	948.3	26.8902	
480	510	0	1	23.5	0	0	0.3216	0.6784	0.0478	918.3	27.1748	
510	540	0	1	22.5	0	0	0.3216	0.6784	0.0478	888.3	27.7721	
540	570	0	1	21.5	0	0	0.3216	0.6784	0.0478	858.3	28.4106	
570	600	1	1	20.5	0.0488	0.0476	0.3216	0.6784	0.0478	828.3	29.0953	
600	630	0	1	18.5	0	0	0.3059	0.6941	0.0479	804.7	29.1337	
630	660	0	0	18.0	0	0	0.3059	0.6941	0.0479	774.7	29.5355	
660	690	1	1	17.5	0.0571	0.0555	0.3059	0.6941	0.0479	744.7	29.9545	
690	720	0	0	16.0	0	0	0.2885	0.7115	0.0483	-	-	
720	750	1	0	16.0	0.0625	0.0605	0.2885	0.7115	0.0483	-	-	
750	780	0	0	15.0	0	0	0.2704	0.7296	0.0485	-	-	
780	810	0	0	15.0	0	0	0.2704	0.7296	0.0485	-	-	
810	840	0	0	15.0	0	0	0.2704	0.7296	0.0485	-	-	
840	870	1	1	14.5	0.0690	0.0665	0.2704	0.7296	0.0485	-	-	
870	900	0	0	13.0	0	0	0.2518	0.7482	0.0486	-	-	
900	930	0	1	12.5	0	0	0.2518	0.7482	0.0486	-	-	
930	960	0	1	11.5	0	0	0.2518	0.7482	0.0486	-	-	
960	990	1	0	11.0	0.0909	0.0867	0.2518	0.7482	0.0486	-	-	
990	1020	1	0	10.0	0.1000	0.0949	0.2289	0.7711	0.0493	-	-	
1020	1050	1	0	9.0	0.1111	0.1048	0.2060	0.7940	0.0494	-	-	
1050	1080	0	0	8.0	0	0	0.1831	0.8169	0.0489	-	-	
1080	1110	0	0	8.0	0	0	0.1831	0.8169	0.0489	-	-	
1110	1140	0	0	8.0	0	0	0.1831	0.8169	0.0489	-	-	
1140	1170	0	1	7.5	0	0	0.1831	0.8169	0.0489	-	-	
1170	1200	0	0	7.0	0	0	0.1831	0.8169	0.0489	-	-	
1200	1230	0	0	7.0	0	0	0.1831	0.8169	0.0489	-	-	

Life Table Survival Estimates												
Interval		Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error			Survival Standard Error	Median Residual Lifetime	Median Standard Error	
[Lower,	Upper]						Survival	Failure				
1230	1260	0	0	7.0	0	0	0.1831	0.8169	0.0489	.	.	
1260	1290	0	0	7.0	0	0	0.1831	0.8169	0.0489	.	.	
1290	1320	0	0	7.0	0	0	0.1831	0.8169	0.0489	.	.	
1320	1350	0	1	6.5	0	0	0.1831	0.8169	0.0489	.	.	
1350	1380	0	0	6.0	0	0	0.1831	0.8169	0.0489	.	.	
1380	1410	1	2	5.0	0.2000	0.1789	0.1831	0.8169	0.0489	.	.	
1410	1440	0	0	3.0	0	0	0.1465	0.8535	0.0510	.	.	
1440	1470	0	0	3.0	0	0	0.1465	0.8535	0.0510	.	.	
1470	1500	0	0	3.0	0	0	0.1465	0.8535	0.0510	.	.	
1500	1530	0	0	3.0	0	0	0.1465	0.8535	0.0510	.	.	
1530	1560	0	0	3.0	0	0	0.1465	0.8535	0.0510	.	.	
1560	1590	0	2	2.0	0	0	0.1465	0.8535	0.0510	.	.	
1590	1620	0	0	1.0	0	0	0.1465	0.8535	0.0510	.	.	
1620	1650	0	0	1.0	0	0	0.1465	0.8535	0.0510	.	.	
1650	1680	0	0	1.0	0	0	0.1465	0.8535	0.0510	.	.	
1680	1710	0	0	1.0	0	0	0.1465	0.8535	0.0510	.	.	
1710	1740	0	0	1.0	0	0	0.1465	0.8535	0.0510	.	.	
1740	1770	0	0	1.0	0	0	0.1465	0.8535	0.0510	.	.	
1770	1800	0	1	0.5	0	0	0.1465	0.8535	0.0510	.	.	
1800	.	0	0	0.0	0	0	0.1465	0.8535	0.0510	.	.	



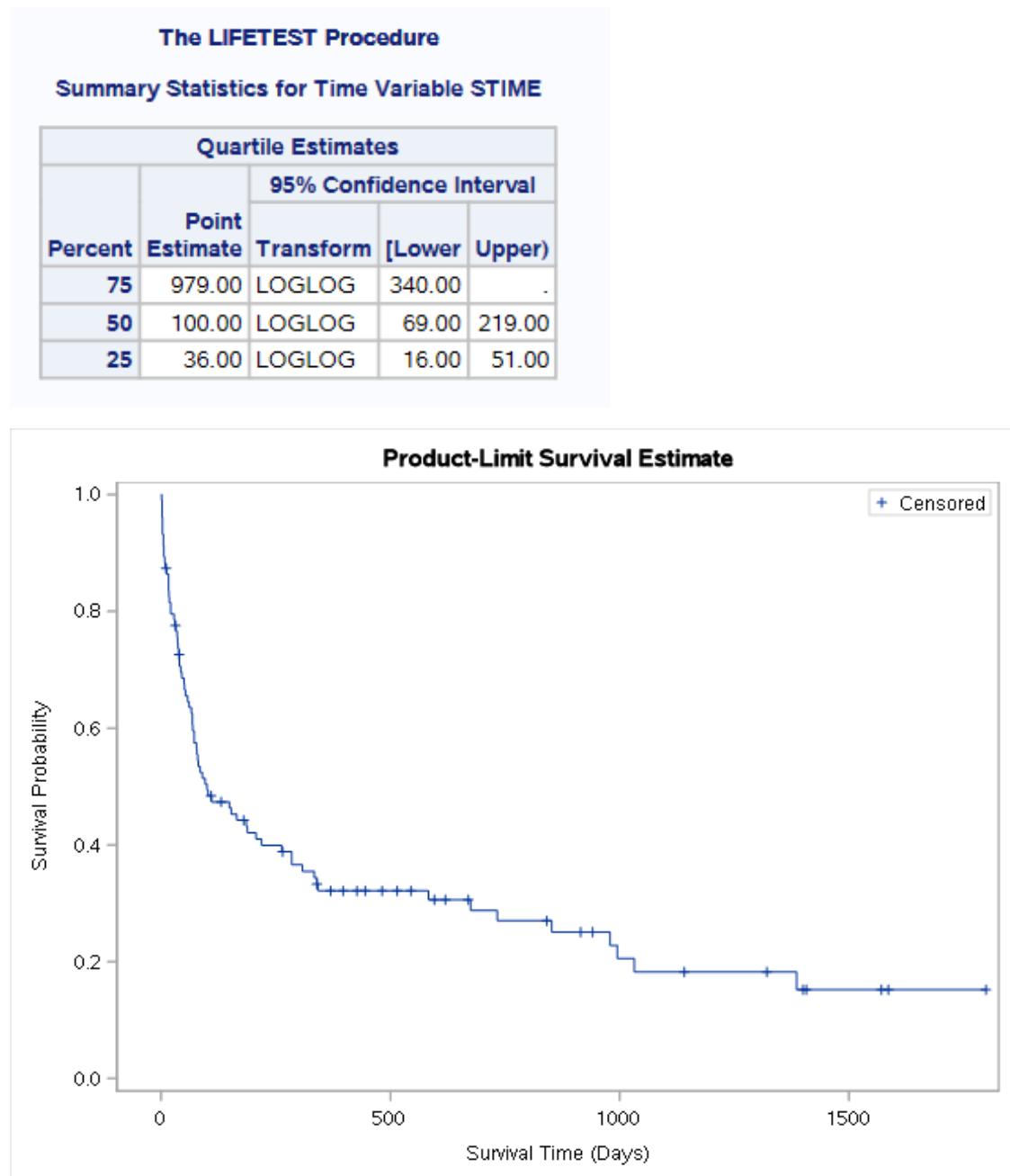
### Méthode Kaplan-Meier

Le tableau des estimateurs ne sera pas reporté (voir intro du document). Pour récupérer ces estimateurs, on peut les récupérer via l'instruction output et les exporter, par exemple, dans un tableur.

```

ods exclude Lifetest.Stratum1.ProductLimitEstimates;
proc lifetest data=trans;
time stime*died(0); run;

```



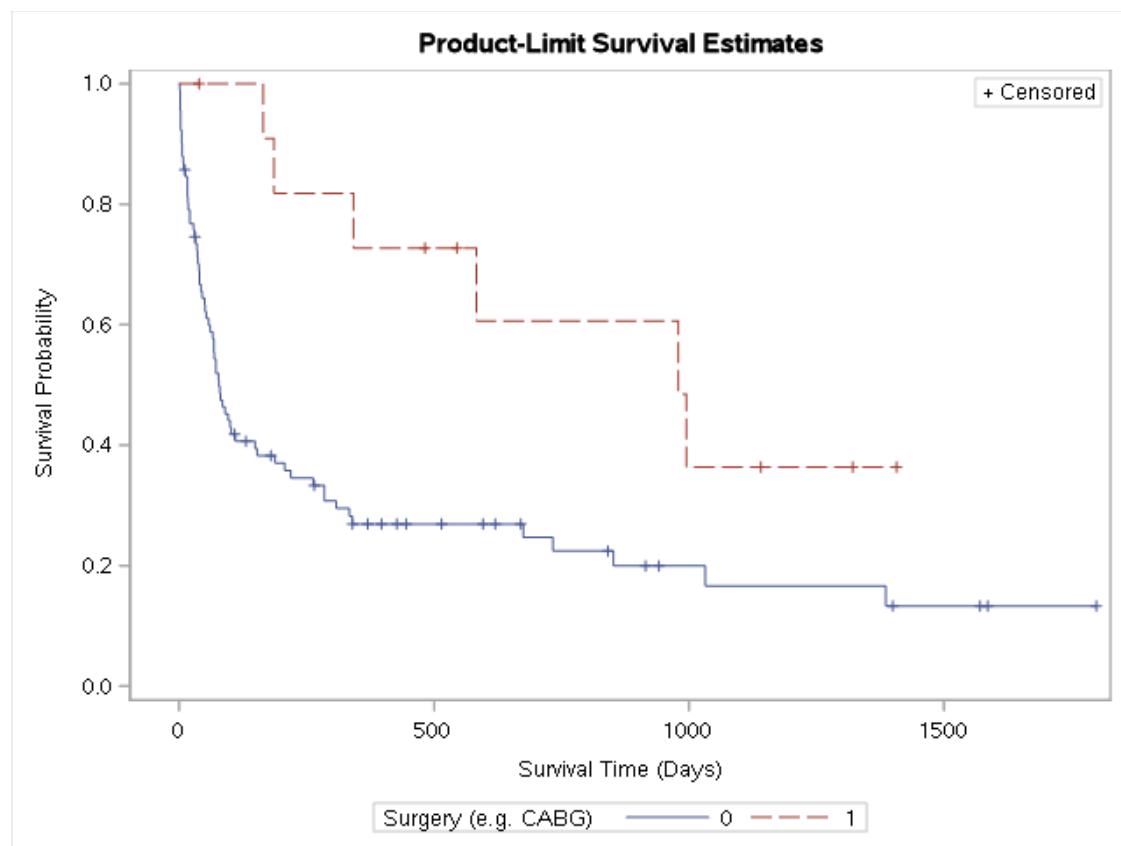
Summary of the Number of Censored and Uncensored Values			
Total	Failed	Censored	Percent Censored
103	75	28	27.18

Warning sur la durée moyenne reportée Sauf exception (pas de censure à droite) ne pas interpréter le tableau donnant la durée moyenne, qui n'a pas été reporté ici. Se reporter à l'estimation des RMST plus bas.

## Comparaison des fonctions de survie

### Tests du log rank

```
ods exclude Lifetest.Stratum1.ProductLimitEstimates Lifetest.Stratum2.  
.ProductLimitEstimates ;  
  
proc lifetest data=trans;  
time stime*died(0);  
strata surgery / test=all;  
run;
```



### The LIFETEST Procedure

Stratum 1: Surgery (e.g. CABG) = 0

Summary Statistics for Time Variable STIME

Percent	Point Estimate	95% Confidence Interval		
		Transform	[Lower	Upper)
75	675.00	LOGLOG	219.00	1386.00
50	78.00	LOGLOG	58.00	149.00
25	30.00	LOGLOG	16.00	43.00

### The LIFETEST Procedure

Stratum 2: Surgery (e.g. CABG) = 1

Summary Statistics for Time Variable STIME

Quartile Estimates				
Percent	Point Estimate	95% Confidence Interval		
		Transform	[Lower]	Upper)
75	.	LOGLOG	979.00	.
50	979.00	LOGLOG	186.00	.
25	342.00	LOGLOG	165.00	995.00

Test of Equality over Strata				
Test	Chi-Square	DF	Pr > Chi-Square	
Log-Rank	6.5900	1	0.0103	
Wilcoxon	8.9898	1	0.0027	
Tarone	8.4624	1	0.0036	
Peto	8.6641	1	0.0032	
Modified Peto	8.7027	1	0.0032	
Fleming(1)	8.6508	1	0.0033	

### Comparaison des RMST

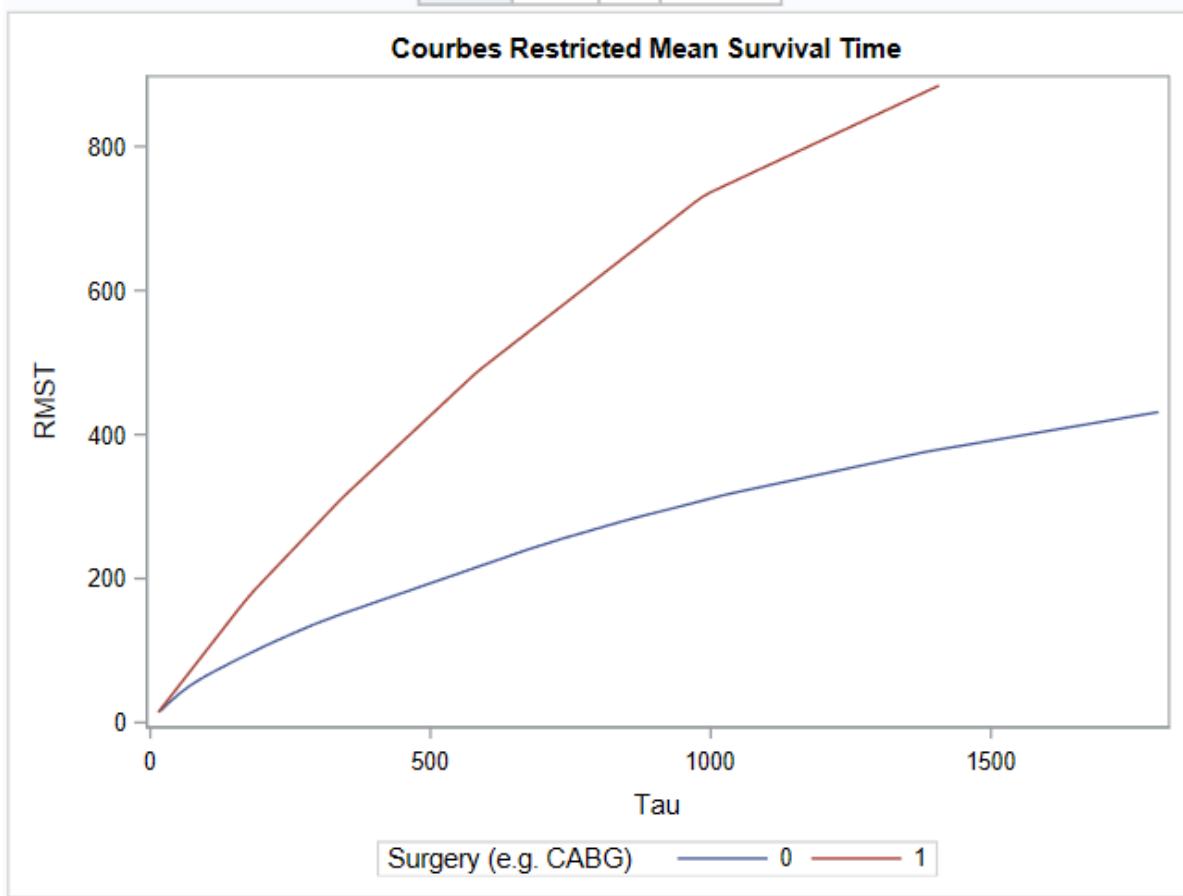
Disponible avec le dernier module stat de Sas base (Sas-Stat 15.1 novembre 2018).

```
ods exclude Lifetest.Stratum1.ProductLimitEstimates;
proc lifetest data=trans rmst plots=(rmst);
time stime*died(0);
strata surgery; run;
```

Informations sur l'analyse RMST	
Tau	1407

Estimations RMST				
Niveau de discréétisation	Surgery (e.g. CABG)	Estimation	Erreur type	
1	0	379.1476	58.6055	
2	1	884.5758	151.9794	

Test RMST d'égalité			
Source	khi-2	DDL	Pr > khi-2
Strata	9.6282	1	0.0019



## Modèle semi paramétrique de Cox

### Estimation du modèle

```
proc phreg data=trans;
model stime*died(0) = year age surgery ;
run;
```

#### The PHREG Procedure

Model Information		
Data Set	WORK.TRANS	
Dependent Variable	STIME	Survival Time (Days)
Censoring Variable	DIED	Survival Status (1=dead)
Censoring Value(s)	0	
Ties Handling	BRESLOW	

Number of Observations Read	103
Number of Observations Used	103

Summary of the Number of Event and Censored Values			
Total	Event	Censored	Percent Censored
103	75	28	27.18

#### Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	596.651	579.089
AIC	596.651	585.089
SBC	596.651	592.042

Testing Global Null Hypothesis: BETA=0				
Test	Chi-Square	DF	Pr > ChiSq	
Likelihood Ratio	17.5617	3	0.0005	
Score	16.6482	3	0.0008	
Wald	15.7002	3	0.0013	

#### Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
YEAR	1	-0.11951	0.06737	3.1468	0.0761	0.887	Year of Acceptance
AGE	1	0.02955	0.01353	4.7683	0.0290	1.030	Age
SURGERY	1	-0.98469	0.43626	5.0946	0.0240	0.374	Surgery (e.g. CABG)

## Tests de l'hypothèse PH

### Test de Grambsch Therneau sur les résidus de Schoenfeld

Le test est exécuté directement dans l'instruction `phreg` (ajouter `zph`). L'option `global` permet de récupérer le résultat du test omnibus (attention rejette facilement  $H_0$  - hypothèse PH respectée - lorsque le nombre de degré de liberté est élevé).

```
ods select PHReg.zphTest;  
  
proc phreg data=trans zph(global nopol);  
model stime*died(0) = year age surgery ;  
run;
```

The PHREG Procedure

zph Tests for Nonproportional Hazards						
Transform	Predictor Variable	Correlation	ChiSquare	Pr > ChiSquare	t Value	Pr >  t
RANK	YEAR	0.1626	2.0370	0.1535	1.41	0.1634
RANK	AGE	0.1052	1.0667	0.3017	0.90	0.3690
RANK	SURGERY	0.2435	3.7290	0.0535	2.14	0.0353
RANK	_Global_	.	7.7422	0.0517	.	.

Par défaut SAS utilise la transformation  $f(t) = t$  (idem Stata). Pour obtenir l'option par défaut de R  $f(t) = 1 - KM(t)$ :

```
ods select PHReg.zphTest;  
proc phreg data=trans zph(global nopol transform=km);  
model stime*died(0) = year age surgery ;  
run;
```

The PHREG Procedure

zph Tests for Nonproportional Hazards						
Transform	Predictor Variable	Correlation	ChiSquare	Pr > ChiSquare	t Value	Pr >  t
1-KM	YEAR	0.1634	2.0576	0.1515	1.42	0.1612
1-KM	AGE	0.1114	1.1957	0.2742	0.96	0.3414
1-KM	SURGERY	0.2520	3.9943	0.0457	2.22	0.0292
1-KM	_Global_	.	8.1906	0.0422	.	.

## *Introduction d'une interaction avec la durée*

Principe : Estimation d'un modèle de Cox en présence d'une intéraction posée sur la durée avec des indicatrices

La covariable doit être sous forme d'indicatrice (binaire: (0,1)). Ce qui est le cas ici avec la variable surgery.

Exemple avec une covariable X à 3 modalités codée 1,2,3.  
Estimation du modèle de Cox avec l'instruction *class* (ref: X=1)

```
proc phreg data=base;
class X(ref="1");
model variable_dur*variable_cens(0) = X; run;
```

Estimation du modèle de Cox avec indicatrices

```
data base; set base;
X1 = X=1;
X2 = X=2;
X3 = X=3; run;

proc phreg data=base;
model variable_dur*variable_cens(0) = X2 X3; run;
```

Pour l'exemple, la variable d'intéraction (*surgeryt* = *surgery* × *stime*) est générée, pour le temps de l'estimation seulement, après l'instruction *model*.

```
ods select PHReg.ParameterEstimates;
proc phreg data=trans ;
model stime*died(0) = year age surgery surgeryt ;
surgeryt = surgery*stime;
run;
```

### The PHREG Procedure

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
YEAR	1	-0.12295	0.06686	3.3815	0.0659	0.884	Year of Acceptance
AGE	1	0.02886	0.01346	4.5980	0.0320	1.029	Age
SURGERY	1	-1.75154	0.67446	6.7442	0.0094	0.174	Surgery (e.g. CABG)
surgeryt	1	0.00223	0.00110	4.0828	0.0433	1.002	

Rappel : l'estimateur pour la variable *surgeryt* n'est pas un rapport de risques mais un rapport de rapports de risques en t

## *Introduction d'une variable dynamique (binaire)*

**Warning:** opération en « aveugle »

Contrairement à R et Stata, la base n'a pas à être splittée, on ne peut pas vérifier si la variable dynamique a été correctement créée. La variable dynamique, qui peut être appréhendée comme une variable en interaction avec la durée, est générée après l'instruction *model*.

Ici la tvc prendra la valeur 1 lorsque *stime>=wait*, 0 sinon. Comme pour les interactions avec la durée, la variable tvc est créée de manière temporaire.

```
ods select PHReg.ParameterEstimates;
```

```
proc phreg data=trans;
model stime*died(0) = year age surgery tvc ;
tvc = transplant=1 and stime>=wait;
run;
```

The PHREG Procedure

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
YEAR	1	-0.12020	0.06736	3.1839	0.0744	0.887	Year of Acceptance
AGE	1	0.03042	0.01391	4.7829	0.0287	1.031	Age
SURGERY	1	-0.98023	0.43655	5.0418	0.0247	0.375	Surgery (e.g. CABG)
tvc	1	-0.08305	0.30484	0.0742	0.7853	0.920	

## Modèle logistique à temps discret

### *Mise en forme de la base et variables d'analyse*

On utilise une boucle pour répliquer les lignes sur la valeur de la durée. La nouvelle variable de durée (t) sous forme de compteur est générée automatiquement.

```
data td; set trans;
do t=1 to mois;
  output;
  end; run;

data td; set td;
if t<mois then died=0;
t2=t*t;
t3=t2*t; run;
```

### *Estimation du modèle*

#### *Fonction continue de la durée*

```
ods select Logistic.FitStatistics;
proc logistic data=td;
model died(ref="0") = t t2 t3 year age surgery ; run;
```

<b>Number of Observations Read</b>	1127
<b>Number of Observations Used</b>	1127

<b>Response Profile</b>		
<b>Ordered Value</b>	<b>DIED</b>	<b>Total Frequency</b>
1	0	1052
2	1	75

Probability modeled is DIED='1'.

<b>Model Convergence Status</b>			
Convergence criterion (GCONV=1E-8) satisfied.			

<b>Model Fit Statistics</b>		
<b>Criterion</b>	<b>Intercept Only</b>	<b>Intercept and Covariates</b>
AIC	553.368	474.673
SC	558.396	509.865
-2 Log L	551.368	460.673

<b>Testing Global Null Hypothesis: BETA=0</b>				
<b>Test</b>	<b>Chi-Square</b>	<b>DF</b>	<b>Pr &gt; ChiSq</b>	
<b>Likelihood Ratio</b>	90.6949	6	<.0001	
<b>Score</b>	96.4611	6	<.0001	
<b>Wald</b>	74.7266	6	<.0001	

<b>Analysis of Maximum Likelihood Estimates</b>					
<b>Parameter</b>	<b>DF</b>	<b>Estimate</b>	<b>Standard Error</b>	<b>Wald Chi-Square</b>	<b>Pr &gt; ChiSq</b>
<b>Intercept</b>	1	7.0827	5.3077	1.7806	0.1821
<b>t</b>	1	-0.3721	0.0824	20.3901	<.0001
<b>t2</b>	1	0.0142	0.00502	8.0344	0.0046
<b>t3</b>	1	-0.00017	0.000079	4.4660	0.0346
<b>YEAR</b>	1	-0.1327	0.0738	3.2338	0.0721
<b>AGE</b>	1	0.0333	0.0147	5.1530	0.0232
<b>SURGERY</b>	1	-1.0109	0.4486	5.0783	0.0242

### *Fonction discrète de la durée*

Pour l'exemple on va regrouper la durée par ses quartiles. Pour chaque individu, on conserve seulement une observation dans chaque quartile.

```
proc rank data=td out=td2 groups=4;
var t;
ranks tq4;
run;

data td2; set td2;
id2=put(id, 3.);
tq42=put(tq4, 1.);
g=id2 || tq42; run;

proc sort data=td2; by id tq4; run;

data td2; set td2;
by g;
if LAST.g; run;
```

```
proc logistic data=td2;
class tq4 / param=ref;
model died(ref="0") = tq4 year age surgery; run;
```

Response Profile		
Ordered Value	DIED	Total Frequency
1	0	124
2	1	75

Probability modeled is DIED='1'.

Class Level Information				
Class	Value	Design Variables		
tq4	0	1	0	0
	1	0	1	0
	2	0	0	1
	3	0	0	0

#### Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	265.682	236.835
SC	268.976	259.888
-2 Log L	263.682	222.835

#### Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	40.8472	6	<.0001
Score	37.1530	6	<.0001
Wald	31.1592	6	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
tq4	3	14.0947	0.0028
YEAR	1	4.4598	0.0347
AGE	1	6.1053	0.0135
SURGERY	1	4.5858	0.0322

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	11.5156	6.5518	3.0893	0.0788
tq4	0	0.4790	0.5625	0.7252	0.3945
tq4	1	-0.5566	0.6331	0.7730	0.3793
tq4	2	-1.4102	0.7540	3.4975	0.0615
YEAR	1	-0.1961	0.0929	4.4598	0.0347
AGE	1	0.0456	0.0185	6.1053	0.0135
SURGERY	1	-1.0773	0.5031	4.5858	0.0322

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
tq4 0 vs 3	1.614	0.536	4.862
tq4 1 vs 3	0.573	0.166	1.982
tq4 2 vs 3	0.244	0.056	1.070
YEAR	0.822	0.685	0.986
AGE	1.047	1.009	1.085
SURGERY	0.340	0.127	0.913

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	77.2	Somers' D	0.544
Percent Discordant	22.7	Gamma	0.545
Percent Tied	0.1	Tau-a	0.257
Pairs	9300	c	0.772

## Modèles paramétrique de type AFT

On utilise la procédure proc *lifereg* et on indique le type de distribution.

Weibull

```
proc lifereg data=trans;
model stime*died(0) = year age surgery /D=WEIBULL;
run;
```

### The LIFEREG Procedure

Model Information		
Data Set	WORK.TRANS	
Dependent Variable	Log(STIME)	Survival Time (Days)
Censoring Variable	DIED	Survival Status (1=dead)
Censoring Value(s)	0	
Number of Observations	103	
Noncensored Values	75	
Right Censored Values	28	
Left Censored Values	0	
Interval Censored Values	0	
Number of Parameters	5	
Name of Distribution	Weibull	
Log Likelihood	-188.6278016	

Number of Observations Read	103
Number of Observations Used	103

Fit Statistics	
-2 Log Likelihood	377.256
AIC (smaller is better)	387.256
AICC (smaller is better)	387.874
BIC (smaller is better)	400.429

Fit Statistics (Unlogged Response)	
-2 Log Likelihood	976.337
Weibull AIC (smaller is better)	986.337
Weibull AICC (smaller is better)	986.955
Weibull BIC (smaller is better)	999.510

Algorithm converged.

Type III Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
YEAR	1	1.7673	0.1837
AGE	1	6.1828	0.0129
SURGERY	1	6.3906	0.0115

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
				Lower	Upper		
Intercept	1	-3.0220	8.7284	-20.1294	14.0854	0.12	0.7292
YEAR	1	0.1620	0.1218	-0.0768	0.4008	1.77	0.1837
AGE	1	-0.0615	0.0247	-0.1100	-0.0130	6.18	0.0129
SURGERY	1	1.9703	0.7794	0.4427	3.4980	6.39	0.0115
Scale	1	1.7983	0.1667	1.4995	2.1566		
Weibull Shape	1	0.5561	0.0516	0.4637	0.6669		

*Log-logistique*

```
proc lifereg data=trans;
model stime*died(0) = year age surgery /D=LLOGISTIC;
run;
```

### The LIFEREG Procedure

Model Information		
Data Set	WORK.TRANS	
Dependent Variable	Log(STIME)	Survival Time (Days)
Censoring Variable	DIED	Survival Status (1=dead)
Censoring Value(s)	0	
Number of Observations	103	
Noncensored Values	75	
Right Censored Values	28	
Left Censored Values	0	
Interval Censored Values	0	
Number of Parameters	5	
Name of Distribution	LL logistic	
Log Likelihood	-183.0393686	

Number of Observations Read	103
Number of Observations Used	103

Fit Statistics	
-2 Log Likelihood	366.079
AIC (smaller is better)	376.079
AICC (smaller is better)	376.697
BIC (smaller is better)	389.252

Fit Statistics (Unlogged Response)	
-2 Log Likelihood	965.160
LL logistic AIC (smaller is better)	975.160
LL logistic AICC (smaller is better)	975.778
LL logistic BIC (smaller is better)	988.333

Algorithm converged.

Type III Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
YEAR	1	4.2192	0.0400
AGE	1	4.0189	0.0450
SURGERY	1	10.8277	0.0010

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-10.4034	8.3410	-26.7515	5.9446	1.56	0.2123
YEAR	1	0.2408	0.1172	0.0110	0.4705	4.22	0.0400
AGE	1	-0.0427	0.0213	-0.0845	-0.0010	4.02	0.0450
SURGERY	1	2.2747	0.6913	0.9198	3.6296	10.83	0.0010
Scale	1	1.1979	0.1161	0.9906	1.4486		

## Risques concurrents

### *Non paramétrique*

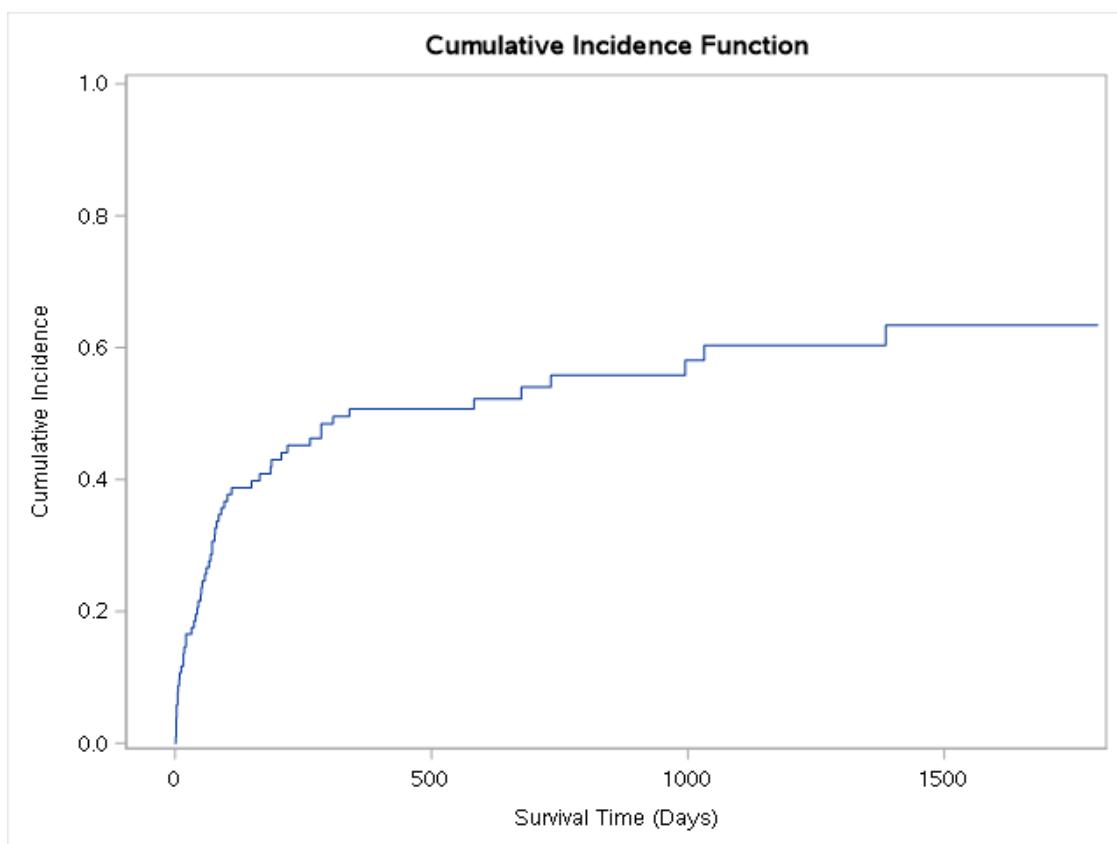
On indique en option la cause d'intérêt avec `eventcode=valeur`, les autres étant considérées comme des risques concurrents.

```
proc lifetest data=trans plots=CIF;
time stime*compet(0) / eventcode=1; run;
```

Summary of Failure Outcomes				
Failed Events	Competing Events	Censored	Total	
56	19	28	103	

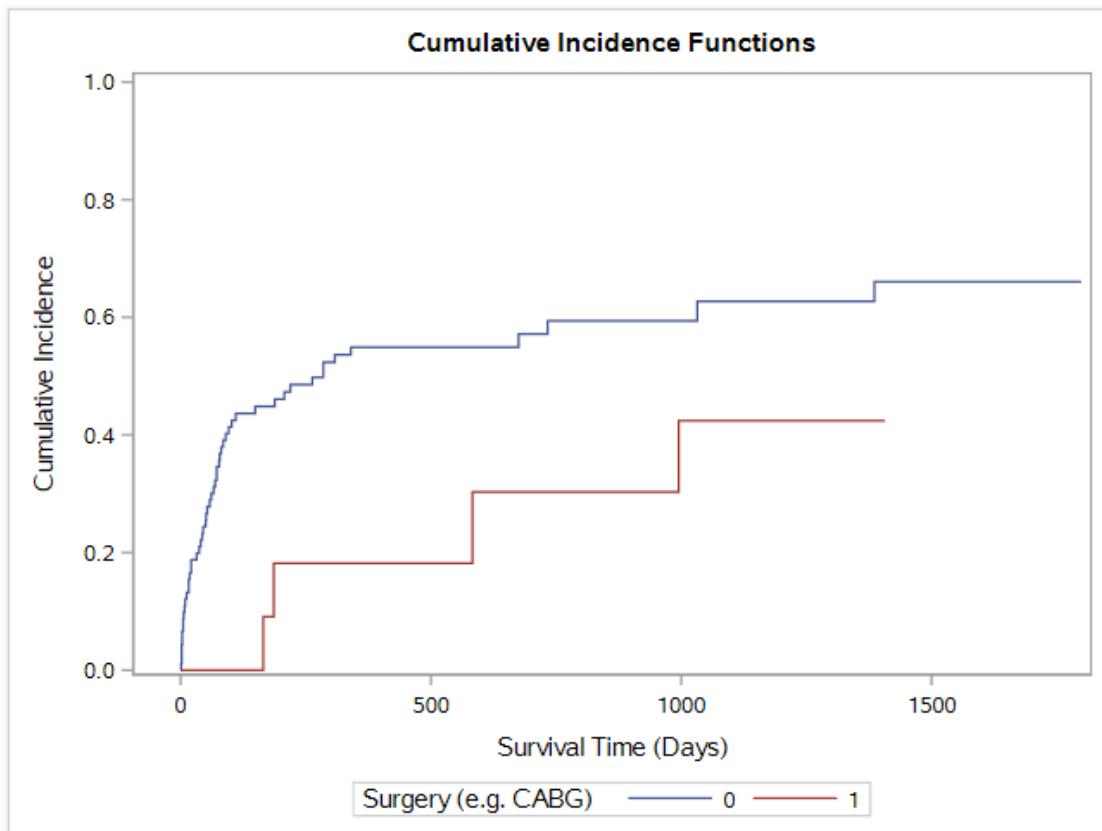
Cumulative Incidence Function Estimates				
STIME	Cumulative Incidence	Standard Error	95% Confidence Interval	
0	0	0	-	-
1	0.00971	0.00971	0.000847	0.0480
2	0.0388	0.0191	0.0126	0.0895
3	0.0583	0.0232	0.0237	0.1153
5	0.0777	0.0265	0.0362	0.1399
6	0.0874	0.0280	0.0427	0.1519
8	0.0971	0.0293	0.0495	0.1638
9	0.1068	0.0306	0.0564	0.1755
12	0.1166	0.0318	0.0636	0.1872
16	0.1362	0.0340	0.0782	0.2103
18	0.1461	0.0350	0.0858	0.2217
21	0.1657	0.0369	0.1011	0.2442
32	0.1756	0.0378	0.1090	0.2554
37	0.1856	0.0386	0.1169	0.2667
40	0.1957	0.0395	0.1251	0.2780
43	0.2058	0.0403	0.1333	0.2892
45	0.2158	0.0410	0.1416	0.3004
50	0.2259	0.0417	0.1500	0.3115
51	0.2360	0.0424	0.1584	0.3226
53	0.2461	0.0430	0.1669	0.3335
58	0.2562	0.0436	0.1755	0.3444
61	0.2662	0.0442	0.1841	0.3553
66	0.2763	0.0448	0.1928	0.3661
69	0.2864	0.0453	0.2015	0.3768
72	0.3066	0.0462	0.2192	0.3982
77	0.3167	0.0466	0.2281	0.4088
78	0.3267	0.0470	0.2370	0.4193
81	0.3368	0.0474	0.2460	0.4298
85	0.3469	0.0478	0.2551	0.4402
90	0.3570	0.0481	0.2642	0.4506
96	0.3671	0.0484	0.2733	0.4610
102	0.3771	0.0487	0.2825	0.4713
110	0.3874	0.0490	0.2919	0.4819
149	0.3980	0.0493	0.3015	0.4926
165	0.4085	0.0496	0.3111	0.5034
186	0.4193	0.0498	0.3210	0.5143

Cumulative Incidence Function Estimates				
STIME	Cumulative Incidence	Standard Error	95% Confidence Interval	
188	0.4301	0.0501	0.3309	0.5253
207	0.4408	0.0503	0.3409	0.5361
219	0.4516	0.0505	0.3509	0.5470
263	0.4624	0.0507	0.3610	0.5577
285	0.4846	0.0510	0.3817	0.5798
308	0.4957	0.0511	0.3922	0.5908
340	0.5068	0.0512	0.4027	0.6017
583	0.5221	0.0519	0.4160	0.6178
675	0.5401	0.0531	0.4308	0.6372
733	0.5580	0.0540	0.4460	0.6561
995	0.5808	0.0560	0.4634	0.6813
1032	0.6036	0.0574	0.4819	0.7054
1386	0.6340	0.0606	0.5028	0.7393



Pour récupérer le test de Gray, on utilise l'instruction *strata*.

```
proc lifetest data=trans plots=CIF;
time stime*compet(0) / eventcode=1;
strata surgery; run;
```



Gray's Test for Equality of Cumulative Incidence Functions		
Chi-Square	DF	Pr > Chi-Square
3.5544	1	0.0594

## Modèles

### Modèle de Fine-Gray

```
proc phreg data=trans;
model stime*compet(0) = year age surgery / eventcode=1 ;
run;
```

Model Information		
Data Set	WORK.TRANS	
Dependent Variable	STIME	Survival Time (Days)
Status Variable	COMPET	
Event of Interest	1	
Competing Event	2	
Censored Value	0	

Number of Observations Read	103
Number of Observations Used	103

Summary of Failure Outcomes			
Total	Event of Interest	Competing Event	Censored
103	56	19	28

Convergence Status			
Convergence criterion (GCONV=1E-8) satisfied.			

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	466.621	455.384
AIC	466.621	461.384
SBC	466.621	467.460

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Wald	11.5295	3	0.0092

Analysis of Maximum Likelihood Estimates							
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
YEAR	1	-0.07235	0.07127	1.0306	0.3100	0.930	Year of Acceptance
AGE	1	0.03702	0.01765	4.3978	0.0360	1.038	Age
SURGERY	1	-0.86888	0.44883	3.7477	0.0529	0.419	Surgery (e.g. CABG)

*Modèle logistique multinomial à temps discret*

```
data td; set trans;
do t=1 to mois;
    output;
    end;
run;
data td; set td;
if t<mois then compet=0;
t2=t*t
run;

proc logistic data=td;
model compet(ref="0") = t t2 year age surgery / link=glogit;
run;
```

Model Information	
Data Set	WORK.TD
Response Variable	COMPET
Number of Response Levels	3
Model	generalized logit
Optimization Technique	Newton-Raphson

Number of Observations Read	1127
Number of Observations Used	1127

Response Profile		
Ordered Value	COMPET	Total Frequency
1	0	1052
2	1	56
3	2	19

Logits modeled use COMPET='0' as the reference category.

Model Convergence Status		
Convergence criterion (GCONV=1E-8) satisfied.		

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	640.263	574.011
SC	650.318	634.339
-2 Log L	636.263	550.011

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	86.2526	10	<.0001
Score	83.7257	10	<.0001
Wald	65.8555	10	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
t	2	31.5814	<.0001
t2	2	15.8351	0.0004
YEAR	2	4.3364	0.1144
AGE	2	6.4002	0.0408
SURGERY	2	5.0701	0.0793

Analysis of Maximum Likelihood Estimates						
Parameter	COMPET	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	5.6442	5.9110	0.9118	0.3396
Intercept	2	1	11.3083	9.8625	1.3147	0.2516
t	1	1	-0.2034	0.0415	24.0681	<.0001
t	2	1	-0.2023	0.0691	8.5693	0.0034
t2	1	1	0.00317	0.000896	12.4905	0.0004
t2	2	1	0.00298	0.00153	3.7775	0.0519
YEAR	1	1	-0.1284	0.0817	2.4734	0.1158
YEAR	2	1	-0.2036	0.1382	2.1710	0.1406
AGE	1	1	0.0439	0.0175	6.3108	0.0120
AGE	2	1	0.0110	0.0246	0.2005	0.6543
SURGERY	1	1	-1.1465	0.5387	4.5305	0.0333
SURGERY	2	1	-0.6139	0.7799	0.6196	0.4312

Odds Ratio Estimates				
Effect	COMPET	Point Estimate	95% Wald Confidence Limits	
t	1	0.816	0.752	0.885
t	2	0.817	0.713	0.935
t2	1	1.003	1.001	1.005
t2	2	1.003	1.000	1.006
YEAR	1	0.879	0.749	1.032
YEAR	2	0.816	0.622	1.070
AGE	1	1.045	1.010	1.081
AGE	2	1.011	0.964	1.061
SURGERY	1	0.318	0.111	0.913
SURGERY	2	0.541	0.117	2.496

# R

## Librairies utilisés pour l'analyse

- *Non Paramétrique:* **survival**, **survRM2**
- *Semi paramétrique, temps discret:* **survival**, fonction **uncount** (package **tidyR**) , fonction **glm**, fonction **quantcut** (package **gtools**)
- *Paramétrique :* **survival** , **flexsurv**
- *Risques concurrents:* **cmprsk**

Autres: **survminer** et **jtools** (+ *RecordLinkage*) pour améliorer certains outputs (graphiques et résultats de régression).

## Installations des librairies

Les dernières versions de certains packages peuvent être installées via Github (ex: **survminer**).

```
#install.packages("survival")
#install.packages("survminer")
#install.packages("flexsurv")
#install.packages("survRM2")
#install.package(tidyR)
#install.packages("gtools")
#install.packages("jtools")
#install.packages("miceadds")
#install.packages("RecordLinkage")
#install.packages("cmprsk")
library(survival)
library(survminer)
library(flexsurv)
library(survRM2)
library(tidyR)
library(gtools)
library(jtools)
library(RecordLinkage)
library(cmprsk)
```

## Chargement de la base transplantation

```
library(readr)
trans =
read_csv("https://raw.githubusercontent.com/mthevenin/analyse_duree/master/bases")
```

# Analyse non paramétrique

## Méthode actuarielle

La fonction disponible du paquet *discsurv*, lifetable ne permet pas de définir des intervalles, contrairement à Sas ou Stata. Les estimateurs sont systématiquement calculés sur des largeurs égales à 1, et il n'a pas de calcul des durées sur les différents quantiles de la courbe de survie. Elle ne sera donc pas présentée.

## Méthode Kaplan-Meier

Le package *survival* est le principal outil d'analyse des durées. Le package *survminer* permet d'améliorer la présentation des graphiques.

### Estimation des fonctions de survie

#### Fonction *survfit*

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

On peut renseigner les variables permettant de calculer la durée et non la variable de durée elle-même.

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

Estimation sans comparaison de groupes :

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

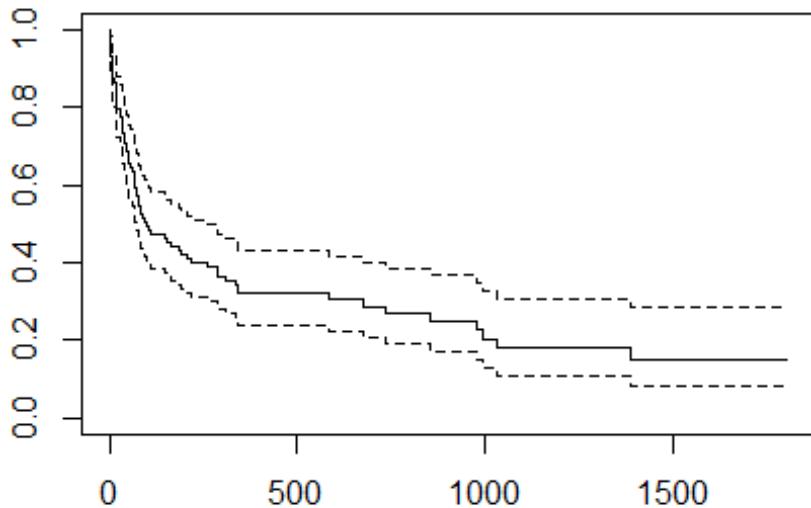
## Call: survfit(formula = Surv(stime, died) ~ 1, data = trans)
##
##      n  events  median 0.95LCL 0.95UCL
##    103      75      100      72      263

summary(fit)

## Call: survfit(formula = Surv(stime, died) ~ 1, data = trans)
##
##   time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1    103      1    0.990 0.00966    0.9715    1.000
##     2    102      3    0.961 0.01904    0.9246    0.999
##     3     99      3    0.932 0.02480    0.8847    0.982
##     5     96      2    0.913 0.02782    0.8597    0.969
##     6     94      2    0.893 0.03043    0.8355    0.955
##     8     92      1    0.883 0.03161    0.8237    0.948
```

##	9	91	1	0.874 0.03272	0.8119	0.940
##	12	89	1	0.864 0.03379	0.8002	0.933
##	16	88	3	0.835 0.03667	0.7656	0.910
##	17	85	1	0.825 0.03753	0.7543	0.902
##	18	84	1	0.815 0.03835	0.7431	0.894
##	21	83	2	0.795 0.03986	0.7208	0.877
##	28	81	1	0.785 0.04056	0.7098	0.869
##	30	80	1	0.776 0.04122	0.6989	0.861
##	32	78	1	0.766 0.04188	0.6878	0.852
##	35	77	1	0.756 0.04250	0.6769	0.844
##	36	76	1	0.746 0.04308	0.6659	0.835
##	37	75	1	0.736 0.04364	0.6551	0.827
##	39	74	1	0.726 0.04417	0.6443	0.818
##	40	72	2	0.706 0.04519	0.6225	0.800
##	43	70	1	0.696 0.04565	0.6117	0.791
##	45	69	1	0.686 0.04609	0.6009	0.782
##	50	68	1	0.675 0.04650	0.5902	0.773
##	51	67	1	0.665 0.04689	0.5796	0.764
##	53	66	1	0.655 0.04725	0.5690	0.755
##	58	65	1	0.645 0.04759	0.5584	0.746
##	61	64	1	0.635 0.04790	0.5479	0.736
##	66	63	1	0.625 0.04819	0.5374	0.727
##	68	62	2	0.605 0.04870	0.5166	0.708
##	69	60	1	0.595 0.04892	0.5063	0.699
##	72	59	2	0.575 0.04929	0.4857	0.680
##	77	57	1	0.565 0.04945	0.4755	0.670
##	78	56	1	0.554 0.04958	0.4654	0.661
##	80	55	1	0.544 0.04970	0.4552	0.651
##	81	54	1	0.534 0.04979	0.4451	0.641
##	85	53	1	0.524 0.04986	0.4351	0.632
##	90	52	1	0.514 0.04991	0.4251	0.622
##	96	51	1	0.504 0.04994	0.4151	0.612
##	100	50	1	<b>0.494 0.04995</b>	<b>0.4052</b>	<b>0.602</b>
##	102	49	1	0.484 0.04993	0.3953	0.592
##	110	47	1	0.474 0.04992	0.3852	0.582
##	149	45	1	0.463 0.04991	0.3749	0.572
##	153	44	1	0.453 0.04987	0.3647	0.562
##	165	43	1	0.442 0.04981	0.3545	0.551
##	186	41	1	0.431 0.04975	0.3440	0.541
##	188	40	1	0.420 0.04966	0.3336	0.530
##	207	39	1	0.410 0.04954	0.3233	0.519
##	219	38	1	0.399 0.04940	0.3130	0.509
##	263	37	1	0.388 0.04923	0.3027	0.498
##	285	35	2	0.366 0.04885	0.2817	0.475
##	308	33	1	0.355 0.04861	0.2713	0.464
##	334	32	1	0.344 0.04834	0.2610	0.453
##	340	31	1	0.333 0.04804	0.2507	0.442
##	342	29	1	0.321 0.04773	0.2401	0.430
##	583	21	1	0.306 0.04785	0.2252	0.416
##	675	17	1	0.288 0.04830	0.2073	0.400
##	733	16	1	0.270 0.04852	0.1898	0.384
##	852	14	1	0.251 0.04873	0.1712	0.367
##	979	11	1	0.228 0.04934	0.1491	0.348
##	995	10	1	0.205 0.04939	0.1279	0.329
##	1032	9	1	0.182 0.04888	0.1078	0.308
##	1386	6	1	0.152 0.04928	0.0804	0.287

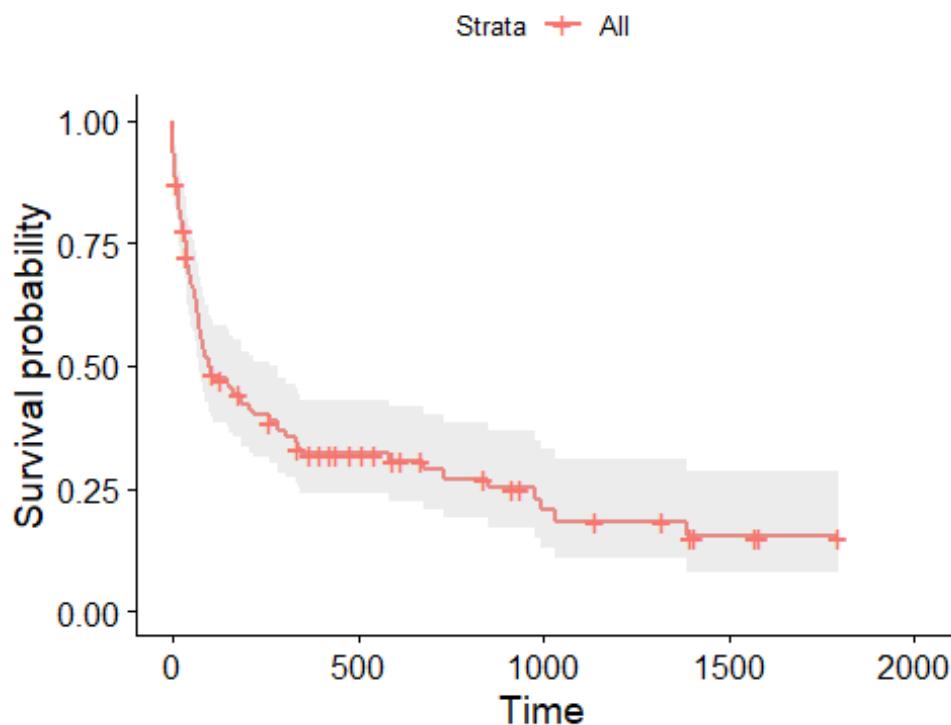
```
plot(fit)
```



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

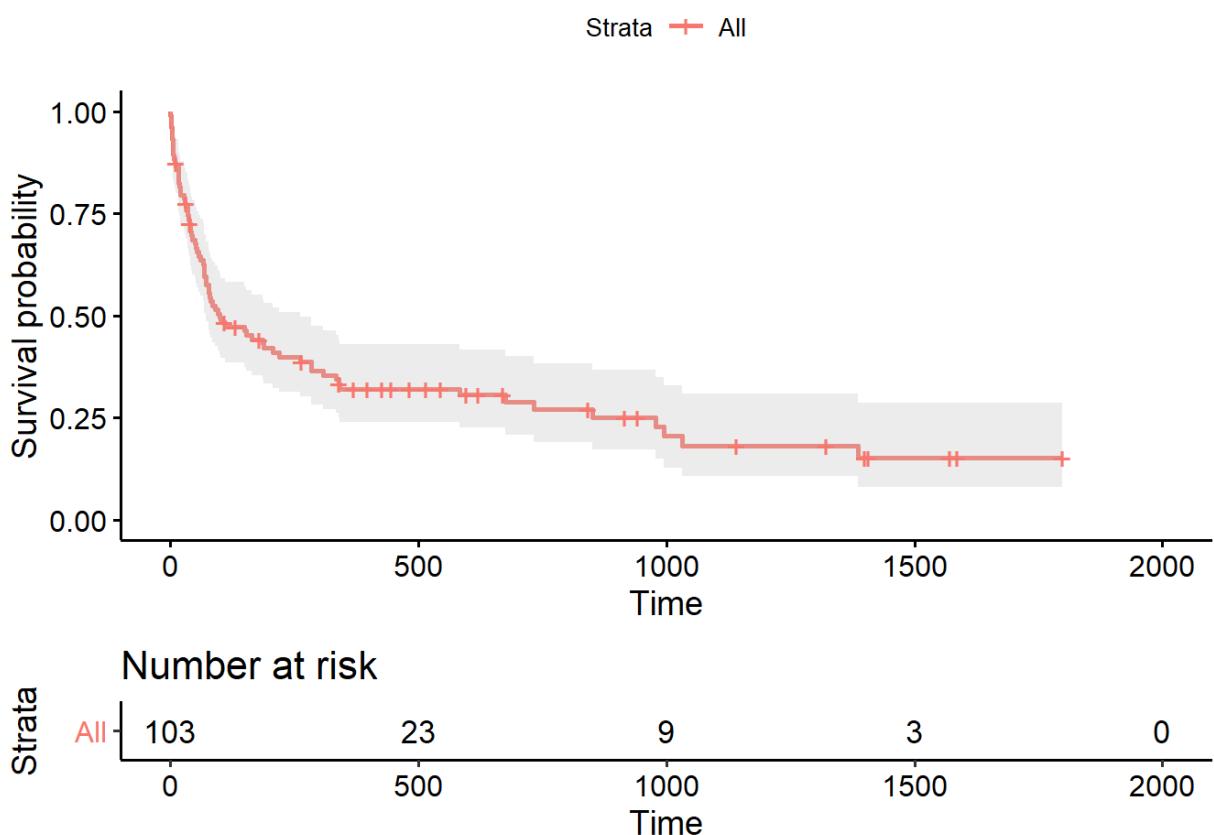
On peut obtenir des graphes de qualité avec les fonctions du package *survminer*. Avec la fonction *ggsurvplot* :

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



On peut ajouter la population encore soumise au risque pour plusieurs points de la durée.

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



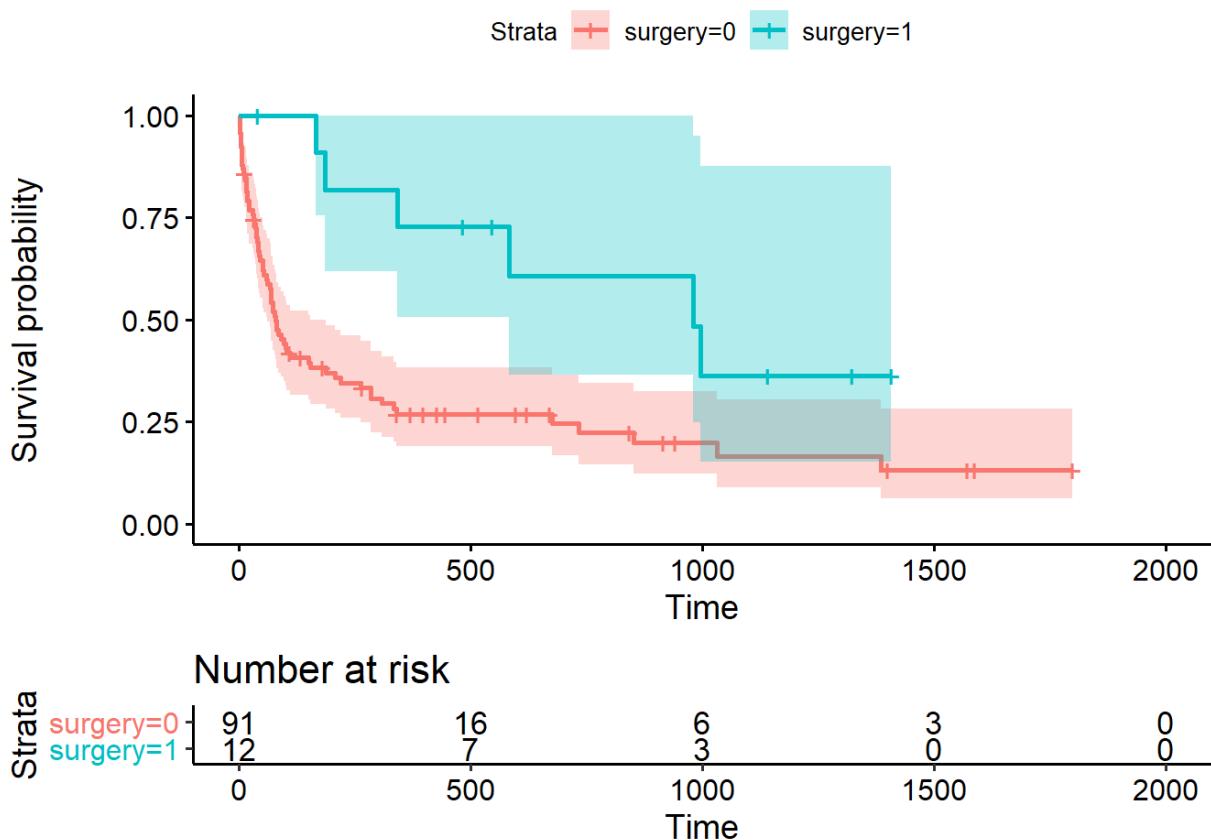
## Comparaison des fonctions de survie

On va comparer les fonctions de survie pour la variable surgery.

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

## Call: survfit(formula = Surv(stime, died) ~ surgery, data = trans)
##
##      n events median 0.95LCL 0.95UCL
## surgery=0 91      69      78      61     153
## surgery=1 12       6     979     583      NA

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



### Tests du log-rank

On utilise la fonction `survdiff`, et on sélectionne le test de Peto-Peto (`rho=1`). La syntaxe est quasiment identique à la fonction `survfit`.

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

## Call:
## survdiff(formula = Surv(stime, died) ~ surgery, data = trans,
##           rho = 1)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## surgery=0 91     45.28    39.12     0.968      8.65
## surgery=1 12      2.03     8.18     4.630      8.65
##
##  Chisq= 8.7 on 1 degrees of freedom, p= 0.003
```

Ici la variable est binaire. Si on veux tester deux à deux les niveaux d'une variable catégorielle à plus de deux modalité, on peut utiliser la fonction `pairwise_survdiff` (syntaxe identique que `survdiff`) de la librairie `survminer`.

### Comparaison des RMST

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

```
trans$arm=trans$surgery

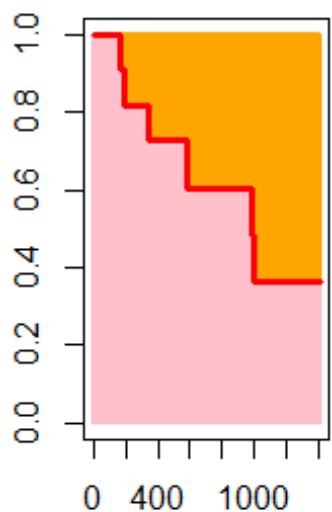
a=rmst2(trans$stime, trans$died, trans$arm, tau=NULL)
print(a)

##
## The truncation time, tau, was not specified. Thus, the default tau
1407 is used.
##
## Restricted Mean Survival Time (RMST) by arm
##          Est.      se lower .95 upper .95
## RMST (arm=1) 884.576 151.979   586.702 1182.450
## RMST (arm=0) 379.148  58.606   264.283  494.012
##
## 
## Restricted Mean Time Lost (RMTL) by arm
##          Est.      se lower .95 upper .95
## RMTL (arm=1) 522.424 151.979   224.550  820.298
## RMTL (arm=0) 1027.852  58.606   912.988 1142.717
##
## 
## Between-group contrast
##          Est. lower .95 upper .95      p
```

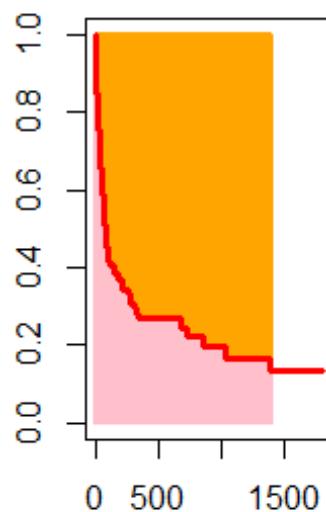
```
## RMST (arm=1)-(arm=0) 505.428   186.175   824.682 0.002  
## RMST (arm=1)/(arm=0)  2.333     1.483     3.670 0.000  
## RMTL (arm=1)/(arm=0)  0.508     0.284     0.909 0.022
```

```
plot(a)
```

**arm=1**



**arm=0**



RMST: 884.58

RMST: 379.15

## Modèle semi paramétrique de Cox

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

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

### *Estimation du modèle*

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

Syntaxe:

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

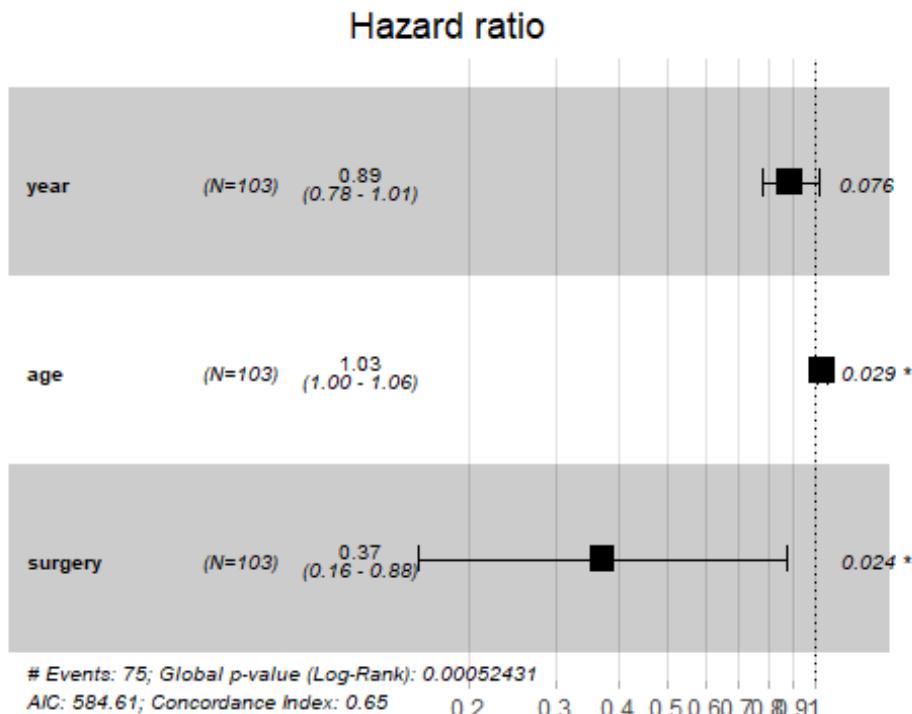
```
coxfit = coxph(formula = Surv(stime, died) ~ year + age + surgery,
data = trans)
summary(coxfit)

## Call:
## coxph(formula = Surv(stime, died) ~ year + age + surgery, data = trans)
##
##    n= 103, number of events= 75
##
##              coef exp(coef)  se(coef)      z Pr(>|z| )
## year     -0.11963   0.88725  0.06734 -1.776   0.0757 .
## age      0.02958   1.03002  0.01352  2.187   0.0287 *
## surgery -0.98732   0.37257  0.43626 -2.263   0.0236 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## year      0.8872     1.1271    0.7775    1.0124
## age       1.0300     0.9709    1.0031    1.0577
## surgery   0.3726     2.6840    0.1584    0.8761
##
## Concordance= 0.653  (se = 0.032 )
## Likelihood ratio test= 17.63  on 3 df,  p=5e-04
## Wald test             = 15.76  on 3 df,  p=0.001
## Score (logrank) test = 16.71  on 3 df,  p=8e-04
```

La table des résultats reporte le logarithme des RR (coef) ainsi que les RR (exp(coef)). Il est intéressant de regarder la valeur de concordance (Harrel's) qui donne des indications sur la qualité de l'ajustement (proche de l'AUC/ROC).

On peut représenter sous forme graphique les résultats avec la fonction `ggforest` de `survminer`

```
ggforest(coxfit)
```



### Tests de l'hypothèse PH

#### Résidus de Schoenfeld

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

Avec `transform="km"`

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

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

Avec *transform="identity"* ( $f(t) = t$ )

```
cox.zph(coxfit, transform="identity")  
  
##          chisq df      p  
## year      4.54  1 0.033  
## age       1.71  1 0.191  
## surgery   4.92  1 0.027  
## GLOBAL    9.47  3 0.024
```

### Introduction d'une interaction

Lorsque la covariable n'est pas continue, elle doit être transformée en indicatrice. Vérifier que les résultats du modèle sont bien identiques avec le modèle estimé précédemment (ne pas oublier d'omettre le niveau en référence).

Ici la variable *surgery* est déjà sous forme d'indicatrice (0,1).

La variable d'interaction est *tt(nom-variable)*, la fonction de la durée (ici forme linéaire simple) est indiquée en option de la fonction: *tt = function(x, t, ...)*  $x*t$ .

```
coxfit2 = coxph(formula = Surv(stime, died) ~ year + age + surgery +  
tt(surgery), data = trans, tt = function(x, t, ...) x*t)  
summary(coxfit2)  
  
## Call:  
## coxph(formula = Surv(stime, died) ~ year + age + surgery + tt(surg  
## ery),  
##        data = trans, tt = function(x, t, ...) x * t)  
##  
##      n= 103, number of events= 75  
##  
##              coef exp(coef)  se(coef)      z Pr(>|z|)  
## year      -0.123074  0.884198  0.066835 -1.841  0.06555 .  
## age       0.028888  1.029310  0.013449  2.148  0.03172 *  
## surgery   -1.754738  0.172953  0.674391 -2.602  0.00927 **  
## tt(surgery) 0.002231  1.002234  0.001102  2.024  0.04299 *  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
##              exp(coef) exp(-coef) lower .95 upper .95  
## year          0.8842     1.1310   0.77564   1.0080  
## age           1.0293     0.9715   1.00253   1.0568  
## surgery       0.1730     5.7819   0.04612   0.6486  
## tt(surgery)   1.0022     0.9978   1.00007   1.0044  
##  
## Concordance= 0.656  (se = 0.032 )  
## Likelihood ratio test= 21.58 on 4 df,  p=2e-04  
## Wald test          = 16.99 on 4 df,  p=0.002  
## Score (logrank) test = 19 on 4 df,  p=8e-04
```

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

### Introduction d'une variable dynamique (binaire)

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

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

#### *Etape 1*

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

#### *Etape 2*

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

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

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

## Call:
## coxph(formula = Surv(stime0, stime, died) ~ year + age + surgery,
##        data = tvc)
##
##          coef exp(coef) se(coef)     z      p
## year    -0.11963   0.88725  0.06734 -1.776 0.0757
## age      0.02958   1.03002  0.01352  2.187 0.0287
## surgery -0.98732   0.37257  0.43626 -2.263 0.0236
##
## Likelihood ratio test=17.63 on 3 df, p=0.0005243
## n= 3573, number of events= 75
```

### *Etape 3*

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

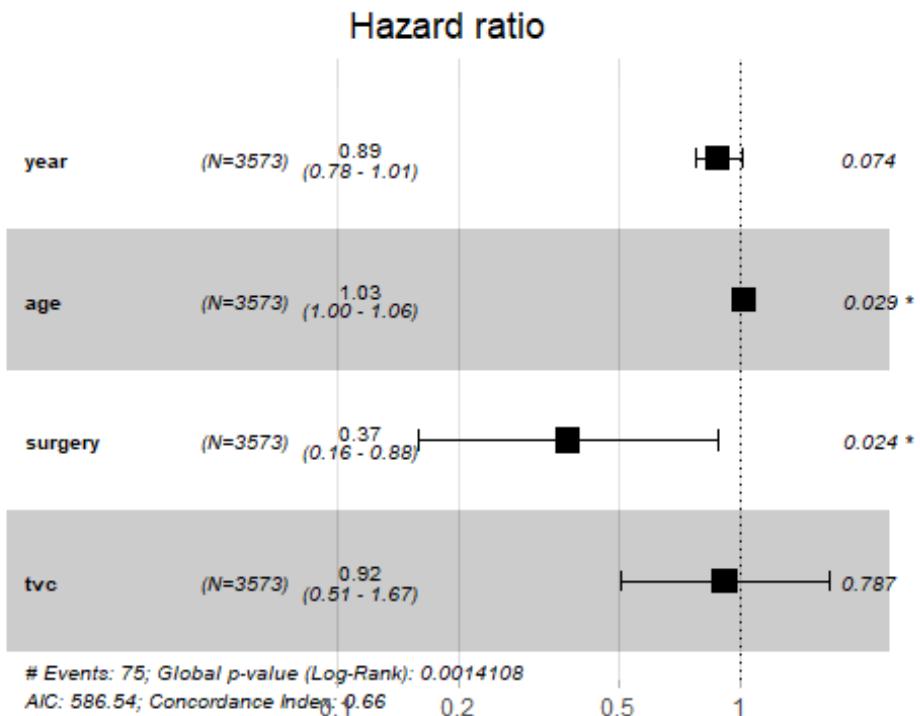
### *Estimation du modèle*

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

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

## Call:
## coxph(formula = Surv(stime0, stime, died) ~ year + age + surgery +
##         tvc, data = tvc)
##
## n= 3573, number of events= 75
##
##          coef exp(coef)  se(coef)      z Pr(>|z|)    
## year     -0.12032  0.88664  0.06734 -1.787  0.0740 .  
## age      0.03044   1.03091  0.01390  2.190  0.0285 *  
## surgery -0.98289   0.37423  0.43655 -2.251  0.0244 *  
## tvc      -0.08221   0.92108  0.30484 -0.270  0.7874  
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95    
## year      0.8866      1.128    0.7770    1.0117  
## age       1.0309      0.970    1.0032    1.0594  
## surgery   0.3742      2.672    0.1591    0.8805  
## tvc       0.9211      1.086    0.5068    1.6741  
##
## Concordance= 0.659  (se = 0.032 ) 
## Likelihood ratio test= 17.7  on 4 df,  p=0.001
## Wald test           = 15.79  on 4 df,  p=0.003
## Score (logrank) test = 16.74  on 4 df,  p=0.002

ggforest(tvcfit)
```



### Analyse en temps discret (régression logistique)

Pour la durée, on va utiliser la variable mois (en fait regroupement sur 30 jours). La fonction ***uncount*** du package *tidyverse* permettra de splitter la base au temps d'observation.

```
library("tidyverse")

dt = uncount(trans,mois)
dt = dt[order(dt$id),]

## # id| year| age| died| stime| surgery| transplant| wait| compet|
## | -- :| --- :| --- :| --- :| --- :| --- :| --- :| --- :| --- :
## | 1| 67| 30| 1| 50| 0| 0| 0| 1|
## | 1| 67| 30| 1| 50| 0| 0| 0| 1|
## | 2| 68| 51| 1| 6| 0| 0| 0| 1|
## | 3| 68| 54| 1| 16| 0| 1| 1| 1|
## | 4| 68| 40| 1| 39| 0| 1| 36| 2|
## | 4| 68| 40| 1| 39| 0| 1| 36| 2|
## | 5| 68| 20| 1| 18| 0| 0| 0| 1|
## | 6| 68| 54| 1| 3| 0| 0| 0| 2|
## | 7| 68| 50| 1| 675| 0| 1| 51| 1|
## | 7| 68| 50| 1| 675| 0| 1| 51| 1|
## | 7| 68| 50| 1| 675| 0| 1| 51| 1|
```

La variable mois a été supprimée, on va générer une variable type compteur pour mesurer la durée à chaque point d'observation. On va également recréer la variable (renommée  $T$ ).

```
dt$x=1
dt$t = ave(dt$x,dt$id, FUN=cumsum)
dt$T = ave(dt$x,dt$id, FUN=sum)
```

##	id	year	age	died	stime	surgery	transplant	wait	compet	x	t	T
##	1	67	30	1	50	0	0	0	1	1	1	2
##	1	67	30	1	50	0	0	0	1	1	2	2
##	2	68	51	1	6	0	0	0	1	1	1	1
##	3	68	54	1	16	0	1	1	1	1	1	1
##	4	68	40	1	39	0	1	36	2	1	1	2
##	4	68	40	1	39	0	1	36	2	1	2	2
##	5	68	20	1	18	0	0	0	1	1	1	1
##	6	68	54	1	3	0	0	0	2	1	1	1
##	7	68	50	1	675	0	1	51	1	1	1	23
##	7	68	50	1	675	0	1	51	1	1	2	23
##	7	68	50	1	675	0	1	51	1	1	3	23

Si un individu est décédé, died=1 est reporté sur toute les lignes (idem qu'avec la variable dynamique). On va modifier la variable tel que died=0 si  $t < T$

```
dt$died[dt$t<dt$T]=0
```

##	id	year	age	died	stime	surgery	transplant	wait	compet	x	t	T
##	1	67	30	0	50	0	0	0	1	1	1	2
##	1	67	30	1	50	0	0	0	1	1	2	2
##	2	68	51	1	6	0	0	0	1	1	1	1
##	3	68	54	1	16	0	1	1	1	1	1	1
##	4	68	40	0	39	0	1	36	2	1	1	2
##	4	68	40	1	39	0	1	36	2	1	2	2
##	5	68	20	1	18	0	0	0	1	1	1	1
##	6	68	54	1	3	0	0	0	2	1	1	1
##	7	68	50	0	675	0	1	51	1	1	1	23
##	7	68	50	0	675	0	1	51	1	1	2	23
##	7	68	50	0	675	0	1	51	1	1	3	23

## *Estimation avec durée comme variable continue*

Avec un effet quadratique d'ordre 2.

```
dt$t2=dt$t^2
dtfit = glm(died ~ t + t2 + year + age + surgery, data=dt, family="binomial")
summ(dtfit)

## MODEL INFO:
## Observations: 1127
## Dependent Variable: died
## Type: Generalized linear model
## Family: binomial
## Link function: logit
##
## MODEL FIT:
## <U+03C7>2(5) = 84.70, p = 0.00
## Pseudo-R2 (Cragg-Uhler) = 0.19
## Pseudo-R2 (McFadden) = 0.15
## AIC = 478.67, BIC = 508.83
##
## Standard errors: MLE
## -----
##           Est.   S.E.   z val.      p
## -----
## (Intercept) 7.74  5.22  1.48  0.14
## t          -0.20  0.04 -5.63  0.00
## t2          0.00  0.00  3.98  0.00
## year        -0.15  0.07 -2.04  0.04
## age          0.03  0.01  2.35  0.02
## surgery     -1.00  0.45 -2.24  0.02
## -----
```

Remarque sur la présentation: la fonction *summ* est intégrée au package *jtools*. Elle ne fonctionne qu'avec le package *Recordlinkage* qui doit être installé et chargé.

### *Estimation avec durée comme variable discrète*

On va créer une variable discrète regroupant la variable t en quartile (pour l'exemple seulement, tous types de regroupement est envisageable).

On va utiliser la fonction *quantcut* du package *gtools*.

```
dt$ct4 <- quantcut(dt$t)
table(dt$ct4)

## 
##   [1,4]  (4,11] (11,23] (23,60]
##   299     275     282     271
```

On va générer un compteur et un total d'observations sur la strate regroupant *id* et *ct4*.

```
dt$n = ave(dt$x, dt$id, dt$ct4, FUN=cumsum)
dt$N = ave(dt$x, dt$id, dt$ct4, FUN=sum)
```

On conserve la dernière observation pour chaque id.

```
dt2 = subset(dt, n==N)
```

### *Estimation du modèle*

```
fit = glm(died ~ ct4 + year + age + surgery, data=dt2, family=binomial)
summ(fit)

## MODEL INFO:
## Observations: 197
## Dependent Variable: died
## Type: Generalized linear model
##   Family: binomial
##   Link function: logit
##
## MODEL FIT:
## <U+03C7>2(6) = 39.30, p = 0.00
## Pseudo-R2 (Cragg-Uhler) = 0.25
## Pseudo-R2 (McFadden) = 0.15
## AIC = 236.48, BIC = 259.46
##
## Standard errors: MLE
## -----
##           Est.  S.E.  z val.    p
## -----
## (Intercept) 12.45  6.65   1.87  0.06
## ct4(4,11]   -1.03  0.42  -2.47  0.01
## ct4(11,23]  -1.62  0.54  -2.96  0.00
```

```

## ct4(23,60]      -0.48   0.60    -0.80   0.42
## year          -0.20   0.09    -2.18   0.03
## age            0.05   0.02     2.53   0.01
## surgery        -1.11   0.50    -2.21   0.03
## -----

```

## Modèles paramétrique usuels

On utilise la fonction `survreg` du package `survival`

### *Modèle de Weibull*

*De type AFT*

```

weibull = survreg(formula = Surv(stime, died) ~ year + age + surgery,
  data = trans, dist="weibull")
summary(weibull)

##
## Call:
## survreg(formula = Surv(stime, died) ~ year + age + surgery, data =
##   trans,
##   dist = "weibull")
##           Value Std. Error      z      p
## (Intercept) -3.0220    8.7284 -0.35  0.729
## year        0.1620    0.1218  1.33  0.184
## age         -0.0615    0.0247 -2.49  0.013
## surgery      1.9703    0.7794  2.53  0.011
## Log(scale)   0.5868    0.0927  6.33 2.5e-10
##
## Scale= 1.8
##
## Weibull distribution
## Loglik(model)= -488.2  Loglik(intercept only)= -497.6
## Chisq= 18.87 on 3 degrees of freedom, p= 0.00029
## Number of Newton-Raphson Iterations: 5
## n= 103

```

### De type PH (risques proportionnels)

La paramétrisation PH n'est pas possible avec la fonction `survreg`. Il faut utiliser le package `flexsurv`, qui permet également d'estimer les modèles paramétriques disponibles avec `survival`.

```
library(flexsurv)
```

Pour estimer le modèle de Weibull de type PH, on utilise en option `dist="weibullPH"`.

```
flexsurvreg(formula = Surv(stime, died) ~ year + age + surgery, data = trans, dist="weibullPH")

## Call:
## flexsurvreg(formula = Surv(stime, died) ~ year + age + surgery,
##             data = trans, dist = "weibullPH")
##
## Estimates:
##          data mean   est      L95%     U95%       se    exp(est)
## shape        NA 5.56e-01 4.64e-01 6.67e-01 5.16e-02        NA
## scale        NA 5.37e+00 4.27e-04 6.75e+04 2.59e+01        NA
## year    7.06e+01 -9.01e-02 -2.20e-01 3.97e-02 6.62e-02 9.14e-01
## age     4.46e+01 3.42e-02 7.09e-03 6.13e-02 1.38e-02 1.03e+00
## surgery 1.17e-01 -1.10e+00 -1.95e+00 -2.45e-01 4.34e-01 3.34e-01

##          L95%     U95%
## shape        NA        NA
## scale        NA        NA
## year    8.03e-01 1.04e+00
## age     1.01e+00 1.06e+00
## surgery 1.43e-01 7.83e-01
##
## N = 103, Events: 75, Censored: 28
## Total time at risk: 31938
## Log-likelihood = -488.1683, df = 5
## AIC = 986.3366
```

## Risques concurrents

Package `cmprsk` pour l'analyse non paramétrique et le modèle de Fine-Gray.

La variable de censure/événement, `compet`, correspond à la variable `died` avec une modalité supplémentaire simulée. On suppose l'existence d'une cause supplémentaire au décès autre qu'une malformation cardiaque et non strictement indépendante de celle-ci.

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

```

# variable compet


```

### Non paramétrique: incidences cumulées

On utilise la fonction **cuminc** du package *cmprsk*.

*Pas de comparaison*

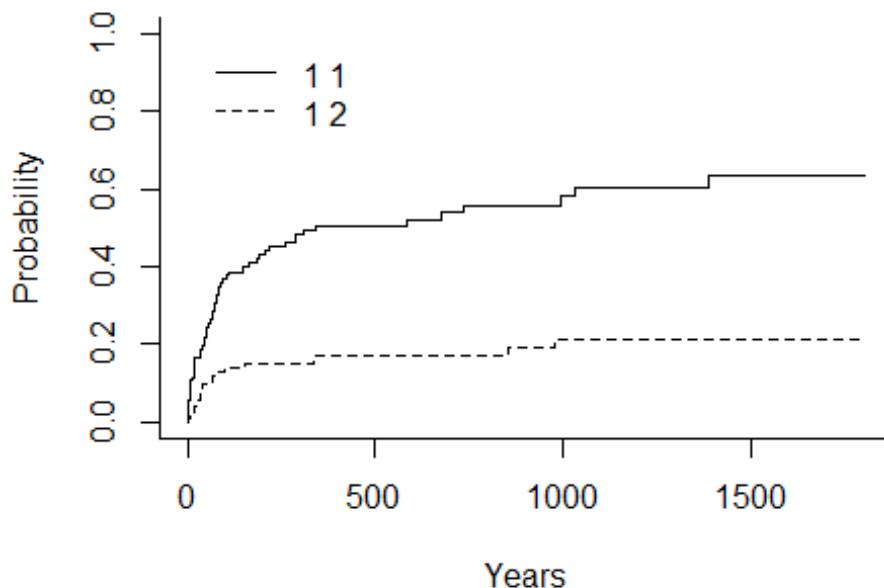
```

ic = cuminc(compet$stime, compet$compet)
ic

## Estimates and Variances:
## $est
##      500     1000     1500
## 1 1 0.5067598 0.5808345 0.6340038
## 1 2 0.1720161 0.2140841 0.2140841
##
## $var
##      500     1000     1500
## 1 1 0.002619449 0.003131847 0.003676516
## 1 2 0.001473283 0.002203770 0.002203770

plot(ic)

```



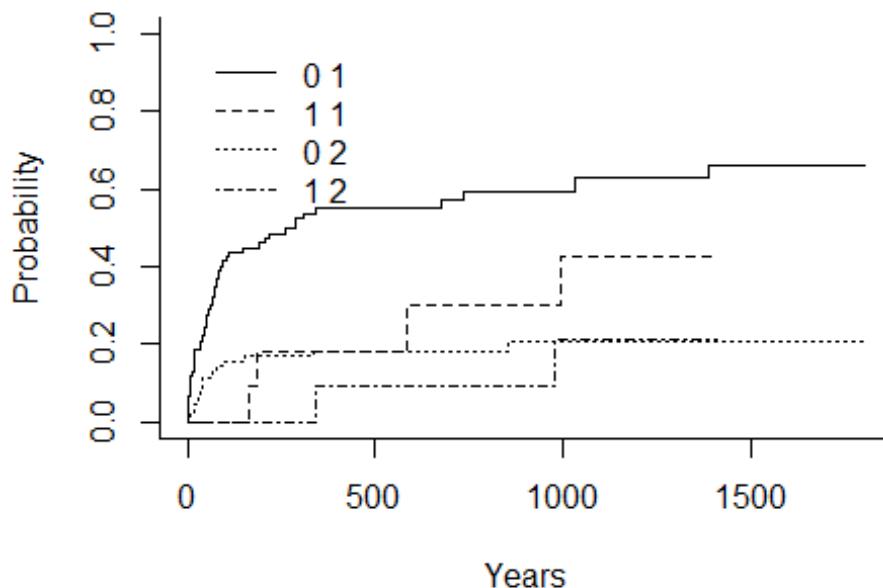
### *Comparaison (variable surgery)*

Le test de Gray est automatiquement exécuté.

```
ic = cuminc(compet$stime, compet$compet, group=compet$surgery, rho=1)
ic

## Tests:
##      stat      pv df
## 1 4.604792 0.03188272 1
## 2 0.272147 0.60189515 1
## Estimates and Variances:
## $est
##           500      1000      1500
## 0 1 0.54917896 0.5940358 0.6604903
## 1 1 0.18181818 0.4242424 NA
## 0 2 0.18168014 0.2066006 0.2066006
## 1 2 0.09090909 0.2121212 NA
##
## $var
##           500      1000      1500
## 0 1 0.002955869 0.003335897 0.004199157
## 1 1 0.014958678 0.033339569 NA
## 0 2 0.001727112 0.002271242 0.002271242
## 1 2 0.008449138 0.022024737 NA

plot(ic)
```



## Modèles

### *Modèle de Fine-Gray*

Attention à l'interprétation des « risks ratio ».

Fonction *crr* du package *cmprsk*.

Peu pratique les covariables doivent être introduite sous forme de matrice (n observations \* k variables). Pour les covariables discrètes, prévoir la forme binaire et l'omission de la catégorie de référence.

```

c <- compet[c("year", "age", "surgery")]
c = as.matrix(c)
summary(crr(stime, compet$competing, c))

## Competing Risks Regression
##
## Call:
## crrftime = compet$stime, fstatus = compet$competing, cov1 = c)
##
##          coef exp(coef) se(coef)     z p-value
## year    -0.0724    0.930   0.0713 -1.02  0.310
## age      0.0370    1.038   0.0176  2.10  0.036
## surgery -0.8688    0.419   0.4488 -1.94  0.053
##
```

```

##          exp(coef) exp(-coef)   2.5% 97.5%
## year        0.930      1.075  0.809  1.07
## age         1.038      0.964  1.002  1.07
## surgery     0.419      2.384  0.174  1.01
##
## Num. cases = 103
## Pseudo Log-likelihood = -228
## Pseudo likelihood ratio test = 11.2 on 3 df,

```

### *Modèle logistique multinomial à temps discret*

On va de nouveau utiliser la variable mois (temps discret). Le modèle sera estimé à l'aide la fonction *multinom* du package *nnet*, les p-values doivent-être programmées, l'output ne donnant que les erreurs-types.

```
#install.packages("nnet")
library(nnet)
```

### *Transformation de la base*

```

compet$T = compet$mois
td      = uncount(compet, mois)
td$x    = 1
td$t    = ave(td$x, td$id, FUN=cumsum)
td$t2   = td$t*td$t
td$e    = ifelse(td$t<td$T, 0, td$compet)

```

### *Estimation*

Pour estimer le modèle, on utilise la fonction *mlogit* du package *nnet*. Les p-values seront calculées à partir d'un test bilatéral (statistique z).

```

competfit = multinom(formula = e ~ t + t2 + year + age + surgery, dat
a = td)

## # weights: 21 (12 variable)
## initial value 1238.136049
## iter 10 value 579.836377
## iter 20 value 387.824428
## iter 30 value 277.775477
## iter 40 value 275.151304
## iter 50 value 275.005454
## final value 275.005419
## converged

summary(competfit)

```

```

## Call:
## multinom(formula = e ~ t + t2 + year + age + surgery, data = td)
##
## Coefficients:
## (Intercept)          t          t2        year        age        surgery
## 1   5.646796 -0.2034545 0.003168301 -0.1284492 0.04389670 -1.1471457
## 2  11.316123 -0.2023134 0.002981492 -0.2036701 0.01100051 -0.6137982
##
## Std. Errors:
## (Intercept)          t          t2        year        age        surgery
## 1  0.007756736 0.04130290 0.0008964768 0.01166713 0.01688253 0.5319373
## 2  0.013113980 0.06898341 0.0015340216 0.01613615 0.02404470 0.7612701
##
## Residual Deviance: 550.0108
## AIC: 574.0108

z = summary(competfit)$coefficients/summary(competfit)$standard.error
s
p <- (1 - pnorm(abs(z), 0, 1)) ^ 2
p

## (Intercept)          t          t2        year        age        surgery
## 1           0 8.396729e-07 0.0004090588      0 0.009318953 0.03104129
## 2           0 3.359389e-03 0.0519462468      0 0.647310003 0.42008041

```

## Stata

Ouverture de la base.

```
webuse set "https://raw.githubusercontent.com/mthevenin/analyse_duree/master/bases/"
webuse "transplantation_m", clear
webuse set
```

```
list in 1/10
```

	id	year	age	died	stime	surgery	transp~t	wait	mois	compet
1.	15	68	53	1	1	0	0	0	1	1
2.	43	70	43	1	2	0	0	0	1	1
3.	61	71	52	1	2	0	0	0	1	1
4.	75	72	52	1	2	0	0	0	1	1
5.	6	68	54	1	3	0	0	0	1	2
6.	42	70	36	1	3	0	0	0	1	1
7.	54	71	47	1	3	0	0	0	1	1
8.	38	70	41	1	5	0	1	5	1	1
9.	85	73	47	1	5	0	0	0	1	1
10.	2	68	51	1	6	0	0	0	1	1

## Analyse non paramétrique

### Méthode actuarielle

Contrairement à la formation, l'estimation sera faite sur des intervalles de 30 jours.

```
ltable stime died, interval(30) graph
```

Interval	Beg.	Total	Deaths	Lost	Survival	Std. Error	[95% Conf. Int.]
0	30	103	22	1	0.7854	0.0406	0.6926 0.8531
30	60	80	14	2	0.6462	0.0475	0.5449 0.7305
60	90	64	12	0	0.5250	0.0498	0.4232 0.6171
90	120	52	5	1	0.4741	0.0499	0.3738 0.5677
120	150	46	1	1	0.4636	0.0499	0.3637 0.5575
150	180	44	2	0	0.4426	0.0498	0.3435 0.5369
180	210	42	3	1	0.4106	0.0495	0.3132 0.5053
210	240	38	1	0	0.3998	0.0494	0.3030 0.4945
240	270	37	1	1	0.3888	0.0492	0.2928 0.4836
270	300	35	2	0	0.3666	0.0488	0.2720 0.4614
300	330	33	1	0	0.3555	0.0486	0.2618 0.4502
330	360	32	3	1	0.3216	0.0478	0.2308 0.4157
360	390	28	0	1	0.3216	0.0478	0.2308 0.4157
390	420	27	0	1	0.3216	0.0478	0.2308 0.4157
420	450	26	0	2	0.3216	0.0478	0.2308 0.4157

480	510	24	0	1	0.3216	0.0478	0.2308	0.4157
510	540	23	0	1	0.3216	0.0478	0.2308	0.4157
540	570	22	0	1	0.3216	0.0478	0.2308	0.4157
570	600	21	1	1	0.3059	0.0479	0.2155	0.4010
600	630	19	0	1	0.3059	0.0479	0.2155	0.4010
660	690	18	1	1	0.2885	0.0483	0.1982	0.3849
720	750	16	1	0	0.2704	0.0485	0.1807	0.3681
840	870	15	1	1	0.2518	0.0486	0.1629	0.3506
900	930	13	0	1	0.2518	0.0486	0.1629	0.3506
930	960	12	0	1	0.2518	0.0486	0.1629	0.3506
960	990	11	1	0	0.2289	0.0493	0.1404	0.3304
990	1020	10	1	0	0.2060	0.0494	0.1192	0.3093
1020	1050	9	1	0	0.1831	0.0489	0.0992	0.2873
1140	1170	8	0	1	0.1831	0.0489	0.0992	0.2873
1320	1350	7	0	1	0.1831	0.0489	0.0992	0.2873
1380	1410	6	1	2	0.1465	0.0510	0.0645	0.2602
1560	1590	3	0	2	0.1465	0.0510	0.0645	0.2602
1770	1800	1	0	1	0.1465	0.0510	0.0645	0.2602

### Récupération des quartiles de la durée

Installation de la commande `qlt`

```
net install qlt, from("https://mthevenin.github.io/analyse_duree/ado/qlt/")
replace
help qlt
```

```
ltable stime died, interval(30) saving(base, replace)
use base, clear

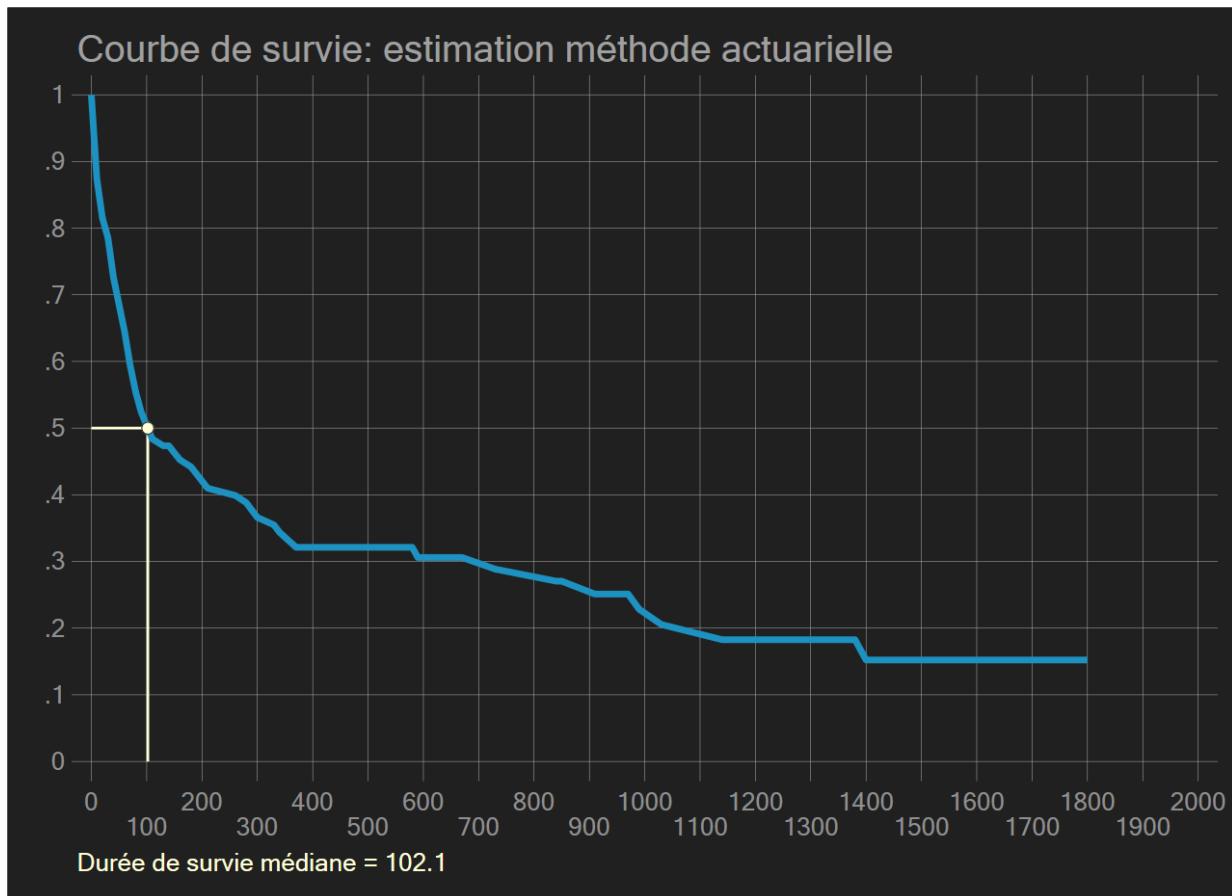
qlt

Durée pour différents quantiles de la fonction de survie
Définition des bornes Stata-ltable
S(t)=0.90: t=      .
S(t)=0.75: t=    7.623
S(t)=0.50: t=   74.729
S(t)=0.25: t= 849.325
S(t)=0.10: t=      .
```

Avec la définition des bornes des intervalles de Sas

```
qlt, sas

Duree pour differents quantiles de la fonction de survie
Definition des bornes Sas-lifetest
S(t)=0.90: t= 13.977
S(t)=0.75: t= 37.623
S(t)=0.50: t= 104.729
S(t)=0.25: t= 906.993
S(t)=0.10: t=      .
```



### **Méthode Kaplan-Meier**

Les données doivent être mises en mode analyse de durée (*help stset*)

```
qui use "D:\Marc\SMS\FORMATIONS\2017\analyse biographique\A distribuer\transplantation.dta", clear
stset stime, f(died)

    failure event:  (assumed to fail at time=stime)
obs. time interval:  (0, stime]
exit on or before:  failure

-----
103  total observations
    0  exclusions
-----
103  observations remaining, representing
103  failures in single-record/single-failure data
31,938  total analysis time at risk and under observation
                    at risk from t =      0
                    earliest observed entry t =      0
                    last observed exit t = 1,799
```

**list in 1/10**

+-----+

	id	year	age	died	stime	surgery	transp~t	wait	_st	_d	_t	_t0
1.	1	67	30	1	50	0	0	0	1	1	50	0
2.	2	68	51	1	6	0	0	0	1	1	6	0
3.	3	68	54	1	16	0	1	1	1	1	16	0
4.	4	68	40	1	39	0	1	36	1	1	39	0
5.	5	68	20	1	18	0	0	0	1	1	18	0
6.	6	68	54	1	3	0	0	0	1	1	3	0
7.	7	68	50	1	675	0	1	51	1	1	675	0
8.	8	68	45	1	40	0	0	0	1	1	40	0
9.	9	68	47	1	85	0	0	0	1	1	85	0
10.	10	68	42	1	58	0	1	12	1	1	58	0

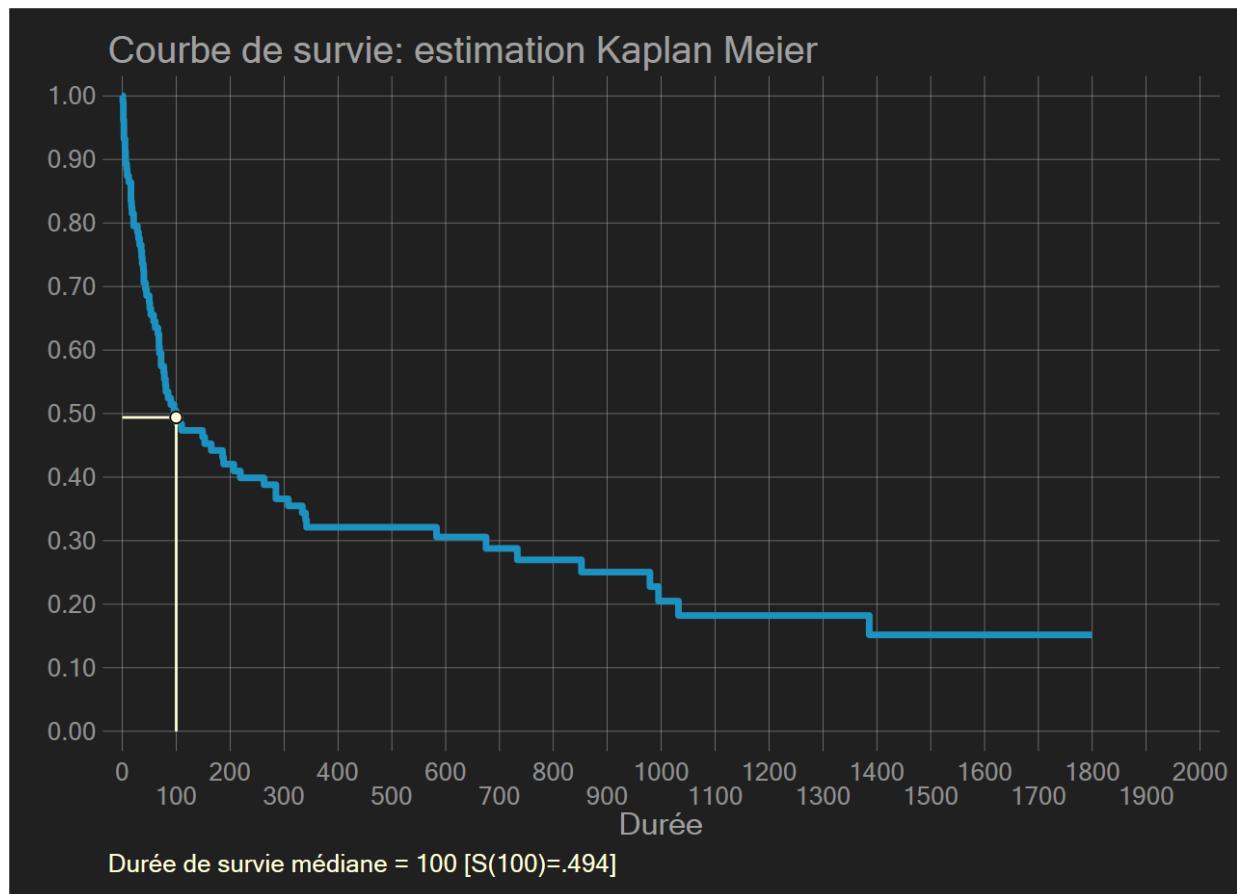
### Estimation de la fonction de survie

sts list

failure _d: died analysis time _t: stime							
Time	At Risk	Fail	Lost	Survivor Function	Std. 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

sts graph

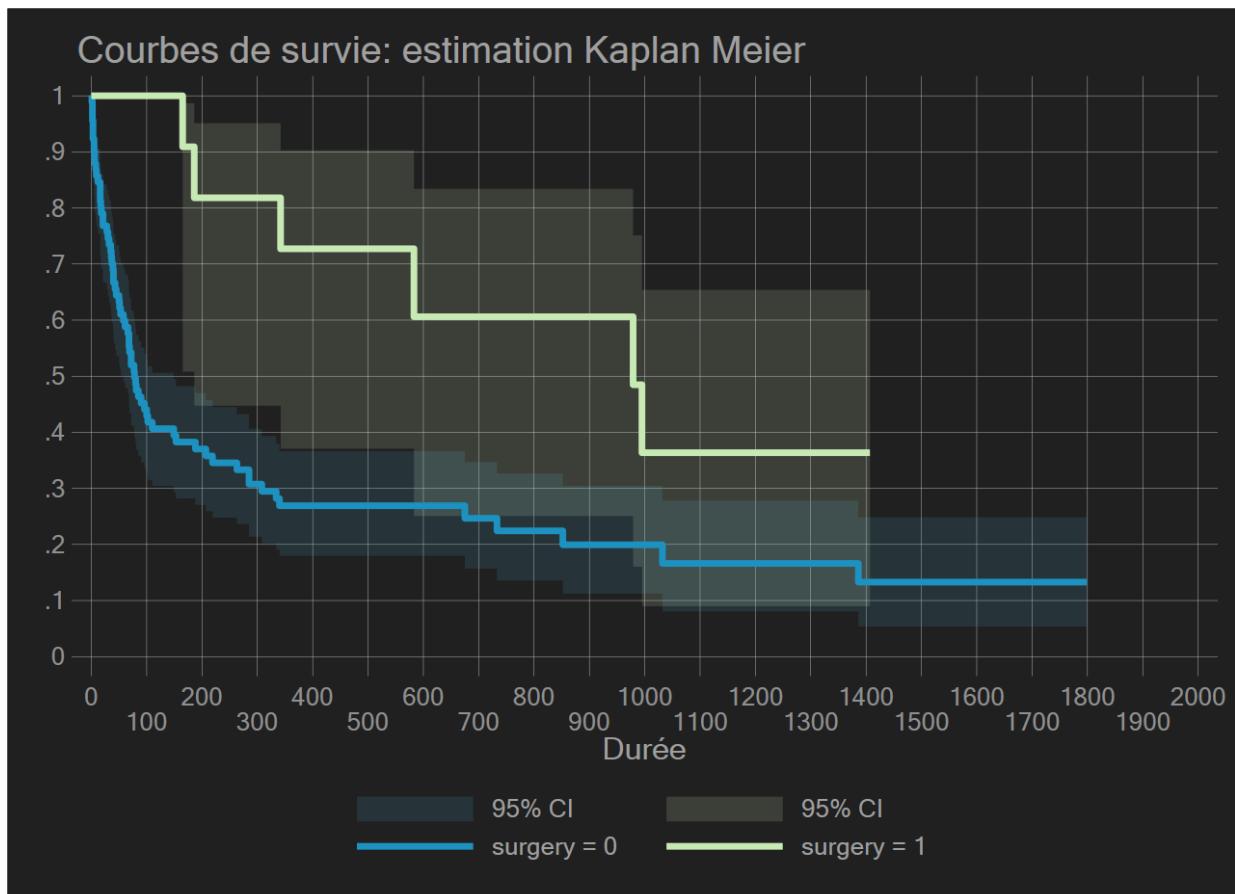


### ***Comparaison des fonctions de survie***

#### ***Tests du log rank***

On va comparer les fonctions de survie pour la variable *surgery*.

```
sts graph, by(surgery)
```



Tests du log rank: fonction ***sts test***. On affichera ici plusieurs variantes du test.

```
local test `" "l" "w" "tw" "p" "'"
foreach test2 of local test {
  sts test surgery, `test2'
}
```

Log-rank test for equality of survivor functions

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

chi2(1) = 6.59  
Pr>chi2 = 0.0103

failure \_d: died  
analysis time \_t: stime

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

failure \_d: died  
 analysis time \_t: stime

#### Tarone-Ware test for equality of survivor functions

surgery	Events observed	Events expected	Sum of ranks
0	69	60.34	73.105398
1	6	14.66	-73.105398
Total	75	75.00	0

chi2(1) = 8.46  
 Pr>chi2 = 0.0036

failure \_d: died  
 analysis time \_t: stime

#### Peto-Peto test for equality of survivor functions

surgery	Events observed	Events expected	Sum of ranks
0	69	60.34	6.0505875
1	6	14.66	-6.0505875
Total	75	75.00	0

chi2(1) = 8.66  
 Pr>chi2 = 0.0032

### *Comparaison des rmst*

Installation de la commande strmst2:

```
ssc install strmst2
```

arm1 = opération  
 arm0 = pas d'opération

## strmst2 surgery

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	310.168	43.160	225.576 394.760

Between-group contrast (arm 1 versus arm 0)

Contrast	Estimate	[95% Conf. Interval]	P> z
RMST (arm 1 - arm 0)	424.590	149.641 699.539	0.002
RMST (arm 1 / arm 0)	2.369	1.513 3.710	0.000

## Risques proportionnels

*Modèle semi paramétrique de Cox*

*Estimation du modèle*

Avec la correction d'Efron

**stcox year age surgery, nolog noshow efron**

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

**stcox year age surgery, nolog noshow efron nohr**

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

### Test de l'hypothèse PH

#### Test Grambsch-Therneau sur les résidus de Schoenfeld

\* f(t)=t - par défaut

**estat phtest, detail**

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

\* f(t)= 1-km - solution par défaut de R

**estat phtest, detail km**

Test of proportional-hazards assumption

Time: Kaplan-Meier

	rho	chi2	df	Prob>chi2
year	0.15920	1.96	1	0.1620
age	0.10907	1.15	1	0.2845
surgery	0.25096	3.96	1	0.0465
global test		7.99	3	0.0462

Intéraction avec une fonction de la durée :  $f(t) = t$

**stcox year age surgery, nolog noshow efron nohr tvc(surgery) texp(\_t)**

Cox regression -- Efron method for ties

No. of subjects =	103	Number of obs =	103
-------------------	-----	-----------------	-----

```

No. of failures =          75
Time at risk     =      31938
Log likelihood   = -287.32903
                                         LR chi2(4)      =      21.58
                                         Prob > chi2    =     0.0002

-----+
_t | Coef.  Std. Err.      z  P>|z|  [95% Conf. Interval]
-----+
main |
  year | -0.1231  0.0668  -1.84  0.066  -0.2541  0.0079
        age |  0.0289  0.0134   2.15  0.032  0.0025  0.0552
        surgery | -1.7547  0.6744  -2.60  0.009  -3.0765 -0.4330
-----+
tvc  |
  surgery |  0.0022  0.0011   2.02  0.043  0.0001  0.0044
-----+
Note: Variables in tvc equation interacted with _t.

```

### ***Introduction d'une variable dynamique (binaire)***

Transformation de la base en format long aux temps d'évènement :

#### ***Modèle de Cox***

##### **Etape 1**

```

stset stime, f(died) id(id)

      id: id
failure event: died != 0 & died < .
obs. time interval: (stime[_n-1], stime]
exit on or before: failure

-----+
103  total observations
  0  exclusions
-----+
103  observations remaining, representing
103  subjects
  75  failures in single-failure-per-subject data
31,938  total analysis time at risk and under observation
                     at risk from t =      0
                     earliest observed entry t =      0
                     last observed exit t =  1,799

```

## Etape 2

```
stssplit, at(failure)
(62 failure times)
(3,470 observations (episodes) created)
```

```
sort id _t
```

```
list in 1/23
```

	id	year	age	died	stime	surgery	transp~t	wait	mois	compet	_st	_d	_t	_t0
1.	1	67	30	.	1	0	0	0	2	1	1	0	1	0
2.	1	67	30	.	2	0	0	0	2	1	1	0	2	1
3.	1	67	30	.	3	0	0	0	2	1	1	0	3	2
4.	1	67	30	.	5	0	0	0	2	1	1	0	5	3
5.	1	67	30	.	6	0	0	0	2	1	1	0	6	5
6.	1	67	30	.	8	0	0	0	2	1	1	0	8	6
7.	1	67	30	.	9	0	0	0	2	1	1	0	9	8
8.	1	67	30	.	12	0	0	0	2	1	1	0	12	9
9.	1	67	30	.	16	0	0	0	2	1	1	0	16	12
10.	1	67	30	.	17	0	0	0	2	1	1	0	17	16
11.	1	67	30	.	18	0	0	0	2	1	1	0	18	17
12.	1	67	30	.	21	0	0	0	2	1	1	0	21	18
13.	1	67	30	.	28	0	0	0	2	1	1	0	28	21
14.	1	67	30	.	30	0	0	0	2	1	1	0	30	28
15.	1	67	30	.	32	0	0	0	2	1	1	0	32	30
16.	1	67	30	.	35	0	0	0	2	1	1	0	35	32
17.	1	67	30	.	36	0	0	0	2	1	1	0	36	35
18.	1	67	30	.	37	0	0	0	2	1	1	0	37	36
19.	1	67	30	.	39	0	0	0	2	1	1	0	39	37
20.	1	67	30	.	40	0	0	0	2	1	1	0	40	39
21.	1	67	30	.	43	0	0	0	2	1	1	0	43	40
22.	1	67	30	.	45	0	0	0	2	1	1	0	45	43
23.	1	67	30	1	50	0	0	0	2	1	1	1	50	45

## Etape 3

```
gen tvc = transplant==1 & wait<=_t
sort id _t
list id transplant wait tvc _d _t _t0 if id==10 , noobs
```

id	transp~t	wait	tvc	_d	_t	_t0
10	1	12	0	0	1	0
10	1	12	0	0	2	1
10	1	12	0	0	3	2
10	1	12	0	0	5	3
10	1	12	0	0	6	5
10	1	12	0	0	8	6
10	1	12	0	0	9	8
10	1	12	1	0	12	9
10	1	12	1	0	16	12
10	1	12	1	0	17	16

10	1	12	1	0	18	17
10	1	12	1	0	21	18
10	1	12	1	0	28	21
10	1	12	1	0	30	28
10	1	12	1	0	32	30
-----						
10	1	12	1	0	35	32
10	1	12	1	0	36	35
10	1	12	1	0	37	36
10	1	12	1	0	39	37
10	1	12	1	0	40	39
-----						
10	1	12	1	0	43	40
10	1	12	1	0	45	43
10	1	12	1	0	50	45
10	1	12	1	0	51	50
10	1	12	1	0	53	51
-----						
10	1	12	1	1	58	53
-----+-----						

### Estimation du modèle

```
stcox year age surgery tvc, nolog noshow efron nohr
```

Cox regression -- Efron method for ties

No. of subjects =	103	Number of obs =	3,573
No. of failures =	75		
Time at risk =	31938	LR chi2(4) =	17.70
Log likelihood =	-289.27014	Prob > chi2 =	0.0014

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
year	-0.1203	0.0673	-1.79	0.074	-0.2523 0.0117
age	0.0304	0.0139	2.19	0.029	0.0032 0.0577
surgery	-0.9829	0.4366	-2.25	0.024	-1.8385 -0.1273
tvc	-0.0822	0.3048	-0.27	0.787	-0.6797 0.5153

## Modèle (logistique) à temps discret

Variable de durée = mois

Mise en forme

```
expand mois
bysort id: gen t=_n
gen e = died
replace e=0 if t<mois

* list in 1/31
```

	id	year	age	died	stime	surgery	transp~t	wait	mois	compet	t	e
1.	1	67	30	1	50	0	0	0	2	1	1	0
2.	1	67	30	1	50	0	0	0	2	1	2	1
3.	2	68	51	1	6	0	0	0	1	1	1	1
4.	3	68	54	1	16	0	1	1	1	1	1	1
5.	4	68	40	1	39	0	1	36	2	2	1	0
6.	4	68	40	1	39	0	1	36	2	2	2	1
7.	5	68	20	1	18	0	0	0	1	1	1	1
8.	6	68	54	1	3	0	0	0	1	2	1	1
9.	7	68	50	1	675	0	1	51	23	1	1	0
10.	7	68	50	1	675	0	1	51	23	1	2	0
11.	7	68	50	1	675	0	1	51	23	1	3	0
12.	7	68	50	1	675	0	1	51	23	1	4	0
13.	7	68	50	1	675	0	1	51	23	1	5	0
14.	7	68	50	1	675	0	1	51	23	1	6	0
15.	7	68	50	1	675	0	1	51	23	1	7	0
16.	7	68	50	1	675	0	1	51	23	1	8	0
17.	7	68	50	1	675	0	1	51	23	1	9	0
18.	7	68	50	1	675	0	1	51	23	1	10	0
19.	7	68	50	1	675	0	1	51	23	1	11	0
20.	7	68	50	1	675	0	1	51	23	1	12	0
21.	7	68	50	1	675	0	1	51	23	1	13	0
22.	7	68	50	1	675	0	1	51	23	1	14	0
23.	7	68	50	1	675	0	1	51	23	1	15	0
24.	7	68	50	1	675	0	1	51	23	1	16	0
25.	7	68	50	1	675	0	1	51	23	1	17	0
26.	7	68	50	1	675	0	1	51	23	1	18	0
27.	7	68	50	1	675	0	1	51	23	1	19	0
28.	7	68	50	1	675	0	1	51	23	1	20	0
29.	7	68	50	1	675	0	1	51	23	1	21	0
30.	7	68	50	1	675	0	1	51	23	1	22	0
31.	7	68	50	1	675	0	1	51	23	1	23	1

## Paramétrisation avec durée continue

### Les critères d'information

```
gen t2=t^2
gen t3=t^3

logit e t      , nolog

Logistic regression                               Number of obs     =    1,127
                                                LR chi2(1)      =     50.85
                                                Prob > chi2     =     0.0000
Log likelihood = -250.26058                      Pseudo R2       =     0.0922

-----+
          e |   Coef.   Std. Err.      z   P>|z|   [95% Conf. Interval]
-----+
          t |  -0.1007   0.0185    -5.45   0.000    -0.1370    -0.0645
      _cons |  -1.6436   0.1724    -9.53   0.000    -1.9815    -1.3057
-----+-----+-----+-----+-----+-----+-----+
```

```
estat ic

Akaike's information criterion and Bayesian information criterion

-----+
          Model |           N   ll(null)   ll(model)      df        AIC        BIC
-----+
          . |     1,127  -275.6841  -250.2606      2    504.5212  514.5758
-----+
```

Note: BIC uses N = number of observations. See [R] BIC note.

```
logit e t t2      , nolog

Logistic regression                               Number of obs     =    1,127
                                                LR chi2(2)      =     65.25
                                                Prob > chi2     =     0.0000
Log likelihood = -243.05761                      Pseudo R2       =     0.1183

-----+
          e |   Coef.   Std. Err.      z   P>|z|   [95% Conf. Interval]
-----+
          t |  -0.2172   0.0357    -6.08   0.000    -0.2872    -0.1471
          t2 |   0.0034   0.0008     4.50   0.000     0.0019    0.0049
      _cons |  -1.2326   0.1925    -6.40   0.000    -1.6098    -0.8554
-----+-----+-----+-----+-----+-----+-----+
```

### estat ic

```
Akaike's information criterion and Bayesian information criterion
```

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	1,127	-275.6841	-243.0576	3	492.1152	507.1972

Note: BIC uses N = number of observations. See [R] BIC note.

**logit e t t2 t3 , nolog**

Logistic regression

Number of obs	=	1,127
LR chi2(3)	=	72.86
Prob > chi2	=	0.0000
Pseudo R2	=	0.1321

Log likelihood = -239.25267

e	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
t	-0.4038	0.0819	-4.93	0.000	-0.5643 -0.2434
t2	0.0157	0.0050	3.14	0.002	0.0059 0.0254
t3	-0.0002	0.0001	-2.31	0.021	-0.0003 -0.0000
_cons	-0.8250	0.2406	-3.43	0.001	-1.2965 -0.3536

**estat ic**

Akaike's information criterion and Bayesian information criterion

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	1,127	-275.6841	-239.2527	4	486.5053	506.6146

Note: BIC uses N = number of observations. See [R] BIC note.

### Estimation du modèle

**logit e t t2 t3 year age surgery, nolog**

Logistic regression

Number of obs	=	1,127
LR chi2(6)	=	90.69
Prob > chi2	=	0.0000
Pseudo R2	=	0.1645

Log likelihood = -230.33671

e	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
t	-0.3721	0.0824	-4.52	0.000	-0.5335 -0.2106
t2	0.0142	0.0050	2.83	0.005	0.0044 0.0241
t3	-0.0002	0.0001	-2.11	0.035	-0.0003 -0.0000
year	-0.1327	0.0738	-1.80	0.072	-0.2773 0.0119
age	0.0333	0.0147	2.27	0.023	0.0046 0.0621
surgery	-1.0109	0.4486	-2.25	0.024	-1.8902 -0.1317

_cons	7.0827	5.3077	1.33	0.182	-3.3203	17.4856
-------	--------	--------	------	-------	---------	---------

### Paramétrisation avec durée discrète

Pour l'exemple seulement, on prendra des intervalles découpés sur les quartiles de la durée.

```
xtile ct4=t, n(4)
bysort id ct4: keep if _n==_N
```

```
tab ct4 e
```

quantiles 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

```
logit e i.ct4 year age surgery, nolog
```

```
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	-1.0334	0.4189	-2.47	0.014	-1.8543 -0.2124
3	-1.6152	0.5449	-2.96	0.003	-2.6831 -0.5473
4	-0.4789	0.5993	-0.80	0.424	-1.6535 0.6957
year	-0.2032	0.0932	-2.18	0.029	-0.3859 -0.0206
age	0.0469	0.0185	2.53	0.011	0.0106 0.0831
surgery	-1.1102	0.5026	-2.21	0.027	-2.0952 -0.1252
_cons	12.4467	6.6537	1.87	0.061	-0.5943 25.4877

## Modèles paramétriques

Commande *streg*

### Modèle AFT

#### Weibull

Par défaut, le modèle de Weibull est exécuté sous paramétrisation PH. Pour une paramétrisation type AFT, ajouter l'option *time*.

```
webuse set "https://raw.githubusercontent.com/mthevenin/analyse_duree/master/bases/"
webuse "transplantation_m", clear
webuse set

stset stime, f(died)
streg year age surgery , dist(weibull) time nolog noshow
estat ic

Weibull AFT regression

No. of subjects =           103                      Number of obs     =      103
No. of failures =          75
Time at risk     =      31938
                                         LR chi2(3)      =      18.87
Log likelihood   =    -188.6278
                                         Prob > chi2     =     0.0003

-----
          _t |      Coef.    Std. Err.      z     P>|z| [95% Conf. Interval]
-----+
       year |    0.1620    0.1218     1.33    0.184    -0.0768    0.4008
        age |   -0.0615    0.0247    -2.49    0.013    -0.1100   -0.0130
      surgery |    1.9703    0.7794     2.53    0.011     0.4427    3.4980
      _cons |   -3.0220    8.7284    -0.35    0.729   -20.1294   14.0854
-----+
      /ln_p |   -0.5868    0.0927    -6.33    0.000    -0.7685   -0.4051
-----+
         p |    0.5561    0.0516
        1/p |    1.7983    0.1667
-----+
Akaike's information criterion and Bayesian information criterion

-----+
      Model |          N  ll(null)  ll(model)      df      AIC      BIC
-----+
       . |      103  -198.0632  -188.6278      5  387.2556  400.4292
-----+
Note: BIC uses N = number of observations. See [R] BIC note.
```

## Loglogistique

```
streg year age surgery , dist(loglog) nolog noshow  
estat ic
```

Loglogistic AFT regression

No. of subjects =	103	Number of obs =	103		
No. of failures =	75				
Time at risk =	31938	LR chi2(3) =	21.69		
Log likelihood =	-183.03937	Prob > chi2 =	0.0001		
<hr/>					
_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
year	0.2408	0.1172	2.05	0.040	0.0110 0.4705
age	-0.0427	0.0213	-2.00	0.045	-0.0845 -0.0010
surgery	2.2747	0.6913	3.29	0.001	0.9198 3.6296
_cons	-10.4034	8.3410	-1.25	0.212	-26.7515 5.9446
/lngamma	0.1805	0.0970	1.86	0.063	-0.0095 0.3706
gamma	1.1979	0.1161			0.9906 1.4486

Akaike's information criterion and Bayesian information criterion

Model	N	ll(null)	ll(model)	df	AIC	BIC
.	103	-193.8865	-183.0394	5	376.0787	389.2524

## Risques concurrents

*Non paramétrique: estimation des IC*

Installer les commandes *stcompet* et *stcomlist*

```
ssc install stcompet  
ssc install stcomlist
```

Le risque d'intérêt est *compet=1*, le risque concurrent est *compet=2*

```
stset stime, failure(compet==1)
stcomlist, competit1(2)
```

failure: competit == 1  
competing failures: competit == 2

Time	CIF	SE	[95% Conf. Int.]	
1	0.0097	0.0097	0.0009	0.0477
2	0.0388	0.0190	0.0127	0.0892
3	0.0583	0.0231	0.0239	0.1149
5	0.0777	0.0264	0.0363	0.1395
6	0.0874	0.0278	0.0429	0.1515
8	0.0971	0.0292	0.0497	0.1634
9	0.1068	0.0304	0.0566	0.1751
12	0.1166	0.0316	0.0638	0.1868
16	0.1362	0.0338	0.0785	0.2099
18	0.1461	0.0349	0.0860	0.2212
21	0.1657	0.0367	0.1014	0.2437
32	0.1756	0.0376	0.1093	0.2550
37	0.1856	0.0384	0.1173	0.2662
40	0.1957	0.0393	0.1254	0.2775
43	0.2058	0.0400	0.1337	0.2888
45	0.2158	0.0408	0.1420	0.2999
50	0.2259	0.0415	0.1503	0.3110
51	0.2360	0.0422	0.1588	0.3221
53	0.2461	0.0428	0.1673	0.3330
58	0.2562	0.0434	0.1759	0.3439
61	0.2662	0.0440	0.1845	0.3548
66	0.2763	0.0445	0.1932	0.3656
69	0.2864	0.0450	0.2020	0.3763
72	0.3066	0.0459	0.2197	0.3976
77	0.3167	0.0464	0.2286	0.4082
78	0.3267	0.0467	0.2376	0.4187
81	0.3368	0.0471	0.2466	0.4292
85	0.3469	0.0475	0.2556	0.4396
90	0.3570	0.0478	0.2648	0.4500
96	0.3671	0.0481	0.2739	0.4604
102	0.3771	0.0484	0.2831	0.4707
110	0.3874	0.0487	0.2925	0.4812
149	0.3980	0.0489	0.3021	0.4920
165	0.4085	0.0492	0.3118	0.5027
186	0.4193	0.0495	0.3217	0.5137
188	0.4301	0.0497	0.3316	0.5246
207	0.4408	0.0499	0.3417	0.5354
219	0.4516	0.0501	0.3517	0.5462
263	0.4624	0.0502	0.3618	0.5570
285	0.4846	0.0505	0.3826	0.5791
308	0.4957	0.0506	0.3931	0.5900
340	0.5068	0.0507	0.4037	0.6009
583	0.5221	0.0514	0.4171	0.6168
675	0.5401	0.0524	0.4322	0.6361
733	0.5580	0.0532	0.4477	0.6548
995	0.5808	0.0548	0.4659	0.6795

<b>1032</b>	<b>0.6036</b>	<b>0.0559</b>	<b>0.4851</b>	<b>0.7031</b>
<b>1386</b>	<b>0.6340</b>	<b>0.0583</b>	<b>0.5083</b>	<b>0.7357</b>

```
failure:  compet == 2  
competing failures:  compet == 1
```

Time	CIF	SE	[95% Conf. Int.]
3	0.0097	0.0097	0.0009 0.0477
6	0.0194	0.0136	0.0038 0.0619
16	0.0292	0.0166	0.0079 0.0761
17	0.0391	0.0191	0.0128 0.0897
28	0.0489	0.0213	0.0182 0.1029
30	0.0587	0.0232	0.0240 0.1157
35	0.0686	0.0250	0.0302 0.1286
36	0.0786	0.0267	0.0367 0.1411
39	0.0885	0.0282	0.0435 0.1534
40	0.0986	0.0296	0.0504 0.1658
68	0.1188	0.0322	0.0650 0.1901
80	0.1288	0.0334	0.0724 0.2020
100	0.1389	0.0345	0.0800 0.2138
153	0.1495	0.0356	0.0880 0.2261
334	0.1605	0.0368	0.0964 0.2392
342	0.1720	0.0381	0.1052 0.2526
852	0.1913	0.0417	0.1175 0.2787
979	0.2141	0.0460	0.1320 0.3094

### *Modèle cause-specific (Cox)*

## Attention: non relié aux IC

stcox year age surgery, nohr

Cox regression -- Breslow method for ties

No. of subjects = 103 Number of obs = 103  
 No. of failures = 56  
 Time at risk = 31938 LR chi2(3) = 15.88  
 Log likelihood = -214.46905 Prob > chi2 = 0.0012

<u>t</u>	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
year	-0.1033	0.0774	-1.33	0.182	-0.2550 0.0485
age	0.0385	0.0163	2.36	0.018	0.0065 0.0704
surgery	-1.1099	0.5290	-2.10	0.036	-2.1468 -0.0730

## Modèle de Fine-Gray

La commande **stcrreg** est installée avec les commandes de base. Relié directement aux IC, la définition du risque diffère du risque instantané usuel (risque de sous répartition).

```
stcrreg year age surgery, compete(compet=2) nohr

failure _d: compet == 1
analysis time _t: stime

Competing-risks regression
No. of obs = 103
No. of subjects = 103
Failure event : compet == 1
No. failed = 56
Competing event: compet == 2
No. competing = 19
No. censored = 28

Wald chi2(3) = 11.42
Log pseudolikelihood = -227.69531 Prob > chi2 = 0.0097

-----
| Robust
_t | Coef. Std. Err. z P>|z| [95% Conf. Interval]
-----+
year | -0.0724 0.0716 -1.01 0.312 -0.2128 0.0679
age | 0.0370 0.0177 2.09 0.037 0.0022 0.0718
surgery | -0.8688 0.4510 -1.93 0.054 -1.7528 0.0153
-----+
```

## Modèle logistique multinomial

Attention: non relié aux IC Pour la variable de durée on utilise la variable *mois*

```
expand mois
bysort id: gen t=_n
gen t2=t*t

gen e = competit
replace e=0 if t<mois

mlogit e t t2 year age surgery

Multinomial logistic regression
Number of obs = 1,127
LR chi2(10) = 86.25
Prob > chi2 = 0.0000
Pseudo R2 = 0.1356

Log likelihood = -275.00542

-----
e | RRR Std. Err. z P>|z| [95% Conf. Interval]
-----+
0 | (base outcome)
-----+
1 |
t | 0.8159 0.0338 -4.91 0.000 0.7522 0.8850
t2 | 1.0032 0.0009 3.53 0.000 1.0014 1.0049
year | 0.8795 0.0718 -1.57 0.116 0.7494 1.0321
-----+
```



# Python

Deux paquets d'analyse: principalement *lifelines* (km, cox, aft...) et *statsmodels* (estimation logit en temps discret, kaplan-Meier, Cox). Le package *statsmodels* est également ne mesure d'estimer des courbes de séjour de type Kaplan-Meier et des modèles à risque proportionnel de Cox. Le package *lifelines* couvre la quasi totalité des méthodes standards, à l'exception des les risques concurrents.

## *Chargement des librairies*

```
import numpy as np
import pandas as pd
import patsy as pt
import lifelines as lf
import matplotlib.pyplot as plt
```

## *Importation de la base*

```
trans = pd.read_csv("https://raw.githubusercontent.com/mthevenin/analyse_duree/master/bases/transplantation.csv")
trans.head(10)
trans.info()
```

## Package *lifelines*

<https://lifelines.readthedocs.io/en/latest/>

### Non Paramétrique: Kaplan Meier

#### *Estimateur KM et durée médiane*

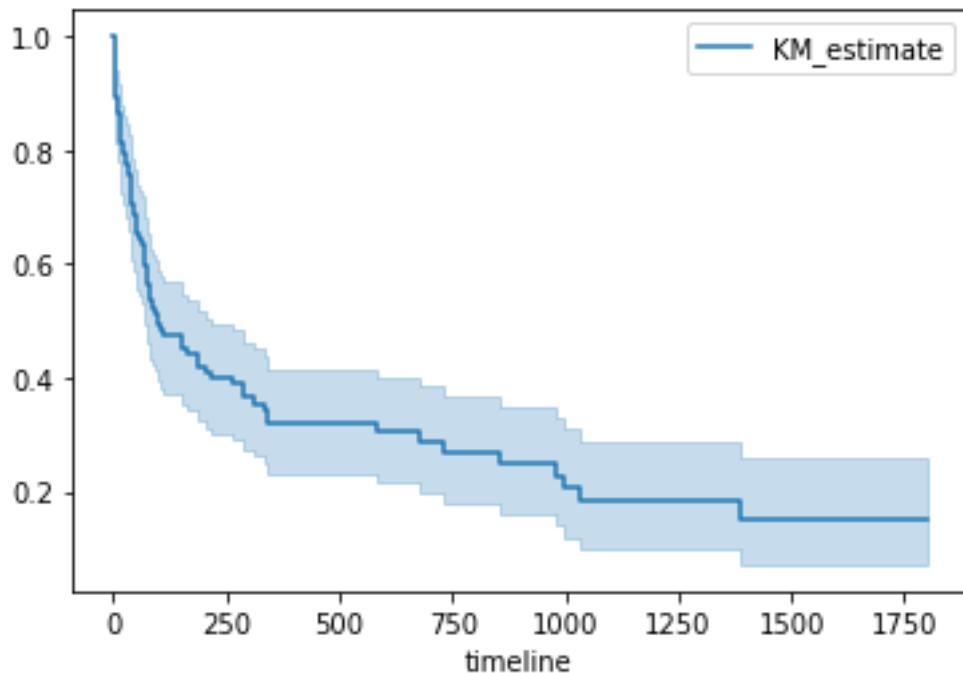
```
T = trans['stime']
E = trans['died']

from lifelines import KaplanMeierFitter
kmf = KaplanMeierFitter()
kmf.fit(T,E)
print(kmf.survival_function_)
a = "DUREE MEDIANE:"
b = kmf.median_survival_time_
print(a,b)
```

```
kmf.plot()
```

timeline	KM_estimate
0.0	1.000000
1.0	0.990291
2.0	0.961165
3.0	0.932039
5.0	0.912621
...	...
1400.0	0.151912
1407.0	0.151912
1571.0	0.151912
1586.0	0.151912
1799.0	0.151912

[89 rows x 1 columns]  
DUREE MEDIANE: 100.0

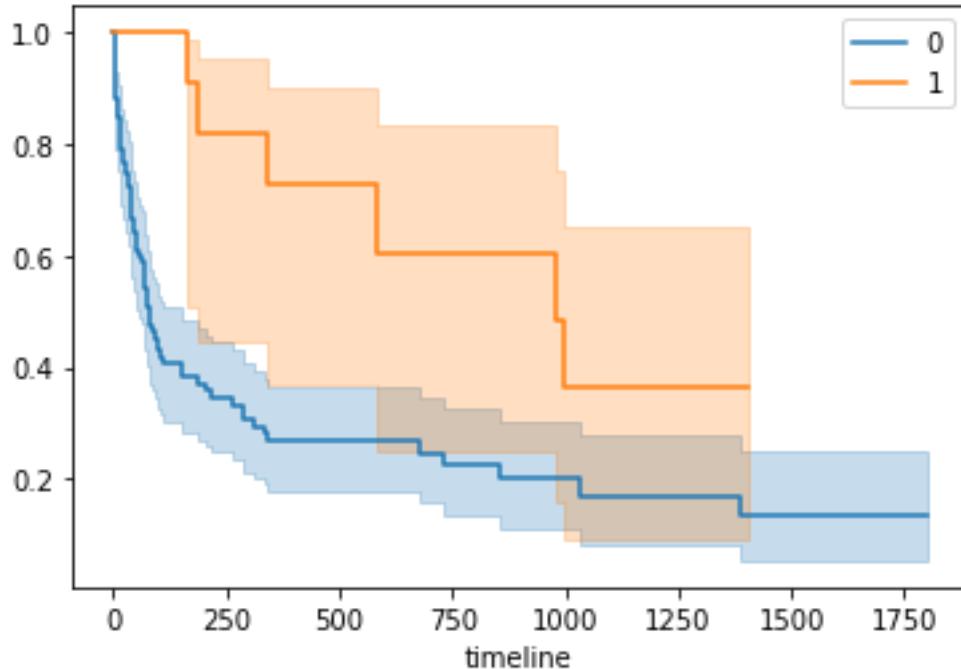


*Comparaison des fonctions de survie*

```

ax = plt.subplot(111)
kmf = KaplanMeierFitter()
for name, grouped_df in trans.groupby('surgery'):
    kmf.fit(grouped_df['stime'], grouped_df['died'], label=name)
    kmf.plot(ax=ax)

```



```

from lifelines.statistics import multivariate_logrank_test
results = multivariate_logrank_test(trans['stime'], trans['surgery'],
                                     trans['died'])
results.print_summary()

<lifelines.StatisticalResult: multivariate_logrank_test>
    t_0 = -1
    null_distribution = chi squared
    degrees_of_freedom = 1
        test_name = multivariate_logrank_test

    ---
    test_statistic      p   -log2(p)
        6.59  0.01      6.61

```

## Semi paramétrique: Cox

### *Estimation*

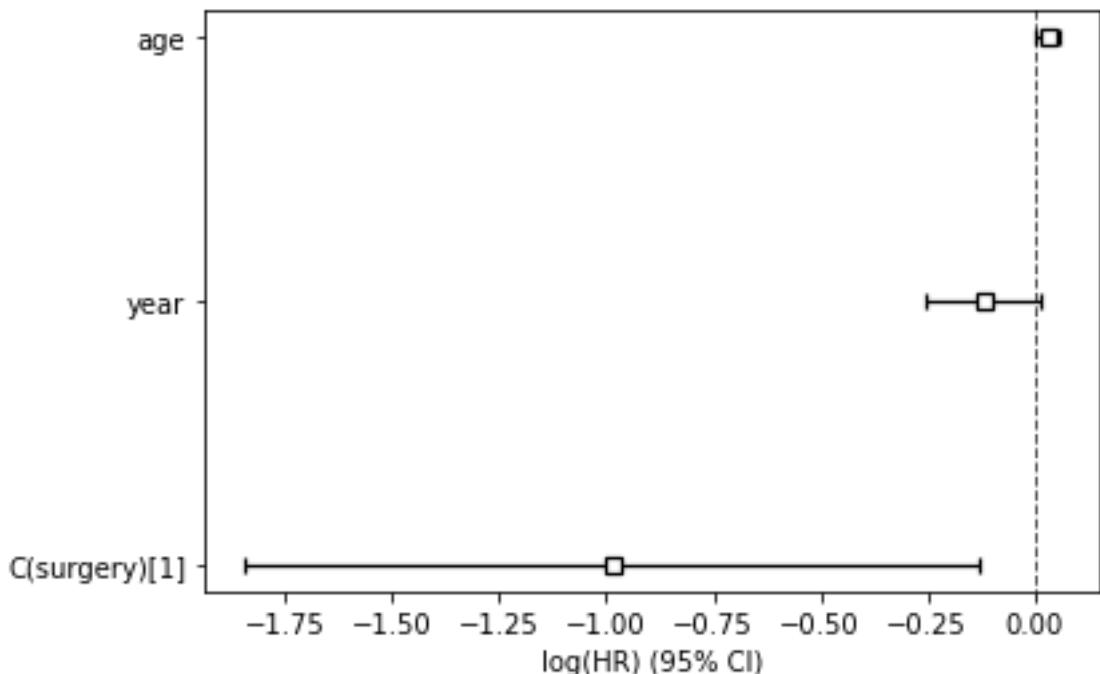
```
model = 'year + age + C(surgery) -1'
X = pt.dmatrix(model, trans, return_type='dataframe')
design_info = X.design_info
YX = X.join(trans[['stime','died']])
YX.drop(['C(surgery)[0]'], axis=1, inplace=True)
YX.head()

from lifelines import CoxPHFitter
cph = CoxPHFitter()
cph.fit(YX, duration_col='stime', event_col='died')
cph.print_summary()
cph.plot()

<lifelines.CoxPHFitter: fitted with 103 total observations, 28 right-censored observations>
    duration col = 'stime'
    event col = 'died'
    baseline estimation = breslow
    number of observations = 103
number of events observed = 75
partial log-likelihood = -289.31
    time fit was run = 2021-04-21 13:24:52 UTC

    ---
            coef  exp(coef)   se(coef)
C(surgery)[1] -0.99      0.37      0.44
year          -0.12      0.89      0.07
age           0.03      1.03      0.01

            z      p    -log2(p)
C(surgery)[1] -2.26  0.02      5.40
year          -1.78  0.08      3.72
age           2.19  0.03      5.12
    --
Concordance = 0.65
Partial AIC = 584.61
log-likelihood ratio test = 17.63 on 3 df
-log2(p) of ll-ratio test = 10.90
```



### Tests hypothèse PH

Test PH: Résidus de Schoenfeld Méthode 1

```
cph.check_assumptions(YX,p_value_threshold=0.05)
```

The ``p\_value\_threshold`` is set at 0.05. Even under the null hypothesis of no violations, some covariates will be below the threshold by chance. This is compounded when there are many covariates.

Similarly, when there are lots of observations, even minor deviances from the proportional hazard assumption will be flagged.

With that in mind, it's best to use a combination of statistical tests and visual tests to determine the most serious violations. Produce visual plots using ``check\_assumptions(..., show\_plots=True)`` and looking for non-constant lines. See link [A] below for a full example.

```
<lifelines.StatisticalResult: proportional_hazard_test>
  null_distribution = chi squared
  degrees_of_freedom = 1
    test_name = proportional_hazard_test
```

		test_statistic	p	-log2(p)
C(surgery)[1]	km	4.01	0.05	4.47
	rank	3.74	0.05	4.23
age	km	1.18	0.28	1.86
	rank	1.06	0.30	1.72
year	km	2.07	0.15	2.73

rank	2.08	0.15	2.75
------	------	------	------

1. Variable 'C(surgery)[1]' failed the non-proportional test: p-value is 0.0452.

*Test PH: Résidus de Schoenfeld Méthode 2*

```
from lifelines.statistics import proportional_hazard_test
zph = proportional_hazard_test(cph, YX, time_transform='all')
zph.print_summary()
```

```
<lifelines.StatisticalResult: proportional_hazard_test>
  null_distribution = chi squared
  degrees_of_freedom = 1
    test_name = proportional_hazard_test
```

```
---
              test_statistic      p   -log2(p)
C(surgery)[1] identity       5.54 0.02     5.75
               km            4.01 0.05     4.47
               log           3.69 0.05     4.19
               rank          3.74 0.05     4.23
age          identity       1.61 0.20     2.29
               km            1.18 0.28     1.86
               log           0.61 0.44     1.20
               rank          1.06 0.30     1.72
year         identity       0.80 0.37     1.43
               km            2.07 0.15     2.73
               log           1.34 0.25     2.02
               rank          2.08 0.15     2.75
```

*Test PH: interaction avec une fonction de la durée*

```
from lifelines.utils import to_episodic_format
from lifelines import CoxTimeVaryingFitter
```

*Transformation de la base YX*

```
long = to_episodic_format(YX, duration_col='stime', event_col='died')
```

*Création de la variable d'interaction*

```
long['surgery_t'] = long['C(surgery)[1]'] * long['stop']
```

## *Estimation*

```
ctv = CoxTimeVaryingFitter()
ctv.fit(long,
        id_col='id',
        event_col='died',
        start_col='start',
        stop_col='stop',)
ctv.print_summary(4)

<lifelines.CoxTimeVaryingFitter: fitted with 31938 periods, 103 subjects, 75 events>
    event col = 'died'
number of subjects = 103
number of periods = 31938
number of events = 75
partial log-likelihood = -287.3290
time fit was run = 2021-04-21 13:32:40 UTC

---
            coef      exp(coef)      se(coef)
C(surgery)[1] -1.7547      0.1730      0.6743
age             0.0289      1.0293      0.0134
year            -0.1231      0.8842      0.0668
surgery_t       0.0022      1.0022      0.0011

            z          p      -log2(p)
C(surgery)[1] -2.6022  0.0093      6.7542
age              2.1479  0.0317      4.9785
year             -1.8415 0.0656      3.9312
surgery_t       2.0239  0.0430      4.5402
---
Partial AIC = 582.6581
log-likelihood ratio test = 21.5846 on 4 df
-log2(p) of ll-ratio test = 12.0103
```

## Modèle à temps discret (régression logistique)

### *Ajustement continu*

Modèle logistique estimé avec le paquet `statsmodel`. La fonction `to_episodic_format` de `Lifelines` permet de mettre en forme la base.  
Pour la durée, on utilisera ici la variable **mois** (regroupement de stime par intervalle de 30 jours).

```
#type R formule => ce qu'on utilisera
import statsmodels.formula.api as smf
#type python#
import statsmodels.api as sm
```

*Transformation de la base en format long*

```
trans = pd.read_csv("https://raw.githubusercontent.com/mthevenin/analyse_duree/master/bases/transplantation.csv")

td.drop(['id'], axis=1, inplace=True)
td['dur'] = td['mois']
td = to_episodic_format(td, duration_col='mois', event_col='died')
```

*Recherche de la fonction d'ajustement*

```
td['t2'] = td['stop']**2
td['t3'] = td['stop']**3

fit1 = smf.glm(formula= "died ~ stop", data=td, family=sm.families.Binomial()).fit()

fit2 = smf.glm(formula= "died ~ stop + t2", data=td, family=sm.families.Binomial()).fit()

fit3 = smf.glm(formula= "died ~ stop + t2 + t3", data=td, family=sm.families.Binomial()).fit()
```

*Comparaison des AIC*

```
print("AIC pour ajustement t1")
print(fit1.aic)
print("AIC pour ajustement durée t1 + t2")
print(fit2.aic)
print("AIC pour ajustement durée t1 + t2 + t3")
print(fit3.aic)
```

AIC pour ajustement t1  
512.1039235968562

AIC pour ajustement durée t1 + t2  
508.1014573009212

AIC pour ajustement durée t1 + t2 + t3  
506.1882809518765

## Estimation du modèle

```
tdfit = smf.glm(formula= "died ~ stop + t2 + t3 + year + age + surgery", da  
ta=td, family=sm.families.Binomial()).fit()  
tdfit.summary()  
  
<class 'statsmodels.iolib.summary.Summary'>  
"""  
Generalized Linear Model Regression Results  
=====  
Dep. Variable: died No. Observations: 1164  
Model: GLM Df Residuals: 1157  
Model Family: Binomial Df Model: 6  
Link Function: logit Scale: 1.0000  
Method: IRLS Log-Likelihood: -240.20  
Date: Wed, 21 Apr 2021 Deviance: 480.39  
Time: 15:44:21 Pearson chi2: 1.30e+03  
No. Iterations: 7  
Covariance Type: nonrobust  
=====  
coef std err z P>|z| [0.025 0.975]  
-----  
Intercept 6.3097 5.201 1.213 0.225 -3.884 16.503  
stop -0.2807 0.077 -3.635 0.000 -0.432 -0.129  
t2 0.0096 0.005 2.083 0.037 0.001 0.019  
t3 -0.0001 6.97e-05 -1.493 0.135 -0.000 3.26e-05  
year -0.1263 0.072 -1.747 0.081 -0.268 0.015  
age 0.0337 0.014 2.330 0.020 0.005 0.062  
surgery -1.0050 0.447 -2.250 0.024 -1.880 -0.130  
=====  
"""
```

## Ajustement discret

Création des intervalles pour l'exemple (quantile de la durée en mois)

```
td['ct4'] = pd.qcut(td['stop'],[0, .25, .5, .75, 1.])  
td['ct4'].value_counts(normalize=True)*100  
td.ct4 = pd.Categorical(td.ct4)  
td['ct4'] = td.ct4.cat.codes  
  
(0.999, 4.0] 27.233677  
(11.0, 23.0] 24.484536  
(4.0, 11.0] 24.398625  
(23.0, 61.0] 23.883162  
Name: ct4, dtype: float64
```

Pour chaque individu, on conserve une seule observation par intervalle.

```
td2 = td  
td2['t'] = td2['ct4']  
td2 = td2.sort_values(['id', 'stop'])
```

```
td2 = td2.groupby(['id','ct4']).last()
td2.head(20)
```

### *Estimation*

```
td2fit = smf.glm(formula= "died ~ C(t) + year + age + surgery", dat
a=td2, family=sm.families.Binomial()).fit()
td2fit.summary()
```

```
<class 'statsmodels.iolib.summary.Summary'>
"""
Generalized Linear Model Regression Results
=====
Dep. Variable: died No. Observations: 200
Model: GLM Df Residuals: 200
Model Family: Binomial Df Model: -1
Link Function: logit Scale: 1.0000
Method: IRLS Log-Likelihood: -112.26
Date: Wed, 21 Apr 2021 Deviance: 224.52
Time: 15:54:03 Pearson chi2: 229.
No. Iterations: 5
Covariance Type: nonrobust
=====
      coef    std err        z   P>|z|    [0.025]    [0.975]
-----
Intercept  11.8018    6.617     1.784    0.074    -1.167    24.770
C(t)[T.1]  -0.9078    0.408    -2.227    0.026    -1.707   -0.109
C(t)[T.2]  -1.8451    0.587    -3.141    0.002    -2.996   -0.694
C(t)[T.3]  -0.3224    0.578    -0.557    0.577    -1.456    0.811
year       -0.1947    0.093    -2.104    0.035    -0.376   -0.013
age        0.0468    0.018     2.543    0.011     0.011    0.083
surgery    -1.1025    0.503    -2.192    0.028    -2.088   -0.117
"""
"""

```

### Modèle paramétrique de type AFT

```
from lifelines import WeibullAFTFitter, LogLogisticAFTFitter
```

### Weibull

```
aftw = WeibullAFTFitter()
aftw.fit(YX, duration_col='stime', event_col='died')
aftw.print_summary()

<lifelines.WeibullAFTFitter: fitted with 103 total observations, 28 r
ight-censored observations>
    duration col = 'stime'
    event col = 'died'
```

```

number of observations = 103
number of events observed = 75
    log-likelihood = -488.17
    time fit was run = 2021-04-21 13:55:14 UTC

---
            coef  exp(coef)   se(coef)
lambda_ C(surgery)[1]  1.97      7.17     0.78
          year       0.16      1.18     0.12
          age      -0.06      0.94     0.02
          _intercept -3.02      0.05     8.73
rho_    _intercept   -0.59      0.56     0.09

            z      p   -log2(p)
lambda_ C(surgery)[1] 2.53  0.01     6.45
          year      1.33  0.18     2.44
          age      -2.49  0.01     6.28
          _intercept -0.35  0.73     0.46
rho_    _intercept   -6.33 <0.005    31.93
---
Concordance = 0.65
AIC = 986.34
log-likelihood ratio test = 18.87 on 3 df
-log2(p) of ll-ratio test = 11.75

```

### *Loglogistique*

```

aftl = LogLogisticAFTFitter()
aftl.fit(YX, duration_col='stime', event_col='died')
aftl.print_summary()

<lifelines.LogLogisticAFTFitter: fitted with 103 total observations,
28 right-censored observations>
    duration col = 'stime'
        event col = 'died'
    number of observations = 103
number of events observed = 75
    log-likelihood = -482.58
    time fit was run = 2021-04-21 13:55:58 UTC

---
            coef  exp(coef)   se(coef)
alpha_ C(surgery)[1]  2.27      9.72     0.69
          year       0.24      1.27     0.12

```

```

age              -0.04      0.96      0.02
intercept       -10.41     0.00      8.34
beta_ _intercept -0.18      0.83      0.10
                           z      p   -log2(p)
alpha_ C(surgery)[1] 3.29 <0.005    9.96
year             2.05  0.04      4.65
age              -2.01  0.04      4.49
_intercept      -1.25  0.21      2.24
beta_ _intercept -1.86  0.06      4.00
---
Concordance = 0.66
AIC = 975.16
log-likelihood ratio test = 21.69 on 3 df
-log2(p) of ll-ratio test = 13.69

```

## Package statsmodels

<https://www.statsmodels.org/dev/duration.html>

Le package permet d'estimer des fonction de séjour de type Kaplan-Meier et des modèles de Cox.

### Fonctions de séjour Kaplan-Meier

```

km = sm.SurvfuncRight(trans["stime"], trans["died"])
km.summary()

      Surv prob  Surv prob SE  num at risk  num events
Time
1      0.990291  0.009661      103      1.0
2      0.961165  0.019037      102      3.0
3      0.932039  0.024799      99      3.0
5      0.912621  0.027825      96      2.0
6      0.893204  0.030432      94      2.0
...
852    0.250655  0.048731      14      1.0
979    0.227868  0.049341      11      1.0
995    0.205081  0.049390      10      1.0
1032   0.182295  0.048877      9      1.0
1386   0.151912  0.049277      6      1.0

```

Les test du log-rank sont disponibles avec la fonction *survdiff* (nom idem R). Au niveau graphique, la programmation semble un peu lourde et mériterait d'être simplifiée (donc non traitée).

### Comparaison de S(t) à partir des tests du log-rank

Résultat: (statistique de test, p-value)

*Test non pondéré*

```
sm.duration.survdiff(trans.stime, trans.died, trans.surgery)  
(6.590012323234387, 0.010255246157888975)
```

*Breslow*

```
sm.duration.survdiff(trans.stime, trans.died, trans.surgery, weight_type='gb')  
(8.989753779902495, 0.0027149757927903417)
```

*Tarone-Ware*

```
sm.duration.survdiff(trans.stime, trans.died, trans.surgery, weight_type='tw')  
(8.462352726451392, 0.0036257256194570653)
```

Modèle de Cox

```
mod = smf.phreg("stime ~ year + age + surgery ",trans, status='died'  
, ties="efron")  
rslt = mod.fit()  
print(rslt.summary())
```

Results: PHReg

```
=====
```

	PH Reg	Sample size:	103			
Model:		Num. events:	75			
Dependent variable:	stime					
Ties:	Efron					
	log HR	log HR SE	HR	t	P> t	[0.025 0.975]
year	-0.1196	0.0673	0.8872	-1.7765	0.0757	0.7775 1.0124
age	0.0296	0.0135	1.0300	2.1872	0.0287	1.0031 1.0577
surgery	-0.9873	0.4363	0.3726	-2.2632	0.0236	0.1584 0.8761

```
=====
```

Confidence intervals are for the hazard ratios